This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

S2Gaussian: Sparse-View Super-Resolution 3D Gaussian Splatting

Yecong Wan[1] Mingwen Shao[1*] Yuanshuo Cheng[1] Wangmeng Zuo[2]
[1]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China)
[2]Faculty of Computing, Harbin Institute of Technology
smw278@126.com, {yecongwan, cys1294414023, cswmzuo}@gmail.com

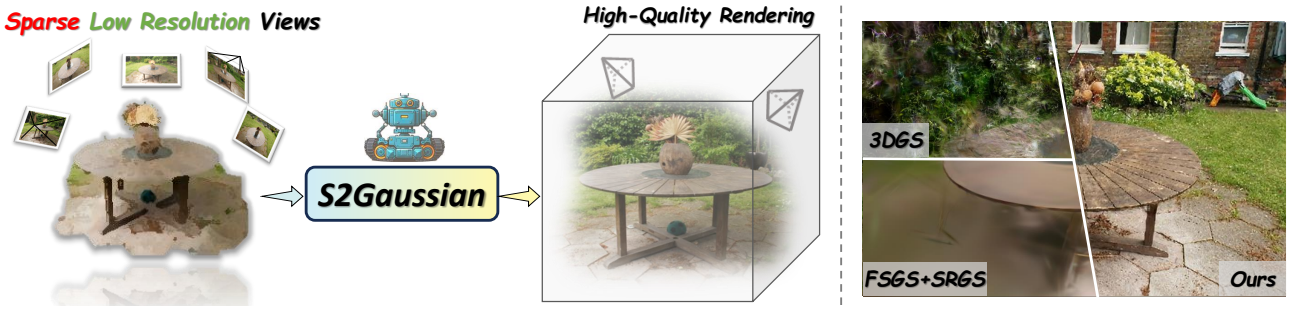

Figure 1. We present **S2Gaussian**, a novel framework capable of reconstructing high-quality 3D scenes for immersive rendering with only sparse and low-resolution input views. S2Gaussian demonstrates superior performance and yields high-fidelity and high-resolution scene reconstruction with sharp geometric and detailed textures, thus enjoying better functionality and practicality in realistic applications.

Abstract

In this paper, we aim ambitiously for a realistic yet challenging problem, namely, how to reconstruct high-quality 3D scenes from sparse low-resolution views that simultaneously suffer from deficient perspectives and clarity. Whereas existing methods only deal with either sparse views or low-resolution observations, they fail to handle such hybrid and complicated scenarios. To this end, we propose a novel Sparse-view Super-resolution 3D Gaussian Splatting framework, dubbed S2Gaussian, that can reconstruct structure-accurate and detail-faithful 3D scenes with only sparse and low-resolution views. The S2Gaussian operates in a two-stage fashion. In the first stage, we initially optimize a low-resolution Gaussian representation with depth regularization and densify it to initialize the high-resolution Gaussians through a tailored Gaussian Shuffle Split operation. In the second stage, we refine the high-resolution Gaussians with the super-resolved images generated from both original sparse views and pseudo-views rendered by the low-resolution Gaussians. In which a customized blur-free inconsistency modeling scheme and a 3D robust optimization strategy are elaborately designed to mitigate multi-view inconsistency and eliminate erroneous updates caused by imperfect supervision. Extensive experiments demonstrate superior results and in particular establishing new state-of-the-art performances with more consistent geometry and finer details. Project Page https://jeasco.github.io/S2Gaussian/.

1. Introduction

The rapid advancement of virtual reality and metaverse technologies has significantly heightened the demand for realistic 3D scene reconstruction, which holds vast potential across various application domains, including medicine, education, entertainment, etc. Among the recent notable developments, Neural Radiance Fields (NeRF) [25] have demonstrated remarkable capability in synthesizing photo-realistic images; however, they still face considerable rendering and training costs, despite ongoing improvements [4, 10, 11, 26]. More recently, 3D Gaussian Splatting (3DGS) [18] has emerged as a prominent approach, characterized by its high quality, rapid reconstruction speed, and support for real-time rendering. Subsequent research efforts [5, 6, 23, 32, 39, 45] have focused on extending the applicability of 3DGS across various scenarios.

Unfortunately, these methods typically rely heavily on well-captured dense and high-resolution images for impressive novel view synthesis, which is cumbersome and some-

[1*] Corresponding authors.

times impractical in real-world applications. To address this challenge, a variety of approaches have emerged that focus on reconstructing 3D scenes from sparse [7, 8, 22, 27, 35, 40, 42, 50] and low-resolution views [9, 12, 21, 31, 43, 44], yielding worth-celebrating successes. However, these two obstacles, i.e., sparsity and low resolution, have long been treated as isolated problems and tackled separately, whereas input views in practical applications often suffer from both deficient perspectives and clarity. Particularly in contexts such as robotics and internet-collected imagery, where the available views are usually sparse and compounded by low resolutions due to environmental constraints and hardware transfer limitations. This interplay of sparsity and low-resolution poses significant challenges for existing reconstruction frameworks. Furthermore, merely incorporating existing methods remains subject to inherent incompatibility, as super-resolution necessitates dense supervision to recover fine details while sparse-view regularization not only fails to introduce details but tends to lead smoothing.

