A toolkit to benchmark point cloud quality metrics with multi-track evaluation criteria

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

▶ **To cite this version:**

Ali Ak, Emin Zerman, Maurice Quach, Aladine Chetouani, Giuseppe Valenzise, et al.. A toolkit to benchmark point cloud quality metrics with multi-track evaluation criteria. 2024 IEEE International Conference on Image Processing (ICIP 2024), IEEE, Oct 2024, ABU DHABI, United Arab Emirates. hal-04615285v2

HAL Id: hal-04615285
https://hal.science/hal-04615285v2

Submitted on 11 Feb 2025

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.

A TOOLKIT TO BENCHMARK POINT CLOUD QUALITY METRICS WITH MULTI-TRACK EVALUATION CRITERIA

Ali Ak[*], *Emin Zerman*[‡], *Maurice Quach*[§], *Aladine Chetouani*[¶], *Giuseppe Valenzise*[§], *Patrick Le Callet*[*†]

[*]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[†]Institut Universitaire de France (IUF)
[‡]Department of Computer and Electrical Engineering, Mid Sweden University, Sundsvall, Sweden
[§]Université Paris-Saclay, CNRS, CentraleSupélec, L2S (UMR 8506), Gif-sur-Yvette, France
[¶]Université d'Orléans, Orléans, France

ABSTRACT

Point clouds (PCs) gained popularity as a representation for 3D objects and scenes and are widely used in numerous applications in augmented and virtual reality domains. Concurrently, quality assessment of PCs became even more relevant to improve various aspects of these imaging pipelines. To stimulate further growth and interest in point cloud quality assessment (PCQA), we created a large-scale PCQA dataset (called "BASICS") which provides the research community with a relevant and challenging dataset to develop reliable objective quality metrics, and we organized the PCVQA grand challenge at ICIP 2023. In this paper, we provide a track-based evaluation methodology for benchmarking visual quality metrics, mirroring the PCVQA grand challenge evaluation scenarios designed to mimic real-life applications. Furthermore, we provide a state-of-the-art benchmark for the point cloud quality metrics. The track-based benchmarking approach shows that there is room for improvement in certain research directions, drawing attention to open problems in the PCQA domain.

***Index Terms*—** point cloud, quality assessment, quality of experience, track-based evaluation, benchmark

1. INTRODUCTION

Immersive multimedia technologies have developed significantly over the last decade thanks to improvements in devices and methodologies to capture, process, and display such content [1]. Fueled by recent developments, point clouds (PCs) gained popularity as a representation of immersive media. PCs enable representing the geometry and color information of 3-dimensional objects, up to millions of points. In addition, many other attributes can be stored for each point, further increasing the amount of data that needs to be stored and transmitted.

Due to the high dimensionality of the PCs, there is an inevitable need for efficient compression algorithms where the efficiency is measured with perceptual quality evaluation. Although the ideal way to measure the perceptual quality is through humans in the loop (*i.e. subjective quality evaluation*), it is time-consuming and expensive [2]. Therefore, subjective PC quality datasets are mainly used to develop PC objective Quality Assessment (PCQA) metrics that predict the perceptual quality [3, 4, 5].

This work was partially supported by the Knowledge Foundation, Sweden, with grant number 2019-0251.

Objective quality metrics are often developed for certain tasks in mind. For example, some metrics measure the aesthetic quality of an image, and others might focus on fidelity (*e.g., how much an image is distorted after compression*). Since these two metrics target clearly different tasks, the way to evaluate their performance must be different. The choice of evaluation data, tasks, and figures of merit impacts the development of more accurate and ecologically valid objective quality metrics.

With the aim of promoting research in the PCQA domain by providing the community with a large-scale dataset and ecologically valid use cases to develop upon, ICIP23 PCVQA Grand Challenge was organized[1]. The comprehensive analyses of the collected metric submissions over the BASICS dataset [4] and the results of these analyses make significant contributions to the PCQA domain and Quality of Experience (QoE) domain in general.

The ICIP23 PCVQA Grand Challenge utilized a track-based evaluation approach, which is not unheard of in grand challenges [6, 7]. Nevertheless, in this case, the track-based approach brought a multifaceted evaluation approach that can be converted into a toolkit for the use of the scientific community. This paper aims to achieve exactly this particular goal by providing a tool for easier benchmarking of visual quality metrics. The contributions of this work can be summarized as follows:

- A detailed benchmark of 10 PCQA metrics on three unique evaluation dimensions, shedding light on the importance of multi-criteria evaluation.
- An open-source Python toolkit to apply the same set of detailed analyses with ease. The toolkit is representation-agnostic, and it can easily be applied to other visual quality assessment problems, such as images, video, light fields, etc.
- Highlighting the open questions and rooms for improvement in the PCQA domain.

