BASICS: Broad Quality Assessment of Static Point Clouds in a Compression Scenario

Ali Ak, Emin Zerman, Maurice Quach, Aladine Chetouani, Aljosa Smolic, Giuseppe Valenzise, Patrick Le Callet

▶ **To cite this version:**

Ali Ak, Emin Zerman, Maurice Quach, Aladine Chetouani, Aljosa Smolic, et al.. BASICS: Broad Quality Assessment of Static Point Clouds in a Compression Scenario. IEEE Transactions on Multimedia, 2024, pp.1-13. 10.1109/TMM.2024.3355642 . hal-04280304v1

HAL Id: hal-04280304
https://hal.science/hal-04280304v1

Submitted on 10 Nov 2023 (v1), last revised 19 Jan 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BASICS: Broad Quality Assessment of Static Point Clouds in a Compression Scenario

Ali Ak, *Member, IEEE,* Emin Zerman, Maurice Quach, *Member, IEEE,* Aladine Chetouani, *Senior Member, IEEE,* Aljosa Smolic, *Senior Member, IEEE,* Giuseppe Valenzise, *Senior Member, IEEE,* Patrick Le Callet, *Fellow, IEEE*

***Abstract*—Point clouds have become increasingly prevalent in representing 3D scenes within virtual environments, alongside 3D meshes. Their ease of capture has facilitated a wide array of applications on mobile devices, from smartphones to autonomous vehicles. Notably, point cloud compression has reached an advanced stage and has been standardized. However, the availability of quality assessment datasets, which are essential for developing improved objective quality metrics, remains limited. In this paper, we introduce BASICS, a large-scale quality assessment dataset tailored for static point clouds. The BASICS dataset comprises 75 unique point clouds, each compressed with four different algorithms including a learning-based method, resulting in the evaluation of nearly 1500 point clouds by 3500 unique participants. Furthermore, we conduct a comprehensive analysis of the gathered data, benchmark existing point cloud quality assessment metrics and identify their limitations. By publicly releasing the BASICS dataset, we lay the foundation for addressing these limitations and fostering the development of more precise quality metrics.**

***Index Terms*—Point cloud quality, 3D models, point cloud compression, subjective quality assessment, dataset.**

I. Introduction

DIGITAL imaging technologies have revolutionized the capability to capture real-world environments and recreate them in different temporal or spatial contexts. This capacity has extended to the realm of 3D scenes through the integration of computer graphics and photogrammetry techniques. Presently, we can capture intricate 3D objects and scenes using an array of tools, ranging from solely RGB cameras to RGB cameras supplemented with additional sensors [1]–[5]. Within the domain of 3D modeling, two primary representations have gained prominence: colored point clouds and textured 3D meshes [6]. In this paper, our focus is directed toward point clouds, a representation widely used in numerous applications, in particular augmented and virtual reality. The acquisition of point clouds is typically accomplished through diverse means, including stereo-camera arrays [3], LiDAR sensors [4] and conventional cameras [5]. Point cloud acquisition produces huge amounts of data, calling for compression techniques for efficient transmission and storage. However, evaluating the performances of compression algorithms is time consuming and expensive. While the field of research in this area has witnessed notable expansion in recent years [6]–[14], there are still many open questions and problems.

Existing datasets in the field have notable shortcomings in terms of diversity, scale, consistency, and data accessibility. These limitations constitute substantial challenges for research into learning-based approaches. The absence of diversity is frequently attributed to several factors, including the recurrent use of the same point clouds across various datasets, limited geometric complexity and semantic categories, as well as the utilization of the same compression algorithms as the sole source of distortion types. This lack of diversity is further compounded by the relatively modest scale of the existing datasets, rendering them ill-suited for the development of learning-based quality metrics. Consistency issues also can be observed in existing datasets, stemming from factors such as the utilization of unnormalized point clouds and inconsistent rendering specifications. A detailed overview of the existing datasets is presented in Table I, which summarizes the aforementioned deficiencies. Another substantial shortcoming is the limited data availability and incompleteness, e.g., due to copyright constraints or repository management practices, as well as the absence of individual opinion scores, standard deviations and confidence intervals, which are typically needed to develop more sophisticated metrics. The accessibility of data in existing datasets is summarized in Table II.

In conclusion, the need for a new dataset is evident due to aforementioned deficiencies in existing datasets. Our proposal seeks to address these shortcomings comprehensively, providing a dataset that encompasses all essential characteristics required for the development of more accurate quality metrics. Furthermore, it provides further insights into the performance of point cloud quality metrics when applied to point clouds encoded with learning-based compression algorithms.

The contributions of this work are threefold:

- We present, and make publicly available, a broad point cloud quality assessment dataset featuring 75 unique point clouds that hold semantic relevance within the context of telepresence scenarios (as detailed in Section II).
- We compare the performances of various state-of-the-art methods for point cloud compression (as outlined in Section IV).
- We provide an exhaustive benchmark of the state-of-the-art point cloud quality metrics, including both point-based and image-based assessment (as outlined in Section V).

A. Ak and P. Le Callet are with Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France.

E. Zerman is with Department of Computer and Electrical Engineering, Mid Sweden University, Sundsvall, Sweden.

M. Quach and G. Valenzise are with CNRS, CentraleSupelec, Laboratoire des Signaux et Systèmes (UMR 8506), Université Paris-Saclay, Gif-sur-Yvette, France.

A. Chetouani is with Université d'Orléans, Orléans, France.

A. Smolic is with Lucerne School of Computer Science and Information Technology, Rotkreuz, Switzerland.

Manuscript submitted on 06 October 2023.

TABLE I
STATISTICAL SUMMARY OF EXISTING PC QUALITY ASSESSMENT DATASETS

Dataset	nb SRC	nb PPC	nb obs per PPC	total nb obs	Display	Visualization	Subj. test method	Temporal dimension	Distortions
BASICS (proposed)	75	1494	60	3600	2D	Passive	DSIS (side-by-side)	Static	Compression: GPCC, VPCC, GEOCNN
vsenseVVDB2 [6]	8	136	23	23	2D	Passive	ACR	Dynamic	Compression: GPCC, VPCC
Perry et al. [7]	6	90	-	73	2D	Passive	DSIS (side-by-side)	Static	Compression: GPCC, VPCC
da Silva Cruz et al. [8]	8	48	-	50	2D	Passive	DSIS	Static	Octree pruning Projection-based compression from 3DTK
Yang et al. [9]	10	420	16	64	2D	Passive	ACR	Static	Octree pruning, random point down-sample Color noise, geometric gaussian noise
Su et al. [10]	20	740	-	60	2D	Passive	DSIS (side-by-side)	Static	Octree pruning, geometric gaussian noise Compression: SPCC, VPCC, LPCC
NBU-PCD1.0 [11]	10	160	-	-	2D	-	-	Static	Octree pruning
SIAT-PCQD [12]	20	340	38	76	HMD	Interactive	DSIS	Static	Compression: VPCC
Subramanyam et al. [13]	8	64	-	52	HMD	Interactive	ACR-HR	Dynamic	Compression: VPCC
PointXR [14]	5	40	20	40	HMD	Interactive	DSIS	Static	Compression: GPCC

TABLE II
PUBLIC AVAILABILITY OF THE EXISTING DATASETS

Dataset	Point Clouds	Subjective Annotations
BASICS (proposed)	✓	Individual Scores
vsenseVVDB2 [6]	✓	Individual Scores
Perry et al. [7]	Broken URL	✗
Su et al. [10]	✓	(D)MOS only
da Silva Cruz et al. [8]	✗	✗
Yang et al. [9]	✓	(D)MOS only
NBU-PCD1.0 [11]	✗	✗
SIAT-PCQD [12]	Broken URL	✗
Subramanyam et al. [13]	✗	Individual Scores
PointXR [14]	✓	Individual Scores

The BASICS dataset has been made publicly available[1] under the Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA 4.0) license with the aim of fostering continued research within the field. The repository contains individual and mean opinion scores, as well as pristine and compressed point clouds. Additionally, the dataset's GitHub page[2] provides the scripts required for the evaluation of point cloud quality metrics. The BASICS dataset has been successfully employed in the grand challenge on point cloud quality assessment[3] organized by some of the authors at the 2023 IEEE International Conference on Image Processing (ICIP2023).

II. THE BASICS DATASET

The BASICS dataset has been designed and collected with two main objectives in mind: on one hand, providing diverse and extensive data to train learning-based point cloud quality assessment metrics; and on the other hand, creating challenging test conditions to benchmark existing quality metrics. In this section, we describe the various stages of the dataset generation process.