To tackle the aforementioned problems, in this paper, we propose a novel 3D reconstruction framework S2Gaussian, capable of reconstructing geometrically precise and richly detailed 3D scenes with only sparse and low-resolution views. Specifically, S2Gaussian consists of two main phases: the high-resolution Gaussians (HR GS) initialization stage and the HR GS optimization stage. In the HR GS initialization stage, we first optimize a low-resolution Gaussian representation using the sparse views as well as the estimated depth information. The optimized low-resolution Gaussians are then utilized to initialize the high-resolution Gaussians through a tailored Gaussian Shuffle Split operation which can provide more compact Gaussian primitives to facilitate fine-grained details reconstruction in high-resolution scenes. In the HR GS optimization stage, we use super-resolved images generated by pre-trained super-resolution model to refine the high-resolution Gaussians, leveraging both the original sparse views and the pseudo-views rendered from the low-resolution Gaussians. Notably, a customized blur-free inconsistency modeling scheme and a 3D robust optimization strategy are introduced to mitigate the impacts of multi-view inconsistency and imperfect supervision, ultimately leading to more detailed and higher-quality scene reconstruction. Comprehensive experiments substantiate the superiority of our S2Gaussian beyond alternatives on a variety of benchmarks.

In conclusion, the main contributions are as follows:

- We propose an innovative two-stage framework, termed S2Gaussian, designed to reconstruct 3D scenes that are both structurally accurate and rich in detail using only sparse and low-resolution input views.
- A dedicated Gaussian Shuffle Split operation is designed to initialize more compact Gaussian primitives for representing fine-grained details and textures in high-resolution scenes.
- We introduce a tailored blur-free inconsistency modeling scheme alongside a 3D robust optimization strategy to address multi-view inconsistencies and rectify erroneous updates arising from imperfect supervision.
- Our S2Gaussian reconstructs more fine-grained and high-quality Gaussian representations significantly outperforming existing methods across multiple benchmarks.

2. Related Work

2.1. Novel View Synthesis using Radiance Fields

Techniques for novel view synthesis typically involve learning a 3D representation using a limited number of input views and generating images from arbitrary novel perspectives. Recent advancements in Neural Radiance Fields (NeRF) [25] has demonstrated encouraging progress in this field, which learns an implicit neural scene representation for novel view synthesis via coordinate-based neural networks and volume rendering function. Subsequent research has focused extensively on enhancing NeRF's rendering quality [1–3, 33], improving efficiency [4, 10, 11, 26], advancing scene understanding [19, 48, 49], and facilitating 3D content generation [15, 28, 30]. Unfortunately, while NeRF significantly improves the quality of novel view rendering, its expensive training time as well as slow rendering speed hinder broader practical applications. Recently, Kerbl et al. [18] introduced a groundbreaking approach with their 3D Gaussian Splatting (3DGS) method, boosting rendering efficiency by utilizing explicit Gaussian representations combined with differentiable rasterization techniques. Building upon 3D Gaussian Splatting representation, tremendous following efforts [6, 23, 32, 39, 45] concentrate extending 3DGS under various scenarios. Nevertheless, these methods typically require dense and high-quality input views for impressive novel view synthesis, and they will tend to overfit the training data and compromise representations when provided with only sparse or degraded views, leading to a substantial decline in performance for the intended scene reconstruction.

2.2. Sparse Novel View Synthesis

To enhance the availability in practical usage, numerous algorithms have been developed to surmount the reliance on dense perspectives. For instance, Depth-NeRF [8] and Reg-NeRF [27] advocate utilizing depth constraints to improve the accuracy of the sparse scene representation. FreeNeRF [42], on the other hand, employs a frequency regularization strategy to manage the frequency range and penalize density fields near the camera. Additionally, SparseNeRF [35] leverages depth priors from inaccurate observations, implementing local depth ranking and spatial continuity constraints to bolster neural radiance fields. More recently, a variety of 3DGS-based approaches have emerged

[7, 22, 40, 50], paving the way for new blueprints in the field. These methods typically focus on various strategies that utilize depth prior to constrain the optimization of Gaussian primitives, such as SfM depth alignment [7], Pearson correlation depth distribution loss [50], global-local depth normalization [22], and patch-based depth correlation [40]. Unlike these approaches, we spotlight more arduous sparse reconstruction under low-resolution conditions, where available information is more scarce.

2.3. Super-Resolution Novel View Synthesis

Super-resolution novel view synthesis aims to reconstruct high-resolution 3D scenes utilizing only low-resolution multi-view inputs. Unlike anti-aliasing methods [1, 2, 45] which are incapable of rendering intricate details of the novel views and are limited by the input image's information level, this field is more focused on fine-grained and detailed high-resolution rendering of potential high-resolution scenes. As a pioneer in the field, NeRF-SR [34] optimizes high-resolution NeRF through a super-sampling strategy, ensuring that the values of low-resolution pixels match the mean value of high-resolution sub-pixels. Follow-up works are designed to utilize either high-resolution reference images [16] or pre-trained 2D models [12, 21, 43] to generate multi-view consistent details. Concurrently, SRGS [9] and GaussianSR [44] are proposed to utilize 2D super-resolution models for texture injection or diffusion prior exploration, respectively. SuperGaussian [31] and SM [20] further prove two profile frameworks to repurpose video upsampling models to achieve 3D super-resolution.

Different from the aforementioned methods, in this work, we ambitiously tackle the challenge of joint sparse-view and super-resolution novel view synthesis by proposing a novel two-stage framework, which cleverly solves the problem of perspectives and clarity deficient to reconstruct geometry-accurate and fine-detailed 3D scenes, making it more desirable and feasible in real-world applications.

3. Methodology

3.1. Preliminaries

3D Gaussian Splatting (3DGS) as an emerged powerful technique for novel view synthesis, models a 3D scene using a collection of G colored Gaussians:

$$g_i(x) = \exp\left(-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)\right), \tag{1}$$

where $1 \leq i \leq G$, $\mu_i \in \mathbb{R}^3$ represents the mean or center of the Gaussian, and $\Sigma_i \in \mathbb{R}^{3\times 3}$ denotes Gaussian's covariance matrix, specifying its shape and size, which is described by scaling vector $s \in \mathbb{R}^3_+$ and rotation vector $q \in \mathbb{R}^4$ for implementation. Additionally, each Gaussian is characterized by an opacity $\sigma_i \in \mathbb{R}_+$ and a view-dependent color $c_i(v) \in \mathbb{R}^{k_c}$ (where k_c is changed) for rasterization.

