Low-Complexity Patch-based No-Reference Point Cloud Quality Metric exploiting Weighted Structure and Texture Features

Michael Neri, *Member IEEE*, Federica Battisti, *Senior Member, IEEE* .

***Abstract*—During the compression, transmission, and rendering of point clouds, various artifacts are introduced, affecting the quality perceived by the end user. However, evaluating the impact of these distortions on the overall quality is a challenging task. This study introduces PST-PCQA, a no-reference point cloud quality metric based on a low-complexity, learning-based framework. It evaluates point cloud quality by analyzing individual patches, integrating local and global features to predict the Mean Opinion Score. In summary, the process involves extracting features from patches, combining them, and using correlation weights to predict the overall quality. This approach allows us to assess point cloud quality without relying on a reference point cloud, making it particularly useful in scenarios where reference data is unavailable. Experimental tests on three state-of-the-art datasets show good prediction capabilities of PST-PCQA, through the analysis of different feature pooling strategies and its ability to generalize across different datasets. The ablation study confirms the benefits of evaluating quality on a patch-by-patch basis. Additionally, PST-PCQA's light-weight structure, with a small number of parameters to learn, makes it well-suited for real-time applications and devices with limited computational capacity. For reproducibility purposes, we made code, model, and pretrained weights available at https://github.com/michaelneri/PST-PCQA.**

***Index Terms*—No-reference, point cloud, deep learning, low-complexity, quality assessment**

I. Introduction

In recent years, thanks to the increasing capability of 3D acquisition systems, point clouds have emerged as one of the most popular formats for immersive media [1]. Point clouds consist of a collection of points defined by geometric coordinates and optional attributes such as color and reflectivity. They provide the users with a more immersive experience than 2D content thanks to a realistic visualization and the possibility of interaction [2].

Point clouds might undergo several distortions during acquisition, transmission, and display [3]. Acquisition distortions refer to errors and inaccuracies that occur during the capture of 3D data, typically from sensors like Light Detection and Ranging (LiDAR) or structured light cameras. When transmitting these data over networks, compression is often necessary to reduce the file size, thus introducing artifacts like noise, resolution loss, or geometric inaccuracies [4], [5]. During the display phase, hardware limitations or rendering algorithms may further affect the quality, potentially resulting in visual inconsistencies or inaccuracies [5]. While compression artifacts [6] are the most common distortions affecting the rendered point cloud, several other types of distortions can occur, which usually affect geometry and color consistency by introducing noise [7], degrading the overall visual quality of the content.

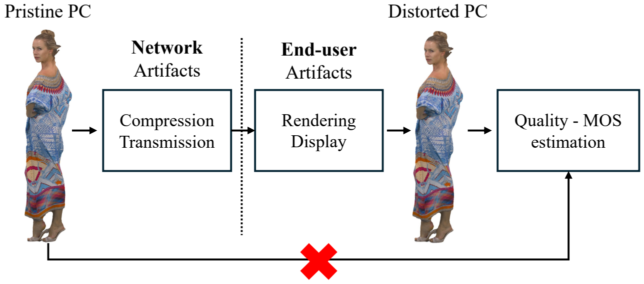

Fig. 1. Description of the no-reference point cloud quality assessment task. From acquisition to rendering, the pristine point cloud is subject to several distortions that may impact the quality perceived by the user.

Given that human observers are the primary users of point clouds in numerous applications, employing subjective quality assessment emerges as the most direct and dependable method for evaluating the quality of point clouds [8]. Despite its significance, subjective quality evaluation poses challenges due to its time-consuming nature and high cost. For the practical implementation of quality-focused point cloud systems, there is a strong demand for objective Point Cloud Quality Assessment (PCQA) models capable of accurately predicting subjective quality assessments [9].

Objective quality estimators can be categorized into three classes: Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR). FR metrics assess quality by comparing the target against an unaltered original, requiring complete access to original data. RR methods need only partial original data, e.g., compression parameters, thus balancing accuracy with data accessibility. NR architectures, instead, evaluate quality without any reference to the original, offering flexibility in real scenarios but potentially at the cost of precision (Figure 1). Moreover, the availability of pristine information may be difficult at the end user device, especially in broadcasting and telecommunication scenarios, thus motivating the

M. Neri is with the with the Faculty of Information Technology and Communication Sciences, Tampere University, Korkeakoulunkatu 1, 33720, Tampere, Finland. (e-mail: michael.neri@tuni.fi)

F. Battisti is with the Department of Information Engineering, University of Padova, Via Gradenigo 6/b, 35131, Padova, Italy. (e-mail: federica.battisti@unipd.it).

The work of F. Battisti was carried out within the “HEAT – Hybrid Extended reAliTy” Project GA 101135637 funded by the EU Horizon Europe Framework Programme (HORIZON).

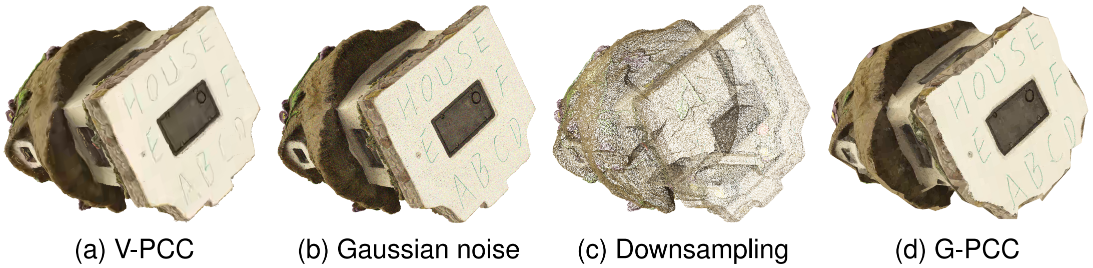

Fig. 2. Examples of compression techniques applied to the *House* point cloud from WPC [12] dataset: (a) V-PCC with 'geometryQP'= 35 and 'textureQP'= 45; (b) Gaussian noise with standard deviation = 0 for points' coordinates and 16 for points' RGB values; (c) downsampling uniformly dividing the point cloud in 2^8 segments; (d) G-PCC with trisoup, 'NodeSizeLog2'= 4 and RAHT quantization step= 64.

development of no-reference metrics for immersive multimedia [10]. However, currently the literature lacks no-reference methods for efficient estimation of the quality of distorted point clouds [8], [11].

State-of-the-art NR PCQA metrics present several challenges:

- *non-ML methods exhibit low correlation between predicted and ground truth quality scores [13], [14];*
- *methods exploiting specific features extracted from the point clouds (e.g., texture and structure) are mainly working in a global way on the entire point cloud [15];*
- *deep learning-based NR PCQA require a significant amount of computational resources and available datasets (also needed to reduce generalization issues) [16]–[18].*

In this work, we introduce PST-PCQA, a low-complexity learning-based NR PCQA for non-sparse point clouds that outperforms state-of-the-art architectures. In more detail, PST-PCQA splits point cloud into patches from which texture and structure features are extracted. Those features are then combined to predict the overall quality. This approach is lightweight, i.e., the number of learnable parameters (1.8M) is the lowest in the state-of-the-art, with a total decrease of 93% with respect to the most efficient approach. This characteristic is crucial in devices where the computational load is limited and in applications where system response time should be real-time.

