2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Omnidirectional Image Quality Assessment with TransVGG and Fused Saliency Guidance

Xican Tan[1], Jing Yu[2], Keke Tong[1], Shengfeng Lou[3], Jingsong Meng[1], Chuang Ma[4] and Wenzhi Chen[5](✉)

[1] International College, Chongqing University of Posts and Telecommunications, Chong Qing 400044, China
[2] Blockchain Security Research Center, Hengxin Technology Ltd, Shanghai 200336, China
[3] School of Computer Science and Technology (School of Artificial Intelligence), Zhejiang Sci-Tech University, Hang Zhou 310000, China
[4] School of Software Engineering, Chongqing University of Posts and Telecommunications, Chong Qing 400044, China
[5] School of Computer Science and Technology (School of Artificial Intelligence), Chongqing University of Posts and Telecommunications, Chong Qing 400044, China
s230201014@stu.cqupt.edu.cn

Abstract. Most existing omnidirectional image quality assessment (OIQA) models focus on locally salient regions within viewports, neglecting the critical guiding role of global saliency in holistic quality evaluation. This limitation restricts their performance when processing complex images. To tackle this, we propose a TransVGG-based OIQA framework guided by fused saliency map. First, the SalBiNet360 network is employed to generate fused saliency maps that combine local and global saliency information, simulating human viewing behavior during omnidirectional image observation. Then a collaborative architecture integrating Swin-Transformer and VGG has been designed to synergistically extract global and local features, thereby resolving the insufficiency of diverse guidance information. To enhance long-sequence data processing, the Mamba model is utilized for efficient omnidirectional image comprehension. Then a parallel hybrid attention mechanism is introduced to retrieve semantic features from saliency feature and guide the global understanding module. Experiments carried out on two OIQA datasets demonstrate that the proposed model outperforms advanced methods in performance.

Keywords: Omnidirectional Image Quality Assessment, Parallelized Channel-and-spatial Attention Mechanism, TransVGG, Fused Saliency Guidance.

1 Introduction

The proliferation of virtual/augmented reality (VR/AR) has made omnidirectional images (OIs) critical for immersive experiences, with applications in education, healthcare, entertainment, and autonomous driving (e.g., Tesla’s road monitoring). However, OIs often suffer from blurring, artifacts and geometric distortions during processing, degrading users’ experience and causing discomfort (e.g., dizziness) [1]. So accurate quality assessment is essential for optimizing VR pipelines and device performance.

The current landscape of objective OIQA techniques encompasses two primary frameworks: full-reference (FR) and no-reference (NR) methodologies [2]. FR-based metrics, typified by peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), operate under the premise of direct access to unimpaired source images. In contrast, NR methods eliminate this dependency, making them more practical for real-world scenarios due to lower computational costs and higher stability. The existing NR-OIQA frameworks are commonly classified into two distinct paradigms: projection-domain techniques and viewport-oriented approaches.

Although many NR-OIQA methods [3,4] have shown high performance on various datasets, there is still room for improvement. For instance, saliency-guided models like SGC [6] focus solely on local viewport regions, ignoring the guiding role of global saliency in holistic quality assessment. Additionally, models such as PICS [5] downsample high-resolution OIs to handle long-sequence features, leading to information loss. Traditional Vision Transformers (ViTs) and graph convolutional networks (GCNs) (e.g., VGCN [7]) struggle to balance global-local feature extraction, limiting their ability to capture comprehensive saliency semantics.

To address these issues, an OIQA model based on TransVGG saliency guidance is proposed. First, we introduce the ResNet-based SalBiNet360 network to generate the fused saliency map for OIs. This saliency map incorporates a global saliency map on the basis of the local saliency map, and finally fuses them into an overall saliency map. Secondly, a combined module of Swin-Transformer and VGG is adopted to extract the attention-grabbing components within the fused saliency map. In light of the long-sequence characteristics of OIs, the Mamba model is introduced to interpret the semantic features in OIs. For the dynamic selection of saliency features, a parallelized channel-and-spatial attention mechanism is employed to dynamically select appropriate saliency image semantics, followed by the fusion and guidance of OI features. Experiments conducted on two OIQA datasets show that its performance surpasses that of most models.

