

MM-PCQA: Multi-Modal Learning for No-reference Point Cloud Quality Assessment

Zicheng Zhang, Wei Sun, Xiongkuo Min, *Member, IEEE*,
Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai, *Senior Member, IEEE*

Abstract—The visual quality of point clouds has been greatly emphasized since the ever-increasing 3D vision applications are expected to provide cost-effective and high-quality experiences for users. Looking back on the development of point cloud quality assessment (PCQA) methods, the visual quality is usually evaluated by utilizing single-modal information, i.e., either extracted from the 2D projections or 3D point cloud. The 2D projections contain rich texture and semantic information but are highly dependent on viewpoints, while the 3D point clouds are more sensitive to geometry distortions and invariant to viewpoints. Therefore, to leverage the advantages of both point cloud and projected image modalities, we propose a novel no-reference point cloud quality assessment (NR-PCQA) metric in a multi-modal fashion. In specific, we split the point clouds into sub-models to represent local geometry distortions such as point shift and down-sampling. Then we render the point clouds into 2D image projections for texture feature extraction. To achieve the goals, the sub-models and projected images are encoded with point-based and image-based neural networks. Finally, symmetric cross-modal attention is employed to fuse multi-modal quality-aware information. Experimental results show that our approach outperforms all compared state-of-the-art methods and is far ahead of previous NR-PCQA methods, which highlights the effectiveness of the proposed method.

Index Terms—point cloud quality assessment, no-reference, multi-modal

I. INTRODUCTION

POINT cloud has been widely adopted in practical applications such as virtual/augmented reality [1], medical 3D reconstruction [2], automatic driving [3], and video post-production [4] due to its ability of representing the 3D world. Consequently, plenty of works have been carried out to deal with point cloud classification [5], [6], [7], [8], [9], [10], [11], detection [3], and segmentation [12], [13]. However, point cloud quality assessment (PCQA) has gained less attention. PCQA aims to accurately predict the visual quality level of point clouds, which is vital for providing useful guidelines for simplification operations and compression algorithms in applications such as metaverse and virtual/augmented reality (VR/AR) [14] to not negatively impact users' quality of experience (QoE). Moreover, the point clouds representing vivid objects/humans are usually more complex in geometric structure and contain large amounts of dense points along

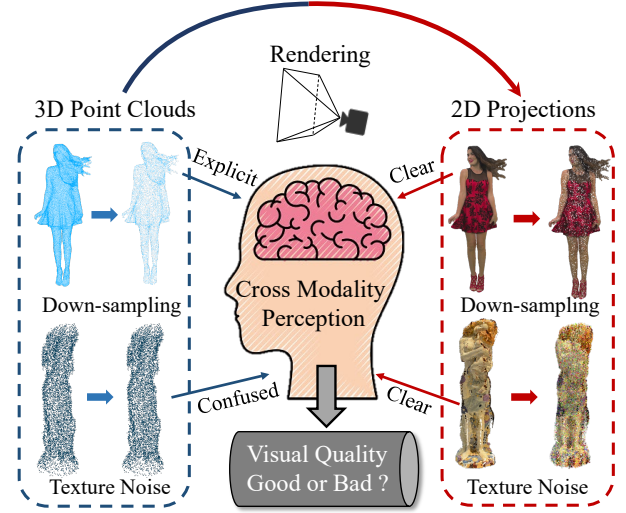


Fig. 1. Examples of reflected distortions. The point clouds can explicitly reveal the geometry down-sampling distortion while failing to recognize texture noise unless the projections are involved, which raises the need for multi-modal perception. The projections are directly rendered from the distorted 3D point clouds via the Open3D [15].

with color attributes, which makes the PCQA problem very challenging.

Generally speaking, PCQA methods can be divided into full-reference PCQA (FR-PCQA), reduced-reference (RR-PCQA), and no-reference PCQA (NR-PCQA) methods according to the involvement extent of the reference point clouds. However, the pristine reference point clouds are not always available in many practical situations, thus NR-PCQA has a wider range of applications. Hence, we focus on the NR-PCQA in this paper. Reviewing the development of NR-PCQA, NR-PCQA metrics are either based on the point cloud features extracted by statistical models [16] and end-to-end neural networks [17] or based on the projected image features obtained via hand-crafted manners [18], [19], [20] or 2D convolutional neural networks (CNN) [21], [22]. Such methods fail to jointly make use of the information from 3D point clouds along with 2D projections, thus resulting in unsatisfactory performance.

Therefore, to boost the performance of PCQA, we propose a multi-modal learning strategy for NR-PCQA, which extracts quality-aware features not only from the 3D point clouds but also from the 2D projections. There are two main reasons to adopt this strategy. First, point clouds can be perceived in both 2D/3D scenarios. We can view point clouds from

Zicheng Zhang, Wei Sun, Xiongkuo Min, Yu Fan, and Guangtao Zhai are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China. E-mail: {zcc1998, sunguwei, minxiongkuo, zhaiguangtao}@sjtu.edu.cn.

Quan Zhou, Jun He, and Qiyuan Wang are with the Bilibili Inc., Shanghai, China. E-mail: {zhouquan, hejun, wangqiyuan}@bilibili.com.