The main contributions of this paper are as follows:

- the definition of a new lightweight NR PCQA, namely Patch-based Structure and Texture (PST)-PCQA, that exploits texture and structure features of the point cloud;
- a patch-wise quality estimation strategy. This approach allows the adoption of learned weights per patch and to improve explainability;
- an extensive analysis on state-of-the-art datasets for NR PCQA to demonstrate the effectiveness of the approach. Comparisons with other methods in the literature are carried out.

The remainder of this paper is structured as follows: Section II details the relevant works in the literature. Section III illustrates the proposed approach from the extraction of features to the final prediction of the quality. Section IV reports the performance of our NR PCQA metric on 3 state-of-the-art datasets, providing insights on the design rationale of the approach. Finally, Section V draws the conclusions with possible future directions of the work.

II. Related Works

In this section, state-of-the-art metrics for the quality assessment of point clouds are presented. First, FR and RR approaches are introduced. Then, an in-depth description of existing NR metrics is provided.

A. Full- and reduced-reference metrics

In [19], the first attempt to assess the quality of colored point clouds was proposed. In more detail, the model PCQM inspects both geometry-based (e.g., mean curvature) and color-based features (e.g., lightness, chroma, and hue) to predict the mean opinion score (MOS) of a distorted point cloud. In this direction, in [20], the authors proposed TDESM, a FR metric which employs 3D Difference of Gaussian filters on both reference and distorted point clouds to extract similarity features from edge information. The authors in [12] proposed a FR metric that exploits 2D projections of the point cloud, which are then analyzed using IW-SSIM for predicting the overall quality. Instead, Zhang *et al.* [21] devised TCDM, a space-aware vector autoregressive model that defines the quality of a distorted point as the difficulty of transforming it into its corresponding reference.

With the advent of deep learning, the research community has started investigating the use of neural networks for quality assessment. An example of using learning-based methods for assessing the quality of point clouds was designed in [22]. Specifically, a Visual Geometry Group (VGG)-like Convolutional Neural Network (CNN) randomly extracts patches from both pristine and distorted point clouds to analyze both structure and color characteristics. MOS was then predicted by means of multilayer perceptrons (MLPs). In [23], the authors devised Multiscale Potential Energy Discrepancy (MPED) in which the differences between pristine and distorted point clouds were measured via a multiscale potential energy approach, inspired by classical physics.

To the best of our knowledge, two reduced reference metrics are available in the state-of-the-art. In [24] the authors exploited information of the artifact type (e.g., Video-based Point Cloud Compression (V-PCC) compression parameters) to predict the MOS of the distorted point cloud. Moreover,

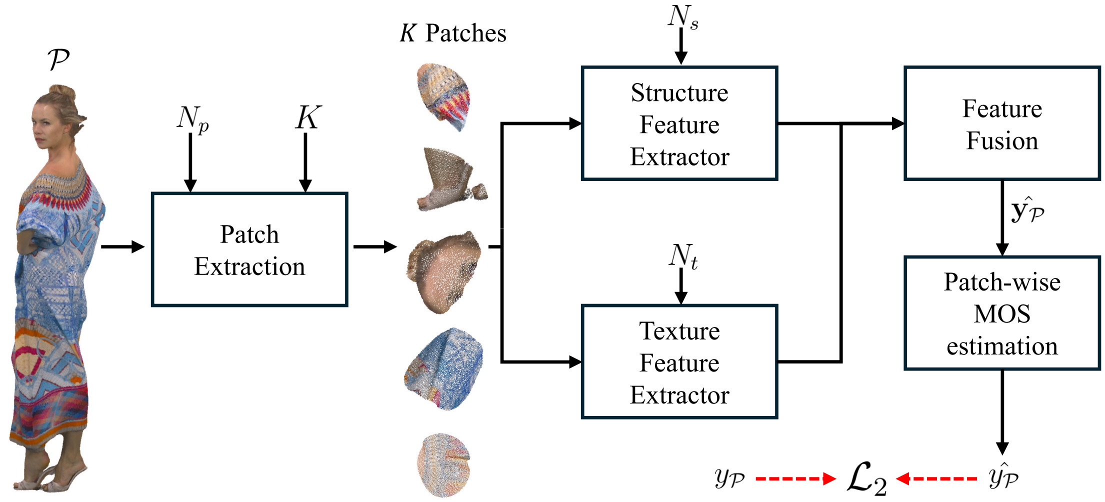

Fig. 3. Description of the proposed approach.

in [25] V-PCC parameters are estimated from the original and the distorted point cloud to predict the MOS.

As stated before, FR and RR metrics are effective but hardly applicable in real scenarios due to the unavailability of the pristine point cloud at the receiver.

B. No-reference metrics

Similarly to FR and RR metrics, the research community had initially started investigating the quality of distorted point clouds by extracting hard-engineered features, i.e., characteristics coming from by domain knowledge and expertise. For instance, in [26], the authors proposed BQE-CVP, a NR metric that extracts features from the distorted point cloud such as geometric and color information. MOS is then predicted using a Random Forest (RF). Similarly, in [27] the distorted point cloud was projected into quality-related geometry and color feature domains in order to apply Natural Scene Statistics (NSS) and entropy-based features. Finally, a Support Vector Regression (SVR) model was devised to regress the quality of the input point cloud. In this direction, Liu *et al.* [28] analyzed the relationship between V-PCC texture quantization parameters and perceptual coding distortion, providing the basis for the definition of a bitstream-layer NR model. However, extracting hand-crafted and compression-based features yielded poor performance. Hence, learning-based methods employing neural networks have shown improved performances with respect to traditional methods thanks to their ability to automatically extract relevant features for predicting point clouds' quality.

The first relevant work using deep learning was proposed in [9], where 2D projections of the distorted point cloud were processed by several CNNs, whose features were concatenated to predict the overall quality. Similarly, in [29], IT-PCQA was devised to predict the MOS of distorted point clouds by inspecting multi-perspective images. Training of the Deep Neural Network (DNN) was carried out as a domain adaptation problem, exploiting the subjective scores available for 2D natural images datasets in the state-of-the-art and transferring this knowledge to the NR PCQA task.

Together with the release of a large-scale NR PCQA dataset, the authors in [8] proposed a 3D CNN which exploited sparse convolutions directly on points, namely ResSCNN. This is the first approach that tackles the computational complexity problem of NR metrics in this field, as ResSCNN encompassed only 1.2M learnable parameters. However, similarly to prior works, its performance on well-known datasets were insufficient to be directly employed in real applications.

In [17] the authors proposed EEP-3DQA which employs lightweight Swin-Transformer [30] as the backbone for feature extraction to predict the quality of both point clouds and mesh models. Similarly to [9], [29], projections of the distorted 3D model are extracted from six standard viewpoints and then analyzed by the DNN. An example of using Graph Convolutional Network (GCN) in PCQA was devised in [31] which attentively analyzes the structural and textural perturbations within point clouds. Moreover, the approach involves a multi-task framework that predicts both distortion type and degree, increasing its sensitivity to several distortion types.

