MOSAIC: Multimodal Perceptual Cue Fusion for Full-Reference Point Cloud Quality Assessment

Jiaxin Li, Tong Chen, Zehong Li, Dandan Ding, and Zhan Ma

***Abstract*—Full-reference point cloud quality assessment (FR-PCQA) has access to the original point cloud, but reliable quality prediction still depends on how geometric, photometric, and view-dependent discrepancies are represented and fused. Motivated by David Marr's three-level account of vision, we formulate FR-PCQA as a multimodal perceptual inference problem and propose MOSAIC, a training-free cue-fusion framework. MOSAIC extracts local 3D structural cues, 3D color cues, and projected 2D saliency cues from reference and distorted point clouds. Instead of aggregating all cues in a single stage, MOSAIC first combines geometry and color fidelity into an intrinsic 3D fidelity score, and then incorporates 2D saliency consistency to reflect attention-related perceptual importance. A bounded adaptive weighting strategy produces the final perceptual fidelity score without mean opinion score (MOS)-supervised training or regression-based score integration. Experiments on five public PCQA databases show that MOSAIC achieves the best average performance among training-free metrics with complete results, obtains the best results on several datasets and distortion groups, and remains competitive with representative learning-based models while avoiding supervised score fitting. Materials are accessible at https://njuvision.github.io/MOSAIC.**

***Index Terms*—Point Cloud Quality Assessment, Multimodal Cue Fusion, 3D Geometry, 3D Color, 2D Saliency**

I. Introduction

Point clouds have become a practical representation for immersive applications such as Virtual/Augmented/eXtended Reality (VR/AR/XR), where they compactly describe 3D objects and scenes. In real-world systems, however, acquisition, compression, transmission, and rendering can introduce visible geometry and color artifacts. Point cloud quality assessment (PCQA) therefore plays an important role in predicting subjective quality, such as the mean opinion score (MOS), and in guiding the optimization of end-to-end visual experience.

Depending on the availability of the original signal, PCQA methods are commonly categorized as full-reference (FR), reduced-reference (RR), or no-reference (NR). With complete reference information, FR-PCQA methods generally provide more stable and reliable quality prediction [1].

Complete reference information, however, does not specify how distortions should be represented or pooled. Direct geometry and color metrics are efficient, but they mainly capture point-wise, low-level local discrepancies and often treat geometry and appearance separately [2]–[4]. Subsequent metrics address this limitation by using local neighborhoods, graph or multiscale analysis, and saliency-aware local processing [1], [5]–[8]. Regression-based and deep-learning methods can further improve correlation with subjective scores by learning score mappings [9]–[13], but their performance often depends on MOS-supervised fitting, specific training, or model design. Despite these advances, many methods still emphasize one primary type of perceptual evidence, such as 3D-domain discrepancies, projected 2D-view quality, or learned score mappings. The remaining question is how to organize complementary structural, appearance, and view-level evidence without MOS-supervised quality regression.

Vision theory and psychophysical studies suggest that human vision recovers scene layout and surface structure by combining several cues, while attention affects which regions dominate quality judgments [14]–[17]. Point-cloud quality perception is therefore a mixed 3D–2D process: intrinsic 3D structure and attributes provide object-level fidelity, whereas 2D projections determine view-dependent visibility and attention. This raises the central question of this work: *how can complete reference information be organized into a training-free and interpretable fusion of 3D structural, 3D appearance, and 2D attention-related cues?*

To address this issue, we revisit FR-PCQA from Marr's three-level perspective of vision [14]. At the computational level, the task is to infer perceived quality from the discrepancy between a reference and a distorted 3D representation. At the representational level, this discrepancy should be described by perceptually meaningful cues, including geometric structure, appearance consistency, and attention-related visual importance. At the implementational level, these cues can be realized through measurable descriptors extracted from point clouds and their projected views.

Motivated by the complementary roles of intrinsic 3D fidelity and view-level saliency consistency, we propose MOSAIC, a training-free and interpretable FR-PCQA metric. MOSAIC jointly represents 3D geometric structure, 3D color appearance, and projected 2D saliency as complementary perceptual cues for quality prediction.

The main contributions are summarized as follows:

- We formulate FR-PCQA from Marr's three-level perspective as a multimodal perceptual cue-fusion problem. This formulation clarifies that perceived point cloud quality should be represented by complementary evidence from 3D geometric structure, 3D color appearance, and 2D attention-related saliency.

Copyright © 2026 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

Jiaxin Li, Tong Chen, and Zhan Ma are with the School of Electronic Science and Engineering, Nanjing University, Nanjing, Jiangsu 210023, China (e-mail: lijiaxin@smail.nju.edu.cn, chentong@nju.edu.cn, mazhan@nju.edu.cn).

Zehong Li and Dandan Ding are with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China (e-mail: lizehong@stu.hznu.edu.cn, DandanDing@hznu.edu.cn).

TABLE I
FREQUENTLY USED NOTATIONS AND ACRONYMS IN THIS PAPER.

Term	Description
PCQA	Point cloud quality assessment
FR-PCQA	Full-reference point cloud quality assessment
HVS	Human visual system
MOS	Mean opinion score
MSE	Mean squared error
PSNR	Peak signal-to-noise ratio
D1	Point-to-point geometric distance
D2	Point-to-plane geometric distance
KNN	K-nearest neighbors
SVR	Support vector regression
G-PCC	Geometry-based point cloud compression
V-PCC	Video-based point cloud compression
RGB / YUV / LAB	Color spaces used for point-cloud attributes
PLCC / SROCC / RMSE	Correlation and error evaluation metrics

- We propose MOSAIC, a training-free FR-PCQA metric that converts these perceptual cues into modality-wise fidelity scores. Instead of developing new geometric descriptors, MOSAIC uses established structural descriptors as interpretable geometric evidence and fuses them with color and projected 2D saliency through a two-stage adaptive rule.
- We conduct extensive evaluations on five public PCQA databases and distortion-wise benchmarks. MOSAIC achieves the best average performance among training-free metrics with complete results, obtains the best results on several datasets and distortion groups, and remains competitive with representative regression- and learning-based methods without MOS-supervised score fitting.

Table I lists frequently used notations in this work.

II. RELATED WORK

Projection-based PCQA can reuse mature image quality assessment (IQA) tools by rendering point clouds into multiple 2D views [18]–[23]. However, the resulting scores are sensitive to viewpoint and rendering settings. Therefore, many FR-PCQA studies still investigate native 3D-space metrics that directly compare geometry and color discrepancies. These methods fall into two groups: training-free metrics and learning-based models.

A. Training-Free FR-PCQA

Training-free FR-PCQA metrics compute quality scores from reference-distorted differences without MOS-supervised fitting. Early metrics mainly measure direct point-wise discrepancies. Point-to-point (p2po) [2] and point-to-plane (p2pl) [3] distances quantify geometric errors, while luminance or color PSNR measures attribute distortion [4]. These metrics are simple and efficient, but geometry and color are treated largely as separate low-level signals.

Later methods moved from isolated point errors to local structural comparison. PCQM [5] extracts local geometric and color descriptors from matched point neighborhoods and combines selected curvature- and lightness-related features into a handcrafted full-reference score. PointSSIM [24] extends structural similarity to point clouds by comparing local statistics of geometry, normal, curvature, or color features. GraphSIM [6] constructs local graphs around key points, treats color attributes as graph signals, and combines graph-based color similarity with geometric weighting; MS-GraphSIM [7] evaluates graph similarity at multiple neighborhood radii. MPED [25] describes local distortions through multiscale potential energy discrepancy. GLQI [26] combines global sampling-density and luminance-distribution changes with local depth/luminance gradients and surface-hole features computed on tangent-plane projection graphs, using reference-content complexity for normalization. More recently, PQI [1] refined the same neighbor-to-neighbor paradigm by selecting scale-wise key points, constructing local neighborhoods with a fixed number of neighboring points, and measuring geometry and attribute similarity at multiple scales.

Saliency-assisted methods such as PQSM [8] incorporate saliency into local 3D similarity measurement and quality-score pooling. MOSAIC instead compares the saliency maps of the reference and distorted projected views and treats their consistency as an independent modality-level fidelity cue, rather than using reference saliency only as a spatial weight for pooling local distortion.

Beyond saliency-aware local-neighborhood metrics such as PQSM, recent methods further organize quality evidence through broader perceptual hypotheses. TCDM [27] formulates quality assessment through transformational complexity: it partitions the reference and distorted point clouds into local 3D Voronoi patch pairs and uses a space-aware vector autoregressive model to estimate self-complexity, transformational complexity, and prediction-related terms. PHM [28] adopts a perception-guided hybrid strategy according to distortion degree. It uses two branches: a visible-difference branch for high-quality samples based on texture-complexity-compensated luminance PSNR, and an appearance-degradation branch that combines graph signal variation with weighted co-occurrence statistics of spectral graph wavelet coefficients. Their contribution is not simply a stronger local descriptor, but a different organization of quality evidence: TCDM asks whether recovery complexity can reflect perceptual degradation, whereas PHM asks whether different distortion levels should trigger different HVS-inspired strategies.

