Explicit_NeRF_QA: A Quality Assessment Database for Explicit NeRF Model Compression

Yuke Xing#1, Qi Yang*2, Kaifa Yang#1, Yilin Xu#1, Zhu Li*3
Cooperative MediaNet Innovation Center, Shanghai Jiao Tong University, Shanghai, China
* Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, USA
1{xingyuke-v, sekiroyyy, yl.xu}@sjtu.edu.cn
2littlleempty@gmail.com
3lizhu@umkc.edu

***Abstract*—In recent years, Neural Radiance Fields (NeRF) have demonstrated significant advantages in representing and synthesizing 3D scenes. Explicit NeRF models facilitate the practical NeRF applications with faster rendering speed, and also attract considerable attention in NeRF compression due to its huge storage cost. To address the challenge of the NeRF compression study, in this paper, we construct a new dataset, called Explicit_NeRF_QA. We use 22 3D objects with diverse geometries, textures, and material complexities to train four typical explicit NeRF models across five parameter levels. Lossy compression is introduced during the model generation, pivoting the selection of key parameters such as hash table size for InstantNGP and voxel grid resolution for Plenoxels. By rendering NeRF samples to processed video sequences (PVS), a large scale subjective experiment with lab environment is conducted to collect subjective scores from 21 viewers. The diversity of content, accuracy of mean opinion scores (MOS), and characteristics of NeRF distortion are comprehensively presented, establishing the heterogeneity of the proposed dataset. The state-of-the-art objective metrics are tested in the new dataset. Best Person correlation, which is around 0.85, is collected from the full-reference objective metric. All tested no-reference metrics report very poor results with 0.4 to 0.6 correlations, demonstrating the need for further development of more robust no-reference metrics. The dataset, including NeRF samples, source 3D objects, multiview images for NeRF generation, PVSs, MOS, is made publicly available at the following location: https://github.com/LittlericeChloe/Explicit_NeRF_QA.**

***Index Terms*—Neural Radiation Field, Subjective Quality Assessment, Objective Quality Assessment, NeRF Model Compression, Explicit NeRF Model**

I. Introduction

Neural Radiance Fields (NeRF) [1] is a method to reconstruct 3D scenes from 2D sparse multiview images, which has an impressive ability to synthesize novel views with high realism. NeRF variants could be categorized into two types: implicit and explicit. Implicit NeRF methods encode 3D scenes implicitly by sampling a large number of spatial points in the rendering process based on MLP networks, resulting in high computational cost and slow convergence speed. In contrast, explicit NeRF methods divide 3D scenes into voxel grids and store the local features in each grid cell, which significantly reduces the computational load during the training and rendering process, thus greatly promoting the development of NeRF practical applications.

One disadvantage of explicit NeRF models is that they inevitably increase the memory and storage overhead. Therefore, researchers have proposed multifarious strategies to reduce the size of explicit NeRF methods pivoting to a series of compression methods [2]–[6]. Similarly to image/video compression, NeRF compression also incurs distortion and impacts perceptual quality. To find the most suitable compression strategies, effective NeRF quality assessment (NeRF-QA) metrics are needed. However, because of the lack of research on NeRF-QA, most researches still use traditional image/video quality metrics to assess NeRF models' quality, which neglect unique distortions of the NeRF models and are resulting in inaccurate prediction. Therefore, a new dataset with diverse types of NeRF distortion samples and substantial scale is firstly needed for designing and evaluating NeRF-QA metrics. An effective metric can facilitate the study of NeRF compression, revealing that a NeRF-QA dataset is consequently necessary for the study of NeRF compression.

As we know, there are three studies on NeRF-QA dataset. [7] collected 14 real scenes, together with the scenes in LLFF [8], they rendered NeRF models into videos and performed experiments to collect subjective scores. [9] use 4 real scenes from "Tanks and Temples" [10] and 4 synthetic scenes from NeRF_Synthetic [1]. For each kind of scene, they chose 4 NeRF methods to synthesize videos and obtain perception quality scores. [11] considered 8 real scenes and 8 synthetic scenes on several NeRF methods. The above works are the pioneer of NeRF-QA and contribute significantly to the NeRF study. However, the above datasets have the disadvantages of small-scale or scanty distortion types. Considering the rapid development of the NeRF study, it is necessary to construct a new dataset that provides diverse content, rich distortions, and trustworthy mean opinion scores (MOS).

