Dual-Level Blind Omnidirectional Image Quality Assessment Network Based on Human Visual Perception

Deyang Liu[1], Lu Zhang[2], Lifei Wan[3], Wei Yao[4], Jian Ma[5]*, Youzhi Zhang[6]
School of Computer and Information, Anqing Normal University, Anqing, 246000, China[1,2,3,6]
School of Teacher Education, Anqing Normal University, Anqing, 246000, China[4]
School of Computer Science, Fudan University, Shanghai 200433, China[5]

***Abstract*—With the rapid development of virtual reality (VR) technology, a large number of omnidirectional images (OIs) with uncertain quality are flooding into the internet. As a result, Blind Omnidirectional Image Quality Assessment (BOIQA) has become increasingly urgent. The existing solutions mainly focus on manually or automatically extracting high-level features from OIs, which overlook the important guiding role of human visual perception in this immersive experience. To address this issue, a dual-level network based on human visual perception is developed in this paper for BOIQA. Firstly, a human attention branch is proposed, in which the transformer-based model can efficiently represent attentional features of the human eye within a multi-distance perception image pyramid of viewport. Then, inspired by the hierarchical perception of human visual system, a multi-scale perception branch is designed, in which hierarchical features of six orientational viewports are considered and obtained by a residual network in parallel. Additionally, the correlation features among viewports are investigated to assist the multi-viewport feature fusion, in which the feature maps extracted from different viewports are further measured for their similarity and correlation by the attention-based module. Finally, the output values from both branches are regressed by fully connected layer to derive the final predicted quality score. Comprehensive experiments on two public datasets demonstrate the significant superiority of the proposed method.**

***Keywords*—Omnidirectional image quality assessment; dual-level network; human visual perception; human attention; multi-scale**

I. Introduction

Virtual reality (VR), as the most popular immersive multimedia, can offer a unique 360-degree visual experience which sets it apart from traditional two-dimensional (2D) formats. Users can explore omnidirectional images (OIs) by wearing VR devices such as head-mounted displays (HMDs). However, the qualities of OIs are degraded during the processes such as stitching, projecting, encoding, and transmitting, which further influence the user experiences, even cause motion sickness. Therefore, the quality evaluation of OIs plays a significant role in guiding OI processing and ensuring a high quality of experience.

In the past few years, many objective OIQA methods have been proposed, including full-reference (FR) type and blind/no-reference (B/NR) type. For FR type, the peak-signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) [1] are respectively adopted for OIQA [2, 3, 4, 5, 6]. However, obtaining the undistorted OIs is challenging due to the complexity of image processing, making FR-OIQA challenging in practical applications. Therefore, it is crucial to develop blind/no-reference omnidirectional image quality assessment (BOIQA/NR-OIQA) methods that can evaluate the quality of OIs without reference images. Regarding the NR-OIQA type, many approaches [7, 8, 9, 10] commonly involve filtering to analyze the frequency domain information or natural scene statistics (NSS) to find statistical regularities in OIs. However, the manual feature designing is challenging [11], which limits the robustness of those methods. To relieve this issue, many data-driven approaches are proposed, which are able to learn inherent relationships between the predicted values and the ground-truth labels. These methods typically consist of two steps: feature extraction and quality regression. Specifically, Convolutional Neural Networks (CNNs) are firstly used to extract high-level features from OIs. Then, fully connected layers are employed for regression to obtain the predicted quality scores. However, most data-driven solutions are directly transferred from 2D IQA methods, in which the features of OIs are extracted in EquiRectangular Projection (ERP) format. Moreover, those approaches do not consider the human visual perception during OIQA. Although several approaches [12, 13, 14, 15] try to extract the viewport (VP) images from OIs to replace the ERP as the inputs, the human visual perception is still under-explored.

Generally, people tend to pay more attention to some contents of interest rather than the entire VP with HMD. This means that the regions of interest in VP are more likely to contribute to the quality rating [15]. Furthermore, the objects in nature are usually captured by the human eyes at various scales [16], which means that the human visual perception of an OI is formed through multiple views from different directions at various viewing distances.

Based on the above analysis, we can conclude that the quality of immersive media experiences is more susceptible to subjective visual perception by humans. However, recent works on OIQA primarily analyze images and overlook the active nature of human visual perception in this process. To address this gap and further to enhance the OIQA performance, this paper proposes a dual-level BOIQA network based on human visual perception. The proposed method tries to explore the human visual perception from two aspects including the human attention and the multi-scale perception. Specifically, the proposed method is a dual-level model, which is composed of three parts: human attention branch (HAB), multiscale perception branch (MPB) and quality regression (QR). For the

HAB, to emphasize the regions of interest, an improved Vision Transformer (ViT) [17] is integrated with the residual CNNs, which enables the proposed network to capture attention-based features within the VP images without disrupting the hierarchical perception. In HAB, the CNNs are responsible for obtaining high-level feature maps of each VP image, while the ViT calculates the attention weights. Furthermore, in order to explore more information within the VP region, we also introduce an image pyramid to represent different viewing distances of each VP in HAB. Regarding the MPB, we first establish a parallel structure to extract multi-scale information from each VP in cubemap projection (CMP) format. Then, to explore the content correlations between VPs at different positions in an OI, we develop a correlation feature fusion module to establish the long-range dependencies among VPs. Finally, the obtained dual-level perception features are regressed through the QR module to predict the final quality scores. Extensive experimental results have validated the effectiveness of the proposed approach. The contributions of the proposed method are listed as follows:

- We propose a BOIQA network based on the human visual perception, in which the region of interest can better be highlighted in a VP region and the multi-scale information can be obtained from low-level to high-level based on multiple views.
- We establish the multiple viewing distances image pyramid of the front VP and obtain the attention-based features from it to explore more information within the VP region. Moreover, we fuse the multi-scale features extracted from each VP in CMP and the obtained attention-based features to explore the content correlations between VPs.
- Comparisons with the state-of-the-art metrics on two public databases demonstrate the strength of our method.

