MMFN-PCQA: A NOVEL MUTI-MODAL FUSION NETWORK FOR NO-REFERENCE POINT CLOUD QUALITY ASSESSMENT

DIAN-JUN XU[1], XIAO-NA ZENG[1], ZE-XI LI[1], XIAN-WEI ZHENG[1]

[1]The School of Mathematics, Foshan University, Foshan 528000, China
E-MAIL: mulingxiao0612@qq.com, xnzeng1030@163.com, 3524501558@qq.com, alex.w.zheng@hotmail.com

Abstract:

Point cloud quality assessment (PCQA) is essential for numerous 3D vision applications. Existing no-reference (NR) approaches typically rely on either 3D geometry or 2D texture features alone, limiting their accuracy. To address these limitations, we propose MMFN-PCQA, a novel multi-modal fusion network for NR-PCQA. MMFN-PCQA integrates both 3D geometric structures and 2D texture information through a dual-branch transformer architecture. Key innovations include a Local-Feature Self-Attention (LFSA) module, which enhances fine-grained local geometric details, and a hybrid U-net Partitioning Segmentation (UPS) module to effectively capture texture information from 2D projections. An Adaptive Cross-Modal Fusion (ACMF) mechanism selectively integrates these modalities via multi-head cross-attention and gated residual aggregation, ensuring robust yet discriminative feature fusion. Experimental results show that our method achieves competitive performance compared with state-of-the-art approaches.Moreover, MMFN-PCQA not only significantly improves assessment accuracy but also retains a lightweight, practical architecture suitable for real-world deployment.

Keywords:

Multi-modal fusion; Point cloud quality accessement; Local-feature self-attention; Adaptive fusion

1. Introduction

Point Cloud Quality Assessment (PCQA) is a crucial research topic in computer vision and multimedia. Its goal is to objectively evaluate point cloud data quality to guide subsequent processing, compression, transmission, and applications. With rapid advances in 3D data acquisition, point clouds — as the primary 3D data representation — are widely used in areas such as autonomous driving, virtual reality, and 3D reconstruction [1]. However, during acquisition, compression, and transmission, point clouds often suffer geometric deformations, texture distortions, and noise. These issues degrade data reliability and limit the effectiveness of point clouds in critical applications. Consequently, PCQA has become fundamental to ensuring the success of 3D technologies by addressing these quality issues.

Traditional PCQA methods are generally classified into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) categories. FR-PCQA relies on high-quality reference point clouds, whereas RR-PCQA uses manually crafted low-resolution reference images for assessment [2]. Although effective in certain scenarios, these methods require additional computational resources or high-quality references, limiting their practical adoption. In contrast, NR-PCQA, which requires no reference data, has attracted increasing attention due to its greater applicability. Existing NR-PCQA algorithms typically exploit hand-crafted statistical features from point clouds [3, 4] or end-to-end neural networks [5], or they derive features from 2-D projections using hand-engineered descriptors [6–8] or CNN backbones [9–11]. These single-modality approaches often fail to exploit the complementary strengths of point clouds (3-D geometry) and projection images (2-D texture), resulting in suboptimal performance.

With the rise of multimodal learning, researchers have started to fuse heterogeneous modalities to obtain more reliable quality predictions. Recent multimodal PCQA frameworks demonstrate that combining image textures with point-cloud geometry can yield a more comprehensive assessment [12]. Nevertheless, how best to extract, align, and fuse cross-modal features while keeping the overall model lightweight remains an open challenge.