In [32] the authors proposed to process static and dynamic views from a moving camera to have a more comprehensive assessment of point cloud quality. Specifically, VQA_PC consists in rotating the camera around the point cloud, extracting spatial and temporal features using deep learning models, and combining them to predict the overall quality of the distorted point cloud. Following the same approach, Wang *et al.* [11] designed MOD-PCQA that exploits multiscale feature extraction to evaluate point cloud quality from various observational distances. The DNN incorporates a three-branch network structure designed to extract features from different scales, enhancing the model's ability to capture and analyze the perceptual quality of point clouds.

A combination of learning-based and traditional features was proposed in [15]. Specifically, the architecture named MFE-Net integrates an adaptive feature extraction (AFE) module for local hand-crafted feature extraction, a local quality acquisition (LQA) model for deep feature learning, and a global quality acquisition (GQA) layer that aggregates these assessments into the global predicted MOS. Another proposed

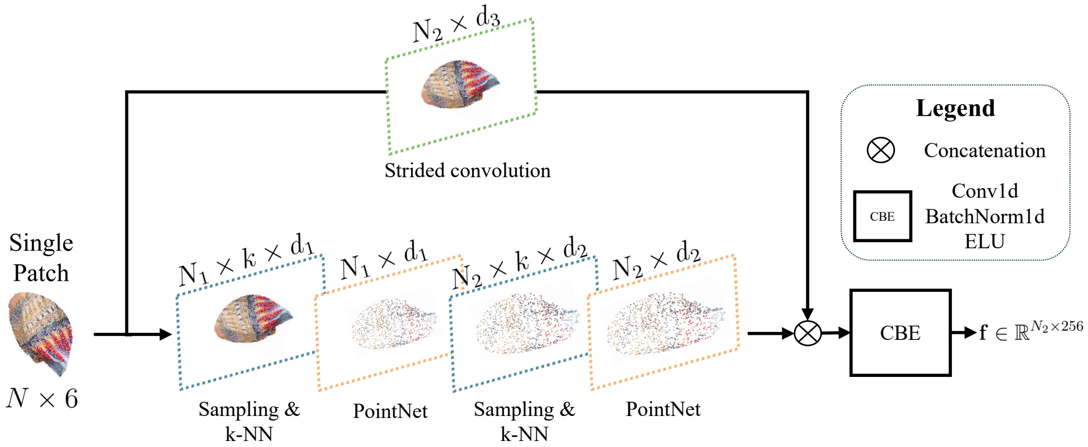

Fig. 4. SFE and TFE neural architectures.

NR metric is Plain-PCQA [18], where spatial geometric properties and texture details are jointly analyzed.

Recent works, rather than processing 2D projections, employed patches from the point cloud to extract features and regress the MOS [16], [33].

Based on these new approaches, we propose a low-complexity deep learning metric that outperforms existing state-of-the-art models in predicting the MOS of non-sparse distorted point clouds. Specifically, PST-PCQA splits the point cloud into patches to separately extract structure and texture features, which are then integrated to estimate the overall quality. Moreover, thanks to its lightweight design, this model can be effectively employed in environments with limited resources, differently from other learning-based approaches employing DNN.

III. Proposed Approach

The objective of this work is to estimate the quality of a generic point cloud as perceived by an average observer, the MOS, without having information of its pristine version. Specifically, a point cloud is denoted as a set of N points that represents the surface of a 3D object, i.e., $\mathcal{P} = \{\mathbf{p_i}, i = 1, \ldots, N\}$. A single point is described as a vector containing its spatial coordinates and color information, $\mathbf{p_i} = [x_i, y_i, z_i, r_i, g_i, b_i]$. The set $\mathcal{P}$ can be denoted as a $N \times 6$ matrix, where N is in the order of millions. Our approach aims at mapping the input point cloud and its quality $f : \mathbb{R}^{N\times 6} \rightarrow \mathbb{R}^{+}$ such as

$$f(\mathcal{P}) = y_{\mathcal{P}}, \qquad (1)$$

where $y_{\mathcal{P}} \in \mathbb{R}^{+}$ is the MOS of the point cloud $\mathcal{P}$.

Figure 3 illustrates all the steps of the proposed architecture. Initially, a preprocessing step is applied to the distorted point cloud to obtain K patches with N_p points. Then, these portions of point cloud are fed to the Structure Feature Extractor (SFE) and the Texture Feature Extractor (TFE) modules to provide patch-wise features, analyzing both their structure and color patterns. Finally, a patch-wise and a global prediction of the quality of the distorted point cloud is provided as output.

A. Patch extraction

Using point clouds patches can enhance computational efficiency and facilitate feature extraction as smaller data segments allow for more in-depth analysis and processing. In fact, raw point clouds, especially those with high densities, may have millions of points, which can overwhelm memory and processing capabilities. Exploiting patches of point cloud can fasten the feature extraction process, whose characteristics can be combined by the MOS prediction module for both local and global analysis. Following this rationale, all points' coordinates $[x_i, y_i, z_i]$ of the point cloud $\mathcal{P}$ are first normalized in the range 1 and 2001, i.e., in a sphere with radius 1000, for training stability and generalization purposes [33]. Then, FPS [34] and k-NN [35] are employed to obtain K centers and to sample N_p points from the point cloud to compose the patches. Finally, all patches are concatenated along the channel dimension to compose the tensor with shape $K \times N_p \times 6$.

B. Structure and texture feature extractors

SFE and TFE neural networks adopt a layered feature extraction process, inspired by the hierarchical feature learning framework of PointNet++ [36]. Specifically, they utilize the sampling, grouping, and PointNet (SGP) layers [36] to create an *abstraction layer*, thus obtaining a transformed representation of the point cloud. This mechanism facilitates localized point cloud analysis that can be used to predict the MOS. Our method differs from [36] in few key aspects:

- we adopt grouped convolutions to reduce the total number of trainable parameters;
- we replace Farthest Point Sampling (FPS) in the sampling phase with random sampling to enhance diversity and improve generalization capabilities;
- we employ the ELU [37] activation function instead of LeakyReLU. In fact, ELU leads to faster learning and to significantly better generalization performance than vanilla ReLUs and LeakyReLU on networks with more than 5 layers.

