Inferring Point Cloud Quality via Graph Similarity

Qi Yang, Zhan Ma, Senior Member, IEEE, Yiling Xu, Member, IEEE, Zhu Li, Senior Member, IEEE, and Jun Sun

Abstract—Objective quality estimation of media content plays a vital role in vast applications. Though numerous metrics have been successfully devised for 2D image or video, it still lacks of a counterpart for emerging 3D point clouds with unstructured and sparsely distributed points. We propose the GraphSIM – an objective metric to accurately predict the subjective quality of point cloud with superimposed geometry and color impairments. Motivated by the facts that human vision system is more sensitive to the high spatial-frequency components (e.g., contours, edges), and weighs more to the local structural variations rather individual point intensity, we first extract geometric keypoints by resampling the reference point cloud geometry information to form the object skeleton; we then construct local graphs centered at these keypoints for both reference and distorted point clouds, followed by collectively aggregating color gradient moments (e.g., zeroth, first, and second) that are derived between all other points and centered keypoint in the same local graph for significant feature similarity (a.k.a., local significance) measurement; Final similarity index is obtained by pooling the local graph significance across all color channels and by averaging across all graphs. Our GraphSIM is validated using two large and independent point cloud assessment datasets that involve a wide range of impairments (e.g., re-sampling, compression, additive noise), reliably demonstrating the state-of-the-art performance for all distortions with noticeable gains in predicting the subjective mean opinion score (MOS), compared with those point-wise distance-based metrics adopted in standardization reference software. Ablation studies have further shown that GraphSIM is generalized to various scenarios with consistent performance by examining its key modules and parameters. Models and associated materials will be made to public at https://njuvision.github.io/GraphSIM

Index Terms—Objective quality assessment, human perception, graph signal processing, point cloud

1 INTRODUCTION

With the advancements of 3D capturing and rendering technologies [1], point cloud has emerged as a promising format for representing 3D object and scene realistically [2]. Point cloud is a composite of a large number of unstructured 3D points associated with corresponding attributes (e.g., RGB color, normal, opacity, etc). These points are often scattered in a 3D space sparsely. Recently, we have witnessed a great deal of efforts that are devoted to drive the point cloud-based applications, such as resampling [3], enhancement [4], [5], saliency detection [6], classification [7], [8], segmentation [9], [10], and compression [2], [11]–[15]. A variety of noises would be inevitably induced by these processing techniques, impairing the reconstruction quality perceived by the human visual system (HVS).

Observation. To the best of our knowledge, most explorations are still applying the point-wise error measurements, such as the peak signal-to-noise ratio (PSNR) of Haubsdorff distance. Such point-wise distance-based evaluations, however, do not accurately reflect the perceptual sensation of our HVS as revealed in many pioneering assessment studies [16]. For 2D images, researchers have developed the structural similarity index (SSIM) [17] for better exploiting the HVS characteristics when assessing the image quality. Whereas, it still lacks of an efficient objective metric for point cloud quality evaluation. One reason is that, for a given point cloud, its scattered 3D points are unstructured without explicit connections, resulting in difficulties on how to arrange points for quantitative and effective quality measurement; it also implicitly involves the superimposed subjective impacts of 3D geometry and associated attributes [1] that are not typically existed for 2D image/video; and furthermore, associated processing, e.g., compression, may introduce a noticeable variations of the total number of points in the same point cloud, leading to issues on how to perform the fair comparison (especially using point-wise metrics). All of these observations of point cloud are greatly distant from the existing studies for 2D image or video, setting obstacles for its efficient quality assessment and modeling.

Perception. Our work is highly motivated by the perceptual intuitions. Our vision system exhibits feedforward visual information extraction and aggregation from the retina (e.g., object/scene sensing) to the primary visual cortex (e.g., content understanding) [18], [19]. Light that is reflected or emitted from the object/scene to our binocular retina, is segregated into different subbands or channels to stimulate respective neurons with nonlinear processing heavily involved, where some channels are aggregated, and some might be suppressed layer-by-layer. It leads to the phenomena that the HVS is highly frequency selective (or masking) and more sensitive to high spatial-frequency, such as the geometric structure (e.g., edge/orientation), contrast, 

1. In this work, we mainly emphasize on the color attributes, e.g., RGB or its variants in other color spaces.

This work was supported by the National Natural Science Foundation of China (61971282), Scientific Research Plan of the Science and Technology Commission of Shanghai Municipality (18511105400) and Ministry of Science and Technology of China (YS2018YFE020131).

Q. Yang, Y. Xu, J. Sun are from Cooperative Medianet Innovation Center, Shanghai Jiaotong University, Shanghai, 200240, China, (e-mail: yang.littleleq@sjtu.edu.cn, yl.xu@sjtu.edu.cn)

Z. Ma is from Nanjing University, Nanjing, Jiangsu, 210093, China (email: mazhan@nju.edu.cn)

Z. Li is from University of Missouri-Kansas City, Kansas, 64110, America (email: lizhu@umkc.edu)
etc and compound saliency [20]. On the other hand, our eyes are not directly sensing the individual point intensity, but rather the connected local neighbor structures due to the low-pass spread functionality of our eye optics [21]. Thus, we can make a reasonable hypothesis that the overall quality sensation is a weighted synthesis of the impacts from individual channel components (e.g., structure, color, etc). It inspires us to find an efficient way to decompose the perception-related key components for effective quality assessment.

One seminal work is the SSIM that measures the image quality using the similarities of respective luminance, contrast, and structure components. Because of the well-structured pixel sampling for 2D image/video, these components can be easily derived by statistical moments of pixel distribution, such as mean, variance, and co-variance, block-by-block [17]. As briefed previously, point cloud does not give explicit relationships among unstructured 3D points, making it difficult to extract appropriate features for quality measurement. Though dimensional reduction, such as the 3D-to-2D projection, can be applied to inherit existing 2D image metrics for weighted prediction [22], it still does not characterize the 3D points distribution very well with inferior subjective correlation. Recently, we have witnessed an explosive growth of graph signal processing (GSP) based applications [23]–[25], where graph offers great capacity to model high-dimensional visual data (e.g., 3D point cloud) by implicitly embedding the local neighbor connections to characterize the importance. Thus, we choose to utilize the GSP techniques to systematically attack the point cloud quality modeling problem.

Our Approach. We model the point cloud quality by considering the geometry and color attributes jointly. First, we extract the keypoints of a point cloud using its geometry information, by which we sort of construct the 3D object skeleton (e.g., contours, edges); We choose a simple yet efficient graph-based resampling method in [8] to fulfill the purpose. Note that these keypoints are from the original reference point cloud geometry, which is then utilized in both reference and impaired point cloud to construct local graphs. This ensures the common 3D structure for fair comparison. Local graph is generated for each keypoint by setting it as the spherical center and connecting all available neighbors within predefined distance; We respectively extract the zeroth, first and second moments of color gradients that are calculated between the spherical center (a.k.a., keypoint) and other effective neighbors in each graph (of both reference and distorted samples), for measuring the local graph significance after feature aggregation and similarity calculation. Such local graph significance is then used to derive the final similarity index that is consecutively pooled across different color channels (e.g., RGB, YUV or Gaussian Color Model - GCM [25]), and weighted among all local graphs. We call this approach GraphSIM in short.

We have evaluated the GraphSIM using two fairly large, independent and publicly accessible Point Cloud Quality Assessment databases, e.g., SJTU-PCQA [27] and IRPC [28], in comparison to other popular metrics used in existing applications. The SJTU-PCQA database, having 420 processed point cloud samples rated with individual mean opinion score (MOS), is developed using the common test point cloud sequences suggested by the experts from well-known Moving Picture Experts Group (MPEG) and industrial leaders. These original samples are augmented with additive noise (e.g., geometry or color attribute), re-sampling, compression (e.g., octree-based) as well as their combined artifacts for assessment. These artifacts well represent the actual noises induced in practical point cloud applications. Our results have reported reliable and superior performance of MOS prediction with averaged Pearson linear correlation coefficient (PLCC), Spearman rank-order correlation coefficient (SROCC), and Root mean squared error (RMSE) at respective 0.89, 0.88 and 1.13 for People samples.

