Point Cloud Quality Assessment via Multi-View Structure-Aware Feature Fusion

Jian Xiong[1] *, **Lingxia Jiang**[1], **Xianzhong Long**[1], **Miaohui Wang**[2], **Hao Gao**[1] *

[1]Nanjing University of Posts and Telecommunications
[2]Shenzhen University
{jxiong, 1022010437, lxz}@njupt.edu.cn, wang.miaohui@gmail.com, tsgaohao@gmail.com

Abstract

Point cloud quality assessment (PCQA) is essential for reliable 3D visual applications. While point-based methods face challenges in characterizing distortions due to point cloud disorder, projection-based approaches offer better efficiency but suffer from geometric distortion insensitivity and texture representation blind spots. This study proposes SAF-Net, a multi-view structure-aware feature fusion network for PCQA. We first identify two key limitations in projection-based methods: insufficient geometric distortion perception and representation blind spots (RBS) in texture images. To address these issues, SAF-Net innovatively integrates object mask maps and local binary pattern (LBP) maps. The mask maps enhance geometric distortion perception by extracting edge sharpness and curvature variations, while LBP maps capture essential structural information to overcome RBS and align with human visual system (HVS) sensitivity. SAF-Net employs a hybrid CNN-ViT architecture to balance local feature extraction and global context modeling, along with a progressive fusion strategy to optimize cross-modal feature interaction. Extensive experiments demonstrate the superior performance of SAF-Net on multiple benchmarks, establishing new state-of-the-art results in PCQA.

Introduction

As a crucial representation of 3D objects, point clouds find extensive applications in virtual reality, immersive communication, etc. Point Cloud Quality Assessment (PCQA), which quantifies perceptual quality consistent with the human visual system (HVS), plays a pivotal role in enhancing the reliability and user experience of 3D applications. Deep learning-based PCQA methods leverage the powerful representational capacity of deep neural networks (DNN) to automatically learn quality-aware features for complex distortions, mainly including point-based methods, projection-based methods, and multimodal fusion-based methods.

Point-based methods directly process raw point clouds for feature extraction, yet face two fundamental challenges: (1) the inherent disorder of point clouds complicates spatial topology modeling, and (2) computational constraints typically limit processing to local patches, hindering comprehensive quality assessment.

*Corresponding authors: Jian Xiong and Hao Gao

Projection-based methods transform unordered point clouds into regular 2D images for feature learning using convolutional neural networks (CNNs) (He et al. 2015) or Vision Transformers (ViTs) (Dosovitskiy et al. 2021), offering three key advantages: *computational efficiency* by avoiding direct processing of unstructured data, *comprehensive scene coverage* through multi-view projection, and *transfer learning potential* via pre-trained models. However, they suffer from inherent limitations: depth loss and occlusion during projection that disrupt 3D topological structures, leading to impaired perception of geometric distortions. Particularly, the absence of background information in projected images leads CNN features to over-emphasize object contours while neglecting internal distortions like point density anomalies, creating "representational blind spots (RBS)".

Multimodal fusion-based methods attempt to address these geometric perception deficiencies by jointly processing projected images and point cloud patches. Nevertheless, they introduce new challenges: (1) substantially increased computational overhead from dual-modal processing, and (2) unresolved optimal fusion strategies for heterogeneous data representations, with insignificant performance advantages demonstrated over single-modal methods.

Therefore, considering the advantages of projection-based methods in computational efficiency, comprehensive scene coverage, and transfer learning potential, this study focuses on the projection-based PCQA paradigm and proposes a multi-view structure-aware feature fusion network, namely SAF-Net. First, this work analyzes two critical limitations in projection-based PCQA methods: insufficient geometric distortion perception and the RBS issue. To address these limitations, we innovatively integrate object mask maps and local binary pattern (LBP) maps with the projected texture images. The mask maps enhance geometric distortion perception by extracting edge sharpness and curvature variations, while LBP maps capture local structural information to overcome the RBS issue and align with HVS sensitivity. SAF-Net employs a CNN-ViT hybrid architecture that leverages CNN's effectiveness in local feature extraction and ViT's superiority in global feature encoding. A two-stage feature fusion mechanism is developed to optimize cross-modal interaction and improve quality perception performance. Experimental results demonstrate that SAF-Net achieves superior performance across multiple benchmark datasets. Contribu-

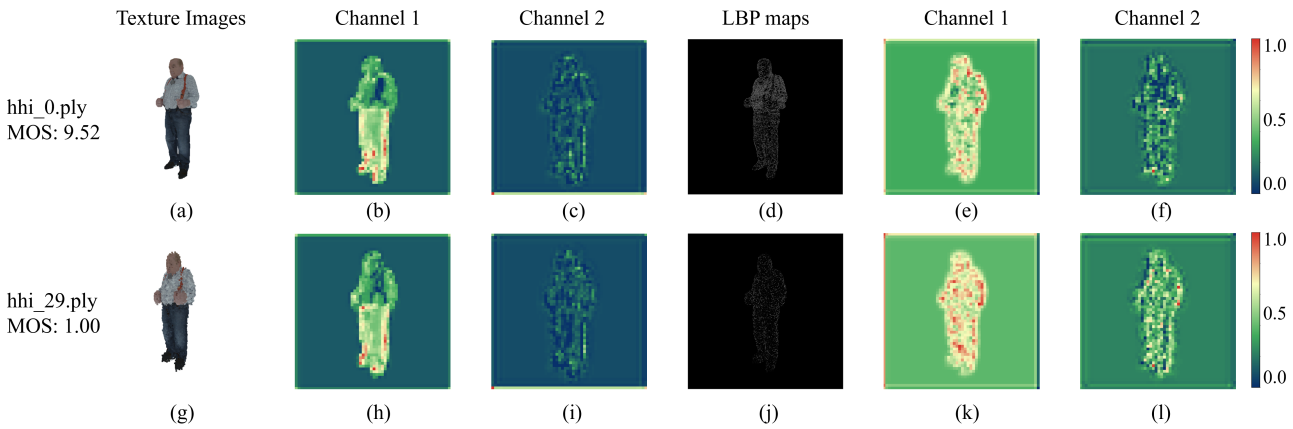

Figure 1: Texture images, LBP maps, and corresponding feature maps for point clouds with different quality levels.

tions can be summarized as follows.

