Visual-Saliency Guided Multi-modal Learning for No Reference Point Cloud Quality Assessment

Xuemei Zhou
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Delft University of Technology
Delft, The Netherlands
xuemei@cwi.nl

Irene Viola
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
irene.viola@cwi.nl

Ruihong Yin
University of Amsterdam
Amsterdam, The Netherlands
r.yin@uva.nl

Pablo Cesar
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Delft University of Technology
Delft, The Netherlands
p.s.cesar@cwi.nl

Abstract

As 3D immersive media continues to gain prominence, Point Cloud Quality Assessment (PCQA) is essential for ensuring high-quality user experiences. This paper introduces ViSam-PCQA, a no-reference PCQA metric guided by visual saliency information across three modalities, which facilitates the performance of the quality prediction. Firstly, we project the 3D point cloud to acquire 2D texture, depth, and normal maps. Secondly, we extract the saliency map based on the texture map and refine it with the corresponding depth map. This refined saliency map is used to weight low-level feature maps to highlight perceptually important areas in the texture channel. Thirdly, high-level features from the texture, normal, and depth maps are then processed by a Transformer to capture global and local point cloud representations across the three modalities. Lastly, saliency along with global and local embeddings, are concatenated and processed through a multi-task decoder to derive the final quality scores. Our experiments on the SJTU, WPC, and BASICS datasets show high Spearman rank order correlation coefficients/Pearson linear correlation coefficients of 0.953/0.962, 0.920/0.920 and 0.887/0.936 respectively, demonstrating superior performance compared to current state-of-the-art methods. The code is available at https://github.com/cwi-dis/ViSam-PCQA_MM2024Workshop.

CCS Concepts

• **Human-centered computing → Visualization design and evaluation methods**; • **Computing methodologies → Perception**; *Model development and analysis*; **Image processing**.

Keywords

No Reference, Point Cloud Quality Assessment, Projection, Visual Saliency, Multi-Modal

ACM Reference Format:
Xuemei Zhou, Irene Viola, Ruihong Yin, and Pablo Cesar. 2024. Visual-Saliency Guided Multi-modal Learning for No Reference Point Cloud Quality Assessment. In *Proceedings of the 3rd Workshop on Quality of Experience in Visual Multimedia Applications (QoEVMA '24), November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3689093.3689183

1 Introduction

Digital representation of 3D models has gained increased interest and has been used in prevalent 3D computer vision applications such as social Virtual Reality (VR) [27, 32], cultural heritage [25] and architectural models [17]. Point clouds play a crucial role in various real-world applications. Hence, predicting the visual quality of point clouds accurately and efficiently, in a way that correlates well with the Human Vision System (HVS), is highly desired. The intricate geometrical structure and densely packed points of point clouds, complete with attributes such as color, normal, and transparency, allow for detailed representations of environments, objects, and humans. While this richness of information is valuable, it presents challenges for the efficiency and accuracy of Point Cloud Quality Assessment (PCQA) metrics. The different factors that contribute to the visual quality of point clouds are not fully understood, adding to the complexity of developing effective PCQA metrics.

Depending on the availability of reference information: PCQA methods can be broadly divided into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) PCQA methods [47]. However, the pristine reference point clouds are not always available in real-world applications. According to the domain in which the PCQA metric is computed, we categorize PCQA metric into two categories: point-based and image-based. Point-based metrics evaluate the quality of the distorted point cloud directly on the 3D point cloud itself, each point has its own quality score. Image-based metrics evaluate the point cloud quality based on 2D projections. The point-based model can better capture the geometry topology

QoEVMA '24, November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-1204-3/24/11
https://doi.org/10.1145/3689093.3689183

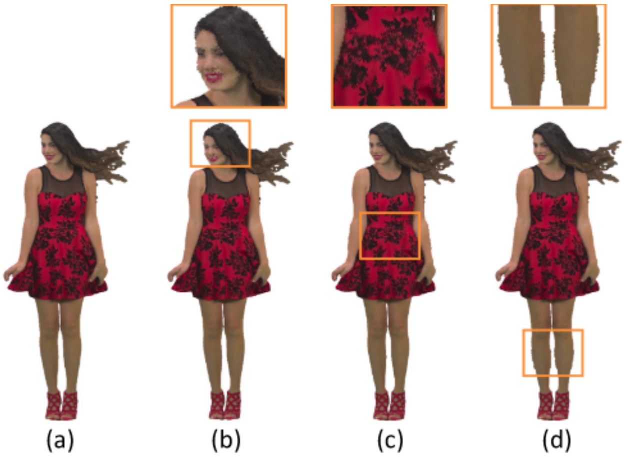

Figure 1: Illustration to show the perceptual impact of distortion in different areas on *redandblack* point cloud. (a) is the reference version. (b)-(d) depict the effects of introducing geometry and color Gaussian noise with equal intensity on the face, dress, and legs, respectively. Notably, (c) exhibits nearly identical perceptual quality as the reference point clouds, attributed to the chaotic background texture that effectively masks the distortion. (d) ranks second in perceptual quality, while (a) is observed to have the least favorable perceptual quality.

of the point cloud at the cost of high computational cost, while image-based models can take advantage of the well-developed Image Quality Assessment (IQA) algorithms while introducing the additional distortion due to the projection process. Considering the above reasons, we focus on the image-based NR-PCQA in this paper.

PQA-Net [22] takes 6 orthographic projections of point clouds as inputs, features are extracted after Convolution Neural Network (CNN) blocks, and they share a distortion identification and a quality prediction module that assist in obtaining final quality scores. IT-PCQA [38] utilizes the rich prior knowledge in images and builds a bridge between 2D and 3D perception in the field of quality assessment, a hierarchical feature encoder and a conditional discriminative network is proposed to extract effective latent features and minimize the domain discrepancy. pmBQA [36] proposes an image-based blind quality indicator via multi-modal learning by using four homogeneous modalities (i.e., texture, normal, depth and roughness). MM-PCQA [43] partitions point clouds into submodels for local geometry representation and renders them into 2D projections for texture. Geometry and texture features are extracted separately using point-based and image-based neural networks. A symmetric cross-modal attention module is used for integrating quality-aware information. IT-PCQA [38] reveals the potential connection between different types of media content in the field of quality assessment. PQA-Net [22] and pmBQA [36] use the multitask decoder and multiple modality-related features on 2D; MM-PCQA [43] proves the effectiveness of cross-modality perception for PCQA.