In addition, another IRPC database [28] is dedicated for studying the compression noises induced by the emerging MPEG standard compliant point cloud compression (PCC) approaches [2], e.g., G-PCC (geometry-based PCC), and V-PCC (video-based PCC). Corresponding PLCC, SROCC and RMSE are 0.93, 0.78 and 0.22 for joint People and Inanimate contents, furthering the generalization of proposed GraphSIM to standard compliant PCC technologies.

All of these have presented the consistent performance gains against traditional point-wise distance-based metrics used in MPEG standardization. We later have also examined the robustness of GraphSIM by dissecting and reassembling its components for study, such as re-sampling mechanism, graph scale, color space, model parameters, pooling methods, etc, further revealing the great generalization of proposed model into various scenarios.

Contribution. The novelties of this work are given below:

- To the best of our knowledge, this GraphSIM is the first one to assess the point cloud quality via GSP techniques, where we aggregate color gradient moments (e.g., zeroth, first and second) of local graphs for similarity measurement.
- Local graphs are generated using the same keypoints from the original point cloud as the spherical centers and their corresponding connected neighbors in respective reference and impaired point cloud. This is motivated by the low-pass spread functionality of our eye optics [17], [21].
- These keypoints are extracted from the original reference point cloud using its geometry information, mimicking the high spatial-frequency selective phenomenon (i.e., more sensitive to the geometric edges, contours) of our HVS [29].
- We test GraphSIM using independent SJTU-PCQA [27] and IRPC [28] datasets, demonstrating the state-of-the-art accuracy in predicting the subjective MOS. Ablation studies further reveal the model generalization to practical scenarios.

The reminder of this paper proceeds as follows: Sec. 2 briefly the graph representation of point cloud as well as key graph operands for subsequent processing; Sec. 3 extracts keypoints of a point cloud using the geometry information for local graph construction, color gradient moments aggregation and similarity derivation as discussed in Sec. 4. Experimental studies are conducted in Sec. 5 to demonstrate the state-of-the-art performance of proposed model for MOS prediction, while ablation studies further evident the ro-
businesst and reliable efficiency for model generalization. Concluding remarks are drawn in Sec. 8.

2 POINT CLOUD VIA GRAPH REPRESENTATION

We briefly introduce key concepts of graph signal processing applied in this paper for re-sampling and local graph construction. A great article on graph signal processing can be found in [30].

2.1 Graph

Let a point cloud \( \mathbf{P} \) have \( N \) points and each point have \( K \) attributes, i.e., \( \mathbf{P} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}^T \in \mathbb{R}^{N \times K} \). In this work, we consider the 3D geometry (e.g., \((x, y, z)\) coordinates), and another three-dimensional color attributes (e.g., RGB), leading to the sextuple representation of the \( i \)-th point \( \mathbf{X}_i = (x_i, y_i, z_i, R_i, G_i, B_i) \). We further use \( \mathbf{X}_i^O = (x_i, y_i, z_i) \) and \( \mathbf{X}_i^I = (R_i, G_i, B_i) \) to separate geometry and color channels respectively, i.e., \( \mathbf{X}_i = [\mathbf{X}_i^O, \mathbf{X}_i^I] \). \( N \) is usually a large number, e.g. \( N = 729,133 \) for MPEG point cloud “RedandBlack”.

Though points are typically unstructured without explicit neighbor connections for a point cloud, our eye optics will implicitly apply point spread function to connect local neighbors for geometry synthesis. Such mechanism can be simulated using a low-pass filter [17] (e.g., Gaussian model). Therefore, we propose to construct the graph representation of a point cloud by encoding the local neighbor connection weights into an adjacency matrix \( \mathcal{W} \in \mathbb{R}^{N \times N} \). Each effective connection between two points having positive weight is referred to as “edge”. We can call the “point” as “vertex” following the convention in graph theory. We formulate the edge/connection weight between \( \mathbf{X}_i \) and \( \mathbf{X}_j \) using the geometric distance as

\[
W_{\mathbf{X}_i, \mathbf{X}_j} = \begin{cases} 
  e^{-\frac{\|\mathbf{X}_i^O - \mathbf{X}_j^O\|^2}{\sigma^2}}, & \text{if } \|\mathbf{X}_i^O - \mathbf{X}_j^O\|_2 \leq \tau; \\
  0, & \text{else},
\end{cases}
\]

(1)

where \( \mathbf{X}_i^O, \mathbf{X}_j^O \) are 3D coordinates of \( \mathbf{X}_i \) and \( \mathbf{X}_j \), \( \sigma \) represents the variance of graph nodes and \( \tau \) is the Euclidean distance threshold used for clustering neighbor points into the same graph. Finally, a point cloud \( \mathbf{P} \) can be represented using graphs involving points and their neighbor connections, noted as \( \mathcal{G}_{\mathbf{P}, \mathcal{W}(\mathbf{P})} \).

2.2 Operand

We first introduce the diagonal Degree matrix \( \mathcal{D} \) used for measuring the edge weighting density of connections attached to each vertex. For \( \mathbf{X}_i \), its connection density is \( d_i = \sum_{j} W_{\mathbf{X}_i, \mathbf{X}_j} \); Overall, we have

\[
\mathcal{D} = \text{diag}(d_1, \ldots, d_N) \in \mathbb{R}^{N \times N}.
\]

(2)

Given an edge between \( \mathbf{X}_i \) and \( \mathbf{X}_j \), i.e., \( e = (\mathbf{X}_i, \mathbf{X}_j) \), its graph edge derivative of \( f \) can be derived from graph Laplacian regularizer [31],

\[
\frac{\partial f}{\partial \mathbf{X}_i} := W_{\mathbf{X}_i, \mathbf{X}_j} [f(\mathbf{X}_i) - f(\mathbf{X}_j)],
\]

and thus corresponding graph gradient of \( f \) is the collection set of all edge derivatives connected to \( \mathbf{X}_i \) as

\[
\nabla \mathbf{X}_i f := \left\{ \frac{\partial f}{\partial \mathbf{X}_i} \right\}_{e \in E}, \text{ s.t. } e = (\mathbf{X}_i, \mathbf{X}_j) \in \mathcal{N}_i,
\]

(4)

with \( E \) as a set of edges connected to \( \mathbf{X}_i \).
It then leads to the **Graph Laplacian matrix**, e.g.,
\[
L = D - W, \tag{5}
\]
which is a *difference* operand on graphs. For any signal \( f \in \mathbb{R}^N \), its Laplacian operation is
\[
L(f) = \sum_{x_i \in \mathcal{N}_i} W_{x_i,x_j} \cdot \left[ f(x_i) - f(x_j) \right], \tag{6}
\]
where \( \mathcal{N}_i \) is a set of neighbors attached to \( x_i \), and \( f \) can be the RGB color intensity, normal or other attributes as well.

### 3 Point Cloud Re-sampling

As suggested in [17] and other neuroscience developments, our HVS weighs more to the structural information of perceived object or scene. For a 3D object, our vision system would first capture the general 3D structure, rather individual point intensity. Such 3D structure is the discriminative **geometric skeleton** (e.g., edges, contours) of the object. Corresponding points with this skeleton are referred to as the “geometric keypoints”.

We could obtain these keypoints via point cloud resampling. These keypoints should form the edges, contours, and skeleton for quality assessment, which mostly belong to the high spatial-frequency band that is sensitive to the HVS. We therefore choose a simple yet efficient high-pass graph filtering method in [3] to fulfill this task. Graph filtering is briefed below. Please refer to [3] for more details.

