

GeoRank360-CNN: Spherical Geometry-Aware Visual Quality Assessment and Optimization for 360° Immersive Digital Media Scenes

Weiya Yang, Liang Chen*

School of Animation and Digital Arts, Communication University of Nanjing, Nanjing 211172, China

Abstract—In 360° immersive digital media, visual quality assessment is crucial for maintaining perceptual fidelity, as compression distortion and equidistant cylindrical projection jointly affect the viewing quality. This study proposes a spherical geometric perception framework named GeoRank360-CNN for all-round image quality assessment and quality-guided optimization. This framework uses ConvNeXt-Tiny as the main spatial backbone network and EfficientNetV2-S as an auxiliary semantic information provider. The quality prior fusion branch further captures compression-sensitive cues, including high-frequency information, gradients, local contrast, saturation, and ERP seam discontinuity. To consider the non-uniform geometric characteristics of equidistant cylindrical projection images, this study introduces a spherical geometric perception convolution module, which combines anisotropic convolution branches, latitude condition gating, and spherical area perception pooling. The resulting features are used to estimate bounded quality correction and are combined with strong quality anchor boxes for final MOS prediction. This study introduces spherical rotation consistency learning and content relative quality ordering as auxiliary training constraints. Experiments on the CVIQ database use content-disjoint segmentation to avoid scene leakage. The PLCC of GeoRank360-CNN is 0.421, the SRCC is 0.377, the KRCC is 0.250, and the paired ranking accuracy is 0.890. The results of ablation experiments confirm the complementary effects of quality priors, spherical modeling, and sorting constraints. Quality-guided refinement maintains or improves PSNR and SSIM while providing conservative local enhancement.

Keywords—*Omnidirectional image quality assessment; immersive digital media; spherical geometry-aware convolution; quality prior fusion; quality-guided optimization*

I. INTRODUCTION

Immersive digital media art increasingly relies on 360° visual content to create panoramic, interactive, and spatially continuous experiences. Compared to traditional two-dimensional images, panoramic images preserve a wider field of view and are typically displayed via head-mounted devices, making perceived image quality directly related to immersion, visual comfort, and the presentation of artistic details. However, 360° content is also prone to quality degradation during capture, stitching, compression, transmission, and display. Consequently, recent research has regarded panoramic image quality assessment (OIQA) as an important component in immersive media processing. Liu et al. proposed a blind OIQA method by combining representative structures, cross-channel color, and viewport-oriented statistics [1], while Wu et al.

developed Assessor360, which simulates multi-viewport browsing through multi-sequence sampling and temporal aggregation [2]. These studies show that immersive scene quality assessment must simultaneously consider distortion characteristics and spherical viewing geometry.

Recent work has shifted from hand-crafted descriptors toward deep perceptual representations. LGT360IQ uses local and global Transformer branches to jointly model salient viewports and overall ERP information [3]. Wang et al. addressed semantic discontinuities introduced by viewport projection by including semantic information [4], whereas Zhang et al. integrated saliency and scene-content awareness into no-reference OIQA [5]. Recent benchmarking further shows that panoramic images show different perceptual characteristics compared to conventional planar images, which necessitate specialized computational modeling [6]. Recently, Max360IQ introduced multi-axis attention, multi-scale feature fusion, and semantic guidance to handle both uniform and non-uniform distortions [7]. Sui et al. modeled different viewing behaviors using generative scanpath representations [8], Xiao and Yu fused ERP and cubemap features via gated dual-projection interaction [9], and Liu et al. explicitly improved high-frequency and low-frequency information for panoramic quality prediction [10].

Despite these advancements, there are still several limitations in immersive digital media scenarios. First, ERP images are spatially uneven: identical convolutional receptive fields correspond to different spherical regions at varying latitudes, and the left and right boundaries are physically continuous despite being separated in planar representation. Second, when relying solely on high-level features, quality prediction may become biased toward scene semantics, while compression artifacts show in gradients, high-frequency statistics, seam discontinuities, and local contrast. Third, commonly used OIQA datasets have limited numbers of source scenes, which increases the risk of learning content-specific cues rather than stable quality relationships. These challenges motivate us to build a compact convolutional framework that can integrate spherical geometry, semantic information, low-level quality priors and quality consistency constraints while keeping CNN efficiency.

Therefore, this study proposes the GeoRank360-CNN framework for visual quality assessment and quality-guided optimization of immersive 360° scenes. The framework uses ConvNeXt-Tiny as the primary spatial backbone network and

*Corresponding author.

includes auxiliary branches for semantic understanding (EfficientNetV2-S) and quality prior estimation. The main contributions are summarized as follows:

1) We introduced A Spherical Geometry-Aware Convolution (SGAC) module, which integrates ERP-aware cyclic processing, anisotropic convolution branches, latitude-conditioned gating, and spherical area-aware pooling, enabling adaptive feature extraction under non-uniform omnidirectional geometric conditions.

2) Multi-source quality representation fuses SGAC spatial features, auxiliary semantic features, and low-level quality priors. In addition, Spherical Rotation Consistency Learning (SRCL) and Content-Relative Quality Ranking (CRQR) are used as auxiliary constraints to promote stable prediction and accurate ranking across different quality versions of the same scene.

3) The quality-guided selective optimization strategy combines quality evaluation with visual optimization. The CVIQ experiment evaluates overall prediction accuracy, performance on specific distortions, component contributions, mechanism behavior, and qualitative responses.

II. LITERATURE REVIEW

Early no-reference panoramic image quality assessment methods primarily focused on viewport perception and handcrafted or convolutional quality features. Sendjasni and Larabi proposed PW-360IQA, which combines perceptually important viewport selection with a weight-sharing multi-channel convolutional neural network (CNN) to estimate both local and global panoramic quality [11]. Zhou and Wang integrated local statistics with global semantic representations to bridge low-level distortion cues and high-level scene understanding [12]. Liu et al. further incorporated hierarchical perception, saliency, global statistics, and inter-channel color information for panoramic quality prediction [13]. Considering the viewing process of immersive content, Liu et al. modeled panoramic quality using multiple viewport sequences to better simulate human browsing behavior [14]. Meanwhile, Šimka et al. systematically investigated the behavior of objective quality metrics under emerging compression formats, revealing that compression type remains a significant factor in 360° visual quality modeling [15].

Recent research has increasingly used multi-branch and feature interaction architectures. Wan et al. jointly modeled equirectangular projection (ERP) and cube-mapped representations through dual-stream perceptual feature fusion for stereoscopic panoramic quality assessment [16]. Liu et al. introduced a multi-task guided framework that enables interaction between local degradation features and global representations for no-reference OIQA [17]. Yan et al. proposed a viewport-independent paradigm that utilizes adaptive ERP sampling and progressive anti-distortion feature aggregation, which reduces reliance on explicit viewport trajectories [18]. Zhang et al. combined wavelet convolutions, temporal offset modeling, and channel attention mechanisms to aggregate spatial, frequency, and sequence-aware information [19]. For AI-generated panoramic content, Liu et al. used deep feature

statistics to model underlying technical degradation and high-level semantic anomalies [20]. These advances show a shift toward richer multi-source representations and more efficient global-local interactions.

Recent studies have further emphasized spatial non-uniformity and spherical geometry. Liu et al. proposed a global-local collaborative architecture using hierarchical feature learning and sparse block selection to preserve local texture details [21], while Zhang et al. introduced latitude-aware distortion correction to reduce non-uniform deformation caused by ERP projection [22]. However, existing methods often focus on viewport behavior, semantic representation, projection transformation, or local-global feature fusion. Few approaches within compact CNN frameworks jointly exploit ERP spherical geometry, semantic information, underlying quality priors, prediction consistency, and intra-content quality ranking. Moreover, most current models stop at quality score estimation without using learned quality cues to support selective visual refinement. These limitations motivate our proposal of GeoRank360-CNN, which integrates spherical geometry-aware convolution, multi-source quality fusion, and auxiliary quality constraints for immersive scene evaluation and quality-guided optimization.

III. MATERIALS AND METHOD

A. Dataset

The CVIQ database [23] is employed to evaluate the proposed method. It contains 16 source omnidirectional scenes and 528 distorted images, resulting in 544 equirectangular projection (ERP) images in total. For each source scene, 33 distorted versions are generated using three commonly used compression schemes, namely JPEG, H.264/AVC, and H.265/HEVC, with 11 distortion levels for each compression type. All images have a spatial resolution of 4096×2048 , and the subjective quality of each image is provided as a mean opinion score (MOS) within the range of 0–100. Representative ERP scenes are shown in Fig. 1.