The aforementioned NR metrics mainly consider the projected images of the point cloud or are completed with 3D point cloud modality. However, they do not consider the impact of visual saliency on improving the prediction accuracy of media content [18]. As illustrated in Figure 1, the impact of distortion in different regions of the point cloud (*RedandBlack*) is evident. Recent developments have seen certain metrics incorporating visual saliency into their design paradigms [19]. Some directly extract visual saliency on the 3D point cloud [4, 14], while others employ existing saliency prediction models on 2D projections [7], subsequently re-projecting them onto the 3D point cloud. Visual saliency is utilized either as a quality indicator or as a weight map for pooling extracted handcrafted features [34], with the aim of selecting features under the guidance of visual saliency. In contrast, our approach utilizes the saliency map from a pre-trained 2D saliency prediction model, to guide the selective learning of low-level features, which are extracted by the image encoder. This aims to automatically identify visually salient areas that aid in perceptual quality prediction. Specifically, we propose incorporating depth-related priors into the 2D saliency map to inherently provide a sense of depth for point clouds. Additionally, the low-level feature maps extracted by a CNN-based image encoder, which preserve spatial information, are weighted with the refined saliency map pixel-wisely. The high-level features, which contain semantic information, are processed through a cross-modality attention mechanism to obtain local correspondence and global feature. By concatenating the corrected visual saliency with the local and global embeddings, we generate the final score through two branches: quality score regression and distortion type classification.

As shown in Figure 1, the perceptual quality of point clouds is dependent on distortion type since the HVS has different tolerances for different distortions, and where the distortion is located can have a huge impact on the overall quality of point clouds [40]. Thus, the proposed visual saliency guided multi-modal learning can estimate the perceptual quality of point clouds effectively and comprehensively. The main contributions of this paper are summarized as follows:

- We propose a Visual Saliency guided multimodal NR PCQA (ViSam-PCQA) metric. Visual saliency from the pre-trained model is treated as pseudo-ground-truth, used to correct low-level features that contain learned attention and spatial information. The spatial information is crucial when weighing the visual saliency map with the feature maps from texture, depth and normal maps.
- We utilize the cross-modality attention to obtain the local correspondence among modalities and global features within the same modality, which can compensate for the stereo spatial information loss during the 3D-to-2D projection.
- Extensive experimental evaluations demonstrate that ViSam-PCQA outperforms other state-of-the-art methods. Ablation studies elucidate the distinct contributions of each component within the framework, with a particular emphasis on highlighting the crucial role played by the corrected visual saliency.

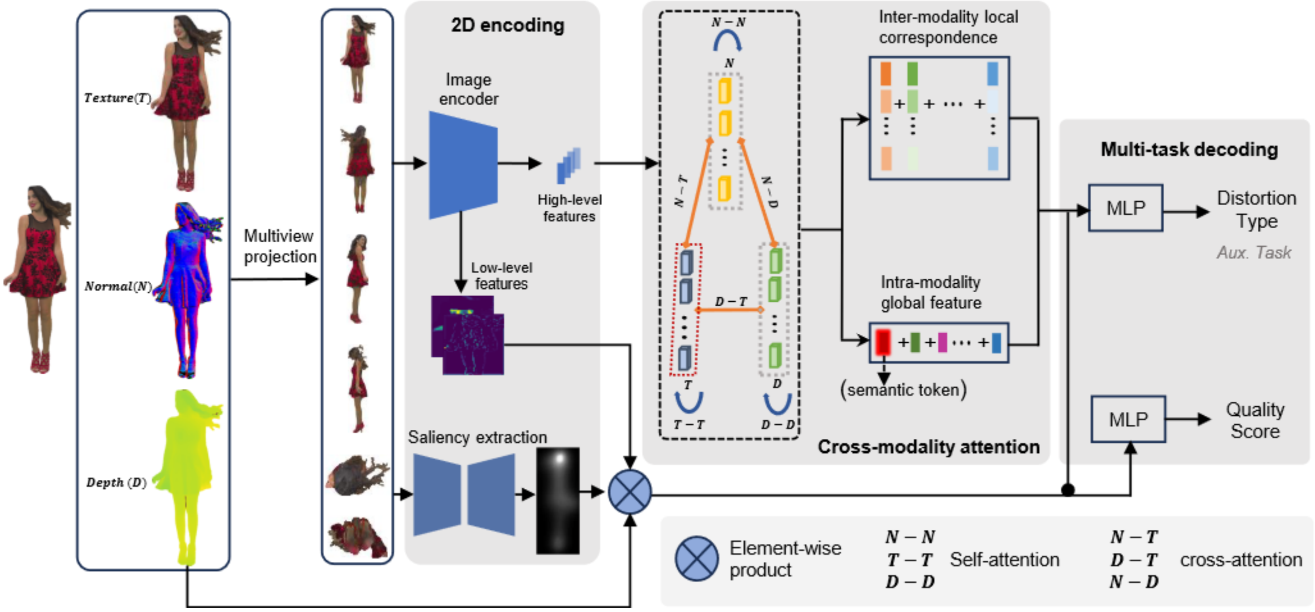

Figure 2: The overall framework of our proposed method.