Let \( A \in \mathbb{R}^{N \times N} \) be a **graph shift operator**, which can be formulated using the adjacency matrix \( W \), or transition matrix \( D^{-1}W \), or graph Laplacian matrix \( L \). A linear, shift-invariant graph filter is a polynomial function of \( A \),
\[
h(A) = \sum_{l=0}^{L-1} h_l A^l = h_0 I + h_1 A + \ldots + h_{L-1} A^{L-1}, \tag{7}
\]
where \( h_l \) is \( l \)-th coefficient, and \( L \) is the length of graph filter. For \( h(A) \), a Haar-like graph filter is selected to implement the high-pass filtering, e.g.,
\[
h_{HH}(A) = I - A
\]
\[
= V \begin{bmatrix}
1 - \lambda_1 & 0 & \ldots & 0 \\
0 & 1 - \lambda_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1 - \lambda_N
\end{bmatrix} V^{-1}. \tag{8}
\]
Here, \( A = D^{-1}W \), \( \lambda_i \) and \( V \) are the eigenvalues, eigenvectors of \( A \). Thus, the frequency response of \( \tilde{x}_i \) is
\[
F(\tilde{x}_i) = h_{HH}(A) \cdot \tilde{x}_i = \tilde{x}_i - \sum_{x_j \in \mathcal{N}_i} A_{x_i,x_j} \cdot \tilde{x}_j, \tag{9}
\]
which is then utilized to order the points in spatial frequency domain for sampling.

Applying re-sampled high-frequency keypoints for subsequent quality assessment not only fits the perceptual intuition, but also significantly reduce the computational complexity which is of practical interests in applications. Other resampling methods can be applied as well in GraphSIM. More discussions are given in Sec. 6.

Note that geometric keypoints are extracted from the reference point cloud only using its geometry information. They are then leveraged to construct local graphs for both reference and distorted point clouds. This ensures the fair comparison with the common 3D geometric structure.

#### 4.1 Keypoints Resampling.

Given a reference point cloud \( \tilde{P} \), we first re-sample it to derive the geometric keypoints,
\[
\tilde{P}^*_s = \Psi(\tilde{P}_r, L) \in \mathbb{R}^{\beta \times \delta}, \beta \ll N, \tag{10}
\]
where \( \Psi(\cdot) \) is a high-pass graph filter suggested in Sec. 3, \( L \) is filter length, and \( \beta \) is the number of effective keypoints after resampling. We show results under different \( L \)s and \( \beta \)s in Fig. I. It reveals that graph-based resampling filter tries to retain points close to the object contour, edges (e.g., hair, skirt hemlines and facial features) along with the decrease of \( \beta \). As reported in [3], the bigger \( L \) comes with the larger reception field, better filtering performance, and higher complexity. We choose \( \beta = N/1000 \), and \( L = 4 \) for subsequent steps to well balance the complexity and efficiency. Other resampling methods and \( \beta \) selection will be further explored in Sec 5.

#### 4.2 Local Graph construction

Each keypoint in \( \tilde{P}_s \) is used as the center to construct **local graph** in both reference \( \tilde{P}_r \) and distorted point clouds \( \tilde{P}_d \). For \( k \)-th keypoint \( \tilde{s}_k \) in \( \tilde{P}_s \), we cluster its neighbors using the Euclidean distance of corresponding geometry components in both \( \tilde{P}_r \) and \( \tilde{P}_d \), i.e.,
\[
\begin{aligned}
\bar{V}(\tilde{x}_r, \tilde{s}_k) &= \{ \tilde{x}_r \} \subset \tilde{P}_r, \| \tilde{x}_r - \tilde{s}_k \|_2^2 \leq \theta, \\
\bar{V}(\tilde{x}_d, \tilde{s}_k) &= \{ \tilde{x}_d \} \subset \tilde{P}_d, \| \tilde{x}_d - \tilde{s}_k \|_2^2 \leq \theta.
\end{aligned} \tag{11}
\]

\( \bar{V}(\tilde{x}_r, \tilde{s}_k) \) and \( \bar{V}(\tilde{x}_d, \tilde{s}_k) \) represent the groups of neighbors of \( \tilde{s}_k \) in reference and distorted content. We then follow [1] to construct connected local graphs with \( \tilde{s}_k \) as the spherical center. In general, the selection of \( \tau \) in [1] depend on \( \theta \), and \( \sigma \) is a function of \( \tau \) used for the adjacency matrix \( W \)s and corresponding graphs.

#### 4.3 Color Gradient Features

We first use zeroth, first and second moments of color gradients to respectively represent the mean, mean and variance features, by which we try to well illustrate the color distribution in a local graph. As will be shown in later discussion, each feature has its unique effect in distortion measurement.

##### 4.3.1 Zeroth Moment: Gradient Mass \( m_g \)

We use [3] to derive the color gradients for vertex \( \tilde{s}_k \) in local graph that are summed up for the zeroth moment based gradient mass \( m_g \) measurement,
\[
m_g = \| \nabla \tilde{s}_k \| = \sum_{\tilde{x}_j \in \mathcal{N}_k} W_{\tilde{x}_j, \tilde{s}_k} [f(\tilde{x}_j) - f(\tilde{s}_k)]. \tag{12}
\]
As aforementioned, graph vertex $\tilde{s}_k$ corresponds to a specific keypoint of a sampled point cloud, which is the center of a constructed graph. Thus, (12) can be utilized to derive the $m_g$ for each graph against its center keypoint. Note we only consider the points that have an effective connection with $\tilde{s}_k$, i.e., $\forall \tilde{X}_j \in \tilde{N}_{\tilde{s}_k}, W_{\tilde{X}_j, \tilde{s}_k} \neq 0$.

For 2D image, relative changes of local pixel intensity can be easily captured by the vision system, reflecting the quality sensation distortion. Such change can be quantitatively measured using local contrast, such as the variance-based contrast used in SSIM [17]. Similarly, we propose to calculate the color gradient in (12) to represent such local color intensity variations. It is then weighted by the Euclidean distance factor to jointly consider the superimposed impairments of geometric and colormetric attributes. Note that (12) is used in [31] as weighted graph total variation for image denoising.

$m_g$ can reflect point density variations if we assume the unit graph (or after normalization). This is devised to deal with the scenario that the number of effective points in a graph changes due to the injection of impairments, such as downsampling, compression. The $m_g$ term, however, is barely utilized in metrics for 2D image/video. This is mainly because pixels in image or video blocks are usually impaired with intensity variations with the same number of appearance. On the contrary, points may emerge or vanish in point cloud due to a variety of computations.

Nevertheless, $m_g$ is not capable of efficiently handling the geometric displacement, perceptual inconsistency, etc. For instance, for a local graph, its $m_g$ will be the same if the point locations changes due to motion displacement (e.g., rotation) but point intensities are kept without change. On the other hand, as observed in subsequent standard compliant PCC approaches, point density is improved after reconstruction, but perceptual sensation is almost the same, compared with the ground truth. This also leads to the inconsistency with the $m_g$ measurement. Thus, we further introduce the first and second moments of graph color gradients to improve the quality measurement.

### 4.3.2 First Moment: Gradient Mean $\mu_g$

We first extend $m_g$ to derive the mean by simple normalization:

$$\mu_g = \frac{1}{N} \| \nabla \tilde{s}_k \tilde{f} \|,$$

(13)

where $N = |\tilde{N}_{\tilde{s}_k}|$ showing the total number of points which have effective connection with key points in a graph. It supplements the $m_g$ to infer the quality if $m_g$ cannot give sufficient discriminative difference for evaluating the perceptual sensation.

$\mu_g$ measures the averaged color gradient difference between the keypoint and its neighbors in a specific local graph, revealing the averaged local contrast variations.

### 4.3.3 Second Moment: Gradient Variance and Co-variance

In order to calculate gradient variance and co-variance, point matching is first performed by using the point-wise Euclidean distance. Such point-wise matching is also used in existing metrics [2] for fair and quantitative comparison.

For two graphs centered at $\tilde{s}_k$, e.g., reference graph $G_{\tilde{r}, \tilde{s}_k}$ and distorted graph $G_{\tilde{d}, \tilde{s}_k}$, we choose the one having the less points as the baseline, and then regulate another one by point matching to guarantee the same number of points. We utilize the nearest distance search to perform the point matching.