Figure 4 depicts the structure of SFE and TFE that only differ in the number of input points. A patch with shape $N_t \times 6$

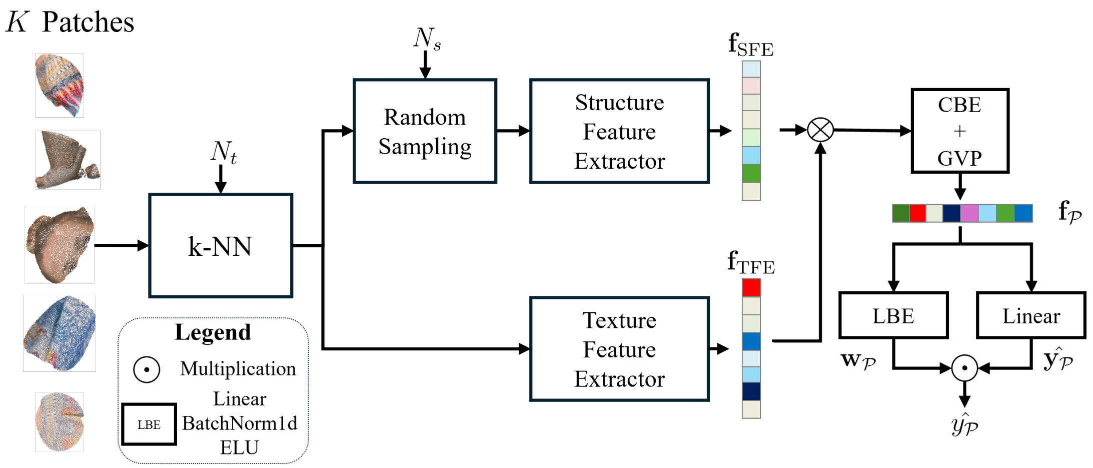
K Patches
N_t
k-NN
N_s
Random Sampling
Structure Feature Extractor
Texture Feature Extractor
$\mathbf{f}_{\text{SFE}}$
$\mathbf{f}_{\text{TFE}}$
CBE + GVP
$\mathbf{f}_{\mathcal{P}}$
LBE
Linear
$\mathbf{w}_{\mathcal{P}}$
$\hat{\mathbf{y}_{\mathcal{P}}}$
$\hat{y_{\mathcal{P}}}$
Legend
Multiplication
LBE
Linear BatchNorm1d ELU

Fig. 5. SFE and TFE common structure.

is fed to the TFE, whereas its downsampled version $N_s \times 6$ is elaborated by the SFE. By doing so, it is possible to simultaneously analyze patch's structural and texture features.

Each patch is analyzed by two sequential SGP layers. During the first pass, the patch is downsampled to N_1 points, grouped into k centers by means of the k-Nearest Neighbour (k-NN) [35] algorithm, and processed by the PointNet [36] layer, yielding the first transformed representation of the input patch, with shape $N_1 \times d_1$. The last *abstraction layer* samples N_2 points, extracts k centers, and projects each point to a d_2-dimensional space, resulting in a $N_2 \times d_2$ matrix.

In addition, the point cloud patch is analyzed by a learnable convolution with stride $s = \lfloor N/d_2 \rfloor$. The result of each branch is concatenated each other to obtain the patch's features $\mathbf{f} \in \mathbb{R}^{N_2 \times 256}$.

Finally, patch-wise features from both branches are arranged to compose texture and structure features of the input point cloud $\mathcal{P}$, namely $\mathbf{f}_{\text{SFE}} \in \mathbb{R}^{K \times N_2 \times 256}$ and $\mathbf{f}_{\text{TFE}} \in \mathbb{R}^{K \times N_2 \times 256}$ respectively.

C. Patch-wise and global quality estimation

To fuse the extracted characteristics from both SFE and TFE and yield a feature vector per patch $\mathbf{f}_{\mathcal{P}} \in \mathbb{R}^{K \times 512}$, we employ Global Variance Pooling (GVP) and a combination of Conv1d, batch normalization, and ELU (CBE) as follows:

$$f_{\mathcal{P}} = \text{CBE}(\text{GVP}(\mathbf{f}_{\text{SFE}} \otimes \mathbf{f}_{\text{TFE}})), \tag{2}$$

where $\otimes$ denotes the concatenation function. Then, a LBE (linear-batchnorm1d-elu) and a linear layer are employed to predict patch-wise weights $\mathbf{w}_{\mathcal{P}} \in \mathbb{R}^K$ and scores $\hat{\mathbf{y}_{\mathcal{P}}} \in \mathbb{R}^K$ from $\mathbf{f}_{\mathcal{P}}$

$$\begin{cases} \mathbf{w}_{\mathcal{P}} = \text{LBE}(\mathbf{f}_{\mathcal{P}}) \\ \hat{\mathbf{y}_{\mathcal{P}}} = \text{Linear}(\mathbf{f}_{\mathcal{P}}). \end{cases} \tag{3}$$

The predicted point cloud MOS $\hat{y}_{\mathcal{P}}$ is then obtained by combining patch-wise scores with their weights

$$\hat{y}_{\mathcal{P}} = \mathbb{E}_K[\mathbf{w}_{\mathcal{P}} \cdot \hat{\mathbf{y}_{\mathcal{P}}}] \tag{4}$$

where $\mathbb{E}_K[\cdot]$ refers to the expected value across patches.

Figure 5 shows how the prediction of the point cloud quality $y_{\mathcal{P}}$ is estimated from its features $\mathbf{f}_{\text{SFE}}$ and $\mathbf{f}_{\text{TFE}}$. The model is trained by minimizing the Mean Squared Error (MSE) of both patch-wise and global MOS estimation with respect to the ground truth

$$\mathcal{L}(\hat{\mathbf{y}_{\mathcal{P}}}, \hat{y}_{\mathcal{P}}, y_{\mathcal{P}}) = \alpha \mathcal{L}_2(\hat{\mathbf{y}_{\mathcal{P}}}, y_{\mathcal{P}}) + \beta \mathcal{L}_2(\hat{y}_{\mathcal{P}}, y_{\mathcal{P}}) \tag{5}$$

where $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$ are two scalars for balancing the patch-wise and global MOS estimation errors, respectively.

IV. Experimental Results

A. Datasets

To assess the performance of PST-PCQA with respect to architectures in the literature, three state-of-the-art NR PCQA datasets are analyzed.

WPC [12]. It includes 20 original reference point clouds, each subject to five distortions: Gaussian noise, downsampling, and three point cloud compression coding techniques proposed by MPEG (Geometry-based Point Cloud Compression (G-PCC) *octave*, G-PCC *trisoup*, and V-PCC) [40]. These distortions present a wide range of geometric and textural variations, offering substantial examples for learning. For every original reference point cloud, 37 distorted versions are created, leading to a total of 740 distorted point clouds (calculated as 37 distortions multiplied by 20 original samples) within the WPC database, all derived from 20 original reference point clouds. WPC contains inanimate everyday objects (e.g., office supplies) with diverse geometric and textural complexity.

