Viewport Perception based Blind Stereoscopic Omnidirectional Image Quality Assessment

Yubin Qi, Gangyi Jiang, *Senior Member, IEEE*, Mei Yu, Yun Zhang, *Senior Member, IEEE* and Yo-Sung Ho, *Fellow, IEEE*

***Abstract*—Compared with traditional 2D images, stereoscopic omnidirectional images (SOIs) usually have more complex perceptual factors due to the particularities of imaging and display, making the objective quality assessment of SOIs challenging. In this paper, we construct a large and diverse subjective SOIs database named as NBU-SOID for further research demand. And then, we propose a viewport perception based blind SOIs quality assessment (VP-BSOIQA) method by considering the impacts of viewport, user behavior and stereoscopic perception on human visual system, which is mainly composed of binocular perception model (BPM) and omnidirectional perception model (OPM). In the BPM, a binocular combination perception map is generated by the dimension reduction of stereopair and the weighting of binocular energy to reflect the binocular masking effect. In the OPM, several viewports are first created to ensure the consistency of evaluation objects. Then, the intra-viewport and inter-viewport weighting factors are designed with the common influences of visual attention and peripheral vision sensitivity to aggregate the novel multi-orientation structural features extracted from all potential viewports. Experimental results on the NBU-SOID and SOLID databases demonstrate that BPM and OPM can be robustly combined with the existing 2D image quality assessment (IQA) methods, thus averagely achieving 10.2% and 12.2% performance gain in terms of SRCC, respectively. In addition, the proposed VP-BSOIQA method outperforms the state-of-the-art blind IQA methods in predicting the quality of SOIs.**

***Index Terms*—Blind image quality assessment, binocular vision, stereoscopic omnidirectional images, viewport perception.**

I. Introduction

Virtual reality (VR), as efficient interactive technology, is capable of providing users with an immersive environment, and has been successfully applied in many fields, such as medical treatment, game, tourism, education and sports events [1]. Stereoscopic omnidirectional images (SOIs), as the most popular visual contents in VR applications, are composed of a pair of omnidirectional images simulating the human eyes [2]. Since the perceptual quality of SOIs is closely related to user visual experience, it is highly demanded to explore SOI quality assessment (SOIQA) methods, which can be applicable to optimize the SOI signal processing algorithms, such as SOI generation, coding, streaming and rendering [3]. Unfortunately, although many studies of image quality assessment (IQA) had been conducted over the last decade, there is a large gap in predicting the perceptual quality of SOIs.

Generally, IQA methods include the subjective and the objective ones. The first directly utilizes human opinion score as the final evaluation, but it is usually inconvenient and laborious in practical applications [4]. Hence, developing objective IQA methods highly consistent with the subjective evaluation is indispensable in image processing field. The existing objective IQA methods can be categorized into the full-reference (FR), reduced-reference (RR) and blind/ no-reference (NR), according to whether reference images are available or not [5]. The FR methods need full information of the reference images and the RR methods only use the partial information, while the NR methods predict the quality of distorted image without any reference images [6]. Some classical FR-IQA methods, such as peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [7], visual information fidelity (VIF) [8], riesz-transform based feature similarity (RFSIM) [9] and gradient magnitude similarity deviation (GMSD) [10] were proposed. However, they were designed and performed well only for the ordinary 2D images.

Nevertheless, in many practical applications, blind IQA (BIQA) methods are demanded due to lack of the reference information. For the most existing BIQA models based on supervised learning, a common two-stage framework is used to predict the quality of the distorted image [11]. Concretely, several perceptual features (*e.g.*, the natural scene statistics (NSS) in spatial [12] or gradient domains [13]) are extracted from test image to identify the distortion types, then a quality regression function is learned to map the feature space to quality scores by machine learning algorithms. These BIQA methods largely depend on the accuracy and quantity of human subjective data in training, so they are also regarded as opinion-aware BIQA (OA-BIQA) methods. Recently, given that it is cumbersome and expensive to obtain the collections of

This work was supported by the Natural Science Foundation of China under Grant Nos. 61671247, 61871258, 62071266 and 61931022, Natural Science Foundation of Zhejiang Province Grant No. Y21F010010, and it was also sponsored by the K. C. Wong Magna Fund of Ningbo University. *(Corresponding author: Gangyi Jiang)*

Y. Qi, G. Jiang and M. Yu are with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: qiyubin123@126.com, jianggangyi@nbu.edu.cn, yumei2@126.com).

