MULTI-MODAL EVALUATION OF 3D POINT CLOUDS IMAGES: A NOVEL NO-REFERENCE APPROACH USING A MULTI-STREAM ATTENTIVE ARCHITECTURE

Marouane Tliba[1]*, Aladine Chetouani*[1]*, Giuseppe Valenzise*[2] *and Fréderic Dufaux*[2]

[1]Laboratoire PRISME, Université d'Orléans, Orléans, France
[2]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes

ABSTRACT

Most existing Point Cloud Quality Assessment (PCQA) methods do not consider the local structures among points, which can impact the overall perceived quality. In this paper, we introduce a novel and efficient no-reference objective metric for PCQA that takes into account the intrinsic feature affinities of points using a fully attention-based network, which results in extracting relevant information to the local structures of the 3D content. In addition, we employ information from two modalities: a suitable 2D projection of the PC and a relevant subset of the native 3D point cloud data. The rationale is that each modality may be more sensitive to different distortion types and thus contribute to the overall quality assessment. To evaluate the performance of our method, we conducted experiments on well-known 3D Point Clouds Quality Assessment benchmarks for PC compression. Our results demonstrate that our multi-modal attention-based PCQA metric is competitive with state-of-the-art methods in terms of both effectiveness and reliability. In particular, our method is able to capture local structures and provide more accurate quality assessments, even better than most full-reference metrics, with a moderate computational cost.

***Index Terms*—** 3D Point Clouds, Image Quality Assessment, Graph Neural Network, Deep Learning.

1. INTRODUCTION

Over the past few years, there has been a rapid development in computer graphics technologies, leading to promising immersive applications with a higher level of real-world depiction. As a result, 3D point clouds have become increasingly important in various applications, ranging from AR/VR, autonomous driving, to telepresence. Point clouds are a set of points in 3D space.Working with point clouds can also present challenges due to the scattered irregularity of points in 3D space[1].

However, due to the large amount of information required to represent a 3D scene accurately, point clouds can consist of thousands or even millions of points. Thus, lossy compression schemes are often utilized to reduce their size for practical purposes. This has made point cloud quality assessment (PCQA) increasingly important. Depending on the availability of the reference image, objective quality metrics can be grouped into three categories: Full-Reference, Reduced-Reference, and No-Reference. Full-Reference (FR) methods [2, 3, 4, 5] require access to the original point cloud in order to compare it to the distorted point cloud and calculate a quality score. Among FR methods, feature-based PC quality metrics extract the geometry with the associated attributes from point-wise level in a global or local way. Examples include PC-MSDM [6] that extends the 2D SSIM metric [7] to PC by considering local curvature statistics, the Geotex [8] metric that exploits the Local Binary Pattern (LBP) [9] descriptors, PCQM [10] that combines the geometry and color features, and GraphSIM [11] that extracts key-points at high frequencies that are sensitive to the perception, and construct local graphs for local significance similarity feature computation. In projection-based PC Quality Metrics, the 3D points or their associated features are projected into 2D regular grids [12, 13]. No-Reference (NR) methods, such as the one proposed in this paper, do not require any information from the original point cloud and only use the distorted version for evaluation. These methods are well-suited for real-time applications because they do not require access to the original data. However, their quality assessment may be less accurate compared to Full-Reference and Reduced-Reference methods.

Our proposed method for no-reference point cloud quality assessment addresses a significant limitation in previous approaches, which often lack the ability to effectively consider the complex structural relationships within native 3D point cloud regions[14, 1]. By utilizing a multi-level self and cross-attention, we capture the local semantic affinities of points, providing a more comprehensive representation for the evaluation of its visual quality. Moreover, inspired by the multimodal nature of 3D PCs — either in fully 3D environment or using a series of fixed 2D projected screens, we extend our approach to leverage this multi-modal information. Specifically, we augment our method by using the a single representative view of the 3D point cloud from the front region, i.e., the one with the maximum semantic information. This provides our network with complementary information to infer a better connection of the distributed 3D irregular representations. Also, 2D images can be more sensitive to noise than 3D data, e.g., surface holes caused by compression. Therefore, capturing this kind of noise in the 2D domain is easier and can increase the accuracy of the quality metric.

The contributions of this paper are summarized as follows:

1. We present an efficient end-to-end method for PC quality assessment integrating multimodal information. Our method operates directly on the 3D point cloud features, and 2D projections. This enables a more empirical and comprehensive evaluation of the point cloud's quality.
2. Our method is designed to capture the local semantic affinities in point representations and creates connectivity between distant features using self-attention. This facilitates better feature extraction from local regions.
3. We aggregate the representation from multiple modalities, 3D geometry and color features in point clouds, as well as the complementary projected view by interpreting the cross-correlation of deep features as an efficient way to find information correspondence.