1) We identify the limitations of projected texture images for PCQA, namely insufficient perception of geometric distortions and the RBS issue.

2) We propose a multi-view structure-aware feature fusion network, i.e., SAF-Net, that integrates object mask maps and LBP maps with texture images for projection-based PCQA.

3) Extensive experiments demonstrate the state-of-the-art (SOTA) performance of SAF-Net on multiple benchmarks.

Related Work

The point-based methods directly process raw point clouds for feature extraction. However, the inherent disorderliness of point clouds increases the difficulty of modeling spatial topology. Moreover, constrained by computational resources, existing methods are typically limited to processing local point cloud patches, making it challenging to comprehensively capture holistic quality characteristics (Xiong et al. 2023). For example, sparse convolution is used to model local correlations for nonempty voxels (Liu et al. 2023b). A fundamental limitation arises from the point sparsity constraining effective receptive field expansion. Graph signal processing approaches have been used to model local correlations in point clouds (Liu et al. 2022; Zhang, Yang, and Xu 2021), however, spatial topological relationships are often ignored in hierarchical feature aggregation. Graph convolution is used to exploit the graph topology to achieve information interaction and feature aggregation between unordered points (Shan et al. 2024a; Wang et al. 2023). However, the required nearest-neighbor searches and sparse matrix operations introduce computational bottlenecks that limit the scale of the processed point cloud patches.

The projection-based methods capture comprehensive scene information through multi-view projection and then utilize DNNs to extract quality-related features from 2D texture images for quality prediction. These methods typically leverage classic image feature encoders (Simonyan and Zisserman 2015) or IQA models to obtain quality-aware features. For efficiency, lightweight CNN models are adopted to improve computational efficiency (Yang et al. 2022a; Liu et al. 2021b). Moreover, due to the limited scale of annotated subjective datasets, which poses challenges for data-driven model training, domain adaptation techniques have been introduced to transfer knowledge from large-scale image subjective datasets to PCQA (Yang et al. 2022a). Additionally, self-supervised learning methods, such as contrastive learning (Shan et al. 2024b) and masked autoencoders (Shan et al. 2024c), have been explored to enhance model learning. In summary, these methods retain comprehensive scene information while avoiding the computational overhead of processing unordered points, and they facilitate transfer learning from 2D pre-trained models. Nevertheless, projection-induced destruction of topological structures may impair their ability to model geometric distortions accurately.

The multimodal fusion-based methods mitigate the geometric perception limitations of the projection-based methods by simultaneously processing projected images and point cloud patches. These methods typically employ 3D feature encoders, e.g., PointNet++ (Qi et al. 2017) and PointMLP (Ma et al. 2022), and 2D image encoders (e.g., ResNet (He et al. 2015) and ViT) to model geometric distortions and textural distortions, respectively (Chai et al. 2024; Zhang et al. 2022b), followed by feature fusion to predict the final quality score. However, the point cloud is partitioned into small patches to preserve local geometric structures, which compromises computational efficiency (Wu et al. 2025; Zhang et al. 2022b). Moreover, the multimodal feature fusion commonly relies on late-fusion strategies (e.g., cross-attention) (Chai et al. 2024; Zhang et al. 2022b; Wu et al. 2025). Although some studies have explored intermediate-fusion strategies to preserve the self-attention learning of individual modalities (Liu et al. 2024), the optimal fusion strategy for heterogeneous data representations remains unresolved, and these methods have yet to demonstrate significant performance advantages over single-modal approaches.

Motivation

Limitations of Projection-based PCQA

Previous projection-based methods primarily process projected texture images of point clouds to extract quality-

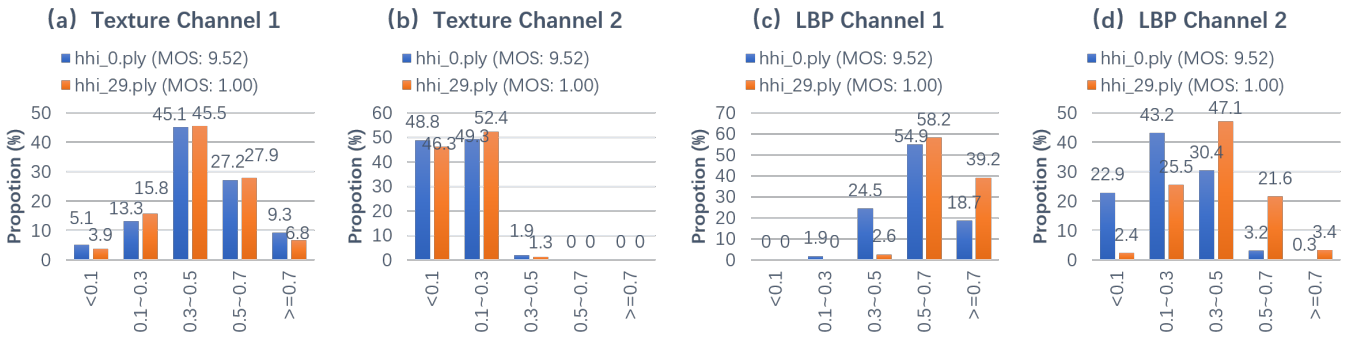

Figure 2: Histograms of the feature maps extracted from the texture images and LBP maps.

related features. However, the depth loss and occlusion during projection disrupt the 3D topological structure resulting in impaired perception of geometric distortions. Moreover, the projected images of point clouds typically lack background information, causing texture-based features to overly concentrate on object contours while neglecting internal details. This results in insensitivity to structural distortions caused by point density anomalies or geometric deformations, forming a "*representation blind spot*" (RBS).

