Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration

Qi Yang, Hao Chen, Zhan Ma, *Senior Member, IEEE*, Yiling Xu, *Member, IEEE*, Rongjun Tang, and Jun Sun

***Abstract*—Point cloud is emerged as a promising media format to represent realistic 3D objects or scenes in applications, such as virtual reality, teleportation, etc. How to accurately quantify the subjective point cloud quality for application-driven optimization, however, is still a challenging and open problem. In this paper, we attempt to tackle this problem in a systematic means. First, we produce a fairly large point cloud dataset where ten popular point clouds are augmented with seven types of impairments (e.g., compression, photometry/color noise, geometry noise, scaling) at six different distortion levels, and organize a formal subjective assessment with tens of subjects to collect mean opinion scores (MOS) for all 420 processed point cloud samples (PPCS). We then try to develop an objective metric that can accurately estimate the subjective quality. Towards this goal, we choose to project the 3D point cloud onto six perpendicular image planes of a cube for the color texture image and corresponding depth image, and aggregate image-based global (e.g., Jensen-Shannon (JS) divergence) and local features (e.g., edge, depth, pixel-wise similarity, complexity) among all projected planes for a final objective index. Model parameters are fixed constants after performing the regression using a small and independent dataset previously published. The proposed metric has demonstrated the state-of-the-art performance for predicting the subjective point cloud quality compared with multiple full-reference and no-reference models, e.g., the weighted peak signal-to-noise ratio (PSNR), structural similarity (SSIM), feature similarity (FSIM) and natural image quality evaluator (NIQE). The dataset is made publicly accessible at http://smt.sjtu.edu.cn or http://vision.nju.edu.cn for all interested audiences.**

***Index Terms*—3D-to-2D projection, point cloud, image features, quality assessment.**

I. Introduction

RECENTLY, thanks to the fast development of high-quality acquisition technologies, point cloud is emerged as an attractive media format for the realistic representations of the 3D objects and scenes in applications using a large amount of discrete and unordered points [1]. Each point contains its geometric information, as well as the photometric attributes (e.g., RGB colors). Similar as the conventional video or image content, point cloud data is also expected to be compressed for efficient storage and communication. Thus, the well-known Moving Picture Experts Group (MPEG) and Joint Photographic Experts Group (JPEG) have established expert groups to study and explore key techniques for the signaling and compression of the point cloud.

Though an international recommendation was concluded for point cloud compression [1], several critical issues have still remained open. One major challenge in this context is how to measure the point cloud quality *perceptually*, *quantitatively* and *efficiently*. We wish that such quality metric can well reflect the subjective sensation of our human visual system (HVS), can produce discriminative index scores for contents with various impairments, and can be derived easily for application-driven optimization.

A. Observations

However, existing studies or metrics on the point cloud quality often overlook one or more aforementioned key facts. For example, those metrics that are already applied during the MPEG point cloud compression (PCC) technology standardization, including the point-to-point (p2point) [2], point-to-plane (p2plane) [3] and point-to-mesh (p2mesh) [4], are mostly leveraging the Euclidean distance for error measurement between corresponding point vectors of original and impaired point clouds. The JPEG committee also has performed a serial subjective experiments [5] to study the performance of these point-based metrics. They have found that these objective models do not accurately correlated with the subjective scores. Oftentimes, related mean squared error (MSE) or equivalent PSNR is then calculated to quantitatively measure the distortion. Such measurement directly mimics the MSE or PSNR metric widely used for 2D image/video content, without taking the subjective characteristics of our HVS [6] into account.

In the meantime, we have also noticed that a number of explorations have been made recently on the subjective quality assessment of point cloud [7]–[11], leading to the consensus

Manuscript received February 29, 2020; revised July 18, 2020 and September 9, 2020; accepted October 15, 2020. Date of publication October 23, 2020; date of current version October 19, 2021. This work was supported in part by the National Key Research and Development Project of China Science, and Technology Exchange Center (2018YFE0206700, 2018YFB1802201), in part by the National Natural Science Foundation of China (61971282), and in part by the Scientific Research Plan of the Science and Technology Commission of Shanghai Municipality (18511105400). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lu Fang. *(Corresponding authors: Zhan Ma; Yiling Xu.)*

Qi Yang, Yiling Xu, Rongjun Tang, and Jun Sun are with the Cooperative Medianet Innovation Center, Shanghai Jiaotong University, Shanghai 200240, China (e-mail: yang_littleqi@sjtu.edu.cn; yl.xu@sjtu.edu.cn; thekey@sjtu.edu.cn; junsun@sjtu.edu.cn).

Hao Chen and Zhan Ma are with the Electrical Engineering, Nanjing University, Nanjing, Jiangsu 210093, China (e-mail: chenhao1210@nju.edu.cn; mazhan@nju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2020.3033117.

Digital Object Identifier 10.1109/TMM.2020.3033117

that popular metrics, such as the Root mean square (RMS) distance, or Hausdorff distance, that are utilized to measure the geometric similarity of 3D points, can not accurately describe the impairments of point cloud with both geometric and photometric distortions.

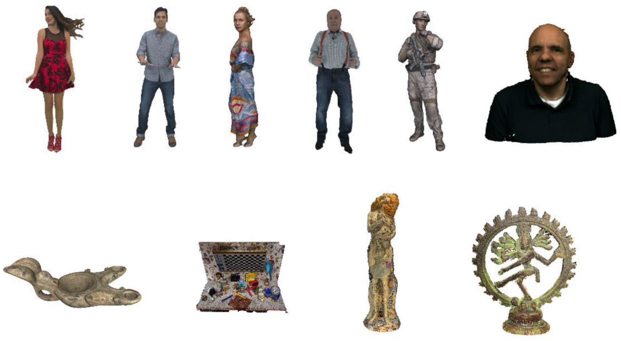

Fig. 1. Snapshots illustration of original point clouds used for subjective quality assessment. First row: “RedandBlack,” “Loot,” “LongDress,” “Hhi_ULLIWegner,” “Soldier,” “Ricardo”; Second row: “Romanoillamp,” “ULB_ Unicorn,’,’ “Statue Klimt,”“Shiva”. These point cloud models are from the common test dataset of international standardization groups for compression technology development. “Ricardo” is used for *training* in assessment.

B. *Our Approach*

In principle, point cloud distortions may come from the geometry component, or photometric attributes, or both of them, leading to drastically different perceptual sensations. To well understand and address the problem, we first organized a fairly large subjective quality assessment of point cloud with seven types of impairments (e.g., color noise, geometry noise, scaling, compression) at six different levels. Ten point cloud sequence samples were chosen from the common test datasets [12]–[15] that were selected by a number of experts and used for PCC standardization, covering a variety of content characteristics for generalization. Subjective assessment closely followed the standard ITU-R BT. 500 [16] where *Single Stimulus/Absolute Category Rating* protocol was involved to collect the mean opinion scores (MOS), with at least sixteen subjects for each processed point cloud sample (PPCS) after outliers removal. Here, PPCS is referred to as the point cloud sample with one specific impairment, which is the same as the processed video samples (PVS) used in video quality assessment [17]. In total, we have $10\times7\times6 = 420$ PPCS in this database for subjective rating.

We further attempted to develop appropriate metric for predicting these collected subjective MOSs. Currently, we had tried to devise a full-reference (FR) metric that could be potentially and directly utilized in the application system for quality related optimization, such as the compression, denoising, etc. As revealed in many successful metrics developed for image and video, such as MS-SSIM (Multi-scale Structural Similarity) [18], [19], VSI (Visual Saliency-induced Index) [20], edge saliency structure [21], binocular theory [22], free-energy principle [23], MSEA (Multiscale Edge Attension) [24], salient trajectories degradation [25], Cross-dimensional perceptual [26], etc, 2D content features, e.g., structures, saliency, edges, etc, were often extracted and fused as a single index to predict the MOS. These motivated us to attack this problem to leverage the existing and successful image features.

More specifically, we first projected the 3D point cloud onto six perpendicular cubic faces (e.g., front, back, left, right, upper, bottom plane/view in Fig. 3(b)), and then weigh the features of both color texture and depth images from different projection planes for the final quality index. Such 3D-to-2D projection-based exploration was well motivated by the facts that: 1) the identical projection mechanism is applied in MPEG standardized video-based point cloud compression (V-PCC) recommendation, demonstrating the state-of-the-art compression efficiency [1]; 2) 2D image features that were exclusively exploited in the past could be well leveraged; 3) Unequal quality weights could be easily adapted across different projection planes for application driven optimization (e.g., view-dependent streaming of point cloud [27]–[29]); 4) both geometric and photometric information of the point cloud were retained by the combination of projected color texture and depth images for quality measurement.

Previous visual neuroscience studies [30] suggested that *masking* effects were dominated in our HVS, leading to discriminative sensations of different stimuli. For example, we were more aware of luminance distortion than the chrominance component. A well-known example was the color space conversion from the RGB domain to YUV[1] for almost all image/video compression standards. Thus, we first converted the point cloud (a.k.a., projected color texture images equivalently) in RGB domain to the Gaussian color space that was more consistent with our HVS [31], [32]. As also studied by David Marr almost four decades ago [33], it reported that our vision system captures the object or scene progressively from the low-level local orientations (e.g., edges) to the mid- and high-level global structure (e.g., depth and contour), leading to the reasonable justification that both global and local distortions would impair our perceptual sensation. Thus, we then proposed to apply the Jensen-Shannon (JS) divergence for global similarity measurement, and utilized edge depth maps to aggregate the raw pixel similarity before being normalized to the local similarity. Normalization factor leveraged the image complexity to alleviate its masking impact. Finally, we offered an objective metric that weighed contributions from all projected planes, to produce the final quality index.

We split the proposed database into two parts, training and testing. We combine the training part with a small, and independent subjective point cloud assessment dataset published in [11][2] as training pool to train model parameters, and then applied the model to testing part for performance evaluation. In addition, we have performed the same plane dependent weighting strategy for well-known indexes for comparison, including PSNR, SSIM, MS-SSIM and etc. Results had demonstrated that our model had

[1] http://softpixel.com/~cwright/programming/colorspace/yuv/

[2] This dataset only provided the compression-induced artifacts. For model generalization, we had included a few PPCSs that were impaired with photometric and geometric noises from our dataset.

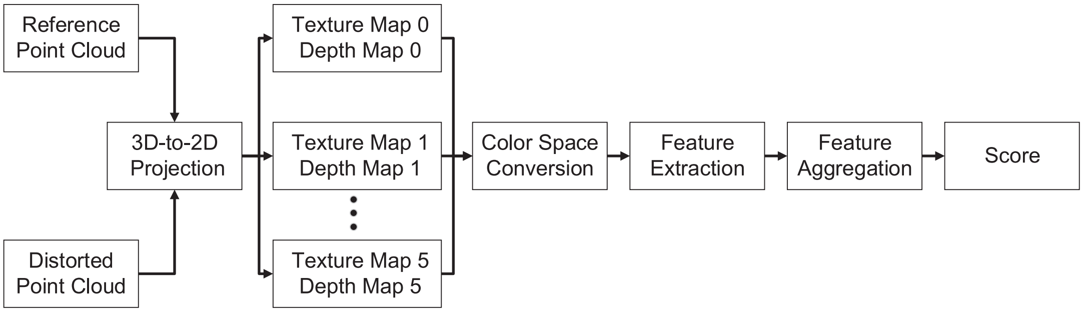

Fig. 2. **Objective Point Cloud Quality Metric.** Illustrative diagram of quality assessment from 3D-to-2D projection, color space conversion, image plane-based global and local feature extraction, and feature aggregation.