2. BACKGROUND: ICIP2023 CHALLENGE DESIGN

This section summarizes the ICIP 2023 PCVQA Grand Challenge, as this challenge created the foundations of the proposed evaluation methodology.

The PCVQA Grand Challenge consisted of 2 stages: the development stage and the test stage. During the development stage, participants were given access to the PCs and mean opinion scores

[1]https://sites.google.com/view/icip2023-pcvqa-grand-challenge/

Table 1. Characterization of the five tracks used in the ICIP 2023 PCVQA Grand Challenge.

	Track-1	Track-2	Track-3	Track-4	Track-5
Comparison range	Inter&Intra - SRC	Inter&Intra - SRC	Inter&Intra - SRC	Inter&Intra - SRC	Intra-SRC
Quality range	Broad Quality	Broad Quality	High Quality	High Quality	Broad Quality
Reference availability	Full Reference	No Reference	Full Reference	No-Reference	Full-Reference

(MOS) in the training set and the PCs in the validation set. Participants used CodaLab[2] to submit their predictions on the validation set. In the test phase, participants submitted their model in a docker[3], and each model was evaluated with the same system on the secret test set.

2.1. Dataset

The BASICS [4] dataset[4] was used for the challenge. It consists of 75 unique PCs (SRCs) from three semantic categories. Each PC was compressed with GPCC, VPCC, and GeoCNN [8] (a learning-based PC compression algorithm) at various compression levels (20 Processed Point Cloud (PPC) per SRC), resulting in 1500 PCs. On Prolific[5], a subjective experiment was conducted with the video renderings of the PCs to collect subjective opinion scores on the PC quality. We invite interested readers to refer to the related publication [4] for more information.

The dataset was split into training, validation, and test parts to be used in the development and test phases. 45 SRCs (900 PPCs) were used for training, 15 (300 PPCs) for validation, and 15 (300 PPCs) for the test.

2.2. Tracks

PCVQA Grand Challenge consists of 5 tracks, summarized through Table 1. The tracks are designed around three dimensions that allow for mimicking unique use cases. The traditional and widely applied evaluation scenario corresponds to Tracks 1 and 2 in our challenge. For all tracks, we provided the same training set to the participants. In all tracks, all metrics are evaluated based on Reference Availability (i.e. Full-Reference (FR) or No-Reference (NR)). Participants were free to submit different metrics to different tracks.

The first dimension is "comparison range", which is defined by limiting the evaluations to inter- or intra-SRC comparisons. Evaluating metric performances for only intra-SRC comparisons (i.e., Track 5) allows us to determine how well the metrics are at discriminating the quality difference between the PPCs derived from the same SRC. This evaluation criterion is valuable for use cases such as fine-tuning compression and enhancement algorithms, training machine learning models for end-to-end applications, and any other use case where the fidelity of the output is the primary concern over aesthetics. Evaluating metric performances only on intra-SRC comparisons is achieved by relying on Krasula's method [9]. Traditional correlation measures (such as Spearman's Rank Order Correlation (SROCC) and Pearson's Linear Correlation Coefficients (PLCC) are not suitable for this evaluation scenario.

The second dimension is the "quality range". The broad quality range covers the whole MOS range of $[1, 5]$ whereas the high quality range is defined as $[3.5, 5]$. Objective quality metrics that show high accuracy in the broad range may not necessarily perform the same in the high quality range [10]. Therefore high quality range evaluation is crucial for applications that aim to deliver top-tier content, such as high quality streaming and digital twins. To this end, we conducted an analysis to assess the accuracy of quality metrics specifically on the high quality part of the dataset, where the MOS is greater than or equal to 3.5.

Finally, the final dimension is the reference availability. If a metric accesses the reference information for evaluation, it is classified as FR. Otherwise, it is NR. We used separate tracks based on reference availability to keep the evaluation fair to NR metrics as it is a more complicated task.