B. Learning-Based FR-PCQA

Learning-based FR-PCQA methods use subjective labels or learned features to improve prediction. One line uses handcrafted descriptors as input features and learns a mapping to subjective scores. PointPCA [9] builds objective quality descriptors from PCA-based local neighborhood analysis and uses them for quality regression. PointPCA+ [29] compares local geometry in a reference-derived PCA coordinate system and YCbCr texture statistics, followed by recursive feature elimination and random-forest quality regression. FRSVR [10] computes five full-reference metrics covering geometry distortion, color distortion, and point-count changes, and integrates them using support vector regression (SVR). SGW-PCQA+ [11] uses spectral graph wavelet representations with

supervised score integration. More recently, the Feature Selection Model (FSM) [30] benchmarks features from representative FR metrics, applies recursive feature elimination with SVR or ridge regression, and combines selected features from PCQM, MS-GraphSIM, and the PSNR of D2-based MSE through ridge regression. Since these methods rely on MOS-labeled data for quality regression, their reported performance depends on the training and testing splits.

Another line learns visual representations or quality mappings from point-cloud patches, projected views, point-cloud/mesh descriptors, or multiscale features [12], [31]. Tliba *et al.* [31] extract local deep features from reference and distorted point-cloud patches and measure visual degradation through cross-correlation between the learned embeddings. CT-PCQA [12] renders point clouds into projected views and extracts visual features with CNN and Transformer backbones, including ConvNeXt [32] and Swin Transformer [33]. The model supports both full-reference and no-reference prediction settings.

Large multimodal model (LMM)-based NR-PCQA has recently been explored in LMM-PCQA [34]. It converts quality labels into textual quality levels and combines LMM outputs with multiscale structural features. This idea inspired FR-MLLM, a full-reference method based on multimodal large language models (MLLMs) [13]. FR-MLLM investigates rendered-view prompting, response similarity, relative-quality response scores, fine-tuned absolute-quality scoring, and SVR-based score integration with conventional FR metrics. These LMM- and MLLM-based studies are promising because they bring high-level semantic and cross-modal priors into PCQA, but they also depend on rendering prompts, response parsing, model fine-tuning, and often supervised score integration.

C. Remarks

The above studies show that FR-PCQA has evolved from point-wise error measurement to local structural comparison, perceptual modeling, and supervised representation learning. Cross-domain representations have also been explored in PCQA. Plain-PCQA [35] integrates projected visual and native 3D geometric features through a Plane-Point Interaction Transformer for no-reference prediction. In the full-reference setting, Li *et al.* [36] combine projection-domain perceptual and fidelity features with native 3D quality and structural descriptors through gradient boosting regression. CF-PCQA [37] combines four native 3D geometry/color descriptors with spatial and wavelet similarities of matched patch projections through linear regression. Both methods learn the mapping from complementary 3D/2D descriptors to subjective quality using MOS labels. In contrast to these MOS-supervised methods, MOSAIC adopts a training-free formulation in which reference–distorted saliency-map consistency across rendered views serves as a complementary fidelity cue. Specifically, it first combines native 3D geometry and color fidelity, and then incorporates projected 2D saliency consistency through two-stage fusion, making the roles of intrinsic 3D fidelity and view-level attention-related evidence explicit.

III. Method

MOSAIC follows Marr's view by extracting perceptually meaningful cues of 3D geometric structure, 3D color appearance, and 2D visual saliency before fusing them for FR-PCQA. The overall framework is shown in Fig. 1. For a point cloud $\mathbf{P} = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_N\} \in \mathbb{R}^{N\times 6}$, each point $\vec{p}_i$ contains 3D coordinates $\vec{g}_i$ and three-channel color attributes $\vec{c}_i$, i.e., $\vec{p}_i = \{\vec{g}_i, \vec{c}_i\} = \{(x_i, y_i, z_i), (r_i, g_i, b_i)\}, i \in [1, N]$. N is the total number of points.

A. Point Alignment

Given a reference point cloud $\mathbf{P}_r \in \mathbb{R}^{N\times 6}$ and a distorted point cloud $\mathbf{P}_d \in \mathbb{R}^{M\times 6}$, each point in $\mathbf{P}_r$ is treated as a key point. Its counterpart in $\mathbf{P}_d$ is identified by nearest-neighbor search using Euclidean distance in 3D space. The matched points form the aligned distorted point cloud $\mathbf{P}_{d'} \in \mathbb{R}^{N\times 6}$, which is used for subsequent point-wise feature comparison.

B. Multimodal Perceptual Representation

1) 3D Structural Representation via Geometry Features:
Local shape feature. Geometric distortions degrade local spatial structures such as surfaces and edges, thereby affecting perceptual quality. Following prior work [9], [38], [39], MOSAIC uses eigenvalue-based descriptors as interpretable 3D structural cues. These descriptors are not introduced as standalone novel shape features; rather, they provide a stable structural representation that can be compared between reference and distorted point clouds and then fused with color and saliency cues.

For each point $\vec{p}_i$, a local neighborhood $\{\vec{g}_i^k\}_{k=1}^{K_g}$ is constructed using K_g-nearest neighbors. The corresponding covariance matrix is computed as

$$\mathrm{Cov}_i = \frac{1}{K_g}\sum_{k=1}^{K_g}\left(\vec{g}_i^k - \vec{g}_i\right)\left(\vec{g}_i^k - \vec{g}_i\right)^T. \tag{1}$$

Eigen-decomposition of Cov_i yields three ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Using these eigenvalues, the local curvature $G_i^{(1)}$ and linearity $G_i^{(2)}$ are defined as

$$G_i^{(1)} = \frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}, \quad G_i^{(2)} = \frac{\lambda_1-\lambda_2}{\lambda_1}. \tag{2}$$

The same descriptors are extracted from both the reference and resampled distorted point clouds, denoted as $\{G_{ri}^{(1)}, G_{ri}^{(2)}\}$ and $\{G_{d'i}^{(1)}, G_{d'i}^{(2)}\}$, respectively.

Local point density feature. Geometric resampling or noise changes both point positions and local density. To capture this effect, we compute the squared Euclidean distances between each point and its K_p nearest neighbors, and use their average as the local point density feature for both reference and distorted point clouds:

$$\begin{aligned} G_{ri}^{(3)} &= \frac{1}{K_p}\sum_{k=1}^{K_p}||\vec{g}_{ri}^k - \vec{g}_{ri}||_2^2, \\ G_{d'i}^{(3)} &= \frac{1}{K_p}\sum_{k=1}^{K_p}||\vec{g}_{d'i}^k - \vec{g}_{d'i}||_2^2. \end{aligned} \tag{3}$$

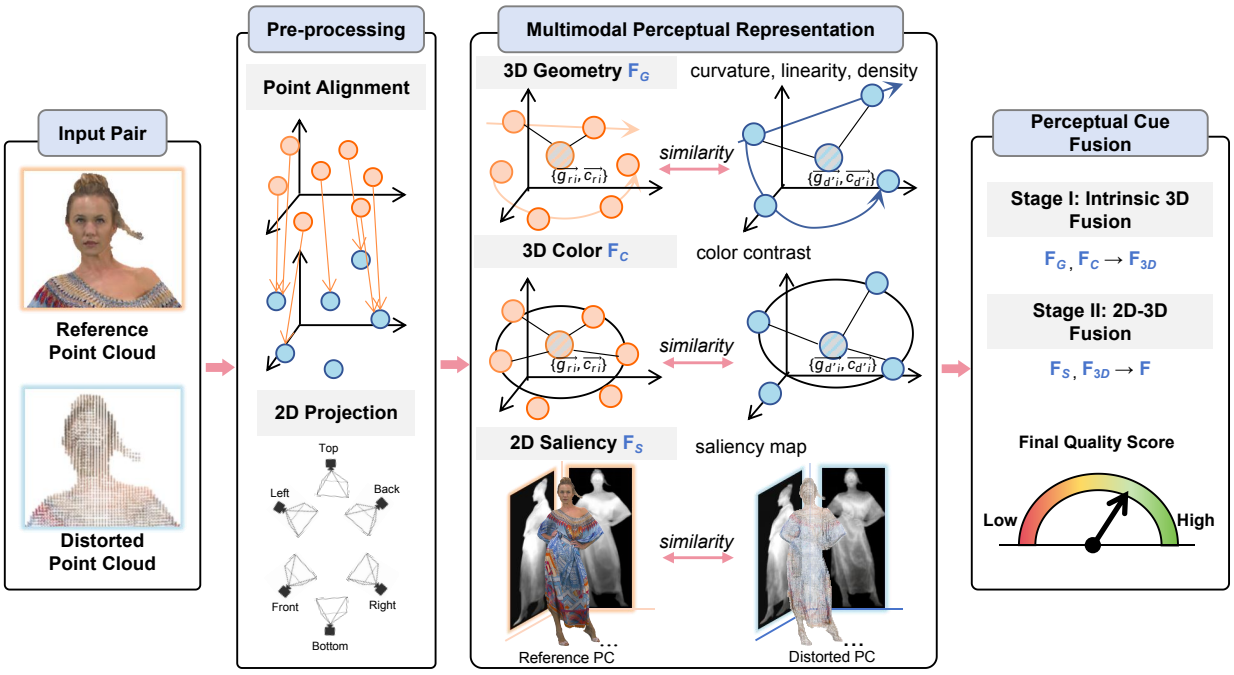


Fig. 1. Overview of the proposed MOSAIC framework from Marr's three-level perspective. The computational goal is to predict perceptual quality from discrepancies between reference and distorted point clouds. The representational level comprises geometric, color, and saliency cues. The implementational level realizes these cues through 3D feature extraction, 2D saliency estimation, and two-stage adaptive fusion.

