



Blind Omnidirectional Image Quality Assessment Based on Swin Transformer with Scanpath-Oriented

Xufeng Tang^{1,2,3}, Ping An^{1,2,3(✉)}, and Chao Yang^{1,2,3}

¹ Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, Shanghai 200444, China
anping@shu.edu.cn

² Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China

³ School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

Abstract. With the emergence of 5th generation mobile communication technology, the demand for Virtual Reality (VR) applications is on the rise worldwide. As one of the technologies related to visual content in VR, the quality evaluation of omnidirectional images has become an important issue. Inspired by the transformer, we propose a novel blind omnidirectional image quality assessment method. Firstly, we predict the path that the human eye follows when viewing omnidirectional images through headsets, and extract the area with the longest gaze duration on the path as the viewport. Then, to consider the intrinsic structural features of each pixel within each viewport, we use the Swin Transformer to extract viewport features. Finally, to establish a general scene perception and accurately evaluate immersive experiences, we construct a spatial viewport map for the entire perceptual scene. The graph structure performs reasoning on the overall relationship based on the spatial perception path. Experimental results demonstrate that our proposed model outperforms the current state-of-the-art Image Quality Assessment metrics, as evidenced by its superior results on two public databases.

Keywords: omnidirectional image · Swin Transformer · natural scene similarity · blind quality assessment

1 Introduction

With the rapid advancement of social productivity and information technology in recent years, human perception of real-world scenes is no longer confined to a limited field of view. This has resulted in a growing global demand for VR applications [1], and the VR industry is presented with new opportunities and

This work was supported in part by the NSFC under Grant 62071287, 61901252, 62020106011, 62171002, and Science and Technology Commission of Shanghai Municipality under Grant 22ZR1424300 and 20DZ2290100.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Lu et al. (Eds.): ICIQ 2023, LNCS 14359, pp. 352–364, 2023.
https://doi.org/10.1007/978-3-031-46317-4_28

challenges. VR technology has revolutionized traditional media by freeing it from the constraints of traditional screens. With the help of Head Mounted Display (HMD) [2], users can experience a 360° immersive view and watch videos from any angle by simply rotating their heads. As a visual imaging technology, it offers users interactive services that provide an in-depth and immersive experience, making it the most popular technology for displaying vision without any blind spots.

However, compared with traditional images, omnidirectional images (OI) [3] require capturing 360° views and typically demand high resolutions such as 4 K, 8 K, or higher to satisfy users' Quality of Experience (QoE). Therefore, such images are often heavily compressed for transmission and storage purposes [4]. During the process of immersive content acquisition, it is inevitable to encounter image distortion. As a result, it will lead to a degradation in the quality of the final image displayed to the user.

At the same time, visual degradation in VR applications can result in a reduced quality of experience for users. To address this issue, No-Reference Omnidirectional Image Quality Assessment (NR-OIQA) has been developed to enable humans to perceive visual distortion in omnidirectional images and improve the quality of the visual experience. Consequently, designing a feasible objective quality evaluation algorithm for the omnidirectional images holds significant practical and theoretical value.

Based on this, we propose a scanpath-oriented deep learning network for blind omnidirectional image quality assessment. Initially, the scanning path of the omnidirectional image is employed as a reference to derive the trajectory of the human eye's gaze within the head-mounted device. The viewports are extracted based on this trajectory. Secondly, taking into account the fact that existing CNN-based OIQA methods are limited by the receptive field and cannot establish global contextual connections, we employ the Swin Transformer to extract features for judging viewport quality. Finally, to construct a global correlation of viewports based on scanning paths, we use a graph-based approach. Notably, we extract the Natural Scene Statistics (NSS) features from each viewport, which effectively represents the similarity and correlation between viewports.