II. Related Works

Generally, OIQA methods can be classified into two categories: subjective methods and objective methods. Subjective OIQA method involves participants directly providing subjective quality scores for the OIs they view in an HMDs. However, it is time-consuming and impractical for batch applications. By contrast, objective OIQA method is more suitable for practical production applications. The objective OIQA method can be further divided into two categories: traditional OIQA metrics and deep learning-based OIQA metrics. This section will emphatically review the objective OIQA methods.

A. Traditional OIQA Metrics

Many works have extended the traditional common used IQA metrics to OIQA. For example, the evaluation schemes based on PSNR transfer the calculation from planar format to spherical format while still inheriting the main idea of per-pixel comparison in PSNR. Moreover, the evaluation schemes based on SSIM mainly focus on the ERP format of panoramic images and analyze metrics such as sharpness, contrast, and brightness.In [18], statistical characteristics of panoramic images were obtained using the adjacent pixels correlation (APC) features and blind quality prediction of panoramic images was then performed using support vector regression (SVR).

The methods that use the ERP as the evaluation basis are mostly borrowed directly from 2D-IQA and have made corresponding improvements for panoramic images. However, they still overlook the unique media characteristics of panoramic images and the geometric distortions present in ERP. Recent works have focused on extracting natural statistical information from other representations of panoramas. Zheng et al. [19] firstly converted the panoramic image from the ERP format to a segmented spherical projection format. They then utilized a heat map as a weighting factor to perceive features in both the two-level and equatorial regions. Zhou et al. [9] achieved panoramic image quality assessment score by analyzing multi-frequency information and statistically evaluating the local and global naturalness presented in both ERP and VP formats. Jiang et al. [8] explored the color information of each VP image unit in the rotated Cubemap Projection (CMP) format through tensor decomposition and piecewise exponential fitting. The above-mentioned works achieved satisfactory performance results by designing hand-crafted features through techniques such as machine learning. However, these manual features based approaches are evidently cumbersome and not easily comprehensive, which reduces the robustness of the proposed method.

B. Deep Learning-based OIQA Metrics

Deep learning-based OIQA approaches benefit from powerful model architectures that can capture more quality-relevant features within the images. Thanks to the guidance of large amounts of labeled data, this kind of methods often outperforms traditional methods. In [20, 21], Kim et al. proposed an adversarial learning-based human perception guider, which improves the prediction capability of deep learning models for quality scores by enhancing the human perception guider's discriminative ability for predicted scores and subjective quality score labels.

Although the aforementioned methods have achieved satisfactory results, they have not considered the differences between immersive media experience and traditional planar images perception. This restricts the feature representation capability of deep models. To address this issue, recent VP-based end-to-end models have been developed to accurately simulate the scene content that can be perceived by the human eye while viewing panoramic images at a moment. Considering the limited field of view of the human eye in head-mounted devices, Li et al. [12] firstly proposed a VP-based assessment scheme and combined it with CNNs for feature extraction. Sun et al. [13] proposed a multi-stream network that utilized the modified ResNet-34 to extract features from each VP in the rotated CMP format. Xu et al. [14] proposed a solution with local and global branches. The local branch utilized ResNet-18 to simultaneously extract internal features from multiple VP images and established connections between them using graph convolution. The global branch extracted feature information from the panoramic ERP format using the VGG [22] network.

These deep learning-based models possess powerful feature extraction capabilities and quality score fitting abilities. However, there is still significant room for improvement in terms of their consistency with the HVS. Therefore, in this paper, we draw inspiration from human visual perception and develop an end-to-end model to investigate the impact of

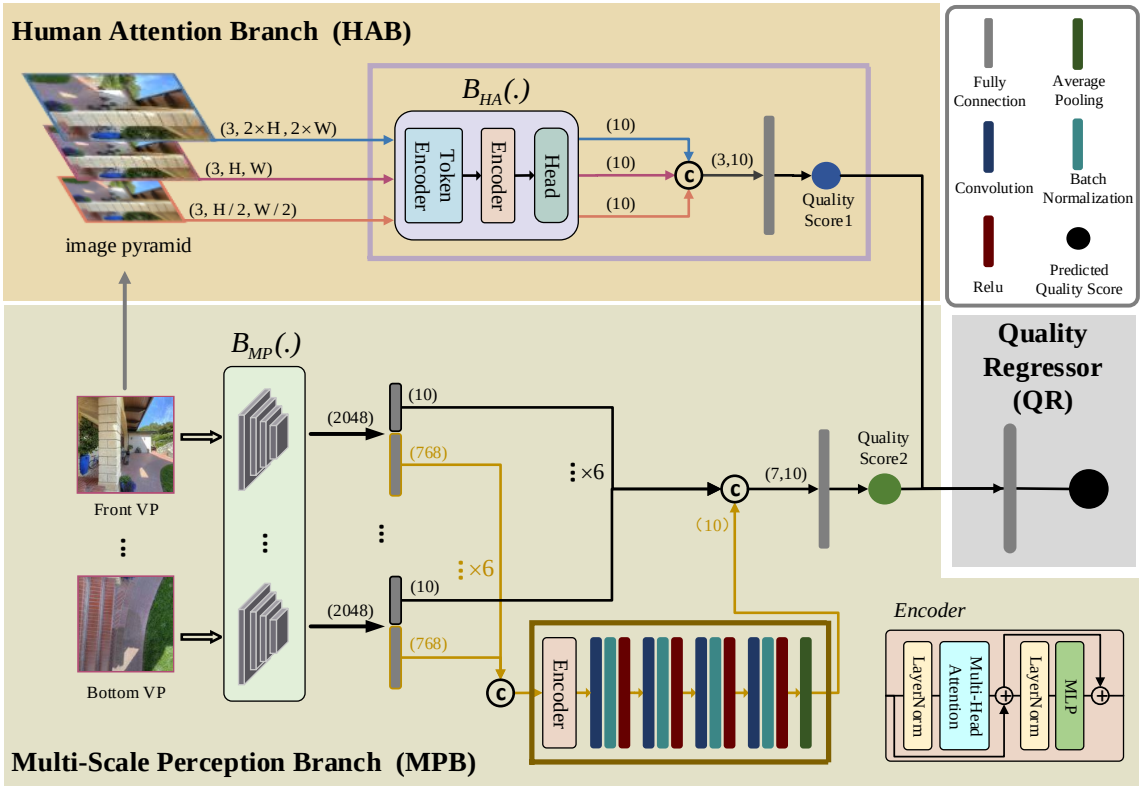
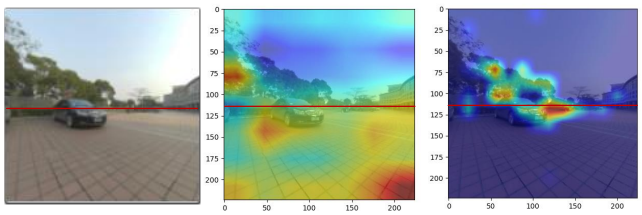