Table 2. Evaluation criteria used for each track

	Track-1	Track-2	Track-3	Track-4	Track-5
SROCC	✓	✓	✓	✓	-
PLCC	✓	✓	✓	✓	-
D/S AUC	✓	✓	✓	✓	✓
B/W CC	✓	✓	✓	✓	✓
Runtime	✓	✓	✓	✓	✓

2.3. Evaluation Criteria

Five different criteria were used to evaluate the performance of the submissions. Two correlation measures (SROCC, PLCC), "Different vs Similar" and "Better vs Worse CC" from the Krasula's method, and runtime complexity. No fitting function was applied prior to evaluation. Prior to evaluation, subjective scores were processed with the state of the art MOS Recovery algorithm ZREC [11] to remove bias and inconsistencies associated with individual opinions. A brief explanation of each evaluation criteria is given below.

Correlation Measures: PLCC measures the prediction accuracy of the objective metrics and SROCC measures the prediction monotonicity [12]. For both correlation coefficients, the values are in the range $[0, 1]$, and higher values indicate a better correlation.

Krasula's method [9]: For the "Different vs Similar" analysis, pairs of PCs are categorized into two groups as pairs with (*i.e., different*) and without (*i.e., similar*) statistically significant differences. For a given pair of PPC, the Tukey's honest significance difference test [13] is used to measure the statistical significance. We assume that the absolute differences in metric predictions for "different" pairs should be larger than the "similar" pairs. Area Under the ROC Curve (AUC) of the Receiving Operating Characteristics (ROC) of the metric score differences between the two categories is used to quantify the metric performance. It is denoted as "D/S AUC" and its values are in the range $[0, 1]$ where higher values indicate a better performance.

In the "Better vs Worse" analysis, different pairs from the D/S analysis are used. The goal is to measure the metrics performance on how well they distinguish the better PC in pairs with statistically significant difference. Metric performances then can be expressed as

[2]https://codalab.lisn.upsaclay.fr

[3]https://www.docker.com/

[4]BASICS Dataset Link: https://zenodo.org/doi/10.5281/zenodo.8324545

[5]https://www.prolific.com

the correct classification percentage. It is denoted as "B/W CC" and its values are in the range $[0, 1]$ where higher values indicate better performance.

Runtime Complexity: Runtime complexity was assessed in terms of milliseconds required to run the metric on a PC on average. For a fair assessment of the runtime complexities of the metrics, we used the same system configurations. No additional process was run in parallel and each metric was evaluated individually. Lower runtimes are more desirable.

Table 2 shows which criteria are used in each track. Note that runtime complexity is only used in the test phase. In the test phase, the models were ranked based on the available criteria for each track. Similar to Borda count [14], the models with ranking $[1, 2, 3, 4, 5]$ will receive $[4, 3, 2, 1, 0]$ points respectively for each criteria. Then, for each track, the participants were ranked based on the collected points.

3. TRACK-BASED BENCHMARKING TOOLKIT

In this section, we introduce the track-based benchmarking tool, which evaluates the selected metrics in three main tracks, following the example of ICIP23 PCVQA Grand Challenge. The benchmarking tool is made publicly online for researchers to use[6].

3.1. Preprocessing for Subjective Quality Data

Both "Different vs Similar" and "Better vs Worse" measures of Krasula's method [9] require statistical significance analysis for the subjective quality data to be able to identify whether the selected pair of stimuli are statistically significantly different from one another (i.e., different) or not (i.e., similar).

The statistical significance of the subjective scores can be found in two different ways. The first option is to employ a one-way analysis of variance (ANOVA) to find out the variance and use Tukey's honestly significant difference (HSD) criterion to account for the multiple comparison bias [22]. This approach works only if the individual subjective opinion scores are available. Alternatively, the z-scores can be calculated from the MOS and standard deviation values, followed by Tukey's HSD.

After the preprocessing is done, the objective visual quality metric scores and the subjective quality scores (along with significance information) can be passed onto the Python toolkit for track-based benchmarking.

3.2. Evaluation over a Specific Quality Range

The first track is evaluation over a specified quality range, which can be the whole quality range (e.g., [1, 5] or [0, 100] depending on the initial quality scale) or a specific quality range (e.g., [3.5, 5] or [0, 20], etc.), for example, high quality range ([3.5, 5]) as it is done in the PCVQA Grand Challange.

This track does not have any other restrictions when it comes to reference availability or comparison range. So, it corresponds to the Track-1, Track-2, Track-3, and Track-4 of the ICIP23 PCVQA Grand Challenge.

