Response to Reviewer #1：

The paper proposes a novel deep learning framework for stereoscopic omnidirectional images (SOIs) by integrating spherical and perceptual features. the paper is well written with detailed discussion on results. the authors need to consider the following modifications:

Comments 1

The abstract must be rewritten to motivate readers to further read the paper.

Response:

Thanks for your comment.

We have rewritten the abstract to better highlight the significance of SOIQA, our methodological novelty, and key results. Now, it clearly reflects the motivations and advantages of our approach.

In revised manuscript, the abstract is rewritten as:

"Stereoscopic omnidirectional images (SOIs) have gained significant attention for their immersive viewing experience by providing binocular depth with panoramic scenes. However, evaluating their visual quality remains challenging due to its unique spherical geometry, binocular disparity, and viewing conditions. To address these challenges, this paper proposes a dual-branch deep learning framework that integrates spherical structural features and perceptual binocular cues to assess the quality of SOIs without reference. Specifically, the global branch leverages spherical convolutions to capture wide-range spatial distortions, while the local branch utilizes a binocular difference module based on discrete wavelet transform to extract depth-aware perceptual information. A feature complementarity module is introduced to fuse global and local representations for final quality prediction. Experimental evaluations on two public SOIQA datasets—NBU-SOID and SOLID—demonstrate that the proposed method achieves state-of-the-art performance, with PLCC/SROCC values of 0.926/0.918 and 0.918/0.891, respectively. These results validate the effectiveness and robustness of our approach in stereoscopic omnidirectional image quality assessment tasks."

Comment 2

Details about the design architecture must be summarized. detailed explanation of architecture should be in the architecture section.

Response:

Thanks for your comments.

We agree that a summary of the overall architecture would enhance the clarity and readability of the method section. To address this, we have added a concise paragraph at the beginning of Section 3 (i.e., before Figure 1) that summarizes the overall design of the proposed framework, including its four main components: SOI preprocessing, global feature extraction via spherical convolutions, local binocular perception modeling via BDM, and feature fusion through the FCM module. The detailed explanation of each sub-module remains in their respective subsections (Sections 3.1–3.4), following the summary.

In revised manuscript, the following paragraph has been added:

"The overall architecture of the proposed framework is shown in Figure 1. It comprises four main components: (1) an SOI preprocessing module to generate different projection samples; (2) a global feature extraction branch based on spherical convolutions that models the structural distortion

characteristics over the sphere; (3) a local feature extraction branch that captures perceptual and binocular cues via a binocular difference module (BDM) enhanced with frequency-aware representations; and (4) a feature complementarity module (FCM) that adaptively fuses global and local features to perform final quality regression. These components are designed to jointly capture both the geometry-specific and perception-specific properties of SOI distortions."

Comment 3

A summary about results should be added to abstract.

Response:

Thanks for your comments.

We have revised the abstract to include a concise summary of the experimental results. Specifically, we now report the PLCC and SROCC values achieved by our method on two benchmark SOIQA datasets (NBU-SOID and SOLID), which reflect its superior performance compared to existing approaches.

This addition helps to clearly communicate the effectiveness of our proposed method and makes the abstract more informative and complete.

The result summary can be found in the last two sentences of the revised abstract as:

"Experimental evaluations on two public SOIQA datasets—NBU-SOID and SOLID—demonstrate that the proposed method achieves state-of-the-art performance, with PLCC/SROCC values of 0.926/0.918 and 0.918/0.891, respectively. These results validate the effectiveness and robustness of our approach in stereoscopic omnidirectional image quality assessment tasks."

Comment 4

Figure 2 looks irrelevant.

Response:

Thanks for your comments.

We realized that the role of Figure 2 may not have been sufficiently clear in the original manuscript. Actually, Figure 2 illustrates the key motivation for employing spherical convolution in our framework. Specifically, it shows how spherical sampling avoids the geometric distortions introduced by ERP (equirectangular projection), thus enabling more consistent and geometry-aware feature extraction across the spherical domain.

