A. Ferrarotti et al., "Analysis of Objective 3D Mesh Quality Metrics for Cultural Heritage," 2025 17th International Conference on Quality of Multimedia Experience (QoMEX), Madrid, Spain, 2025, pp. 1-4, doi: 10.1109/QoMEX65720.2025.11219903.

Analysis of Objective 3D Mesh Quality Metrics for Cultural Heritage

Anna Ferrarotti[1], Isabel Rodríguez[2], Javier Usón[2], Sara Baldoni[3], Jesús Gutiérrez[2], Daniel Berjón[2], Francisco Morán[2], Federica Battisti[3], Narciso García[2], Marco Carli[1] and Julián Cabrera[2]

[1]Roma Tre University, Rome, Italy
[2]Universidad Politécnica de Madrid, Madrid, Spain
[3]University of Padova, Padua, Italy

***Abstract*—Extended reality technologies are increasingly used in cultural heritage for preserving and accessing sites and artworks, where 3D model acquisition and rendering are key. Despite progress in reconstruction methodologies, a standardized approach to quality assessment is still missing. This study aims to evaluate objective quality metrics —both image-based and model-based, Full Reference and No Reference— applied to 3D models generated using the Structure from Motion algorithm. By varying parameters such as the number of images, number of triangles, and texture resolution, we examine the impact of these factors on metric outcomes, aiming to assess their reliability in cultural heritage applications.**

***Index Terms*—3D meshes, Quality Assessment, Objective Metrics**

I. Introduction

Extended Reality (XR) technologies are increasingly used in entertainment [1], health care [2], and workplace training [3], with growing adoption in cultural heritage for monitoring [4], preserving, and improving access to historical sites [5], [6]. These applications often rely on reconstructed 3D models of monuments and artworks, which provide the basis for immersive and accurate virtual experiences. Despite the advancements in 3D reconstruction and the increasing availability of quality metrics, there is no standard method to evaluate the perceptual quality of reconstructed models. Existing metrics can be classified as Full Reference (FR) [7]–[9] or No Reference (NR) [10]–[12], and as image-based or model-based, depending on the features exploited for quality assessment. In cultural heritage contexts, reference models are often unavailable, limiting the use of FR metrics. Thus, NR metrics—both image- and model-based—are more suitable. Yet, these metrics offer limited guidance for consumer-level 3D mesh generation, where reconstruction parameters, such as the number of images, number of triangles, and texture resolution, can vary. Subjective evaluations could help, but are time-consuming and might result in inconsistencies due to individual biases and varying perceptions, if the number of involved participants is insufficient. This study focuses on image-based objective metrics and compares them with FR model-based metrics using simulated meshes reconstructions, assessing their responsiveness to parameter changes. The goal is to understand how well current metrics can reflect the differences among 3D models with varying reconstruction parameters. This preliminary work, focused only objective metrics, paves the way towards an automatic pipeline for 3D meshes quality evaluation, helping users designing XR experiences for cultural heritage applications.

II. Proposed Method

While more recent techniques like Neural Radiance Field (NeRF) [13] and 3D Gaussian Splatting [14] offer high-fidelity results [15], their integration into standard 3D pipelines remains complex. Thus, we opted for traditional mesh-based reconstruction, still widely used in cultural heritage applications. The impact of key parameters such as number of images, number of triangles, and texture resolution was tested in terms of quality scores obtained through state-of-the-art quality metrics. Six models were selected from the BASICS dataset [16], which provides point clouds derived from Sketchfab[1]. We focused on the *inanimate objects* and *buildings and landscapes* categories, ensuring relevance to cultural heritage contexts. The chosen models were: (i) Horn of Salt Diggers Brotherhood of Wieliczka, (ii) Palace of Fine Arts, (iii) Mexico City Metropolitan Cathedral, (iv) Schwarzenbach- houses with interior, (v) Roman Temple of Evora, (vi) Kriegerdenkmal. Each selected model was rendered in Blender[2], and 1000 captures were acquired from viewpoints uniformly distributed over a spherical surface, following a spiral from base to top, using a fixed field of view ($42°$) and resolution (3840×2160 Pixels, $\simeq 8.3$ MP). 3D reconstruction was carried out using a pipeline involving OpenMVG [17] and OpenMVS [18]. OpenMVG was used to estimate camera poses and generate a sparse point cloud via Structure from Motion (SfM). Then, Multi-View Stereo (MVS) algorithms in OpenMVS were employed to produce depth maps from the calibrated images. These depth maps were projected to form dense point clouds, which were converted into 3D meshes using Poisson surface reconstruction [19]. Finally, the obtained meshes were simplified through quadric edge collapse decimation and textured using a photometric mapping method [20]. We generated 64

