

GaussianSpa: An “Optimizing-Sparsifying” Simplification Framework for Compact and High-Quality 3D Gaussian Splatting

Yangming Zhang* Wenqi Jia*
 Dept. of Computer Science
 University of Texas at Arlington
 {yxyz0925, wxj1489}@mavs.uta.edu

Wei Niu
 School of Computing
 University of Georgia
 wniu@uga.edu

Miao Yin†
 Dept. of Computer Science
 University of Texas at Arlington
 miao.yin@uta.edu



Figure 1. We present GaussianSpa, enabling high-quality and compact view synthesis with superior rendering of details. Compared to the existing state-of-the-art method, Mini-Splatting [17], our GaussianSpa captures detail-rich textures and low-frequency background more accurately with fewer Gaussians.

Abstract

3D Gaussian Splatting (3DGS) has emerged as a mainstream for novel view synthesis, leveraging continuous aggregations of Gaussian functions to model scene geometry. However, 3DGS suffers from substantial memory requirements to store the large amount of Gaussians, hindering its efficiency and practicality. To address this challenge, we introduce **GaussianSpa**, an optimization-based simplification framework for compact and high-quality 3DGS. Specifically, we formulate the simplification objective as a constrained optimization problem associated with the 3DGS training. Correspondingly, we propose an efficient “optimizing-sparsifying” solution for the formulated problem, alternately solving two independent sub-problems and gradually imposing substantial sparsity onto the Gaussians in the 3DGS training process. We conduct quantitative and qualitative evaluations on various datasets, demonstrating the superiority of GaussianSpa over existing state-of-the-art approaches. Notably, GaussianSpa achieves an **average PSNR improvement of 0.9 dB on the real-world**

Deep Blending dataset with $10\times$ fewer Gaussians compared to the vanilla 3DGS. Our project page is available at <https://noodle-lab.github.io/gaussianspa>.

1. Introduction

Novel view synthesis has become a pivotal area in computer vision and graphics, advancing applications such as virtual reality, augmented reality, and immersive media experiences [18]. NeRF [32] has recently gained prominence in this domain because it can generate high-quality, photorealistic images from sparse input views by representing scenes as continuous volumetric functions based on neural networks. However, NeRF requires substantial computational resources and long training times, making it less practical for real-time and large-scale applications.

3D Gaussian Splatting (3DGS) [22] has emerged as a powerful alternative, leveraging continuous aggregations of Gaussian functions to model scene geometry and appearance. Unlike NeRF, which relies on neural networks to approximate volumetric radiance fields, 3DGS directly represents scenes using a collection of Gaussians. This approach effectively captures details and smooth transitions, offering

*Yangming Zhang is the leading co-first author, while Wenqi Jia is the secondary co-first author. †Miao Yin is the corresponding author.

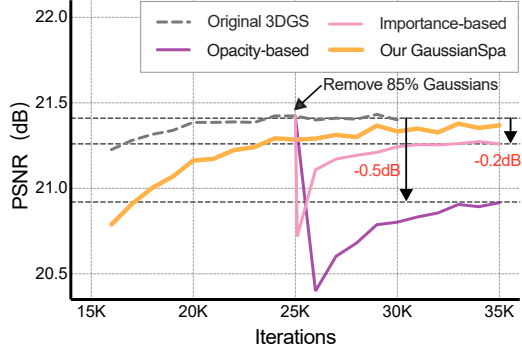


Figure 2. **PSNR curves of hand-crafted criteria-based pruning methods.** Gaussians are removed by 85% at iteration 25K.

faster training and rendering. 3DGS achieves superior visual fidelity [25] compared to NeRF while reducing computational overhead, making it more suitable for interactive applications that demand quality and performance.

Despite its strengths, 3DGS suffers from significant memory requirements that hinder its practicality. The main issue is the massive memory consumption required to store numerous Gaussians, each with parameters like position, covariance, and color. In densely sampled scenes, the sheer volume of Gaussians leads to memory usage that exceeds the capacity of typical hardware, making it challenging to handle higher-resolution scenes and limiting its applicability in resource-constrained environments.

Existing works, e.g., Mini-Splatting [17], LightGaussian [16], LP-3DGS [51], EfficientGS [28], and RadSplat [37], have predominantly focused on mitigating this issue by removing a certain number of Gaussians. Techniques such as pruning and sampling aim to discard unimportant Gaussians based on hand-crafted criteria such as opacity [22, 50], importance score (hit count) [16, 17, 37], dominant primitives [28], and binary mask [26, 51]. However, such single-perspective heuristic criteria may lack robustness in dynamic scenes or varying lighting. Moreover, the sudden one-shot removal may cause permanent loss of Gaussians crucial to visual synthesis, making it challenging to recover the original performance after even long-term training, as shown in Figure 2. Mask-based pruning methods also suffer from this issue due to weak sparsity enforcement on the multitude of Gaussians. Consequently, while these methods can alleviate memory and storage burdens to some extent, they often lead to sub-optimal rendering outcomes with loss of details and visual artifacts, thereby compromising the quality of the synthesized views.

In this paper, we present an optimization-based simplification framework, GaussianSpa, for compact and high-quality 3DGS. In the proposed framework, we formulate 3DGS simplification as a constrained optimization problem under a target number of Gaussians. Then, we propose an

efficient “optimizing-sparsifying” solution for the formulated problem by splitting it into two simple sub-problems that are alternately solved in the “optimizing” step and the “sparsifying” step. Instead of permanently removing a certain number of Gaussians, GaussianSpa incorporates the “optimizing-sparsifying” algorithm into the training process, gradually imposing a substantial sparse property onto the trained Gaussians. Hence, our GaussianSpa can simultaneously enjoy maximum information preservation from the original Gaussians and a desired number of reduced Gaussians, providing compact 3DGS models with high-quality rendering. Overall, our contributions can be summarized as follows:

- We propose a general 3DGS simplification framework that formulates the simplification objective as an optimization problem and solves it in the 3DGS training process. In solving the formulated optimization problem, our proposed framework gradually restricts Gaussians to the target sparsity constraint without explicitly removing a specific number of points. Hence, GaussianSpa can maximally maintain and smoothly transfer the information to the sparse Gaussians from the original model.
- We propose an efficient “optimizing-sparsifying” solution for the formulated problem, which can be integrated into the 3DGS training with negligible costs, separately solving two sub-problems. In the “optimizing” step, we optimize the original loss function attached by a regularization with gradient descent. In the “sparsifying” step, we analytically project the auxiliary Gaussians onto the constrained sparse space.
- We comprehensively evaluate GaussianSpa via extensive experiments on various complex scenes, demonstrating improved rendering quality compared to existing state-of-the-art approaches. Particularly, with as high as $10\times$ fewer number of Gaussians than the vanilla 3DGS, GaussianSpa achieves an average 0.4 dB PSNR improvement on the Mip-NeRF 360 [2] and Tanks&Temples [24] datasets, 0.9 dB on Deep Blending [20] dataset. Furthermore, we conduct various visual quality analyses, demonstrating the superiority of GaussianSpa in rendering high-quality views and capturing rich details.

2. Related Work

2.1. Novel View Synthesis

Novel View Synthesis (NVS) [1] generates images of a 3D scene from unseen viewpoints based on existing multi-view data. A seminal work in this field is NeRF [32], which uses an implicit neural network to represent a scene as a continuous 5-dimensional function, mapping 3D coordinates and viewing directions to color and density. However, its reliance on single narrow-ray sampling poses challenges in addressing aliasing and blur artifacts. More recently, 3D

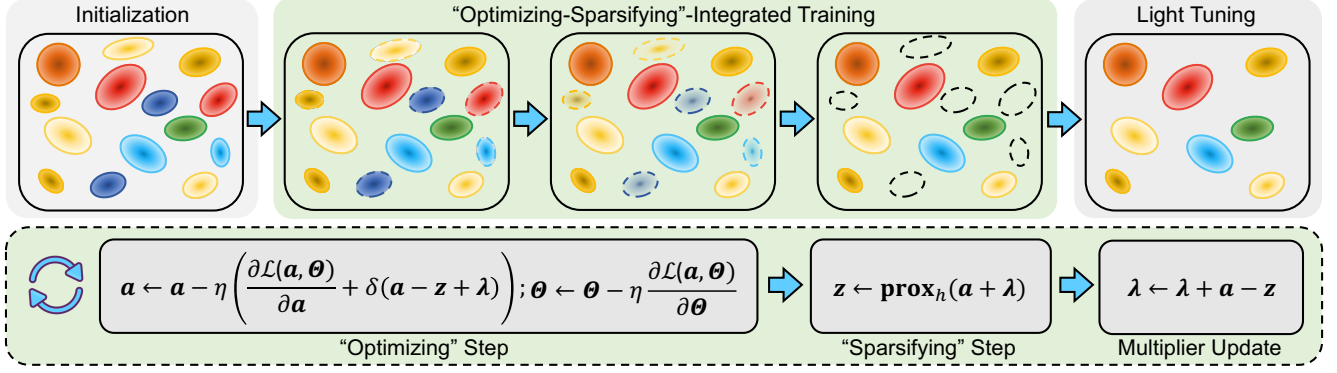


Figure 3. Overall workflow of our proposed GaussianSpa framework.

Gaussian Splatting (3DGS), which leverages continuous aggregations of Gaussian functions to model scene geometry, has emerged as a point-based alternative, demonstrating significant improvements in rendering quality and speed.

2.2. Efficient 3DGS

3DGS significantly improves rendering quality and computational efficiency compared to traditional methods such as NeRF [32], which are often resource-intensive and computationally slow. However, the demands of real-time rendering and the reconstruction of unbounded, large-scale 3D scenes underscore the requirement for model efficiency [7]. Optimization for efficient 3DGS focuses on four main categories: Gaussian simplification, quantization, spherical harmonics optimization, and hybrid compression.

Gaussian Simplification. Gaussian simplification involves removing Gaussians to compress 3DGS models into compact forms, significantly reducing memory costs. Prior approaches [10, 17, 23, 27, 29, 31, 37, 43, 51] mainly focus on pruning redundant Gaussians according to hand-crafted importance criteria. Mini-Splatting [17] addresses overlapping and reconstruction issues with blur splitting, depth reinitialization, and stochastic sampling. Radsplatting [37] enhances robustness by using a max operator to compute importance scores from ray contributions. Taming 3DGS [31] applies selective densification based on pixel saliency and gradients, while LP-3DGS [51] learns a binary mask for efficient pruning.

Quantization. Quantization [47] aims to compress data using discrete entries without appreciable degradation in quality, reducing data redundancy. Several works [9, 19, 33, 35, 36, 39, 48] apply quantization for compressing 3DGS models. For example, EAGLES [19] quantizes attributes like spherical harmonics (SH) coefficients. Vector quantization, a form of quantization, compresses data by mapping vectors to a learned codebook. CompGS [35] leverages K-means-based vector quantization to reduce storage and rendering time, while RDO-Gaussian [48] combines pruning and entropy-constrained vector quantization

for efficient compression with minimal quality loss.

Spherical Harmonics Optimization. Among 3DGS parameters, spherical harmonics (SH) coefficients take up a significant portion, requiring (45+3) floating-point values per Gaussian for a degree of 3 [22], which accounts for around 80% of the total attribute volume. Prior works [16, 26] optimize SH coefficients to reduce storage overhead. For instance, EfficientGS [28] retains zeroth-order SH for all Gaussians, increasing the order when necessary.

Hybrid Compression. In addition to methods focused on individual aspects of 3DGS compression, several recent approaches [16, 26, 28, 49] combine multiple techniques for hybrid compression to achieve further storage reduction. LightGaussian [16] uses a global importance score for pruning and applies vector quantization to compress SH coefficients. EfficientGS [28] selectively increases SH order and densifies non-steady Gaussians. CompactGaussian [26] introduces a learnable mask for pruning, uses residual vector quantization for scale and rotation, and replaces SH with a grid-based neural field for color representation.