Fig. 1. The overall framework structure of our dual-level network

simulating human attention [23] and multi-scale [24] perception on panoramic quality assessment.

III. Proposed Method

The proposed dual-level BOIQA network contains three modules, namely human attention branch (HAB), multi-scale perception branch (MPB) and quality regression (QR). The overall framework structure is illustrated in Fig. 1. The HAB focusses on extracting the high-level information from a multiple viewing distances image pyramid based on attention mechanism. The MPB aims to explore multi-scale perception features of each VP and explore the correlation information among those VPs. The QR is utilized to predict perceptual quality scores.



(a) simple image (b) ResNet-50 (c) Hybrid model of ViT and ResNet-50

Fig. 2. the visualization of attention from the output features to a VP example from CVIQD database.The solid red line shows the position of the equator. (a) the simple image; (b) the learned feature maps of (a) only with ResNet-50; (c) the learned feature maps of (a) with the hybrid model of ResNet-50 and ViT.

A. Human Attention Network

Fig. 2 shows the visualization result of the attention weights on a VP example under different OIQA models. It is evident that the hybrid model combining CNNs and ViT pays more attention to the equatorial region and salient objects compared to a pure CNNs model. This aligns well with the attention habits of the HVS. Based on the above analysis, to obtain information that better aligns with human visual attention characteristics, the HAB branch is designed to extract internal features of VP images based on attention mechanisms. To further explore the comprehensive perception of the VP under different viewing distances, we also introduced a multiple viewing distances image pyramid of the front VP as the input of the HAB branch, which is shown in Fig. 1.

For the front VP initialized with a resolution of $H \times W$ through the center cropping operation, we increase the center cropping resolution to $2H \times 2W$ to represent a larger field of view with a longer distance. Similarly, we decrease the center cropping resolution to $\frac{H}{2} \times \frac{W}{2}$ to represent a smaller field of view with a closer distance. Therefore, the input image pyramid V_f^l, V_f^o, V_f^s of the front VP can be established with a multiple distances representation. Specifically, the V_f^o represents the VP in its original resolution, V_f^l represents a version with a higher resolution, and V_f^s represents a version with a lower resolution.

To fully explore the information based on human visual attention from this VP pyramid, we integrate the ResNet-50 and an improved ViT as the backbone for feature extraction. Specifically, each layer of the pyramid is firstly fed into the ResNet-50 in parallel. The semantic features of each view can

be obtained and then being converted into token forms. In our method, the improved ViT network consists of two stages. The first stage tires to compute the attention weight among tokens within each view by multi head attention (MHA). The second stage is used to further adjust the dimensionality of the obtained feature maps based on view-content attention through the Head module which is composed of fully connected layers. Finally, we fuse the extracted high-level features of each view at different distances, and obtain the quality score of this branch with a regression operation. This process can be expressed as:

$$\begin{cases} F_f^l, F_f^o, F_f^s = B_{HA}(V_f^l, V_f^o, V_f^s), \\ Q_1 = Linear(Cat(F_f^l, F_f^o, F_f^s)), \end{cases} \tag{1}$$

where $B_{HA}(.)$ represents the feature extraction network of the HAB, F_f^l, F_f^o, F_f^s separately represent the extracted features from the image pyramid under different distances. Each extracted features of the pyramid has a dimensionality of 10. $Cat(.)$ and $Linear(.)$ respectively donate the concatenate operation and the fully connected layer. Q_1 is the obtain predicted quality score of this process.

B. Multi-Scale Perception Network

In general, the user's comprehensive quality perception of a panoramic image is influenced by multiple VPs at different positions. It is necessary to perform multi-scale quality perception across various positions in the panoramic image and explore the correlation information between these VPs in terms of both location and content.

In this work, we have established a multi-scale perceptual branch as an auxiliary branch. Firstly, a group of VP images $V_u, V_d, V_l, V_r, V_f, V_b$ are achieved from a panoramic image at six directions (up, down, left, right, front, and back). As mentioned above, human visual perception is a hierarchical process that involves perceiving texture, contours, and high-dimensional semantics. Therefore, in this branch, ResNet-50 with residual structure is adopted as the backbone for feature extraction of each VP. The residual network is capable of capturing multi-scale perceptual features from low-level to high-level in each directional VP, which aligns well with the multi-scale perception of HVS. This process can be represented as:

$$F_i^m = B_{MP}(V_i), i \in \{u, d, l, r, f, b\}, \tag{2}$$

For each VP image V_i, multi-scale perceptual features F_i^m are simultaneously extracted through the feature extraction network. B_{MP} represents the backbone of this branch. It is worth noting that the multi-scale features obtained here have a dimensionality of 2048.

