INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 4
MPEG VIDEO CODING

ISO/IEC JTC 1/SC 29/WG 4 m66420
January 2024, Online

Title: [INVR]EE2.1-Related: 3D Gaussian Splatting for Visual Representation
Source: Sungkyunkwan University (SKKU)
Authors: Jaeyeol Choi, Jun-Hyeong Park, Jong-Beom Jeong, Eun-Seok Ryu

1 Introduction

The demand for real-time processing of radiance fields model is increasing nowadays. 3D INVR models like NeRF, Instant-NGP, TensoRF, ReRF, and K-Planes, which have been discussed in the MPEG INVR group, face limitations in real-time inference of 2D images. Among the existing NeRF variants, [1] and [2] have the advantage of enabling real-time rendering, but they either significantly lack quality or have limitations in the achievable resolution. The 3D Gaussian Splatting[3] paper claims to support high-quality 1080p image rendering at speed over 30fps. This document aims to apply and verify 3D Gaussian Splatting in the 3D INVR task.

This document provides an explanation of the 3D Gaussian Splatting technique and analyzes its advantages and disadvantages. Following this, the performance of 3D Gaussian Splatting is compared with other models according to experiments following the test condition of EE2.1.

2 3D Gaussian Splatting

2.1 Explanation of 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) is a method that uses multiple three-dimensional Gaussian probability distributions to represent a scene. The points sampled based on each Gaussian distribution assume an ellipsoid shape. The position of each ellipsoid is determined by the mean of the distribution, while the covariance matrix is decomposed to scaling and rotation matrices. Color of the ellipsoid is expressed through spherical harmonics, and transparency is represented by a value α. After rasterization of each ellipsoid, the loss is calculated by comparing it with the original image. The overall pipeline is differentiable, enabling optimization through stochastic gradient descent. Moreover, 3DGS greatly reduces rendering time by using a GPU-based optimized rasterization module.

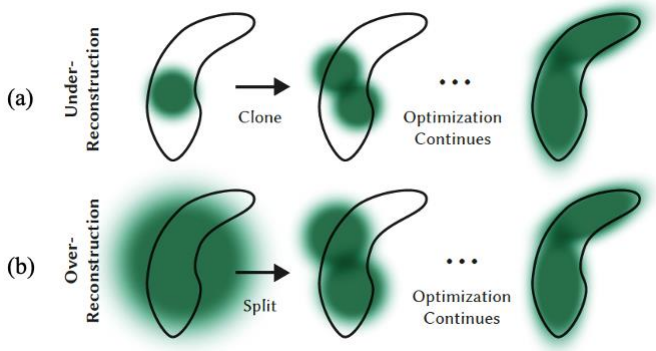


Figure 1. Adaptive control of Gaussians

There are several distinctive aspects of 3DGS to note. Firstly, at the start of training, the initial 3D Gaussians are initialized using a point cloud estimated through structure from motion (SfM)[4]. While NeRF-related methods typically acquire camera positions and rotations of 2D images using SfM, 3DGS additionally obtains a sparse point cloud through internal

calls to *colmap*. Secondly, during the optimization process, apart from the mean, covariance of 3D Gaussians, coefficients of spherical harmonics, and opacity, there are also operations involving the removal, cloning, and splitting of Gaussians. A Gaussian is removed if its opacity falls below a threshold. As shown in Figure 1(a), if the Gaussian features do not completely fill a reference region, it is defined as under-reconstruction and the gaussian is cloned. Similarly, as shown in Figure 1(b), if a Gaussian covers too broad compared to feature, it splits. This allows for more accurate representation of detailed elements.

2.2 Comparison with NeRF-based Approach

In the rasterization process, the screen is divided into 16x16 tiles, and for each tile, the relevant Gaussians are selected and sorted using GPU radix sort based on view depth. This rasterization, which transforms the 3D model into 2D image, is parallelized for each tile, enabling real-time rendering. NeRF-based models compose rays for each pixel corresponding to the screen and perform queries on the neural network for points sampled from these rays. In contrast, 3DGS projects the Gaussians corresponding to each area (tile) in order or proximity. Additionally, it does not use a deep neural network, and the explicit representation of visual components such as the position, shape, and color of ellipsoids also effectively reduces the inferring time.

3 Experiments

Experiments are conducted to evaluate the rendering speed and quality of the 3DGS model for still images. The experiment follows the test conditions of 'INVR exploration experiments document[5]'. Frame no.0 of M-CG1 (Mirror), frame no.16 of M-NC1 (SKKU_VRroom1D)[6], frame no.0 of M-NC2 (HauntedLamp)[7] are used for the experiment. For test set, view06 and 08 for Mirror, v11 and v27 for SKKU_VRroom1D, v06 for HauntedLamp was utilized. The average values of the metrics of two test views are recorded in the Table1, 2. All the rest were used for training set. The M-CG2 (Garage) sequence was temporarily excluded from the experiment due to difficulties in obtaining a sparse point cloud by SfM for initializing the Gaussians. For comparative analysis, the 3D INVR models K-Planes[8] and ReRF[9], as well as the immersive video standard MIV, were employed.

