UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI

This is the Author's *accepted* manuscript version of the following contribution:

M. Hamidi, S. Porcu, A. Floris and L. Atzori, "MVAW-PCQA: A No-reference Point Cloud Quality Assessment via Multi-View Adaptive Weighting," *2025 17th International Conference on Quality of Multimedia Experience (QoMEX)*, Madrid, Spain, 2025, pp. 1-7.

© 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

The publisher's version is available at:

http://dx.doi.org/10.1109/QoMEX65720.2025.11219985

When citing, please refer to the published version.

This full text was downloaded from UNICA IRIS https://iris.unica.it/

MVAW-PCQA: A No-reference Point Cloud Quality Assessment via Multi-View Adaptive Weighting

MohammadAli Hamidi[1,2], Simone Porcu[1,2], Alessandro Floris[1,2], and Luigi Atzori[1,2]
[1]DIEE, University of Cagliari, 09123 Cagliari, Italy
[2]CNIT, University of Cagliari, 09123 Cagliari, Italy
{mohammadali.hamidi, simone.porcu, alessandro.floris84, l.atzori}@unica.it

***Abstract*—Point cloud quality assessment (PCQA) is a critical research area focused on evaluating the perceptual Quality of Experience (QoE) of point clouds to enhance visual experiences of immersive multimedia applications for end users. To prevent the complex computations on 3D data applied by model-based methods, projection-based models have been developed to estimate the QoE by analysing 2D projection views of the point cloud. In this paper, we propose a novel projection-based No-Reference (NR) PCQA method, called Multi-View Adaptive Weighting Point Cloud Quality Assessment (MVAW-PCQA), to predict the QoE of distorted point clouds using six 2D projection views as the input of a convolutional neural network (CNN) architecture. First, multi-view involves independently extracting features from multiple projection views of a point cloud, guaranteeing view-specific features are learned without prematurely mixing spatial information, and preserving the unique contributions of each projection view to the final quality prediction. Then, an adaptive weighting fusion mechanism combines the features extracted from the different projection views by learning their relative importance. This design enables the model to focus on the most informative projections for predicting the point cloud quality. The experimental results demonstrate that our method outperforms state-of-the-art NR-PCQA methods on the SJTU-PCQA dataset in terms of root mean square error (RMSE) and correlation coefficients (Pearson, Spearman, and Kendall), while adopting a lightweight design with a reasonable number of parameters for the trained neural network.**

***Index Terms*—Point Cloud Quality Assessment, Quality of Experience, Projection-based model, No-reference quality model.**

I. Introduction

Immersive technologies are increasingly integrated into everyday applications, enhancing user experiences in domains such as video conferencing, live entertainment, and virtual events. The growing accessibility of extended reality (XR) headsets has allowed a broader audience to interact with immersive content in more natural and compelling ways. A key enabler of such experiences is the use of point clouds, which allow for realistic 3D representations of dynamic scenes and human subjects. For example, 3D reconstructions of performers, such as singers during concerts, can be pre-recorded and streamed as volumetric video, enabling users to experience events in an interactive, spatially immersive manner.

To stream immersive content efficiently through the network, protocols like Dynamic Adaptive Streaming over HTTP (DASH) have been extended to support point cloud sequences [1]. These protocols can handle the processing, encoding, and adaptive streaming of point cloud data over the network to client devices, where the content is decoded and rendered in real time for immersive consumption. In this context, the objective point cloud quality assessment (PCQA) becomes crucial to drive rate adaptation strategies, help select the optimal representation for transmission, and ultimately ensure a satisfactory Quality of Experience (QoE) for the end user.

Objective QoE PCQA models can be broadly categorized into three types: Full Reference (FR), Reduced Reference (RR), and No Reference (NR) [2]. FR models require access to both the original and distorted point cloud; RR models rely on specific features extracted from both the original and distorted point cloud; NR models, on the other hand, operate solely on the received (distorted) point cloud. Since NR models do not require any information from the original point cloud to assess the quality of the distorted one, these methods are the most lightweight and practical in real-time PCQA applications, reducing computational load and bandwidth requirements across the streaming pipeline.

