NOVA-3DGS: No-reference Objective VAlidation for 3D Gaussian Splatting

Valentina Piras[1,2] Amedeo F. Bonatti[3] Carmelo De Maria[3] Paolo Cignoni[2] Francesco Banterle[2]

[1]Department of Computer Science, University of Pisa, Italy, [2]ISTI-CNR, Italy
[3]Department of Information Engineering and Research Center E. Piaggio of the University of Pisa, Pisa, Italy

Abstract

In recent years, radiance field methods, and in particular 3D Gaussian Splatting (3DGS), have distinguished themselves in the field of image-based rendering and scene reconstruction techniques, gaining significant success in academia and being cited in numerous research papers. Like other methods, 3DGS requires a large and diverse dataset of images for network training as a fundamental step to ensure effectiveness and high-quality results. Consequently, the acquisition phase is highly time-consuming, especially considering that a portion of the acquired dataset is not actually used for training but is reserved for testing. This is necessary because all commonly used metrics for evaluating the quality of 3D reconstructions, such as PSNR and SSIM, are reference-based metrics; i.e., requiring a ground truth. In this work, we present NOVA, a study focused on no-reference evaluation of 3DGS renders, based on key metrics in this field: PSNR and SSIM.

CCS Concepts

• ***General and reference*** → ***Metrics;*** • ***Computing methodologies*** → ***Rendering;*** • ***Computer systems organization*** → *Neural networks;*

1. Introduction

3D Gaussian Splatting (3DGS) [KKLD23] has revolutionized novel view synthesis by enabling explicit scene representation, real-time high-quality rendering, and reduced training time. It is widely used in many different applications. Validating reconstruction quality is crucial, typically relying on full-reference (FR) metrics like PSNR and SSIM. However, FR metrics have two main limitations. First, they can only assess quality from viewpoints where images were captured, as ground truth is required. A no-reference (NR) metric would enable evaluation from any viewpoint. Second, reserving images for testing lowers reconstruction quality. Using all acquired images for optimization would yield better results with the same resources while allowing NR evaluation from previously unseen viewpoints. To address these issues, we present NOVA, a study on NR methods for validating 3DGS results. This work converts the main FR metrics (PSNR and SSIM) into an NR metric specifically tailored for 3DGS renderings, leveraging PSNR's strong correlation with human perception in neural view synthesis [LWH*24].

2. Related Works

Recently, there has been growing interest in NR metrics; however, to the best of our knowledge, none have attempted to convert a FR metric into a NR metric. NeRF-NQA [QLC*24] is a NR quality assessment method specifically designed for neural view synthesis techniques, such as NeRF and its variants. However, NeRF-NQA is not tailored for 3DGS, which employs different scene representations and rendering strategies. Puzzle Similarity [HCD24] is a NR metric designed to locate artifacts in novel views of 3D reconstructions. Their method exhibits a stronger correlation with human assessments compared to traditional FR metrics. However, it remains unclear whether the metric generalizes well to different datasets used for 3DGS-based reconstructions.

3. Method Overview

In our approach, we want to infer the quality values that the FR metrics compute using only the distorted image [BAM*23]; see Figure 1. We first applied the 3DGS method to all datasets to generate renderings. In this way each dataset was divided into a training set, used for 3D reconstruction optimization, and a testing set, used for quality assessment. We used a total of three datasets: the dataset from the 3DGS project [KKLD23] (13 scenes; training set: 302 images; testing set: 38 images), the fieldwork scenes dataset by Liang *et al.* [LWH*24] (9 scenes; training set: 899 images; testing set: 1079 images), and the standard dynamic range (SDR) dataset by Piras *et al.* [PBM*24] (9 scenes; training set: 323 images; testing set: 49 images). During the training phase, we used ground truth and corresponding distorted (reconstruction) images from the 3DGS training set. The network was trained on two FR metrics (PSNR and SSIM) for 128 epochs with three models (NorVDP-Net++, ResNet-18, and a Transformer pretrained by Dosovitskiy

DOI: 10.2312/egp.20251024

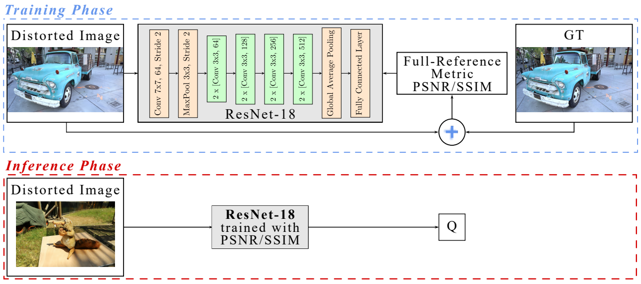

Figure 1: *An illustration of the training and inference phases. The truck photo is sourced from the Tanks and Temples dataset [KPZK17].*