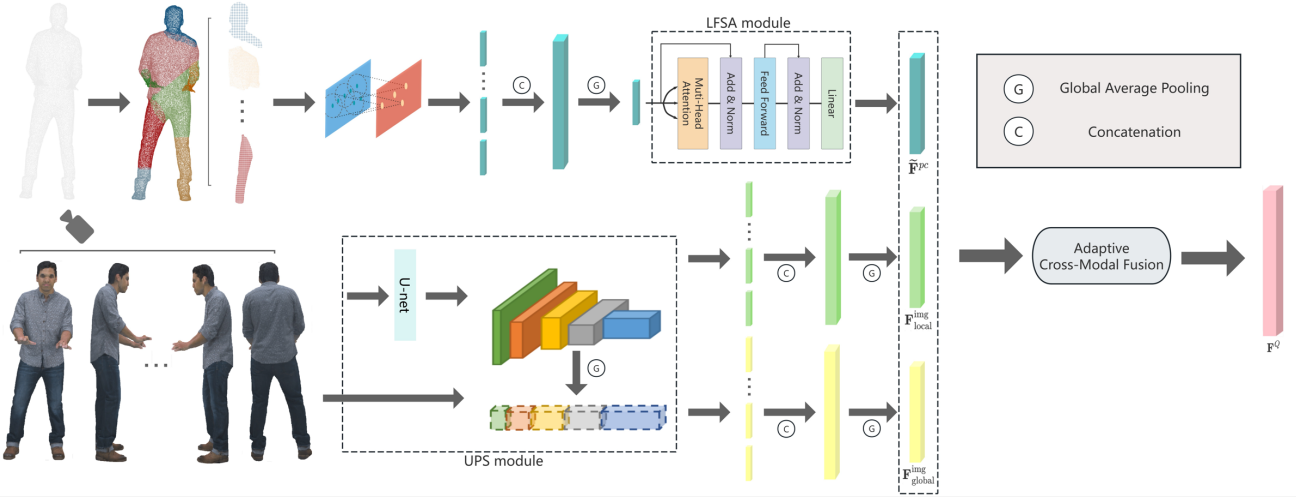

FIGURE 1. MMFN-PCQA Architecture Diagram.

To address current limitations, we propose MMFN-PCQA, a transformer-based NR-PCQA framework (Figure. 1). A dual-stream transformer lets point-cloud and image tokens learn independently, mirroring the human visual system, while an Adaptive Cross-Modal Fusion module uses cross-attention to share only the most relevant geometric and image cues. This design preserves modality-specific learning yet captures rich inter-modal correlations, yielding highly discriminative quality features.

The main contributions of this paper are as follows

- We design a point-cloud encoder enhanced by Local-Feature Self-Attention, together with an image encoder that blends ResNet-50 and U-Net, allowing our framework to capture fine-grained 3-D geometry and rich 2-D texture cues in a unified manner.
- We introduce ACMF, which uses multi-head cross-attention (MCA) and gated residual aggregation to adaptively integrate the two modalities in an HVS-inspired manner, balancing modality-specific refinement with effective information exchange.
- Experiment results show that our MMFN-PCQA achieves state-of-the-art performance on three benchmarks.

2. Methods

In this section, we present the technical details of our MMFN-PCQA framework for point clouds. As illustrated in Figure. 1, the framework consists of three main components: 1) Multi-modal feature extraction; 2) Adaptive Cross-Modal Fusion; and 3) A loss function for quality prediction.

2.1 Multi-modal feature extraction

To guarantee a comprehensive assessment, we design a dual-branch feature-extraction module composed of a point-cloud encoder θ_{pc} for the 3-D modality and an image encoder θ_{img} for the 2-D modality. Relying on a single modality would miss critical cues; therefore, our approach leverages 3-D structural information together with 2-D texture details for more robust evaluation.

Point Cloud Feature Extraction. In the 3-D branch, each point cloud is first down-sampled by Farthest Point Sampling (FPS) to obtain N_δ representative points $\{\delta_m\}_{m=1}^{N_\delta}$. For every sampled point, the N_s nearest neighbors are retrieved using the k-Nearest-Neighbors (KNN) algorithm [12], forming a set of local sub-models that are fed into a PointNet++-based [13] encoder θ_{pc} to produce patch-level features F^{pc}:

$$S^{pc} = \left\{ \underset{k=N_s}{\mathrm{KNN}}(\delta_m) \right\}_{m=1}^{N_\delta} \tag{1}$$

$$F^{pc} = \{\theta_{pc}(S_l^{pc})\}_{l=1}^{N_\delta} \tag{2}$$

where S^{pc} is the set of point-cloud sub-models, $KNN(\cdot)$ denotes the k-nearest-neighbor operation, and δ_m is the m-th farthest-sampled point.

Traditional PointNet++ aggregates channel information only via max-pooling at the back-end and thus struggles to discriminate fine-grained differences such as subtle noise or deformations. To address this, we introduce a Local Feature Self-Attention (LFSA) module (see Figure. 1) that performs channel-wise self-reweighting with a residual path, thereby suppressing over-smoothing while preserving local details:

$$\widetilde{F}^{pc} = \mathrm{LFSA}\left(\frac{1}{N_\delta}\sum_{i=1}^{N_\delta} F_i^{pc}\right) \tag{3}$$

where F_i^{pc} is the quality-aware embedding of the l-th sub-model S_l^{pc}, and $\widetilde{F}^{pc}$ is the LFSA-enhanced point-cloud feature before fusion.