To faithfully reconstruct the 3D scene, 3DGS also employs densification operation, i.e., the split, clone, and prune operation, to adaptively regulate the number and density of Gaussian primitives. For simplicity, we deploy the same adaptive density control as the original 3DGS and will not elaborate further in the subsequent presentation.

3.2. S2Gaussian Overview

The schematic illustration of the proposed S2Gaussian is depicted in Fig. 2. S2Gaussian is primarily composed of two phases i.e., the HR GS initialization stage and the HR GS optimization stage. In the HR GS initialization stage, we first optimize a low-resolution Gaussian representation using the sparse views as well as the predicted depth information, which is further utilized to initialize the high-resolution Gaussian representation by densifying and detailing the Gaussian primitives through our tailored Gaussian Shuffle Split operation. In the HR GS optimization stage, we further optimize the high-resolution Gaussians with the super-resolved images generated by pre-trained super-resolution model from both original sparse views and the pseudo-views rendered by LR Gaussians. In particular, a dedicated blur-free inconsistency modeling scheme and a 3D robust optimization strategy are advanced to eliminate the effects of multi-view inconsistency and imperfect supervision to reconstruct more detailed and high-quality 3D scenes.

3.3. Stage 1: HR GS Initialization

Directly reconstructing a high-resolution 3D scene from sparse low-resolution viewpoints is extremely challenging, as it is not only prone to overfitting specific viewpoints but also lacks high-resolution detail constraints. To alleviate such a reconstruction dilemma, we advocate initially reconstructing a low-resolution Gaussian representation as an auxiliary to facilitate the optimization of high-resolution Gaussians. On the one hand, the low-resolution Gaussian can be used to initialize high-resolution Gaussians to provide reliable base Gaussian ellipsoids and 3D scene structures. On the other hand, it can provide pseudo-views as additional supervision to avoid overfitting on specific viewpoints as well as provide more potential details and textures for high-resolution scenes.

Concretely, we first obtain the monocular depth maps for the sparse low-resolution training views by utilizing the handy pre-trained depth estimation model [29]. Then we utilize the RGB images as well as the depth maps to co-optimize a low-resolution Gaussian representation with mature depth regularization technologies [22, 50], where Pearson correlation loss [50] is employed in this paper as a baseline. Thereafter, a tailored Gaussian Shuffle Split regime is proposed to densify and detail the sparse Gaussian primitives so as to initialize HR Gaussians for better high-resolution fine-grained details representation.

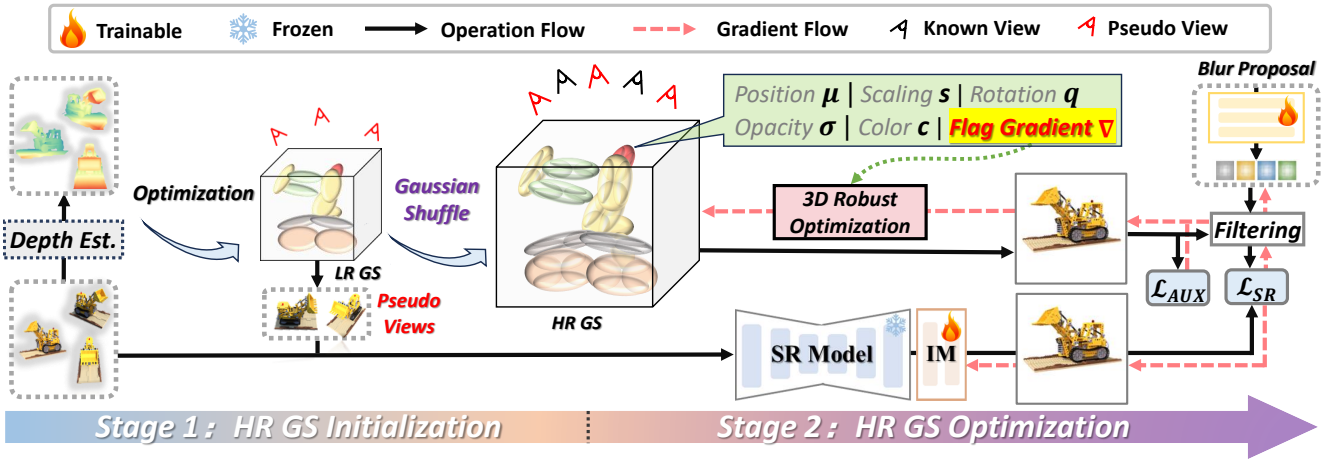

Figure 2. **Overview of S2Gaussian**. The S2Gaussian initially optimizes an LR GS and densifies it to initialize the HR GS through a tailored Gaussian Shuffle Split operation. Then, the original sparse views along with the pseudo views rendered by the LR GS are super-resolved together to refine the high-resolution texture with 3D robust optimization. In which an inconsistency modeling module (IM) and a blur proposal module are incorporated to mitigate inconsistency and blurriness, aiming to create 3D scenes with high-fidelity texture details.

Gaussian Shuffle Split. A precise and detailed initialization is crucial for reconstructing scene details [18]. Although the structure and layout of the entire 3D scene are basically guaranteed after the optimization of the low-resolution views, the initialized sparse and coarse Gaussian ellipsoids struggle to simulate high-resolution details which require denser Gaussians for finer representation [41]. Even equipped with adaptive density control [18], it falls short in our settings as there are neither dense perspectives nor high-quality details for supervision. To address this issue, we propose a training-free local Gaussian densification strategy *Gaussian Shuffle Split* which can provide more mobilizable Gaussian primitives by replacing the original large Gaussian with six small Gaussians to facilitate more comprehensive simulation of fine-grained details and textures in high-resolution scenes.