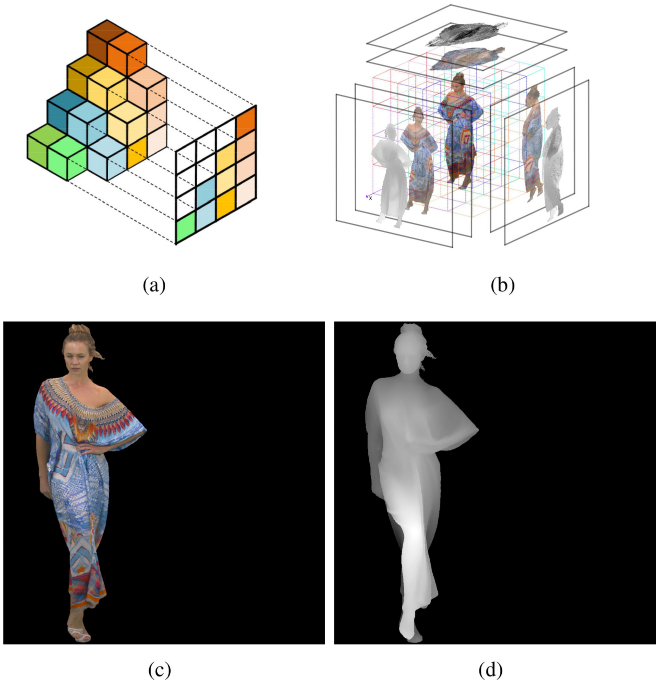
(a)
(b)
(c)
(d)

Fig. 3. **3D-to-2D Projection.** (a) Perspective projection mechanism; (b) Projected texture and depth images upon six faces of a cube (This figure comes from [49]); snapshots of a (c) color texture image in RGB space, and a (d) depth image, for the front view at the first projection plane ($i = 0$). "LongDress" is exemplified.

provided the state-of-the-art efficiency in predicting the subjective quality. However, we had also noticed that, compared with the image/video assessments and modeling, subjective point cloud quality studies were still in its infancy, requiring substantial efforts for further exploration. Thus, we would like to make the dataset in this work publicly accessible for all interested audience at http://smt.sjtu.edu.cn or https://vision.nju.edu.cn.

C. Contributions

The novelty of this paper are briefed as follows:

- We provide a subjective point cloud assessment dataset with 420 PPCSs, e.g., ten original point clouds with seven types of impairments (e.g., compression, photometric/geometric noise, scaling) at six different distortions levels, covering a wide range of content characteristics for application scenarios.
- We offer an objective metric to quantitatively predict the subjective MOS of a point cloud, via the weighted combination of global (e.g., JS divergence) and local features (e.g., edge, depth, pixel-wise similarity, complexity) from all perpendicular projection planes. To the best of our knowledge, this metric is the *first* one with explicit function forms to quantify the subjective quality.
- These features are extracted and aggregated using both the color texture image and depth images, to well reflect the geometric and photometric distortion in original 3D point cloud.
- We split the proposed dataset into two parts, e.g., training and testing. We merge the testing part with a small dataset published in [11] to derive the model parameters and keep them as fixed constants for performance evaluation. Experiments reveal that our metric provided the state-of-the-art efficiency in predicting the MOS.

The rest of this paper proceeds as follows: related works are reviewed in Section II, followed by the subjective assessment discussion in Section III. Section IV provides the development of the objective quality metric, with the performance evaluation in Section V. Ablation studies are then offered to further analyze the efficiency of proposed metric from its components in Section VI. Finally, concluding remarks are drawn in Section VII.

II. Related Work

This section reviews the explorations directly related to the quality studies of point cloud, including the subjective assessment, objective metrics, etc.

A. Subjective Assessment

Assessing the point cloud quality subjectively was relatively new and just emerged in recent years. Zhang *et al.* [9] studied the perceptual quality of 3D point cloud models with noise-induced photometric (color) and geometric (shape) distortions. Though it visualized the MOS trend with respect to the noise intensity, it did not conclude a closed-form function form that could be easily applied for quantitative optimization. Javaheri *et al.* [11] put their focus on the compression-induced distortion using Octree and Graph transform-based methods, and had analyzed the correlations between collected MOSs and existing geometric distance measurements (e.g., RMS and Hausdorff).

Notable efforts on assessing the point cloud quality had been devoted by experts in EPFL, leading to a number of pioneering publications in [7], [8], [10], [34]–[37]. These studies had covered a wide range of topics, from the rating setup, database, to the objective metrics performance comparison. It revealed that none of existing objective quality metrics could predict the subjective MOS very well, and encouraged the further investigations to explore better metric. Recently, Alexiou *et al.* [36] attempted to correlate the perceptual quality with respect to the weighted objective index (such as PSNR, MS-SSIM [18], etc) of projected image planes, where experimental studies had shown that there was no clear MOS prediction performance improvement when having more than six projection planes. Such observations have also supported our explorations in this work that we have utilized six perpendicular projection planes of a cube to facilitate the 3D-to-2D decomposition. Such projection mechanism was also adopted by MPEG V-PCC to enable the high-efficiency point cloud compression. On the other hand, our efforts had been laid on devising an appropriate closed-form model for predicting the 3D point cloud quality with various distortion impairments.

B. Objective Metrics

Most studies in the past had emphasized on the geometric distortion measurement of point cloud object, such as the point-to-point (p2point) [2], point-to-plane (p2plane) [3] and point-to-mesh (p2mesh) [4]. The p2point measured the degree of distortion by quantifying the distance (error) vector between corresponding points. The p2plane improved the p2point using the error vector projected along with the normal orientation; and the p2mesh first reconstructed the surface and then measured the distance from the point to the surface. Its efficiency, however, was heavily dependent on the surface reconstruction algorithm. Both p2point and p2plane were reported when evaluating the point cloud compression efficiency in MPEG reference software.

More recently, several new point cloud quality assessment models were derived based on point-based metrics. Alexiou *et al.* [35] proposed to use angular difference of point normals to measure the geometrical distortions. Meynet *et al.* [38] inherited the research results of Mesh via using local curvature statistics to reflect the point cloud surface distortion. Considering that Hausdorff distance is too sensitive to noise, Javaheri *et al.* [39] proposed a generalized Hausdorff distance by employing the Kth lowest distance rather than using the biggest distance. The results show that the generalized Hausdorff distance is more robust when dealing with compression-based distortion.

Above-mentioned metrics only consider point cloud geometrical distortion, while color information also plays an important role in human visual perception. Therefore, Meynet *et al.* [40] pooled point cloud curvature and color lightness together via optimally-weighted linear combination. Viola *et al.* [41] used color histogram to drive objective metrics. Alexiou *et al.* [42] explored four types of attributions, e.g., geometry, normal vectors, curvature values and colors, and normalized these features in the form of SSIM [19]. Yang *et al.* [43] proposed to use graph signal processing to extract point cloud color gradient and realize robust quality prediction. However, these models present relatively high computation complexity due to some operations, such as surface fitting and graph construction.

TABLE I
POINT CLOUD SAMPLE ILLUSTRATION

name	#points	axis range		
		$[x_{\min}, x_{\max}]$	$[y_{\min}, y_{\max}]$	$[z_{\min}, z_{\max}]$
RedandBlack	729133	[182,575]	[10,987]	[121,353]
Loot	784142	[28,380]	[7,999]	[119,473]
LongDress	806806	[151,397]	[5,1012]	[87,523]
Hhi_ULLIWegner	900153	[0,61875]	[0,64057]	[0,170135]
Solider	1059810	[29,389]	[7,1023]	[31,436]
Ricardo	960703	[127,870]	[507,803]	[0,605]
ULB_Unicorn	2000297	[0,2826]	[0,1503]	[0,1663]
Romanoillamp	1286052	[-28,59]	[-126,46]	[-77,0]
Statue Klimt	499886	[0,921]	[0,3112]	[0,899]
Shiver	1010591	[0,1081]	[0,969]	[0,517]

These newly proposed point cloud metrics were closely related to the structural similarity applied in 2D image/video contents. As aforementioned, 3D point cloud could be projected to a number of 2D planes. Thus, a straightforward means was to apply the weighted quality indices of these image planes [35], [37]. However, there is information loss during projection (see Sec. VI), therefore how to derive an effective point cloud quality assessment model over 2D planes still requires further exploration.

As reported and advocated in those publications, it still requires substantial efforts to develop objective metrics to accurately predict the subjective 3D point cloud quality. One aspect is to generate sufficient large point cloud databases with subjective scores (MOSs), such as the [11]; and the other aspect is using these datasets to devise analytical objective metric for subjective quality prediction. Our work in this paper has tried to attack the point cloud quality modeling problem in both aspects.

III. SUBJECTIVE QUALITY ASSESSMENT

To thoroughly understand the point cloud quality, we first perform the subjective quality assessment to collect MOSs with respect to different impairments.

A. Point Cloud Database

Ten different point cloud content, including six human body models (e.g., "RedandBlack," "Loot," "LongDress," "Hhi_ULLIWegner," "Soldier," and "Ricardo") and four inanimate objects (e.g., "ULB_Unicorn," "Romanoillamp," "Statue Klimt," "Shiva"), are chosen as the original samples. These point clouds are selected by a number of experts to cover a variety of content characteristics for generalization in prospective applications, and are utilized for the compression standardization by the MPEG and JPEG point cloud groups [12]–[14], [44] and for the product development by industry leaders [15]. Table I gives the basic information of these point clouds (e.g., number of points, dimensional ranges of x, y and z axes), and Fig. 1 illustrates the snapshots.

Distortions will be inevitable when processing the point clouds, including the acquisition noise, re-sampling, compression, etc. They can be applied individually, such as Octree-based compression (OT), Color noise (CN), Geometry Gaussian noise (GGN), Downscaling (DS); or superimposed, such as the Downscaling and Color noise (DS+CN or D+C), Downscaling and Geometry Gaussian noise (DS+GGN or D+G), Color noise and Geometry Gaussian noise (CN+GGN or C+G). To avoid misunderstanding, we refer to the original point cloud with impairments to as the PPCS. More specifically, each original point cloud is processed with seven different types of distortions at six different levels, with details shown below.