Image Feature Extraction. For the 2-D branch, N_{frames} rendered views I_i are generated. A pre-trained ResNet-50 [14] serves as the backbone image encoder θ_{img}; Features from each view are extracted and aligned via global average pooling to yield a global projection descriptor F_{global}^{img}:

$$F_{global}^{img} = \frac{1}{N_{frames}}\sum_{i=1}^{N_{frames}} \theta_{img}(I_i) \tag{4}$$

where N_{frames} is the number of rendered views.

To preserve local texture, we design an U-net Partitioning Segmentation (UPS) module that fuses U-Net [15] with ResNet-50 to produce block-level representations:

$$F_{local}^{img} = \mathrm{UPS}(I_i), i = 1, ..., N_{frames} \tag{5}$$

2.2 Adaptive Cross-Modal Fusion

To seamlessly integrate 2-D and 3-D modality, we propose an Adaptive Cross-Modal Fusion (ACMF) module. It employs a multi-head cross-attention (MCA) strategy to enhance global and local image features, followed by a Gated-Residual-Aggregation (GRA) that further blends image and point-cloud representations (see Figure.2).

Secifically, MCA enhances local image features F_{local}^{img} and global image features F_{global}^{img}, and then inputs them together with the original local image features and global

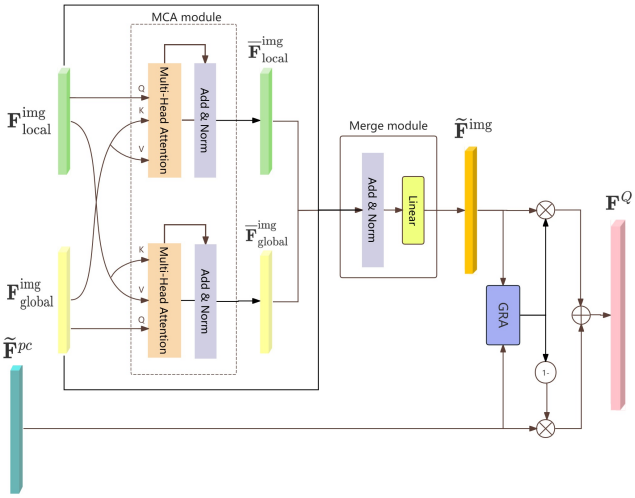

FIGURE 2. ACMF Module.

image features into the Merge module for a more refined image representation $\widetilde{F}^{img}$:

$$\overline{F}_{local}^{img}, \overline{F}_{global}^{img} = \mathrm{MCA}(F_{local}^{img}, F_{global}^{img}) \tag{6}$$

$$\widetilde{F}^{\mathrm{img}} = \mathrm{Merge}\{F_{local}^{img}, F_{global}^{img}, \overline{F}_{local}^{img}, \overline{F}_{global}^{img}\} \tag{7}$$

Finally, $\widetilde{F}^{img}$ and the self-enhanced point-cloud feature $\widetilde{F}^{pc}$ are adaptively weighted by the GRA gate and combined with residual links to produce a high-capacity quality feature F^Q:

$$\begin{aligned} F^Q &= w \odot \widetilde{F}_{img} + (1-w) \odot \widetilde{F}_{pc} + \widetilde{F}_{img} + \widetilde{F}_{pc}, \\ w &= \sigma\left(g\left([\widetilde{F}_{img}], [\widetilde{F}_{pc}]\right)\right) \end{aligned} \tag{8}$$

where $g([\cdot],[\cdot])$ is the gating mechanism within the GRA module; it adaptively adjusts the fusion weight of each modality channel and, together with the residual links, strengthens information flow.

2.3 Loss function

Following common practice, we simply use two-fold fully-connected layers to regress the quality features F^Q into predicted quality scores. In our quality assessment framework, we care not only about the numerical accuracy of these predictions but also about preserving the correct relative ordering of samples. Accordingly, our loss function comprises two components: the Mean Squared Error

(MSE) and a ranking loss. The MSE term encourages the predicted scores to closely match the ground-truth quality labels, and is defined as:

$$L_{MSE} = \frac{1}{n}\sum_{n=1}^{n}(q_n - q_n')^2 \tag{9}$$

where q_n is the predicted quality scores, q_n' is the quality labels of the point cloud, and n is the size of the mini-batch.

