3DGS.zip: A survey on 3D Gaussian Splatting Compression Methods

Milena T. Bagdasarian[1], Paul Knoll[1], Florian Barthel[1,2], Anna Hilsmann[1], Peter Eisert[1,2], and Wieland Morgenstern[1]

[1]Fraunhofer Heinrich Hertz, HHI
[2]Humboldt University of Berlin

Abstract

We present a work-in-progress survey on 3D Gaussian Splatting [9] compression methods, focusing on their statistical performance across various benchmarks. This survey aims to facilitate comparability by summarizing key statistics of different compression approaches in a tabulated format. The datasets evaluated include TanksAndTemples [10], MipNeRF360 [1], DeepBlending [7], and SyntheticNeRF [15]. For each method, we report the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and the resultant size in megabytes (MB), as provided by the respective authors. This is an ongoing, open project, and we invite contributions from the research community as GitHub issues or pull requests. Please visit `http://w-m.github.io/3dgs-compression-survey/` for more information and a sortable version of the table.

1 Scope of this survey

In this survey, we focus on compression methods for 3D Gaussian Splatting (3DGS), aiming to optimize memory usage while preserving visual quality and real-time rendering speed. We provide a comprehensive comparison of various compression techniques, with quantitative results for the most commenly used datasets summarized in a tabulated format. Our goal is to ensure transparency and reproducibility of the included approaches. Additionally, we offer a brief explanation of each pipeline and discuss main compression approaches. Rather than covering all existing 3DGS methods, our focus is specifically on their compression techniques; for a broader overview of 3DGS methods and applications, we refer readers to [5, 23]. While we include many common approaches shared between neural radiance field (NeRF) [14] compression and 3DGS compression, we direct readers to [3, 12] for NeRF-specific compression methods.

Survey Table

Method	Rank	TanksAndTemples				MipNeRF360				DeepBlending				SyntheticNeRF			
		PSNR↑	SSIM↑	LPIPS↓	Size MB↓	PSNR↑	SSIM↑	LPIPS↓	Size MB↓	PSNR↑	SSIM↑	LPIPS↓	Size MB↓	PSNR↑	SSIM↑	LPIPS↓	Size MB↓
HAC-lowrate	4.6	24.04	0.846	0.187	8.1	27.53	0.807	0.238	15.3	29.98	0.902	0.269	4.3	33.24	0.967	0.037	1.2
HAC-highrate	4.6	24.40	0.853	0.177	11.2	27.77	0.811	0.230	21.9	30.34	0.906	0.258	6.3	33.71	0.968	0.034	1.9
gsplat-1.00M	5.2	24.03	0.857	0.163	15.4	27.29	0.811	0.229	15.3								
IGS-Low	5.7	23.70	0.836	0.227	8.4	27.33	0.809	0.257	12.5	30.63	0.904	0.293	6.3	33.36	0.971	0.036	1.8
IGS-High	5.9	24.05	0.849	0.210	12.5	27.62	0.819	0.247	25.4	32.33	0.924	0.253	7.7	34.18	0.975	0.032	2.7
Morgenstern et al. w/o SH	6.0	25.27	0.857	0.217	8.2	27.02	0.803	0.232	16.7	30.50	0.908	0.261	5.5	31.75	0.961	0.040	2.0
Morgenstern et al.	7.5	25.63	0.864	0.208	21.4	27.64	0.814	0.220	40.3	30.35	0.909	0.258	16.8	33.70	0.969	0.031	4.1
Navaneet et al. 32K	7.8	23.44	0.838	0.198	13.0	27.12	0.806	0.240	19.0	29.90	0.907	0.251	13.0				
Navaneet et al. 16K	8.0	23.39	0.836	0.200	12.0	27.03	0.804	0.243	18.0	29.90	0.906	0.252	12.0				
RDO-Gaussian	8.4	23.34	0.835	0.195	11.5	27.05	0.802	0.239	22.4	29.63	0.902	0.252	17.2	33.12	0.967	0.034	2.2
Reduced3DGS	8.9	23.57	0.84	0.188	14.0	27.10	0.809	0.226	29.0	29.63	0.902	0.249	18.0				
Compressed3D	10.7	23.32	0.832	0.194	17.3	26.98	0.801	0.238	28.8	29.38	0.898	0.253	25.3	32.94	0.967	0.033	3.7
Scaffold-GS	10.9	23.96	0.853	0.177	87.0	28.84	0.848	0.220	156.0	30.21	0.906	0.254	66.0				
Compact3DGS+PP	11.5	23.32	0.831	0.202	20.9	27.03	0.797	0.247	29.1	29.73	0.900	0.258	23.8	32.88	0.968	0.034	2.7
EAGLES	11.9	23.37	0.84	0.20	29.0	27.23	0.81	0.24	54.0	29.86	0.91	0.25	52.0				
Compact3DGS	12.9	23.32	0.831	0.201	39.4	27.08	0.798	0.247	48.8	29.79	0.901	0.258	43.2	33.33	0.968	0.034	5.5
LightGaussian	13.0	23.11	0.817	0.231	22.0	27.28	0.805	0.243	42.0					32.72	0.965	0.037	7.8
EAGLES-Small	13.7	23.10	0.82	0.22	19.0	26.94	0.80	0.25	47.0	29.92	0.90	0.25	33.0				

Note: The best methods in each category are highlighted (fist , second , third). The ranks represent the average rankings of the methods across all available datasets. The quality metrics PSNR, SSIM, and LPIPS are equally weighted with the model size, meaning they each contribute one-sixth to the ranks, while the size contributes half.