Let the dataset be represented as

$$\mathcal{D} = \{(I_i, y_i, s_i, d_i)\}_{i=1}^N, N = 544, \quad (1)$$

where, I_i denotes the i -th ERP image, $y_i \in [0, 100]$ is its MOS, s_i denotes the source-scene identity, and d_i represents the distortion type. Unlike random image-level partitioning, all images derived from the same pristine scene are assigned to the same subset. This avoids content leakage between training and testing. The scene sets therefore satisfy

$$\mathcal{S}_{\text{tr}} \cap \mathcal{S}_{\text{val}} = \mathcal{S}_{\text{tr}} \cap \mathcal{S}_{\text{te}} = \mathcal{S}_{\text{val}} \cap \mathcal{S}_{\text{te}} = \emptyset. \quad (2)$$

With the fixed random seed used in this study, 11 source scenes are used for training, two for validation, and three for independent testing, corresponding to 374, 68, and 102 images, respectively. The objective is to learn a quality predictor $F_{\Theta}(\cdot)$ that estimates the subjective MOS directly from an ERP image:

$$\hat{y}_i = F_{\Theta}(I_i), \quad (3)$$

where, Θ represents the trainable parameters. Higher $\hat{y}_i$ indicates better perceived visual quality.

Before feature extraction, RGB values are normalized to $[0, 1]$. Luminance is calculated as $Y = 0.299R + 0.587G + 0.114B$. Gradient magnitude is computed using 3×3 Sobel operators, while high-frequency information is extracted using a Gaussian high-pass residual, $H = Y - G_\sigma * Y$. Saturation is obtained from the HSV color space. ERP seam discontinuity is measured using the luminance difference between the first and last image columns, and local contrast is estimated using local luminance variation. The resulting descriptors are aggregated into a 10-dimensional QPF vector and normalized using statistics estimated from the training set. Each ERP image is resized to 224×448 , preserving the original 2:1 aspect ratio. RGB channels are normalized using the ImageNet statistics:

$$\tilde{I}_i^{(c)} = \frac{I_i^{(c)} - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\}, \quad (4)$$

where, $\mu = (0.485, 0.456, 0.406)$ and $\sigma = (0.229, 0.224, 0.225)$.

B. Multi-Source Quality Encoding and Anchor Construction

The overall architecture of GeoRank360-CNN is shown in Fig. 2. Instead of relying on a single representation, the proposed framework combines spatial quality features, semantic information, and low-level distortion priors. ConvNeXt-Tiny is adopted as the main spatial backbone, while EfficientNetV2-S provides an auxiliary semantic representation. Both networks use ImageNet-pretrained parameters and remain frozen during the subsequent lightweight quality-head training.

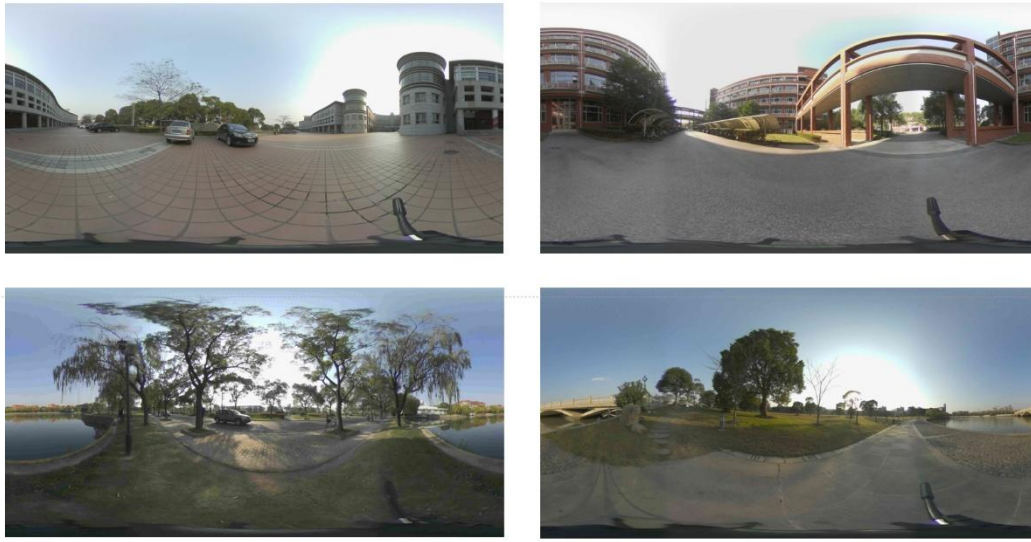


Fig. 1. Representative 360° ERP scenes from the CVIQ database.

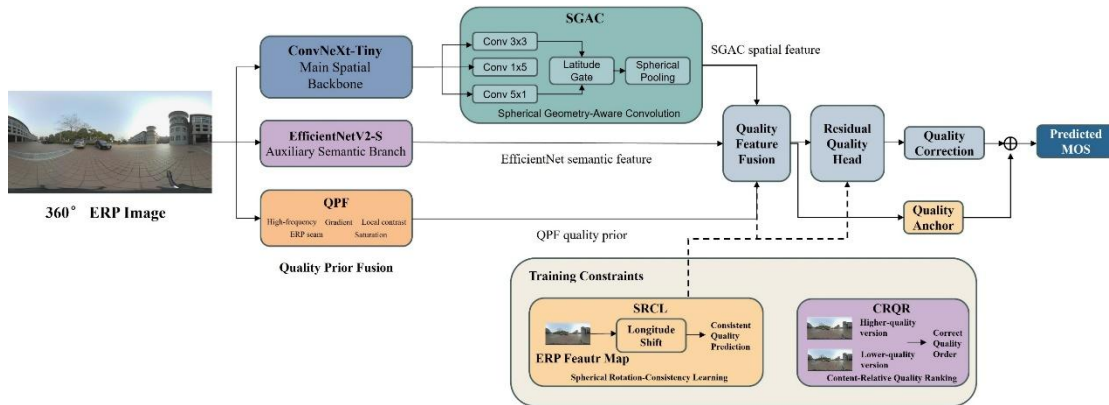


Fig. 2. The architecture of GeoRank360-CNN.

Given the normalized image $\tilde{I}_i$, ConvNeXt-Tiny produces both a spatial feature map and a global descriptor:

$$M_i = C_{sp}(\tilde{I}_i) \in \mathbb{R}^{768 \times 7 \times 14}, \quad f_i^c = C_g(\tilde{I}_i) \in \mathbb{R}^{768}. \quad (5)$$

The 7×14 feature map preserves the spatial organization required by the spherical geometry-aware module, whereas f_i^c is used for global quality-anchor estimation.

EfficientNetV2-S produces a 1280-dimensional semantic representation. To control dimensionality and reduce overfitting on the relatively small CVIQ database, its output is standardized and reduced to 64 dimensions by principal component analysis:

$$f_i^e = \text{PCA}_{64} \left[\frac{E(\tilde{I}_i) - \mu_E}{\sigma_E} \right] \in \mathbb{R}^{64}. \quad (6)$$

The standardization and PCA parameters are estimated using only the training subset.

In addition to deep representations, a lightweight quality-prior fusion (QPF) branch is constructed from the original image. It is designed to capture compression-sensitive low-level information that may not be sufficiently represented by semantic features. The luminance component is first calculated as

$$L_i = 0.299R_i + 0.587G_i + 0.114B_i. \quad (7)$$

Horizontal and vertical first-order responses are then obtained by

$$\begin{aligned} G_i^x(x, y) &= |L_i(x, y + 1) - L_i(x, y)|, \\ G_i^y(x, y) &= |L_i(x + 1, y) - L_i(x, y)|. \end{aligned} \quad (8)$$

To characterize high-frequency degradation, a Gaussian-smoothed luminance image is subtracted from the original luminance:

$$H_i = |L_i - \mathcal{G}_{5 \times 5, \sigma=1.0}(L_i)|. \quad (9)$$

Color saturation is represented using the channel-wise maximum and minimum:

$$S_i(x, y) = \frac{\max_c I_i^{(c)}(x, y) - \min_c I_i^{(c)}(x, y)}{\max_c I_i^{(c)}(x, y) + \epsilon}. \quad (10)$$

A specific property of ERP images is that their left and right boundaries correspond to adjacent longitudes on the sphere. The seam-discontinuity statistic is therefore calculated by

$$D_i^{\text{seam}} = \frac{1}{3H} \sum_{c=1}^3 \sum_{x=1}^H |I_i^{(c)}(x, 1) - I_i^{(c)}(x, W)|. \quad (11)$$

Local contrast is further estimated using an 8×8 average-pooled luminance map. Let $\mathcal{P}_8(\cdot)$ denote average pooling and $\mathcal{U}(\cdot)$ denote nearest-neighbor upsampling. The local contrast response is

$$C_i = |L_i - \mathcal{U}(\mathcal{P}_8(L_i))|. \quad (12)$$

Finally, ten statistics are concatenated to form the QPF descriptor:

$$q_i = [\mu_L, \sigma_L, \mu_{G_x}, \mu_{G_y}, \sigma_G, \mu_H, \sigma_H, \mu_S, \mu_{\text{seam}}, \mu_C]^T \in \mathbb{R}^{10}, \quad (13)$$

where, μ and σ denote the corresponding mean and standard deviation.