2D perspective via projecting them on the screen or directly watch point clouds in 3D format using the VR equipment. Thus multi-modal learning is able to cover more range of practical situations. Second, different types of distortions have divergent visual influence on different modalities. As shown in Fig. 1, the structural damage and geometry down-sampling are more obvious in the point cloud modality while the image modality is more sensitive to texture distortions caused by color quantization and color noise. Moreover, it is easier to extract quality-aware semantic information from the image modality. Thus, the proposed multi-modal learning fashion can make up for the deficiencies and take advantage of both modalities.

Further, considering that the local patterns such as smoothness and roughness are very important for quality assessment, we first propose to split the point cloud into sub-models rather than sampling points for analysis, which has been previously adopted for extracting the low-level pattern features of the point cloud [23], [24]. The image projections are acquired by rendering the point clouds from random viewpoints with a fixed viewing distance to maintain the consistency of the texture scale. Then a point cloud encoder and an image encoder are used to encode the point clouds and projected images into quality-aware embeddings respectively, which are subsequently strengthened by symmetric cross-modality attention. Finally, the quality-aware embeddings are decoded into quality values with fully-connected layers. The main contributions of this paper are summarized as follows:

- We propose a novel Multi-Modal learning framework for NR-PCQA (MM-PCQA) to interactively use the information from both the point cloud and image modalities. To the best of our knowledge, we are the first to introduce multi-modal learning into the PCQA field.
- To preserve the local patterns that are vital for visual quality, we propose to split the point cloud into sub-models rather than sampling points as the input of the point cloud encoder. To better incorporate the multi-modal features, we employ cross-modal attention to model the mutual relationship between the quality-aware features extracted from two modalities.
- Extensive experiments show that MM-PCQA achieves the best performance among the compared state-of-the-art methods (even including the FR-PCQA methods). The ablation studies reveal the contributions of different modalities, the patch-up strategy, and cross-modal attention, demonstrating the effectiveness of the proposed framework.

II. RELATED WORK

A. Quality Assessment for Point Cloud

In the early years of PCQA development, some simple point-based FR-PCQA methods are proposed by MPEG, such as p2point [25] and p2plane [26]. To further deal with colored point clouds, point-based PSNR-yuv [27] is carried out. Since the point-level difference is difficult to reflect complex distortions, many well-designed FR-PCQA metrics are proposed to employ structural features and have achieved considerable

performance, which includes PCQM [18], GraphSIM [19], PointSSIM [20], etc.

To cover more range of practical applications, some NR-PCQA methods have been proposed as well. Chetouani *et al.* [23] extract patch-wise hand-crafted features and use classical CNN models for quality regression. PQA-net [21] utilizes multi-view projection for feature extraction. Zhang *et al.* [16] use several statistical distributions to estimate quality-aware parameters from the geometry and color attributes' distributions. Fan *et al.* [22] infer the visual quality of point clouds via the captured video sequences. Liu *et al.* [17] employ an end-to-end sparse CNN for quality prediction. Yang *et al.* [28] further transfer the quality information from natural images to help understand the point cloud rendering images' quality via domain adaptation. The mentioned methods are all based on single-modal information, thus failing to jointly incorporate the multi-modal quality information.

B. Multi-modal Learning for Point Cloud

Many works [29], [30], [31] have proven that multi-modal learning is capable of strengthening feature representation by actively relating the compositional information across different modalities such as image, text, and audio. Afterwards, various works utilize both point clouds and image data to improve the understanding of 3D vision, which are mostly targeted at 3D detection. Namely, AVOD [6] and MV3D [11] make region proposals by mapping the LiDAR point clouds to bird's eye view. Then Qi *et al.* [32] and Wang *et al.* [7] attempt to localize the points by proposing 2D regions with 2D CNN object detector and then transforming the corresponding 2D pixels to 3D space. Pointpainting [8] projects the points into the image segmentation mask and detects 3D objects via LiDAR-based detector. Similarly, PI-RCNN [9] employs semantic feature maps and makes predictions through point-wise fusion. 3D-CVF [10] transfers the camera-view features using auto-calibrated projection and fuses the multi-modal features with gated fusion networks. More recently, some works [33], [34] try to make use of multi-modal information in a self-supervised manner and transfer the learned representation to the downstream tasks.

III. PROPOSED METHOD

The framework overview is clearly exhibited in Fig. 2. The point clouds are first segmented into sub-models and put into the point cloud encoder θ_P . The projected images are directly rendered from the colored point clouds and put into the image encoder θ_I . Subsequently, the quality-aware encoder features are optimized with the assistance of symmetric cross-modality attention. Finally, the features are concatenated and decoded into the quality values via the quality regression.

A. Preliminaries

Suppose we have a colored point cloud $\mathbf{P} = \{g_{(i)}, c_{(i)}\}_{i=1}^N$, where $g_{(i)} \in \mathbb{R}^{1 \times 3}$ indicates the geometry coordinates, $c_{(i)} \in \mathbb{R}^{1 \times 3}$ represents the attached RGB color information, and N stands for the number of points. The point cloud modality $\hat{\mathbf{P}}$