2 Datasets and Evaluation Statistics

2.1 Datasets

Performance and quality assessment of 3D Gaussian Splatting algorithms is typically performed on multiple datasets. These datasets provide 3D scenes or objects with various properties, such as varying levels of detail, lighting conditions, and complexities, which allow for comprehensive evaluation of the algorithms.
In our survey we include Tanks and Temples [10], MipNerf360 [1], Deep Blending [7] as real-world datasets, and Synthetic NeRF [15] as a synthetic dataset. From Tanks and Temples we include "truck" and "train" two unbounded outdoor scenes wich have a centered view point. The MipNerf360 dataset also has a centered view point but includes in- and outdoor scenes. The following scenes are included: "bicycle", "bonsai", "counter", "flowers", "garden", "kitchen", "room", "stump", "treehill". From the Deep Blending dataset we include "Dr Johnson" and "Palyroom" two indoor scenes with a viewpoint directed outward. The synthetic scenes: chair, drums, ficus, hotdog, lego, material, mic, ship stem from the SyntheticNeRF dataset.

2.2 Evaluation Statistics

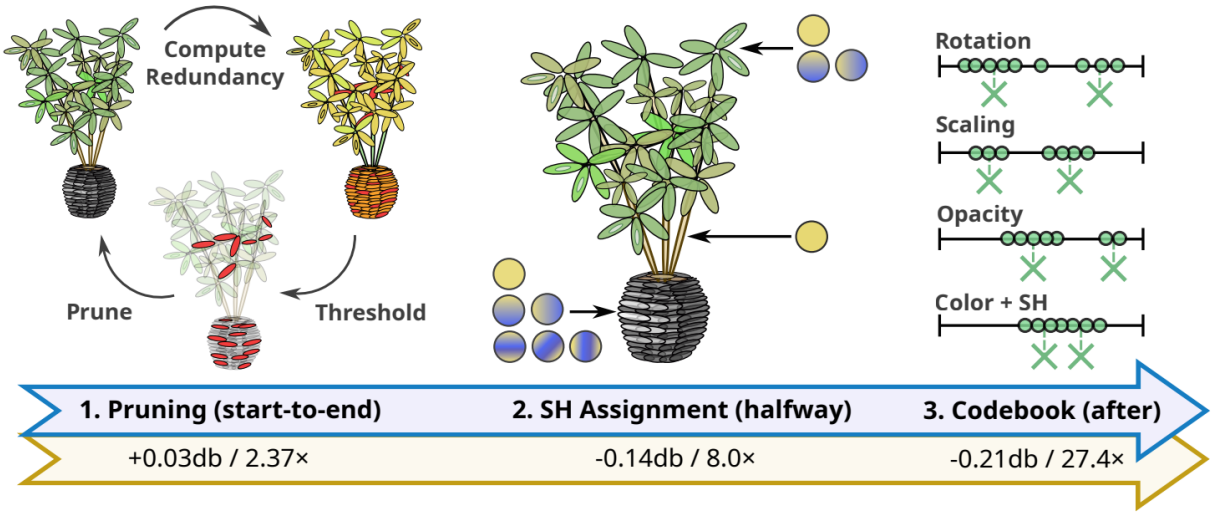
3 Description of Included Compression Approaches

In the following sections, we present a comprehensive overview of state-of-the-art 3D Gaussian Splatting (3DGS) compression methods. Although each method shares the common goal of minimizing memory usage while preserving rendering quality and speed, their strategies differ significantly. The summaries aim to distill the key concepts behind each approach, enabling better comparison and serving as a valuable reference for future research in the field of 3DGS compression.

3.1 Reducing the Memory Footprint of 3D Gaussian Splatting

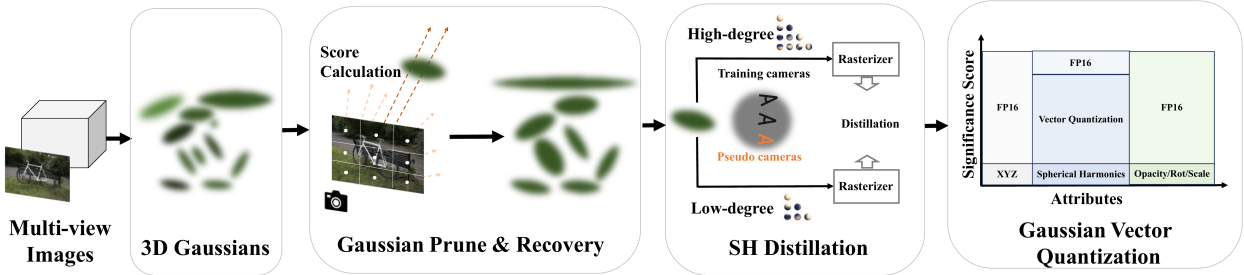
This approach addresses three main issues contributing to large storage sizes in 3D Gaussian Splatting (3DGS). To reduce the number of 3D Gaussian primitives, the authors introduce a scale- and resolution-aware redundant primitive removal method. This extends opacity-based pruning by incorporating a redundancy

score to identify regions with many low-impact primitives. To mitigate storage size due to spherical harmonic coefficients, they propose adaptive adjustment of spherical harmonic (SH) bands. This involves evaluating color consistency across views and reducing higher-order SH bands when view-dependent effects are minimal. Additionally, recognizing the limited need for high dynamic range and precision for most primitive attributes, they develop a codebook using K-means clustering and apply 16-bit half-float quantization to the remaining uncompressed floating point values. [19]