1) OT: Compression noise is exemplified using the octree pruning method provided in well-known Point Cloud Library (PCL) (http://pointclouds.org/downloads/). Octree pruning removes leaf nodes to adjust tree resolution for compression purpose. Here, we have experimented different compression levels by removing points at 13%, 27%, 43%, 58%, 70% and 85%. It is difficult to guarantee the point removal percentage at the exact number. Thus we allow ±3% deviation.
2) CN: Color noise, or photometric noise is applied to the photometric attributes (RGB values) of the points. We inject the noise for 10%, 30%, 40%, 50%, 60%, and 70% points that are randomly selected, where noise levels are respectively and again randomly given within ±10, ±30, ±40, ±50, ±60, and ±70 for corresponding points (e.g., 10% random points with ±10 noise, 30% random points with ±30, and so on so forth). Noise is equally applied to R, G, B attributes. Cropping is used if the noisy intensity $\tilde{p} = p + n$, is out of the range of [0, 255], e.g., if $\tilde{p} < 0$, $\tilde{p} = 0$; and if $\tilde{p} > 255$, $\tilde{p} = 255$.
3) GGN: We apply Gaussian distributed geometric shift to each point randomly. In this study, all the points will be augmented with a random geometric shift that is within 0.05%, 0.1%, 0.2%, 0.5%, 0.7%, 1.2% of the bounding box.
4) DS: We randomly downsample the point clouds by removing 15%, 30%, 45%, 60%, 75%, 90% points from the original point clouds. We directly utilize the downscaling function pcdownscample() offered by the Matlab software.
5) DS+CN or D+C: We combine aforementioned DS and CN where the downsampling process is firstly applied and then the color noise is added in a consecutive order, e.g., 15% DS and 10% random points with ±10 noise, 30% DS and 30% random points with ±30 noise, and so on so forth).
6) DS+GGN or D+G: GGN and DS are superimposed. The DS process is firstly applied before augmenting the GGN consecutively, e.g., 15% DS with 0.05% GGN, 30% DS with 0.1% GGN, and so on so forth).
7) CN+GGN or C+G: Both GGN and CN are superimposed. The GGN is firstly applied, and then is the CN, e.g., 0.05% GGN and 10% random points with ±10 noise, 0.1% GGN and 30% random points with ±30 noise, and so on so forth).

Note that octree-based compression OT is not superimposed with any other distortions. This is mainly because applying the OT itself already includes the effects of quantization noise, and scaling for both geometry and photometric components by leaf points pruning.

It is also worth to point out that we have applied a slightly different scheme for color noise CN implementation (without geometric variations). Normally we would add random noise to all pixels for images or video frames. Nowadays, point cloud generation is often utilizing multiple-cameras, e.g., RGB cameras, depth cameras, etc. These cameras would present different noise sources. Thus, we choose to add random noise to random selected points. There are numerous combinations for such noise injection. To reduce the complexity of experiments, we have coupled the range of noise (e.g., ±30, but the actual noise level is random within this range) and the percentage of points to be selected for noise augmentation.

It includes at least $36 = 6\times 6$ permutation options if we combine any two individual noises. Instead, we choose to augment different noise at the same distortion level for aforementioned D+C, D+G, and C+G. In this way, our method is to sub-sample all possible permutations, but still could cover the entire noise range.

B. Subjective Rating

Because single stimulus can avoid the influence of media scenario and collect more justice results, it has been widely used in 3D media quality assessment, e.g., [45][46]. Therefore in this paper, we also adopted the *single stimulus* method for subjective rating, and all the steps are compliant with the ITU-R Recommendation BT. 500 [16]. We use the CloudCompare [47] as our point cloud rendering software for collecting the MOS. Specifically, the zoom rate is set as 1:1, all the samples are cached in the software and listed in “DB Tree” (a window in the software). Moving from one sample to another can be easily implemented by click the titles presented in “DB Tree”. To facilitate the recording of scores, we set up a special score recorder. After scoring, experiment organizer will switch samples and presents initial view to the observers. The computer used in subjective experiments is equipped with a popular Dell SE2216H monitor (21.5 inches and the resolution is 1920×1080), and configured with an Intel i5-6300 HQ CPU, 8-GigaBytes RAM and 1-TeraBytes hard drive. All participants are asked to sit on a adjustable chair to ensure their eyes at the same height as the center of the screen. The viewing distance is about 3× height of the rendered point cloud (≈ 0.75 meter). The evaluation process is conducted indoors, under a normal lighting condition. Besides, we do not use any shadow or color enhancement during rendering, and keep the color attributes as the raw input source.

Each PPCS is rendered on the display in a full screen mode. Operations, such as zoom-in and zoom-out are not allowed in this study, but rotation is enabled to let the users simulate the free-view navigation. Every individual PPCS takes about 15 seconds for each participant, leaving another 5 seconds for rating before proceeding to the next PPCS. Raw scores are given in the range of [1, 10], associated with five quality scales (e.g., 1-2: bad, 3-4: poor, 5-6: fair, 7-8: good and 9-10: excellent) [16].

For each subject, the entire subjective experiment is divided into *training* and *testing*. Training session lasts about 5 minutes (≈15 PPCSs), by which we use "Ricardo" with certain impairments that are ordered along with the distortion intensity levels, to let the subjects familiarize themselves with the point cloud content, distortion-induced quality range, and operational steps to navigate the content. Such training would alleviate the unexpected noise in testing session, benefiting the subsequent data analysis and modeling. Testing session uses the remaining 378 = 9×7×6 PPCSs to collect subjective scores. However, it is impossible to have a single subject to complete the entire 378 PPCSs (a.k.a., 120 minutes). Most subjects will feel dizzy and tired beyond 30-minute rating duration, which would severely reduce the robustness of the subjective assessment, leading to unreliable rating outcomes. To tackle the problem, we have applied the *random sampling* mechanism to divide these 378 samples into four sub-groups, each of which will have 94 or 95 PPCSs.

There are sixty-four people in total participating into the subjective experiments. All these subjects are between 18 and 30 years old with normal vision or after correction. Most of them have no experience of subjective experiments, and are naïve with the point clouds. Following the convention [16], [17], each PPCS is assessed with at least 16 individuals.

C. Data Post-Processing

We process the collected raw subjective scores following the methods described in [17]. Z-scores are first calculated to normalize the raw ratings of each subject:

$$Z_{m,i} = \frac{X_{m,i} - \mu(\mathbf{X}_i)}{\sigma(\mathbf{X}_i)}, \tag{1}$$

where $X_{m,i}$ and $Z_{m,i}$ denote the raw rating and Z-score of m-th point cloud from i-th participant, respectively. $\mathbf{X}_i$ is a vector containing all PPCS ratings from i-th participant. $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation operands, respectively.

Screening method described in BT. 500 [16] is deployed to remove the outliers whose scores are inconsistent with other participants. As a result, six participants are marked as outliers and removed; the remaining 58 participants are kept with their ratings for MOS derivation.

After screening, we scale the Z-scores back to [1-10] scale, using:

$$\begin{aligned}\hat{X}_{m,i} = &\ (\mu(\mathbf{X}_{\max}) - \mu(\mathbf{X}_{\min})) \\ &\times \frac{Z_{m,i} - Z_{i,\min}}{Z_{i,\max} - Z_{i,min}} + \mu(\mathbf{X}_{\min}),\end{aligned} \tag{2}$$

where $\mathbf{X}_{\max}$ and $\mathbf{X}_{\min}$ are the maximum and minimum scores of a given point cloud content with all distortions from all participants, $Z_{i,\max}$ and $Z_{i,\min}$ are the maximum and minimum Z-scores of the i-th participant for the same point cloud. After applying (2), the scores from all the participants have a common range of $[\mu(\mathbf{X}_{\min}), \mu(\mathbf{X}_{\max})]$. In our subjective test data, $\mu(\mathbf{X}_{\min}) = 1$ and $\mu(\mathbf{X}_{\max}) = 10$. Then, the MOS value for a particular PPCS is derived by averaging the common scores $\hat{X}_{m,i}$ for all participants after all the steps.

IV. Modeling the Point Cloud Quality Via 3D-to-2D Projection

This section leverages the subjective MOSs collected in last section to develop appropriate objective metric for perceptual point cloud quality prediction. Existing objective metrics developed for image and video quality estimation are mostly dependent on the content characteristics, such as edge, saliency, complexity, etc. In order to leverage these well-developed mechanisms for 2D image/video, we have projected the 3D point clouds onto perpendicular image planes for subsequent feature extraction and aggregation for quality prediction.

Figure 2 gives an overview of the systematic explorations in this paper, including the *3D-to-2D projection* to convert the original 3D point cloud to six pairs of color texture and depth images, *color space conversion* to map native RGB samples to the Gaussian color space that is more consistent to our HVS, global (e.g., JS divergence) and local (e.g., depth, edge, pixel-wise similarity, complexity) *feature extraction*, and *feature aggregation* for final index score.

A. 3D-to-2D Perpendicular Projection

Each 3D point cloud is mapped to six perpendicular 2D image planes ($0 \leq i \leq 5$) via the perspective projection, as shown in Fig. 3. These image planes, e.g., referred to as the "front," "back," "left," "right," "top" and "bottom" planes,[3] correspond to six faces of a cube. Upon each projection plane, at each 2D location, its depth and RGB colors are captured, resulting in associated depth and color texture image pairs. The same projection mechanism is also utilized in MPEG V-PCC, where projected texture and depth images are organized and encoded using the High-Efficiency Video Coding [1], [48]. We follow this convention to perform the feature extractions for all projected images.

Oftentimes, the front view after projection is assumed to offer the most attractive or interesting information, such as the front faces of these point clouds in Fig. 1. In our experimental platform, we set this as the initial view for point cloud rendering. Specifically, we present the initial view of each sample in the proposed database in Fig. 4.

We perform point cloud projection via matlab [49] and implement it as follows: First, identifying the projection plane, for example, X-O-Y; then, for points of sample, we plot them on the plane as texture image via its x coordinates and y coordinates, and reserve its z coordinates as depth information in the depth map. If multiple points share same x and y while different z, we keep the point which are closer to the projection plane (e.g., smaller z), other points are obscured. The resolution of the projected texture and depth images are 1280*1280. This function

[3] These six projection planes are also sometimes referred to as the corresponding views, e.g., "front view," "back view," "left view," "right view," "top view" and "bottom view".

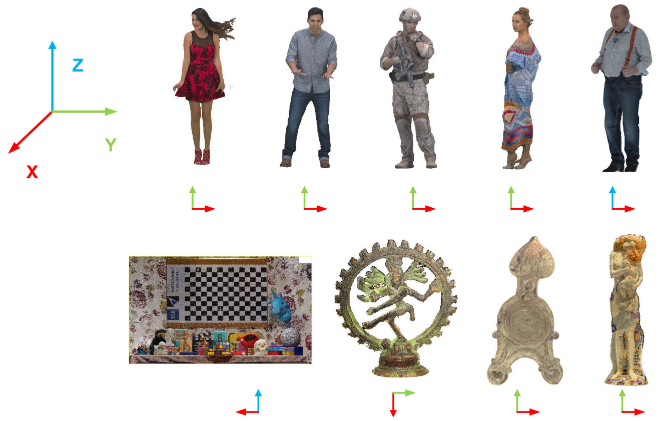

Fig. 4. Initial view of samples in the proposed database. The mini coordinate system under each sample illustrates the projection plane with the initial view obtained. For example, for "RedandBlack," we can obtain its texture image of initial view via projecting it to Y-O-X plane.

output these six images under fixed sequence, in order to establish relationship between projection planes and six views, we set labels for each simple to convenient follow-up processing.

B. Color Space Conversion

We often wish to extract features that can be combined to efficiently represent the sensations of the visual stimuli. A good example is having them from uncorrelated domain (e.g., color space) for subsequent weighted fusion. However, native RGB color space typically presents high correlation in between. Hence, a color space conversion is usually applied. An analogous example is the RGB to YUV conversion that is widely adopted in video compression standards. Cross-color correlation can be well decomposed, leading to more efficient and compact image representation in YUV domain than in native RGB space.