A. Material selection

In order to construct a point cloud quality assessment dataset that encompasses semantic diversity tailored for telepresence applications, we gathered a total of 75 point clouds, distributed across three semantic categories. Despite the broad scope of these categories, they remain highly relevant within the context of telepresence applications. The categories we have defined are "Humans & Animals (HA)", "Inanimate Objects (IA)", "Buildings & Landscapes (BL)", collectively covering a comprehensive set of subjects such as animals, humans, everyday items, vehicles, architectural structures and natural landscapes. Figure 1 offers a visual glimpse into each of these categories.

For the data collection and material selection part, our primary objective was to collect a comprehensive repository of publicly available point clouds that could be freely redistributed. However, numerous data sources impose strict restrictions on redistribution. Consequently, we acquired the 3D models from two sources: collaborator studios (i.e., V-SENSE studio[4] and XD Productions[5]) as well as an online repository for 3D model sharing called SketchFab[6]. Even in this case, availability of point clouds was somewhat limited. To address this, we gathered 3D meshes and generated point clouds via sampling the mesh surface, as detailed in Section II-B.

In total, 104 models were handpicked by three authors of this paper considering the semantic categories described above. Following the removal of materials that exhibited significant similarity, models with highly reflective surfaces and imperfect texturing, as well as those with loose semantic associations to their respective categories, the total count of models was reduced to a final selection of 75. For each model, information regarding its source (a SketchFab URL if applicable), file format, statistics, and licensing details can be accessed in the dataset repository[1].

B. Pre-processing

In order to mitigate potential distortions stemming from other sources, the collected 3D models require a pre-processing and a conversion into point clouds before any further processing. This was necessary due to the heterogeneous formats in which these models were originally available.

Models that were already in point cloud PLY format required minimal attention, except for voxelization. Conversely,

[1]Dataset Link: https://doi.org/10.5281/zenodo.8324546
[2]GitHub: https://github.com/kyillene/BASICS-Public
Dataset and Github page will be made available after initial reviews.
[3]https://sites.google.com/view/icip2023-pcvqa-grand-challenge/
[4]https://v-sense.scss.tcd.ie/
[5]https://www.xdprod.com/
[6]https://www.sketchfab.com/

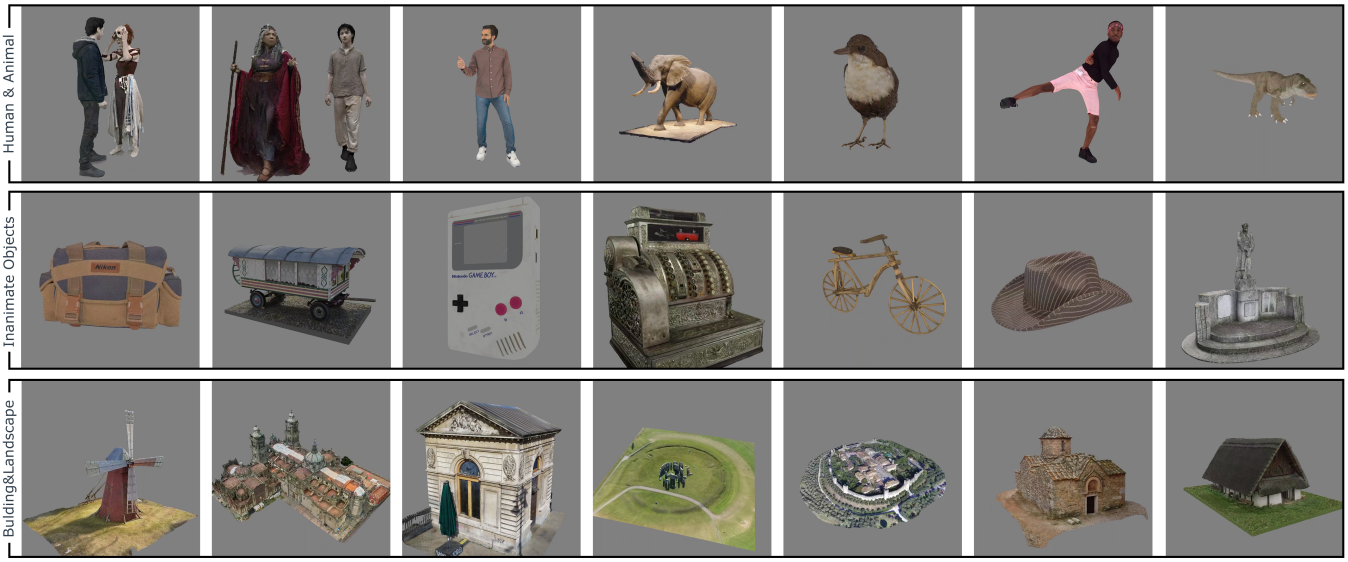

Fig. 1. Sample renderings of 7 point clouds from each of the 3 semantic categories in the dataset (i.e., 7 of the total of 25 for each category).

3D meshes underwent through several steps. These steps are further discussed below.

1) Making 3D meshes uniform: The collected 3D mesh models were in various formats. All 3D meshes were converted into OBJ format, to streamline the pre-processing chain, using Blender[7] and Meshlab[8].

2) Cleaning 3D meshes: Some of the 3D meshes had either parts that had transparent or reflective properties (e.g., glasses in some models), which could not be replicated well during point cloud rendering. Some other meshes had incomplete parts (e.g., trees, some of the building façades) which would decrease the users' quality of experience and introduce other sources of distraction and distortion. To avoid such effects, these parts were removed or cleaned in Blender.

The mesh files were then unified into a single OBJ file, so that the sampling process in the pipeline could be done with ease. Next, the material properties (which are described in the .mtl files) are checked to eliminate any other reflective properties of the materials, which could not be reproduced correctly in the point cloud format. After all these operations, the 3D meshes were ready for the point cloud sampling step.

3) Sampling point clouds: Using CloudCompare[9], point clouds were sampled from the 3D meshes' surfaces. During this operation 10 million points were sampled on the surfaces of the said meshes. The sampling operation extracted the location, color, and normal attributes for each point in the point clouds. At the end of this stage, all 3D models were in (originally or after conversion) point cloud format.

4) Point cloud voxelization: We perform point cloud voxelization using 10 bit quantization. That is, the spatial coordinates are normalized such that they are integers between 0 and 1023. This has two main advantages: first, the coordinates are in a range that is predictable for point cloud processing but also with respect to rendering and second, we use voxelized coordinates in combination with cube based rendering to improve stability, predictability and quality of renderings.

[7]https://www.blender.org/
[8]https://www.meshlab.net/
[9]https://www.cloudcompare.org/main.html

C. Compression

Compression is a crucial stage for various point cloud storage and transmission applications, including streaming and telepresence. It is also the most realistic distortion source for point clouds. In this context, one needs to compress point clouds to transfer them to the remote receiver under current bandwidth limitations.

We selected four compression algorithms:

- GPCC-Octree-Predlift (noted as GPCC-Predlift),
- GPCC-Octree-RAHT (noted as GPCC-Raht),
- VPCC, and
- GeoCNN.

VPCC and GPCC were selected as they are part of MPEG standardization efforts, and they are among the most commonly used compression methods. GeoCNN was selected to introduce artifacts of a learning-based compression to BASICS database. Sample results are shown in Figure 2.

GPCC [15] compresses the point cloud using the octree structure which can find the occupancy of the points in 3D space without projection to 2D space. GPCC uses either octree or trisoup (Triangle soup) approaches – based on a pruned octree. As octree and trisoup only focus on geometry compression, attributes (such as color) are compressed using region-adaptive hierarchical transform (RAHT) and predicting/lifting transform (indicated as pred/lift or predlift).

