Textured Mesh Quality Assessment: Large-scale Dataset and Deep Learning-based Quality Metric

Yana Nehmé, Johanna Delanoy, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, Guillaume Lavoué

Textured Mesh Quality Assessment: Large-Scale Dataset and Deep Learning-based Quality Metric

YANA NEHMÉ, Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, France
JOHANNA DELANOY, Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, France
FLORENT DUPONT, Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, France
JEAN-PHILIPPE FARRUGIA, Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, France
PATRICK LE CALLET, Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, France
GUILLAUME LAVOUÉ, Univ Lyon, Centrale Lyon, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205, ENISE, France

Over the past decade, 3D graphics have become highly detailed to mimic the real world, exploding their size and complexity. Certain applications and device constraints necessitate their simplification and/or lossy compression, which can degrade their visual quality. Thus, to ensure the best Quality of Experience (QoE), it is important to evaluate the visual quality to accurately drive the compression and find the right compromise between visual quality and data size. In this work, we focus on subjective and objective quality assessment of textured 3D meshes. We first establish a large-scale dataset, which includes 55 source models quantitatively characterized in terms of geometric, color, and semantic complexity, and corrupted by combinations of 5 types of compression-based distortions applied on the geometry, texture mapping and texture image of the meshes. This dataset contains over 343k distorted stimuli. We propose an approach to select a challenging subset of 3000 stimuli for which we collected 148929 quality judgments from over 4500 participants in a large-scale crowdsourced subjective experiment. Leveraging our subject-rated dataset, a learning-based quality metric for 3D graphics was proposed. Our metric demonstrates state-of-the-art results on our dataset of textured meshes and on a dataset of distorted meshes with vertex colors. Finally, we present an application of our metric and dataset to explore the influence of distortion interactions and content characteristics on the perceived quality of compressed textured meshes.

CCS Concepts: • **Computing methodologies** → **Perception**; *Mesh models*; *Texturing*; Image-based rendering; **Neural networks**.

Additional Key Words and Phrases: Computer Graphics, Perception, 3D Mesh, Texture, Visual Quality Assessment, Subjective Quality Evaluation, Objective Quality Evaluation, Dataset, Perceptual Metric, Deep Learning, Crowdsourcing

1 INTRODUCTION

The use of 3D graphical data is growing for the general public with the proliferation of acquisition technologies (3D scanners, 360 °cameras, MRI, etc.), intuitive 3D modeling tools, 3D printers, and affordable virtual and mixed reality Head-Mounted Displays -HMD- (Oculus Rift, HTC Vive, Microsoft HoloLens, etc.). All of these technologies make the size and complexity of 3D data explode. The resulting 3D scenes are huge and extremely detailed: they may contain several million geometric primitives, associated with a wide range of appearance attributes, intended to reproduce a realistic material appearance.

Extended reality -XR- (i.e. Augmented and Virtual Reality AR/VR) is seen as the next potential computing platform. The great advantage of XR technologies is that they provide 6 Degrees of Freedom (6DoF) allowing realistic interactions and a high level of immersion. However, the visualization and interaction in XR of large and complex 3D scenes remains an unsolved issue to date due to two major challenges: (1) the complexity of a 3D scene that can be displayed on a HMD is substantially smaller than that on a standard screen, (2) for networked applications, latency problems may occur when streaming the 3D scene on the client device. This problem is growing as more online VR/AR applications consider 3D data stored on remote servers.

To adapt the complexity of 3D content for HMDs (notably for autonomous devices, such as the Occulus Quest 2) and to avoid latency due to transmission, simplification and compression are inevitable. These operations reduce the amount of data (reduce the Level of Details (LoD) and the size of 3D data) and by extension the costs in processing, storage, and transmission. However, such operations are lossy and result in visual degradations that may affect the perceived quality of the 3D scene and, in turn, the user's Quality Of Experience (QoE). It is therefore essential to define measures to accurately assess the impact of these distortions in order to find the right compromise between visual quality and data size/LoD. For this purpose, quality assessment methodologies are required.

The perceptual quality can be assessed using subjective studies and objective metrics. Objective metrics consist in algorithms designed to automatically predict the visual quality loss (i.e. the level of annoyance of visual artifacts). On the other hand, subjective studies, aka. user studies, involve inviting a group of participants to assess the visual quality of test data. These subjective experiments provide the most reliable way to create ground-truth datasets useful for understanding human psychological behavior (when perceiving multimedia content) as well as for benchmarking and tuning objective quality metrics.

Public quality assessment metrics and datasets for 3D graphics are lacking, especially for meshes with color and/or texture attributes. Indeed, existing datasets are rather small and have limited generalization ability, making them not challenging enough to test

Authors' addresses: Yana Nehmé, yana.nehme@gmail.com, Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, F-69621 Villeurbanne, France; Johanna Delanoy, johanna.delanoy@insa-lyon.fr, Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, F-69621 Villeurbanne, France; Florent Dupont, Florent.Dupont@liris.cnrs.fr, Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France; Jean-Philippe Farrugia, jean-philippe.farrugia@univ-lyon1.fr, Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France; Patrick Le Callet, patrick.lecallet@univ-nantes.fr, Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France; Guillaume Lavoué, glavoue@liris.cnrs.fr, Univ Lyon, Centrale Lyon, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205, ENISE, F-42023 Saint Etienne, France.

and benchmark quality metrics and insufficient to train and drive learning-based ones.

In this work, we produce a valuable large-scale quality assessment dataset of textured meshes, with more than 343k distorted meshes generated from the compression of 55 source models. To ensure the generalization ability of our dataset, (1) we devised a set of measures to quantitatively characterize the source models in order to avoid biases related to the selection of these models. (2) We involved mixed compression-based distortions of different natures, such as texture compression, geometry quantization, UV map quantization, LoD and texture sub-sampling. (3) We proposed an approach for the selection of a challenging subset of 3000 stimuli that was subject-rated in a large-scale crowdsourced experiment.

Leveraging our annotated dataset, we propose an image-based deep learning quality metric for 3D graphics, called Graphics-LPIPS (Learned Perceptual Image Patch Similarity). It operates on patches of rendered stimuli that are fed to a convolutional neural network with learning weights on top to extract features. Features are then fused and pooled to predict the quality of the patch. The global quality score of the stimulus is obtained by averaging the quality of local patches. Our metric outperforms other image quality metrics in terms of correlation and classification ability on our dataset of textured meshes and on an existing dataset of meshes with vertex colors.

Finally, our metric allowed us to annotate the remaining stimuli in our dataset to subsequently explore the influence of distortion interactions and content characteristics on the perceived quality of 3D graphics.

The main contributions of our work are as follows:

- We provide the largest dataset of textured meshes with over 343k stimuli generated from 55 source models quantitatively characterized in terms of geometric, color, and semantic complexity to ensure their diversity. The dataset covers a wide range of compression-based distortions applied with different strengths. The database can be used to train no-reference quality metrics and develop rate-distortion models for meshes.
- From the established dataset, we carefully selected a challenging subset of 3000 stimuli that we annotated in a large-scale subjective experiment in crowdsourcing involving over 4500 participants. To the best of our knowledge, it is the largest quality assessment dataset of textured meshes associated with subjective scores to date. This database is valuable for training and benchmarking quality metrics.
- We propose an image-based metric, named Graphics-LPIPS, for assessing the quality of rendered 3D graphics. It employs convolutional neural networks. Our metric demonstrates state-of-the-art results on two different datasets.
- Leveraging our whole dataset and metric, we provide an in-depth analysis on the effect of each distortion and their combinations on the perceived quality of textured meshes. We also evaluate the influence of the geometric and color complexity of the model on the perception of distortions.

The datasets and the source code of the metric are publicly available [1].

The rest of this paper is organized as follows: Section 2 reviews previous work on subjective and objective quality assessment of 3D graphics. Section 3 describes our dataset and the set of measures we propose for 3D content characterization. Section 4 details the subjective experiment and the process adopted to select the subset of test stimuli. We describe in section 5 our learning-based quality metric and evaluate its performance. Section 6 presents an application of the metric and the dataset. Concluding remarks and perspectives are provided in section 8.

2 RELATED WORK

In this section, we review previous work on subjective and objective quality assessment of graphical 3D content. We provide an overview of existing datasets and metrics for predicting the visual impact of distortions applied on such data. Note that, 3D data can be represented in different ways (3D meshes, point clouds, voxels), with and without appearance attributes (color, texture, material, etc.). We are specifically interested in 3D meshes and point clouds with color/texture attributes.

2.1 Subjective quality assessment

Subjective quality tests involving 3D models were initially introduced on meshes, more precisely on geometry-only models, to assess the artifacts induced by the simplification, smoothing, watermarking and compression [Christaki et al. 2018; Corsini et al. 2007; Lavoué 2009; Torkhani et al. 2015; Vanhoey et al. 2017; Váša and Rus 2012; Watson 2001]. Little work considered meshes with color attributes (vertex color or texture) [Guo et al. 2016; Gutiérrez et al. 2020; Nehmé et al. 2020; Nehmé et al. 2021b; Pan et al. 2005; Zerman et al. 2020].

Subjective quality experiments involving point clouds have become prevalent over the last six years. The pioneering studies were conducted on colorless point cloud content [Alexiou and Ebrahimi 2017; Alexiou et al. 2017; Javaheri et al. 2017; Javaheri et al. 2019; Su et al. 2019]. Subsequently, the majority of the studies focused on evaluating the quality of colored point clouds [Alexiou et al. 2019; Liu et al. 2021a; Perry et al. 2020; Subramanyam et al. 2020; Torlig et al. 2018; Wu et al. 2021; Yang et al. 2020; Zerman et al. 2019, 2020].

The aforementioned works considered different subjective methodologies, different ways to display the 3D models to participants (still images, animations, interactive scenes) and different test equipment (2D screen, augmented reality and virtual reality headsets).

The experimental methodologies used were inspired by existing image/video subjective methodologies. They are mainly derived from single stimulus methods, in which participants see only one stimulus and rate its quality [Corsini et al. 2007; Gutiérrez et al. 2020; Subramanyam et al. 2020; Torkhani et al. 2015; Viola et al. 2022; Zerman et al. 2020], double stimulus methods in which participants rate the visual degradation after seeing the reference and distorted stimuli [Lavoué 2009; Nehmé et al. 2021b; Perry et al. 2020; Su et al. 2019; Torlig et al. 2018; Watson 2001], and pairwise comparison methods in which participants choose the better quality stimulus

[1] https://github.com/MEPP-team/Graphics-LPIPS

from two stimuli presented to them [Alexiou et al. 2019; Christaki et al. 2018; Guo et al. 2016; Vanhoey et al. 2017; Váša and Rus 2012]. Recently, a couple of comprehensive/comparative studies [Alexiou and Ebrahimi 2017; Nehmé et al. 2020] evaluated the impact of the subjective methodologies on the obtained quality scores. They found that the double stimulus method is the most suitable to assess the quality of 3D graphics.

Some researchers have implemented real-time interactive scenes, allowing participants to freely interact (rotate and zoom) with the 3D models being rated (real-time interactive inspection) [Alexiou and Ebrahimi 2017; Alexiou et al. 2017, 2020; Gutiérrez et al. 2020; Mekuria et al. 2017; Subramanyam et al. 2020; Torlig et al. 2018; Wu et al. 2021; Yang et al. 2020]. These experiments are conducted on desktop devices as well as in immersive environments with varying Degrees of Freedom (DoF). Other researchers have controlled the viewpoints visualized by the participants to ensure the same user experience (passive inspection). They used 2D still images or predefined camera paths to generate videos of the models [Guo et al. 2016; Javaheri et al. 2017; Nehmé et al. 2021b; Pan et al. 2005; Rogowitz and Rushmeier 2001; Su et al. 2019; Zerman et al. 2019, 2020]. This approach avoids cognitive overload that can alter human judgments. The effect of adopting different modes of inspection for subjective quality assessment is still unclear/an open question as very few comparisons have been made to date [Torkhani et al. 2015; Viola et al. 2022].

Most of the reported experiments for both meshes and point clouds were conducted on desktop settings. Only the studies presented in [Alexiou et al. 2020; Christaki et al. 2018; Nehmé et al. 2020; Nehmé et al. 2021b; Subramanyam et al. 2020; Viola et al. 2022; Wu et al. 2021] considered a VR environment, and those presented in [Alexiou et al. 2017; Gutiérrez et al. 2020] a AR environment. An early attempt of 3D tele-immersion was reported in [Mekuria et al. 2017]. So far, no work has been done to understand the impact of display devices on the perceived quality of 3D content.

The experiments presented above were conducted in laboratories. In recent years, CrowdSourcing (CS) experiments have gained popularity, as they are relatively fast and are therefore more practical for evaluating large-scale datasets. A recent study investigated whether a crowdsourcing test can achieve the accuracy of a laboratory test for the quality assessment of 3D graphics [Nehmé et al. 2021a]. The results showed that under controlled conditions and with a proper participant screening approach, a crowdsourcing experiment based on the double stimulus method can be as accurate as a laboratory experiment (based on the same methodology). Another crowdsourcing study evaluated the perception of compression distortions on point clouds [Lazzarotto et al. 2021].

Several works presented above have publicly released their datasets. Table 1 lists the publicly available quality assessment datasets for 3D content with color/texture attributes. There is a lack of large-scale 3D content datasets, especially those for meshes with color attributes, either in the form of vertex colors or texture maps. Existing datasets are rather small: they contain only few hundreds distorted stimuli, which is not sufficient to drive deep learning metrics that rely on the richness and generality of datasets. In this work, we produce the largest quality assessment dataset of textured meshes to date. In total, more than 343k distorted meshes were generated, of which 3000 are associated with subjective Mean Opinion Scores (MOS) derived from a large-scale subjective experiment conducted in crowdsourcing. Quality scores of the remaining stimuli were predicted (Pseudo-MOS) using a proposed quality metric based on deep learning, trained and tested on the subset of annotated stimuli. Our large dataset allowed us to analyze the impact of the distortions and model characteristics on the perceived quality of textured meshes.

Table 1. Publicly available quality assessment datasets for meshes and point clouds.

