Van Damme *et al.*
EURASIP Journal on Image and Video Processing (2024) 2024:42
https://doi.org/10.1186/s13640-024-00655-y

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Analyzing the applicability of psychometric QoE modeling for projection-based point cloud video quality assessment

Sam Van Damme[1*], Jeroen van der Hooft[1], Filip De Turck[1] and Maria Torres Vega[2]

*Correspondence:
Sam.VanDamme@UGent.be

[1] IDLab, Department of Information Technology (INTEC), Ghent University - imec, Ghent, Belgium
[2] eMedia Research Lab, Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium

Abstract

Point cloud video delivery will be an important part of future immersive multimedia. In it, objects represented as sets of points are embedded within a video which is streamed and displayed to remote users. This opens possibilities towards remote presence scenarios such as tele-conferencing, remote education and virtual training. Due to its infeasibly high bandwidth requirements, encoding is unavoidable. The introduced artifacts and network degradations can have an important but unpredictable impact on the end-user's Quality of Experience (QoE). Thus, real-time quality monitoring and prediction mechanisms are key to allow for fast countermeasures in case of QoE decrease. Since current state-of-the-art research is focusing on either continuous QoE monitoring of traditional video streaming services or objective delivery optimizations of point cloud content without any QoE validation, we believe this work brings a valuable contribution to current literature. Therefore, we present a no-reference (NR) QoE model, consisting of KMeans clustering and sigmoidal mapping, that works on video-level, group-of-pictures (GOP)-level and frame-level granularity. Results show the value of the sigmoidal mapping across all granularity levels. The clustering algorithm shows its value at the video-level and in the role of an outlier detector on the more fine-grained levels. Satisfying results are yet obtained with correlation values often going above 0.700 on GOP- and frame-level while maintaining root mean squared error (RMSE) below 10 on a 0–100 scale. In addition, a Command Line Interface (CLI) Video Metric Tool is presented that allows for easy and modular calculation of NR metrics on a given video.

Keywords: Point cloud video, Quality-of-Experience, No Reference, Near-continuous modeling, Video Metric Tool

1 Introduction

Point cloud video delivery will be one of the key aspects of immersive multimedia. To enable this, dense sets of six-dimensional points (three space coordinates + 3 color channels) need to be streamed over the network to present them to the remote end-user. As such, a plethora of possibilities open up towards remote presence scenarios such as more immersive tele-conferencing, remote education and virtual training. Due to the high bandwidth requirements of this content, in the order of Gbps, encoding is key. The

introduced encoding artifacts, together with any degradations present in the network, such as losses, delay or jitter, can have a severe impact on the end-user's quality of experience (QoE). As such, a real-time client-side QoE mechanism is essential to monitor quality and to allow for fast counteraction in case any severe QoE drops are detected.

Although subjective evaluations are still the most reliable way of evaluating QoE, they come with inherent costs, duration and complexity which are incompatible with a real-time system. Therefore, objective evaluation methods are needed. For the specific case of point cloud content, two different approaches can be distinguished: geometric and projection-based quality evaluation. Geometric approaches make an assessment of the distorted point cloud figure (or its mesh representation) by comparing it to its original, undistorted counterpart. As such, point-to-point or point-to-plane Peak Signal-to-Noise Ratio (PSNR)-based metrics, among others, can be applied to estimate the quality of the given object [1–4]. Projection-based evaluations make a quality assessment based on the current Field-of-View (FoV) rather than the figure as a whole. In other words, the three-dimensional figure is first projected to a two-dimensional plane after which traditional 2D video metrics can be applied for assessing quality. Note that projection-based metrics therefore pose an important advantage in comparison to geometric approaches w.r.t. establishing the (near) real-timeliness of quality evaluation, as existing 2D No Reference (NR)-metrics are already designed in computationally efficient ways. As such, their well-established and highly optimized pipelines for traditional video pose a unique opportunity to reuse their value for (near) real-time point cloud evaluation. However, research towards their accuracy and modeling with respect to subjective perception is still rather limited. This is important, however, as one cannot straightforwardly assume that existing metrics and models for natural video will also apply to specific types of content such as synthetic content or volumetric media. *Twitch*-alike gaming videos, for example, have shown to allow for much more straightforward modeling approaches than is the case for natural video content [5], which often rely on rather complex Machine Learning (ML) or Deep Learning (DL) modeling. Among these projection-based metrics, Full Reference (FR) metrics such as Netflix' Video Multimethod Assessment Fusion (VMAF) take human intervention out of the equation while still maintaining high accuracy in terms of correlation to subjective perception scores. This is the case for natural video content, but also for the specific case of point cloud video with correlation values to Mean Opinion Scores (MOS) often up to 0.9 [6–10]. As such VMAF is, among others, typically used as an objective FR benchmark and substitute for time consuming subjective MOS in the evaluation of projection-based point cloud quality [11–13]. As these metrics rely on a comparison with the original, undistorted content, however, they are infeasible in live-streaming scenarios as the reference content cannot be accessed without suffering from encoding and network degradation as well. As such, client-side NR metrics, which act solely upon the distorted stream to measure blur, noise, blockiness, etc. are relied on to fulfill the aforementioned task. To this end, this work presents a NR QoE model for projection-based point cloud video, consisting of KMeans clustering and sigmoidal mapping. In our previous work [13], we presented a first exploration and preliminary results on this matter. Encouraged by these, this paper presents the fine-tuned model and accompanying full-fetched analysis on multiple granularity levels (video, group-of-pictures (GOP), and frame-level). The latter is of utter importance to enable

QoE monitoring in a (near) real-time fashion while maintaining high accuracies towards subjective scores or well-established subjective benchmarks such as VMAF. In that way, the necessary adaptations and optimizations can be made in a fast manner to keep the end-user's perception and immersion to a maximum. As most traditional NR metrics have highly optimized implementations in order to evaluate on a per-frame base, they are extremely suited to fulfill this task, provided they can meet the necessary accuracy requirements as well. In addition, the necessary tools to easily calculate aforementioned metrics are provided in the form of a Command Line Interface (CLI) *Video Metric Tool*. The tool allows for customization on which metrics to include or exclude in the calculation and new metrics or alternative implementations can easily be plugged into the existing framework.

Therefore, our contributions are fivefold:

1. We present an NR QoE model, consisting of KMeans clustering and sigmoidal mapping, that works on video-level, GOP-level and frame-level granularity.
2. A full-fetched analysis of the dataset [13] is conducted, providing in-depth insights on multiple granularity levels (GOP- and frame-level analysis).
3. An analysis towards the applicability of multiple objective benchmarks is provided, with respect to their correlations and root mean squared error (RMSE) towards subjective scores.
4. Investigation of the extent to which a one-for-all QoE model is feasible and when per-video models are more appropriate.
5. Presentation of a video quality metric CLI tool that allows to calculate a set of NR metrics on a given video.

The remainder of this paper is organized as follows. Section 2 presents an overview of the most important literature in terms of geometric and projection-based point-cloud quality metrics as well as objective quality modeling and benchmarking. Section 3 discusses the followed methodology, followed by Sect. 4 in which the created *Video Metric Tool* is presented. Section 5 presents the dataset as well as the obtained results, which are further discussed in Sect. 6. Section 7, at last, lists the most prominent findings of this work and gives some pointers towards future research directions.

2 Background and related work

This section will briefly discuss research related to the work presented in this paper. First, an overview of geometric approaches is provided, followed by objective, projection-based point cloud quality metrics. Next, a brief discussion on existing mechanisms for fine-grained QoE monitoring and modeling is provided, as well as a thorough overview on applied benchmarks in literature. At last, the most important takeaways from existing literature are listed.

2.1 Geometry-based point cloud quality metrics

Alexiou et al. [14] created a geometric metric by calculating the average angular similarity between nearest neighboring points. Their results show Pearson Linear Correlation Coefficients (PLCCs) between 0.89 and 0.99 to subjective MOS depending on the type of

distortion being applied and whether Absolute Category Rating (ACR) or Double Stimulus Impairment Scale (DSIS) is being assessed.

In a second study [15], they adapted Structural Similarity Index Measure (SSIM) for use in point clouds based on geometry, normal vectors, curvature values, and colors. Similar to SSIM, features are extracted per neighborhood after applying dispersion statistics. The authors show that maximum PLCCs to MOS between 0.80 and 0.90 are achieved, depending on the considered dataset.

Diniz et al. [16] derived a Reduced Reference (RR) point cloud quality assessment model based on local patterns. In this model, each pixel is assigned a binary code by thresholding the difference in intensity with its surrounding pixels. The quality of the point cloud is then determined by the difference between the histograms of the original and the distorted content, mapping this distance to a predicted MOS using a third-order polynomial relationship. Results show that PLCCs to MOS varying between 0.67 and 0.88 can be achieved, depending on the considered content.

In a second study [17], they present an objective FR visual quality assessment metric for static point clouds, named *BitDance*, which uses color and geometry texture descriptors. The proposed method first extracts the statistics of color and geometry information of the reference and test point clouds. Then, it compares the color and geometry statistics in terms of distance and combines them in a logistic mapping to estimate the perceived quality. Their results show an average PLCC of 0.84 to subjective scores when averaged over multiple datasets.

Viola et al. [4] created a RR quality metric by extracting color statistics, creating histograms and correlograms from both the original and the distorted sequence. The distance between both is used to predict the subjective MOS by applying a curve-fitting approach.

Following up on this work, the authors created a second RR metric based on a weighted combination of feature differences in terms of geometry, luminance and normal [18]. Evaluating both metrics, PLCCs up to 0.90 for the subjective MOS are achieved for a single publicly available dataset.

Nehme et al. [19] proposed a FR metric for the quality assessment of 3D meshes, which works entirely on the mesh domain. The proposed metric integrates both geometry and color information, using statistics on curvature (e.g., contrast and structure) and color (e.g., chroma and hue comparison) on local neighborhoods corresponding to the original and the distorted content. While individual features result in PLCCs for the MOS between 0.30 and 0.70 only, the overall metric results in a PLCCs between 0.86 and 0.91, depending on the considered dataset.

Tian et al. [20] propose a geometric point-to-plane distance as a measure of geometric distortions on point cloud compression. To this end, the intrinsic resolution of the point clouds is proposed as a normalizer to convert the mean square errors to PSNR numbers. In addition, the perceived local planes are investigated at different scales of the point cloud. As such, they create a metric that is independent of the size of the point cloud and shows to better track the perceived quality than the point-to-point approach.

Javaheri et al. [21] present a geometric point cloud quality assessment metric based on a generalization of the Hausdorff distance. This generalization is realized by computing the distance over a subset of data after ranking all the values. Their results show that this

generalization leads to an improvement in correlation towards subjective scores, and even tends to outperform the MPEG adopted geometry quality metrics when decoded point clouds with different types of coding distortions are considered.

In a second study [2], they propose novel improved PSNR-based metrics by exploiting the intrinsic point cloud characteristics such as resolution and the rendering process that must occur before visualization. Their results show improvements of up to 32% in PLCC for certain cases.

In a third study [22], a scale-invariant point cloud geometric quality assessment metric is proposed as the correspondence between a point and a distribution of points based on the Mahalanobis distance. Their results show PLCC gains up to 31.9% and Spearman Rank Order Correlation Coefficient (SROCC) gains up to 22.8% with regard to existing MPEG metrics.

Liu et al. [23] propose a RR linear perceptual quality model for V-PCC encoding based on geometry and color quantization whose coefficients can easily be computed from two features extracted from the original point cloud. Their subjective quality test results show that the proposed model outperforms certain state-of-the-art FR objective measures in terms of both PLCC and SROCC with correlation values up to 0.91.

Chen et al. [24] propose a Layered Projection-based Point Cloud Quality Metric (LP-PCQM). The distorted point cloud and its original version are layered such that geometry and color features of layers can be extracted. The geometry feature is obtained using the projection-based method and the color features are extracted upon RGB by using the point-based method. Finally, the LP-PCQM is a weighted linear combination of an optimal subset of these pooled geometry and color features of layers. Their results show PLCCs between 0.71 and 0.90 depending on the dataset.

2.2 Projection-based point cloud quality metrics

In the last few years, a significant amount of research has been conducted on measuring the objective quality of *omnidirectional video* [25–27]. In the context of *point cloud video*, however, research is still at an early stage, with standards for evaluation procedures still to be agreed upon. Nevertheless, a decent amount of studies can be mentioned in the field of projection-based quality modeling for point-cloud video, which this work is focusing on.

Recently, several attempts have been made to tailor projection-based metrics to volumetric video. Alexiou et al. [28] research projection-based objective quality assessment of point cloud imaging by investigating the impact of the number of viewpoints employed to assess the visual quality of a content. This is done while discarding information that does not belong to the object under assessment, such as background color. In addition, they propose to assign weights to the projected views based on user interactivity information. Their results show that employing a larger number of projected views does not necessarily lead to better predictions of visual quality, while user interactivity information can improve the performance.

Yang et al. [7] presented a FR metric based on the projection of the point cloud on the six perpendicular planes of a cube. For each of the resulting images, features are extracted from the color and depth information in terms of texture similarity, Jensen–Shannon (JS) divergence, and edges. Combining feature values from all six planes, a

single quality index is then derived. Results show PLCCs with subjective MOS ranging from 0.66 to 0.97, depending on the considered content and the introduced encoding distortions. In a second study [29], *GraphSIM* is proposed: a metric to accurately and quantitatively predict the human perception of point clouds with superimposed geometry and color impairments. Based on the characteristics of the Human Visual System (HVS), local graphs are constructed around geometric keypoints which are used to compute three moments of color gradients. Afterwards, the similarity index is calculated by pooling the local graph significance across all color channels and averaging across all graphs. Their results show PLCCs between 0.89 and 0.98 to MOS depending on the dataset and the content type.

Torlig et al. [30] propose a framework for quality assessment of point clouds by means of rendering software that allows for real-time voxelization and projection of the 3D point clouds onto 2D planes, while allowing interaction between the user and the projected views. These projected images are then employed by two-dimensional objective quality metrics, in order to predict the perceptual quality of the displayed stimuli. Their benchmarking results, using subjective ratings that were obtained through experiments in two test laboratories, show high predictive power.

Yang et al. [31] propose a new metric called *Volu-FMAF* based on the results of a comprehensive user study to understand the effectiveness of popular perceptual quality metrics for volumetric video. It combines volumetric VMAF with viewpoint related features. To this end, a Support Vector Regression (SVR) with Radial Basis Function (RBF) kernel is trained with subjective DSIS MOS as the benchmark. In addition, they present a novel neural-based volumetric video streaming framework *RenderVolu* and design a distortion-aware rendered image super-resolution network, called *RenDA-Net*. Their results show a boost in perceptual quality of 171% to 190% while achieving a 108× speedup in encoding efficiency compared to state-of-the-art approaches.

Fan et al. [32] propose a novel DL NR volumetric video quality assessment method based on multi-view learning. This is realized by projecting the volumetric videos to 2D video sequences from various viewpoints. Next, a set of quality-aware features is extracted from the projected video sequences by means of a 3D-Convolutional Neural Network (CNN) backbone. Based on these, a regressor is designed that fuses the features from the multiple viewpoints and joins them into quality scores. The results show that their method outperforms state-of-the-art objective volumetric video quality assessment metrics on the V-SENSE VVDB2 database [33], with PLCCs and SROCCs up to 0.901 and 0.865, respectively, and RMSEs down to 7.927.

Van der hooft et al. [1] presented an objective and subjective quality evaluation of point cloud streaming for multiple scenarios in terms of bandwidth, rate adaptation, viewport prediction and user motion. Their results show high correlation with MOS for traditional video metrics such as PSNR, SSIM and Video Quality Metric (VQM). They further indicated that the subjective perception of point cloud video lays within a very small interval of the total range of the objective metrics, which might be a result of the inclusion of (too much) background during the quality metric calculation.