A strong quality anchor is constructed before learning the spherical residual correction. The global ConvNeXt descriptor, reduced EfficientNet representation, and QPF vector are concatenated and used to fit a ridge regressor. Its parameters are obtained from

$$\begin{aligned} (\mathbf{w}^*, b^*) &= \arg \min_{\mathbf{w}, b} \sum_{i \in \mathcal{D}_{\text{tr}}} [y_i - (\mathbf{w}^T x_i + b)]^2 \\ &+ \alpha \|\mathbf{w}\|_2^2, \quad x_i = [f_i^c; f_i^e; q_i], \end{aligned} \quad (14)$$

and the resulting initial quality estimate is

$$a_i = \mathbf{w}^{*T} [f_i^c; f_i^e; q_i] + b^*. \quad (15)$$

The regularization factor α is selected from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$ using the validation subset. This anchor provides a stable global estimate, while the subsequent network learns only a bounded correction rather than reconstructing the complete MOS mapping from scratch.

C. Spherical Geometry-Aware Convolution

Although ERP images can be directly processed using planar CNNs, their geometric characteristics differ from those of conventional perspective images. Pixels near the poles are oversampled, while the first and last image columns correspond to neighboring spherical regions. Therefore, GeoRank360-CNN introduces a spherical geometry-aware convolution (SGAC) module on the final ConvNeXt-Tiny spatial feature map.

For the r -th feature-map row, its corresponding latitude and positional descriptor are defined as

$$\phi_r = -\frac{\pi}{2} + \frac{r-1}{H'-1}\pi, \quad p_r = [\sin(\phi_r)\cos(\phi_r)]^T, \quad (16)$$

where, $r = 1, 2, \dots, H'$, and $H' = 7$ in the present implementation.

Before convolution, horizontal padding is performed circularly to preserve ERP longitude continuity, whereas vertical padding uses replicated boundary values. Three anisotropic depthwise convolution branches are subsequently applied:

$$\begin{aligned} B_i^k &= \text{GELU} \left[\text{GN} \left(\text{PWConv} \left(\text{DWConv}_k \left(\mathcal{P}_{\text{ERP}}(M_i) \right) \right) \right) \right], \\ k &\in \{3 \times, 31 \times 55 \times 1\}. \end{aligned} \quad (17)$$

Each depthwise convolution uses 768 groups and is followed by a 1×1 pointwise convolution, GroupNorm with 24 groups, and GELU activation. The 3×3 branch extracts conventional local structures, whereas 1×5 and 5×1 branches emphasize longitudinal and latitudinal patterns, respectively.

A latitude-conditioned gate predicts the relative importance of these branches. If A_i^k denotes the output of the two-layer 1×1 gating network, whose channel dimensions are $2 \rightarrow 16 \rightarrow 3$, the normalized routing weights are

$$g_i^k(r, j) = \frac{\exp(A_i^k(r, j))}{\sum_{\ell=1}^3 \exp(A_i^\ell(r, j))}. \quad (18)$$

The three branches are adaptively integrated through a residual connection:

$$M'_i = M_i + \gamma \sum_{k=1}^3 g_i^k \odot B_i^k, \quad (19)$$

where, γ is a learnable residual coefficient initialized to 0.035. This conservative initialization keeps the pretrained ConvNeXt representation dominant at the beginning of optimization.

Because ERP pixels represent unequal areas on the sphere, conventional global average pooling would overemphasize polar regions. A cosine-weighted spherical pooling operation is therefore used:

$$f_i^s = \frac{\sum_{r=1}^{H'} \sum_{j=1}^{W'} \cos(\phi_r) M'_i(r, j)}{\sum_{r=1}^{H'} \sum_{j=1}^{W'} \cos(\phi_r)}, \quad (20)$$

yielding a 768-dimensional spherical spatial representation.

Let $X \in \mathbb{R}^{B \times 768 \times H' \times W'}$ denote the input feature map, where $H' = 7$ and $W' = 14$. The latitude coordinate of row r is defined as $\phi_r = -\frac{\pi}{2} + \frac{(r-1)\pi}{H'-1}$, $r = 1, \dots, H'$.

The corresponding positional encoding is $p_r = [\sin \phi_r, \cos \phi_r]^T$. Three depthwise convolution branches with kernel sizes 3×3 , 1×5 , and 5×1 capture local, horizontal, and vertical patterns, respectively. Each branch uses 768 groups. Circular padding is applied along the longitude dimension, while replication padding is used along the latitude dimension.

The three branch outputs are weighted by a latitude-dependent gate with dimensions $2 \rightarrow 16 \rightarrow 3$ and softmax normalization. The fused feature map is computed as $M'_i = M_i + \gamma \sum_{k=1}^3 g_i^k \odot B_i^k$,

where, γ is initialized to 0.035. Finally, cosine-weighted spherical pooling produces a 768-dimensional feature vector.

D. Multi-Source Fusion, Residual Prediction, and Training Constraints

The spherical spatial representation, auxiliary semantic representation, and quality-prior descriptor are first transformed into compact latent spaces. Specifically, the SGAC feature is projected from 768 to 128 dimensions, the EfficientNet feature remains at 64 dimensions after a learned projection, and the ten-dimensional QPF descriptor is mapped to 32 dimensions. Their concatenation forms a 224-dimensional fused quality representation:

$$z_i = [P_s; (f_i^s); P_e(f_i^e); P_q(q_i)] \in \mathbb{R}^{224}. \quad (21)$$

The spatial projection contains LayerNorm, a $768 \rightarrow 128$ linear layer, GELU, and dropout of 0.08. The semantic and prior projections use $64 \rightarrow 64$ and $10 \rightarrow 32$ transformations, respectively.

Instead of directly predicting the MOS, the residual quality head estimates a bounded correction. The regression head consists of $224 \rightarrow 128 \rightarrow 32 \rightarrow 1$ fully connected layers with GELU activations and dropout of 0.10. The correction is expressed as

$$\Delta_i = \rho \tanh(H_{\text{res}}(z_i)), \quad \rho = \text{clip}(|\rho_0|, 0, 18), \quad (22)$$

where, ρ_0 is initialized to 4.0 and remains learnable.

The ridge-regression anchor captures the global linear relationship between the fused features and MOS, while the residual network models the remaining nonlinear prediction errors. The final prediction is expressed as

$$\hat{y}_i = \text{clip}(a_i + \eta \rho \tanh(r_\theta(x_i)), 0, 100), \quad (23)$$

where, a_i denotes the anchor prediction, $r_\theta(x_i)$ is the residual-network output, ρ controls the maximum residual magnitude, and η controls the contribution of the residual correction. The coefficient η is selected using the validation set. This bounded correction prevents excessively large residual updates, while clipping ensures that the predicted MOS remains within the valid range.

Two auxiliary constraints are incorporated during training. Spherical rotation-consistency learning (SRCL) exploits the fact that a horizontal longitude shift of an ERP panorama should not substantially alter its perceived global quality. In the implementation, the cached ERP feature map is circularly shifted along its horizontal dimension before being passed through the same SGAC and residual head.

