

# Compressing 3D Gaussian Splatting by Noise-Substituted Vector Quantization

Haishan Wang<sup>1</sup>, Mohammad Hassan Vali<sup>1</sup>, and Arno Solin<sup>1</sup>

Department of Computer Science, Aalto University, Espoo, Finland  
`{haishan.wang, mohammad.vali, arno.solin}@aalto.fi`

**Abstract.** 3D Gaussian Splatting (3DGS) has demonstrated remarkable effectiveness in 3D reconstruction, achieving high-quality results with real-time radiance field rendering. However, a key challenge is the substantial storage cost: reconstructing a single scene typically requires millions of Gaussian splats, each represented by 59 floating-point parameters, resulting in approximately 1 GB of memory. To address this challenge, we propose a compression method by building separate attribute codebooks and storing only discrete code indices. Specifically, we employ noise-substituted vector quantization technique to jointly train the codebooks and model features, ensuring consistency between gradient descent optimization and parameter discretization. Our method reduces the memory consumption efficiently (around 45×) while maintaining competitive reconstruction quality on standard 3D benchmark scenes. Experiments on different codebook sizes show the trade-off between compression ratio and image quality. Furthermore, the trained compressed model remains fully compatible with popular 3DGS viewers and enables faster rendering speed, making it well-suited for practical applications.

**Keywords:** Gaussian Splatting · Compression · Vector Quantization

## 1 Introduction

In computer graphics, 3D scene reconstruction has captured great attention from both academia and industry due to its wide range of applications. A key objective in this domain is novel view synthesis (NVS), which aims to generate novel images from new viewpoints based on a set of input images. Early approaches based on the multi-view stereo, such as structure-from-motion [20], provided robust and fundamental solutions to this task before the advent of deep learning. Neural radiance field (NeRF, [15]) introduced neural networks to map spatial features to optical information. The latest advancement in this field is 3D Gaussian splatting (3DGS, [10]), which represents 3D scenes using a set of differentiable Gaussian primitives, often called splats. This technique significantly expands the boundaries of the domain by enabling high-fidelity reconstruction alongside real-time rendering, even for complex scenes. Consequently, 3DGS has been applied to various fields, including autonomous driving [28], AI-generated content [26], Simultaneous Localization and Mapping (SLAM) [24,14], and so on.

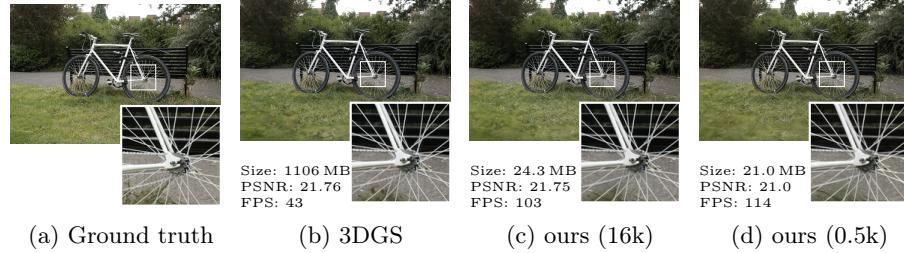


Fig. 1: We reduce the storage requirements by implementing an advanced VQ for 3DGS. It reduces file sizes and accelerates rendering speed, while maintaining high reconstruction quality. The reported frames per second (FPS) metrics were measured using an Nvidia RTX 4070 GPU.

While 3DGS offers structure simplicity, computational efficiency and continuous rendering through explicit scene modelling, its high memory consumption remains a main limitation to further applications. Typically, representing a single scene requires millions of Gaussian splats, each characterized by 59 floating-point attributes, leading to substantial memory usage (*e.g.*, approximately 1.1 GB for the BICYCLE scene in the Mip-NeRF360 dataset; see Fig. 1). Recent studies have revealed strong correlations between Gaussian attributes and a high dependency among Gaussian splats [1], indicating substantial information redundancy. These findings suggest the feasibility of employing compression techniques to reduce memory consumption with minimal impact on rendering performance.

Various 3DGS compression techniques have been proposed to reduce memory consumption while maintaining rendering quality. These approaches generally fall into two categories: *(i)* Machine learning (ML) methods, introduce hierarchical structures, predictive models, or neural networks to reduce redundancy. *(ii)* Signal processing (SP) methods, which apply vector quantization, pruning, and entropy coding to optimize memory usage. While ML-based methods offer strong compression, they introduce computational overhead due to their reliance on view-dependent neural networks and implicit representations, limiting applications requiring real-time rendering or explicit modelling. On the other hand, SP-based methods often struggle with optimization inconsistencies caused by the incompatibility between discretization and gradient descent. A key challenge remains: how to efficiently compress 3DGS while maintaining reconstruction quality, preserving GS advantages (real-time rendering speed and explicit scene modelling).

In this paper, we address this challenge by introducing **NSVQ-GS**, a Noise-Substituted Vector Quantization (NSVQ, [22]) method that ensures optimization consistency while achieving an optimal balance between high compression ratios and high-fidelity reconstruction. Instead of treating quantization as a hard selection process, NSVQ models the quantization error by adding a noise term to the input vector such that it retains the statistical properties of the original

error and thus enables direct optimization of the codebooks with gradient-based optimization.

Our main **contributions** are summarized as follows.