$\vec{g}_{ri}$ and $\vec{g}_{d'i}$ are the coordinates of the i-th point, and $\vec{g}_{ri}^{k}$ and $\vec{g}_{d'i}^{k}$ represent the coordinates of its k-th nearest neighbor; K_p is the number of nearest neighbors used to compute local density. Together with curvature and linearity, this density term makes the structural cue sensitive to both surface-shape changes and sampling-distribution changes, which are often affected differently by compression, downsampling, and noise.

2) Appearance Representation Using Color Features: MOSAIC uses the native RGB color space. Although alternative spaces (*e.g.*, YUV or LAB) have been explored in prior studies [5], [10], RGB directly represents the stored color attributes in many point-cloud datasets and avoids additional color-space conversion. As reported in Table IX, RGB also achieves the best overall performance within our framework. Since human perception is sensitive to local chromatic contrast, we extract a local RGB residual as the color feature for each point $\vec{p}_i$:

$$\mathbf{C}_{ri} = \left[C_{ri}^{(1)},\ C_{ri}^{(2)},\ C_{ri}^{(3)}\right] = \frac{1}{K_c}\sum_{k=1}^{K_c}(\vec{c}_{ri}^{k} - \vec{c}_{ri}), \tag{4}$$

$$\mathbf{C}_{d'i} = \left[C_{d'i}^{(1)},\ C_{d'i}^{(2)},\ C_{d'i}^{(3)}\right] = \frac{1}{K_c}\sum_{k=1}^{K_c}(\vec{c}_{d'i}^{k} - \vec{c}_{d'i}), \tag{5}$$

where $\mathbf{C}_{ri}$ and $\mathbf{C}_{d'i}$ denote the color features composed of R, G, B components for reference and the aligned distorted point cloud, respectively; $\vec{c}_{ri}$ and $\vec{c}_{d'i}$ represent the color of the i-th point, while $\vec{c}_{ri}^{k}$ and $\vec{c}_{d'i}^{k}$ correspond to the colors of its k-th nearest neighbor. K_c denotes the number of neighboring points used to evaluate local color contrast.

3) Attention-related Saliency Representation: Distortions can alter the predicted saliency patterns of rendered point clouds. MOSAIC compares these patterns as a complementary fidelity cue, rather than directly measuring distortion severity in salient regions. Since a stable saliency layout may coexist with substantial local distortion, this cue is combined with native 3D geometry and color fidelity. Saliency has also been explored in prior PCQA work [40].

Following the orthogonal six-view projection (O6P) setting in [19], we first normalize both reference and distorted point clouds to a fixed projection grid and rasterize them onto the six faces of the axis-aligned bounding cube. These six perpendicular views cover the front/back, left/right, and top/bottom directions with deterministic canonical coverage. We then extract saliency maps from each projected view using the established image saliency detection algorithm [41], yielding saliency representations S_r^v and S_d^v for reference and distorted point clouds, respectively, with $v \in \{1, \ldots, 6\}$ as the view index.

We use projected 2D saliency for two reasons. First, human observers typically perceive rendered point clouds through image-like views, making view-based saliency a natural approximation of visual attention. Second, extracting saliency in the projected domain avoids the ambiguity of defining attention directly on irregular 3D point sets.

Unlike saliency-guided local metrics that transfer saliency back to 3D neighborhoods for local descriptor comparison and score pooling, MOSAIC keeps saliency in the projected-view domain and treats projected 2D saliency-map consistency as an independent modality-level fidelity cue. This choice is motivated by the observation that distortions may alter the saliency distribution itself, which provides complementary information beyond geometry and color discrepancies.

C. Modality-wise Perceptual Fidelity Measurement

We then compute point-to-point similarity between the reference and aligned distorted point clouds for each geometry and color feature. The geometry and color fidelity scores F_G and F_C are obtained by averaging over N points and three features:

$$\mathrm{F}_G = \frac{1}{3}\frac{1}{N}\sum_{l=1}^{3}\sum_{i=1}^{N}\frac{2G_{ri}^{(l)}G_{d'i}^{(l)}+T}{(G_{ri}^{(l)})^2+(G_{d'i}^{(l)})^2+T}, \tag{6}$$

$$\mathrm{F}_C = \frac{1}{3}\frac{1}{N}\sum_{l=1}^{3}\sum_{i=1}^{N}\frac{2C_{ri}^{(l)}C_{d'i}^{(l)}+T}{(C_{ri}^{(l)})^2+(C_{d'i}^{(l)})^2+T}, \tag{7}$$

where $T = 0.001$ is a small constant used to avoid numerical instability.

For the saliency modality, we calculate the structural similarity index (SSIM) [42] for each view pair $(S_r^v, S_d^v), v \in \{1,\ldots,6\}$. Following [40], the view scores are averaged to obtain the final saliency fidelity score F_S:

$$\mathrm{F}_S = \frac{1}{6}\sum_{v=1}^{6}\mathrm{SSIM}(S_r^v, S_d^v). \tag{8}$$

D. Two-stage Perceptual Cue Fusion

After obtaining the geometry, color, and saliency fidelity scores F_G, F_C, and F_S, MOSAIC performs fusion in two stages. The first stage combines F_G and F_C to obtain the intrinsic 3D fidelity score F_{3D}, and the second stage combines F_{3D} with F_S to produce the final perceptual fidelity score F. The corresponding formulations are given in Eqs. (9) and (10).

$$\mathrm{F}_{3D} = \frac{w_{\mathrm{F}_G}\mathrm{F}_G + w_{\mathrm{F}_C}\mathrm{F}_C}{w_{\mathrm{F}_G} + w_{\mathrm{F}_C} + \epsilon}, \tag{9}$$

$$\mathrm{F} = \frac{w_{\mathrm{F}_{3D}}\mathrm{F}_{3D} + w_{\mathrm{F}_S}\mathrm{F}_S}{w_{\mathrm{F}_{3D}} + w_{\mathrm{F}_S} + \epsilon}, \tag{10}$$

where w_{F_G}, w_{F_C}, $w_{\mathrm{F}_{3D}}$, and w_{F_S} denote the weights associated with geometry, color, fused 3D, and saliency fidelity scores, respectively; F_{3D} represents the fused intrinsic 3D fidelity score, F denotes the final perceptual fidelity score, and $\epsilon = 10^{-8}$ is a small denominator constant used to avoid division by zero.

Perceptual studies on multisensory cue integration show that the human visual system does not simply average all cues with fixed weights. Instead, the contribution of each cue is related to its reliability and uncertainty [43]. Human visual responses to stimulus intensity also tend to be nonlinear and saturated, and such behavior can be approximated by sigmoid-like functions [44]. These studies do not directly provide a pooling formula for PCQA, but they suggest how the modality weights should behave in our metric. A modality with higher fidelity should receive a larger weight because it is more consistent between the reference and distorted point clouds. The weighting function should also vary smoothly with the fidelity score, thereby avoiding abrupt changes in the fused quality prediction caused by small score variations. In addition, bounded weights ensure that no single highly consistent modality completely dominates the fusion.

Based on this idea, MOSAIC treats each modality fidelity score, denoted as F_m below, as a simple cue-consistency measure. A low F_m is still kept in the fusion rather than removed. Since the final score is a weighted average of the fidelity values, this low-fidelity modality still lowers the final quality score. The sigmoid weight should therefore be read as cue reliability, not as distortion importance. A degraded cue is penalized through its lower fidelity value in the weighted average, while the sigmoid prevents an unstable low-consistency cue from taking over the fusion.