3.2 LightGaussian: Unbounded 3D Gaussian Compression with 15x Reduction and 200+ FPS

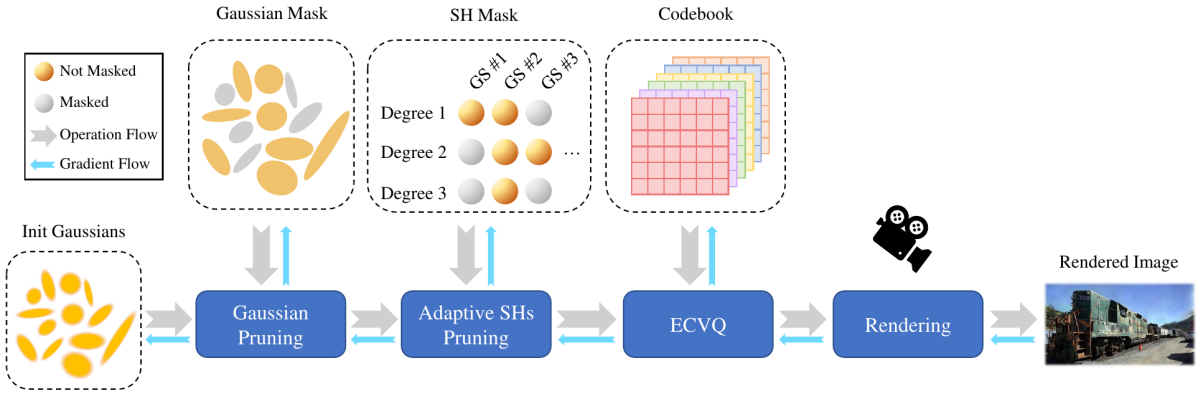
LightGaussian aims to transform 3D Gaussians to a more efficient and compact form, avoiding the scalablity issues that arrises from the large number of SfM (Structure from Motion) points for unbounded scenes. Inspired by Network Pruning, the method identifies Gaussians that minimally contribute to scene reconstruction and employs a pruning and recovery process, thereby efficiently reducing redundancy in Gaussian counts while maintaining visual effects. Additionally, LightGaussian utilizes knowledge distillation and pseudo-view augmentation to transfer spherical harmonics efficients to a lower degree. Furthermore, the authors propose a Gaussian Vector Quantization based on the global significance of Gaussians to quantize all redundant attributes, achieving lower bitwidth representations with minimal accuracy losses. [4]



3.3 End-to-End Rate-Distortion Optimized 3D Gaussian Representation

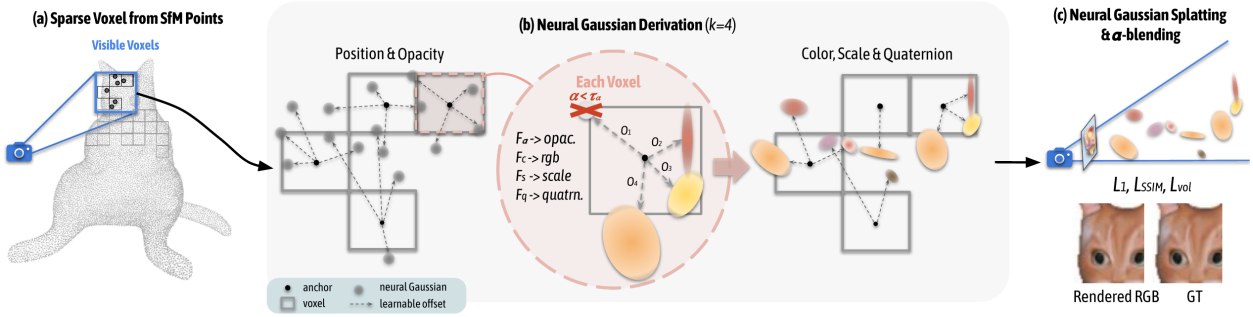
This paper introduces RDO-Gaussian, an end-to-end Rate-Distortion Optimized 3D Gaussian representation. The authors achieve flexible, continuous rate control by formulating 3D Gaussian representation learning as a joint optimization of rate and distortion. Rate-distortion optimization is realized through dynamic pruning and entropy-constrained vector quantization (ECVQ). Gaussian pruning involves learning a mask to eliminate redundant Gaussians and adaptive SHs pruning assigns varying SH degrees to each Gaussian

based on material and illumination needs. The covariance and color attributes are discretized through ECVQ, which performs vector quantization. [21]