To address your concern, we have revised the manuscript in two aspects. Firstly, we have Added a dedicated explanation in Section 3.2 to clearly state the purpose of Figure 2 and its relevance to SOI feature modeling. Meanwhile, figure caption has been improved to highlight the differences between ERP-based and spherical sampling.

Detail revision is shown as follows.

1) Explanation of the purpose of Figure 2:

"As shown in Figure 2, compared with standard 2D convolution on ERP images, spherical convolution preserves the spatial continuity and geometric consistency of omnidirectional content. This sampling scheme enables the network to more accurately model structural distortions distributed across the sphere, which is crucial for SOI quality assessment."

2) Revised figure caption of Figure 2:

"Figure 2. Comparison of sampling patterns in ERP projection and spherical convolution."

Comment 5

There is a typo in Tables 3,4,5,6 label.

Response:

Thanks for your comments.

We are deeply sorry for our careless editorial mistake. In revised manuscript, the table caption of Table 3, 4, 5 and 6 have been rewritten. Specifically, according to the review comments, for better illustration of the experimental results, Table 6 in original manuscript has been reorganized as two tables with different performance indicators. Eventually, the detail table captions in revised manuscript are as follows:

Table 3. Performance of framework with different views

Table 4. Performance of different frequency feature adaptive module

Table 5. Performance of different convolution kernels

Table 6. Performance comparison with existing method for individual distortion in PLCC

Table 7. Performance comparison with existing method for individual distortion in SROCC

Comment 6

add more results in conclusions section and add conclusions as points.

Response:

Thanks for your comments.

In revision, we have rewritten the entire conclusion section in bullet-point format to clearly highlight the main findings and contributions of our work.

Specifically, the following aspects have been mentioned in conclusions section of revised manuscript:

1) Included key experimental results (PLCC/SROCC on NBU-SOID and SOLID) to emphasize the performance of our model;

2) Summarized major technical contributions, including the dual-branch design and feature complementarity module;

3) Stated known limitations of the current approach (e.g., fixed viewports, no eye tracking, no temporal modeling);

4) Outlined directions for future work, such as integrating gaze information and applying the model to SOI videos.

The conclusion section is revised as:

In this paper, we proposed a dual-branch deep learning framework for blind quality assessment of stereoscopic omnidirectional images (SOIs). Based on the experimental results and design contributions, we concluded this work as follows:

- **Performance Summary:**

1) The proposed framework integrates spherical structural features and binocular perceptual cues via dedicated branches and employs a feature complementarity module (FCM) to adaptively fuse global and local features. This design enables more robust and accurate quality prediction, aligning well with the nature of SOIs.

2) Extensive experiments on the NBU-SOID and SOLID datasets demonstrated that our method achieves superior performance, with PLCC/SROCC values of 0.926/0.918 and 0.918/0.891, respectively. Compared to state-of-the-art models, our approach shows improved generalization across different distortion types and viewing conditions.

- **Limitations:**

1) The proposed framework assumes a fixed distortion distribution during training and lacks an explicit distortion-aware mechanism. It treats all types of distortions uniformly without dynamically adapting its feature extraction or fusion strategies. This may limit its generalization when applied to real-world scenarios with novel or mixed distortion types.

2) The proposed framework uses a fixed set of six viewports uniformly sampled from the spherical content. While this strategy ensures spatial coverage, it does not account for variations in user attention or perception.

3) The proposed framework is designed for static stereoscopic omnidirectional images and does not incorporate temporal modeling. As such, it may not fully capture perceptual variations or consistency across SOI video sequences.

- **Future Work:**

1) We will explore dynamic viewpoint adaptation guided by saliency prediction or eye-tracking data to better match subjective perception patterns.

2) We will investigate integrating temporal cues, motion-aware representations, and recurrent or transformer-based architectures to extend our model to SOI video quality assessment.