Recent studies in [31], [32] have also suggested that Gaussian color model (GCM) is more consistent with our HVS than RGB space. Thus, in the next derivations, we will first covert the color texture images into corresponding Gaussian images via,

$$\begin{bmatrix} \widehat{E} \\ \widehat{E}_{\lambda} \\ \widehat{E}_{\lambda\lambda} \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{3}$$

Here, $\widehat{E}$, $\widehat{E}_{\lambda}$ and $\widehat{E}_{\lambda\lambda}$ represent luminance and two chrominance channels Cs (e.g., $C \in [\widehat{E}, \widehat{E}_{\lambda}, \widehat{E}_{\lambda\lambda}]$) of Gaussian color model. Snapshots of the front view of "Soldier" are given in Fig. 5 with native RGB color in subplot 5(a), and respective Gaussian color components. As shown, luminance component $\widehat{E}$ in Fig. 5(b) retain the spatial details of texture images, leaving color characteristics and features embedded in $\widehat{E}_{\lambda}$ and $\widehat{E}_{\lambda\lambda}$ of Fig. 5(c) and 5(d).

In next paragraphs, we will give more details regarding the feature extractions from the Gaussian color images, e.g., both luminance and chrominance components.

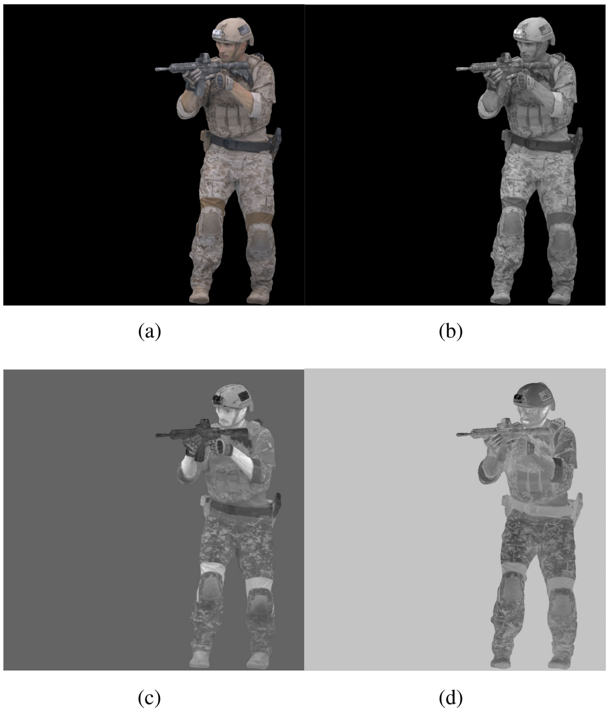

Fig. 5. **Color space Conversion.** Snapshots illustration of front view of "Soldier" point cloud. (a) Projected texture image in RGB space; (b) Luminance component $\widehat{E}$, (c) Chrominance component $\widehat{E}_{\lambda}$, and (d) Chrominance component $\widehat{E}_{\lambda\lambda}$ of Gaussian color model.

C. Feature Extraction

For each projection plane, we have both 2D texture and depth images to represent the photometric and geometric information of the original point cloud, as shown in Fig. 3(c) and 3(d). We have attempted to utilise features extracted from these images for objective metric development.

Perceptual quality sensation is complicated and often superimposed with many effects (e.g., masking, frequency selectivity, etc). As also reported in [33], local stimuli will be first perceived (such as the local orientation, edge, depth, etc), and then global stimuli are augmented (e.g., structural contours of object/scene that is possibly generated by connecting aforementioned local features). Thus, we have tried to make a hypothesis that the overall perceptual quality are dominantly contributed by the *local* and *global* parts, i.e.,

$$Q_{\texttt{tot}} = F_{\texttt{local}}^{\eta_l} \cdot F_{\texttt{global}}^{\eta_g}. \tag{4}$$

In this basic model, we simply use η_l and η_g to quantitatively weigh the contributions from the local and global features. Without losing the generality, we start with the i-th image plane. But to ensure the simple presentation, we omit the index i for image plane indication during model derivations.

1) Local Features: We first try to extract appropriate local features for quality prediction. Generally, users will be more sensitive to the distortions to the edge area [24], and to the objects that are closer in distance (e.g., small depth). Thus, we

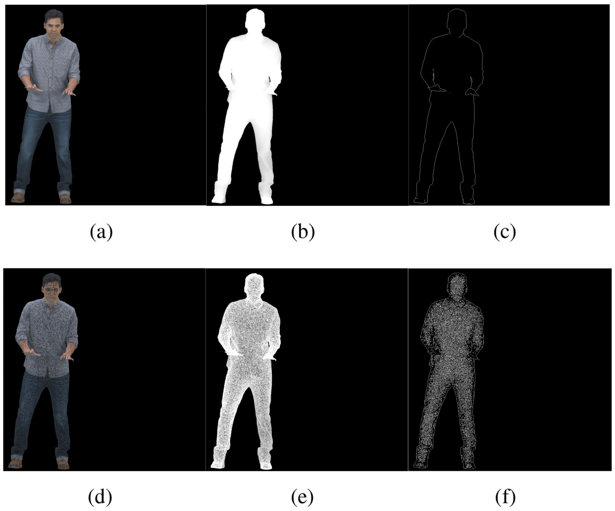

Fig. 6. **Depth-Edge Map.** First row is the reference after projection, including (a) color texture image in RGB space, (b) depth map, (c) depth-edge map ω_{de}; and the second row is the impaired one with down-scaling artifacts, having respective down-scaled (d) RGB image, (e) depth map, (f) depth-edge map ω_{de}. Down-scaled images are scaled-up at the same size for comparative illustration. "Loot" is used as an example.

aggregate the *depth-edge map* upon the texture similarity between the reference and impaired image signal, as a local feature representation.

More specifically, depth-edge map ω_{de} is derived by applying the *Sobel* detector ($\mathbb{S}$) on both reference and impaired depth images (e.g., $\mathbf{D}_r$ and $\mathbf{D}_d$), and max-pooling [50], e.g.,

$$\omega_{de}(x,y) = \max\left(\mathbb{S} * \mathbf{D}_r(x,y), \mathbb{S} * \mathbf{D}_d(x,y)\right). \tag{5}$$

This simple max-pooling that is widely used in recently emerged deep neural networks, is used here to extract sharp difference between the reference and impaired signal for distortion measurement. Note that each element in this edge-depth map ω_{de} is a binary value, e.g., 0 or 1. An illustrative comparison is given in Fig. 6, where the upper part is the reference and bottom part is the impaired one processed with downscaling operation. Depth-edge map can clearly reveal the distortion areas. Other artifacts have a similar pattern, showing that depth-edge map is an effective way to extract important information for distortion measurement.

On the other hand, texture similarity image $\mathbf{S}_C$ between the reference $\mathbf{C}_r$ and distorted $\mathbf{C}_d$ signals at channel C can be calculated following the popular pixel-wise approach suggested by [51], [52], i.e.,

$$\mathbf{S}_C(x,y) = \frac{2\mathbf{C}_r(x,y)\cdot\mathbf{C}_d(x,y) + T}{\mathbf{C}_r^2(x,y) + \mathbf{C}_d^2(x,y) + T}. \tag{6}$$

Note Eq. (6) applied on the pixel value. Different color channels can be derived using the same way. In this paper, we have one luminance channel and two chrominance channels. Usually, T is a small and non-negative constant to prevent numerical instability [19]. We set it as 0.0001 in this paper.

We then aggregate the texture similarity using the binary depth-edge map for a specific channel C to have the depth-edge activated similarity index f_C^{des}, e.g.,

$$f_C^{des} = \frac{\sum_{(x,y)\in\Omega} \omega_{de}(x,y) \cdot \mathbf{S}_C(x,y)}{\sum_{(x,y)\in\Omega} \omega_{de}(x,y)}, \tag{7}$$

with Ω for the complete pixel locations, and $C \in [\widehat{E}, \widehat{E}_\lambda, \widehat{E}_{\lambda\lambda}]$.

Oftentimes, our HVS will be masked by certain range of frequency from the underlying content. For example, we may have the same depth-edge activated similarity index f_C^{des} for different content. But the one having very complex spatial details is less sensitive to the distortion than the other having clear and simple spatial pattern [53]–[55]. Thus, a content complexity factor is introduced to adapt the F_C^{des}, where we utilize the method in [56] to measure the content complexity of a projected image at a specific channel C, e.g.,

$$\alpha_C = -\log \int p_C(\rho) \cdot (\log p_C(\rho))\, d\rho, \tag{8}$$

where $p_C(\rho)$ stands for the probability density function of ρ in the error map $\mathbf{e}_C$ between the reference image and its associated filtered correspondence at channel C. Bilateral filtering processing is performed with parameters derived using neighbors in a autoregressive means. More details can be found in [56].

By combining Eq. (7) and (8), it leads to the local feature index for a specific image channel C, and at a given i-th projection plane, e.g.,

$$f_{\texttt{local}}(C) = \frac{f_C^{des}}{(\alpha_C)^{\epsilon_C}}. \tag{9}$$

Here, ϵ_C is used to generalize the impact of the complicate complexity masking phenomenon in a rather simple way.

2) Global Feature: Kullback-Leibler (KL) divergence [57] between the approximated posterior against the true posterior is used to describe the difference sensed by our visual system for a reference and its corresponding distorted version. It has also been utilized in free energy theory [58], [59] to quantify the image/video quality as in [56]. In other words, KL divergence measures the degree of deviations from the target probability distribution to the other reference distribution. It is asymmetric, and also referred to as the *self similarity* [60].

We extend this asymmetric KL divergence to the symmetric Jensen shannon (JS) divergence [61] via

$$JS(p||q) = \frac{1}{2}KL\left(p||\frac{p+q}{2}\right) + \frac{1}{2}KL\left(q||\frac{p+q}{2}\right), \tag{10}$$

where p and q are two probability distributions. JS divergence is symmetrical with $JS(p||q) = JS(q||p)$, and its value is of [0, 1] with "0" for the most similar and "1" for the most diverged scenarios. Thus, we can represent the global similarity using (1 - JS).

Though we can mathematically derive JS divergences for both luminance and chrominance components, global feature is mainly determined by the spatial distribution of the luminance part. Given that our HVS is also more sensitive to the luminance component and chrominance parts do not introduce noticeable gains if being included, we will utilize the luminance

JS divergence as the *global feature* for quality prediction, i.e.,

$$f_{\texttt{global}} = 1 - JS\left(p_{\widehat{E}_r} || p_{\widehat{E}_d}\right), \tag{11}$$

with $p_{\widehat{E}_r}$ and $p_{\widehat{E}_d}$ for the respective probability distributions of reference and distorted texture images in same projection orientation. Our experimental studies have shown that JS divergence offers higher correlation with the collected MOS than original KL measurements.

D. Objective Quality Index

We then introduce the image plane weighting factor κ_i to linearly weigh the contributions from six image planes, resulting in

$$F_{\texttt{global}} = \sum\nolimits_i \kappa_i \cdot f_{\texttt{global}}(i), \tag{12}$$

$$F_{\texttt{local}} = \sum\nolimits_C \gamma_C \cdot \left(\sum\nolimits_i \kappa_i \cdot f_{\texttt{local}}(i, C)\right), \tag{13}$$

with $i \in [0, 5]$ and $C \in [0, 2]$ as image plane index and color channel index, and γ_C as the channel weighting factor.