For simplicity, we use $\tilde{G}_{\tilde{r}, \tilde{s}_k}$ and $\tilde{G}_{\tilde{d}, \tilde{s}_k}$ to represent the graphs after point matching. Subsequently, we derive the variance and co-variance upon these two graphs having same number of scattered points.

Following the gradient calculation, we have the edge weighted gradient for $\tilde{X}_j$ as

$$g_{\tilde{s}_j, \tilde{s}_k} = W_{\tilde{s}_j, \tilde{s}_k} \cdot (f(\tilde{X}_j) - f(\tilde{s}_k)).$$

(14)

It comprises the weighted gradient distribution of all connected points for a specific graph, as

$$\tilde{g}_{\tilde{s}_k} := \left\{ g_j = g_{\tilde{X}_j, \tilde{s}_k} \right\}, s.t. \tilde{X}_j \in \tilde{N}_{\tilde{s}_k}. $$

(15)

We use $g_j$ for simplicity, and $\tilde{N}_{\tilde{s}_k}$ is for either $\tilde{G}_{\tilde{r}, \tilde{s}_k}$ or $\tilde{G}_{\tilde{d}, \tilde{s}_k}$.

It then leads to the variance derivation of edge weighted gradients as,

$$\sigma^2_g := \frac{\sum (g_j - \bar{g})^2}{N},$$

(16)

where $g_j$ represents $j$-th element in $\tilde{g}_{\tilde{s}_k}$, $\bar{g}$ represents averaged gradient of $\tilde{g}_{\tilde{s}_k}$, and $N = |\tilde{N}_{\tilde{s}_k}|$. Similarity, we calculate the co-variance as

$$c_{\tilde{g}_{\tilde{s}_k}, \tilde{g}'_{\tilde{s}_k}} = \text{cov}(\tilde{g}_{\tilde{s}_k}, \tilde{g}'_{\tilde{s}_k}) = E[\tilde{g}_{\tilde{s}_k} \cdot \tilde{g}'_{\tilde{s}_k}] - E[\tilde{g}_{\tilde{s}_k}] \cdot E[\tilde{g}'_{\tilde{s}_k}], $$

(17)

with $\tilde{g}_{\tilde{s}_k}$ and $\tilde{g}'_{\tilde{s}_k}$ representing the weighted gradient distribution of respective reference and impaired point clouds. For simplicity, we note $c_{\tilde{g}_{\tilde{s}_k}, \tilde{g}'_{\tilde{s}_k}}$ as $c_g$.

### 4.4 GraphSIM

For each color channel, we have mass $m_g$, mean $\mu_g$ and variances $c_g$ derived by the statistical movements of the gradient distribution for both graphs from the reference and impaired point cloud. These features represent the local graph significance that can be utilized for quality measurement quantitatively. As inspired by the SSIM, we propose to fuse these tree features-based similarity together to have a general index for color channel $C$, i.e.,

$$S_{C, \tilde{s}_k} = \text{SIM}_{m_g} \cdot \text{SIM}_{\mu_g} \cdot \text{SIM}_{c_g},$$

(18)

where

$$\text{SIM}_{m_g} = \frac{2m^r \cdot m^d + T_0}{(m^r)^2 + (m^d)^2 + T_0},$$

(19)

$$\text{SIM}_{\mu_g} = \frac{2\mu^r \cdot \mu^d + T_1}{(\mu^r)^2 + (\mu^d)^2 + T_1},$$

(20)

$$\text{SIM}_{c_g} = \frac{c^r + T_2}{\sigma^r \cdot \sigma^d + T_2},$$

(21)

$T_0, T_1$ and $T_2$ are small no-zero constants to prevent numerical instability. Note that $\text{SIM}_{m_g}, \text{SIM}_{\mu_g}$ and $\text{SIM}_{c_g}$ are in the range of [-1,1]. We use the superscripts $r$ and $d$ to indicate the reference and distorted samples, respectively.

We extend [18] to represent the local graph quality by aggregating it across all color components. Note that we
TABLE 1: Sample Point Clouds Illustration.

| name       | #points | axis range          |  |
|------------|---------|---------------------|---|
| RedandBlack| 729133  | [122.575, 109.387]  | 121.353 |
| Loot       | 784142  | [28.380, 7.999]     | 119.473 |
| Soldier    | 1059810 | [29.389, 7.1023]    | 31.436  |
| LongDress  | 806806  | [151.397, 5.1012]   | 87.523  |
| Hhi        | 900153  | [0.61875, 0.64057]  | 0.17135 |

Table 1 gives the basic information of these point clouds (e.g., number of points, dimensional ranges of x, y and z axes), as well as the illustrative snapshots in Fig. 2(a). These original point clouds are recommended by the experts in MPEG for compression standardization, well covering a variety of content characteristics [2].

Each native point cloud sample is augmented with seven different types of impairments under six levels, including four individual distortions, Octree-based compression (OT), Color noise (CN), Geometry Gaussian noise (GGN), Down-sampling (DS), and three superimposed distortions, such as Downsampling and Color noise (D+C), Downsampling and Geometry Gaussian noise (D+G), Color noise and Geometry Gaussian noise (C+G). These impairments, covering the resampling, intensity and geometric noise, and compression, are used to well simulate the artifacts that might be induced in practical applications. Please refer to [27] for more details.

can represent the point cloud in different color spaces, such as the RGB, YUV, Gaussian Color Model (GCM) [26], [32], etc. Referring to the overall PSNR calculation of YUV content [33], it suggests the color channel weighting factors as Y:U:V = 6:1:1, assuming more sensitive perception of the luminance components. We apply this widely-adopted factors for pooling, i.e.,

\[ S_{x_k} = \frac{1}{\gamma} \sum_C \gamma_C \cdot |S_{x_k,C}|, \]

(22)

where \( \gamma_C \) is the pooling factor of \( C \)-channel reflecting the importance of individual color channel during visual perception, \( \gamma = \sum_C \gamma_C \). We will first demonstrate the efficiency of GraphSIM in GCM space in Sec. 5, followed by further discussions on color spaces in Sec. 6. And we also quantitatively analyze the influence of different pooling methodologies for deriving (18) and (22).

In the end, we can have the overall point cloud quality by averaging all local graph similarities:

\[ Q = \frac{1}{\beta} \sum_{s_k \in P} S_{x_k}, \]

(23)

with \( \beta \) as the total number of keypoints (and corresponding constructed local graphs) defined previously.

5 EXPERIMENTAL EVALUATIONS

This section evaluates the GraphSIM and other five state-of-the-art metrics for point cloud quality prediction, using two publicly accessible point cloud database, SJTU-PCQA database in [27] and IRPC database in [28].

5.1 Subjective Point Cloud Assessment Database

5.1.1 SJTU-PCQA Database

We use five high-quality human body point cloud samples in People category, e.g., “RedandBlack”, “Loot”, “Soldier”, “LongDress”, and “Hhi”. Table 1 gives the basic information of these point clouds (e.g., number of points, dimensional ranges of x, y and z axes), as well as the illustrative snapshots in Fig. 2(a). These original point clouds are recommended by the experts in MPEG for compression standardization, well covering a variety of content characteristics [2].

Each native point cloud sample is augmented with seven different types of impairments under six levels, including four individual distortions, Octree-based compression (OT), Color noise (CN), Geometry Gaussian noise (GGN), Down-sampling (DS), and three superimposed distortions, such as Downsampling and Color noise (D+C), Downsampling and Geometry Gaussian noise (D+G), Color noise and Geometry Gaussian noise (C+G). These impairments, covering the resampling, intensity and geometric noise, and compression, are used to well simulate the artifacts that might be induced in practical applications. Please refer to [27] for more details.

5.1.2 IRPC Database

We further adopt two high-quality Inanimate samples, e.g., “Facade”, “House”, and two high-quality People samples, “LongDress”, “Loot”, used in IRPC for additional evaluation shown in Fig. 2(b). This dataset is independently collected [28] with the emphasis on the point cloud compression distortions.

Each native sample is augmented using three different compression methods (e.g., OT, G-PCC and V-PCC) with three compression levels (e.g., High quality - HQ, Medium quality - MQ, and Low quality - LQ). More details can be found in [28]. We use G-PCC and V-PCC coded samples for evaluation since OT-based compression is already included in SJTU-PCQA dataset.

