Balancing Representation Abstractions and Local Details Preservation for 3D Point Cloud Quality Assessment

Marouane Tliba, Aladine Chetouani, Giuseppe Valenzise, Frédéric Dufaux

▶ **To cite this version:**

Marouane Tliba, Aladine Chetouani, Giuseppe Valenzise, Frédéric Dufaux. Balancing Representation Abstractions and Local Details Preservation for 3D Point Cloud Quality Assessment. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024), IEEE, Apr 2024, Séoul, South Korea. hal-04413630

HAL Id: hal-04413630
https://hal.science/hal-04413630v1

Submitted on 20 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BALANCING REPRESENTATION ABSTRACTIONS AND LOCAL DETAILS PRESERVATION FOR 3D POINT CLOUD QUALITY ASSESSMENT

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

3D Point Clouds (PCs) have become a valuable tool for representing intricate 3D information. Assessing the quality of PCs remains a challenging task, especially when striving for optimal immersive experiences. This paper introduces a novel metric and training approach that leverages projection-based views to evaluate the quality of 3D content. Our approach addresses a critical issue related to the intrinsic bias of deep networks for image recognition towards building hierarchical representations including only the global semantic, at the expense of local details. This bias is a limiting factor in tasks like 3D point cloud quality assessment where instances of the same content with varying degrees and types of degradation can possess strikingly similar representations. We propose a novel point cloud quality metric using a dual supervised and unsupervised training strategy to balance semantic understanding and preservation of critical perceptual quality-relevant information. The results demonstrate the effectiveness and reliability of our solution compared to state-of-the-art metrics on two standard 3D PCs quality assessment benchmarks (3D PCQA). The source code is available at https://github.com/mtliba/PQCA-ICASSP2024

Index Terms— 3D Point Clouds, Image Quality Assessment, Representation Learning , Self-Supervised Learning.

1. INTRODUCTION

In recent years, the widespread adoption of 3D imaging technologies and 3D Point Clouds (PCs) [1] has stimulated the development of applications such as augmented reality and the metaverse [2], offering enhanced detail and realism [3]. However, these point clouds often suffer degradations during acquisition and transmission, and require in practice lossy compression to manage storage and bandwidth constraints. This makes reliable no-reference quality metrics crucial for ensuring real-time immersive experiences. While subjective tests remain the gold standard for evaluating 3D PCs quality, they typically involve subjects viewing 2D projections of the PCs [4, 5]. Integrating this practice to develop a No-reference metric serves a dual purpose: it mirrors the subjective assessment setup for obtaining the ground truth Mean Opinion Score (MOS) and leverages deep learning techniques developed for 2D images, facilitating advancements in objective 3D Quality Measurements.

To this end, various objective metrics have been developed for 3D Point Clouds Quality Assessment (PCQA). These methods can be categorized into three main groups: Point-based, Feature-based, and Projection-based metrics. **Point-based metrics**, including Point-to-Point (Po2Po) [6], Point-to-Plane (Po2Pl) [7], Plane-to-Plane (Pl2Pl) [8], evaluate quality by comparing geometric and feature distances between the reference and distorted point clouds. **Feature-based metrics** analyze point-wise geometry and associated attributes, either globally or locally. Examples include PCQM [9], a metric combining geometry and color features; and PointSSMIM [10], which explores the applicability of the Structural Similarity (SSIM) index in higher-dimensional, irregular spaces.

The concept of Projection-based metrics involves the projection of a 3D point cloud onto a 2D plane to evaluate its quality [11, 12, 13, 14, 15]. While this projection introduces minor distortions, leveraging advancements in 2D computer vision remains a significant step forward [16, 17]. However, there are limitations when applying deep networks designed for image recognition in this context. The noise in projected point clouds is relatively subtle and affects only a small portion of the 2D plane. Moreover, the hierarchical, abstracted representations built by deep recognition models become problematic for tasks like 3D PCQA. Conventional models tend to abstract away from local information [17], resulting in a concerning issue: images with various degradation levels and types can have remarkably similar representations, making fine-grained perceptual distinctions nearly impossible.