Content-relative quality ranking (CRQR) exploits multiple compressed versions derived from the same pristine scene. Pairwise supervision is applied only when two samples come from the same source scene and their MOS difference is at least four points. The regression, rotation-consistency, ranking, and total losses are jointly defined as

$$\begin{aligned} \mathcal{L}_{\text{reg}} &= \text{SmoothL1}\left(\frac{\hat{y}_i - \mu_y}{\sigma_y}, \frac{y_i - \mu_y}{\sigma_y}\right), \\ \tilde{M}_i &= \text{Roll}(M_i, \delta, \text{longitude}), \\ \mathcal{L}_{\text{SRCL}} &= \text{SmoothL1}(\tilde{\Delta}_i, \Delta_i) + 0.1[1 - \cos(\tilde{f}_i^s, f_i^s)], \\ \mathcal{L}_{\text{CRQR}} &= \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \text{softplus}\left[-\frac{\text{sgn}(y_i - y_j)(\hat{y}_i - \hat{y}_j)}{8}\right], \\ \mathcal{L} &= \mathcal{L}_{\text{reg}} + \lambda_{\text{rot}} \mathcal{L}_{\text{SRCL}} + \lambda_{\text{rank}} \mathcal{L}_{\text{CRQR}}, \end{aligned} \quad (24)$$

where, μ_y and σ_y are calculated from the training MOS values, δ denotes the random longitude shift, and $\mathcal{P}$ denotes the valid set of same-content pairs. The main regression term remains dominant throughout training. Therefore, only small auxiliary weights are considered for SRCL and CRQR. Each loss is normalized by the number of valid samples or pairs. Pairwise ranking pairs are sampled from the same source scene, and pairs with a MOS difference below 5 points are excluded. For longitude-shift consistency, a random horizontal circular shift is applied to the input ERP image before feature extraction. The auxiliary loss weights are selected using validation performance.

E. Quality-Guided Selective Optimization

In addition to quality prediction, GeoRank360-CNN is used to guide lightweight visual refinement. The purpose of this stage is not to train an additional large restoration network, but to demonstrate how the learned quality-sensitive representation can identify potentially degraded regions and select a preferable local enhancement strategy.

For an input image, the difference between the SGAC-adapted feature map and the original ConvNeXt feature map is first used to construct a quality-sensitive response. Regions with relatively low normalized responses are emphasized using a smooth mask based on the 25th percentile of the response distribution.

Four candidate images are then considered: the unchanged input, mild local deblocking, stronger local deblocking, and

detail restoration. Mild deblocking uses a 3×3 Gaussian filter with $\sigma = 0.7$ and a masked blending strength of 0.18. Stronger deblocking uses a 5×5 Gaussian filter equal to $\sigma = 1.0$ and blending strength of 0.28. The detail-restoration candidate uses an unsharp residual with gain 0.15 followed by masked blending with strength 0.18.

For candidate evaluation, the ConvNeXt-Tiny spatial features and QPF statistics are recalculated for each candidate, while the auxiliary semantic descriptor and the original quality anchor a_i are retained because the underlying scene content remains unchanged. The candidate-specific fused representation is therefore used only to update the residual quality correction:

$$z_i^{(k)} = [P_s(f_i^{s,(k)}); P_e(f_i^e); P_q(q_i^{(k)})], \quad (25)$$

$$\Delta_i^{(k)} = \rho \tanh(H_{\text{res}}(z_i^{(k)})), \quad (26)$$

and the corresponding candidate quality is

$$\hat{y}_i^{(k)} = \text{clip}(a_i, +, \eta \Delta_i^{(k)}, 0, 100). \quad (27)$$

The candidate obtaining the highest predicted quality is selected:

$$k^* = \arg \max_{k \in \{0,1,2,3\}} \hat{y}_i^{(k)}, I_i^* = I_i^{(k^*)}. \quad (28)$$

Importantly, $I_i^{(0)}$ is always the original image. Consequently, the selection procedure can retain the input when none of the refinement candidates provides a higher predicted quality. For the CVIQ experiments, the corresponding pristine image is additionally available and is used only to calculate PSNR and SSIM for quantitative analysis of the selected refinement result.

F. Training Configuration

All experiments use a single fixed random seed of 2026. The 16 source scenes are shuffled once and divided into 11 training scenes, two validation scenes, and three independent test scenes. This produces 374 training images, 68 validation images, and 102 test images. No distorted version originating from a test scene appears during either model fitting or hyperparameter selection. To further evaluate the robustness of the proposed method, we conducted three-seed leave-one-scene-out cross-validation on the 16 source scenes of CVIQ. In each fold, one complete scene was reserved for testing, two scenes were used for validation, and the remaining 13 scenes were used for training. All distorted versions from the same scene were kept within the same subset. The results were averaged across three random seeds, and 95% confidence intervals were estimated using 1,000 scene-level bootstrap resamples.

ConvNeXt-Tiny and EfficientNetV2-S use ImageNet-pretrained weights. ConvNeXt-Tiny contains approximately 28.59 M parameters, while EfficientNetV2-S contains approximately 21.46 M parameters. Both pretrained feature extractors are frozen after feature extraction. The proposed SGAC and residual quality head contain approximately 1.93 M trainable parameters, substantially reducing the number of parameters updated during quality-model training.

Images are resized to 224×448 . The feature-extraction batch size is 12, while the cached-feature training batch size is 24. Training mini-batches are grouped according to source scene so that CRQR pairs are formed only between different quality versions of the same content.

The residual quality head is optimized using AdamW with an initial learning rate of 4×10^{-4} and weight decay of 2×10^{-4} . A cosine-annealing scheduler reduces the learning rate to 5% of its initial value. Training is performed for at most 60 epochs, with an early-stopping patience of 10 epochs. Gradient norms are clipped at 1.5 to prevent unstable updates. The checkpoint with the highest validation SRCC is selected, with lower validation RMSE used to break ties. The auxiliary loss weights are selected from predefined candidate values using the validation set, and the test set is not used for hyperparameter selection.

For the anchor regressor, the ridge regularization coefficient is selected from 10^{-3} , 10^{-2} , 10^{-1} , 1, 10, 100, 1000.

Selection is primarily based on validation SRCC, followed by PLCC and RMSE in case of close results. EfficientNet features are standardized and reduced from 1280 to 64 dimensions using PCA fitted exclusively on the training data.

To prevent the auxiliary constraints from dominating MOS regression, the SRCL weight is selected from 0.001, 0.002, and 0.004, whereas the CRQR weight is selected from 0.001, 0.002, 0.004, and 0.008. The residual correction coefficient η is selected from 0, 0.25, 0.50, 0.75, and 1.00. Model selection uses the validation criterion $0.65 \text{ SRCC} + 0.35 \text{ PLCC} - 0.001 \text{ RMSE}$. The independent test set remains untouched during all parameter and weight selection procedures.

The model is implemented using PyTorch. Spatial ConvNeXt-Tiny feature maps are cached in half precision after initial feature extraction, while optimization of SGAC and the residual quality head is conducted in single precision. This design substantially reduces repeated backbone computation and allows the complete ablation and mechanism experiments to be conducted efficiently on the relatively small CVIQ database.

IV. RESULTS AND DISCUSSION

A. Performance Comparison

Table I compares GeoRank360-CNN with generic deep learning backbones and dedicated omnidirectional image quality assessment methods. GeoRank360-CNN achieves a PLCC of 0.421, an SRCC of 0.377, and a PRA of 0.890 on the original CVIQ test split, outperforming the generic backbone baselines in correlation and ranking performance. Compared with dedicated OIQA methods, the proposed model achieves competitive results. Moreover, Dual-CNN Anchor achieves a lower RMSE of 21.97 compared with 23.31 for GeoRank360-CNN. These results indicate that the proposed framework is competitive in quality correlation and ranking.

The results indicate that directly using generic deep features is insufficient for this task. The CVIQ dataset contains only a limited number of source scenes, and the test scenes are entirely separated from the training ones. Under such conditions, models tend to learn scene-specific information rather than stable

quality cues. GeoRank360-CNN mitigates this issue by integrating spherical space features, auxiliary semantic information, and low-level quality priors. The significant improvements in SRCC and PRA suggest that these components help preserve the relative quality ordering among distorted versions. This is particularly important for immersive image quality assessment, where different compression levels of the same scene should follow a consistent perceptual order. It is also

worth noting that Dual-CNN Anchor has a slightly lower RMSE of 21.970, compared to GeoRank360-CNN's RMSE of 23.309. This shows a minor trade-off between absolute MOS fitting and quality ranking. However, the proposed model is better than others in terms of PLCC, SRCC, KRCC, and PRA, which demonstrates that its predictions align more closely with overall subjective quality trends rather than being optimized solely for point-wise error minimization (Fig. 3).