Interestingly, both G-PCC and V-PCC increase the points greatly after reconstruction, leading to the point density growth in a local graph. In Table 3 we have noticed that it even can double the original points, e.g., G-PCC compressed “Facade” at HQ level. For V-PCC, additional 40% 50% points are reported. To the best of our knowledge, such phenomenon is barely observed for 2D image/video.

3. Correspondingly, compression relevant quantization parameters are respectively set at low, medium and high levels.
TABLE 2: Model performance (PLCC, SROCC and RMSE) for different point clouds augmented with a variety of impairments. Averaged results are given in People categories. Note that M-p2po, M-p2pl, H-p2po and H-p2pl only measure the geometric distortion.

| Name       | Type   | Points | Δ %  | Name     | Type   | Points | Δ %  |
|------------|--------|--------|------|----------|--------|--------|------|
| Soldier    | 10-bit | 596058 | -    | House    | 10-bit | 488745 | -    |
| G-PCC      | 10-bit | 596131 | +94% | G-PCC    | 10-bit | 488995 | +75% |
| MQ         | 10-bit | 488795 | +15% | MQ       | 10-bit | 488995 | +36% |
| V-PCC      | 10-bit | 264680 | +4%  | V-PCC    | 10-bit | 264680 | +40% |

5.2 Gaussian Color Decomposition.

We first utilize the Gaussian Color Model (GCM) to demonstrate the model efficiency. This is mainly because GCM is suggested to be more closely related to the color sensation of our HVS [25]. Other color spaces are discussed in subsequent ablation studies. Normally, the GCM function decomposes the native RGB signal via

\[
\begin{bmatrix}
\hat{E} \\
\hat{E}_\lambda
\end{bmatrix} =
\begin{bmatrix}
0.63 & 0.27 \\
0.30 & -0.35
\end{bmatrix}
\begin{bmatrix}
R \\
G
\end{bmatrix},
\]  

where \(\hat{E}, \hat{E}_\lambda\) are respective luminance and two chrominance components.

5.3 Model Parameters.

In total, we have a set of parameters associated with different processing stages that need to be determined for GraphSIM.

- \(\beta, L\) in Resampling: \(\beta = N/1000\) and \(L = 4\). These numbers are used to well balance the efficiency and complexity.
- \(\theta, \tau, \sigma\) in Local Graph Construction: Given a local graph centered at \(\hat{s}_k\), we set \(\theta = \frac{1}{\sqrt{B}}\) to cluster neighbors, where \(B = \min(X_y Y_x Z_z)\) with \(X_x = X_{\text{max}} - X_{\text{min}}, Y_y = Y_{\text{max}} - Y_{\text{min}}\) and \(Z_z = Z_{\text{max}} - Z_{\text{min}}\) as respective bounding box scale of \(x-, y-,\) and \(z\)-axis of reference point; We then determine the \(\tau\) using the largest Euclidean distance from the 50 nearest neighbors of all the points in \(\hat{V}(\hat{x}, \hat{s}_k)\) or \(\hat{V}(\hat{x}, \hat{s}_k)\) in [11]. If there are less than 50 neighbors (e.g., < 50 points in \(\hat{V}(\hat{x}, \hat{s}_k)\) or \(\hat{V}(\hat{x}, \hat{s}_k)\)), the largest Euclidean distance is set as \(\tau\); Finally, we have \(\sigma = \tau^2/2\). It is worth
to point out that we actually connect the $\tau$ and $\theta$ of local graph with the scale of point cloud (e.g., bounding box size). This allows us to perform the normalization evenly (without bias) to the same scale for fair comparison.

- $T_0$, $T_1$, $T_2$, $\gamma_C$ in Similarity Pooling: $T_0$, $T_1$, and $T_2$ are set as 0.001 following the suggestion in [17]. We set $[\gamma_E, \gamma_{E_x}, \gamma_{E_{\lambda}}] = [6, 1, 1]$ to reflect the different importance of various color components where normally, luminance is more sensitive to the HVS. The numbers, e.g., 6, 1, and 1, follow the conventions widely used in compression society to derive the overall PSNR of all YUV channels where weighting coefficients for $Y$, $U$, and $V$ are 6, 1, and 1, i.e., $[\gamma_Y, \gamma_U, \gamma_V] = [6, 1, 1]$.

These parameters are either fixed constants, or can be easily derived according to the signal statistics (e.g., point cloud bounding box scale, color space, sampling ratio), suggesting that our GraphSIM is fairly lightweight and straightforward for practical applications.

### 5.4 Performance Evaluation.

We compare our GraphSIM with another five state-of-the-art metrics adopted in MPEG point cloud compression software [34], e.g.,

- PSNR-MSE-P2point (M-p2po)
- PSNR-MSE-P2plane (M-p2pl)
- PSNR-Hausdorff-P2point (H-p2po)
- PSNR-Hausdorff-P2plane (H-p2pl)
- PSNR$_{\text{YUV}}$: $\text{PSNR}_{\text{YUV}} = (6 \times \text{PSNR}_Y + \text{PSNR}_U + \text{PSNR}_V)/8$ [33]

Note that first four metrics only give the geometry measurement without color components using either point-to-point (p2po) or point-to-plane (p2pl) error-based Hausdorff or RMSE (root mean squared error) distances, while PSNR$_{\text{YUV}}$ calculates the overall YUV distortion of matched points in reference and distorted contents.

To ensure the consistency between subjective scores (e.g., MOS) and objective predictions from various models, we map the objective predictions of different models to the same dynamic range following the recommendations suggested by the video quality experts group (VQEG) [35], [36], to derive popular PLCC for prediction accuracy, SROCC for prediction monotonicity, and RMSE for prediction consistency for evaluating the model performance. The larger PLCC or SROCC comes with the better model performance. On the contrary, the lower RMSE is better. More details can be found in [35].

### 5.4.1 SJTU-PCQA Database

In this part, we present the performance of GraphSIM and other state-of-the-art metrics over SJTU-PCQA database. As reported in Table 2 our GraphSIM consistently offers the leading performance (mostly ranked at the top), in People categories for all impairments, with model performance evaluated using (PLCC, SROCC, RMSE) as (0.89, 0.88, 1.13). In comparison, model performances are (0.89, 0.79, 1.11) for metric M-p2po, (0.74, 0.66, 1.66) for M-p2pl, (0.80, 0.70, 1.49) for H-p2po, (0.71, 0.66, 1.83) for H-p2pl and (0.71, 0.71, 1.74) for PSNR$_{\text{YUV}}$, respectively.

**Consistency.** The GraphSIM shows robust performance across various contents and impairments, and demonstrates reliable correlations with subjective MOSs. On the contrary, other metrics varies quite significantly for different cases. For example, objective metric M-p2po presents comparable PLCC to GraphSIM for “Loot” (0.87 vs 0.87), “Soldier” (0.92 vs 0.91) and overall People category (0.89 vs 0.89), but exhibits obvious inferior correlation for SROCC, e.g., (0.77 vs 0.88), (0.80 vs 0.89) and (0.79 vs 0.88), respectively. One potential cause for such unreliable variation is due to the normalization scheme. As for M-p2po metric, though we calculate the distance between paired points in respective reference and distorted point cloud, distance itself varies significantly for different points clouds because of the dimensional range scale differences. For example, the axis range $[(x_{\text{min}}, x_{\text{max}}), (y_{\text{min}}, y_{\text{max}}), (z_{\text{min}}, z_{\text{max}})]$ of “Hhi” is $[(0.61875), (0.64057), (0.170135)]$, while the axis range of “Loot” is $[(28,380), (7,999), (119,473)]$. This would apparently lead to the variations of distance measurement in several order of magnitudes. To overcome this issue, a simple fix is proposed to normalize the distance-based MSE using the...
maximum dimensional range \( p \), e.g., \( p = \max\{x_{\text{max}} - x_{\text{min}}, y_{\text{max}} - y_{\text{min}}, z_{\text{max}} - z_{\text{min}}\} \). It, however, still can not avoid the bias if the distance error is not aligned with axis having the maximum range. All of these suggest that point-wise distance-based metrics are difficult to be generalized for reliable performance measurement.