As illustrated in Fig. 3, given a Gaussian primitive with attributes $\{\mu, s, q, \sigma, c\}$, we first generate six duplicates, each center shifted along one of the six directions aligned with the original Gaussian's three principal axes, i.e., the positive and negative directions of each axis. The shift for each axis is set to α (0.5 by default) times the corresponding axis's scaling $s = [s_1, s_2, s_3]$ value. Thus the new position of the six sub-Gaussians can be obtained:

$$\begin{aligned}
&\mu_{sub1} = \mu + R \otimes [\alpha s_1, 0, 0]^T, \mu_{sub2} = \mu + R \otimes [-\alpha s_1, 0, 0]^T, \\
&\mu_{sub3} = \mu + R \otimes [0, \alpha s_2, 0]^T, \mu_{sub4} = \mu + R \otimes [0, -\alpha s_2, 0]^T, \\
&\mu_{sub5} = \mu + R \otimes [0, 0, \alpha s_3]^T, \mu_{sub6} = \mu + R \otimes [0, 0, -\alpha s_3]^T,
\end{aligned} \tag{2}$$

where R represents the rotation matrix calculated from q to align with the original rotation of the Gaussian. Additionally, for each shifted sub-Gaussian, we scale their s value corresponding to the offset axis to $\frac{1}{4}$ of its original value, i.e., reducing the scale along that axis to shrink the size. The remaining two s values are scaled by an another factor of λ relative to their original values:

$$\begin{aligned}
s_{sub1} = s_{sub2} &= [s_1/4, s_2/\lambda, s_3/\lambda], \\
s_{sub3} = s_{sub4} &= [s_1/\lambda, s_2/4, s_3/\lambda], \\
s_{sub5} = s_{sub6} &= [s_1/\lambda, s_2/\lambda, s_3/4],
\end{aligned} \tag{3}$$

where λ is experimentally set to 1.9 which we found this ratio ensures that the combined representation of the sub-Gaussians better approximate the original Gaussian, preserving the integrity of the 3D scene. All other attributes, i.e., rotation q, opacity α, and color c remain identical to the original Gaussian ellipsoid. The combined six sub-Gaussians are utilized to substitute the original large Gaussians for densified 3D representation.

We deployed Gaussian Shuffle Split only for Gaussian primitives with opacity greater than 0.5 since these are more likely to be responsible for representing object surfaces and key structures which require a more detailed representation in high-resolution scenes. The opacity of all primitives is set close to zero after the Gaussian Shuffle Split to facilitate the automatic elimination of redundant Gaussians with low opacity.

3.4. Stage 2: HR GS Optimization

With the initialized high-resolution Gaussians, in this stage, we aim to continue optimizing it with the super-resolved images generated from both original sparse views and the pseudo-views rendered by the LR Gaussians. We synthesize pseudo views by interpolating virtual views between two known views as in [50]. Considering the prohibitively expensive data costs associated with 3D super-resolution models and their computational complexity, we advocate for utilizing off-the-shelf pre-trained 2D super-resolution models to enhance low-resolution views. Unfortunately, directly

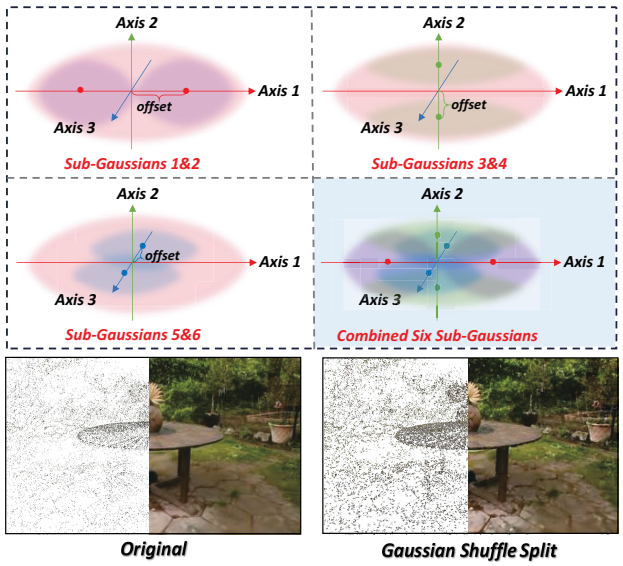

Figure 3. Illustration of Gaussian Shuffle Split that utilizes a combination of six compact Gaussians to replace the original large Gaussian primitive, which can densify the 3D scene virtually without damaging the original 3D representation.

leveraging 2D super-resolution model cannot ensure multi-view consistency, and the low-resolution pseudo-views inevitably exhibit artifacts since there are areas that have not been optimized before. The Gaussian primitives, in their attempt to accommodate these inconsistencies and erroneous representations, will lead to the optimized 3D scene converging toward inaccurate structures and blurriness.

During optimization, two crucial aspects contribute fundamentally to these problems, i.e., inconsistent supervision and erroneous updates caused by artifacts. Accordingly, a customized blur-free inconsistency modeling scheme and a 3D robust optimization strategy are proposed in the following to alleviate potential inaccurate reconstruction.