Our contributions are listed as follows:

- We propose a novel approach for extracting viewports from omnidirectional images by leveraging a model of scanning paths. A graph structure is constructed, which represents the complete viewing path of the omnidirectional image. It enables us to simulate the information interaction among different viewports and model the overall viewing process dynamically.
- We propose employing NSS features to calculate feature similarity and correlation across various viewports, with the objective of constructing an affinity matrix.
- We propose a novel deep learning model that integrates Swin Transformer with the graph structure to predict quality scores for omnidirectional images. This model facilitates both local and global feature interactions within and

across viewports. Our network outperforms existing Full-Reference and No-Reference methods on two benchmark databases.

2 Related Works

In this section, we introduce various methods about No-Reference omnidirectional image quality assessment. Then we make an overview of the recent related works on vision transformers.

2.1 NR-OIQA

NR-OIQA aims to objectively and accurately evaluate visual quality without reference images. Recently, deep learning technologies promote the development of NR-OIQA. Kim et al. [5] proposed a CNN-based adversarial learning method, which is called DeepVR-IQA. They partitioned an omnidirectional image into patches and employed an adversarial network to estimate their local quality and weight. Then the weighted quality scores are aggregated to obtain the final score.

Tian et al. [6] utilized a pseudo-reference viewport and employed spherical convolution to eliminate projection distortion. The final prediction score is obtained by merging the quality scores from two branches.

From the perspective of mitigating geometric distortion, Sun et al. [7] used a multi-channel CNN framework to predict the quality score of omnidirectional images. On this basis, Zhou et al. [8] incorporated a distortion discrimination-assisted network to promote OIQA learning tasks. However, the inherent differences between viewpoints as well as the interactive information between them are being overlooked.

To better illustrate the dependency of various viewports in 360° images, Xu et al. [9] first introduced graph convolutional networks into OIQA and modeled the spatial positional relationship of viewports in omnidirectional images. However, they only consider the spatial position of the viewports in the construction of the graph but ignore its content characteristics. To this end, Fu et al. [10] developed an adaptive hypergraph convolutional network (AHGCN) for NR-OIQA. In addition to the location-based features, the content-based features are also taken into consideration, which are generated based on their content similarity.

While the spatial and content characteristics of viewports are taken into account, the influence of viewport distortion is overlooked. Therefore, we propose to use NSS features sensitive to distortion to construct the correlation between viewports with the Swin Transformer. The NSS features are also used in [11, 12], and [13] to achieve high consistency with human perception.

2.2 IQA Based on Swin Transformer

Inspired by the success of the Transformer [14] in various NLP tasks [15], an increasing number of methods based on the Transformer [16] have appeared

in CV tasks, including no-reference omnidirectional image quality assessment tasks. Compared to frequently employed CNN models, the Swin Transformer introduces a shifted-window self-attention mechanism that facilitates the establishment of contextual connections. Conversely, CNNs possess restricted receptive fields, which restrict their attention to global features. In the task of IQA, both local and global quality perceptions are critical. Evaluators of image quality are sensitive not only to the quality of the current viewport but also to the previously viewed viewport, as this can affect their overall quality perception. Inspired by this fact, we use the Swin Transformer to establish local information interaction within the viewport. Also, a graph structure is used to construct feature transfer between viewports.

3 Method

In this section, we introduce the proposed OIQA method. Figure 1 illustrates the overall architecture. Our method uses a generative model to extract viewports from 360° images, producing realistic scanpaths. We implement the visual viewport interaction based on human eye perception and generate a perception score.

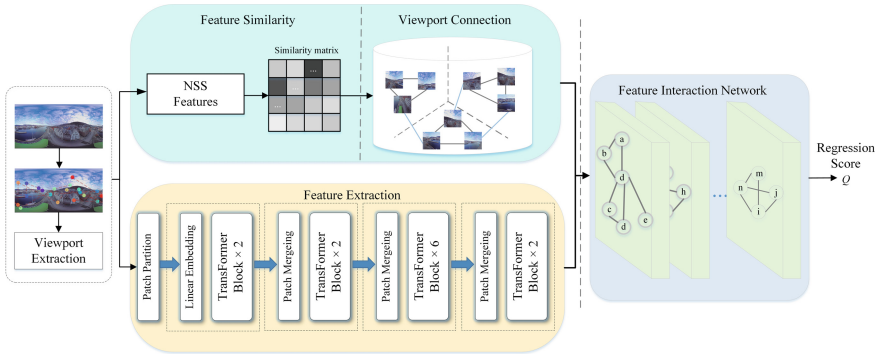


Fig. 1. The architecture of our proposed model. Viewports are firstly extracted from the distorted omnidirectional image in ERP format and input into the feature extraction module. The semantic features will be sent to the feature interaction network together with the extracted relevance matrix and regress the final perception score.