Dataset	3D Representation	Attributes	# Stimuli rated
LIRIS Textured Mesh [Guo et al. 2016]	Mesh	Texture maps	• 100×2 renderings • 36×2 renderings
3D Meshes with Vertex Colors [Nehmé et al. 2021b]	Mesh	Vertex colors	480
M-PCCD [Alexiou et al. 2019]	Point cloud	Colored	• 240 • 40 & 30
IRPC [Javaheri et al. 2019]	Point cloud	• 2×Colorless • Colored	• 54 • 54
WPC [Su et al. 2019]	Point cloud	Colored	740
VsenseVVDB [Zerman et al. 2019]	Point cloud	Colored	32
VsenseVVDB2 [Zerman et al. 2020]	• Point cloud • Mesh	• Colored • Texture maps	• 136 • 28
ICIP2020 [Perry et al. 2020]	Point cloud	Colored	96
PointXR [Alexiou et al. 2020]	Point cloud	Colored	40
SJTU-PCQA [Yang et al. 2020]	Point cloud	Colored	378
SIAT-PCQD [Wu et al. 2021]	Point cloud	Colored	340
LB-PCCD [Lazzarotto et al. 2021]	Point cloud	Colored	105
2DTV-VR-QoE [Viola et al. 2022]	Point cloud	Colored	72
LS-PCQA [Liu et al. 2021a] Available after publication	Point cloud	Colored	• 1320 (MOS) • 22704 (Pseudo-MOS)
Our dataset	Mesh	Texture maps	• 3000 (MOS) • 340750 (Pseudo-MOS)

2.2 Objective quality assessment

Simple geometric measures, such as Hausdorff distance [Aspert et al. 2002], Root Mean Squared (RMS) error, and Peak Signal-to-Noise Ratio (PSNR), are only weakly correlated with the human vision since they are based on pure geometric distances and ignore perceptual information [Lavoué et al. 2016; Zerman et al. 2019]. Therefore, many perceptually-driven visual quality metrics have been proposed for meshes and point clouds.

The most popular of these metrics are based on top-down approaches. They treat the Human Visual System (HVS) as a black box and identify changes in content features induced by distortions to estimate perceived quality. They are mostly full-reference and work in a similar way by first establishing the correspondence between the reference and degraded models, after which a set of local feature errors are computed locally (over a neighborhood around each point/vertex) and then pooled into a global quality score. Early metrics developed evaluate only geometric distortions such as MSDM2 [Lavoué 2011], DAME [Váša and Rus 2012], FMPD [Wang et al.

2012], and TPDM [Torkhani et al. 2014] for meshes, and the point-to-point, point-to-plane and plane-to plane distances [Alexiou and Ebrahimi 2018; Tian et al. 2017] and PC-MSDM [Meynet et al. 2019] for point clouds. The following works pioneered the development of metrics for 3D content with color attributes, some of which were designed for meshes [Guo et al. 2016; Nehmé et al. 2021b; Tian and AlRegib 2008] and the majority for point clouds such as PCQM [Meynet et al. 2020], Hist_Y [Viola et al. 2020], GraphSIM [Yang et al. 2020; Zhang et al. 2021], Point-to-distribution [Javaheri et al. 2020].

With the rise of machine learning, a new category of quality metrics has emerged. These metrics are based on a purely data-driven approach, and do not rely on any explicit model. They learn/optimize the weights of geometric and color descriptors using mainly regression [Chetouani 2018; Nehmé et al. 2021b; Yildiz et al. 2020]. More recently, deep learning approaches are gaining in popularity. They allow, among other benefits, the emergence of no-reference methods [Nouri et al. 2017]. Convolutional Neural Networks (CNNs) were investigated and adjusted to assess the quality of both meshes and point clouds. In [Abouelaziz et al. 2017], the CNN was fed with perceptual hand-crafted geometric features extracted from the 3D mesh and presented as 2D patches. An end-to-end sparse CNN was designed to develop a no-reference quality metric for colored point clouds [Liu et al. 2021a]. A recent work computed the perceptual loss for point clouds using an auto-encoder architecture based on convolution layers [Quach et al. 2021].

The field of quality assessment of 3D content, especially those with color attributes (either in the form of texture maps or vertex/point colors), can still be considered to be in its early stages compared to that of images. Thus, many image-based approaches have been proposed for the quality assessment of 3D data, whether meshes [Caillaud et al. 2016; Zhu et al. 2010] or point clouds [He et al. 2021; Wu et al. 2021; Yang et al. 2020], using existing well-known Image Quality Metrics (IQMs), such as SSIM (and its derivatives) [Wang et al. 2004], iCID [Preiss et al. 2014], HDR-VDP2 [Mantiuk et al. 2011], VIF [Sheikh and Bovik 2006], etc. That is, IQMs are applied to projected views (rendered snapshots) of 3D models allowing a complete capture of geometric and color distortions as reflected in the final rendering as well as environmental and lighting conditions. Lately, several authors have exploited CNNs to assess the quality of 3D content using image-based approaches. For instance, PQA-Net [Liu et al. 2021c] is a deep neural network for no-reference quality evaluation of point clouds that extracts features from multiple views using CNNs. The features are then fused and fed to a distortion identifier and a quality predictor. Another network, proposed in [Tao et al. 2021] for colored point clouds, uses a feature extractor composed of sequential CNNs to extract multiscale features from geometry and color patches separately. The final quality score is then obtained as a weighted average across all patches. For meshes, a recent metric was devised by extracting feature vectors from 3 different CNN models and combining them [Abouelaziz et al. 2020]. It uses a patch-selection strategy based on mesh saliency to give more importance to perceptual relevant/attractive regions. The existing works considered geometry-only meshes (without color/texture attributes). In this work, we propose a learning-based quality metric for textured meshes.

Over the last years, convolutional neural networks have successfully rivaled traditional image quality metrics [Bosse et al. 2018; Gao et al. 2017; Talebi and Milanfar 2018; Zhang et al. 2018]. Readers can refer to [Tariq et al. 2020] for a comprehensive study determining why deep features are good predictors of image quality. For 3D content, it is still difficult/challenging to develop quality metrics based on deep learning, mainly due to the lack of large and rich datasets of 3D objects, especially those with color attributes, as mentioned previously in subsection 2.1. Given our large-scale dataset of 3000 textured mesh, we propose a learning-based metric, called Graphics-LPIPS, for assessing the quality of rendered 3D graphics. The metric is an extension of the LPIPS metric [Zhang et al. 2018] originally designed for images and perceptual similarity tasks, which we adapted for 3D graphics and quality assessment tasks. Our metric employs pre-trained CNN with learning linear weights on top that we fed with patches of rendered images of 3D models. Our metric provides a good stability and excellent results in terms of correlation and classification ability on two different datasets.

3 DATASET GENERATION

We produced a large-scale textured meshes quality assessment dataset composed of 343750 distorted meshes derived from 55 source models each associated with 6250 distorted versions. Distortions represent combination of level of details simplification, and texture and geometry compression. The dataset covers a wide range of geometric, color and semantic characteristics. Indeed, each source model has been carefully selected and characterized as will be shown in this section.

3.1 3D source model selection

We collected 55 textured 3D models from SketchFab[2]. The selection was done manually and carefully to get high quality textured meshes with creative commons licenses. Table 2 lists the models, their number of vertices and semantic category, while Figure 1 illustrates them.

Some models were cleaned up to repair topological and geometrical defects (zero-area triangles, non-manifold geometry, holes, etc.). In addition, all models were converted to a unique format: the meshes are provided as OBJ (+ the material file), and the textures as JPEG images of 2K resolution (normalized texture size: 2048×2048). The textures encode surface colors (i.e. diffuse map); other information such as surface normals, roughness, and ambient occlusion are ignored. For models with multiple texture images, these were baked into one single image. More details about the data preparation can be found in supplemental material.

3.2 Content characterization

Our goal is to create a high diversity/generality dataset and avoid biases related to the selection of source models. To this end, we proposed an approach to quantitatively characterize the geometric, color and semantic complexity of 3D models.

The characterization of 3D models is not as straightforward as it seems due to the multimodal nature of these data (geometry, color/texture and material information). In the field of images and

[2]https://sketchfab.com/features/free-3d-models

Fig. 1. The 3D graphic source models constituting our database.

videos, the content characterization and classification is usually based on the Spatial perceptual Information (SI) [Engelke et al. 2009; ITU-T P.910 2008; Yu and Winkler 2013]. It is an indicator of edge energy, i.e. it emphasizes regions of high spatial frequency that correspond to edges: an image I is converted to gray-scale, then filtered with horizontal and vertical Sobel kernels. The standard deviation (std) over the pixels of the Sobel-filtered image is finally computed (Eq. 1).

$$SI = std_{space}[Sobel(I)] \tag{1}$$

Since there is no standard on how to characterize 3D content and inspired by the SI of images, we proposed two new measures to characterize the color and the geometry of textured 3D models. We also introduced a third measure, based on the visual attention complexity [Abid et al. 2020] (presented later in this subsection), to characterize the models regarding their semantic aspect. Our measures are based on rendered images, which means they are sensitive to the chosen viewpoint. However, for most objects, our human judgment will be mainly focused on the most informative viewpoint i.e. the one containing the most color, geometry, and semantic information. One way to reflect this tendency would be to compute our measures on a variety of viewpoints and take their maximum

Table 2. List of source models, their number of vertices and semantic category.

Model ID	#Vertices	Semantic category	Model ID	#Vertices	Semantic category
#1	357364	Animal statue	**#29**	119038	Food
#2	123189	Human statue	**#30**	114145	Decoration
#3	395490	Furniture	**#31**	16803	Musical instrument
#4	99984	Machine	**#32**	358684	Sculpture
#5	198683	Decoration	**#33**	155931	Animal statue
#6	258490	Human statue	**#34**	486850	Building
#7	483746	Human statue	**#35**	150006	Bust
#8	250723	Sculpture	**#36**	249439	Mean of transport
#9	189633	Animal	**#37**	130016	Sculpture
#10	109929	Machine	**#38**	304435	Bust
#11	134472	Electronic device	**#39**	100890	Sculpture
#12	17560	Musical instrument	**#40**	124936	Plant
#13	75950	Animal skeleton	**#41**	74819	Animal
#14	209609	Animal	**#42**	150498	Decoration
#15	77924	Animal statue	**#43**	62989	Decoration
#16	669346	Animal	**#44**	36204	Animal
#17	393652	Animal	**#45**	650778	Sculpture
#18	27611	Food	**#46**	4999	Food
#19	308944	Sculpture	**#47**	110819	Book
#20	185416	Sculpture	**#48**	56902	Decoration
#21	149906	Apparel	**#49**	92223	Mean of transport
#22	74662	Animal	**#50**	260670	Decoration
#23	20144	Plant	**#51**	60710	Mean of transport
#24	297988	Decoration	**#52**	635206	Animal statue
#25	151002	Sculpture	**#53**	150000	Decoration
#26	98763	Machine	**#54**	299976	Animal
#27	77027	Mean of transport	**#55**	125002	Animal
#28	306933	Animal skeleton			

value. For the sake of simplicity, and because we will further use this main viewpoint to train our network, we manually chose the main viewpoint for each model (shown in Figure 1) and compute our measures on renderings from this viewpoint. We conducted experiments (detailed in the supplemental material) that show that this manual viewpoint selection leads to very similar results to max-pooling measures taken from 4 viewpoints around the objects. The proposed characterization strategy can thus be applied in both cases (automatic viewpoint sampling + max-pooling or manual viewpoint selection) with similar results.

A. Geometric characterization To enhance the geometry of the model, we compute a rendering that only takes into account the shading of the model without any of its texture information. Then, we calculate the SI of the rendered snapshot. We thus obtain SI_{Geo} an objective indicator/measure that assesses the geometric complexity of the 3D model. See Figure 2.a.

B. Color characterization To assess the textural/color characteristics, we consider this time a rendering without any shading; just the colors from the texture are displayed. We apply the Sobel filter on the rendered snapshot and remove the silhouette detected by the filter because it rather reflects geometric information than color information, and is already taken into account by SI_{Geo}. The color characterization is thus defined by the SI_{Col} measure computed on the obtained filtered image. See Figure 2.b.

C. Semantic characterization Semantic information cannot be characterized by the SI. We therefore employed the Visual Attention

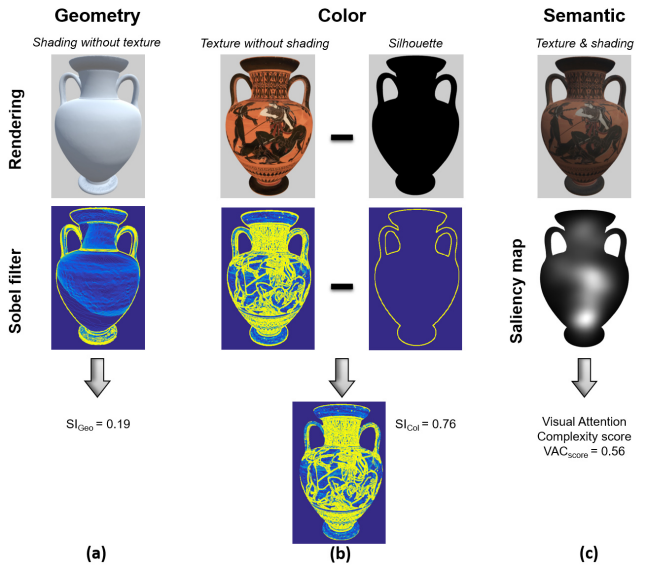


Fig. 2. Characterization of the geometry, color, and semantic of textured 3D models.

Complexity (VAC) measure, recently proposed in [Abid et al. 2020], which is perfectly adapted for this task. The VAC measure consists in evaluating the dispersion of salient regions of the rendered 3D model. Rendered images of 3D models associated with low VAC scores indicate that there are highly salient regions that attract human visual attention (focused gaze behavior), while images with high VAC scores indicate that the saliency is diffused and not focused on one region (overall gazing behavior).
The VAC is computed as follows and illustrated in Figure 2.c: First, we render the 3D model with shading and texture attributes under its main viewpoint. Once the snapshot of the final rendered object is generated, we compute the saliency map (which represents the probability of gazing at each pixel) using the "Salicon" computational model as recommended in [Abid et al. 2020]. Finally, we compute a conditional entropy on the saliency map. Thus, we obtain the VAC_{score} which characterizes the visual attention complexity of the 3D model and is closely linked to a semantic value.