After obtaining the multi-scale features F_i^m obtained from multiple VP images, most methods propose to concatenate those high-dimensional features and perform quality regression. Conversely, in order to further capture the inter-viewpoint correlation information, we apply a fully connected layer to convert the multi-scale features corresponding to each VP into tokens with a dimensionality of 768. Subsequently, we perform element-wise multiplication operations based on attention mechanism among those tokens to obtain the correlational features. The specific formula representation is as follows.

$$\begin{cases} T_i^m = Linear(F_i^m) \\ F^c = E_{MP}(T_i^m) \end{cases}, \ i \in \{u, d, l, r, f, b\}, \tag{3}$$

Here, $Linear(.)$ represents the fully connected operation, T_i^m denotes the token corresponding to each VP after undergoing this fully connected operation. E_{MP} represents the Encoder network based on multi-head attention, and F^c represents the final correlation feature map among those VPs.

In order to further fuse the obtained correlation features and consider the spatial relationship between each VP, we further introduce a correlation feature fusion module. This module consists of four convolutional blocks and an average-pooling. Each convolutional block includes a convolution (Conv) layer, a batch normalization (BN) layer, and a Rectified Linear Unit (ReLU) activation. The convolutional operation calculates the internal correlations of the feature map F^c using a 3×3 receptive field, integrating the content-based correlation information between different VPs. This locally nested convolutional structure also helps compensate for the positional correlation between VPs that may be overlooked in the previous computations. Finally, an average-pooling operation is applied to obtain the fused correlation features through regression. The specific process is illustrated by the following equation.

$$F^{c'} = C_{MP}(F^c), \tag{4}$$

where C_{MP} represents the correlation feature fusion module, $F^{c'}$ is the achieved fused correlation features, whose dimensionality is 10.

We perform final feature regression on the multi-scale information obtained from each VP and the corresponding fusion information between them. Specifically, we first adjust each multi-scale feature map F_i^m to dimensionality 10. Then, we concatenate those adjusted feature maps with the multi-scale feature $F^{c'}$. Next, a fully connected operation is applied to the concatenated feature map for feature regression. This process can be described as:

$$\begin{cases} F_i^{m'} = Linear(F_i^m) \\ Q_2 = Linear(Cat(F_i^{m'}, F^{c'})) \end{cases} \quad i \in \{u, d, l, r, f, b\} \tag{5}$$

where $F_i^{m'}$ represents the multi-scale features of each VP after dimension adjustment. $Linear(.)$ and $Cat(.)$ represent the fully connected operation and concatenation operation, respectively. Q_2 denotes the quality score obtained from the final regression of this multi-scale perception branch.

C. Quality Regressor

The quality regressor consists of two steps. Firstly, we conduct concatenation operation of HAB and MPB. Afterwards, the predicted score is obtained by the final layer of fully connected. The training loss is described as follows:

$$\begin{cases} Q = Linear(Cat(Q_1, Q_2)) \\ L = |Q - MOS|^2, \end{cases} \tag{6}$$

where Q is the final predicted quality score. The MOS is the ground-truth label of OI, which also means the subjective

TABLE I. OVERALL PERFORMANCE COMPARISONS ON THE OIQA AND CVIQD DATABASES. THE BEST RESULTS ARE DENOTED IN BOLD.

	Database	OIQA			CVIQD		
Type	Methods	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
FR	PSNR	0.5812	0.5226	1.7005	0.7008	0.6239	9.9599
	S-PSNR	0.5997	0.5399	1.6721	0.7083	0.6449	9.8564
	WS-PSNR	0.5819	0.5263	1.6994	0.6729	0.6107	10.3283
	CPP-PSNR	0.5683	0.5149	1.7193	0.6871	0.6265	10.1448
	SSIM	0.8718	0.8588	1.0238	0.9002	0.8842	6.0793
	MS-SSIM	0.7710	0.7379	1.3308	0.8521	0.8222	7.3072
	FSIM	0.9014	0.8938	0.9047	0.9340	0.9152	4.9864
	DeepQA	0.9044	0.8973	0.8914	0.9375	0.9292	4.8574
NR	BRISQUE	0.8424	0.8331	1.1261	0.8376	0.8180	7.6271
	BMPRI	0.6503	0.6238	1.5874	0.7919	0.7470	8.5258
	DB-CNN	0.8852	0.8653	0.9717	0.9356	0.9308	4.9311
	MC360IQA	0.9267	0.9139	0.7854	0.9429	0.9428	4.6506
	VGCN	0.9584	0.9515	0.5967	0.9651	0.9639	3.6573
	Ours	**0.9598**	**0.9530**	**0.5862**	**0.9680**	**0.9664**	**3.5014**

quality score. The L represents the loss between S and MOS, the $|.|$ represents absolute value operation.

IV. EXPERIMENTAL RESULTS

Our experiment uses two popular public datasets, namely CVIQD [25] and OIQA[26]. They both include 16 original panoramic images with different types and degrees of distortion. The former includes 528 compressed images generated by JPEG, H.264/AVC and H.265/HEVC standards, and the subjective score label of it ranges from 1 to 100. The latter contains 320 distorted images generated by four distortion types: JPEG compression (JPEG), JPEG2000 compression (JP2K), Gaussian blur (GB) and Gaussian white noise (GN), and the subjective ground-truth label of it ranges from 1 to 10.