Although mask-based methods impose sparsity relying on mask loss, they and other methods both lead to irreversible performance loss in the Gaussian simplification process by performing a one-shot pruning on the Gaussians, as shown in Figure 2. Consequently, performance often worsens when applying SH optimization and quantization due to accumulated errors. Our work focuses on *Gaussian simplification* by formulating it as an optimization problem and imposing substantial sparsity onto Gaussians, mitigating performance issues with improved rendering quality.

3. Methodology: GaussianSpa

3.1. Background of 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) explicitly represents scenes using a collection of point-based 3D continuous Gaussians. Specifically, each Gaussian G is characterized by its covari-

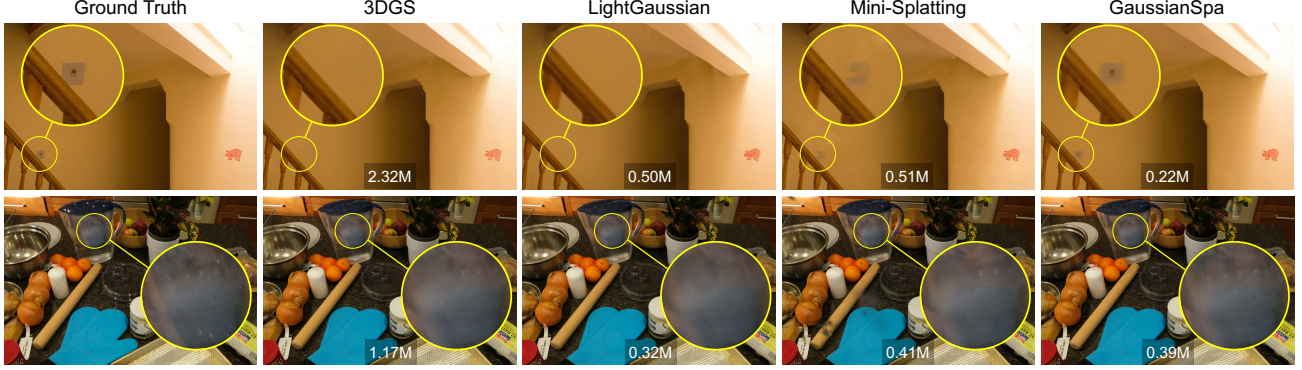


Figure 4. **Visual quality results on the DrJohnson and Counter scenes, compared to existing simplification approaches and vanilla 3DGS.** The numbers of remaining Gaussians are displayed. It is observed that our GaussianSpa recovers details closest to the ground truth in the actual rendering outcomes with a significantly reduced number of Gaussians.

ance matrix Σ and center position μ as

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

where $\mathbf{x}$ is an arbitrary location in the 3D scene. The covariance matrix Σ is generally composed of $\Sigma = \mathbf{R}\mathbf{S}\mathbf{S}^\top \mathbf{R}^\top$ with a rotation matrix $\mathbf{R}$ and a scale matrix $\mathbf{S}$, ensuring the positive definite property. Each Gaussian also has an associated opacity a and spherical harmonics (SH) coefficients.

When rendering a 2D image from the 3D scene at a certain camera angle, 3DGS projects 3D Gaussians to that 2D plane and calculates the projected 2D covariance matrix by

$$\Sigma' = \mathbf{J}\mathbf{W}\Sigma\mathbf{W}^\top \mathbf{J}^\top, \quad (2)$$

where $\mathbf{W}$ is the view transformation, and $\mathbf{J}$ is the Jacobian of the affine approximation of the projective transformation [52]. Then, for each pixel on the 2D image, the color C is computed by blending all N ordered Gaussians contributing to the pixel as

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j). \quad (3)$$

Here, c_i and α_i denote the view-dependent color computed from the corresponding SH coefficients and the rendering opacity calculated by the Gaussian opacity a_i , respectively, corresponding to the i -th Gaussian.

In the training process [22], the Gaussians are initialized from a sparse point cloud generated by Structure-from-Motion (SfM) [44]. Then, the number of Gaussians is adjusted by a densification algorithm [22] according to the reconstruction geometry. Afterward, each Gaussian's attributes, including the center position, rotation matrix, scaling matrix, opacity, and SH coefficients, are optimized by minimizing the reconstruction errors using the standard gradient descent. Specifically, the loss $\mathcal{L}$ is given by

$$\mathcal{L} = (1 - \rho)\mathcal{L}_1 + \rho\mathcal{L}_{\text{D-SSIM}}, \quad (4)$$

where $\mathcal{L}_1$ and $\mathcal{L}_{\text{D-SSIM}}$ are the ℓ_1 norm and the structural dissimilarity between the rendered output and the ground truth, respectively. ρ is a parameter that controls the tradeoff between the two losses.

3.2. Problem Formulation: Simplification as Constrained Optimization

To mitigate irreversible information loss caused by existing one-shot criterion-based pruning, our key innovation is to gradually impose substantial sparsity onto the Gaussians and maximally preserve information in the training process. Mathematically, considering a general 3DGS training, we introduce a sparsity constraint on the Gaussians to the training objective, $\min \mathcal{L}$. Thus, the 3DGS training is reformulated as another constrained optimization problem, the primary formulation of our proposed simplification framework, i.e.,

$$\begin{aligned} \min \mathcal{L}, \\ \text{s.t. } \mathcal{N}(\mathbf{G}) \leq \kappa. \end{aligned} \quad (5)$$

The introduced constraint $\mathcal{N}(\cdot) \leq \kappa$ restricts the number of Gaussians no greater than a target number κ .

