VIP-PCQA: A Multi-Modal Framework for No-reference Point Cloud Quality Assessment

Kang Fu, Zicheng Zhang, Huiyu Duan*, Xiaohong Liu, Xiongkuo Min*, Jia Wang, and Guangtao Zhai*
Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University

***Abstract*—Point clouds often suffer from geometric and color noise, as well as compression artifacts, during their production, storage, and transmission. Therefore, accurately and automatically evaluating the quality of point clouds is crucial for optimizing storage and compression strategies. This paper introduces the VIP-PCQA, a novel framework that combines Video, Image, and Point cloud modalities for no-reference Point Cloud Quality Assessment. The framework begins by rendering projection videos and normal images from point clouds, followed by sampling patches and computing statistical features related to color and geometry. Subsequently, a video encoder, two image encoders, and a point cloud encoder are employed to extract modality-specific features. Finally, these features are fused to regress the quality score. Experimental results on three publicly available benchmark databases demonstrate that VIP-PCQA achieves outstanding performance with excellent generalization capabilities. An ablation study further highlights the indispensable contribution of each modality to the framework's success. The code is released on https://github.com/ZedFu/VIP-PCQA.**

***Index Terms*—3D model, multi-modal, point cloud, quality assessment**

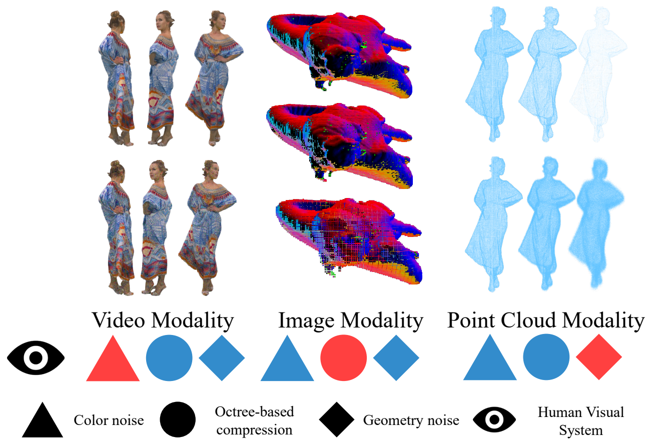

Fig. 1. The different modalities have different sensitivities to different degradation, this is why we choose these modality to extract quality aware features. Red indicates that this modality is more sensitive to the degradation, while blue indicates that it is less sensitive.

I. Introduction

Point clouds, as a fundamental 3D representation, are widely applicable in various domains, including virtual/augmented reality, automatic driving, and facial expression modeling [1]. However, representing 3D models accurately often requires large numbers of points, leading to high memory consumption. Consequently, storing and transmitting point clouds necessitates compression, which can degrade their visual quality. Point Cloud Quality Assessment (PCQA) aims to predict the quality of point clouds, a task closely tied to human visual system (HVS). This makes PCQA as a crucial component for enhancing compression strategies and improving the visual fidelity of point cloud data. According to the reliance on reference point clouds, PCQA methods can be divided into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) methods. In practical scenarios, such as 3D reconstruction, the availability of the pristine reference point cloud is often limited, making NR-PCQA methods more broadly applicable. Depending on the domain in which the PCQA metric is computed, PCQA methods can also be divided into model-based and projection-based categories. Model-based PCQA methods directly utilize the geometric and color attributes of distorted point clouds to predict quality scores, making them more sensitive to down sampling and geometric Gaussian noise. In contrast, projection-based PCQA methods use projection video/image of distorted point clouds to quantify quality, making them more sensitive to color noise and video-based point cloud compression (V-PCC). MM-PCQA [2] represents the first attempt to integrate point-based and projection-based methods, achieving notable performance. However, this approach relies on separate and static projections, which is inconsistent with how people actually view point clouds [3]. Additionally, methods such as [4], [5] leverage normal or depth images to capture local geometric information, significantly enhancing performance.

Therefore, we propose a novel multi-modal framework for NR-PCQA, named VIP-PCQA, which extracts the perception quality features from the video modality, the image modality, and the point cloud modality. Fig 1 illustrates the different sensitivities of each modality to various degradation. The video modality is more sensitive to color noise, while the structural damage and geometry down sampling are more obvious in the point cloud modality. The geometric compression and color quantization are more noticeable in the normal image modality. Consequently, the proposed multi-modal framework compensates for the limitations of each modality by leveraging their unique strengths.