For GPCC, it was noticed during the pilot tests that the trisoup generated uneven structure and holes on the reconstructed point clouds. Therefore, in this database, only octree is used for GPCC geometry coding, together with the two attribute coding methods: Octree-Predlift and Octree-Raht. For compression, the six quality parameter (QP) values from common test conditions (CTC)[citation?] were used. The worst

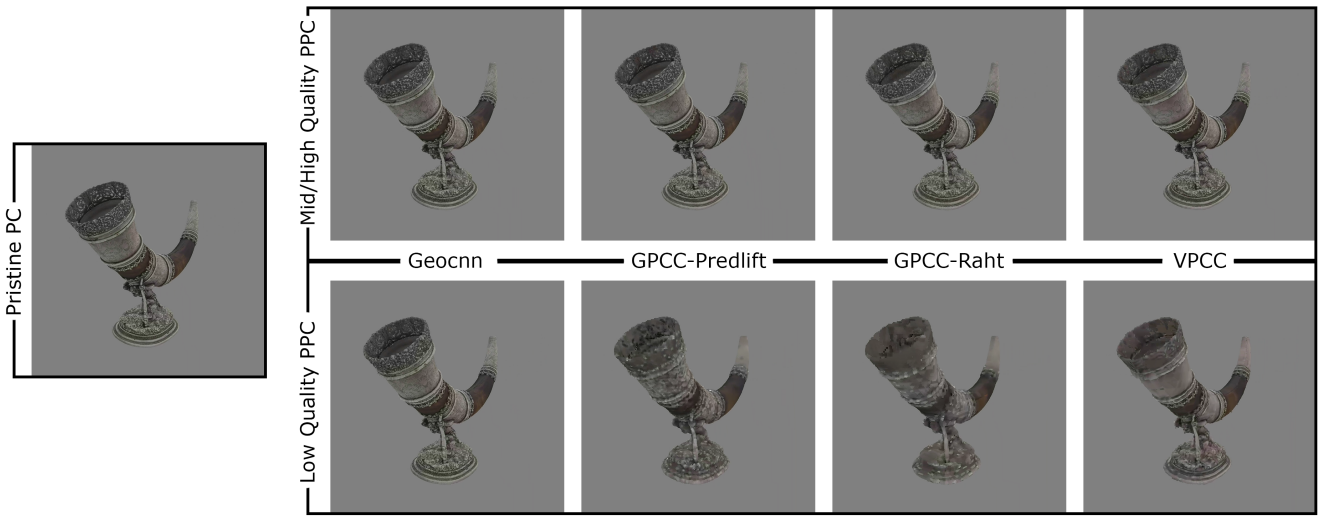

Fig. 2. Sample frames from the video renderings of a selected processed point cloud (PPC), showing results for each compression algorithm.

quality level (i.e., r1) was discarded in the pilot tests because its quality was too bad and this would affect the subjective quality experiment adversely by changing the rating scale and the participants' votes. With this change, the number of quality levels for GPCC became five: $r_{GPCC} \in \{2, 3, 4, 5, 6\}$.

VPCC [15] is the video-based point cloud compression approach, in which the point cloud is projected to the sides of a cube and the projection is coded using traditional video compression methods, such as HEVC/H.265 or VVC/H.266. The projection is done for both color information and also the depth information (i.e., the inherent 3D structure of points in 3D). Utilizing the inherent temporal coding capabilities of traditional video codecs, VPCC can effectively compress dynamic point cloud sequences.

For VPCC, we used the "all-intra" mode, and the compression levels were taken from the "longdress" sequence QP as determined in the CTC. Nevertheless, it was noticed that the given bitrates in the CTC seemed to yield a much higher quality than GPCC, therefore, another quality level below the ones given in CTC was added. This quality level is called *rate_zero* with $gQP = 36$ and $tQP = 47$. With this change, the number of quality levels for VPCC became six: $r_{VPCC} \in \{0, 1, 2, 3, 4, 5\}$.

GeoCNN [16] compresses voxelized point clouds by first performing block partitioning. Then, each block is passed to a variational autoencoder where the encoder transforms the input binary occupancy voxel grid to a latent space. The latent space is then quantized and entropy coded using a learned entropy model. After entropy decoding the bitstream, the latent space is transformed back to a voxel grid containing predicted occupancy probabilities. The probabilities are then thresholded to binary values which yields the decoded block. With the result of each block, the entire decompressed point cloud is obtained. Four different quality levels were used for GeoCNN. $r_{GeoCNN} \in \{1, 2, 3, 4\}$.

III. Subjective quality assessment

We conducted a large-scale subjective experiment using the Prolific [17] crowdsourcing platform with 3654 participants. 1494 processed point clouds (PPC) from 75 original point clouds (SRC) were generated for the experiment with each PPC were evaluated by 60 unique participants on average. Consequently, we successfully accumulated approximately 90,000 subjective opinion scores. This section describes the details regarding the crowdsourcing study.

A. Generating Visual Stimuli

In a voxelized point cloud, a voxel is occupied if at least one point of the point cloud is within the voxel. Each voxel is rendered as a cube spanning its volume. This is different from the more common point-based method ("point" as OpenGL point primitive) where points are rendered as screen-aligned squares of a given window space size. The main drawback of the point-based rendering is its susceptibility to scaling issues: as the viewer zooms in or out, the points can either become smaller or larger, impacting the perceived density of the point cloud. Furthermore, point-based rendering often results in flicker artifacts, particularly when spatial overlaps occur during perspective changes. Cube-based rendering addresses these issues, ensuring consistent rendering quality across various perspectives and effectively eliminating flicker artifacts.

In practice, we specify cube sizes. For octree-based methods, the size of the cube is defined based on the number of removed octree levels n_r. With n_l bit quantization, the maximum is n_l levels. Thus, we specify the size of the cube as $2^{n_l - n_r}$: that is, a size of 1 when lossless, a size of 2 when removing one octree level, a size of 4 when removing two octree levels, etc…For other methods, the cube size is determined empirically to ensure water-tight rendering.

We employed a helix-like rendering trajectory, as illustrated in Figure 3. The frontal orientation of each point cloud was manually designated, and the rendering trajectory consistently

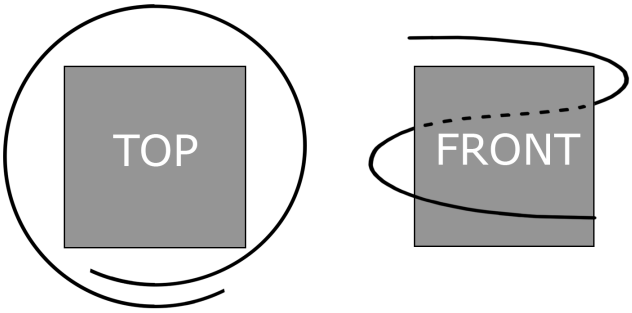

Fig. 3. Visualization of the rendering trajectory from top and front views.

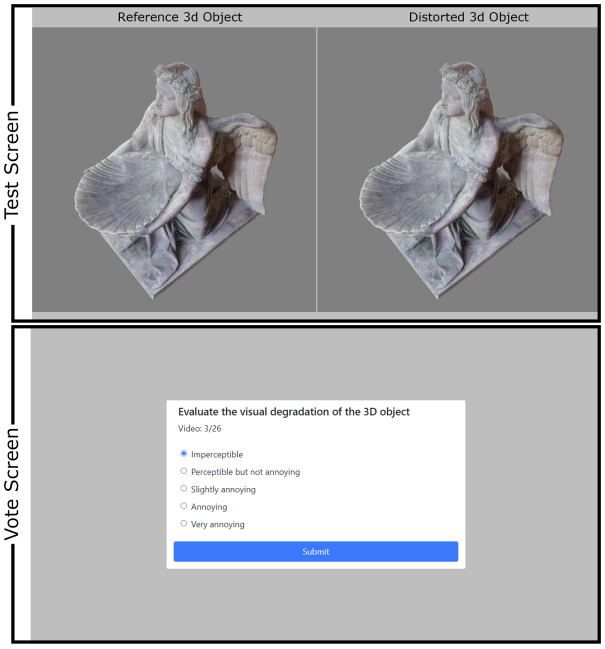

Fig. 4. Sample screenshots from the experiment. Rendered point cloud videos were shown side-by-side (above), and each stimulus was followed by a voting screen (below).

started from this predefined front side. To ensure that the front side of the point cloud remained visible either at the beginning or at the end of the rendered video, we introduced a slight overlap between the trajectory's starting and ending points. Certain point clouds are unnatural to observe from lower angles (e.g., landscape, buildings). Therefore, each point cloud was individually categorized as either "low", "mid" or "high" to determine the initial elevation of the rendering trajectory. While moving on the rendering trajectory, the camera was always oriented toward the center of the point cloud.

B. Methodology

Subjective quality assessment of point cloud content can be broadly categorized as interactive and passive [18]. In the interactive paradigm, observers have the liberty to examine the point cloud from any desired point of view, often within the context of augmented reality or virtual reality applications. In contrast, the passive approach involves rendering point clouds in the form of a conventional video with a predetermined camera trajectory. Although both paradigms have their own advantages and drawbacks, there is no statistically significant difference between the subjective opinions collected with each approach [19]. In order to minimize the variance between observer opinions and facilitate a more practical data collection through crowdsourcing, we opted for the passive approach [20].