2. PROPOSED METHOD

In this section, we present our proposed design, and how it can serve as an efficient and robust base feature extraction network for no-reference point cloud quality assessment. Our network is novel in two key aspects: First, it utilizes an attention mechanism to establish soft, learnable links between the representation of 3D points at multiple levels of the 3D architecture. In particular, these links are created using self-attention, and combining the geometry and color features space through cross-attention. Second, it employs multi-modal learning to exploit the sensitivity of different modalities to specific distortions and to improve correspondence between distant patches within a region. The 2D projection serves as a complement to the 3D method by filling in any missing information and better mapping affinities between distributed regions. As depicted in Fig. 1, the overall pipeline of our method consists of three main steps: 3D pre-processing, features extraction, multi-modal fusion and quality estimation.

2.1. Pre-processing

Pre-processing is a critical stage in our approach. It consists in dividing the PC into vertical slices (partitions) of points. This partitioning has two primary objectives: (i) enable parallel processing of points, and (ii) meet the memory requirements for GPU processing, as some PCs may have an important size (e.g millions of points). In order to balance the computation load, the number of partitions varies between 8 and 24 according to the size of the original PC. Each partition is then divided to form local patches. To achieve this, we first select a set of non-overlapping centroids and then apply the k-nearest neighbor clustering method around them. In order to reduce the memory print, we selected only half of the ensemble of patches for each partition. We note here that, unlike other methods, we avoid applying a sampling on the point cloud local regions. This is to preserve local information related to the sensitivity of our downstream-task. Each patch is then fed into the 3D model for feature extraction and representation modeling. At the same time, we select the most representative front view of the projected point cloud to pass it to the 2D network.

2.2. Multi-Modal Quality Estimation

Feature extraction

Our model design draws inspiration from three existing architectures: PointNet [15] for PCs processing, vision transformer [16] for image classification, and transformer for language modeling [17].

First, we employ a principal characteristic of PointNet in extracting information from point sets by using a permutation-invariant function. To do so, we use a network with two parallel streams for color and geometry independently. This network transforms the input point geometry and color information into a higher-dimensional space using a *feature embedding* layer. To capture richer local structure information, we build on the *feature embedding* of the geometry stream by introducing multi-head *self*-attention, which draws the semantic affinities between neighboring points.

Second, to combine the geometry and the color representation, we employed multi-head *cross*-attention. This results not only in adding the local connectivity information between adjacent point representations, but also completing it with corresponding color features. The resulting attentive connectivity of points' representation is updated dynamically at each network level, capturing different levels of semantic structure, as the point embedding is updated. Our method thus combines the strengths of both PointNet and Transformer while introducing novel features for improved performance.

Finally, our network only processes a part of the point cloud (about 50% of the original points). Since this input reduction might impact negatively the final quality evaluation, we complement the 3D representation with information from another modality, i.e., using features from the 2D images domain. To this end, we used a vision transformer (ViT)[16, 18] encoder to extract features from the 2D projected view. Afterward, we apply cross-attention [18] between both modalities' representations for fusion and to provide an enhanced multi-modal signal capturing higher-level PC features. We detail the 3D and 2D feature extraction in the following.

2.2.1. 3D Streams Network

We consider the geometry and color as independent information, corresponding to two different processing streams. Depending on which stream we consider, the input points $\mathbf{x}_i$ bring different information. For the geometry stream, $\mathbf{x}_i = (x_i, y_i, z_i)$ contains the 3-dimensional coordinates of the points, while for the color stream, $\mathbf{x}_i = (r_i, g_i, b_i)$ represents the RGB attribute information. In this work, we test only the use of Two-Stream Network, but it is possible to include additional features in other streams.