To better understand RBS, we selected two point clouds from the SJTU-PCQA dataset (Yang et al. 2021) with the same scene but different quality levels, and compared the feature maps generated by the ResNet50 encoder (He et al. 2015) for the projected images. Figure 1 (a)-(c) show a projected image of hhi_0.ply (MOS = 9.52) and the corresponding feature maps of Channel 1 and Channel 2 from the first convolutional layer, respectively. Figure 1 (g)-(i) show a projected image of hhi_29.ply (MOS = 1) in the same view and the corresponding feature maps of the same channels, respectively. It shows that, regardless of point cloud quality, strong feature responses are predominantly distributed along the outer contour regions, while internal regions exhibit weaker responses, leading to reduced sensitivity to internal distortions. Further statistical results, as shown in Figure 2 (a) and (b), demonstrate that the histograms of the high- and low-quality feature maps are nearly identical, indicating that relying solely on texture-based features is insufficient for accurate PCQA.

Structure-aware Feature Fusion

To address these limitations and align with HVS sensitivity to structural information, e.g., contrast (Wang et al. 2004) and gradient (Xue et al. 2014), the proposed method incorporates LBP and mask maps as complementary features.

The core idea of LBP lies in its binary encoding of local neighborhood patterns, which quantifies gray-scale variations in projected images. Due to its sensitivity to low-contrast regions, LBP effectively captures internal distortions that traditional methods overlook. Figure 1 (d) and (j) show the LBP maps of (a) and (g), respectively. We also extracted features using ResNet50. Figure 1 (e) and (f) are the feature maps extracted from (d). Similarly, Figure 1 (k) and (l) are the feature maps extracted from (j). The resulting feature maps show that strong responses are uniformly distributed across internal regions rather than concentrated on contours. Moreover, the statistical results shown in Figure 2 (c) and (d), demonstrate that high- and low-quality point clouds exhibit different response patterns in the LBP-based feature maps: high-quality point clouds yield lower feature values, whereas low-quality ones produce higher values. Thus, incorporating LBP helps mitigate the RBS issue in texture-based projection methods.

Additionally, the lack of depth information in texture projections limits geometric distortion modeling. The object mask map explicitly describes the outer contour of objects and thus plays a crucial role in enhancing the modeling of geometric distortions, such as spikes, noise, and irregular curvature changes. Moreover, this foreground-background separation ensures that subsequent feature extraction and model training focus only on the region of interest, thereby improving the robustness of quality assessment.

Proposed Method

Overview of the Framework

Figure 3 shows the diagram of the proposed SAF-Net, which consists of three fundamental components. The data preprocessing module begins by generating multiview texture projections of the input point cloud, from which two complementary representations are computed: mask maps that explicitly encode geometric boundaries and LBP maps that capture structural texture variations. For feature encoding, we employ a hybrid CNN-ViT architecture that collaboratively combines convolutional layers for efficient local feature extraction with ViT for global context modeling, with each input modality processed through dedicated encoder pathways to produce hierarchical feature representations. The feature fusion module employs a progressive fusion strategy that enhances the information interaction of each feature through two-stage feature fusion.

Data Preprocessing

For a given point cloud **P**, the texture images are generated through a six-view orthogonal projection. The symbol T_v represents a projected texture image, and $v = 1, 2, ...6$ is

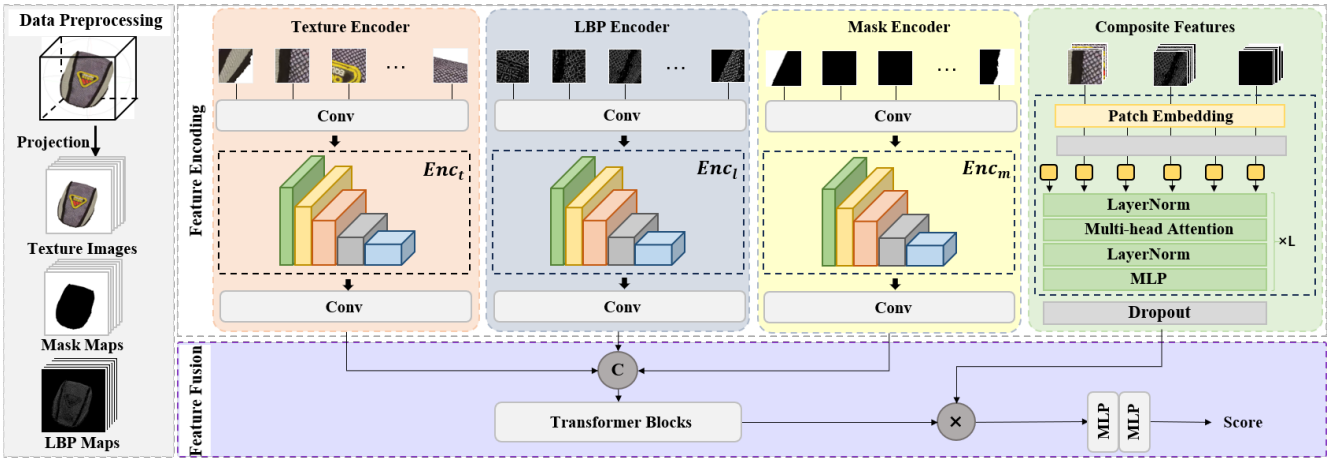

Figure 3: The diagram of the proposed structure-aware feature fusion network (SAF-Net).

the projection view, i.e., $T_v \in \mathbb{R}^{3\times H\times W}$. Here, H and W denote the height and width of the images.