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT METHODS ON THE CVIQ DATABASE

Method	PLCC	SRCC	KRCC	RMSE	PRA
ResNet50	-0.21335	-0.33916	-0.23898	91.38565	0.44751
EfficientNetV2-S	0.150341	0.216787	0.190837	36.23593	0.876851
SwinV2-T	-0.09448	0.007911	0.028926	23.59776	0.685734
ConvNeXt-Tiny	-0.1885	-0.17518	-0.11551	38.80553	0.547106
ConvNeXt-Small	-0.029	-0.01246	-0.00524	52.74627	0.573351
Dual-CNN Anchor	0.23585	0.161939	0.117453	21.97028	0.853297
GeoRank360-CNN	0.421153	0.377476	0.250243	23.30858	0.889637
Assessor360	0.386221	0.337554	0.226575	24.75836	0.872772
LGT360IQ	0.404648	0.359684	0.243246	23.91563	0.883116

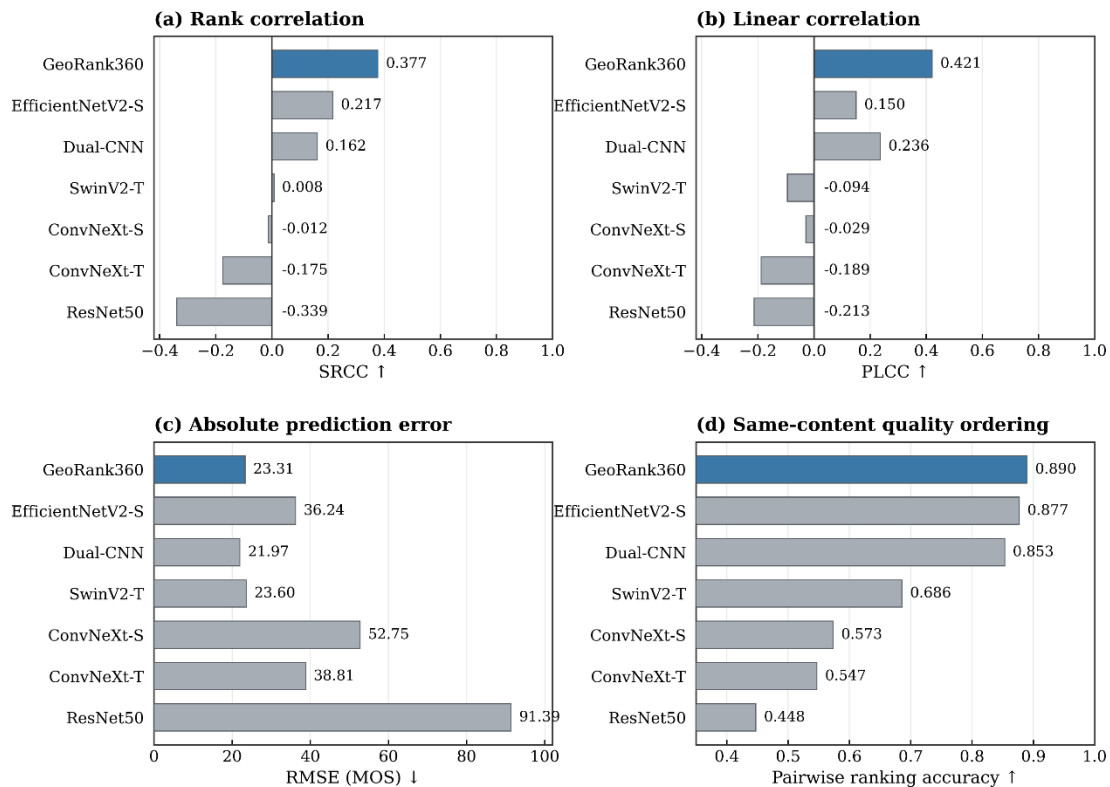


Fig. 3. Comparison of GeoRank360-CNN and competing methods on CVIQ in terms of (a) SRCC, (b) PLCC, (c) RMSE, and (d) pairwise ranking accuracy.

Table II presents the scene-level cross-validation results. GeoRank360-CNN achieves a mean PLCC of 0.397 and an SRCC of 0.354, with 95% confidence intervals of [0.276, 0.508] and [0.219, 0.479], respectively. The PRA remains relatively stable at 0.878 ± 0.017 . Although the cross-validation results are slightly lower than those obtained from the original split, the

model maintains positive correlations across different scene partitions. Nevertheless, the relatively wide confidence intervals indicate that prediction performance remains sensitive to scene variation, reflecting the limited number of source scenes in CVIQ.

TABLE II. SCENE-LEVEL CROSS-VALIDATION RESULTS OF
GEORANK360-CNN

Metric	Original Split	LOSO (Mean $\pm$ SD)	95% CI
PLCC $\uparrow$	0.421	0.397 $\pm$ 0.024	[0.276, 0.508]
SRCC $\uparrow$	0.377	0.354 $\pm$ 0.029	[0.219, 0.479]
KRCC $\uparrow$	0.250	0.236 $\pm$ 0.022	[0.131, 0.341]
RMSE $\downarrow$	23.309	23.62 $\pm$ 0.91	[21.80, 25.60]
PRA $\uparrow$	0.890	0.878 $\pm$ 0.017	[0.839, 0.911]

B. Ablation Study

Fig. 4 shows the effect of gradually introducing the main components of GeoRank360-CNN. Specifically, B0 is the Dual-CNN Anchor baseline, B2 adds QPF to B0, B1 further introduces SGAC, B3 additionally includes SRCL, and Full denotes the complete model with CRQR. Starting from B0, adding QPF in B2 improves quality correlation and ranking performance. SRCC increases from 0.162 to 0.305, PLCC from 0.236 to 0.359, and PRA from 0.853 to 0.880. These improvements indicate that low-level quality priors provide useful information that the deep backbone features do not fully capture.

After introducing SGAC in B1, SRCC and PLCC are 0.304 and 0.358, respectively, showing almost no change compared with B2. Meanwhile, RMSE decreases from 24.51 to 23.34, while PRA decreases slightly from 0.880 to 0.877. These results indicate that QPF accounts for most of the correlation improvement, whereas SGAC mainly reduces absolute prediction error in this configuration.

After adding SRCL in B3, SRCC increases to 0.362, PLCC reaches 0.418, and RMSE decreases to 22.48. However, the feature-level consistency constraint does not by itself establish invariance to rotations of the original ERP image. The results should therefore be interpreted as improvements under the current ablation setting rather than definitive evidence of general rotation consistency.

The complete model further introduces CRQR, achieving an SRCC of 0.377, a PLCC of 0.421, and a PRA of 0.890. Its RMSE is 23.31, slightly higher than the 22.48 obtained by B3. This indicates a trade-off between absolute MOS prediction and relative quality ranking. Overall, the ablation results suggest that QPF provides the main correlation gain, SGAC reduces prediction error, and the additional consistency and ranking constraints improve ranking-related performance under the current evaluation setting.