- **Compact 3DGS representation via NSVQ.** We introduce a discrete feature encoding method that maintains optimization consistency, avoid the clustering algorithms for code assignment, and achieve high compression ratios while preserving reconstruction quality across various bitrates.
- **Efficient compression with real-time rendering and compatibility.** Our approach reduces memory usage and enables faster rendering, while keeping full compatibility with all existing 3DGS applications, such as web-based 3D visualization, 3D editing, and robotic vision.
- **State-of-the-art performance.** We demonstrate state-of-the-art results on standard benchmarks in the category of signal processing (SP)-based GS compression, without reliance on any neural networks.

## 2 Background and Related Work

We provide the necessary background on 3DGS, focusing on its parameter structure and rendering process, then review the existing GS compression approaches.

### 2.1 3D Gaussian Splatting

The 3D scene is modelled by 3DGS as a set of Gaussian splats. Each Gaussian splat consists of 6 attributes: 3D spatial coordinates  $\mathbf{x} \in \mathbb{R}^3$ , opacity  $o \in \mathbb{R}$ , scaling and rotation parameters  $\mathbf{s} \in \mathbb{R}^3, \mathbf{r} \in \mathbb{R}^4$  which jointly represent the covariance matrix  $\Sigma = \mathbf{R}\mathbf{S}\mathbf{S}^\top\mathbf{R}^\top \in \mathbb{R}^{3 \times 3}$ , colours  $\mathbf{c} \in \mathbb{R}^3$  and spherical harmonics (SH) coefficients  $\mathbf{c}^{sh} \in \mathbb{R}^{45}$  of order 3 (the dimensions of SH depend on the order, 3<sup>rd</sup> order is the convention trade-off between performance and cost). The scaling and rotation matrices  $\mathbf{S}, \mathbf{R}$  are recovered by corresponding parameters  $\mathbf{s}, \mathbf{r}$ . The pixel-wise colour rendering of 3DGS keeps the image formation of pixel-based  $\alpha$ -blending and volumetric rendering in NeRF [10]. The pixel colour  $C$  is calculated by  $\alpha$ -blending:

$$C = \sum_{i=1}^{|\mathcal{N}|} \tilde{\mathbf{c}}_i \alpha_i \prod_{j=1}^{n-1} (1 - \alpha_j), \quad (1)$$

where  $\mathcal{N}$  refers to all Gaussians splats visible from the viewpoint of the current pixel, which are sorted by depth,  $\tilde{\mathbf{c}}_i$  denotes the colour recovered from colours  $\mathbf{c}_i$  and spherical harmonics  $\mathbf{c}_i^{sh}$ , and  $\alpha_i$  is the alpha blending term obtained by scaling the opacity by the Gaussian distribution

$$\alpha_i = o_i \exp \left( -\frac{1}{2} (\mathbf{x}' - \mu'_i) \Sigma_i'^{-1} (\mathbf{x}' - \mu'_i)^\top \right),$$

where  $\mathbf{x}', \mu'_i$  denote the projected coordinates of the pixel and the Gaussian splat. The covariance matrix after 2D-projection is  $\Sigma' = \mathbf{J}\mathbf{W}\Sigma\mathbf{W}^\top\mathbf{J}^\top$ , where  $\mathbf{J}, \mathbf{W}$  denote the Jacobian of the affine approximation of projection and viewing transformation.

## 2.2 Compression Approaches for 3DGS

3DGS has achieved remarkable success in the 3D reconstruction domain with a wide range of applications. However, the high storage costs limit its widespread adoption. Over the past few years, researchers have developed many methods to address this limitation, which can be classified into two fundamental strategies: **compaction**, which reduces the number of splats through adaptive density control (ADC) and improved heuristic, and **compression**, which optimizes the organization of attributes to minimize redundancy. This paper focuses on developing a GS compression method that is compatible with most existing compaction techniques.

Existing compression approaches can be categorized into two types: Signal processing-based (**SP-based**) and machine learning-based (**ML-based**).

**SP-based methods** often employ techniques such as vector quantization (VQ), which discretize high-dimensional continuous feature spaces into a set of representative codewords [6,18,19,17]. Besides VQ techniques, LightGaussians [6] adaptively distils SH parameters and improves ADC by removing Gaussian splats with minimal global significance of reconstruction. Compressed3D [18] introduces space-filling curves for efficient coordinates information storage. Reduced3DGS [19] estimates splat redundancy in a scale- and resolution-aware manner for ADC, selects SH bands adaptively, and suggests half-floating data representation. CompGS [17] utilizes periodic K-means clustering for codebook assignment and incorporates an opacity regularization to control splats amount.

Despite these advancements, the issue of *gradient collapse* [21] in VQ-based methods has not received sufficient attention. For example, Reduced3DGS avoids training on VQ, while LightGaussians and Compressed3D fix the codebook assignment during training. CompGS addresses this issue using a straight-through estimator (STE) which copies the gradients through VQ function. These limitations motivate us to propose NSVQ-GS.

**ML-based methods** leverage techniques from the machine learning domain, such as Self-organizing Maps [16,25] or hash grids [4]. Some methods utilize simple multilayer perceptrons (MLPs) as decoders for Gaussian attributes, particularly for colour-related features [8,12]. A prominent example of ML-based methods is Scaffold-GS [13], which introduces anchor points for the hierarchical structure of GS attributes. For each anchor, the model reconstructs a group of neighbouring splats by low-dimensional embeddings via several shared view-dependent decoders. This idea has inspired several follow-up works. For instance, HAC [4] designs an adaptive quantization module on anchor attribute values, and predicts anchor attributes by querying anchor coordinates in the hash grid. ContextGS [23] refines anchors reconstruction from coarse to fine granularity using autoregressive models with quantized hyperpriors.