The contributions are summarized as follows:

- Motivated by the Human Visual System (HVS), the SalBiNet360 network is employed to generate local and global saliency maps, which are then fused into an integrated saliency map. The effectiveness of this saliency map generator in replicating human visual perception is validated through ablation experiments.
- We use a combination of Swin-Transformer and VGG to extract features from the fused saliency map. TransVGG is pre-trained on the ImageNet dataset for multiple rounds to improve the performance of the model. The features output by VGG in the three groups are fed into the parallelized channel-and-spatial attention mechanism for feature selection.
- To address the understanding of long-sequence data in OIs, the Mamba model is first pre-trained and then applied to analyze such data. The superiority of our approach is validated through comprehensive experiments on two existing OIQA databases, where it outperforms numerous state-of-the-art metrics.

2 Related Work

2.1 FR-OIQA

Early Full-Reference Image Quality Assessment (FR-IQA) methods mainly rely on simple statistical features, such as Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR). MSE measures the difference between reference images and distorted images by calculating the average of squared differences of corresponding pixel values, while PSNR, defined based on MSE, quantifies the signal-to-noise ratio of images [8]. However, these methods ignore the characteristics of the Human Visual System (HVS), leading to significant deviations between evaluation results and human subjective perception. To better consider the characteristics of HVS, the Structural Similarity Index (SSIM) was proposed [9]. SSIM evaluates image quality by calculating the brightness, contrast, and structural similarity of images, which can more accurately reflect human perception of image structural information. Since then, many improved methods based on SSIM have emerged. Due to the spherical characteristics of panoramic images, numerous studies have been dedicated to extending existing FR-IQA metrics to panoramic images. Yu et al. [10] proposed a Spherical-PSNR (S-PSNR) method, which calculates PSNR by uniformly sampling a set of points on the sphere. Sun et al. [11] introduced Weighted Spherical PSNR (WS-PSNR), assigning weights according to stretched regions. Chen et al. [12] proposed Spherical Structural Similarity Index (S-SSIM) to calculate image similarity in spherical format. Zhou et al. [13] designed a weighted-to-spherically-uniform SSIM metric, which can alleviate the pixel redundancy problem in stretched regions. Nevertheless, FR-OIQA methods face the challenge of difficult access to distortion-free reference panoramic images, limiting their applications in real-world scenarios.

2.2 NR-OIQA

In the research of No-Reference Omnidirectional Image Quality Assessment (NR-OIQA) methods, with the rise of Convolutional Neural Networks (CNN), deep learning-based NR-OIQA methods have continuously emerged. Kim et al. [14] designed adversarial networks to evaluate local quality, determining weights based on latitude and longitude. PICS [15] proposed a complete network to address the visual inconsistency of ERP-format images. Sun et al. [16] proposed a Multi-Channel Network (MC360IQA) to extract features from six viewports in CMP format. Xu et al. [7] proposed a Viewport-Oriented Graph Convolutional Network (VGCN) to explore local viewport spatial relationships, combining with DBCNN [17] to evaluate the global information of panoramic images. Fu et al. [3] developed an Adaptive Hypergraph Convolutional Network (AHGCN) to capture spatial and content correlations between viewports. Qiu and Shao [18] proposed a Saliency-Guided Network (SGC) for OIQA. Sui [19] proposed a viewport generation-based OIQA model. SAL360IQA [20] emphasizes key regions by integrating the saliency of panoramic images. Zhou et al. [21] utilized panoramic image saliency maps to assign weights for different viewing directions.

3 Models

The architecture of this model, as shown in Fig. 1, is primarily structured with three core modules. The Saliency Map Generation (SMG) module is devised to mimic the perceptual experience of human image viewing. Simultaneously, the Global Understanding Module (GUM) is established to perceive the semantic information within OIs. Furthermore, the Salient Features Guidance Module (SFGM) enables feature extraction and guidance, ultimately yielding an image quality score.