However, the above problem cannot be solved because the restriction on the number of Gaussian functions is not differentiable. Recalling the rendering process with Eq. 3, the contribution of the i -th Gaussian to a particular pixel is finally determined by its opacity a_i . Alternatively, the original sparsity constraint directly on the Gaussians can be explicitly transformed into a sparsity constraint on the opacity variables. As each Gaussian only has one opacity variable, we can use a vector $\mathbf{a}$ to represent all the opacities and easily restrict its number of non-zeros with ℓ_0 norm, $\|\mathbf{a}\|_0$. For better presentation, we use a variable set Θ to represent other GS variables except for the opacity $\mathbf{a}$. Hence, the constrained optimization problem can be reformulated as

Algorithm 1 Procedure of “Optimizing-Sparsifying”

Input: Gaussian opacity $\mathbf{a}$, 3DGS variables Θ , target number of Gaussians κ , penalty parameter δ , feasibility tolerance ϵ , maximum iterations T .

Output: Optimized $\mathbf{a}$ and Θ .

```

1:  $\mathbf{z} \leftarrow \mathbf{a}, \lambda \leftarrow 0$ ;
2:  $t \leftarrow 0$ ;
3: while  $\|\mathbf{a} - \mathbf{z}\|^2 > \epsilon$  and  $t \leq T$  do
4:   Update  $\mathbf{a}$  and  $\Theta$  with Eq. 14;  $\triangleright$  “Optimizing” Step
5:   Update  $\mathbf{z}$  with Eq. 16;  $\triangleright$  “Sparsifying” Step
6:   Update  $\lambda$  with Eq. 17;  $\triangleright$  Multiplier Update
7:    $t \leftarrow t + 1$ ;
8: end while

```

$$\begin{aligned} \min_{\mathbf{a}, \Theta} \mathcal{L}(\mathbf{a}, \Theta), \\ \text{s.t. } \|\mathbf{a}\|_0 \leq \kappa. \end{aligned} \quad (6)$$

With this reformulation, we have simplified the primary problem to another form that our proposed “optimizing-sparsifying” can efficiently solve.

3.3. “Optimizing-Sparsifying” Solution

In this subsection, we present an efficient “optimizing-sparsifying” solution for problem Eq. 6. The key idea is to split it into two simple sub-problems separately solved in an “optimizing” step and a “sparsifying” step alternately.

In order to deal with the non-convex constraint, $\|\mathbf{a}\|_0 \leq \kappa$, we first introduce an indicator function,

$$h(\mathbf{a}) = \begin{cases} 0, & \|\mathbf{a}\|_0 \leq \kappa, \\ +\infty, & \text{otherwise,} \end{cases} \quad (7)$$

to the minimization objective and remove the sparsity constraint. Thus, problem Eq. 6 becomes an unconstrained optimization form, i.e.,

$$\min_{\mathbf{a}, \Theta} \mathcal{L}(\mathbf{a}, \Theta) + h(\mathbf{a}). \quad (8)$$

Here, the first term is the original 3DGS training objective, which is easy to handle, while the second term is still non-differentiable. This prevents the above problem from being solved using gradient-based solutions under existing training frameworks such as PyTorch [41]. To decouple the two terms, we further introduce an auxiliary variable $\mathbf{z}$, whose size is identical to $\mathbf{a}$, and convert Eq. 8 as

$$\begin{aligned} \min_{\mathbf{a}, \mathbf{z}, \Theta} \mathcal{L}(\mathbf{a}, \Theta) + h(\mathbf{z}), \\ \text{s.t. } \mathbf{a} = \mathbf{z}. \end{aligned} \quad (9)$$

The above problem becomes another constrained problem with an equality constraint. Thus, the corresponding dual augmented Lagrangian [4] form can be given by

$$L(\mathbf{a}, \mathbf{z}, \Theta, \lambda; \delta) = \mathcal{L}(\mathbf{a}, \Theta) + h(\mathbf{z}) + \frac{\delta}{2} \|\mathbf{a} - \mathbf{z} + \lambda\|^2 + \frac{\delta}{2} \|\lambda\|^2, \quad (10)$$

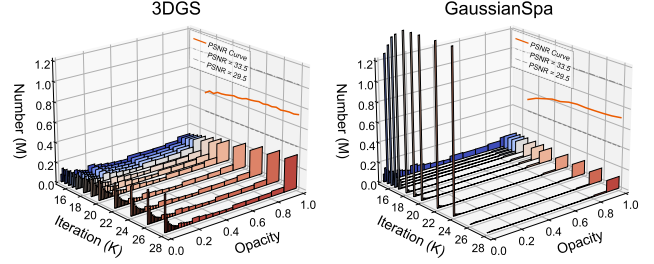


Figure 5. Evolution of opacity distribution and PSNR for 3DGS and GaussianSpa. The “optimizing-sparsifying” starts at iteration 15K. At iteration 25K, GaussianSpa removes “zero” Gaussians, followed by a light tuning.

where λ is the dual Lagrangian multiplier, and δ is the penalty parameter. Correspondingly, we can separately optimize over the 3DGS variables, $\mathbf{a}, \Theta$, and the auxiliary variable, $\mathbf{z}$, in the augmented Lagrangian, then solve two split sub-problems in an “optimizing” step and a “sparsifying” step alternately.

► “Optimizing” Step: $\min_{\mathbf{a}, \Theta} \mathcal{L}(\mathbf{a}, \Theta) + \frac{\delta}{2} \|\mathbf{a} - \mathbf{z} + \lambda\|^2. \quad (11)$

In the “optimizing” step, the minimization objective, Eq. 11, targets only 3DGS variables and contains two terms, which are the original 3DGS loss function and a quadratic regularization that enforces the opacity $\mathbf{a}$ close to the exactly sparse variable $\mathbf{z}$. Since both terms are differentiable, the sub-problem in this step can be directly optimized with gradient descent. In other words, the solution for this sub-problem is consistent with *optimizing* 3DGS variables with additional gradients, which are computed by

$$\frac{\partial L}{\partial \mathbf{a}} = \frac{\partial \mathcal{L}(\mathbf{a}, \Theta)}{\partial \mathbf{a}} + \delta(\mathbf{a} - \mathbf{z} + \lambda), \quad (12)$$

$$\frac{\partial L}{\partial \Theta} = \frac{\partial \mathcal{L}(\mathbf{a}, \Theta)}{\partial \Theta}. \quad (13)$$

Hence, the opacity $\mathbf{a}$ and other GS variables Θ , can be updated by

$$\mathbf{a} \leftarrow \mathbf{a} - \eta \frac{\partial L}{\partial \mathbf{a}}, \quad \Theta \leftarrow \Theta - \eta \frac{\partial L}{\partial \Theta}, \quad (14)$$

respectively. Here, η is the learning rate during training.