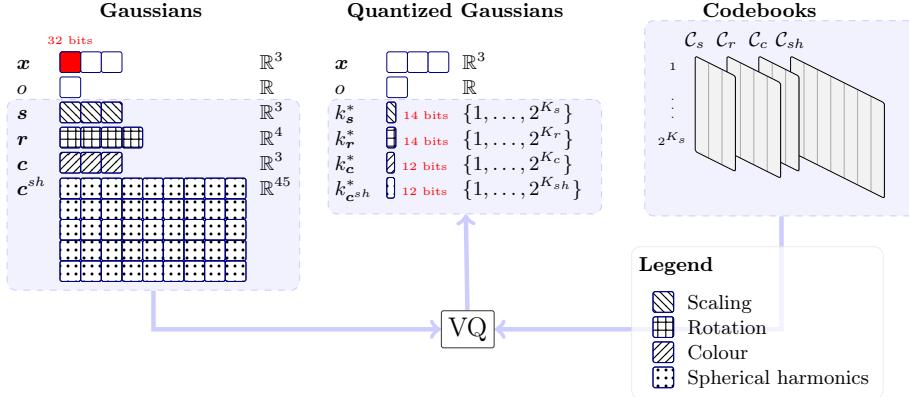


Fig. 2: Overview of the efficient reduction on storage requirement by our NSVQ-GS (16k). A single unit box represents 32 bits. Substituting Gaussian splats with their quantized counterparts and codebooks saves substantial memory consumption.

### 3 Methods

In this section, we explain the NSVQ technique [22] that optimizes VQ codebook by gradient-based optimizers and then, we present the training process of our proposed method NSVQ-GS.

#### 3.1 Noise Substitution in Vector Quantization

Vector quantization (VQ, [7]) is a classical signal processing technique that is used to compress a continuous data distribution with a limited discrete set of representative vectors called a codebook. Each codebook vector represents a subset of the data distribution, such that it is the closest codebook vector to all data samples in the subset. Given an input  $\mathbf{t} \in \mathbb{R}^{1 \times D}$  and a codebook  $\mathcal{C} = \{\mathbf{z}_k \in \mathbb{R}^{1 \times D} \mid k \in 1, \dots, 2^K\} \in \mathbb{R}^{2^K \times D}$  of bitrates  $K \in \mathbb{N}$ , the hard quantized input  $\mathbf{t}_q$  is computed as

$$\mathbf{t}_q = \mathbf{z}_{k^*}, \quad k^* = \arg \min_{k \in \{1, \dots, 2^K\}} \|\mathbf{t} - \mathbf{z}_k\|_2, \quad (2)$$

where  $k^*$  is the index of the closest code from  $\mathcal{C}$  to the input  $\mathbf{t}$  and  $\|\cdot\|_2$  refers to the Euclidean distance.

According to Eq. (2), the VQ is nondifferentiable and therefore cannot propagate gradients during the backward pass in neural network training. This issue, known as the *gradient collapse* problem [21], prevents effective learning. A common approach used to address this problem is the straight-through estimator (STE, [3]), which copies the gradients unchanged over the VQ function during backpropagation. Despite its simplicity, STE introduces several limitations: it incurs additional hyper-parameter tuning, modifies the optimization hyperplane

due to the inclusion of supplementary loss terms in the training objective, and fails to account for quantization effects during training.

Noise substitution in vector quantization (NSVQ, [22]) is another solution to *gradient collapse* that leads to faster convergence, more accurate gradients, and less hyper-parameter tuning than STE. NSVQ simulates the quantization by adding noise to the input vector such that the noise has the original quantization magnitude but in a random direction. NSVQ quantizes a given input vector  $\mathbf{t}$  as

$$\tilde{\mathbf{t}}_q = \mathbf{t} + \|\mathbf{t} - \mathbf{t}_q\|_2 \cdot \frac{\mathbf{e}}{\|\mathbf{e}\|_2}, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3)$$

where  $\mathbf{t}_q = \mathbf{z}_{k^*}$  is the hard quantized version of input (see Eq. (2)). Since  $\tilde{\mathbf{t}}_q$  is a differentiable function of input  $\mathbf{t}$  and selected codebook vector of  $\mathbf{z}_{k^*}$ , it can be used directly in end-to-end training of neural networks to backpropagate gradients through non-differentiable VQ function.

*Codebook collapse* [21, 5] is a common challenge in the training of VQ codebooks, where a subset of codebook vectors remain unused for quantization. As a result, these codebook vectors are not updated and remain inactive throughout training. To address this challenge, we adopt the *codebook replacement* procedure proposed in [22], *i.e.*, after a predefined number of training batches, inactive codebook vectors, those used less than a threshold are replaced with a permutation of a randomly selected set of actively used ones.

### 3.2 Proposed Method

The memory consumption of 3DGS arises mainly from the substantial amount of splats and associated attributes. For an efficient representation of them, we employ NSVQ to quantize four Gaussian attributes: colours, SH, scaling and rotation parameters. This approach strikes an optimal balance between compression efficiency (memory reduction) and minimal additional degradation to reconstruction quality. During model storage and rendering, the quantized features are modeled by the codebooks and the corresponding indices. For instance, with setting codebook bitrates as 10, the 45-dimensional SH features, originally stored as 32-bit floating-point values requiring 1,440 bytes, can be replaced by a single index requiring only 1.25 bytes. The detailed training process of proposed NSVQ-GS is described below.