In this paper, we create a new dataset called Explicit_NeRF_QA. We use 22 synthetic scenes as reference to generate NeRF models. All NeRF methods belong to mainstream and widely studied explicit models, i.e., InstantNGP [12], DVGO [13], Plenoxels [14], and TensoRF [15]. Each NeRF model has five quality levels corresponding to different compression levels by controlling key model parameters, such as the hash table size for InstantNGP. In all, there are $22 \times 4 \times 5 = 440$ samples in Explicit_NeRF_QA. Both reference and distorted NeRF samples are rendered into processed video sequences (PVS) to conduct subjective experiments. We split the dataset into two parts: 20 samples used as training sessions and 420 samples as rating sessions. The double stimulus impairment scale is used to collect subjective scores. After removing outliers, we analyze the content diversity and MOS accuracy, and also illustrate that the NeRF models have distinct distortion types from the traditional 2D image, such as fogging, ripple, and wave. In addition, NeRF models have different response to object materials, specular reflective material is the most challenging type for NeRF generation and lossy compression. We test 15 state-of-the-art (SOTA) objective metrics. The highest Person and Speraman correlations are around 0.85, which are reported from full-reference metrics, while all the tested no-reference metrics show very low correlations, indicating the need for further development of more robust no-reference metrics for NeRF models.

II. Dataset Construction

This section presents the selection of source content, the generation of multiview images, the selection and training of NeRF models, and the subjective experiments.

A. Source Content Selection

The performance of NeRF models is correlated with material categories, therefore, we additionally consider material diversity in source content selection. Specifically, we use 22 different contents as source, of which eight from previous work [1] and 14 new high-quality 3D scenes carefully selected from [16] with diverse categories, materials, geometry and texture complexity. The 22 selected scenes are shown in Fig. 1.

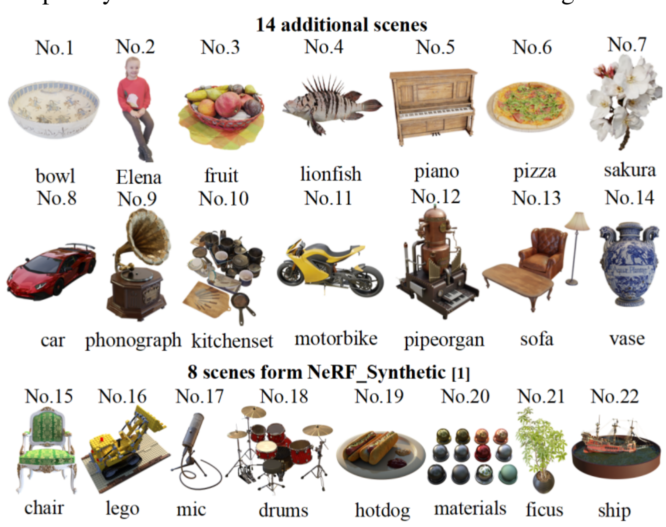


Fig. 1: The source content in Explicit_NeRF_QA

For the 14 new high-quality 3D scenes, we categorize their materials into three types based on the properties of light reflection: specular reflective materials (SRM), diffuse reflective materials (DRM), and glossy reflective materials (GRM): 1) **SRM**: Objects possess highly smooth surfaces that can produce clear reflections. The reflected light highly depends on the position of observer and light source. 2) **DRM**: Objects feature rough surfaces that scattering incident light uniformly in multiple directions. The color and brightness remain consistent viewed from any angle. 3) **GRM**: Objects have common surfaces (between smooth and rough) that can produce blurred reflections, the clarity depends on the surface's roughness.

Among the 14 additinoal 3D scenes in Fig. 1, No.1-7 belong to DRM, No.8 and 9 belong to SRM, and No.10-14 belong to GRM.

B. Multiview Image Generation

The generated multiview images consist of three parts based on the requirements and rules of training NeRF models: training, testing and validation set. The training set is used as input for training NeRF models, testing set is used to calculate objective metrics, and validation set is used to render videos for subjective experiments. For training and testing sets, we randomly select 100 viewpoints on the sphere or hemisphere around objects. Training and testing sets do not have repeated viewpoints. For validation set, we select 300 viewpoints located spirally at the upper hemisphere circling the object with uniform angular intervals. Each set includes multiview images and the corresponding camera parameter files.