► “Sparsifying” Step: $\min_{\mathbf{z}} h(\mathbf{z}) + \frac{\delta}{2} \|\mathbf{a} - \mathbf{z} + \lambda\|^2. \quad (15)$

In the “sparsifying” step, the minimization objective, Eq. 15, targets only the auxiliary variable, $\mathbf{z}$, while the term of indicator function dominates Eq. 15. This sub-problem essentially has a form of the proximal operator [40], prox_h , associated with the indicator function $h(\cdot)$. Thus, the solution of this sub-problem can be directly given by

$$\mathbf{z} \leftarrow \text{prox}_h(\mathbf{a} + \lambda). \quad (16)$$

Method	Mip-NeRF 360				Tanks&Temples				Deep Blending			
	PSNR↑	SSIM↑	LPIPS↓	#G/M↓	PSNR↑	SSIM↑	LPIPS↓	#G/M↓	PSNR↑	SSIM↑	LPIPS↓	#G/M↓
3DGS [22]	27.45	0.810	0.220	3.110	23.63	0.850	0.180	1.830	29.42	0.900	0.250	2.780
CompactGaussian [26]	27.08	0.798	0.247	1.388	23.32	0.831	0.201	0.836	29.79	0.901	0.258	1.060
LP-3DGS-R [51]	27.47	0.812	0.227	1.959	23.60	0.842	0.188	1.244	-	-	-	-
LP-3DGS-M [51]	27.12	0.805	0.239	1.866	23.41	0.834	0.198	1.116	-	-	-	-
EAGLES [19]	27.23	0.809	0.238	1.330	23.37	0.840	0.200	0.650	29.86	0.910	0.250	1.190
Mini-Splatting [17]	27.40	0.821	0.219	0.559	23.45	0.841	0.186	0.319	30.05	0.909	0.254	0.397
Taming 3DGS [31]	27.31	0.801	0.252	0.630	23.95	0.837	0.201	0.290	29.82	0.904	0.260	0.270
CompGS [35]	27.12	0.806	0.240	0.845	23.44	0.838	0.198	0.520	29.90	0.907	0.251	0.550
GaussianSpa	27.85	0.825	0.214	0.547	23.98	0.852	0.180	0.269	30.37	0.914	0.249	0.335
									30.33	0.912	0.254	0.256

Table 1. **Quantitative results on multiple datasets, compared with existing state-of-the-art works.** Mini-Splatting [17] results are replicated using official code. 3DGS results are reported from [19]. “#G/M” represents number of million Gaussians.

The proximal operator maps the opacity α to the auxiliary variable z , which contains at most κ non-zeros. This operator results in a *projection* solution [5], *sparsifying* ($\alpha + \lambda$), by projecting it onto a set that satisfies $\|z\|_0 \leq \kappa$. More specifically, the projection keeps top- κ elements and sets the rest to zeros.

► **Overall Procedure.** After the “optimizing” step and the “sparsifying” step, the dual Lagrangian multiplier is updated by

$$\lambda \leftarrow \lambda + \alpha - z. \quad (17)$$

Overall, the “optimizing,” “sparsifying,” and multiplier update steps are performed alternately in the 3DGS training process until convergence, as summarized in Algorithm 1. Generally, we consider our solution convergences once the condition, $\|\alpha - z\|^2 \leq \epsilon$, is satisfied, or the iterations reach the predefined maximum number.

Remark. The “optimizing” step optimizes over the original Gaussian variables and pushes the Gaussian opacity α close to the exactly sparse auxiliary variable z using a gradient-based approach, simultaneously satisfying the 3DGS performance and the sparsity requirements. The “sparsifying” step essentially prunes the auxiliary variable z , which can be considered the exactly sparse version of α . As those steps are performed alternately, summarized in Algorithm 1, the process eventually converges to a sparse set that maximizes the performance. Upon the convergence, most Gaussians become “zeros,” and the information can be preserved and transferred to the “non-zero” Gaussians.

3.4. Overall Workflow of GaussianSpa

With the proposed “optimizing-sparsifying”-integrated training process presented in the previous subsection, the Gaussians exhibit substantial sparsity characteristics. In other words, a certain number of Gaussian opacities are close to zeros, meaning those Gaussians have almost no contribution to the rendering, which can be directly removed. As the redundant Gaussians are appropriately

removed, our simplified 3DGS model can exhibit superior performance after a light tuning, even slightly higher than the original 3DGS. The overall workflow of our GaussianSpa is illustrated in Figure 3.

We visualize the evolution of Gaussian opacity distribution along with the PSNR, trained on the Room scene, as shown in Figure 5. The opacity distribution of GaussianSpa is significantly distinct from the original 3DGS – there is a clear gap between the “zero” Gaussians and the remaining Gaussians in GaussianSpa, while the PSNR keeps increasing. This means GaussianSpa has successfully imposed a substantial sparsity property and preserved information to the “non-zero” Gaussians. At iteration 25K, we remove all “zero” Gaussians and perform a light tuning that further improves the performance, finally obtaining a compact model with high-quality rendering.