et al. [DBK*21]), three batch sizes (1, 4, 8), and with/without a sigmoid activation. Furthermore, the trainings were performed on random patches (512×512 for NorVDPNet++ and ResNet-18, 224×224 for the Transformer) of input full-resolution images (13 scenes; average resolution 1315×854 pixels). For PSNR the mean training time was 1h 35m 21s and never exceeded 1h 46m 38s, while for SSIM the mean training time was 1h 46m e 17s and never exceeded 1h 58m 40s. Trainings were performed on a single NVIDIA RTX 3080 GPU equipped with 10GB of memory. Then, during the inference phase we used both the training and testing set of the 3DGS dataset [KKLD23], and we also performed cross-dataset evaluation using the testing sets of Liang *et al.* [LWH*24] and Piras *et al.* [PBM*24] datasets. To assess the quality of the generated indices, we calculated the Pearson's linear correlation coefficient (PLCC), the Spearman's rank order correlation coefficient (SROCC), and the root mean square error (RMSE) between the generated quality values and those computed by the FR metrics using the ground truth.

PSNR ↑

Model	Sigmoid	Batch Size	Kerbl *et al.* training dataset PLCC ↑	SROCC ↑	RMSE ↓	Kerbl *et al.* testing dataset PLCC ↑	SROCC ↑	RMSE ↓	Liang *et al.* testing dataset PLCC ↑	SROCC ↑	RMSE ↓	Piras *et al.* testing dataset PLCC ↑	SROCC ↑	RMSE ↓
NorVDPNet++	Off	1	0.945	0.935	1.408	0.915	0.909	1.597	0.124	0.050	9.887	0.420	0.439	4.261
		4	0.906	0.883	1.816	0.884	0.876	1.856	0.276	0.234	8.689	0.461	0.460	4.154
		8	0.891	0.867	1.829	0.866	0.841	2.039	0.179	0.106	9.338	0.242	0.220	4.626
	On	1	0.895	0.880	2.038	0.880	0.836	1.893	0.178	0.086	8.650	0.324	0.367	4.595
		4	0.878	0.858	2.098	0.862	0.840	1.996	0.250	0.188	8.749	0.264e-04	-0.028	5.066
		8	0.905	0.889	1.875	0.902	0.885	1.748	0.266	0.229	8.825	0.221	0.186	4.644
ResNet-18	Off	1	0.264	0.246	4.055	0.104	0.008	3.956	-0.217	-0.150	8.623	0.189	0.292	4.993
		4 ★	0.894	0.882	1.817	0.874	0.878	1.933	0.447	0.436	8.533	0.667	0.691	3.495
		8	0.879	0.866	2.029	0.873	0.877	2.268	0.312	0.328	8.891	0.664	0.705	3.570
	On	1	0.259	0.288	4.121	0.170	0.229	3.859	0.382	0.233	8.019	0.378	0.465	5.021
		4	0.887	0.877	2.482	0.885	0.884	2.212	0.343	0.347	7.253	0.658	0.664	3.856
		8 ★	0.901	0.891	2.261	0.898	0.894	2.005	0.469	0.415	7.266	0.674	0.699	3.776
Transformer	Off	1	0.994	0.994	0.494	0.868	0.858	2.014	0.323	0.320	9.655	0.341	0.295	4.490
		4	0.994	0.994	0.474	0.860	0.842	2.066	0.100	0.113	10.199	0.241	0.288	4.638
		8	0.995	0.994	0.437	0.838	0.828	2.159	0.399	0.508	10.794	0.410	0.412	4.336
	On	1	0.997	0.998	0.316	0.886	0.886	1.915	0.532	0.483	8.500	0.350	0.292	4.831
		4 ♦	0.998	0.999	0.273	0.894	0.906	1.835	0.495	0.410	7.933	0.261	0.304	4.800
		8 ♦	0.998	0.998	0.297	0.882	0.908	1.929	0.594	0.522	8.036	0.242	0.275	4.819