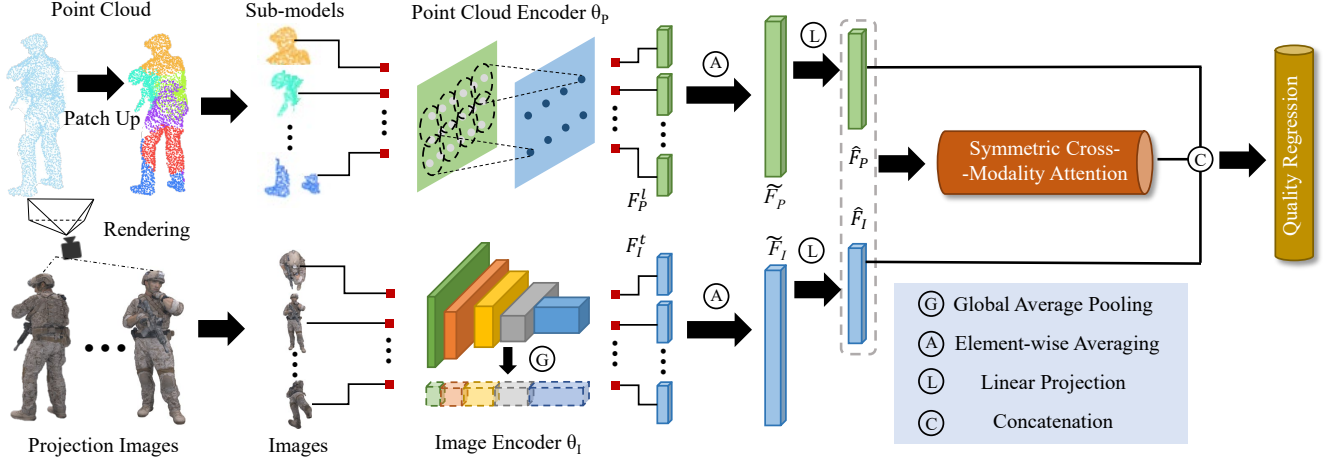


Fig. 2. The framework of the proposed method.

is obtained by normalizing the original geometry coordinates and the image modality \mathbf{I} is generated by rendering the colored point cloud \mathbf{P} into 2D projections. Note that $\hat{\mathbf{P}}$ contains no color information.

B. Point Cloud Feature Extraction

It is natural to transfer mainstream 3D object detectors such as PointNet++ [35] and DGCNN [36] to visual quality representation of point clouds. However, different from the point clouds used for classification and segmentation, the high-quality point clouds usually contain millions of dense points, which makes it difficult to extract features directly from the source points unless utilizing down-sampling. However, the common sampling methods are aimed at preserving semantic information whereas inevitably damaging the geometry patterns that are crucial for quality assessment. To avoid the geometry error caused by the down-sampling and preserve the smoothness and roughness of local patterns, we propose to split the point cloud into several local sub-models for geometric structure feature representation. Specifically, given a normalized point cloud $\hat{\mathbf{P}}$, we employ the farthest point sampling (FPS) to obtain N_δ anchor points $\{\delta_m\}_{m=1}^{N_\delta}$. For each anchor point, we utilize K nearest neighbor (KNN) algorithm to find N_s neighboring points around the anchor point and form such points into a sub-model:

$$S = \{\mathbf{KNN}(\delta_m)\}_{m=1}^{N_\delta}, \quad (1)$$

where S is the set of sub-models, $\mathbf{KNN}(\cdot)$ denotes the KNN operation and δ_m indicates the m -th farthest sampling point. We use $N_\delta \times N_s > N$ to ensure that the sub-models cover sufficient number of points in the source point cloud. An illustration of the patch-up process is exhibited in Fig. 3. It can be seen that the local sub-models are capable to preserve the local patterns. We randomly select N_P sub-models from the generated N_δ sub-models for geometric feature extraction. Then a point cloud feature encoder $\theta_P(\cdot)$ is employed to map the selected sub-models to quality-aware embedding space:

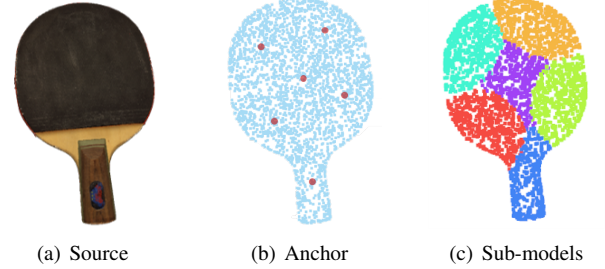


Fig. 3. An example of the patch-up process. (a) represents the source point cloud, (b) describes the anchor points after the farthest point sampling, and (c) shows the patch-up results after KNN operation. Color information is not included in the sub-models.

$$F_P = \{\theta_P(S_l)\}_{l=1}^{N_P},$$

$$\tilde{F}_P = \frac{1}{N_P} \sum_{l=1}^{N_P} F_P^l, \quad (2)$$

where $F_P^l \in \mathbb{R}^{1 \times C_P}$ indicates the quality-aware embedding for the l -th selected sub-model S_l , C_P represents the number of output channels of the point cloud encoder $\theta_P(\cdot)$, and $\tilde{F}_P \in \mathbb{R}^{1 \times C_P}$ is the pooled results after average fusion.