Y. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: yun.zhang@siat.ac.cn).

Y.-S. Ho is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: hoyo@gist.ac.kr).

distorted images with their corresponding subjective opinion scores, there have appeared some popular opinion-unaware BIQA (OU-BIQA) methods that do not require to be trained on a subjective assessment database, such as NIQE [14], dipIQ [15], BPRI [16], HOSA [17] and UCA [18]. These BIQA methods achieved outstanding performance on several benchmark 2D-IQA databases.

However, compared with traditional 2D images, SOIs have their unique perceptual characteristics in some aspects, such as the storage form, field of view (FoV) and rendering device. Firstly, considering that SOIs are captured in spherical format and cannot be directly stored and transmitted, the projection from the spherical format to 2D plane (*e.g.*, equi-rectangular projection (ERP) format [19]) is required while compressing the SOIs with the existing image/video coding standard. Secondly, the whole SOI usually cover the range of 360°×180° FoV, while users can only observe a small part of the SOI in the form of viewport at a short period. Thirdly, instead of being viewed on 2D screen, SOIs are viewed with the head mounted display (HMD), making users have a sense of immersion and depth. Therefore, the traditional 2D-IQA methods are not accurate for SOIs by simply evaluating them with ERP format. Meanwhile, SOIQA is more complicate than a simple combination of stereoscopic IQA (SIQA) and omnidirectional IQA (OIQA).

Based on the above analyses, a viewport perception based blind SOIQA (VP-BSOIQA) method is proposed which considers comprehensively the impacts of viewport, user behavior and stereoscopic perception in the immersive environments. The main contributions are:

1) A large and diverse subjective SOIQA database named as NBU-SOID is built with the purpose of verifying the effectiveness of the proposed VP-BSOIQA method, and it will be made available online for the further research demand.

2) Binocular perception model (BPM) and omnidirectional perception model (OPM) are proposed to simulate the binocular and omnidirectional visual characteristics of human visual system (HVS) in the immersive environments, respectively.

3) This work presents a way to develop some novel metrics specialized for SOIs by combining the existing 2D-IQA metrics with the proposed two perception models. As an example, a new multi-orientation structural feature extraction approach is given to form the proposed VP-BSOIQA method.

The rest of the paper is organized as follows: The related works and motivations are described in Section II. The subjective evaluation experiment is presented in Section III. The proposed method and experimental results are characterized in Section IV and Section V, respectively. Finally, the conclusion and future work are demonstrated in Section VI.

II. Related Works and Motivations

In this section, the related works are first reviewed, including the existing SIQA, OIQA and SOIQA methods, and then the motivations about the proposed VP-BSOIQA method is elaborated.

A. Related Works

1) *SIQA Methods*: Different from 2D-IQA metrics, SIQA metrics additionally take into account the binocular visual characteristics. When the similar monocular stimuli fall in the corresponding retinal areas of the left and right eyes, binocular fusion occurs and enters into a single stable binocular perception. When the quality of the visual contents presented to the left and right eyes are significantly different, perception is alternately inhibited between the left and right views, resulting in binocular rivalry eventually [20]. Initially, Gorley *et al*. [21] designed a SIQA metric by simply averaging the quality of the left and right views predicted by 2D-IQA metrics. Unfortunately, it cannot accurately evaluate the asymmetrically distorted stereoscopic images, because it has not fully considered the influence of binocular visual perception. Binocular vision is a process that fuses simple and complex cells. The monocular stimuli of the left and right eyes firstly pass through the interocular gain control pathway, and then combine with each other to form a single cyclopean perception [22]. Generally, the visual stimuli are defined by the brightness variations, where the left and right views are merged into a cyclopean image to characterize the binocular interaction. Chen *et al*. [23] introduced the cyclopean images considering parallax and binocular visual characteristics into SIQA. Jiang *et al*. [24] predicted the quality of stereoscopic image by learning the monocular perception features based on non-negative matrix factorization and chromatic visual features. Liu *et al*. [25] fused the left and right views into a novel synthesized cyclopean image and extracted the NSS features to identify distortion. Since the perceptual characteristics of SOIs are more complicated than those of the conventional stereoscopic images, the above SIQA metrics cannot predict the quality of SOIs accurately by direct application.