2 Related Work

Image-based PCQA. In the image-based approaches, firstly used in [8] for point clouds, the rendered models are mapped onto planar surfaces, on which conventional IQA metrics are applied to provide a quality score. The prediction accuracy of IQA metrics on 2D views of point clouds is initially examined in [30]. Yet, enabling a large number of views in denser camera arrangements may lead to redundancies and extra computational costs, without guaranteeing performance improvements, as indicated in [2]. Yang *et al*. [37] introduce a metric based on a weighted combination of global and local features, which are extracted from 6 orthographic texture and depth images. Wu *et al*. [35] obtain the same projections, and weight the contributions of each face based on the ratio of the size of a plane to the sum of the area of six planes on the bounding box. A final score is obtained as a weighted average between quality scores from geometry and texture information. In [21], point clouds undergo translations, rotations and scaling before projection. They suggested that the principle of information content-weighted pooling provides a good framework and proposed the use of IW-SSIM on the projected views. The mentioned methods are mainly based on the projected texture information, without considering the geometry structure of the point cloud.

Visual Saliency for PCQA. Bourbia *et al*. [7] present an NR approach that incorporates the advantage of the transformer encoder architecture and the visual saliency to predict the perceived visual quality of distorted point clouds. They project the point cloud into multi-view and weight each view with its corresponding calculated saliency map through a pointwise multiplication to detect the regions of interest, and then regress the weighted sub-images to a quality score. However, the weighted sub-images are not guaranteed to be the saliency areas correlated to the perceptual quality. RR-CAP [46] makes the first attempt to simplify reference and distorted point clouds into projected saliency maps with a downsampling operation in an RR manner. The objective quality scores of distorted point clouds are produced by combining content-oriented similarity and statistical correlation measurements based on the saliency maps. PQSM [34] introduces a 3D point cloud saliency map generating method, which integrates depth information to enhance geometric representation. Three structural descriptors capturing geometry, color, and saliency discrepancies are used to construct local neighborhoods. A saliency-based pooling strategy refines the descriptors, yielding a comprehensive quality score. Laazouf *et al*. [15] firstly compute a 3D saliency map for each distorted point cloud. Then, a threshold-based filter is used to select the most salient points. Estimates of their statistical properties (Entropy, Standard deviation, Skewness, Kurtosis, Median and Mean) form a features vector from both geometrical and perceptual attributes. The support vector regressor is utilized to regress the feature vector as a quality score. These three non-learning metrics consider mainly one modality.

3 Proposed Method

The framework overview is exhibited in Figure 2. The point clouds are first projected into three different modalities, texture map, depth map and normal map. Then we use an image encoder θ_I to extract the low-level features and high-level features, respectively. Since the primary cues for visual attention often come from the 2D projections captured by the retina [6], depth and normal information are crucial for spatial perception and object localization. We use a pre-trained visual saliency model on the texture image and use its output to correct the low-level feature after an image encoder. At the same time, the texture image, depth image, and normal image are put into the same image encoder to get the semantic feature. Subsequently, the semantic features are put into an intra-and-inter modality attention module to get the global and local features of the point cloud. Finally, the global and local features are concatenated to the distortion type classification branch to learn the distortion-oriented features. The corrected visual saliency with the distortion-oriented features are concatenated and decoded into the quality values via the quality regression branch.

3.1 Pre-processing

Consider one point cloud denoted as $\mathcal{P} = \{p_{(1)}, p_{(2)}, ..., p_{(i)}\}_{i=1}^{N} \in R^{N\times 6}$, where each point $p_{(i)=[p_i^G, p_i^T]} = [x, y, z, G, R, B]$ indicates the geometry coordinates and the RGB color information, N stands for the number of points belonging to the point cloud. Let $\mathcal{P}$ be orthogonally projected onto M different 2D planes around the bounding box, resulting in M texture maps, $\mathcal{T} \in R^{H\times W\times 3}$, M depth maps, $\mathcal{D} \in R^{H\times W\times 1}$, and M normal maps, $\mathcal{N} \in R^{H\times W\times 3}$, where $m \in M = |\{\text{up, down, left, right, front, back}\}|$ and $H \times W$ denotes the resolution of m_{th} projected image after removing the background. For texture map $\mathcal{T}$, we calculate the 2D saliency map based on the current state-of-the-art perceptual saliency detection algorithm [23], which is defined as $\mathcal{V} = \{I_{i,m}\}_{i=1}^{H\times W} \in R^{H\times W\times 1}$, where $I_{i,m}$ denotes for the importance value of the i_{th} pixel from the m_{th} texture map.

3.2 Corrected Saliency Map Generation

We select the CNN-based image encoder that can retain 2D spatial information [16] at the shallow layers to extract the low-level features from only the texture image. TranSalNet [23] is used to extract the salient area of the texture image, which is defined as

$$\mathrm{V}_m = \phi(\mathcal{T}_m), \tag{1}$$

ϕ is the pre-trained TranSalNet model, $\mathrm{V}_m \in R^{H\times W\times 1}$ is the extracted saliency map. Intuitively, in the stereo scenes, human has a preference to the area that is closer to themself [48]. So after obtaining the saliency map, the corresponding depth image is laid on the up of it, which is expressed as

$$\mathrm{V}_{dm} = \mathrm{V}_m \bigotimes \mathcal{D}_m, \tag{2}$$

where $\otimes$ is the element-wise product. Subsequently, we utilize the depth-guided saliency map V_{dm} to weight the low-level feature map of all channels. By producing the element-wise product of the pseudo-ground-truth visual saliency and the learned visual saliency through the network [28], the effect of pseudo-ground-truth saliency maps considering the HVS for intervening in the saliency maps learned by the network is achieved, resulting in the corrected saliency maps,

$$\hat{F}_V = Avg(\mathrm{V}_{dm} \bigotimes \mathrm{L}_m^C), \tag{3}$$

L_m^C is the shallow layer output of the CNN based image encoder with C channels, $Avg()$ is the average pooling along the feature map and multi-view channels. Figure 3 shows the initial saliency map through the pre-traind model, the depth-related saliency map, the feature map and the corrected saliency map of the first and last channel, respectively. Notably, we computed the saliency map across the entire projection and integrated this global prior with the feature maps from all channels. Each channel captures distinct salient areas based on different filters, as observed in Figure 3. For instance, the first channel highlights the texture on the dress as salient, while the last channel emphasizes the contour of the projected image. We enable the network to autonomously learn the allocation of importance with the global saliency prior.