For the video modality, we first rotate the camera around the point cloud along two orthogonal circular pathways to capture two projection videos, ensuring coverage of quality-

This work was supported in part by the National Natural Science Foundation of China under Grant (62401365, 62225112, 62271312, 62132006, 623B2073). * Corresponding authors (Email :{huiyuduan, minxiongkuo, zhaiguangtao}@sjtu.edu.cn)

aware content. Next, following strategies commonly used in Video Quality Assessment (VQA) methods [6]–[9], we extract spatial and temporal features from selected key frames and video clips using the Swin Transformer [10] and SlowFast R50 [11]. For the image modality, we render normal images from six perpendicular projections in a cube of the point cloud and utilize the Swin Transformer to extract relevant features. For the point cloud modality, we extract 3D Natural Scene Statistics (3DNSS) from the distorted point cloud as described in [12], and sample patches containing local patterns such as smoothness and roughness. Local features are then extracted from these patches using PointNet++ [13]. Finally, we employ the simple fully connected layers to regress the concatenated quality features into final quality scores. Extensive experimental results demonstrate that the proposed method outperforms existing NR-3DQA methods on three PCQA benchmark.

II. Related Work

A. Model-based PCQA

Model-based PCQA directly use the attributes of point clouds, such as geometry and color, to predict quality scores. Early methods [4], [14]–[17] calculate and analysis the similarity of various aspect between reference and distorted point cloud to predict quality of distorted ones. Afterwards, 3D-NSS [12] utilizes some classic Natural Scene Statistics (NSS) distributions to quantify the quality of distortion point clouds. ResSCNN [18] employs an end-to-end convolutional neural network (CNN) to extract quality features and predict quality score of distortion point clouds. Model-based methods are relatively insensitive to the distortion likes Video-based Point Cloud Compression and color noise, so many works are now turning to projection-based PCQA.

B. Projection-based PCQA

The early projection-based FR-PCQA methods calculate the quality scores by projections rendered from distorted point clouds. These work [5], [19] employ conventional Image Quality Assessment (IQA) metrics to these projections and depth images on six perpendicular projections for quality evaluation. The development of deep learning networks has further enhanced the performance of projection-based methods. VQA-PC [3] deal with PCQA as the VQA problem by evaluating point cloud from moving camera videos. Yang *et al.* [20] introduces domain adaptation to use natural images to help model understand the quality of point cloud rendering images. GMS-3DQA [21] propose a multi-projection grid mini-patch sampling strategy to improve the efficiency of PCQA. Recently, LMM-PCQA [22] introduces Large Multi-modality Models (LMM) to evaluate the quality of point clouds through their projections, achieving outstanding performance.

III. Proposed Method

The framework overview is clearly exhibited in Fig 2. The distorted point clouds are converted to projection videos, normal images, patches, and 3DNSS features through preprocessing. Afterwards, these data are fed to different modules to extract corresponding features. Finally, the features are concatenated and mapped to the quality scores via the quality regression.

A. Data preprocessing

Suppose we have a distorted point cloud $\mathbf{P} = \{g(i), c(i)\}_{i=1}^{K}$, where $g(i) \in \mathbb{R}^{1\times3}$ and $c(i) \in \mathbb{R}^{1\times3}$ indicate the geometry coordinate and corresponding RGB color information respectively. K stands for the number of points. In order to get projection videos, we use camera in 2 orthogonal circular pathways to render projection videos, these two pathways can be represented as:

$$\theta_A : \begin{cases} x^2 + z^2 = R^2, \\ y = 0, \end{cases} \quad \theta_B : \begin{cases} x^2 + y^2 = R^2, \\ z = 0, \end{cases} \tag{1}$$

where x, y, and z indicate the coordinate in the Cartesian coordinate system with $O = \frac{1}{K}\sum_{k=1}^{K} g(k)$ as the center, R is the radius of the circular pathways. With the help of Open3d [23], we can render corresponding video sequence $\mathbf{V}_\varsigma$:

$$\mathbf{V}_\varsigma = \Psi_\varsigma(P), \quad \varsigma \in (A, B), \tag{2}$$

where Ψ_ς represents the video capture operation in the corresponding pathway. For the image modality, we define 6 perpendicular viewpoints of the given point cloud, corresponding to the 6 surfaces of a cube:

$$\mathbf{I} = \psi(\mathbf{P}), \tag{3}$$

where $\mathbf{I} = \{I_i | i = 1, \ldots, 6\}$ is the set of 6 normal projections and $\psi(\cdot)$ indicates the normal projections capture process. For the point cloud modality, we can get 3DNSS features following [12]:

$$F_{ns} = \mathbf{3DNSS}(\mathbf{P}), \tag{4}$$

where $\mathbf{3DNSS}(\cdot)$ and F_{ns} are the extract operation and features of 3DNSS respectively. Additionally, we can get the geometry normalized point cloud $\hat{\mathbf{P}} = \{\hat{g(k)}\}_{i=1}^{K}$, and then we utilize the farthest point sampling to sample K_a anchor points. We then employ K nearest neighbor (KNN) algorithm to find K_n neighboring points around each anchor points:

$$\mathbf{S} = \{S_i = \mathbf{KNN}(\hat{g(i)}) | i = 1 \ldots, K_a\}, \tag{5}$$

where $\mathbf{S}$ is the set of patches and $\mathbf{KNN}(\cdot)$ indicates the KNN operation.

B. Video modality features extraction

Following previous VQA methods [6], [7], [24], [25], we extract the quality-aware features of the projection videos from the spatial and temporal domains. In the spatial domain, we employ a image encoder $\mathbf{E}_{vs}(\cdot)$ to extract spatial features F_{vs} from the key frames sampled from projection videos:

$$F_{vs}^{i\varsigma} = \mathbf{E}_{vs}(\kappa_i^\varsigma), \tag{6}$$

where κ_k^ς is the i-th key frame from projection video $\mathbf{V}_\varsigma, \varsigma \in (A, B)$, whose number of frames and frame rate are l and r respectively. we can split the projection video into clips $\mathbf{C} =$

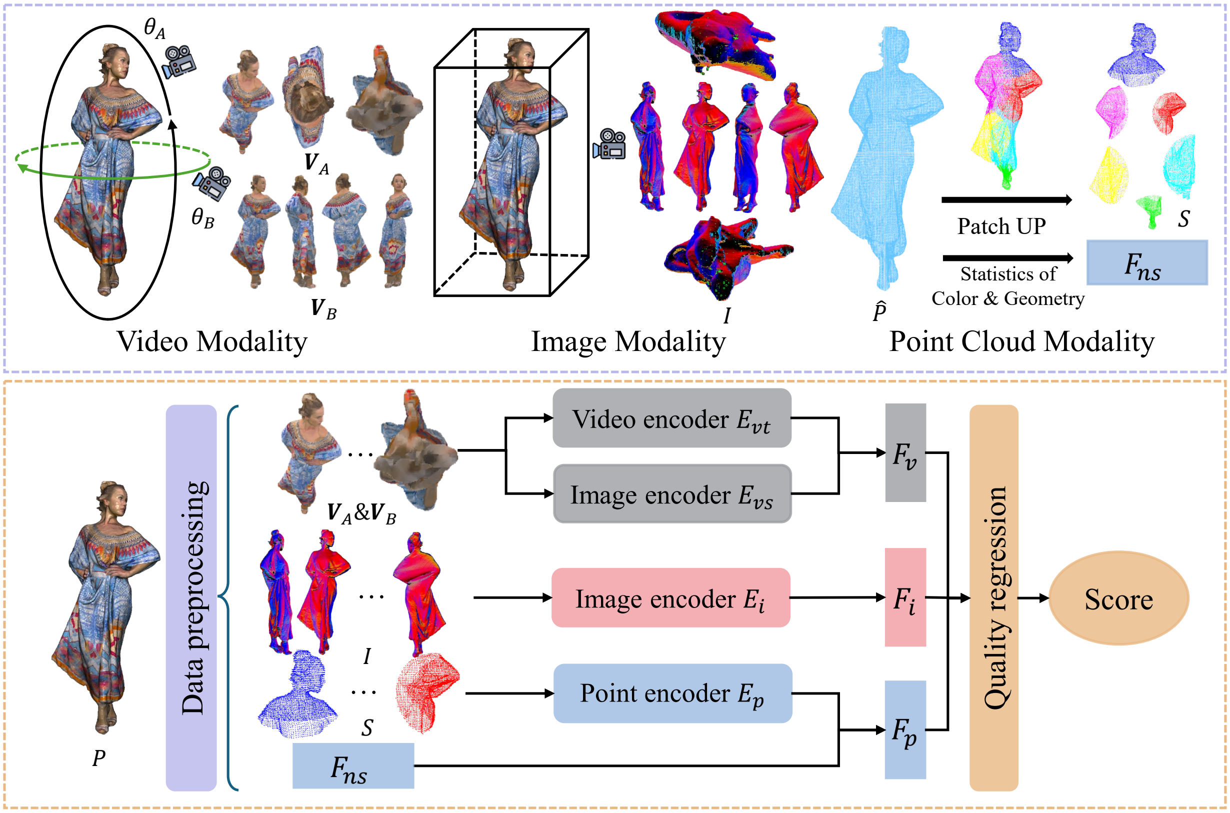

Fig. 2. The data preprocessing and framework of our proposed method. We obtain the data in video, image, and point cloud modality from point clouds in data preprocessing and our method extract features from these data to predict the quality score.