C. NeRF Model Selection and Training

1) NeRF Model Selection: Four prevalent and representative explicit NeRF models are selected as anchor methods to generate NeRF samples:

InstantNGP [12]: InstantNGP employs a multi-level hash encoding structure to map multi-resolution voxel grids into different hash tables, which can realize store learnable feature vectors with more compact method.

DVGO [13]: DVGO proposes two initialization algorithms to prevent the scene geometry from falling into local optima, and introduces a post-activation voxel grid interpolation method that generates more clear boundaries at lower grid resolutions.

Plenoxels [14] : With each grid storing volume density and spherical harmonics coefficients, Plenoxels directly optimizes these parameters through gradient backpropagation and regularization methods without relying on any neural network.

TensoRF [15]: Viewing the voxel grid as a 4D tensor of size (X, Y, Z, C), TensoRF introduces a Vector-Matrix decomposition method and factorizes the 4D tensor into multiple and compact low-rank tensor components, thus reducing storage space.

2) NeRF Model Training: To build a large-scale NeRF-QA dataset that includes various distortion types and quality levels, we propose five parameter levels (PL) for each NeRF method to produce samples with hierarchical size. The key parameters selected for each NeRF model are shown in Table I.

D. PVS Generation

To conduct subjective experiments, each NeRF sample is processed into PVS. By employing predefined camera trajectories in the validation set, 300 frames are rendered that circle

TABLE I: PL of NeRF models. L01-L05 corresponding to the five PLs.

Model	Instantngp	Plenoxels	DVGO	TensoRF
Key Parameters	Hash Table Size T	Voxel Grid Resolution	Voxel Grid Resolution (coarse, fine)	Voxel Grid Resolution, Component Count
L01	2^6	32^3	$[10^3, 10^3]$	$[15^3, \text{Comp:}48]$
L02	2^8	64^3	$[10^3, 32^3]$	$[25^3, \text{Comp:}48]$
L03	2^{10}	128^3	$[15^3, 32^3]$	$[40^3, \text{Comp:}96]$
L04	2^{13}	256^3	$[32^3, 64^3]$	$[100^3, \text{Comp:}96]$
L05	2^{19}	512^3	$[100^3, 160^3]$	$[300^3, \text{Comp:}192]$

the object with uniform angular intervals at a resolution of 800×800. Subsequently, with a frame rate of 30 and constant rate factor of 10 to guarantee visually lossless encoding [17], these images are compiled into PVS through FFMPEG utilizing libx265. Each PVS has a duration of 10 seconds.

E. Subjective Experiment

1) Training and Evaluation: To guarantee the reliability of the collected subjective scores, we employ the "lionfish" (No.11 in Fig. 1) to establish a training session, following the same method proposed in [18]. In the rating session, an 11-level impairment scale proposed by ITU-TP.910 [19] is used as the rating method. A double stimulus impairment scale method is utilized, with PVS of the reference and distorted NeRF samples side-by-side displayed on the left and right of the monitor. The experiment was carried out using a 27-inch AOC Q2790PQ monitor in an indoor laboratory environment under standard lighting conditions. To align with the PVSs' format, the monitor's display resolution is set to 1600×800. To prevent visual fatigue caused by too long experiment time, 420 PVSs are randomly divided into 6 smaller groups.

2) Outlier Removal: To filter outliers from the raw subjective scores, two consecutive steps are used. First, we add an extremely low-quality PVS and a duplicated PVS to each rating session, known as trapping samples. After collecting subjective scores, outliers are identified and excluded based on the ratings to these trapping samples. Subsequently, ITU-R BT.500 [20] is applied for the secondary detection and elimination of outliers. Finally, two outliers are identified and removed from the initial subjective scores.

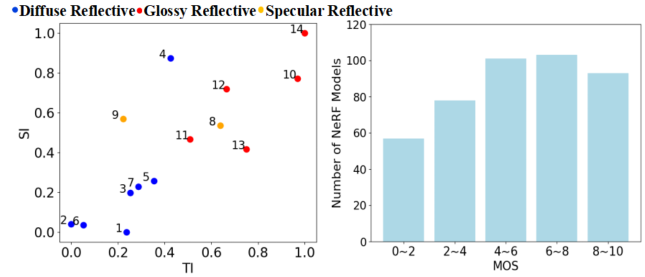
III. Dataset Validation

A. Diversity of Content

To validate the diversity of the source content, we measure spatial perceptual information (SI), temporal perceptual information (TI) [19], and the diversity of materials of the newly selected 14 scenes. SI and TI use reference PVSs as input, and the results are shown in Fig. 2. The relatively uniform distribution of the scatter points indicates the diversity of the source content.