3) We will attempt to explore distortion-invariant learning, domain generalization techniques, or meta-learning strategies to address distortion uncertainty and improve performance under mixed or novel distortion conditions."

Comment 7

a nomenclature is needed as there is a lot of abbreviations in text.

Response:

Thanks for your comments.

Indeed, there are many abbreviation in our manuscript, which may introduce obstacles for the readers to understand this work. Hence, for better readability, in revised manuscript, we have update and optimize the nomenclature of this work. The revised nomenclature is as follows:

Abbreviations

The following abbreviations are used in this manuscript:

SOI	Stereoscopic omnidirectional image
SOIQA	Stereoscopic omnidirectional image quality assessment
FR	Full reference
NR	No reference
CNN	Convolutional Neural Network
HMD	Head mounted display
FoV	Field of view
ERP	Equirectangular projection
BDM	Binocular difference module
DWT	Discrete wavelet transforms
FFAM	Frequency feature adaptive module
FCM	Feature complementarity module
MOS	Mean opinion score
PLCC	Pearson linear correlation coefficient
SROCC	Spearman rank correlation coefficient

RMSE	Root mean squared error

Comment 8

the authors should use two tables instead of table 6 one for PLCC and another for SROCC.

Response:

Thanks for your comments.

To better demonstrate the performance comparison results between the method proposed in this paper and existing quality evaluation methods, according to the reviewers' comments, we have divided Table 6 from the original manuscript into two tables in the revised manuscript, namely Table 6 and Table 7. Specifically, Table 6 is used to illustrate the comparison of PLCC (Pearson Linear Correlation Coefficient) metrics among different methods, while Table 7 presents the comparison results of SROCC (Spearman Rank Order Correlation Coefficient) between different metrics. Through the performance metrics comparison in these two tables, it can be seen that the method proposed in this paper has superior evaluation performance.

The revised tables are as follows:

Table 6. Performance comparison with existing method for individual distortion in PLCC

Types	Metrics	NBU-SOID					SOLID	
		JPEG	JPEG2000	HEVC	Sym	Asym	Sym	Asym
2D FR-IQA	PSNR	0.732	0.851	0.798	0.854	0.673	0.804	0.503
	SSIM	0.885	0.887	0.860	0.893	0.773	0.965	0.751
	GMSD	0.960	0.926	0.897	0.921	0.794	0.883	0.710
FR-OIQA	S-PSSNR	0.771	0.875	0.838	0.879	0.718	0.812	0.591
	CPP-PSNR	0.808	0.871	0.828	0.898	0.713	0.804	0.554
	WS-PSNR	0.739	0.869	0.819	0.865	0.688	0.799	0.576
	WS-SSIM	0.894	0.885	0.854	0.894	0.775	0.973	0.785
2D NR-IQA	BRISQUE	0.919	0.877	0.683	0.837	0.613	0.875	0.745
	OG	0.925	0.896	0.858	0.818	0.815	0.871	0.718
	NIQE	0.778	0.807	0.579	0.741	0.550	0.801	0.669
	HOSA	0.790	0.849	0.602	0.811	0.663	0.658	0.413
SIQA	SINQ	0.964	0.876	0.863	0.839	0.796	0.870	0.805
OIQA	SSP-OIQA	0.927	0.887	0.865	0.831	0.747	0.839	0.761
SOIQA	VP-BSOIQA	**0.973**	**0.949**	**0.948**	0.893	0.873	**0.982**	0.907
	Propose	0.962	0.934	0.883	**0.921**	**0.938**	0.948	**0.912**

Table 7. Performance comparison with existing method for individual distortion in SROCC