By injecting (9), (11) into respective (13), (12), and then (4), we could finally reach at the overall quality metrics at

$$\begin{aligned} Q_{\texttt{tot}} = & \left(\sum\nolimits_i \kappa_i \cdot \left(1 - JS\left(p_{\widehat{E}_r(i)} || p_{\widehat{E}_d(i)}\right)\right)\right)^{\eta_g} \\ & \times \left(\sum\nolimits_C \gamma_C \cdot \left(\sum\nolimits_i \kappa_i \cdot \frac{f_C^{ds}(i)}{(\alpha_C(i))^{\epsilon_C(i)}}\right)\right)^{\eta_l}. \end{aligned} \tag{14}$$

Our extensive simulations have shown that $\epsilon_C(i)$ can be replaced by a constant ϵ for all projection planes and color channels without noticeable performance degradation. Hence, we have parameters $\kappa_i, i \in [0, 5], \gamma_C, C \in [\widehat{E}, \widehat{E}_\lambda, \widehat{E}_{\lambda\lambda}], \epsilon, \eta_g$ and η_l.

V. Model Performance Evaluation

This section examines the efficiency of the proposed objective model for subjective point cloud quality estimation.

A. Model Parameters

In order to apply the proposed model (14) for subjective quality prediction in practice, we first have to determine the model parameters quantitatively.

Projection plane weighting factors κs. In current testbed setup, initial view of each point cloud is its salient front face (shown in Fig. 1), which is then projected to the corresponding image plane as the "front view". Other views are projected accordingly to another five perpendicular image planes. Note because the samples in the proposed database have different coordinate scale, we resize their coordinate according to the actual size of rendering area in the CloudCompare. Specifically, for each sample, there is a certain bounding box after rendered by CloudCompare, e.g., for "RedandBlack," the bounding box is $[280, 700, 160]_{x,y,z}$. In order to ensure the fair subject assessment, we enforce the same-size presentation of projected image and its corresponding 3D model. Note that projected images may be resized to fit the range of the bounding box that is specified for each point cloud in CloudCompare. During the rating process, we allow the subject to freely rotate the point cloud with sufficient time for content consumption, e.g., 15 seconds. Thus, we assume that the overall quality sensation can be represented as the aggregated sensation among all image planes, and weighting factor κ_i can be simply referred to as the normalized duration spent upon a specific plane. The same weighting strategy is also reported in [36]. For our assessment, we have observed that users would spend the most time on the front view, and then the back view, followed by the left and right views, and the least time for upper and bottom views. According to our simulations, we set $\kappa = [0.5, 0.2, 0.1, 0.1, 0.05, 0.05]$, representing coefficients of respective "front," "back," "right," "left," "top," and "down" planes (see Fig. 1).

Another promising setup of κ can be saliency-driven scheme [20], where the weighting coefficients can be adapted by the saliency distribution. This fits the fundamental intuition that the overall perceptual sensation is highly correlated with the content saliency to which users would pay more attention. However, it still lacks of efficient algorithms to quantify the point cloud saliency. Thus, we defer this as our future study.

η_g, η_l, γ_C and ϵ. There is another database developed for subjective point cloud quality assessment in [11] where only compression distortions (e.g., Octree and Graph transform with three difference levels) are considered for a subset of point cloud sequences (e.g., six out of ten in Table I using "Statue Klimt," "Shiva," "Egyptian mask," "Loot," "Longdress" and "RedandBlack"). In addition, we split the proposed database into two parts, training part and testing part (e.g., "Hhi_ULLIWegner" and "ULB_Unicorn" are selected as training part) to avoid overfitting. Combined with samples in [11], we have 6×3×2 + 2×7×6 = 120 PPCSs in total to form a *training* pool for model parameter derivation of (14) via least squared error-based fitting. We will keep these parameters as fixed constants, i.e., $\eta_g = 1$, $\eta_l = 1, \epsilon = 0.4, \gamma_{\widehat{E}} = 0.1, \gamma_{\widehat{E}_\lambda} = 25, \gamma_{\widehat{E}_{\lambda\lambda}} = 15$ and apply them to our dataset with testing PPCSs (e.g., 7×7×6 = 294) for performance evaluation ("Ricardo" is used for subjective experiment training and not included for evaluation).

B. Performance Evaluation

To ensure the consistency between the objective and subjective evaluation scores (e.g., MOS or DMOS) of the various quality assessment models, the video quality experts group (VQEG) [6], [68] recommends to map the dynamic range of the scores from objective quality assessment models into a common scale using

$$Q_i = k_1\left(\frac{1}{2} - \frac{1}{1 + e^{k_2(s_i - k_3)}}\right) + k_4 s_i + k_5. \tag{15}$$

Here, s_i is the calculated objective score of the i-th distorted point cloud from a quality assessment model (such as (14)), Q_i is the corresponding mapped score. k_1, k_2, k_3, k_4 and k_5 are the regression model parameters to be fitted by minimizing the sum of squared differences between the objective and subjective evaluation scores.

Three different performance metrics commonly used in quality assessment society are offered to quantify the efficiency of our proposed metric in (14), including Pearson linear correlation

TABLE II
PERFORMANCE EVALUATION ON SUBJECTIVE POINT CLOUD QUALITY PREDICTION

		PLCC							SROCC							RMSE						
	Distortion:	OT	CN	GGN	DS	D+C	D+G	C+G	OT	CN	GGN	DS	D+C	D+G	C+G	OT	CN	GGN	DS	D+C	D+G	C+G
People	PSNR	0.49	0.70	0.84	0.55	0.45	0.60	0.85	0.43	0.71	0.79	-0.34	-0.18	0.39	0.79	1.28	1.33	1.34	1.94	2.19	2.06	1.34
	PSNR-HVS-M [63]	0.61	0.68	0.95	0.53	0.46	0.60	0.93	0.50	0.72	0.90	-0.31	-0.16	0.43	0.90	1.17	1.37	0.78	1.96	2.17	2.07	0.97
	SSIM [19]	0.78	0.58	0.95	0.41	0.00	0.48	0.95	0.62	0.72	0.92	-0.29	-0.14	0.44	0.91	0.92	1.51	0.74	2.11	2.45	2.26	0.83
	MS-SSIM [18]	0.74	0.43	0.95	0.22	0.32	0.43	0.95	0.65	0.72	0.93	-0.18	-0.03	0.46	0.92	0.99	1.68	0.75	2.26	2.36	2.33	0.79
	IW-SSIM [64]	0.78	0.59	**0.98**	0.77	0.14	0.74	**0.98**	0.64	0.70	0.95	-0.36	-0.16	0.63	**0.96**	0.93	1.51	**0.51**	1.46	2.43	1.75	**0.49**
	FSIM [52]	**0.80**	0.53	0.75	0.46	0.46	0.56	0.77	0.60	0.69	0.71	0.01	0.25	0.55	0.74	**0.87**	1.57	1.65	2.05	2.18	2.14	1.62
	VIF [65]	**0.80**	0.69	**0.98**	0.80	0.72	**0.95**	**0.98**	**0.73**	0.59	**0.96**	**0.77**	0.67	**0.91**	**0.96**	0.88	1.35	0.55	1.39	1.71	**0.81**	0.52
	NIQE [66]	0.78	0.15	0.50	**0.89**	0.55	0.75	0.43	0.55	-0.02	0.23	-0.01	-0.25	-0.13	0.30	0.92	1.84	2.14	**1.06**	2.05	1.70	2.30
	IL-NIQE [67]	0.41	0.25	0.50	0.42	0.67	0.35	0.12	0.33	0.39	-0.19	0.06	0.15	-0.06	0.03	1.34	1.80	2.15	2.10	1.82	2.49	2.54
	OG-IQA [68]	0.38	0.18	0.56	0.42	0.54	0.43	0.37	-0.09	0.04	0.10	-0.23	0.52	0.24	0.08	1.36	1.83	2.08	2.10	2.07	2.33	2.39
	Proposed	0.66	**0.97**	0.83	0.78	**0.92**	0.77	0.96	0.64	**0.95**	0.82	0.63	**0.94**	0.74	0.93	1.10	**0.48**	1.38	1.44	**0.95**	1.66	0.70
Inanimate	PSNR	0.55	0.56	0.68	0.38	0.41	0.66	0.64	-0.36	-0.12	0.41	-0.16	-0.12	-0.18	0.36	1.94	0.80	1.86	1.85	2.11	1.88	1.86
	PSNR-HVS-M [63]	0.70	0.51	0.60	0.37	0.51	0.51	0.60	-0.21	0.34	0.49	-0.17	-0.08	0.31	0.44	1.64	0.83	2.02	1.86	1.98	2.15	1.93
	SSIM [19]	0.47	0.35	0.62	0.48	0.47	0.58	0.64	0.11	0.34	0.51	-0.06	0.03	0.43	0.49	2.03	0.90	1.99	1.75	2.04	2.04	1.86
	MS-SSIM [18]	0.46	0.27	0.62	0.48	0.46	0.51	0.64	0.09	0.34	0.43	-0.08	-0.02	0.29	0.40	2.03	0.92	1.99	1.75	2.04	2.15	1.87
	IW-SSIM [64]	0.75	0.44	0.82	0.75	0.70	0.89	0.90	0.13	0.43	**0.82**	0.26	0.34	**0.79**	0.85	1.53	0.86	1.45	1.33	1.63	1.15	1.06
	FSIM [52]	0.89	0.37	0.73	0.70	0.22	0.65	0.77	0.45	0.34	0.68	0.21	0.31	0.55	0.63	1.04	0.89	1.74	1.43	2.24	1.90	1.55
	VIF [65]	0.94	0.51	**0.88**	0.41	0.63	**0.88**	0.87	0.60	0.36	0.81	0.24	0.30	0.78	0.80	0.77	0.83	1.22	1.83	1.78	**1.21**	1.18
	NIQE [66]	0.89	0.58	0.45	0.58	0.48	0.58	0.52	0.70	0.43	0.03	0.43	0.50	0.54	0.18	1.04	0.78	2.26	1.63	2.02	2.03	2.07
	IL-NIQE [67]	0.58	0.58	0.52	0.61	0.62	0.48	0.49	0.25	0.44	0.22	0.22	0.31	0.34	0.24	1.86	0.78	2.16	1.58	1.80	2.19	2.12
	OG-IQA [68]	0.67	0.60	0.53	0.57	0.54	0.58	0.51	0.41	**0.57**	0.13	0.19	0.44	-0.21	0.45	1.76	0.77	2.14	1.64	1.94	2.05	2.09
	Proposed	**0.95**	**0.64**	0.74	**0.81**	**0.91**	0.73	**0.94**	**0.94**	0.50	0.67	**0.83**	**0.91**	**0.79**	**0.94**	**0.66**	**0.74**	**1.71**	**1.17**	**0.93**	1.72	**0.84**
All	PSNR	0.35	0.53	0.67	0.12	0.25	0.28	0.66	0.07	0.21	0.61	-0.18	0.03	0.15	0.62	1.80	1.33	1.86	2.17	2.36	2.51	1.89
	PSNR-HVS-M [63]	0.35	0.53	0.79	0.27	0.11	0.31	0.80	0.20	0.54	0.73	-0.17	0.04	0.25	0.75	1.76	1.33	1.55	2.11	2.40	2.43	1.51
	SSIM [19]	0.48	0.44	0.79	0.20	0.15	0.36	0.80	0.44	0.51	0.73	-0.15	0.07	0.27	0.77	1.65	1.41	1.53	2.14	2.39	2.38	1.49
	MS-SSIM [18]	0.48	0.35	0.74	0.00	0.27	0.36	0.75	0.44	0.52	0.69	-0.11	0.09	0.26	0.73	1.66	1.47	1.68	2.19	2.33	2.38	1.67
	IW-SSIM [64]	0.53	0.43	0.80	0.03	0.19	0.21	0.87	0.47	0.52	0.79	-0.12	0.13	0.40	0.83	1.60	1.42	1.50	2.18	2.37	2.50	1.23
	FSIM [52]	0.78	0.41	0.49	0.02	0.36	0.38	0.53	0.63	0.50	0.47	0.01	0.26	0.33	0.56	1.18	1.43	2.18	2.18	2.25	2.36	2.13
	VIF [65]	**0.79**	0.54	**0.91**	**0.74**	0.68	**0.91**	0.91	**0.76**	0.43	**0.89**	0.52	0.50	**0.85**	0.88	**1.15**	1.32	**1.06**	**1.47**	1.76	**1.03**	1.03
	NIQE [66]	0.73	0.20	0.30	0.60	0.27	0.41	0.26	0.54	0.18	0.06	0.18	0.05	0.11	0.03	1.28	1.54	2.40	1.75	2.32	2.33	2.43
	IL-NIQE [67]	0.36	0.41	0.37	0.36	0.36	0.30	0.29	0.12	0.18	0.01	0.07	0.09	0.06	0.04	1.75	1.43	2.34	2.04	2.26	2.43	2.41
	OG-IQA [68]	0.45	0.16	0.45	0.37	0.37	0.27	0.12	0.27	0.23	-0.02	0.23	-0.02	-0.22	0.12	1.68	1.55	2.24	2.03	2.30	2.45	2.50
	Proposed	**0.79**	**0.85**	0.76	0.73	**0.89**	0.69	**0.95**	**0.76**	**0.79**	0.73	**0.58**	**0.88**	0.68	**0.93**	1.16	**0.81**	1.64	1.50	**1.10**	1.85	**0.81**