Both the color and geometry streams consist of three *feature embedding* layers, each comprising a series of 1-D convolutions interspersed with nonlinear functions. Following each feature embedding layer in the geometry stream, a multi-head self-attention layer is applied to draw the semantic affinities between neighboring points. Afterward, we employ cross-multi-

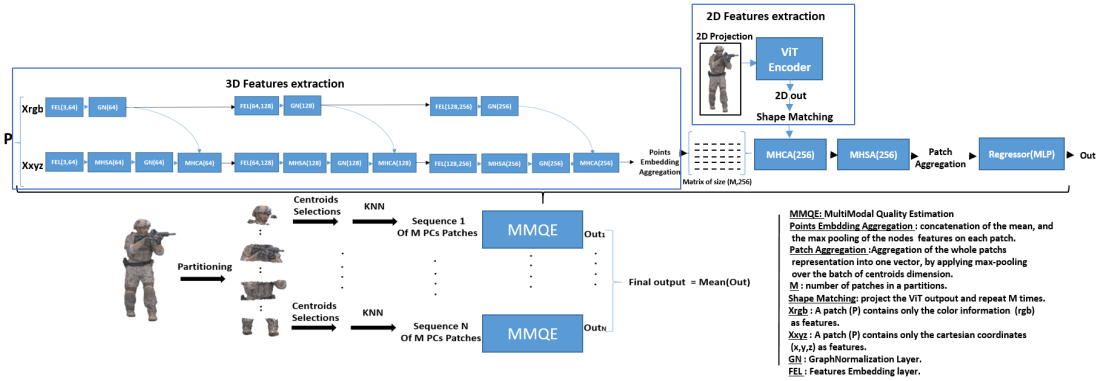

Fig. 1. General pipeline of our proposed method

head attention to align the learned color information with the geometrical representation.

Formally, for a given input point cloud partition composed of M patches $\{P^0, P^1, P^2, .., P^M\}$, each patch P^i is represented as a matrix of size $\mathbb{R}^{N\times F}$, where N is the number of points in the patch and F is the number of features per point. On each layer, a standard *feature embedding* layer $f_{\mathbf{\Theta}} : \mathbb{R}^F \rightarrow \mathbb{R}^{F'}$ is applied to produce a new representation of the provided X^i_{xyz} and X^i_{rgb} referring the geometry and color raw inputs or a subsequently produced representations, for each of the two streams. A multi-head self-attention layer $MHSA_{\mathbf{\Theta}} : \mathbb{R}^{F'} \rightarrow \mathbb{R}^{F'}$ is then applied to the geometry stream output $f(X_{xyz})$, producing an updated representation X'_{xyz}. Mathematically, the proposed $MHSA_{\mathbf{\Theta}}$ can be expressed as follows:

$$\mathbf{q} = X_{xyz}W_q, \mathbf{k} = X_{xyz}W_k, \mathbf{v} = X_{xyz}W_v \tag{1}$$

$$\mathbf{A} = \frac{softmax(\mathbf{q}\mathbf{k}^T)}{\sqrt{\frac{C}{h}}} \tag{2}$$

$$MHSA_{\mathbf{\Theta}}(f(X_{xyz})) = \mathbf{A}\mathbf{v} \tag{3}$$

where $W_q, W_k, W_v \in R^{C(C/h)}$ are learnable parameters, C and h are the embedding dimension and number of heads.

Notice that the X'_{xyz} representation can be interpreted as an embedding induced from a graph propagation layer, since the attention scores between points create a sort of soft connection simulating an adjacency matrix. More precisely, the actual connections between points could be obtained by setting a threshold on attention scores. Therefore, to accelerate convergence, we incorporate a GraphNorm operation [19] on both streams. Although the color representation X'_{rgb} can not be deemed to have an origin from a graph-similar function, we find that applying the same sort of normalization on the two streams is useful to keep a fixed scale of the features. The GraphNorm operation is a variation of InstanceNorm [20], tailored for graph normalization, and includes a learnable parameter α that determines how much of the channel-wise average to retain in the shift operation. The operation is expressed as follows:

$$\mathbf{x}' = \frac{\mathbf{x}' - \alpha \cdot \mathbb{E}[\mathbf{x}']}{\sqrt{\text{Var}[\mathbf{x}' - \alpha \cdot \mathbb{E}[\mathbf{x}']] + \epsilon}} \cdot \gamma + \beta \tag{4}$$

where γ and β are learnable affine parameters that are similar to those used in other normalization techniques.

Afterward, the representations produced by the two streams, X'_{rgb} and X'_{xyz}, are finally fused using a multi-head-cross attention layer $MHSCA_{\mathbf{\Theta}} : \mathbb{R}^{F'} \rightarrow \mathbb{R}^{F'}$. We note that cross-attention is essentially an extension of self-attention in which attention from one distribution is used to highlight the extracted features in another distribution. It can thus be used as an intuitive information fusion technique. The fusion here is carried out by measuring the similarity between the queries $\mathbf{q}$ (a linear projection of X'_{rgb}) and the keys $\mathbf{k}$ (a linear projection of X'_{xyz}) as well as using it to adjust the values vector $\mathbf{v}$ (a linear projection of X'_{xyz}). Consequently, in contrast to Eq. (3), the $MHCA_{\mathbf{\Theta}}$ function takes two inputs $f(X'_{xyz})$, and $f(X'_{rgb})$. The resulting vector is considered to be the new updated version of X'_{xyz} that is used for the subsequent network embedding layer. We refer the reader to [17, 18] for more details about the computation of cross-attention. To sum up, in the two-stream network, at each level a combination between the geometry and the color stream is applied to update the geometry representation with information about the corresponding color features.