Several methodologies can be found in the literature and recommendations for subjective quality assessment of traditional image and video sequences [21]. Commonly employed methodologies include, but are not limited to, Absolute Category Rating (ACR), Double Stimulus Impairment Scale (DSIS), Two-Alternative Forced Choice (2AFC). Several studies compared the accuracy and reliability of each methodology for diverse types of multimedia content. In the realm of traditional images and videos, it is shown that the pair comparison methodology tends to be more accurate due to straightforward experiment procedure and there is no statistically significant difference between ACR and DSIS methodology [22]. However, it's worth noting that the pair comparison methodology may become impractical when dealing with a substantial number of test conditions due to the exponential increase in required comparisons [23]. On the other hand, the recent study by Nehme *et al.* [24] suggests that the DSIS method is more accurate than ACR for 3D graphical content. The rationale behind this assertion is that in ACR experiments, participants unfamiliar with the pristine models may struggle to discern various types of distortions. DSIS methodology enhances accuracy by presenting both the reference and the distorted model before the rating phase, allowing for a more informed evaluation. In line with these findings, we adopted the DSIS methodology with a side-by-side presentation format, as recommended by Nehme *et al.* [24]. A screenshot depicting the stimuli presented to the participants is featured in Figure 4.

C. Test Procedure

Previous studies have indicated that crowdsourcing experiments can yield results of comparable accuracy to traditional laboratory experiments across various Quality of Experience (QoE) tasks and with diverse experiment designs [20], [25]. In light of these findings, we chose to leverage the Prolific crowdsourcing platform to recruit participants and conduct our subjective experiment. Prolific ensures transparency and ethical participation by clearly communicating to participants that they are taking part in a research study. The experiment requirements are thoughtfully balanced to benefit both researchers and participants alike [26]. This approach not only provides access to a broad pool of participants but also expedites the data collection process.

Test sessions & Duration: In crowdsourcing settings where participants lack supervision during the experiment, it is essential to limit both the number of stimuli and the duration of the test compared to laboratory experiments [25]. To accommodate this requirement, we divided the experiment into 60 sessions, each containing 25 stimuli and 2 "dummies".

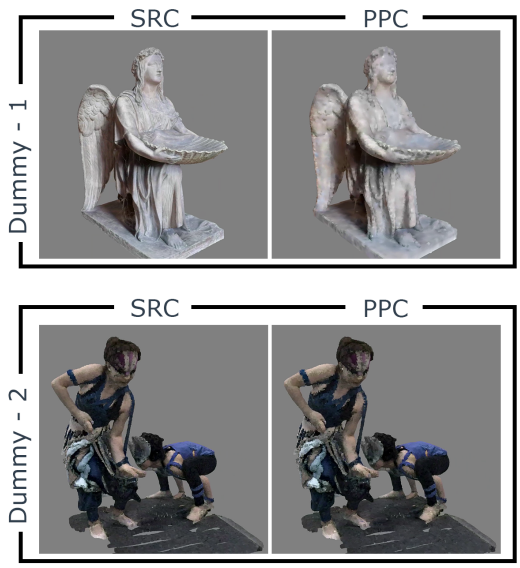

Fig. 5. Selected frames from the SRC and PPC video renderings of the 2 dummies used in all playlists.

Sample frames from the two dummy stimuli are presented in Figure 5. For training purposes, one dummy from the highly compressed stimuli and one dummy from the least compressed stimuli were uniformly presented to every participant. These dummy stimuli remained consistent for all participants, and participants were not informed that these stimuli were included for training purposes. In total, each participant rated 27 stimuli of 10-seconds rendered videos. With unlimited voting time after each stimulus presentation, the average duration of the test sessions amounted to approximately 5 minutes and 30 seconds.

Participants & Requirements: We recruited 60 participants (50% female - 50% male) on average per session, 3654 participants in total. The age of the participants range from 18 to 70. Each participant was compensated for their time and effort in line with the Prolific requirements. Moreover, to uphold the integrity of the experiment and guarantee that all stimuli were presented as intended, participants were constrained to use specific browsers operating in full-screen mode at a resolution of 1080p. Additionally, participants were required to meet specific qualifications, namely completing at least 200 submissions on the Prolific platform with a 100% approval rate. These stringent prerequisites helped ensure the reliability and commitment of the participants.

IV. Subjective Experiment Results

This section presents our analysis on the collected subjective quality scores. In Section IV-A, we discuss the observer reliability and provide an overview of the results obtained from the observer screening tools we have applied. Additionally, in Section IV-B, we present our findings on the performance of compression algorithms.

A. Observer Screening

In addition to the recruitment requirements of the participants (see Section III-C), the "dummy stimuli" described above were also used as trap questions to detect participants engaging in malicious behavior. Moreover, post-experiment observer screening tools were employed to enhance the reliability of the subjective opinion scores.

As described in Section III, the dataset underwent evaluation by a total of 3,654 participants, divided into 60 smaller playlists, each receiving an average of 60 participants. Within each playlist, we included 2 dummy stimuli (depicted in Figure 5) designed to calibrate participants' expectations regarding the extent of distortions present in the experiment. Participants were not informed that these initial 2 stimuli were for training purposes and evaluated them like regular stimuli. Source and processed point cloud renderings were displayed side-by-side like the rest of the stimuli in the experiment (see Figure 4 for what is displayed to observers). We selected the first dummy with the highest level compression in the dataset (resulting in a low-quality PPC), and the second dummy with the lowest level of compression (resulting in a high-quality PPC). This deliberate selection allowed us to expose each participant to both ends of the quality scale, aligning their expectations regarding the extent of degradation. Notably, if a subject rated the first dummy as "imperceptible" due to its clear distortions, we excluded them from the experiment, removing a total of 22 subjects.

As post-experiment observer screening, we applied the two common ITU standards as well as a recently proposed model called ZREC [21], [27], [28][10]. Our objective in employing these methodologies was to assess the validity of the collected subjective opinion scores and establish reliable MOS. While the three methods differ fundamentally, their primary aim is to reduce confidence intervals (CI) by identifying and, if necessary, excluding outliers from the collected data. Although there is a high correlation between the MOS obtained through each method, the average 95% CIs differ significantly.

Initially, we calculated the average 95% CI by taking the mean of the 95% CI of all stimuli without applying any post-screening methods, referred to as raw-MOS. Subsequently, we implemented the outlier rejection as recommended in ITU.BT500 [21]. Notably, BT500 identified only 43 subjects as outliers out of the total 3,633 participants. As a result, the impact on both the MOS and the 95% CI was minimal, with the 95% CI (0.1982) remaining almost as high as the raw-MOS 95% CI. Conversely, we observed a more substantial shift in the acquired MOS when using P913-12.6 and ZREC, which yielded much lower confidence interval values. Specifically, we calculated an average 95% CI of 0.1697 and 0.1463 for P913-12.6 and ZREC, respectively. Based on these findings, we utilized the MOS acquired via ZREC due to the lower CI. To promote further investigation, we provide both the raw opinion scores and estimated MOS using ZREC in the public repository of the dataset.

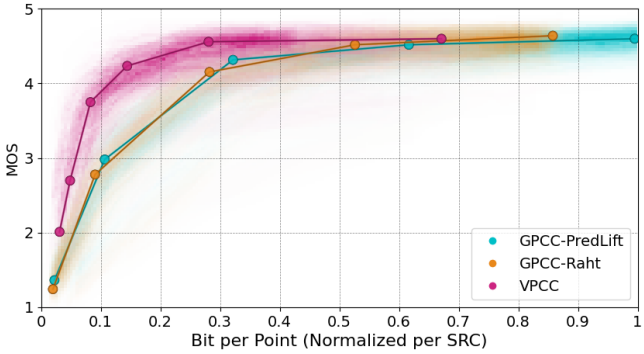

Fig. 6. Heatmaps represent the MOS-BPP curve densities of each compression algorithm. Centers of the 2 dimensional Gaussian distribution fit over each compression level for each algorithm is plotted as a line.