SIAT-PCQD [39]. The SIAT-PCQD database comprises 20 reference point clouds, which undergo several preprocessing steps like subsampling, rotation, and scaling to achieve 10-bit geometric precision. Each point cloud is then subject to distortions using different geometry parameters (ranging from 20 to 32 in 4 increments) and texture parameters (ranging from 27 to 42 in 5 increments) through the V-PCC [40] coding method, resulting in 17 distinct distorted versions per reference point cloud. Consequently, the database encompasses a total of 340 distorted point clouds. SIAT-PCQD contains both human figures and objects. The human category consists

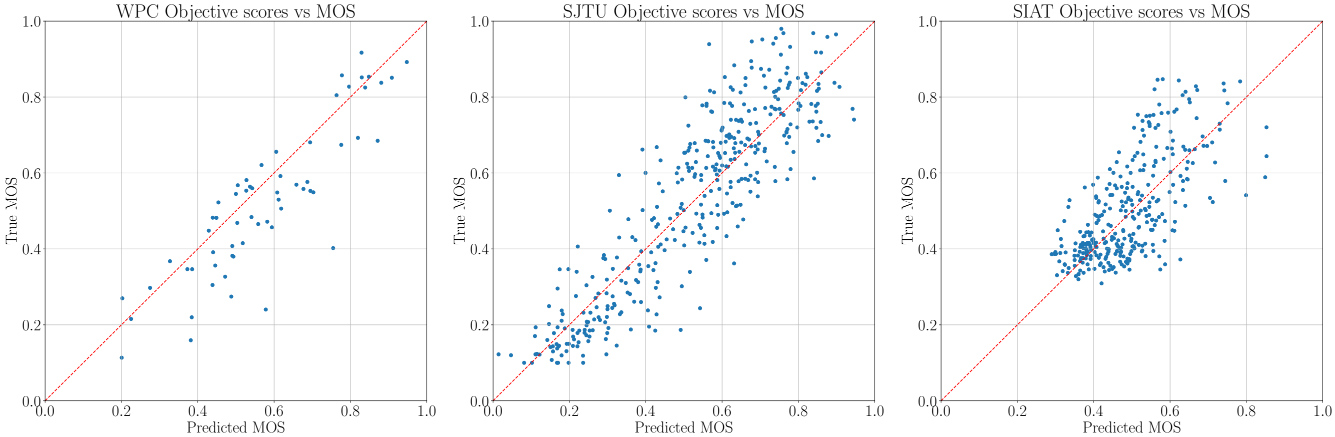

Fig. 6. Scatter plots between normalized predicted and ground truth MOS for WPC [12], SJTU-PCQA [38], and SIAT-PCQD [39] datasets, respectively. Identity line is plotted in red for comparison with ideal quality estimator.

of six full-body figures and four upper-body figures, while objects include ten different instances (e.g., building).

SJTU-PCQA [38]. It comprises 10 publicly accessible reference point clouds, each subject to 7 types of synthetic distortions with 6 intensity levels. These distortions include octree-based compression, color noise, geometry Gaussian noise, downscaling, and combinations thereof. Consequently, the SJTU-PCQA database features a total of 378 distorted point clouds, obtained from 9 samples multiplied by 7 distortions and then by 6 levels. This dataset has been included to analyze the performance of our approach on a dataset with few samples, thus evaluating its convergence stability. SJTU-PCQA includes six human models and four inanimate objects.

All the distorted versions of the same point clouds are either in the training or in the testing dataset to avoid data leakage. Pooling selection and ablation studies are carried out on WPC [12] since, according to the literature, it is the most difficult NR PCQA dataset for data-driven approaches. In fact, WPC database incorporates more intricate distortions and exploits more levels of degradation, modeling real use cases.

B. *Metrics*

The criteria for evaluating the relationship between predicted scores and quality labels are Spearman Rank Order Correlation Coefficient (SRCC), Kendall Rank Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Squared Error (RMSE). A high-performing model is indicated by SRCC, KRCC, and PLCC values approaching 1, and a RMSE value near 0.

C. *Implementation details*

In this work, for fair comparison with state-of-the-art approaches, we split the datasets as follows:

- **WPC:** We follow training and testing split as in [9];
- **SIAT-PCQD:** Leave-one-out 20 cross validation has been implemented;
- **SJTU-PCQA:** Leave-one-out 10 cross validation has been adopted.

Following [33], $K = 16$ patches with $N_p = 14900$ points are extracted from the distorted point clouds. Then, $N_s = 1024$ points are randomly sampled from the patch for analyzing its structure. Differently, $N_t = 8192$ points are selected by means of the k-NN algorithm [35]. In both TFE and SFE, dimensionality of features are set to $d_1 = 128$ and $d_2 = 256$, with number of points $N_1 = 512$ and $N_2 = 256$, respectively. During SGP layers, the number of groups in the k-NN algorithm is $k = 32$. Overall, the number of trainable parameters of the learning-based metric is 1.8M, highlighting the low-complexity of the approach. The model is trained for 400 epochs with batches of size 4. A cosine annealing learning rate is employed with initial learning rate $\eta_{\max} = 0.001$ with a maximum number of steps $T_{\max} = 400$. Both terms of the loss function in Eq. (5) equally contribute to the backpropagation algorithm ($\alpha = 1$ and $\beta = 1$). The overall implementation and evaluation of PST-PCQA has been designed in Python 3.10 in a workstation with a CUDA-enabled graphic processing unit (NVIDIA RTX 4070). Pytorch-Lightning and Weights&Biases are utilized for training and logging, respectively. Further implementation details are available at https://github.com/michaelneri/PST-PCQA.

D. *Analysis on the type of pooling function*

Generally, the choice of pooling functions significantly affects the precision of the prediction in learning-based approaches [41]. To this aim, in Table I an analysis on which type of feature extraction, both first- (e.g., Global Max Pooling (GMP) and Global Average Pooling (GAP)) and second-order (e.g., GVP) statistical moments, works better on WPC [12] is presented. When combined, GAP + GVP and GMP + GVP show improved performance over GAP and GMP alone in terms of SRCC and KRCC. However, it is worth highlighting the effectiveness of GVP for this task, with the highest correlation to MOS in terms of PLCC, SRCC, and KRCC. This suggests that combining pooling methods can capture a broader range of features that may correlate with human perception, but the combination might not always lead to improvement.

TABLE I
PERFORMANCE OF DIFFERENT FEATURE POOLING ON THE WPC DATASET.

Pooling	PLCC ↑	SRCC ↑	KRCC ↑	RMSE ↓
GMP	0.8501	0.8295	0.6506	10.4842
GAP	0.8657	0.8511	0.6703	**10.2192**
GAP + GVP	0.8462	0.8133	0.6298	10.4975
GMP + GVP	0.8712	0.8463	0.6696	10.2588
GVP	**0.8821**	**0.8624**	**0.6854**	10.5769

E. Analysis on the number of patches K

Although PST-PCQA feature extraction architecture does not change with respect to the number of patches K, the patch-wise MOS estimation module includes batch normalization across patches. Hence, to evaluate the effect of K, we provide the performance of PST-PCQA with respect to diverse values of $K = \{2, 4, 8, 16, 32\}$ in Table II. From the results it is worth noting that having few patches ($K = \{2, 4\}$) impacts the performance of PST-PCQA. Comparable results are achieved with $K = \{8, 16\}$ patches whereas having more than $K = 16$ yields an inefficient model both in terms of correlation between true and predicted MOS and of computational complexity.

TABLE II
PERFORMANCE OF DIFFERENT NUMBER OF PATCHES K ON THE WPC DATASET.