**Training process** The entire process comprises four phases as illustrated in Fig. 3: the warm-up, pruning, vector quantization, and fine-tuning.

Consider a set of  $N$  Gaussians splats, denoted as  $\{G_i\}_{i=1}^N$ . Each splat  $G$  contains features  $G = (\mathbf{x}, \mathbf{o}, \mathbf{s}, \mathbf{r}, \mathbf{c}, \mathbf{c}^{sh})$  as described in Section 2.1. The training process begins with a **warm-up phase**, spanning the first 15k iterations. During this phase, the training procedure aligns precisely with the standard 3DGS [10], including the adaptive density control for splat densification and pruning.

The **pruning phase** occurs between 15k and 20k iterations, during which Gaussian splats are further pruned using opacity regularization, as introduced in CompGS [17]. In this phase, the training objective incorporates an additional

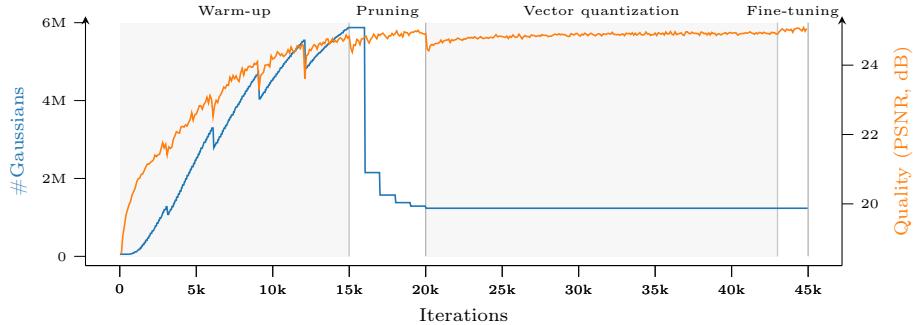


Fig. 3: Overview of the training process consisting of four phases. During **warm-up**, the model learns the 3D information by increasing the number of Gaussian splats. The **pruning** stage reduces the number of Gaussians while maintaining reconstruction performance. Density control is applied only until the end of the pruning phase. In the **vector quantization** phase, reconstruction quality initially degrades but recovers after sufficient training. Finally, the **fine-tuning** removes constraints imposed by noise substitution, further refining the final results.

regularization loss term, defined as  $\mathcal{L}_{\text{opacity}} = \sum_{i=1}^N o_i$ . Splats with low opacity are subsequently removed to enhance the overall compaction.

The **vector quantization phase** begins after the initial 20k iterations. In this work, we quantize all parameters except for the coordinates  $\mathbf{x}$  and opacity  $o$ , as quantizing these parameters would result in obvious quality degradation. We construct codebooks  $\mathcal{C}_s \in \mathbb{R}^{2^{K_s} \times 3}, \mathcal{C}_r \in \mathbb{R}^{2^{K_r} \times 4}, \mathcal{C}_c \in \mathbb{R}^{2^{K_c} \times 3}, \mathcal{C}_{sh} \in \mathbb{R}^{2^{K_{sh}} \times 45}$  for four attributes associated with the covariance matrix and colours, where  $K_s, K_r, K_c, K_{sh} \in \mathbb{N}$  denote the respective codebook bitrates. The codebook entries are initialized as the centroids of clustered feature distributions by K-means. The quantized Gaussians splats are then represented as:  $\tilde{G}_q = (\mathbf{x}, o, \tilde{\mathbf{s}}_q, \tilde{\mathbf{r}}_q, \tilde{\mathbf{c}}_q, \tilde{\mathbf{c}}_q^{sh})$ , where  $(\tilde{\mathbf{s}}_q, \tilde{\mathbf{r}}_q, \tilde{\mathbf{c}}_q, \tilde{\mathbf{c}}_q^{sh})$  are quantized features obtained from codebooks  $\mathcal{C}_s, \mathcal{C}_r, \mathcal{C}_c, \mathcal{C}_{sh}$  using Eq. (3). During this phase, the model utilises  $\tilde{G}_q$  for  $\alpha$ -blending as described in Eq. (1). Both model parameters and all codebooks are trained jointly, with periodic replacement of codebook entries to increase the codebook entries usage.

In the final 2k iterations, the model undergoes a **fine-tuning phase** with frozen code assignment, which means the code indices  $k_{\mathbf{x}}^*$  in Eq. (2) are fixed, without the need for a nearest code search. Meanwhile, the noise-substitution in quantized features is skipped, meaning the model is trained with hard-quantized features  $G_q = (\mathbf{x}, o, \mathbf{s}_q, \mathbf{r}_q, \mathbf{c}_q, \mathbf{c}_q^{sh})$ .

In the end, the model is stored as  $(\{\hat{G}_i\}_{i=1}^N, \{\mathcal{C}_i\}_{i \in \{s, r, c, sh\}})$ , where the quantized attributes are replaced by their corresponding codebook entries:  $\hat{G} = (\mathbf{x}, o, k_{\mathbf{s}}^*, k_{\mathbf{r}}^*, k_{\mathbf{c}}^*, k_{\mathbf{c}^{sh}}^*)$  which represent the final quantized Gaussian splats. Each index, ranging from 1 to  $2^K$ , requires  $K$  bits. Consequently, all indices are stored in a compact bitstream format within a binary file.