B. Performance of compression algorithms

Although BASICS dataset contains 4 compression algorithms, in this analysis, we excluded GeoCNN. Since GeoCNN only compresses geometry and is complemented with uncompressed color information for visualization, it is not fair to directly compare it with GPCC-Predlift, GPCC-Raht and VPCC, which code both geometry and color. When comparing the performance of GPCC-Predlift, GPCC-Raht and VPCC codecs, we employed two complementary approaches that analyze and visualize the MOS vs Bit-Per-Point (BPP) behavior of each compression algorithm. Due to large differences over BPP values per SRC, we first normalized all BPP values per SRC. We achieved this by dividing each PPC with the highest BPP value of the PPC across all compression algorithms that belongs to the corresponding SRC. Consequently, we could operate within a BPP range of 0 to 1, allowing a relative comparison of the compression algorithms.

The first approach draws inspiration from DenseLines [29], a method originally devised to visualize vast quantities of time series data. To construct a Mean Opinion Score-Bit-Per-Point (MOS-BPP) curve density map for each compression algorithm, we partitioned the grid into square cells. Subsequently, for a given compression algorithm and SRC, we assigned a value of 1 to each cell intersecting with the polygon defined by the MOS and 95% CI values of each PPC. For each SRC, we normalized the marked cell values by dividing them by the total number of intersecting cells, effectively illustrating the effect of lower 95% CI and BPP values. Then, for each compression algorithms, we aggregated the normalized cell values from all SRCs and performed a linear mapping to scale the results into [0,1] range. Outputted values then plotted as overlaying heatmaps (with increasing transparency from 0 to 1), as showcased in Figure 6.

In the second approach, we employed a 2D Gaussian distribution to model each compression level within each compression algorithm. Subsequently, we extracted the centers of these 2D Gaussian distributions and utilized them to generate lines that captured the average MOS-BPP curves for each compression algorithm. These results were then plotted as lines superimposed upon the curve density heatmaps, providing a concise representation of each compression algorithm's performance.

As depicted in Figure 6, there is no significant difference between the GPCC-Predlift and GPCC-Raht. At lower BPP values, VPCC exhibits superior performance compared to both GPCC variants. However, as BPP values increase, the distinction between these compression algorithms becomes less pronounced. It's worth noting that while there may be occasional exceptions to these observations, they are minor in nature and do not substantially deviate from the trends observed in the analysis. We encourage interested readers to check the dataset public repository, where the MOS-BPP plots for each SRC is provided.

V. Benchmark of Objective Quality Metrics

Point cloud objective quality metrics can be categorized into three classes, based on the input to the metrics: image-based, color-based and geometry-based. Image-based metrics take the rendered point cloud image or image sequences as input and assess the quality of the point clouds. Geometry-based metrics primarily rely on the geometric information stored at each point in the point cloud, without considering color attributes. Color-based metrics, on the other hand, use the color information of each point to assess point cloud quality. Some metrics, such as PCQM [30], have the capability to utilize both geometry and color information for quality assessment.

Furthermore, each metric can be categorized into three based on the presence of reference point cloud information as full-reference (FR), reduced-reference (RR) and no-reference (NR). FR metrics access all information from the reference point cloud in addition to the distorted point cloud. RR metrics can access only partial information (features) from the reference point clouds. NR metrics assess the quality of the point cloud without any access to the reference point cloud.

In this section, we benchmark 14 image-based, 9 color-based and 17 geometry-based metrics from the literature. Some of these metrics were omitted from the results due to minor differences to their variants. An introduction to the selected metrics is provided in Section V-A. We employ various figures of merit and evaluation scenarios to assess the performance of these metrics, which are outlined in Section V-B. The results of the metric evaluations are presented in Section V-C, Section V-D, and Section V-E.

A. Selected Metrics

For all image-based metrics, average pooling over 30 fps video renderings has been used to predict the final quality as recommended in [31]. Image-based metrics have been computed on the rendered frames used during the subjective test, and include simple measures such as MSE, PSNR, SSIM [32], MS-SSIM [33], and 11 other more sophisticated metrics. Feature Similarity Index (FSIM [34]) and its color-dependent variant, FSIMc [34], fall under the category of full-reference metrics. They rely on phase congruency and gradient magnitude to locally quantify image quality, utilizing

[10]https://github.com/kyillene/ZREC

phase congruency as a weighting function to yield a single quality score. Gradient Magnitude Similarity Deviation (GMSD [35]) is another full-reference metric that employs pixel-wise gradient magnitude similarity to predict image quality. D-JNDQ [36] is a learning-based full-reference metric trained on the first Just Noticeable Difference (JND) points of JPEG compression artifacts. It combines a white-box optical and retinal pathway model with a Siamese neural network to predict image quality. MW-PSNR [37], [38] relies on morphological wavelet decomposition and the Mean Squared Error (MSE) of the wavelet sub-bands. Our evaluation includes both full-reference (MW-PSNR-FR) and reduced-reference (MW-PSNR-RR) versions of this metric. ADM2 [39] assesses image quality by separating detail losses and additive impairments. It encompasses features also used in the Video Multi-method Assessment Fusion (VMAF) metric [40]. VIF [41] quantifies the information present in the reference image and how much of this reference information can be extracted from the distorted image. It is another feature used in the VMAF metric [40]. VMAF [40], proposed by Netflix, fuses several image-based features, including ADP2 and VIF, along with a simple temporal feature to evaluate video quality. FVVDP [42] models the response of the human visual system to differences across the temporal domain and the visual field.

In addition to image-based metrics, several geometry-based metrics were also evaluated over the dataset. In the last decade, three fundamental approaches were proposed to evaluate PC quality focusing on point-to-point [43] (p2point), point-to-plane [44] (p2plane) and plane-to-plane [45] (pl2plane) differences in 3D space. The p2point and p2plane metrics are computed using either mean square error (MSE) or peak peak signal-to-noise ratio (PSNR). In this context, the term "plane" refers to the surface of a point defined by its normal vector. The pl2plane metrics are computed using either MSE or root mean square (RMS). While categorized as a geometry-based metric, PCQM [30] uses a linear combination of several geometry-based (curvature comparison, curvature contrast, and curvature structure) and color-based features (lightness comparison, lightness contrast, lightness structure, chroma comparison, and hue comparison) to assess the visual quality of a point cloud. PointSSIM [46] offers three geometry-based variants with slight differences in both implementation and performance.

Moreover, color differences between the points in reference and distorted PC can be quantified with PSNR to estimate the visual quality of the distorted PC. We applied this metric on Y, U, and V channels separately and referred as Color-Y-PSNR, Color-U-PSNR, and Color-V-PSNR respectively. 3 color-based variants of the PointSSIM [46] metric were also included in the evaluation.

B. Evaluation Criteria

We evaluate the performance of the selected metrics in three different scenarios. In Section V-C, we aim to assess the perceptual quality of the PPC covering the *entire quality range*. This is the most generic and traditional evaluation scenario for quality metrics. In Section V-D, we evaluate the performance of the metrics in the *high-quality range*. This scenario is particularly relevant for applications such as content production, high-quality streaming, and digital twins, where maintaining the highest visual fidelity is crucial. In Section V-E, we evaluate the metric performances for their *sensitivity to quality differences* within different versions of the same point cloud content. This evaluation scenario is especially suitable to optimization scenarios such as point cloud compression and enhancement, and as loss functions in end-to-end PC learning pipelines. For broad and high-quality range evaluation scenarios, we employ traditional correlation measures, while Krasula's method [47] is used for the Intra-SRC evaluation scenario.

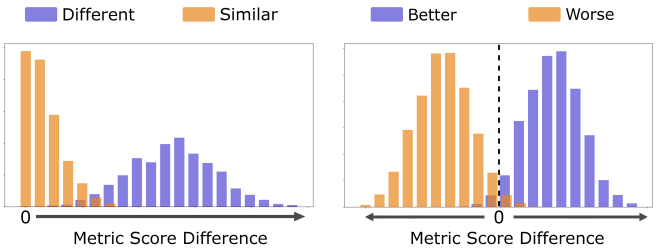

Fig. 7. Ideal distributions of metric score differences for "Different vs Similar" and "Better vs Worse" analysis. A greater metric score difference is expected for different pairs in "Different vs Similar" analysis. For "Better vs Worse" analysis, metric score differences are expected to be positive and negative respectively for better and worse pairs.

Correlation measures: Pearson's linear correlation coefficient (PLCC) measures the prediction accuracy of the objective metrics and Spearman's rank-order correlation coefficient (SROCC) measures the strength of prediction monotonicity [48]. Following the recommendations [21], [48], a 5-parameter logistic function was fitted prior to evaluation. For both PLCC and SROCC, the values are in the range [0, 1] and higher values indicate a better correlation.