We therefore use a sigmoid function to map each fidelity score to a modality weight:

$$w_{F_m} = \frac{1}{1+\exp(-\alpha F_m)}, \quad F_m \in \{\mathrm{F}_G, \mathrm{F}_C, \mathrm{F}_{3D}, \mathrm{F}_S\}. \tag{11}$$

where α controls the slope of the nonlinear mapping. We set $\alpha = 1$ in all experiments to keep the weighting stable and avoid extra dataset-specific tuning. Since the sigmoid mapping is monotonic, it maps each weight to $(0, 1)$. Thus, a more consistent cue receives a larger but bounded weight. The saturation of the sigmoid also reduces unnecessary weight changes when a cue already has high fidelity. Here, the sigmoid is not used as an exact model of neural responses. It is used as a simple regularized mapping for adaptive cue fusion. The ablation in Table VI and the parameter sensitivity analysis in Table XII further examine this choice.

IV. Experiments

A. Experimental Settings

Datasets: We evaluate MOSAIC on five publicly accessible PCQA datasets, including SJTU [19], WPC [20], ICIP [45], LS-PCQA [46], and MJ-PCCD [47]. These datasets cover diverse distortion types and content characteristics, providing benchmarks for evaluating prediction accuracy across diverse contents and distortions.

- The **SJTU dataset** [19] comprises 9 reference point clouds and 378 distorted versions generated using seven distortion types, including octree compression, color noise, geometry Gaussian noise, downsampling, and their combinations, applied at six distortion levels.
- The **WPC dataset** [20] consists of 20 reference point clouds, each subjected to five types of distortions, including Gaussian noise, downsampling, and compression noise generated by MPEG G-PCC (Trisoup), G-PCC (Octree), and V-PCC, producing 740 distorted samples.
- The **ICIP dataset** [45] contains six human-body point clouds compressed using MPEG G-PCC (Octree), G-PCC (Trisoup), and V-PCC codecs at five bitrates, resulting in 90 impaired samples.
- The **LS-PCQA dataset** [46] includes 930 distorted point clouds created from 85 reference point clouds using 31 distortion types, where Part I is annotated with MOSs.
- The **MJ-PCCD dataset** [47] includes six reference point clouds and 213 distorted samples generated using G-PCC, V-PCC, and learning-based JPEG Pleno standard codecs.

TABLE II
PERFORMANCE EVALUATION OF TRAINING-FREE FR-PCQA METRICS ON DIFFERENT DATASETS. THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED USING RED BOLD, GREEN UNDERLINED, AND BLUE ITALIC TEXT, RESPECTIVELY. A DASH INDICATES THAT THE RESULT IS UNAVAILABLE OR NOT DIRECTLY REPORTED. AVERAGE VALUES ARE ARITHMETIC MEANS OVER DATASETS AND ARE REPORTED ONLY WHEN ALL DATASET RESULTS ARE AVAILABLE.

	SJTU [19]			**WPC** [20]			**ICIP** [45]			**LS-PCQA** [46]			**MJ-PCCD** [47]			**Average**		
	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
M-p2po [2]	0.728	0.801	1.717	0.578	0.566	18.714	0.947	0.935	0.365	0.398	0.185	0.763	0.434	0.314	1.054	0.617	0.560	4.523
M-p2pl [3]	0.652	0.661	1.893	0.454	0.446	20.422	0.954	0.948	0.341	0.351	0.166	0.779	0.396	0.370	1.074	0.561	0.518	4.902
PSNRyuv [4]	0.649	0.646	1.847	0.551	0.536	19.132	0.868	0.868	0.564	0.390	0.386	0.766	0.416	0.310	1.064	0.575	0.549	4.675
PCQM [5]	0.860	0.847	1.237	0.750	0.743	15.164	0.845	0.866	0.608	0.404	0.362	0.761	0.581	0.442	0.953	0.688	0.652	3.745
PointSSIM [24]	0.846	0.844	1.295	0.551	0.541	19.132	0.687	0.621	0.825	0.248	0.149	0.806	0.491	0.431	1.019	0.565	0.517	4.615
GraphSIM [6]	0.856	0.842	1.256	0.687	0.678	16.649	0.906	0.889	0.481	0.252	0.243	0.805	0.574	0.546	0.958	0.655	0.640	4.030
MS-GraphSIM [7]	0.899	0.890	1.031	0.709	0.704	16.172	0.482	0.021	1.001	0.327	0.307	0.786	–	–	–	–	–	–
MPED [25]	0.901	0.892	1.053	0.626	0.627	17.874	0.958	*0.953*	0.328	*0.540*	0.501	*0.700*	0.534	0.440	0.990	0.712	0.683	4.189
PQI [1]	0.912	0.906	*0.996*	*0.813*	*0.808*	*13.364*	0.966	0.955	0.292	0.437	0.379	0.748	0.582	0.481	0.952	0.742	0.706	3.270
TCDM [27]	**0.930**	**0.910**	**0.895**	0.806	0.804	13.557	0.942	0.935	0.383	0.335	0.325	0.784	**0.885**	**0.837**	**0.546**	*0.780*	*0.762*	*3.233*
PHM [28]	*0.906*	0.888	1.025	0.838	0.833	12.510	**0.969**	**0.958**	**0.280**	0.546	*0.500*	0.697	*0.780*	*0.690*	*0.733*	0.808	0.774	3.049
MOSAIC	0.912	*0.898*	0.994	**0.844**	**0.834**	**12.292**	*0.965*	0.955	*0.297*	**0.622**	**0.598**	**0.652**	0.831	0.755	0.651	**0.835**	**0.808**	**2.977**

TABLE III
PERFORMANCE EVALUATION ON DIFFERENT DISTORTION TYPES. THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED USING RED BOLD, GREEN UNDERLINED, AND BLUE ITALIC TEXT, RESPECTIVELY. AVERAGE VALUES ARE ARITHMETIC MEANS OVER THE FIVE DISTORTION GROUPS.

	G-PCC Octree			**G-PCC Trisoup**			**V-PCC**			**Downsampling**			**Gaussian Noise**			**Average**		
	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
M-p2po [2]	0.489	0.173	0.220	0.531	0.511	0.208	0.508	0.487	0.181	0.087	0.713	0.286	0.147	0.145	0.210	0.352	0.406	0.221
M-p2pl [3]	0.498	0.176	0.219	0.517	0.497	0.210	0.558	0.544	0.174	0.087	0.649	0.286	0.147	0.041	0.210	0.361	0.381	0.220
PSNRyuv [4]	0.700	0.675	0.180	0.618	0.621	0.193	0.464	0.382	0.186	0.631	0.631	0.223	0.729	0.682	0.145	0.628	0.598	0.185
PCQM [5]	**0.886**	**0.838**	**0.117**	0.791	0.792	0.150	*0.718*	*0.698*	*0.146*	0.886	0.840	0.133	0.893	0.869	0.096	0.835	0.807	*0.128*
PointSSIM [24]	0.449	0.134	0.225	0.567	0.562	0.203	0.606	0.592	0.167	0.714	0.701	0.201	0.747	0.710	0.141	0.617	0.540	0.187
GraphSIM [6]	0.830	0.781	0.141	0.785	0.785	0.152	0.499	0.436	0.182	**0.951**	0.905	**0.089**	0.876	0.853	0.102	0.788	0.752	0.133
MPED [25]	0.535	0.494	0.213	0.548	0.545	0.206	0.688	0.696	0.153	0.933	0.903	0.103	0.855	0.820	0.110	0.712	0.692	0.157
PQI [1]	0.776	0.743	0.159	0.737	0.742	0.166	0.582	0.560	0.171	*0.948*	**0.913**	*0.091*	0.916	0.901	0.085	0.792	0.772	0.134
TCDM [27]	0.686	0.619	0.183	0.846	0.845	0.121	0.744	0.742	0.141	0.943	0.894	0.096	0.897	*0.884*	0.094	*0.823*	*0.797*	0.127
PHM [28]	0.704	0.667	0.179	**0.874**	**0.869**	**0.120**	0.581	0.535	0.171	0.950	*0.908*	0.090	**0.929**	**0.913**	**0.079**	0.808	0.778	*0.128*
MOSAIC	*0.798*	*0.766*	*0.152*	*0.819*	*0.816*	*0.141*	**0.781**	**0.769**	**0.131**	*0.948*	0.909	0.092	*0.906*	0.872	*0.090*	**0.850**	**0.826**	**0.121**

Implementation Details: For neighborhood construction, $K_g = 100$, $K_p = 10$, and $K_c = 10$ are used for geometry (Eq. (1)), point density (Eq. (3)), and color feature (Eqs. (4) and (5)) extraction, respectively. MOSAIC adopts KNN-based neighborhoods rather than radius-based neighborhoods (see Sec. IV-D3) to balance structural coverage and computational cost. Following prior studies [38], [39], a relatively large K_g is used for local shape estimation, while smaller neighborhoods are used for density and color to preserve fine-scale point-spacing changes and local chromatic contrast. The nonlinear fusion parameter α is set to 1 in Eq. (11). We conduct our experiments on an Intel Xeon Silver 4314 CPU with 512 GB of RAM. All evaluated methods run on the CPU without GPU acceleration.