coefficient (PLCC) for prediction accuracy, Root mean squared error (RMSE) for prediction consistency, Spearman rank-order correlation coefficient (SROCC) for prediction monotonicity. The PLCC is defined as

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{n}(Q_i - \bar{Q})(m_i - \bar{m})}{\sqrt{\sum_{i=1}^{n}(Q_i - \bar{Q})^2(m_i - \bar{m})^2}}, \tag{16}$$

where m_i denotes the subjective score (MOS or DMOS) of the i-th distorted PPCS, $\bar{m}$ and $\bar{Q}$ denote the mean values of m_i and Q_i, respectively. Similarly, the RMSE metric is

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(Q_i - m_i)^2}, \tag{17}$$

and SROCC evaluation is given by

$$\mathrm{SROCC} = 1 - \big(6\sum\nolimits_{i=1}^{N} d_i^2\big)/\big(N(N^2-1)\big), \tag{18}$$

with d_i as the difference between the i-th PPCS's ranks in the objective and subjective evaluations respectively, and N for the total number of samples. The higher the values of the PLCC or SROCC, the better the performance of the model. On the contrary, the lower RMSE value, the better performance of the model.

Table II evaluates the efficiency of our proposed model for the seven different impairments. For comparative studies, we have also included the PSNR, PSNR-HVS-M [62], SSIM [19], MS-SSIM [18], IWSSIM [63], FSIM [51], VIF [64] as full reference examples, and NIQE [65], IL-NIQE [66] and OG-IQA [67] as no reference representatives. Final scores are weighted using the same plane-dependent coefficients κ. Some of recent studies trend to test metrics on different categories separately [11], [36], e.g., "People" and "Inanimate". The reason is that these two categories present obvious spatial morphology difference, e.g., "People" are mostly cylindrical point cloud, while "Inanimate" is more irregular. In order to illustrate the robustness of different objective models, we first test their performance under each category, and then test overall performance. We use red, blue and cyan to highlight the best performance for different content categories (e.g., People, Inanimate and All).

Overall, the proposed model ranks at the first place for 34 times, followed by VIF 24 times, IW-SSIM 7 times, FSIM 2 times, NIQE 2 times, and OG-IQA 1 time. The VIF presents better efficiency for "People" class (e.g., ranked as the first for 10 times), while the number of proposed model is 6. In contrast, the proposed model shows apparent leading performance in "Inanimate" content with 16 times ranked as the first. This is mainly because our model has included the depth information to reflect the geometric distortion that is not considered in conventional metrics. We have also noticed that among all distortions, the best results for CN presents relatively low score, e.g., 0.57 as of SROCC of CN for "Inanimate" samples. We have revisited these contents, and found out that these "Inanimate" contents

TABLE III
PERFORMANCE EVALUATION OF (14) USING DIFFERENT COLOR SPACE AND EDGE DETECTOR. PLCC, SROCC, AND RMSE NUMBERS ARE DIFFERENT FROM THOSE SHOWN IN TABLE II BECAUSE WE HAVE USED THE RANDOMLY SAMPLED SUBSET FOR COMPARATIVE STUDY

Space	Operator	PLCC	SROCC	RMSE
GCM	Canny	0.5992	0.5889	1.8785
	Prewitt	0.6031	0.5996	1.8716
	Sobel	**0.6076**	**0.6020**	**1.8635**
RGB	Canny	0.4367	0.4110	2.1108
	Prewitt	0.4291	0.4442	2.1237
	Sobel	0.4557	0.4401	2.0886
YUV	canny	0.5903	0.5808	1.8940
	Prewitt	0.5921	0.5899	1.8908
	Sobel	0.5970	0.5912	1.8823

are often involved with much higher noises in its native presentation, especially for “Shiva” and “Statue Klimt”. These noisy textures would affect the perceptual sensation, and introduce unreliable rating scores. The actual reason might be related to the content characteristics. “Shiva” and “Statue Klimt” are belong to historical relics, so their surfaces were somewhat corroded over time. Given the fast advances of high-performance acquisition technologies and restoration technologies, we expect that the future point cloud can be high fidelity with high dense and less noise.

VI. ABLATION STUDIES

Additional analysis is provided in the following paragraphs for in-depth understanding of our proposed objective metric in (14).

Color space. Color space conversion is applied to remove the correlation between native RGB samples for efficient feature extraction. In addition to the GCM used in this work, there are other popular conversion methods, such as the RGB to YUV. Here, we have provided the additional experiments to show the model performance when the color space is YUV or the native RGB.

In this comparison, we have randomly sampled the several point cloud contents with all impairments included, and calculated the PLCC, SROCC and RMSE. All the other steps are kept the same following the aforementioned model derivation. Table III has reported the GCM-based color space decomposition offers the best performance.

Edge detector. There are many different edge detectors, such as the *Robert*, *Prewitt* and *Canny* and *Sobel*. Here, we have also shown in Table III that Sobel-based edge detection is the most efficient means to generate depth-edge maps for subsequent aggregation.

Projection planes. In [36], authors use 6, 12, 42 and 162 projection planes to test the performance difference under PSNR, SSIM [19], MS-SSIM [18] and VIF [64]. The results show that there is not much difference when increasing the number of projection planes over six. To demonstrate this conclusion, we obtain more projection image via rotating point cloud (e.g., Fig. 7, rotating point cloud along the center axis which is parallel to the initial view plane). We adopt 15°, 30°, 45°, 60° and 75° as rotation angles. Under each angle, we enforce the cubic projection with six perpendicular planes with rotated point cloud, and we can obtain extra 4 projection images. Note we do not consider top and bottom views because rotating along center axis will not change top and bottom view. In order to facilitate comparison, we set the weight of initial view as 0.5, top and bottom view as 0.05, other view share the rest weights.

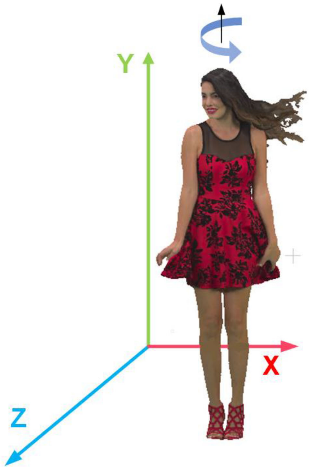

Fig. 7. Point cloud rotation. Via rotating different angles, more projection image an be obtained through perpendicular projection. In this paper, we adopt 15°, 30°, 45°, 60° and 75° as rotation angles and 26 projection images are collected in total. The rotation axis is parallel to the initial view plane, e.g., for “RedandBlack,” the initial view plan is Y-O-X, the direction of the rotation axis (e.g., the black arrow) is parallel to the Y axis which is also the center of the bounding box.

TABLE IV
PERFORMANCE EVALUATION OF USING DIFFERENT PROJECTION PLANS. K REPRESENTS THE NUMBER OF PROJECTION PLANES

Model	PLCC	SROCC	RMSE
$K = 6$	0.6076	0.6020	1.8635
$K = 10$	0.6229	0.6081	1.8355
$K = 14$	0.6291	0.6110	1.8239
$K = 18$	0.6313	0.6137	1.8197
$K = 22$	0.6332	0.6146	1.8161
$K = 26$	0.6340	0.6153	1.8145

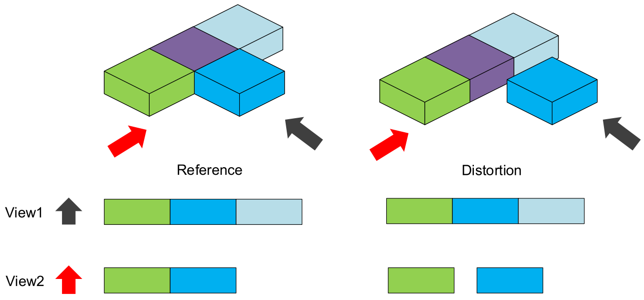

Fig. 8. Projection loss.

Results in Table IV show that with the increase of projection planes, the performance of the proposed model is improved. Specially, the value of PLCC increased from 0.6076 to 0.6340, SROCC from 0.6020 to 0.6153, RMSE from 1.8635 to 1.8145. The reason is that projection can lead to information loss, more

TABLE V
PERFORMANCE EVALUATION OF MODEL COMPONENTS

Model	PLCC	SROCC	RMSE
$F_{\texttt{global}}$	0.2020	0.1590	2.2980
$F_{\texttt{local}}$	0.6047	0.5963	1.8687
Proposed($\alpha_C^{\epsilon C} = 1$)	0.5911	0.5961	1.9574
Proposed(total)	0.6076	0.6020	1.8635

TABLE VI
PERFORMANCE EVALUATION ON MOS AND DMOS

Model	MOS			DMOS		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
PSNR	0.2481	0.2512	2.3354	0.2199	0.3604	2.3711
PSNR-HVS-M	0.2382	0.2615	2.3413	0.2105	0.3610	2.3761
SSIM	0.3654	0.2789	2.2440	0.4483	0.3724	2.1726
MS-SSIM	0.3659	0.2592	2.2437	0.4692	0.3602	2.1465
IW-SSIM	0.4339	0.3285	2.1720	0.4899	0.4033	2.1189
FSIM	0.3196	0.3019	2.2843	0.3886	0.3944	2.2395
VIF	0.5243	0.5647	2.0653	**0.6519**	0.5804	1.8430
NIQE	0.3262	-0.1149	2.2788	0.0461	-0.0369	2.4280
IL-NIQE	0.2703	-0.0478	2.3210	0.2255	0.0331	2.3679
OG-IQA	0.1163	0.0214	2.3943	0.0736	0.0093	2.4240
Proposed	**0.6076**	**0.6020**	**1.8635**	0.6426	**0.6062**	**1.8371**

projection planes can compensate this problem. We use Fig. 8 as example to explain it.