3.3. Codec-Specific Evaluation

The second track is codec-specific evaluation, which mimics the point of view of codec developers. Since the developers mainly focus on evaluating their own codec after algorithmic changes are made, the codec is not evaluated as part of a bigger dataset. Instead, a metric (or set of metrics) was run particularly for the codec in question, disregarding other processing methods (e.g., capture, compression, transmission, and display processing which creates artifacts).

3.4. Intra-SRC Evaluation

The third track is intra-SRC evaluation (that is evaluation within the same visual source content). As also mentioned above, this track provides insight into the metrics on how well the metrics are at discriminating the quality difference between the PPCs derived from the same SRC. This level of scrutiny might be really important in applications such as high-end security applications, enhancement algorithms, and identifying the sources of errors while developing visual processing methods. This track corresponds to the Track-5 of the ICIP23 PCVQA Grand Challenge.

3.5. Advantages and Limitations

The proposed track-based benchmarking toolkit is advantageous in getting more insight into metric performances by exposing the metric in question to different use cases and different challenges. Rather than only relying on correlation values, the scientific community has been trying to find new methods that could provide more insight into visual quality metric performance. This includes converting the correlation problem into a classification problem [23], generating discriminability measures for objective quality metrics ($\Delta VQM, \tau_{0.05}$) [23, 22], and finding the metric performance considering the classification problem [22, 9]. The proposed track-based evaluation method can point out the deficiencies of a metric, which can consolidate research efforts to the unsolved parts of the broader QA problem.

Limitations can be seemingly reduced performance for some metrics that are not trying to address the QA problem in all the tracks. That is, metrics developed for specific purposes might come out as underachieving in other tracks, which could be misleading.

Despite the limitations, the proposed track-based evaluation approach combines the strengths of different approaches into one toolkit. In the following sections, we showcase a benchmark for point cloud quality metrics which shows how the results can be analysed and validates the proposed toolkit simultaneously.

4. OVERVIEW OF QUALITY METRICS

To demonstrate the benchmarking tool, we relied on 10 metrics (5 FR and 5 NR) including the 2 top-performing FR metrics from the BASICS [4] dataset and the 3 FR and 5 NR metrics from the test phase of the ICIP23 PCVQA Grand Challenge. In this section, we will briefly introduce these metrics before presenting the results.

4.1. FR Metrics

RWatanabe-FR [17] is a point-based metric that relies on geometry and color features extracted from the PC and its graph representation. In addition, it penalizes high differences in the number of points between the reference and distorted PCs. Geometry features consist of point2point [24] and point2plane[25] features while color features are based on the difference between global and local color variation on the graph. A Support Vector Regression (SVR) algorithm is used to quantify the distortions.

[6]The track-based benchmarking toolkit is freely and publicly available at: https://github.com/kyillene/MTB-PCQA

Table 3. Metric performances in terms of SROCC and PLCC over the test set of BASICS [4] dataset. Metric performances over broad and high quality ranges as well as codec-specific performances are given and indicated in each column. Metrics are categorized as FR and NR metrics, and each category is ordered based on the broad quality range rankings.

	Broad Quality		High Quality		GPCC-Predlift		GPCC-Raht		VPCC		GeoCNN	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
ZZhang-NR [15]	0.8817	0.9088	0.6458	0.6237	0.8691	0.9587	0.8847	0.9547	0.8265	0.8169	0.8114	0.7988
QZhou-NR [16]	0.7932	0.8062	0.5536	0.4005	0.8789	0.9249	0.8464	0.9080	0.7349	0.6538	0.6656	0.5730
RWatanabe-NR [17]	0.7637	0.7994	0.4563	0.4275	0.8563	0.8846	0.8223	0.8607	0.5786	0.5666	0.7427	0.8194
OMessai-NR [18]	0.5492	0.5923	0.2983	0.1692	0.7331	0.7724	0.6922	0.7451	0.2282	0.1759	0.3703	0.2581
YZhang-NR	0.3899	0.5154	0.0888	0.0993	0.6360	0.7167	0.5792	0.6714	0.1625	0.1049	0.1482	0.3065
RWatanabe-FR [17]	0.8726	0.9169	0.5453	0.5073	0.8912	0.9732	0.9037	0.9733	0.8745	0.8877	0.6742	0.6353
XZhou-FR [19]	0.8717	0.9092	0.5983	0.4618	0.8916	0.9746	0.9185	0.9712	0.8095	0.7855	0.6821	0.7794
ZZhang-FR [15]	0.8725	0.8974	0.6421	0.6104	0.8807	0.9561	0.8834	0.9497	0.7931	0.7812	0.8142	0.7775
PCQM [20]	0.7139	0.7615	0.2163	0.2434	0.8036	0.8692	0.8476	0.8660	0.6732	0.5871	0.1219	0.4283
PointSSIM [21]	0.6493	0.7156	0.2515	0.3171	0.7502	0.8566	0.8117	0.8584	0.4965	0.4185	0.5775	0.6199