[1]https://sketchfab.com
[2]https://www.blender.org

TABLE I: Mean ± SD evaluated over the 300 test captured images for the selected 3D model reconstruction parameters. The best result is **bold underlined**, the second best in **bold**, the third in underlined, and the worst is unformatted. A ↑ indicates that higher values correspond to better quality, while ↓ signifies the opposite. Texture resolution is indicated in MP.

Parameter		Metric						
		PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	BRISQUE ↓	Hausdorff ↓	L^2 ↓
		Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD
Number of images	50	24.037 ± 2.931	**0.634** ± 0.143	0.124 ± 0.075	**6.710** ± 1.533	70.974 ± 12.962	**0.011** ± 0.003	0.001 ± 0.000
	125	**24.105** ± 3.029	**0.636** ± 0.147	0.121 ± 0.077	**6.759** ± 1.600	69.708 ± 13.500	**0.011** ± 0.004	0.001 ± 0.000
	250	**24.088** ± 3.021	0.632 ± 0.147	**0.120** ± 0.077	6.789 ± 1.615	**69.156** ± 13.656	**0.013** ± 0.005	0.001 ± 0.000
	500	23.907 ± 2.909	0.618 ± 0.145	**0.119** ± 0.077	6.802 ± 1.621	**68.701** ± 13.913	0.014 ± 0.008	0.001 ± 0.000
Number of triangles	5k	22.797 ± 2.458	0.587 ± 0.121	0.127 ± 0.078	**6.759** ± 1.617	71.682 ± 13.629	0.012 ± 0.005	0.001 ± 0.000
	25k	24.062 ± 3.002	0.634 ± 0.149	0.120 ± 0.077	**6.767** ± 1.595	69.789 ± 13.464	0.012 ± 0.006	0.001 ± 0.000
	50k	**24.518** ± 3.048	**0.647** ± 0.153	**0.117** ± 0.076	6.762 ± 1.585	**68.917** ± 13.476	0.012 ± 0.006	0.001 ± 0.000
	100k	**24.759** ± 2.958	**0.651** ± 0.149	**0.118** ± 0.074	6.771 ± 1.575	**68.151** ± 13.329	0.012 ± 0.006	0.001 ± 0.000
Texture resolution	1	22.132 ± 2.381	0.482 ± 0.108	0.181 ± 0.086	7.002 ± 1.187	81.499 ± 9.242	-	-
	4	23.748 ± 2.667	0.607 ± 0.115	0.133 ± 0.069	6.750 ± 1.485	69.811 ± 10.605	-	-
	16	**25.031** ± 2.845	**0.707** ± 0.108	**0.090** ± 0.051	**6.650** ± 1.773	**64.168** ± 12.433	-	-
	32	**25.226** ± 2.897	**0.724** ± 0.108	**0.079** ± 0.046	**6.657** ± 1.820	**63.060** ± 12.918	-	-

TABLE II: P-values for the Kruskal-Wallis test, asterisks indicate statistical significance. A - indicates that the test was not performed.

Parameter	Metric						
	PSNR	SSIM	LPIPS	BRISQUE	NIQE	Hausdorff	L^2
Number of images	0.9648	0.7447	0.9300	0.4001	0.8794	0.0349*	0.8122
Number of triangles	< 0.0001*	0.0036*	< 0.0001*	0.0666	0.9872	0.9824	< 0.0001*
Texture resolution	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	0.0024*	-	-