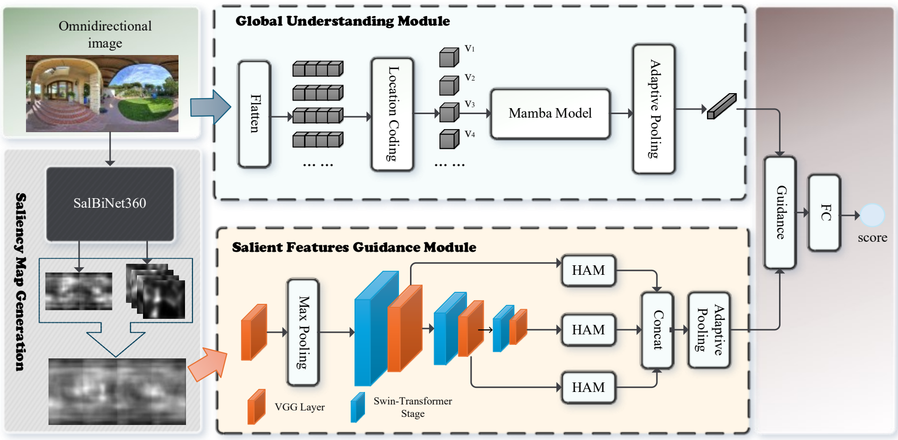


Fig. 1. The network framework diagram of the TransVGG model

3.1 Saliency Map Generation (SMG)

To address the limitation of insufficient local saliency guidance in quality assessment [4,18], our framework incorporates a modified SalBiNet360 architecture [22] based on ResNet50. This approach constructs a dual-branch network where one branch predicts global saliency and the other refines local details, combining their outputs through a weighted fusion mechanism to get final saliency map.

For global saliency prediction, the network splits after the second ResNet50 convolutional block. The entire OI is processed through three multi-scale contextual modules (MCMs) applied to the last three convolutional blocks' features. These modules, inspired by hierarchical feature integration strategies, extract multi-scale contextual information using kernel sizes of varying dimensions. A multi-level decoder then aggregates these hierarchical features to generate the global saliency map. Local saliency prediction involves projecting the 360° image onto six cubic faces (90° FOV each) to reduce distortion. The local subnetwork processes these rectilinear projections using a single MCM, designed to minimize redundancy in non-salient regions. After channel reduction and residual fusion, a single-level decoder restores the resolution, and reprojected maps are averaged to form the local saliency output.

The final saliency map is generated via linear combination:

$$S_{Fused} = \alpha S_G + \beta S_L \tag{1}$$

where α and β balance global and local contributions. This fusion strategy ensures comprehensive coverage of both global scene context and localized visual attention.

3.2 Global Understanding Module (GUM)

Due to the large scale of the original OIs, which contain a great deal of texture and semantic information, the Mamba model, which is good at handling long sequences, is adopted to process and understand OIs. First, to enable the model to obtain position encoding more effectively, the images are flattened and then fed into a long-sequence encoding process. The position encoding is obtained by flattening the images and then adding position information to each image patch F_{pos} .

Subsequently, the data enters the core processing part of the Mamba model. First, the data goes through normalization to stabilize its distribution and then undergoes convolution:

$$F_{conv} = Conv(Normalized(F_{pos})) \tag{2}$$

After introducing non-linearity through the SiLU activation function, the data enters the Selective State Space Model (SSM). The Selective SSM can dynamically select which states need to be retained and propagated and which can be suppressed based on the input. When processing OI data, it can effectively capture the semantic associations between different regions of the image, even if these regions are spatially separated.