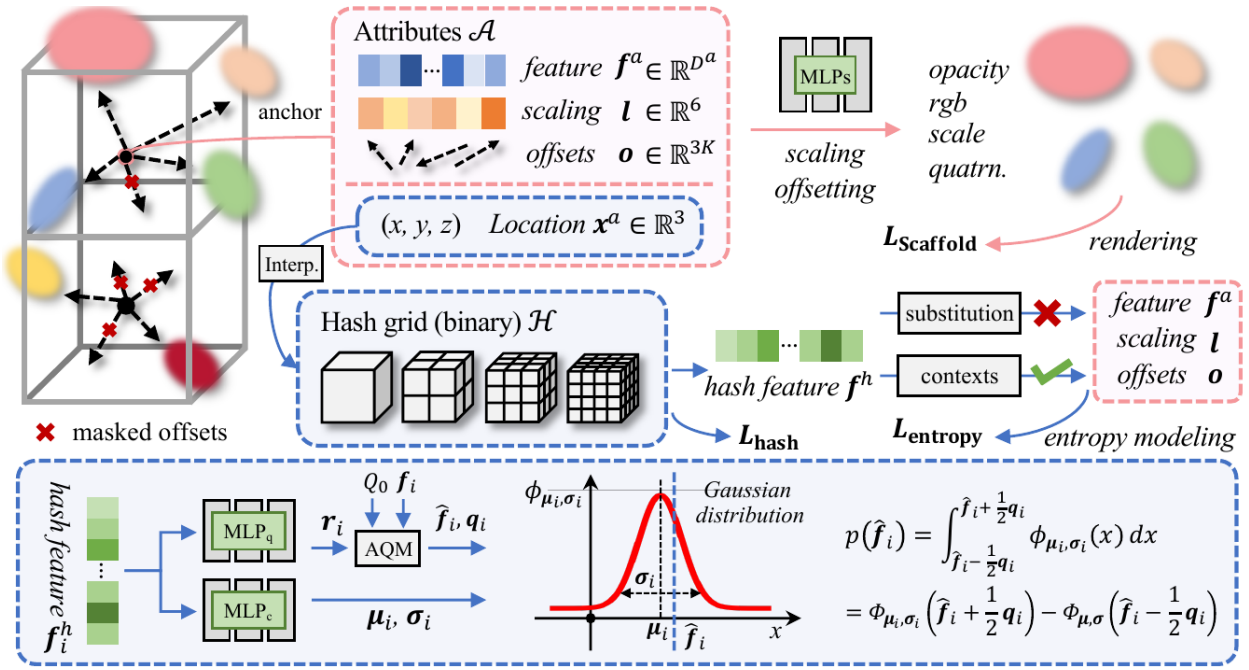
3.4 Scaffold-GS: Structured 3D Gaussians for View-Adaptive Rendering

Scaffold-GS introduces anchor points that leverage scene structure to guide the distribution of local 3D Gaussians. Attributes like opacity, color, rotation, and scale are dynamically predicted for Gaussians linked to each anchor within the viewing frustum, enabling adaptation to different viewing directions and distances. Initial anchor points are derived by voxelizing the sparse, irregular point cloud from Structure from Motion (SfM), forming a regular grid. To refine and grow the anchors, Gaussians are spatially quantized using voxels, with new anchors created at the centers of significant voxels, identified by their average gradient over N training steps. Random elimination and opacity-based pruning regulate anchor growth and refinement. [13]



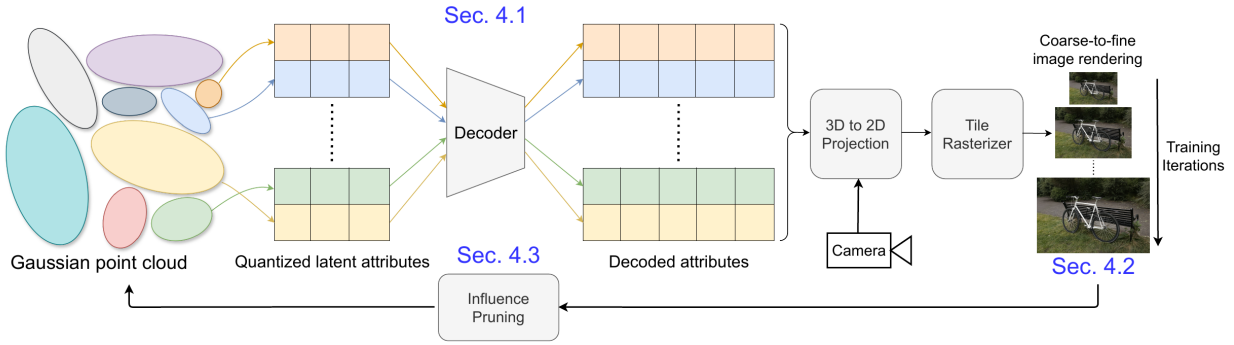
3.5 HAC: Hash-grid Assisted Context for 3D Gaussian Splatting Compression

The paper proposes a Hash-grid Assisted Context (HAC) framework for compressing 3D Gaussian Splatting (3DGS) models by leveraging the mutual information between attributes of unorganized 3D Gaussians (anchors) and hash grid features. Using Scaffold-GS as a base model, HAC queries the hash grid by anchor location to predict anchor attribute distributions for efficient entropy coding. The framework introduces an Adaptive Quantization Module (AQM) to dynamically adjust quantization step sizes. Furthermore, this method employs adaptive offset masking with learnable masks to eliminate invalid Gaussians and anchors, by leveraging the pruning strategy introduced by Compact3DGS and additionally removing anchors if all the attached offsets are pruned. [2]