B. Analysis of MOS

To verify whether the proposed database covers a wide range of quality levels, the MOS distribution is presented in Fig. 2. For each score segment, Explicit_NeRF_QA has at least 50 samples, showcasing that the MOSs are evenly distributed in the whole quality range.



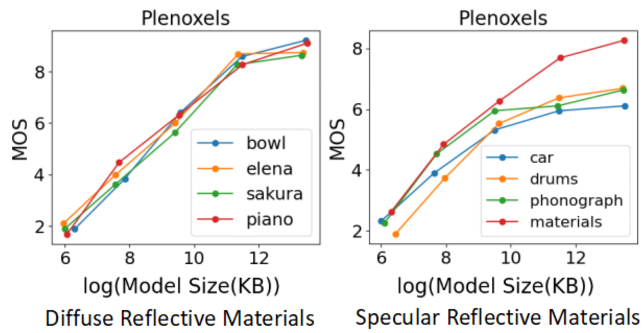


Fig. 3: MOS vs log(model size(KB))

To verify the accuracy of the MOS, we examine the correlation between MOS and bitrate. Fig. 3 illustrates eight contents with two different materials: we use results of Plenoxels as examples, and picked four objects belonging to DRM (i.e.,bowl, elena, sakura, piano) and SRM (i.e., car, drums, phonograph, materials). Given a NeRF model, a larger bitrate is supposed to indicate more detailed visual features, thus higher MOS. The results report that the curves present perfect monotonicity, validating the accuracy of the MOS scores and the reasonable settings of our parameter levels. For DRM, the MOS grow gradually and eventually reach 8-10 scores. For SRM, most samples' MOS do not have notable growth with the increasing of bitrate after reaching scores around 5 to 7. Current NeRF models are not perfect and still under study, they unavoidable introduce some artifacts when reconstructing 3D representations from multiview images. Considering that human observers are highly sensitive to the distortion produced on smooth surfaces, training NeRF models on materials with clear reflections is more challenging.

IV. NeRF Distortion Analysis

By observing 440 NeRF PVSs and gathering feedback from participants in subjective experiments, we found that NeRF has some distortion types that are not available in other media forms, such as point clouds [21] and meshes [22], which are one of the main challenges in conducting objective quality assessments based on NeRF. We comprehensively and systematically summarize 9 unique distortion types specific to NeRF:

(1) Transparency Distortion: Certain areas of an image appear more transparent than the corresponding reference image, allowing the content behind them to be partially visible, which reduces the realism of the synthesized view.

(2) Floating Distortion: Typically appearing as discrete, blurry spots, blocks, or fog-like structures, with an unnatural visual break from their background.

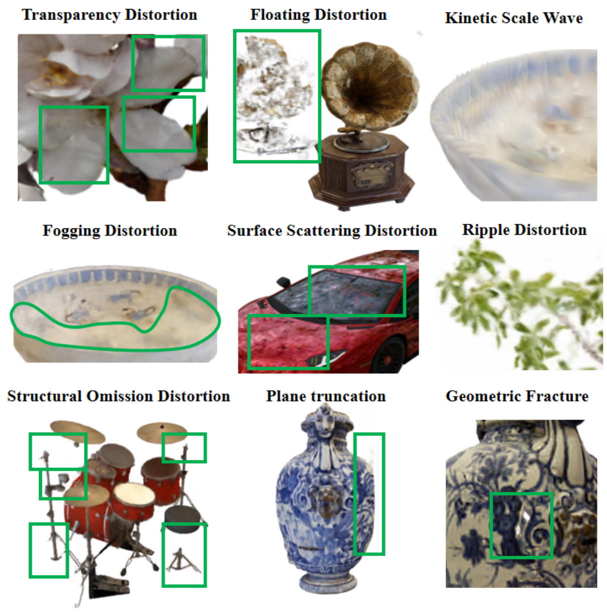


Fig. 4: Examples for NeRF Distortions

(3) Kinetic Scale Wave Distortion: A scaly appearance and repetitive texture appearance when static, and a wavy effect when in motion, giving viewers a sense of blurriness and dizziness, significantly reducing perceived quality.

(4) Fogging Distortion: Characterized by the appearance of fog-like or smoky, opaque, or semi-transparent coverings on the surface or nearby space of the rendered objects, obscuring the details underneath and decrease surface details and textures visual clarity.