After three consecutive (*feature embedding*, *MHSA GraphNorm*, *MHCA*) blocks of layers, a **Points Embedding Aggregation** is applied on each patch independently. Here a Max and Mean pooling operations are applied on the features channel dimension. Afterward, the result of the two poolings is concatenated to one vector, so each point cloud partition with M patch is transformed into M sequences of $2F'$-dimensional vectors.

2.2.2. 2D Stream and Multi-modal Fusion

To enhance the features of the 3D Network, we add a 2D stream in parallel to provide complementary information to the 3D representation. The combination of the 2D and 3D streams is achieved using cross-attention, which incorporates both the

attention mechanism and cross-correlation operations. This decision is motivated by two factors. Firstly, while the projection process introduces some distortion, certain types of noise are more accurately captured on 2D images, due to the capability of neural networks to address local connectivity on a regular grid. Secondly, computing the cross attention of 3D features with respect to 2D features provides more insight into the location of patches within a region (PC's partitions) and their semantic affinities.

ViT for Features Encoding: Visual Transformer (ViT) [16] is a transformer-based architecture that operates on patches of the input image, enabling it to model long-range dependencies and capture global context information and object relationships. The model consists of transformer blocks with multi-head self-attention layers and feed-forward neural networks, and includes a special CLS token [16] that is processed along with the image patch embeddings. The output of the final transformer block corresponding to the CLS token, is used as the image representation for the downstream task. For this study, we utilized a pre-trained ViT model on the Imagenet1K dataset, which consist of 12 layers and 12 attention heads. The patch size was set to 16×16, and the 3D model image was cropped to a size of 384×384.

Cross-Attention For Multi-modal Fusion: Both representation modalities from 2D network and 3D network are passed to a cross-attention layer for features fusion. First, the 2D CLS are repeated M times, and projected to meet the dimension size of 3D representation M sequence of $2F'$-dimension vectors. Afterward, for the cross-attention operation, we consider the representations coming from the 2D network as Queries, and linear transformations of 3D representation as Key and Value. This operation measures the cross-correlation between the features of both modalities and uses the resulting output to rectify the 3D patches' features. In addition to its fusion purpose, the mulimodal cross-attention operation also serves to find correspondences between the 3D and 2D representations, thus enhancing the local information of 3D features of each independent patch by knowing its corresponding location on the 2D regular grid.

2.3. Feature Aggregation and Quality Estimation

In order to aggregate the representations obtained from each point cloud partition and capture the affinities between them, we use a **Patch Aggregation** method. This involves also a multi-head self-attention layer followed by max pooling to produce a vector representing each partition. Finally, a shallow multi-layer perceptron is used to estimate the quality score, and the overall score is obtained by computing the mean of the sequence of partition scores.

3. EXPERIMENTS

3.1. Training and Implementation Details

We trained our model end-to-end using the mean square error (MSE) as the loss function. The goal of the network is to create a mapping function between the input point clouds and the mean opinion score (MOS) quality. The loss function is defined as:

$$\mathcal{L} = \text{MSE}\left(\text{mean}\left(\sum_i Out_i\right), \mathcal{Y}\right), \tag{5}$$

where Out_i refers to each predicted partition score, and $\mathcal{Y}$ refers to the MOS.

3.2. Evaluation Protocol and Result Analysis

To evaluate the effectiveness of our model, we conducted experiments on a publicly available benchmark that uses subjective scores and adopts different emerging compression schemes, ICIP20 [21]. ICIP20 includes 6 reference point clouds, each compressed using 5 levels and 90 degraded versions were derived through three types of compression.We used a 6-fold cross-validation protocol to train and test our model on ICIP20, with 5 reference point clouds used for training and one for testing at each iteration. Prior to each fold end-to-end multi-stream training, we first fine-tuned the ViT model by adding a regression head that will removed later. The 2D images used for this part of the training are obtained by projecting point cloud images of each corresponding fold.