We applied these measures on our 55 selected models. The measures were computed on snapshots of the models having the same resolution. The obtained measures, normalized between 0 and 1, are shown in Figure 3. As it can be seen, the source models of our dataset have various geometric, color and semantic characteristics. Figure 3.a shows a good distribution on the SI_{Geo}/SI_{Col} plane: models located at the right-bottom corner of the plane, such as the model #42, present a rich texture with very low geometric complexity. On the contrary, the model #35 in the top-left corner of the plot is monochrome but has sharp edges and many small geometric details to depict its hair and face.
Figure 3.b shows that our models cover different semantic aspects. For instance, model #1 and #35 have a low VAC_{score} indicating that the visual attention of the participants will probably be focused on the salient regions of these models that are the epigraph (the writing more generally) and the face, respectively. On the other hand, there are no particularly salient regions on models #14 and #49. These models exhibit low visual attention complexity.

The set of measures we proposed reveals the characteristics of 3D models in terms of geometry, color and semantics. These measures are extremely fast to compute and work consistently for both coarse and dense 3D data (regardless of the 3D data representation and the number of vertices).

3.3 Distortions

From the 55 source models, we created 343750 distorted versions generated from combinations of 5 real-world distortions. The distortions represent lossy compression and simplification algorithms applied on the geometry, texture mapping and texture image.

- **Level of Detail (LoD) simplification**: we applied a surface simplification algorithm based on iterative edge collapse and quadric error metric. This algorithm takes into account both geometry and texture mapping [Cignoni et al. 2008; Garland and Heckbert 1998]. We generated 10 levels of simplification ($LoD_{simpL} \in [L1, L10]$) by uniformly reducing the number of mesh faces so that the mesh of the most degraded level ($LoD_{simpL} = L10$) has around 2000 faces (regardless of the source model). Thus, $\Delta_{simpL} = (NbFaces_0 - NbFaces_{min})/10$ where Δ_{simpL} is the number of faces removed at each LoD_{simpL} level, $NbFaces_0$ is the number of faces of the source model, and $NbFaces_{min} = 2000$.
- **Quantization**: we uniformly quantized the position of the vertices as well as the coordinates of the texture using Draco[3], an open-source library for compressing and decompressing 3D geometric meshes and point clouds. To generate the compressed meshes, we considered 5 levels for each attribute:
  - The quantization bits for the position attribute qp range from 7 to 11 bits ($qp \in [7, 11]$).
  - The quantization bits for the texture coordinates attribute (a.k.a UV map) qt range from 6 and 10 bits ($qt \in [6, 10]$). The UV map represents the parametrization defined to map the texture image onto the model surface.
- **Texture map sub-sampling**: we reduced the size of the texture maps by resampling them using the Lanczos low pass filter. We generated 5 texture sizes (T_S): 2048×2048 (the original size), 1440×1440, 1024×1024, 712×712, 512×512.
- **Texture compression**: we used the JPEG compression which is the most commonly used algorithm for lossy 2D image compression. We selected 5 texture qualities (T_Q) obtained by varying the compression level: 90 (the best quality considered but the least effective compression), 75, 50, 25, 10 (the lowest texture quality and the highest compression).

We note that for each distortion type, the degradation levels were selected to cover a range of high, medium and low quality distorted meshes.

By combining/mixing all geometry and texture distortions, we obtained 10 $LoD_{simpL} \times 5$ $qp \times 5$ $qt \times 5$ $T_S \times 5$ $T_Q = 6250$ distortions per model, a.k.a. Hypothetical Reference Circuits (HRCs) according

[3] https://github.com/google/draco

to [VQEG 2010]. HRCs denote the processing operations applied to the source models to obtain the distorted versions. Each HRC is associated with a size (in Bytes), resulting from the compression of the source model with the corresponding distortion parameters (using JPEG for the texture and Draco encoder for the connectivity, geometry and UV maps). Thus, the size of a stimulus (in Bytes) is equal to the sum of the size of its compressed texture and its compressed 3D model.

The distortions were applied systematically with the same levels to all models. Our dataset thus contains 343 750 distorted stimuli (55 source models × 6250 HRCs) that span a great diversity in terms of visual content and quality. Figure 4 shows some visual examples of distorted stimuli along with their distortion parameters. More examples are provided in the supplementary material.

4 SUBJECTIVE EXPERIMENT

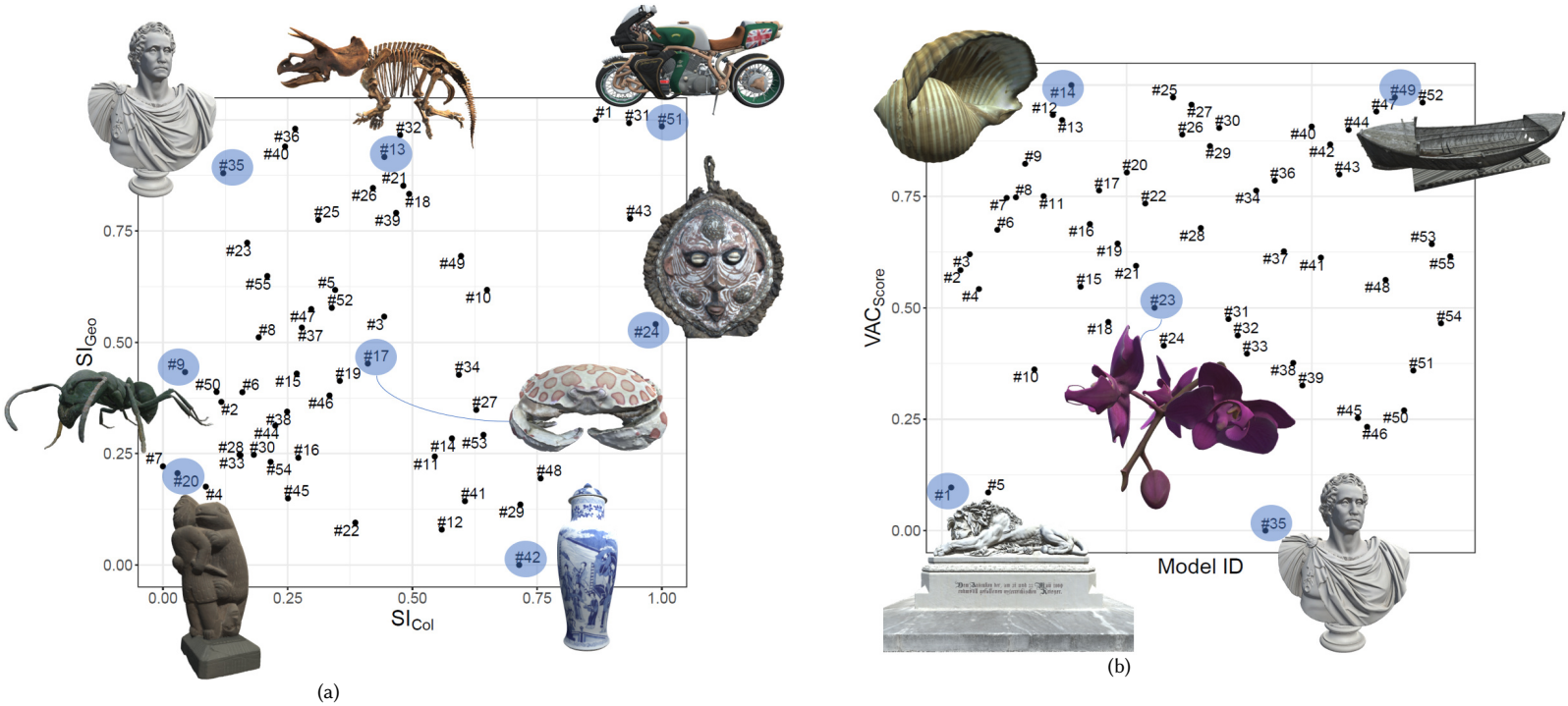
We conducted a large-scale subjective quality assessment experiment in CrowdSourcing (CS), wherein 4513 participants were involved to rate the perceived quality of a generalized and challenging subset of 3000 stimuli carefully selected from the dataset presented above. Over 148k quality scores were collected. This section describes our extensive online subjective study.

4.1 Test stimuli selection

Our dataset contains more than 343k stimuli. Participants cannot be asked to rate the quality of such a large amount of data. We therefore had to select a subset of stimuli to rate in the subjective experiment. The selection of this subset is a crucial step since we aim to use it later to train a learning-based metric. Thus, we seek to obtain an unbiased, generalized and challenging subset, which leads to several selection criteria: the selected subset had to contain all the source models, as well as an equal distribution of HRCs (combinations of distortions created). In addition, the subset of stimuli must equitably cover the entire range of quality (from imperceptible to very annoying distortions) to have a balanced representation of the visual quality. Last but not least, we want this subset to be challenging for objective quality metrics.

We selected 3000 stimuli from 343750 (which represents about 0.9% of the total dataset). To do so, we developed the following approach.

- First, we predicted the Mean Opinion Score (MOS) of all the 343750 stimuli, using existing objective quality metrics that we calibrated on an existing subjectively-rated quality assessment dataset: we used the dataset of meshes with vertex colors [Nehmé et al. 2021b]. We fitted two logistic regression models (mapping functions) between the MOSs of this dataset and the following two objective quality metrics: (1) HDR-VDP2 [Mantiuk et al. 2011] since it provided the best performance among the Image Quality Metrics (IQMs) tested on this dataset [Nehmé et al. 2021b] and (2) LPIPS (Learned Perceptual Image Patch Similarity) [Zhang et al. 2018] since it is a commonly used IQM, based on pre-trained CNNs, with many successful applications [Huang et al. 2018; Yang et al. 2018].
  We computed HDR-VDP2 and LPIPS on snapshots of the stimuli in our dataset rendered from their main viewpoints (defined in



(a) (b)

Fig. 3. (a) Geometry and color spatial information and (b) visual attention complexity for the selected source models.

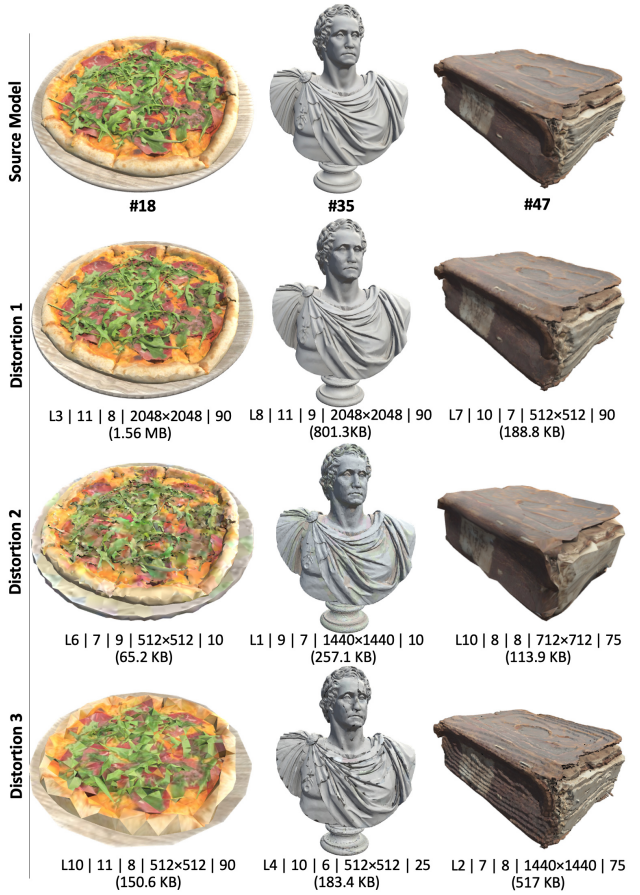


Fig. 4. Some examples of distorted models associated with their size (in KB), from barely visibles distortions (Distortion 1) up to very annoying ones (Distortion 3). Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q.

subsection 3.2), and then predicted their MOS using the obtained regression models. As in [Liu et al. 2021a; Wu et al. 2020], we refer to the predicted MOSs as Pseudo-MOSs. Thus, we obtained 2 pseudo-MOS values per stimulus (pseudo-MOS$_{HDRVDP}$ and pseudo-MOS$_{LPIPS}$).

We used two metrics (instead of one) to be able to sample stimuli for which the metrics do not agree on their quality (as shown later in this subsection and in Figure 5), resulting in a more challenging subset of stimuli.

- Second, to ensure a good and equitable coverage of the whole visual quality range and to get a subset of challenging stimuli, we regularly sampled the plane formed by the 2 pseudo-MOSs, shown in Figure 5, while respecting 2 constraints: considering HDR-VDP2 as the pivot metric, we uniformly sampled the area of each quality range. For each sampled point, we selected the "closest" set of stimuli (in terms of distance), and then chose the one that ensures an equal/equitable distribution of (1) all source models (e.g., as many degraded stimuli for model ID_i as for model ID_j) and (2) all levels of each distortion(e.g, almost as many stimuli are selected with a $qp = 7$ as those with a $qp = 10$).

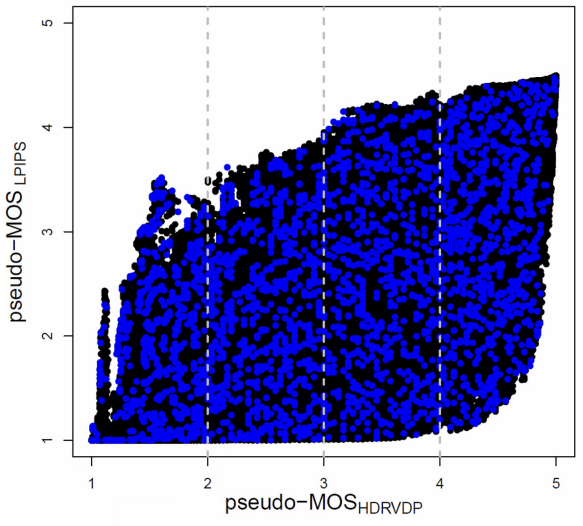


Fig. 5. Selection of the test stimuli by constrained sampling of the plane formed by 2 pseudo-MOSs. The black dots refer to the pseudo-MOS values of all stimuli in the dataset, while the blue dots refer to those selected for the subjective experiment.

4.2 Rendering

To adequately assess the visual quality of 3D content, it is important that the object moves so that the observer can see the whole object and not focus on one single viewpoint [Rogowitz and Rushmeier 2001]. Thus to avoid manual selection of multiple relevant viewpoints for each model, we animated all models in our dataset with a full rotation (360 °) around their vertical axis. We rendered the dynamic test stimuli in a neutral room (light gray walls), without shadows and under a directional light coming from the top right of the room. All models were approximately the same size and rendered with a lambertian material; mipmapping was activated.