$\{C_i|i = 1, \ldots, K_k = l/r\}$ and the key frame are the first frame in each clip.

In the temporal domain, we use a pretrained 3D-CNN model to extract temporal features from the downsampled low resolution projection videos, this is because pretrained 3D-CNN model is able to capture the effect of content changes and extract the motion information that is highly correlated with the HVS [7] and the temporal features are not sensitive to the resolution. the temporal features F_{vt} can be derived as:

$$F_{vt}^{i\varsigma} = \mathbf{E}_{vt}(C_i^{\varsigma}), \tag{7}$$

where C_i^{ς} is the i-th clips from projection video $\mathbf{V}_{\varsigma}, \varsigma \in (A, B)$, $\mathbf{E}_{vt}$ indicates the pretrained 3D-CNN model. Finally, we fusion the spatial features and temporal features to get the video modality features F_v:

$$F_v = \frac{1}{2 * K_k} \sum_{k=1}^{K_k} F_{vs}^{kA} \oplus F_{vt}^{kA} + F_{vs}^{kB} \oplus F_{vt}^{kB}, \tag{8}$$

where $\oplus$ indicates the concatenation operation.

C. Image modality features extraction

We employ a image encoder $E_i(\cdot)$ to extract quality-aware features from normal projections set $\mathbf{I}$:

$$F_i = \frac{1}{6} \sum_{k=1}^{6} E_i(I_k), \tag{9}$$

where F_i represents the image modality features, which is more sensitive to point cloud compression and local geometry distortion.

D. Point modality features extraction

For the set of patches obtained in preprocessing, we employ a point cloud encoder $\mathbf{E}_p(\cdot)$ to convert the patches to quality aware embedding space:

$$F_{pt} = \frac{1}{K_a} \sum_{k=1}^{K_a} \mathbf{E}_p(S_k), \tag{10}$$

where F_{pt} is the average of the quality aware features of obtained patches and is capable to express the local geometry of point clouds. Finally we concatenate the 3DNSS features and local geometry features to get point modality features F_p:

$$F_p = F_{pt} \oplus F_{ns}. \tag{11}$$

E. Quality Regression & Loss Function

We simply use several fully connected layers **FC** to regress the concatenated perceptual quality features into quality scores:

$$q = \mathbf{FC}(F_v \oplus F_i \oplus F_p), \tag{12}$$

following previous works [2], [6], we use the Mean Squared Error (MSE) loss and rank loss to compose our loss, which is because the quality assessment task emphasizes not only the accuracy of the predicted quality values but also the importance of preserving quality rankings. The MSE loss is aim to ensure the predicted values closely align with the quality labels, which is defined as:

$$L_{mse} = \frac{1}{N}\sum_{n=1}^{N}(q_n - q_n^{'})^2, \tag{13}$$

where q_n and $q_n^{'}$ are predicted and labeled quality scores respectively, and N indicates the size of the mini-batch. The rank loss helps the model differentiate the relative quality of point clouds, making it particularly effective for evaluating point clouds with similar quality levels. We use the differentiable rank function proposed in [6] to calculate rank loss:

$$L_{rank}^{ij} = max(0, |q_i - q_j| - e(q_i, q_j) \cdot (q_i^{'} - q_j^{'})), \tag{14}$$

$$e(q_i, q_j) = \begin{cases} 1 & q_i \geq q_j, \\ -1 & q_i < q_j, \end{cases} \tag{15}$$

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

$$L_{rank} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} L_{rank}^{ij} \tag{16}$$

The final loss function can be calculated as the weight sum of MSE loss and rank loss:

$$Loss = \lambda_1 L_{mse} + \lambda_2 L_{rank}, \tag{17}$$

where λ_1 and λ_2 are hyperparameters to control the proportion of above losses.