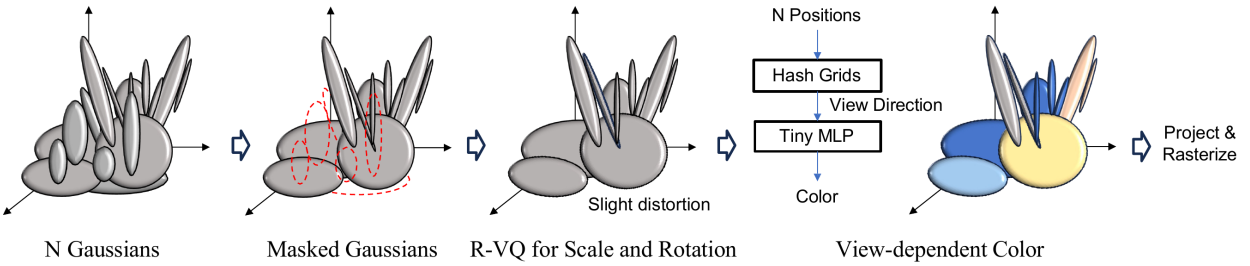
3.6 EAGLES: Efficient Accelerated 3D Gaussians with Lightweight EncodingS

The authors of this approach observed that in 3DGS, the color and rotation attributes account for over 80% of memory usage; thus, they propose compressing these attributes via a latent quantization framework. Additionally, they quantize the opacity coefficients of the Gaussians, improving optimization and resulting in fewer floaters or visual artifacts in novel view reconstructions. To reduce the number of redundant Gaussians resulting from frequent densification (via cloning and splitting), the approach employs a pruning stage to identify and remove Gaussians with minimal influence on the full reconstruction. For this, an influence metric is introduced, which considers both opacity and transmittance. [6]



3.7 Compact 3D Gaussian Representation for Radiance Field

This approach introduces a Gaussian volume mask to prune non-essential Gaussians and a compact attribute representation for both view-dependent color and geometric attributes. The volume-based masking strategy combines opacity and scale to selectively remove redundant Gaussians. For color attribute compression, spatial redundancy is exploited by incorporating a grid-based (Instant-NGP) neural field, allowing efficient representation of view-dependent colors without storing attributes per Gaussian. Given the limited variation in scale and rotation, geometric attribute compression employs a compact codebook-based representation to identify and reuse similar geometries across the scene. Additionally, the authors propose quantization and entropy-coding as post-processing steps for further compression. [11]



3.8 gsplat

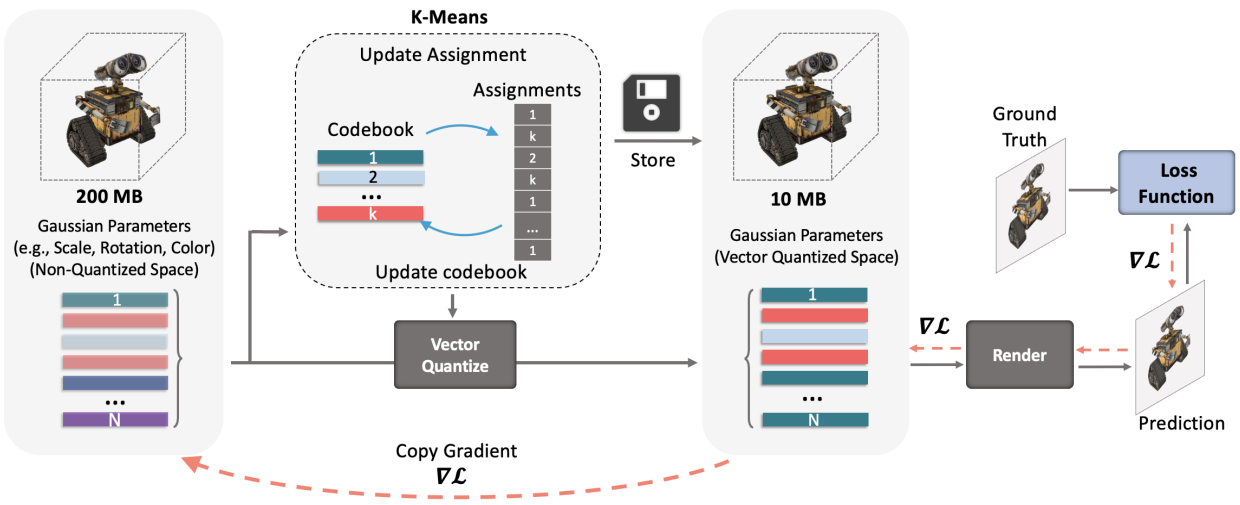
This approach combines 3D Gaussian Splatting as Markov Chain Monte Carlo (3DGS-MCMC) with compression techniques from the

Self-Organizing Gaussians

paper and Making Gaussian Splats more smaller. It is implemented in gsplat, an open-source library for CUDA-accelerated differentiable rasterization of 3D gaussians with Python bindings. The library is inspired by the SIGGRAPH paper "3D Gaussian Splatting for Real-Time Rendering of Radiance Fields", but gsplat is faster, more memory efficient, and with a growing list of new features. [8]