3.3 Multi-modal Feature Extraction

We next use the same CNN-based image encoder Ψ to extract the high level features from the texture map, depth map, and normal map, separately, resulting in:

$$\mathrm{H}_m^K = \Psi(K_m), \tag{4}$$

where $k \in \{\mathcal{T}, \mathcal{D}, \mathcal{N}\}$, $\mathrm{H}_m^K \in R^d$ is a d-dimension representation.

3.4 Global Feature Aggregation and Local Feature Correspondence via Transformer

Considering three distinct modalities and an image encoder handling input as image patches, we utilize the Transformer architecture to extract both global and local features. To extract the global features within each modality, the self-attention module is applied to each modality. Besides, similar to BERT [12] and ViT [10], we introduce a learnable semantic token in the self-attention module. The semantic token is shared among the multi-view projections, serving as a global feature of the whole point cloud. For the correspondence among different modalities, we use a symmetrical attention module to explore the relationship of the image patches from different modalities. Here, we take 3 modality-related features as input, and obtain the intra-attention global features F_α^k is defined as

$$\mathrm{F}_\alpha^k = \Theta(\mathrm{Z}^k, \mathrm{Z}^k), \tag{5}$$

$\mathrm{Z}^k = [T_s^k, H_1^k, H_2^k, ..., H_M^k] \in R^{(1+M)\times d}$, $T_s \in R^d$ is the sementic token. $\mathrm{F}_\alpha^k \in R^d$ is a d dimensional representation. The inter-modality local features among 3 different modalities F_β^a, are defined as

$$\begin{aligned}\mathrm{F}_\beta^{H^\mathcal{T}, H^\mathcal{D}} &= \Theta^*(H^\mathcal{T}, H^\mathcal{D}),\\ \mathrm{F}_\beta^{H^\mathcal{T}, H^\mathcal{N}} &= \Theta^*(H^\mathcal{T}, H^\mathcal{N}),\\ \mathrm{F}_\beta^{H^\mathcal{N}, H^\mathcal{D}} &= \Theta^*(H^\mathcal{N}, H^\mathcal{D}),\end{aligned} \tag{6}$$

Likewise, $\mathrm{F}_\beta^{H^\mathcal{T}, H^\mathcal{D}}$, $\mathrm{F}_\beta^{H^\mathcal{T}, H^\mathcal{N}}$ and $\mathrm{F}_\beta^{H^\mathcal{N}, H^\mathcal{D}}$ are d dimensional representation. The modality-related feature can express the local relationship among the modalities. For example, the local region of texture distortions related to facial features might exhibit a stronger association with the front view rather than the back view of texture distortion. The global feature is the feature map derived from the semantic token. The final quality embedding can be concatenated by the intra-modal global features and the inter-modal local features obtained by:

$$\hat{F}_Q = \hat{F}_g \oplus \hat{F}_l, \tag{7}$$

where $\oplus$ indicates the concatenation operation, and $\hat{F}_Q$ represents the final quality-aware features, the global feature $\hat{F}_g$ and local feature $\hat{F}_l$ are defined as follows:

$$\hat{F}_g = \mu(\mu([H_1^k, H_2^k, ..., H_M^k]) + \mathrm{F}_\alpha^k), \tag{8}$$

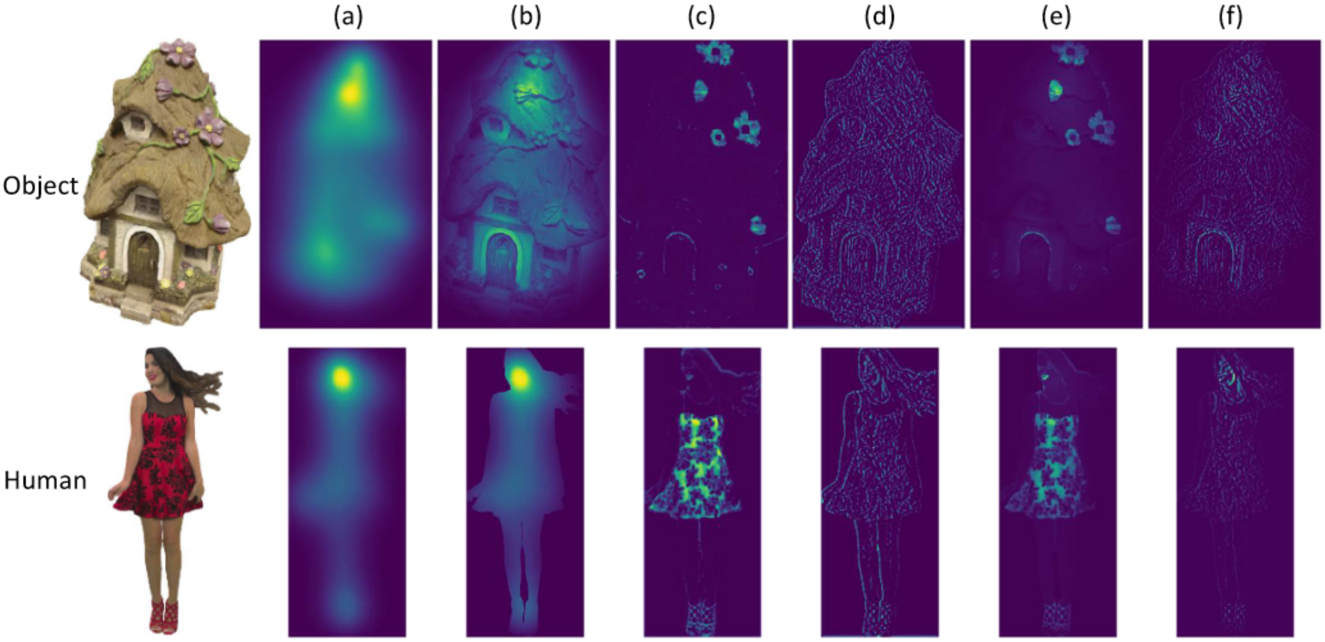

Figure 3: Examples of the visual saliency related operations. From (a) to (f) are the saliency map detected by TranSalNet; the depth-guided saliency map; the output of the 1st/256th channel of layer1 in ResNet; the corrected saliency map for the 1st/256th channel, respectively.