reconstructed versions per model by varying three parameters: number of input images (50, 125, 250, 500), number of employed triangles (5k, 25k, 50k, 100k), and texture resolution (1, 4, 16, 32 MP). For quality evaluation, we adopted both image-based and model-based objective metrics. For each reconstructed model, 300 views not used in the reconstruction phase were rendered for comparison. FR metrics, i.e., Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [21], and Learned Perceptual Image Patch Similarity (LPIPS) [22], were computed by comparing rendered views from the same camera positions on both the reference and reconstructed models. For NR evaluation, we used Naturalness Image Quality Evaluator (NIQE) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [23], which rely on natural scene statistics to estimate image quality without a reference. To ensure accurate image-based comparison, masks were generated to exclude background pixels. Since reconstructed models may exhibit variations in pose, orientation, and scale, alignment was performed by comparing the SfM-estimated camera parameters with those used for synthetic rendering. This transformation allowed for the alignment of reconstructed meshes with their original counterparts. For each view, the union of the masks from reference and reconstructed images was used to compute metrics. In fact, using the intersection of the reference and reconstructed masks means examining the best-case scenario and can yield misleadingly good results in reconstructions where the two masks are very different. Moreover, the Intersection Over Union (IoU) of the masks across all views was evaluated, achieving an average IoU of 98.44% (±0.79%), which validates this masking methodology. Additionally, we incorporated two geometry-based metrics, i.e., the Hausdorff and L^2 distances [24], which directly compare the 3D structure of the reconstructed mesh to the reference model. These offer a complementary perspective to 2D metrics by measuring spatial deviations in 3D space.

III. Results

Table I presents the mean and Standard Deviation (SD) of both image-based and model-based quality metrics for the 300 test images. In terms of the number of images, the metrics produce mixed results. LPIPS and NIQE suggest that quality improves with more captures, while Hausdorff distance and NIQE indicate the opposite. PSNR, SSIM, and L^2 show no consistent pattern. Regarding the number of triangles, most metrics—such as PSNR, BRISQUE, and LPIPS—demonstrate a general improvement in perceived quality with increasing triangle counts. All metrics identify the 5k triangle models as the lowest quality. The highest quality is in most cases attributed to the 100k triangle models, except for LPIPS, which favors the 50k triangle configuration. Model-based metrics (Hausdorff and L^2) show a similar trend. Notably, NIQE displays a contrasting behavior, suggesting a decline in quality as the number of triangles increases. As for texture resolution, the metrics unanimously (except NIQE) recognize an improvement in quality with higher texture resolutions. NIQE identifies the 16 MP resolution as optimal, followed by 32 MP, diverging from the general trend. To further investigate these differences, statistical tests were carried out. First, the normality of the data distribution was assessed using D'Agostino and Pearson, Kolmogorov–Smirnov, and Shapiro–Wilk tests [25]–[27]. Since the data did not follow a Gaussian distribution, non-parametric Kruskal–Wallis tests [28] were employed to determine whether the reconstruction parameters produced statistically significant differences in metric values. For parameters showing signifi-

TABLE III: P-values for the paired-Wilcoxon tests with Bonferroni correction, asterisks indicate statistical significance. N/A has been used to indicate the cases in which the Kruskal-Wallis test did not report statistical differences. A - indicates that the test was not applicable. Texture resolution is indicated in MP.

Parameter	Comparison	Metric						
		PSNR	**SSIM**	**LPIPS**	**BRISQUE**	**NIQE**	**Hausdorff**	**L^2**
Number of images	50 vs 125	N/A	N/A	N/A	N/A	N/A	1.0000	N/A
	50 vs 250	N/A	N/A	N/A	N/A	N/A	0.2199	N/A
	50 vs 500	N/A	N/A	N/A	N/A	N/A	0.0135*	N/A
	125 vs 250	N/A	N/A	N/A	N/A	N/A	0.6443	N/A
	125 vs 500	N/A	N/A	N/A	N/A	N/A	0.0051*	N/A
	250 vs 500	N/A	N/A	N/A	N/A	N/A	0.7250	N/A
Number of triangles	5k vs 25k	< 0.0001*	< 0.0001*	< 0.0001*	N/A	N/A	N/A	< 0.0001*
	5k vs 50k	< 0.0001*	< 0.0001*	< 0.0001*	N/A	N/A	N/A	< 0.0001*
	5k vs 100k	< 0.0001*	< 0.0001*	< 0.0001*	N/A	N/A	N/A	< 0.0001*
	25k vs 50k	< 0.0001*	< 0.0001*	< 0.0001*	N/A	N/A	N/A	< 0.0001*
	25k vs 100k	< 0.0001*	< 0.001*	0.0420*	N/A	N/A	N/A	< 0.0001*
	50k vs 100k	< 0.0001*	0.6424	1	N/A	N/A	N/A	< 0.0001*
Resolution	1 vs 4	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	-	-
	1 vs 16	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	-	-
	1 vs 32	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	< 0.001*	-	-
	4 vs 16	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	0.0723	-	-
	4 vs 32	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	0.1748	-	-
	16 vs 32	< 0.0001*	< 0.0001*	< 0.0001*	< 0.0001*	0.5398	-	-