A. Experimental Settings

Our experiments were conducted with 11^{th} Gen Intel(R) Core(TM) CPU i7-11700F @ 2.50GHz, 16 GB RAM, NVIDIA RTX 3060. The batch size was set to 4 and the learning strategy was RMSprop [27] whose learning rate is initialized to 0.0001. The rotation angle for the rotated CMP was fixed to 4 serving as data augmentation and the VP image resolution $H \times W$ is set to 256×256. Each database is split into training and testing sets according to the standard ratio of 8:2. This means that the distorted images corresponding to 3 reference images are randomly selected as testing set and the remaining are regarded as the training set. During the training phase, we use the pretrain results of ImageNet to the HAB and the MPB's backbone. By transferring the model training parameters from a large dataset to our task-specific dataset, we can achieve significant benefits. For the Backbone of HAB, the number of the MHA is set to 8 and the number of encoder blocks is set to 11. Finally, we adopt three standard assessment methods: Pearsons linear correlation coefficient (PLCC), Spearman's rank order correlation coefficient (SROCC) and root mean squared error (RMSE) to assess the model performances. The former two respectively evaluate the prediction results based on rank correlation and linear correlation. A value closer to 1 indicates a better prediction result. The latter measures the discrepancy between the predicted and ground-truth values, with a value closer to 0 indicating a better prediction result. We also used a five-parameter logistic function to fit the predicted quality scores and the ground-truth labels:

$$y = \beta_1 (\frac{1}{2} - \frac{1}{1 + \exp{(\beta_2 (x - \beta_3))}}) + \beta_4 x + \beta_5, \quad (7)$$

where x refers to the predicted quality score and y represents the mapped score. β_1 to β_5 are five parameters.

B. Performance Evaluation

1) Comparison Metrics: In order to illustrate the effectiveness of our model, the comparison algorithms includes FR and NR OIQA metrics. The FR-OIQA metrics include PSNR, S-PSNR [2], WS-PSNR [3], CPP-PSNR [4], SSIM [1], MS-SSIM [5], FSIM [6] and DeepIQA [28]. The NR-OIQA contain BRISQUE [29], BMPRI [30], DB-CNN [31], MC360IQA [13], DDA-BOIQA [32] and VGCN [14].

The performance comparison results on the OIQA and CVIQD datasets are shown in Table I. Among these FR-OIQA methods, the PSNR-related algorithms which have weaker correlation with the HVS exhibit poorer performance compared to these state-of-the-art objective algorithms. It is a breakthrough that the SSIM takes into account the brightness, contrast, and structural features associated with the HVS. However, the evaluation results are still limited and the performances are inferior to deep learning-based FR-OIQA methods. The reason lies in that deep learning-based methods directly consider the internal relationship between images and subject scores, while other methods mainly focus on one or two features of the OI. For NR methods, these algorithms generally outperform FR-OIQA algorithms. BRISQUE, BMPRI, and DB-CNN are implemented for OIQA specifically targeting ERP format of OIs. Specifically, BRISQUE and BMPRI are implemented based on handcrafted feature designs, while DB-CNN is implemented based on a data-driven model. Furthermore, the MC360IQA and VGCN models consider the VP images into their CNN

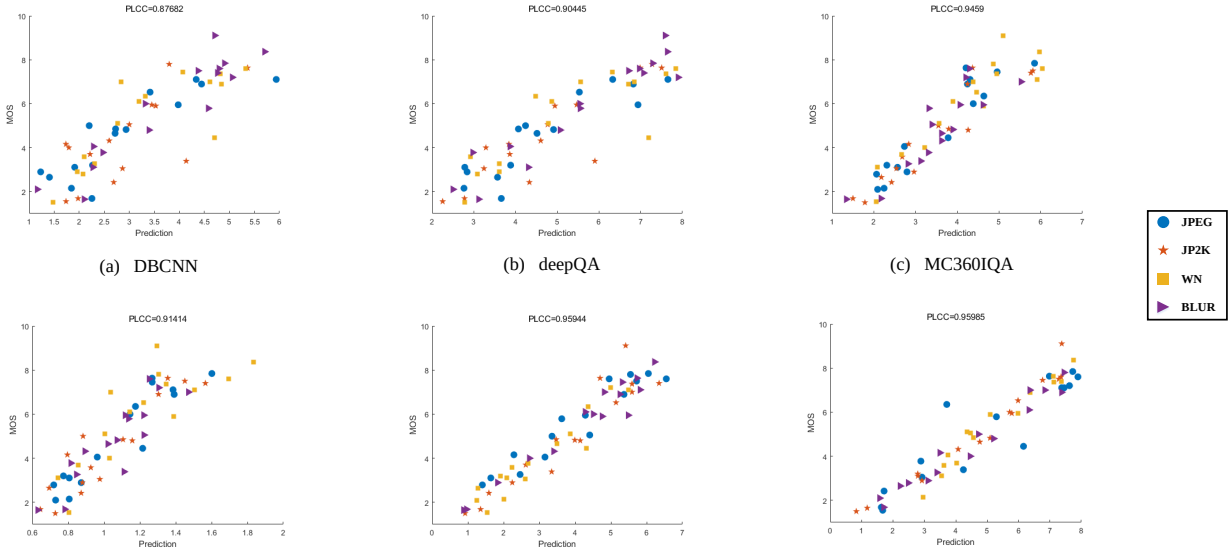
TABLE II. PERFORMANCE COMPARISON ON OIQA DATABASE. THE BEST RESULT IS ANNOTATED WITH BOLD, AND THE SECOND-BEST RESULT IS ANNOTATED WITH UNDERLINE.