The masks maps are generated to better model geometric distortions, such as surface burrs and curvature variations. Here, the mask map is a binary representation that distinguishes the foreground (projected object) from the background in a given texture image. Since the background pixels in the texture maps are set to 255 during projection, the corresponding mask map can be efficiently computed as,

$$M_v(x,y) = \begin{cases} 1, & \text{if grayscale}(T_v(x,y)) = 255, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where (x, y) denotes the pixel coordinates, and grayscale$(\cdot)$ is the grayscale operations, i.e., $M_v \in \mathbb{R}^{1\times H\times W}$.

The LBP map is a texture descriptor that encodes local structural patterns by comparing the intensity of each pixel with its neighboring pixels. In this work, we employ the basic LBP operator with a 3×3 neighborhood (radius=1, 8 sampling points) to compute the LBP maps from the grayscale texture images. For a given pixel at position (x_c, y_c) in the texture images, its LBP code is computed as,

$$L_v(x_c,y_c) = \sum_{k=0}^{7} S(i_k - i_c)\cdot 2^k, \quad (2)$$

where i_c and i_k denote the grayscale pixel values of the center pixel and its neighboring pixels, respectively. The formula $S(\cdot)$ is a thresholding function that binarizes intensity differences, i.e.,

$$S(d) = \begin{cases} 1, & \text{if } d \geq 0, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Feature Encoding

The feature encoding stage adopted a hybrid CNN-ViT encoding architecture. It leverages convolutional layers to extract fine-grained local information while utilizing ViT's self-attention mechanism to capture global contextual relationships, thereby achieving "local-global" feature fusion. Three CNN-based image encoders are used to independently encode individual features, i.e., texture images $\mathbf{T} = \{T_v\}$, mask maps $\mathbf{M} = \{M_v\}$, and LBP maps $\mathbf{L} = \{L_v\}$. This approach helps maintain the independence of feature encoding across different modalities. Furthermore, the proposed method combines texture images, mask maps, and LBP maps as input to the ViT encoder, leveraging the long-range dependency modeling of the self-attention mechanism to achieve global fusion of composite features.

Individual Feature Encoding ResNet50 (He et al. 2015) is employed as backbone of the CNN-based encoders. The features extracted from the final convolutional layer serve as the output. To better exploit the potential of transfer learning, the image encoders are initialized with weights pre-trained on ImageNet. Furthermore, to enhance the model's generalization and reduce dependency on the source domain data distribution, the obtained features are sequentially processed through a dropout layer and a convolutional layer.

$$F_t = \text{Conv}(\text{Drop}(\text{Enc}_t(\text{Conv}(\mathbf{T})))), \quad (4)$$

$$F_m = \text{Conv}(\text{Drop}(\text{Enc}_m(\text{Conv}(\mathbf{M})))), \quad (5)$$

$$F_l = \text{Conv}(\text{Drop}(\text{Enc}_l(\text{Conv}(\mathbf{L})))), \quad (6)$$

where Conv$(\cdot)$ and Drop$(\cdot)$ denote a convolutional layer and a dropout layer, and Enc_t, Enc_m, and Enc_l are the image encoders for texture images, mask maps, and LBP maps, respectively. The symbols F_t, F_m, and F_l are the corresponding individual features, respectively.

Composite Feature Encoding To enable the input images to be processed by the ViT model, the input images are divided into non-overlapping $p \times p$ patches, and then a patch embedding is performed, i.e.,

$$X_i = \text{PE}(p_i^t) + \text{PE}(p_i^m) + \text{PE}(p_i^l), i = 1, 2, ..., N. \quad (7)$$

where PE$(\cdot)$ denotes the patch embedding operation, X_i represents the sum of the patch embeddings of the three inputs. The symbols i and N are the indices and the total number of the small patches, respectively, i.e., $N = HW/p^2$.