We designed the experiment based on the Double Stimulus Impairment Scale (DSIS) method, in which observers see the source model and the same model impaired side by side and rate the impairment of the second stimulus in relation to the source model using a five-level impairment scale, displayed after the presentation of each pair of stimuli. This method has proven to be the most accurate and stable for evaluating the quality of 3D graphics [Nehmé et al. 2020]. Since the experiment is conducted in crowdsourcing, we generated videos of the rendered stimuli in order to limit the participant's interactions with the 3D objects, since we have no full control over the participant's test environment. The only interaction required by the participant is the selection of the score when rating. The videos were all in 650×550 resolution (so that the videos of the source and degraded models fit simultaneously on a screen with a minimum resolution of 1920×1080) with a frame rate of 30 fps and encoded using H.264 encoder (mp4 container) at a bitrate of 5 Mbps to ensure

imperceptibility of compression artifacts. Videos are 8 seconds long, which is the time it takes for models to complete the full rotation.

4.3 Experimental environment

To obtain reliable and controlled results, we used our own web platform to present the subjective experiment to participants. This platform has proven its effectiveness in achieving the accuracy of a laboratory test and producing reliable results [Nehmé et al. 2021a]. The crowdsourcing service was used only to recruit participants. Our platform, illustrated in Figure 6, is suitable for presenting videos according to the DSIS method. Only a web browser with an MPEG-4 decoder is required to run the experiment; no other software needs to be installed. The platform first checks the compatibility of the participant's device: minimum screen resolution of 1920×1080, page zoom level, maintain full screen mode throughout the experiment. The test instructions are then displayed to the participant with a progress bar, at the bottom of this page, showing the status of the loading process of all the video pairs that will be used in the test. This way, the videos of the source and distorted models are played simultaneously during the test, without any latency or unintended interruptions. When the loading is completed a start button appears leading to the test. These steps/windows are illustrated in the supplementary material. The pairs of videos are displayed in a random order to each participant. Participants cannot replay the videos or provide their score until the videos have been played completely. There is no time limit for voting and videos of the stimuli are not shown during that time. At the end of the experiment, participants will receive unique codes allowing them to get their remuneration.

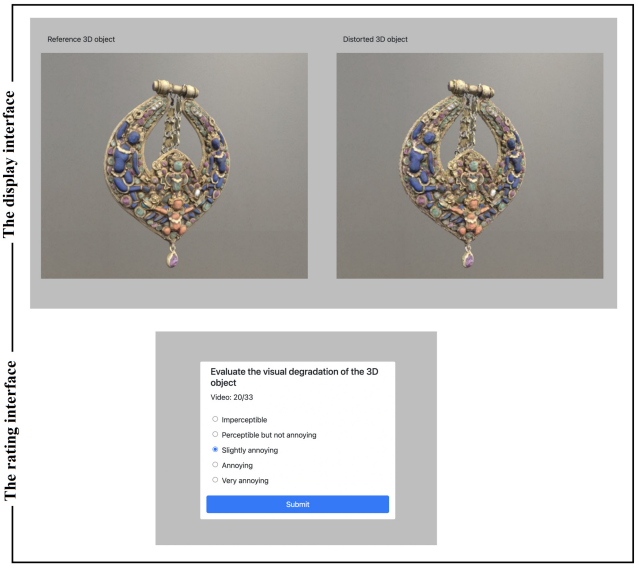


Fig. 6. The graphical interface of the subjective experiment platform.

4.4 Test sessions & participants

Creation of test sessions. The number of stimuli and the duration of the experiment should be limited as much as possible in crowdsourcing to keep participants motivated and to avoid unreliable results [Jiménez et al. 2018; Redi et al. 2015; Reimann et al. 2021]. Therefore, we divided our subset of 3000 test stimuli into 100 playlists. Each participant rates one playlist, i.e., only 30 stimuli. This way, we stay within the duration of the experiment in [Nehmé et al. 2021a] and within that of a pilot experiment we conducted to evaluate the number of scores needed per stimulus to have the same confidence intervals as for a lab study (see supplementary material). The test stimuli were fairly distributed among the playlists so that each playlist contains a maximum diversity of source models (a source model is repeated a maximum of 2 times in a playlist). Additionally, each playlist spans the full range of distortions and all playlists have nearly the same pseudo-MOS distribution.

As in [Ghadiyaram and Bovik 2016], we injected 3 Golden Units (GU, a.k.a trapping stimuli) into each playlist to facilitate the detection of unreliable participants later. The golden units (GUs) included (1) a very poor quality stimulus (GU_{poor}) , (2) a very high quality stimulus (GU_{high}) and (3) a stimulus displayed twice (GU_{rep1} and (GU_{rep2})) to assess the participant's consistency (coherence of his/her scores). Participants who fail to answer correctly the golden units are considered outliers and their scores are discarded.

Training. The experiment started with training. In order to familiarize participants with the task and the rating scale, we selected 5 stimuli not included in the 3000 stimuli test and all referring to the same source model. Each stimulus represented one level of the five-level scale of the DSIS method. After displaying each pair of training videos for 8 sec, the rating interface is displayed for 5 sec and the proposed score assigned to this distortion is highlighted. Once the training is completed the actual test began.

Duration. The test session of our experiment consisted of 33 pairs of videos to rate (1 playlist of 30 stimuli and 3 golden units) and lasted about 10-12 minutes: informed consent + loading videos + instructions + 5 training stimuli × (8s video length + 5s Rating) + 33 test stimuli × (8s video length + ∼ 4s Rating).

Participants. We ran our experiment until each playlist was fully rated by 45 participants. To set the number of participants per playlist, we referred to the CS experiment in [Nehmé et al. 2021a], but we also conducted another pilot experiment with 30 stimuli (selected from this dataset) that were rated by 60 participants. We assessed the evolution of the confidence intervals according to the number of participants (see supplementary material).

It took us about 5 days to collect all the data: 148929 quality judgments were collected. A total of 4513 participants completed the experiment: 2501 males and 2012 females. They were from 67 different countries and aged between 18 and 80. All participants were naive about the purpose of the experiment. The recruiting process of the participants was conducted using Prolific[4], as the results of the CS experiment in [Nehmé et al. 2021a] highlight the reliability and seriousness of the Prolific participants.

[4]https://www.prolific.co/

In the remainder of the paper, subjective scores are mapped on a discrete numerical scale from 1 to 5, following the ITU recommendation [ITU-R BT.500-13 2012] as follows: Imperceptible: 5, Perceptible but not annoying: 4, Slightly annoying: 3, Annoying: 2, Very annoying: 1.

Participants screening. As in [Hoßfeld et al. 2014], participants were filtered by combining the following two screening strategies: (1) the ITU-R BT.500-13 screening procedure [ITU-R BT.500-13 2012], which detected 159 outliers. This procedure is summarized in [Mantiuk et al. 2012] as follows: "*The procedure involves counting the number of trials in which the result of the observer lies outside* $\pm 2\times$ *standard deviation range and rejecting those observers for which (i) more than* 5% *of the trials are outside that range; and (ii) the trials outside that range are evenly distributed so that the absolute difference between the counts of trials exceeding the lower and the upper bound of that range is not more than* 30%."
(2) the golden units (GU) analysis, which revealed 110 outliers distributed as follows:

- 24 participants rated the distortion of the very poor quality stimulus as imperceptible or perceptible but not annoying ($s_i^{GU_{poor}} \geq 4$, where $s_i^{GU_{poor}}$ denotes the score assigned by participant i to GU_{poor}).
- 39 participants rated the very good quality GU as annoying or very annoying ($s_i^{GU_{high}} \leq 2$).
- 32 participant gave inconsistent scores to the third GU showed twice ($\left|s_i^{GU_{rep1}} - s_i^{GU_{rep2}}\right| \geq 3$).
- 7 participants rated $s_i^{GU_{poor}} = 3$ & $s_i^{GU_{high}} = 3$.
- 8 participants scored ($s_i^{GU_{high}} = 3 \mid s_i^{GU_{poor}} = 3$) & $\left|s_i^{GU_{rep1}} - s_i^{GU_{rep2}}\right| = 2$.

Of the participants who failed to evaluate the golden units, 14 were also detected by the ITU-R BT.500-13 screening procedure. As a result, 255 out of 4513 participants were rejected (5.6%). Only the scores of the remaining participants will be used in the following sections. We compute the Mean Opinion Score (MOS) from these scores by averaging the scores given by different participants on each stimuli.

We present in Figure 7 the distribution of raw scores and MOSs obtained for the subset of 3000 stimuli evaluated by the participants. It can be seen that the subjective scores distributed across the whole quality range. Figure 8 shows the distribution of Confidence Intervals (CIs) of the MOSs for each source model. No loose confidence intervals were found (most confidence intervals are below 0.3), demonstrating good agreement between participants' ratings across the stimuli of the different models.

5 GRAPHICS-LPIPS: A PERCEPTUAL QUALITY METRIC FOR 3D GRAPHICS BASED ON CNN

As discussed in the related work section, there is a lack of quality metrics for 3D graphics with color attributes, especially those based on deep learning approaches. Given our large-scale dataset of textured mesh of which 3000 stimuli are associated with subjective

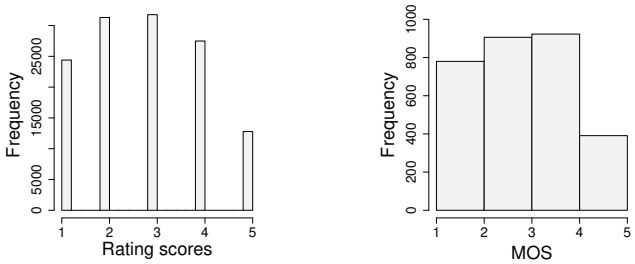


Fig. 7. Distribution of (a) raw scores and (b) MOSs for the subset of 3000 stimuli rated in the subjective experiment.

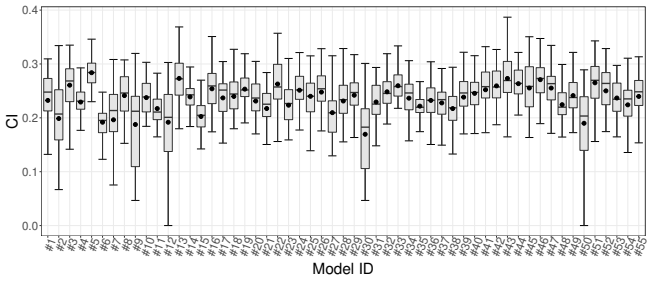


Fig. 8. Boxplots of CIs obtained for the source models. Mean values are displayed as circles.

scores, we were able to create a deep-learning quality metric for such data.

We considered the Learned Perceptual Image Patch Similarity (LPIPS) metric [Zhang et al. 2018] which employs a deep neural network to evaluate the perceptual similarity between 2 image patches, and adapted it to 3D data, then trained it using our dataset. The choice of LPIPS was motivated by its many successful applications [Huang et al. 2018; Yang et al. 2018], the simplicity of the approach and the fact that it is based on an in-depth study across different architectures.

5.1 Description of Graphis-LPIPS

The base principle of LPIPS is to use pre-trained neural networks to extract deep features from two image patches, x (the reference) and x_0 (the distorted patch). The two inputs are treated in parallel by two siamese CNNs that share weights and that we denote by F. The feature difference between the two patches ($F(x) - F(x_0)^2$) can then be mapped to a perceptual difference by going through a 1×1 convolution layer that learns the appropriate weights ω for each channel. Compared to the original LPIPS, we added a 1×1 convolution layer with weights ω_0 which allows to further calibrate the model to our perceptual data. Finally, the result is spatially averaged over all pixels from the patch. The obtained score $d(x, x_0)$ represents a perceptual difference between the original patch x and the distorted one. The architecture of our network is shown in Figure 9. In accordance to the recommendation of the authors, we use the pre-trained AlexNet with fixed weights as feature extractor

and only learn the weights ω and ω_0 of the convolution layers on top.

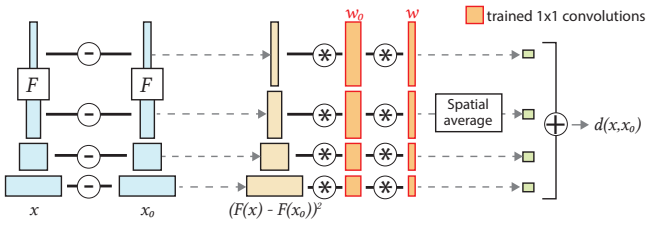


Fig. 9. Graphics-LPIPS architecture: to compute a distance $d(x, x_0)$ between two patches x and x_0, features are first extracted from the layers of the network F and normalized in the channel dimension. The feature difference then goes through two layers of 1×1 convolution in order to reach a 1-channel measure. Finally, we average across spatial dimension and sum across all layers.

The LPIPS metric was originally trained and tested for perceptual similarity tasks (a.k.a Two Alternative Force Choice 2AFC), in which participants were asked to choose which of two distorted patches (x_0, x_1) is more similar to the reference x. A small network was thus added by the authors at training time to map the obtained perceptual distances to the collected preference score. Differently from the original LPIPS, our dataset is composed of a MOS scores per image and we want our metric to predict this overall quality score. We thus modify the remaining of LPIPS pipeline in the following ways: (1) we divide our stimuli into patches that are suitable for distance computation, (2) we pool the perceptual distances obtained on all these patches to get an image score that can be compared to the MOS, (3) we devise a training strategy to ensure that the loss can be computed per image (and not only per-patch).

Given a distorted image I (along with its collected quality score MOS_I) and a reference image I^r, we sample corresponding patches x_i and x_i^r. We use the previously described pipeline to obtain the perceptual score $d(x_i^r, x_i)$ of each pair of patches. Similarly to [Bosse et al. 2018; Kang et al. 2014], we opted for the "pooling by simple averaging" to get the score of each image, the estimated overall image quality $\hat{Q}$ of I is thus computed as:

$$\hat{Q}_I = \frac{1}{N_p} \sum_{i=1}^{Np} d(x_i^r, x_i) \tag{2}$$

where N_p denotes the number of patches sampled from I, x_i refers to a patch from the distorted image I, and x_i^r is its corresponding patch on the reference image I^r. Note that since our stimuli are renderings of distorted meshes, the perceived visual alterations are expected to be relatively well distributed on the image, without strong local distortions. For this reason, averaging the local scores over the image is a reasonable strategy and gives the best result compared to other pooling functions. Results with other pooling strategies (L2, L3 or max pooling) are reported in the supplementary material.