Moreover, in point cloud streaming, the content itself poses unique challenges. Unlike traditional 2D media, point clouds are significantly more complex to handle due to their large size, sparse and irregular structure, and high sensitivity to distortions introduced during acquisition, compression, or transmission. These issues are particularly critical in streaming pipelines based on DASH, where geometry-based encoding formats (e.g., G-PCC [3]) significantly increase the bitrate and decoding complexity. As a result, model-based PCQA approaches, which rely on full 3D data to assess the point cloud quality, end up inheriting the same processing and bandwidth overheads they are meant to mitigate. Although some of these methods have demonstrated strong performance in assessing the perceived quality of point clouds, many of

This work has been partially supported by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP C29J24000300004, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART"), by the European Union under the Italian NRRP of NextGenerationEU, "Sustainable Mobility Center" Centro Nazionale per la Mobilità Sostenibile, CNMS, CN_00000023), and by Italian NRRP - M4C1 - Inv. 3.4 and M4C1 - Inv. 4.1, Ministerial Decree no. 351/2022.

This work has also been partially funded by the European Union (SPIRIT, 101070672). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The SPIRIT project has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

them suffer from high computational complexity due to the use of deep 3D convolutions [4], graph operations [5], or multimodal fusion techniques [6]. Additionally, they often require elaborate preprocessing pipelines to align and normalize different input modalities, which can hinder practical deployment in real-time or resource-constrained environments.

To address these limitations, we propose the multi-view Adaptive Weighting Point Cloud Quality Assessment (MVAW-PCQA) model, a novel NR-PCQA objective model that relies exclusively on 2D projection views, eliminating the need for direct processing of 3D point cloud data. Our approach employs a pre-trained Convolutional Neural Network (CNN) to process six orthogonal projection views of the point cloud through a multi-view strategy. Multi-view means the network processes multiple PVs belonging to a point cloud independently to ensure consistent feature distributions, guaranteeing the network learns view-specific features without prematurely mixing spatial information. To enhance the representational power of the model, we introduce an adaptive weighting fusion mechanism that combines features extracted from the different projection views by learning their relative importance. This design enables the model to focus on the most informative projections. The resulting model offers a new PCQA solution, outperforming state-of-the-art PCQA methods on the SJTU-PCQA dataset. Moreover, the proposed approach maintains a compact design with a reasonable number of parameters for the trained neural network.

The paper is structured as follows. Section II discusses the related work in PCQA. In Section III, we present the proposed method, whereas Section IV evaluates and compares the model performance with state-of-the-art solutions. Finally, Section V concludes the paper.

II. Related Work

Objective PCQA methods are based on mathematical algorithms to estimate the quality of distorted point clouds and can be categorized into full-reference (FR), reduced-reference (RR), and no-reference (NR) according to the extent of the available original point cloud data. FR methods, such as GraphSIM, PointSSIM, and PCQM, need the complete original point cloud to estimate the quality of the distorted one. The GraphSIM method [7] computes the quality of the distorted point cloud as a function of geometry and color distortion using the graph signal gradient. PointSSIM [8] applies a 3D adaptation of the well-known image quality metric Structural Similarity (SSIM) index to point clouds, whereas PCQM [9] is based on a weighted linear combination of geometry- and colour-based features to estimate the quality of compressed PCs. RR approaches extract specific features from both the original and distorted point clouds, but are very limited in the literature [10].

NR models are preferred to FR and RR models because they do not require any information from the original point cloud, which makes them the most lightweight and practical in real-time applications. Indeed, several NR PCQA models have been proposed in the literature that can be broadly categorized according to the type of input data they process: 2D projections (projection-based), 3D geometry (model-based), or a combination of both (hybrid) [2]. Projection-based models have gained popularity due to their alignment with the MPEG V-PCC compression standard [11], which converts point clouds into sequences of 2D patches representing geometry and texture. The IT-PCQA model [12] explores the relationship between natural images and 3D projections using hierarchical CNN features combined with adversarial domain adaptation. PQA-Net [13] adopts a multi-view projection strategy and processes the resulting images through a CNN-based architecture that includes both distortion classification and quality regression. Zhang et al. [14] propose a dynamic view capture framework, where rotating camera paths generate multiple projections analyzed with both 2D-CNN and 3D-CNN networks. A more recent approach, MS-PCQE [15], leverages projection scale diversity and introduces a dual-branch Vision Transformer architecture, combining focal-length-aware feature interaction with mask-aware attention mechanisms to enhance quality prediction.