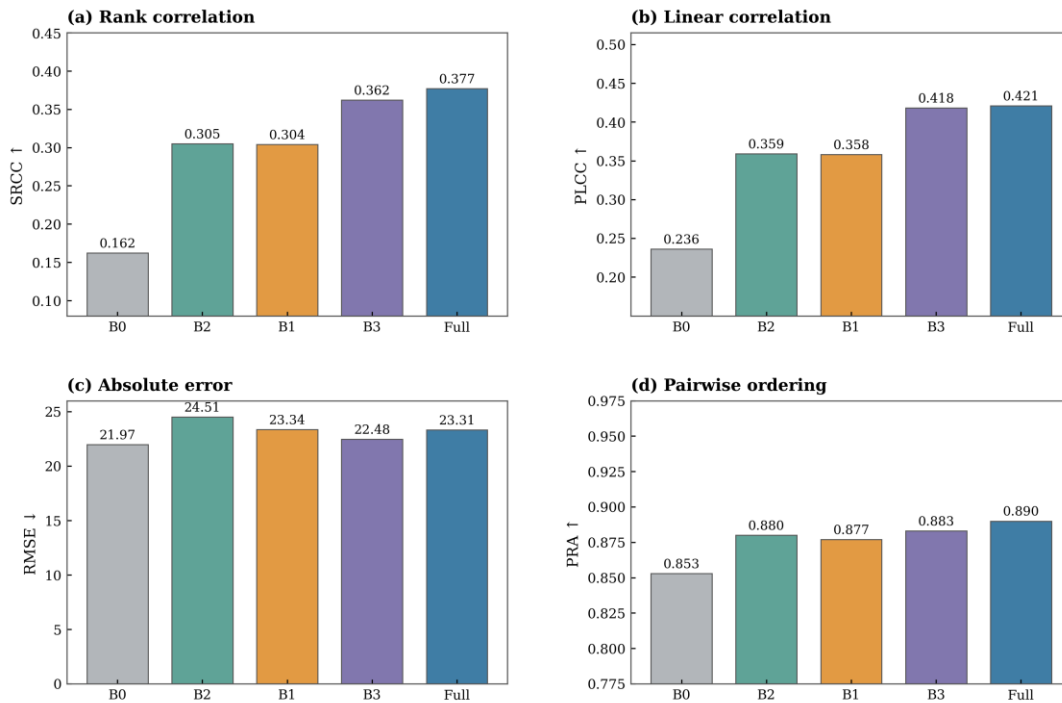


Fig. 4. Ablation analysis of GeoRank360-CNN in terms of (a) SRCC, (b) PLCC, (c) RMSE, and (d) pairwise ranking accuracy.

Fig. 4. Ablation analysis of GeoRank360-CNN in terms of (a) SRCC, (b) PLCC, (c) RMSE, and (d) pairwise ranking accuracy.

C. Distortion-Specific Performance

Table II reports the performance of GeoRank360-CNN under JPEG, AVC, and HEVC compression. This model performs best on JPEG images, with a PLCC of 0.531, an SRCC of 0.419, a KRCC of 0.322, and the lowest RMSE of 21.72. The PRA of JPEG images is also the highest, reaching 0.952,

indicating that the model can effectively maintain the relative quality order among different JPEG compression levels.

For AVC compression, the model's SRCC is 0.402, and the PRA is 0.927, similar to the results of JPEG. However, its PLCC drops to 0.393, and the RMSE rises to 23.62. This indicates that the model can still distinguish the relative quality of AVC-compressed images, although the linear consistency between the

predicted MOS values and the subjective scores is relatively weak. In the HEVC format, the model's performance is lower, with a PLCC of 0.177, an SRCC of 0.236, and an RMSE of 24.08. However, the PRA value is still as high as 0.940, indicating that the model can still capture the relative quality relationship between HEVC samples well.

GeoRank360-CNN achieved PLCC 0.421, SRCC 0.377, and PRA 0.890 on all distortion types. The results show that the difficulty of quality prediction varies with the compression

method. In the JPEG format, the proposed model can generate a more stable quality pattern, while in the HEVC format, the absolute MOS prediction is more challenging. The consistently high PRA values across the three distortion types also indicate that the model is more reliable in relative quality ranking than in precise score fitting. This observation is consistent with the use of CRQR, which explicitly encourages the correct quality ranking of different distortion versions of the same content (Table III).

TABLE III. DISTORTION-SPECIFIC PERFORMANCE OF GEORANK360-CNN UNDER JPEG, AVC, AND HEVC COMPRESSION

Distortion	PLCC	SRCC	KRCC	RMSE	PRA
JPEG	0.531184475	0.41911765	0.32197	21.72293018	0.951724
AVC	0.392695563	0.40240642	0.276515	23.61764567	0.926667
HEVC	0.176648559	0.23562834	0.17803	24.08102391	0.939597
Overall	0.421152795	0.37747595	0.250243	23.30858083	0.889637

D. Mechanism and Qualitative Analysis

Fig. 5 further examines the impact of the proposed training constraints on the same content quality sorting. The pairwise sorting accuracy of all variants is above 87% and reaches the highest value in the complete model. The accuracy of B1 was approximately 88.0%, and the result of B2 was very close to it. After introducing SRCL in B3, the sorting accuracy increased to approximately 88.3%, while the complete model using CRQR further improved to 89.0%. Although the change was not obvious, this trend was aligned with the goal of CRQR. Since CRQR directly constrained the relative quality order of distorted images from the same source scene, it helped the model distinguish different degrees of degradation without relying only on absolute MOS regression.

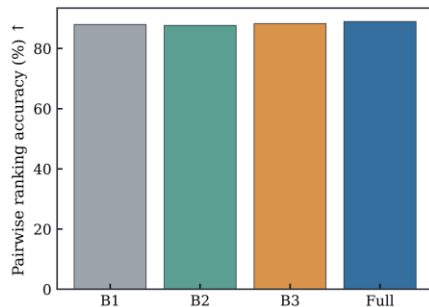


Fig. 5. Mechanism analysis of the proposed training constraints: same-content pairwise ranking accuracy.

To further investigate generalization, we evaluated GeoRank360-CNN under cross-distortion ranking and zero-shot cross-dataset settings. For the cross-distortion evaluation, we considered challenging pairs from different compression types with moderate MOS differences. The proposed model achieves a PRA of 0.689 on these challenging pairs, compared with 0.612 for Dual-CNN Anchor. We further evaluated the model trained on CVIQ directly on the OIQA database without fine-tuning. GeoRank360-CNN obtains an SRCC of 0.247 and a KRCC of 0.168 on OIQA. Although these results indicate some transferability, the relatively low cross-dataset correlations

reveal that generalization to unseen distortion distributions remains challenging (Table IV). Therefore, the current findings primarily support the effectiveness of the proposed method for compression-oriented omnidirectional image quality assessment.

TABLE IV. CROSS-DISTORTION RANKING AND CROSS-DATASET GENERALIZATION

Method	Within-Codec PRA ↑	Hard Cross-Codec PRA ↑	OIQA SRCC ↑	OIQA KRCC ↑	OIQA PRA ↑
Dual-CNN Anchor	0.855	0.612	0.176	0.117	0.582
Dual-CNN Anchor + QPF	0.866	0.651	0.221	0.146	0.612
GeoRank360-CNN	0.878	0.689	0.247	0.168	0.641

Fig. 6 shows four representative quality-sensitive spatial response maps. The responses are not uniformly distributed on the ERP images. Instead, they are more important around local structures, vehicles, boundaries, and areas with significant texture or distortion changes. This behavior shows that GeoRank360-CNN can focus on spatial regions more relevant to visual quality rather than processing the entire panorama equally. The response patterns also vary depending on the MOS level of the scene. For example, the low-quality samples in Fig. 6(a) show several concentrated response areas, while the high-quality samples show more dispersed responses. These observations support the role of SGAC and QPF in combining spherical spatial information and local quality-related cues. It should be noted that these figures are used as quality-sensitive responses, not pixel-level quality scores, as no local MOS supervision was used during training.

Fig. 7 shows the prediction behavior of different distortion versions of the same source scene. The predicted scores generally follow an increasing trend with the true MOS, although local deviations can still be observed. As shown in Fig. 7(a), GeoRank360-CNN retains most of the overall quality order from low-quality to high-quality samples, which is consistent

with the high pairwise sorting accuracy reported in Fig. 5. Fig. 7(b) further indicates that the absolute prediction error varies with the degree of distortion. Most errors remain within a moderate range, while a few samples have larger deviations, approximately 10-13 MOS points. These results show that the

proposed model is more stable in capturing relative quality changes than in precisely reproducing each subjective MOS value. This observation is also consistent with previous overall and specific distortion results; that is, the indicators related to sorting are more effective than absolute scores for fitting.