IV. Experiment

A. Experiment Protocol

1) Databases: We test the performance of proposed VIP-PCQA on three commonly used databases: SJTU-PCQA [5], WPC [26], and WPC2.0 [27]. The SJTU-PCQA database applies 7 types of distortion (color noise, compression, geometric shift, down-sampling, and three mixed distortions) with 6 levels to 9 reference point clouds and generates 378 distorted point clouds. The WPC database contains 20 reference point clouds and each point cloud is degraded into 37 distorted stimuli by 4 distortions (Gaussian white noise, down sampling, Geometry-based Point Cloud Compression (G-PCC), and V-PCC), a total of 740 distorted point clouds in the WPC database. The WPC2.0 database has 16 reference point clouds, each undergoing 25 degradation (V-PCC) settings, leading to 400 distorted point clouds.

2) Evaluation Criteria: Four common consistency evaluation criteria are use to judge the correlation between the predicted scores and quality annotations, including Spearman Rank Correlation Coefficient (SRCC), Kendall's Rank Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Squared Error (RMSE). An effective model should achieve SRCC, KRCC, and PLCC values approaching 1, while maintaining a RMSE value close to 0. Followed by [**?**], [28]–[31], we adopt a four-parameter logistic function to map the predicted scores into the subjective quality scores before calculating the criteria.

3) Compared Methods: For the above PCQA databases, We compare our proposed method with the following methods:

- FR quality assessment methods: MSE-p2point (MES-p2po) [14], Hausdorff-p2point (HD-p2po) [14], MSE-p2plane (MSE-p2pl) [4], Hausdorff-p2plane (HD-p2pl) [4], PSNR-yuv [32], PCQM [15], GraphSIM [16], and PointSSIM [17].
- NR quality assessment methods: BRISQUE [33], NIQE [34], IL-NIQE [35], IT-PCQA [20], ResSCNN [18], PQA-net [36], 3D-NSS [12], GMS-3DQA [21], MM-PCQA [2], and LMM-PCQA [22].

It is worth noting that BRISQUE, NIQE, IL-NIQE are image-based quality assessment method and validated on the 6 perpendicular projections and the average scores are recorded.

4) Experiment Setup: For the image encoders $\mathbf{E}_{vs}$ and $\mathbf{E}_i$, we both use the architecture of Swin transformer tiny [10] and initialize it with the weights pretrained on the ImageNet-1k [37]. The SlowFast R50 [11] is utilized as the temporal feature extractor $\mathbf{E}_{vt}$. Additionally, we use the point++ [13] as the point cloud encoder $\mathbf{E}_p(\cdot)$. The Adam optimizer with the initial learning rate $5e-5$ and batch size 4 are used for training the proposed model. The learning rate decays with a multiplicative factor of 0.9 for every 5 epochs and the number of epochs is set at 50. Specifically, we set the number of anchor points K_a and neighboring points K_n as 6 and 2048. The normal projection and projection videos are rendered in the resolution of 2048×2048 and 512×512 respectively. In addition, we remove the white background in the normal projection and key frames and crop it in the resolution of 224×224 as input. The weights λ_1 and λ_2 for L_{mse} and L_{rank} are both set as 1.

Following the practices in [2], [21], [24], [38], [39], we employ the k-fold cross validation strategy in the experiments on the above databases to accurately estimate the performance of the proposed VIP-PCQA. The 9-fold, 5-fold, and 4-fold cross validation is selected for SJTU-PCQA, WPC, and WPC2.0 respectively. The average performance is recorded as the final results.

B. Experimental Results

The final experimental results on the SJTU-PCQA, WPC, and WPC2.0 databases are shown in Table I, from which we can make several useful conclusions: 1) The proposed method demonstrates outstanding performance on the aforementioned databases, achieving the best results on the SJTU-PCQA and WPC databases and ranking second on the WPC2.0 database. 2) From the SJTU-PCQA database to WPC database and WPC2.0 database, the performance of most methods has dropped significantly, which may be due to WPC database and WPC 2.0 database use more complex and finer-grained degradation to reference point clouds, which is more difficult to PCQA task. 3) Based on the previous conclusions, we attempt to explain why the result of our methods on the WPC2.0 database is not as good as LMM-PCQA. Since WPC2.0 has more finer-grained degradation and less data, it requires higher generalization of the method. LMM-PCQA is based on LMM, which uses the Internet-scale data for

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART APPROACHES ON THE SJTU-PCQA, WPC, AND WPC2.0 DATABASES. BEST IN **RED**, SECOND IN BLUE.