Evaluation Criteria: For all methods, predicted scores are mapped to MOS values using the five-parameter logistic function recommended by the ITU VQEG (Video Quality Expert Group) [48]:

$$\hat{y} = \beta_1 \left(0.5 - \frac{1}{1 + e^{\beta_2 (y - \beta_3)}} \right) + \beta_4 y + \beta_5, \qquad (12)$$

where y represents the predicted quality score, and $\beta_1, \ldots, \beta_5$ are optimized by minimizing the squared errors between the mapped scores $\hat{y}$ and the subjective MOS values. We report the Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Correlation Coefficient (SROCC), and root mean square error (RMSE), which measure prediction accuracy, monotonicity, and error, respectively. Higher PLCC and SROCC and lower RMSE indicate better performance. For the Average columns, PLCC, SROCC, and RMSE are averaged only when all corresponding dataset results are available. These averages provide an overview of whether a metric remains stable across several databases. For RMSE, the average should still be read together with dataset-wise RMSE because its absolute value depends on the subjective-score scale and logistic fitting range of each database.

B. Comparisons with Training-free Metrics

We compare MOSAIC with MSE-p2point (M-p2po) [2], MSE-p2plane (M-p2pl) [3], PSNRyuv [4], PointSSIM [24], PCQM [5], GraphSIM [6], MS-GraphSIM [7], MPED [25], PQI [1], TCDM [27], and PHM [28]. For training-free metrics, objective scores are evaluated directly on each dataset.

1) Dataset-wise Evaluation: Table II shows that no single metric dominates all datasets. TCDM gives the best results on SJTU and MJ-PCCD, while PHM is particularly strong on ICIP. The Average columns provide a compact summary over datasets with complete results and show that MOSAIC achieves the best average PLCC/SROCC/RMSE. MOSAIC

TABLE IV
AVERAGE POINT COUNTS AND COMPUTATIONAL TIME OF REPRESENTATIVE FR-PCQA METRICS ON FIVE PUBLIC DATABASES. RUNTIME IS REPORTED IN SECONDS PER POINT-CLOUD PAIR.

Database	Average number of points		Runtime (s)								
	Reference	Distorted	PSNRyuv [4]	PCQM [5]	PointSSIM [24]	GraphSIM [6]	MPED [25]	PHM [28]	TCDM [27]	PQI [1]	MOSAIC
SJTU	1,008,541	706,684	10.30	3.55	4.24	11.68	7.60	24.04	5.22	4.71	6.41
WPC	1,359,252	1,926,565	18.81	4.08	8.00	22.38	11.70	38.60	8.77	7.88	15.08
ICIP	870,573	613,458	5.54	3.55	2.86	10.11	4.73	17.88	4.88	3.98	5.23
LS-PCQA	789,833	710,084	5.87	3.62	3.08	10.22	4.84	16.88	5.05	4.14	5.55
MJ-PCCD	2,930,424	4,353,776	38.56	43.38	21.07	49.46	19.20	92.55	35.50	21.62	22.11

also obtains the best performance on WPC and LS-PCQA, with (PLCC/SROCC/RMSE) values of (0.844/0.834/12.292) and (0.622/0.598/0.652), respectively. These two datasets are particularly useful for assessing performance on diverse contents and distortions, as they are large and challenging [20], [46].

For the datasets where MOSAIC is not the best method, its performance remains close to the leading result. On MJ-PCCD, MOSAIC ranks second on all three criteria and clearly improves over PHM and PQI. On SJTU, it ties PQI in PLCC and gives the second-best RMSE. On ICIP, PHM and PQI are slightly better, which is consistent with their strong behavior on compression-oriented databases.

2) Distortion-wise Evaluation: To evaluate robustness under different distortion mechanisms, we conduct a distortion-wise analysis by integrating samples from SJTU [19], WPC [20], ICIP [45], and MJ-PCCD [47]. Five representative distortion types are considered. G-PCC Octree Compression includes 182 distorted point clouds, with 80 samples from WPC, 30 from ICIP, and 72 from MJ-PCCD. G-PCC Trisoup Compression includes 270 samples, with 240 from WPC and 30 from ICIP. V-PCC Compression includes 279 samples, with 180 from WPC, 30 from ICIP, and 69 from MJ-PCCD. Downsampling includes 114 samples, with 54 from SJTU and 60 from WPC. Gaussian Noise includes 234 samples, with 54 from SJTU and 180 from WPC.

Table III shows that MOSAIC remains in the leading group across different distortion mechanisms. The Average columns show the best average PLCC/SROCC/RMSE across the five distortion groups, while individual categories reveal complementary strengths among PCQM, TCDM, PHM, PQI, and MOSAIC. These results indicate consistent performance across the evaluated distortion types: MOSAIC does not rely on distortion-specific switching or supervised fitting, but still remains competitive under compression, downsampling, and noise distortions.

The main exception is G-PCC Octree, where PCQM performs best. The gap mainly appears in lossless-geometry and lossy-attribute cases, which is consistent with the metrics' formula designs. Following its original formulation, PCQM represents local geometry/color evidence as distance-based distortion features [5]. The final score is a linear combination, $\mathrm{PCQM} = \sum_i w_i f_i$; when geometry is nearly unchanged, it is driven mainly by attribute-distance terms. By contrast, TCDM [27] builds geometry- and color-channel complexity similarities from self- and transformational-complexity terms, multiplies them, and then combines this product with a prediction-based feature. PHM [28] multiplicatively fuses visible-difference and appearance-degradation terms, and MOSAIC averages normalized geometry, color, and saliency fidelities with bounded adaptive weights. These similarity- or product-style formulations are generally robust, but they may be less discriminative for subtle attribute-only changes when the geometry cue remains almost perfect. This distance-versus-similarity behavior deserves further study.

These results indicate that MOSAIC, TCDM, and PHM form a leading group of recent training-free FR-PCQA metrics. A common reason is that all three revisit PCQA from a new perceptual or computational perspective, rather than only refining conventional point-wise errors. TCDM views quality degradation through transformational complexity. PHM uses a hybrid strategy to model different HVS responses at different distortion levels. MOSAIC treats FR-PCQA as multimodal cue fusion among 3D structure, 3D appearance, and projected 2D saliency. Within this group, MOSAIC's value lies in consistent performance across the evaluated datasets and distortion types through an interpretable two-stage fusion design, without supervised fitting or distortion-specific switching.

3) Computational Complexity: We evaluate computational efficiency on five datasets using the implementations summarized in Table IV.

We report average runtime per point-cloud pair and average reference and distorted point counts for each database. Runtimes are generally higher on databases with larger point clouds, with the highest values observed on MJ-PCCD. This trend also reflects differences in point density and spatial distribution; therefore, it does not isolate the effect of point count. For MOSAIC, the unweighted arithmetic mean of the five database-level average runtimes is 10.88 s, placing it in the middle range among the evaluated implementations. Its mean runtime is lower than those of PSNRyuv, PCQM, GraphSIM, PHM, and TCDM, but higher than those of PointSSIM, MPED, and PQI. Because the methods use different software environments (C/C++, MATLAB, and Python), these measurements provide practical runtime references rather than an implementation-independent efficiency ranking.

C. Comparison with Regression- & Learning-Based Methods

Regression-based and learning-based FR-PCQA methods are useful references, but we keep them separate from the main

TABLE V
PLCC/SROCC COMPARISON BETWEEN MOSAIC AND REPRESENTATIVE REGRESSION- OR LEARNING-BASED FR-PCQA REFERENCES ON OVERLAPPING DATASETS.

Method	SJTU [19]	WPC [20]	ICIP [45]	MJ-PCCD [47]
MOSAIC	0.912/0.898	0.844/0.834	0.965/0.955	0.831/0.755
PointPCA [9]	0.909/0.879	0.891/0.885	–	–
FRSVR [10]	–	0.811/0.796	0.956/0.935	0.709/0.620
SGW-PCQA+ [11]	–	0.873/0.869	–	–
FR-MLLM [13]	0.935/0.915	0.889/0.881	–	–
CT-PCQA (FR) [12]	0.929/0.910	0.881/0.878	–	–

training-free comparison in Table II. Their reported scores may also reflect subjective-label fitting, evaluation splits, learned representations, or SVR-based score integration. Accordingly, Table V is used to provide context rather than a strict ranking.

Table V uses the best PLCC/SROCC values reported in the original papers and keeps only datasets shared with our evaluation. The comparison is therefore restricted to shared datasets, although the reported values still come from each paper's own protocol. Under this practical comparison, MOSAIC is competitive with regression- and learning-based references. Its reported PLCC/SROCC values are slightly higher than those of PointPCA on SJTU and FRSVR on WPC, ICIP, and MJ-PCCD, and are relatively close to those of SGW-PCQA+, FR-MLLM, and CT-PCQA on the shared SJTU/WPC datasets. These methods can benefit from subjective-label fitting or learned representations, whereas MOSAIC uses deterministic training-free cues.