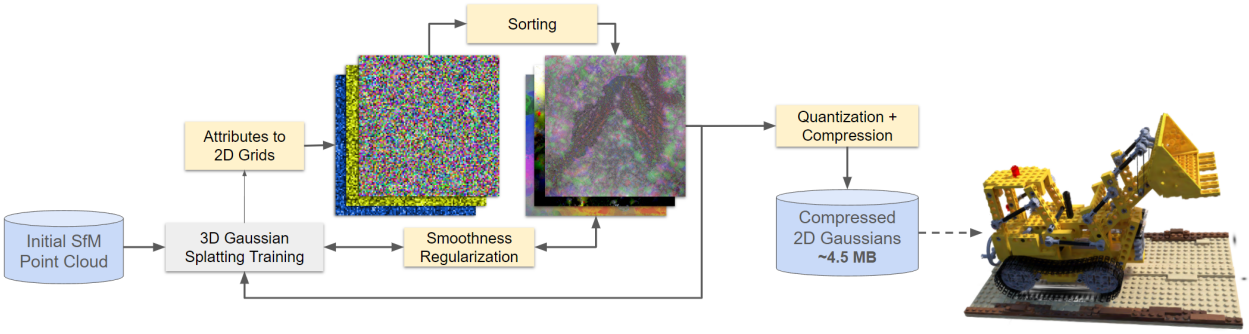
3.9 Compact3D: Compressing Gaussian Splat Radiance Field Models with Vector Quantization

This approach introduces a vector quantization method based on the K-means algorithm to quantize the Gaussian parameters in 3D Gaussian splatting, as many Gaussians may share similar parameters. Only a small codebook is stored along with the index of the code for each Gaussian, resulting in a large reduction in the storage of the learned radiance fields and a reduction of the memory footprint at rendering time. Additionally, the indices are further compressed by sorting the Gaussians based on one of the quantized parameters and storing the indices using a method similar to Run-Length-Encoding (RLE). To reduce the number of Gaussians, this method applies a regularizer to encourage zero opacity, before pruning Gaussians with opacity smaller than a threshold. [17]



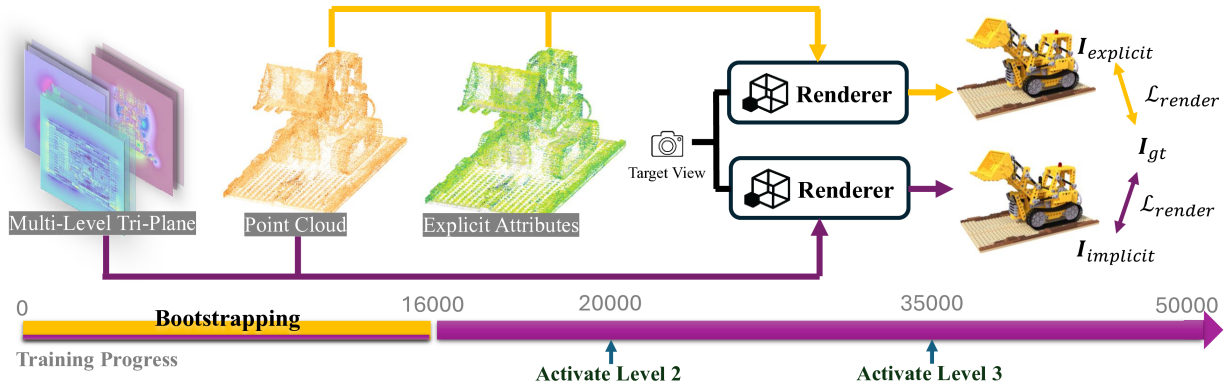
3.10 Compact 3D Scene Representation via Self-Organizing Gaussian Grids

Compressing 3D data is challenging, but many effective solutions exist for compressing 2D data (such as images). The authors propose a new method to organize 3DGS parameters into a 2D grid, drastically reducing storage requirements without compromising visual quality. This organization exploits perceptual redundancies in natural scenes. They introduce a highly parallel sorting algorithm, PLAS, which arranges Gaussian parameters into a 2D grid, maintaining local neighborhood structure and ensuring smoothness. This solution is particularly innovative because no existing method efficiently handles a 2D grid with millions of points. During training, a smoothness loss is applied to enforce local smoothness in the 2D grid, enhancing the compressibility of the data. The key insight is that smoothness needs to be enforced during training to enable efficient compression. [16]



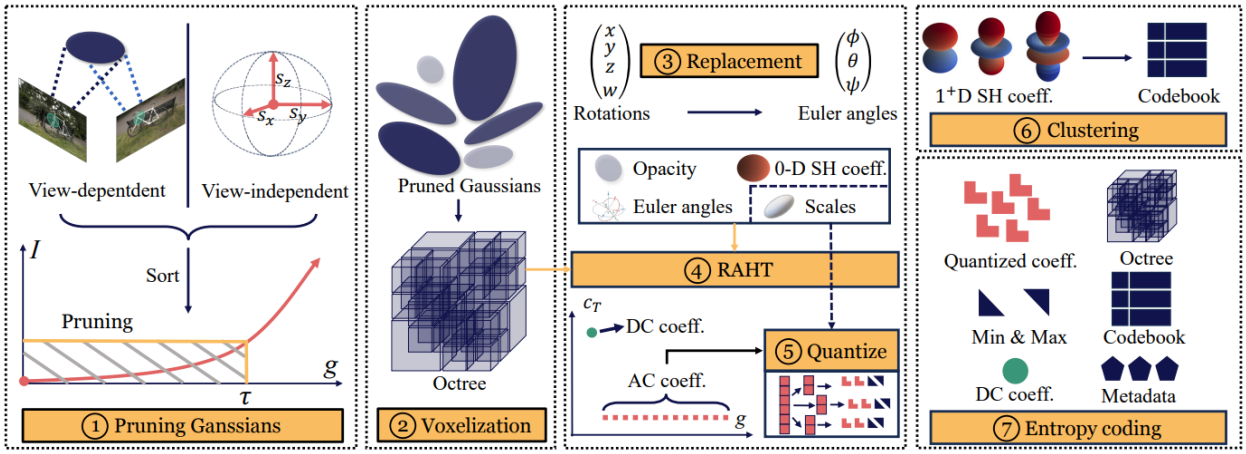
3.11 Implicit Gaussian Splatting with Efficient Multi-Level Tri-Plane Representation