Type	Modality	Method	SJTU-PCQA			WPC			Average		
			PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
FR	P	PCQM	0.885	0.864	0.709	0.500	0.743	0.560	0.818	0.804	0.634
	P	GraphSIM	0.845	0.878	0.695	0.616	0.583	0.419	0.731	0.731	0.557
	P	PointSSIM	0.714	0.687	0.496	0.467	0.454	0.328	0.591	0.571	0.412
	P	MS-GraphSIM	0.898	0.874	-	0.563	0.540	-	0.731	0.707	-
NR	P	3D-NSS	0.781	0.782	0.602	0.645	0.644	0.472	0.713	0.713	0.537
	P	IT-PCQA	0.865	0.828	0.643	0.576	0.568	0.301	0.721	0.698	0.472
	P	ResSCNN	0.893	0.859	0.681	0.453	0.436	0.299	0.673	0.648	0.490
	P	ψ-Net	0.900	0.870	-	0.780	0.760	-	0.840	0.760	-
	P	DualGradient	0.940	0.930	0.790	0.810	0.800	0.620	0.875	0.865	0.705
	P	PKT-PCQA	0.912	0.932	-	0.560	0.557	-	0.736	0.745	-
	P	GPA-Net	0.882	0.873	-	-	-	-	-	-	-
	P	D^3-PCQA	0.800	0.820	-	0.810	0.790	-	0.805	0.805	-
	P	TCDM	0.930	0.910	-	0.807	0.804	-	0.869	0.857	-
	P	CANet	0.942	0.923	0.789	0.895	0.898	**0.731**	0.919	0.911	0.760
	I	PQA-Net	0.859	0.837	0.630	0.712	0.703	0.494	0.786	0.770	0.562
	I	GMS-3DQA	0.929	0.889	0.727	0.764	0.768	0.592	0.847	0.829	0.660
	I	GC-PCQA	0.930	0.911	0.755	0.809	0.805	0.625	0.870	0.858	0.690
	I	PAME	0.905	0.889	-	0.781	0.763	-	0.843	0.826	-
	I	EEP-PCQA	0.936	0.910	0.764	0.830	0.826	0.642	0.883	0.868	0.703
	I	CoPA	0.913	0.897	-	0.785	0.779	-	0.849	0.838	-
	I	MOD-PCQA	0.953	0.931	0.794	0.873	0.875	0.695	0.913	0.903	0.745
	P+I	MM-PCQA	0.922	0.914	0.776	0.836	0.842	0.624	0.829	0.878	0.700
	P+I	CMDC-PCQA	0.944	0.930	0.780	0.842	0.849	0.655	0.893	0.890	0.718
	P+I	MFT-PCQA	0.930	0.915	0.770	0.838	0.839	0.649	0.884	0.877	0.710
	P+I	Plain-PCQA	0.930	0.913	0.760	0.878	0.879	0.695	0.904	0.896	0.728
	I	Ours	**0.963**	**0.949**	**0.823**	**0.896**	**0.898**	0.722	**0.930**	**0.924**	**0.773**

Table 1: Performance Comparison of SAF-Net and State-of-the-Art Methods on the SJTU-PCQA and WPC Datasets.

Subsequently, a learnable CLS token is prepended to the sequence to aggregate global information. Moreover, a position embedding is also added to the patch embeddings to retain positional information.

$$z_0 = [X_{class}; X_1; X_2; ...; X_N] + E_{pos}. \tag{8}$$

Finally, feature fusion is performed using L layers of Transformer blocks, where each block consists of multi-head self-attention (MSA) and multilayer perceptron (MLP), with residual connections employed throughout. Note that layernorm (LN) is applied before each block.

$$z_l = \text{MLP}(\text{LN}(\text{MSA}(\text{LN}(z_{l-1})))), \quad l = 1, 2, ...L. \tag{9}$$

After passing through L consecutive Transformer blocks, we obtain the composite features $F_c = z_L$.

Feature Fusion

Following the feature encoding stage, we obtain four distinct types of features: texture features F_t, mask features F_m, structural features F_l, and composite features F_c. To fully exploit these features and provide more precise and comprehensive feature support for the final quality regression task, we propose a dual-stage feature fusion strategy.

We first fuse the three individual features using two layers of Transformer blocks, referred to as a local feature, i.e.,

$$F_{local} = \psi(F_t \oplus F_m \oplus F_l), \tag{10}$$

where $\oplus$ denotes a concatenate operation.

Subsequently, we further integrate the composite feature F_c and the local feature F_{local}.

$$F_{local} = \phi(F_{local} * F_c), \tag{11}$$

where $*$ denotes an element-wise product.

This hierarchical fusion mechanism can ensure effective aggregation of both the individual and composite features, enhancing the model's discriminative capability for quality assessment.

Loss Function

The Huber loss function is used to train the proposed model. It is a robust loss function that combines the advantages of mean squared error (MSE) and mean absolute error (MAE). The Huber loss function is a segmented function that regulates the sensitivity to outliers by means of a threshold parameter with the following formula,

$$L_\delta(y, \bar{y}) = \begin{cases} \frac{1}{2}(y - \bar{y})^2, & \text{if } |y - \bar{y}| \le \delta, \\ \delta \cdot |y - \bar{y}| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases} \tag{12}$$

where y denotes the MOS value, $\bar{y}$ denotes the prediction value of the model, and δ is a parameter that determines the threshold for switching from MSE loss to MAE loss.

Type	Modal.	Method	PLCC	SRCC	KRCC
FR	P	PCQM	0.692	0.683	0.493
	P	GraphSIM	0.751	0.741	0.553
	P	PointSSIM	0.471	0.480	0.298
NR	P	3D-NSS	0.608	0.561	0.412
	P	ResSCNN	0.720	0.750	-
	I	PQA-Net	0.643	0.619	0.461
	I	GMS-3DQA	0.800	0.800	0.619
	P+I	MM-PCQA	0.809	0.809	0.620
	P+I	CMDC-PCQA	0.810	0.818	0.633
	I	**Ours**	**0.896**	**0.898**	**0.720**

Table 2: Performance Comparison on WPC2.0 Dataset.

Type	Modal.	Method	PLCC	SRCC
FR	P	PCQM	0.439	0.510
	P	GraphSIM	0.281	0.320
	P	PointSSIM	0.178	0.180
	P	MS-GraphSIM	0.348	0.389
NR	P	ResSCNN	0.624	0.595
	P	GPA-Net	0.619	0.592
	P	IT-PCQA	0.347	0.326
	I	PQA-Net	0.592	0.588
	P+I	MM-PCQA	0.597	0.581
	P+I	MFT-PCQA	0.741	0.714
	I	**Ours**	**0.880**	**0.884**

Table 3: Performance Comparison on LS-PCQA Dataset.