Impairment Superimposition. Except for the GraphSIM and PSNR\(_{YUV}\), other four metrics only consider the geometrical distortion and cannot handle color attribute impairments, e.g., CN (color noise), contrast change, etc, at all (see Table 2). This limits the generalization of these distance-based metrics, e.g., M-p2po, M-p2pl, H-p2po, and H-p2pl, for evaluating the superimposed distortions. The same problem is observed for PSNR\(_{YUV}\) as well. The PSNR\(_{YUV}\) presents the best performance for CN distortion on average (e.g., “People(ave)”), having (PLCC, SROCC, RMSE) = (0.97, 0.94, 0.48). It, however, offers the worst performance for OT impairment with (PLCC, SROCC, RMSE) = (0.54, 0.52, 1.52). This is mainly because PSNR\(_{YUV}\) only calculates the color intensity difference for two matched points from both reference and impaired point clouds, without considering the geometric impacts. As quantitatively listed in Table 2, PSNR\(_{YUV}\) offers relatively poor performance for cases with geometric distortion, such as GGN with (PLCC, SROCC, RMSE) = (0.86, 0.82, 1.34), DS with (PLCC, SROCC, RMSE) = (0.74, 0.74, 1.56), D+G with (PLCC, SROCC, RMSE) = (0.85, 0.77, 1.37)). In contrast, GraphSIM provides the (PLCC, SROCC, RMSE) at (0.97, 0.96, 0.62), (0.97, 0.91, 0.55) and (0.99, 0.96, 0.43) for corresponding GGN, DS, and D+G, respectively.

Scatter Plot. For better illustration, we have also provides the scatter plots shown in Fig. 6 for all six models. Though some metrics provide better performance in certain types of impairments (e.g., M-p2po for “C+G” distortion), they are not reliable and consistent. This is also reflected from the scatter plots. All point-wise distance-based metrics could not offer competitive performance with our GraphSIM in predicting the subjective MOS, where most predictions are away from the “\( y = x \)” axis (e.g., perfect-prediction line).

5.4.2 IRPC Database

We further the evaluations of GraphSIM using another IRPC database shown in Table 4. Overall, the GraphSIM provides the noticeable performance margin with (PLCC, SROCC, RMSE) = (0.93, 0.78, 0.22) for all samples in both Inanimate and People categories, as shown in Table 4.

Performance consistency is still a critical issue for other metrics. For example, though PSNR\(_{YUV}\) shows the similar PLCC and RMSE as the GraphSIM, it has severely degradation in SROCC measurement even for People contents. We then retrieve the MOS and respective objective scores in Table 5. As we can see, PSNR\(_{YUV}\) exhibits larger variations across different content. For example, PSNR\(_{YUV}\) of “Loot” at LQ scale even offers higher objective score than it of “Longdress” at the HQ scale. This comes from the lack of inappropriate geometric scale normalization since PSNR\(_{YUV}\) only applies the error measurement between matched points. The same inconsistency are observed for PSNR\(_{YUV}\) in evaluating the OT artifacts shown in Table 2.

Additionally, we have found that except for proposed GraphSIM, other five metrics demonstrate very poor TABLE 4: Model performance (PLCC, SROCC and RMSE) for different point clouds encoded using G-PCC and V-PCC.

| Metric | G-PCC | V-PCC |
|--------|-------|-------|
| MOS   |       |       |
| GraphSIM | (0.97, 0.97, 0.61) | (0.99, 0.99, 0.63) |
| PSNR\(_{YUV}\) | (0.96, 0.96, 0.63) | (0.98, 0.98, 0.65) |
| SROCC | (0.87, 0.87, 0.55) | (0.92, 0.93, 0.56) |
| RMSE  | (0.61, 0.61, 0.25) | (0.71, 0.71, 0.27) |

TABLE 5: Objective score (PSNR\(_{YUV}\), GraphSIM) and MOS of People samples in IRPC [28].

| Level | Longdress | Loot |
|-------|-----------|------|
| MOS   | PSNR\(_{YUV}\) | GraphSIM |
| G-PCC | (0.93, 0.93, 0.22) | (0.97, 0.97, 0.61) |
| V-PCC | (0.98, 0.98, 0.65) | (0.99, 0.99, 0.63) |

SROCC index for Inanimate samples. It further evidences that point-wise distance-based error measurement is not capable of reliably characterizing the point cloud quality.

6 Ablation Studies

This section have examined the GraphSIM by disecting and reassembling its modules to demonstrate its generalization and efficiency.

6.1 Color Space

In Sec. 5 we have exemplified the efficiency of GraphSIM assuming the GCM-based color channel decomposition. It mainly follows the suggestions that GCM well correlates with the color sensation of our HVS [26]. In practices, we may use other color spaces, such as RGB and YUV that are typically applied in compression societies. We set the same color weighting factors for YUV and GCM spaces, e.g., \( \gamma_Y = 6 \), \( \gamma_U = 1 \) and \( \gamma_V = 1 \), given that luminance component is more sensitive [33]. For RGB space, \( \gamma_R = 1 \), \( \gamma_G = 2 \) and \( \gamma_B = 1 \). It follows the observations that green color components are more sensible to our vision system. The exact weighting factor setting of RGB is motivated by the fact that, in typical imaging CMOS, we often have two green pixels associated one red and one blue pixel. Table 6 lists the model performance averaged for all sequences in People category across GCM, RGB, and YUV. Other modules in GraphSIM are kept as suggested in Sec. 5. As reported, our GraphSIM has shown consistent performance across various color spaces, again ensuring the model generalization to different applications.
TABLE 6: Model performance with various color spaces.

| Color Space | PLCC | SROCC | RMSE |
|-------------|------|-------|------|
| RGB         | 0.8942 | 0.8890 | 1.1050 |
| YUV         | 0.8851 | 0.8882 | 1.1087 |
| GCM         | 0.8901 | 0.8840 | 1.1251 |

TABLE 7: Model performance with different resampling mechanism: SJTU-PCQA People category is exemplified with other contents having the similar outcomes.

| Method       | β       | PLCC | SROCC | RMSE |
|--------------|---------|------|-------|------|
| Random       | N/1e3   | 0.8836 | 0.8751 | 1.1561 |
|              | N/2e3   | 0.8827 | 0.8739 | 1.1602 |
|              | N/4e3   | 0.8773 | 0.8725 | 1.1744 |
| High-pass [3]| N/1e3   | 0.8882 | 0.8835 | 1.1341 |
|              | N/2e3   | 0.8913 | 0.8841 | 1.1192 |
|              | N/4e3   | 0.8836 | 0.8773 | 1.1469 |
|              | N/4e4   | 0.8898 | 0.8835 | 1.1264 |

6.2 Local Graph

Resampling. Keypoints resampling play a vital role in GraphSIM for maintaining the 3D geometric structure used in subsequent graph similarity measurement. We have exemplified graph filtering-based high spatial-frequency (Haar-alike high-pass filter) resampling [3] previously. Here we introduce an additional random resampling for comparative study. In the meantime, we provide more simulations with respect to different resampling ratios (e.g., βs) for two methods as well. Table 7 has shown the reliable performance (e.g., with outstanding PLCC, SROCC, and RMSE reported) of GraphSIM for both methods at various sampling ratios. Random sampling introduces unstable performance from β = N/1e3 to N/1e4. This might be due to the reason that random sampled keypoints are not well covering the geometric structure, or frequency band sensitive to the perception. On the other hand, high-pass filtering retains the consistent performance across βs. It, to some extent, implies that as long as we can have the keypoints to accurately reflect the geometric structure (e.g., contours, edges), the number of keypoints can be sufficient smaller than the total points in the native point cloud. The Haar-alike high-pass filter suggested in [3] apparently meets this criteria. We expect that a better high-pass resampling would further improve the overall performance. But, given the outstanding efficiency shown in Table 7, it has already demonstrated the model generalization to various resampling methodologies.