Krasula's method [47]: It involves two stages of analysis: "Different vs Similar" and "Better vs Worse". In the "Different vs Similar" analysis, pairs of PPC from the dataset are split into two categories as pairs with (*i.e., different*) and without (*i.e., similar*) statistically significant differences. For a given pair of PPC, one way ANOVA followed by Tukey's honest significance difference test [49] is used to measure the statistical significance of the differences. The assumption is that the absolute differences in metric predictions for "different" pairs should be larger than the "similar" pairs. Receiver Operating Characteristic (ROC) analysis is used to quantify the performance, and expressed as Area Under the ROC Curve (AUC). In the "Better vs Worse" analysis, pairs identified as "different" in the first stage are used. In this stage, the aim is to measure how well the metrics identify the better PPC in pairs with statistically significant difference. Metric performance in this stage can be expressed as the correct classification percentage as well as AUC values, similar to the first stage. An illustrative example of how metric score differences should be distributed for each stage is depicted in Figure 7. In "Different vs Similar" analysis, higher metric score differences

TABLE III
COLUMNS "ALL" PRESENT THE PEARSON AND SPEARMAN CORRELATION COEFFICIENTS BETWEEN THE LISTED METRIC PREDICTIONS AND MOS (WITH ZREC [28]) OF THE ALL PPC IN THE **BROAD QUALITY RANGE**. MOREOVER, METRIC PERFORMANCES FOR EACH COMPRESSION ALGORITHM ARE ALSO INDIVIDUALLY REPORTED IN THE CORRESPONDING COLUMNS.

		All		GeoCNN		GPCC Predlift		GPCC Raht		VPCC	
Category	Metric	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
Image Based	MSE	0.2387	0.2226	0.4099	0.4183	0.3349	0.2578	0.3194	0.2387	0.1306	0.0649
	PSNR	0.2488	0.2269	0.4220	0.4231	0.3395	0.2621	0.3238	0.2429	0.1149	0.0691
	SSIM [32]	0.6119	0.5431	0.5496	0.5373	0.7491	0.6145	0.7599	0.6365	0.4034	0.3675
	MS-SSIM [33]	0.5481	0.4607	0.4944	0.4813	0.6744	0.5286	0.6847	0.5473	0.3553	0.2951
	FSIM [34]	0.6335	0.5612	0.5724	0.5714	0.7610	0.6308	0.7683	0.6501	0.4042	0.3748
	FSIMc [34]	0.6334	0.5608	0.5713	0.5703	0.7607	0.6298	0.7679	0.6495	0.4040	0.3745
	GMSD [35]	0.6716	0.6113	0.6331	0.6336	0.7988	0.6718	0.7944	0.6854	0.4556	0.4390
	D-JNDQ [36]	0.6732	0.6313	**0.7215**	**0.7229**	0.7891	0.6762	0.8017	0.7001	0.4352	0.4402
	MW-PSNR-FR [37]	0.3479	0.3296	0.4590	0.4577	0.4471	0.3766	0.4366	0.3668	0.1956	0.1624
	MW-PSNR-RR [38]	0.5110	0.4975	0.5704	0.5647	0.6250	0.5591	0.6275	0.5600	0.3011	0.3107
	ADM2 [39]	0.7283	0.6520	0.6434	0.6216	0.8408	0.6887	0.8362	0.7073	0.5513	0.5386
	VIF-scale3 [41]	0.6492	0.5947	0.5962	0.6035	0.7705	0.6525	0.7745	0.6717	0.4311	0.4249
	VMAF [40]	**0.7419**	**0.6686**	0.6572	0.6391	**0.8540**	**0.7091**	**0.8532**	**0.7310**	**0.5541**	**0.5412**
	FVVDP [42]	0.6936	0.6417	0.6312	0.6457	0.8197	0.6983	0.8300	0.7255	0.4676	0.4631
Color Based	Color-Y-PSNR	0.5376	0.5282	0.2166	0.2448	0.7407	0.7018	0.7567	0.7208	0.4251	0.4335
	Color-U-PSNR	0.5432	0.5105	0.2585	0.1254	0.6814	0.6529	0.6602	0.6291	0.3838	0.3751
	Color-V-PSNR	0.5729	0.5411	0.3034	0.2661	0.7014	0.6755	0.6893	0.6614	0.4428	0.4238
	PointSSIM-ColorAB [46]	**0.7291**	0.6907	0.5855	0.3936	0.8021	0.7962	0.8249	0.8278	**0.7378**	**0.7675**
	PointSSIM-ColorBA [46]	0.7241	**0.6928**	**0.6044**	**0.4312**	**0.8033**	**0.7984**	**0.8257**	**0.8288**	0.7287	0.7611
	PointSSIM-ColorSym [46]	0.7250	0.6919	0.6021	0.4296	0.8028	0.7975	0.8253	0.8279	0.7317	0.7633
Geometry Based	p2point-MSE [43]	0.8427	0.7759	0.6060	0.6114	**0.9718**	0.8863	**0.9707**	0.9002	0.7221	0.7154
	p2point-PSNR [43]	0.6827	0.4850	0.2798	0.2257	0.7697	0.5876	0.7822	0.6164	0.5172	0.4294
	p2plane-MSE [44]	0.8866	**0.8370**	**0.6865**	**0.6415**	0.9681	**0.8886**	0.9678	**0.9028**	0.7915	0.8068
	p2plane-PSNR [44]	0.7001	0.5164	0.3216	0.3113	0.7576	0.5794	0.7708	0.6071	0.5880	0.4816
	pl2plane-Mean [45]	0.1393	0.1272	0.3390	0.1193	0.1730	0.1368	0.1753	0.1610	0.1029	0.0994
	pl2plane-RMS [45]	0.1197	0.1002	0.2205	0.0553	0.1511	0.1195	0.1565	0.1448	0.0942	0.0777
	pl2plane-MSE [45]	0.1189	0.1002	0.3334	0.2193	0.1508	0.1195	0.1562	0.1448	0.0942	0.0777
	PCQM [30]	**0.8878**	0.8102	0.4475	0.2965	0.9510	0.8746	0.9584	0.8953	**0.8507**	**0.8332**
	PointSSIM-GeomAB [46]	0.7760	0.7196	0.5497	0.5469	0.9067	0.8510	0.9091	0.8784	0.5439	0.5527
	PointSSIM-GeomBA [46]	0.7644	0.7145	0.5572	0.5783	0.9075	0.8477	0.9107	0.8757	0.5102	0.5148
	PointSSIM-GeomSym [46]	0.7731	0.7226	0.5566	0.5767	0.9094	0.8510	0.9116	0.8782	0.5677	0.5493

are expected for "Different" pairs and lower differences for "Similar" pairs. In the "Better vs Worse" analysis, for pairs categorized as "Better", positive metric score differences are expected, indicating that the first point cloud in the pair is better than the second one. Conversely, for pairs categorized as "Worse", negative metric score differences are expected, indicating that the first point cloud in the pair is worse than the second one.

C. Broad Quality Range Evaluation

Broad quality range evaluation scenario is the generic and commonly used use-case in the literature and it typically involves calculating metric performances with the traditional correlation measures (*e.g.,* PLCC, SROCC, etc.). The entire dataset is used for this evaluation, and resultes reported as PLCC and SROCC values between the metric predictions and the collected MOS.

Table III presents the PLCC and SROCC of each metric in this evaluation scenario. The metrics are categorized into three group based on input type that they are operating on, as previously discussed in Section V-A. The first two column show the metrics' PLCC and SROCC scores on the entire dataset. Additionally, metric performances were evaluated for individual compression algorithms and the results are presented in subsequent columns as indicated above each column.

PCQM [30] and p2plane-MSE [44] exhibit the best performance on the entire dataset among the selected metrics, despite their poorer performance in predicting GeoCNN compression distortions. Among color-based metrics, we again notice a similar pattern on the accuracy of metrics when it comes to GeoCNN compression distortions. PointSSIM [46] variants perform relatively better than other metrics in this category.

Simple image-based metrics (*e.g., MSE, PSNR, SSIM [32], MS-SSIM [33]*) have low accuracy across all compression categories and consequently on the whole dataset. VMAF [40] shows the best performance among image-based metrics in the whole dataset. We also observe a general trend among image-based metrics towards a lack of accuracy on VPCC compression distortions. On another note, we observe that D-JNDQ [36] performs the best to predict GeoCNN distortions among the selected metrics, despite not being retrained on the dataset.