cant results (p-value < 0.05), post-hoc pairwise comparisons were conducted using Wilcoxon tests with Bonferroni correction. Table II presents the p-values from the Kruskal–Wallis tests for all considered metrics and parameters, while Table III shows the post-hoc analysis results. The Kruskal–Wallis test identified a significant effect of the number of images only for the Hausdorff distance. Post-hoc tests revealed significant differences between reconstructions using 50 and 500 images, and between 125 and 500 images. This suggests that geometric structure, as captured by model-based metrics, is more sensitive to this parameter than image-based ones. However, the lack of consistent results across other metrics may be due to the uniform spatial distribution of camera positions used in this study. A non-uniform distribution could have resulted in more pronounced differences. Significant differences related to the number of triangles were observed for PSNR, SSIM, LPIPS, and L^2. In particular, PSNR and L^2 detected statistically significant differences across all triangle count pairs. SSIM and LPIPS reported significant differences between all but the 50k and 100k configurations. These findings highlight that FR metrics can detect changes in mesh complexity, although the sensitivity decreases at higher triangle counts. All image-based metrics demonstrated significant variations with respect to texture resolution, as shown in Table II. However, post-hoc results (Table III) revealed that NIQE only detected a difference between the lowest resolution (1 MP) and the higher ones, while failing to distinguish between medium and high resolutions. In conclusion, the statistical analysis supports the reliability of objective metrics in detecting quality differences due to texture resolution and mesh complexity, whereas sensitivity to the number of input images remains limited under uniform acquisition conditions.

IV. Conclusions and future works

This work presents a preliminary evaluation of objective metrics for 3D mesh quality assessment in cultural heritage applications. Six models were reconstructed using SfM and analyzed using both image-based and model-based metrics under varying reconstruction parameters. Results show that image-based metrics are sensitive to texture resolution but less effective in detecting differences related to the number of images or triangles, especially for NR metrics. Although standard metrics such as PSNR and SSIM detect changes in texture resolution and mesh complexity, their reliance on reference images limits their use in cultural heritage scenarios. NR metrics, instead, were mainly sensitive to texture resolution and failed to capture other differences in the reconstruction parameters. However, one of the main limitations of this study is the absence of a subjective evaluation, which prevents verifying the correlation between metric outputs and human perception. Moreover, no investigation was conducted on recent 3D-specific NR quality metrics. Nevertheless, this research represents an initial step towards the development of automatic evaluation tools that are aligned with human perceptual criteria. Such tools would not only enhance the reliability of 3D reconstruction assessments,but also enable non-experts to contribute to immersive cultural heritage applications. Future work will focus on conducting subjective experiments to validate the perceptual relevance of objective metrics and on extending the analysis to real-world 3D reconstructions, as the current study is limited to simulated reconstructions.

Acknowledgment

This work has been partially supported by projects PID2020-115132RB (SARAOS) and PID2023-148922OA-I00 (EEVOCATIONS), funded by MCIN/AEI/10.13039/501100011033 of the Spanish Government, HORIZON-IA-1010702-50 (XReco), funded by the European Union, and UNICO-5G I+D TSI-063000-2021-80 (DISRADIO–Pilotos), funded by the Ministry of Digital Transformation of the Spanish Government and the NextGenerationEU (RRTP).

REFERENCES

[1] E. S. de Lima, B. M. Silva, and G. T. Galam, "Adaptive virtual reality horror games based on machine learning and player modeling," *Entertainment Computing*, vol. 43, p. 100515, 2022.

[2] S. Barteit, L. Lanfermann, T. Bärnighausen, F. Neuhann, and C. Beiersmann, "Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review," *JMIR serious games*, vol. 9, no. 3, p. e29080, 2021.