C. Image Feature Extraction

Inspired by the success of PCQA methods via utilizing the image modality [21], [37], we propose to use a 2D image encoder for feature extraction. N_I image projections are rendered from the distorted colored point clouds from random viewpoints but with a fixed viewing distance to keep the texture consistent. Considering that the perceived texture visual quality varies at different scales [38], [39], we propose to use hierarchical 2D CNN backbones to make use of multi-scale quality-aware features to better model the human perception of texture quality. Assume that the 2D CNN backbone has j layers, the hierarchical features can be described as:

$$\theta_I(x) = \alpha_1(x) \oplus \alpha_2(x) \cdots \oplus \alpha_j(x),$$

$$\alpha_k(x) = \text{GAP}(L_k(x)), k \in \{1, \dots, j\}, \quad (3)$$

where $\theta_I(\cdot)$ represents the hierarchical image encoder, $\oplus(\cdot)$ indicates the concatenation operation, $\text{GAP}(\cdot)$ represents the global average pooling operation, $L_k(x)$ stands for the feature maps obtained from k -th layer, and $\alpha_k(x)$ denotes the average pooled features from $L_k(x)$. Then we embed the rendered 2D images into quality-aware space with the hierarchical 2D image encoder:

$$\begin{aligned} F_I &= \{\theta_I(I_t)\}_{t=1}^{N_I}, \\ \tilde{F}_I &= \frac{1}{N_I} \sum_{t=1}^{N_I} F_I^t, \end{aligned} \quad (4)$$

where $F_I^t \in \mathbb{R}^{1 \times C_I}$ denotes the quality-aware embedding for the t -th image projection I_t , C_I represents the number of output channels of the 2D image encoder $\theta_I(\cdot)$, and $\tilde{F}_I \in \mathbb{R}^{1 \times C_I}$ is the pooled results after average fusion.

D. Symmetric Cross-Modality Attention

As shown in Fig 1, the single modality may be incomplete to cover sufficient information for quality assessment, thus we propose a symmetric attention transformer block to investigate the interaction between the visual quality features gathered from the point cloud and image modalities. Given the intra-modal features $\tilde{F}_P \in \mathbb{R}^{1 \times C_P}$ and $\tilde{F}_I \in \mathbb{R}^{1 \times C_I}$ from the point clouds and images respectively, we adjust them to the same dimension with linear projection:

$$\hat{F}_P = W_P \tilde{F}_P, \quad \hat{F}_I = W_I \tilde{F}_I, \quad (5)$$

where $\hat{F}_P \in \mathbb{R}^{1 \times C'}$ and $\hat{F}_I \in \mathbb{R}^{1 \times C'}$ are adjusted features, W_P and W_I are learnable linear mappings, and C' is the number of adjusted channels. To further explore the discriminate components among the modalities, the multi-head attention module is utilized:

$$\begin{aligned} \Gamma(Q, K, V) &= (h_1 \oplus h_2 \cdots \oplus h_n)W, \\ h_\mu &= \beta(QW_\mu^Q, KW_\mu^K, VW_\mu^V)|_{\mu=1}^n, \\ \beta(Q, K, V) &= \text{softmax}(QK^T/\sqrt{d})V, \end{aligned} \quad (6)$$

where $\Gamma(\cdot)$ indicates the multi-head attention operation, $\beta(\cdot)$ represents the attention function, h_μ is the μ -th head, and W , W_Q , W_K , W_V are learnable linear mappings. As illustrated in Fig. 4, both point-cloud-based and image-based quality-aware features participate in the attention learning of the other modality. The final quality embedding can be concatenated by the intra-modal features and the guided multi-modal features obtained by the symmetric cross-modality attention module:

$$\hat{F}_Q = \hat{F}_P \oplus \hat{F}_I \oplus \Psi(\hat{F}_P, \hat{F}_I), \quad (7)$$

where $\oplus(\cdot)$ indicates the concatenation operation, $\Psi(\cdot)$ stands for the symmetric cross-modality attention operation, and \hat{F}_Q represents the final quality-aware features ($\Psi(\hat{F}_P, \hat{F}_I) \in \mathbb{R}^{1 \times 2C'}$, $\hat{F}_Q \in \mathbb{R}^{1 \times 4C'}$).

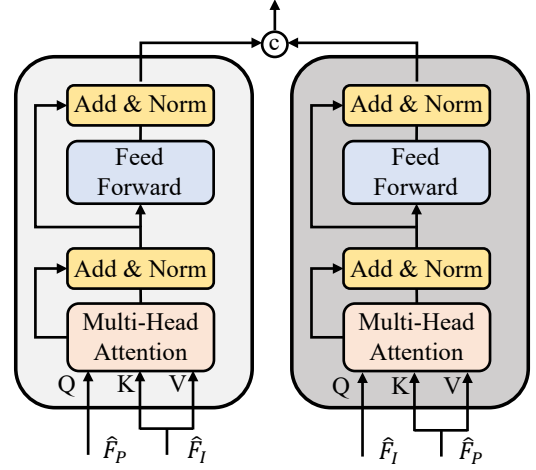


Fig. 4. Illustration of the symmetric cross-modality attention module (SCMA). The point cloud embedding \hat{F}_P is used to guide the attention learning of image embedding and the image embedding \hat{F}_I is used to guide the attention learning of point cloud embedding respectively.