3.1 Viewport Extraction

When a 360° image is viewed in a VR device, the visual content is displayed as a flat section that touches the sphere created by the viewing angle. Also, when evaluating the quality of a 360° image, viewers look around the 360° image from multiple perspectives. Based on this, we employ a technique that mimics the human visual perception process by examining the scanning path data of an

omnidirectional image as seen through human eyes. We use the model, which is proposed in [17] to directly process the equirectangular project (ERP) format. The predicted gaze points are shown in Fig. 1 when viewing the omnidirectional image in HMD.

Figure 2 illustrates the process of viewport extraction. We set the viewing angle to 90° , which consists of the FOV of the most popular VR devices. Then, given a distorted omnidirectional image V_d and select N central points to extract viewports. The viewport sets are denoted as $\{V_i\}_{i=1}^N$. Then, we obtain N viewports covering the 90° FOVs.

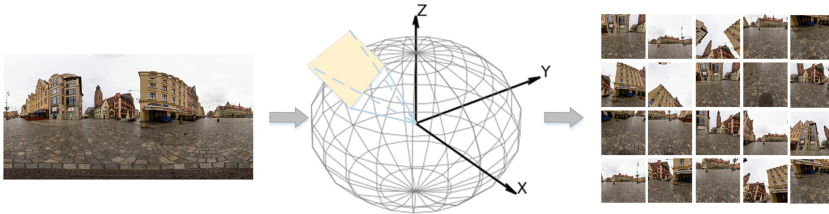


Fig. 2. The process of viewports extraction.

3.2 Graph Nodes Constructed by Swin Transformer

The Swin Transformer utilizes a shifted-window-based local attention computation method to achieve a hierarchical Transformer architecture. So we use it to extract the semantic feature. It consists of multiple Swin Transformer blocks. Figure 3 shows two successive blocks.

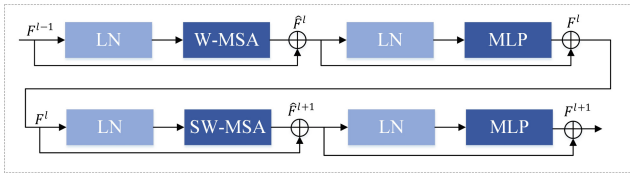


Fig. 3. Swin Transformer Block.

The window-based multi-head self-attention (W-MSA) module and the shifted window-based multi-head self-attention (SW-MSA) module are employed in two consecutive transformer blocks. Prior to every MSA module and MLP layer, a LayerNorm (LN) layer is employed for normalization, and residual connections are applied after each module. Based on such a window division mechanism, continuous Swin Transformer blocks can be calculated as:

$$\hat{F}^l = W\text{-MSA}(\text{LN}(F^{l-1})) + F^{l-1}, \quad (1)$$

$$F^l = \text{MLP}(\text{LN}(\hat{F}^l)) + \hat{F}^l, \quad (2)$$

$$\hat{F}^{l+1} = \text{SW-MSA}(\text{LN}(F^l)) + F^l, \quad (3)$$

$$F^{l+1} = \text{MLP}(\text{LN}(\hat{F}^{l+1})) + \hat{F}^{l+1}, \quad (4)$$

\hat{F}^l and F^l denote the output of the l_{th} block of the (S)W-MSA and MLP, respectively. N viewports $\{V_i\}_{i=1}^N$ are sampled and sent to the Swin Transformer. The number of blocks in each stage is 2, 2, 6, 2. We represent the feature of N viewports as $V = \{v_1, v_2, \dots, v_N\}$. The feature of each viewport represents a node of the graph.