Other projection-based methods move away from end-to-end deep learning and focus instead on regression-based quality estimation. Van Damme et al. [16] propose a linear regression model fitted with handcrafted NR metrics, combined with a sigmoid mapping that is adjusted per content class. Weil et al. [17] introduce a Gradient Boost regressor that estimates perceived quality based on compression parameters, frame rate, and viewing distance. Nguyen et al. [18] adapt the ITU-T P.1203 model to the context of dynamic point cloud streaming by tuning its coefficients based on subjective evaluations. In addition, the bitstreamPCQ model [19] bypasses the need for decoded projections by analytically modeling the effects of texture and geometry quantization using encoding parameters only.

Model-based NR PCQA models operate directly on the 3D point cloud data. The 3D-NSS model [20] extracts geometric and color features such as curvature, anisotropy, and LAB color values, applying statistical descriptors and training a support vector regressor. GPA-Net [5] introduces a graph-based convolutional kernel, GPAConv, capable of capturing structure-aware distortions, followed by a multi-task decoder that estimates both distortion type and severity. ResSCNN [4] proposes a sparse convolutional architecture that processes geometry and color features encoded as sparse tensors, enabling hierarchical feature extraction. GQI [21] relies on CNNs trained on local patches, integrating geometric distances, curvature, and grayscale information to compute a global quality score.

A notable example of a hybrid architecture is MM-PCQA [6], which combines 3D and 2D modalities. This model extracts texture features from image projections and geometry features from point cloud segments, merging them via a symmetric cross-modal attention module. The design captures complementary information from both domains, enhancing prediction accuracy at the cost of increased computational complexity. Another noteworthy hybrid architecture is Plain-

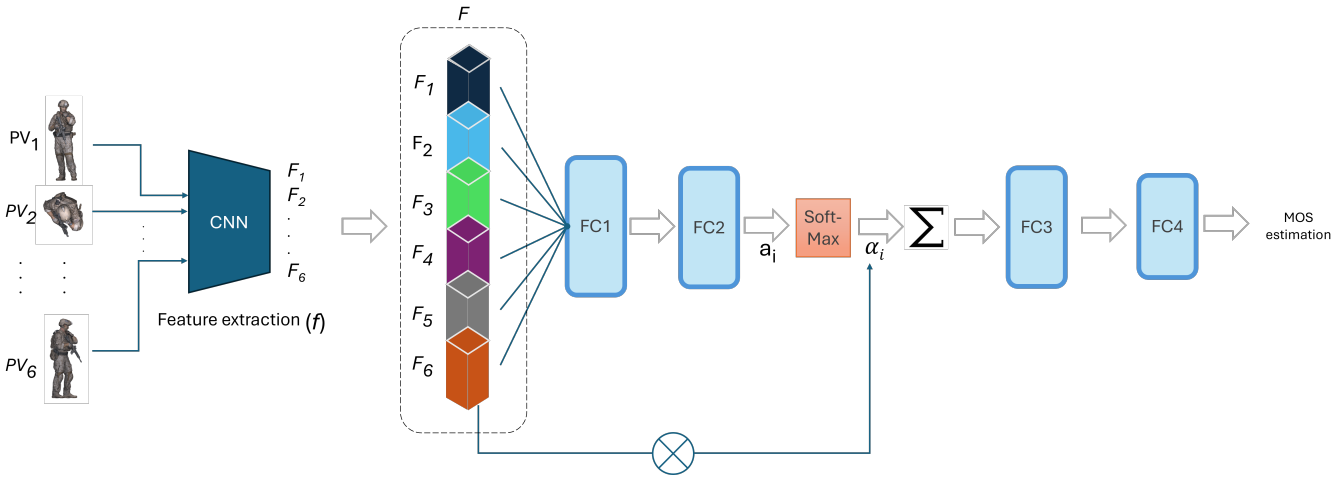

Fig. 1. The architecture of the proposed MVAW-PCQA method.

PCQA [22], a deep learning-based model that jointly analyzes visual and geometric attributes of point clouds. The architecture features two parallel branches: a no-reference branch that extracts perceptual features from visual components, and a degraded-reference branch that incorporates geometric information derived from a simplified reference. This design allows the model to adapt to different levels of reference availability while capturing complementary aspects of visual and structural quality.