The Mean Square Error (MSE) is used as the minimization criterion. The loss function to train our network is then:

$$E_I = (\hat{Q}_I - MOS_I)^2 \tag{3}$$

where E_I is the loss over an image.

Training data. As training data, we use for each stimulus of our dataset a snapshot taken from its main viewpoint to which we assigned the obtained MOS. Thus, we have 3000 annotated images representing our 3000 degraded stimuli. The image size, 650×550, is the video resolution of the stimuli seen by the participants in the subjective experiment. We divided (patchified) the images into small overlapping patches of size 64×64 sampled every 32 pixels. We removed patches containing less than 65% stimulus information (i.e., the percentage of background pixels in the patch is greater than 35%). We got an average of 60 patches per stimulus.

80 % of the stimuli in the dataset (about 2400) are used for training and 20 % for testing. The dataset is randomly split by source model ensuring that no 3D model is used for both training and testing. As a result 44 source models out of 55 were included in the training while the rest were used for testing. We used k-fold cross-validation and generated 5 different splits. We report the performances over the 5 folds in the next subsection and chose a representative fold with average performances for all the subsequent evaluations.

Training. As the distances computed for patches of the same image are combined for the optimization of the network (Eq. 2 and Eq. 3), we cannot treat each patch as a separate sample (in other words, the patches of the same image can not be distributed over different batches). Thus, each batch was made to contain N_I images, each represented by N_p patches, resulting in a batch size of $N_I \times N_p$ patches. The backpropagated error is the average loss over the images in a batch (mean(E_I) computed in Eq. 3). During training, patches are randomly sampled every epoch to ensure that as many different image patches as possible are used in training. At inference time, we use all the patches from the image to make the prediction.

Our final model was trained for 10 epochs (5 epochs at initial learning rate 10^{-4} and 5 epochs with linear decay). Each batch contained $N_I = 4$ images (stimuli), each represented by $N_p = 150$ patches. The maximum number of patches for one image in our dataset being 149, all the patches from each image are thus seen at each iteration. For images that are represented by less than 150 patches, we repeat the patches until reaching this number.

We refer to the version of LPIPS adapted for 3D Graphics as *Graphics-LPIPS* and compare its performances to those of the original LPIPS in the following subsection.

5.2 Results and Evaluation

Figure 10 summarizes the performance of *Graphics-LPIPS* and compares it to state-of-the-art Image Quality Metrics (IQMs), including the original LPIPS, on the test set of each of our five folds (each fold containing around 600 stimuli obtained from 11 source models). We report the mean and standard deviation over the five folds, the results on each individual fold can be found in supplemental material. Plots illustrating the subjective scores versus the metric values on a representative fold are shown in Figure 11. As for our metric, the IQMs were computed on the snapshots taken from the main

viewpoint of the stimuli. The parameters of the IQMs are provided in the supplementary material.

We evaluated the performance of the metrics in terms of correlations and classification abilities. For correlation measures, the Pearson Linear Correlation (*PLCC*) and the Spearman Rank Order Correlation (*SROCC*) were chosen. To account for saturation effects associated with human senses, we computed PLCC after a logistic regression that establishes a nonlinear mapping between subjective and metric scores. For the classification ability analysis, we considered that proposed in [Krasula et al. 2016] which consists of assessing (1) the ability of the metric to distinguish between significantly different and similar pairs of stimuli and (2) the ability of the metric to predict which stimulus is of better quality in a pair of stimuli. This analysis takes into account the uncertainty of the subjective scores and considers the Area Under the Curve (AUC) as the performance measure. We denote AUC_{DS} and AUC_{BW} respectively for the 2 analysis scenarios.

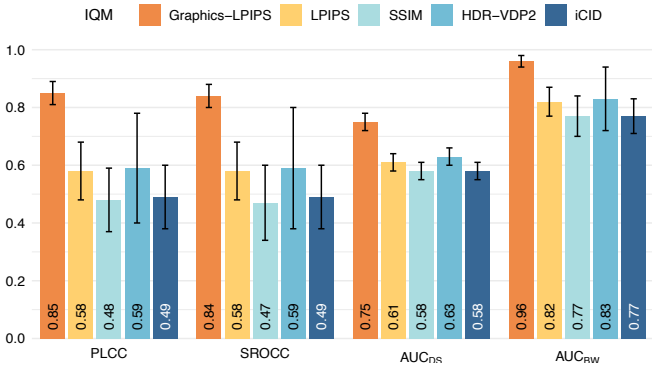


Fig. 10. Performance of our metric (Graphics-LPIPS) compared to state-of-the-art image metrics. The reported numbers are averages over our five folds while the error bars show the standard deviation over the folds.

The proposed metric shows a much better classification ability and correlation with MOSs than IQMs. Statistical tests on the logistic regression residuals yield p-values << 0.0001. Additionally, our metric exhibits more consistent results over the different folds (smaller standard deviation), showing that it generalizes better to a wide variety of images. The poor performance of the IQMs reflects the challenging aspect of our dataset. We believe that this is related to (1) the process of selecting the 3000 stimuli, which samples a lot of stimuli for which two quality metrics did not agree (see section 4.1) and to (2) the large variability of source models and distortion combinations (mixed distortions) present in this dataset.

5.3 Validation on a dataset of 3D meshes with vertex colors

We evaluated the performance of our metric on the dataset of meshes with vertex colors reported in [Nehmé et al. 2021b], to assess its robustness. This dataset is composed of 480 animated stimuli, generated from 5 source models corrupted by simplifications and compressions applied on geometry and color: uniform quantizations applied on either (1) geometry or (2) color, simplifications that take

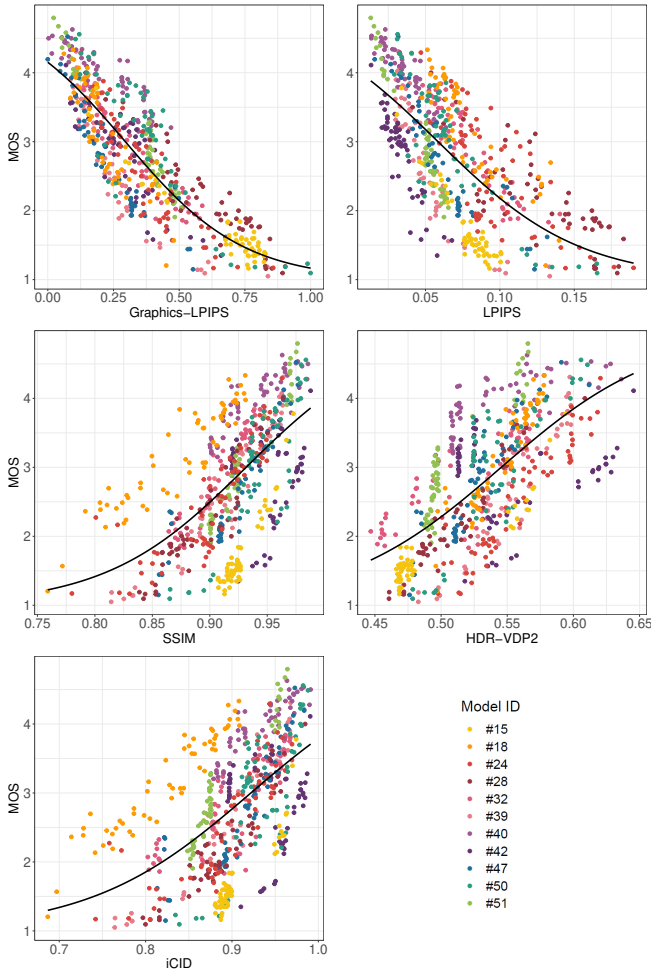


Fig. 11. MOS vs. quality metric values for the test set of textured meshes. Each point represents a distorted stimulus identified by its source model. The curve shows the logistic regression.

into account either the (3) geometry only or (4) both geometry and color. Each distortion was applied with 4 different strengths adjusted manually in order to cover the whole range of visual quality from imperceptible to high levels of impairment. Each stimulus was displayed in 3 viewpoints and animated with 2 short rigid motions (rotations and zooms). The dataset was obtained through a subjective study in VR based on the DSIS method.

As each stimulus in this dataset was rated in 3 different viewpoints, we computed Graphics-LPIPS on snapshots taken from each viewpoint. We did not consider the influence of animations. Thus, for a given stimulus, we averaged its mean opinion scores (MOSs) over the two animations. The database used is therefore composed of 240 stimuli. We included the results of IQMs computed on these snapshots as well as the results of the Color Mesh Distortion Measure (CMDM), which is a model-based quality metric for 3D meshes with colors attributes (i.e. works entirely on the mesh domain). It incorporates perceptually-relevant geometry and color features and

is based on a data-driven approach [Nehmé et al. 2021b]. CMDM was computed only over the visible parts of the 3D model (visible vertices) in each viewpoint.

Table 3. Performance comparison of different metrics on a dataset of meshes with vertex colors. For metrics marked with a *, the values are reprinted from the supplementary material of [Nehmé et al. 2021b].

	Graphics-LPIPS	$CMDM_{vis}$*	SSIM*	HDR-VDP2*	iCID*
PLCC	**0.89**	**0.89**	0.79	0.81	0.86
SROCC	**0.88**	0.87	0.8	0.83	0.87

Although our metric was trained on a different dataset with different models and different distortions and even different color representation (textures and not vertex colors), its performance is comparable4to that of CMDM which was learned on this dataset. This shows the good robustness of our metric and validates (1) that it did not just learn the distortions that are specific to our dataset and (2) its ability to differentiate and rank stimuli from different source models and different distortions. Moreover, Table 3 also shows that our metric can be computed on different viewpoints of the 3D object (even if it is not necessarily the main viewpoint) and still provide good results in correlation with MOSs.

Furthermore, we noticed that IQMs exhibit poorer performance on our dataset of textured meshes than on the dataset of meshes with vertex colors (see Figure 10 and Table 3). This difference in IQMs performance between the two datasets illustrates once again the challenging aspect of our dataset.

5.4 Robustness to changes in lighting and material

Lighting. One of the potential drawbacks of an image-based metric like ours is its potential dependency on the rendering conditions. Our metric is supposed to be able to evaluate the quality of a 3D model, regardless of how it is illuminated. To test whether our metric is robust to changes in the lighting conditions, we conducted the two following experiments:
(1) We moved the directional light toward the left side of the object (Figure 12.left) and toward the bottom (Figure 12.right). We then rendered the images with these new conditions and used the same network as before (*without any retraining*) to compute results on the test set of our representative fold. Figure 12 provides the correlation results according to the angle from the canonical axis. For horizontal variations, the performance remains almost identical and even increases slightly for grazing angles. For vertical variations, the performance decreases slightly when the illumination becomes close to front lighting (≈ 0.01 decrease in Pearson correlation). Given the fact that front lighting tends to completely mask geometric details, the observed robustness of our metric remains excellent, although the network never saw those lighting conditions before.
(2) We also tested two completely different lightings: one dimmed light coming from the left (Figure 13.a) and a spot light coming from the front of the object (Figure 13.b). In line with previous results, the first configuration leads to very good scores ($PLCC = 0.87$), showing even better performances than with the original illumination. The second condition leads to more degraded results ($PLCC = 0.81$). Still, although this configuration is completely different from the original one, its performances are still outperforming other IQMs.

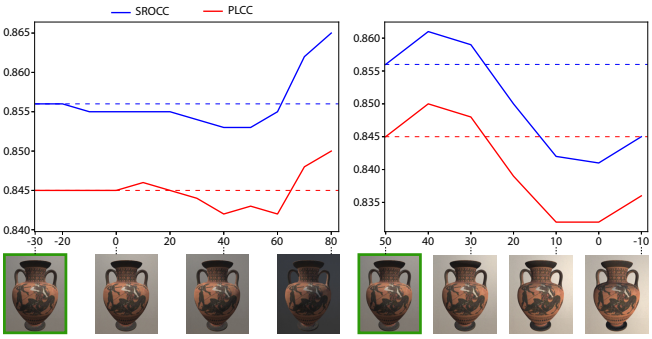


Fig. 12. Performance of our metric when the light direction changes horizontally (left) or vertically (right). The angles in abscissa refer to the angles from the canonical axis (0 corresponds an horizontal line pointing straight to the front). The reference lighting lies on the left, its performances are depicted with a dash line. Samples of the rendered images are shown at the bottom, the reference lighting is circled in green.

Material. Additionally, we evaluated how our metric behaves when changing the material property of an object. Note that, contrary to lighting conditions, our metric is not supposed to be robust against material change since the material intrinsically defines the appearance of the object itself. We rendered the objects with a glossy material (glossiness = 0.8, metallic = 0 in Unity PBR model, see Figure 13.c) as well as with a metallic one (glossiness = 0.6, metallic = 0.8, see Figure 13.d). In that case, the performances of our metric decrease significantly with Pearson correlations of respectively 0.72 and 0.74. As mentioned above, this is expected since materials play a key role in the perception of an object, more specular material can lead to amplifying the visual effects of geometric deformation while minoring the texture ones. These results open interesting perspectives on the influence of the material in the perception of compression artifacts, that would require new subjective studies.

(a) dimmed light PLCC = 0.87 SROCC = 0.87 | (b) spotlight PLCC = 0.81 SROCC = 0.80 | (c) glossy PLCC = 0.72 SROCC = 0.72 | (d) metallic PLCC = 0.74 SROCC = 0.73

Fig. 13. Different conditions of lighting (dimmed light and spotlight) or material (glossy and metallic) along with their prediction scores (correlations).

5.5 View-independent approach

In order to have an automatic version of our metric that does not require manual selection of a main viewpoint for each 3D model, we

considered another training scenario, using a set of snapshots of the model taken from different viewpoints. It may be especially relevant in our case since all the stimuli in our dataset were animated with a full rotation (360 °) during the subjective test (see subsection 4.2). For each stimuli, we thus generated 4 snapshots taken from 4 camera positions regularly sampled on its bounding box and prepared the data as for the above network. We used the same training and testing set as our representative fold and used the same training parameters, randomly sampling $N_p = 300$ patches from all possible viewpoints. The results on the test set are reported in Table 4. For the IQMs, the global quality score of a stimulus is the average of the IQM values computed on its 4 snapshots.

Table 4. Performance comparison of different metrics on the test set of our textured mesh dataset, when several viewpoints are considered per stimulus.