In Fig. 8, different cubes represent different points. Assuming there is a kind of distortion which causes the blue point moves forward along the direction of view1 (e.g., the black arrow). The projected image under view1 cannot perceive this kind of distortion, while view2 (e.g., the red arrow) can record this distortion. This example first shows that it is necessary to capture point cloud image under multiple view, and also illustrates the information loss during projection. Any projection based metrics cannot detect this distortion based on view1. Considering only using certain number of projection planes in practice, some distortions are overlooked which can limit the overall performance of projection based metrics to some extent. However, there are only 4.3%, 2.2% and 2.6% growth rate of PLCC, SROCC and RMSE with using extra 20 projection planes. So, we proposed to use six perpendicular planes to trade off performance against complexity.

Individual feature contribution. There are three important factors in Eq. (14), e.g., global feature $F_{\texttt{global}}$, local feature $F_{\texttt{local}}$ and complex factor α_C. We present the contribution of three components to overall performance in Table V. The results illustrate that the contribution of local feature is the most prominent, while the utilization of global feature and complex factor can improve model robustness.

Difference MOS analysis. In Sec. V-B, we discover that the primary quality difference (such as the noise on "Shiva" and "Statue Klimt") can lead to a relatively poor objective performance. Thus, we convert MOS to difference MOS (DMOS) and re-evaluate objective metrics in Table II. The results are shown in Table VI. The performance of most metrics is improved in terms of DMOS. Note when we test the proposed model in the form of DMOS, we use the same parameters as MOS. According to results in Table VI, the proposed model still ranks first place under most indices, while it presents competitive performance with VIF under PLCC of DMOS.

TABLE VII
PERFORMANCE EVALUATION OF DEPTH MAP

Model	PLCC	SROCC	RMSE
C1	0.2740	0.2390	2.2566
C2	0.6076	0.6020	1.8635

Depth map analysis. From Table VI and Table V we can find, only using local feature can achieve better performance than VIF, e.g., PLCC, SROCC, RMSE of VIF(MOS) is (0.5243, 0.5647, 2.065) while $F_{\texttt{local}}$ is (0.6047, 0.5963, 1.8687). Actually, the calculation of $F_{\texttt{local}}$ is quite simple, and we propose to use depth information when pooling image map to involve spatial distortion, e.g., Eq. (5) and Eq. (7). In this part, we further test the effectiveness of depth map. We list the performance of two different pooling strategies in Table VII, e.g., pooling without depth map (C1), and pooling with depth map (C2).

Table VII presents the extraordinary effectiveness of depth map. Under C1, the PLCC, SROCC and RMSE of the proposed model is (0.2740, 0.2390, 2.2566), which is weaker than most of image quality assessment models in Table VI. After pooling by depth map (e.g., C2), the performance of the proposed model is improved to (0.6076, 0.6020, 1.8635). Regarding the effectiveness of depth map, we suggest the researchers to fully consider the depth map in the study of projection based point cloud quality assessment metrics.

Point-based metrics. As we introduced in Sec. II, there are several point-based metrics have been proposed, including p2point [2], p2plane [69], PSNR-YUV [37], PCQM [40]. We first briefly introduce these point-based metrics and then test them on the proposed database.

- **p2point**. P2point is based on Euclidean distance between a pair of points. First, for each point in the reference point cloud, e.g., a_j, identifying a corresponding point in distorted point cloud via nearest neighbor search, e.g., b_j. Then, compute the error vector $E(i, j)$ by connecting a_j and b_j.
- **p2plane**. P2plane is an extension of p2point. Error vector $\widehat{E}(i, j)$ of p2plane is the projected version of the $E(i, j)$ along the normal direction N_j, e.g., $\widehat{E}(i, j) = E(i, j) \cdot N_j$.
- **PSNR-YUV**. PSNR-YUV is derived by point color difference in YUV space based on same point matching strategy in p2point, e.g., PSNR-YUV$=(6 * \mathrm{PSNR_Y} + 1 * \mathrm{PSNR_U} + 1 * \mathrm{PSNR_V})/8$.
- **PCQM**. PCQA is based on local curvature and lightness. Please refer to [40] for more details.

For p2point and p2plane, we can use two different pooling strategies to obtain point cloud overall quality distortion, e.g., mean square error (MSE) or Hausdorff distance. Besides, in [69] authors proposed a normalization method over p2point and p2plane to realize comparison between multiple point clouds. Specifically, using point cloud bounding box to convert Euclidean distance to PSNR value. The reason is that p2point and p2plane can be affected by point cloud coordinate scale. e.g., $E(i, j)$ and $\widehat{E}(i, j)$ can present significant difference for different point cloud. Therefore, there are 11 point-based metrics are tested on the proposed database, e.g., MSE-p2point,

TABLE VIII
PERFORMANCE EVALUATION OF POINT-BASED METRICS ON PROPOSED DATABASE

Model	PLCC	SROCC	RMSE
MSE-p2point	0.0466	0.7009	2.4081
MSE-p2plane	0.0462	0.6881	2.4081
Hausdorff-p2point	0.6548	0.6189	1.8221
Hausdorff-p2plane	0.6325	0.6233	1.8673
PSNR-MSE-p2point	0.6699	0.7181	1.7898
PSNR-MSE-p2plane	0.6270	0.6669	1.8779
PSNR-Hausdorff-p2point	0.5988	0.5831	1.9307
PSNR-Hausdorff-p2plane	0.6129	0.5983	1.9048
PSNR-YUV	0.6311	0.6207	1.8701
PCQM	**0.8603**	**0.8465**	**1.2291**
Proposed	0.6076	0.6020	1.8635

MSE-p2plane, Hausdorff-p2point, Hausdorff-p2plane, PSNR-MSE-p2point, PSNR-MSE-p2plane, PSNR-Hausdorff-p2point, PSNR-Hausdorff-p2plane, PSNR-YUV, and PCQM.

The results in Table VIII illustrate the performance of above 11 metrics and the proposed model. We can find PCQM present obvious superiority. The reason is that PCQM extract features based on both geometry and color information simultaneously, while p2point and p2plane only consider geometrical distortion, and PSNR-YUV only uses color differences. Besides, the normalization method proposed by [69] can realize great improvement over MSE based p2point and p2plane, e.g., the PLCC, SROCC and RMSE of MSE-p2point is (0.0466, 0.7009, 2.4081) while PSNR-MSE-p2point is (0.6699, 0.7181, 1.7898).

The proposed model outperforms MSE-p2point and MSE-p2plane, and realize competitive performance with other point-based metrics except PCQM. The advantage of the proposed model is low complexity. Specifically, for a point cloud with n points, the projection complexity from a specific viewpoint is $O(n)$. Assuming that, for i-th projected texture image there are m_i projected points, we have $\sum_i^6 m_i \approx n$ or $m_i < n$. The derivations of global, local and texture complexity features need to transverse all points, each of which has the complexity about $O(m)$, respectively. In the end, the total complexity of the proposed model is roughly at $O(n)$. For PCQM, it needs to perform the nearest-neighbor search for each point, and then extract features from quadric fitted surfaces for deriving the quality index. Thus, its overall complexity is about $O(n^2)$. The major complexity contribution in point-based metrics is from its neighbor search to iterative examine close points in proximity. On the contrary, the proposed method can use projection to arrange points without neighbor search, leading to the significant complexity reduction.

Drawbacks. We would like to point out that this work is just a preliminary exploration on the emerging point cloud quality studies. Though our objective metric has demonstrated the state-of-the-art efficiency in comparison to the conventional PSNR, SSIM, and MS-SSIM, there are still many aspects for further investigations. For example, there is much room left to improve the objective metric, e.g., conventional image/video quality metric usually having correlation ranking over 0.95, but our objective metric showing the correlation ranking mostly less than 0.9 (or even less than 0.7). Thus, our database and model are made publicly accessible to the society for further development.

VII. CONCLUSION

We have explored the 3D-to-2D projection-based mechanism to develop the objective point cloud quality metric in this paper. This metric has shown the state-of-the-art efficiency in predicting the subjective point cloud quality at various impairments, in comparison to the conventional popular image-based measurements (e.g., PSNR, SSIM, MS-SSIM).

In this paper, we have provided a fairly large subjective point cloud assessment database (e.g., ten original point cloud content with seven types of impairments at six different distortion levels) with 420 samples and associated subjective MOSs. Additionally, we have offered an closed-form objective metric via the image feature aggregation by weighting global (e.g., JS divergence) and local (e.g., complexity normalized depth-edge aggregated similarity) features extracted from color texture and depth images of all projection planes.

There are many interesting avenues for future exploration. Instead of applying the image-based 2D feature extraction and aggregation, it is natural to have the 3D features extracted directly from the point cloud for quality assessment. A potential method is leveraging the graph theory. On the other hand, saliency has been well and explosively studied for the 2D image/video, and even for the immersive content. Previous studies have shown that our perceptual sensation is also highly correlated with the content saliency. Thus, another interesting topic is developing efficient saliency detection for point cloud and connecting the saliency with the perceptual quality.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their constructive comments to improve this manuscript.

REFERENCES

[1] S. Schwarz *et al.*, "Emerging MPEG standards for point cloud compression," *IEEE J. Emerg. Sel. Top. Circuits Syst.*, vol. 9, no. 1, pp. 133–148, Mar. 2019.

[2] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: Measuring error on simplified surfaces," in *Proc. Comput. Graph. Forum*, vol. 17, no. 2. Wiley Online Library, 1998, pp. 167–174.

[3] R. Mekuria, Z. Li, C. Tulvan, and P. Chou, *Evaluation Criteria for Point Cloud Compression*. ISO/IEC MPEG n16332, Geneva, Switzerland, Feb. 2016.

[4] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, *Evaluation Metrics for Point Cloud Compression*. ISO/IEC JTC m74008, Geneva, Switzerland, Jan. 2017.

[5] S. Perry, A. Pinheiro, E. Dumic, and L. A. da Silva Cruz, "Study of subjective and objective quality evaluation of 3D point cloud data by the JPEG committee," *Electron. Imag.*, vol. 2019, no. 10, pp. 312–1, 2019.

[6] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[7] E. Alexiou, "On subjective and objective quality evaluation of point cloud geometry," in *Proc. 9th Int. Conf. Qual. Multimedia Experience*, 2017, pp. 1–3.

[8] E. Alexiou and T. Ebrahimi, "On the performance of metrics to predict quality in point cloud representations," in *Proc. Appl. Digit. Image Process. XL*, vol. 10396, 2017, p. 103961H.

[9] J. Zhang, W. Huang, X. Zhu, and J.-N. Hwang, "A subjective quality evaluation for 3D point cloud models," in *Proc. Int. Conf. Audio, Lang. Image Process.*, 2014, pp. 827–831.