D. Ablation Studies

This section probes the potential of MOSAIC by isolating its internal design choices, including multimodal cue fusion, saliency formulation, neighborhood construction, color space, projection strategy, keypoint selection, and parameter sensitivity.

TABLE VI
MULTIMODAL CUE FUSION STRATEGIES.

Scheme	Fusion	WPC [20]			LS-PCQA [46]		
		PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
Only 3D	Adaptive	0.793	0.775	13.973	0.523	0.493	0.709
One-stage	Uniform	0.835	0.823	12.622	0.564	0.533	0.687
	Adaptive	0.836	0.825	12.573	0.586	0.560	0.674
Two-stage	Uniform	0.844	0.833	12.296	0.592	0.566	0.671
	Adaptive	**0.844**	**0.834**	**12.292**	**0.622**	**0.598**	**0.652**

1) Multimodal Cue Fusion: Table VI evaluates the strategy on how to fuse multimodal cues. The "Only 3D" baseline uses intrinsic 3D geometry and color cues without projected 2D saliency. The one-stage variants combine all cues in a single step, whereas the two-stage variants first fuse geometry and color into an intrinsic 3D fidelity score and then incorporate saliency. Uniform and adaptive denote linear averaging and sigmoid-based weighting, respectively. We choose WPC and LS-PCQA for this ablation because they are large-scale and challenging, and they provide a clear view of the contribution of each design choice.

Replacing the "Only 3D" variant with multimodal fusion improves PLCC/SROCC from 0.793/0.775 to 0.844/0.834 on WPC, and from 0.523/0.493 to 0.622/0.598 on LS-PCQA. This indicates that projected 2D saliency provides perceptual information that is not fully captured by intrinsic 3D geometry and color fidelity.

Among the evaluated fusion schemes, the two-stage adaptive variant achieves the best or tied-best results for all three criteria on WPC and LS-PCQA. Compared with one-stage adaptive fusion, it increases PLCC/SROCC from 0.836/0.825 to 0.844/0.834 and reduces RMSE from 12.573 to 12.292 on WPC. On LS-PCQA, PLCC/SROCC increase from 0.586/0.560 to 0.622/0.598, and RMSE decreases from 0.674 to 0.652. Since both variants use the same cues, these results support the contribution of the two-stage fusion design on the evaluated databases.

2) Saliency Formulation: We compare the proposed saliency-map consistency formulation (MOSAIC-SMC) with a reference-saliency-weighted pooling variant (MOSAIC-RSW), while keeping the intrinsic 3D geometry/color branch, the six projection views, and the final fusion rule unchanged. SMC averages SSIM between the reference and distorted saliency maps over the six views (Eq. (8)). RSW instead pools a local SSIM map between the projection images using normalized reference saliency weights, then averages the view scores. Thus, SMC measures changes in predicted saliency, whereas RSW measures local quality weighted by reference saliency. A stable saliency layout can yield high SMC despite local degradation, and thus SMC serves as a complementary fidelity cue.

MOSAIC-SMC performs better on SJTU, WPC, ICIP, and MJ-PCCD, whereas MOSAIC-RSW performs better on LS-PCQA. The largest SMC gain occurs on WPC. Structural artifacts there may alter projected boundaries and predicted saliency, including regions receiving low reference saliency weights. In LS-PCQA, appearance distortions may reduce local image similarity while leaving the saliency layout relatively stable, favoring RSW. These results support retaining SMC within MOSAIC's two-stage fusion, while also showing that its advantage depends on the distortion characteristics.

3) Neighborhood Construction: In the local shape feature extraction (Eq. (1)), neighborhoods can be constructed using either KNN-based or radius-based strategies. Table VIII shows that KNN-based neighborhoods generally provide more stable performance across datasets, while radius-based neighborhoods are comparable only in some cases. Runtime is also important for large-scale point clouds. In our implementation, radius-based neighborhood construction incurs substantially higher search cost, with an average runtime approximately 16× higher than the KNN-based strategy on SJTU dataset. MOSAIC therefore adopts KNN-based neighborhoods to balance prediction accuracy and computational efficiency.

4) Color Space: MOSAIC adopts RGB because it is consistent with the stored color attributes of point-cloud datasets. We also evaluate MOSAIC using YUV and CIE-LAB (LAB), as reported in Table IX. RGB achieves the best overall

TABLE VII
COMPARISON OF SALIENCY-MAP CONSISTENCY AND REFERENCE-SALIENCY-WEIGHTED POOLING. THE BEST RESULT FOR EACH DATABASE AND EVALUATION CRITERION IS SHOWN IN **BOLD**.

	SJTU [19]			**WPC** [20]			**ICIP** [45]			**LS-PCQA** [46]			**MJ-PCCD** [47]			**Average**		
	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
MOSAIC-SMC	**0.912**	**0.898**	**0.994**	**0.844**	**0.834**	**12.292**	**0.965**	**0.955**	**0.297**	0.622	0.598	0.652	**0.831**	**0.755**	**0.651**	**0.835**	**0.808**	**2.977**
MOSAIC-RSW	0.856	0.833	1.254	0.658	0.629	17.267	0.934	0.931	0.406	**0.692**	**0.670**	**0.600**	0.760	0.687	0.761	0.780	0.749	4.058

TABLE VIII
ON NEIGHBORHOOD CONSTRUCTION FOR LOCAL SHAPE FEATURES. EACH CELL REPORTS PLCC/SROCC/RMSE.

Dataset	**KNN**	**Radius**
SJTU [19]	0.912/0.898/0.994	0.880/0.873/1.154
WPC [20]	0.844/0.834/12.292	0.826/0.817/12.911
ICIP [45]	0.965/0.955/0.297	0.964/0.954/0.302
LS-PCQA [46]	0.622/0.598/0.652	0.589/0.566/0.672
MJ-PCCD [47]	0.831/0.755/0.651	0.838/0.753/0.638

performance on most datasets. Although YUV and LAB are slightly better on LS-PCQA, these gains are marginal and come with noticeable degradation on other datasets. Moreover, using RGB avoids the need for color space conversion, which can introduce additional computational cost and potential numerical errors. We therefore keep RGB as the default color representation in this work.

5) Projection Protocol: Table X examines how the projected 2D saliency cue is affected when only the 3D-to-2D view-generation protocol is changed. In all variants, MOSAIC uses the RGB projection images for saliency extraction and keeps the subsequent saliency comparison and cue fusion unchanged. The three protocols are selected as representative view-generation families in projection-based PCQA: O6P represents deterministic fixed-grid orthogonal projection, FVC represents static virtual-camera rendering, and MCV represents trajectory-based moving-camera rendering.

- **O6P** (orthogonal six-view projection) [19] normalizes the point cloud to a fixed projection grid and rasterizes it onto the six faces of the axis-aligned bounding cube, namely front/back, left/right, and top/bottom. This coordinate-grid projection gives deterministic coverage without virtual-camera placement. O6P is the default projection protocol in MOSAIC.
- **FVC** (fixed virtual-camera projection) [49] renders RGB views from fixed virtual cameras around the point cloud. The point cloud is normalized before rendering, and the virtual cameras keep a fixed viewing distance so that the projected texture scale remains stable.
- **MCV** (moving-camera video projection) [50] generates frames by rotating the camera around the point cloud along circular trajectories, forming video clips for spatial-temporal VQA feature extraction; here, only the sampled RGB frames are used as saliency views.

TABLE IX
PERFORMANCE WITH VARIOUS COLOR SPACES. EACH CELL REPORTS PLCC/SROCC/RMSE. THE BEST VALUES WITHIN EACH DATASET AND METRIC ARE MARKED IN **BOLD**.

Dataset	**RGB**	**YUV**	**LAB**
SJTU [19]	**0.912/0.898/0.994**	0.882/0.849/1.142	0.856/0.817/1.256
WPC [20]	**0.844/0.834/12.292**	0.799/0.786/13.775	0.756/0.741/15.001
ICIP [45]	**0.965/0.955/0.297**	0.940/0.934/0.387	0.882/0.890/0.534
LS-PCQA [46]	0.622/0.598/0.652	**0.651**/0.638/**0.632**	0.640/**0.645**/0.640
MJ-PCCD [47]	**0.831/0.755/0.651**	0.781/0.734/0.731	0.741/0.676/0.785

O6P gives the most balanced performance: it is best on SJTU, WPC, and MJ-PCCD and also obtains the best Average PLCC/SROCC/RMSE of 0.835/0.808/2.977. FVC slightly improves ICIP, and MCV slightly improves LS-PCQA, indicating that alternative view-generation rules may fit certain content distributions. However, these gains are not stable across datasets when the rest of MOSAIC is kept unchanged. We therefore keep O6P as the default projection protocol because it gives the strongest average performance and uses deterministic fixed-grid coverage.