Table 1: Benchmark comparison with baseline methods. Values are bold as the best results among all SP-based methods. (Baselines are collected from the benchmark [1].) The model size unit is MB.

Methods	Mip-NeRF 360				Tanks and Temples				Deep Blending				
	PSNR↑	SSIM↑	LPIPS↓	Size↓	PSNR↑	SSIM↑	LPIPS↓	Size↓	PSNR↑	SSIM↑	LPIPS↓	Size↓	
SP-based	<b>NSVQ-GS (16k) (ours)</b>	<b>27.28</b>	0.807	0.239	<b>16.38</b>	<b>23.62</b>	<b>0.842</b>	0.190	<b>11.02</b>	<b>29.90</b>	<b>0.906</b>	<b>0.249</b>	<b>11.49</b>
	CompGS (16k)	27.03	0.804	0.243	18	23.39	0.836	0.200	12	<b>29.90</b>	<b>0.906</b>	0.252	12
	Reduced3DGS	27.10	<b>0.809</b>	<b>0.226</b>	29	23.57	0.840	<b>0.188</b>	14	29.63	0.902	<b>0.249</b>	18
	Compact3DGS	27.08	0.798	0.247	48.8	23.32	0.831	0.201	39.4	29.79	0.901	0.258	43.2
	LightGaussians	<b>27.28</b>	0.805	0.243	42	23.11	0.817	0.231	22	—	—	—	—
	Compressed3D	26.98	0.801	0.238	28.8	23.32	0.832	0.194	17.28	29.38	0.898	0.253	25.3
ML-based	HAC (lowrate)	27.53	0.807	0.238	15.26	24.04	0.846	0.187	8.1	29.98	0.902	0.269	4.35
	SOG	27.08	0.799	0.230	38.42	23.56	0.837	0.186	21.72	29.26	0.894	0.268	16.92
	ContextGS (lowrate)	27.62	0.808	0.237	12.68	24.12	0.849	0.186	9.443	30.09	0.907	0.265	3.485
3DGS	27.21	0.815	0.214	734	23.14	0.841	0.183	411	29.41	0.903	0.243	676	

## 4 Experiments

We evaluate NSVQ-GS in 3DGS compression, aiming to demonstrate two key aspects: compression efficiency—how well we reduce storage costs while maintaining model fidelity, and rendering performance—how the compressed models perform in reconstruction. We follow the benchmarking protocols established in 3DGS.zip [1], with main comparisons to the closest prior work, CompGS [17].

### 4.1 Settings

**Data sets** We evaluate our method for real-world 3D scene reconstruction tasks on the standard benchmark data sets, following the conventions established in 3DGS [10]. The benchmark consists of three data sets: Mip-NeRF360 [2] (9 scenes), Tanks & Temples [11] (2 scenes), and Deep Blending [9] (2 scenes). These data sets cover a diverse range of real-world scenarios, including both unbounded outdoor environments and complex indoor settings. Train and test data split adheres to the methodology suggested by Mip-NeRF360 [2], where the test set comprises every 8<sup>th</sup> image (*i.e.*, images with indices satisfying  $i \bmod 8 \equiv 0$ ), while the remaining images are allocated as the training set.

**Implementation** Our training process consists of four phases: (1) warm-up phase for the first 15k iterations, (2) pruning phase during 15k–20k iterations, (3) vector quantization phase where Gaussian splats are trained jointly with codebooks during 20k–43k iterations, and (4) the fine-tuning phase with frozen codebook assignment during the 43k–45k iterations. In all experiments, we keep the convention established in CompGS [17], by setting the bitrates of codebook  $K_s = K_r = 4K_c = 4K_{sh}$ . The codebook bitrates notation follows a power-of-two scaling, where ‘16k’ means  $K_s = 2^{14} = 16384$ , ‘8k’ means  $K_s = 2^{13} = 8192$ , and so on. In practice, all experiments are conducted on a computing server equipped with a Nvidia A100 GPU and 32 GB memory. Our programming implementation is based on the pytorch-1.12.0 package and several submodules from the 3DGS codebase [10]. On average, training one scene with ‘4k’ bitrate settings requires around 100 minutes.