The update and reset gates are specifically designed to modulate the merging extent of current input features with historical state representations:

$$h = \tanh(W_h \odot h_{prev}) + W_x F_{conv} + b_h \tag{3}$$

$$h = z \odot h_{prev} + (1-z) \odot h \tag{4}$$

In this formulation, h_{prev} corresponds to the latent state from the prior temporal step, while $\tilde{h}$ designates the proposed intermediate state. The variable h, conversely, embodies the revised latent representation computed at the current time step. The tanh function refers to the hyperbolic tangent activation function, while W_h, W_x is the bias vector. z is the update gate and b_h is bias term. Through the control of update and reset gates, the Selective SSM dynamically determines which states to retain or propagate and which to suppress based on input, enabling effective capture of long-range semantic correlations between spatially distant regions in images.

The selected features $F_{selected}$ are then fed into an adaptive pooling layer to normalize their dimensions, facilitating subsequent feature fusion:

$$F_{out} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{H} F_{selected}(i,j,c) \tag{5}$$

3.3 Salient Features Guidance Module (SFGM)

First, the saliency map undergoes max pooling for feature extraction and dimensionality reduction before being fed into a hybrid backbone network combining Swin-Transformer and VGG. The Swin-Transformer serves as the prefix network for three groups, while pre-trained VGG acts as the suffix network for each group to extract fused global and local information from the saliency map. The outputs $F_{\mathrm{i}}(i \in \{1,2,3\})$ of each VGG layer in the three groups are then fed into a parallel Hybrid Attention Mechanism (HAM) to focus on critical information within the saliency maps, as shown in Fig. 2. Specifically, for each group of features:

$$F_{\max} = \max(F_i) \tag{6}$$

$$F_{avg} = AvgPool(F_i) \tag{7}$$

Then adding the results F_{max} and F_{avg} together, they are respectively used as the inputs of the multi-layer perceptron (MLP) and output as F_{mlp}. At the same time, the features of the two are concatenated as F_{concat} and then fed into the convolutional layer, thus realizing the parallelization of channel and spatial attention. Subsequently, the convolutional result F_{conv} and the output of the MLP F_{mlp} are input into the Sigmoid function to gain the attention weights a_{ch} and a_{sp}. After weighted summation of the attention weights, they are multiplied by the original features to purposefully emphasize the eye-catching parts in the features.

$$F_{sa} = (a_{ch} + a_{sp}) \odot F_i \tag{8}$$

Where ⊙ represents element-wise multiplication. After the three groups of features are processed by the HAM, they are fused. To achieve the guidance of saliency, the features of the OI F_{out} and the saliency map feature F_{sa} are continuously and dynamically fused. Then, a fully connected layer (FC) and a Multi-Layer Perceptron (MLP) are used for quality prediction to obtain the final image score.

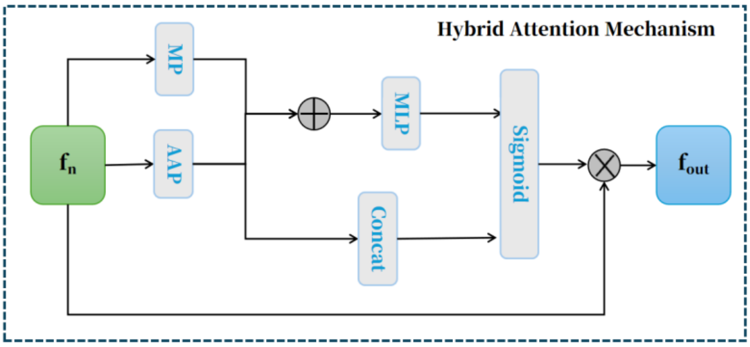


Fig. 2. Parallelized channel-and-spatial attention mechanism network framework.

4 Experiments

4.1 Datasets

1. The OIQA database consists of 336 original OIs in the equirectangular format. Among them, there are 16 undistorted original images and 320 distorted images. The distorted images are obtained by superimposing five-level distortions on the OIs. The mean opinion score (MOS) ranges from 1 to 10.
2. The CVIQ database provides a total of 544 original OIs, including 16 OIs and 528 distorted images. The distorted images are generated by 11-level compression distortions of three types. MOS ranges from 0 to 10.

4.2 Metrix

- Spearman Rank Correlation Coefficient (SRCC)

$$SRCC = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{9}$$

where N is the number of image samples, and d_i represents the rank difference between the subjective and objective evaluations of the i~th image.