Types	Metrics	NBU-SOID					SOLID	
		JPEG	JPEG2000	HEVC	Sym	Asym	Sym	Asym
2D FR-IQA	PSNR	0.744	0.853	0.785	0.870	0.690	0.843	0.423
	SSIM	0.879	0.888	0.844	0.865	0.778	0.927	0.756
	GMSD	0.939	0.925	0.888	0.895	0.794	0.889	0.588
FR-OIQA	S-PSSNR	0.788	0.882	0.831	0.898	0.729	0.829	0.416
	CPP-PSNR	0.816	0.877	0.825	0.893	0.723	0.829	0.407

	WS-PSNR	0.750	0.874	0.815	0.890	0.710	0.824	0.403
	WS-SSIM	0.886	0.890	0.846	0.870	0.777	0.934	0.784
2D NR-IQA	BRISQUE	0.879	0.688	0.604	0.693	0.471	0.759	0.473
	OG	0.907	0.834	0.771	0.782	0.697	0.819	0.566
	NIQE	0.743	0.793	0.510	0.698	0.525	0.657	0.536
	HOSA	0.766	0.849	0.564	0.779	0.601	0.738	0.414
SIQA	SINQ	0.949	0.793	0.777	0.791	0.723	0.794	0.682
OIQA	SSP-OIQA	0.881	0.848	0.801	0.715	0.627	0.779	0.693
SOIQA	VP-BSOIQA	**0.957**	0.889	**0.875**	0.862	0.833	0.927	0.777
	Propose	0.942	**0.933**	0.872	**0.893**	**0.909**	**0.937**	**0.878**

Comment 9

more detailed comparison with previous work in literature is needed.

Response:

Thanks for your comments.

Revised manuscript included more comprehensive comparisons with prior SOIQA methods, both in the Related Works (Section 2) and in the Experimental Results and Analysis sections (Sections 4.6–4.7).

To be specific, firstly, we have added a new paragraph in Section 2 to highlight the limitations of representative prior methods, including their lack of binocular modeling or geometric alignment. The detail revision is as follows:

"In contrast to existing stereoscopic omnidirectional image quality assessment (SOIQA) methods which typically focus on 2D projections or fixed viewpoint patches, our method fully exploits the stereoscopic and spherical characteristics in a unified dual-branch architecture. Notably, some existing methods usually treats the left and right views separately, ignoring binocular interaction. Other existing methods is limited to viewport-based quality fusion without explicitly modeling binocular disparity or spherical distortion. In contrast, our model leverages spherical convolutions for geometry-aware representation and a binocular difference module (BDM) for perceptual disparity modeling, resulting in more robust and accurate quality predictions."

Furthermore, in the experimental analysis, we expanded the discussion to explain why our method outperforms these baselines, emphasizing the contribution of spherical convolutions and the binocular difference module. Detail revision is as follows:

"Compared with traditional methods, our method consistently achieves higher PLCC and SROCC scores across both datasets. This improvement can be attributed to our use of spherical convolutions, which better preserve the spatial structure of SOIs, and to the explicit modeling of binocular perceptual asymmetry. Moreover, unlike traditional methods which requires pre-defined distortion labels, our approach remains effective under unknown and mixed distortion scenarios, highlighting its generalization ability."

Comment 10

Figure 5 needs more detailed explanation about its functionality.

Response:

Thanks for your comments.

We agree that the functionality of Figure 5 was not clearly explained in the original version. To address this issue, in revised manuscript, two revisions have been implemented. Firstly, we have added a detailed paragraph in Section 3.4 describing the internal operations of the Feature Complementarity Module (FCM), including spatial alignment via transpose convolutions and gated fusion using sigmoid activations. The revision is as follows:

"The functionality of the Feature Complementarity Module (FCM), as illustrated in Figure 5, is to effectively integrate global and local feature streams. Specifically, features from both branches are first aligned in spatial resolution using transpose convolution operations. Then, channel- and spatial-wise interactions are enabled through a gated fusion mechanism implemented with sigmoid activations. This allows the model to adaptively emphasize perceptually salient features from both the spherical and binocular perspectives. Finally, the fused representation is fed into a fully connected regression head to predict the final quality score."