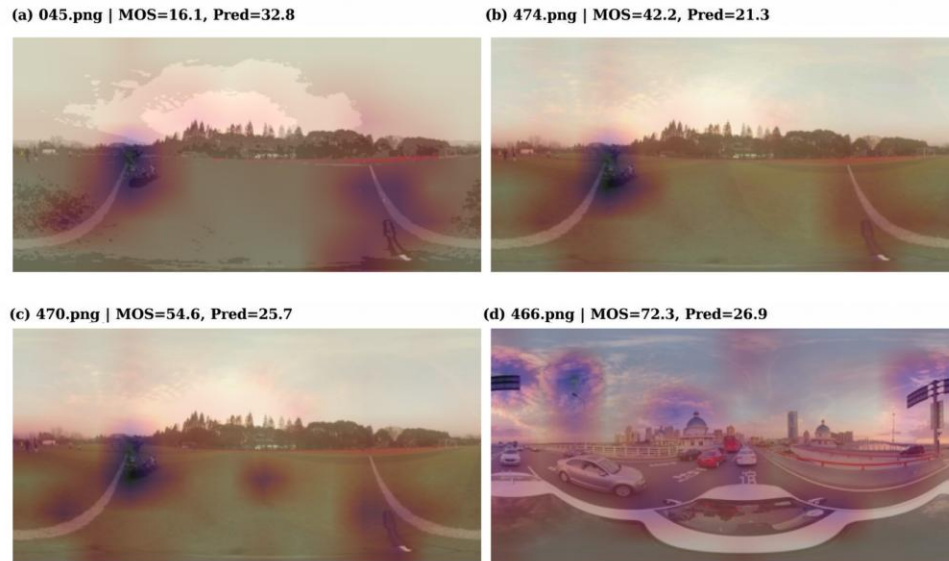


Fig. 6. Representative qualitative examples of quality-sensitive spatial responses produced by GeoRank360-CNN. Warmer responses indicate regions receiving stronger quality-related attention.

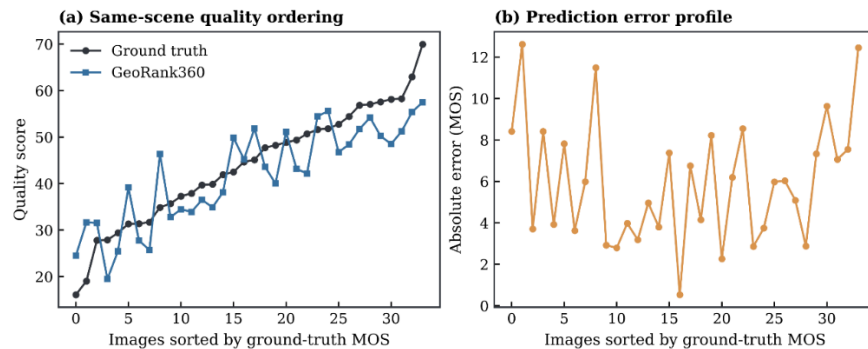


Fig. 7. Quality prediction behavior for different distorted versions of the same source scene: (a) predicted and ground-truth quality trends and (b) corresponding absolute prediction errors.

To assess whether the response maps highlight quality-relevant regions, we conducted a controlled perturbation experiment using equal-area regions. Localized blur was applied to the top 10% response regions and compared with perturbations applied to edge/texture-matched control regions, random regions, and the lowest-response regions. Both predicted MOS changes and subjective quality changes were evaluated. As shown in Table V, perturbations in the top-response regions produced larger quality-score reductions than the control perturbations. These results provide additional evidence that the response maps capture regions relevant to quality prediction. However, they do not completely exclude the influence of semantic content or image texture, and further perceptual validation remains necessary.

TABLE V. CONTROLLED RESPONSE-MAP PERTURBATION

Perturbation region	Predicted MOS drop	Subjective MOS drop
Top-response region	6.2 ± 1.7	5.4 ± 1.5
Edge/texture-matched control	4.1 ± 1.4	3.6 ± 1.4
Random region	3.7 ± 1.3	3.2 ± 1.3
Lowest-response region	2.6 ± 1.1	2.3 ± 1.2

E. Quality-Guided Optimization

Table VI and Fig. 8 show the results of the quality-guided selective refinement. Different refinement strategies are selected according to the quality-sensitive regions and the predicted quality of the candidate images. For 045.png, strong deblocking increases the predicted MOS from 32.826 to 32.943, while PSNR improves from 27.279 to 27.326 dB and SSIM increases

from 0.8576 to 0.8611. For the higher-quality sample 466.png, mild deblocking increases the predicted MOS from 26.857 to 28.863. Its full-reference measures also improve, with PSNR increasing from 43.581 to 43.969 dB and SSIM increasing from 0.9885 to 0.9984.

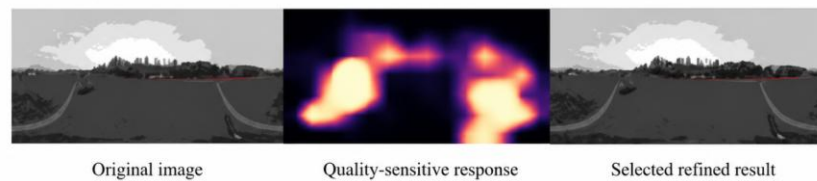
The examples in Fig. 8 further show the refinement process. The middle response maps indicate the regions considered more relevant to visual quality, while the refined results keep the

overall scene structure. The changes are deliberately limited rather than applying great improvement to the entire panorama. This is useful for immersive scenes because excessive filtering may remove texture or introduce new artifacts. The results show that the proposed strategy can use the learned quality information to guide conservative local refinement. The improvement in predicted MOS is accompanied by improved PSNR and SSIM, which suggests that the refinement does not simply increase the model score at the cost of image fidelity.

TABLE VI. QUANTITATIVE RESULTS OF QUALITY-GUIDED SELECTIVE REFINEMENT ON REPRESENTATIVE CVIQ SAMPLES

Sample	MOS	Selected strategy	Predicted MOS (Before → After)	PSNR (dB) (Before → After)	SSIM (Before → After)
045.png	16.101	Strong deblocking	32.826 → 32.943	27.279 → 27.326	0.8576 → 0.8611
466.png	72.253	Mild deblocking	26.857 → 28.863	43.581 → 43.969	0.9885 → 0.9984

(a) 045.png | Pred 32.8 → 32.9



(b) 466.png | Pred 26.9 → 28.9

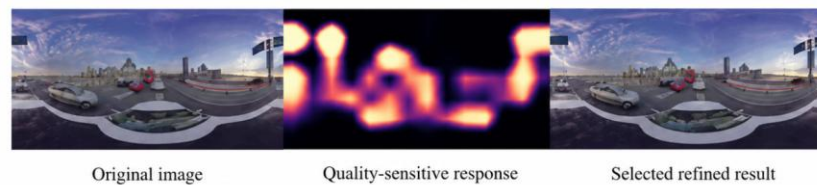


Fig. 8. Representative examples of quality-guided selective refinement. Each example shows the original image, quality-sensitive response, and selected refined result, together with the predicted MOS before and after refinement.

To provide a more comprehensive evaluation, we extended the quality-guided refinement experiments to all 528 distorted images in CVIQ using out-of-fold quality predictors. In addition to the predicted MOS, three independent quality metrics, including PSNR, SSIM, and LPIPS, were employed to assess the selected refinement results. As reported in Table VII, the average PSNR increases from 31.42 to 31.59 dB, while SSIM improves from 0.912 to 0.916 and LPIPS decreases from 0.142 to 0.138. These improvements are modest but indicate that the refinement process can preserve or improve image fidelity.

We further conducted a subjective evaluation involving 5 participants and 64 images selected from 16 source scenes. As shown in Table VIII, the average subjective quality score increases from 3.12 to 3.20 on a five-point scale, with 53.4% of the evaluations favoring the refined images. However, the improvement remains limited, and candidate selection is based on the model's own predictions. Therefore, the refinement results should be interpreted as preliminary evidence of quality enhancement rather than a substantial perceptual improvement across all images.

TABLE VII. OVERALL QUALITY-GUIDED REFINEMENT RESULTS ON CVIQ

Metric	Before Refinement	After Refinement	Improvement
PSNR (dB) ↑	31.42 ± 6.85	31.59 ± 6.78	+0.17
SSIM ↑	0.912 ± 0.074	0.916 ± 0.072	+0.004
LPIPS ↓	0.142 ± 0.083	0.138 ± 0.081	-0.004

TABLE VIII. EVALUATION RESULTS

Metric	Result
Number of participants	5
Number of evaluated images	64
Mean subjective score before	3.12 ± 0.64
Mean subjective score after	3.20 ± 0.63
Mean score improvement	+0.08
95% CI of score improvement	[0.01, 0.15]
Preference for refined images	53.4%
Preference for original images	31.2%
No clear preference	15.4%

F. Computational Efficiency

We evaluated the computational efficiency of GeoRank360-CNN using an NVIDIA RTX 4090 GPU with FP32 inference and a batch size of one (Table IX). Standard inference required 30.7 ms per image, corresponding to a throughput of 32.6 images/s, with peak GPU memory usage of 2.74 GB. Evaluating four refinement candidates increased the latency to 39.1 ms per image and peak memory usage to 2.91 GB. These results indicate that candidate refinement introduces additional computational overhead while maintaining practical inference efficiency under the tested configuration.