E. Quality Regression & Loss Function

Following common practice, we simply use two-fold fully-connected layers to regress the quality features \hat{F}_Q into predicted quality scores. For the quality assessment tasks, we not only focus on the accuracy of the predicted quality values but also lay importance on the quality rankings [40], [41]. Therefore, the loss function employed in this paper includes two parts: Mean Squared Error (MSE) and rank error. The MSE is utilized to keep the predicted values close to the quality labels, which can be derived as:

$$L_{MSE} = \frac{1}{n} \sum_{\eta=1}^n (q_\eta - q'_\eta)^2, \quad (8)$$

where q_η is the predicted quality scores, q'_η is the quality labels of the point cloud, and n is the size of the mini-batch. The rank loss can better assist the model to distinguish the quality difference when the point clouds have close quality labels. To this end, we use the differentiable rank function described in [42], [43] to approximate the rank loss:

$$\begin{aligned} L_{rank}^{ij} &= \max(0, |q_i - q_j| - e(q_i, q_j) \cdot (q'_i - q'_j)), \\ e(q_i, q_j) &= \begin{cases} 1, & q_i \geq q_j, \\ -1, & q_i < q_j, \end{cases} \end{aligned} \quad (9)$$

where i and j are the corresponding indexes for two point clouds in a mini-batch and the rank loss can be derived as:

$$L_{rank} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n L_{rank}^{ij}, \quad (10)$$

Then the loss function can be calculated as the weighted sum of MSE loss and rank loss:

$$Loss = \lambda_1 L_{MSE} + \lambda_2 L_{rank} \quad (11)$$

where λ_1 and λ_2 are used to control the proportion of the MSE loss and the rank loss.

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART APPROACHES ON THE SJTU-PCQA AND WPC DATABASES. ‘P’ AND ‘I’ STAND FOR THE POINT CLOUD AND IMAGE MODALITIES RESPECTIVELY. BEST IN **red** AND SECOND IN **blue**.

Type	Methods	Modal	SJTU-PCQA				WPC			
			SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
FR	MSE-p2po	P	0.72	0.81	0.56	1.36	0.45	0.48	0.31	19.89
	HD-p2po	P	0.71	0.77	0.54	1.44	0.27	0.39	0.19	20.89
	MSE-p2pl	P	0.62	0.59	0.48	2.28	0.32	0.26	0.22	22.82
	HD-p2pl	P	0.64	0.68	0.45	2.12	0.28	0.27	0.16	21.98
	PSNR-yuv	P	0.79	0.81	0.61	1.31	0.44	0.53	0.31	19.31
	PCQM	P	0.86	0.88	0.70	1.08	0.74	0.74	0.56	15.16
	GraphSIM	P	0.87	0.89	0.69	1.03	0.58	0.61	0.41	17.19
	PointSSIM	P	0.71	0.74	0.54	1.54	0.46	0.47	0.33	20.13
RR	PCMRR	P	0.47	0.68	0.33	1.72	0.31	0.33	0.22	21.42
NR	BRISQUE	I	0.39	0.42	0.29	2.09	0.27	0.30	0.19	21.86
	NIQE	I	0.13	0.24	0.10	2.26	0.11	0.22	0.09	23.14
	IL-NIQE	I	0.08	0.16	0.05	2.33	0.09	0.14	0.08	24.01
	ResSCNN	P	0.81	0.82	0.61	1.11	0.55	0.56	0.44	17.12
	PQA-net	I	0.83	0.85	0.63	1.07	0.70	0.71	0.49	15.02
	3D-NSS	P	0.72	0.74	0.52	1.66	0.64	0.64	0.45	16.47
	MM-PCQA(Ours)	P+I	0.91	0.92	0.78	0.77	0.83	0.83	0.64	12.84

IV. EXPERIMENTS

A. Databases

To test the performance of the proposed method, we employ the subjective point cloud assessment database (SJTU-PCQA) [37] and the Waterloo point cloud assessment database (WPC) proposed by [44] for validation. The SJTU-PCQA database includes 9 reference point clouds and each point cloud is corrupted with seven types of distortions (compression, color noise, geometric shift, down-sampling, and three distortion combinations) under six strengths, which generates $378 = 9 \times 7 \times 6$ distorted point clouds in total. The WPC database contains 20 reference point clouds and augmented each point cloud with four types of distortions (down-sampling, Gaussian white noise, Geometry-based Point Cloud Compression (G-PCC), and Video-based Point Cloud Compression (V-PCC)), which generates $740 = 20 \times 37$ distorted point clouds.

B. Implementation Details

The Adam optimizer [45] is utilized with weight decay $1e-4$ and the initial learning rate is set as $5e-5$. The model is trained for 50 epochs by default. We set the point cloud sub-model size N_s as 2048 and set $N_\delta = N/N_s + 1$. The projected images with the resolution of $1920 \times 1080 \times 3$ are randomly cropped into image patches at the resolution of $224 \times 224 \times 3$ as the inputs. During the experiment, 6 sub-models and 4 projected images are selected at random for each point cloud. The PointNet++ [35] is utilized as the point cloud encoder and the hierarchical ResNet50 [46] is used as the image encoder, where the ResNet50 is initialized with the pre-trained model on the ImageNet database [47]. The multi-head attention module employs 8 heads and the feed-forward dimension is set as 2048. The weights λ_1 and λ_2 for L_{MSE} and L_{rank} are both set as 1.