Blur-Free Inconsistency Modeling. To mitigate the inconsistency effort of the single-image super-resolution model, we employ a learnable inconsistency modeling module (IM), i.e., two residual blocks [13], after the pre-trained super-resolution model to simulate the inconsistency across different views $I_{SR}^{IM} = I_{SR} + IM(I_{SR})$, thus avoiding making Gaussians to represent such inconsistency. However, we empirically found that this module tends to lose details during optimization to win better consistency, resulting in the loosing of sharp textures. Therefore, we further propose a blur proposal module (BP) to blur the rendered image which is then constrained with modified super-resolution image I_{SR}^{IM}, thereby compensating for the loss of fine-grained details and textures caused by IM. The blur proposal module, i.e., a four-layer convolutional network, predicts per-pixel blur kernels $\mathcal{B}_k \in \mathbb{R}^{K\times K\times H\times W}$ (K is the blur kernel size, 5 for $4\times$ task) with the gradient detached rendered image R_{HR}^{detach} as input $\mathcal{B}_k = BP(R_{HR}^{detach})$. Then the blurred high-resolution image is obtained via $R_{HR}^{blur} = R_{HR} * \mathcal{B}_k$ which is utilized to compute loss with I_{SR}^{IM}, avoiding the smoothing imposed by directly constraining R_{HR}:

$$\mathcal{L}_{SR} = (1-\beta)\mathcal{L}_1(R_{HR}^{blur}, I_{SR}^{IM}) + \beta\mathcal{L}_{D\text{-}SSIM}(R_{HR}^{blur}, I_{SR}^{IM}), \tag{4}$$

where β is set to 0.2 similar to the original 3DGS [18].

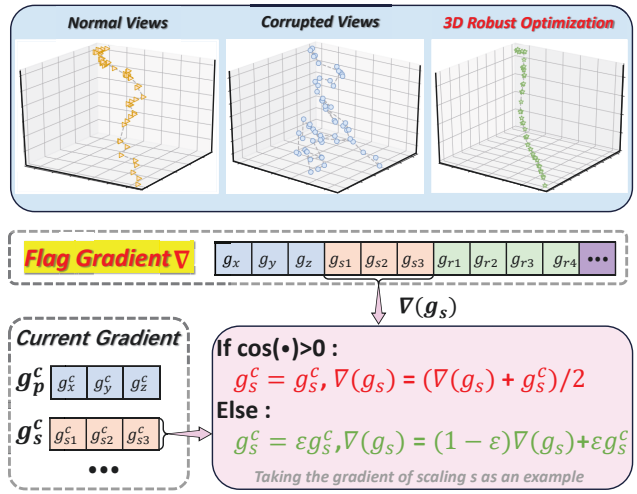

Figure 4. *Top:* Visualization of the gradient variation under different quality views supervision and our proposed 3D robust optimization. *Bottom:* Illustration of the proposed 3D robust optimization strategy that takes the gradient of scaling s as an example.

3D Robust Optimization. Albeit the above scheme could surmount the detrimental effects of multi-view inconsistencies, the optimization of Gaussians still unavoidably suffers from improper supervision of pseudo-views where some areas are not sufficiently optimized. To this end, we empirically observed that optimizing Gaussian primitives utilizing normal high-quality views results in stable and overall consistent gradients (top left of Fig. 4) whereas corrupted views can lead to significant perturbation and gradient confusion (top middle of Fig. 4), ultimately resulting in blurry rendering. Intrigued by this observation, we aim to design a more robust optimization strategy to mitigate the fluctuations caused by imperfect supervision.

More specifically, we augment the canonical Gaussian primitives

$$\{\mu, s, q, \sigma, c\} \cup \{\nabla\}, \tag{5}$$

with an additional per-Gaussian attribute flag gradient $\nabla \in \mathbb{R}^{11+k_c}$. Where ∇ is used to record the gradient trends of other attributes of the Gaussian, i.e., μ, s, q, σ, and the changed k_c SH coefficients. As shown in Fig. 4, for each Gaussian in one optimization step, we first compute the cosine similarity between the current gradient for each attribute g_i^c (i indicate different attribute) and the corresponding flag gradient $\nabla(g_i)$ stored in ∇. If the cosine similarity is greater than zero, indicating that the gradient update trends are aligned, the gradient g_i^c will be directly used to

Table 1. Quantitative comparison on Blender ×4 (8 views), LLFF ×4 (3 views), and Mip-NeRF360 ×4 (24 views). The best, second best, and third best entries are marked in red, orange, and yellow, respectively.

Method	Blender ×4 (8 views)				LLFF ×4 (3 views)				Mip-NeRF 360 ×4 (24 views)			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
NeRF-SR [34]	12.41	0.744	0.515	93.79	9.28	0.226	0.617	169.58	10.26	0.269	0.628	151.62
RegNeRF [27]	20.68	0.841	0.129	41.28	15.78	0.432	0.422	116.74	17.28	0.417	0.449	105.45
SparseNeRF [35]	21.32	0.844	0.136	42.62	16.32	0.446	0.428	127.64	17.38	0.432	0.446	109.27
3DGS [18]	20.65	0.831	0.141	42.58	13.01	0.272	0.489	138.53	16.25	0.342	0.474	108.32
SRGS [9]	22.64	0.862	0.116	36.24	17.58	0.472	0.291	102.43	18.37	0.457	0.397	91.62
Mip-Splatting [45]	22.78	0.868	0.119	33.63	15.65	0.433	0.372	114.03	18.08	0.448	0.421	86.17
DNGaussian [22]	20.97	0.844	0.135	41.15	16.27	0.430	0.346	129.26	17.21	0.422	0.452	112.22
FSGS [50]	21.45	0.854	0.133	40.36	18.35	0.505	0.247	97.07	18.64	0.458	0.394	94.21
FSGS [50]+SRGS [9]	22.68	0.865	0.114	35.42	19.49	0.590	0.175	62.63	19.83	0.513	0.448	84.12
S2Gaussian (Ours)	**24.19**	**0.879**	**0.089**	**28.47**	**20.45**	**0.654**	**0.139**	**45.89**	**22.05**	**0.687**	**0.296**	**43.51**