and

$$\hat{F}_l = \mu(H^k + \mathrm{F}_\beta^{H^\mathcal{T}, H^\mathcal{D}} + \mathrm{F}_\beta^{H^\mathcal{T}, H^\mathcal{N}} + \mathrm{F}_\beta^{H^\mathcal{N}, H^\mathcal{D}}), \tag{9}$$

in which μ is the mean operation along multi-view channel. The multi-task decoder consists of a Multi Layer Perception (MLP)-based classifier and regressor. The regressor and the classifier are a two- and three-layer ReLU-MLP respectively.

$$\begin{aligned} \hat{Q} &= \vec{D}(\hat{F}_Q), \\ \vec{P} &= \vec{D}(\hat{F}_Q \oplus \hat{F}_V), \\ \hat{P} &= softmax(\vec{P}), \end{aligned} \tag{10}$$

where $\hat{Q}$ is the predicted quality score, $\hat{P} = \{\hat{p_1}, \hat{p_2}, ..., \hat{p_E}\}$ is the predicted probability over E distortion types, and $\vec{P}$ is the output of the fully connected layers for distortion type classification before softmax.

For the quality regression task, we focus on minimizing the average prediction error of all training samples and lay importance on the ranking of the quality as [43]. Therefore, the loss function for the regression task includes two parts: MSE and ranking error, which can be derived as:

$$\mathcal{L}_1 = \frac{1}{n} \sum_{e=1}^{n} (q_e - q_e')^2, \tag{11}$$

where q_e is the predicted quality scores, q_e' is the ground truth labels of the point cloud, and n is the size of the mini-batch. The rank loss can better assist the model in distinguishing the quality ranking even the point clouds in a mini-batch have similar quality levels. To this end, we use the differentiable rank function described in [29] to approximate the rank loss:

$$\begin{aligned} \mathcal{L}_2^{ij} &= \max\left(0, \left|q_i - q_j\right| - e\left(q_i, q_j\right) \cdot \left(q_i' - q_j'\right)\right), \\ e\left(q_i, q_j\right) &= \left\{ \begin{array}{c} 1, q_i \geq q_j, \\ -1, q_i < q_j, \end{array} \right. \end{aligned} \tag{12}$$

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

$$\mathcal{L}_2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{L}_2^{ij}, \tag{13}$$

cross-entropy loss $\mathcal{L}_3$ is used for distortion type classification. Then, the loss function can be calculated as the weighted sum of MSE loss, rank loss and distortion type classification loss:

$$Loss = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 \tag{14}$$

where λ_1, λ_2 and λ_3 are used to control the proportion of the MSE loss, the rank loss and distortion type classification loss.

4 Experiments

4.1 Datasets

- SJTU: It has 9 reference point clouds and 378 distorted samples. Each reference PC is impaired by 7 different types of distortion under 6 levels. Detailed distortion types include Octree-based compression, Color Noise (CN), Geometric Gaussian Noise (GGN), downsampling, and combinations of the CN, GGN and downsampling. SJTU includes 5 human body models and 4 inanimate objects.

Table 1: Performance comparison with state-of-the-art approaches on the SJTU, WPC and BASICS datasets. Best in bold and second with underline. State-of-the-art results for NR-PCQA are cited from the literature, employing varied training strategies and splits, without independent validation by the authors.

Type	Modal Number	Methods	SJTU Dataset		WPC Dataset		BASICS Dataset	
			SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
FR	1	PointSSIM [3]	0.687	0.714	0.454	0.467	0.692	0.725
	1	MSE-p2po [24]	0.729	0.812	0.456	0.485	0.799	0.005
	1	PSNR-yuv [30]	0.795	0.817	0.449	0.530	0.510	0.543
	1	PCQM [26]	0.864	0.885	0.743	0.750	0.810	0.888
	1	GraphSIM [39]	0.878	0.845	0.583	0.616	0.773	0.801
	1	PointPCA [5]	0.907	0.932	0.890	0.894	0.866	0.926
NR	2	IT-PCQA [38]	0.630	0.580	0.540	0.550	0.310	0.302
	1	3D-NSS [42]	0.714	0.738	0.648	0.651	0.617	0.657
	4	pmBQA [36]	0.900	0.932	0.912	0.898	/	/
	2	MM-PCQA [43]	0.910	0.923	0.841	0.856	0.831	0.882
	1	GMS-3DQA [44]	0.911	0.918	0.831	0.834	0.855	0.930
	1	Wang's [33]	0.930	0.940	0.800	0.810	/	/
	1	PKT-PCQA [20]	0.932	0.912	0.557	0.560	/	/
	3	ViSam-PCQA(Ours)	**0.953**	**0.962**	**0.920**	**0.920**	**0.887**	**0.936**

- WPC: It contains 20 reference point clouds, and each is degraded under five types of distortions with different levels, leading to 740 distorted samples. Distortions include downsampling, Gaussian noise contamination, G-PCC (Octree), G-PCC (Trisoup) and V-PCC. WPC dataset collects objects including snacks, fruits and vegetables, etc.
- BASICS: The BASICS dataset [1] comprises 75 point clouds from 3 different semantic categories: (i) Humans & Animals, (ii) Inanimate Objects, and (iii) Buildings & Landscapes. Each point cloud is compressed with 3 compression methods from the MPEG standardization field, i.e., Octree-RAHT, Octree-Predlift and V-PCC; 1 learning-based algorithm, i.e., GeoCNN, at varying compression levels, resulting in 1494 processed point clouds. BASICS dataset is the current largest available dataset for PCQA with human annotated labels.