The ranking loss further helps the model discriminate between point clouds whose quality labels are similar. To this end, we adopt the differentiable ranking function from [16] to approximate the ranking loss:

$$\begin{aligned} L_{ij}^{rank} &= \max\left(0, |q_i - q_j| - e(q_i, q_j)\cdot(q_i' - q_j')\right) \\ e\left(q_i, q_j\right) &= \begin{cases} 1, q_i \geq q_j \\ -1, q_i < q_j \end{cases} \end{aligned} \tag{10}$$

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

$$L_{rank} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}^{rank} \tag{11}$$

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

$$Loss = \lambda_1 L_{MSE} + \lambda_2 L_{rank} \tag{12}$$

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

3 Experiment

3.1 Database preparation

To validate the effectiveness of the proposed MMFN-PCQA, we conduct experiments on three public benchmarks-SJTU-PCQA [17], WPC [18], and LS-PCQA [5]. SJTU-PCQA comprises nine pristine point clouds, each degraded by seven distortion types at six severity levels, producing 378 distorted samples. WPC contains 20 reference point clouds and 740 distorted counterparts generated with four kinds of artifacts. LS-PCQA providing 24024 distorted point clouds derived from 104 references and synthesized with 33 distinct distortion types.

3.2 Implementation details

The Adam optimizer is utilized with weight decay 1e-4, the initial learning rate is set as 5e-5, and the batch size is set as 4. The model is trained for 50 epochs by default. Specifically, We set the point cloud sub-model size N_s as 2048, set the number of sub-models $N_\delta = 6$, and set the number of image projections $N_{\text{frames}} = 6$.

The projected images with the resolution of 1920×1080×3 are randomly cropped into image patches at the resolution of 224×224×3 as the inputs (the white background is removed from the projected images).

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.

3.3 Competitors and evaluation criteria

For the sake of comprehensively investigating the prediction performance and overall result of MMFN-PCQA, 14 state-of the-art quality assessment methods are selected for comparison, including 8 FR-PCQA methods and 6 NR-PCQA methods.

The FR-PCQA methods include MSEp2point (MSE-p2po) [9], Hausdorffp2point (HD-p2po) [9], MSE-p2plane (MSE-p2pl) [20], Hausdorff-p2plane (HD-p2pl) [20],PSNR-yuv [19], PCQM [9], GraphSIM [7], and PointSSIM [8].

NRPCQA methods include IT-pcqa [21], ResSCNN [5], GDBF [22], GPA-Net [23], TCM [24] and DisPA [25].

To evaluate how faithfully our metric reflects human judgments, we report the Spearman rank-order correlation coefficient (SROCC), Pearson linear correlation coefficient (PLCC), and root-mean-square error (RMSE). Each database is partitioned by content, meaning that folds never share the same reference point clouds. We follow the K-fold cross-validation protocol from [12], averaging the results to reduce randomness. As summarized in Table 1, MMFN-PCQA achieves the best performance on all three benchmarks, consistently outperforming competing NR-PCQA methods.

3.4 Ablation study

To validate the effectiveness of each component of our method, we present the ablation study results as follows. In the case of w/o ACMF, we simply integrate the features from both modalities via concatenation. For the

TABLE 1. Performance comparison with state-of-the-art approaches on the LS-PCQA, SJTU-PCQA, and WPC-PCQA databases. P and I stand for the point cloud and image modalities respectively. Best in red and second in blue.