optimize the parameter, and the flag gradient value $\nabla(g_i)$ will be updated to the average one:

$$g_i^c = g_i^c,\ \nabla(g_i) = (\nabla(g_i) + g_i^c)/2. \tag{6}$$

Conversely, if the cosine similarity is less than or equal to zero, indicating potential disturbances, the current gradient will be scaled down by a factor of ϵ to slow the parameter updates, and $\nabla(g_i)$ will be updated as follows:

$$g_i^c = \epsilon g_i^c,\ \nabla(g_i) = (1-\epsilon)\nabla(g_i) + \epsilon g_i^c, \tag{7}$$

where ϵ is set to 0.1 to attenuate potentially erroneous updates and preserve the ability to shift update trends. In practice, such strategy can be combined with any gradient-based optimizer, e.g., Adam, by simply passing the modified gradients to the respective optimizer. Notably, ∇ will be eliminated after the optimization and will not affect the rendering speed and storage consumption.

As illustrated in the top right of Fig. 4, our proposed 3D robust optimization scheme can significantly improve the stability of Gaussian primitive optimization by eliminating erroneous updates and perturbations thus reconstructing more accurate and high-quality 3D representations.

Total Objective. Apart from $\mathcal{L}_{SR}$, an auxiliary loss combined with total variation (TV) loss and sub-pixel constraint is also deployed to stabilize the training:

$$\mathcal{L}_{AUX} = \mathcal{L}_{TV}(R_{HR}) + \mathcal{L}_1(R_{HR}\downarrow, I_{LR}), \tag{8}$$

where $\downarrow$ indicates area average downsampling and I_{LR} denotes the low-resolution images before super-resolution.

In summary, the final loss $\mathcal{L}$ for high-resolution Gaussians optimization is defined as:

$$\mathcal{L} = \mathcal{L}_{AUX} + \mathcal{L}_{SR}. \tag{9}$$

4. Experiments

Following prior literature [21, 44], we primarily experiment on 4 × super-resolution tasks in the maintext since most real-world 2D super-resolution models [36, 46] are mainly trained on × 4 data, more diagnostic experiments on other scaling factors are provided in the Suppl.

4.1. Implementation Details

We implemented S2Gaussian upon the original 3DGS repository and all experiments are conducted on a single RTX3090 GPU. For the optimization of LR GS, the pre-trained DPT [29] model is adapted for depth estimation and the total iterations are set to 10,000. During the HR GS optimization stage, the Gaussian primitives are further optimized with additional 10,000 iterations. As for the off-the-shelf 2D super-resolution model, we opt for powerful diffusion-based ResShift [46] as our backbone. We conduct experiments on three benchmark datasets with SfM for initialization, advanced techniques like DUSt3R [37] can facilitate the initialization but that is not the main purpose of our current work.

Blender Datasets [25] contain 8 detailed objects synthesized by Blender with a resolution of 800 × 800. Each object initially includes 100 training images and 200 testing images. We uniformly sample 8 images from the original 100 training images and use the low-resolution of 200 × 200 pixels for training.

LLFF Datasets [24] consist of 8 forward-facing real-world scenes. Following the previous setup, we select every eighth image as the test set, and evenly sample 3 sparse views from the remaining images for training. We leverage low-resolution images at 252 × 189 pixels for training and render at a higher resolution of 1008 × 756 pixels.

Mip-NeRF 360 Datasets [2] consist of 9 real-world scenes with 5 outdoors and 4 indoors. Similar to LLFF, we use 1/8 of the full view for testing and uniformly sample 24 images from the remaining views for training. We utilize the images with downsampling rate of 4 (about 1K resolution) as the target size and further downsample the training images by a factor 4 to obtain sparse low-resolution inputs.

The following measures are used to evaluate the quantitative performance: Peak Signal-to-Noise Ratio (PSNR) [17], Structural Similarity Index (SSIM) [38], LPIPS (VGG) [47], and Fréchet Inception Distance (FID) [14].

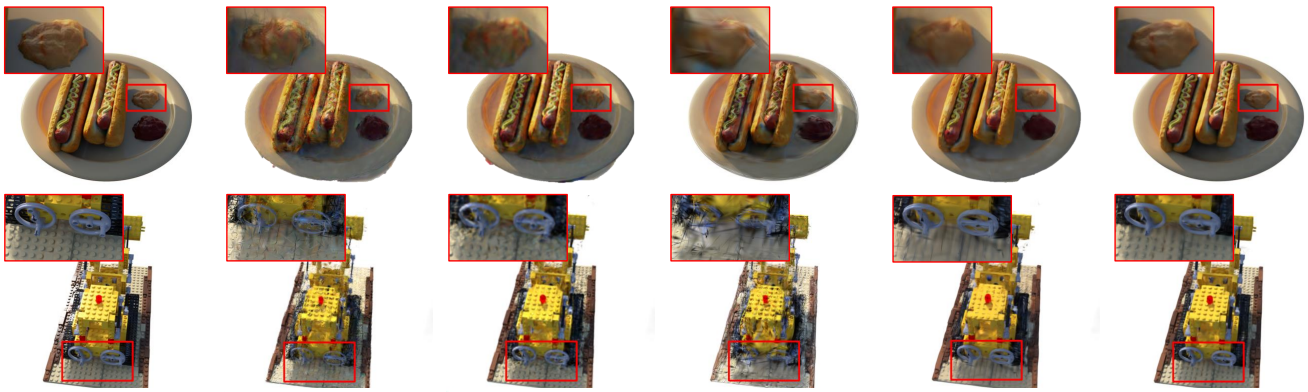

Figure 5. Qualitative comparisons on Blender ×4 datasets with 8 input views. Our method produces more visually appealing results, successfully reconstructing intricate thin structures with fine-grained details. **Best viewed at screen!**