4.2 Evaluation Criteria

The evaluation of performance relies on three standard criteria including Spearman Rank Order Correlation Coefficient (SROCC), Pearson Linear Correlation Coefficient (PLCC). Additionally, the logistic regression recommended by standardization organization [31] is used to map the dynamic range of the scores from the predicted score into the quality label range. Higher values of SROCC, PLCC indicate better performance in terms of correlation with human opinion.

4.3 Implementation Details

All the projections are rendered with the assistance of Open3D [45], the number of projections for each modality is naturally set to 6. Adam optimizer [13] is utilized with weight decay 1e-4, the initial learning rate is set as 5e-5, the batch size is set as 18, and the model is trained for 100 epochs. The projected images are randomly cropped into image patches at the resolution of 224×224 for all modalities and corresponding saliency maps. The ResNet50 [11] is used as the image encoder, which is initialized with the ImageNet dataset [9]. The multi-head attention module employs 8 heads and the feed-forward dimension is set as 2048. The weights λ_1, λ_2 and λ_3 for $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ are set as 1.

Table 2: Ablation study of ViSam-PCQA for key components, i.e., corrected visual saliency, multi modalities that include both the depth and normal map, and distortion type, DT is short for Distortion Type.

Settings	SJTU Dataset		
	SROCC	PLCC	ACC
ViSam-PCQA	**0.953**	0.962	**0.762**
(*Visual Saliency*)			
/wo corrected saliency maps	0.951	0.962	0.751
(*Modality*)			
/wo depth & normal maps	0.942	0.952	0.659
(*Distortion Type*)			
/wo DT classification	0.950	**0.965**	/

For relatively small datasets SJTU (378) and WPC (740), the k-fold cross validation strategy is employed to accurately estimate the performance of the proposed method. 9-fold and 5-fold cross validation is selected for SJTU and WPC, respectively. The average performance is recorded as the final result. For the BASICS dataset, we divide the dataset into train-validation-test as the ratio of 6:2:2. There is no content overlap between the training and testing sets. For the FR-PCQA methods that require no training, we simply validate them on the same testing sets and record the average performance.

4.4 Overall Performance

14 state-of-the-art PCQA methods are selected for comparison, which consist of 6 FR-PCQA methods and 8 NR-PCQA methods. The FR-PCQA methods include PointSSIM [3], MSE-p2point (MSE-p2p) [24], PSNR_YUV [30], PCQM [26], GraphSIM [39], and PointPCA

[5], these metrics construct and evaluate on a point-to-point comparison or local neighborhood to include the structural information. The NR-PCQA methods include: GMS-3DQA [44], which takes the projections from only the texture. 3D-NSS [42], PKT-PCQA [20], and Wang's metric [33] evaluate the quality directly on the point cloud, the last two adopt multi-task learning which includes distortion type classification, distortion level regression/classification, and quality regression, respectively. IT-PCQA [38], MM-PCQA [43], pmBQA [36] resolve the PCQA problem with more than one modality.

The results, as detailed in Table 1, highlight ViSam-PCQA's superior performance across all evaluation criteria on both SJTU and WPC datasets, representing a significant advancement. Notably, the SROCC/PLCC witnessed an increase of 2.2%/5.2% and 0.87%/2.4% when compared with the second-best metric for SJTU and WPC datasets, respectively. Moreover, our model outperforms all FR-PCQA metrics, underscoring its ability to capture essential point cloud characteristics and align closely with the HVS. Summarizing the outcomes, several key conclusions can be drawn: 1) The incorporation of additional modalities (pmBQA) and heightened modality complexity (MM-PCQA) does not consistently result in performance enhancement, suggesting the existence of redundant information that may confound the network. 2) In contrast to models like GPA-Net and Wang's metric, which integrate two auxiliary tasks (distortion type classification and distortion degree regression), our emphasis on the quality regression task with visual saliency related features, suggests that an excessive refinement of auxiliary tasks may not necessarily bolster prediction accuracy. 3) The consistent performance observed on SJTU, WPC and BASICS datasets, despite variations in distortion types and content, underscores the robustness of ViSam-PCQA.

4.5 Ablation Study

The SJTU dataset encompasses a variety of contents, including both human figures and inanimate objects, and exhibits a broad range of distortion types. To gain a deeper understanding, we conducted ablation studies on the SJTU dataset by systematically removing key components one at a time.

Impacts of the corrected saliency maps. Quality assessment should align with human perception. Saliency maps highlight regions in an image that are perceptually more important or salient, it directs attention to parts of the point cloud that may have a more significant visual impact. The SROCC performance has a slight increase on SJTU. Additionally, incorporating the pseudo-ground-truth saliency map with the learning process enables to capture the details and variations in quality that might be overlooked by a uniform weighting approach guided only by the quality score regression. We can see the visual saliency helps the auxiliary task, the accuracy of distortion type classification improves 1.4% on SJTU. This, in turn, can lead to a more nuanced and accurate quality evaluation.

Impacts of the modalities. Combining information from both depth and normal contributes to a more realistic visual representation of the scene [41]. The depth map provides information about the distance of each point in the point cloud from the camera, which can help in assessing the surface details and detecting discontinuities. Normal maps encode surface normals at each point, which can aid in evaluating the smoothness and geometric fidelity. Removing such information will result in an inaccurate estimation for an overall perceptual experience. All criteria performance drops (1.2%, 1.0% and 13.5% for SROCC, PLCC and ACC) for the SJTU dataset. Leveraging depth/normal maps in PCQA provides a multi-faceted approach to evaluating geometric accuracy, surface details, and visual realism.

Table 3: Cross-dataset evaluation among SJTU, WPC and BASICS datasets. Note the model is validated on the test dataset with all the contents.