Ref	Modal	Methods	LS-PCQA			SJTU-PCQA			WPC-PCQA		
			SRCC ↑	PLCC ↑	RMSE ↓	SRCC ↑	PLCC ↑	RMSE ↓	SRCC ↑	PLCC ↑	RMSE ↓
FR	P	MSE-p2pl	0.287	0.444	0.745	0.6277	0.594	2.2815	0.3281	0.2695	22.8226
	P	HD-p2pl	0.269	0.401	0.762	0.6441	0.6874	2.1255	0.2827	0.2753	21.9893
	P	PointSSIM	0.3155	0.5346	0.7564	0.6867	0.7136	1.7001	0.4542	0.4667	20.2733
	P	HD-p2po	0.269	0.403	0.761	0.7157	0.7753	1.4475	0.2786	0.3972	20.899
	P	MSE-p2po	0.325	0.528	1.58	0.7294	0.8123	1.3613	0.4558	0.4852	19.8943
	P	PSNR-yuv	0.4876	0.507	0.6341	0.7950	0.817	1.3151	0.4493	0.5304	19.3119
	P	PCQM	0.3911	0.4044	0.7527	0.8644	0.8853	1.0862	0.7434	0.7499	15.1639
	P	GraphSIM	0.332	0.355	0.778	0.8783	0.8449	1.0321	0.5831	0.6163	17.1939
NR	I	IT-pcqa	0.2611	0.3319	1.1484	0.3654	0.3952	2.8096	0.2866	0.3	22.866
	P	ResSCNN	0.4785	0.4349	1.1082	0.4446	0.4087	2.7915	0.4832	0.4956	18.7378
	P+I	GLDBF	-	-	-	0.85	0.873	1.482	0.775	0.792	11.459
	P	GPA-Net	0.602	0.628	-	0.875	0.886	-	0.758	0.769	-
	P	TCDM	0.408	0.433	0.75	0.91	0.93	0.891	0.804	0.807	13.525
	P+I	MM-PCQA	0.581	0.597	1.89	0.876	0.898	1.09	0.761	0.774	14.9
	I	DisPA	0.631	0.625	1.6	0.919	0.908	0.89	0.79	0.788	13
	P+I	ours	0.6877	0.7013	0.5708	0.9047	0.9319	0.8548	0.8165	0.8124	13.1788

TABLE 2. Ablations on WPC-PCQA database. Best in red.

Modal	SRCC ↑	PLCC ↑	RMSE ↓
w/o image modality	0.4575	0.4954	19.7359
w/o point cloud modality	0.7328	0.7344	15.26
w/o ACMF	0.773	0.7758	14.2499
MMFN-PCQA (ours)	0.8165	0.8124	13.1788

w/o image modality, the image block-level features are not fed into the network. Similarly, for the w/o point cloud modality, the LFSA-enhanced point cloud features are not utilized. From Table 2, we can see that each component of MMFN-PCQA benefits the final performance. Both point cloud modality and image modality make contributions to the model, and the ACMF further improve the performance, which validates the effectiveness of our fusion strategy.

Acknowledgements

This article is supported by the National Nature Science Foundation of China (Grant No: 61901116, Grant No: 61471229 and Grant No:62001115), the Guangdong Basic and Applied Basic Research Foundation (Grant No: 2023A1515140111, Grant No: 2019A1515010789, Grant No: 2021A1515012289 and Grant No: 2019A1515110136) and the Jihua Laboratory Scientific Project (Grant No: X210101UZ210).

References

[1] Wuyuan Xie, Yunheng Liu, Kaiming Wang and Miaohui Wang, "LLM-Guided Cross-Modal Point Cloud Quality Assessment: A Graph Learning Approach," in IEEE Signal Processing Letters, Vol 31, pp. 2250-2254, Aug. 2024.

[2] 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", Proceedings of the 32nd ACM International Conference on Multimedia, New York, pp. 7783-7792, October, 2024.

[3] Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu and Wenhan Zhu, "A No-Reference Visual

Quality Metric For 3D Color Meshes", 2021 IEEE ICMEW, Shenzhen, pp. 1-6, July, 2021.

[4] 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," in IEEE Transactions on Circuits and Systems for Video Technology, Vol 32, No. 11, pp. 7618-7631, Nov. 2022.

[5] 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, Vol 19, No. 80, pp. 1 - 26, Feb. 2023.

[6] G. Meynet, Y. Nehmé, J. Digne and G. Lavoué, "PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds," 2020 Twelfth International Conference on QoMEX, Athlone, Ireland, pp. 1-6, May, 2020.

[7] Qi Yang, Zhan Ma, Yiling Xu, Zhu Li and Jun Sun, "Inferring Point Cloud Quality via Graph Similarity," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 44, No. 6, pp. 3015-3029, June. 2022.

[8] E. Alexiou and T. Ebrahimi, "Towards a Point Cloud Structural Similarity Metric," 2020 IEEE ICMEW, London, UK, pp. 1-6, July, 2020.

