A Novel No-Reference Point Clouds Quality Metric using Transformer Similar Architecture

Marouane Tliba, Aladine Chetouani, Giuseppe Valenzise, Frédéric Dufaux

A Novel No-Reference Point Clouds Quality Metric using Transformer Similar Architecture

Marouane TLIBA[1] Aladine CHETOUANI[1] Giuseppe VALENZISE[2] Frederic DUFAUX[2]

[1]Laboratoire PRISME, Universite d'Orleans, Orleans, France

[2]Universite Paris-Saclay, CNRS, CentraleSupelec, Laboratoire des signaux et systèmes

Résumé – Avec l'augmentation de la popularité des expériences immersives, le nuage de points est devenu la méthode préférée pour représenter les médias 3D. Cependant, il existe plusieurs distorsions potentielles qui peuvent affecter le contenu 3D tout au long des processus d'acquisition et de rendu. De plus, la transmission efficace de contenu volumétrique sur des systèmes de communication traditionnels peut avoir un impact négatif sur la qualité perçue du contenu livré. Pour évaluer avec précision l'étendue de cette dégradation, des métriques de qualité sont nécessaires. Ce travail présente une nouvelle métrique de qualité sans référence basée sur l'apprentissage profond qui peut évaluer le nuage de points directement, sans nécessiter de pré-traitement important. Cette métrique peut être utilisée pour une évaluation en temps réel à la fois au niveau de la transmission et du rendu. Notre approche proposée utilise une conception de modèle novatrice qui comprend des couches d'attention croisée et d'auto-attention pour apprendre les affinités sémantiques locales et maintenir la combinaison optimale d'informations de géométrie et de couleur à plusieurs niveaux, de l'extraction de fonctionnalités de base à la modélisation de représentation profonde.

Abstract – With the increasing popularity of immersive experiences, point cloud has emerged as the preferred method for representing 3D media. However, there are several potential distortions that can impact the 3D content throughout the acquisition and rendering processes. Additionally, transmitting volumetric content efficiently over traditional communication systems can negatively impact the perceived quality of the delivered content. To accurately assess the extent of this degradation, quality metrics are necessary. This paper presents a new deep-based no-reference quality metric that can evaluate the point cloud directly, without requiring extensive pre-processing. This metric can be used for real-time evaluation at both the transmission and rendering levels. Our proposed approach utilizes a novel model design that includes cross and self-attention layers to learn local semantic affinities and maintain the optimal combination of geometry and color information across multiple levels, from basic feature extraction to deep representation modeling.

1 Introduction

Point clouds (PC) have become increasingly important in recent years due to their ability to capture detailed and accurate representations of real-world objects and environments. One of the key benefits of using point clouds is that they can capture a high level of detail, including small features and irregularities that may be difficult to capture with other 3D data formats. However, 3D point clouds can only accurately represent a 3D scene by incorporating a large amount of information, which can reach thousands or even millions of points. Therefore, lossy compression schemes are often used, and became an inevitable application. To optimize the development of immersive 3D experiences and achieve the best visual quality for a given bit-rate, accurate quality metrics are needed. As a result, the field of point cloud quality assessment (PCQA) has received significant attention from researchers in recent years [9][13]. Point cloud perceptual quality can be assessed using subjective or objective methods. To this end, several effective Full reference objective approaches have been proposed. These approaches can be classified into three main groups : Point-based [10], [5]. Feature-based[8]. And Projection-based.

Some No-reference approaches have been also proposed especially the ones based on deep learning. However, these methods still present notable limitations and disadvantages, for instance, in order to adapt the irregular structure of scattered 3D point clouds, it implies a heavy pre-processing as 2D images projection or transformation (voxelization). As a result, the introduced pre-processing turning also the No-reference quality assessment time consuming and empirically hard to achieve. Recently, new efficient deep-based metrics such as PointNet-SSNR [12] and PointNet-DCCFR[11] have been released. They exploit the intrinsic features of point cloud data. As a drawback, these methods focus on distinct portions of the point cloud, assuming that the perceived quality is the same over the whole point cloud.

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 a point cloud regions. By utilizing a multi-level self and cross-attention, we are able to capture the local semantic affinities of points, providing a more comprehensive representation for the evaluation of its visual quality. To sum up, the main contributions of this paper are summarized as follows :

1. We present a novel efficient end-to-end deep-based method for PCs quality assessment (PCQA) that operates directly on the whole point cloud, without the need for any projection or other transformation. This allows for 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 connecti-

vity between distant features using self-attention. This facilitates better feature extraction from local regions.

3. We propose a new system for processing geometry and color features in point clouds using a parallel two-stream architecture. The two-stream networks operate independently, while keep interacting with each other dynamically at multiple levels using a cross-attention network design.

2 Proposed Method

As depicted in Fig. 1, the overall pipeline of our method consists of three main steps : starting with the pre-processing, then features extraction and representation mapping using self and cross-attention mechanisms, and finally quality estimation.