Table 1. Results obtained on ICIP20 dataset using 6-fold cross validation

Model	PLCC ↑	SROCC ↑
po2point MSE	**0.946**	0.934
po2plane MSE	**0.959**	0.951
PSNR po2point MSE	0.868	0.855
color Y MSE	0.876	0.892
color Y PSNR	0.887	0.892
pl2plane AVG	0.922	0.910
pl2plane MSE	0.925	0.912
PCQM	0.796	0.832
GraphSim	0.931	0.893
PointNet-SSNR	0.908	0.955
PointNet-DCCFR	**0.947**	**0.973**
PointNet-Graph	**0.946**	**0.973**
Ours	**0.945**	**0.978**

Table 1 presents the results of our method on the ICIP20 dataset and compares them to state-of-the-art methods, with mean correlations calculated over all folds for our and other method. Our results demonstrate a strong correlation with the subjective ground truth, showing a clear gap for both PLCC and SROCC when compared to most existing methods. Our proposed method outperforms all other methods with the highest SROCC score achieving a correlation equal to **0.976**, and sharing the highest PLCC score with the full-reference po2planeMSe method, with a correlation equal to **0.959**. It's worth noting that our approach surpasses most full-reference methods [10, 6]. Notably, all listed methods in the table are full-reference except for PointNet-SSNR[14] and PointNet-Graph [22], which are no-reference.

4. CONCLUSION

In this paper, we introduce a novel no-reference quality metric for point clouds using a learning-based approach. Our network uses multi-modal input: we augment 3D point coordinates and attributes with a 2D projection of the point cloud, with the goal to extract complementary features for PC quality assessment. We also use self- and cross-attention to capture local relations across points. Our method achieves competitive results in predict MOS of compressed PCs.

5. REFERENCES

[1] M. Tliba, A. Chetouani, G. Valenzise, and F. Dufaux, "Representation learning optimization for 3d point cloud quality assessment without reference," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 3702–3706.

[2] R. Mekuria, Z. Li, C. Tulvan, and P. A. Chou, "Evaluation criteria for PCC (point cloud compression)," in *in ISO/IEC MPEG Doc*. N16332, 2016., vol. II, pp. 803–806.

[3] D. Tian et al., "Geometric distortion metrics for point cloud compression," in *IEEE ICIP*, 2017.

[4] E. Alexiou and T. Ebrahimi, "Point cloud quality assessment metric based on angular similarity," in *IEEE ICME-W*, 2018.

[5] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: measuring errors on simplified surfaces,," in *Computer Graphics Forum*, 1998, vol. 17, pp. 167–174.

[6] G. Meynet, J. Digne, and G. Lavoué, "Pc-msdm: A quality metric for 3d point clouds," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[7] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] R. Diniz, P. G. Freitas, and M. C. Q. Farias, "Towards a point cloud quality assessment model using local binary patterns," in *QoMEX*, 2020, pp. 1–6.

[9] M. Pietikäinen and G. Zhao, "Two decades of local binary patterns: A survey," *CoRR*, vol. abs/1612.06795, 2016.

[10] J. D. G. Meynet, Y. Nehmé and G. Lavoué, "Pcqm: A full-reference quality metric for colored 3d point clouds," 2020.

[11] 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*, pp. 3015–3029, 2022.

[12] A. Chetouani, M. Quach, G. Valenzise, and F. Dufaux, "Convolutional neural network for 3d point cloud quality assessment with reference," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–6.

[13] A. Chetouani, M. Quach, G. Valenzise, and F. Dufaux, "Deep learning-based quality assessment of 3d point clouds without reference," in *2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2021, pp. 1–6.

[14] M. Tliba, A. Chetouani, G. Valenzise, and F. Dufaux, "Point cloud quality assessment using cross-correlation of deep features," in *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, 2022, pp. 63–68.

[15] C. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," 2017.

[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[17] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[18] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 347–356.

[19] T. Cai, S. Luo, K. Xu, D. He, T. y Liu, and L. Wang, "Graphnorm: A principled approach to accelerating graph neural network training," *International Conference on Machine Learning*, 2020.

[20] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *ArXiv*, vol. abs/1607.08022, 2016.

[21] S. Perry et al., "Quality evaluation of static point clouds encoded using mpeg codecs," *2020 IEEE ICIP*, pp. 3428–3432.

[22] M. Tliba, A. Chetouani, G. Valenzise, and F. Dufaux, "Efficient Deep-Based Graph Metric for Point Cloud Quality Assessment," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, June 2023, IEEE.