4. Experiments

4.1. Experimental Settings

Datasets, Baselines, and Metrics. We evaluate the rendering performance on multiple real-world datasets, including Mip-NeRF 360 [3], Tanks&Temples [24], and Deep Blending [20]. On all datasets, we compare our proposed GaussianSpa with vanilla 3DGS [22] and existing state-of-the-art Gaussian simplification approaches, including CompactGaussian [26], LP-3DGS [51], EAGLES [19], Mini-Splatting [17], Taming 3DGS [31], and CompGS [35]. We follow the standard practice to quantitatively evaluate the rendering performance, reporting results using the peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and learned perceptual image patch similarity (LPIPS). In addition to quantitative results, we comprehensively evaluate GaussianSpa with visual analyses such as ellipsoid view distribution and point cloud visualization against the existing state-of-the-art approaches.

Implementation Details. We conduct experiments under the same environment specified in the original 3DGS

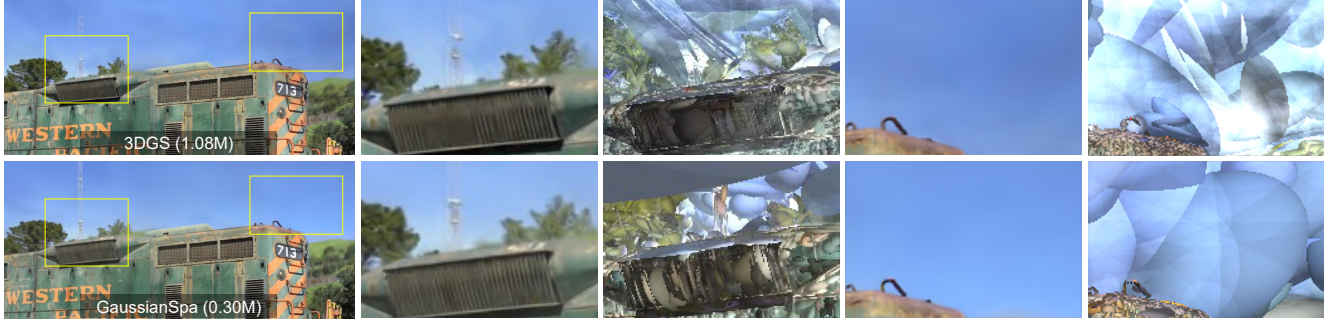


Figure 6. **Visualization for rendered Gaussian ellipsoids.** (Add'l Figure 12 in the Supplementary Material). GaussianSpa renders high-frequency (e.g., air outlet of the train) and low-frequency (e.g., large-area sky) areas adaptively using appropriate numbers of Gaussians, respectively, more precisely matching the original textures, compared to the state-of-the-art Mini-Splatting [17] and the vanilla 3DGS.

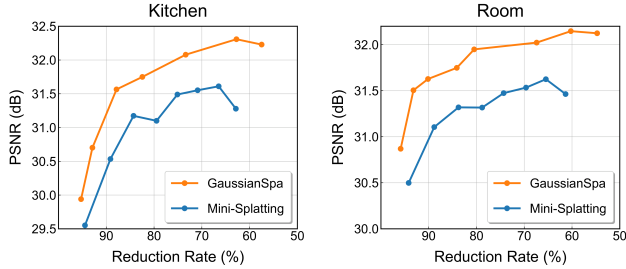


Figure 7. **Quality-Reduction Rate curves on the (left) Kitchen and (right) Room scenes.** GaussianSpa outperforms Mini-Splatting [17] at multiple reduction rates.

[22] with the PyTorch framework. Our experimental server has two AMD EPYC 9254 CPUs and eight NVIDIA GTX 6000 Ada GPUs. We use the same starting checkpoint files as Mini-Splatting [17] before our “optimizing-sparsifying” process to ensure fair comparisons. GaussianSpa starts “optimizing-sparsifying” at iteration 15K and removes “zero” Gaussians at iteration 25K.

4.2. Quantitative Results

The quantitative results are summarized in Table 1, compared with various existing approaches. The table shows that our proposed GaussianSpa outperforms the baseline 3DGS across all three metrics with substantially fewer Gaussians. To be specific, after reducing the Gaussian points by $6\times$, GaussianSpa still improves the PSNR by 0.4 dB over the vanilla 3DGS. Compared to other criterion-based simplification methods, including EAGLES [19], Mini-Splatting [17], Taming 3DGS [31], and CompGS [35], as well as the learnable mask-based approaches, CompactGaussia [26] and LP-3DGS [51], our GaussianSpa exhibits a significant improvement under a less number of Gaussians. Particularly, GaussianSpa achieves a PSNR improvement as high as 0.7 dB with even $3\times$ fewer Gaussians, compared against LP-3DGS [51]. These results quantitatively demonstrate the superiority of our GaussianSpa.

We plot quality curves for the Gaussian reduction rate

on the Room and Kitchen in Figure 7 (more results are in Figure 14 in the Supplementary Material), showing the robustness of GaussianSpa for different numbers of remaining Gaussians. GaussianSpa outperforms the state-of-the-art method, Mini-Splatting [17], by an average of 0.5 dB PSNR improvement across multiple Gaussian reduction rates.

4.3. Visual Quality Results

Figure 4 shows the rendered images on the Drjohnson and Counter scenes [3, 20] for visual quality evaluation, in comparison to LightGaussian [16], Mini-Splatting [17], and the vanilla 3DGS. Together with Figure 1, these outcomes demonstrate GaussianSpa’s superior rendering quality on diverse scenes over existing approaches, consistent with the quantitative results in the previous subsection. For instance, as shown in Figure 4, GaussianSpa accurately renders the outlet behind the railings for the Playroom scene, while LightGaussian [16] and Mini-Splatting [17] lose this object. Furthermore, GaussianSpa fully recovers the original details of the zoom-in object in the Counter scene, while other methods outcome blurry surfaces.