While model-based and hybrid methods generally achieve high accuracy, they require full 3D decoding and involve significant computational overhead, which may hinder real-time applications. In contrast, projection-based approaches, particularly those aligned with V-PCC encoding and streaming via protocols such as DASH, provide a more efficient solution. By operating on 2D projections already available in the rendering pipeline, these models reduce the need for complex preprocessing or heavy resource consumption. Therefore, the model presented in this work is designed to address these limitations, introducing a new architecture that processes six orthogonal projections of the point cloud through a convolutional backbone to extract representative features. These features are then combined using an adaptive weighting mechanism that emphasizes the most informative views for quality prediction. This projection-based strategy eliminates the need for geometry reconstruction or multimodal integration, while maintaining architectural efficiency.

III. Proposed Approach

We propose a projection-based NR-PCQA method, called MVAW-PCQA, to predict the QoE of distorted point clouds using only their 2D projection views as the input. MVAW-PCQA is based on a deep learning architecture (shown in Fig. 1) that extracts features from each projection view representing the point cloud and applies an adaptive weighting mechanism to give more importance to features extracted from the most significant projection views, i.e., those including the most relevant information for predicting the point cloud quality.

A. MVAW-PCQA Architecture

Let's consider a colored 3D point cloud $\mathcal{PC}$, composed of a set of N points,

$$\mathcal{PC} = \{p_n = (x_n, y_n, z_n) \in \mathbb{R}^3 \mid n = 1, 2, \ldots, N\}, \quad (1)$$

where (x_n, y_n, z_n) are the 3D coordinates of each point p_n. We generate six orthogonal 2D projection views (*PVs*), denoted as $\{PV_1, PV_2, \ldots, PV_6\}$, each corresponding to a front, back, left, right, top, and bottom perspective of the 3D $\mathcal{PC}$, respectively, as illustrated in Fig. 2. These *PVs* preserve both the visual texture and color information of the $\mathcal{PC}$, providing a rich 2D representation of the 3D structure.

The architecture of the proposed method is shown in Fig. 1, whose inputs are the six *PVs* obtained from the 3D $\mathcal{PC}$. First, we utilized a CNN backbone for extracting features from the six *PVs*. The CNN was implemented using the ConvNeXt-T neural network [23] due to its strong trade-off between accuracy and computational efficiency, delivering high performance with low inference cost. Pre-trained on ImageNet, ConvNeXt-T was adapted to the PCQA task by replacing its final classification layer with a fully connected layer that outputs a 256-dimensional latent representation.

Let the CNN-based feature extraction function be denoted as:

$$F_i = f(PV_i), \quad i \in \{1, 2, \ldots, P\}, \quad F_i \in \mathbb{R}^D, \quad (2)$$

where PV_i represents the i-th projection view, and the scalar $P = 6$ denotes the total number of projection views extracted from each 3D point cloud. In our implementation, each *PV* is processed independently to obtain a feature vector $F_i \in \mathbb{R}^D$, where $D = 256$. All feature vectors F_i are then vertically stacked to form the global feature matrix $F = [F_1, F_2, F_3, F_4, F_5, F_6]^T \in \mathbb{R}^{P \times D}$.

To avoid common fusion strategies [24]–[26] based on uniform weighting or heuristic rules, which may lead to

suboptimal performance due to the varying importance of each projection, we propose a learnable adaptive weighting mechanism as follows:

- We process each *PV* using a two-layer Multi-Layer Perceptron (MLP) with ReLU activation:

$$a_i = W_2\,\sigma(W_1 F_i + b_1) + b_2, \quad i = \{1, \ldots, P\}, \tag{3}$$

where:
  - $W_1 \in \mathbb{R}^{256\times128}$ and $b_1 \in \mathbb{R}^{128}$ are trainable parameters of the first linear layer FC1;
  - σ is the ReLU activation function;
  - $W_2 \in \mathbb{R}^{128\times1}$ and $b_2 \in \mathbb{R}^{1}$ are trainable parameters of the second linear layer FC2.
- To get an importance score of the different vectors F_i, a softmax function is applied along the projection dimension:

$$\alpha_i = \frac{\exp(a_i)}{\sum_{j=1}^{P}\exp(a_j)}, \quad \text{such that} \sum_{i=1}^{P}\alpha_i = 1. \tag{4}$$

The α_i weights emphasize the most informative projections while downweighting the least relevant ones. The fused feature representation is then computed as a weighted sum:

$$F_{fused} = \sum_{i=1}^{P}\alpha_i F_i, \quad F_{fused} \in \mathbb{R}^{D}, \tag{5}$$

which highlights the most representative feature vector F_i and reduces the influence of the least representative feature vectors.