K	PLCC ↑	SRCC ↑	KRCC ↑	RMSE ↓
2	0.8075	0.7906	0.6251	13.1335
4	0.8208	0.8087	0.6235	12.5506
8	0.8772	**0.8779**	**0.6965**	10.7112
16	**0.8821**	0.8624	0.6854	**10.5769**
32	0.8438	0.8266	0.6390	11.7893

F. Results on all the datasets

We compare the results of PST-PCQA with different types of state-of-the-art approaches:

- FR: p2p_{MSE} [42], p2p_{H} [42], $\text{p2plane}_{\text{MSE}}$ [43], $\text{p2plane}_{\text{H}}$ [43], PSNR_{Y} [44]. IW-SSIM [12], PCQM [19], MPED [23], TCDM [21].
- RR: PCM_{R} [45] and Liu *et al.* [25].
- NR[1]: BRISQUE [14], NIQE [13], ResCNN [8], 3D-NSS [27], PQA-Net [9], VQA-PC [32], MM-PCQA [16], EEP-3DQA [17], BEQ-CVP [26], SGT-PCQA [5], and Plain-PCQA [18].

Table III shows the comparison of performance between the proposed approach and the state-of-the-art on WPC [12] and SJTU-PCQA [38] datasets. It is worth noticing the superiority of our approach on both datasets in mostly all the metrics with respect to NR models, indicating a strong correlation with MOS. Precisely, our approach outperforms other architectures in terms of PLCC and RMSE on the WPC dataset, whereas it surpasses the state-of-the-art on SJTU-PCQA in all the metrics. In addition, PST-PCQA is also outperforming both RR and FR approaches, emphasizing its practical applicability in scenarios where the pristine point cloud is unavailable.

Table IV depicts the results of our method with respect to approaches in the literature on the SIAT-PCQD [39] dataset, showing similar performance to SGT-PCQA [5]. However, it is important to highlight the difficulty of learning-based approaches on this dataset due to the limited number of point cloud per distortions. In fact, best performance are obtained with hand-crafted features [5], [26] and machine learning regressors, such as RF. In our setup, 4 folds were unsuccessful (PLCC $\approx 50\%$), whereas the others reached an average PLCC performance of 90%. This behavior is mainly caused by the model overfitting to the training set. To address this, designing data augmentation techniques or semi-supervised learning approaches in this field could further enhance performance.

To visually inspect the correlation between predicted and true MOS, Figure 6 displays the scatter plots across the three analyzed datasets.

G. Cross-corpus generalization

To assess the generalization capabilities of the proposed approach, we train on a source dataset, e.g., WPC, and test to a different dataset, e.g., SJTU-PCQA and viceversa. Table V depicts the results of PST-PCQA with respect to state-of-the-art approaches, demonstrating its ability to model Human Vision System (HVS) of point clouds in out-domain scenarios. In fact, PST-PCQA achieves the highest generalization performance in cross-dataset scenarios. For example, when trained on SJTU-PCQA and tested on WPC, it achieves an SRCC of 0.2737 and a PLCC of 0.3797, outperforming other methods such as VQA-PC (SRCC = 0.2733, PLCC = 0.3067). However, the generalization scores remain relatively low, reflecting the challenging nature of cross-corpus evaluation.

H. Ablation study

To demonstrate the effectiveness of each component of the proposed approach, an ablation study has been carried out and the results are depicted in Table VI. It is important to highlight the impact of patch-wise loss $\mathcal{L}_2(\hat{\mathbf{y}_{\mathcal{P}}}, y_{\mathcal{P}})$, which acts as a regularizer for the model. Moreover, our approach without local weighting performs worse, with a decrease of the performance of 4.6% in terms of PLCC, validating our contribution. Finally, we evaluate the contribution of each stream, namely TFE and SFE. The results show that the approach based on the fusion of the features obtained from TFE a SFE yields the best performance. However, it is worth noting that when PST-PCQA includes only one of the two streams, the obtained results are comparable. This demonstrates that combining features with diverse point densities improves the quality estimation.

I. Complexity comparison

Table VII compares the number of parameters of the proposed approach with the top-3 learning-based architectures in the literature. It is worth noting that PST-PCQA has the lowest number of learnable parameters, demonstrating its low-complexity nature. As stated in [46], a reduced number of

[1] COPP-Net [33] is not included due to incorrect training, validation, and testing splits, yielding incomparable results.

TABLE III
EXPERIMENTAL RESULTS ON WPC AND SJTU-PCQA. **BOLD** AND UNDERLINE NOTATIONS HAVE BEEN USED FOR HIGHLIGHTING THE BEST AND SECOND PERFORMANCE, RESPECTIVELY. (−) MEANS NO DATA IS AVAILABLE.

		WPC [12]				SJTU-PCQA [38]			
		PLCC ↑	SRCC ↑	KRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	KRCC ↑	RMSE ↓
FR	$\mathrm{p2p_{MSE}}$ [42]	0.4853	0.4559	0.3182	19.8943	0.8228	0.7294	0.5617	1.3290
	$\mathrm{p2p_{H}}$ [42]	0.3972	0.2786	0.1944	20.8993	0.8005	0.7159	0.5450	1.3634
	$\mathrm{p2plane_{MSE}}$ [43]	0.2440	0.3282	0.2250	22.8226	0.6697	0.6278	0.4825	1.6961
	$\mathrm{p2plane_{H}}$ [43]	0.3842	0.2959	0.2071	21.0416	0.7779	0.6952	0.5302	1.4372
	$\mathrm{PSNR_{Y}}$ [44]	0.6166	0.5823	0.4164	17.9001	0.8124	0.7871	0.6116	1.3224
	IW-SSIM [12]	0.8504	0.8481	−	12.0600	0.7949	0.7833	−	1.4224
	PCQM [19]	0.6162	0.5504	0.4409	17.9027	0.8600	0.8470	−	1.2370
	MPED [23]	0.7000	0.6780	−	16.3740	0.8221	0.7436	0.5799	1.2866
	TCDM [21]	0.8070	0.8040	−	13.5250	0.9300	0.9100	−	0.8910
RR	$\mathrm{PCM_{R}}$ [45]	0.3926	0.3605	0.2543	20.9203	0.6699	0.5622	0.4091	1.7589
NR	BRISQUE [14]	0.2614	0.3155	0.2088	21.1730	0.4214	0.3975	0.2966	2.0930
	NIQE [13]	0.1136	0.2225	0.0953	23.1410	0.2420	0.1379	0.1009	2.2620
	ResCNN [8]	0.4531	0.4362	0.2987	20.2591	0.7821	0.7911	0.5224	1.3651
	3D-NSS [27]	0.6284	0.6309	0.4573	18.1706	0.7819	0.7813	0.6023	1.7740
	IT-PCQA [29]	0.7950	0.7800	−	−	0.5800	0.6300	−	−
	PQA-Net [9]	0.6671	0.6368	0.4684	16.6758	0.8586	0.8372	0.6304	1.0719
	VQA-PC [32]	0.8001	0.8012	0.6237	13.5570	0.8702	0.8611	0.6811	1.1012
	MM-PCQA [16]	0.8556	0.8414	0.6513	12.3506	0.9226	0.9102	0.7838	0.7716
	EEP-3DQA [17]	0.8296	0.8264	0.6422	12.7451	0.9363	0.9095	0.6811	1.1010
	MOD-PCQA [11]	0.8733	<u>0.8752</u>	**0.6952**	11.0600	<u>0.9534</u>	<u>0.9311</u>	<u>0.7939</u>	<u>0.7124</u>
	Plain-PCQA [18]	<u>0.8783</u>	**0.8793**	<u>0.6951</u>	<u>10.8308</u>	0.9302	0.9133	0.7603	0.8607
	PST-PCQA (ours)	**0.8821**	0.8624	0.6854	**10.5769**	**0.9593**	**0.9514**	**0.8049**	**0.6630**

TABLE IV
EXPERIMENTAL RESULTS ON SIAT-PCQD. (−) MEANS NO DATA IS AVAILABLE.