Experimental Results

Experimental Settings

The proposed model was tested on four PCQA benchmark datasets: SJTU-PCQA (Yang et al. 2021), WPC (Liu et al. 2023a), WPC2.0 (Liu et al. 2021a), and LS-PCQA (Liu et al. 2023b). Furthermore, following the methodology in (Chetouani et al. 2021; Fan et al. 2022), a rigorous cross-validation approach derived from original point clouds was used to validate the method. Specifically, the SJTU-PCQA, WPC, LS-PCQA, and WPC2.0 datasets employ 9/5/5/4 -fold, respectively.

In SAF-Net, the CNN-based image encoder adopts ResNet50 as its backbone, with output channel dimensions of [64, 256, 512, 1024, 2048] at each stage, respectively. The ViT model consists of 12 transformer layers, in which the number of self-attention heads is 12, and the embedding dimension is 768. Input images are partitioned into 16×16 patches as the basic processing units.

The proposed model is implemented using PyTorch and evaluated on NVIDIA RTX 4090 GPUs. The optimization process utilizes the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 5e-5, weight decay coefficient of 1e-4, and learning rate halving every 5 epochs. Training proceeds for a maximum of 300 epochs with a minibatch size of 8. The parameter δ in the loss function is set to 0.5.

Each point cloud is projected to images at 512×512 resolution. To enhance the generalization of the model, random cropping with a size of 224×224 is applied. For comprehensive quantification of the correlation between predicted quality scores and subjective ratings, this study employs three evaluation metrics: PLCC (Pearson Linear Correlation Coefficient), SRCC (Spearman Rank Order Correlation Coefficient), and KRCC (Kendall's Rank Correlation Coefficient).

Performance Comparison with SOTA Methods

We compared SAF-Net with SOTA methods, including full-reference (FR) and no-reference (NR) ones. The FR methods comprise: PCQM (Meynet et al. 2020), GraphSIM (Yang et al. 2022b), PointSSIM (Alexiou and Ebrahimi 2020), and MS-GraphSIM (Zhang, Yang, and Xu 2021). NR methods fall into three categories: Point-based methods include 3D-NSS (Zhang et al. 2022a), IT-PCQA (Yang et al. 2022a), ResSCNN (Liu et al. 2023b), PKT-PCQA (Liu et al. 2022), GPA-Net (Shan et al. 2024a), ψ-Net (Xiong et al. 2023), DualGradient (Wang et al. 2023), D^3-PCQA (Liu et al. 2025), TCDM (Zhang et al. 2024a), and CANet (Xiong et al. 2025). Projection-based methods encompass PQA-Net (Liu et al. 2021b), GC-PCQA (Chen et al. 2024), GMS-3DQA (Zhang et al. 2024b), PAME (Shan et al. 2024c), EEP-3DQA (Zhang et al. 2023), CoPA (Shan et al. 2024b), and MOD-PCQA (Wang, Gao, and Li 2024). Multimodal fusion-based methods include MM-PCQA (Zhang et al. 2022b), CMDC-PCQA (Wu et al. 2025), MFT-PCQA (Liu et al. 2024), and Plain-PCQA (Chai et al. 2024).

Table 1 shows the comparisons on the SJTU-PCQA and WPC datasets. The results demonstrate that SAF-Net significantly outperforms all the compared methods across both datasets. Specifically, among the projection-based methods compared, MOD-PCQA achieves the best average performance, yet SAF-Net shows improvements of 1.9% (PLCC), 2.3% (SRCC), and 3.7% (KRCC). For the point-based methods, CANet delivers the best average performance, but SAF-Net still surpasses it by 1.4% (PLCC), 1.9% (SRCC), and 1.7% (KRCC), respectively. In addition, the results show that the multimodal fusion-based methods do not exhibit a significant performance advantage over the projection-based and point-based methods. In particular, SAF-Net outperforms the best multimodal fusion-based method Plain-PCQA by 2.9% (PLCC), 3.1% (SRCC), and 6.2% (KRCC).

Table 2 shows the results tested on the WPC2.0 dataset, which was collected for the quality evaluation of V-PCC compressed point clouds. The results demonstrate that SAF-Net substantially outperforms all the compared methods. Specifically, it achieves significant performance improvements of 10.6% (PLCC), 9.8% (SRCC), and 13.7% (KRCC) over the SOTA CMDC-PCQA method, indicating superior adaptability for the evaluation of compressed point clouds.

Table 3 shows the performance comparisons conducted on the LS-PCQA dataset, which contains more diverse distortion types and larger sample sizes compared to the other datasets. In particular, SAF-Net exhibits remarkable performance gains of 18.9% (PLCC) and 23.8% (SRCC) over the best-performing competitor MFT-PCQA, further validating the enhanced adaptability of SAF-Net for the evaluation of point clouds with multiple types of distortion.

Train	Test	PQA-Net	GPA-Net	ResSCNN	MM-PCQA	**Ours**
LS-PCQA	SJTU	0.342	0.556	0.546	0.581	**0.791**
LS-PCQA	WPC	0.266	0.433	0.466	0.454	**0.721**
WPC	SJTU	0.235	0.553	0.572	0.612	**0.670**
WPC	LS-PCQA	0.285	0.346	0.339	0.515	**0.539**
SJTU	LS-PCQA	0.281	0.333	0.338	**0.373**	0.331
SJTU	WPC	0.22	0.418	0.269	0.409	**0.595**

Table 4: Experimental Results of Cross-dataset Validation (PLCC).