In this paper, we introduce a novel 3D PCQA metric trained using an innovative strategy combining supervised and unsupervised learning that effectively leverages deep recognition models. This strategy comprises two key components: **Maximizing Intermediate-Layer Similarity**: We enhance the similarity between intermediate layers and the final representation, enriching the feature representation latent space with essential low-level information crucial for quality assessment. **Quality-Driven Self-Supervised Loss**: We introduce a tailored self-supervised loss component to meet quality assessment requirements, achieved through the use of quality-invariant augmentations during training.

2. PROPOSED METHOD:

Our approach aims to create an innovative dual training strategy for 3D PCQA using 2D projected views using supervised

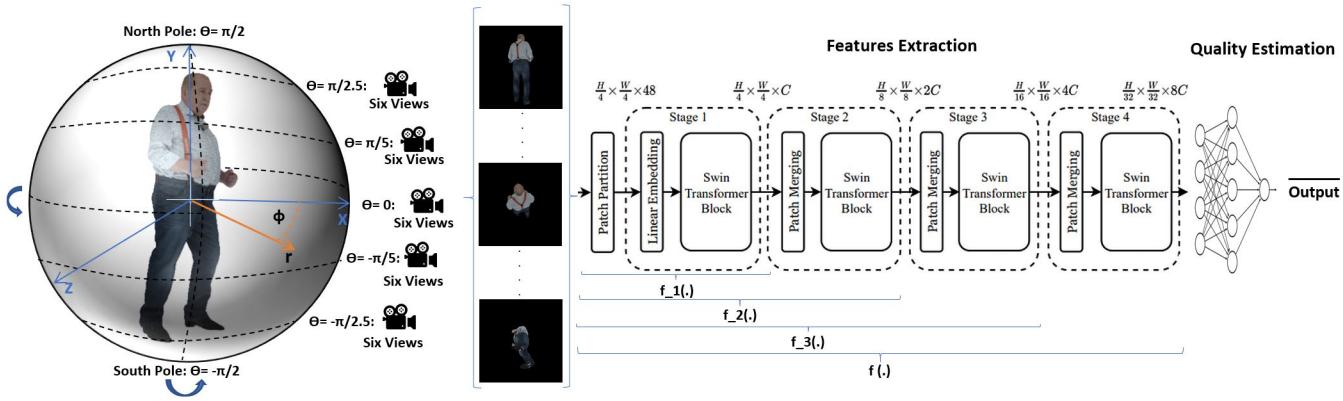

Fig. 1. Overview of our proposed framework for 3D Point Cloud Perceptual Quality prediction. The pipeline illustrates the preprocessing, feature extraction, and quality estimation stages: the quality is obtained as the mean overall views.

and unsupervised learning. We focus on maintaining global semantics while preserving important details. By maximizing intermediate-layer similarity with the global representation, we ensure that crucial early-layer details are consistently represented and refined in the final output. Additionally, we use self-supervised learning with random view rotations to maintain view semantics without affecting perceptual quality characteristics. As depicted in Fig. 1, our method can be summarized in the following steps: view projection, feature extraction, and a novel training strategy.

2.1. 3D Point Cloud Views Projection

Our initial step involves converting distorted 3D point cloud (PC) objects into various 2D viewpoints using perspective projection, simulating human visual perception during quality assessment. We utilize multiple virtual cameras at various angles to capture the PCs from different perspectives. The 3D object's centroid is marked as the origin of a spherical coordinate system (r, θ, ϕ), where r represents the radius, adjusted based on each PC's dimensions for precise capture. θ and ϕ are elevation and azimuth angles, respectively, ranging from [0, 2π]. The virtual camera coordinates are systematically defined: azimuth angles vary by $\frac{\pi}{3}$, and elevation angles are assigned values of 0, $\frac{\pi}{5}$, $\frac{\pi}{2.5}$, $-\frac{\pi}{5}$, and $-\frac{\pi}{2.5}$. This setup results in five distinct camera trajectories, each capturing six unique views, totaling 30 images for each distorted 3D point cloud. In summary, given a 3D Point Cloud PC, we obtain rendered projections as $X = \psi(PC)$, where X is the set of 30 rendered projections $(x_j, j = 1, \ldots, 30)$, and $\psi(\cdot)$ represents the rendering process.