Figure 6. Qualitative comparisons on LLFF ×4 datasets with 24 input views. Our method still generates the most accurate super-resolution 3D scenes, preserving both photo-realistic texture and geometric-level details. **Best viewed at screen!**

4.2. Quantitative and Qualitative Comparisons

To rigorously validate the effectiveness of our proposed S2Gaussian, we conduct a comparative analysis against a range of prior methods, including NeRF-SR [34], RegNeRF [27], SparseNeRF [35], vanilla 3DGS [18], SRGS [9], Mip-Splatting [45], DNGaussian [22], and FSGS [50]. Besides, we also include an enhanced version of FSGS with SRGS to provide a standard upsampling baseline. While NeRF-SR and FSGS+SRGS employ super-resolution techniques directly, the remaining methods are trained on low-resolution input views and subsequently rendered at higher resolutions.

Quantitative Evaluation. Tab. 1 presents quantitative comparison results for 4 × sparse-view super-resolution novel view synthesis task on the Blender [25], LLFF [24], and Mip-NeRF 360 datasets [2]. It is observed that our S2Gaussian delivers remarkable performance gains and outperforms all competitive method significantly in terms of PSNR, SSIM, LPIPS, and FID metrics. Especially for Mip-NeRF 360 datasets, S2Gaussian exceeds the top-performing combinatorial method FSGS+SRGS by an impressive margin of 2.22 dB PSNR, underscoring its superior capability in reconstructing large, real-world scenes. In addition to delivering higher reconstruction fidelity, reflected by PSNR and SSIM, S2Gaussian also achieves substantial improvements in perceptual metrics LPIPS and FID, validating the effectiveness of S2Gaussian in converging towards greater perception quality with faithful and fidelity details.

Qualitative Evaluation. We also demonstrate visual comparisons in Fig. 5, Fig. 6, and Fig. 7. As suggested, both 3DGS and Mip-Splatting struggle with the extremely sparse view problem, while FSGS is limited to characterizing high-resolution details. Despite the combinatorial method FSGS+SRGS can reconstruct a reasonable overall structure but suffers from inconsistency and sparsity leading to missing details and blurring. In comparison, S2Gaussian recovers more coherent fine structures with natural textures and harmonious global tone while other methods tend to introduce improper structures and unnatural details, especially for the extremely sparse scenes (Fig. 6) as well as large-scale scenarios (Fig. 7).

4.3. Ablation Studies

In Tab. 2, we conduct ablation experiments on the components introduced in S2Gaussian. The effectiveness of each

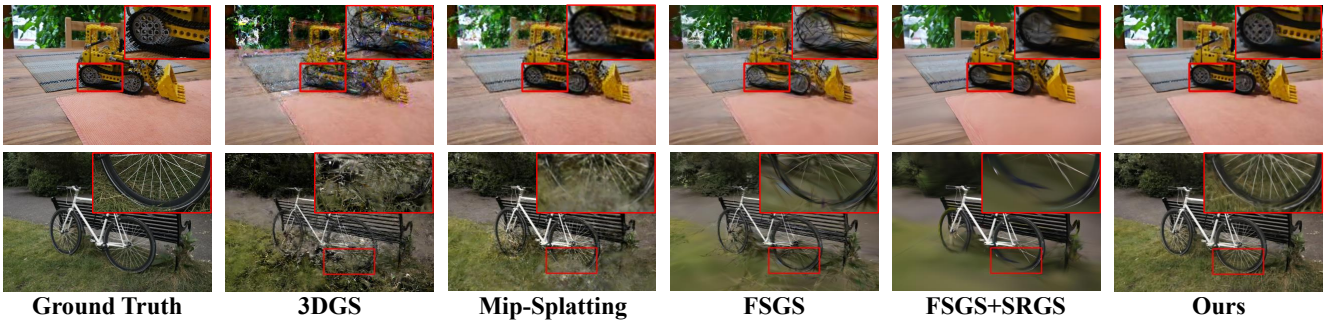

Figure 7. Qualitative comparisons on Mip-NeRF360 ×4 datasets with 24 input views. Our method consistently delivers more coherent fine structures than other methods in large-scale scenes. **Best viewed at screen!**

proposed component in S2Gaussian is evaluated by gradually integrating them into the model, revealing their individual contributions to the overall performance. More detailed analyses are as below.

Table 2. Ablation studies on Mip-NeRF 360 ×4 (24 views).

Variant	PSNR↑	SSIM↑	LPIPS↓	FID↓
Baseline	19.79	0.520	0.429	87.22
+ Two-Stage	20.06	0.528	0.415	80.64
+ Gaussian Shuffle Split	20.62	0.559	0.376	71.29
+ Blur-Free Inconsistency Modeling	21.12	0.624	0.346	59.63
+ 3D Robust Optimization	22.05	0.687	0.296	43.51

Effect of Gaussian Shuffle Split. As shown in Tab. 3, fewer sub-Gaussians lead to suboptimal detail modeling (lower PSNR), while 6 offers the best trade-off, while more sub-Gaussians will not only increase the optimization complexity but also introduce extra memory overhead. Additionally, 6 sub-Gaussians can be symmetrically located along the three principal axes of the original Gaussian, perfectly approximating the original Gaussian representation.

Table 3. Ablation of sub-Gaussian number on Mip-NeRF 360.

Number	4	5	**6**	7	8	9
Storage Memory (M)	67	89	**112**	137	169	202
Final PSNR	21.76	21.95	**22.05**	22.03	22.09	22.08

Effect of Blur-Free Inconsistency Modeling. As illustrated in Fig. 8, IM models the inconsistency of super-resolved images, and while this would introduce blurring, we model this effect explicitly through blur kernel learning, thus reconstructing results that are blur-free and with fine structural details.