Finally, the fused feature vector F_{fused} is passed through another two-layer MLP with ReLU activation to provide the predicted MOS:

$$MOS_p = W_4\,\sigma(W_3 F_{fused} + b_3) + b_4, \tag{6}$$

where:

- $W_3 \in \mathbb{R}^{256\times128}$ and $b_3 \in \mathbb{R}^{128}$ are trainable parameters of the first linear layer FC3;
- σ is the ReLU activation function;
- $W_4 \in \mathbb{R}^{128\times1}$ and $b_4 \in \mathbb{R}^{1}$ are trainable parameters of the second linear layer FC4.

IV. Experimental Results

To investigate the QoE prediction performance of the proposed MVAW-PCQA method, we have conducted experiments on the SJTU-PCQA dataset and compared the achieved performance with that of state-of-the-art methods.

A. Dataset

The SJTU-PCQA dataset [27] is a large-scale PCQA benchmark dataset comprising nine distinct PCs with varying content characteristics, each subjected to seven types of distortions across six different distortion levels. The considered distortions include Octree-based compression, color noise, downscaling, geometry Gaussian noise, downscaling combined with color noise, downscaling combined with geometry Gaussian noise, and color noise combined with geometry Gaussian noise.

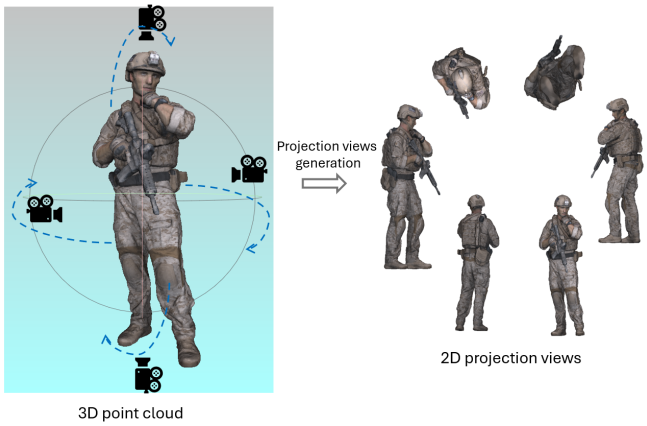

Fig. 2. Generation of the six 2D projection views (front, back, left, right, top, and bottom) from the 3D point cloud.

To evaluate subjective quality perception, the authors conducted a comprehensive user study, involving 64 participants aged 18 to 30 years, who provided quality scores on a 10-point scale, mapped to five quality categories: 1–2 (Bad), 3–4 (Poor), 5–6 (Fair), 7–8 (Good), and 9–10 (Excellent) accorgid to the ITU-R Recommendation BT.500-11 [28]. Thus, the dataset provides a collection of MOS for 378 different samples (9 contents × 7 distortions × 6 levels).

B. Implementation details

The proposed approach was implemented on a workstation featuring an Intel I9-14900K CPU, an NVIDIA RTX 4090 GPU, and 64 GB of RAM. The model was developed using the PyTorch deep learning framework. The initial learning rate of the CNN backbone layers and the regression layer were set to 5×10^{-5} and 5×10^{-4}, respectively. We observed that applying a lower learning rate to the pre-trained backbone facilitated more stable optimization and faster convergence using the ADAM optimizer [29]. The batch size was set to 4, and training followed a 15-epoch linear warm-up strategy, with a maximum of 200 training epochs. The final model contains approximately 29 million parameters.

For optimization, we employed the Adam optimizer with a weight decay of 1×10^{-4} in every 8 epochs of training for all learning rates. To improve generalization, we applied data augmentation techniques, including random cropping (224 × 224 × 3) and random (90°) rotation, to introduce data variations during training.

K-fold cross-validation was used as a key method for evaluating model performance. Following [30], K was set to 9 for the SJTU-PCQA dataset, ensuring that 8 PCs, each encompassing a combination of applied distortions, were used for training, while the remaining one was reserved for validation. The MVAW-PCQA was trained to minimize the Mean Squared Error (MSE) between the predicted and ground truth MOS values.