Tex.	LBP	Mask	Comp.	PLCC	SRCC	KRCC
	✓	✓	✓	0.887	0.891	0.709
✓		✓	✓	0.877	0.878	0.694
✓	✓		✓	0.853	0.855	0.666
✓	✓	✓		0.862	0.864	0.678
			✓	0.828	0.830	0.644
✓	✓	✓	✓	**0.893**	**0.896**	**0.716**

Table 5: Ablation Experiments of Input Features.

Modal.	Method	Param. (M)	Inference (ms)	Memory (MB)
P+I	MM-PCQA	52.93	989.25	795.53
P	CANet	3.11	500.00	1258.15
P	DualGradient	16.38	1010.00	9910.88
I	GMS-PCQA	27.58	7.60	112.36
I	**Ours**	195.53	17.15	171.84

Table 6: Model Complexity and Computation Efficiency.

Cross-dataset Validation

To validate the domain generalization capability of the proposed method, we conducted cross-dataset experiments on the benchmark datasets. The results are shown in Table 4. When trained on LS-PCQA, SAF-Net achieved PLCC scores of 0.791 on SJTU-PCQA and 0.721 on WPC, representing improvements of 36.1% and 54.7% over the suboptimal methods, respectively. With WPC as the training set, SAF-Net obtained PLCC values of 0.670 (SJTU-PCQA) and 0.539 (WPC), outperforming the second-best method MM-PCQA by 9.5% and 4.7%, respectively. In the setting where SJTU-PCQA served as the training set and WPC as the test set, SAF-Net exhibited a substantial performance gain of 42.3% compared to GPA-Net, the suboptimal approach. These results consistently confirm the superior domain generalization of the proposed method over existing methods.

However, when encountering significant domain shifts (training on SJTU-PCQA and testing on LS-PCQA), both SAF-Net and the comparison methods demonstrate limited generalization performance with PLCC scores below 0.4. This performance limitation primarily stems from the substantial discrepancies between SJTU-PCQA and LS-PCQA in terms of both dataset scale and distortion diversity.

Ablation Study

Table 5 shows the PLCC results of ablation studies on the WPC dataset. Specifically, we evaluated the importance of input features for the model by conducting the following experiments: (1) sequentially removing individual features, i.e., texture images, mask maps, and LBP maps (the 2^{nd}-4^{th} rows), (2) removing the composite features (the 5^{th} row), and (3) using only the composite features (the 6^{th} row).

When any individual input feature was removed, the performance significantly declined (average drop of 2.3%), with the absence of mask features having the most impact (performance decreased by 4.5%). This indicates that each feature contributes to the model's predictions.

When the composite features were removed or only the composite features were used, the model performance also exhibited notable degradation. Particularly, when only the combined features were retained, the performance decreased by 7.3% of the baseline. This strongly validates the necessity of the structure-aware feature fusion in SAF-Net.

Model Complexity and Computational Efficiency

We have compared model complexity and computational efficiency of SAF-Net with representative methods. The results are shown in Table 6. Regarding model parameters, the proposed method reaches 195.53 M, showing an increase compared to the other models, which is primarily due to the introduction of independent local feature encoding and multi-feature fusion mechanisms. In terms of computational efficiency, SAF-Net demonstrates significant advantages, requiring only 17.15 ms per inference while maintaining GPU memory usage at 171.84 MB. Notably, compared to point-based and multimodal fusion-based methods, the proposed projection-based method not only maintains high computational efficiency but also improves the PCQA accuracy.

Conclusion

Projection-based methods have become one of the main paradigms for PCQA due to its advantages in terms of computational efficiency, comprehensive scene coverage, and transfer learning potential. This study analyzes two major limitations of projected texture images in PCQA: insufficient geometric distortion perception and representation blind spots. On this basis, we propose SAF-Net by innovatively integrating object mask maps, LBP maps, and texture images. The mask maps enhance geometric distortion perception by extracting edge sharpness and curvature variations, while LBP maps capture essential structural information to overcome RBS and align with HVS sensitivity. The results show that SAF-Net has achieved the SOTA results.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62371254.

References

Alexiou, E.; and Ebrahimi, T. 2020. Towards a Point Cloud Structural Similarity Metric. In *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 1–6.

Chai, X.; Shao, F.; Mu, B.; Chen, H.; Jiang, Q.; and Ho, Y.-S. 2024. Plain-PCQA: No-reference point cloud quality assessment by analysis of plain visual and geometrical components. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 6207–6223.

Chen, W.; Jiang, Q.; Zhou, W.; Shao, F.; Zhai, G.; and Lin, W. 2024. No-Reference Point Cloud Quality Assessment via Graph Convolutional Network. *Trans. Multi.*, 27: 2489–2502.

Chetouani, A.; Quach, M.; Valenzise, G.; and Dufaux, F. 2021. Deep Learning-Based Quality Assessment Of 3d Point Clouds Without Reference. In *2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 1–6.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Fan, Y.; Zhang, Z.; Sun, W.; Min, X.; Liu, N.; Zhou, Q.; He, J.; Wang, Q.; and Zhai, G. 2022. A No-reference Quality Assessment Metric for Point Cloud Based on Captured Video Sequences. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, 1–5.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

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

Liu, Q.; Liu, Y.; Su, H.; Yuan, H.; and Hamzaoui, R. 2022. Progressive knowledge transfer based on human visual perception mechanism for perceptual quality assessment of point clouds. *arXiv preprint arXiv:2211.16646*.

Liu, Q.; Su, H.; Duanmu, Z.; Liu, W.; and Wang, Z. 2023a. Perceptual Quality Assessment of Colored 3D Point Clouds. *IEEE Transactions on Visualization and Computer Graphics*, 29(8): 3642–3655.