	Graphics-LPIPS	SSIM	HDR-VDP2	iCID
PLCC	**0.84**	0.69	0.68	0.68
SROCC	**0.83**	0.67	0.69	0.67
AUC_{DS}	**0.74**	0.64	0.64	0.63
AUC_{BW}	**0.96**	0.9	0.88	0.88

Comparing the results of Figure 10 and Table 4, we observe that the performance of our metric decreases slightly when considering a view-independent approach. This indicates that our manual selection of the main viewpoint for the source models is indeed relevant and helps the network. It also indicates that the perceptual pooling is not uniform: some parts or viewpoints of the objects have a stronger influence on the overall quality perceived by the observer. This effect depends on the metrics, as the performance of SSIM and iCID improves when considering multiple viewpoints per stimulus, while that of HDR-VDP remains stable.

6 APPLICATION

This section presents an application of the proposed metric and dataset. We used Graphics-LPIPS to annotate our whole dataset of textured meshes and study the influence of several factors on the visual quality.

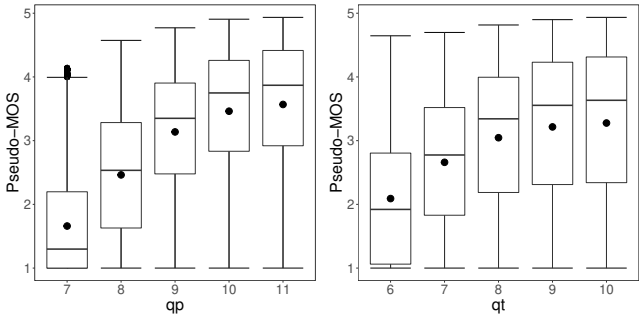
Indeed, our subjective experiment involved only 3000 stimuli out of 343750 (i.e. only 3000 stimuli have a MOS value). To annotate the remaining stimuli of the dataset, we applied Graphics-LPIPS to predict their MOS, referred to as pseudo-MOS. The pseudo-MOSs distribution of all stimuli of the dataset is provided in the supplementary material.

Annotating the entire dataset allowed us to explore the influence of the different compression parameters and their interactions on the subjective scores and thus on the perceived quality. We were also able to evaluate the impact of model characteristics on the perception of distortions. The subsequent subsections detail these analyzes.

6.1 Influence of each distortion on perceived quality

The perceptual quality of textured 3D content depends on both the geometry and color distortions. Since the distortions in our dataset are of different natures (quantization, sub-sampling, simplification) and affect different aspects of the 3D model (geometry and texture), we believe that their impact on the perceived quality is therefore very different. In this subsection, we provide an in-depth analysis of the effect of each distortion on the perceived quality. We also determine which distortions affect the quality scores the most. To do so, we ran a Multivariate Analysis of Variance (ANOVA: $LoD_{simpL} \times qp \times qt \times T_Q \times T_S$) on the quality score of the entire dataset. All of the five distortions affect significantly the perceived quality (p-value << 0.0001), the full ANOVA table can be found in supplemental. .

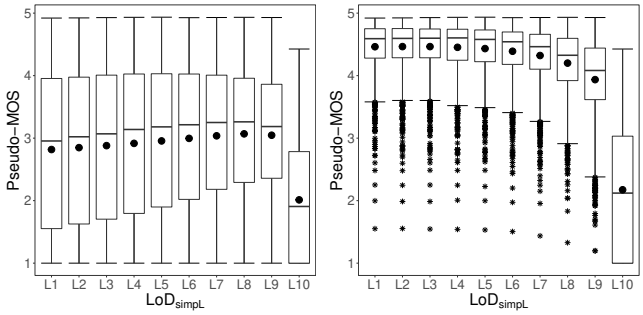
6.1.1 Influence of the geometry and texture coordinate quantization. Figures 14a and 14b show the impact of the quantization parameters on the visual quality of 3D models. As expected, quantizing the vertex position or texture coordinates with too few bits can seriously degrade model quality. The advantage of using fewer quantization bits is the size reduction of the compressed files, however the resulting visual quality is vastly different from that of the original source model. Therefore, choosing the optimal/correct quantization parameters for an application depends on the intended quality as well as the size constraint. This is known as Rate-Distortion (RD) optimization.



(a) (b)

Fig. 14. Boxplots of MOSs obtained for the quantization of the (a) vertices' positions qp and (b) texture coordinates qt. Mean values are displayed as circles.

6.1.2 Influence of the LoD simplification. When looking at Figure 15a, it appears that the most simplified stimuli ($L7$, $L8$, $L9$) rated slightly better than the less simplified stimuli ($L1$, $L2$, $L3$). This is counter-intuitive and could led us to think that simplifying the models with high strength did not introduce markedly visible impairments. This is not strictly true: it is actually highly dependent on the geometry quantization level. In fact, if we consider only the subset of the least geometry quantized stimuli ($qp = 11$ & $qt = 10$), we see that the MOS logically decreases as the simplification level increases (see Figure 15b). There is thus a significant interaction between the geometry quantization of the model and its levels of detail (p-value << 0.0001). Subsection 6.2.1 details this point.
Note that for the most simplified level $L10$, it is a bit peculiar: for $L10$, the models are brutally/roughly simplified to have about 2000 faces. This is very degrading (regardless of the qp and qt values), especially for dense models with the highest number of vertices.



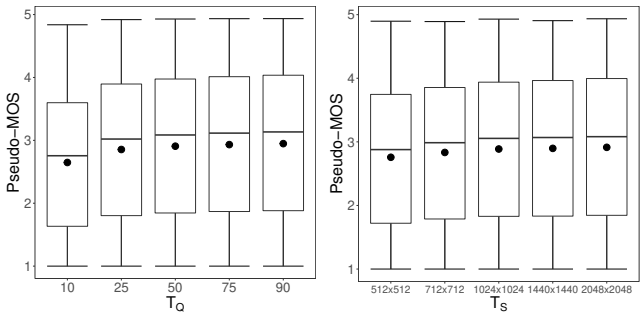
(a) (b)

Fig. 15. Boxplots of MOSs obtained for the LoD simplification LoD_{simpL} (a) for all the stimuli and (b) for the least geometry quantized stimuli ($qp = 11$ & $qt = 10$). Mean values are displayed as circles.

6.1.3 Influence of the texture compression and sub-sampling. Figures 16a and 16b show the impact of the two distortions applied to the texture map (JPEG compression T_Q and sub-sampling T_S) on the perceived quality. While their effect is statistically significant, the impact of these distortions on the MOS is not as obvious as that of texture coordinates quantization (shown in Figure 14b). Figure 16a shows that for $T_Q \geq 50$, the increase of the texture quality does not seem to affect the overall perceived quality.

For texture sub-sampling, Figure 16b shows that increasing the texture resolution T_S more than 712×712 did not overall influence the perceived quality. This may seem logical since the resolution of the stimulus videos shown in the experiment was 650×650. However, we recall that texel density depends on the surface area and even though render resolution is smaller than the texture resolution it does not mean that distortions will always be imperceptible (particularly when the UV mapping is non-uniform). The impact of texture sub-sampling T_S is emphasized when considering its interaction with texture compression T_Q (see subsection 6.2.2).

Thus for our visualization conditions, it seems that we can push the JPEG compression level and sub-sample the texture heavily without impacting the overall quality of the stimulus. This allows to drastically reduce the size of the compressed data.



(a) (b)

Fig. 16. Boxplots of MOSs obtained for the texture (a) compression T_Q and (b) sub-sampling T_S. Mean values are displayed as circles.

6.2 Influence of distortion interactions on perceived quality

Based on the results of the previous subsection, we believe that the impact of the combinations of the different types of distortions differ from the cumulative impact of each distortion applied alone. This subsection presents the distortion interactions that have the most impact on the perceived quality of textured meshes. Other interesting interactions are provided in the supplementary material.

6.2.1 Interaction of LoD simplification and position quantization. The perception of geometry quantization artifacts depends strongly on the level of details of the stimulus (significant interaction with a p-value << 0.0001). Figure 17 illustrates this interaction.

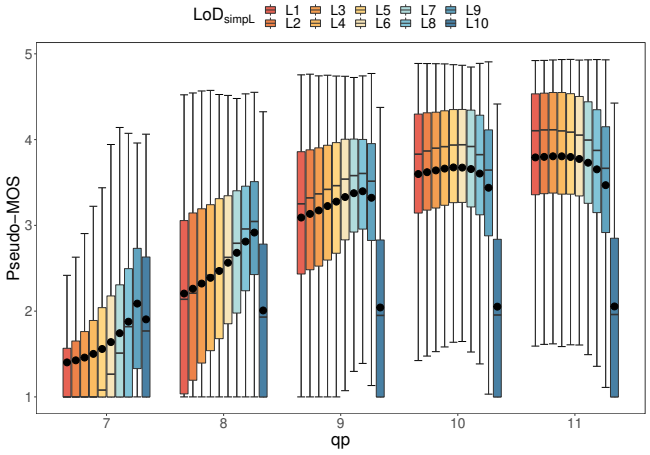


Fig. 17. Boxplots of MOSs illustrating the interaction between the LoD simplification LoD_{simpL} and the quantization of the model's positions qp.

When quantizing stimulus's positions with too few bits ($qp \in \{7, 8, 9\}$), the MOS increases as the simplification level LoD_{simpL} increases (i.e., the number of vertices decreases). The reason is that, the local geometry alteration (local contrast alteration) caused by a strong quantization is more visible on dense meshes (less simplified, $LoD_{simpL} = L1$) than on coarse meshes ($LoD_{simpL} = L9$). Figure 18 illustrates a visual example in which we can see how the effect of the quantization of the vertex positions is much more visible on the dense model ($LoD_{simpL} = L1$). This is due to the fact that the frequency of artifacts created by geometry quantization is higher on a dense mesh than on a simplified mesh; this is what makes the artifacts more visible.

6.2.2 Interaction of the texture compression and sub-sampling. There is a significant interaction (p-value << 0.0001) between the compression and sub-sampling applied on texture images. Figure 19 shows its impact on the perceived quality. For the lowest texture quality ($T_Q = 10$), the MOS increases as the texture size T_S increases. Overall, compression artifacts are less visible on larger textures. The reason is that the blocking artifacts caused by the JPEG compression are bigger/larger on screen for smaller textures.

(a) Source model Model #3 (b) **L1**|**7**|10|2048 × 2048|90 Pseudo-MOS = 1 (c) **L9**|**7**|10|2048 × 2048|90 Pseudo-MOS = 2.4

Fig. 18. Visual example illustrating the interaction between the LoD simplification LoD_{simpL} and the position quantization qp regarding the perceived quality. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_S|T_Q$. The geometric quantization artifacts ($qp = 7$) are more visible on the dense mesh (b) than on the simplified mesh (c).

We can also notice that stimuli with medium or low compressed textures ($T_Q \geq 50$) obtained almost the same MOSs regardless the texture size. This is coherent with what we observed in Figure 16a.

Those results show that, for the viewing conditions of our experiment, the perception of texture compression artifacts is subject to significant masking effects, probably due to the texture mapping, shading, and rasterization processes. Those masking effects makes the JPEG artifacts significantly less visible than for a *natural* 2D image directly displayed on the screen.

6.3 Influence of content characteristics on perceived quality

The content has a concealing effect on the perception of the distortions, which is consistent with the characteristics of the human visual system [Karunasekera and Kingsbury 1995]. Indeed, for the same distortion parameters, the perceived quality may not be the same depending on the models and their characteristics.

In this section, we evaluate the influence of content characteristics on the perception of distortions and thus on quality. To do so, we use the content characterization measures we developed in Section 3.2 (SI_{Geo} and SI_{Col}). We group our 55 models into 5 clusters based on their geometric and color complexity. Thus, the first cluster "$SI_{Geo}1$" contains the first 11 models with the least complex geometry (lowest SI_{Geo} values), while "$SI_{Geo}5$" designates the 11 models with the most geometric details (highest SI_{Geo} values). Similarly, "$SI_{Col}1$" denotes the first 11 source models with the least color details while "$SI_{Col}5$" refers to the models with the richest texture.

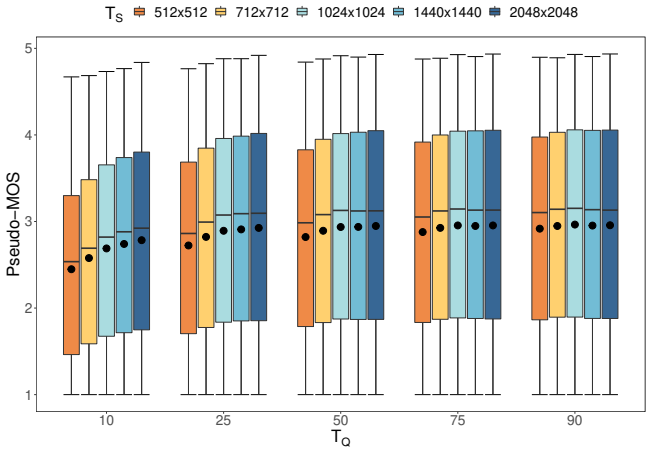


Fig. 19. Boxplots of MOSs illustrating the interaction between the texture compression T_Q and sub-sampling T_S.

Our clusters are well dispersed in the SI_{Geo}/SI_{Col} plane. An histogram representation of Figure 3.a., provided in the supplementary material, illustrates that.

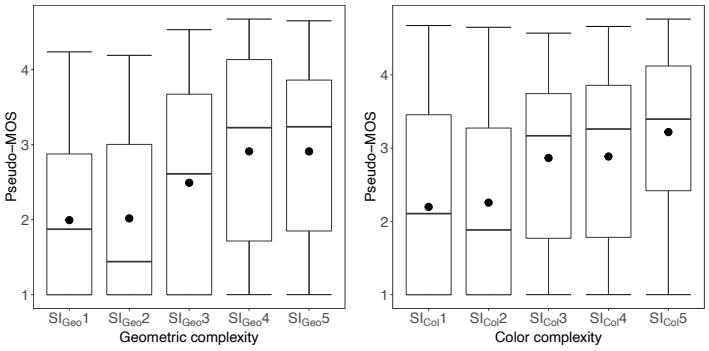
6.3.1 Influence of geometric and color complexity on the perception of position quantization. To evaluate the influence of the characteristics of a model on the perception of the degradations generated by the quantization of the position of its vertices qp, we considered the subset of stimuli having a strongly quantized geometry and the levels of all other distortions fixed at their best levels (giving the best quality), i.e. we considered the stimuli having: $qp \in \{7,\ 8,\ 9\}$ & $LoD_{simpL} \in \{L1,\ L2,\ L3\}$ & $qt \in \{9,\ 10\}$ & $T_Q \in \{75,\ 90\}$ & $T_S \in \{1440 \times 1440,\ 2048 \times 2048\}$.