Finally, as in [15], we adopted the following logistic function to eliminate the nonlinearity,

$$Q_p = \frac{\lambda_1 - \lambda_2}{1 + \exp\left(-\frac{Q_f - \lambda_3}{|\lambda_4|}\right)} + \lambda_2, \quad (7)$$

where Q_p and Q_f represent the mapped and predicted MOS scores, respectively, and λ_i (with $i = 1, 2, 3, 4$) are parameters optimized via curve fitting.

C. Comparing with the State-of-the-art

Table I presents the evaluation results of the proposed MVAW-PCQA on the SJTU-PCQA dataset compared to state-of-the-art FR and NR PCQA approaches in terms of four key performance metrics, including Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), Kendall Rank Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE). Notably, our proposed approach surpasses all FR and NR state-of-the-art approaches across all performance metrics. In particular, a relevant performance improvement is observed over the top-performing (MS-PCQE) in NR-PCQA methods with higher SRCC (0.9306 vs. 0.9180) and PLCC (0.9466 vs. 0.9326), as well as surpassing MM-PCQA in terms of higher KRCC (0.7947 vs. 0.7838) and a lower RMSE (0.7219 vs. 0.7716).

Moreover, Table II compares some of the PCQA methods in terms of the number of parameters utilized by the implemented neural networks adopted in the prediction models. It can be seen that the proposed MVAW-PCQA model has more parameters (29M) than some PCQA methods, including PQA-Net (0.29M), ResSCNN (1.23M), and MS-PCQE (14.27M). However, MVAW-PCQA has a number of parameters comparable to that of Plain-PCQA (28.5M) and significantly lower than MM-PCQA (58.37M). Despite its relatively high parameter count, MVAW-PCQA achieves superior performance across all evaluation metrics on the SJTU-PCQA dataset, outperforming all existing methods in terms of SRCC, PLCC, KRCC, and RMSE. This demonstrates that the MVAW-PCQA architecture effectively leverages its higher capacity to capture perceptual quality cues, resulting in more accurate and reliable predictions.

Moreover, Table III shows a complementary experiment with a computational analysis to assess the practical feasibility of the proposed framework under real-world deployment constraints. We report the average inference time per fold, peak GPU memory usage, and floating-point operations (FLOPs) per sample for the SJTU-PCQA dataset using the ConvNeXt-T backbone. For the SJTU-PCQA dataset, the model achieves an average total inference time of 1.74 seconds for all samples per fold, with 0.2 GB peak GPU memory usage, and approximately 26.82 GFLOPs per sample. These values suggest the model is computationally efficient and lightweight, making it suitable for edge deployment scenarios.

Finally, as reported in [23], although MVAW-PCQA adopts a relatively large model in terms of parameter count, it is built upon the ConvNeXt-T architecture, which has been specifically optimized for computational efficiency. Unlike transformer-based models that rely on attention mechanisms with irregular memory access patterns [31], [32], ConvNeXt-T is composed entirely of convolutional operations that are highly parallelizable and benefit from efficient GPU implementation, resulting in lower FLOPs [23] and faster inference, even when the total parameter count exceeds that of other neural network. Consequently, MVAW-PCQA inherits these advantages, making it both accurate and practical for deployment in resource-constrained environments. These results validate the framework as a high-capacity yet efficient model that achieves state-of-the-art performance while maintaining a favorable balance between accuracy and computational cost.

TABLE I
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART APPROACHES ON THE SJTU-PCQA DATASET.

Type	Approaches	SRCC	PLCC	KRCC	RMSE
FR	GraphSIM [7]	0.8783	0.8449	0.6947	1.0321
	M-p2po [33]	0.7294	0.8123	0.5617	1.3613
	HD-p2po [34]	0.7157	0.7753	0.5447	1.4475
	M-p2pl [35]	0.6277	0.5940	0.4825	2.2815
	PB-PCQA [27]	0.6020	0.6076	-	1.8635
	PSNR-yuv [36]	0.7950	0.8170	0.6196	1.3151
	PCQM [9]	0.8644	0.8853	0.7086	1.0862
	MS-SSIM [37]	0.6888	0.4082	0.4995	2.2154
	PointSSIM [8]	0.6867	0.7136	0.4964	1.7001
NR	3D-NSS [20]	0.7144	0.7382	0.5174	1.7686
	ResSCNN [4]	0.8590	0.8931	0.6812	1.0373
	PQA-net [13]	0.8372	0.8586	0.6304	1.0719
	GPA-Net [5]	0.8750	0.8860	-	-
	Plain-PCQA [22]	0.9133	0.9302	0.7603	0.8607
	MM-PCQA [6]	0.9103	0.9226	0.7838	0.7716
	MS-PCQE [15]	0.9180	0.9326	0.7740	0.8241
	MVAW-PCQA	**0.9306**	**0.9466**	**0.7947**	**0.7219**