D. High Quality Range Evaluation

High-quality range evaluation is crucial for applications that aim to deliver top-tier content, such as high-quality streaming and digital twins. It's important to note that metrics that perform well in the general quality range may not exhibit the same level of accuracy in the high-quality range. Therefore, we conducted an analysis to assess the accuracy of quality metrics

TABLE IV
COLUMNS PRESENT THE PEARSON AND SPEARMAN CORRELATION COEFFICIENTS BETWEEN THE LISTED METRIC PREDICTIONS AND MOS (WITH ZREC [28]) OF THE PPC IN THE **HIGH-QUALITY RANGE** WHERE $MOS \geq 3.5$.

Category	Metric	PLCC	SROCC
Image Based	MSE	0.1187	0.0920
	PSNR	0.1476	0.1397
	SSIM [32]	0.2901	0.2119
	MS-SSIM [33]	0.2697	0.1600
	FSIM [34]	0.3100	0.2508
	FSIMc [34]	0.3100	0.2502
	GMSD [35]	0.3503	0.2993
	D-JNDQ [36]	**0.4120**	**0.3771**
	MW-PSNR-FR [37]	0.1548	0.1490
	MW-PSNR-RR [38]	0.2795	0.2633
	ADM2 [39]	0.3171	0.2750
	VIF-scale3 [41]	0.2697	0.1600
	VMAF [40]	0.3466	0.3067
	FVVDP [42]	0.3598	0.3278
Color Based	Color-Y-PSNR	0.2751	0.2728
	Color-U-PSNR	0.1808	0.1808
	Color-V-PSNR	0.2192	0.2051
	PointSSIM-ColorAB [46]	0.4530	0.3938
	PointSSIM-ColorBA [46]	0.4545	**0.4052**
	PointSSIM-ColorSym [46]	**0.4547**	0.4004
Geometry Based	p2point-MSE [43]	0.4898	0.4221
	p2point-PSNR [43]	0.1077	0.0774
	p2plane-MSE [44]	**0.5708**	**0.5418**
	p2plane-PSNR [44]	0.1067	0.1512
	pl2plane-Mean [45]	0.0292	0.0511
	pl2plane-RMS [45]	0.0770	0.0939
	pl2plane-MSE [45]	0.0494	0.0642
	PCQM [30]	0.5038	0.4775
	PointSSIM-GeomAB [46]	0.4418	0.4527
	PointSSIM-GeomBA [46]	0.4466	0.4617
	PointSSIM-GeomSym [46]	0.4455	0.4615

specifically on the high-quality part of the dataset, where the MOS is greater than or equal to 3.5. This evaluation helps identify which metrics excel in scenarios where maintaining exceptionally high quality is a priority.

In this evaluation scenario, metric performances are relatively low overall, as indicated in Table IV. However, the relative order of the metrics in terms of their performances remains relatively consistent with the broad quality range evaluation. p2plane-MSE [44] performs the best overall while D-JNDQ [36] is the best performing image-based metric and the PointSSIM-ColorBA [46] is the best performing color-based metric for this use-case.

E. Intra-SRC Evaluation

Intra-SRC evaluation focuses on assessing metrics' performance when comparing PPC derived from a single pristine SRC at various compression levels. It allows us to gauge how well metrics can discriminate between different compression levels originating from the same SRC, helping us optimize processes that rely on such discrimination. This evaluation scenario is valuable for applications like fine-tuning compression and enhancement algorithms, training machine learning models for end-to-end processing, and other situations where fidelity is a primary concern.

Prior to the analysis, we preprocess the subjective scores as described in Krasula's method [47]. First, 20 PPCs were paired within each source point cloud, generating $(20 \times (20-1)/2)$ pairs per SRC. In total, we end up with 14143 pairs. Afterwards, a one-way ANOVA test is applied to individual scores collected for each stimulus in each pair, followed by Tukey's Honest test. 5019 pairs among the total 14143 were identified as "Similar" whereas 9124 contains a statistically significant different between the two PPC and thus identified as "Different". From those "Different" pairs, we split them into two roughly equal-sized groups as "Better" and "Worse" depending on the order of the pair. There are 4075 "Better" and 5049 "Worse" pairs.

Due to space limitations of the manuscript, we report the result of the analysis only on the selected 6 metrics among the initial list presented in Section V-A. The rest of the results can be acquired via the provided scripts in the GitHub repository of the dataset.

Different vs Similar Analysis: The top row in Figure 8 presents the results of the analysis as histograms of metric score differences for "Different" and "Similar" pairs. We expect better-performing metrics to provide metric score distributions similar to the ideal case as depicted in Figure 7. Additionally, performance of each metric quantified with AUC values, reported under each metric name. We observe a better performance from PCQM, providing a higher AUC value and a very similar distribution to the ideal case. Statistical significance tests on this task also reveals that PCQM performs significantly better than all other metrics in "Different vs Similar" task.

Better vs Worse Analysis: Similar to the previous stage, bottom row of the Figure 8 presents the results as histograms of metric score differences and quantifies the performance of each metric with AUC and CC values. We observe that most metrics perform relatively well on identifying "Better" and "Worse" pairs apart. PCQM shows a very similar distribution to the ideal case depicted in Figure 7 as also reflected by the AUC and CC values.

VI. CONCLUSION

We conducted a large-scale crowdsourcing study on point cloud compression quality assessment. To the best of our knowledge, this is the largest publicly available point cloud quality assessment dataset. It contains 75 source point clouds, each compressed with 4 different compression algorithms resulting in nearly 1500 processed point clouds. More than 3500 naive observers participated to the experiment.

Our study revealed several noteworthy insights regarding objective quality metrics' performances. While most point cloud objective quality metrics perform well in predicting GPCC distortions, the majority of the metrics still struggle with VPCC distortions. Furthermore, an even larger majority fall short in assessing the quality GeoCNN distortions, a learning-based compression algorithm. This highlights a pressing need for improved quality metrics capable of accurately evaluating learning-based compression distortions. Given the scarcity of learning-based compression algorithms in publicly available datasets (see Table I), the need for better quality metrics that can accurately predict learning-based compression distortions is revealed in this work.

Additionally, we observed significant room for improvement in metrics designed for high-quality content. The correlations

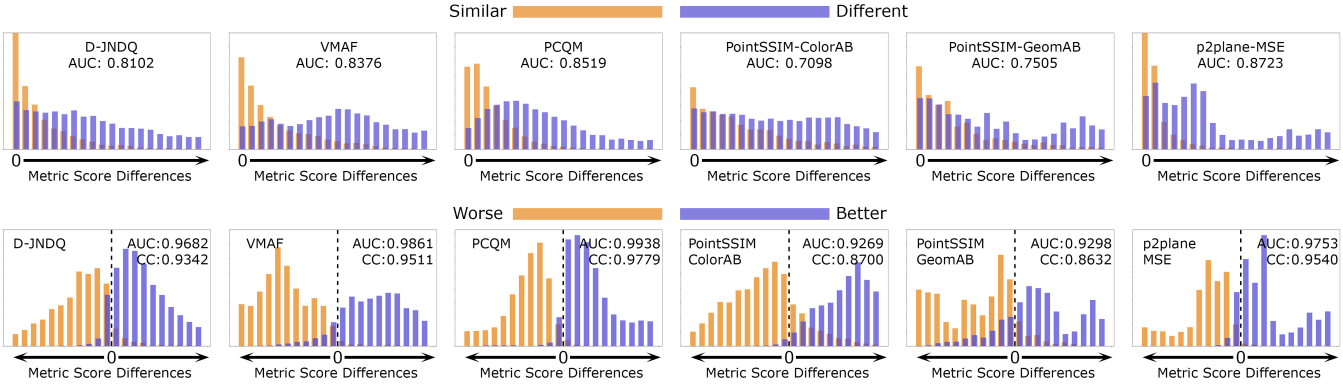

Fig. 8. The plots in the top row depict the metric score differences for pairs categorized as "Different" and "Similar" for the **Intra-SRC evaluation**. Each metric's score differences are individually normalized within minimum and maximum ranges. The height of the bars represents the number of occurrences, and each bar ranges between 0 and 1500. The metric names are indicated at the top of each plot, and the AUC values are reported below each metric name. Similarly, the plots in the bottom row show the metric score differences for pairs categorized as "Better" and "Worse" for the Intra-SRC evaluation. Again, the metric score differences are individually normalized within minimum and maximum ranges. The height of the bars denotes the number of occurrences, and each bar ranges between 0 and 800. The metric names are indicated at the top left corner of each plot and the AUC and CC are reported in the top right corner of each plot.

between the MOS and the best-performing metric predictions (p2plane-MSE in this scenario) remained below 0.60. On a more positive note, our intra-SRC evaluation scenario yielded more promising results. The best-performing metrics demonstrated more promising results for distinguishing the higher-quality point cloud. Nevertheless, there remains room for improvement in predicting the statistical significance of quality differences between point clouds.