2.2. Feature Extraction

We employ the Swin Transformer [18] for feature extraction, known for capturing long-range details and providing rich hierarchical semantic representations. The model processes input images, splitting them into non-overlapping patches. We use the "Base" variant, dividing images into 4×4 patches, embedding them into C equal to 128-dimensional vectors. These patches go through stages with specific layer counts $\{2, 2, 18, 2\}$ to hierarchically extract features. During attention, a 12×12 window balances attention specificity and computational efficiency. This choice of the Swin Transformer aligns with our goal of preserving relevant details and maintaining global context. Integrating its capabilities of capturing a long range of dependencies with our novel training strategy, we tackle the challenge of accentuating crucial low-level details information, while retaining a global context for an effective 3D PCQA using 2D projection views.

2.3. Training Strategy

Our proposed dual training strategy consists of optimizing supervised and unsupervised losses. The goal of the **supervised loss** is to create a mapping function parameterized by θ that captures the relationship between the feature representation latent space $f(X; \theta)$ and the distribution of the subjective quality scores S, through a shallow Multi-Layer Perceptron (MLP). Our supervised loss over n PCs is :

$$L_{\text{supervised}} = \frac{1}{n \times 30} \sum_{j=1}^{n \times 30} \|\text{MLP}(f(x_j; \theta)) - S_j\|^2 \quad (1)$$

Our **unsupervised loss** plays a crucial role in shaping a more robust feature representation latent space tailored for our downstream task of PCQA, as illustrated in Fig. 2. This loss function aims to maximize the similarity between intermediate stages and the final representations, leveraging Negative Cosine Similarity (NCS) loss as depicted in graphs (a) and (b) Fig. 2. In graph (a), we align the intermediate stages with the final representation, while in graph (b), we perform the reverse process. Here, $f_i(x : \theta)$ corresponds to the feature representation of each stage i. It is worth noting that NCS here is applied between partially shared networks, where f encapsulates the parameters of f_i, also, both f and f_i are applied to the same input x and not to an augmented version. Following [19, 20], each stage similarity loss can be expressed as symmetrized NCS with a stop gradient setting [20]:

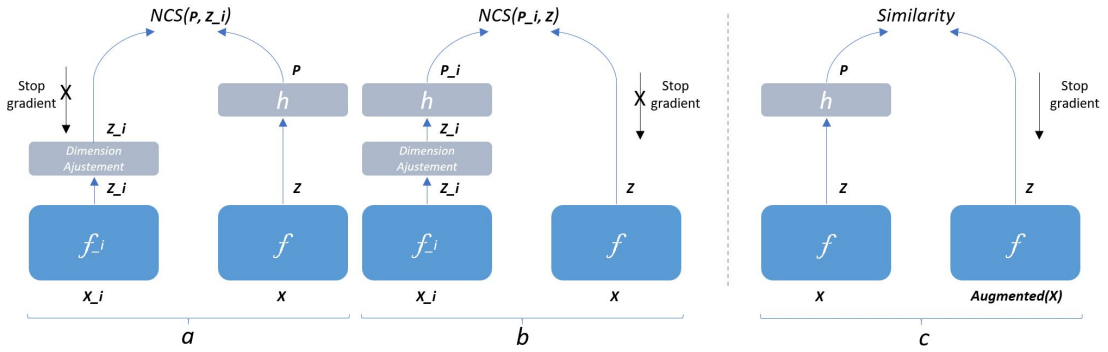

Fig. 2. Illustration of our unsupervised loss strategy. Graphs (a) and (b) show the use of Negative Cosine Similarity (NCS) to align intermediate and final representations. While (a) aligns intermediate stages to the final stage, (b) does the reverse, the dimension adjustment map z_i to the dimension of z. Graph (c) applies the Simsiam loss on a projected view and its augmentation.

$$L_{\text{i}} = \frac{1}{2}\left(\text{NCS}(p, \text{stopgrad}(z_i)) + \text{NCS}(p_i, \text{stopgrad}(z))\right) \quad (2)$$

where the NCS is expressed as: $NCS(a,b) = -\frac{a \cdot b}{\|a\|_2 \cdot \|b\|_2}$, p is the result of the prediction MLP head, denoted as h [19], which is used to match the output of f and f_i. z represents the final representation, and z_i is the intermediate representation. This loss encourages the similarity between the feature representations at different stages of the network and the last representation, fostering a richer latent space encapsulating low-level information which are perceptually important.