- Pearson Linear Corrletion Coefficient (PLCC)

$$PLCC = \frac{\sum_{i=1}^{N}(s_i-\mu_{s_i})(o_i-\mu_{o_i})}{\sqrt{\sum_{i=1}^{N}(s_i-\mu_{s_i})^2\sum_{i=1}^{N}(o_i-\mu_{o_i})^2}} \tag{10}$$

where s_i and o_i denote the i~th subjective and mapped objective quality values respectively, and μ_{s_i} and μ_{o_i} represent the corresponding average values of s_i and o_i respectively.

4.3 Evaluation

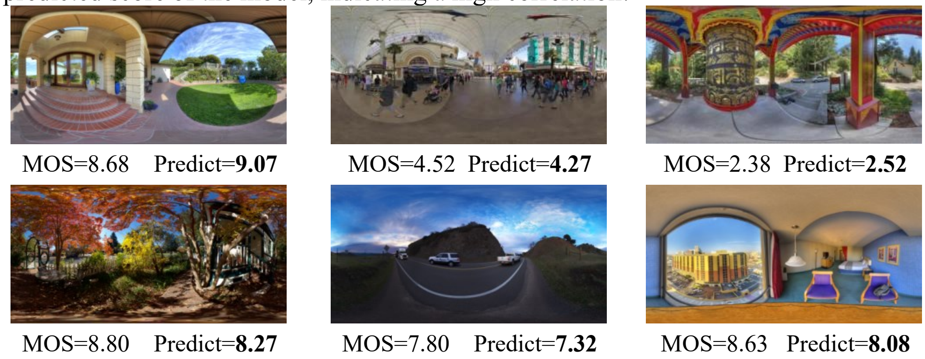
In the experiment, our method is implemented in the Pytorch framework and the experiment is conducted on a computer with Intel i7-12700H CPU with 15.6 GB RAM and an NVIDIA GeForce RTX3060 GPU. The training and test datasets are divided at a ratio of 8:2. Training of the model is conducted for 300 epochs on both CVIQD and OIQA datasets to achieve complete convergence. To mitigate the uncertainty of the training-test split, this random splitting and cross-validation process is result and the average of the ten results is taken as the final performance.

As shown in Table 1, compared with the Full-Reference (FR) model S-SSIM, the TransVGG model has an improvement of approximately 6.2% in the SRCC index and about 2.4% in the PLCC index on the CVIQD dataset. On the OIQA dataset, the SRCC index of the TransVGG model is increased by about 4.1%, and the PLCC index is increased by approximately 0.2%. When compared with the No-Reference (NR) Assessor360 model, the TransVGG model shows an improvement of roughly 0.4% in the SRCC index on the CVIQD dataset. On the OIQA dataset, the SRCC and PLCC indices of the TransVGG model are improved by about 1.0% and approximately 0.2% respectively. Overall, the TransVGG model has improvements to varying degrees in multiple indices compared with other models, demonstrating its performance advantages.

Table 1. Performance evaluations of the proposed methodology versus FR-OIQA and NR-OIQA baselines are presented on the CVIQD and OIQA datasets.

Types	Methods	CVIQD		OIQA	
		SRCC	PLCC	SRCC	PLCC
FR	PSNR	0.802	0.920	0.511	0.565
	SSIM	0.674	0.667	0.385	0.289
	S-PSNR	0.708	0.708	0.540	0.600
	S-SSIM	0.931	0.944	0.617	0.652
NR	DBCNN	0.949	0.963	0.927	0.931
	MC360IQA	0.914	0.951	0.919	0.925
	MFILGN	0.640	0.695	0.642	0.673
	AHGCN	0.962	0.964	0.959	0.965
	VGCN	0.967	0.977	0.955	0.964
	GSR	0.944	0.962	0.945	0.954
	SGC	0.956	0.953	0.964	0.966
	Assessor360	0.964	**0.977**	0.964	0.964
	Proposed.	**0.971**	0.970	**0.965**	**0.967**

As shown in Fig**.** 3, after the model is trained, several random images are tested. MOS stands for the score of the image after scoring, and Predict is the score of the image after the model prediction. The results show that the MOS of each OIs is close to the predicted score of the model, indicating a high correlation.