3.3 Graph Edges Constructed by NSS

Considering that the extracted viewport is independent, it cannot simulate the process of viewing the omnidirectional image. Additionally, there exist variations in visual distortions across different viewports. We use NSS features that are crucial to the perceptual quality of OI as the edge of the graph structure to represent the similarity and correlation between different viewports.

To measure the loss of naturalness in viewports, it is necessary to compute the local mean subtracted and contrast normalized (MSCN) coefficients. These coefficients can be used to analyze the statistical features. For each distorted ERP map and viewports, the MSCN coefficients are calculated by:

$$\hat{D}^z(i, j) = \frac{D^z(i, j) - \mu(i, j)}{\sigma(i, j) + C} \quad (5)$$

where i and j represent the spatial coordinates. $\hat{D}^z(i, j)$ means the MSCN coefficients. $\mu(i, j)$ and $\sigma(i, j)$ represent the local mean and the standard deviation.

Then the generalized Gaussian distribution model is employed to model the statistic feature.

Figure 4 shows the difference between the MSCN distribution of different viewports. It is clear that the FOV information exhibits superior features and a greater capacity for expressing noise-related features compared to ERP images.

In order to construct viewports' correlation based on NSS features, we calculate the feature similarity through Eq. (6).

$$s_{i,m} = \frac{g_i \cdot g_m}{\|g_i\|_2 \cdot \|g_m\|_2} \quad (6)$$

where $i, m \in \{1, 2, \dots, N\}$, and g_i, g_m represent the NSS features of the viewport i and m , respectively. $s_{i,m}$ denotes the natural feature similarity between two viewports on a spherical domain.

Considering that the feature similarity between viewports will change with different distortion types and different distortion levels, we use the average of feature similarities across multiple viewports as the feature similarity threshold.

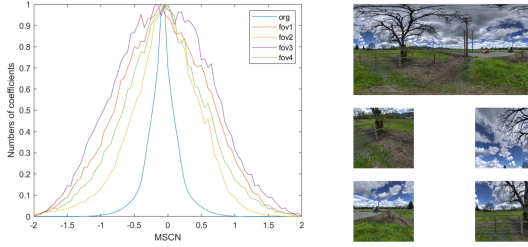


Fig. 4. The MSCN distribution of different viewports and the ERP image.

We calculate N viewports with the most similar NSS features by the following formula:

$$A_{i,m}(v_i, v_m) = \begin{cases} 1, & s_{i,m} \geq \text{average}(s_{i,m}) \\ 0, & s_{i,m} < \text{average}(s_{i,m}) \end{cases} \quad (7)$$

where $A_{i,m}$ is the affinity matrix representing whether there is information interaction between different viewports.

3.4 Quality Prediction

With the representation of the node feature vector $V = \{v_1, v_2, \dots, v_N\}$ and the affinity matrix A , the perception process based on omnidirectional scanning path is constructed. Each node feature is represented as a 768-dimensional feature vector to input. And then the quality of the omnidirectional images can be predicted by the network, which is composed of 5-layer graph convolutions. The process of interacting and updating the node information can be expressed as:

$$\mathbf{H}^{(l+1)} = f\left(\text{BN}\left(\hat{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)\right) \quad (8)$$

where \hat{A} is the adjacency matrix after normalization. The Softplus activation function $f(\cdot)$ [18] is used with batch normalization $\text{BN}(\cdot)$. The resulting feature matrices H^l are obtained by applying activations to the trainable weight matrix W^l . To match the number of hierarchical feature nodes of the Swin Transformer, the output dimension of each layer's feature nodes is 384, 192, 96, 48, 1. We then obtain the score of each viewport and leverage information from each viewport to produce accurate quality score Q .

4 Experimental Results

In this section, we provide an introduction to the databases utilized in our experiments, along with pertinent implementation details. We then compare the performance of our network with other metrics on a single and across databases. Finally, we conduct an ablation study and a cross-database evaluation to demonstrate the robustness and effectiveness of our model.