		JPEG			JP2K			WN			BLUR		
		PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
FR	PSNR	0.6941	0.7060	1.6141	0.8632	0.7821	1.1316	0.9547	0.9500	0.5370	0.9282	0.7417	0.8299
	S-PSNR	0.6911	0.6148	1.6205	0.9205	0.7250	0.8757	0.9503	0.9357	0.5620	0.8282	0.7525	1.0910
	WS-PSNR	0.7133	0.6792	1.5713	0.9344	0.7500	0.9128	0.9626	0.9500	0.4890	0.8190	0.7668	1.1172
	CPP-PSNR	0.6153	0.5362	1.7693	0.8971	0.7250	0.9904	0.9276	0.9143	0.6739	0.7969	0.7185	1.1728
	SSIM	0.9077	0.9008	0.9406	0.9783	0.9679	0.4643	0.8828	0.8607	0.8474	**0.9926**	0.9777	0.2358
	MS-SSIM	0.9102	0.8937	0.9288	0.9492	0.9250	0.7052	0.9691	0.9571	0.4452	0.9251	0.8990	0.7374
	FSIM	0.8938	0.8490	1.0057	0.9699	0.9643	0.5454	0.9170	0.8893	0.7197	0.9914	**0.9902**	**0.2544**
	DeepQA	0.8301	0.8150	1.2506	**0.9905**	**0.9893**	**0.3082**	0.9709	**0.9857**	0.4317	0.9623	0.9473	0.5283
NR	BRISQUE	0.9160	0.9392	0.8992	0.7397	0.6750	1.5082	0.9818	0.9750	0.3427	0.8663	0.8508	0.9697
	BMPRI	0.9361	0.8954	0.7886	0.8322	0.8214	1.2428	0.9673	0.9821	0.4572	0.5199	0.3807	1.6584
	DB-CNN	0.8413	0.7346	1.2118	0.9755	0.9607	0.4935	0.9772	0.9786	0.3832	0.9536	0.8865	0.5875
	MC360IQA	0.9459	0.9008	0.7272	0.9165	0.9036	0.8966	0.9718	0.9464	0.4251	0.9526	0.9580	0.5907
	VGCN	**0.9540**	**0.9294**	**0.6720**	0.9771	0.9464	0.4772	0.9811	0.9750	0.3493	0.9852	0.9651	0.3327
	Ours	0.9475	0.9133	0.7167	0.9885	0.9821	0.3390	**0.9888**	0.9714	**0.2690**	0.9859	0.9777	0.3251

model, resulting in significant performance improvements. It is because the VP images are similar to the perception of human eyes. Our algorithm exhibits significantly higher performance compared to most deep learning based algorithms in terms of accuracy and monotonicity on those two databases. It is evident that the proposed dual-level network based on human visual perception is more consistent with the subject quality perception.

2) Performance Validity of Individual Distortion Types: As illustrated in Table II and Table III, we also conducted comparative experiments of individual distortion types on OIQA and CVIQD. In general, our algorithm exhibits the best comprehensive performance for most of the distortions. The scatter plots in Fig. 3 and Fig. 4 depict the correlation between MOS and the predictions for individual distortion types on the two databases. These plots provide additional evidence to support the superiority of our approach. Specifically, our algorithm achieves top performance in both WN and AVC distortion types, and it closely follows the top-performing algorithm in JPEG and JP2K distortions. For example, as shown in Fig. 4, our algorithm exhibits an SROCC value in JPEG that is only 0.0161 lower than the top-performing VGCN, and is only 0.0062 lower than DeepQA on the OIQA. This further demonstrates the strong robustness of our algorithm in compression distortion types. Additionally, in terms of HEVC distortion, our algorithm achieves the best PLCC and RMSE values in NR-OIQA. It is noteworthy that SSIM and FSIM in FR-OIQA actually achieved the best results in this distortion type. The reason is that HEVC distortion typically includes color inaccuracies or artifacts, while FSIM primarily measures quality degradation by assessing the similarity in luminance, contrast, and structural aspects between the reference and distorted images. Therefore, FSIM is more sensitive to color changes. Additionally, FSIM benefits from having a reference image for comparison, which enhances its ability to identify artifacts such as pseudo-imaging. This also indicates the effectiveness of these schemes for a certain type of distortion.

3) Ablation Study: In order to further demonstrate the effectiveness of each module, we separately removed each component of our model to conduct ablation experiments on two datasets. The experimental results are presented in Table IV. We separately adopt the human attention Branch (HAB) and the multi-scale perception branch (MPB) to predict the perceptual quality based on human visual perception in the dual-level network. In this section, we compare the performance with or without these two branches to respectively demonstrate the validity of each branch. We can conclude that both branches have strong quality prediction capabilities. However, the dual-stream network proposed in this paper, which combines features from these two branches, exhibits superior quality perception abilities, particularly in improving the SROCC values. Moreover, the influence of HAB is more pronounced on OIQA, mainly due to the diverse resolutions of OIs in this dataset, which are effectively addressed by the image pyramid utilized in the HAB. In our implementation, attentional features are obtained from an image pyramid in HAB. It is necessary to investigate how feature extraction based on attention mechanism affects overall performance. Therefore, we replace the original backbone in HAB with ResNet-50 and test the performance of the overall architecture. The results show that the performance of the CNN-based backbone is inferior to the backbone used in this paper, which further demonstrate the necessity of considering attentional features in HAB. In addition, we also conducted an ablation study for the correlation feature fusion module of VPs in MPB. As compared to the original implementation, the results on both databases showed slight improvements, which further proves that there is contextual and positional correlation information between different viewports within a panoramic image.