XZhou-FR [19] (also called PointPCA+) uses Principal Component Analysis (PCA) over the extracted features. 16 geometry and 6 color features were extracted. Recursive Feature Elimination (RFE) algorithm is used to identify the most relevant set of features. Similarly to its predecessor PointPCA, a total quality score is obtained via learning-based fusion of individual predictions from geometry and texture descriptors.

ZZhang-FR [15] is a projection-based metric that relies on a cube-like projection and extracts features from the projected views via popular vision backbones. The similarity between the feature maps of reference and distorted projections is then used to estimate the quality of the distorted PC.

PCQM [20] is a point-based metric which uses several geometry and color features with a simple linear model mapping the feature space to perceptual quality scores. It was shown to be performing relatively well in the BASICS [4] benchmark.

PointSSIM [21] provides structural similarity scores for a given PC in comparison to its pristine reference. Structural similarity scores are obtained per attribute. The feature maps are computed by statistical dispersion estimators.

4.2. NR Metrics

ZZhang-NR [15] is another projection-based metric similar to the authors' FR implementation. In ZZhang-NR, features are extracted only from the distorted PC projections and inputted to the fully connected layers.

QZhou-NR [16] (also called BPQA) leverages the green learning paradigm [26]. It consists of three modules. The first module calculates the color saliency of points and is used in the 3D-to-2D patch projection module to generate multiple maps in module 2. These maps are then fed into the green learning module where channel-wise Saab transform is utilized.

RWatanabe-NR [17] utilizes geometry features based on PCA as well as the graph total variation features which captures both geometry and color features. Similar to their FR implementation, an SVR model is adopted to quantify the PC quality.

OMessai-NR [18] is a lightweight metric that utilizes vision transformers and deformable convolutional networks. Geometry and color information with frequency magnitude maps are inputted to a deep learning model named Deep CNN-ViT which consists of deformable convolution, depth-wise convolution, and vision transformer.

5. MULTI-TRACK EVALUATION RESULTS

In this section, we evaluate and discuss the performance of the metrics over different quality ranges, for different codecs, and in the intra-SRC comparison scenario.

5.1. Evaluation over Quality Range

Table 3 presents the metric performances in terms of SROCC and PLCC. The first two columns present the results in the broad quality range while the third and fourth columns present the correlations in high quality range. In broad quality range, we observe relatively high performances from the FR metrics and few of the NR metrics. However, we cannot obtain the same accuracy in metric predictions in the high quality range.

Moreover, we can visually inspect the difference in correlations in Figure 1. Pink (●) points represent the high quality range stimuli while the rest of the stimuli are represented with blue (●) color. When looked at in isolation, it is evident that metric predictions are far from ideal in the high quality range. Metrics often cannot distinguish the perceptual difference between high quality stimuli. For example, RWatanabe-FR rates most of the stimuli close to 4.5 with few exceptions despite the stimuli span the range $[3.5, 5]$. As a result, we observe a low performance in the high quality range despite the their superior performance in the broad quality range.

This shows that there is still room for improvement in point cloud quality assessment for high quality content transmission (i.e., high quality, high bitrate) or offline applications that still need to be compressed but not transmitted (e.g., cultural heritage and education applications).

5.2. Codec-Specific Evaluation

The last 8 columns in Table 3 present correlation coefficients between the metric predictions and MOS for PCs that are compressed by each codec in the BASICS dataset.