To assess the impact of geometry complexity, we eliminated the stimuli with rich textures ($SI_{Col}4$ and $SI_{Col}5$) in order to dissociate the influence of geometry and color and to avoid a possible masking effect of one on the other. According to ANOVA, a significant interaction exists between the geometric complexity of the model and the visual impact of the position quantization (p-value << 0.0001). Figure 20a shows that the geometric information can masks the geometry alteration caused by the quantization of the vertices position: For the same quantization level qp, meshes with complex geometry ($\in \{SI_{Geo}4,\ SI_{Geo}5\}$) obtained higher MOSs than those with less complex geometry ($\in \{SI_{Geo}1,\ SI_{Geo}2\}$).

Regarding the impact of color complexity, the results presented in Figure 20b show that for the same level of quantization, models with rich texture ($\in \{SI_{Col}4,\ SI_{Col}5\}$) were judged to be of higher quality than those having simpler texture ($\in \{SI_{Col}1,\ SI_{Col}2\}$) . These results corroborate those observed for point clouds, reported in [Liu et al. 2021b].

Thus, we can say that both geometry and color mask the geometric degradations of a quantized 3D model.

6.3.2 Influence of geometric and color complexity on the perception of texture coordinates quantization. Following the same approach described in 6.3.1, we varied $qt \in \{6,\ 7,\ 8\}$ and set the levels of all



(a) (b)

Fig. 20. Boxplots of the MOSs illustrating the influence of (a) geometric complexity SI_{Geo} and (b) color complexity SI_{Col} of the models on the perceived degradation of geometric quantization qp.

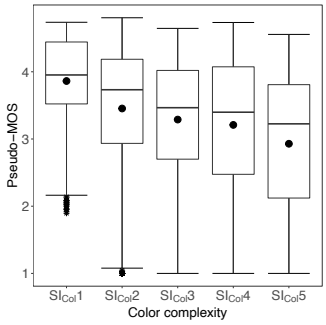


Fig. 21. Boxplots of the MOSs illustrating the influence of the color complexity SI_{Col} of the models on the perceived degradation of texture coordinates quantization qt.

other distortions at their best levels ($LoD_{simpL} \in \{L1,\ L2,\ L3\}$ & $qp \in \{10,\ 11\}$ & $T_Q \in \{75,\ 90\}$ & $T_S \in \{1440 \times 1440,\ 2048 \times 2048\}$) in order to evaluate whether the model's geometric and color characteristics can mask the impairments caused by quantizing its UV map.

Figure 21 clearly shows that models with close-to-uniform textures (i.e., low contrast, with no high frequencies or noticeable structure, $\in \{SI_{Col}1,\ SI_{Col}2\}$) are less sensitive to the UV map quantization than those with colorful and detail-rich textures ($\in \{SI_{Col}4,\ SI_{Col}5\}$).

When analyzing the interaction between the geometric complexity SI_{Geo} of the models and the quantization of their UV map qt, we realized that the influence of this interaction is more complex to evaluate, yet it is significant (p-value $<< 0.0001$). The boxplots of the MOSs illustrating this interaction, as well as some visual examples are provided in the supplementary material. We noticed that the impact of the UV map quantization on the visual quality depends not only on the geometric and color complexity of the model but also on the amount of texture seams (the level of fragmentation of its texture atlas): quantization artifacts are more visible on models exhibiting a large number of texture seams (texture atlas highly fragmented and/or not efficiently packed). Further work is still needed to study the effect of texture seams. We speculate that the set of quality measures, reported in [Maggiordomo et al. 2020], to characterize the quality of the surface parametrization, notably the "UV Occupancy" measure which assesses the quality of the atlas packing and the "Atlas Crumbliness and Solidity" measure which captures the severity of texture seams in a given UV map, could be a good starting point.

7 LIMITATIONS

In the present work, the material information of each object is limited to one single diffuse texture, which is then mapped onto a Lambertian material for rendering. This pipeline thus does not integrate other texture maps representing physically-based material information like metalness and roughness, nor microgeometry information like normals or bumps. We made this technical choice for the sake of simplicity, given that this simple material representation remains reasonably realistic for many use cases and spans an already huge space of distortion parameters. As illustrated in Section 5.4, the nature of the material (e.g., its specularity) obviously influences the visual impact of compression artifacts. In that respect, new user studies are needed to understand and model this influence.

Our metric implementation makes the hypothesis that the perceived distortion of a global model can be modeled as an average of the local distortions from patches sampled either on a principal view or on multiple views. Even if we showed that this pooling provided better results than some other choices (L2, L3, or max pooling), it is highly probable that the reality of our visual system is quite different. Complex attention mechanisms must be at work, as raised by the results of our view-independent approach. We believe that one solution to capture this complexity would be to learn this pooling function, by letting the network learn where to focus its attention. Similarly, the view-independent version of our metric is based on the automatic computation of 4 views sampled around the object. The integration of a data-driven view selection model (such as proposed in [Secord et al. 2011]) could certainly improve the results.

The proposed deep-learning metric has been shown to outperform standard state-of-the-art image quality metrics. More extensive comparisons could be conducted by including metrics dedicated to meshes and point clouds (e.g., [Meynet et al. 2020; Nehmé et al. 2021b]) as well as no-reference ones (e.g. [Liu et al. 2021c]).

The proposed dataset is large but based on only 55 source models. An extension could be considered by increasing the variety of source models (e.g., adding 300 sources) and limiting the number of distortions per source to keep subjective testing feasible.

8 CONCLUSION AND FUTURE WORK

We produced a valuable large-scale textured meshes quality assessment dataset, with more than 343k distorted meshes derived from 55 source models corrupted by combinations of 5 real-world distortions, related to compression and simplification, applied on the geometry and texture. The source models cover a good diversity in visual contents. Indeed, we proposed three measures, based on

spatial information and visual attention complexity, to quantitatively characterize the geometric, color and semantic complexity of the model. A subset of 3000 stimuli were rated in a crowdsourcing subjective experiment. This subset was selected to equitably cover the entire quality range, and to be challenging for objective quality metrics.

Our dataset served us to develop a new image-based quality assessment metric for 3D graphics based on CNN. The metric, called Graphics-LPIPS, can be seen as an extension of LPIPS [Zhang et al. 2018]. It is computed on rendered snapshots of the 3D models. It employs a Siamese network fed with reference patches and distorted patches. We employed the AlexNet architecture with learning linear weights on top. The overall quality of the model is derived by averaging local patch qualities. The metric outperformed other image quality metrics in terms of correlations with subjective scores and classification abiltity on our textured mesh dataset. Our metric also demonstrated a good robustness as it provided the best results on a dataset of meshes with vertex colors.

After validating the performance of Graphics-LPIPS, we used it to predict the quality scores of stimuli in our dataset not included in the subjective experiment. Annotating the entire dataset allowed us to analyze the influence of each distortion as well as that of their combinations on the perceived quality. We also determine which distortions affect the quality scores the most. We found a strong perceptual interaction between the geometry quantization of the mesh and its level of details. Indeed, quantization artifacts are less visible on coarse meshes. Regarding the texture compression, we showed that the quality level of the JPEG compression algorithm applied to the texture can be reduced to very low values while maintaining the quality of the final rendered stimulus.

Furthermore, we evaluated the influence of the geometry and color complexity on the perception of distortions. We observed that both color and geometry can mask the geometric degradations of a quantized 3D model. Models with close-to-uniform textures are less sensitive to UV map quantization; however, the impact of this distortion on the visual quality depends also on the amount of texture seams.

Further potential applications. Our dataset of 340K stimuli associated with pseudo-MOSs can be used to train no-reference quality metrics, but not only. Another real-world use case/application of this dataset can be in the rate-distortion control and optimization. This is possible because each of our stimuli is associated with a quality score and a file size, resulting from the compression methods we used for the source model and the texture. Thus, our dataset could allow to propose an analytical perceptual rate-distortion model capable of maximizing the visual quality of the reconstructed textured meshes subjected to a target bitrate.

Future works. As mentioned in Section 7, some parts of the 3D objects probably have a stronger impact on the overall perceived quality than others. Therefore, we believe that an important improvement to our metric Graphics-LPIPS is to integrate and learn an attention model that would estimate the perceptual weight of each patch (from each view) on the global perceived quality. Still as mentioned in Section 7, a natural follow-up of our work is to replicate this study for 3D objects associated with more complex appearance models, e.g. represented by GGX parametrizations including normal, diffuse, roughness, and specular maps.

The source code of our metric and the datasets of textured meshes along with the subjective scores (individual quality scores, MOSs and Pseudo-MOSs) is publicly available online [5].

ACKNOWLEDGMENTS

This work was supported by French National Research Agency as part of ANR-PISCo project (ANR-17-CE33-0005).

REFERENCES

M. Abid, M. Perreira Da Silva, and P. Le Callet. 2020. Perceptual Characterization of 3D Graphical Contents Based on Attention Complexity Measures. *QoEVMA'20: Proceedings of the 1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications* (2020), 31–36. https://doi.org/10.1145/3423328.3423498

Ilyass Abouelaziz, Aladine Chetouani, Mohammed El Hassouni, Longin Jan Latecki, and Hocine Cherifi. 2020. No-reference mesh visual quality assessment via ensemble of convolutional neural networks and compact multi-linear pooling. *Pattern Recognition* 100 (2020).

Ilyass Abouelaziz, Mohammed El Hassouni, and Hocine Cherifi. 2017. A convolutional neural network framework for blind mesh visual quality assessment. *2017 IEEE International Conference on Image Processing (ICIP)* (2017), 755–759. https://doi.org/10.1109/ICIP.2017.8296382

E. Alexiou and T. Ebrahimi. 2017. On the performance of metrics to predict quality in point cloud representations. In *Applications of Digital Image Processing XL*, Andrew G. Tescher (Ed.), Vol. 10396. International Society for Optics and Photonics, SPIE, 282 – 297.

E. Alexiou and T. Ebrahimi. 2018. Point Cloud Quality Assessment Metric Based on Angular Similarity. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME.2018.8486512

E. Alexiou, E. Upenik, and T. Ebrahimi. 2017. Towards subjective quality assessment of point cloud imaging in augmented reality. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. 1–6. https://doi.org/10.1109/MMSP.2017.8122237

E. Alexiou, I. Viola, T. M. Borges, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi. 2019. A comprehensive study of the rate-distortion performance in MPEG point cloud compression. *APSIPA Transactions on Signal and Information Processing* 8 (2019), e27.

E. Alexiou, N. Yang, and T. Ebrahimi. 2020. PointXR: A Toolbox for Visualization and Subjective Evaluation of Point Clouds in Virtual Reality. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123121

N. Aspert, D. Santa-Cruz, and T. Ebrahimi. 2002. MESH: measuring errors between surfaces using the Hausdorff distance. In *Proceedings. IEEE International Conference on Multimedia and Expo*. https://doi.org/10.1109/ICME.2002.1035879

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2018. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. *IEEE Transactions on Image Processing* 27, 1 (2018), 206–219. https://doi.org/10.1109/TIP.2017.2760518

F. Caillaud, V. Vidal, F. Dupont, and G. Lavoué. 2016. Progressive Compression of Arbitrary Textured Meshes. *Computer Graphics Forum* 35, 7 (Oct. 2016), 475–484.

A. Chetouani. 2018. Three-dimensional mesh quality metric with reference based on a support vector regression model. *Journal of Electronic Imaging* 27, 4 (2018), 1 – 9. https://doi.org/10.1117/1.JEI.27.4.043048

Kyriaki Christaki, Emmanouil Christakis, and Petros Drakoulis. 2018. Subjective Visual Quality Assessment of Immersive 3D Media Compressed by Open-Source Static 3D Mesh Codecs. *25th International Conference on MultiMedia Modeling (MMM)* (2018), 1–12.

Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. 2008. MeshLab: an Open-Source Mesh Processing Tool. In *Eurographics Italian Chapter Conference*. The Eurographics Association. https://doi.org/10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136

Massimiliano Corsini, Elisa Drelie Gelasca, Touradj Ebrahimi, and Mauro Barni. 2007. Watermarked 3-D mesh quality assessment. *IEEE Transactions on Multimedia* 9 (2007), 247–256.

Ulrich Engelke, Maulana Kusuma, Hans-Jürgen Zepernick, and Manora Caldera. 2009. Reduced-reference metric design for objective perceptual quality assessment in wireless imaging. *Signal Processing: Image Communication* 24, 7 (2009), 525–547. https://doi.org/10.1016/j.image.2009.06.005

[5] https://github.com/MEPP-team/Graphics-LPIPS

Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. 2017. DeepSim: Deep similarity for image quality assessment. *Neurocomputing* 257 (2017), 104–114. https://doi.org/10.1016/j.neucom.2017.01.054 Machine Learning and Signal Processing for Big Multimedia Analysis.

Michael Garland and Paul S. Heckbert. 1998. Simplifying Surfaces with Color and Texture Using Quadric Error Metrics. *Proceedings of the Conference on Visualization '98* (1998), 263–269.

Deepti Ghadiyaram and Alan C. Bovik. 2016. Massive Online Crowdsourced Study of Subjective and Objective Picture Quality. *IEEE Transactions on Image Processing* 25, 1 (2016), 372–387. https://doi.org/10.1109/TIP.2015.2500021

Jinjiang Guo, Vincent Vidal, Irene Cheng, Anup Basu, Atilla Baskurt, and Guillaume Lavoué. 2016. Subjective and Objective Visual Quality Assessment of Textured 3D Meshes. *ACM Transactions on Applied Perception* 14 (10 2016), 1–20. https://doi.org/10.1145/2996296

Jesús Gutiérrez, Toinon Vigier, and Patrick Le Callet. 2020. Quality Evaluation of 3D Objects in Mixed Reality For Different Lighting Conditions. *Electronic Imaging* 2020 (01 2020). https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-128

Zhouyan He, Gangyi Jiang, Zhidi Jiang, and Mei Yu. 2021. Towards A Colored Point Cloud Quality Assessment Method Using Colored Texture And Curvature Projection. In *2021 IEEE International Conference on Image Processing (ICIP)*. 1444–1448. https://doi.org/10.1109/ICIP42928.2021.9506762

Tobias Hoßfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. 2014. Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing. *IEEE Transactions on Multimedia* 16, 2 (2014), 541–558. https://doi.org/10.1109/TMM.2013.2291663

Xun Huang, Ming-Yu Liu, Serge Belongi, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-Image Translation. *ECCV 2018* 11207 (2018), 179–196. https://doi.org/10.1007/978-3-030-01219-9_11

ITU-R BT.500-13. 2012. Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service. *International Telecommunication Union* (2012).