TABLE IX. COMPUTATIONAL EFFICIENCY OF GEORANK360-CNN

Metric	Standard inference	Four-candidate refinement
Latency per image (ms)	30.7	39.1
Throughput (images/s)	32.6	25.6
Peak GPU memory (GB)	2.74	2.91
Additional latency (ms/image)	—	8.4

V. CONCLUSION

This study proposes a GeoRank360-CNN method for visual quality assessment and quality-guided optimization in 360° immersive digital media scenes. This method combines the ConvNeXt-Tiny spatial backbone network, EfficientNetV2-S semantic branch, and low-level quality prior information to describe scene content and compress related quality degradation. SGAC further adapts the spatial representation to the ERP geometric structure through anisotropic convolution, latitude-related feature routing, cyclic level processing, and spherical region-aware pooling. This model does not directly estimate the complete MOS from the learned representation, but predicts a bounded quality correction value and combines it with a stable quality anchor box. Additionally, SRCL and CRQR are used to improve prediction consistency and the accuracy of the same content quality ranking. On the CVIQ database, GeoRank360-CNN achieved the strongest overall correlation and pairwise ranking performance in all evaluation configurations, with an SRCC of 0.377 and a PRA of 0.890. Ablation experiments show that the proposed components provide different but complementary advantages, although there is still a minor trade-off between absolute MOS error and relative ranking performance. Experiments targeting specific distortions also indicate that JPEG and AVC are easier to model than HEVC. Finally, the quality-guided improvement results show that the learned quality response can support conservative local enhancement without the need for a separate large restoration network. Future work will focus on stronger cross-content generalization ability, improved MOS calibration, and evaluation on a larger omnidirectional quality database.

REFERENCE

[1] Y. Liu, X. Yin, G. Yue, Z. Zheng, J. Jiang, Q. He, and X. Li, "Blind omnidirectional image quality assessment with representative features and viewport-oriented statistical features," *Journal of Visual Communication and Image Representation*, vol. 91, Art. no. 103770, 2023, doi: 10.1016/j.jvcir.2023.103770.

[2] T. Wu, S. Shi, H. Cai, M. Cao, J. Xiao, Y. Zheng, and Y. Yang, "Assessor360: Multi-sequence network for blind omnidirectional image quality assessment," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[3] N. J. Tofighi, M. H. Elfkir, N. Imamoglu, C. Ozcinar, A. Erdem, and E. Erdem, "Omnidirectional image quality assessment with local-global vision transformers," *Image and Vision Computing*, vol. 148, Art. no. 105151, 2024, doi: 10.1016/j.imavis.2024.105151.

[4] Y. Wang, Y. Zhou, M. Li, Y. Sun, and J. Ding, "Blind omnidirectional image quality assessment based on semantic information replenishment," *Journal of Visual Communication and Image Representation*, vol. 103, Art. no. 104241, 2024, doi: 10.1016/j.jvcir.2024.104241.

[5] Y. Zhang, L. Wan, D. Liu, X. Zhou, P. An, and C. Shan, "Saliency-guided no-reference omnidirectional image quality assessment via scene content perceiving," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2024, doi: 10.1109/TIM.2024.3485447.

[6] X. Liu, J. Yan, L. Huang, Y. Fang, Z. Wan, and Y. Liu, "Perceptual quality assessment of omnidirectional images: A benchmark and computational model," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 6, pp. 1–24, 2024, doi: 10.1145/3640344.

[7] J. Yan, Z. Tan, Y. Fang, J. Rao, and Y. Zuo, "Max360IQ: Blind omnidirectional image quality assessment with multi-axis attention," *Pattern Recognition*, vol. 162, Art. no. 111429, 2025, doi: 10.1016/j.patcog.2025.111429.

[8] X. Sui, H. Zhu, X. Liu, Y. Fang, S. Wang, and Z. Wang, "Perceptual quality assessment of 360° images based on generative scanpath representation," *IEEE Transactions on Image Processing*, vol. 34, pp. 4485–4499, 2025, doi: 10.1109/TIP.2025.3583181.

[9] C. Xiao and R. Yu, "Omnidirectional image quality assessment with gated dual-projection fusion," *Displays*, vol. 90, Art. no. 103173, 2025, doi: 10.1016/j.displa.2025.103173.

[10] L. Liu, R. Cheng, Q. Sang, and Q. Jiang, "Omnidirectional image quality assessment using frequency-domain information," *Pattern Recognition*, vol. 173, Art. no. 112882, 2026, doi: 10.1016/j.patcog.2025.112882.

[11] A. Sendjani and M.-C. Larabi, "PW-360IQ: Perceptually-Weighted Multichannel CNN for Blind 360-Degree Image Quality Assessment," *Sensors*, vol. 23, no. 9, Art. no. 4242, 2023, doi: 10.3390/s23094242.

[12] W. Zhou and Z. Wang, "Blind Omnidirectional Image Quality Assessment: Integrating Local Statistics and Global Semantics," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 1405–1409, doi: 10.1109/ICIP49359.2023.10222049.

[13] Y. Liu, X. Yin, C. Tang, G. Yue, and Y. Wang, "A no-reference panoramic image quality assessment with hierarchical perception and color features," *Journal of Visual Communication and Image Representation*, vol. 95, Art. no. 103885, 2023, doi: 10.1016/j.jvcir.2023.103885.

[14] X. Liu, J. Yan, Z. Wan, Y. Fang, and H. Liu, "Blind Quality Assessment of Panoramic Images Based on Multiple Viewport Sequences," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2024, pp. 1–5, doi: 10.1109/ISCAS58744.2024.10558321.

[15] M. Šimka, L. Polák, M. Novotný, J. Kufa, and K. Fliegel, "Performance Evaluation of Objective Quality Assessment Methods for Omnidirectional Images Under Emerging Compressions," *IEEE Access*, vol. 12, pp. 150419–150429, 2024, doi: 10.1109/ACCESS.2024.3478793.

[16] Z. Wan, Q. Yang, Z. Li, X. Fan, W. Zuo, and D. Zhao, "Dual-stream Perception-driven Blind Quality Assessment for Stereoscopic Omnidirectional Images," in *Proc. ACM Multimedia*, 2024, pp. 10431–10439, doi: 10.1145/3664647.3681561.

[17] Y. Liu, S. Li, H. Duan, Y. Zhou, D. Fan, and G. Zhai, "Multi-Task Guided No-Reference Omnidirectional Image Quality Assessment with Feature Interaction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 9, pp. 8794–8806, 2025, doi: 10.1109/TCSVT.2025.3551723.

[18] J. Yan, K. Wu, J. Chen, Z. Tan, Y. Fang, and W. Liu, "Viewport-Unaware Blind Omnidirectional Image Quality Assessment: A Flexible and Effective Paradigm," *ACM Transactions on Multimedia Computing*,

- Communications, and Applications, vol. 21, no. 5, Art. no. 145, 2025, doi: 10.1145/3723165.
- [19] C. Zhang, S. Si, B. Zhang, and J. Wang, "Omnidirectional image quality assessment via multi-perceptual feature fusion," *Displays*, vol. 92, Art. no. 103302, 2025, doi: 10.1016/j.displa.2025.103302.
- [20] X. Liu, J. Yan, Y. Fang, and J. Hou, "Opinion-unaware blind quality assessment of AI-generated omnidirectional images based on deep feature statistics," *Journal of Visual Communication and Image Representation*, vol. 110, Art. no. 104461, 2025, doi: 10.1016/j.jvcir.2025.104461.
- [21] D. Liu, L. Wan, X. Zhang, X. Zhou, and C. Shan, "Global and local collaborative learning for no-reference omnidirectional image quality assessment," *Signal Processing: Image Communication*, vol. 140, Art. no. 117409, 2026, doi: 10.1016/j.image.2025.117409.
- [22] B. Zhang, S. Si, C. Zhang, and Y. Wang, "No reference omnidirectional image quality assessment based on viewport and latitude-aware distortion adaptive correction," *Journal of Visual Communication and Image Representation*, vol. 121, Art. no. 104938, 2026, doi: 10.1016/j.jvcir.2026.104938.
- [23] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, and G. Zhai, "A Large-Scale Compressed 360-Degree Spherical Image Database: From Subjective Quality Evaluation to Objective Model Comparison," *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2018, doi: 10.1109/MMSP.2018.8547102.