[3] B. Xie, H. Liu, R. Alghofaili, Y. Zhang, Y. Jiang, F. D. Lobo, C. Li, W. Li, H. Huang, M. Akdere, C. Mousas, and L.-F. Yu, "A review on virtual reality skill training applications," *Frontiers in Virtual Reality*, vol. 2, p. 645153, 2021.

[4] S. Doukianou and V. Lalioti, "Ethical Extended Reality: Bridging Technology and Cultural Heritage," in *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. IEEE, 2024, pp. 301–306.

[5] P. Singh, N. Pahuja, M. Kansal, S. Gurung, U. Shukla, and S. Gupta, "Enhancing Tourism Experiences and Preserving Cultural Heritage with AR and VR," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*. IEEE, 2024, pp. 225–231.

[6] L. Cecere, F. Colace, M. De Santo, A. Lorusso, D. Santaniello, and C. Valentino, "Overview of Cultural Heritage Education and Emerging Technologies," in *2024 IEEE International Humanitarian Technologies Conference (IHTC)*. IEEE, 2024, pp. 1–7.

[7] S. Lee, J. Kang, S. Lee, W. Lin, and A. C. Bovik, "3D-PSSIM: Projective Structural Similarity for 3D Mesh Quality Assessment Robust to Topological Irregularities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[8] Y. Nehmé, J. Delanoy, F. Dupont, J.-P. Farrugia, P. Le Callet, and G. Lavoué, "Textured Mesh Quality Assessment: Large-scale Dataset and Deep Learning-based Quality Metric," *ACM Transactions on Graphics*, vol. 42, no. 3, pp. 1–20, 2023.

[9] Y. Nehmé, F. Dupont, J.-P. Farrugia, P. Le Callet, and G. Lavoué, "Visual quality of 3d meshes with diffuse colors in virtual reality: Subjective and objective evaluation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 3, pp. 2202–2219, 2020.

[10] Z. Zhang, W. Sun, X. Min, T. Wang, W. Lu, and G. Zhai, "No-Reference Quality Assessment for 3D Colored Point Cloud and Mesh Models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7618–7631, 2022.

[11] Z. Zhang, W. Sun, H. Wu, Y. Zhou, C. Li, Z. Chen, X. Min, G. Zhai, and W. Lin, "GMS-3DQA: Projection-Based Grid Mini-patch Sampling for 3D Model Quality Assessment," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 6, pp. 1–19, 2024.

[12] M. El Hassouni and H. Cherifi, "Learning graph features for colored mesh visual quality assessment," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3381–3385.

[13] 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*, vol. 65, no. 1, pp. 99–106, 2021.

[14] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Trans. Graph.*, vol. 42, no. 4, Jul. 2023. [Online]. Available: https://doi.org/10.1145/3592433

[15] A. Guédon and V. Lepetit, "SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5354–5363.

[16] A. Ak, E. Zerman, M. Quach, A. Chetouani, A. Smolic, G. Valenzise, and P. Le Callet, "BASICS: Broad quality assessment of static point clouds in a compression scenario," *IEEE Transactions on Multimedia*, vol. 26, pp. 6730–6742, 2024.

[17] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "OpenMVG: Open multiple view geometry," in *International Workshop on Reproducible Research in Pattern Recognition*. Springer, 2016, pp. 60–74.

[18] "OpenMVS: Open Multi-View Stereo Reconstruction Library," https://cdcseacave.github.io/, 2015, accessed 2025-04-03.

[19] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, vol. 7, no. 4, 2006.

[20] R. Pagés, D. Berjón, F. Morán, and N. García, "Seamless, static multitexturing of 3D meshes," *Computer Graphics Forum*, vol. 34, no. 1, pp. 228–238, 2015.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: from Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

[23] 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, 2012.

[24] N. Aspert, D. Santa-Cruz, and T. Ebrahimi, "Mesh: Measuring errors between surfaces using the hausdorff distance," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 1. IEEE, 2002, pp. 705–708.

[25] R. D'agostino and E. S. Pearson, "Tests for departure from normality. empirical results for the distributions of b 2 and $\sqrt{b}$," *Biometrika*, vol. 60, no. 3, pp. 613–622, 1973.

[26] A. Kolmogorov, "On the empirical determination of a distribution function," *Breakthroughs in Statistics: Methodology and Distribution*, pp. 106–113, 1992.

[27] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[28] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.