2.1 Pre-processing

Pre-processing is a critical stage in our approach, aimed at optimizing the point cloud (PC) processing pipeline and consists of 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 sufficient number of centroids and then apply the k-nearest neighbor clustering method. We note that the selected distant centroids ensure to have an ensemble of patches that covers the whole partition' points. Each of the patches is then fed into our model for feature extraction and representation modeling. By applying this pre-processing method, we ensure that we provide our model with a complete single enclosing all prominent features, and thus increase the model's ability to learn coherent local information from coherent regions.

2.2 Proposed Transformer Model

2.2.1 Overview

Our model design draws inspiration from two existing architectures : PointNet [4] and Transformer [14]. In particular, we employ a principal characteristic of PointNet in extracting information from point sets by using a permutation-invariant function. To do so, we used two parallel streams network 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 built on the *feature embedding* of the geometry stream by introducing multi-head self-attention which draws the semantic affinities between neighboring points.

In order 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 level of the network, capturing different levels of semantic affinities, as the point embedding is updated. Our method thus combines the strengths of both PointNet and Transformer while introducing novel features for improved performance.

2.2.2 Two-stream Network

We consider the geometry and color as independent information, corresponding to two different processing streams. Depending on which stream we consider, 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, points represent the RGB attribute information. Notice that it is possible to include additional features in other streams.

Both the color and geometry streams consist of three *feature embedding* layers, each of which is comprised of a series of 1-D convolutions that are 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-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 stander *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.

The X'_{xyz} representation can be considered as an embedding induced from a graph propagation layer. As the attention scores between the points create a sort of soft connection simulating the adjacency matrix. More precisely, the real connection between points could be obtained by setting an attention scores threshold. Therefore, to accelerate convergence, we incorporate a GraphNorm operation [1] 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 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 :

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

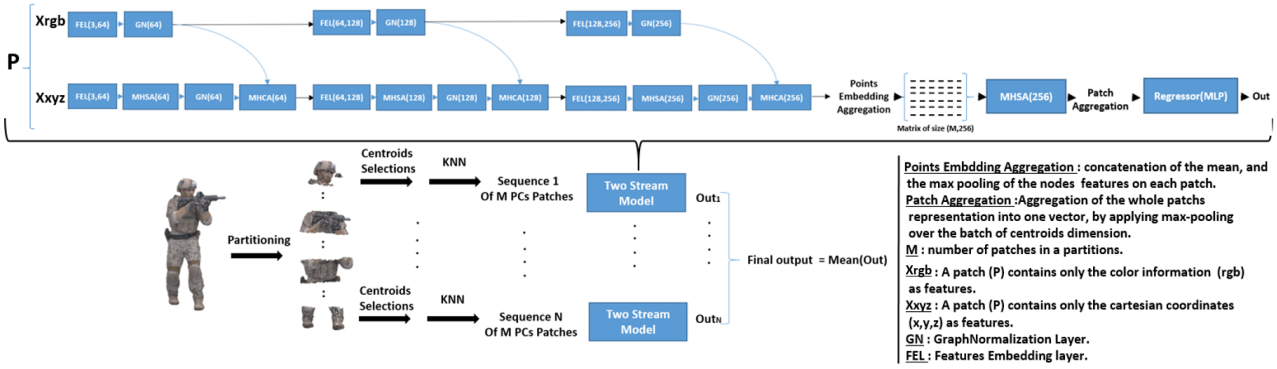

FIGURE 1 : General pipeline of our proposed method

TABLE 1 : Results obtained by training the model on ICIP20 and testing it on PointXR

Model	PLCC ↑	SROCC ↑
po2pointMSE	0.887	0.978
po2planeMSE	0.855	0.942
PSNRpo2pointMSE	**0.983**	**0.978**
PSNRpo2planeMSE	0.972	0.950
PointNet-DCCFR	0.981	0.964
PointGraph	0.967	**0.988**
Our	**0.969**	**0.990**

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 $MHSCA_{\Theta} : \mathbb{R}^{F'} \rightarrow \mathbb{R}^{F'}$. We note that cross-attention is only a varied form of self-attention and intuitive information fusion method in which attention from one distribution is used to highlight the extracted features in another distribution. 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 *Equation* (3), $MHCA_{\Theta}$ functions 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 [14, 2] 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 pooling is concatenated to one vector, so each point cloud partition with M patch, is transformed into M sequence of $2F'$-dimension vectors.

2.2.3 Feature Aggregation and Quality Estimation

In order to aggregate the representations obtained from each point cloud partition and capture the affinities between them. We used a **Patches Aggregation** method. This involved a multi-head self-attention layer [14], [6] followed by max pooling to produce a vector representing for each partition. Finally, A shallow multi-layer perceptron was used to estimate the quality score, and the overall score was obtained by computing the mean of the sequence of partition scores.