Instead of basing the judgment only on the whole dataset and for broad quality range, codec-specific evaluation helps identify the hard to address codecs and the weaknesses of certain metrics, as it is done by many other benchmarking articles. For example, PointSSIM has quite acceptable results for GPCC-Predlift and GPCC-Raht codec. Nevertheless, PointSSIM's performance on VPCC shows that this metric might not be very dependable for VPCC codec. Similarly, PCQM's performance on the learning-based GeoCNN codec might

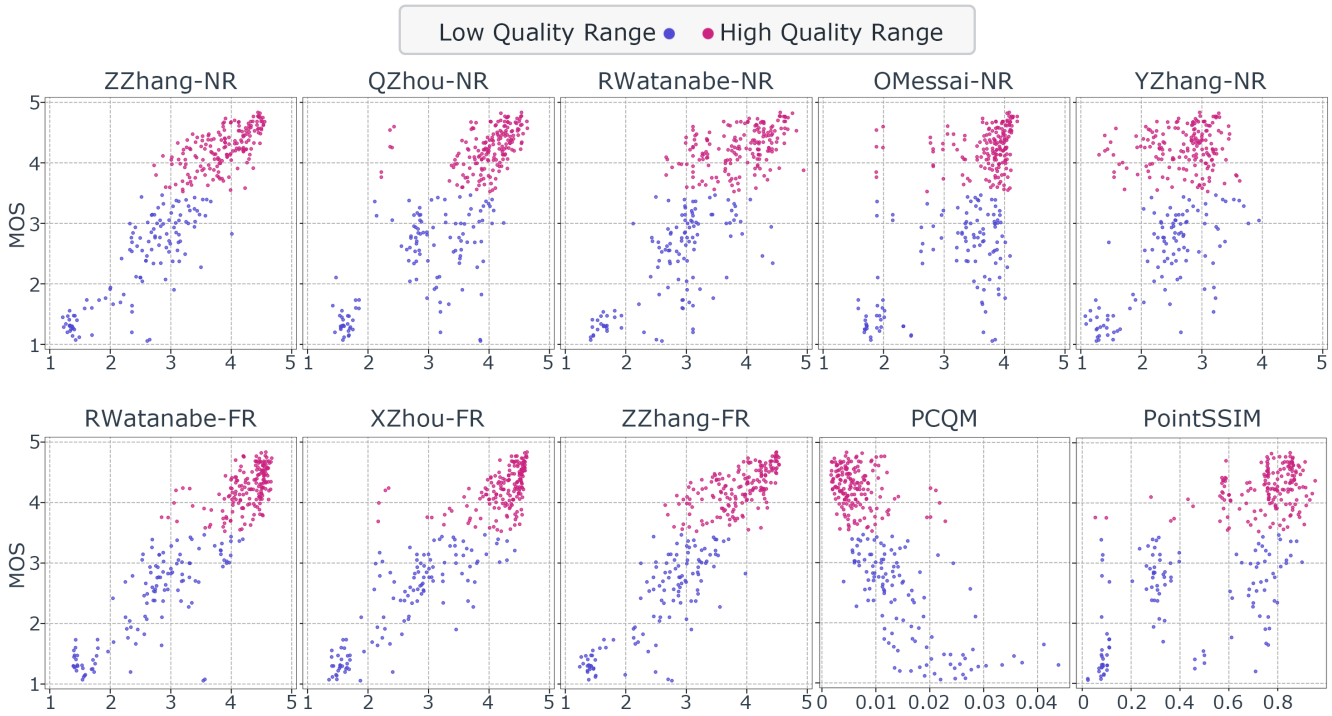

Fig. 1. Metric predictions plotted against the MOS. Metrics are ordered from left to right based on their rankings in the broad range quality. NR and FR metrics are displayed in the rows above and below, respectively. Data points are color-coded based on the quality range for better readability.

indicate that PCQM might be more relevant to use on G-PCC and V-PCC.

5.3. Intra-SRC Evaluation

Intra-SRC evaluation can be seen as a subset of the Inter-SRC evaluation and thus it can be considered a relatively simpler task. Despite its relative simplicity, metric performances in this track are not at desirable levels. Figure 2 presents the distribution of metric score differences for D/S (top row) and B/W (bottom row) tasks. In the D/S task, we expect higher absolute metric score differences for pairs with statistically significant differences (*i.e. Different pairs*) and lower for pairs without statistically significant differences (*i.e. Similar pairs*). As shown in Figure 2, all metrics show a significant overlap between the absolute metric score differences of different and similar pairs (i.e., top row), which is far from the ideal case (the leftmost plot). In the B/W task, results show that the evaluated metrics can identify the better PC in pairs with statistically significant differences.