TABLE II
COMPARISON OF THE NUMBER OF PARAMETERS IN PCQA NEURAL NETWORKS.

Method	Number of params
PQA-Net [13]	0.29 M
ResSCNN [4]	1.23 M
MS-PCQE [15]	14.27 M
Plain-PCQA [22]	28.50 M
MM-PCQA [6]	58.37 M
MVAW-PCQA	29 M

V. CONCLUSION

This paper presents a novel projection-based NR PCQA method, MVAW-PCQA, which relies solely on 2D projection views to estimate the perceptual quality of point clouds. A pre-trained CNN backbone is employed as a feature extractor, processing six projection views of the point cloud using the multi-view approach to extract view-specific features, enhancing spatial consistency across projection views. Furthermore, an adaptive weighting fusion mechanism is introduced to dynamically assign importance weights to each view, highlighting the most informative projections while down-weighting the less relevant ones. This leads to a more comprehensive and discriminative feature representation. The results demonstrated that our proposed approach could outperform all FR and NR

TABLE III
AVERAGE INFERENCE COST PER FOLD FOR SJTU-PCQA DATASET.

Inference Time (s)	GPU Memory (GB)	FLOPs (G)
1.74	0.2	26.82

state-of-the-art approaches across all evaluation metrics on the SJTU-PCQA dataset, including RMSE, PLCC, SRCC, and KRCC. In particular, relevant performance enhancements were observed for the MOS prediction, with a reduction in RMSE by 0.05, and increases in PLCC by 0.014, SRCC by 0.012, and KRCC by 0.01. As part of future work, we plan to evaluate our method on additional benchmark datasets, such as WPS [38], WPC2.0 [39], and SIAT-PCQD [40].

The achieved performance highlights the potential of MVAW-PCQA to benefit XR applications, particularly in scenarios where selecting and streaming the most appropriate point cloud quality is essential to ensuring an optimal QoE for end users. In particular, MVAW-PCQA can be effectively applied to guide adaptive bitrate decisions during point cloud streaming, enabling perceptually optimized delivery in real-time immersive environments.

REFERENCES

[1] M. Hosseini and C. Timmerer, "Dynamic Adaptive Point Cloud Streaming," in *Proceedings of the 23rd Packet Video Workshop*, ser. PV '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 25–30.

[2] S. Porcu, C. Marche, and A. Floris, "No-Reference Objective Quality Metrics for 3D Point Clouds: A Review," *Sensors*, vol. 24, no. 22, 2024.

[3] O. Nakagami and P. G-PCC, "PCC WD G-PCC (Geometry-Based PCC)," *ISO/IEC JTC1/SC29/WG11 MPEG, Standard*, no. 17771, 2018.

[4] Y. Liu, Q. Yang, Y. Xu, and L. Yang, "Point Cloud Quality Assessment: Dataset Construction and Learning-based No-reference Metric," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 2s, Feb. 2023.

[5] Z. Shan, Q. Yang, R. Ye, Y. Zhang, Y. Xu, X. Xu, and S. Liu, "GPA-Net:No-Reference Point Cloud Quality Assessment With Multi-Task Graph Convolutional Network," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 8, pp. 4955–4967, 2024.

[6] Z. Zhang, W. Sun, X. Min, Q. Zhou, J. He, Q. Wang, and G. Zhai, "MM-PCQA: Multi-Modal Learning for No-reference Point Cloud Quality Assessment," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, 2023.

[7] 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*, vol. 44, no. 6, pp. 3015–3029, 2022.

[8] 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.

[9] 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.

[10] 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.

[11] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis, and V. Zakharchenko, "Emerging MPEG Standards for Point Cloud Compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2019.