We analyze the proposed GaussianSpa by visualizing the rendered Gaussian ellipsoids and final point clouds in Figure 6 and Figure 8, respectively. Those visualized results demonstrate that GaussianSpa can adaptively control the density to match the details when sparsifying the Gaussians, ensuring high-quality sparse 3D representation. From Figure 6 and Figure 12, we can observe that GaussianSpa uses fewer but larger Gaussians to represent the blue sky and uses dense Gaussians to render complex carpet textures, providing precise rendering of details in the zoom-in images. In contrast, Mini-Splatting [17] cannot fully present the details, leading to blurry rendered textures. Figure 8 further demonstrates the superiority of GaussianSpa in simplification over existing methods. GaussianSpa adaptively reduce the Gaussians in the low-frequency areas and preserve more Gaussians in high-frequency areas that generally need denser Gaussians to synthesize, resulting in sparse Gaussian point clouds that precisely outline the scene contours.

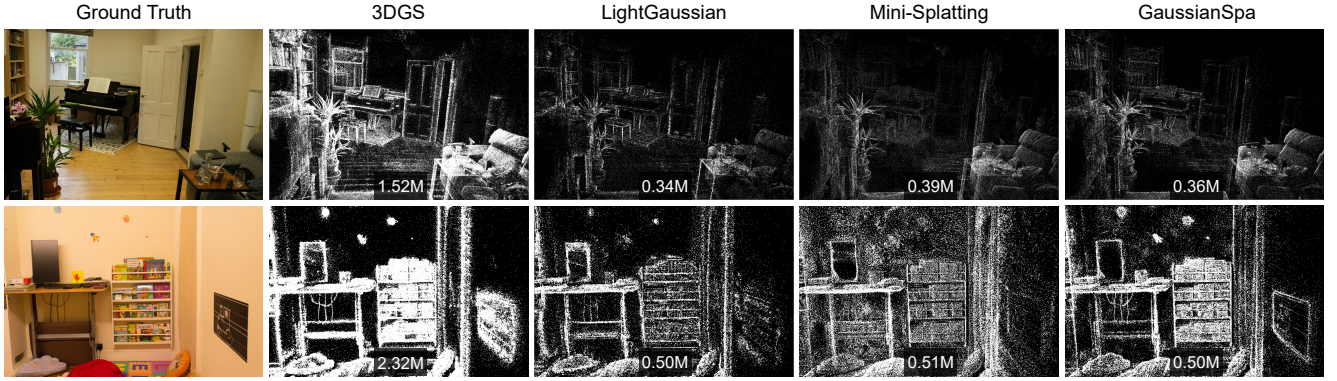


Figure 8. **Visualized point clouds for the Room and DrJohnson scenes.** With GaussianSpa, sparse Gaussians concentrate on high-frequency areas after our “optimizing-sparsifying”-integrated training process, ensuring the capability to capture detail-rich textures and shapes. On the contrary, the point clouds produced by other approaches cannot outline the contours.

5. Discussion

5.1. Generality of GaussianSpa

The proposed GaussianSpa is a general Gaussian simplification framework. Existing importance criteria can be applied to project the auxiliary variable z onto the sparse space in the “sparsifying” step, Eq. 15, obtaining more compact 3DGS models with improved rendering performance. For example, we employ the criterion from LightGaussian [16] and plot the PSNR curves at multiple Gaussian reduction rates in Figure 14 in the Supplementary Material. It is observed that GaussianSpa consistently outperforms LightGaussian [16] with an average of 0.4 dB improvement for the Playroom and Kitchen scenes.

5.2. Storage Compression

While GaussianSpa focuses on simplifying and optimizing the number of Gaussians for a sparse 3DGS representation, existing compression techniques such as SH optimization and vector quantization can be applied as add-ons to the simplified 3DGS model. Table 2 (in Supplementary Material) shows the final storage cost for the Mip-NeRF 360 dataset by incorporating the SH distillation and vector quantization methods proposed by LightGaussian [16]. This comparison demonstrates that GaussianSpa can achieve the lowest storage consumption while maintaining superior rendering quality over existing hybrid compression methods, such as EfficientGS [28] and LightGaussian [16].

5.3. Relation to Other Optimization Algorithms

The “optimizing-sparsifying” solution in our proposed GaussianSpa falls under optimization methods that project a target variable onto a constrained space, specifically the ℓ_p ball (where $p = 0$ in our case). This approach is similar to foundational optimization algorithms in compressed sensing [13] and sparse learning [6], such as Ridge regres-

sion [21] with the ℓ_2 regularization and Lasso regression [45] with the ℓ_1 regularization. The regularization acts as a penalty, combined with quadratic loss, reformulating the problem as a convex optimization task.

However, genuine sparsity requires ℓ_0 regularization, which leads to a nonconvex, NP-hard problem [12, 34, 38, 46]. Lasso (where $p = 1$) provides the tightest convex approximation but sacrifices exact sparsity [8]. Greedy algorithms like Matching Pursuit [30, 42] offer another approach, though they lack guaranteed global optima.

Iterative thresholding (IT) [11, 14, 15] offers a different solution, using matrix multiplications and scalar shrinkage steps to yield a simple yet effective structure. Furthermore, a closed-form solution for the global minimizer can be obtained in cases with a unitary matrix [14], even for nonconvex scenarios. Inspired by this, we introduce a duplicated variable (i.e., z) and project it onto the ℓ_0 ball, optimizing over two variables alternately, similar to the alternating direction method of multipliers [5]. Leveraging the unitary case’s simplicity, the “sparsifying” step aligns with this scenario and thus admits an explicit solution. Thus, our algorithm benefits from ℓ_0 -induced sparsity while achieving fast convergence guided by the explicit solution.

6. Conclusion

In this paper, we present an optimization-based simplification framework, GaussianSpa, for compact and high-quality 3DGS. GaussianSpa formulates the simplification objective as a constrained optimization problem with a sparsity constraint on the Gaussian opacities. Then, we propose an “optimizing-sparsifying” solution to efficiently solve the formulated problem during the 3DGS training process. Our comprehensive evaluations on multiple datasets with quantitative results and qualitative analyses demonstrate the superiority of GaussianSpa in rendering quality with fewer Gaussians compared to the state of the arts.