TABLE 2 : 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
PSNR po2point HAU	0.548	0.456
PSNR po2plane HAU	0.580	0.547
color Y MSE	0.876	0.892
color Cb MSE	0.683	0.694
color Cr MSE	0.594	0.616
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**
PointGraph	**0.946**	**0.973**
Our	**0.959**	**0.976**

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. The model was implemented in PyTorch and trained using the Adam optimizerwith an initial learning rate of 0.0001 and a batch size of one. The number of epochs varied depending on the training folds, ranging from 80 to 200 in the ICIP dataset.

3.2 Evaluation Protocol and Result Analysis

To evaluate the effectiveness of our model, we conducted experiments on two publicly available benchmarks that use subjective scores and adopt different emerging compression schemes, ICIP20 [7] and PointXR [3]. ICIP20 includes 6 reference point clouds, each compressed using 5 levels and 90 degraded versions were derived through three types of compression. PointXR includes 5 point clouds, each compressed using G-PCC with octree coding for geometry compression and Lifting and RAHT for color compression, resulting in 45 degraded versions.

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. To evaluate the generalization ability of our method to predict the quality on unknown PCs, we also trained on ICIP20 and tested on PointXR using Pearson and Spearman correlations. Results are reported in Table 2 and Table 1, with mean correlations calculated over all folds. We compared our method's performance to state-of-the-art methods using the same protocol.

Table 2 presents the results of our method on the ICIP20 dataset and compares them to state-of-the-art methods. 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 po2planeMSe method with a corelation equal to **0.959**. It's worth noting that our method even surpasses most of the full-reference methods. Notably, all listed methods in the table are full-reference except for PointNet-SSNR and PointNet-Graph, which are no-reference.

The cross-dataset evaluation results are presented in Table 1, demonstrating high correlations achieved by our method and outperforming some of the compared methods. These results exhibit the generalization capability of our metric in predicting the quality of unknown point clouds, indicating the proposed method's consistent performance across different validation sets. As larger annotated datasets for point clouds quality become available in the future, further validation of our approach can be conducted.

4 Conclusion

In this paper, we introduce a novel no-reference quality metric for point clouds using a learning-based approach. Our network employs multi-head self and cross-attention mechanisms to capture the local semantic affinities and establish probabilistic connectivity between points. Since the perceptual quality degradation can occur on both geometry and color levels, we propose a two-stream architecture that processes color and geometry distortion in parallel, and interact dynamically at multiple levels of the network. Our method achieves state-of-the-art correlations on two benchmarks for point clouds quality assessment .

Références

[1] Tianle CAI, Shengjie LUO, Keyulu XU, Di HE, Tie yan LIU et Liwei WANG : Graphnorm : A principled approach to accelerating graph neural network training. 2020.

[2] Chun-Fu Richard CHEN, Quanfu FAN et Rameswar PANDA : Crossvit : Cross-attention multi-scale vision transformer for image classification. *In 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 347–356, 2021.

[3] N. Yang E. ALEXIOU et T EBRAHIMI : Pointxr : A toolbox for visualization and subjective evaluation of point clouds in virtual reality. 2020.

[4] C. Qi el AL : Pointnet++ : Deep hierarchical feature learning on point sets in a metric space. 2017.

[5] D. Tian et AL : Geometric distortion metrics for point cloud compression. *In IEEE ICIP*, 2017.

[6] M Tliba et AL : Satsal : A multi-level self-attention based architecture for visual saliency prediction. *IEEE Access*, 10:20701–20713, 2022.

[7] S. Perry et AL : Quality evaluation of static point clouds encoded using mpeg codecs. pages 3428–3432, 2020.

[8] J. Digne G. MEYNET, Y. Nehmé et G. LAVOUÉ : Pcqm : A full-reference quality metric for colored 3d point clouds. 2020.

[9] Yana NEHMÉ, Jean-Philippe FARRUGIA, Florent DUPONT, Patrick LECALLET et Guillaume LAVOUÉ : Comparison of subjective methods, with and without explicit reference, for quality assessment of 3d graphics. 2019.

[10] C. Tulvan R. MEKURIA, Z. Li et P. CHOU : Evaluation criteria for pcc (point cloud compression). *In in ISO/IEC MPEG Doc*, volume II, pages 803–806. N16332, 2016.

[11] Marouane TLIBA, Aladine CHETOUANI, Giuseppe VALENZISE et Frédéric 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*, pages 63–68, 2022.

[12] Marouane TLIBA, Aladine CHETOUANI, Giuseppe VALENZISE et Fréderic DUFAUX : Representation learning optimization for 3d point cloud quality assessment without reference. *In 2022 IEEE International Conference on Image Processing (ICIP)*, pages 3702–3706, 2022.

[13] Marouane TLIBA, Aladine CHETOUANI, Giuseppe VALENZISE et Fréderic DUFAUX : Pcqa-graphpoint : Efficient deep-based graph metric for point cloud quality assessment. *In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[14] Ashish VASWANI *et al.* : Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017.