This method introduces a hybrid representation for splatting-based radiance fields, where Gaussian primitives are separated into explicit point cloud and implicit attribute features. The attribute features are encoded using a multi-resolution multi-level tri-plane architecture integrated with a residual-based rendering pipeline. It employs a level-based progressive training scheme for joint optimization of point clouds and tri-planes, starting with coarse attributes and refining them with higher-level details. Spatial regularization and a bootstrapping scheme are applied to enhance the consistency and stability of the Gaussian attributes during training. [22]



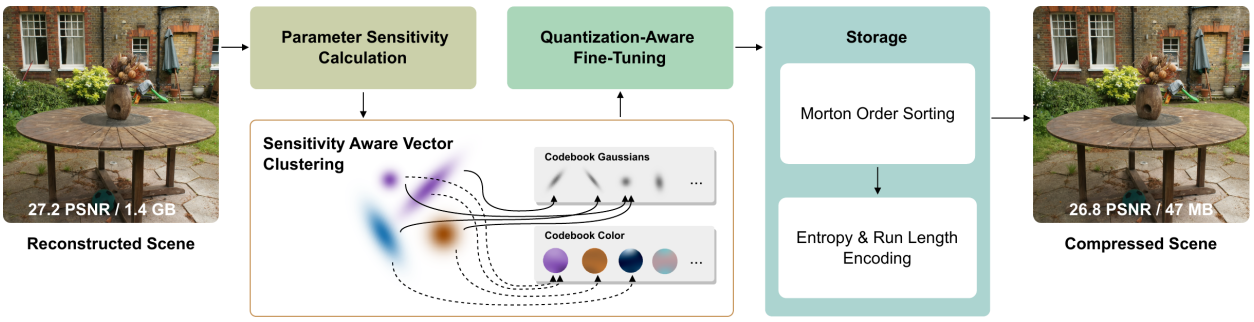
3.12 MesonGS: Post-training Compression of 3D Gaussians via Efficient Attribute Transformation

MesonGS employs universal Gaussian pruning by evaluating the importance of Gaussians through forward propagation, considering both view-dependent and view-independent features. It transforms rotation quaternions into Euler angles to reduce storage requirements and applies region adaptive hierarchical transform (RAHT) to reduce entropy in key attributes. Block quantization is performed on attribute channels by dividing them into multiple blocks and perform quantization for each block individually, using vector quantization for compressing less important attributes. Geometry is compressed using an octree, and all elements are packed with the LZ77 codec. A finetune scheme is implemented post-training to restore quality. [24]



3.13 Compressed 3D Gaussian Splatting for Accelerated Novel View Synthesis

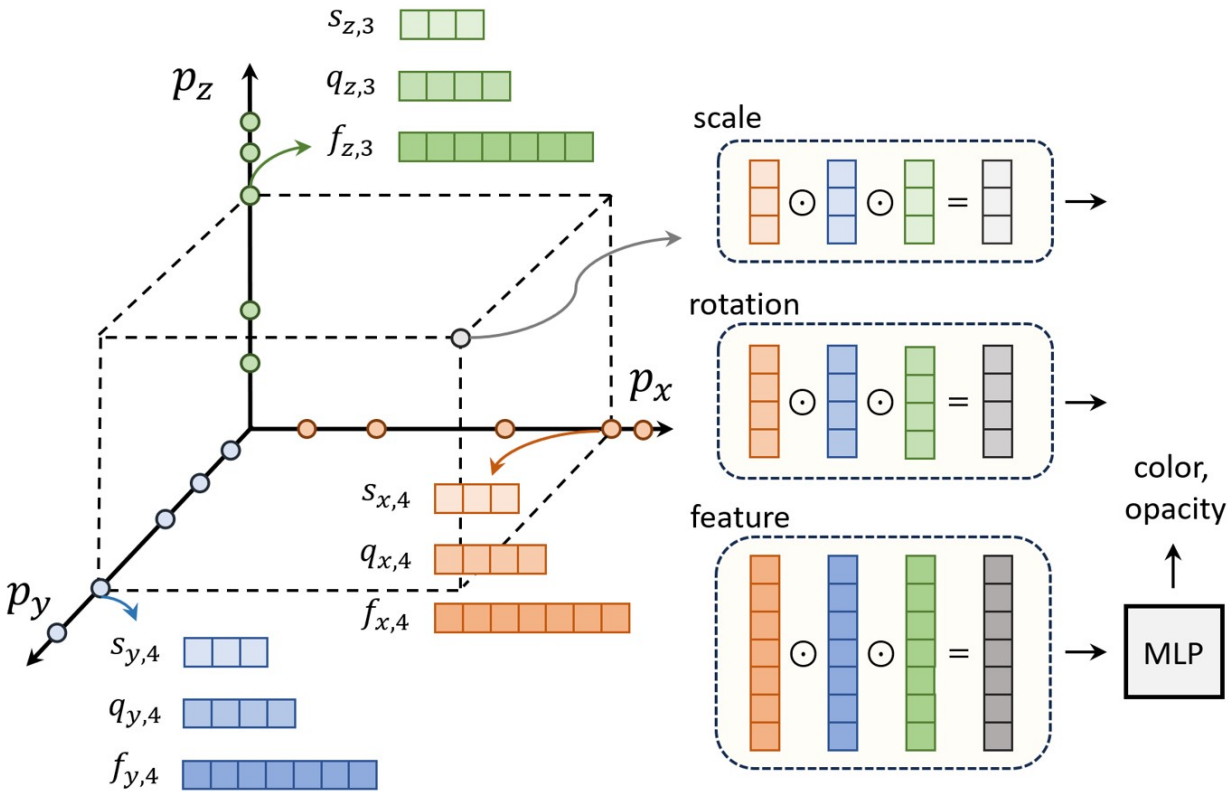
The authors propose a compressed 3D Gaussian splat representation consisting of three main steps: 1. sensitivity-aware clustering, where scene parameters are measured according to their contribution to the training images and encoded into compact codebooks via sensitivity-aware vector quantization; 2. quantization-aware fine-tuning, which recovers lost information by fine-tuning parameters at reduced bit-rates using quantization-aware training; and 3. entropy encoding, which exploits spatial coherence through entropy and run-length encoding by linearizing 3D Gaussians along a space-filling curve. Furthermore, a renderer for the compressed scenes utilizing GPU-based sorting and rasterization is proposed, enabling real-time novel view synthesis on low-end devices. [18]