[12] 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.

[13] 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.

[14] Z. Zhang, W. Sun, Y. Zhu, X. Min, W. Wu, Y. Chen, and G. Zhai, "Evaluating Point Cloud from Moving Camera Videos: A No-Reference Metric," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[15] X. Chai and F. Shao, "MS-PCQE: Efficient No-Reference Point Cloud Quality Evaluation via Multi-Scale Interaction Module in Immersive Communications," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2024.

[16] S. V. Damme, M. T. Vega, J. van der Hooft, and F. D. Turck, "Clustering-based Psychometric No-Reference Quality Model for Point Cloud Video," in *Proceedings - International Conference on Image Processing, ICIP*. IEEE Computer Society, 2022, pp. 1866–1870.

[17] J. Weil, Y. Alkhalili, A. Tahir, T. Gruczyk, T. Meuser, M. Mu, H. Koeppl, and A. Mauthe, "Modeling Quality of Experience for Compressed Point Cloud Sequences based on a Subjective Study," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, 2023, pp. 135–140.

[18] M. Nguyen, S. Vats, and H. Hellwagner, "No-Reference Quality of Experience Model for Dynamic Point Clouds in Augmented Reality," in *MHV*. Association for Computing Machinery (ACM), 2 2024, pp. 90–91.

[19] Q. Liu, H. Su, T. Chen, H. Yuan, and R. Hamzaoui, "No-Reference Bitstream-Layer Model for Perceptual Quality Assessment of V-PCC Encoded Point Clouds," *IEEE Transactions on Multimedia*, vol. 25, pp. 4533–4546, 2023.

[20] Z. Zhang, W. Sun, X. Min, W. Zhu, T. Wang, W. Lu, and G. Zhai, "A No-Reference Evaluation Metric for Low-Light Image Enhancement," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[21] 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 and Expo Workshops (ICMEW)*, 2021, pp. 1–6.

[22] X. Chai, F. Shao, B. Mu, H. Chen, Q. Jiang, and Y.-S. Ho, "Plain-PCQA: No-Reference Point Cloud Quality Assessment by Analysis of Plain Visual and Geometrical Components," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6207–6223, 2024.

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[24] S. A. Golestaneh and K. Kitani, "No-reference image quality assessment via feature fusion and multi-task learning," *arXiv preprint arXiv:2006.03783*, 2020.

[25] W.-x. Tao, G.-y. Jiang, Z.-d. Jiang, and M. Yu, "Point cloud projection and multi-scale feature fusion network based blind quality assessment for colored point clouds," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 5266–5272.

[26] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on image processing*, vol. 27, no. 1, pp. 206–219, 2017.

[27] 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*, vol. 23, pp. 3877–3891, 2021.

[28] ITU, "Methodology for the Subjective Assessment of the Quality of Television Pictures ITU-R Recommendation BT.500-11, Tech. Rep." 2000.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] Y. Fan, Z. Zhang, W. Sun, X. Min, N. Liu, Q. Zhou, J. He, Q. Wang, and G. Zhai, "A no-reference quality assessment metric for point cloud based on captured video sequences," in *2022 IEEE 24th international workshop on Multimedia signal processing (MMSP)*. IEEE, 2022, pp. 1–5.

[31] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Advances in neural information processing systems*, vol. 32, 2019.

[32] S. Saha and L. Xu, "Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies," *arXiv preprint arXiv:2503.02891*, 2025.

[33] R. Mekuria, Z. Li, C. Tulvan, and P. A. Chou, "Evaluation criteria for pcc (point cloud compression)," 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:57014333
[34] 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, 2017.
[35] 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.
[36] 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 *Optical Engineering + Applications*, 2018.
[37] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, 2003, pp. 1398–1402 Vol.2.
[38] 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*, pp. 1–1, 2022.
[39] Q. Liu, H. Yuan, R. Hamzaoui, H. Su, J. Hou, and H. Yang, "Reduced reference perceptual quality model with application to rate control for video-based point cloud compression," *IEEE Transactions on Image Processing*, 2021.
[40] X. Wu, Y. Zhang, C. Fan, J. Hou, and S. Kwong, "Subjective Quality Database and Objective Study of Compressed Point Clouds With 6DoF Head-Mounted Display," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4630–4644, 2021.