Furthermore, we maximize the similarity of representations for the same view with various rotation angles by employing the SimSiam loss [19]:

$$L_{\text{simsiam}} = \frac{1}{2}\left(\text{NCS}(p1, \text{stopgrad}(z2)) + \text{NCS}(p2, \text{stopgrad}(z1))\right). \quad (3)$$

In our approach, we have chosen to utilize rotation as the sole augmentation strategy due to its quality-invariant properties, ensuring that the global representation consistently reflects quality. This process is highlighted in Fig. 2, graph (c). Here the encoder f shares weights between the two views [19]. $p = h(f(x1))$ and $z = f(x2)$. The prediction head h [19] transforms the output of one view and matches the other rotated view. The overall optimization objective for the dual training can be expressed as a combination of the supervised and unsupervised losses, with a weighting parameter λ :

$$L = L_{\text{supervised}} + \lambda \left(L_{\text{simsiam}} + \sum_{i=1}^{4} L_{\text{i}} \right). \quad (4)$$

3. RESULT ANALYSIS

We evaluated the effectiveness of our model on two public benchmarks that use subjective scores and involve various degradation types, including compression and noise. **ICIP2020** [21] consists of six reference point clouds, generating 90 degraded versions. These variations involve three main compression techniques: V-PCC, G-PCC with triangle soup coding, and octree coding. **SJTU** [28] consists of 9 reference point clouds, generating a total of 378 distorted versions arising from six degradation modalities: Octree-based compression (OT), Color Noise (CN), Downscaling (DS),(GG)Geometry Gaussian noise, and combination as (D+C), (D+G), (C+G).

Table 1. Results obtained on ICIP20[21] dataset using 6-folds

index	Model	PLCC ↑	SROCC ↑
m,f	po2point MSE [6]	0.946	0.950
m,f	po2plane MSE [7]	0.945	0.959
m,f	color Y PSNR [22]	0.887	0.892
m,f	pl2plane AVG [8]	0.922	0.910
m,f	pl2plane MSE [8]	0.925	0.912
m,f	PCQM [9]	0.796	0.832
m,f	GraphSim [23]	0.931	0.893
m,n	PointNet-SSNR [24]	0.908	0.955
m,f	PointNet-DCCFR [25]	0.947	0.973
m,n	PointNet-Graph [26]	0.946	0.973
p,n	MultiModal [27]	0.945	**0.978**
p,n	Swin Supervised	0.918	0.914
p,n	Ours	**0.965**	0.970

We applied k-fold cross-validation by dividing the database into k equally sized segments. The model was trained on $k-1$ segments and tested on one. The procedure is repeated k times with different test segments. The final performance metric was calculated as the average of these k-fold cross-validation results. We conducted a comprehensive evaluation of our approach by comparing it with state-of-the-art methods, This comparison encompassed deep learning-based and not-deep-based methods, operating either on 3D point clouds, or 2D projected views (indexed as m, and p, respectively; f, n indexes refer to Full, and No reference respectively). The selection of deep-based methods was based on those already trained and tested in their original benchmark settings on ICIP, and SJTU using k-fold cross-validation. Pearson Correlation Coefficient (PLCC) and Spearman Correlation Coefficient (SROCC) are considered to report the quality prediction ability of all meth-