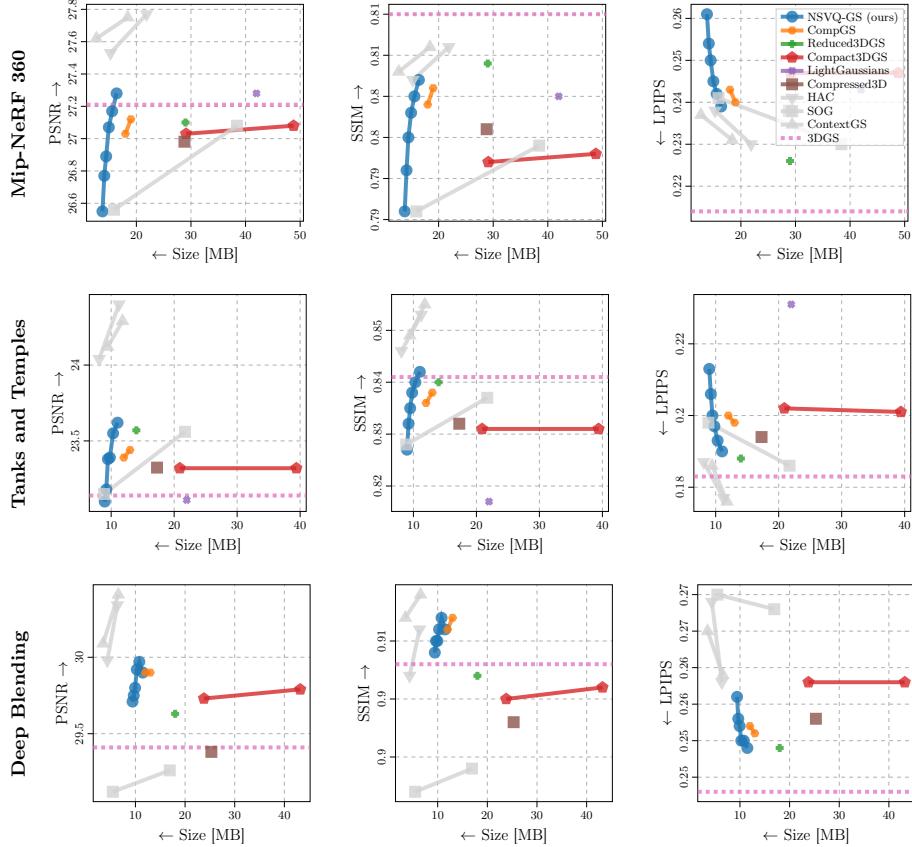


Fig. 4: The comparison of reconstruction quality across baselines. In each sub-figure, the  $x$ -axis denotes the model size in Megabytes,  $y$ -axis denotes the metrics on the reconstruction result. Each method comprising different sub-methods is plotted as multiple connected points. Our **NSVQ-GS** performs best within the category of SP-based methods, whereas ML-based methods (in gray) can boost performance further while losing some of the appealing 3DGS properties.

**Metrics** To evaluate the performance of NVS, we utilize widely recognized metrics. The Peak Signal-to-Noise Ratio (PSNR) quantifies the ratio between the maximum possible power of signals (*i.e.*, the ground truth image) and the power of corrupting noise. The Structural Similarity Index Measure (SSIM) evaluates perceptive quality by accounting for luminance, contrast, and structure degradation in synthetic images. The Learned Perceptual Image Patch Similarity (LPIPS) computes image similarity using a pre-defined NN designed to align with human perception [27]. Beyond image-based reconstruction quality, we also report model size (in megabytes) as a metric to evaluate the compression efficiency, reflecting the representation compactness.

**Baselines** We compare our methods with baseline methods, including original 3DGS (30k) [10], five SP-based methods: CompGS [17], Reduced3DGS [19], Compact3DGS [12], LightGaussians [6] and Compressed3D [18], and three additional ML-based methods: HAC [4], SOG [16], and ContextGS [23].

## 4.2 Results

**Quantitative results** We evaluate the reconstruction performance using four quantitative metrics summarized in [Table 1](#) and visualized in [Fig. 4](#).

In [Table 1](#), the methods are categorized into three clusters: SP-based methods, ML-based methods and original 3DGS (30k). The results demonstrate that our NSVQ-GS (16k) reaches an optimal balance between compression efficiency and reconstruction quality across all SP-based baselines. Compared to 3DGS (30k), our NSVQ-GS attains higher PSNRs across all data sets while utilizing only 2.2% of the storage consumption on average.

However, it is important to note that the table presents only one sub-method for each method, whereas each model may encompass multiple sub-methods, reflecting varying trade-offs between compression ratio and reconstruction quality. To facilitate a more comprehensive visualization, the scatter plots in [Fig. 4](#) include all sub-methods. Specifically, the sub-methods of our NSVQ-GS differ in codebook bitrates, ranging from ‘0.5k’ to ‘16k’. This comparison demonstrates that our model outperforms all other SP-based GS compression baselines, particularly the best-performing SP-based baseline, CompGS. It is observed that reducing bitrates leads to a degradation in reconstruction with decreasing storage benefits, as the primary storage consumption is attributed to non-quantized features. Therefore, to achieve a better compression model, it is advisable to retain relative high codebook bitrates and focusing on optimizing non-quantized features. However, the SSIM and LPIPS of NSVQ-GS are generally worse than the PSNR compared to 3DGS on all datasets, possibly due to the lack of locality prior and global sense in our compression in our technique.

**Qualitative results** A qualitative comparison was conducted among ground truth, 3DGS, CompGS (16k) and our NSVQ-GS (16k), as illustrated in [Fig. 6](#). Despite utilizing only approximately 2.2% of the memory, NSVQ-GS efficiently reconstructs scenes with high visual fidelity as 3DGS. Both 3DGS, CompGS (16k), and ours (16k) exhibit limitations in capturing the fine details of ground in flower and treehill scenes. Another comparison, presented in [Fig. 5](#), underscores the advantages of NSVQ-GS over the best SP-based compression method, CompGS, at extremely low bitrates ‘0.5k’, where the parameters are set to  $K_s = K_r = 512, K_c = K_{sh} = 128$ . The inability of CompGS to accurately reconstruct sharp details at such low bitrates is likely due to the STE solution for gradient collapse, which merely copies the gradients during training while disregarding quantization effects. These comparisons highlight the performance of NSVQ-GS in maintaining reconstruction quality, even under tight compression constraints.