Figure 2. Results of test view v06 from M-CG1 (Mirror) sequence

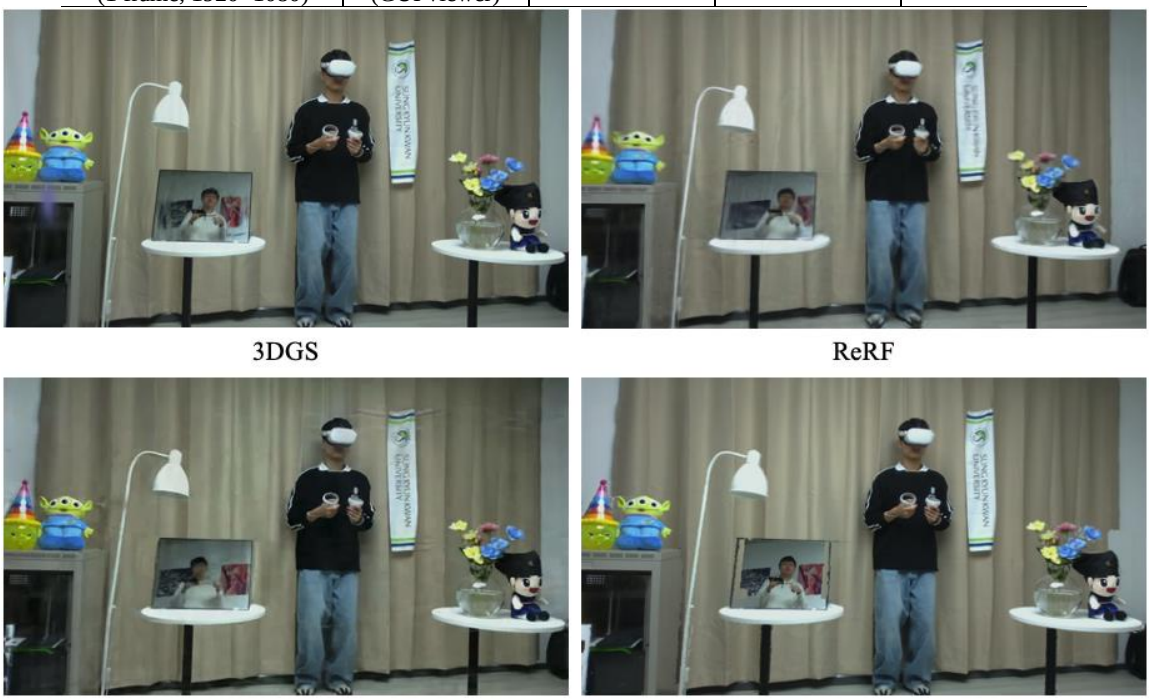
3DGS ReRF

K-Planes MIV DSDE

Figure 3. Results of test view v08 from M-CG1 (Mirror) sequence

Table 1. Performance analysis of M-CG1 (Mirror) sequence

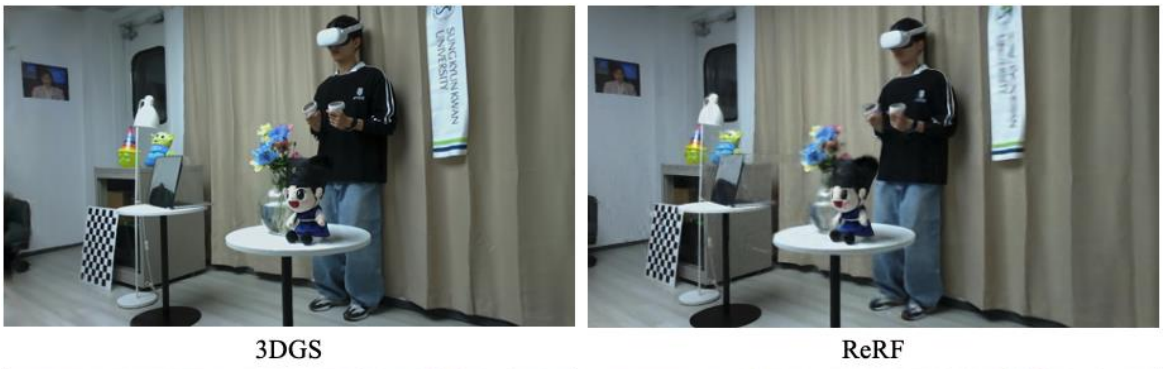
		3DGS	ReRF	K-Planes	MIV DSDE
Objective Quality	**PSNR ↑**	32.06	30.80	26.74	29.60
	SSIM ↑	0.936	0.887	0.836	0.933
	LPIPS ↓	0.131	0.096	0.302	0.106
Rendering Time (1 frame, 1920×1080)		25.67 ms (GUI viewer)	8.28 s	2.08 s	44.69 s