In our own, previous work [13], we communicated a first set of preliminary, video-level results on the clustering-based NR QoE assessment model for point cloud video being presented in this work. Using this approach, PLCCs up to 0.977 and RMSEs down

to 0.077 on a 0-to-1 scale are obtained at video-level. In the following section, this approach will be discussed in more detail and a full-fetched analysis with results on multiple granularity levels (GOP and video-level) will be presented.

2.3 Fine-grained QoE monitoring, modeling, and benchmarking

When it comes to (near) continuous, per-GOP or per-frame monitoring and modeling of QoE for point cloud content, very little research is available. On one hand, a lot of studies focus on optimizations of point cloud content delivery in terms of rate control [34–38], transmission [39], error concealment [40] or coding [41] rather than continuous monitoring. Here, the influence on end-user QoE is most of the times not taken into account and studies that do so often do not validate their QoE predictions with real subjective scores or accurate objective benchmarks such as VMAF [37, 38]. Research that does focus on (near) continuous QoE modeling, though, are mostly centered around traditional video streaming services [42–47] or on specific use-cases such as tele-conferencing [48], live broadcasting [49] or cloud gaming [50]. To the best of our knowledge, no works currently exist that apply this research on continuous QoE monitoring and modeling to the specific case of point cloud video delivery.

To enable such a system, it is of high importance to select an appropriate metric for benchmarking, such that models can be trained and evaluated on fine granularity levels where subjective scoring is infeasible. Preferably this is also a projection-based or traditional video-metric that can operate at frame-level to allow for real-time monitoring. Furthermore, the linearity, monotony and accuracy of the metric compared to subjective scores at a more coarse level should be taken into account, such that one can hypothesize these will hold at finer granularity. In this respect, Ak et al. [51] show VMAF to have the best performance along image-based metrics for static point clouds. PLCCs and SROCCs of 0.742 and 0.669 are obtained. These show to be outperformed, though, by geometry-based metrics such as p2plane-Mean Squared Error (MSE) and Point- Centered Quarter Method (PCQM), which result in correlations well-above 0.8. Unfortunately, no RMSE values were reported to further analyze accuracy. Furthermore, these metrics tend to show rather stable behavior alongside encoders, where image-based metrics tend to drop performance for V-PCC. On the downside, it is important to realize that geometry-based metrics require the availability of the full point cloud object in every frame of the video rather than a single projection. As such, these induce considerably, and maybe even infeasible, computational overhead when to be applied in a real-life practical implementation.

In the analysis of Lazzarotto et al. [52] on encoding distortions in static point clouds, Feature Similarity Indexing Method (FSIM) is identified as the best performing metric on average over all metric types. Respective PLCCs, SROCCs and RMSEs of 0.876, 0.790 and 0.566 are obtained. VMAF follows shortly with values of 0.862, 0.770 and 0.595. Furthermore, GraphSIM, MS-PointSSIM, PCQM, Information content Weighted SSIM (IW-SSIM) and Multi-Scale SSIM (MS-SSIM) are showing similar performance. Important to mention, however, is that most of these metrics tend to show reduced performance on V-PCC distortions compared to G-PCC and JPEG Pleno. Only FSIM, PCQM and MS-PointSSIM show rather stable in this respect.

Yang et al. [53] conducted a correlation analysis for static color meshes. Among the video-based metrics, VMAF shows the best average performance compared to PSNR and SSIM with average PLCC and SROCC of 0.70 and 0.51 and an RMSE of 1.0. Although some point-based metrics, such as PCQM-PSNR, show better performance on average, it is worth mentioning that the accuracy of these metrics is typically very sensitive to the difference in point density between the reference and the distorted object. Furthermore, the calculation of such benchmarks may induce a considerable computational overhead, as was stated earlier.

In a subsequent study [54], they confirm VMAF to be the best performing video-based metric in comparison with PSNR and SSIM, be it with much lower average PLCC and SROCC than reported before (0.48 and 0.53, respectively). This time, similar performance is recorded for PCQM-PSNR, while YUV-PSNR is showing better performance with respective average PLCC and SROCC of 0.59 and 0.65. However, the same remark with regard to computational complexity can be restated in this case. As one can notice, rather low correlations are reported in general, which the authors attribute to the challenging content of the SJTU-TMQA database being analyzed.

In a third study on dynamic meshes [10], they obtain rather high PLCCs and SROCCs for VMAF with respective values of 0.92 and 0.90. These are comparable with PSNR (0.93, 0.91), SSIM(0.96, 0.95), MS-SSIM (0.94, 0.92) and FSIM (0.95, 0.94). Furthermore, the point-based metric PCQM once again shows comparable performance with respective correlations of 0.94 and 0.93. Unfortunately, no RMSE values were reported to further investigate metric accuracy.

Wien et al. [9] presented a similar analysis for dynamic meshes. Here, VMAF also shows to be the best performing image-based metric in terms of PLCC, SROCC, Kendall Rank Correlation Coefficient (KRCC) and RMSE to DSIS MOS, when compared to PSNR, SSIM, MS-SSIM and VQM. VMAF reaches respective correlation values of 0.84,0.85 and 0.67 while limiting RMSE to 1.1. Only MS-SSIM is able to somewhat stay in the neighbourhood of VMAF with respective correlations of 0.79, 0.83 and 0.64 and an RMSE of 1.3. Furthermore it is also interesting to notice how both VMAF and MS-SSIM, and even straightforward PSNR, are able to easily outperform point-based metrics.

In our own work, at last, we previously calculated PLCCs of SSIM and VMAF to both ACR and DSIS MOS on a set of point cloud videos [8]. Here, it was shown that both indicate similar results, although VMAF shows somewhat better performance with respective average PLCCs of 0.88 and 0.93 compared to 0.80 and 0.85 for SSIM. In this work, this analysis will be broadened with SROCC and RMSE as additional evaluation metrics.

2.4 Conclusion

As can be noticed from this discussion, there is little to no consensus on which approach to use for objective quality estimation of volumetric media. Specific research towards projection-based NR metrics and modeling is still rather limited, especially for the specific case of point cloud video delivery. The presented approaches are highly varying in terms of input features and are not always as computationally friendly. In addition, neither of them are investigating whether similar accuracy can be reached by benchmarking

against objective FR metrics such as VMAF instead of subjective MOS. This is required, however, to be able to rule out the human factor in order to create a fully automated quality assessment system even when previously unseen content is entering the database (which would require a costly subjective study for every new sequence otherwise).

In this respect, the above discussion has revealed that in general, VMAF and FSIM tend to show best performance when image-based metrics are considered. SSIM-based metrics tend to show similar performance in terms of correlations, although they often lack in accuracy (i.e., RMSE) due to their narrow working range as discussed in our earlier work [8]. Caution is advised, though, as the performance of these image-based metrics tends to severely vary depending on the encoding artifacts being present. Therefore, it is required to carefully evaluate the dataset at hand to make a well-supported choice on the applied benchmark. Some geometry-based metrics, such as PCQM and its variations, tend to show similar or even somewhat improved performance compared to image-based metrics. Nevertheless, the gains show to be limited compared to the additional computational overhead they induce. Furthermore, their performance typically depends on the difference in point density between reference and distorted point cloud figures, such that their generalizability can be questioned. For this reason, the remainder of this work will focus on image-based metrics for benchmarking, which match the projection-based purpose of the presented methodology.

As such, we believe that this work addresses an important hiatus in the state-of-the-art, therefore providing a valuable contribution. First, we create and evaluate a predictive and objective NR quality assessment model for point cloud video streaming on a (near) continuous assessment level. This model is based on a straightforward white-box approach, i.e., a sigmoidal mapping of a weighted linear combination, that poses low computational requirements while maintaining satisfying accuracy. It will be accompanied by a thorough evaluation on available benchmarks in order to identify the most suitable one in terms of linearity, monotony and accuracy towards subjective scores.

As a side contribution, we present an easy-to-use CLI tool for the calculation of NR metrics on a given input video. The tool allows for customization on which metrics to include or exclude in the calculation and new metrics or alternative implementations can easily be plugged into the existing framework.

3 Methods

This section will describe the methodology followed in this work. First, it presents how the presented model fits into the broader architecture of QoE management for (point cloud) video streaming. Next, the set of metrics included in the model will be described. At last, the modeling approach will be discussed including its two major parts: clustering and psychometric modeling.

3.1 QoE management architecture

To clearly indicate how the proposed model fits into the broader QoE management paradigm, a brief and conceptual overview of the latter for the use-case of point-cloud video streaming is first provided. Figure 1 shows a typical (point cloud) video streaming architecture. A given set of point cloud figures is stored at a *database* on the server side. There, a *point cloud encoder* foresees multiple quality representations of the given

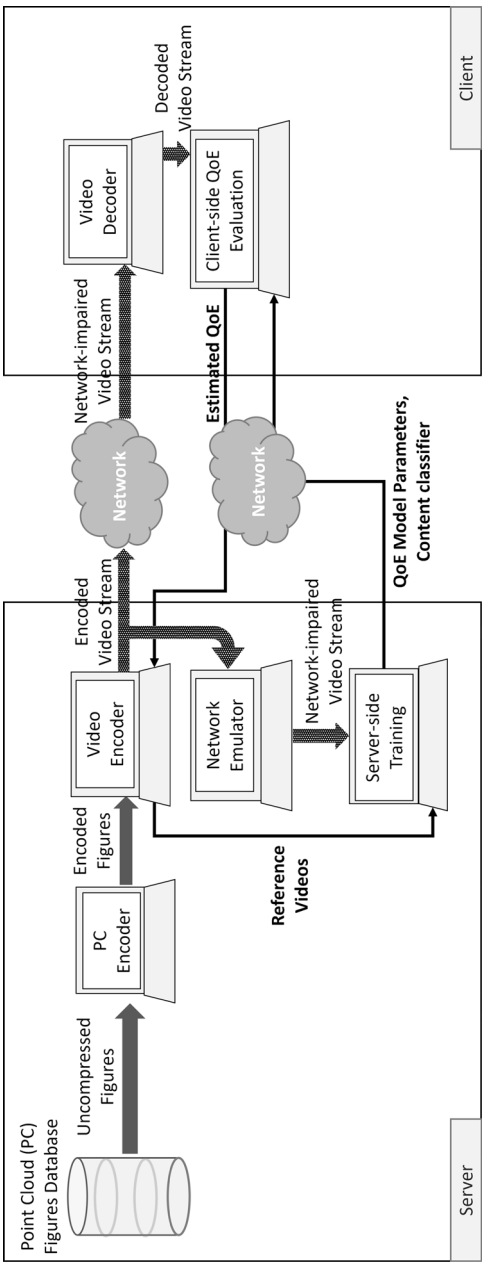


Fig. 1 The envisioned system architecture

figures. Based on the content, the current network conditions and the real-time estimated end-user QoE, *the video encoder* combines multiple figures at their appropriate quality representations into one scene, projects them on the 2D plane in a regular pixel-based format based on the current viewport, and encodes the whole into the video bitstream. This is done by tweaking the encoding parameters in terms of bitrate and bandwidth allocation scheme based on the client-side QoE estimation and the current network status. This can for example be established using a Reinforcement Learning (RL) approach where the QoE estimation is used as a reward to train an optimizer on how to react to certain changes in the network conditions to keep QoE to an optimum. The resulting encoded video stream is sent over the *network* to the client side, where it is decoded by the *video decoder*. This decoded video representation is used for *client-side QoE estimation*, of which the output is sent back over the network to the server-side video encoder for decision-making. As such, a *QoE management loop* [55] is constructed in which the client-side QoE assessments are fed back to the server for quality monitoring. Here, this feedback is used for further optimization of encoding and network control of streamed content, therefore positively affecting end-user QoE after which the loop can repeat itself. The contribution of this work is situated in the QoE modeling, as an accurate estimation of end-user QoE will lead the encoder to the most optimal decisions to maintain the ground-truth end-user QoE, therefore acting as the fuel of this quality management loop. The parameters for this client-side QoE estimation model, together with a content classifier (if relevant for the particular model), are retrieved from the *server-side training* module, which uses a *network emulator* to simulate network distortions at client side. Furthermore, the undistorted reference content is also included to be able to calculate a FR objective metric as a ground truth benchmark, such that model training can take place fully automated without any human intervention or assessment.

3.2 Quality metrics

To allow the QoE estimation model to work on a per-frame base, the NR quality metrics are limited to spatial quality estimators. Temporal metrics such as Temporal Information (TI)[56], Mean Motion Intensity (MMI) [57] and jerkiness [57] are therefore not considered. In addition, these metrics should be low complexity pixel-based features, which can be run on light-weight devices in real-time as new frames arrive. As a result, the following set of NR metrics (as proposed in *Van Damme et al.* [13]) is included, which have already proven their suitability to (near) real-time quality assessment of natural videos in previous studies [58–60]. Note that, in case of per-video or per-frame analysis, the per-frame values are averaged over the respective number of frames afterwards (except for Spatial Information (SI) where the maximum is taken by definition):

- *Noise*. Noisy pixels are identified by thresholding on the difference between the local derivative and the average derivative. Both the *average noise (NOI)* (the average difference divided by the total amount of noisy pixels) and *noise ratio (NRT)* (the number of noisy pixels to the total number of pixels) are calculated [61]
- *Blur*. Blurred pixels are identified by thresholding on the difference between a pixel and the corresponding pixel in the derivative image. Both the *average blur (BLU)* (the

average difference divided by the total amount of blurred pixels) and the *blur ratio (BRT)* (number of blurred pixels to the total amount of edge pixels, after applying an edge detection algorithm) are calculated [61].
- *Blockiness (BLK)*. Calculated by analyzing the inner and outer edges of 8x8 subblocks on both the vertical and horizontal Sobel-filtered versions of the frame. As such, an *inner* and *edge blockiness* level is determined, of which the average difference over all blocks describes the blockiness value of the frames [62].
- *SI*. Measurement for the degree of spatial detail, calculated by taking the standard deviation of the pixel intensities of a Sobel-filtered version of each frame [56].
- *Bandwidth (BW)*. The bandwidth at which the video is being sent.

In terms of benchmarking, we evaluated four well-known image-based metrics concerning their PLCC, SROCC, and RMSE towards both ACR and DSIS MOS:

- Netflix's VMAF. This FR quality metric is forged out of four features (ANSNR, DLM, VIF and MI) using an SVR approach with subjective MOS as the benchmark [63]. We chose it due to its proven strong correlation with subjective MOS for point cloud video in a wide variety of compression and network conditions as shown in previous research [8–10]. Therefore, VMAF is used as a direct benchmark, without any further logistic mapping.
- The VQM, as implemented by the Institute for Telecommunication Sciences (ITS) [64].
- SSIM. The well-established perception-based model that considers image degradation as perceived change in structural information [65].
- PSNR. The logarithm of the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation.

Note that all of these metrics allow for per-GOP and per-frame quality benchmarking contrary to subjective evaluation methods such as ACR or DSIS. Furthermore, it is also important to mention that other deep-learning-based perceptual metrics exist, such as Learned Perceptual Image Patch Similarity (LPIPS) [66] and Deep Image Structure and Texture Similarity (DISTS) [67]. Although they have proven to be promising for predicting image quality perception in terms of correlation with subjective scores, their applicability on a per-GOP and per-video level is still to be proven. Moreover, contrary to VMAF, VQM, SSIM, and PSNR, they have not yet been evaluated on the specific case of point cloud content.