6) Keypoint Selection: Table XI compares five sampling protocols while keeping all other parts of MOSAIC unchanged. **All** is the default setting in MOSAIC and uses every reference point. **FPS** uses farthest point sampling with the $N/1000$ seed ratio adopted in PHM's Voronoi-patch construction [28]. **RS** randomly selects 400 seeds, and **FastFPS** first forms a random candidate subset and then applies FPS; these two rows are 400-seed control variants motivated by the seed budget used in TCDM's Voronoi partition [27]. **HGS** uses the high-pass graph sampling idea from GraphSIM [6] to select geometry-aware keypoints, here with $N/10000$ sampled points.

Using all points gives the best performance on SJTU, WPC, ICIP, and MJ-PCCD, and the Average columns further show its overall advantage. Sparse keypoint strategies slightly improve LS-PCQA, but they substantially degrade the other four datasets. This behavior is consistent with the role of sampling in the original methods: graph or Voronoi metrics can use sparse seeds as patch centers, whereas MOSAIC measures point-aligned local geometry, color, and saliency fidelities. If too many points are removed, local color variation, density changes, and attention-related evidence may be lost. The slight LS-PCQA gain suggests that sparse seeds can sometimes suppress noisy fine-scale matches, but the overall results favor full point coverage. Since the all-point setting still presents moderate runtime in Table IV, it is used as the default strategy in this work.

7) Parameter Sensitivity: To examine whether MOSAIC relies on parameter tuning, we conduct a one-at-a-time sensitivity analysis for K_g, K_p, K_c, and α, as reported in Table XII. In each non-default row, only the parameter named in the first column is changed, while the other three parameters remain

TABLE X
PROJECTION PROTOCOLS FOR THE SALIENCY CUE. AVERAGE VALUES ARE ARITHMETIC MEANS OVER THE FIVE DATASETS; BEST VALUES ARE IN **BOLD**.

	SJTU [19]			**WPC** [20]			**ICIP** [45]			**LS-PCQA** [46]			**MJ-PCCD** [47]			**Average**		
	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
O6P [19]	**0.912**	**0.898**	**0.994**	**0.844**	**0.834**	**12.292**	0.965	0.955	0.297	0.622	0.598	0.652	**0.831**	**0.755**	**0.651**	**0.835**	**0.808**	**2.977**
FVC [49]	0.899	0.885	1.063	0.817	0.804	13.204	**0.969**	**0.957**	**0.281**	0.632	0.607	0.644	0.740	0.656	0.787	0.811	0.782	3.196
MCV [50]	0.894	0.881	1.086	0.810	0.798	13.431	0.966	0.953	0.293	**0.636**	**0.608**	**0.642**	0.696	0.629	0.840	0.800	0.774	3.258

TABLE XI
ON KEYPOINT SELECTION. AVERAGE VALUES ARE ARITHMETIC MEANS OVER THE FIVE DATASETS; BEST VALUES ARE IN **BOLD**.

		SJTU [19]			**WPC** [20]			**ICIP** [45]			**LS-PCQA** [46]			**MJ-PCCD** [47]			**Average**		
Method	**Keypoints**	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
All	N	**0.912**	**0.898**	**0.994**	**0.844**	**0.834**	**12.292**	**0.965**	**0.955**	**0.297**	0.622	0.598	0.652	**0.831**	**0.755**	**0.651**	**0.835**	**0.808**	**2.977**
FPS [28]	$N/1000$	0.745	0.744	1.620	0.751	0.745	15.149	0.816	0.822	0.656	**0.647**	**0.633**	**0.635**	0.828	0.670	0.656	0.757	0.723	3.743
RS	400	0.742	0.739	1.628	0.752	0.748	15.105	0.818	0.823	0.653	0.641	0.623	0.638	0.824	0.661	0.663	0.755	0.719	3.737
FastFPS	400	0.732	0.732	1.653	0.753	0.749	15.088	0.821	0.828	0.650	0.641	0.622	0.639	0.819	0.654	0.672	0.753	0.717	3.740
HGS [6]	$N/10000$	0.716	0.719	1.695	0.740	0.735	15.413	0.806	0.814	0.673	0.634	0.611	0.643	0.782	0.613	0.730	0.736	0.698	3.831

TABLE XII
ONE-AT-A-TIME PARAMETER SENSITIVITY ANALYSIS. AVERAGE VALUES ARE ARITHMETIC MEANS OVER THE FIVE DATASETS.

	SJTU [19]			**WPC** [20]			**ICIP** [45]			**LS-PCQA** [46]			**MJ-PCCD** [47]			**Average**		
	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
Default	**0.912**	**0.898**	**0.994**	**0.844**	**0.834**	**12.292**	**0.965**	**0.955**	**0.297**	**0.622**	**0.598**	**0.652**	**0.831**	**0.755**	**0.651**	**0.835**	**0.808**	**2.977**
$K_g = 10$	0.892	0.882	1.096	0.846	0.837	12.231	0.969	0.964	0.279	0.628	0.604	0.648	0.807	0.729	0.691	0.828	0.803	2.989
$K_g = 50$	0.912	0.896	0.997	0.842	0.832	12.364	0.969	0.962	0.281	0.624	0.601	0.650	0.825	0.745	0.661	0.834	0.807	2.991
$K_p = 50$	0.902	0.890	1.047	0.832	0.822	12.726	0.968	0.958	0.285	0.614	0.586	0.657	0.865	0.774	0.587	0.836	0.806	3.060
$K_p = 100$	0.901	0.889	1.053	0.831	0.822	12.752	0.968	0.957	0.285	0.609	0.582	0.660	0.871	0.775	0.576	0.836	0.805	3.065
$K_c = 50$	0.897	0.876	1.074	0.842	0.835	12.376	0.960	0.947	0.319	0.614	0.600	0.657	0.815	0.754	0.678	0.826	0.802	3.021
$K_c = 100$	0.876	0.849	1.170	0.834	0.828	12.642	0.952	0.942	0.349	0.606	0.596	0.661	0.801	0.744	0.700	0.814	0.792	3.104
$\alpha = 0.5$	0.912	0.898	0.994	0.844	0.834	12.285	0.965	0.955	0.297	0.607	0.586	0.661	0.830	0.754	0.652	0.832	0.805	2.978
$\alpha = 3$	0.912	0.898	0.994	0.844	0.834	12.303	0.965	0.955	0.297	0.640	0.612	0.639	0.831	0.754	0.651	0.838	0.811	2.977

fixed at the default setting. This design is intentional. The goal is not to search for the best parameter combination, but to test whether the proposed cue extraction and fusion remain stable when one design factor is perturbed. A joint search over K_g, K_p, K_c, and α would mix the effects of local shape scale, density scale, color scale, and fusion slope, and could introduce dataset-specific tuning that is inconsistent with the training-free setting.

The default setting remains competitive across datasets, indicating that the metric is not highly sensitive to moderate parameter changes. Reducing the geometry neighborhood size K_g to 10 or 50 causes only limited fluctuations on most datasets, suggesting that the structural representation is stable once sufficient local geometry is captured. Increasing the density neighborhood size K_p improves MJ-PCCD but degrades SJTU, WPC, and LS-PCQA, indicating that overly large density neighborhoods may smooth fine-scale sampling changes. A similar effect appears for the color neighborhood size K_c, where larger neighborhoods dilute local chromatic differences. For the nonlinear mapping parameter, $\alpha = 3$ gives slightly higher average PLCC/SROCC, mainly due to the LS-PCQA gain, while $\alpha = 1$ remains nearly tied in average RMSE and avoids a sharper weighting curve. We therefore use the default parameters for all datasets.

V. CONCLUSION

This paper examined FR-PCQA through the lens of multimodal perceptual cue organization. The main insight is that 3D geometry and color describe how the point cloud itself is degraded, whereas projected 2D saliency consistency captures distortion-induced changes in predicted saliency across rendered views. In MOSAIC, we first integrate local geometric structure and color appearance into an intrinsic 3D fidelity representation, then incorporate projected 2D saliency consistency as a complementary view-dependent cue. Experiments on five public PCQA datasets show that this training-free formulation is sufficient to place MOSAIC in the leading group of recent FR-PCQA metrics, with strong average performance, clear gains on WPC and LS-PCQA, and competitive results against representative learning-based references without MOS-supervised score fitting.

The current formulation also has several scope limitations. First, the saliency branch uses fixed six-view orthogonal projections, which keeps the metric deterministic but does not model user-specific viewing trajectories, dynamic head motion, or viewpoint-dependent occlusion changes in interactive rendering. Second, we fix neighborhood sizes and fusion slope across all datasets to preserve reproducibility, although point clouds with very different densities, scales, or acquisition pipelines may benefit from adaptive parameter selection. Future work will therefore explore view-adaptive or trajectory-aware saliency cues, density-aware neighborhood construction, and more efficient implementations for large-scale point clouds.