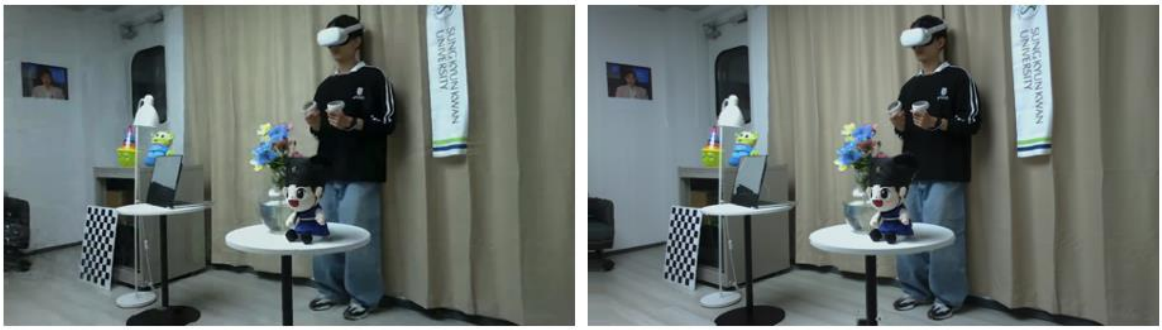
3DGS ReRF

K-Planes MIV DSDE

Figure 4. Results of test view v11 from M-NC1 (SKKU_VRroom1D) sequence



3DGS ReRF



K-Planes MIV DSDE

Figure 5. Results of test view v27 from M-NC1 (SKKU_VRroom1D) sequence

Table 2. Performance analysis of M-NC1 (SKKU_VRroom1D) sequence

		3DGS	ReRF	K-Planes	MIV DSDE
Objective Quality	**PSNR ↑**	28.53	29.03	27.53	28.61
	SSIM ↑	0.930	0.892	0.897	0.898
	LPIPS ↓	0.168	0.159	0.347	0.203
Rendering Time (1 frame, 1920×1080)		27.61 ms (GUI viewer)	8.21 s	2.11 s	83.40 s

3DGS ReRF

K-Planes MIV DSDE

Figure 6. Results of test view v06 from M-NC2 (HauntedLamp) sequence

Table 3. Performance analysis of M-NC2 (HauntedLamp) sequence

		3DGS	ReRF	K-Planes	MIV DSDE
Objective Quality	**PSNR ↑**	29.19	32.33	27.54	24.36
	SSIM ↑	0.900	0.880	0.837	0.724
	LPIPS ↓	0.191	0.165	0.236	0.338
Rendering Time (1 frame, 1920×1080)		31.44 ms (GUI viewer)	12.35 s	2.16 s	38.27 s

3DGS demonstrated the highest PSNR and SSIM values for the Mirror dataset. For natural datasets such as SKKU_VRroom1D and HauntedLamp, while the PSNR was lower than ReRF, the SSIM results were higher. In this way, 3DGS showed similar objective quality compared to state-of-the-art INVR models.

For closer inspection of the rendered images, 3DGS proved effective in restoring high-frequency areas. For instance, in Figure 2, the pattern behind the mirror, and in Figure 5, the letters on the placard, are well-represented, demonstrating its ability to handle complicated features. However, some artifacts were observed, such as in the face of the man reflected in the mirror in Figure 2 and the purple ellipsoid on the left side of Figure 5. When stability and overall consistency in quality are more important than the restoration of detailed elements, ReRF appears to be a more suitable choice.

The most significant advantage of 3DGS compared to other models is its rendering speed. 3DGS is able to render images of size 1920×1080 within 25~35ms. A video capturing the real-time rendering response to keyboard input using the 3DGS GUI viewer is made available in the attached folder.

4 Conclusion

This document introduces the new explicit radiance model, 3DGS, and analyzes the experimental results on the 3D INVR dataset. It highlights the advantages of high-quality real-time rendering capabilities, suggesting that investigating various methods applying 3D Gaussian Splatting should be worthwhile.

5 References

[1] Chen, Zhiqin, et al. "Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[2] Cao, Junli, et al. "Real-Time Neural Light Field on Mobile Devices." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[3] Kerbl, Bernhard, et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering." ACM Transactions on Graphics 42.4, 2023.

[4] J. L. Sch¨onberger and J.-M. Frahm, “Structure-from-motion revisited,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] "Exploration experiments on implicit neural visual representation", Standard ISO/IEC JTC 1/SC 29/WG 4, MPEG/n0426, 2023.

[6] "[INVR] EE2.1-Related: Report with New Natural INVR Video Contents: SKKU_VRroom", Standard ISO/IEC JTC 1/SC 29/WG 4, MPEG/m64721, 2023.

[7] "[INVR] [INVR] Proposition of a Multi-camera Dataset: HauntedLamp", Standard ISO/IEC JTC 1/SC 29/WG 4, MPEG/m64774, 2023.

[8] Fridovich-Keil, Sara, et al. "K-planes: Explicit radiance fields in space, time, and appearance." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[9] Wang, Liao, et al. "Neural Residual Radiance Fields for Streamably Free-Viewpoint Videos." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.