Type	Methods	SJTU-PCQA				WPC				WPC2.0			
		SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
FR	MSE-p2po	0.7294	0.8123	0.5617	1.3613	0.4558	0.4852	0.3182	19.8943	0.4315	0.4626	0.3082	19.1605
	HD-p2po	0.7157	0.7753	0.5447	1.4475	0.2786	0.3972	0.1943	20.8990	0.3587	0.4561	0.2641	18.8976
	MSE-p2pl	0.6277	0.5940	0.4825	2.2815	0.3281	0.2695	0.2249	22.8226	0.4136	0.4104	0.2965	21.0400
	HD-p2pl	0.6441	0.6874	0.4565	2.1255	0.2827	0.2753	0.1696	21.9893	0.4074	0.4402	0.3174	19.5154
	PSNR-yuv	0.7950	0.8170	0.6196	1.3151	0.4493	0.5304	0.3198	19.3119	0.3732	0.3557	0.2277	20.1465
	PCQM	0.8644	0.8853	0.7086	1.0862	0.7434	0.7499	0.5601	15.1639	0.6825	0.6923	0.4929	15.6314
	GraphSIM	0.8783	0.8449	0.6947	1.0321	0.5831	0.6163	0.4194	17.1939	0.7405	0.7512	0.5533	14.9922
	PointSSIM	0.6867	0.7136	0.4964	1.7001	0.4542	0.4667	0.3278	20.2733	0.4810	0.4705	0.2978	19.3917
NR	BRISQUE	0.3975	0.4214	0.2966	2.0937	0.2614	0.3155	0.2088	21.1736	0.0820	0.3353	0.0487	21.6679
	NIQE	0.1379	0.2420	0.1009	2.2622	0.1136	0.2225	0.0953	23.1415	0.1865	0.2925	0.1335	22.5146
	IL-NIQE	0.0837	0.1603	0.0594	2.3378	0.0913	0.1422	0.0853	24.0133	0.0911	0.1233	0.0714	23.9987
	IT-PCQA	0.8651	0.8283	0.6430	1.1661	0.4870	0.4329	0.3006	19.8960	0.5661	0.5432	0.3477	18.7224
	ResSCNN	0.8600	0.8100	-	-	-	-	-	-	0.7500	0.7200	-	-
	PQA-net	0.8372	0.8586	0.6304	1.0719	0.7026	0.7122	0.4939	15.0812	0.6191	0.6426	0.4606	16.9756
	3D-NSS	0.7144	0.7382	0.5174	1.7686	0.6479	0.6514	0.4417	16.5716	0.5077	0.5699	0.3638	17.7219
	GMS-3DQA	0.9108	0.9177	0.7735	0.7872	0.8308	0.8338	0.6457	12.2292	0.8272	0.8218	0.6277	12.9904
	MM-PCQA	0.9103	0.9226	0.7838	0.7716	0.8414	0.8556	0.6513	12.3506	0.8023	0.8024	0.6202	13.4289
	LMM-PCQA	**0.9376**	**0.9404**	**0.8002**	**0.7175**	**0.8825**	**0.8739**	**0.7064**	**11.8171**	**0.8614**	**0.8634**	**0.6723**	**10.6924**
	Ours	**0.9459**	**0.9582**	**0.8176**	**0.6661**	**0.8902**	**0.8899**	**0.7160**	**10.2990**	**0.8565**	**0.8555**	**0.6745**	**10.9981**

pretraining and contains several billions of parameters, so it has better generalization performance. 4) The results of BRISQUE, NIQE, and IL-NIQE are the worst, which may be because these methods are designed for natural scene images based on statistical regularities in natural images, which is quite different from the geometry and color distortions in the projection of point clouds.

C. Ablation Study

In order to investigate the contributions of different modalities and validate the rationality of each modalities covered by our approach, we undertake an ablation study and the result are shown in Table II. From the experimental results, we can draw the following conclusions: 1) The video modality, image modality, and point cloud modality all contribute to the final result, among which the video modality contributes the most, which may be because the video modality can better reflect various distortions in the point cloud. 2) The performance on the WPC2.0 database and WPC database drops more than that on the SJTU-PCQA database in the case of missing same modality, this indicates that fusing multiple modalities improve the model to cope with more complex and finer-grained degradation.

D. Cross-Dataset Study

We conduct the cross-database evaluation to test the generalization ability of proposed methods. Since the WPC database has the largest number of point clouds, so we trained our model on it and evaluated on SJTU-PCQA database and WPC2.0 database. From the experimental results shown in Table III, we can find that: Our proposed method has good generalization performance and second only to LMM-PCQA, which employs the LMM to evaluate the projections of point clouds. This is understandable, because the LMM is pretrained on huge scale data and has several billions of parameters.