6. DISCUSSION & CONCLUSION

The proposed track-based benchmarking methodology highlights the strengths and weaknesses of different visual quality assessment metrics. We define several tracks based on comparison range, quality range, and reference availability dimensions. By designing the evaluation scenario with the proposed dimensions, various use cases can be replicated. In addition, we propose to evaluate metric performances for specific distortion types, *i.e., codecs*. We provide a Python toolkit containing the preprocessing, evaluation, and data visualization scripts to run the same analyses on other datasets.

By conducting extensive analyses on the performance of objective quality metric predictions, we show that each dimension used for designing the tracks has a great influence on the metric performance. Metric performances drop significantly when tested on high quality range. Considering this together with the higher performance for broad quality, one can deduct that the tested metrics are designed for the broad quality range. For use cases aiming to deliver high quality content, such as digital twins, relying on the high quality range for evaluation is crucial.

Each metric relies on unique sets of features to predict the perceptual quality. Due to this fundamental difference, we may observe significant differences in metric performances with the codec-specific evaluations. For example, RWatanabe-FR [17] has the highest accuracy in broad quality range while performing poorly on assessing GeoCNN distortions. When the use case concerns only a specific set of distortions rather than a more generalized approach, evaluation criteria should be adjusted accordingly.

Intra-SRC evaluations show that the metrics trained on the broad quality range are not suitable for fidelity-based tasks such as compression pipeline optimization. Metrics struggle at determining statistically significant differences in pairs of stimuli.

To summarize, although it provides an easy mean to compare the metrics, it is not enough to benchmark metric performances only over the broad quality range. Metrics should be designed and evaluated for different use cases. The evaluation scenario should be designed according to the targeted use case by including related distortion types and selecting appropriate figures of merit. We show in detail how each dimension in the evaluation scenario impacts the

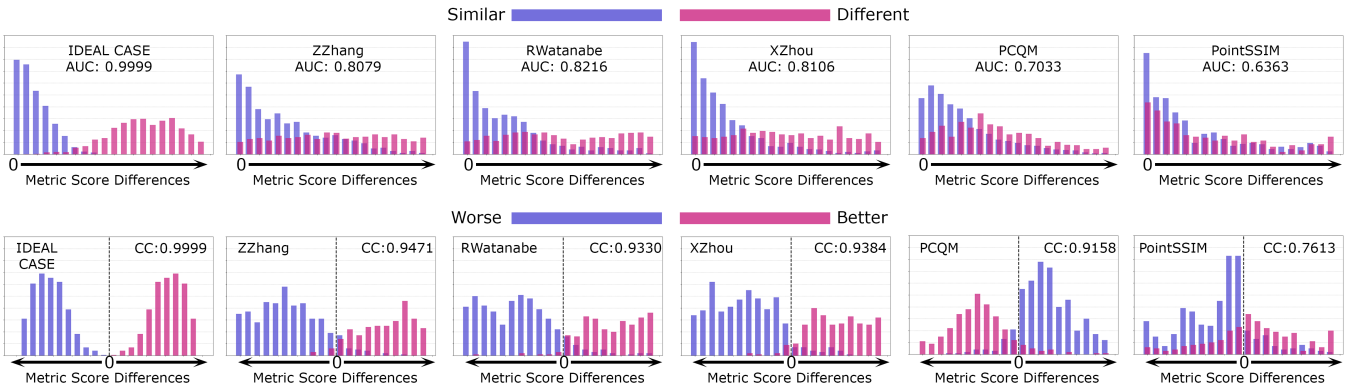

Fig. 2. Distribution of prediction differences for each FR metric shown as histograms. Absolute prediction differences are used for the Different vs Similar analysis at the top row. The bottom row shows the Better vs Worse analysis results. Ideal distributions are shown at the left for each analysis.

outcome of the benchmarking. Although the analyses were done on the PCQA task with the BASICS dataset, the proposed benchmark can be easily extended to other domains with the provided Python toolbox.

7. REFERENCES

[1] Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar, *Immersive Video Technologies*, Academic Press, 2022.

[2] Evangelos Alexiou, Yana Nehmé, Emin Zerman, Irene Viola, Guillaume Lavoué, Ali Ak, Aljosa Smolic, Patrick Le Callet, and Pablo Cesar, "Subjective and objective quality assessment for volumetric video," in *Immersive Video Technologies*, Giuseppe Valenzise, Martin Alain, Emin Zerman, and Cagri Ozcinar, Eds., chapter 18, pp. 501–552. Academic Press, 2023.