4.1 Databases

Two databases of omnidirectional images are utilized in the experiment: OIQA Database [19] and CVIQD Database [20].

OIQA Database: The database consists of 16 original images and their corresponding 320 degraded images. The degraded images include JPEG compression (JPEG), JPEG2000 compression (JP2K), Gaussian blur (BLUR), and Gaussian white noise (WN).

CVIQD Database: This database includes 16 reference images and 528 corresponding distorted images. Three encoding techniques are used to compress images, namely JPEG, H.264/AVC, and H.265/HEVC.

4.2 Implementation Details

Our model was executed on an NVIDIA GeForce RTX 3090 GPU with PyTorch. Our model uses Swin-Tiny as the backbone, which is pre-trained on ImageNet [21]. Each viewport image is resized to 256×256 and the batch size is set to 2. We utilize the Adam optimizer. The learning rate is set as 1×10^{-5} . The split ratio of the training set and test set in the database is 8:2. To avoid any overlap between the training and test data, distorted images that correspond to the same reference image have been assigned to the same set. The training loss we use is Mean Square Error (MSE) Loss. The final perception score is generated by predicting the mean score of 20 viewports.

Table 1. Overall performance comparison on CVIQD and OIQA databases. Best performance in bold.

Database		CVIQD			OIQA		
Metric		SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
FR-IQA	PSNR	0.6239	0.7008	9.9599	0.5226	0.5812	1.7005
	SSIM [22]	0.8842	0.9002	6.0793	0.8588	0.8718	1.0238
	CPP-PSNR [23]	0.6265	0.6871	10.1448	0.5149	0.5683	1.7193
	S-PSNR [24]	0.6449	0.7083	9.8564	0.5399	0.5997	1.6721
	MS-SSIM [25]	0.8222	0.8521	7.3072	0.7379	0.7710	1.3308
	WS-PSNR [27]	0.6107	0.6729	10.3283	0.5263	0.5819	1.6994
NR-IQA	BRISQUE [12]	0.8180	0.8376	7.6271	0.8331	0.8424	1.1261
	DB-CNN [26]	0.9308	0.9356	4.9311	0.8653	0.8852	0.9717
	MC360IQA [7]	0.9428	0.9429	4.6506	0.9139	0.9267	0.7854
	VGCN [9]	0.9639	0.9651	3.6573	0.9515	0.9584	0.5967
	AHGCN [10]	0.9623	0.9643	3.6990	0.9590	0.9649	0.5487
	proposed	0.9699	0.9619	3.5999	0.9702	0.9709	0.5292

4.3 Overall Performance on Individual Databases

We compare our method with the state-of-the-arts on the OIQA and CVIQD databases. Spearman’s Rank Order Correlation Coefficient (SROCC), Pearson Correlation Coefficient (PLCC), and Root Mean Squared Error (RMSE) are used to evaluate the performance of our model. The results of our comparison are presented in Table 1, where we highlight the top performances in boldface. Data of other methods are all from [9, 10]. Our method exhibits superior performance on both databases when compared to six FR-IQA methods and five NR-IQA methods. This is attributed to the effective modeling of human perceptual quality in our proposed approach. In comparison to VGCN and AHGCN on OIQA, our method not only exhibits superior monotonicity but also achieves higher accuracy. However, on CVIQD, our model achieved a slightly lower accuracy of 0.9619 when compared to VGCN and AHGCN.

4.4 Cross Database Validation

Table 2. Cross-database performances of the proposed model.