Following the practices in [23], [22], the k-fold cross validation strategy is employed for the experiment to accurately estimate the performance of the proposed method. Specifically, for the SJTU-PCQA database, 1 group of point cloud is chosen as the testing set and the remaining 8 groups of point cloud

are taken as the training set. Such procedure is repeated 9 times with each of the point cloud groups used exactly once as the testing set. Similarly, we randomly split the WPC database into 5 folds and each fold consists of 4 groups of point clouds. Then a similar experiment procedure as discussed above is applied. In all, 9-fold cross validation is conducted on the SJTU-PCQA database and 5-fold cross validation is conducted on the WPC database. The average performance is recorded as the final results. It’s worth noting that there is no content overlap between the training and testing sets. We strictly retrain the available baselines with the same database split set up to keep the comparison fair. What’s more, for the FR-PCQA and RR-PCQA methods that require no training, we simply validate them on the same testing sets and record the average performance.

C. Competitors and Evaluation Criteria

15 state-of-the-art quality assessment methods are selected for comparison, which consist of 8 FR-PCQA methods, 1 RR-PCQA method, and 6 NR-PCQA methods. The FR-PCQA methods include MSE-p2point (MSE-p2po) [25], Hausdorff-p2point (HD-p2po) [25], MSE-p2plane (MSE-p2pl) [26], Hausdorff-p2plane (HD-p2pl) [26], PSNR-yuv [27], PCQM [18], GraphSIM [19], and PointSSIM [20]. The RR-PCQA method includes PCMRR [48]. The NR-PCQA methods include BRISQUE [49], NIQE [50], IL-NIQE [51], ResSCNN [17], PQA-net [21], and 3D-NSS [16]. Note that BRISQUE, NIQE, IL-NIQE are image-based quality assessment metrics and are validated on the same projected images. Furthermore, to deal with the scale differences between the predicted quality scores and the quality labels, a five-parameter logistic function is applied to map the predicted scores to subjective ratings, as suggested by [40], [52].

Four mainstream evaluation criteria in the quality assessment field are utilized to compare the correlation between the predicted scores and MOSs, which include Spearman Rank Correlation Coefficient (SRCC), Kendall’s Rank Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient

TABLE II

CONTRIBUTIONS OF THE MODALITIES, WHERE ‘P’ STANDS FOR ONLY USING POINT CLOUD FEATURES, ‘I’ STANDS FOR USING ONLY IMAGE FEATURES, ‘P+I’ REPRESENTS USING BOTH MODALITIES’ FEATURES THROUGH SIMPLE CONCATENATION, AND ‘SCMA’ INDICATES THE SYMMETRIC CROSS-MODALITY ATTENTION IS USED. BEST IN BOLD.

Modal	SJTU-PCQA		WPC	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
P	0.84	0.89	0.52	0.55
I	0.88	0.89	0.79	0.80
P+I	0.91	0.90	0.81	0.82
P+I+SCMA	0.91	0.92	0.83	0.83

(PLCC), Root Mean Squared Error (RMSE). An excellent quality assessment model should obtain values of SRCC, KRCC, PLCC close to 1 and RMSE to 0.

1) *Performance Discussion*: The experimental results are listed in Table I, from which we can make several useful inspections: a) Our method MM-PCQA presents the best performance among the two databases and outperforms the compared NR-PCQA methods by a large margin. For example, MM-PCQA surpasses the second place NR-PCQA method PQA-net by 0.08 (0.91 vs. 0.83) on the SJTU-PCQA database and by 0.13 (0.83 vs. 0.70) on the WPC database in terms of SRCC. By jointly utilizing intra-modal and cross-modal features, MM-PCQA enforces the model to relate the compositional quality-aware patterns and mobilize under-complementary information between the image and point cloud modalities to optimize the quality representation, thus gaining higher performance; b) There is a significant performance drop from the SJTU-PCQA database to the WPC database since the WPC database contains more complex distortions, being more challenging for PCQA models. MM-PCQA achieves a relatively smaller performance drop than most compared methods. Particularly, the SRCC and PLCC values of MM-PCQA drop by 0.08 and 0.09 respectively. However, all compared PCQA methods except 3D-NSS experience a larger performance drop over 0.1 on both SRCC and PLCC values. Therefore, we can conclude that MM-PCQA gains more robustness over more complex distortions; c) The image-based handcrafted NR methods BRISQUE, NIQE, and IL-NIQE obtain the poorest performance. This is because such methods are carried out for evaluating the natural scene images, which have a huge gap from the projected images rendered from the point clouds. In all, the overall experimental performance firmly validates our motivation that multi-modal learning can help the model better understand the visual quality of point clouds.

D. Contributions of the Modalities

As described above, MM-PCQA jointly employs the features from the point cloud and image modalities and we hold the hypothesis that multi-modal learning helps the model to gain better quality representation than single-modality. Therefore, in this section, we conduct ablation studies to validate our hypothesis. The performance results are presented in Table II. The performance of both single-modal-based model is inferior to the multi-modal-based model, which suggests that both

TABLE III

CONTRIBUTIONS OF THE PATCH-UP STRATEGY, WHERE ‘P’ STANDS FOR ONLY USING POINT CLOUD FEATURES, ‘P+I’ REPRESENTS USING BOTH POINT CLOUD AND IMAGE FEATURES, ‘FPS’ INDICATES USING FARTHEST POINT SAMPLING STRATEGY, AND ‘PATCH-UP’ INDICATES USING PATCH-UP STRATEGY. BEST IN BOLD.