3.3 K-Means classification and psychometric QoE modeling

The proposed model [13] (Fig. 2) calculates the perceived quality as a sigmoidal fitting of a weighted linear combination of NR metrics Q_{NR} towards an objective FR benchmark Q_{FR}, where the different weights and parameters are determined based on the particular class of the video. This classification is needed as literature has shown that different types of videos can rely on totally different types of NR metrics for quality estimation [5].

Whenever a new video is received at the client side (indicated in black in Fig. 2), the set of NR-metrics Q_{NR} is calculated on each of the incoming frames. Once a

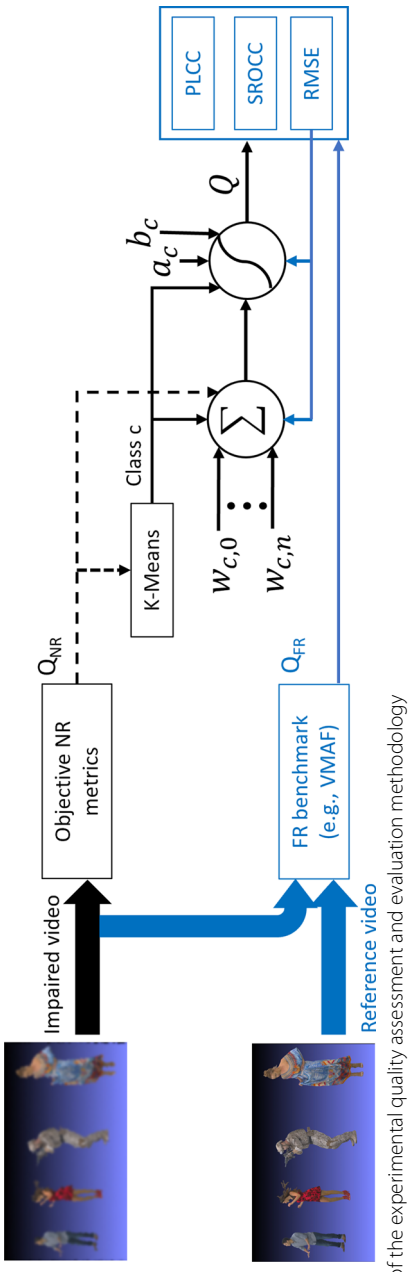


Fig. 2 Block diagram of the experimental quality assessment and evaluation methodology

sufficiently large portion of the video (e.g., one GOP) is received, the obtained metrics are averaged (apart from SI, where the maximum is taken by definition [69]) to characterize the given video. The obtained characterization is fed to a *K-Means* classifier in order to obtain the class c of the given video. K-Means is chosen for its fast training and evaluation times as well as its intuitive interpretation. This classifier is pre-trained at server side using already available sequences. Note that K-Means also allows for easy updating of this classification whenever new content is added to the server-side database. Based on the resulting class c, the weights $w_{c,i}$ for a linear combination of the NR-metrics x_i as well as the parameters a_c and b_c of a sigmoidal mapping (Eq. 1) are determined. Note that the latter is based upon the well-known Quality of Service (QoS)–QoE relationship proposed by *Fiedler et al.* [68] (Fig. 3). The subsequent calculation of both results in the quality prediction Q, which lays within the [0, 100] interval:

$$Q = \sigma_c\left(w_{c,0} + \sum_{i=1}^{7} w_{c,i} \cdot x_i\right) \text{with } \sigma_c(x) = \frac{100}{1 + e^{a_c x - b_c}}. \tag{1}$$

At server side (indicated in blue in Fig. 2), both the weights $w_{c,i}$ and the sigmoid parameters a_c and b_c are trained by minimizing the MSE against the FR benchmark Q_{FR} which is known to correlate strongly to subjective scores (e.g., VMAF [8]). Note that this metric cannot be calculated at client side as the undistorted content is unavailable. Furthermore, Q_{FR} is also used for evaluation of the obtained models (e.g., in terms of PLCC, SROCC and RMSE). Note that the calculation of Q_{FR} on new content is only needed if it fundamentally differs from the current dataset. This can be done on the server side, however, where computational and time-related requirements are less stringent. By sending the appropriate weights and parameters as well as the classifier to the client at video request, this provides all necessary tools for client-side quality estimation.

In order to further explore the applicability of the objective NR model for quality assessment of streamed point cloud video, we fundamentally enlarge the video-level analysis presented in our previous work [13]. In addition, we extend the model to also

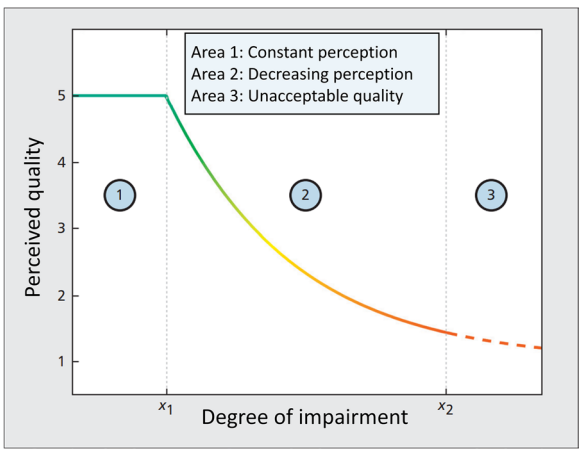


Fig. 3 Typical, general relationship between the degree of impairment and the perceived quality in a multimedia service [68]

cover GOP- and frame-level granularity. In addition, it will be investigated to what extent the preference towards a content-independent one-for-all model can hold for smaller time frames and when one has to rely on per-video models.

4 Implementation

To enable a straightforward frame-level calculation of aforementioned metrics, a *Python*-based CLI *Video Metric Tool*[1] is created to provide the calculation of these metrics for a given input video. The CLI (*video_metric_tool.py* with *argparse.py* as a backbone) takes the path to a given video as an input. Optionally, a list of metrics to include or exclude from the full list (Sect. 3.2) can be provided, e.g., to speed up calculations. The tool provides an output *.csv*-file (of which path and name can be specified) with a per-frame calculation of the NR metrics based on the implementations referenced in Sect. 3.2. Note that TI, Motion Intensity (MI) and jerkiness are also available in the NR-library (*nrlib.py*) for calculation, but were not used in the analysis of this work. The file *analyser.py* is called by the command-line application and is responsible for reading in the given video frame-by-frame and to return the resulting *.csv*-file. In addition, it translates the NR metrics given as an input to the according functionality in *nrlib.py*. Note that this mapping behavior also allows for the easy addition of additional metrics or alternative implementations by adding the corresponding implementations in *nrlib.py* and altering the mapping in *analyser.py* accordingly.

5 Evaluation

This section describes the evaluation of the proposed QoE estimation procedure. First, the dataset used for this research will be described, including the subjective annotations (i.e., MOS) and the way they were gathered. Furthermore, the results of the modeling approach on a video-, GOP- and frame-level are presented.

5.1 Dataset

For the evaluation of the model, we used the subjectively annotated point cloud dataset from our previous work [13, 70]. It consists of 16 source videos (Table 1), each between 18 and 50 s of length and with a framerate of 30 fps. Each sequence contains the generated viewport of a scene consisting of four point cloud objects from the 8i dataset [71], each with a different setup of the figures (line, semi-circle, circle and square) and a different type of camera movement (panning, zoom, rotation and zig-zag).

The point cloud objects were encoded using the Video-based Point Cloud Compression (VPCC) encoder [72] with MPEG's five reference quality representations [73], resulting in bitrates between 2.4 Mbps and 53.5 Mbps (compared to uncompressed bitrates in the order of Gbps). The resulting point cloud objects are then merged together in different scenes. Four different types of scenes are distinguished (Fig. 4): one in which the objects are standing next to each other in a straight line, an analogue one in which they are positioned on a semi-circle, a setup in which the objects are facing each other in a circle and a constellation in which the objects are placed on the corners of a square (all facing the same direction). By combining these constellations with different types of

[1] https://github.ugent.be/samdamme/VideoMetricTool.

(a) Circle (b) Line

(c) Semi-circle (d) Square

Fig. 4 Screenshots of the multiple scene configurations, i.e. **a** Circle, **b** Line, **c** Semi-circle and **d** Square

Table 1 Summary of the 16 source videos in the dataset in terms of point cloud constellation, duration and camera movement. Videos 1–3 are subjectively annotated with both ACR and DSIS MOS

Video	Setup	Duration	Camera movement
1	Line	24s	Pan left-to-right and back (angle)
2	Semi-circle	18s	Zoom-in/zoom-out + rotate to next (*loot* and *redandblack*)
3	Semi-circle	18s	Zoom-in/zoom-out + rotate to next (*soldier* and *longdress*)
4	Circle	24s	Outside rotation
5	Circle	24s	Outside rotation + zoom-in/zoom-out
6	Circle	24s	Outside rotation + zoom-in/zoom-out in between figures
7	Line	24s	Pan left-to-right and back (angle)
8	Line	24s	Pan left-to-right and back (frontal)
9	Line	24s	Rotate left-to-right and back
10	Semi-circle	50s	Zoom-in/zoom-out + rotate to next
11	Semi-circle	24s	Rotate left-to-right and back
12	Semi-circle	24s	Rotate + pause on figure
13	Square	24s	Outside rotation
14	Square	24s	Outside rotation + zoom-in/zoom-out
15	Square	24s	Outside rotation + zoom-in/zoom-out
16	Square	24s	Zig-zag

camera movement (panning, zoom, rotation, zig-zag), a total of 16 FoVs was generated and rendered using the MPEG point cloud compression renderer[74].

Furthermore, a network was emulated using *Mininet* [75] with a single client connected to a HTTP/1.1 enabled *Jetty* server [76]. The available bandwidth was fixed on different discrete values (15, 20, 60, 100, 140 and ∞ Mbps) using traffic control on a shared link. The buffer size was varied between 0 and 4 s using discrete steps of 1 s. The size of the visual area [77] was used to prioritize objects within the scene. Objects are then ranked based on their priority similar to the approach taken by Hosseini [78], where the Euclidean distance is used to distinguish between objects. After ranking these objects according to their priority, the available bandwidth is assigned to the four point cloud objects using three different schemes:

- *Uniform:* Starting with the highest ranked point cloud object, the quality of the different objects is increased one representation at the time. On the upside, this results in similar quality for all objects, creating a more smooth FoV. On the downside, one might allocate valuable bandwidth to objects never being consumed.
- *Greedy:* The highest ranked point cloud object is given the highest possible quality before considering the next object. This is a rather useful approach when a limited amount of objects is in scope, but might significantly reduce the user's perceived quality when multiple objects are considered.
- *Hybrid:* Here, the collection of objects is divided in two sets: the set of objects within the FoV and the set of objects outside of it. First, the quality of the objects within the FoV is improved uniformly until either the highest quality is assigned or no more bandwidth remains. Next, objects outside the FoV are considered and their qualities are improved one by one using the greedy approach. As such, the advantages of the two previous approaches are combined which is especially useful when the user targets a specific group of subjects.

Note that these quality allocation schemes determine the quality of each specific point cloud object before projection to the 2D plane, contrary to tiling-based schemes that determine the quality of each tile of the viewport after projection has taken place. Furthermore, the latency was set to 37 ms, a reference value for 4G networks [1]. As a result, a total of 453 sequences is obtained.

5.2 Calculation of the quality metrics

For each of the sequences, the same seven objective NR metrics (BLU, BRT, NOI, NRT, BLO, SI and BW) , PSNR, SSIM, VQM and VMAF are calculated analogously as discussed in Sect. 3.2. The NR metrics were implemented in Python following the implementation of our previous research [5] and using the tool created for this work as described in Sect. 4. Note that their mathematical definitions are provided in Appendix A at the end of this manuscript. VMAF was calculated with the freely available *GitHub* tool [79], resulting in values within the [0, 100] interval. It was chosen due to its proven strong correlation with subjective MOS for point cloud video, as proven in previous research [8–10]. Therefore, VMAF is used as a direct benchmark, without any further logistic mapping. VQM [64] is calculated using the freely available tool. To this end, the standardized National Telecommunications and Information Administration (NTIA) General Model was followed, using full-reference calibration. The respective *Python OpenCV* [80] and *SciKit-image* [81] implementations are followed for PSNR and SSIM. All metrics are calculated using their default settings. Note that each of these benchmarks allow for per-GOP and per-frame quality benchmarking contrary to subjective evaluation methods such as ACR or DSIS. For the GOP level, the resulting set of per-frame metrics is chunked in parts with a size of 1 GOP (= 30 frames). For each chunk, the 30 per-frame values of each metric are averaged to obtain the per-GOP metrics (apart from SI, which uses the maximum by definition). Analogously, metrics are averaged over the whole video (or the maximum is taken in the case of SI) to obtain the video-level metrics.

In addition, this dataset is also partially annotated with subjective MOS [1]. These were collected for all sequences of source videos 1–3, both in ACR (26 subjects) and DSIS (28 subjects) fashion.

5.3 Evaluation approach

The evaluation approach can be divided into three important parts: (i) the K-Means classifier, (ii) the evaluation of possible benchmarks and (iii) the actual modeling as a combination of a linear regressor and a sigmoidal mapping.

5.3.1 K-Means classifier

The K-Means classifier used the same implementation as in our previous work [13], which is based on *Python*'s *SciKit Learn* library [82]. The algorithm is ran 10 times with a maximum of 300 iterations per run. The relative tolerance is set to 0.0001. As K-Means is distance based, NR metrics not laying within the [0, 1] interval (BW, BLU, NOI, SI) are first normalized using a min–max scaler before putting them into the algorithm. Depending on the granularity under scrutiny, either the per-frame, per-GOP or per-video metrics are fed to the algorithm. The number of clusters is determined for each granularity level by optimizing the *Silhouette Score (SS)* [83]. To this end, the K-Means algorithm is run multiple times with the number of clusters varying from 1 to 16 (= the number of videos). The optimal number of clusters is determined as the run in which the maximal SS is obtained. The resulting, optimal number of clusters and according SS for each granularity level are shown in Table 2. Note that if the multiple configurations of a given video end up in different clusters after the K-Means algorithm, they are completely assigned to the cluster with the largest number of sequences to allow for easy cross-validation in the modeling step. Furthermore, the GOP size is set to 30, which corresponds to 1 s of video playback.

5.3.2 Benchmarks

The performance of the calculated benchmarks is evaluated in terms of linearity, monotony and accuracy by respective means of PLCC, SROCC and RMSE. These are calculated with respect to both ACR and DSIS MOS. To this end, evaluations are performed on each individual video as well as on average.

5.3.3 QOE model evaluation

The performance of the model on each video is cross-validated by leaving the particular video out as a test set and optimizing the model parameters (by minimizing MSE) on the remaining videos in the cluster. When evaluating video 6, for example, which is in a

Table 2 The optimal number of clusters and according SS at each granularity level

Level	Optimal number of clusters	SS
Video	4	0.606
GOP	7	0.348
Frame	2	0.395

cluster with videos 4,5 and 13, the training set consists of all data from videos 4, 5 and 13 while the test set encompasses all configurations from video 6. In case no clustering is applied, all videos are considered, i.e., all configurations from videos 1–5 and 7–16 make up the training set while all configurations from video 6 are included in the test set. Note that in case a video ends up in a separate cluster on its own after the clustering step, or when a per-video model is constructed, a 5-fold cross-validation is performed on the different configurations of that particular video. Here, the composition of each fold is determined at random. Note that the normalization parameters of BW, BLU, NOI and SI are recalculated at each iteration of the cross-validation on the training set only to avoid data leakage.