[3] Emin Zerman, Cagri Ozcinar, Pan Gao, and Aljosa Smolic, "Textured mesh vs coloured point cloud: A subjective study for volumetric video compression," in *Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020.

[4] Ali Ak, Emin Zerman, Maurice Quach, Aladine Chetouani, Aljosa Smolic, Giuseppe Valenzise, and Patrick Le Callet, "BASICS: Broad quality assessment of static point clouds in compression scenarios," *IEEE Transactions on Multimedia*, 2024.

[5] Xinju Wu, Yun Zhang, Chunling Fan, Junhui Hou, and Sam Kwong, "SIAT-PCQD: Subjective point cloud quality database with 6DoF head-mounted display," 2021.

[6] Fan Yu, Shiliang Zhang, Pengcheng Guo, Yihui Fu, Zhihao Du, Siqi Zheng, Weilong Huang, Lei Xie, Zheng-Hua Tan, DeLiang Wang, et al., "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9156–9160.

[7] He Wang, Pengcheng Guo, Yue Li, Ao Zhang, Jiayao Sun, Lei Xie, Wei Chen, Pan Zhou, Hui Bu, Xin Xu, et al., "Icmc-asr: The icassp 2024 in-car multi-channel automatic speech recognition challenge," *arXiv preprint arXiv:2401.03473*, 2024.

[8] Maurice Quach, Giuseppe Valenzise, and Frédéric Dufaux, "Improved deep point cloud geometry compression," in *IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020.

[9] Lukáš Krasula, Karel Fliegel, Patrick Le Callet, and Miloš 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.

[10] Deepthi Nandakumar, Yongjun Wu, Hai Wei, and Avisar Ten-Ami, "On the accuracy of video quality measurement techniques," in *IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019.

[11] Jingwen Zhu, Ali Ak, Patrick Le Callet, Sriram Sethuraman, and Kumar Rahul, "Zrec: Robust recovery of mean and percentile opinion scores," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2630–2634.

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

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

[14] JC de Borda, "Mémoire sur les élections au scrutin," *Histoire de l'Académie Royale des Sciences*, 1781.

[15] Zicheng Zhang, Yingjie Zhou, Wei Sun, Xiongkuo Min, and Guangtao Zhai, "Simple baselines for projection-based full-reference and no-reference point cloud quality assessment," 2023.

[16] Qingyang Zhou, Aolin Feng, Tsung-Shan Yang, Shan Liu, and C.-C. Jay Kuo, "BPQA: A blind point cloud quality assessment method," in *IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, 2023.

[17] Ryosuke Watanabe, Shashank N. Sridhara, Haoran Hong, Eduardo Pavez, and Antonio Ortega, "ICIP 2023 Challenge: Full-reference and non-reference point cloud quality assessment methods with support vector regression," in *IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, 2023, pp. 3654–3658.

[18] Oussama Messai, Abdelouahid Bentamou, Abbass Zein-Eddine, and Yann Gavet, "Activating frequency and VIT for 3D point cloud quality assessment without reference," in *IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, 2023, pp. 3636–3640.

[19] Xuemei Zhou, Evangelos Alexiou, Irene Viola, and Pablo Cesar, "PointPCA+: Extending pointpca objective quality assessment metric," in *IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, 2023.

[20] Gabriel Meynet, Yana Nehmé, Julie Digne, and Guillaume 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.

[21] Evangelos Alexiou and Touradj Ebrahimi, "Towards a point cloud structural similarity metric," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[22] Emin Zerman, Giuseppe Valenzise, and Frederic Dufaux, "An extensive performance evaluation of full-reference HDR image quality metrics," *Quality and User Experience*, vol. 2, pp. 1–16, 2017.

[23] ITU-T, "Method for specifying accuracy and cross-calibration of video quality metrics (VQM)," ITU-T Recommendation J.149, 2009.

[24] Rufael Mekuria, Zhu Li, Christian Tulvan, and Phil Chou, "Evaluation criteria for PCC (Point Cloud Compression)," ISO/IEC JTC 1/SC29/WG11 Doc. N16332, 2016.

[25] Dong Tian, Hideaki Ochimizu, Chen Feng, Robert Cohen, and Anthony Vetro, "Geometric distortion metrics for point cloud compression," in *IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 3460–3464.

[26] C.-C. Jay Kuo and Azad M. Madni, "Green learning: Introduction, examples and outlook," *Journal of Visual Communication and Image Representation*, vol. 90, pp. 103685, 2023.