(5) Surface Scattering Distortion: Often appears on the surfaces of specular reflective materials, which exhibits textures that look like cotton or cloud-like patterns, contradicting the object's actual smooth or uniform surface texture.

(6) Ripple Distortion: Unnatural ripples or texture distortions make the texture of the objects look wavy or as if they have irregular creases.

(7) Structural Omission Distortion: Characterized by missing macro structures that disrupt the realism and continuity of the entire scene.

(8) Plane Truncation Distortion: It is characterized by abrupt cut-offs and incomplete rendering of 3D surfaces, resulting in objects appearing as if they have been sliced or truncated along certain planes, and making it difficult to appreciate the full geometry and detail of the scene. When in motion, these truncations can create a disruptive visual experience for the viewer.

(9) Geometric Fracture Distortion: Appearance of geometric discontinuities and damaged effects, where the structural integrity of the object's surface is broken, disrupting the continuity of shape and texture.

V. Objective Metrics Testing

The performance of SOTA image/video quality metrics on PVSs are tested in this section. We select 15 objective metrics covering various types, specifically: i) full-reference image metrics, including classic metrics, PSNR, SSIM [23], MS-SSIM [24], IW-SSIM [25], VIF [26] and FSIM [27], deep-learning based metrics, inclduing LPIPS (VGG and AlexNet versions) [28] and DISTS [29]; ii) no-reference image metrics, including BRISQUE [30], CLIP_IQA [31] and NIQE [32]; iii) video metrics, including FovVideoVDP [33], HDR-VQM [34] and VMAF [35].

A. *The Performance of Metrics*

To evaluate the performance of the metrics, the Pearson linear correlation coefficient (PLCC) [36], the Spearman Rank Order Correlation Coefficient (SRCC) [37] and the Root Mean Square Error (RMSE) [36] are calculated. A five-parameter logistic fitting function proposed by the video quality experts group [38] is used to map the dynamic range of scores from the objective metrics to a common scale, in order to ensure consistency between the various objective metrics and MOS. The performance of the metrics is shown in Table II.

TABLE II: The Performance of Metrics

Metrics	Reference	SRCC	PLCC	RMSE
PSNR	✓	0.75	0.76	1.70
SSIM	✓	0.77	0.83	1.46
LPIPS (VGG)	✓	0.74	0.75	1.72
LPIPS (Alex)	✓	0.77	0.81	1.51
MSSSIM	✓	0.76	0.83	1.43
IW-SSIM	✓	0.82	0.91	1.08
VIF	✓	0.79	0.85	1.48
FSIM	✓	0.76	0.79	1.59
DISTS	✓	0.86	0.88	1.24
BRISQUE	×	0.40	0.38	2.38
CLIP-IQA	×	0.64	0.61	2.08
NIQE	×	0.61	0.65	1.99
FovVideoVDP	✓	0.77	0.79	1.60
VQM	✓	0.70	0.73	1.79
VMAF	✓	0.83	0.83	1.44

It can be observed that the best full-reference metric is IW-SSIM with PLCC, SRCC, and RMSE are 0.82, 0.91, 1.1. Its information-weighting mechanism is well suited for evaluating NeRF rendering considering it can distinguish and emphasize distortions that occur in information-rich areas, such as fogging distortion and surface scattering distortion. DISTS has higher PLCC than IW-SSIM, e.g., 0.86, but lower SRCC (0.88). All the non-reference metrics report obviously poor performance than other metrics. VMAF is the best video metric, it also shows good performance on mesh and point cloud PVSs [39]. If NeRF PVS is available, using VMAF is a good choice since this metric has been widely accepted and generally shows robust performance in most cases.

VI. Conclusion

In this work, we create a new dataset called Explicit_NeRF_QA. It consists of 22 3D objects with diverse content and varying levels of compression distortion based on 4 explicit NeRF methods. With accuracy MOS labels, this dataset enables the design of NeRF objective metrics and facilitates the development of NeRF compression technologies.

This dataset undergoes comprehensive analysis to validate its sample diversity, MOS accuracy, and also illustrates typical NeRF distortion characteristics. Based on the evaluation of SOTA metrics, the best full-reference metrics achieve a correlation around 0.85, while all the non-reference metrics are struggle with NeRF quality prediction with only 0.4 to 0.6 correlations. This dataset serves as a catalyst for new research endeavors on the robustness of objective metrics.