5.4 Results and discussion

In this section, the results will be discussed on three levels of granularity: video-level, GOP-level (30 frames), and frame-level. Note that the references videos were excluded from this analysis (apart from the calculation of benchmarks), as their perfect score of 100 in terms of VMAF benchmark would put to much confusion and bias on the finer granularity levels of the model. This decision is justified, as the transmission of a video sequence that exactly matches its reference (so without any encoding or network distortion in place) can be considered non-existent in practical scenarios.

5.4.1 Benchmarks

Table 3 shows the obtained performance values of the four benchmarks under scrutiny towards both flavors of MOS. When looking at ACR MOS, the best performing metric

Table 3 PLCCs, SROCCs and RMSEs of the four benchmarks to both ACR and DSIS subjective MOS. For each MOS flavor, benchmark, and evaluation metric the best score (highest for PLCC and SROCC, lowest for RMSE) is indicated in bold. For PSNR, which is unconstrained by definition, a min–max scaler was applied to make it fall within the [0,100] interval

	Video	**MOS_{ACR}**			**MOS_{DSIS}**		
		PLCC	**SROCC**	**RMSE**	**PLCC**	**SROCC**	**RMSE**
PSNR	1	0.960	0.929	**8.257**	0.963	**0.821**	28.486
	2	0.978	0.750	47.027	0.978	0.857	26.417
	3	**0.989**	0.964	17.649	0.993	**0.964**	26.111
	Avg.	0.976	0.881	**24.311**	0.978	0.881	27.005
SSIM	1	0.955	0.929	52.677	0.963	**0.821**	26.925
	2	**0.989**	0.679	57.929	**0.979**	0.892	37.880
	3	0.981	0.964	58.277	**0.995**	**0.964**	44.445
	Avg.	0.975	0.857	56.294	0.979	0.892	36.417
VQM	1	**0.971**	**0.964**	45.669	0.951	0.750	20.047
	2	0.984	**0.821**	52.573	**0.979**	0.821	32.485
	3	0.981	**0.999**	52.097	0.967	0.893	38.234
	Avg.	**0.979**	**0.928**	50.113	0.966	0.821	30.255
VMAF	1	0.948	0.929	35.507	**0.969**	**0.821**	**10.093**
	2	0.984	0.679	**37.723**	**0.979**	**0.893**	**17.398**
	3	0.986	0.964	33.241	0.993	**0.964**	**19.095**
	Avg.	0.973	0.857	35.490	**0.980**	**0.893**	**15.529**

in terms of PLCC tends to differ depending on the particular video, although it has to be said that all metrics are showing high correlations in general. In terms of SROCC, VQM is showing the best performance especially with respect to video 2. SSIM and VMAF show more struggle on this video with SROCCs below 0.7, while PSNR holds the middle with a value of 0.750. For videos 1 and 3, high SROCCs are obtained by all metrics. With regard to RMSE, both SSIM and VQM are showing rather high values which, taking the high correlation values into consideration, hints at severe over- or under-prediction. PSNR is showing the best performance in this regard, although the performance depends on the video at hand as can be seen from the high value for video 2. VMAF somewhat holds the middle with values below the ones of VQM and SSIM, but clearly higher values for videos 1 and 3 compared to PSNR. Nevertheless, it shows the most constant performance in this respect, therefore acting more reliable than PSNR.

When evaluating performance towards DSIS MOS, VMAF is clearly showing the best performance over all metrics. Only for video 3, a slightly higher PLCC is reported for SSIM. Nevertheless, PSNR, SSIM and VQM are clearly showing higher values in terms of RMSE compared to VMAF.

Based on this discussion, as well as on the results obtained from literature (Sect. 2.3), we decided VMAF to be the most appropriate choice for benchmarking. On one hand, this decision is based on the observation that most works in literature identify VMAF as the best performing among image-based metrics. On the other hand, DSIS MOS seems to be most applied as a subjective evaluation methodology for point cloud content, of which the performance of VMAF is clear on this dataset based on the above results. Nevertheless, it is important to mention that a clear, stable and reliable metric for benchmarking point cloud video is yet to emerge, such that the choice of benchmark can be debated and probably depends on the dataset under scrutiny.

5.4.2 QoE-model: video-level

Table 4 shows detailed results of the one-for-all video-level model. Note that the performance of per-video models is not analyzed in this Section, as there are insufficient data points for model training and validation on video-level for this purpose. The first and most straightforward conclusion is the clear positive impact of clustering to the performance of the model. In terms of PLCC, improvements can be noticed for all cases, with correlation scores going above 0.95 in the majority of the cases. For some of the videos, impacts are especially high as is the case for videos 3 and 12 in the non-mapping case, where improvements of 0.406 and 0.199, respectively, can be observed. In the case with sigmoidal mapping, we see remarkable improvements for videos 3 and 10, with respective increases of 0.523 and 0.190. On average, a clear improvement of 0.116 is observed in the non-sigmoidal case, as well as an improvement of 0.108 in the sigmoidal case.

For the SROCC scores, we also notice improvements for all videos except 7 and 10. Decreases there are limited, however, with differences of 0.007 and 0.044, respectively. The most important improvements can be noticed for videos 1,2 and 13, with increases of 0.249, 0.143 and 0.112, respectively. Note that, despite its increase, the SROCC of video 2 is still limited (0.679) after applying clustering. On average, clustering shows an increase in SROCC of 0.047. Furthermore, it is also worth mentioning that SROCCs

Table 4 PLCCs, SROCCs and RMSEs towards VMAF of the one-for-all video-level model in comparison with the cases without clustering (Cl.) and/or sigmoidal mapping. For each video and each case, the best performing metric (highest for PLCC/SROCC and lowest for RMSE) is indicated in bold

	Video	No clustering			Clustering			
		PLCC	SROCC	RMSE	Cl.	PLCC	SROCC	RMSE
No sigmoidal mapping	1	0.969	0.750	**5.576**	2	**0.996**	**0.999**	13.213
	2	0.820	0.536	17.200	0	0.946	**0.679**	15.138
	3	0.568	0.714	25.958	2	**0.974**	**0.821**	15.047
	4	0.835	0.985	1.803	1	**0.997**	**0.994**	3.001
	5	0.838	0.942	2.403	1	**0.994**	**0.968**	8.021
	6	0.868	0.907	2.601	1	0.986	**0.978**	3.198
	7	0.863	**0.988**	4.226	0	**0.988**	0.981	**1.018**
	8	0.805	0.984	3.234	0	0.985	**0.986**	2.202
	9	0.817	0.959	3.115	0	0.987	**0.972**	1.842
	10	**0.904**	**0.988**	**3.811**	3	0.319	0.944	46.119
	11	0.818	0.966	3.085	0	**0.985**	**0.974**	1.417
	12	0.787	0.982	3.291	0	0.986	**0.990**	3.929
	13	0.827	0.837	2.075	1	**0.998**	**0.949**	2.480
	14	0.843	0.942	2.115	0	**0.973**	**0.948**	8.366
	15	0.847	0.867	2.095	0	0.967	**0.896**	6.918
	16	0.809	0.960	**5.352**	0	**0.991**	**0.977**	19.595
	Avg.	0.826	0.894	11.440	/	0.942	**0.941**	9.469
Sigmoidal mapping	1	0.970	0.750	6.149	2	**0.996**	**0.999**	12.411
	2	0.822	0.536	30.335	0	**0.973**	**0.679**	**10.652**
	3	0.451	0.703	17.227	2	**0.974**	**0.821**	**14.585**
	4	0.934	0.985	**1.454**	1	0.996	**0.994**	3.036
	5	0.910	0.942	**2.194**	1	0.993	**0.968**	7.960
	6	0.910	0.907	**2.478**	1	**0.987**	**0.978**	3.228
	7	0.945	**0.988**	3.703	0	0.987	0.981	1.206
	8	0.907	0.984	2.713	0	**0.992**	**0.986**	**1.394**
	9	0.922	0.959	2.480	0	**0.993**	**0.972**	**1.064**
	10	0.651	**0.988**	20.835	3	0.841	0.944	28.552
	11	0.909	0.966	2.620	0	**0.985**	**0.974**	**0.963**
	12	0.893	0.982	**2.661**	0	**0.997**	**0.990**	3.162
	13	0.917	0.837	**1.883**	1	**0.998**	**0.949**	2.568
	14	0.932	0.942	**1.660**	0	0.957	**0.948**	5.570
	15	0.933	0.867	**1.636**	0	**0.977**	**0.896**	6.185
	16	0.901	0.960	5.485	0	0.988	**0.977**	20.424
	Avg.	0.869	0.894	19.858	/	**0.977**	**0.941**	**7.685**

are the same for the cases with and without sigmoidal mapping as applying a monotone function does not change the mutual ordering of the results.

In terms of RMSE, results only seem to be positively impacted by applying this sigmoidal function when combined with the clustering algorithm. For videos 2 and 10, for example, decreases of 4.486 and 17.567 on a 0-100 scale are observed. In the non-clustering case, however, these same videos show increases in RMSE of 13.135 and 17.024, respectively. On the downside, some of the videos show a small decrease when applying sigmoidal mapping after clustering, such as is the case for videos 4, 6, 7, 13 and 16.

These decreases are marginal, however, and do not weigh up against the big gains sigmoidal modeling can bring to other videos. Despite, there can be noticed that for videos 1, 2, 3 and 10, the resulting RMSEs remain high compared to the other videos while still showing high correlation values. As such, this is an indication of either over- or under-prediction. This could be a result of the fundamental different properties of the videos in terms of blockiness and blur ratio as is illustrated in Fig. 5. This is also reflected in the fact that videos 1, 3 and video 10 end up in their own cluster. Only video 2 was assigned to the major cluster due to its fundamental difference in SI compared to 1, 3 and 10 as is shown in Fig. 6. Reapplying the proposed modeling approach on the clusters that can be derived from Fig. 5 did not improve results, however. As such, it could be the case that more optimal modeling methods can be designed for these outlying videos. This is a subject for further research, though. Nevertheless, a decrease of RMSE with 1.784 is observed on average when applying sigmoidal mapping post-clustering. On the contrary, in the non-clustering case, we even observe an increase in RMSE with 8.418.

Table 5 shows the subjective correlations and RMSEs of the NR metrics, the ground truth VMAF values and the VMAF predictions ($VMAF_p$) to the ACR and DSIS MOS annotations of videos 1–3. For the ACR MOS it can be noticed that the high PLCC values are maintained by the predicted values with only a small decrease of 0.055 for video 3 and even minor increases for videos 1 and 2 compared to VMAF.. On average, the difference is limited to a 0.015 decrease. Interestingly, this performance is comparable to both the BLU, BRT and SI NR metrics. In terms of SROCC, the view is mixed. On average, the predictions show a small increase of 0.036 compared to VMAF. More remarkable is video 2, on one hand, with an increase of 0.320 to an almost perfect 0.999. Video 3, on the other hand, suffers from a 0.214 decrease to an SROCC of 0.750. Also here, similar performance can be observed from BRT, SI and BLU. For RMSE, the most clear observation is the fact that values are rather high for both ground truth and predictions. This is intuitive, however, as VMAF is designed as a FR metric comparing distorted with undistorted content. As such, better performance compared to DSIS MOS can be expected, especially since the quality of undistorted point cloud video content is still below user expectations, therefore resulting in more strict scores [1]. On average, a minor increase of 3.035 RMSE can be observed compared to VMAF. Larger increases can be observed for videos 1 and 2 however, with values of 12.298 and 10.415, respectively. Video 3, in contrary, shows a decrease in RMSE of 13.61. For the NR metrics, rather similar RMSEs are obtained with NOI performing well on video 3 similar to the proposed method while SI and BRT are showing good performance on videos 1 and 2, respectively.

For the DSIS MOS, a similar observation can be made for the PLCCs as for ACR. The high performance compared to VMAF in terms of PLCC is maintained with only a limited decrease of 0.013 on average and 0.005, 0.023 and 0.013 for videos 1-3, respectively. Once again, similar performance can be observed from BLU, BRT, and SI. In terms of SROCC correlations, decreases can be noticed for videos 2 and 3 when comparing to VMAF, be it more limited than in the ACR case with respective differences of 0.072 and 0.035 and 0.036 on average. Also in terms of SROCC, the good performance of BLU, BRT and SI still holds. The DSIS RMSEs, at last, are indeed much lower than their ACR counterpart, as previously discussed. On the downside, videos 1 and 2 do show increases in RMSE of 11.308 and 10.602, respectively, when compared to VMAF, while video 3

Table 5 PLCCs, SROCCs and RMSEs of both the ground truth VMAF and the predicted VMAF ($VMAF_p$), using clustering and sigmoidal mapping, to both ACR and DSIS subjective MOS of videos 1–3. Furthermore, the performance of each NR metric as a single-feature predictor is indicated as well

	Video	MOS_{ACR}			MOS_{DSIS}		
		PLCC	**SROCC**	**RMSE**	**PLCC**	**SROCC**	**RMSE**
NOI	1	− 0.797	− 0.464	40.217	− 0.810	− 0.464	67.102
	2	− 0.492	− 0.036	53.354	− 0.426	− 0.143	33.677
	3	− 0.951	− 0.929	**15.449**	− 0.916	− 0.929	22.641
	Avg.	− 0.747	− 0.476	36.340	− 0.717	− 0.512	41.140
NRT	1	0.115	0.214	36.386	0.081	0.071	63.339
	2	0.656	0.714	30.106	0.708	0.679	50.915
	3	0.369	0.571	29.443	0.481	0.571	44.481
	Avg.	0.380	0.500	31.978	0.423	0.440	52.912
BLU	1	0.944	**0.929**	30.898	0.951	0.821	4.718
	2	0.744	0.750	55.938	0.796	0.857	35.760
	3	0.973	0.893	28.222	0.990	**0.999**	42.978
	Avg.	0.887	0.857	38.353	0.912	0.892	27.819
BRT	1	− 0.951	− 0.857	37.383	− 0.964	− 0.750	64.342
	2	− 0.981	− 0.964	**28.033**	− 0.974	− 0.929	48.829
	3	− 0.981	**− 0.964**	30.231	**-0.994**	**-0.964**	45.268
	Avg.	− 0.971	**− 0.928**	**31.882**	− 0.977	− 0.881	52.813
BLK	1	− 0.125	0.107	41.874	− 0.496	− 0.393	17.092
	2	− 0.508	− 0.714	48.958	− 0.386	− 0.321	29.175
	3	− 0.264	− 0.536	47.613	− 0.184	− 0.464	34.182
	Avg.	− 0.299	− 0.381	46.148	− 0.355	− 0.393	26.816
SI	1	0.931	0.821	**23.436**	**0.984**	**0.964**	**4.011**
	2	0.958	0.703	32.369	0.947	**0.937**	52.982
	3	0.962	0.893	49.692	0.948	0.714	35.081
	Avg.	0.950	0.806	35.166	0.960	0.872	30.691
VMAF	1	0.948	**0.929**	35.507	0.969	0.821	10.093
	2	0.984	0.679	37.723	**0.979**	0.893	**17.398**
	3	**0.986**	**0.964**	33.241	0.993	0.964	19.095
	Avg.	**0.973**	0.857	35.490	**0.980**	**0.893**	**15.529**
$VMAF_p$	1	**0.952**	**0.929**	47.805	0.964	0.821	21.401
	2	**0.990**	**0.999**	48.138	0.956	0.821	28.000
	3	0.931	0.750	19.631	0.980	0.929	**9.034**
	Avg.	0.958	0.893	38.525	0.967	0.857	19.478

shows a similar level of decrease with a value of 10.061. As a result, the average increase is limited to 3.949. In this case, much higher RMSE values can be observed for the NR metrics, such that it is clear that this is where the main value of the video-level model is situated compared to single-feature NR prediction. In summary, while deviations on a per-video basis can be observed, the average performance of the predictor shows to be comparable to ground truth VMAF for both the ACR and DSIS cases. This is also the case for BLU, BRT and SI, apart from RMSE when evaluating against DSIS MOS.