Metric	Tested on CVIQD			Tested on OIQA		
	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
MC360IQA [7]	0.8629	0.8886	6.5526	0.3329	0.4375	1.9012
VGCN [9]	0.9050	0.9241	5.4616	0.7832	0.7911	1.2934
proposed	0.8589	0.7510	7.2613	0.8634	0.8669	1.0271

To substantiate the generalizability of our model, we carried out cross-database experiments using two databases. Specifically, the OIQA database was employed for training purposes, while the CVIQD database was employed for testing purposes, and vice versa. Test results are presented in Table 2, indicating that our model achieves good cross-database performance on OIQA, but performs poorly on the CVIQD database. This is largely attributed to Swin Transformer’s local attention mechanism, which calculates attention only on a portion of the input sequence. This design allows the model to focus more on relevant information when handling different types of noise, thereby reducing its sensitivity to noise.

4.5 Ablation Study

Viewports Sampling Strategy: We conducted two experiments to validate our viewport extraction method. Firstly, we determined the optimal number of viewports by comparing the SROCC and PLCC metrics with varying numbers of viewports. Secondly, we compared the effectiveness of our viewport extraction method with a fixed region approach.

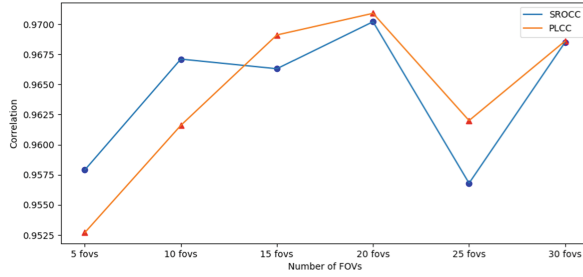


Fig. 5. The comparison of different number of viewports.

Table 3. The influence of the number of viewports.

	5fovs	10fovs	15fovs	20fovs	25fovs	30fovs
SROCC	0.9579	0.9671	0.9663	0.9702	0.9568	0.9685
PLCC	0.9527	0.9616	0.9691	0.9709	0.9620	0.9686

Figure 5 displays a comparison of various number of viewports. The SROCC and PLCC values are presented in line charts. Table 3 provides specific performance results for each number of viewports. The result reveals that 20 viewports achieve the highest SROCC and PLCC values. Based on it, we selected 20 viewports as the optimal number for our experiments.

Table 4. Performance comparison of different viewport extraction methods.

Database	CVIQD			OIQA		
Metric	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
Fixed	0.9415	0.9283	4.8700	0.9530	0.9550	0.6799
Proposed	0.9699	0.9619	3.5999	0.9702	0.9709	0.5292

We conducted a comparative analysis between our proposed viewport extraction method and the fixed area viewport extraction technique to verify its effectiveness. The results, as shown in Table 4, indicate that the fixed viewpoint extraction method is relatively ineffective, while our proposed method demonstrates superior performance.

The Effect of the NSS Features: To validate the effectiveness of constructing viewport information correlation using NSS features, we compared our model without using NSS features, where the model only relied on the Swin Transformer. As shown in Table 5, our method demonstrates better performance on both CVIQD and OIQA databases, especially on the OIQA database, where

SROCC and PLCC both achieved a over 0.02 improvement. This confirms the effectiveness of our use of NSS features.

Table 5. Performance comparison with or without NSS features.

Database	CVIQD			OIQA		
Metric	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
Baseline	0.9613	0.9599	4.0119	0.9449	0.9505	0.6605
Baseline+NSS	0.9699	0.9619	3.5999	0.9702	0.9709	0.5292

5 Conclusion

In this article, we present a deep learning model for the evaluation of omnidirectional image quality. We take into account the fact that the quality of a viewport can have an impact on our perception of subsequent viewports, indicating interdependence between different viewport qualities. Additionally, due to the inherent differences between viewpoints, our perception of quality may also vary across different viewports. We utilize Swin Transformer to facilitate the acquisition of inter-viewpoint information exchange and employ NSS features to determine the similarity and correlation between different viewpoints. This approach enables us to not only model local features but also account for global perception systems, resulting in improved quality regression. Based on experimental results, the proposed model exhibits superior performance compared to the state-of-the-art approaches.