Liu, Q.; Yuan, H.; Hamzaoui, R.; Su, H.; Hou, J.; and Yang, H. 2021a. Reduced Reference Perceptual Quality Model With Application to Rate Control for Video-Based Point Cloud Compression. *IEEE Transactions on Image Processing*, 30: 6623–6636.

Liu, Q.; Yuan, H.; Su, H.; Liu, H.; Wang, Y.; Yang, H.; and Hou, J. 2021b. PQA-Net: Deep no reference point cloud quality assessment via multi-view projection. *IEEE transactions on circuits and systems for video technology*, 31(12): 4645–4660.

Liu, Y.; Shan, Z.; Zhang, Y.; and Xu, Y. 2024. Mft-pcqa: Multi-modal fusion transformer for no-reference point cloud quality assessment. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7965–7969. IEEE.

Liu, Y.; Yang, Q.; Xu, Y.; and Yang, L. 2023b. Point cloud quality assessment: Dataset construction and learning-based no-reference metric. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s): 1–26.

Liu, Y.; Yang, Q.; Zhang, Y.; Xu, Y.; Yang, L.; Xu, X.; and Liu, S. 2025. Once-Training-All-Fine: No-Reference Point Cloud Quality Assessment via Domain-Relevance Degradation Description. *IEEE Transactions on Broadcasting*, 71(2): 616–630.

Ma, X.; Qin, C.; You, H.; Ran, H.; and Fu, Y. 2022. Rethinking Network Design and Local Geometry in Point Cloud: A Simple Residual MLP Framework. arXiv:2202.07123.

Meynet, G.; Nehmé, Y.; Digne, J.; and Lavoué, G. 2020. PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. arXiv:1706.02413.

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

Shan, Z.; Zhang, Y.; Yang, Q.; Yang, H.; Xu, Y.; Hwang, J.-N.; Xu, X.; and Liu, S. 2024b. Contrastive pre-training with multi-view fusion for no-reference point cloud quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25942–25951.

Shan, Z.; Zhang, Y.; Yang, Q.; Yang, H.; Xu, Y.; and Liu, S. 2024c. Pame: Self-supervised masked autoencoder for no-reference point cloud quality assessment. *arXiv preprint arXiv:2403.10061*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

Wang, J.; Gao, W.; and Li, G. 2024. Zoom to Perceive Better: No-Reference Point Cloud Quality Assessment via Exploring Effective Multiscale Feature. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 6334–6346.

Wang, S.; Wang, X.; Gao, H.; and Xiong, J. 2023. Non-local geometry and color gradient aggregation graph model for no-reference point cloud quality assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6803–6810.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wu, B.; Shang, X.; Zhao, X.; Shen, L.; An, P.; and Ma, S. 2025. CMDC-PCQA: No-Reference Point Cloud Quality Assessment via a Cross-Modal Deep-Coupling Framework. *IEEE Transactions on Instrumentation and Measurement*.

Xiong, J.; Jiang, L.; Hu, Q.; Zhang, C.; Xie, J.-c.; and Gao, H. 2025. Cellular Aggregation Graph Convolutional Network for Point Cloud Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Xiong, J.; Wu, S.; Luo, W.; Suo, J.; and Gao, H. 2023. -Net: Point Structural Information Network for No-Reference Point Cloud Quality Assessment. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Xue, W.; Zhang, L.; Mou, X.; and Bovik, A. C. 2014. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. *IEEE Transactions on Image Processing*, 23(2): 684–695.

Yang, Q.; Chen, H.; Ma, Z.; Xu, Y.; Tang, R.; and Sun, J. 2021. Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration. *IEEE Transactions on Multimedia*, 23: 3877–3891.

Yang, Q.; Liu, Y.; Chen, S.; Xu, Y.; and Sun, J. 2022a. No-reference point cloud quality assessment via domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21179–21188.

Yang, Q.; Ma, Z.; Xu, Y.; Li, Z.; and Sun, J. 2022b. Inferring Point Cloud Quality via Graph Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3015–3029.

Zhang, Y.; Yang, Q.; and Xu, Y. 2021. MS-GraphSIM: Inferring Point Cloud Quality via Multiscale Graph Similarity. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, 1230–1238. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.

Zhang, Y.; Yang, Q.; Zhou, Y.; Xu, X.; Yang, L.; and Xu, Y. 2024a. TCDM: Transformational Complexity Based Distortion Metric for Perceptual Point Cloud Quality Assessment. *IEEE Transactions on Visualization and Computer Graphics*, 30(10): 6707–6724.

Zhang, Z.; Sun, W.; Min, X.; Wang, T.; Lu, W.; and Zhai, G. 2022a. No-Reference Quality Assessment for 3D Colored Point Cloud and Mesh Models. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7618–7631.

Zhang, Z.; Sun, W.; Min, X.; Zhou, Q.; He, J.; Wang, Q.; and Zhai, G. 2022b. MM-PCQA: Multi-modal learning for no-reference point cloud quality assessment. *arXiv preprint arXiv:2209.00244*.

Zhang, Z.; Sun, W.; Wu, H.; Zhou, Y.; Li, C.; Chen, Z.; Min, X.; Zhai, G.; and Lin, W. 2024b. GMS-3DQA: Projection-Based Grid Mini-patch Sampling for 3D Model Quality Assessment. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(6).

Zhang, Z.; Sun, W.; Zhou, Y.; Lu, W.; Zhu, Y.; Min, X.; and Zhai, G. 2023. EEP-3DQA: Efficient and Effective Projection-Based 3D Model Quality Assessment. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2483–2488.