		SIAT-PCQD [39]			
		PLCC ↑	SRCC ↑	KRCC ↑	RMSE ↓
FR	$\mathrm{p2p_{MSE}}$ [42]	0.3136	0.3963	0.2761	0.1224
	$\mathrm{p2p_{H}}$ [42]	0.2980	0.3791	0.2620	0.1231
	$\mathrm{p2plane_{MSE}}$ [43]	0.3498	0.4125	0.2947	0.1208
	$\mathrm{p2plane_{H}}$ [43]	0.3218	0.3862	0.2679	0.1221
	$\mathrm{PSNR_{Y}}$ [44]	0.3443	0.3481	0.2318	0.1211
	IW-SSIM [12]	0.8181	0.6966	0.5183	0.0742
	PCQM [19]	0.6539	0.6666	0.4825	0.0994
RR	$\mathrm{PCM_{R}}$ [45]	0.3851	0.3940	−	−
	Liu *et al.* [25]	0.9133	0.9095	−	−
NR	3D-NSS [27]	0.5550	0.5310	−	−
	IT-PCQA [29]	0.7870	0.7920	−	−
	SGT-PCQA [5]	**0.8480**	**0.7950**	−	−
	BEQ-CVP [26]	0.7230	0.6490	−	−
	PST-PCQA (ours)	0.8304	0.7931	**0.5785**	**0.0183**

TABLE V
CROSS-CORPUS GENERALIZATION ANALYSIS.

	WPC → SJTU		SJTU → WPC	
	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑
3D-NSS [27]	0.2117	0.2034	0.1214	0.1313
PQA-Net [9]	0.5411	0.6102	0.2211	0.2334
ResCNN [8]	0.5012	0.4954	0.2301	0.2293
VQA-PC [32]	0.5866	0.6525	0.2733	0.3067
PST-PCQA	**0.7413**	**0.7522**	**0.2737**	**0.3797**

trainable parameters can reduce the likelihood of overfitting, which is particularly beneficial for small datasets. Furthermore, models with fewer parameters demand less computational power and time during both the optimization and inference phase due to the decreased quantity of parameters.

TABLE VI
ABLATION STUDY ON THE WPC DATASET [12].

	PLCC ↑	SRCC ↑	KRCC ↑	RMSE ↓
No $\mathcal{L}_2(\hat{\mathbf{y}_{\mathcal{P}}}, y_{\mathcal{P}})$	0.7773	0.6990	0.5119	13.6713
No LBE	0.8358	0.8587	0.6667	12.1790
No TFE	0.8642	0.8639	0.6783	11.4367
No SFE	0.8751	0.8637	0.6836	10.7429
PST-PCQA	**0.8821**	**0.8624**	**0.6854**	**10.5769**

According to 3GPP [47] specification for Extended Reality (XR) applications, a system processing and broadcasting multimedia content is real-time if the delay is around or lower than 200 ms. In our setup, PST-PCQA runs on a NVIDIA RTX 4070, which is a commercial-off-the-shelf GPU for gaming, with an average inference time of 70 ms per point cloud.

TABLE VII
COMPLEXITY COMPARISON WITH STATE-OF-THE-ART.

Approach	MM-PCQA [16]	Plain-PCQA [18]	EEP-3DQA [17]	**PST-PCQA**
# Params (M)	58.37	28.50	27.54	**1.80**

V. CONCLUSIONS

In this work we propose a novel low-complexity learning-based NR PCQA, namely PST-PCQA, which analyzes the input point cloud in patches. Our approach combines both local and global features to provide a prediction of the point cloud MOS. Extensive experimental results on 3 widely adopted datasets in the state-of-the-art show the effectiveness of PST-PCQA, providing design rationales on feature pooling, cross-corpus generalization capabilities. The ablation study demonstrates that a patch-wise analysis can enhance the performance of PST-PCQA. Furthemore, the reduced number of learnable parameters of PST-PCQA enables its use in real-time and computation-constrained hardware. A possible future investigation can concern the in-depth analysis of the impact of geometry- and texture-based distortions on the predicted MOS. Moreover, as a future work, adaptive 1D kernel convolutions, similarly to their 2D counterpart in [41], i.e., changing kernel values with respect to the content, could be included to improve the generalization ability of PST-PCQA.

References

[1] A. Ak, E. Zerman, M. Quach, A. Chetouani, A. Smolic, G. Valenzise, and P. Le Callet, “BASICS: Broad Quality Assessment of Static Point Clouds in a Compression Scenario,” *IEEE Transactions on Multimedia*, pp. 1–13, 2024.

[2] E. Alexiou, E. Upenik, and T. Ebrahimi, “Towards subjective quality assessment of point cloud imaging in augmented reality,” in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2017.

[3] H. Su, Q. Liu, Y. Liu, H. Yuan, H. Yang, Z. Pan, and Z. Wang, “Bitstream-Based Perceptual Quality Assessment of Compressed 3D Point Clouds,” *IEEE Transactions on Image Processing*, vol. 32, pp. 1815–1828, 2023.

[4] Z. Liu, Q. Li, X. Chen, C. Wu, S. Ishihara, J. Li, and Y. Ji, “Point Cloud Video Streaming: Challenges and Solutions,” *IEEE Network*, vol. 35, no. 5, pp. 202–209, 2021.

[5] R. Tu, G. Jiang, M. Yu, Y. Zhang, T. Luo, and Z. Zhu, “Pseudo-Reference Point Cloud Quality Measurement Based on Joint 2-D and 3-D Distortion Description,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–14, 2023.

[6] M. Yang, Z. Luo, M. Hu, M. Chen, and D. Wu, “A Comparative Measurement Study of Point Cloud-Based Volumetric Video Codecs,” *IEEE Transactions on Broadcasting*, vol. 69, no. 3, pp. 715–726, 2023.

[7] Q. Liang, Z. He, M. Yu, T. Luo, and H. Xu, “MFE-Net: A Multi-Layer Feature Extraction Network for No-Reference Quality Assessment of 3-D Point Clouds,” *IEEE Transactions on Broadcasting*, vol. 70, no. 1, pp. 265–277, 2024.

[8] Y. Liu, Q. Yang, Y. Xu, and L. Yang, “Point Cloud Quality Assessment: Dataset Construction and Learning-Based No-Reference Metric,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2022.

[9] Q. Liu, H. Yuan, H. Su, H. Liu, Y. Wang, Huan Yang, and Junhui Hou, “PQA-Net: Deep No Reference Point Cloud Quality Assessment via Multi-View Projection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4645–4660, 2021.