Neighbor Dimension. Local graph is utilized as the basic unit for similarity derivation, which is derived from the clustered neighbors. Given that τ and σ are dependent on θ, we first examine the impacts of different θ by setting θ = 0.01, 0.05, 0.1, 0.15, and 0.2, as shown in Fig. 4. It reveals that the model performance can be quickly improved by enlarging the neighbor dimension with larger θ, and gets quite stable when θ ≥ 0.05. This ensures the general applicability of the GraphSIM as long as we give a reasonable θ bounded on the point cloud dimensional scale.

Graph Scale. τ is used as threshold to cluster neighbors into the same graph following [1]. Given that τ is dependent on the largest distance of the k-th nearest neighbors, we examine the impacts of different k by setting k = 10, 20, 50, 80, 100, as shown in Fig. 5. Model performance can be gradually improved while k increases. But it begins to stable when k > 50. This also ensures the general applicability of the GraphSIM with reasonable k bounded on the point cloud graph scale.

6.3 Pooling Strategy

In Sec. 4 we first pool three feature-based similarities (e.g., m_y, μ_g and c_g) using multiplication, and then fuse color channels (e.g., R, G, B) using average pooling. Note that multiplication pooling is also used in SSIM index, and averaged pooling is widely used for overall PSNR derivation.

This part has attempted to examine different pooling methods in GraphSIM for in-depth understanding of its capacity. We define the pooling method P1 for feature similarity fusion under the same color channel, and P2 for the pooling across all color channels. Both P1 and P2 can adopt multiplication (M) or averaging (AVE). It then leads to four different combination C = [P1, P2], e.g., C1 = [AVE, AVE], C2 = [M, AVE], C3 = [AVE, M] and C4 = [M, M]. For P1, we distribute the same weighting factors, 1:1:1, for three features, while apply the 6:1:1 weighting factors for P2 assuming the GCM color model.

We use all samples in People category of SJTU-PCQA to study different pooling methods, with results shown in Table 8. Note that SROCC is relatively consistent for different pooling combinations. This is because both M and AVE do not change the monotonicity of test samples. On the contrary, PLCC and RMSE show obvious degradation when applying the C4, while C1 offers the best quantitative result. We believe that M aggravates the prediction error when fusing multiple feature-based similarities together. Assuming there are two samples, e.g., A and B, with respective MOEs as 9 and 8. We extract two features, e.g., f1 and f2, that will be utilized to derive individual similarity, e.g., SIM_f1, SIM_f2, for final objective index. For sample A, SIM_f1 = SIM_f2 = 0.9; while for sample B, SIM_f1 = SIM_f2 = 0.8. Final
TABLE 8: Model performance with various pooling methods.

| Method | PLCC | SRCC | RMSE |
|--------|------|------|------|
| C1     | 0.8913 | 0.8906 | 1.1792 |
| C2     | 0.8901 | 0.8840 | 1.1542 |
| C3     | 0.8840 | 0.8800 | 1.1542 |
| C4     | 0.7995 | 0.8663 | 1.5463 |

objective scores for sample A is 0.81, and is 0.64 for sample B. If we use AVE instead, final scores are 0.9 and 0.8, showing higher correlation with MOS data. Alternatively, if one feature has high similarity, e.g., 0.9, while another one has very low similarity, e.g., 0.1, the results after performing the M or AVE will be much more different (e.g., 0.9 × 0.1 = 0.09 vs. (0.9 + 0.1)/2 = 0.5). Thus, without resorting for complex weighting factors for each individual feature (e.g., MOS data fitting), AVE is more reliable and robust than M for PLCC and RMSE measurement.

7 RELATED WORK

Our work is closely related to the point cloud quality assessment and modeling. We give a brief review here.

A number of pioneering explorations have been made to assess the subjective point cloud quality [16], [38]–[45], from the assessment protocol, user interaction mechanism, distortion impairment, objective metric modeling, etc. On the other hand, a publicly accessible SJTU-PCQA database [27] has been released with 420 processed point cloud samples and associated MOSs. All of them could potentially benefit the society to develop and analyze the point cloud quality.

Point-wise error measurement was first applied to evaluate the geometry distortion of point cloud, such as the point-to-point (p2p) [46], point-to-plane (p2pl) [37], or point-to-mesh (p2m) [47], which could be then converted to the Hausdorff distance or MSE for geometric PSNR derivation. Color distortion can also be measured point-wisely to evaluate the Y or YUV error of geometric matched point pair. All of these are adopted into the MPEG point cloud compression reference software [44] for compression efficiency measurement. Later as analyzed extensively by EPFL lab members in their serial publications [16], [33], [39]–[41], [43], [45], [48], these point-wise distance based metrics are not well correlated with the subjective assessments with unreliable prediction accuracy. [49], [50] also tested the performance of these metrics under the distortion caused by typical compression methods, such as MPEG Point cloud Test Model Category 2 (TMC2) [51] and reached the same conclusion. Motivated by the projection-based approach used for MPEG point cloud compression, Alexiou et al. [45] recently proposed to project 3D point cloud to 2D planes, by which classical image objective quality metrics (e.g., SSIM [17]) could be applied. Experiments showed better efficiency under certain types of impairments, but more deep investigations were highly desired for reliable and consistent quality prediction.

8 CONCLUSION

Point cloud techniques have advanced fast in recent years for virtual reality, telepresence, etc. However, it still lacks of an efficient objective quality metrics that can accurately predict the subjective MOSs, and can be embedded into the system for performance optimization. Existing point-wise distance-based metrics used in MPEG point cloud compression standards [2] are not only unstable across contents and distortions, but also can not well reflect the perceptual sensation of the HVS.

Thus, we have developed the GraphSIM to approach this problem by jointly considering the geometry and color distortions. It includes the point cloud resampling to extract keypoints (e.g., contours, edges) at high spatial-frequency that are more sensitive to the perception, followed by constructing the local graphs centered at extracted keypoints; and color gradient aggregation of each graph for final similarity index pooling across color channels and all graphs. Our GraphSIM has demonstrated consistent, reliable correlation with the subjective MOSs upon two independent point cloud quality assessment datasets, presenting noticeable gains over the state-of-the-art metrics adopted in MPEG point cloud reference software. GraphSIM parameters are either fixed constants, or directly dependent on the input signal (e.g., color space, bounding box scale, etc), making it fairly easy for system implementation. Ablation studies have further supported the model generalization by examining its key modules and model parameters.

There are several interesting avenues for future exploration. For example, how to extend GraphSIM for geometry point cloud (i.e., without color attributes) is worth for deep investigation. Applying the GraphSIM into MPEG point cloud compression technologies to quantitatively optimize the rate-distortion efficiency is another practical and attractive topic.

9 ACKNOWLEDGMENT

We would like to thank anonymous reviewers for their constructive comments to improve this manuscript. In the meantime, we really appreciate the efforts devoted in [28] for developing the IPRC point cloud assessment database.