2) *OIQA and SOIQA Methods*: In recent years, a few quality metrics serving for coding applications were proposed to evaluate omnidirectional visual contents by solving non-uniform sampling, which are also named as sampling-related OIQA methods. Yu *et al*. [26] proposed spherical PSNR (S-PSNR), which mapped the pixels of the original and reconstructed ERP images onto the spherical surface and calculated PSNR with a number of sampling points on the spherical surface. To solve the problem of projection format mismatch between the original and reconstructed signals, craster parabolic projection PSNR (CPP-PSNR) was presented to calculate PSNR in CPP plane [27]. Different from S-PSNR and CPP-PSNR, weighted-to-spherically-uniform PSNR (WS-PSNR) [28] and weighted-to-spherically-uniform SSIM (WS-SSIM) [29] were proposed based on an area ratio factor between the ERP format and the sphere. Although the above-mentioned sampling-related OIQA metrics may overcome the shortcomings of applying the representative 2D-IQA metrics to different projections, their consistencies with perception of human eyes are still low due to the lack of analyzing the visual perception [30].

Therefore, several perception-related OIQA or SOIQA metrics have been proposed. Upenik *et al*. [31] incorporated visual attention data into traditional PSNR metric. Kim *et al*.

[32] proposed a deep learning based OIQA method consisting of the quality predictor and human perception guider. Lim *et al*. [33] predicted the quality score of distorted image using the latent spatial and position features, and then optimized the corresponding score via adversarial learning with human opinion. On the basis of asymmetric mechanism in human brain, Xia *et al*. [34] used the same weight as in [28] to conduct the local binary pattern and statistical analysis, so that the related high frequency and low frequency features are obtained. Yang *et al*. [35] incorporated the luminance masking effect of human vision and the latitude characteristics of ERP images to evaluate the quality of SOIs. To ensure the objective assessment object is closer to the image seen by users with HMD, Sun *et al*. [36] converted the ERP image into six viewpoint images, denoted as cubemap projection (CMP), and designed a multi-channel deep learning model and a quality regressor to predict the quality. Similarly, Zheng *et al*. [37] conducted the transformation process from ERP to segmented spherical projection (SSP) and took into account the perceptual influence caused by visual saliency. Croci *et al*. [38] divided the omnidirectional content into multiple patches using the spherical Voronoi diagram. However, these perception-related OIQA metrics are all conducted on the ERP, CMP or SSP plane, not in line with the actual viewing object, namely viewport.

Since viewport image plays an important role in OIQA and SOIQA metrics, it has been considered in many metrics, but the perceptual impacts caused by user behavior are usually ignored. For instance, Chen *et al*. [39] and Xu *et al*. [40] proposed the multi-view fusion module for integrating all viewport scores according to the content and location of viewports, which is easy but not precise enough. Azevedo *et al*. [41] proposed a viewport-based multi-metric fusion metric for visual quality assessment of omnidirectional videos, but the features extracted from all viewports are not aggregated by any pooling means, which may result in overfitting of quality prediction model. Xu *et al*. [42] and Li *et al*. [43] presented the OIQA metrics based on viewport oriented graph convolutional network and viewport based convolutional neural network, respectively. Even though the above networks consider both auxiliary tasks of viewport proposal and viewport saliency prediction, the accuracy of saliency prediction is usually related to training data, but unfortunately, there is a lack of relevant saliency database of omnidirectional image or SOIs.

B. Motivations

Fig. 1 depicts the typical end-to-end SOIs processing pipeline with three main steps [44]. Firstly, SOIs are formed by stitching several images captured by the circular camera array to cover the whole FoV of 360°×180°. Secondly, due to the incompatibility in storage and transmission, the SOIs are projected from spherical surface to ERP format and then encoded by the existing 2D image/video encoder. Finally, when the coded ERP images are transmitted to the client for viewing, an inverse projection transformation from ERP format to spherical surface and a viewport rendering are required to restore the actual scene. These unique adaptation processes to the immersive environments make the perceptual