TABLE II
CONTRIBUTIONS OF VIDEO MODALITY, IMAGE MODALITY, AND POINT CLOUD MODALITY, WHERE 'W/O VIDEO', 'W/O IMAGE', AND 'W/O POINT CLOUD ' INDICATES EXCLUDING THE VIDEO, IMAGE, OR POINT CLOUD MODALITY RESPECTIVELY. BEST IN **RED**, SECOND IN BLUE.

Modal	SJTU-PCQA		WPC		WPC2.0	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
w/o video	0.8692	0.8874	0.7314	0.7445	0.7180	0.7201
w/o point cloud	0.8924	0.9021	0.7832	0.8093	0.7754	0.7653
w/o image	**0.9137**	**0.9228**	**0.8584**	**0.8579**	**0.8089**	**0.8066**
complete model	**0.9459**	**0.9582**	**0.8902**	**0.8899**	**0.8565**	**0.8555**

TABLE III
THE EXPERIMENTAL RESULTS OF CROSS-DATABASE STUDY, 'WPC→SJTU-PCQA' AND 'WPC→WPC2.0' SIGNIFIES THAT THE MODEL IS TRAINED ON WPC DATABASE AND TESTED ON SJTU-PCQA DATABASE AND WPC2.0 DATABASE RESPECTIVELY. ADDTIONALLY, WE ELIMINATE THOSE POINT CLOUD GROUPS FROM THE WPC DATABASE THAT HAVE REFERENCE COUNTERPARTS IN THE WPC2.0 TESTING SETS. BEST IN **RED**, SECOND IN BLUE.

Model	WPC→SJTU-PCQA		WPC→WPC2.0	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
PQA-net	0.5411	0.6102	0.6006	0.6377
3D-NSS	0.1817	0.2344	0.4933	0.5613
GMS-3DQA	0.7421	0.7611	0.7822	0.7714
MM-PCQA	0.7991	0.7902	0.7917	0.7935
LMM-PCQA	**0.8246**	**0.7999**	**0.8385**	**0.8387**
Ours	**0.8009**	**0.7916**	**0.7941**	**0.8081**

V. CONCLUSION

In this paper, we propose the novel multi-modal learning framework VIP-PCQA for NR-PCQA. The approach begins by rendering projection videos, and normal images, obtaining patches and 3DNSS features from distorted point clouds. Features are then extracted from these data using video, image, and point cloud encoders, which are subsequently fused to predict the final quality score. Experimental evaluations on three benchmark databases and comprehensive ablation studies validate the effectiveness of the proposed approach and underscore the potency of the proposed VIP-PCQA.

References

[1] Sijing Wu, Yunhao Li, Yichao Yan, Huiyu Duan, Ziwei Liu, and Guangtao Zhai, "Mmhead: Towards fine-grained multi-modal 3d facial animation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7966–7975.

[2] Zicheng Zhang, Wei Sun, Xiongkuo Min, Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai, "Mm-pcqa: Multi-modal learning for no-reference point cloud quality assessment," *International Joint Conference on Artificial Intelligence*, 2023.

[3] Zicheng Zhang, Wei Sun, Yucheng Zhu, Xiongkuo Min, Wei Wu, Ying Chen, and Guangtao Zhai, "Treating point cloud as moving camera videos: A no-reference quality assessment metric," *arXiv e-prints*, pp. arXiv–2208, 2022.

[4] Dong Tian, Hideaki Ochimizu, Chen Feng, Robert Cohen, and Anthony Vetro, "Geometric distortion metrics for point cloud compression," in *IEEE International Conference on Image Processing*, 2017, pp. 3460–3464.

[5] Qi Yang, Hao Chen, Zhan Ma, Yiling Xu, Rongjun Tang, and Jun Sun, "Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration," *IEEE Transactions on Multimedia*, 2020.

[6] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 856–865.

[7] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5944–5958, 2022.

[8] Wei Sun, Tao Wang, Xiongkuo Min, Fuwang Yi, and Guangtao Zhai, "Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[9] Fuwang Yi, Mianyi Chen, Wei Sun, Xiongkuo Min, Yuan Tian, and Guangtao Zhai, "Attention based network for no-reference ugc video quality assessment," in *IEEE international conference on image processing (ICIP)*. IEEE, 2021, pp. 1414–1418.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[12] Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, and Guangtao Zhai, "No-reference quality assessment for 3d colored point cloud and mesh models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[13] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[14] R Mekuria, Z Li, C Tulvan, and P Chou, "Evaluation criteria for point cloud compression," *ISO/IEC MPEG*, , no. 16332, 2016.