References

1. Zink, M., Sitaraman, R., Nahrstedt, K.: Scalable 360 video stream delivery: challenges, solutions, and opportunities. *Proc. IEEE* **107**(4), 639–650 (2019)
2. Kim, H.G., Lim, H.T., Ro, Y.M.: Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Trans. Circuits Syst. Video Technol.* **30**(4), 917–928 (2020)
3. Bidgoli, N.M., Azevedo, R.G.D.A., Maugey, T., et al.: OSLO: on-the-sphere learning for omnidirectional images and its application to 360-degree image compression. *IEEE Trans. Image Process.* **31**, 5813–5827 (2022)
4. Zhu, H., et al.: MetaIQA: deep meta-learning for no-reference image quality assessment. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 14131–14140 (2020)
5. Lim, H.T., Kim, H.G., Ra, Y.M.: VR IQA NET: deep virtual reality image quality assessment using adversarial learning. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 6737–6741 (2018)

6. Tian, C., et al.: Viewport-Sphere-Branch Network for Blind Quality Assessment of Stitched 360° Omnidirectional Images. In: IEEE Transactions on Circuits and Systems for Video Technology. <https://doi.org/10.1109/TCSVT.2022.3225172>
7. Sun, W., Min, X., Zhai, G., et al.: MC360IQA: a multi-channel CNN for blind 360-degree image quality assessment. *IEEE J. Sel. Top. Sign. Proc.* **14**(1), 64–77 (2020)
8. Zhou, Y., Sun, Y., Li, L., et al.: Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network. *IEEE Trans. Circuits Syst. Video Technol.* **32**(4), 1767–1777 (2022)
9. Xu, J., Zhou, W., Chen, Z.: Blind omnidirectional image quality assessment with view-port oriented graph convolutional networks. *IEEE Trans. Circuits Syst. Video Technol.* **31**(5), 1724–1737 (2021)
10. Fu, J., et al.: Adaptive hypergraph convolutional network for no-reference 360-degree image quality assessment. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 961–969 (2022)
11. Liu, Y., Yin, X., Yue, G., et al.: Blind omnidirectional image quality assessment with representative features and viewport oriented statistical features. *J. Vis. Commun. Image Represent.* **91**, 103770 (2023)
12. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **21**(12), 4695–4708 (2012)
13. Yan, B., Bare, B., Tan, W.: Naturalness-aware deep no-reference image quality assessment. *IEEE Trans. Multimedia* **21**(10), 2603–2615 (2019)
14. Vaswani, A., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)
15. Lee, J., Toutanova, K.: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint: [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
16. Yang, L., et al.: TVFormer: trajectory-guided visual quality assessment on 360° images with transformers. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 799–808 (2022)
17. Martin, D., Serrano, A., Bergman, A.W., et al.: ScanGAN360: a generative model of realistic Scanpaths for 360° images. *IEEE Trans. Visual Comput. Graphics* **28**(5), 2003–2013 (2022)
18. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. *JMLR Workshop and Conference Proceedings*, pp. 315–323 (2011)
19. Duan, H., et al.: Perceptual quality assessment of omnidirectional images. In: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, pp. 1–5 (2018)
20. Sun, W., et al.: A large-scale compressed 360-degree spherical image database: from subjective quality evaluation to objective model comparison. In: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSp), Vancouver, BC, Canada, pp. 1–6 (2018)
21. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
22. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
23. Xu, M., Li, C., Chen, Z., et al.: Assessing visual quality of omnidirectional videos. *IEEE Trans. Circuits Syst. Video Technol.* **29**(12), 3516–3530 (2019)

24. Yu, M., Lakshman, H., Girod, B.: A framework to evaluate omnidirectional video coding schemes. In: 2015 IEEE International Symposium on Mixed and Augmented Reality, Fukuoka, Japan, pp. 31–36 (2015)
25. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Pacific Grove, CA, USA, Vol. 2, pp. 1398–1402 (2003)
26. Zhang, W., Ma, K., Yan, J., et al.: Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circuits Syst. Video Technol.* **30**(1), 36–47 (2020)
27. Sun, Y., Lu, A., Yu, L.: Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Process. Lett.* **24**(9), 1408–1412 (2017)