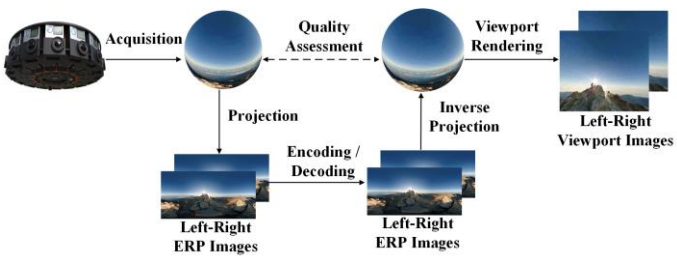
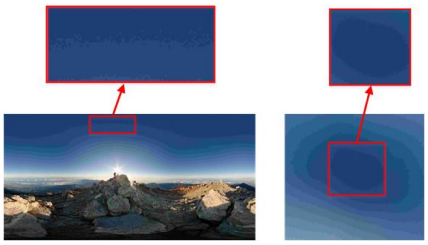


Fig. 1. End-to-end stereoscopic omnidirectional image processing pipeline.



(a) (b)

Fig. 2. Distortion comparison between the ERP and viewport images. (a) ERP image. (b) Viewport image.

Table I
THE PERCEPTUAL FACTORS CONTAINED IN DIFFERENT KINDS OF IQA METRICS.

Types	Perceptual Factors		
	Viewport	User Behavior	Stereoscopic Perception
2D-IQA			
SIQA			✓
OIQA	✓	✓	
SOIQA	✓	✓	✓

characteristics of SOIs are more complex than those of the ordinary 2D/3D images in the following three aspects, *i.e*., viewport, user behavior and stereoscopic perception.

1) *Viewport*: During the viewport rendering process, the compression artifacts in the ERP images will be geometrically changed, especially in the pole regions. Fig. 2 illustrates an example of the distortion comparison between ERP and viewport images. From the locally enlarged images, it can be observed that the original streak artifacts become circular when users view the image in the form of viewport.

2) *User Behavior*: Users can freely change the position of viewport by moving their heads to select their interested viewports while viewing. In addition, for each viewport, some local areas may be drawn more attention by eye movements, especially the ones with salient objects.

3) *Stereoscopic Perception*: When the quality of left and right views of SOIs are similar, binocular fusion will occur. However, if the presented contents are significantly different, binocular rivalry will occur, which may cause severe visual discomfort because HMD is closer to eyes and has larger FoV than the conventional 3D display.

Some key perceptual factors involved in different kinds of IQA metrics are described in Table I. Unfortunately, the existing IQA metrics (including 2D-IQA, SIQA and OIQA metrics) have not comprehensively considered the perceptual characteristics of SOIs, especially the user behavior in the immersive environment. Therefore, an SOIQA that jointly considers these key factors is highly demanded. Furthermore,

to evaluate an objective quality assessment for SOIs, a large-scale public subjective SOIs database is mandatory. However, most existing databases are based on monocular omnidirectional vision, which does not fully suit the performance evaluation of SOIQA metrics. Consequently, in this study, we construct a subjective SOIQA database named as NBU-SOID and propose the VP-BSOIQA method by simultaneously taking into account the above mentioned three perceptual factors.

III. NBU-SOID: Database Construction

The establishment of SOIQA metrics is usually based on a subjective assessment database for analyzing the objective performance, but the sole SOIQA database (*i.e.*, SOLID [45]) at present only contains 6 reference images and 138 distorted images with one distortion type and three depth levels, which is small-scale and inadequate for practical research. Therefore, we extend the distortion types and diversity of the visual contents in our subjective SOIQA database (NBU-SOID [46]), which can be used to verify the effectiveness and robustness of objective IQA algorithms.

A. Source Images Acquisition

Twelve high-quality SOIs are used as the reference sources in NBU-SOID, eight of which are the still frames extracted from stereoscopic omnidirectional videos provided by MPEG [47] and VRQ-TJU [48], and the rest are from Yahoo. It is noteworthy that all the reference images with the resolution of 2560×2560 (or 4096×4096) are represented in the ERP format, where the left and right views of SOIs are packed in the top-bottom pattern. Fig. 3 depicts snapshots of the twelve reference SOIs in NBU-SOID, including both indoor and outdoor scenes.