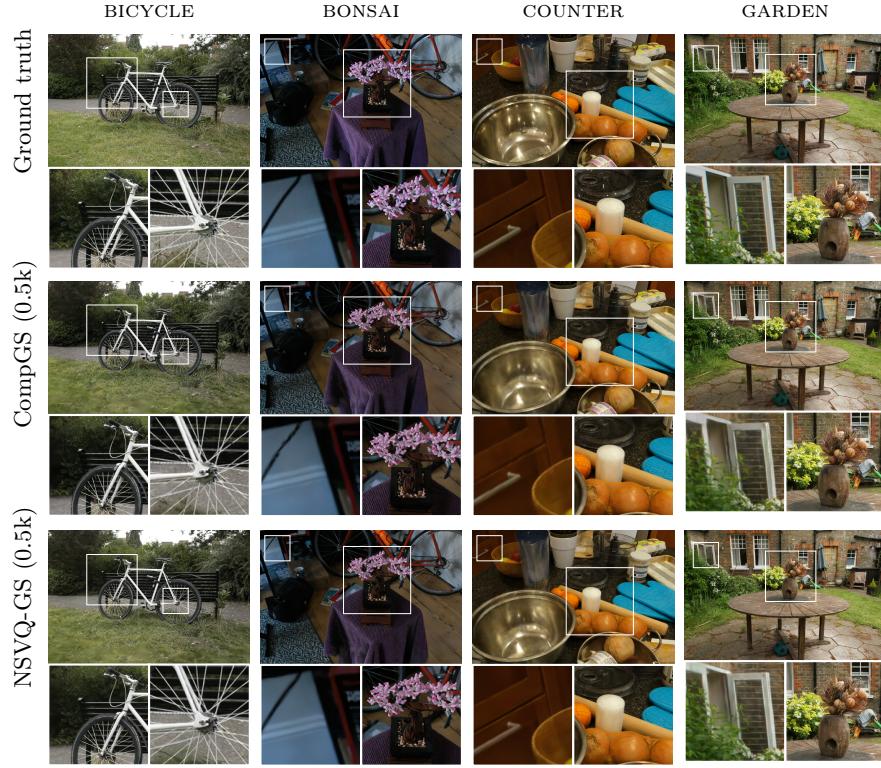


Fig. 5: Qualitative comparison between ground truth, CompGS (0.5k), and NSVQ-GS (0.5k) (ours). Our NSVQ-GS captures difficult sharp boundaries and straight lines better compared to CompGS (see, e.g., BICYCLE). This becomes clearer at stronger compression constraints (low codebook bitrates).

## 5 Conclusion and Discussion

In this paper, we proposed NSVQ-GS, a novel VQ-based model for GS compression. The introduced NSVQ-based technique addresses the challenge of *gradient collapse*, which arises from the inherent inconsistency between the discrete nature of quantization and gradient-descent optimization applied to Gaussian splat features. Our model achieves efficient compression of Gaussian splatting data while maintaining high reconstruction quality, as shown by both quantitative and qualitative evaluations. Furthermore, the streamlined storage structure enhances rendering speed, ensures compatibility with other compaction methods, and preserves the potential for broad industrial applications of 3DGS. It is worth stressing that while some ML-based methods achieve higher compression rates, the trained models lose appealing properties associated with 3DGS, e.g., real-time rendering and model editing capabilities which requires explicit modelling. Thus, we consider advancing SP-based GS compression methods to be an impactful direction for future research. While NSVQ-GS demonstrates

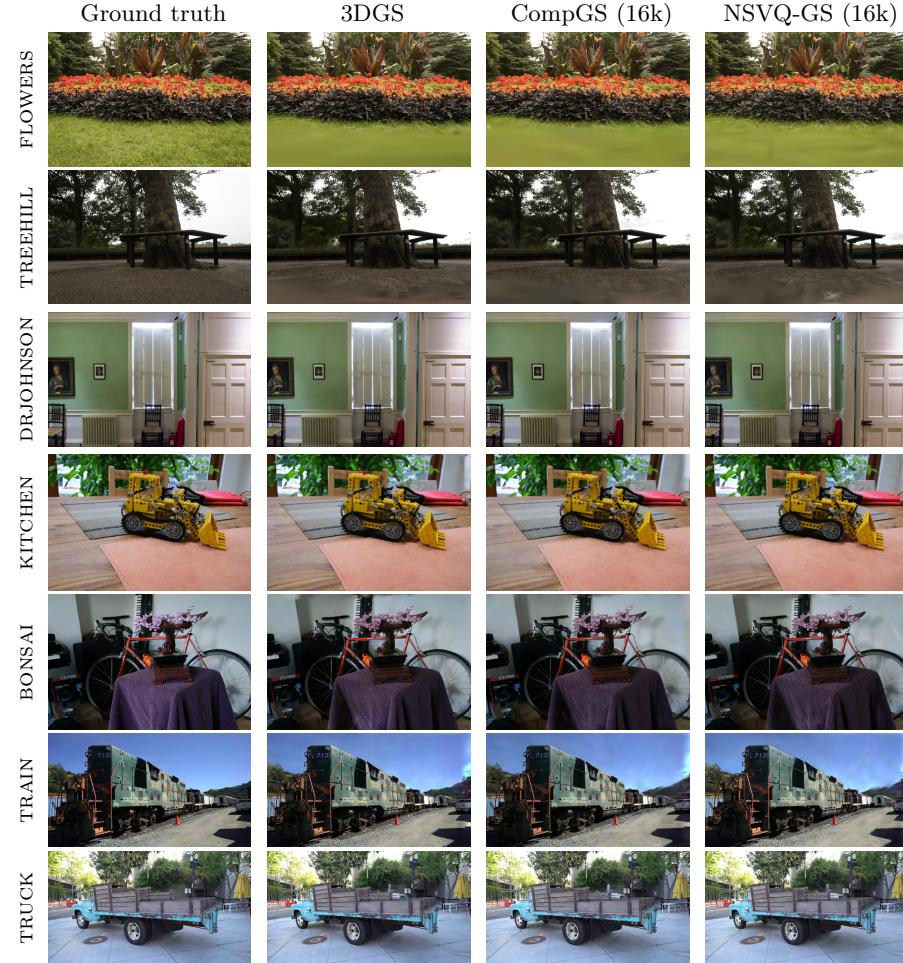


Fig. 6: Qualitative comparison between ground truth, 3DGS, CompGS (16k), and NSVQ-GS (16k) (ours).