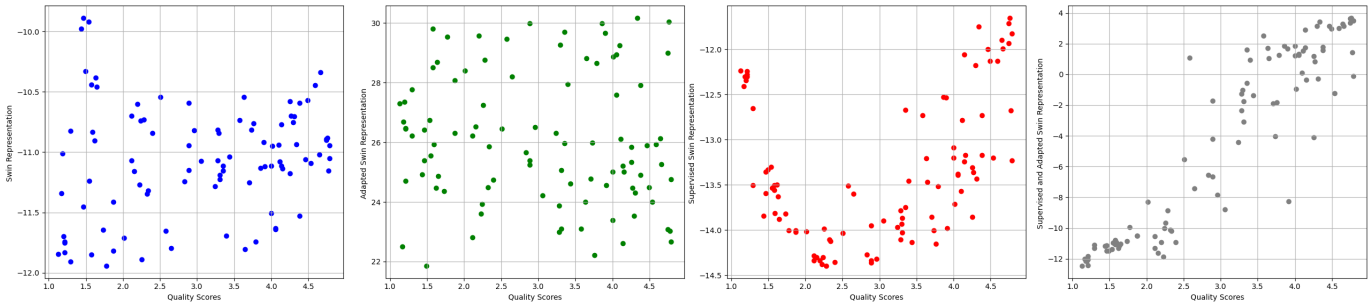

Fig. 3. PCA-based visualization of Swin Transformer features representations distribution from various training strategies on the ICIP20 dataset against the subjective MOS. The four plots, from left to right, display: (1) Trained for image recognition, (2) Our self-supervised training strategy, (3) Supervised training for quality assessment on ICIP, and (4) our combined approach using both supervised and self-supervised losses trained on ICIP for PCQA.

Table 2. Results obtained on SJTU dataset using 9-folds[28]

index	Model	PLCC ↑	SROCC ↑
m,f	po2point MSE [6]	0.812	0.729
m,f	po2plane MSE[7]	0.594	0.628
m,f	color Y PSNR [22]	0.817	0.795
m,f	PCQM[9]	0.885	0.864
m,f	GraphSIM [23]	0.845	0.878
m,f	PointSSIM[10]	0.714	0.687
m,n	PointGraph[26]	0.903	0.873
p,n	3D-NSS[11]	0.714	0.738
p,n	3DResnet [13]	0.861	0.832
p,n	VQA_PC [12]	0.864	0.851
p,n	Ours	**0.915**	**0.908**

ods.

Table 1 presents the performance of our proposed method on the ICIP dataset. Our method demonstrates a high correlation with the subjective MOS ground truth, achieving a PLCC of 0.965 and an SROCC of 0.970. These results outperform all other metrics in terms of PLCC and are competitive in terms of SROCC. To further validate the impact of the unsupervised loss, we conducted a training experiment using only the supervised loss, labeled as "Swin Supervised" in Table 1. The performance noticeably decreased, highlighting a gap in both PLCC and SROCC. Table 2 shows the performance of our method on SJTU which represents an intricate distribution of degradation and noises, as can be seen, our method is achieving a high correlation overcoming state-of-the-art methods on both PLCC and SORCC. The results obtained on both datasets show the effectiveness of our novel proposed training strategy.

To further validate our method, we visualize the network's representations in Fig. 3 using Principal Component Analysis against the subjective MOS. We explored four configurations: (1) pre-trained weights on Imagenet 22k, (2) retraining the network on ICIP2020 with unsupervised loss only, (3) retraining the network on ICIP2020 for quality assessment using supervised loss only, and (4) retraining the network on ICIP2020 for quality assessment using both supervised and unsupervised loss. In configuration (1), the model struggled to discern nuances between different image qualities, resulting in a representation tightly clustered within a small range of 2 points (-12 to -10), lacking a linear correlation with quality scores. Configuration (2) produced a noticeably distributed representation spanning 8 points (from 22 to 30) but showed limited correlation with quality scores, indicating the model's ability to differentiate image qualities but failing to align them accurately with quality scores. Configuration (3) exhibited a slightly correlated representation with quality scores but remained confined to a narrow range (-14.5 to -12). In contrast, configuration (4) displayed a well-distributed representation spanning (-12 to 4) within a larger range of 16 points, and showed a strong linear correlation with quality scores, which has not been noted in all previous configurations. This visualization provides evidence supporting the efficacy of our novel training strategy in improving point cloud quality assessment. Our choice to employ the ICIP dataset for visualization is based on its composition of 3D point cloud (PC) objects predominantly depicting human figures. This shared semantic representation across the dataset aids in discerning how shifts in the network's representation are indeed attributable to variations in quality features, rather than differences in the underlying semantic content.