[10] L. A. da Silva Cruz *et al.*, "Point cloud quality evaluation: Towards a definition for test conditions," in *Proc. 11th Int. Conf. Qual. Multimedia Experience*, 2019, pp. 1–6.

[11] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, “Subjective and objective quality evaluation of compressed point clouds,” in *Proc. 19th Int. Workshop Multimedia Signal Process.*, 2017, pp. 1–6.

[12] “MPEG people datasets,” 2017. [Online]. Available: http://mpegfs.int-evry.fr/MPEG/PCC/DataSets/pointCloud/CfP/decoded/Dynamic_Objects/People/8i/

[13] “MPEG inanimate datasets,” 2017. [Online]. Available: http://mpegfs.int-evry.fr/MPEG/PCC/DataSets/pointCloud/CfP/datasets/Static_Objects_and_Scenes/Inanimate_Objects/

[14] “JPEG pleno database,” http://uspaulopc.di.ubi.pt/

[15] “Microsoft database,” https://jpeg.org/plenodb

[16] Int. Telecommun. Union, Methodology for the Subjective Assessment of the Quality of Television Pictures ITU-R Recommendation BT.500-11, Tech. Rep., 2000.

[17] Z. Ma, M. Xu, Y.-F. Ou, and Y. Wang, “Modeling of rate and perceptual quality of compressed video as functions of frame rate and quantization stepsize and its applications,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 5, pp. 671–682, May 2011.

[18] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, 2003, pp. 1398–1402.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[20] L. Zhang, Y. Shen, and H. Li, “VSI: A visual saliency-induced index for perceptual image quality assessment,” *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[21] K. Gu *et al.*, “Saliency-guided quality assessment of screen content images,” *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

[22] W. Zhou and L. Yu, “Binocular responses for no-reference 3d image quality assessment,” *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1077–1084, Jun. 2016.

[23] L. Xu *et al.*, “Free-energy principle inspired video quality metric and its use in video coding,” *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 590–602, Apr. 2016.

[24] Q. Yang *et al.*, “Modeling the screen content image quality via multiscale edge attention similarity,” *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 310–321, Jun. 2020.

[25] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin, “Quality assessment for video with degradation along salient trajectories,” *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2738–2749, Nov. 2019.

[26] G. Zhai*et al.*, “Cross-dimensional perceptual quality assessment for low bit-rate videos,” *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1316–1324, Nov. 2008.

[27] L. He, W. Zhu, K. Zhang, and Y. Xu, “View-dependent streaming of dynamic point cloud over hybrid networks,” in *Proc. Pacific Rim Conf. Multimedia*. Springer, 2018, pp. 50–58.

[28] S. Schwarz, M. M. Hannuksela, F. S. Vida, and S. P. Nahid, “2D video coding of volumetric video data,” in *Proc. Picture Coding Symp.*, 2018, pp. 61–65.

[29] W. Zhu, Z. Ma, Y. Xu, L. Li, and Z. Li, “View-dependent dynamic point cloud compression,” *IEEE Trans. Circuits Syst. Video Technol.*, p. 1, 2020, doi: 10.1109/TCSVT.2020.2985911.

[30] B. R. Masters, *The New Visual Neurosciences*. John S. Werner and Leo M. Chalupa, Eds., Cambridge, MA, USA: MIT Press, 2014.

[31] J.-M. Geusebroek, R. Van Den Boomgaard, A. W. Smeulders, and A. Dev, “Color and scale: The spatial structure of color images,” in *Eur. Conf. Comput. Vis.*, Springer, 2000, pp. 331–341.

[32] J.-M. Geusebroek, R. Van den Boomgaard, A. W. M. Smeulders, and H. Geerts, “Color invariance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1338–1350, Dec. 2001.

[33] D. Marr, “Vision: A computational investigation into the human representation and processing of visual information,” 1982.

[34] E. Alexious *et al.*, “Point cloud subjective evaluation methodology based on reconstructed surfaces,” in *Proc. Appl. Digit. Image Process. XLI*, vol. 10752, 2018, Paper 107520H.

[35] E. Alexiou and T. Ebrahimi, “Point cloud quality assessment metric based on angular similarity,” in *Proc. Int. Conf. Multimedia Expo*, 2018, pp. 1–6.

[36] E. Alexiou and T. Ebrahimi, “Exploiting user interactivity in quality assessment of point cloud imaging,” in *Proc. Int. Conf. Qual. Multimedia Experience*, 2019, pp. 1–6.

[37] 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 *Proc. Appl. Digit. Image Process. XLI*, vol. 10752, 2018, Paper 107520I.

[38] G. Meynet, J. Digne, and G. Lavoue, “PC-MSDM: A quality metric for 3D point clouds,” in *Proc. 11th Int. Conf. Qual. Multimedia Experience*, 2019, pp. 1–3.

[39] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, “A generalized hausdorff distance based quality metric for point cloud geometry,” in *Proc. 12th Int. Conf. Qual. Multimedia Experience (QoMEX)*, Athlone, Ireland, 2020, pp. 1–6, doi: 10.1109/QoMEX48832.2020.9123087.

[40] G. Meynet, Y. Nehme, and G. Lavoue, “PCQM: A full-reference quality metric for colored 3D point clouds,” in *Proc. 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–6.

[41] I. Viola, S. Subramanyam, and P. Cesar, “A color-based objective quality metric for point cloud contents,” in *Proc. 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–6.

[42] E. Alexiou and T. Ebrahimi, “Towards a point cloud structural similarity metric,” in *Proc. Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–3.

[43] Q. Yang, Z. Ma, Y. Xu, Z. Li, and J. Sun, “Inferring point cloud quality via graph similarity,” 2020, *arXiv:2006.00497*.

[44] “MPEG static object and scenes datasets,” 2017. [Online]. Available: http://mpegfs.int-evry.fr/MPEG/PCC/DataSets/pointCloud/CfP/datasets/Static_Objects_and_Scenes/ULB_Unicorn/

[45] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. C. Bovik, “Study of 3d virtual reality picture quality,” *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 1, pp. 89–102, Jan. 2020.

[46] S. Xie, Y. Xu, Q. Shen, Z. Ma, and W. Zhang, “Modeling the perceptual quality of viewport adaptive omnidirectional video streaming,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3029–3042, Sep. 2020.

[47] “Cloudcompare,” 2019. [Online]. Available: http://www.danielgm.net/cc/

[48] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[49] L. He, W. Zhu, and Y. Xu, “Best-effort projection based attribute compression for 3D point cloud,” in *Proc. 23rd Asia-Pacific Conf. Commun.*, 2017, pp. 1–6.

[50] N. Murray and F. Perronnin, “Generalized max pooling,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2473–2480.

[51] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[52] A. Liu, W. Lin, and M. Narwaria, “Image quality assessment based on gradient similarity,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2011.

[53] M. K. Kundu and S. K. Pal, “Thresholding for edge detection using human psychovisual phenomena,” *Pattern Recognit. Lett.*, vol. 4, no. 6, pp. 433–441, 1986.

[54] F. Yang, Y. Chang, and S. Wan, “Gradient-threshold edge detection based on the human visual system,” *Opt. Eng.*, vol. 44, no. 2, p. 020505, 2005.

[55] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, “Just noticeable difference for images with decomposition model for separating edge and textured regions,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1648–1652, Nov. 2010.

[56] K. Gu *et al.*, “Evaluating quality of screen content images via structural variation analysis,” *IEEE Trans. Visual. Comput. Graph.*, vol. 24, no. 10, pp. 2689–2701, Oct. 2017.

[57] S. Sherman *et al.*, “Solomon kullback, information theory and statistics,” *Bull. Amer. Math. Soc.*, vol. 66, no. 6, pp. 472–472, 1960.

[58] K. Friston, J. Kilner, and L. Harrison, “A free energy principle for the brain,” *J. Physiol.-Paris*, vol. 100, no. 1-3, pp. 70–87, 2006.

[59] K. Friston, “The free-energy principle: A unified brain theory?” *Nat. Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, 2010.

[60] J. R. Hershey and P. A. Olsen, “Approximating the kullback leibler divergence between gaussian mixture models,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, vol. 4, 2007, pp. IV-317–IV-320.

[61] D. Polani, “Kullback-leibler divergence,” *Encyclopedia Syst. Biol.*, pp. 1087–1088, 2013.

[62] N. Ponomarenko *et al.*, “Coefficient contrast masking of DCT basis functions,” in *Proc. 3rd Int. Workshop Video Process. Qual. Metrics*, vol. 4, 2017.

[63] Z. Wang and Q. Li, “Information content weighting for perceptual image quality assessmen,” *Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[64] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” in *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006, doi: 10.1109/TIP.2005.859378.

[65] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a completely blind image quality analyzer,” *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[66] L. Zhang, L. Zhang, and A. C. Bovik, “A feature-enriched completely blind image quality evaluator,” *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[67] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, “Blind image quality assessment by relative gradient statistics and adaboosting neural network,” *Signal Process.-Image Commun.*, vol. 40, pp. 1–15, 2016.

[68] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” 2009. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx

[69] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, “Geometric distortion metrics for point cloud compression,” in *Proc. Int. Conf. Image Process.*, 2017, pp. 3460–3464.

Qi Yang received the B.S. degree in communication engineering from Xidian University, Xi’an, China, in 2017. He is currently working toward the Ph.D. degree in information and communication engineering with Shanghai Jiao Tong University, Shanghai, China, since 2017. His research interests include image processing, 3-D point cloud quality assessment and reconstruction.

Hao Chen received the B.E. degree in electronics and information engineering from Northwestern Polytechnical University, Xi’an, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently on the faculty of Electronic Science and Engineering School, Nanjing University, Nanjing, China. His research focus on real-time video streaming, joint source-channel coding and machine learning. He was the co-recipient of 2019 IEEE Broadcast Technology Society Best Paper Award.

Zhan Ma (Senior Member, IEEE) received the B.S. and M.S. degrees from the 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, NY, USA, in 2011. He is currently on the faculty of Electronic Science and Engineering School, Nanjing University, Nanjing, China. His current research focus on the learning-based video coding, and smart cameras. From 2011 to 2014, he has been with Samsung Research America, Dallas TX, and Futurewei Technologies, Inc., Santa Clara, CA, respectively. His current research interests focus on the next-generation video coding, energy-efficient communication, gigapixel streaming and deep learning. He was the co-recipient of 2018 ACM SIGCOMM Student Research Competition Finalist, 2018 PCM Best Paper Finalist, 2019 IEEE Broadcast Technology Society Best Paper Award, and 2020 IEEE MMSP Image Compression Grand Challenge Best Performing Solution.

Yiling Xu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 1999, 2001, and 2004, respectively. She is a Professor of School of Electronic Information and Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. From 2004 to 2013, she was with Multimedia Communication Research Institute, Samsung Electronics, Inc., Korea. Her main research interests include video processing and transmission, 3-D point cloud compression and network optimization.

Rongjun Tang received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently working toward the M.S. degree in computer and information engineering with the Chinese University of Hong Kong, Hong Kong. His research interests include medical image processing and immersive media quality assessment.

Jun Sun received the B.S. degree from the University of Electronic Sciences and Technology of China, Chengdu, China, in 1989, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 1995, all in electrical engineering. He is currently a Professor and Ph.D. advisor of Shanghai Jiao Tong University. 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 authored or coauthored more than 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.