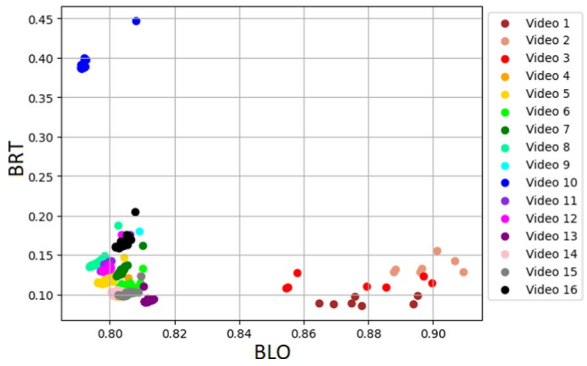


Fig. 5 BLO vs. BRT scatter plot for the 16 videos at video-level

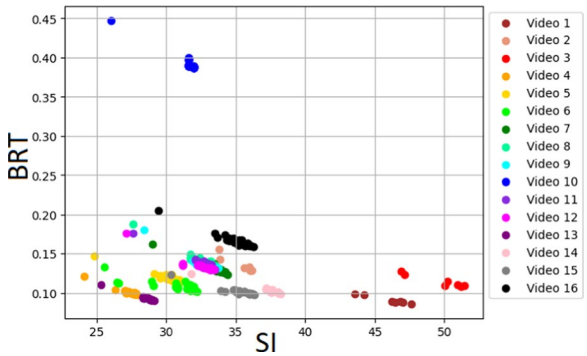


Fig. 6 SI vs. BRT scatter plot for the 16 videos at video-level

5.4.3 QoE-model: GOP-level

Table 6 shows the results of the same one-for-all modeling approach as in the previous Section, but trained on GOP-level (30 frames = 1 s) rather than video-level. In terms of PLCC, the sigmoidal mapping shows to have a mostly positive impact on the results with most of the videos showing an increase in correlation for both the non-clustering and the clustering case. Only videos 3, 6, 10 and 16 show a decrease in PLCC in the non-clustering case, be it with differences of 0.003, 0.070, 0.094 and 0.010, respectively. In the clustering case only video 10, of which the outlying behavior was shown in the previous Section, shows a decrease in PLCC with a marginal difference of 0.004. On average, limited performance increases are observed of 0.017 in the non-clustering case and 0.036 in the clustering case. The influence of this clustering results in a more mixed view regarding PLCC values. In the non-sigmoid case, half of the videos (1,2,5,8,9,11,12 and 16) are showing an increase in PLCC due to clustering, with this effect being most prominent for video 1 with an increase of 0.096. For the other videos a decrease is observed which is, once again, most pronounced for video 10 with a decrease of 0.188. Video 15 also suffers from a decrease worth mentioning, however, with a difference of 0.115. On average, PLCC stays about the same with respective values of 0.597 and 0.595. In the case where post-clustering sigmoidal mapping is applied, results are looking somewhat better with 10 out of 16 videos showing an increase in PLCC as a result of clustering. This is, once again, most pronounced for video 1 with an increase from 0.688 to 0.782.

Table 6 PLCCs, SROCCs and RMSEs towards VMAF of the one-for-all GOP-level model in comparison with the cases without clustering (Cl.) and/or sigmoidal mapping. For each video and each case, the best performing metric (highest for PLCC/SROCC and lowest for RMSE) is indicated in bold

	Video	No clustering			Clustering			
		PLCC	SROCC	RMSE	Cl.	PLCC	SROCC	RMSE
No sigmoidal mapping	1	0.668	0.659	4.246	4	0.764	**0.767**	**3.296**
	2	0.786	0.854	6.583	0	0.822	**0.869**	4.027
	3	**0.601**	**0.584**	8.260	3	0.556	0.546	8.321
	4	0.675	**0.582**	2.919	6	0.658	0.488	2.719
	5	0.490	0.430	5.372	6	0.542	**0.459**	5.457
	6	**0.165**	**0.356**	**8.312**	3	0.133	0.263	25.085
	7	0.678	**0.684**	3.633	4	0.656	0.617	4.722
	8	0.649	0.588	5.515	6	0.715	**0.607**	4.703
	9	0.613	**0.536**	5.351	4	0.641	0.439	4.565
	10	**0.497**	**0.695**	**13.805**	5	0.309	0.554	20.524
	11	0.639	0.579	4.965	6	0.684	**0.626**	4.538
	12	0.569	**0.455**	5.203	4	0.644	**0.455**	4.603
	13	0.714	**0.723**	2.783	5	0.675	0.687	8.921
	14	0.595	**0.696**	4.444	1	0.590	0.680	3.121
	15	0.706	**0.762**	3.825	6	0.591	0.429	3.985
	16	0.505	**0.540**	7.528	2	0.537	0.538	7.045
	Avg.	0.597	**0.608**	**5.797**	/	0.595	0.564	7.227
Sigmoidal mapping	1	0.688	0.659	4.506	4	**0.75**0.782	**0.767**	4.274
	2	0.793	0.854	8.421	0	**0.75**0.850	**0.869**	3.686
	3	0.598	**0.584**	8.805	3	0.582	0.546	**7.856**
	4	**0.725**	**0.582**	2.627	6	0.716	0.488	**2.454**
	5	0.523	0.430	**5.233**	6	**0.75**0.583	**0.459**	5.372
	6	0.095	**0.356**	9.799	3	0.141	0.263	26.031
	7	**0.725**	**0.684**	**3.479**	4	0.719	0.617	4.389
	8	0.693	0.588	5.158	6	**0.75**0.766	**0.607**	**4.368**
	9	0.658	**0.536**	5.006	4	**0.75**0.693	0.439	**4.341**
	10	0.403	**0.695**	14.641	5	0.305	0.554	26.940
	11	0.684	0.579	4.671	6	**0.75**0.736	**0.626**	**4.157**
	12	0.624	**0.455**	4.879	4	**0.75**0.701	**0.455**	**4.283**
	13	**0.765**	**0.723**	**2.449**	5	0.713	0.687	8.201
	14	0.608	**0.696**	4.337	1	**0.75**0.624	0.680	**3.064**
	15	**0.751**	**0.762**	**3.492**	6	0.637	0.429	3.934
	16	0.495	**0.540**	7.712	2	**0.75**0.553	0.538	**6.996**
	Avg.	0.614	**0.608**	5.951	/	**0.75**0.631	0.564	7.522

For the videos showing a decrease in PLCC (3, 4, 7, 10, 13 and 15), differences are rather limited. Only for videos 10, 13 and 15 higher decreases are observed of 0.098, 0.052 and 0.114, respectively.

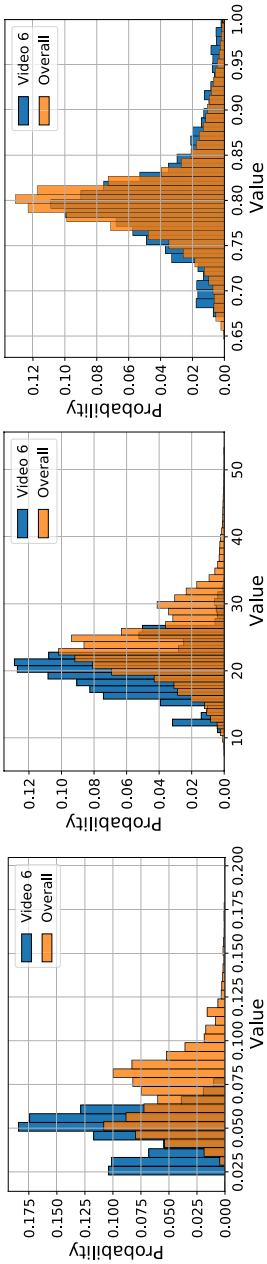
In terms of SROCC, results are not affected by sigmoidal mapping due to its monotonic nature. Clustering shows to have a mostly negative effect on SROCC, with 10 of 16 videos showing a decrease. This is especially pronounced for videos 4, 6, 10 and 15 with respective decreases of 0.094, 0.093, 0.141 and 0.333, respectively. Videos 1, 2, 5, 8, 11 show an increase in correlation which is once again most pronounced for video 1 with

an increase of 0.108. On average, clustering seems to reduce performance, however, with a decrease of 0.044 in SROCC.

When it comes to RMSE, mixed results can also be observed as 7 and 6 out of 16 videos show a decrease when adding clustering in the non-sigmoid and sigmoid cases, respectively. In the non-sigmoid case, most increases in RMSE are limited with exception of videos 6, 10 and 13 where respective differences of 16.773, 11.773 and 6.138 are observed. On average, an increase in RMSE of 1.430 is noticed. In the case with sigmoidal mapping, similar results can be seen with decreases or small increases in RMSE for all videos except 6, 10 and 13. The latter show increases of 16.232, 12.299 and 5.752, respectively. On average, an increase of 1.571 can still be noted though. The impact of sigmoidal mapping to RMSE shows to be mainly beneficial, with 10 out of 16 videos showing better results in the non-clustering case and 13 out of 16 in the clustering case. In the non-clustering case, limited increases in RMSE are observed for videos 1, 2, 3, 6, 10 and 16. In the clustering case, the same is true for videos 1 and 6. Only video 10 is showing deviating behavior with an RMSE increase of 6.416.

It is worth pointing out that video 6 is consistently underperforming in all cases. The most probable explanation for this observation is the deviating distributions of NRT and, to less extent SI, when compared to the average over all videos as is depicted in Fig. 7. There can clearly be seen in the distributions of NRT that while the both show a two-modal behavior, the average distribution is much more spread out than is the case for the spiking behavior of video 6. To a lesser extent the same behavior can be observed for SI while this is not the case for the majority of the other metrics as is illustrated for BLK. As such, the model has more trouble generalizing towards the outlying behavior of video 6 than for other videos which are more in line with the global average.

By means of comparison, Table 7 shows the results of optimizing the model parameters for each video separately rather than on a per-cluster base. Evaluation is performed using a 5-fold cross-validation on a per-video base. The positive impact of sigmoidal mapping should be clear in this case, given the improved performance of each single video. For videos 2, 14 and 16, results are of course identical to Table 6 as these were already isolated in a separate class by the clustering algorithm. But also for most of the other videos, similar performance can be observed. For videos 3, 6, and 10, clear improvements can be noticed though. This is especially remarkable for video 6, which is thoroughly underperforming in the one-for-all modeling approach. While still not splendid, results are clearly more in line with the other videos under scrutiny. As such, it seems that clustering might be more confusing than clarifying to the model in some cases. This could be explained by the higher variability in NR metrics due to the finer granularity such that clusters become more heterogeneous than is the case at video-level. Videos 4 and 12, in contrary, seem to suffer from the loss of information of other videos in the same cluster. When to or not to apply model pre-clustering is therefore an interesting and important further research direction. On average, the per-video modeling approach shows slightly better results than the one-for-all approach although these marginal gains are not weighing up to the additional computational complexity of training a model for every new video in the database. For some specific videos that tend to show outlying behavior with respect to the general model, this is an approach that could be considered, however.



(a) NRT (b) SI (c) BLK

Fig. 7 Histograms showing the distributions of **a** NRT, **b** SI and **c** BLK of video 6 (blue) compared to the average distribution over all videos (orange, excluding video 6)

Table 7 PLCCs, SROCCs and RMSEs towards VMAF of GOP-level models trained on a per-video base. A comparison of the cases with and without sigmoidal mapping is provided. For each video and each case, the best performing metric (highest for PLCC/SROCC and lowest for RMSE) is indicated in bold

Video	No sigmoidal mapping			Sigmoidal mapping		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
1	0.759	**0.792**	2.184	**0.783**	**0.792**	**2.083**
2	0.822	**0.869**	4.027	**0.850**	**0.869**	**3.686**
3	0.750	**0.785**	6.199	**0.767**	**0.785**	**5.993**
4	0.466	**0.581**	2.679	**0.503**	**0.581**	**2.590**
5	0.559	**0.569**	4.919	**0.598**	**0.569**	**4.746**
6	0.450	**0.478**	5.214	**0.518**	**0.478**	**5.141**
7	0.628	**0.704**	3.309	**0.653**	**0.704**	**3.241**
8	0.686	**0.715**	4.139	**0.752**	**0.715**	**3.873**
9	0.440	**0.636**	4.442	**0.461**	**0.636**	**4.322**
10	0.663	**0.740**	8.114	**0.668**	**0.740**	**8.055**
11	0.691	**0.646**	4.243	**0.757**	**0.646**	**3.925**
12	0.443	**0.482**	4.926	**0.478**	**0.482**	**4.785**
13	0.623	**0.719**	2.105	**0.675**	**0.719**	**1.993**
14	0.590	**0.680**	3.121	**0.624**	**0.680**	**3.064**
15	0.697	**0.659**	3.115	**0.748**	**0.659**	**2.902**
16	0.537	**0.538**	7.045	**0.553**	**0.538**	**6.996**
Avg.	0.613	**0.662**	4.361	**0.649**	**0.662**	**4.211**

Table 8 Average PLCCs, SROCCs, and RMSEs of the six NR-metrics, PSNR, and SSIM to the VMAF benchmark at GOP-level. The best performing metric (highest for PLCC/SROCC and lowest for RMSE) is indicated in bold. Note that a min–max scaler was applied to NOI, BLU, SI and PSNR to make them fall within the [0, 100] interval

Metric	PLCC	SROCC	RMSE
NOI	− 0.275	− 0.274	16.168
NRT	− 0.511	− 0.511	76.407
BLU	0.384	0.376	75.931
BRT	− 0.115	− 0.111	71.568
BLK	0.128	0.142	**7.014**
SI	− 0.476	− 0.432	40.532
PSNR	**0.765**	**0.740**	31.420
SSIM	0.755	0.737	14.627
Proposed	0.631	0.564	7.522

Table 8 compares the average performance of the proposed model (including both clustering and sigmoidal mapping) in terms of PLCC, SROCC and RMSE to VMAF with each of the separate NR metrics and the PSNR and SSIM benchmarks. It is clear that the proposed approach is improving upon single-feature prediction with any of the NR metrics. Only NRT is getting in the neighbourhood in terms of correlation values, but with an average RMSE that is about a factor 10 higher than the proposed method. In this respect, BLK as a predictor is showing an RMSE that is even slightly better than the proposed solution, but at the cost of very low correlation values. As such, the proposed method shows to combine the best of both worlds. It is also interesting to notice that in terms of RMSE it is even showing better performance than both of the FR benchmarks

with a value of 7.522 compared to 14.627 for SSIM and 31.420 for PSNR. Although the correlation values of the proposed method are somewhat lower, this is in fact a satisfying result giving the NR nature of the proposed solution compared to the FR approach of both SSIM and PSNR.