[9] Qi Liu, Hui Yuan, Honglei Su, Hao Liu, Yu Wang and Huan Yang, "PQA-Net: Deep No Reference Point Cloud Quality Assessment via Multi-View Projection," in IEEE Transactions on Circuits and Systems for Video Technology, Vol 31, No. 12, pp. 4645-4660, Dec. 2021.

[10] Yu Fan, Zicheng Zhang, Wei Sun, Xiongkuo Min, Ning Liu and Quan Zhou, "A No-reference Quality Assessment Metric for Point Cloud Based on Captured Video Sequences," 2022 IEEE 24th International Workshop on MMSP, Shanghai, China, pp. 1-5, September, 2022.

[11] Zicheng Zhang, Wei Sun, Yucheng Zhu, Xiongkuo Min, Wei Wu and Ying Chen, "Evaluating Point Cloud From Moving Camera Videos: A No-Reference Metric," in IEEE Transactions on Multimedia, Vol. 27, pp. 927-939, Dec. 2025.

[12] 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", Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, pp.1759-1767, August, 2023.

[13] Charles R. Qi , Li Yi, Hao Su, and Leonidas J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space", Conference and Workshop on Neural Information Processing Systems, Long Beach, pp.5099-5108, December, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on CVPR, Las Vegas, NV, USA, pp. 770-778, June, 2016.

[15] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", MICCAI, Munich, pp.234-241, October, 2015.

[16] Wei Sun, Xiongkuo Min, Wei Lu and Guangtao Zhai, "A Deep Learning Based No-Reference Quality Assessment Model for UGC Videos", Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, pp.856–865, October, 2022.

[17] 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," in IEEE Transactions on Multimedia, Vol 23, pp. 3877-3891, Oct. 2021.

[18] Qi Liu, Honglei Su, Zhengfang Duanmu, Wentao Liu and Zhou Wang, "Perceptual Quality Assessment of Colored 3D Point Clouds," in IEEE Transactions on Visualization and Computer Graphics, Vol 29, No. 8, pp. 3642-3655, Aug. 2023.

[19] Eric M. Torlig, E. Alexiou, Tiago A. Fonseca, Ricardo L. de Queiroz, and T. Ebrahimi, "A novel methodology for quality assessment of voxelized point clouds", Proc. SPIE 10752, Applications of Digital Image Processing XLI, 107520I, San Diego, California, United States, September, 2018.

[20] Dong Tian, H. Ochimizu, Chen Feng, R. Cohen and A. Vetro, "Geometric distortion metrics for point

cloud compression," 2017 IEEE ICIP, Beijing, China, pp. 3460-3464, September, 2017.

[21] Qi Yang, Yipeng Liu, Siheng Chen, Yiling Xu, Jun Sun and Zhan Ma, "No-Reference Point Cloud Quality Assessment via Domain Adaptation", 2022 IEEE/CVF Conference on CVPR, New Orleans, LA, USA, pp. 21147-21156, September, 2022.

[22] Zhichao Chen, Shuyu Xiao, Yongfang Wang, Yihan Wang and Hongming Cai, "GLDBF: Global and local dual-branch fusion network for no-reference point cloud quality assessment", Displays, Vol 85, pp. 102882, Dec. 2024.

[23] Ziyu Shan, Qi Yang, Rui Ye, Yujie Zhang, Yiling Xu, Xiaozhong Xu and Shan Liu, "GPA-Net:No-Reference Point Cloud Quality Assessment With Multi-Task Graph Convolutional Network", IEEE Transactions on Visualization and Computer Graphics, Vo1 30, No. 8, pp. 4955-4967, Aug. 2024.

[24] Yujie Zhang, Qi Yang, Yifei Zhou, Xiaozhong Xu, Le Yang and Yiling Xu, "TCDM: Transformational Complexity Based Distortion Metric for Perceptual Point Cloud Quality Assessment," in IEEE Transactions on Visualization and Computer Graphics, Vol 30, No. 10, pp. 6707-6724, Oct. 2024.

[25] Ziyu Shan, Yujie Zhang, Yipeng Liu and Yiling Xu, "Learning Disentangled Representations for Perceptual Point Cloud Quality Assessment via Mutual Information Minimization", The Thirty-eighth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, September, 2024.