SSIM ↑

Model	Sigmoid	Batch Size	Kerbl *et al.* training dataset PLCC ↑	SROCC ↑	RMSE ↓	Kerbl *et al.* testing dataset PLCC ↑	SROCC ↑	RMSE ↓	Liang *et al.* testing dataset PLCC ↑	SROCC ↑	RMSE ↓	Piras *et al.* testing dataset PLCC ↑	SROCC ↑	RMSE ↓
NorVDPNet++	Off	1	0.921	0.825	0.070	0.943	0.831	0.068	-0.204	-0.327	0.194	0.636	0.662	0.076
		4	0.796	0.736	0.088	0.794	0.720	0.083	-0.086	-0.326	0.169	0.065	0.200	0.090
		8	0.781	0.724	0.089	0.810	0.713	0.083	0.011	-0.136	0.168	-0.044	0.081	0.097
	On	1	0.923	0.867	0.055	0.891	0.798	0.058	0.016	-0.449	0.154	0.582	0.546	0.065
		4	0.866	0.833	0.071	0.922	0.883	0.048	-0.101	-0.552	0.131	0.063	0.186	0.090
		8	0.765	0.784	0.096	0.751	0.834	0.079	0.166	0.003	0.128	-0.547	-0.573	0.113
ResNet-18	Off	1	0.679	0.682	0.120	0.773	0.731	0.097	0.467	0.520	0.099	0.608	0.705	0.078
		4	0.903	0.831	0.077	0.891	0.848	0.083	0.450	0.629	0.187	0.790	0.824	0.124
		8	0.917	0.844	0.079	0.938	0.904	0.080	0.527	0.677	0.194	0.804	0.839	0.121
	On	1	0.510	0.597	0.121	0.661	0.646	0.106	0.315	0.493	0.135	0.536	0.705	0.080
		4 ★	0.937	0.904	0.048	0.931	0.920	0.048	0.593	0.651	0.150	0.844	0.843	0.089
		8 ★	0.937	0.912	0.049	0.963	0.947	0.038	0.630	0.698	0.145	0.817	0.833	0.091
Transformer	Off	1	0.990	0.974	0.015	0.863	0.797	0.044	0.090	0.406	0.243	0.369	0.381	0.113
		4	0.986	0.964	0.019	0.879	0.810	0.045	0.089	0.382	0.323	0.290	0.341	0.139
		8	0.988	0.967	0.017	0.892	0.835	0.042	-0.017	0.271	0.320	0.368	0.411	0.133
	On	1 ♦	0.996	0.988	0.009	0.889	0.881	0.039	0.181	0.119	0.163	0.453	0.566	0.096
		4 ♦	0.995	0.984	0.011	0.887	0.862	0.039	0.041	0.056	0.161	0.487	0.624	0.095
		8 ♦	0.996	0.988	0.009	0.900	0.879	0.037	0.278	0.336	0.151	0.451	0.551	0.094

Table 1: *Comparisons of NOVA's predicted quality scores and ground truth for PSNR and SSIM using PLCC, SROCC, and RMSE. Best and second-best values are highlighted in red and blue, respectively. Mean indices across the 4 datasets were computed for each training setting, with ★ marking the best (red) and second-best (blue) mean PLCC and SROCC values, and ♦ for RMSE mean value.*

4. Results

Preliminary results have demonstrated the feasibility of the proposed method concerning the FR metrics PSNR and SSIM. Table 1 shows that the Transformer architecture with sigmoid activation achieves the best results on the training dataset and the lowest median RMSE value across all datasets. However, it can also be observed that the best overall performance, both on familiar and entirely unseen datasets, is obtained with the ResNet-18 architecture using a batch size of 8 and sigmoid activation. This configuration achieves the highest mean PLCC and SROCC values while maintaining a low RMSE (ranking third in terms of best median value).

5. Conclusions and Future Work

With the presented study, NOVA, we assessed the feasibility of using neural networks to convert FR metrics into NR metrics for the quality assessment of 3DGS renderings.

The preliminary results indicate that, for both PSNR and SSIM, the best performance is achieved using the ResNet-18 architecture with sigmoid activation and a batch size of 8. In this configuration, the model performs consistently well, not only on the training dataset but also on entirely new datasets, maintaining a low RMSE. Future work includes exploring additional architectures, FR metrics, and datasets.

References

[BAM*23] Banterle F., Artusi A., Moreo A., Carrara F., Cignoni P.: Nor-vdpnet++: Real-time no-reference iqm. *IEEE Access* (2023), 34544–34553. doi:10.1109/ACCESS.2023.3263496. 1

[DBK*21] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N.: An image is worth 16x16 words: Transformers for image recognition, 2021. arXiv:2010.11929. 2

[HCD24] Hermann N., Condor J., Didyk P.: Puzzle similarity: A no-reference metric for 3d artifacts, 2024. arXiv:2411.17489. 1

[KKLD23] Kerbl B., Kopanas G., Leimkuehler T., Drettakis G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* (2023). doi:10.1145/3592433. 1, 2

[KPZK17] Knapitsch A., Park J., Zhou Q.-Y., Koltun V.: Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics 36*, 4 (2017). doi:10.1145/3072959.3073599. 2

[LWH*24] Liang H., Wu T., Hanji P., Banterle F., Gao H., Mantiuk R., Öztireli C.: Perceptual quality assessment of nerf. *Comput. Graph. Forum* (2024). doi:10.1111/cgf.15036. 1, 2

[PBM*24] Piras V., Bonatti A. F., Maria C. D., Cignoni P., Banterle F.: Hdr imaging for gaussian splatting: Are 8 bits enough? In *STAG - Eurographics Ital. Chapter Conf.* (2024). doi:10.2312/stag.20241341. 1, 2

[QLC*24] Qu Q., Liang H., Chen X., Chung Y. Y., Shen Y.: Nerf-nqa: No-reference quality assessment for nerf. *IEEE TVCG* (2024), 2129–2139. doi:10.1109/TVCG.2024.3372037. 1