[10] K. Lamichhane, M. Neri, F. Battisti, P. Paudyal, and M. Carli, “No-Reference Light Field Image Quality Assessment Exploiting Saliency,” *IEEE Transactions on Broadcasting*, vol. 69, no. 3, pp. 790–800, 2023.

[11] J. Wang, W. Gao, and G. Li, “Zoom to Perceive Better: No-reference Point Cloud Quality Assessment via Exploring Effective Multiscale Feature,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[12] Q. Liu, H. Su, Z. Duanmu, W. Liu, and Z. Wang, “Perceptual Quality Assessment of Colored 3D Point Clouds,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 8, pp. 3642–3655, 2023.

[13] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “Completely Blind” Image Quality Analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-Reference Image Quality Assessment in the Spatial Domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[15] Q. Liang, Z. He, M. Yu, T. Luo, and H. Xu, “MFE-Net: A Multi-Layer Feature Extraction Network for No-Reference Quality Assessment of 3-D Point Clouds,” *IEEE Transactions on Broadcasting*, vol. 70, no. 1, pp. 265–277, 2024.

[16] Z. Zhang, W. Sun, X. Min, Q. Wang, J. He, Q. Zhou, and G. Zhai, “MM-PCQA: Multi-modal learning for no-reference point cloud quality assessment,” in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

[17] Z. Zhang, W. Sun, Y. Zhou, W. Lu, Y. Zhu, X. Min, and G. Zhai, “EEP-3DQA: Efficient and Effective Projection-Based 3D Model Quality Assessment,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2023.

[18] X. Chai, F. Shao, B. Mu, H. Chen, Q. Jiang, and Y. Ho, “Plain-PCQA: No-Reference Point Cloud Quality Assessment by Analysis of Plain Visual and Geometrical Components,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[19] G. Meynet, Y. Nehmé, J. Digne, and G. Lavoué, “PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds,” in *Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020.

[20] Z. Lu, H. Huang, H. Zeng, J. Hou, and K. Ma, “Point Cloud Quality Assessment via 3D Edge Similarity Measurement,” *IEEE Signal Processing Letters*, vol. 29, pp. 1804–1808, 2022.

[21] Y. Zhang, Q. Yang, Y. Zhou, X. Xu, L. Yang, and Y. Xu, “TCDM: Transformational Complexity Based Distortion Metric for Perceptual Point Cloud Quality Assessment,” *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–18, 2023.

[22] A. Chetouani, M. Quach, G. Valenzise, and F. Dufaux, “Convolutional Neural Network for 3D Point Cloud Quality Assessment with Reference,” in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–6.

[23] Q. Yang, Y. Zhang, S. Chen, Y. Xu, J. Sun, and Z. Ma, “MPED: Quantifying Point Cloud Distortion Based on Multiscale Potential Energy Discrepancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6037–6054, 2023.

[24] Y. Liu, Q. Yang, and Y. Xu, “Reduced Reference Quality Assessment for Point Cloud Compression,” in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2022, pp. 1–5.

[25] Q. Liu, H. Yuan, R. Hamzaoui, H. Su, J. Hou, and H. Yang, “Reduced Reference Perceptual Quality Model With Application to Rate Control for Video-Based Point Cloud Compression,” *IEEE Transactions on Image Processing*, vol. 30, pp. 6623–6636, 2021.

[26] L. Hua, G. Jiang, M. Yu, and Z. He, “BQE-CVP: Blind Quality Evaluator for Colored Point Cloud Based on Visual Perception,” in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2021.

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

[28] Q. Liu, H. Su, T. Chen, H. Yuan, and R. Hamzaoui, “No-Reference Bitstream-Layer Model for Perceptual Quality Assessment of V-PCC Encoded Point Clouds,” *IEEE Transactions on Multimedia*, vol. 25, pp. 4533–4546, 2023.

[29] Q. Yang, Y. Liu, S. Chen, Y. Xu, and J. Sun, “No-Reference Point Cloud Quality Assessment via Domain Adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[31] Z. Shan, Q. Yang, R. Ye, Y. Zhang, Y. Xu, X. Xu, and S. Liu, “GPA-Net:No-Reference Point Cloud Quality Assessment with Multi-task Graph Convolutional Network,” *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2023.

[32] Z. Zhang, W. Sun, Y. Zhu, X. Min, W. Wu, Y. Chen, and G. Zhai, “Evaluating Point Cloud from Moving Camera Videos: A No-Reference Metric,” *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[33] J. Cheng, H. Su, and J. Korhonen, “No-Reference Point Cloud Quality Assessment via Weighted Patch Quality Prediction,” *arXiv preprint arXiv:2305.07829*, 2023.

[34] Y. Eldar, M. Lindenbaum, M. Porat, and Y.Y. Zeevi, “The farthest point strategy for progressive image sampling,” *IEEE Transactions on Image Processing*, vol. 6, no. 9, pp. 1305–1315, 1997.

[35] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[37] D. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by Exponential Linear Units (ELUs),” *arXiv preprint arXiv:1511.07289*, 2015.

[38] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun, “Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration,” *IEEE Transactions on Multimedia*, vol. 23, pp. 3877–3891, 2021.

[39] X. Wu, Y. Zhang, C. Fan, J. Hou, and S. Kwong, "Subjective Quality Database and Objective Study of Compressed Point Clouds With 6DoF Head-Mounted Display," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4630–4644, 2021.

[40] 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, and V. Zakharchenko, "Emerging MPEG Standards for Point Cloud Compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2019.

[41] Z. Zhou, J. Li, D. Zhong, Y. Xu, and P. Le Callet, "Deep Blind Image Quality Assessment Using Dynamic Neural Model with Dual-order Statistics," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[42] R. Mekuria, Z. Li, C. Tulvan, and P. Chou, "Evaluation criteria for pcc (point cloud compression)," *ISO/IEC JTC*, vol. 1, pp. N16332, 2016.

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

[44] R. Mekuria, S. Laserre, and C. Tulvan, "Performance assessment of point cloud compression," in *IEEE Visual Communications and Image Processing (VCIP)*, 2017.

[45] I. Viola and P. Cesar, "A Reduced Reference Metric for Visual Quality Evaluation of Point Cloud Contents," *IEEE Signal Processing Letters*, vol. 27, pp. 1660–1664, 2020.

[46] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[47] 3GPP, "5G; Extended Reality (XR) in 5G," Technical Specification (TS) 26.928.331, 3rd Generation Partnership Project (3GPP), 11 2020, Version 16.0.0.

Michael Neri (Member, IEEE) obtained the Ph.D. in Applied Electronics (Roma Tre University) in 2025. He is now a researcher at Tampere University, Faculty of Information Technology and Communication Sciences, Finland. His main research interests are in the area of computer vision, deep learning, and audio processing.

Federica Battisti (Senior Member, IEEE) is Associate Professor with the Department of Information Engineering, University of Padova. Her research interests include multimedia quality assessment and security. She is Editor in Chief for *Signal Processing: Image Communication* (Elsevier) and Vice Chair of the EURASIP Technical Area Committee on Visual Information Processing.