MOS=8.68 Predict=**9.07**

MOS=4.52 Predict=**4.27**

MOS=2.38 Predict=**2.52**

MOS=8.80 Predict=**8.27**

MOS=7.80 Predict=**7.32**

MOS=8.63 Predict=**8.08**

Fig. 3. Comparison of subjective scores of different images and scores predicted by the proposed algorithm

4.4 Saliency Method Comparison

Given the model's substantial reliance on the saliency map's highlighting capabilities, a comparative assessment of three saliency generation techniques is conducted to identify the most optimal approach for this model. The results of this analysis are presented in Table 2. They are SalBiNet360 [22], SalNet360 [23], and U-net [24] respectively. These three methods adopt different architectures and algorithmic ideas when processing saliency maps. SalBiNet360 utilizes a unique bidirectional network structure to enhance the capture of salient information in images; SalNet360 focuses on the feature extraction method of cube projection to generate saliency maps, while U-net adopts an architecture with an encoder-decoder and skip connections.

Table 2. Performance of different saliency map generators in finally evaluation

Dataset	CVIQ		OIQA	
Method	SRCC	PLCC	SRCC	PLCC
SalBiNet360	**0.971**	**0.970**	**0.965**	**0.967**
SalNet360	0.967	0.964	0.956	0.950
U-Net	0.950	0.943	0.942	0.937

In the CVIQ dataset, SalBiNet360's SRCC and PLCC values reach 0.971 and 0.970 respectively. For SalNet360, these values are 0.967 and 0.964, while U-Net has SRCC of 0.950 and PLCC of 0.943. As depicted in Fig**.** 4, when it comes to SRCC and PLCC indicators, the blue and orange columns symbolizing SalBiNet360 outperform those of the other two methods.

Regarding the OIQA dataset, SalBiNet360 shows an SRCC value of 0.965 and a PLCC value of 0.967. SalNet360 has an SRCC of 0.956 and a PLCC of 0.950, and U-Net's SRCC is 0.942 with a PLCC of 0.937. Similarly, in the figure, the green and red columns associated with SalBiNet360 top the list in these two indicators.

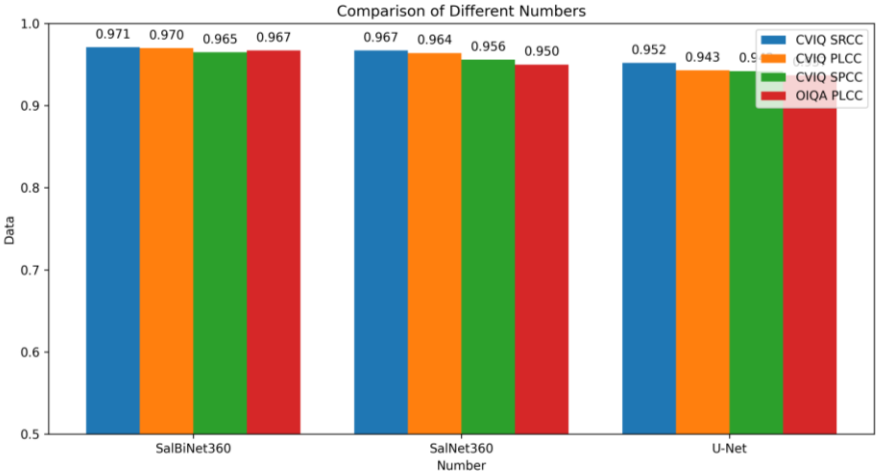


Fig. 4. Performance of different saliency map generators

4.5 Ablation analysis

The ablation experiment sets up three different experimental groups, namely SV1, SV2, and SV3. Among them, SV1 represents the model using only one group of global-local modules, that is, a combination of one group of Swin-Transformer and a VGG Layer; SV2 uses two groups of this combination; and SV3 uses three groups.