5.4.4 QoE-model: frame-level

Table 9 shows the results of the one-for-all modeling approach on frame-level. In terms of the impact of clustering, a mixed result can once again be seen. When it comes to

Table 9 PLCCs, SROCCs and RMSEs towards VMAF of the one-for-all frame-level model in comparison with the cases without clustering (Cl.) and/or sigmoidal mapping. For each video and each case, the best performing metric (highest for PLCC/SROCC and lowest for RMSE) is indicated in bold

	Video	No clustering			Clustering			
		PLCC	SROCC	RMSE	Cl.	PLCC	SROCC	RMSE
No sigmoidal mapping	1	0.656	0.620	4.297	0	0.771	**0.761**	**2.676**
	2	0.788	**0.822**	**7.166**	1	0.786	0.821	8.371
	3	0.591	**0.562**	**8.329**	1	0.582	0.552	8.332
	4	0.657	0.498	3.007	1	0.651	**0.514**	3.103
	5	0.474	0.394	5.547	0	0.548	**0.470**	5.285
	6	0.181	**0.327**	**8.109**	1	**0.183**	0.310	8.388
	7	0.671	0.672	3.660	0	0.691	**0.708**	3.451
	8	0.630	0.534	5.614	1	0.618	**0.549**	6.187
	9	0.587	0.487	5.501	1	0.571	**0.494**	6.126
	10	0.492	**0.679**	13.841	1	**0.508**	0.676	**13.590**
	11	0.628	0.549	5.124	0	0.633	**0.566**	4.860
	12	0.559	0.412	5.330	0	0.636	**0.480**	4.752
	13	0.699	0.649	2.767	1	0.688	**0.652**	2.874
	14	0.589	**0.636**	4.387	0	0.596	0.498	3.696
	15	0.683	**0.726**	3.987	1	0.660	0.723	4.515
	16	**0.463**	**0.410**	**7.941**	1	0.445	0.403	8.039
	Avg.	0.584	0.561	5.913	/	0.598	**0.574**	**5.890**
Sigmoidal mapping	1	0.676	0.620	4.551	0	**0.794**	**0.761**	3.397
	2	**0.799**	**0.822**	9.376	1	0.796	0.821	10.680
	3	**0.593**	**0.562**	8.796	1	0.582	0.552	9.144
	4	**0.707**	0.498	**2.729**	1	0.701	**0.514**	2.777
	5	0.506	0.394	5.408	0	**0.582**	**0.470**	**5.168**
	6	0.110	**0.327**	9.186	1	0.112	0.310	9.287
	7	0.718	0.672	3.508	0	**0.739**	**0.708**	**3.283**
	8	**0.672**	0.534	**5.284**	1	0.658	**0.549**	5.762
	9	**0.631**	0.487	**5.179**	1	0.613	**0.494**	5.707
	10	0.407	**0.679**	14.497	1	0.431	0.676	14.060
	11	0.672	0.549	4.839	0	**0.682**	**0.566**	**4.578**
	12	0.611	0.412	5.022	0	**0.694**	**0.480**	**4.419**
	13	**0.749**	0.649	**2.447**	1	0.740	**0.652**	2.494
	14	0.604	**0.636**	4.241	0	**0.631**	0.498	**3.595**
	15	**0.723**	**0.726**	**3.692**	1	0.700	0.723	4.139
	16	0.462	**0.410**	8.070	1	0.435	0.403	8.260
	Avg.	0.603	0.561	6.052	/	**0.618**	**0.574**	6.047

PLCC there can be noticed that for both the sigmoidal and non-sigmoidal case half of the videos (1, 5-7, 10-12, 14) are showing improvement by including clustering into the modeling algorithm. Especially videos 1, 5 and 12 show decent improvements with respective differences of 0.115, 0.074 and 0.077 in the non-sigmoidal case and 0.118, 0.076 and 0.077 in the sigmoidal case. Furthermore, it should be noted that, for videos showing a decrease in performance, differences are rather limited. On average, small improvements of 0.014 and 0.015 for the respective non-sigmoid and sigmoid cases can be observed. Similar observations can be made for SROCC, where 9 out of 16 videos (1, 4, 5, 7-9, 11-13) show improvements from clustering. Once again, videos 1, 5 and 12 seem to benefit the most from this addition with respective improvements of 0.141, 0.076 and 0.068, respectively. Video 14, on the downside, shows to suffer from a relevant decrease of 0.138 SROCC. On average, however, a limited improvement of 0.013 can be seen. When it comes to RMSE, seven videos (1, 5, 7, 10-12, 14) are positively impacted by applying the clustering algorithm. Here, videos 1, 12 and 14 seem to experience the greatest impact with respective improvements of 1.621, 0.578 and 0.691 in the non-sigmoid case and 1.154, 0.603 and 0.646 and the sigmoid case. Videos 2, 9 and 15 seem to show the largest decreases in performance due to clustering, with respective RMSE increases of 1.205, 0.625 and 0.528 in the non-sigmoid case and 1.304, 0.528 and 0.447 in the sigmoid case. On average, however, small decreases can be observed of 0.023 in the non-sigmoid case and 0.005 if sigmoidal mapping is applied.

The influence of the latter translates to an improvement for the majority of the videos. In terms of PLCC, only videos 6, 10 and 16 show a reduction with respective differences of 0.071, 0.085 and 0.001 in the non-clustering case and 0.071, 0.077 and 0.010 in the clustering case. On average, sigmoidal mapping results in respective improvements of 0.017 and 0.20 in PLCC. Looking at RMSE, similar conclusions can be drawn, with that difference that also videos 1–3 are decreasing in performance for both the non-clustering and the clustering case. Especially for videos 2, 6 and 10 we see relevant increases in RMSE with respective differences of 2.210, 1.077 and 0.656 in the non-clustering case and 2.309, 0.899 and 0.470 in the clustering case. On average, sigmoidal mapping shows small decreases in performance with respective differences of 0.139 and 0.157 in RMSE. Note that for video 6, a much lower performance can once again be observed due to the difference in NRT and SI distributions as explained earlier in Sect. 5.4.3 (Table 9).

Table 10 once again shows the results of optimizing the model parameters for each video separately rather than on a per-cluster base. Similar as in the GOP-level case, sigmoidal mapping is clearly showing its benefit. Only videos 14 and 16 are showing marginal increases in RMSE of 0.025 and 0.002, respectively. When comparing to the one-for-all approach, a mixed view arises. Just below half of the videos (1, 4, 7, 11-14) show a decrease in PLCC for both the sigmoid and the non-sigmoid case when comparing to their best performing counterpart (either with or without clustering) in the one-for-all approach. This is especially true for videos 4, 7, 12 and 13, which show respective decreases of 0.265, 0.083, 0.212 and 0.145 in the non-sigmoidal case and 0.285, 0.107, 0.246 and 0.164 when sigmoidal mapping is applied. Videos 3, 6 and 10, on the contrary, show to benefit from a per-video approach with non-sigmoidal increases of 0.115, 0.325 and 0.129; and 0.129, 0.417 and 0.208 in the sigmoidal case. On average, almost no difference is observed with a 0.001 increase for the non-sigmoid case and a 0.006 decrease

Table 10 PLSSs, SROCCs and RMSEs towards VMAF of frame-level models trained on a per-video base. A comparison of the cases with and without sigmoidal mapping is provided. For each video and each case, the best performing metric (highest for PLCC/SROCC and lowest for RMSE) is indicated in bold

Video	No sigmoidal mapping			Sigmoidal mapping		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
1	0.739	**0.762**	2.341	**0.764**	**0.762**	**2.244**
2	0.828	**0.848**	4.132	**0.849**	**0.848**	**3.845**
3	0.706	**0.727**	6.759	**0.722**	**0.727**	**6.577**
4	0.401	**0.505**	2.902	**0.422**	**0.505**	**2.845**
5	0.572	**0.590**	4.980	**0.598**	**0.590**	**4.862**
6	0.508	**0.479**	5.629	**0.529**	**0.479**	**5.536**
7	0.608	**0.680**	3.437	**0.632**	**0.680**	**3.378**
8	0.685	**0.691**	4.215	**0.748**	**0.691**	**3.934**
9	0.592	**0.695**	4.064	**0.648**	**0.695**	**3.861**
10	0.637	**0.715**	8.659	**0.639**	**0.715**	**8.640**
11	0.584	**0.576**	4.836	**0.630**	**0.576**	**4.669**
12	0.424	**0.469**	5.088	**0.448**	**0.469**	**5.015**
13	0.554	**0.640**	2.312	**0.585**	**0.640**	**2.262**
14	0.529	**0.667**	**3.373**	**0.542**	**0.667**	3.398
15	0.683	**0.687**	3.299	**0.726**	**0.687**	**3.140**
16	0.497	**0.507**	**7.446**	**0.506**	**0.507**	7.448
Avg.	0.597	**0.639**	4.592	**0.624**	**0.639**	**4.478**

Table 11 Average PLCCs, SROCCs, and RMSEs of the six NR-metrics, PSNR, and SSIM to the VMAF benchmark at frame-level. The best performing metric (highest for PLCC/SROCC and lowest for RMSE) is indicated in bold. Note that a min–max scaler was applied to NOI, BLU, SI and PSNR to make them fall within the [0, 100] interval

Metric	PLCC	SROCC	RMSE
NOI	− 0.199	− 0.226	18.639
NRT	− 0.187	− 0.279	76.707
BLU	0.142	0.221	76.252
BRT	− 0.311	− 0.240	72.030
BLK	0.005	0.011	10.728
SI	− 0.004	− 0.141	45.967
PSNR	0.673	0.689	30.677
SSIM	**0.820**	**0.718**	16.272
Proposed	0.618	0.574	**6.047**

in the sigmoid case. In terms of SROCC, most videos seem to show increased performance when using per-video models, as is the case for e.g., videos 3, 5, 6, 8 and 9. These show respective improvements of 0.165, 0.120, 0.152, 0.142 and 0.201. Only videos 4,7, 12, 13 and 15 show limited decreases of 0.009, 0.028, 0.011, 0.012 and 0.039, respectively. On average, a small decrease of 0.065 can be observed w.r.t. the one-for-all model. In terms of RMSE, we see improvements for almost all videos, where video 10 is especially worth mentioning with respective RMSE decreases of 4.931 in the non-sigmoidal case and 5.420 in the sigmoidal case. Video 12 is the only video showing an RMSE increase in

the non-sigmoid case with a difference of 0.336. In the sigmoidal case, limited decreases in performance can be observed for videos 4, 7 and 11 with respective RMSE increases of 0.116, 0.095 and 0.091. On average, RMSEs show a reduction of 1.298 in the non-sigmoid case and 1.569 in the sigmoid case.

In Table 11, the average performance of the proposed method (including clustering and sigmoidal mapping) in terms of PLCC, SROCC, and RMSE to VMAF is depicted in comparison to single-feature prediction by each of the NR metrics as well as the PSNR and SSIM FR benchmarks. One can notice that also at frame-level the proposed method is clearly improving upon each of the NR metrics in terms of correlation values. Also in terms of RMSE a clear improvement can be noticed, with once again only BLK being able to stay in the neighborhood with an obtained RMSE of 10.728 compared to 6.047 for the proposed method. Once again, this RMSE shows to be better that what PSNR and SSIM can obtain with respective RMSEs of 30.677 and 16.272. In terms of PLCC and SROCC the proposed method is only little below the performance of PSNR. Only for SSIM a clear difference can still be noticed. Nevertheless, the obtained performance can be considered satisfying given the NR nature of the proposed model compared to the FR implementations of both PSNR and SSIM.

6 Discussion

Section 5.4 has discussed the most prominent results on each of the three granularity levels. At the video-level, the beneficial influence of clustering combined with a per-cluster sigmoidal mapping is clear. As such, it can easily be applied as a generic prediction mechanism for video-level quality, especially since it predictions match the performance of the ground-truth VMAF in terms of correlation and RMSE to (DSIS) MOS. For most real-time applications, especially when live-streaming is considered, more fine-grained predictions are preferred in order to accurately act upon and end-user drop in QoE. While the sigmoidal mapping aspect of the proposed model is still showing beneficial for the majority of the videos on a GOP- and frame-level, the influence of clustering shows a more mixed view. Especially for certain specific videos such as 3, 6 and 10, a per-video model proves to be much more beneficial than a one-for-all approach. From this observation, one could interpret the clustering algorithm as an outlier identifying strategy. By excluding these outliers from the input data, the non-outlying videos (which still make up the majority of the data) could be modeled using one-for-all model based on sigmoidal mapping. The minority of videos identified as outlying could then be provided with their own personal model. As a result, this would provide a good trade-off between maximizing performance on one hand and limiting computational and practical overhead on the other hand, especially since filtering out outlying videos would help further fine-tuning the one-for-all model. It has to be mentioned, however, that videos such as video 6 that show outlying performance are not always identified as a separate cluster by the model yet. As such further improvement of this algorithm to transform it to an outlier detector is desirable. Exploring additional NR metrics, possibly on the temporal part such as TI, MI or jerkiness could play a role in this. Giving the easy adaptation of the Video Metric Tool presented in Sect. 4 and the fact that the implementations of TI, MI and jerkiness are already provided, this is certainly within reach.

Furthermore, one could state that the results for the majority of the videos, although lower than on video-level, are satisfying with correlations often going above 0.700 while maintaining RMSEs below 10 on a 0–100 scale. Giving the limited complexity of the modeling approach, this opens opportunities towards both further generalization as well as specification for specific types of content by appropriately updating or replacing the clustering prior to the sigmoidal mapping. In case higher per-GOP or per-frame accuracy is required, one can still opt to steer towards more complex modeling approaches based on ML/DL to exploit any further relationships in the data that are not revealed by the psychometric mapping. Nevertheless, when comparing the performance of the proposed method to both th individual NR metrics and alternative FR benchmarks, it is shown that satisfying results are obtained, especially in terms of RMSE. Although correlation values remain somewhat below the performance of the FR metrics, results are satisfying and promising given the NR nature of the presented approach. All with all, it can thus be concluded that a decent methodology and baseline towards further research has been provided, and that further improvements to the presented approach are also within reach.

7 Conclusions and future work

This work has presented an NR quality model for point cloud video, which consists of a combination of KMeans clustering and sigmoidal mapping. Video-, GOP-, and frame-level granularity were considered. The applicability of the model to this granularities is investigated and analysis is performed to the cases in which a per-video model is preferred. In addition, a CLI *Video Metric Tool* is presented that allows for easy and modular calculation of multiple NR-metrics on a given video. Results show that the sigmoidal mapping aspect of the model shows its value across all granularity levels. The clustering algorithm results in a more mixed view. However, the latter would be valuable in the role of an outlier detector to select videos which would benefit from a specific video model, as well as further improving the given model by delineating the most predictable data for a one-for-all approach. Nevertheless, satisfying results are yet obtained with correlation values often going above 0.700 on GOPs- and frame-level while maintaining RMSEs below 10 on a 0–100 scale.

Especially in terms of RMSE, satisfying results are obtained with the proposed model outperforming FR benchmarks. Although correlation values remain somewhat below the performance of these same benchmarks, results are satisfying and promising given the NR nature of the proposed method.

It needs to be mentioned, though, that the presented analysis is limited to a single dataset of point cloud videos. Therefore, future work includes the validation on additional projection-based datasets, preferably with other point cloud objects (e.g., from the *V-SENSE* dataset [33]). Furthermore, it is also worth noticing that the included encoding distortions are limited to V-PCC encoding. As such, the exploration of other encoding mechanisms such as G-PCC would further add to the generalizability of the presented model.

A Mathematical definition of the NR metrics

This Appendix gives the mathematical definitions of each NR metric on a single frame F with size $m \times n$. As stated before, the corresponding value for a GOP or video can be calculated by averaging the the per-frame results except for SI, where the maximum is taken.

A.1 Noise (NOI) and noise ratio (NRT)

This Appendix mathematically defines noise (NOI) and noise ratio (NRT) based on the implementation of Choi et al. [61].