(a) DBCNN (b) deepQA (c) MC360IQA

(d) VGG16 (e) VGCN (f) Ours

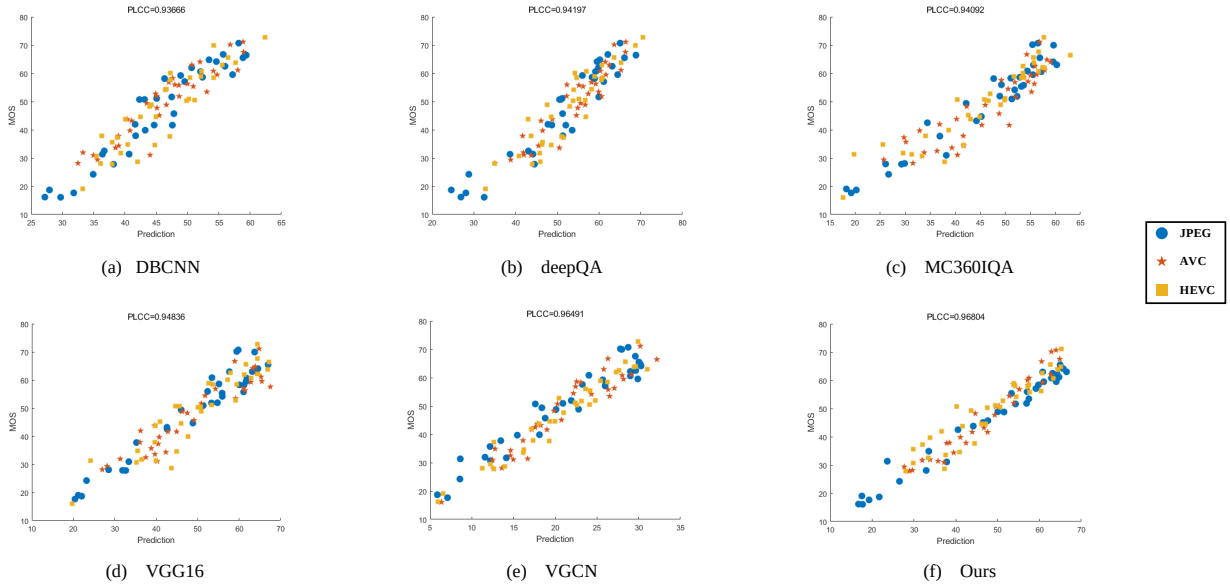
Fig. 3. Scatter plots of MOS values against predictions by OIQA metrics for individual distortion type on the testing set of OIQA Database.

TABLE III. PERFORMANCE COMPARISON ON CVIQD DATABASE. THE BEST RESULT IS ANNOTATED WITH BOLD, AND THE SECOND-BEST RESULT IS ANNOTATED WITH UNDERLINE.

		JPEG			AVC			HEVC		
		PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
FR	PSNR	0.8682	0.6982	8.0429	0.6141	0.5802	10.5520	0.5982	0.5762	9.4697
	S-PSNR	0.8661	0.7172	8.1008	0.6307	0.6039	10.3760	0.6514	0.6150	8.9585
	WS-PSNR	0.8572	0.6848	8.3465	0.5702	0.5521	10.9841	0.5884	0.5642	9.5473
	CPP-PSNR	0.8585	0.7059	8.3109	0.6137	0.5872	10.5615	0.6160	0.5689	9.3009
	SSIM	0.9822	0.9582	3.0468	0.9303	0.9174	4.9029	0.9436	0.9452	3.9097
	MS-SSIM	0.9636	0.9047	4.3355	0.7960	0.7650	8.0924	0.8072	0.8011	6.9693
	FSIM	0.9839	0.9639	2.8928	0.9534	0.9439	4.0327	**0.9617**	**0.9532**	**3.2385**
	DeepQA	0.9526	0.9001	4.9290	0.9477	0.9375	4.2683	0.9221	0.9288	4.5694
NR	BRIAQUE	0.9464	0.9031	5.2442	0.7745	0.7714	8.4573	0.7548	0.7644	7.7455
	BMPRI	0.9874	0.9562	2.5597	0.7161	0.6731	9.3318	0.6154	0.6715	9.3071
	DB-CNN	0.9779	0.9576	3.3862	0.9564	0.9545	3.9063	0.8646	0.8693	5.9335
	MC360IQA	0.9698	0.9693	3.9517	0.9487	0.9569	4.2281	0.8976	0.9104	5.2557
	DDA-BOIQA	0.9570	0.9610	5.6010	0.9530	0.9490	3.8730	0.9290	0.9140	4.5250
	VGCN	**0.9894**	0.9759	**2.3590**	0.9719	0.9659	3.1490	0.9401	0.9432	4.0257
	Ours	0.9878	**0.9803**	2.5285	**0.9780**	**0.9796**	**2.7888**	0.9408	0.9405	4.0023

TABLE IV. ABLATION STUDY RESULTS FOR REMOVING EACH INDIVIDUAL BRANCH OR MODULE ON OIQA AND CVIQD.

	OIQA			CVIQD		
Methods	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
w/o HAB	0.9474	0.9414	0.6686	0.9623	0.9658	3.8786
w/o MPB	0.9572	0.9476	0.6049	0.9694	0.9627	3.4993
w/o attentional features in HAB	0.9482	0.9449	0.6639	0.9655	0.9650	3.7124
w/o correlation features in MPB	0.9569	0.9497	0.6072	0.9634	0.9632	3.8238
Ours	**0.9598**	**0.9530**	**0.5862**	**0.9680**	**0.9664**	**3.5014**



(a) DBCNN (b) deepQA (c) MC360IQA

(d) VGG16 (e) VGCN (f) Ours

Fig. 4. Scatter plots of MOS values against predictions by OIQA metrics for individual distortion type on the testing set of CVIQD Database.

V. Conclusion

In response to the fact that the perception of immersive media quality is more susceptible to subjective visual perception by the human eye, in this paper, we propose an innovative approach that integrates two characteristics of human visual perception, namely attentional perception and multi-scale perception, into the process of acquiring panoramic image features. Specifically, we propose a dual-level network based on human visual perception for blind omnidirectional image quality assessment. By transforming the front viewport image to an image pyramid with multiple viewing distances, the human attention branch is able to capture the high-level information based on attention mechanism. To obtain the features of different viewports from different position, we further establish a module to fuse their correlation information in the multi-scale feature perception branch after parallel extraction of their multi-scale features.

Experimental on two OIQA datasets show that our approach achieves the best performance, further validating the effectiveness of the human visual perception guidance. Of course, our work needs further in-depth research. Our approach only incorporates two essential aspects of human visual perception to assist the omnidirectional image quality assessment process. However, human visual perception is diverse, and the challenge lies in quantifying it effectively in a general end-to-end model. This will be the focus of our future research endeavors.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62171002, 61906118, in part by STCSM under Grant SKLSFO2021-05, in part by University Discipline Top Talent Program of Anhui under Grant gxbjZD2022034, in part by Project on Anhui Provincial Natural Science Study by Colleges and Universities under Grant 2022AH030106, in part by Key research projects in humanities and social sciences under Grant SK2019A0373, in part by Anhui educational science research project under Grant JK22007, and in part by China Postdoctoral Science Foundation under Grant 2022M710745.