As part of our commitment to advancing point cloud quality assessment, we are making our dataset publicly available. This dataset includes mean and individual opinion scores, along with scripts for metric evaluation in various scenarios. It also comprises all point clouds and their associated video renderings, forming what we call the BASICS dataset. We believe that the release of the BASICS dataset will contribute significantly to the improvement of existing point cloud quality metrics, the development of more robust ones, and the resolution of the challenges highlighted in this study.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision) and from Science Foundation Ireland (SFI) under the Grant Number 15/RP/27760.

REFERENCES

[1] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, "High-quality streamable free-viewpoint video," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–13, 2015.

[2] T. Zhou, S. M. Hasheminasab, and A. Habib, "Tightly-coupled camera/lidar integration for point cloud generation from gnss/ins-assisted uav mapping systems," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 336–356, 2021.

[3] O. Schreer, I. Feldmann, S. Renault, M. Zepp, M. Worchel, P. Eisert, and P. Kauff, "Capture and 3d video processing of volumetric video," in *IEEE International conference on image processing (ICIP)*. IEEE, 2019, pp. 4310–4314.

[4] X. Roynard, J.-E. Deschaud, and F. Goulette, "Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification," *The International Journal of Robotics Research*, vol. 37, no. 6, pp. 545–557, 2018. [Online]. Available: https://doi.org/10.1177/0278364918767506

[5] T. Rosnell and E. Honkavaara, "Point cloud generation from aerial image data acquired by a quadrocopter type micro unmanned aerial vehicle and a digital still camera," *Sensors*, vol. 12, no. 1, pp. 453–480, 2012.

[6] E. Zerman, C. Ozcinar, P. Gao, and A. Smolic, "Textured mesh vs coloured point cloud: A subjective study for volumetric video compression," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[7] S. Perry, H. P. Cong, L. A. da Silva Cruz, J. Prazeres, M. Pereira, A. Pinheiro, E. Dumic, E. Alexiou, and T. Ebrahimi, "Quality evaluation of static point clouds encoded using mpeg codecs," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3428–3432.

[8] L. A. da Silva Cruz, E. Dumić, E. Alexiou, J. Prazeres, R. Duarte, M. Pereira, A. Pinheiro, and T. Ebrahimi, "Point cloud quality evaluation: Towards a definition for test conditions," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–6.

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

[10] H. Su, Z. Duanmu, W. Liu, Q. Liu, and Z. Wang, "Perceptual quality assessment of 3d point clouds," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3182–3186.

[11] L. Hua, M. Yu, G. yi Jiang, Z. He, and Y. Lin, "Vqa-cpc: a novel visual quality assessment metric of color point clouds," in *SPIE/COS Photonics Asia*, 2020.

[12] X. Wu, Y. Zhang, C. Fan, J. Hou, and S. Kwong, "Siat-pcqd: Subjective point cloud quality database with 6dof head-mounted display," 2021. [Online]. Available: https://dx.doi.org/10.21227/ad8d-7r28

[13] S. Subramanyam, J. Li, I. Viola, and P. Cesar, "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)*, 2020, pp. 127–136.

[14] E. Alexiou, N. Yang, and T. Ebrahimi, "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)*, 2020, pp. 1–6.

[15] D. B. Graziosi, O. Nakagami, S. Kuma, A. Zaghetto, T. Suzuki, and A. Tabatabai, "An overview of ongoing point cloud compression standardization activities: video-based (V-PCC) and geometry-based (G-PCC)," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.

[16] M. Quach, G. Valenzise, and F. Dufaux, "Improved deep point cloud geometry compression," *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219708225

[17] "Prolific," https://www.prolific.com/, accessed: Nov. 2022. [Online].

[18] E. Alexiou, Y. Nehmé, E. Zerman, I. Viola, G. Lavoué, A. Ak, A. Smolic, P. Le Callet, and P. Cesar, "Chapter 18 - subjective and objective quality assessment for volumetric video," in *Immersive Video Technologies*, G. Valenzise, M. Alain, E. Zerman, and C. Ozcinar, Eds. Academic Press, 2023, pp. 501–552. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323917551000249

[19] I. Viola, S. Subramanyam, J. Li, and P. Cesar, "On the impact of vr assessment on the quality of experience of highly realistic digital humans: A volumetric video case study," *Quality and User Experience*, vol. 7, 05 2022.

[20] Y. Nehmé, P. L. Callet, F. Dupont, J.-P. Farrugia, and G. Lavoué, "Exploring crowdsourcing for subjective quality assessment of 3d graphics," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–6.

[21] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT.500-14, 2019.

[22] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, p. 2478–2491, dec 2012. [Online]. Available: https://doi.org/10.1111/j.1467-8659.2012.03188.x

[23] E. Zerman, V. Hulusic, G. Valenzise, R. Mantiuk, and F. Dufaux, "The relation between MOS and pairwise comparisons and the importance of cross-content comparisons," in *Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2018)*, Burlingame, United States, Jan. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01654133

[24] Y. Nehmé, J.-P. Farrugia, F. Dupont, P. LeCallet, and G. Lavoué, "Comparison of subjective methods, with and without explicit reference, for quality assessment of 3d graphics," in *ACM Symposium on Applied Perception 2019*, ser. SAP '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3343036.3352493

[25] A. Goswami, A. Ak, W. Hauser, P. L. Callet, and F. Dufaux, "Reliability of crowdsourcing for subjective quality evaluation of tone mapping operators," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–6.

[26] S. Palan and C. Schitter, "Prolific.ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[27] ITU-R, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," ITU-R Recommendation Recommendation P.913, 2021.

[28] J. Zhu, A. Ak, P. Le Callet, S. Sethuraman, and K. Rahul, "ZREC : robust recovery of mean and percentile opinion scores," Mar. 2023, working paper or preprint. [Online]. Available: https://hal.science/hal-04017583

[29] D. Moritz and D. Fisher, "Visualizing a million time series with the density line chart," *ArXiv*, vol. abs/1808.06019, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52052189

[30] G. Meynet, Y. Nehmé, J. Digne, and G. Lavoué, "Pcqm: A full-reference quality metric for colored 3d point clouds," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[31] A. Ak, E. Zerman, S. Ling, P. L. Callet, and A. Smolic, "The effect of temporal sub-sampling on the accuracy of volumetric video quality assessment," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

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

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

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

[35] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[36] A. Ak, A. Pastor, and P. Le Callet, "From just noticeable differences to image quality," in *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, ser. QoEVMA '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 23–28. [Online]. Available: https://doi.org/10.1145/3552469.3555712

[37] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR synthesized image quality assessment based on morphological wavelets," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–6.

[38] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR-synthesized image quality assessment based on morphological multi-scale approach," *EURASIP Journal on Image and Video Processing*, vol. 2017, 07 2016.

[39] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *Trans. Multi.*, vol. 13, no. 5, p. 935–949, oct 2011. [Online]. Available: https://doi.org/10.1109/TMM.2011.2152382

[40] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric, 2016," *Dostupno na: http://techblog. netflix. com/2016/06/toward-practical-perceptual-video. html [accessed on 22 Aug 2022]*, 2016.

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

[42] R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney, "Fovvideovdp: A visible difference predictor for wide field-of-view video," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459831

[43] R. Mekuria, Z. Li, C. Tulvan, and P. Chou, "Evaluation criteria for PCC (Point Cloud Compression)," ISO/IEC JTC 1/SC29/WG11 Doc. N16332, 2016.

[44] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, "Geometric distortion metrics for point cloud compression," in *IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 3460–3464.

[45] E. Alexiou and T. Ebrahimi, "Point cloud quality assessment metric based on angular similarity," in *International Conference on Multimedia & Expo (ICME)*, 2018.

[46] ——, "Towards a point cloud structural similarity metric," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020, pp. 1–6.

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

[48] ITU-R, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," ITU-R Recommendation P.1401, 2020.

[49] J. W. Tukey, "Comparing individual means in the analysis of variance." *Biometrics*, vol. 5 2, pp. 99–114, 1949.