$k \leftarrow 3$
$c \leftarrow 0$
$F_{\text{denoised}} \leftarrow \text{averageFilter}_{k \times k}(F)$
▷ Average filter with kernel size $k \times k$
$dF_x \leftarrow \partial F_{\text{denoised}} / \partial x$
▷ Derivative in x-direction
$dF_y \leftarrow \partial F_{\text{denoised}} / \partial y$
▷ Derivative in y-direction
$dF \leftarrow \text{elementWiseMax}(dF_x, dF_y)$
$\mu_x \leftarrow d\bar{F}_x$
▷ Average of dF_x
$\mu_y \leftarrow d\bar{F}_y$
▷ Average of dF_y
for $i = 0..m-1$, $j = 0..n-1$ **do**
▷ Iterate over all pixels
  if $dF_x(i,j) > \mu_x$ **and** $dF_y(i,j) > \mu_y$ **then**
    $dF(i,j) \leftarrow 0$
  end if
end for
$\mu \leftarrow d\bar{F}$
▷ Average of dF
for $i = 0..m-1$, $j = 0..n-1$ **do**
▷ Iterate over all pixels
  if $dF(i,j) \leq \mu$ **then**
    $dF(i,j) \leftarrow 0$
  else
    $c \leftarrow c + 1$
  end if
end for
$\text{NOI} \leftarrow \frac{1}{c} \sum_{i,j} dF$
$\text{NRT} \leftarrow \frac{c}{m \cdot n}$

A.2 Blur (BLU) and blur ratio (BRT)

This Appendix mathematically defines blur (BLU) and blur ratio (BRT) based on the implementation of Choi et al. [61].

$dF_x \leftarrow \partial \mathrm{F}/\partial x$
$dF_y \leftarrow \partial \mathrm{F}/\partial y$
$\mu_x \leftarrow d\bar{F}_x$
$\mu_y \leftarrow d\bar{F}_y$
$E_x \leftarrow O_{m\times n}$
$E_y \leftarrow O_{m\times n}$
▷ $m \times n$ matrices with zeros
$B_x \leftarrow J_{m\times n}$
$B_y \leftarrow J_{m\times n}$
▷ $m \times n$ matrices with ones
$t \leftarrow 0.1$
$c \leftarrow 0$
for $i = 0..m-1$, $j = 0..n-1$ **do**
  if $dF_x(i,j) \leq \mu_x$ **then**
    $dF_x(i,j) \leftarrow 0$
  end if
  if $dF_y(i,j) \leq \mu_y$ **then**
    $dF_y(i,j) \leftarrow 0$
  end if
  if $i > 0$ **and** $i < m-1$ **and** $dF_x(i,j) > dF_x(i-1,j)$ **and** $dF_x(i,j) > dF_x(i+1,j)$ **then**
    $E_x(i,j) = 1$
  end if
  if $j > 0$ **and** $j < n-1$ **and** $dF_y(i,j) > dF_y(i,j-1)$ **and** $dF_y(i,j) > dF_y(i,j+1)$ **then**
    $E_y(i,j) = 1$
  end if
  if $dF_x(i,j) = 0$ **then**
    $B_x(i,j) \leftarrow 1$
  else
    $B_x(i,j) \leftarrow \frac{1}{dF_x(i,j)}|2F(i,j) - dF_x(i,j)|$
  end if
  if $dF_y(i,j) = 0$ **then**
    $B_y(i,j) \leftarrow 1$
  else
    $B_y(i,j) \leftarrow \frac{1}{dF_y(i,j)}|2F(i,j) - dF_y(i,j)|$
  end if
end for
$E \leftarrow E_x \odot E_y$
▷ Element-wise multiplication
$B \leftarrow \mathrm{elementWiseMax}(B_x, B_y)$
for $i = 0..m-1$, $j = 0..n-1$ **do**
  if $B(i,j) < t$ **then**
    $c \leftarrow c + 1$
  end if
end for
$\mathrm{BLU} \leftarrow \frac{1}{c}\sum_{i,j} B$
$\mathrm{BRT} \leftarrow \frac{c}{\sum_{i,j} E}$

A.3 Blockiness (BLK)

This Appendix mathematically defines blockiness (BLK) based on the implementation of Perra et al. [62].

$c \leftarrow 0$
$\mathrm{BLK} \leftarrow 0$
$b \leftarrow 8$
▷ Block size
$k \leftarrow 2.3$
$dF_{\mathrm{Sobel,x}} \leftarrow \mathrm{Sobel}_x(F)$
▷ Sobel in x-direction
$dF_{\mathrm{Sobel,y}} \leftarrow \mathrm{Sobel}_y(F)$
▷ Sobel in y-direction
$dF_{\mathrm{Sobel}} \leftarrow \mathrm{Sobel}(F)$
▷ Regular Sobel
$m_x \leftarrow \max(dF_{\mathrm{Sobel,x}})$
$m_y \leftarrow \max(dF_{\mathrm{Sobel,y}})$
$m_I \leftarrow \max(dF_{\mathrm{Sobel}})$
for $i = 0..\lfloor \frac{m}{b} \rfloor + 1, j = 0..\lfloor \frac{n}{b} \rfloor + 1$ **do**
▷ Iterate over all blocks
$b_i = i \cdot b$
$b_j = j \cdot b$
if $b_i < m - b$ **and** $b_j < n - b$ **then**
$B_x \leftarrow dF_{\mathrm{Sobel,x}}(b_i..b_i + b, b_j..b_j + b)$
$B_y \leftarrow dF_{\mathrm{Sobel,y}}(b_i..b_i + b, b_j..b_j + b)$
$I \leftarrow dF_{\mathrm{Sobel}}(b_i..b_i + b, b_j..b_j + b)$
▷ For each matrix, select block with (b_i, b_j) as upper left indices
$s_x \leftarrow \frac{1}{2bm_x} \sum_j (|B_x(:, 0)| + |B_x(:, b-1)|)$
▷ Left and right edges
$s_y \leftarrow \frac{1}{2bm_y} \sum_i (|B_y(0, :)| + |B_y(b-1, :)|)$
▷ Upper and lower edges
$s \leftarrow \max(s_x, s_y)$
▷ Inner edges
$s_I \leftarrow \frac{1}{4(b-3)m_I} (\sum_i (|I(1, 1..b-2)| + |I(b-2, 1..b-2)|) + \sum_j (|I(2..b-3, 1)| + |I(2..b-3, b-2)|))$
if $s \neq 0$ **or** $s_I \neq 0$ **then**
$\mathrm{BLK} \leftarrow \mathrm{BLK} + 2\frac{|s^k - s_I^k|}{|s^k + s_I^k|}$
$c \leftarrow c + 1$
end if
end if
end for
$\mathrm{BLK} \leftarrow \frac{\mathrm{BLK}}{c}$

A.4 Spatial information (SI)

This Appendix mathematically defines spatial information (SI) based on the definition by the International Telecommunication Union (ITU) [56].

$$\mathrm{SI} \leftarrow \sigma(\mathrm{Sobel}(F))$$
$$\triangleright \text{ with } \sigma \text{ the standard deviation}$$

Abbreviations

ACR	Absolute Category Rating
CLI	Command Line Interface
CNN	Convolutional Neural Network
DISTS	Deep Image Structure and Texture Similarity
DL	Deep Learning
DSIS	Double Stimulus Impairment Scale
FoV	Field-of-View
FR	Full Reference
FSIM	Feature Similarity Indexing Method
FWO	Research Foundation Flanders
GOP	Group-of-Pictures
HVS	Human Visual System
ITS	Institute for Telecommunication Sciences
ITU	International Telecommunication Union
IW-SSIM	Information content Weighted SSIM
JS	Jensen–Shannon
KRCC	Kendall Rank Correlation Coefficient
LPIPS	Learned Perceptual Image Patch Similarity
LP-PCQM	Layered Projection-based Point Cloud Quality Metric
MI	Motion Intensity
ML	Machine Learning
MMI	Mean Motion Intensity
MOS	Mean Opinion Score
MS-SSIM	Multi-Scale SSIM
MSE	Mean Squared Error
NR	No Reference
NTIA	National Telecommunications and Information Administration
PCQM	Point-Centered Quarter Method
PLCC	Pearson Linear Correlation Coefficient
PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
QoS	Quality of Service
RBF	Radial Basis Function
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
RR	Reduced Reference
SI	Spatial Information
SROCC	Spearman Rank Order Correlation Coefficient
SS	Silhouette Score
SSIM	Structural Similarity Index Measure
SVR	Support Vector Regression
TI	Temporal Information
VLAIO	Flanders Innovation and Entrepreneurship
V-PCC	Video-based Point Cloud Compression
VMAF	Video Multimethod Assessment Fusion
VQM	Video Quality Metric

Acknowledgements
Not applicable.

Author contributions
SVD and MTV provided the first conceptualization, research direction and envisioned QoE management architecture. JVDH created the multiple point cloud compression representations, implemented the network emulator and generated the set of point cloud videos. In addition, JVDH leaded the subjective experiments for the partial MOS annotation of the dataset. SVD implemented and calculated NR and VMAF scores for the given video sequences and the created the accompanying ***Video Metric Tool***. SVD also proposed and evaluated the modeling approach and its corresponding analysis and interpretation. SVD provided the drafting of the manuscript, while JVDH, MTV and FDT thoroughly revised and provided feedback. All authors read and approved the final manuscript.

Funding
Part of this research was funded by the ICON project INTERACT, realized in collaboration with imec, with project support from Flanders Innovation and Entrepreneurship (VLAIO). Project partners are imec, Rhinox, Pharrowtech, Dekimo and

TEO. Sam Van Damme and Jeroen van der Hooft are funded by the Research Foundation Flanders (FWO), grant numbers 1SB1822N and 1281021N, respectively. This research is partially funded by the FWO WaveVR project, grant number G034322N.

Data availability

The ***Video Metric Tool***, presented in Sect. 4, is available at the ***VideoMetricTool*** repository to be found at https://github.ugent.be/samdamme/VideoMetricTool.The set of source videos analyzed during the current study, as well as its accompanying ***.csv***-dataset are available via https://cloud.ilabt.imec.be/index.php/s/Dx2rRfQy3CwFM5o. All other data generated or analyzed during this study are included in this published article.

Declarations

Competing interests

The authors declare that they have no Competing interests.

Received: 31 January 2024 Accepted: 17 October 2024
Published online: 19 November 2024

References

1. J. van der Hooft, M. Torres Vega, C. Timmerer, A.C. Begen, F. De Turck, R. Schatz, Objective and Subjective QoE Evaluation for Adaptive Point Cloud Streaming, in *International conference on quality of multimedia experience*. 2020. https://doi.org/10.1109/QoMEX48832.2020.9123081
2. A. Javaheri, C. Brites, F. Pereira, J. Ascenso, Improving PSNR-Based Quality Metrics Performance For Point Cloud Geometry, in *2020 IEEE international conference on image processing (ICIP)*. 2020, pp. 3438–3442. https://doi.org/10.1109/ICIP40778.2020.9191233
3. E. Alexiou, T. Ebrahimi, On subjective and objective quality evaluation of point cloud geometry, in *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. 2017, pp. 1–3. https://doi.org/10.1109/QoMEX.2017.7965681
4. I. Viola, S. Subramanyam, P. Cesar, A Color-Based Objective Quality Metric for Point Cloud Contents, in *International conference on quality of multimedia experience*. 2020. https://doi.org/10.1109/QoMEX48832.2020.9123089
5. S. Van Damme, M. Torres Vega, J. Heyse, F. De Backere, F. De Turck, A low-complexity psychometric curve-fitting approach for the objective quality assessment of streamed game videos. Signal processing: image communication 88. 2020. https://doi.org/10.1016/j.image.2020.115954
6. L.A. da Silva Cruz, E. Dumić, E. Alexiou, J. Prazeres, R. Duarte, M. Pereira, A. Pinheiro, 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. https://doi.org/10.1109/QoMEX.2019.8743258
7. Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, J. Sun, Predicting the perceptual quality of point cloud: A 3D-to-2D projection-based exploration. IEEE transactions on multimedia. 2020. https://doi.org/10.1109/TMM.2020.3033117
8. S. Van Damme, M. Torres Vega, F. De Turck, A Full- and No-Reference Metrics Accuracy Analysis for Volumetric Media Streaming, in *International conference on quality of multimedia experience*. 2021. https://doi.org/10.1109/QoMEX51781.2021.9465420
9. M. Wien, J. Jung, V. Baroncini, Formal Visual Evaluation and Study of Objective Metrics for MPEG Dynamic Mesh Coding, in *2022 10th european workshop on visual information processing (EUVIP)*. 2022, pp. 1–6. https://doi.org/10.1109/EUVIP53989.2022.9922894
10. K. Yang, Q. Yang, J. Jung, Y. Xu, X. Xu, S. Liu, Exploring the Influence of View and Camera Path Selection for Dynamic Mesh Quality Assessment, in *2023 IEEE international conference on multimedia and expo (ICME)*. 2023, pp. 2489–2494. https://doi.org/10.1109/ICME55011.2023.00424
11. M. Rudolph, S. Schneegass, A. Rizk, RABBIT: Live Transcoding of V-PCC Point Cloud Streams, in *Proceedings of the 14th conference on ACM multimedia systems* (Association for Computing Machinery, New York, NY, USA, 2023), MMSys '23, p. 97–107. https://doi.org/10.1145/3587819.3590978
12. Y. Shi, P. Venkatram, Y. Ding, W.T. Ooi, Enabling Low Bit-Rate MPEG V-PCC-encoded Volumetric Video Streaming with 3D Sub-sampling, in *Proceedings of the 14th conference on ACM multimedia systems* (Association for Computing Machinery, New York, NY, USA, 2023), MMSys '23, p. 108–118. https://doi.org/10.1145/3587819.3590981
13. S. Van Damme, M. Torres Vega, J. van der Hooft, F. De Turck, Clustering-Based Psychometric No-Reference Quality Model for Point Cloud Video, in *2022 IEEE International conference on image processing (ICIP)*. 2022, pp. 1866–1870. https://doi.org/10.1109/ICIP46576.2022.9897602
14. E. Alexiou, T. Ebrahimi, Point Cloud Quality Assessment Metric Based on Angular Similarity, in *2018 IEEE international conference on multimedia and expo (ICME)*. 2018, pp. 1–6. https://doi.org/10.1109/ICME.2018.8486512
15. E. Alexiou, T. Ebrahimi, Towards a Point Cloud Structural Similarity Metric, in *IEEE International conference on multimedia expo workshops*. 2020. https://doi.org/10.1109/ICMEW46912.2020.9106005
16. R. Diniz, P.G. Freitas, M.C.Q. Farias, Towards a Point Cloud Quality Assessment Model using Local Binary Patterns, in *International conference on quality of multimedia experience*. 2020. https://doi.org/10.1109/QoMEX48832.2020.9123076
17. R. Diniz, P.G. Freitas, M.C.Q. Farias, Color and geometry texture descriptors for point-cloud quality assessment. IEEE Signal Process Lett 28, 1150–1154 (2021). https://doi.org/10.1109/LSP.2021.3088059
18. I. Viola, P. Cesar, A reduced reference metric for visual quality evaluation of point cloud contents. IEEE Signal Process Lett (2020). https://doi.org/10.1109/LSP.2020.3024065