As can be seen from Table 3 and Fig**.** 5, on the CVIQ dataset, the SRCC value of SV3 is 0.971, which is approximately a 1.8% increase compared to 0.954 of SV2; the PLCC value increases from 0.950 of SV2 to 0.970 of SV3, with an increase of about 2.1%. On the OIQA dataset, the SRCC value of SV3 is 0.961, which is approximately a 1.4% increase compared to 0.948 of SV2; the PLCC value increases from 0.937 of SV2 to 0.963 of SV3, with an increase of about 2.8%.

Table 3. Performance of different number of TransVGG modules

Dataset	CVIQ		OIQA	
Types	SRCC	PLCC	SRCC	PLCC
SV3	**0.971**	**0.970**	**0.965**	**0.967**
SV2	0.954	0.950	0.948	0.937
SV1	0.932	0.943	0.921	0.927

This indicates that by increasing the number of groups of Swin-Transformer and VGG, SV2 can significantly optimize the performance compared to before, enabling the model to learn richer features. However, when it comes to SV3, the improvement range of some indicators slows down, and there are fluctuations in the model performance. Such a trend of performance change provides an important reference for the optimization and selection of the structure of the TransVGG model.

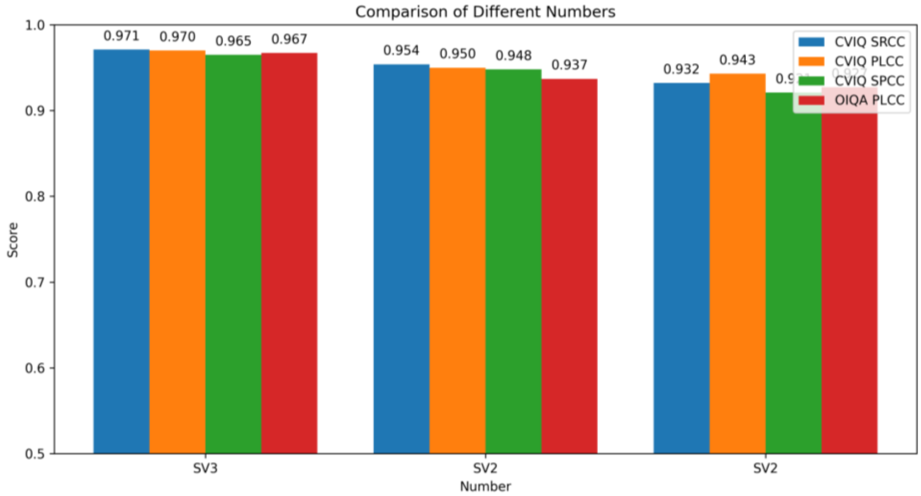


Fig. 5. Performance of the number of different numbers of TransVgg modules

5 Conclusion

In the research of OIQA, this paper has proposed a framework guided by global saliency maps based on TransVGG. To accurately simulate the perceptual experience of humans when viewing images, the model has constructed a Saliency Map Generation (SMG) module, which has effectively located the key salient regions in the image. Meanwhile, a Global Understanding Module (GUM) has been proposed to comprehensively perceive the semantic information of OIs. In addition, the Salient Feature Guided Module (SFGD) has deeply extracted and precisely guided the salient features of the image. This model has addressed the problems of insufficient local salient information and inadequate semantic guidance for saliency. Moreover, it has demonstrated advanced performance and strong robustness in both overall experiments and ablation experiments.