REFERENCES

[1] D. J. Brady, W. Pang, H. Li, Z. Ma, Y. Tao, and X. Cao, “Parallel cameras,” Optica, vol. 5, no. 2, pp. 127–137, 2018.
[2] S. Schwarz, M. Preda, V. Barontini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuca, S. Lasserre, Z. Li et al., “Emerging mpeg standards for point cloud compression,” IEEE J. Emerg. Sel. Topics in Circuits and Systems, vol. 9, no. 1, pp. 133–148, 2018.
[3] S. Chen, D. Tian, C. Feng, A. Vetro, and J. Kovačević, “Fast resampling of three-dimensional point clouds via graphs,” IEEE Trans. Signal Processing, vol. 66, no. 3, pp. 666–681, 2017.
[4] S. Yan, Y. Peng, G. Wang, S. Lai, and M. Zhang, “Weakly supported plane surface reconstruction via plane segmentation guided point cloud enhancement,” IEEE Access, 2019.
[5] Y. Regaya, F. Fadli, and A. Amira, “3D point cloud enhancement using unsupervised anomaly detection,” in 2019 Int. Symposium on Systems Engineering (ISSE’19). IEEE, 2019, pp. 1–6.
[6] T. Zheng, C. Chen, J. Yuan, B. Li, and K. Ren, “Pointcloud saliency maps,” in Proc. IEEE Int. Conf. Computer Vision (ICCV’19), 2019, pp. 1598–1606.
[7] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, “Semantic3d. net: A new large-scale point cloud classification benchmark,” arXiv preprint arXiv:1704.03847, 2017.
[8] Z. Zhang, L. Zhang, X. Tong, P. T. Mathiopoulos, B. Guo, X. Huang, Z. Wang, and Y. Wang, “A multilevel point-cluster-based discriminative feature for als point cloud classification,” IEEE Trans. Geoscience and Remote Sensing, vol. 54, no. 6, pp. 3309–3321, 2016.
[9] T. Rabbani, F. Van Den Heuvel, and G. Vosselman, “Segmentation of point clouds using smoothness constraint,” International archives of photogrammetry, remote sensing and spatial information sciences, vol. 36, no. 5, pp. 248–253, 2006.

[10] L. Li, Z. Li, S. Liu, and H. Li, “Occupancy-map-based rate distortion optimization and partition for video-based point cloud compression,” IEEE Trans. Circuits and Systems for Video Technology, 2020.

[11] E. Alexiou, “On subjective and objective quality evaluation of point cloud geometry,” in 9th Int. Conf. Quality of Multimedia Experience (QoMEX’17). IEEE, 2017, pp. 1–3.

[12] C. Javaheri, C. Brites, F. Pereira, and J. Ascenso, “Point cloud subjective assessment database. [Online]. Available: https://vision.nju.edu.cn/28/fd/c29466a469245/page.htm

[13] E. Alexiou, P. Xu, and T. Ebrahimi, “Towards modelling of visual saliency in point clouds for immersive applications,” in 26th IEEE Int. Conf. Image Processing (ICIP’19), no. CONE, 2019.

[14] J. Zhang, W. Huang, X. Zhu, and J.-N. Hwang, “A subjective quality evaluation for 3d point cloud models,” in Int. Conf. Audio, Language and Image Processing (ALIP’14). IEEE, 2014, pp. 827–831.

[15] M. Khaled, “Pcc test model category 2 v0,” ISO/IEC JTC1/SC29/WG11 MPEG2018/N107520l, Macau, China, Oct, 2018.

[16] Mpeg reference software. [Online]. Available: http://mpeg.chIVER/ft/software/MPEG/PPC/PPC-DIMETRIC.html

[17] E. Alexiou, T. Ebrahimi, M. V. Bernardo, M. Pereira, A. Pinheiro, L. A. D. S. Cruz, C. Duarte, L. G. Dmitrovic, E. Dumi¢, D. Matkovics et al., “Point cloud subjective evaluation methodology based on 2d rendering,” in 10th Int. Conf. Quality of Multimedia Experience (QoMEX’18). IEEE, 2018, pp. 1–6.

[18] M. Khaled, “Pcc test model category 2 v0,” ISO/IEC JTC1/SC29/WG11 MPEG2018/N107524l, Macau, China, Oct, 2017.

[19] J. Zhang, W. Huang, X. Zhu, and J.-N. Hwang, “A subjective quality evaluation for 3d point cloud models,” in Int. Conf. Audio, Language and Image Processing (ALIP’14). IEEE, 2014, pp. 827–831.

[20] M. Khaled, “Pcc test model category 2 v0,” ISO/IEC JTC1/SC29/WG11 MPEG2018/N107524l, Macau, China, Oct, 2017.
Qi Yang  Qi Yang received the B.S. degree in communication engineering from Xidian University, Xian, China, in 2017. He is currently working toward the Ph.D degree in information and communication engineering at Shanghai Jiao Tong University, Shanghai, China, since 2017. His research interests include media quality assessment, computer vision, and distributed computation.

Zhan Ma  Zhan Ma (SM’19) received the B.S. and M.S. from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2004 and 2006 respectively, and the Ph.D. degree from the New York University, New York, in 2011. He is now on the faculty of Electronic Science and Engineering School, Nanjing University, Jiangsu, 210093, China. From 2011 to 2014, he has been with Samsung Research America, Dallas TX, and Futurewei Technologies Inc., Santa Clara, CA, respectively. His current research focuses on the next-generation video coding, energy-efficient communication, gigapixel streaming and deep learning. He is a co-recipient of 2018 ACM SIGCOMM Student Research Competition Finalist, 2018 PCM Best Paper Finalist, and 2019 IEEE Broadcast Technology Society Best Paper Award.

Yiling Xu  Yiling Xu received the B.S., M.S. and Ph.D. from the University of Electronic Science and Technology of China, China, in 1999, 2001 and 2004 respectively. She is a full researcher of School of Electronic Information and Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200145, China. From 2004 to 2013, she was with Multimedia Communication Research Institute of Samsung Electronics Inc, Korea. Her main research interests include architecture design for next generation multimedia systems, dynamic data encapsulation, adaptive cross layer design, dynamic adaption for heterogenous networks and N-screen content presentation.

Zhu Li  Zhu Li is now an Associate Professor with the Dept of Computer Science and Electrical Engineering (CSEE), University of Missouri-Kansas City, and director of the NSF I/UCRC Center for Big Learning (CBL) at UMKC. He received his PhD in Electrical and Computer Engineering from Northwestern University, Evanston in 2004. He was AFOSR SFFP summer visiting faculty at the US Air Force Academy (USAF), 2016 , 2017 , 2018 and 2020, with the UAV Research Center. He was Sr. Staff Researcher/Sr. Manager with Samsung Research America’s Multimedia Standards Research Lab in Richardson, TX, 2012-2015, Sr. Staff Researcher/Media Analytics Lead with FutureWei (Huawei) Technology’s Media Lab in Bridgewater, NJ, 2010 2012, and an Assistant Professor with the Dept of Computing, The Hong Kong Polytechnic University from 2008 to 2010, and a Principal Staff Research Engineer with the Multimedia Research Lab (MRL), Motorola Labs, from 2000 to 2008. His research interests include point cloud and light field compression, graph signal processing and deep learning in the next generation visual compression, image processing and understanding. He has 47 issued or pending patents, 100+ publications in book chapters, journals, and conferences in these areas. He is an IEEE senior member, associated editor for IEEE Trans on Image Processing(2020 ), IEEE Trans on Multimedia (2015-18), IEEE Trans on Circuits and System for Video Technology(2016-19), and Journal of Signal Processing Systems (Springer), since 2015. He serves on the steering committee member of IEEE ICME (2015-18), he is an elected member of the IEEE Multimedia Signal Processing (MMSP), IEEE Image, Video, and Multidimensional Signal Processing (IVMSP), and IEEE Visual Signal Processing and Communication (VSPC) Tech Committees. He is program co-chair for IEEE Int’l Conf on Multimedia and Expo (ICME) 2019, and co-chaired the IEEE Visual Communication and Image Processing (VCIP) in 2017. He received the Best Paper Award at IEEE Int’l Conf on Multimedia and Expo (ICME), Toronto, 2006, the Best Paper Award (DoCoMo Labs Innovative Paper) at IEEE Int’l Conf on Image Processing (ICIP), San Antonio, 2007.

Jun Sun  Jun Sun is currently a professor and Ph.D. advisor of Shanghai Jiao Tong University. He received his B.S. in 1989 from University of Electronic Sciences and technology of China, Chengdu, China, and a Ph.D. degree in 1995 from Shanghai Jiao Tong University, all in electrical engineering. In 1996, he was elected as the member of HDTV Technical Executive Experts Group (TEEG) of China. Since then, he has been acting as one of the main technical experts for the Chinese government in the field of digital television and multimedia communications. In the past five years, he has been responsible for several national projects in DTV and IPTV fields. He has published over 50 technical papers in the area of digital television and multimedia communications and received 2nd Prize of National Science and Technology Development Award in 2003, 2008. His research interests include digital television, image communication, and video encoding.