Training Dataset	Testing Dataset					
	SJTU		WPC		BASICS	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
SJTU	–	–	0.531	0.516	0.488	0.654
WPC	0.788	0.817	–	–	0.608	0.646
BASICS	0.577	0.591	0.393	0.391	–	–

Impacts of the distortion type classification. We assume that the distortion type classification task can facilitate the quality regression task. However, from Table 2 we can see a the SROCC has a slight drop for SJTU datasets after removing the auxiliary task. In summary, depth and normal modalities contribute essential geometric details to enhance the structural integrity of the point cloud. Visual saliency functions as a refinement mechanism, elevating prediction accuracy across all aspects. The efficacy of an additional distortion type classification task is contingent upon the dataset's specific characteristics. Notably, within the proposed framework, the multi-modal completion yields superior performance gains compared to the other two components.

4.6 Cross-dataset Evaluation

To gauge the generalization capability of the proposed ViSam-PCQA, cross-dataset evaluations were conducted. Our approach involves training the model on the entire dataset and testing it on all data from another dataset. The resulting performance metrics, presented in Table 3, demonstrate the model's ability to generalize across different datasets. Notably, ViSam-PCQA exhibits superior generalization compared to other learning-based models, for example, GPA-Net and MM-PCQA, their performance for SJTU→WPC and WPC→SJTU are 0.424/0.431 and 0.535/0.574, 0.430/0.459 and 0.769/0.778 respectively. Surprisingly, training on the WPC dataset and testing on the SJTU dataset yields even better performance than certain FR-PCQA and NR-PCQA metrics on the SJTU test set, indicating a robust generalization tendency. However, training on BASICS and testing on WPC gets the lowest performance, that's mainly because WPC only contains objects, and the BASICS dataset contains learning-based compression distortion types.

5 Conclusion

In this paper, we introduce an innovative visual saliency-guided multi-modal metric for image-based no-reference Point Cloud Quality Assessment (ViSam-PCQA). ViSam-PCQA incorporates a saliency map derived from a pre-trained model into the learning process to enhance prediction accuracy. Specifically, we capture multi-modal information in the form of texture, depth, and normal maps, which are then fed into an image encoder to obtain low- and high-level features, offering a comprehensive description of the point cloud. The low-level feature map from the texture map is refined using a corresponding depth-guided saliency map, enabling the neural network to select the salient region by weighting the feature map instead of directly applying it to the input texture map. The high-level feature undergoes the Transformer to extract global features for individual modalities and establish local correspondence across the three modalities. A quality regression and distortion type classification are employed to generate an overall quality score for the distorted point clouds. The proposed metric is evaluated on three publicly available datasets, demonstrating a substantial improvement across all evaluation criteria and achieving a new state-of-the-art performance.

Acknowledgments

This work was supported through the NWO WISE grant and the European Commission Horizon Europe program, under the grant agreement 101070109, *TRANSMIXR* https://transmixr.eu/. Funded by the European Union. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-7083.

References

[1] Ali Ak, Emin Zerman, Maurice Quach, Aladine Chetouani, Aljosa Smolic, Giuseppe Valenzise, and Patrick Le Callet. 2024. BASICS: Broad quality assessment of static point clouds in a compression scenario. *IEEE Transactions on Multimedia* (2024).

[2] Evangelos Alexiou and Touradj Ebrahimi. 2019. Exploiting user interactivity in quality assessment of point cloud imaging. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX.2019.8743277

[3] Evangelos Alexiou and Touradj Ebrahimi. 2020. Towards a point cloud structural similarity metric. In *ICMEW*. 1–6.

[4] Evangelos Alexiou, Peisen Xu, and Touradj Ebrahimi. 2019. Towards Modelling of Visual Saliency in Point Clouds for Immersive Applications. In *2019 IEEE International Conference on Image Processing (ICIP)*. 4325–4329. https://doi.org/10.1109/ICIP.2019.8803479

[5] Evangelos Alexiou, Xuemei Zhou, Irene Viola, and Pablo Cesar. 2021. PointPCA: Point cloud objective quality assessment using PCA-based descriptors. *arXiv preprint arXiv:2111.12663* (2021).

[6] Ali Borji and Laurent Itti. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 185–207. https://doi.org/10.1109/TPAMI.2012.89

[7] Salima Bourbia, Ayoub Karine, Aladine Chetouani, Mohammed El Hassouni, and Maher Jridi. 2022. No-reference point clouds quality assessment using transformer and visual saliency. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*. 57–62.

[8] Ricardo L. de Queiroz and Philip A. Chou. 2017. Motion-Compensated Compression of Dynamic Voxelized Point Clouds. *IEEE Transactions on Image Processing* 26, 8 (2017), 3886–3895. https://doi.org/10.1109/TIP.2017.2707807

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF CVPR*. 248–255.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE/CVF CVPR*. 770–778.

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[13] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[14] Abdelouahed Laazoufi and Mohammed El Hassouni. 2022. Saliency-based point cloud quality assessment method using aware features learning. In *2022 9th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. 1–5. https://doi.org/10.1109/WINCOM55661.2022.9966464

[15] Abdelouahed Laazoufi and Mohammed El Hassouni. 2022. Saliency-based point cloud quality assessment method using aware features learning. In *2022 9th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE, 1–5.

[16] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. 2016. Deep saliency with encoded low level distance map and high level features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 660–668.

[17] Haojia Lin, Xiawu Zheng, Lijiang Li, Fei Chao, Shanshan Wang, Yan Wang, Yonghong Tian, and Rongrong Ji. 2023. Meta Architecture for Point Cloud Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17682–17691.

[18] Weisi Lin and C-C Jay Kuo. 2011. Perceptual visual quality metrics: A survey. *Journal of visual communication and image representation* 22, 4 (2011), 297–312.

[19] Weisi Lin, Sanghoon Lee, et al. 2022. Visual saliency and quality evaluation for 3D point clouds and meshes: An overview. *APSIPA Transactions on Signal and Information Processing* 11, 1 (2022).