References

[1] 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," *Commun. ACM*, vol. 65, no. 1, p. 99–106, dec 2021.

[2] C. L. Deng and E. Tartaglione, "Compressing explicit voxel grid representations: Fast nerfs become also small," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, January 2023, pp. 1236–1245.

[3] J. Tang, X. Chen, J. Wang, and G. Zeng, "Compressible-composable nerf via rank-residual decomposition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 798–14 809, 2022.

[4] D. Rho, B. Lee, S. Nam, J. C. Lee, J. H. Ko, and E. Park, "Masked wavelet representation for compact neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 20 680–20 690.

[5] 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*, June 2023, pp. 4222–4231.

[6] T. Takikawa, A. Evans, J. Tremblay, T. Müller, M. McGuire, A. Jacobson, and S. Fidler, "Variable bitrate neural fields," in *ACM SIGGRAPH 2022 Conference Proceedings*, ser. SIGGRAPH '22. New York, NY, USA: Association for Computing Machinery, 2022.

[7] H. Liang, T. Wu, P. Hanji, F. Banterle, H. Gao, R. Mantiuk, and C. Oztireli, "Perceptual quality assessment of nerf and neural view synthesis methods for front-facing views," *arXiv preprint arXiv:2303.15206*, 2023.

[8] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: practical view synthesis with prescriptive sampling guidelines," *ACM Trans. Graph.*, vol. 38, no. 4, jul 2019.

[9] P. Martin, A. Rodrigues, J. Ascenso, and M. P. Queluz, "Nerf-qa: Neural radiance fields quality assessment database," in *2023 15th International Conference on Quality of Multimedia Experience*. IEEE, 2023, pp. 107–110.

[10] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, 2017.

[11] P. Martin, A. Rodrigues, J. Ascenso, and M. P. Queluz, "Nerf view synthesis: Subjective quality assessment and objective metrics evaluation," *arXiv preprint arXiv:2405.20078*, 2024.

[12] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, jul 2022.

[13] C. Sun, M. Sun, and H.-T. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5449–5459, 2021.

[14] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 5501–5510.

[15] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 333–350.

[16] "Sketchfab. free 3d models," [Online], https://sketchfab.com/features/free-3d-models/.

[17] K. Yang, Q. Yang, J. Jung, Y. Xu, X. Xu, and S. Liu, "Exploring the influence of view and camera path selection for dynamic mesh quality assessment," in *2023 IEEE International Conference on Multimedia and Expo*, 2023, pp. 2489–2494.

[18] Q. Yang, J. Jung, T. Deschamps, X. Xu, and S. Liu, "Tdmd: A database for dynamic color mesh subjective and objective quality explorations," *arXiv preprint arXiv:2308.01499*, 2023.

[19] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[20] I. R. BT, "500-14, methodologies for the subjective assessment of the quality of television images," *Geneva: International Telecommunication Union*, 2019.

[21] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun, "Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3877–3891, 2021.

[22] B. Cui, Q. Yang, K. Yang, Y. Xu, X. Xu, and S. Liu, "Sjtu-tmqa: A quality assessment database for static mesh with texture map," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024, pp. 7875–7879.

[23] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402 Vol.2.

[25] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[26] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[27] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[29] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[30] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[31] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.

[32] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[33] R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney, "Fovvideovdp: a visible difference predictor for wide field-of-view video," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021.

[34] M. Narwaria, M. P. Da Silva, and P. Le Callet, "Hdr-vqm: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46–60, 2015.

[35] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara *et al.*, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, p. 2, 2016.

[36] F. M. Dekking, *A modern introduction to probability and statistics: Understanding why and how*. Springer Science & Business Media, 2005.

[37] D. Zwillinger and S. Kokoska, *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

[38] J. Antkowiak, T. J. Baina, F. V. Baroncini, N. Chateau, F. FranceTelecom, A. C. F. Pessoa, F. S. Colonnese, I. L. Contin, J. Caviedes, and F. Philips, "Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000," *Final report from the video quality experts group on the validation of objective models of video quality assessment march*, vol. 2000, 2000.

[39] Q. Yang, J. Jung, H. Wang, X. Xu, and S. Liu, "Tsmd: A database for static color mesh quality assessment study," in *2023 IEEE International Conference on Visual Communications and Image Processing*, 2023, pp. 1–5.