ACKNOWLEDGMENT

The authors would like to thank the researchers who created and publicly released the SJTU-PCQA [19], WPC [20],

ICIP [45], LS-PCQA [46], and MJ-PCCD [47] databases, together with their reference point clouds, distorted samples, and subjective quality scores. The authors also gratefully acknowledge the authors of the compared FR-PCQA metrics, including PCQM [5], PointSSIM [24], GraphSIM [6], MS-GraphSIM [7], MPED [25], PQI [1], TCDM [27], and PHM [28], for making source code, evaluation scripts, or reproducible benchmark materials available to the community. These public resources were essential for building fair and transparent comparisons in this study.

REFERENCES

[1] J. Zhang, T. Chen, D. Ding, and Z. Ma, "Revisit point cloud quality assessment: Current advances and a multiscale-inspired approach," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 10, pp. 8886–8899, 2025.

[2] R. Mekuria, Z. Li, C. Tulvan, and P. Chou, "Evaluation criteria for PCC (Point Cloud Compression)," *ISO/IEC MPEG n16332*, Feb. 2016.

[3] 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.

[4] E. M. Torlig, E. Alexiou, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi, "A novel methodology for quality assessment of voxelized point clouds," in *Applications of Digital Image Processing XLI*, vol. 10752. SPIE, 2018, pp. 174–190.

[5] 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.

[6] Q. Yang, Z. Ma, Y. Xu, Z. Li, and J. Sun, "Inferring point cloud quality via graph similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3015–3029, 2022.

[7] Y. Zhang, Q. Yang, and Y. Xu, "MS-GraphSIM: Inferring point cloud quality via multiscale graph similarity," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1230–1238.

[8] Z. Wang, Y. Zhang, Q. Yang, Y. Xu, J. Sun, and S. Liu, "Point cloud quality assessment using 3d saliency maps," *arXiv preprint arXiv:2209.15475*, 2022.

[9] E. Alexiou, X. Zhou, I. Viola, and P. Cesar, "PointPCA: point cloud objective quality assessment using PCA-based descriptors," *Journal on Image and Video Processing*, vol. 2024, no. 1, Aug. 2024.

[10] R. Watanabe, S. N. Sridhara, H. Hong, E. Pavez, K. Nonaka, T. Kobayashi, and A. Ortega, "Full reference point cloud quality assessment using support vector regression," *Signal Processing: Image Communication*, vol. 131, p. 117239, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596524001401

[11] R. Watanabe, K. Nonaka, E. Pavez, T. Kobayashi, and A. Ortega, "Full-reference point cloud quality assessment using spectral graph wavelets," in *IEEE International Conference on Image Processing (ICIP)*, 2024, pp. 3313–3319.

[12] Y. Zhou, Z. Zhang, W. Sun, X. Min, and G. Zhai, "CT-PCQA: A convolutional neural network and transformer combined method for point cloud quality assessment," *Signal Processing: Image Communication*, vol. 138, p. 117371, 2025.

[13] R. Watanabe, T. Konno, H. Sankoh, B. Tanaka, and T. Kobayashi, "Full-reference point cloud quality assessment with multimodal large language models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[14] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1982.

[15] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young, "Measurement and modeling of depth cue combination: In defense of weak fusion," *Vision Research*, vol. 35, no. 3, pp. 389–412, 1995.

[16] D. C. Knill and J. A. Saunders, "Do humans optimally integrate stereo and texture information for judgments of surface slant?" *Vision Research*, vol. 43, no. 24, pp. 2539–2558, 2003.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] 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.

[19] 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.

[20] 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.

[21] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, "Joint geometry and color projection-based point cloud quality metric," *IEEE Access*, vol. 10, pp. 90 481–90 497, 2022.

[22] Q. Liu, H. Yuan, H. Su, H. Liu, Y. Wang, H. Yang, and J. 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.

[23] 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*, vol. 34, no. 7, pp. 6334–6346, 2024.

[24] E. Alexiou and T. Ebrahimi, "Towards a point cloud structural similarity metric," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020.

[25] 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, 2022.

[26] Y. Xu, Y. Zhang, Q. Yang, X. Xu, and S. Liu, "Compressed point cloud quality index by combining global appearance and local details," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 9, pp. 291:1–291:22, 2024.

[27] 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*, vol. 30, no. 10, pp. 6707–6724, 2024.

[28] Y. Zhang, Q. Yang, Y. Xu, and S. Liu, "Perception-guided quality metric of 3d point clouds using hybrid strategy," *IEEE Transactions on Image Processing*, vol. 33, pp. 5755–5770, 2024.

[29] X. Zhou, E. Alexiou, I. Viola, and P. Cesar, "PointPCA+: A full-reference point cloud quality assessment metric with PCA-based features," *Signal Processing: Image Communication*, vol. 135, p. 117262, 2025.

[30] J. Prazeres, R. Rodrigues, M. Pereira, and A. M. G. Pinheiro, "Objective quality evaluation of point clouds: Quality features and fusion metrics," *IEEE Access*, vol. 14, pp. 83 029–83 043, 2026.

[31] M. Tliba, A. Chetouani, G. Valenzise, and F. Dufaux, "Point cloud quality assessment using cross-correlation of deep features," in *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, 2022, pp. 63–68.

[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.

[33] 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, pp. 10 012–10 022.

[34] Z. Zhang, H. Wu, Y. Zhou, C. Li, W. Sun, C. Chen, X. Min, X. Liu, W. Lin, and G. Zhai, "LMM-PCQA: Assisting point cloud quality assessment with LMM," *arXiv preprint arXiv:2404.18203*, 2024.

[35] X. Chai, F. Shao, B. Mu, H. Chen, Q. Jiang, and Y.-S. 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*, vol. 34, no. 7, pp. 6207–6223, 2024.

[36] R. Li, X. Li, T. Tian, Y. Yi, and X. Ma, "A multi-view projection and 3D feature fusion model for full-reference point cloud quality assessment," *Information*, vol. 17, no. 9, p. 823, 2026.

[37] M. Cui, Y. Zhang, C. Fan, R. Hamzaoui, and Q. Li, "Colored point cloud quality assessment using complementary features in 3D and 2D spaces," *IEEE Transactions on Multimedia*, vol. 26, pp. 11 111–11 125, 2024.

[38] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Surface reconstruction from unorganized points," in *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, 1992, pp. 71–78.

[39] H. Knutsson, C.-F. Westin, and M. Andersson, "Representing local structure using tensors ii," in *Image Analysis: 17th Scandinavian Conference*. Springer, 2011, pp. 545–556.
[40] W. Zhou, G. Yue, R. Zhang, Y. Qin, and H. Liu, "Reduced-reference quality assessment of point clouds via content-oriented saliency projection," *IEEE Signal Processing Letters*, vol. 30, p. 354–358, 2023.
[41] H. R. Tavakoli and J. Laaksonen, "Bottom-up fixation prediction using unsupervised hierarchical models," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 287–302.
[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[43] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.
[44] L. N. Vinke, I. M. Bloem, and S. Ling, "Saturating nonlinearities of contrast response in human visual cortex," *The Journal of Neuroscience*, vol. 42, no. 7, pp. 1292–1302, 2022. [Online]. Available: https://doi.org/10.1523/JNEUROSCI.0106-21.2021
[45] S. Perry, H. P. Cong, L. A. da Silva Cruz, J. Prazeres, M. Pereira, A. Pinheiro, E. Dumic, E. Alexiou, and T. Ebrahimi, "Quality evaluation of static point clouds encoded using MPEG codecs," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3428–3432.
[46] 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*, vol. 19, no. 2s, pp. 1–26, 2023.
[47] D. Lazzarotto, M. Testolina, and T. Ebrahimi, "Subjective performance evaluation of bitrate allocation strategies for MPEG and JPEG Pleno point cloud compression," *Journal of Image and Video Processing*, vol. 2024, no. 14, 2024. [Online]. Available: https://doi.org/10.1186/s13640-024-00629-0
[48] J. Antkowiak, J. Baina, V. Baroncini, N. Chateau, A. C. F. Pessoa, S. Colonnese, L. Contin, and J. Caviedes, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Video Quality Experts Group (VQEG), Tech. Rep., Mar. 2000.
[49] 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.
[50] 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*, vol. 27, pp. 927–939, 2025.



Jiaxin Li is with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. Jiaxin Li's research interests include point cloud processing, visual quality assessment, and immersive visual media.



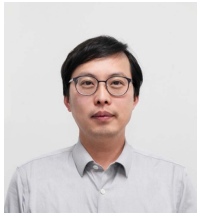
Tong Chen is with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. Tong Chen's research interests include multimedia signal processing, point cloud quality assessment, and perceptual visual computing.



Zehong Li is with Hangzhou Normal University, Hangzhou, China. Zehong Li's research interests include point cloud analysis, visual quality modeling, and 3D media processing.



Dandan Ding is with Hangzhou Normal University, Hangzhou, China. Dandan Ding's research interests include visual perception, multimedia quality assessment, and immersive media processing.



Zhan Ma is with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. Zhan Ma's research interests include image and video processing, point cloud coding, visual quality assessment, and multimedia communication.