ITU-T P.910. 2008. Subjective video quality assessment methods for multimedia applications. *International Telecommunication Union* (2008).

A. Javaheri, C. Brites, F. Pereira, and J. Ascenso. 2017. Subjective and objective quality evaluation of compressed point clouds. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. 1–6. https://doi.org/10.1109/MMSP.2017.8122239

A Javaheri, C Brites, F Pereira, and J Ascenso. 2019. Point Cloud Rendering after Coding : Impacts on Subjective and Objective Quality. *arXiv:1912.09137* (2019), 1–13.

A. Javaheri, C. Brites, F. Pereira, and J. Ascenso. 2020. Mahalanobis Based Point to Distribution Metric for Point Cloud Geometry Quality Evaluation. *IEEE Signal Processing Letters* 27 (2020), 1350–1354. https://doi.org/10.1109/LSP.2020.3010128

Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. 2018. Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing. *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)* (2018), 1–6. https://doi.org/10.1109/QoMEX.2018.8463298

Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional Neural Networks for No-Reference Image Quality Assessment. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 1733–1740. https://doi.org/10.1109/CVPR.2014.224

S.A. Karunasekera and N.G. Kingsbury. 1995. A distortion measure for blocking artifacts in images based on human visual sensitivity. *IEEE Transactions on Image Processing* 4, 6 (1995), 713–724. https://doi.org/10.1109/83.388074

L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma. 2016. On the accuracy of objective image and video quality models: New methodology for performance evaluation. *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)* (2016), 1–6.

Guillaume Lavoué. 2009. A Local Roughness Measure for 3D Meshes and Its Application to Visual Masking. *ACM Trans. Appl. Percept.* 5, 4, Article 21 (2009), 23 pages. https://doi.org/10.1145/1462048.1462052

Guillaume Lavoué. 2011. A Multiscale Metric for 3D Mesh Visual Quality Assessment. *Computer Graphics Forum* 30, 5 (2011), 1427–1437.

Guillaume Lavoué, Mohamed Chaker Larabi, and Libor Vasa. 2016. On the Efficiency of Image Metrics for Evaluating the Visual Quality of 3D Models. *IEEE Transactions on Visualization and Computer Graphics* 22, 8 (2016), 1987–1999.

Davi Lazzarotto, Evangelos Alexiou, and Touradj Ebrahimi. 2021. Benchmarking of objective quality metrics for point cloud compression. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. 1–6.

Qi Liu, Hui Yuan, Raouf Hamzaoui, Honglei Su, Junhui Hou, and Huan Yang. 2021b. Reduced Reference Perceptual Quality Model With Application to Rate Control for Video-Based Point Cloud Compression. *IEEE Transactions on Image Processing* 30 (2021), 6623–6636. https://doi.org/10.1109/TIP.2021.3096060

Qi Liu, Hui Yuan, Honglei Su, Hao Liu, Yu Wang, Huan Yang, and Junhui Hou. 2021c. PQA-Net: Deep No Reference Point Cloud Quality Assessment via Multi-view Projection. *IEEE Transactions on Circuits and Systems for Video Technology* (2021). https://doi.org/10.1109/TCSVT.2021.3100282

Yipeng Liu, Qi Yang, Yiling Xu, and Le Yang. 2021a. Point Cloud Quality Assessment: Dataset Construction and Learning-based No-Reference Approach. *preprint arXiv:2012.11895* (2021).

Andrea Maggiordomo, Federico Ponchio, Paolo Cignoni, and Marco Tarini. 2020. Real-World Textured Things: a Repository of Textured Models Generated with Modern Photo-Reconstruction Tools. *preprint ArXiv:2004.14753* (2020).

Rafał Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. 2011. HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions. *ACM Trans. Graph.* 30, 4, Article 40 (July 2011), 14 pages. https://doi.org/10.1145/2010324.1964935

Rafał K. Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. 2012. Comparison of Four Subjective Methods for Image Quality Assessment. *Computer Graphics Forum* 31, 8 (dec 2012), 2478–2491. http://doi.wiley.com/10.1111/j.1467-8659.2012.03188.x

R. Mekuria, K. Blom, and P. Cesar. 2017. Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 4 (Apr. 2017), 828–842.

Gabriel Meynet, Julie Digne, and Guillaume Lavoué. 2019. PC-MSDM: A quality metric for 3D point clouds. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. 1–3. https://doi.org/10.1109/QoMEX.2019.8743313

Gabriel Meynet, Yana Nehmé, Julie Digne, and Guillaume Lavoué. 2020. PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123147

Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué. 2020. Comparison of Subjective Methods for Quality Assessment of 3D Graphics in Virtual Reality. *ACM Transactions on Applied Perception* 18, 1 (Dec. 2020), 1–23. https://doi.org/10.1145/3427931

Yana Nehmé, Patrick Le Callet, Florent Dupont, Jean-Philippe Farrugia, and Guillaume Lavoué. 2021a. Exploring Crowdsourcing for Subjective Quality Assessment of 3D Graphics. *IEEE International Workshop on Multimedia Signal Processing (MMSP)* (2021).

Yana Nehmé, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. 2021b. Visual Quality of 3D Meshes With Diffuse Colors in Virtual Reality: Subjective and Objective Evaluation. *IEEE Transactions on Visualization and Computer Graphics* 27, 3 (2021), 2202–2219. https://doi.org/10.1109/TVCG.2020.3036153

Anass Nouri, Christophe Charrier, and Olivier Lézoray. 2017. 3D Blind Mesh Quality Assessment Index. *IS&T International Symposium on Electronic Imaging* (Jan. 2017).

Yixin Pan, I Cheng, and A Basu. 2005. Quality metric for approximating subjective evaluation of 3-D objects. *IEEE Transactions on Multimedia* 7, 2 (apr 2005), 269–279. https://doi.org/10.1109/TMM.2005.843364

S. Perry, H. P. Cong, L. A. da Silva Cruz, J. Prazeres, M. Pereira, A. Pinheiro, E. Dumic, E. Alexiou, and T. Ebrahimi. 2020. Quality Evaluation Of Static Point Clouds Encoded Using MPEG Codecs. In *2020 IEEE International Conference on Image Processing (ICIP)*. 3428–3432. https://doi.org/10.1109/ICIP40778.2020.9191308

Jens Preiss, Felipe Fernandes, and Philipp Urban. 2014. Color-image quality assessment: From prediction to optimization. *IEEE Transactions on Image Processing* 23, 3 (2014), 1366–1378. https://doi.org/10.1109/TIP.2014.2302684

Maurice Quach, Aladine Chetouani, Giuseppe Valenzise, and Frederic Dufaux. 2021. A deep perceptual metric for 3D point clouds. *Electronic Imaging* 2021, 9 (2021), 257–1–257–7. https://doi.org/doi:10.2352/ISSN.2470-1173.2021.9.IQSP-257

Judith Redi, Ernestasia Siahaan, Pavel Korshunov, Julian Habigt, and Tobias Hossfeld. 2015. When the Crowd Challenges the Lab: Lessons Learnt from Subjective Studies on Image Aesthetic Appeal. *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia* (2015), 33–38. https://doi.org/10.1145/2810188.2810194

Max Reimann, Ole Wegen, Sebastian Pasewaldt, Amir Semmo, Jürgen Döllner, and Matthias Trapp. 2021. Teaching Data-driven Video Processing via Crowdsourced Data Collection. *Eurographics 2021 - Education Papers* (2021). https://doi.org/10.2312/eged.20211000

Bernice E Rogowitz and Holly E Rushmeier. 2001. Are image quality metrics adequate to evaluate the quality of geometric objects? *Proc SPIE 4299, Human Vision and Electronic Imaging VI* (06 2001). https://doi.org/10.1117/12.429504

Adrian Secord, Jingwan Lu, Adam Finkelstein, Manish Singh, and Andrew Nealen. 2011. Perceptual models of viewpoint preference. *ACM Transactions on Graphics* 30, 5 (Oct. 2011).

H. R. Sheikh and A. C. Bovik. 2006. Image information and visual quality. *IEEE Transactions on Image Processing* 15, 2 (Feb. 2006), 430–444.

H. Su, Z. Duanmu, W. Liu, Q. Liu, and Z. Wang. 2019. Perceptual Quality Assessment of 3D Point Clouds. In *2019 IEEE International Conference on Image Processing (ICIP)*. 3182–3186.

Shishir Subramanyam, Jie Li, Irene Viola, and Pablo Cesar. 2020. Comparing the Quality of Highly Realistic Digital Humans in 3DoF and 6DoF: A Volumetric Video Case Study. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 127–136.

Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011. https://doi.org/10.1109/

TIP.2018.2831899

Wen-xu Tao, Gang-yi Jiang, Zhi-di Jiang, and Mei Yu. 2021. *Point Cloud Projection and Multi-Scale Feature Fusion Network Based Blind Quality Assessment for Colored Point Clouds.* New York, NY, USA, 5266–5272.

Taimoor Tariq, Okan Tarhan Tursun, Munchurl Kim, and Piotr Didyk. 2020. Why Are Deep Representations Good Perceptual Quality Features? *Computer Vision – ECCV 2020* (2020), 445–461.

Dihong Tian and G. AlRegib. 2008. Batex3: Bit allocation for progressive transmission of textured 3-D models. *IEEE Transactions on Circuits and Systems for Video Technology* 18, 1 (2008), 23–35.

Dong Tian, Hideaki Ochimizu, Chen Feng, Robert Cohen, and Anthony Vetro. 2017. Geometric distortion metrics for point cloud compression. In *2017 IEEE International Conference on Image Processing (ICIP)*. 3460–3464. https://doi.org/10.1109/ICIP.2017.8296925

Fakhri Torkhani, Kai Wang, and Jean-Marc Chassery. 2014. A curvature-tensor-based perceptual quality metric for 3D triangular meshes. *Machine Graphics & Vision* 23, 1 (2014), 59–82.

Fakhri Torkhani, Kai Wang, and Jean-Marc Chassery. 2015. Perceptual quality assessment of 3D dynamic meshes: Subjective and objective studies. *Signal Processing: Image Communication* 31, 2 (Feb. 2015), 185–204. https://doi.org/10.1016/j.image.2014.12.008

E. M. Torlig, E. Alexiou, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi. 2018. A novel methodology for quality assessment of voxelized point clouds. In *Applications of Digital Image Processing XLI*, Andrew G. Tescher (Ed.), Vol. 10752. International Society for Optics and Photonics, SPIE, 174 – 190.

K. Vanhoey, B. Sauvage, P. Kraemer, and G. Lavoué. 2017. Visual quality assessment of 3D models: On the influence of light-material interaction. *ACM Transactions on Applied Perception* 15, 1 (2017). https://doi.org/10.1145/3129505

Libor Váša and Jan Rus. 2012. Dihedral Angle Mesh Error: a fast perception correlated distortion measure for fixed connectivity triangle meshes. *Computer Graphics Forum* 31, 5 (2012).

Irene Viola, Shishir Subramanyam, and Pablo Cesar. 2020. A Color-Based Objective Quality Metric for Point Cloud Contents. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123089

Irene Viola, Shishir Subramanyam, Jie Li, and Pablo Cesar. 2022. On the impact of VR assessment on the Quality of Experience of Highly Realistic Digital Humans. arXiv:2201.07701 [cs.MM]

VQEG. 2010. Report on the Validation of Video Quality Models for High Definition Video Content. (June 2010).

Kai Wang, Fakhri Torkhani, and Annick Montanvert. 2012. A Fast Roughness-Based Approach to the Assessment of 3D Mesh Visual Quality. *Computers & Graphics* (2012).

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

Benjamin Watson. 2001. Measuring and predicting visual fidelity. *ACM Siggraph* (2001), 213–220.

Jinjian Wu, Jupo Ma, Fuhu Liang, Weisheng Dong, Guangming Shi, and Weisi Lin. 2020. End-to-End Blind Image Quality Prediction With Cascaded Deep Neural Network. *IEEE Transactions on Image Processing* 29 (2020), 7414–7426. https://doi.org/10.1109/TIP.2020.3002478

Xinju Wu, Yun Zhang, Chunling Fan, Junhui Hou, and Sam Kwong. 2021. Subjective Quality Database and Objective Study of Compressed Point Clouds With 6DoF Head-Mounted Display. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 12 (2021), 4630–4644.

Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. 2018. Pose Guided Human Video Generation. *Computer Vision – ECCV 2018* 11214 (2018), 204–219. https://doi.org/10.1007/978-3-030-01249-6_13

Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun. 2020. Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration. *IEEE Transactions on Multimedia* (2020), 3877–3891.

Qi Yang, Zhan Ma, Yiling Xu, Zhu Li, and Jun Sun. 2020. Inferring Point Cloud Quality via Graph Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/TPAMI.2020.3047083

Zeynep Cipiloglu Yildiz, A. Cengiz Oztireli, and Tolga Capin. 2020. A machine learning framework for full-reference 3D shape quality assessment. *Visual Computer* 36, 1 (2020), 127–139.

Honghai Yu and Stefan Winkler. 2013. Image complexity and spatial information. *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)* (2013), 12–17. https://doi.org/10.1109/QoMEX.2013.6603194

Emin Zerman, Pan Gao, Cagri Ozcinar, and Aljosa Smolic. 2019. Subjective and objective quality assessment for volumetric video compression. *Electronic Imaging* 2019, 10 (2019), 323–1.

Emin Zerman, Cagri Ozcinar, Pan Gao, and Aljosa Smolic. 2020. Textured mesh vs coloured point cloud: A subjective study for volumetric video compression. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 586–595. https://doi.org/10.1109/CVPR.2018.00068

Yujie Zhang, Qi Yang, and Yiling Xu. 2021. *MS-GraphSIM: Inferring Point Cloud Quality via Multiscale Graph Similarity.* Association for Computing Machinery, New York, NY, USA, 1230–1238. https://doi.org/10.1145/3474085.3475294

Qing Zhu, Junqiao Zhao, Zhiqiang Du, and Yeting Zhang. 2010. Quantitative analysis of discrete 3D geometrical detail levels based on perceptual metric. *Computers & Graphics* 34, 1 (2010), 55–65. https://doi.org/10.1016/j.cag.2009.10.004