[20] Qi Liu, Yiyun Liu, Honglei Su, Hui Yuan, and Raouf Hamzaoui. 2022. Progressive Knowledge Transfer Based on Human Visual Perception Mechanism for Perceptual Quality Assessment of Point Clouds. *arXiv preprint arXiv:2211.16646* (2022).

[21] Qi Liu, Honglei Su, Zhengfang Duanmu, Wentao Liu, and Zhou Wang. 2022. Perceptual Quality Assessment of Colored 3D Point Clouds. *IEEE TVCG* (2022).

[22] Qi Liu, Hui Yuan, Honglei Su, Hao Liu, Yu Wang, Huan Yang, and Junhui Hou. 2021. PQA-Net: Deep No Reference Point Cloud Quality Assessment via Multi-View Projection. *IEEE TCSVT* 31, 12 (2021), 4645–4660.

[23] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. 2022. TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing* 494 (2022), 455–467.

[24] R Mekuria, Z Li, C Tulvan, and P Chou. 2016. Evaluation criteria for point cloud compression. *ISO/IEC MPEG* 16332 (2016).

[25] María Antonia Diaz Mendoza, Emiro De La Hoz Franco, and Jorge Eliecer Gómez Gómez. 2023. Technologies for the Preservation of Cultural Heritage—A Systematic Review of the Literature. *Sustainability* 15, 2 (2023), 1059.

[26] Gabriel Meynet, Yana Nehmé, Julie Digne, and Guillaume Lavoué. 2020. PCQM: A full-reference quality metric for colored 3D point clouds. In *QoMEX*. 1–6.

[27] Mario Montagud, Jie Li, Gianluca Cernigliaro, Abdallah El Ali, Sergi Fernández, and Pablo Cesar. 2022. Towards socialVR: evaluating a novel technology for watching videos together. *Virtual Reality* 26, 4 (2022), 1593–1613.

[28] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. 2020. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8839–8848.

[29] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. 2022. A deep learning based no-reference quality assessment model for ugc videos. In *ACM MM*. 856–865.

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

[31] VEQG. [n. d.]. Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment. *[online]. Availabel httpwww.its.bldrdoc.govvqegvqeg-home.aspx* ([n. d.]).

[32] Irene Viola, Jack Jansen, Shishir Subramanyam, Ignacio Reimat, and Pablo Cesar. 2023. VR2Gather: A Collaborative, Social Virtual Reality System for Adaptive, Multiparty Real-Time Communication. *IEEE MultiMedia* 30, 2 (2023), 48–59. https://doi.org/10.1109/MMUL.2023.3263943

[33] Songtao Wang, Xiaoqi Wang, Hao Gao, and Jian Xiong. 2023. Non-Local Geometry and Color Gradient Aggregation Graph Model for No-Reference Point Cloud Quality Assessment. In *ACM MM*. 6803–6810.

[34] Zhengyu Wang, Yujie Zhang, Qi Yang, Yiling Xu, Jun Sun, and Shan Liu. 2022. Point cloud quality assessment using 3D saliency maps. *arXiv preprint arXiv:2209.15475* (2022).

[35] Xinju Wu, Yun Zhang, Chunling Fan, Junhui Hou, and Sam Kwong. 2021. Subjective Quality Database and Objective Study of Compressed Point Clouds With 6DoF Head-Mounted Display. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 12 (2021), 4630–4644. https://doi.org/10.1109/TCSVT.2021.3101484

[36] Wuyuan Xie, Kaimin Wang, Yakun Ju, and Miaohui Wang. 2023. pmBQA: Projection-based Blind Point Cloud Quality Assessment via Multimodal Learning. In *ACM MM*. 3250–3258.

[37] Qi Yang, Hao Chen, Zhan Ma, Yiling Xu, Rongjun Tang, and Jun Sun. 2020. Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration. *IEEE Transactions on Multimedia* (2020), 1–1. https://doi.org/10.1109/TMM.2020.3033117

[38] Qi Yang, Yipeng Liu, Siheng Chen, Yiling Xu, and Jun Sun. 2022. No-Reference Point Cloud Quality Assessment via Domain Adaptation. In *IEEE/CVF CVPR*. 21179–21188.

[39] Qi Yang, Zhan Ma, Yiling Xu, Zhu Li, and Jun Sun. 2020. Inferring point cloud quality via graph similarity. *IEEE TAMI* (2020).

[40] Lin Zhang, Ying Shen, and Hongyu Li. 2014. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing* 23, 10 (2014), 4270–4281.

[41] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. 2017. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5287–5295.

[42] Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, and Guangtao Zhai. 2022. No-reference quality assessment for 3d colored point cloud and mesh models. *IEEE TCSVT* (2022).

[43] Zicheng Zhang, Wei Sun, Xiongkuo Min, Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai. 2023. MM-PCQA: Multi-Modal Learning for No-reference Point Cloud Quality Assessment. In *IJCAI*.

[44] Zicheng Zhang, Wei Sun, Houning Wu, Yingjie Zhou, Chunyi Li, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. 2023. GMS-3DQA: Projection-based Grid Minipatch Sampling for 3D Model Quality Assessment. *arXiv preprint arXiv:2306.05658* (2023).

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

[46] Wei Zhou, Guanghui Yue, Ruizeng Zhang, Yipeng Qin, and Hantao Liu. 2023. Reduced-reference quality assessment of point clouds via content-oriented saliency projection. *IEEE Signal Processing Letters* 30 (2023), 354–358.

[47] Xuemei Zhou, Evangelos Alexiou, Irene Viola, and Pablo Cesar. 2023. PointPCA+: Extending PointPCA Objective Quality Assessment Metric. In *2023 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*. 1–5. https://doi.org/10.1109/ICIPC59416.2023.10328338

[48] Xuemei Zhou, Yun Zhang, Na Li, Xu Wang, Yang Zhou, and Yo-Sung Ho. 2021. Projection invariant feature and visual saliency-based stereoscopic omnidirectional image quality assessment. *IEEE Transactions on Broadcasting* 67, 2 (2021), 512–523.