Model	SJTU-PCQA		WPC	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
P+FPS	0.33	0.34	0.12	0.15
P+Patch-up	0.84	0.89	0.52	0.55
P+I+FPS	0.89	0.89	0.78	0.79
P+I+Patch-up	0.91	0.92	0.83	0.83

TABLE IV

CROSS-DATABASE EVALUATION, WHERE SJTU→WPC INDICATES THE MODEL IS TRAINED ON THE WHOLE SJTU-PCQA DATABASE AND VALIDATED WITH THE DEFAULT TESTING SETUP OF THE WPC DATABASE, AND VICE VERSA. BEST IN BOLD.

Model	SJTU→WPC		WPC→SJTU	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
PQA-net	0.17	0.18	0.54	0.61
3D-NSS	0.15	0.14	0.18	0.23
MM-PCQA	0.20	0.30	0.72	0.79

point cloud and image features make contributions to the final results. After using SCMA module, the performance is further boosted, which validates the effectiveness of cross-modality attention.

E. Ablation for the Patch-up Strategy

To prove that the patch-up strategy is more suitable for PCQA, we present the performance of using farthest point sampling (FPS) strategy as well. To make the comparison even, $12,288 = 6 \times 2048$ points (the same number of points as contained in the default 6 sub-models) are sampled for each point cloud and the results are shown in Table III. It can be seen that the patch-up strategy greatly improves the performance when only point cloud features are used. The reason is that the sampled points are not able to preserve the local patterns that are vital for quality assessment. Moreover, when the point cloud contains more points (even up to millions), the sampled points may not even maintain the main shape of the object unless greatly increasing the number of sampling points.

F. Cross-database Evaluation

The cross-database evaluation is further conducted to test the generalization ability of the proposed MM-PCQA and the experimental results are exhibited in Table IV. From the table, we can find that the proposed MM-PCQA better generalizes learned feature representation from the WPC database and achieves much better performance than the most competitive NR-PCQA methods when validating on the SJTU-PCQA database. Although all the NR-PCQA models trained on the SJTU-PCQA database obtain poor performance, MM-PCQA still yields the best performance.

TABLE V

PERFORMANCE RESULTS OF THE POINT CLOUD BRANCH BY VARYING THE NUMBER OF SUB-MODELS ON THE SJTU-PCQA AND WPC DATABASES. BEST IN BOLD.

Number of 3D sub-models	SJTU-PCQA		WPC	
	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow
2	0.74	0.78	0.46	0.52
4	0.73	0.82	0.45	0.49
6	0.84	0.89	0.52	0.55
8	0.82	0.89	0.50	0.54

TABLE VI

PERFORMANCE RESULTS OF THE IMAGE BRANCH BY VARYING THE NUMBER OF 2D PROJECTIONS ON THE SJTU-PCQA AND WPC DATABASES. BEST IN BOLD.

Number of 2D projections	SJTU-PCQA		WPC	
	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow
2	0.87	0.87	0.75	0.76
4	0.88	0.89	0.79	0.80
6	0.88	0.88	0.79	0.78
8	0.86	0.87	0.78	0.79

G. Number of 2D Projections and 3D sub-models

We try to investigate the contributions of the point cloud and image branch by varying the number of the 3D sub-models and input 2D projections. The performance results are exhibited in Table V and Table VI. We can see that employing 6 sub-models and 4 projections yields the best performance for the point cloud and image branch respectively. With the number increasing, the features extracted from sub-models and projections may get redundant, thus resulting in the performance drop.

H. Limitation

From the above analysis, the major limitation of the proposed model lies in the cross-database (SJTU \rightarrow WPC) evaluation, mainly due to the following reasons: a) The number of training samples for the SJTU-PCQA database is half of the WPC database (378 \rightarrow 740), which means the SJTU-PCQA database contains less diverse content and distortion types. b) The distortion levels of the SJTU-PCQA database are relatively coarse-grained. Thus the quality representation learned from the SJTU-PCQA database is not effective on the WPC database which contains stimuli with fine-grained distortion levels. In all, due to the limited scale of current PCQA databases, the generalization performance of MM-PCQA is still unsatisfactory.

V. CONCLUSION

This paper proposes a novel multi-modal learning approach for no-reference point cloud quality assessment (MM-PCQA). MM-PCQA aims to acquire quality information across modalities and optimize the quality representation. In particular, the point clouds are patched up to preserve the important local geometric structure patterns. Then the projected images are employed to reflect the texture distortions. PointNet++ and hierarchical ResNet50 are utilized as the feature encoders. Symmetric cross-modal attention is further employed to make

full use of the cross-modal information. Experimental results show that MM-PCQA reaches a new state-of-the-art on the SJTU-PCQA and WPC database. Extensive ablation studies further demonstrate the potency of the proposed multi-modal learning framework.