Reference

1. Zhai, G.T., Min, X.K.: Perceptual image quality assessment: a survey. Science China (Information Sciences) **63**(11), 84–135 (2020)
2. Ai, D., Bai, Y.S., Yu, K.X., Yuan, H., Liu, Y.: Recent progress of panoramic image quality assessment methods. Computer Engineering and Applications **58**(24), 1–11 (2012)
3. Fu, J., Hou, C., Zhou, W., Xu, J.H., Chen, Z.B.: 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)
4. Wu, T.H., Shi, S.W., Cai, H.M., Cao, M.D., Xiao, J., Zheng, Y.Q., Yang, Y.J.: Assessor360: Multi-sequence Network for Blind Omnidirectional Image Quality Assessment. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 64957–64970 (2023)
5. Zhou, Y., Ding, Y.Y., Sun, Y.J., et al.: Perceptual Information Completion-Based Siamese Omnidirectional Image Quality Assessment Network. IEEE Transactions on Instrumentation and Measurement 73, 1–10 (2024)
6. Zhang, Y., Wan, L., Liu, D., Zhou, X., An, P., Shan, C.: Saliency-Guided No-Reference Omnidirectional Image Quality Assessment via Scene Content Perceiving. IEEE Transactions on Instrumentation and Measurement 73, 5039115 (2024)
7. Xu, J., Zhou, W., Chen, Z.: Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. IEEE Trans. Circuits Syst. Video Technol. 31(5), 1724–1737 (2021)
8. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)
9. Wang, Z., Bovik, A.C.: Mean squared error: Love it or leave it? - A new look at signal fidelity measures. IEEE Signal Processing Magazine **19**(1), 98–117 (2002)
10. Yu, M., Lakshman, H., Girod, B.: A framework to evaluate omnidirectional video coding schemes. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, pp. 31–36 (2015)
11. Sun, Y., Lu, A., Yu, L.: Weighted-to-spherically-uniform quality evaluation for panoramic video. IEEE Signal Processing Letters **24**(9), 1408–1412 (2017)

12. Chen, S., Zhang, Y., Li, Y., Chen, Z., Wang, Z.: Spherical structural similarity index for objective omnidirectional video quality assessment. In: Proceedings of the IEEE International Conference on Multimedia Expo (ICME), pp. 1–6 (2018)
13. Zhou, Y., Yu, M., Ma, H., Shao, H., Jiang, G.: Weighted-to-spherically-uniform SSIM objective quality evaluation for panoramic video. In: 14th IEEE International Conference on Signal Processing (ICSP), pp. 54–57 (2018)
14. Kim, H.G., Lim, H., 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)
15. Zhou, Y., Ding, Y., Sun, Y., Li, L., Wu, J., Gao, X.: Perceptual information completion-based Siamese omnidirectional image quality assessment network. IEEE Trans. Instrum. Meas. 73, 1–10 (2024)
16. Sun, W., Min, X., Zhai, G., Gu, K., Duan, H., Ma, S.: MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment. IEEE Journal of Selected Topics in Signal Processing (JSTSP) **14**(1), 64–77 (2020)
17. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans. Circuits Syst. Video Technol. 30(1), 36–47 (2020)
18. Zhang, Y., Wan, L., Liu, D., Zhou, X., An, P., Shan, C.: Saliency-Guided No-Reference Omnidirectional Image Quality Assessment via Scene Content Perceiving. IEEE Transactions on Instrumentation and Measurement 73, 5039115 (2024)
19. Sui, X., Zhu, H., Liu, X., Fang, Y., Wang, S., Wang, Z.: Perceptual Quality Assessment of 360° Images Based on Generative Scanpath Representation. Submitted to IEEE Transactions on Image Processing (2024)
20. Sendjasni, A., Larabi, M.-C.: SAL-360IQA: A saliency weighted patch-based CNN model for 360-degree images quality assessment. In: Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW), pp. 1–6 (2022)
21. Zhou, M., et al.: Perception-oriented U-shaped transformer network for 360-degree no-reference image quality assessment. IEEE Trans. Broadcast. 69(2), 396–405 (2023)
22. Chen, D., Qing, C., Xu, X., Zhu, H.: SalBiNet360: Saliency Prediction on 360° Images with Local-Global Bifurcated Deep Network. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 92–100 (2020)
23. Monroy, R., Lutz, S., Chalasani, T., Smolic, A.: SalNet360: Saliency maps for omni-directional images with CNN. Signal Processing: Image Communication 26–34 (2018)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, pp. 234–241 (2015)