advancement, there remains potentials to further improve the compression ratio. One promising direction is to compress unquantized Gaussian features, *e.g.*, quantizing spatial coordinates using space-filling curves. Additionally, the development of compaction models incorporating advanced heuristics could yield even greater compression efficiency.

A reference implementation of the methods is available at <https://github.com/AaltoML/NSVQGS>.

**Acknowledgments.** This work was supported by the Research Council of Finland (362408, 339730) and the Finnish Center for Artificial Intelligence FCAI. We acknowledge the computational resources provided by the Aalto Science-IT project and CSC – IT Center for Science, Finland.

## References

1. Bagdasarian, M.T., Knoll, P., Li, Y.H., Barthel, F., Hilsmann, A., Eisert, P., Morgenstern, W.: 3DGs.zip: A survey on 3D Gaussian splatting compression methods. arXiv preprint arXiv:2407.09510 (2024)
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5470–5479 (2022)
3. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
4. Chen, Y., Wu, Q., Lin, W., Harandi, M., Cai, J.: Hac: Hash-grid assisted context for 3D Gaussian splatting compression. In: European Conference on Computer Vision (ECCV). pp. 422–438. Springer (2025)
5. Dieleman, S., van den Oord, A., Simonyan, K.: The challenge of realistic music generation: modelling raw audio at scale. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31. Curran Associates, Inc. (2018)
6. Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., Wang, Z.: LightGaussian: Unbounded 3D Gaussian compression with 15x reduction and 200+ FPS. arXiv preprint arXiv:2311.17245 (2023)
7. Gersho, A., Gray, R.M.: Vector Quantization and Signal Compression. Springer (1992)
8. Girish, S., Gupta, K., Shrivastava, A.: Eagles: Efficient accelerated 3D Gaussians with lightweight encodings. In: European Conference on Computer Vision (ECCV). pp. 54–71. Springer (2024)
9. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G., Brostow, G.: Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics (ToG) **37**(6) (Dec 2018)
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (ToG) **42**(4), 139–1 (2023)
11. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4) (Jul 2017)
12. Lee, J.C., Rho, D., Sun, X., Ko, J.H., Park, E.: Compact 3D Gaussian representation for radiance field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21719–21728 (2024)
13. Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-gs: Structured 3D Gaussians for view-adaptive rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20654–20664 (2024)
14. Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian Splatting SLAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18039–18048 (2024)
15. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
16. Morgenstern, W., Barthel, F., Hilsmann, A., Eisert, P.: Compact 3D scene representation via self-organizing Gaussian grids. In: European Conference on Computer Vision (ECCV). pp. 18–34. Springer (2024)

17. Navaneet, K., Pourahmadi Meibodi, K., Abbasi Koohpayegani, S., Pirsavash, H.: CompGS: Smaller and faster Gaussian splatting with vector quantization. In: European Conference on Computer Vision (ECCV). pp. 330–349. Springer (2024)
18. Niedermayr, S., Stumpfegger, J., Westermann, R.: Compressed 3D Gaussian splatting for accelerated novel view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10349–10358 (2024)
19. Papantonakis, P., Kopanas, G., Kerbl, B., Lanvin, A., Drettakis, G.: Reducing the memory footprint of 3D Gaussian splatting. Proceedings of the ACM on Computer Graphics and Interactive Techniques **7**(1) (May 2024)
20. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016)
21. Vali, M.H.: Vector Quantization in Deep Neural Networks for Speech and Image Processing. Ph.D. thesis, Aalto University (2025)
22. Vali, M.H., Bäckström, T.: NSVQ: Noise substitution in vector quantization for machine learning. IEEE Access **10**, 13598–13610 (2022)
23. Wang, Y., Li, Z., Guo, L., Yang, W., Kot, A.C., Wen, B.: ContextGS: Compact 3D Gaussian splatting with anchor level context model. arXiv preprint arXiv:2405.20721 (2024)
24. Yan, C., Qu, D., Xu, D., Zhao, B., Wang, Z., Wang, D., Li, X.: GS-SLAM: Dense visual SLAM with 3D Gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19595–19604 (2024)
25. Ye, V., Li, R., Kerr, J., Turkulainen, M., Yi, B., Pan, Z., Seiskari, O., Ye, J., Hu, J., Tancik, M., et al.: gsplat: An open-source library for Gaussian splatting. arXiv preprint arXiv:2409.06765 (2024)
26. Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: GaussianDreamer: Fast generation from text to 3D Gaussians by bridging 2D and 3D diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6796–6807 (2024)
27. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
28. Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.H.: DrivingGaussian: Composite Gaussian splatting for surrounding dynamic autonomous driving scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21634–21643 (2024)