REFERENCES

- [1] Y. Park, V. Lepetit, and W. Woo, "Multiple 3d object tracking for augmented reality," in *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2008, pp. 117–120.
- [2] Q. Cheng, P. Sun, C. Yang, Y. Yang, and P. X. Liu, "A morphing-based 3d point cloud reconstruction framework for medical image processing," *Computer methods and programs in biomedicine*, vol. 193, p. 105495, 2020.
- [3] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.
- [4] R. Mekuria, K. Blom, and P. Cesar, "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828–842, 2016.
- [5] E. Grilli, F. Menna, and F. Remondino, "A review of point clouds segmentation and classification algorithms," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, p. 339, 2017.
- [6] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–8.
- [7] Z. Wang and K. Jia, "Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1742–1749.
- [8] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.
- [9] L. Xie, C. Xiang, Z. Yu, G. Xu, Z. Yang, D. Cai, and X. He, "Pi-rnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12460–12467.
- [10] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection," in *European Conference on Computer Vision*, 2020, pp. 720–736.
- [11] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots," in *European conference on computer vision*, 2020, pp. 68–84.
- [12] M. Cheng, L. Hui, J. Xie, and J. Yang, "Sspc-net: Semi-supervised semantic 3d point cloud segmentation network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1140–1147.
- [13] M. Xu, Z. Zhou, and Y. Qiao, "Geometry sharing network for 3d point cloud classification and segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12500–12507.
- [14] J. D. N. Dionisio, W. G. B. III, and R. Gilbert, "3d virtual worlds and the metaverse: Current status and future possibilities," *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, pp. 1–38, 2013.
- [15] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.
- [16] Z. Zhang, W. Sun, X. Min, T. Wang, W. Lu, and G. Zhai, "No-reference quality assessment for 3d colored point cloud and mesh models," *arXiv preprint arXiv:2107.02041*, 2021.
- [17] Y. Liu, Q. Yang, Y. Xu, and L. Yang, "Point cloud quality assessment: Dataset construction and learning-based no-reference metric," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022.
- [18] G. Meynet, Y. Nehmé, J. Digne, and G. Lavoué, "Pcqm: A full-reference quality metric for colored 3d point clouds," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

- [19] Q. Yang, Z. Ma, Y. Xu, Z. Li, and J. Sun, "Inferring point cloud quality via graph similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [20] E. Alexiou and T. Ebrahimi, "Towards a point cloud structural similarity metric," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020, pp. 1–6.
- [21] Q. Liu, H. Yuan, H. Su, H. Liu, Y. Wang, H. Yang, and J. Hou, "Pqa-net: Deep no reference point cloud quality assessment via multi-view projection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4645–4660, 2021.
- [22] Y. Fan, Z. Zhang, W. Sun, X. Min, W. Lu, T. Wang, N. Liu, and G. Zhai, "A no-reference quality assessment metric for point cloud based on captured video sequences," *arXiv preprint arXiv:2206.05054*, 2022.
- [23] A. Chetouani, M. Quach, G. Valenzise, and F. Dufaux, "Deep learning-based quality assessment of 3d point clouds without reference," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2021, pp. 1–6.
- [24] K. You and P. Gao, "Patch-based deep autoencoder for point cloud geometry compression," in *ACM Multimedia Asia*, 2021, pp. 1–7.
- [25] R. Mekuria, Z. Li, C. Tulvan, and P. Chou, "Evaluation criteria for point cloud compression," *ISO/IEC MPEG*, no. 16332, 2016.
- [26] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, "Geometric distortion metrics for point cloud compression," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3460–3464.
- [27] E. M. Torlig, E. Alexiou, T. A. Fonseca, R. L. de Queiroz, and T. Ebrahimi, "A novel methodology for quality assessment of voxelized point clouds," in *Applications of Digital Image Processing XLI*, vol. 10752, 2018, pp. 174–190.
- [28] Q. Yang, Y. Liu, S. Chen, Y. Xu, and J. Sun, "No-reference point cloud quality assessment via domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 179–21 188.
- [29] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.
- [30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.
- [31] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3884–3892.
- [32] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.
- [33] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo, "Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9902–9912.
- [34] Y. Zhang, J. Chen, and D. Huang, "Cat-det: Contrastively augmented transformer for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 908–917.
- [35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.
- [36] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.
- [37] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun, "Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration," *IEEE Transactions on Multimedia*, 2020.
- [38] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, vol. 2, 2003, pp. 1398–1402.
- [39] W. Sun, T. Wang, X. Min, F. Yi, and G. Zhai, "Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2021, pp. 1–6.
- [40] Y. Fang, L. Huang, J. Yan, X. Liu, and Y. Liu, "Perceptual quality assessment of omnidirectional images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 580–588.
- [41] J. Gu, G. Meng, C. Da, S. Xiang, and C. Pan, "No-reference image quality assessment with reinforcement recursive list-wise ranking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8336–8343.
- [42] S. Wen and J. Wang, "A strong baseline for image and video quality assessment," *arXiv preprint arXiv:2111.07104*, 2021.
- [43] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," *arXiv preprint arXiv:2204.14047*, 2022.
- [44] Q. Liu, H. Su, Z. Duanmu, W. Liu, and Z. Wang, "Perceptual quality assessment of colored 3d point clouds," *IEEE Transactions on Visualization and Computer Graphics*, 2022.
- [45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.
- [48] I. Viola and P. Cesar, "A reduced reference metric for visual quality evaluation of point cloud contents," *IEEE Signal Processing Letters*, vol. 27, pp. 1660–1664, 2020.
- [49] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [50] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.
- [51] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.
- [52] J. Antkowiak, T. Jamal Baina, F. V. Baroncini, N. Chateau, F. FranceT-elecom, A. C. F. Pessoa, F. Stephanie Colonnese, I. L. Contin, J. Caviedes, and F. Philips, "Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000," 2000.