4. CONCLUSION

Our work introduces an innovative metric and training approach for 3D PCs quality assessment. We address the challenge of preserving fine-grained details while maintaining overall image context, deviating from traditional deep network biases. Our approach leverages 2D projections and combines supervised and unsupervised training methods to achieve a balance between semantic understanding and preserving critical visual quality-relevant information. Extensive testing on benchmark datasets demonstrates the potential of our approach for improved 3D PCQA.

5. REFERENCES

[1] Y. Guo et al., "Deep learning for 3d point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[2] A. P. Placitelli et al., "Low-cost augmented reality systems via 3d point cloud sensors," in *2011 Seventh International Conference on Signal Image Technology Internet-Based Systems*, 2011, pp. 188–192.

[3] L. Han et al., "Live semantic 3d perception for immersive augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 2012–2022, 2020.

[4] E. Alexiou et al., "Point cloud subjective evaluation methodology based on 2d rendering," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, 2018, pp. 1–6.

[5] L.A. da Silva Cruz et al., "Point cloud quality evaluation: Towards a definition for test conditions," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–6.

[6] R. Mekuriaet et al., "Evaluation criteria for PCC (point cloud compression)," in *in ISO/IEC MPEG Doc*. N16332, 2016., vol. II, pp. 803–806.

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

[8] E. Alexiou et al., "Point cloud quality assessment metric based on angular similarity," in *IEEE ICME-W*, 2018.

[9] G. Meynet et al., "Pcqm: A full-reference quality metric for colored 3d point clouds," 2020.

[10] E. Alexiou et al., "Towards a point cloud structural similarity metric," in *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[11] Z. Zhang et al., "No-reference quality assessment for 3d colored point cloud and mesh models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7618–7631, 2022.

[12] Z. Zhang et al., "Treating point cloud as moving camera videos: A no-reference quality assessment metric," *arXiv preprint arXiv:2208.14085*, 2022.

[13] Y. Fan et al., "A no-reference quality assessment metric for point cloud based on captured video sequences," *arXiv preprint arXiv:2208.14085*, 2022.

[14] S. Bourbia, A. Karine, A. Chetouani, M. El Hassouni, and M. Jridi, "No-reference 3d point cloud quality assessment using multi-view projection and deep convolutional neural network," *IEEE Access*, vol. 11, pp. 26759–26772, 2023.

[15] A. Chetouani et al., "Convolutional Neural Network for 3D Point Cloud Quality Assessment with Reference," in *IEEE MMSP*, Tampere, Finland, Oct. 2021.

[16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *2016 IEEE (CVPR)*, pp. 770–778, 2015.

[17] M Tliba et al., "Satsal: A multi-level self-attention based architecture for visual saliency prediction," 2022, vol. 10, pp. 20701–20713.

[18] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10002.

[19] X. Chen et al., "Exploring simple siamese representation learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2020.

[20] JB. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," *ArXiv*, vol. abs/2006.07733, 2020.

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

[22] E.M Torlig et al., "A novel methodology for quality assessment of voxelized point clouds," in *Optical Engineering + Applications*, 2018.

[23] Q. Yang et al., "Inferring point cloud quality via graph similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 3015–3029, 2022.

[24] M. Tliba et al., "Representation learning optimization for 3d point cloud quality assessment without reference," in *2022 IEEE (ICIP)*, 2022, pp. 3702–3706.

[25] M. Tliba et al., "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.

[26] M. Tliba et al., "Pcqa-graphpoint: Efficient deep-based graph metric for point cloud quality assessment," in *ICASSP 2023 - 2023 IEEE (ICASSP)*, 2023, pp. 1–5.

[27] M. Tliba et al., "Multi-modal evaluation of 3d point clouds images: A novel no-reference approach using a multi-stream attentive architecture," in *(EUSIPCO 2023)*, 2023.

[28] Q. Yang et al., "Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3877–3891, 2021.