19. Y. Nehmé, F. Dupont, J. Farrugia, P. Le Callet, G. Lavoué, Visual quality of 3D meshes with diffuse colors in virtual reality: subjective and objective evaluation. IEEE Trans Vis Comput Gr 27(3), 2202–2219 (2021). https://doi.org/10.1109/TVCG.2020.3036153
20. D. Tian, H. Ochimizu, C. Feng, R. Cohen, A. Vetro, Geometric distortion metrics for point cloud compression, in *2017 IEEE international conference on image processing (ICIP)*. 2017, pp. 3460–3464. https://doi.org/10.1109/ICIP.2017.8296925
21. A. Javaheri, C. Brites, F. Pereira, J. Ascenso, A Generalized Hausdorff Distance Based Quality Metric for Point Cloud Geometry, in *2020 Twelfth international conference on quality of multimedia experience (QoMEX)*. 2020, pp. 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123087
22. A. Javaheri, C. Brites, F. Pereira, J. Ascenso, Mahalanobis based point to distribution metric for point cloud geometry quality evaluation. IEEE Signal Process Lett 27, 1350–1354 (2020). https://doi.org/10.1109/LSP.2020.3010128
23. Q. Liu, H. Yuan, R. Hamzaoui, H. Su, J. Hou, H. Yang, Reduced reference perceptual quality model with application to rate control for video-based point cloud compression. IEEE Trans Image Process 30, 6623–6636 (2021). https://doi.org/10.1109/TIP.2021.3096060
24. T. Chen, C. Long, H. Su, L. Chen, J. Chi, Z. Pan, H. Yang, Y. Liu, Layered projection-based quality assessment of 3d point clouds. IEEE Access 9, 88108–88120 (2021). https://doi.org/10.1109/ACCESS.2021.3087183
25. R. Shafi, W. Shuai, M.U. Younus, 360-degree video streaming: a survey of the state of the art. Symmetry 12(9), 1491 (2020). https://doi.org/10.3390/sym12091491
26. F. Chiariotti, A survey on 360-degree video: coding, quality of experience and streaming. Comput Res Repos 177, 133–155 (2021). https://doi.org/10.1016/j.comcom.2021.06.029
27. J. Ruan, D. Xie, A survey on QoE-oriented VR video streaming: some research issues and challenges. Electronics 10(17), 2155 (2021). https://doi.org/10.3390/electronics10172155
28. E. Alexiou, T. Ebrahimi, Exploiting user interactivity in quality assessment of point cloud imaging, in *2019 Eleventh international conference on quality of multimedia experience (QoMEX)*. 2019, pp. 1–6. https://doi.org/10.1109/QoMEX.2019.8743277
29. Q. Yang, Z. Ma, Y. Xu, Z. Li, J. Sun, Inferring point cloud quality via graph similarity. IEEE Trans Pattern Anal Mach Intell 44(6), 3015–3029 (2020). https://doi.org/10.1109/TPAMI.2020.3047083
30. E.M. Torlig, E. Alexiou, T.A. Fonseca, R.L. de Queiroz, T. Ebrahimi, A novel methodology for quality assessment of voxelized point clouds. Appl Digit Image Process XLI 10752, 107520I (2018). https://doi.org/10.1117/12.2322741
31. M. Yang, D. Wu, Z. Wang, M. Hu, Y. Zhou, Understanding and Improving Perceptual Quality of Volumetric Video Streaming, in *2023 IEEE international conference on multimedia and expo (ICME)*. 2023, pp. 1979–1984. https://doi.org/10.1109/ICME55011.2023.00339
32. Y. Fan, Z. Zhang, W. Sun, X. Min, J. Lin, G. Zhai, N. Liu, MV-VVQA: Multi-View Learning for No-Reference Volumetric Video Quality Assessment, in *2023 31st European signal processing conference (EUSIPCO)*. 2023, pp. 670–674. https://doi.org/10.23919/EUSIPCO58844.2023.10290018
33. E. Zerman, P. Gao, C. Ozcinar, A. Smolic, Subjective and Objective Quality Assessment for Volumetric Video Compression in *Proc. IS &T Int'l. Symp. on electronic imaging: image quality and system performance XVI*. 2019, pp. 323–1–323–7. https://doi.org/10.2352/ISSN.2470-1173.2019.10.IQSP-323
34. T. Wang, F. Li, P.C. Cosman, Learning-based rate control for video-based point cloud compression. IEEE Trans Image Process 31, 2175–2189 (2022). https://doi.org/10.1109/TIP.2022.3152065
35. L. Li, Z. Li, S. Liu, H. Li, Rate control for video-based point cloud compression. IEEE Trans Image Process 29, 6237–6250 (2020). https://doi.org/10.1109/TIP.2020.2989576
36. F. Shen, W. Gao, A Rate Control Algorithm for Video-based Point Cloud Compression, in *2021 International conference on visual communications and image processing (VCIP)*. 2021, pp. 1–5. https://doi.org/10.1109/VCIP53242.2021.9675449
37. L. Wang, C. Li, W. Dai, S. Li, J. Zou, H. Xiong, Qoe-driven adaptive streaming for point clouds. IEEE Trans Multimed (2022). https://doi.org/10.1109/TMM.2022.3148585
38. C. Zhang, Y. Cao, Z. Liu, R. Yin, Y. Zhu, X. Chen, Trans-RL: A Prediction-Control Approach for QoE-Aware Point Cloud Video Streaming, in *GLOBECOM 2022 - 2022 IEEE global communications conference*. 2022, pp. 1899–1904. https://doi.org/10.1109/GLOBECOM48099.2022.10001399
39. H. Lin, B. Zhang, Y. Cao, Z. Liu, X. Chen, A Deep Reinforcement Learning Approach for Point Cloud Video Transmissions, in *2021 IEEE 94th vehicular technology conference (VTC2021-Fall)*. 2021, pp. 1–5. https://doi.org/10.1109/VTC2021-Fall52928.2021.9625496
40. C.H. Wu, X. Li, R. Rajesh, W.T. Ooi, C.H. Hsu, Dynamic 3D point cloud streaming: distortion and concealment, in *Proceedings of the 31st ACM Workshop on Network and Operating Systems Support for Digital Audio and Video* (Association for Computing Machinery, New York, NY, USA, 2021), NOSSDAV '21, p. 98–105. https://doi.org/10.1145/3458306.3458876
41. C.F. Santos, F. Lopes, A. Pinheiro, L.A. da Silva Cruz, A Sub-Partitioning Method for Point Cloud Inter-prediction Coding, in *2018 IEEE visual communications and image processing (VCIP)*. 2018, pp. 1–4. https://doi.org/10.1109/VCIP.2018.8698661
42. M. Ghosh, D.C. Singhal, R. Wayal, DeSVQ: Deep Learning Based Streaming Video QoE Estimation, in *Proceedings of the 23rd international conference on distributed computing and networking* (Association for Computing Machinery, New York, NY, USA, 2022), ICDCN '22, p. 19–25. https://doi.org/10.1145/3491003.3491023
43. C.G. Bampis, Z. Li, I. Katsavounidis, A.C. Bovik, Recurrent and dynamic models for predicting streaming video quality of experience. IEEE Trans Image Process 27(7), 3316–3331 (2018). https://doi.org/10.1109/TIP.2018.2815842
44. C. Chen, L.K. Choi, G. de Veciana, C. Caramanis, R.W. Heath, A.C. Bovik, Modeling the time—varying subjective quality of http video streams with rate adaptations. IEEE Trans Image Process 23(5), 2206–2221 (2014). https://doi.org/10.1109/TIP.2014.2312613
45. T.N. Duc, C.T. Minh, T.P. Xuan, E. Kamioka, Convolutional neural networks for continuous qoe prediction in video streaming services. IEEE Access 8, 116268–116278 (2020). https://doi.org/10.1109/ACCESS.2020.3004125

46. N. Eswara, S. Ashique, A. Panchbhai, S. Chakraborty, H.P. Sethuram, K. Kuchi, A. Kumar, S.S. Channappayya, Streaming video qoe modeling and prediction: a long short-term memory approach. IEEE Trans Circuits Syst Video Technol 30(3), 661–673 (2020). https://doi.org/10.1109/TCSVT.2019.2895223
47. N. Eswara, K. Manasa, A. Kommineni, S. Chakraborty, H.P. Sethuram, K. Kuchi, A. Kumar, S.S. Channappayya, A continuous qoe evaluation framework for video streaming over http. IEEE Trans Circuits Syst Video Technol 28(11), 3236–3250 (2018). https://doi.org/10.1109/TCSVT.2017.2742601
48. L. Ma, T. Xu, G. Sternberg, A. Balasubramanian, A. Zeira, Model-based QoE prediction to enable better user experience for video teleconferencing, in *2013 IEEE international conference on acoustics, speech and signal processing*. 2013, pp. 2815–2819. https://doi.org/10.1109/ICASSP.2013.6638170
49. P. Chen, L. Li, Y. Huang, F. Tan, W. Chen, QoE Evaluation for Live Broadcasting Video, in *2019 IEEE international conference on image processing (ICIP)*. 2019, pp. 454–458. https://doi.org/10.1109/ICIP.2019.8802978
50. H. Chen, X. Zhang, Y. Xu, J. Ren, J. Fan, Z. Ma, W. Zhang, T-gaming: a cost-efficient cloud gaming system at scale. IEEE Trans Parallel Distrib Syst 30(12), 2849–2865 (2019). https://doi.org/10.1109/TPDS.2019.2922205
51. A. Ak, E. Zerman, M. Quach, A. Chetouani, A. Smolic, G. Valenzise, P. Le Callet, Basics: broad quality assessment of static point clouds in a compression scenario. IEEE Trans Multimed (2024). https://doi.org/10.1109/TMM.2024.3355642
52. D. Lazzarotto, M. Testolina, T. Ebrahimi, Assessing objective quality metrics for jpeg and mpeg point cloud coding. arXiv preprint arXiv:2403.00410. 2024.
53. Q. Yang, J. Jung, H. Wang, X. Xu, S. Liu, TSMD: A Database for Static Color Mesh Quality Assessment Study, in *2023 IEEE international conference on visual communications and image processing (VCIP)*. 2023, pp. 1–5. https://doi.org/10.1109/VCIP59821.2023.10402660
54. B. Cui, Q. Yang, K. Yang, Y. Xu, X. Xu, S. Liu, SJTU-TMQA: A Quality Assessment Database for Static Mesh with Texture Map, in *ICASSP 2024 - 2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2024, pp. 7875–7879. https://doi.org/10.1109/ICASSP48485.2024.10445942
55. M. Torres Vega, C. Perra, F. De Turck, A. Liotta, A review of predictive quality of experience management in video streaming services. IEEE Trans Broadcasting 64(2), 432–445 (2018). https://doi.org/10.1109/TBC.2018.2822869
56. International Telecommunication Union (ITU). ITU-T Rec. P.910: Subjective video quality assessment methods for multimedia applications. 2008
57. S. Borer, A model of jerkiness for temporal impairments in video transmission, in *2010 Second international workshop on quality of multimedia experience (QoMEX)*. 2010, pp. 218–223. https://doi.org/10.1109/QOMEX.2010.5516155
58. M. Shahid, A. Rossholm, B. Lövström et al., No-reference image and video quality assessment: a classification and review of recent approaches. J Image Video Process 2014(40), 2014 (2014). https://doi.org/10.1186/1687-5281-2014-40
59. M. Torres Vega, D.C. Mocanu, A. Liotta, Unsupervised deep learning for real-time assessment of video streaming services. Multimed Tools Appl 76, 22303–22327 (2017). https://doi.org/10.1007/s11042-017-4831-6
60. M. Torres Vega, D.C. Mocanu, S. Stavrou, A. Liotta, Predictive no-reference assessment of video quality. Signal Process: Image Commun 52, 20–32 (2017). https://doi.org/10.1016/j.image.2016.12.001
61. M.G. Choi, J.H. Jung, J.W. Jeon, No-reference image quality assessment using blur and noise. Int J Comput Sci Eng 3(2), 76–80 (2009)
62. C. Perra, A low computational complexity blockiness estimation based on spatial analysis, in *2014 22nd Telecommunications Forum Telfor (TELFOR)*. 2014, pp. 1130–113https://doi.org/10.1109/TELFOR.2014.7034606
63. A. Aaron, Z. Li, M. Manohara, J.Y. Lin, E.C. Wu, C. Kuo, Challenges in Cloud Based Ingest and Encoding for High Quality Streaming Media, in *IEEE International Conference on Image Processing*. 2015. https://doi.org/10.1109/ICIP.2015.7351097
64. Institute for Telecommunication Sciences (ITS). Video Quality Metric (VQM). https://its.ntia.gov/research-topics/video-quality-research/guides-and-tutorials/description-of-vqm-tools.aspx. Accessed 10 Apr 2024
65. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Processg 13(4), 600 (2004). https://doi.org/10.1109/TIP.2003.819861
66. R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2018
67. K. Ding, K. Ma, S. Wang, E.P. Simoncelli, Image quality assessment: unifying structure and texture similarity. IEEE Trans Pattern Anal Mach Intell 44(5), 2567–2581 (2022). https://doi.org/10.1109/TPAMI.2020.3045810
68. M. Fiedler, T. Hossfeld, P. Tran-Gia, A generic quantitative relationship between quality of experience and quality of service. IEEE Network 24(2), 36–41 (2010). https://doi.org/10.1109/MNET.2010.5430142
69. P. Paudyal, F. Battisti, M. Carli, Impact of video content and transmission impairments on quality of experience. Multimed Tools Appl 75(23), 16461–16485 (2016). https://doi.org/10.1007/s11042-015-3214-0
70. J. van der Hooft, M. Torres Vega, T. Wauters, C. Timmerer, A.C. Begen, F. De Turck, R. Schatz, From capturing to rendering: volumetric media delivery with six degrees of freedom. IEEE Commun Mag 58(10), 49–55 (2020). https://doi.org/10.1109/MCOM.001.2000242
71. E. d'Eon, T. Myers, B. Harrison, P.A. Chou. Joint MPEG/JPEG Input. 8i Voxelized Full Bodies - A Voxelized Point Cloud Dataset. 2017. https://jpeg.org/plenodb/pc/8ilabs/ Accessed 10 April 2024
72. Mpeg pcc tmc2. https://github.com/MPEGGroup/mpeg-pcc-tmc2. Accessed 10 Apr 2024
73. S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P.A. Chou, R.A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A.M. Tourapis, V. Zakharchenko, Emerging mpeg standards for point cloud compression. IEEE J Emerg Sel Top Circuits Syst 9(1), 133–148 (2019). https://doi.org/10.1109/JETCAS.2018.2885981
74. MPEG. Mpeg pcc renderer. http://mpegx.int-evry.fr/software/MPEG/PCC/mpeg-pcc-renderer/. Accessed 10 Apr 2024
75. Mininet. https://mininet.org. Accessed 10 Apr 2024
76. Jetty. https://www.eclipse.org/jetty/. Accessed 10 Apr 2024

77. J. van der Hooft, T. Wauters, F. De Turck, C. Timmerer, H. Hellwagner, Towards 6DoF HTTP Adaptive Streaming Through Point Cloud Compression, in *Proceedings of the 27th ACM International Conference on Multimedia* (Association for Computing Machinery, New York, NY, USA, 2019), MM '19, p. 2405–2413. https://doi.org/10.1145/3343031.3350917.
78. M. Hosseini, Adaptive rate allocation for view-aware point-cloud streaming. CoRR abs/1911.00812. 2019. arXiv:1911.00812
79. Netflix. Video Multimethod Assessment Fusion (VMAF). https://github.com/Netflix/vmaf. Accessed 10 Apr 2024
80. OpenCV. Peak-Signal-to-Noise-Ratio (PSNR). https://shimat.github.io/opencvsharp_docs/html/23f56d6b-49ef-3365-5139-e75712c20fe4.htm. Accessed 10 Apr 2024
81. SciKit-image. Structural Similarity Index Measure (SSIM). https://scikit-image.org/docs/stable/api/skimage.metrics.html#skimage.metrics.structural_similarity. Accessed 10 Apr 2024
82. SciKit Learn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. Accessed 10 Apr 2024
83. Silhouette Score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. Accessed 10 Apr 2024

Publisher's Note