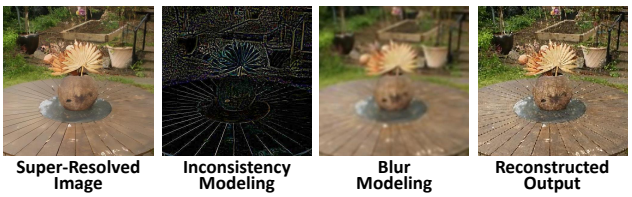

Figure 8. Visualization of the super-resolved image, the residual output of IM, the blurred reconstructed image with predicted blur kernels, and the original reconstructed output.

Effect of 3D Robust Optimization. Fig. 9 showcases visual comparison with or without 3D robust optimization under corrupted training views. It is observed that 3D robust optimization can effectively avoid unfavorable or erroneous supervision to optimize towards higher quality whereas its absence results in Gaussian primitives simulating corrupted representations.

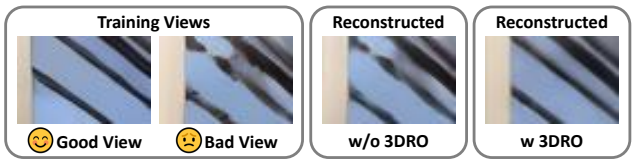

Figure 9. Visual comparison with or without 3D robust optimization (3DRO) under corrupted training views.

5. Concluding Remarks

In this paper, we propose S2Gaussian, an innovative framework that can efficiently reconstruct high-quality 3D scenes with only sparse and low-resolution input views. In contrast to existing frameworks, S2Gaussian can handle both view sparse and clarity deficient simultaneously, which is more functional and practical in real-world scenarios. The proposed framework begins by optimizing a low-resolution Gaussian representation, which is then densified to initialize high-resolution Gaussians via tailored Gaussian Shuffle Split operation. Subsequently, it cleverly solves the view sparsity problem by rendering pseudo-views from low-resolution Gaussians for super-resolution supervision. Wherein a customized blur-free inconsistency modeling scheme and a 3D robust optimization strategy are further proposed to mitigate multi-view inconsistency and eliminate erroneous updates for better geometric structure and fine-grained details reconstruction. Extensive experiments on multiple benchmarks manifest the effectiveness, superiority, and robustness of our method. We expect this work to provide insights into more complicated 3D reconstruction and steer future research on this Gordian knot.

Acknowledgements

The authors are very indebted to the anonymous referees for their critical comments and suggestions for the improvement of this paper. This work was supported by the National Key Research and Development Program of China (Grant No. 2021YFA1000102), the National Natural Science Foundation of China (Grant Nos. 62376285 and 61673396), and Natural Science Foundation of Shandong Province, China (Grant No. ZR2022MF260).

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021.

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022.

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023.

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022.

[5] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024.

[6] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. In *Forty-first International Conference on Machine Learning*, 2024.

[7] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2024.

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022.

[9] Xiang Feng, Yongbo He, Yubo Wang, Yan Yang, Zhenzhong Kuang, Yu Jun, Jianping Fan, et al. Srgs: Super-resolution 3d gaussian splatting. *arXiv preprint arXiv:2404.10318*, 2024.

[10] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5501–5510, 2022.

[11] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14346–14355, 2021.

[12] Yuqi Han, Tao Yu, Xiaohang Yu, Yuwang Wang, and Qionghai Dai. Super-nerf: View-consistent detail generation for nerf super-resolution. *arXiv preprint arXiv:2304.13518*, 2023.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[15] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023.

[16] Xudong Huang, Wei Li, Jie Hu, Hanting Chen, and Yunhe Wang. Refsr-nerf: Towards high fidelity and super resolution view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8244–8253, 2023.

[17] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[19] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023.

[20] Hyun-kyu Ko, Dongheok Park, Youngin Park, Byeonghyeon Lee, Juhee Han, and Eunbyung Park. Sequence matters: Harnessing video models in super-resolution. *arXiv preprint arXiv:2412.11525*, 2024.

[21] Jie Long Lee, Chen Li, and Gim Hee Lee. Disr-nerf: Diffusion-guided view-consistent super-resolution nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20561–20570, 2024.

[22] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20775–20785, 2024.

[23] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20654–20664, 2024.

[24] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019.

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[26] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[27] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

[28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[30] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Hyeonsu Kim, Jaehoon Ko, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023.

[31] Yuan Shen, Duygu Ceylan, Paul Guerrero, Zexiang Xu, Niloy J Mitra, Shenlong Wang, and Anna Früstück. Supergaussian: Repurposing video models for 3d super resolution. *arXiv preprint arXiv:2406.00609*, 2024.

[32] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[33] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.

[34] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6445–6454, 2022.

[35] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023.

[36] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024.

[37] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[39] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024.

[40] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360 {\deg} sparse view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00206*, 2023.

[41] Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20923–20931, 2024.

[42] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8254–8263, 2023.

[43] Youngho Yoon and Kuk-Jin Yoon. Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12428–12438, 2023.

[44] Xiqian Yu, Hanxin Zhu, Tianyu He, and Zhibo Chen. Gaussiansr: 3d gaussian super-resolution with 2d diffusion priors. *arXiv preprint arXiv:2406.10111*, 2024.

[45] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024.

[46] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024.

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[48] Xiaoyun Zheng, Liwei Liao, Jianbo Jiao, Feng Gao, and Ronggang Wang. Surface-sos: Self-supervised object segmentation via neural surface representation. *IEEE Transactions on Image Processing*, 2024.

[49] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding

with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.

[50] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European Conference on Computer Vision*, pages 145–163. Springer, 2025.