References

[1] Z. Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[2] M. Yu, H. Lakshman, and B. Girod. “A framework to evaluate omnidirectional video coding schemes”. In: *2015 IEEE international symposium on mixed and augmented reality*. IEEE. 2015, pp. 31–36.

[3] Y. Sun, A. Lu, and L. Yu. “Weighted-to-spherically-uniform quality evaluation for omnidirectional video”. In: *IEEE signal processing letters* 24.9 (2017), pp. 1408–1412.

[4] V. Zakharchenko, K. P. Choi, and J. H. Park. “Quality metric for spherical panoramic video”. In: *Optics and Photonics for Information Processing X*. Vol. 9970. SPIE. 2016, pp. 57–65.

[5] Z. Wang, E. P. Simoncelli, and A. C. Bovik. “Multiscale structural similarity for image quality assessment”. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[6] L. Zhang et al. “FSIM: A feature similarity index for image quality assessment”. In: *IEEE transactions on Image Processing* 20.8 (2011), pp. 2378–2386.

[7] S. Ling, G. Cheung, and P. Le Callet. “No-reference quality assessment for stitched panoramic images using convolutional sparse coding and compound feature selection”. In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2018, pp. 1–6.

[8] H. Jiang et al. "Multi-Angle Projection Based Blind Omnidirectional Image Quality Assessment". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.7 (2022), pp. 4211–4223. DOI: 10.1109/TCSVT.2021.3128014.

[9] W. Zhou et al. "No-Reference Quality Assessment for 360-Degree Images by Analysis of Multifrequency Information and Local-Global Naturalness". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.4 (2022), pp. 1778–1791. DOI: 10.1109/TCSVT.2021.3081182.

[10] Y. Liu et al. "HVS-Based Perception-Driven No-Reference Omnidirectional Image Quality Assessment". In: *IEEE Transactions on Instrumentation and Measurement* 72 (2023), pp. 1–11. DOI: 10.1109/TIM.2022.3232792.

[11] S. Habib et al. "External Features-Based Approach to Date Grading and Analysis with Image Processing". In: *Emerg. Sci. J* 6.4 (2022), pp. 694–704.

[12] C. Li et al. "Viewport Proposal CNN for 360deo Quality Assessment". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10169–10178. DOI: 10.1109/CVPR.2019.01042.

[13] W. Sun et al. "MC360IQA: The Multi-Channel CNN for Blind 360-Degree Image Quality Assessment". In: *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2019, pp. 1–5. DOI: 10.1109/ISCAS.2019.8702664.

[14] J. Xu, W. Zhou, and Z. Chen. "Blind Omnidirectional Image Quality Assessment With Viewport Oriented Graph Convolutional Networks". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.5 (2021), pp. 1724–1737. DOI: 10.1109/TCSVT.2020.3015186.

[15] A. Sendjasni and M.-C. Larabi. "SAL-360IQA: A Saliency Weighted Patch-Based CNN Model for 360-Degree Images Quality Assessment". In: *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE. 2022, pp. 1–6.

[16] Y. Lu et al. "Blind image quality assessment based on the multiscale and dual-domains features fusion". In: *Concurrency and Computation: Practice and Experience* (2021), e6177.

[17] A. Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[18] W. Ding et al. "No-reference Panoramic Image Quality Assessment based on Adjacent Pixels Correlation". In: *2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. 2021, pp. 1–5. DOI: 10.1109/BMSB53066.2021.9547132.

[19] X. Zheng et al. "Segmented Spherical Projection-Based Blind Omnidirectional Image Quality Assessment". In: *IEEE Access* 8 (2020), pp. 31647–31659. DOI: 10.1109/ACCESS.2020.2972158.

[20] H.-T. Lim, H. G. Kim, and Y. M. Ra. "VR IQA NET: Deep Virtual Reality Image Quality Assessment Using Adversarial Learning". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 6737–6741. DOI: 10.1109/ICASSP.2018.8461317.

[21] H. G. Kim, H.-T. Lim, and Y. M. Ro. "Deep Virtual Reality Image Quality Assessment With Human Perception Guider for Omnidirectional Image". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.4 (2020), pp. 917–928. DOI: 10.1109/TCSVT.2019.2898732.

[22] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[23] E. Z. Ye et al. "DeepImageTranslator V2: analysis of multimodal medical images using semantic segmentation maps generated through deep learning". In: *bioRxiv* (2021), pp. 2021–10.

[24] M. Jesmeen et al. "SleepCon: Sleeping Posture Recognition Model using Convolutional Neural Network". In: *Emerging Science Journal* 7.1 (2022), pp. 50–59.

[25] W. Sun 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)*. IEEE. 2018, pp. 1–6.

[26] H. Duan et al. "Perceptual Quality Assessment of Omnidirectional Images". In: *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2018, pp. 1–5. DOI: 10.1109/ISCAS.2018.8351786.

[27] A. Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013).

[28] J. Kim and S. Lee. "Deep learning of human visual sensitivity in image quality assessment framework". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1676–1684.

[29] A. Mittal, A. K. Moorthy, and A. C. Bovik. "No-reference image quality assessment in the spatial domain". In: *IEEE Transactions on image processing* 21.12 (2012), pp. 4695–4708.

[30] X. Min et al. "Blind image quality estimation via distortion aggravation". In: *IEEE Transactions on Broadcasting* 64.2 (2018), pp. 508–517.

[31] W. Zhang et al. "Blind image quality assessment using a deep bilinear convolutional neural network". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.1 (2018), pp. 36–47.

[32] Y. Zhou et al. "Omnidirectional Image Quality Assessment by Distortion Discrimination Assisted Multi-Stream Network". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.4 (2022), pp. 1767–1777. DOI: 10.1109/TCSVT.2021.3081162.