[15] Gabriel Meynet, Yana Nehmé, Julie Digne, and Guillaume Lavoué, "Pcqm: A full-reference quality metric for colored 3d point clouds," in *International Workshop on Quality of Multimedia*, 2020, pp. 1–6.

[16] Qi Yang, Zhan Ma, Yiling Xu, Zhu Li, and Jun Sun, "Inferring point cloud quality via graph similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[17] Evangelos Alexiou and Touradj Ebrahimi, "Towards a point cloud structural similarity metric," in *International Conference on Multimedia and Expo Workshop*, 2020, pp. 1–6.

[18] Yipeng Liu, Qi Yang, Yiling Xu, and Le Yang, "Point cloud quality assessment: Dataset construction and learning-based no-reference metric," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2022.

[19] Ricardo L De Queiroz and Philip A Chou, "Motion-compensated compression of dynamic voxelized point clouds," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3886–3895, 2017.

[20] Qi Yang, Yipeng Liu, Siheng Chen, Yiling Xu, and Jun Sun, "No-reference point cloud quality assessment via domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2022, pp. 21179–21188.

[21] Zicheng Zhang, Wei Sun, Houning Wu, Yingjie Zhou, Chunyi Li, Xiongkuo Min, Guangtao Zhai, and Weisi Lin, "Gms-3dqa: Projection-based grid mini-patch sampling for 3d model quality assessment," *arXiv preprint arXiv:2306.05658*, 2023.

[22] Zicheng Zhang, Haoning Wu, Yingjie Zhou, Chunyi Li, Wei Sun, Chaofeng Chen, Xiongkuo Min, Xiaohong Liu, Weisi Lin, and Guangtao Zhai, "Lmm-pcqa: Assisting point cloud quality assessment with lmm," *ACM MM*, 2024.

[23] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.

[24] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai, "Perceptual video quality assessment: A survey," *Science China Information Sciences*, vol. 67, no. 11, pp. 211301, 2024.

[25] Huiyu Duan, Guangtao Zhai, Xiaokang Yang, Duo Li, and Wenhan Zhu, "Ivqad 2017: An immersive video quality assessment database," in *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2017, pp. 1–5.

[26] Qi Liu, Honglei Su, Zhengfang Duanmu, Wentao Liu, and Zhou Wang, "Perceptual quality assessment of colored 3d point clouds," *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[27] Qi Liu, Hui Yuan, Raouf Hamzaoui, Honglei Su, Junhui Hou, and Huan Yang, "Reduced reference perceptual quality model with application to rate control for video-based point cloud compression," *IEEE Transactions on Image Processing*, vol. 30, pp. 6623–6636, 2021.

[28] Jochen Antkowiak, TDF Jamal Baina, France Vittorio Baroncini, Noel Chateau, France FranceTelecom, Antonio Claudio França Pessoa, FUB Stephanie Colonnese, Italy Laura Contin, Jorge Caviedes, and France Philips, "Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000," 2000.

[29] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, and Guangtao Zhai, "Finevq: Fine-grained user generated content video quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

[30] Xilei Zhu, Liu Yang, Huiyu Duan, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet, "Esiqa: Perceptual quality assessment of vision-pro-based egocentric spatial images," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2025.

[31] Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, and Patrick Le Callet, "Confusing image quality assessment: Toward better augmented reality experience," *IEEE Transactions on Image Processing (TIP)*, vol. 31, pp. 7206–7221, 2022.

[32] Eric M Torlig, Evangelos Alexiou, Tiago A Fonseca, Ricardo L de Queiroz, and Touradj Ebrahimi, "A novel methodology for quality assessment of voxelized point clouds," in *Applications of Digital Image Processing XLI*, 2018, vol. 10752, pp. 174–190.

[33] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[34] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[35] Lin Zhang, Lei Zhang, and Alan C Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[36] Qi Liu, Hui Yuan, Honglei Su, Hao Liu, Yu Wang, Huan Yang, and Junhui Hou, "Pqa-net: Deep no reference point cloud quality assessment via multi-view projection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4645–4660, 2021.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[38] Huiyu Duan, Xiongkuo Min, Wei Sun, Yucheng Zhu, Xiao-Ping Zhang, and Guangtao Zhai, "Attentive deep image quality assessment for omnidirectional stitching," *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, vol. 17, no. 6, pp. 1150–1164, 2023.

[39] Kang Fu, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, Xiongkuo Min, Jia Wang, and Guangtao Zhai, "Multi-dimensional quality assessment for text-to-3d assets: Dataset and model," *IEEE Transactions on Multimedia (TMM)*, 2025.