References

- [1] Shai Avidan and Amnon Shashua. Novel view synthesis in tensor space. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 1034–1040. IEEE, 1997. 2
- [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. 2
- [3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 6, 7
- [4] Stephen Boyd and Lieven Vandenbergh. *Convex optimization*. Cambridge university press, 2004. 5
- [5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. 6, 8
- [6] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009. 8
- [7] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024. 3
- [8] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001. 8
- [9] Yihang Chen, Qianyi Wu, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. Hac: Hash-grid assisted context for 3d gaussian splatting compression. In *European Conference on Computer Vision*, pages 422–438. Springer, 2025. 3
- [10] 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. 3
- [11] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004. 8
- [12] Geoffrey Mark Davis. *Adaptive nonlinear approximations*. New York University, 1994. 8
- [13] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006. 8
- [14] Michael Elad, Boaz Matalon, Joseph Shtok, and Michael Zibulevsky. A wide-angle view at iterated shrinkage algorithms. In *Wavelets XII*, pages 15–33. SPIE, 2007. 8
- [15] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge university press, 2012. 8
- [16] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *arXiv preprint arXiv:2311.17245*, 2023. 2, 3, 7, 8, 1, 4
- [17] Guangchi Fang and Bing Wang. Mini-splatting: Representing scenes with a constrained number of gaussians. *arXiv preprint arXiv:2403.14166*, 2024. 1, 2, 3, 6, 7, 4
- [18] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022. 1
- [19] Sharath Girish, Kamal Gupta, and Abhinav Shrivastava. Eagles: Efficient accelerated 3d gaussians with lightweight encodings. *arXiv preprint arXiv:2312.04564*, 2023. 3, 6, 7, 1
- [20] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018. 2, 6, 7
- [21] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 8
- [22] 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. 1, 2, 3, 4, 6, 7
- [23] Sieun Kim, Kyungjin Lee, and Youngki Lee. Color-cued efficient densification method for 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 775–783, 2024. 3
- [24] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 2, 6
- [25] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. Point-based neural rendering with per-view optimization. In *Computer Graphics Forum*, pages 29–43. Wiley Online Library, 2021. 2
- [26] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21719–21728, 2024. 2, 3, 6, 7
- [27] Rong Liu, Rui Xu, Yue Hu, Meida Chen, and Andrew Feng. Atomgs: Atomizing gaussian splatting for high-fidelity radiance field. *arXiv preprint arXiv:2405.12369*, 2024. 3
- [28] Wenkai Liu, Tao Guan, Bin Zhu, Lili Ju, Zikai Song, Dan Li, Yuesong Wang, and Wei Yang. Efficientgs: Streamlining gaussian splatting for large-scale high-resolution scene representation. *arXiv preprint arXiv:2404.12777*, 2024. 2, 3, 8, 1
- [29] 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. 3
- [30] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993. 8
- [31] Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Francisco Vicente Carrasco, Markus Steinberger, and Fer-

- nando De La Torre. Taming 3dgs: High-quality radiance fields with limited resources. *arXiv preprint arXiv:2406.15643*, 2024. 3, 6, 7
- [32] 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. 1, 2, 3
- [33] Wieland Morgenstern, Florian Barthel, Anna Hilsmann, and Peter Eisert. Compact 3d scene representation via self-organizing gaussian grids. *arXiv preprint arXiv:2312.13299*, 2023. 3
- [34] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995. 8
- [35] KL Navaneet, Kossar Pourahmadi Meibodi, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Compgs: Smaller and faster gaussian splatting with vector quantization. In *European Conference on Computer Vision*, 2024. 3, 6, 7
- [36] Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10349–10358, 2024. 3
- [37] Michael Niemeyer, Fabian Manhardt, Marie-Julie Rakotosaona, Michael Oechsle, Daniel Duckworth, Rama Gosula, Keisuke Tateno, John Bates, Dominik Kaeser, and Federico Tombari. Radsplat: Radiance field-informed gaussian splatting for robust real-time rendering with 900+ fps. *arXiv preprint arXiv:2403.13806*, 2024. 2, 3
- [38] Yigit Oktar and Mehmet Turkan. A review of sparsity-based clustering methods. *Signal processing*, 148:20–30, 2018. 8
- [39] Panagiotis Papantonakis, Georgios Kopanas, Bernhard Kerbl, Alexandre Lanvin, and George Drettakis. Reducing the memory footprint of 3d gaussian splatting. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 7(1):1–17, 2024. 3
- [40] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014. 5
- [41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5
- [42] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993. 8
- [43] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *arXiv preprint arXiv:2403.17898*, 2024. 3
- [44] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 4
- [45] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996. 8
- [46] Andreas M Tillmann. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letters*, 22(1):45–49, 2014. 8
- [47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3
- [48] Henan Wang, Hanxin Zhu, Tianyu He, Runsen Feng, Jiajun Deng, Jiang Bian, and Zhibo Chen. End-to-end rate-distortion optimized 3d gaussian representation. *arXiv preprint arXiv:2406.01597*, 2024. 3
- [49] Shuzhao Xie, Weixiang Zhang, Chen Tang, Yunpeng Bai, Rongwei Lu, Shijia Ge, and Zhi Wang. Mesongs: Post-training compression of 3d gaussians via efficient attribute transformation. *arXiv preprint arXiv:2409.09756*, 2024. 3
- [50] Runyi Yang, Zhenxin Zhu, Zhou Jiang, Baijun Ye, Xiaoxue Chen, Yifei Zhang, Yuantao Chen, Jian Zhao, and Hao Zhao. Spectrally pruned gaussian fields with neural compensation. *arXiv preprint arXiv:2405.00676*, 2024. 2
- [51] Zhaoliang Zhang, Tianchen Song, Yongjae Lee, Li Yang, Cheng Peng, Rama Chellappa, and Deliang Fan. Lp-3dgs: Learning to prune 3d gaussian splatting. *arXiv preprint arXiv:2405.18784*, 2024. 2, 3, 6, 7
- [52] M. Zwicker, H. Pfister, J. van Baar, and M. Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS '01.*, pages 29–538, 2001. 4