3.14 F-3DGS: Factorized Coordinates and Representations for 3D Gaussian Splatting

The paper introduces a novel 3D Gaussian compression method using structured coordinates and decomposed representations through factorization. Inspired by tensor or matrix factorization techniques, this method generates 3D coordinates via a tensor product of 1D or 2D coordinates, enhancing spatial efficiency. It extends factorization to include attributes like color, scale, rotation, and opacity, which compresses the model size while maintaining essential characteristics. A binary mask is employed to eliminate non-essential Gaussians, significantly accelerating training and rendering speeds.

Note: This paper is currently not included in the survey table because it shows unusually high results in the Tanks and Temples dataset and reports higher results for the original 3DGS than those in the original publication. This raises the possibility that their testing methods may differ from those used in other papers. [20]



References

[1] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. 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] Y. Chen, Q. Wu, J. Cai, M. Harandi, and W. Lin. Hac: Hash-grid assisted context for 3d gaussian splatting compression, 2024, 2403.14530.

[3] Y. Chen, Q. Wu, M. Harandi, and J. Cai. How far can we compress instant-ngp-based nerf? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20321–20330, 2024.

[4] Z. Fan, K. Wang, K. Wen, Z. Zhu, D. Xu, and Z. Wang. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps, 2024, 2311.17245.

[5] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[6] S. Girish, K. Gupta, and A. Shrivastava. Eagles: Efficient accelerated 3d gaussians with lightweight encodings, 2024, 2312.04564.

[7] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018.

[8] J. Hu, R. Li, V. Ye, and A. Kanazawa. gsplat compression, 2024.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[10] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[11] J. C. Lee, D. Rho, X. Sun, J. H. Ko, and E. Park. Compact 3d gaussian representation for radiance field, 2024.

[12] L. Li, Z. Shen, Z. Wang, L. Shen, and L. Bo. Compressing volumetric radiance fields to 1 mb. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4222–4231, 2023.

[13] T. Lu, M. Yu, L. Xu, Y. Xiangli, L. Wang, D. Lin, and B. Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering, 2024.

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[15] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[16] W. Morgenstern, F. Barthel, A. Hilsmann, and P. Eisert. Compact 3d scene representation via self-organizing gaussian grids, 2024, 2312.13299.

[17] K. Navaneet, K. P. Meibodi, S. A. Koohpayegani, and H. Pirsiavash. Compact3d: Compressing gaussian splat radiance field models with vector quantization, 2024, 2311.18159.

[18] S. Niedermayr, J. Stumpfegger, and R. Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis, 2024.

[19] P. Papantonakis, G. Kopanas, B. Kerbl, A. Lanvin, and G. Drettakis. Reducing the memory footprint of 3d gaussian splatting. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 7(1):1–17, May 2024.

[20] X. Sun, J. C. Lee, D. Rho, J. H. Ko, U. Ali, and E. Park. F-3dgs: Factorized coordinates and representations for 3d gaussian splatting, 2024, 2405.17083.

[21] H. Wang, H. Zhu, T. He, R. Feng, J. Deng, J. Bian, and Z. Chen. End-to-end rate-distortion optimized 3d gaussian representation, 2024, 2406.01597.

[22] M. Wu and T. Tuytelaars. Implicit gaussian splatting with efficient multi-level tri-plane representation. *arXiv preprint arXiv:2408.10041*, 2024.

[23] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, pages 1–30, 2024.

[24] S. Xie, W. Zhang, C. Tang, Y. Bai, R. Lu, S. Ge, and Z. Wang. Mesongs: Post-training compression of 3d gaussians via efficient attribute transformation. In *European Conference on Computer Vision*. Springer, 2024.