B. Distortion Simulation

The twelve reference SOIs in NBU-SOID are processed with JPEG, JPEG2000 or HEVC intra codec, so as to their distorted versions with different degrees of compression artifacts. In particular, distortion intensity is achieved by controlling quality (q), bits per pixel (bpp) and quantization parameter (qp) for JPEG, JPEG2000 and HEVC, respectively. Afterwards, an expert screening is conducted to select the control parameters representing five scales of visual quality. Both symmetrical and asymmetrical distortions, *i.e.*, the left and right views of SOIs are applied with the even or uneven distortion intensities, are also considered in NBU-SOID. Table II lists the selected parameters representing different distortion levels, where symmetrical distortion levels are labeled as 1-5, and asymmetrical distortion levels are labeled as 6-11, respectively. In total, there are 180 symmetrically distorted SOIs and 216 asymmetrically distorted SOIs generated in NBU-SOID.

C. Subjective Evaluation Procedure

Subjective evaluation experiments have been conducted in Ningbo University. Thirty subjects aged from 20 to 28 years old including professionals and non-professionals have been invited to participate in the experiment. The whole subjective

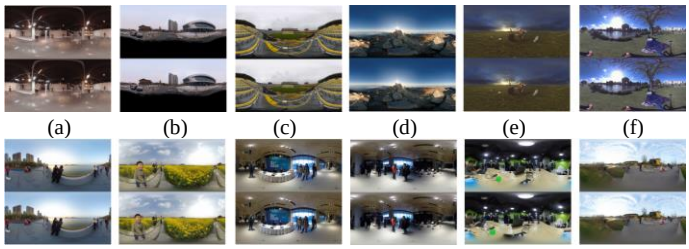


Fig. 3. Original reference images in NBU-SOID. (a) Dancing. (b) Drive. (c) Stadium. (d) Tour. (e) Cartoon. (f) Hiking. (g) Riverside. (h) Field. (i) Sign in. (j) Chat. (k) Wait. (l) Tourist.

TABLE II
THE PARAMETER SETTING OF DISTORTION LEVELS IN NBU-SOID. THE ABBREVIATIONS 'SYM' AND 'ASYM' DENOTE SYMMETRICAL AND ASYMMETRICAL DISTORTION, RESPECTIVELY.

Types	Level	Control Parameters					
		JPEG (q)		JPEG2000 (bpp)		HEVC (qp)	
		Left	Right	Left	Right	Left	Right
Sym	1	10	10	0.05	0.05	22	22
	2	30	30	0.1	0.1	27	27
	3	50	50	0.3	0.3	32	32
	4	70	70	0.6	0.6	37	37
	5	90	90	1.0	1.0	42	42
Asym	6	10	50	0.05	0.3	22	27
	7	10	90	0.05	1.0	22	32
	8	30	50	0.1	0.3	22	42
	9	30	70	0.1	0.6	27	37
	10	50	90	0.3	1.0	32	37
	11	70	90	0.6	1.0	32	42

experiment had taken about 20 days to ensure that the experimental data are fully valid and reliable. The evaluation procedure mainly consists of three stages. Firstly, before the subjective evaluation, we conducted verbal instructions to subjects to let them know the experimental procedures. Then, since most subjects are not professional in SOIQA, a set of additional reference SOIs and the corresponding distorted images were created and used for training subjects. Finally, we conducted the subjective test according to absolute category rating with hidden reference (ACR-HR) [49], where subjects were asked to sit on a rotatable chair and view SOIs with HMD (*i.e.*, HTC Vive Pro). The overall resolution of HTC Vive Pro is 2880×1600 pixels, which gives 1440×1600 pixels per eye, and the provided FoV in the horizontal direction is 110°. Moreover, voting was performed after each viewing by using five-grade quality scale with the following levels: "5-Excellent", "4-Good", "3-Fair", "2-Poor", and "1-Bad". To prevent subjects from fatigue, we controlled the duration of each test session for less than 30 minutes, and gave the subjects sufficient rest time between sessions.

D. Data Processing and Analysis

To implement the reliable assessment, the screening criteria described in [50] have been strictly followed, and five participants are removed as outliers. Then, the subjective scores are obtained by calculating the difference mean opinion score (DMOS) based on the approach in [51], where higher DMOS values represent better visual quality. Fig. 4 illustrates the

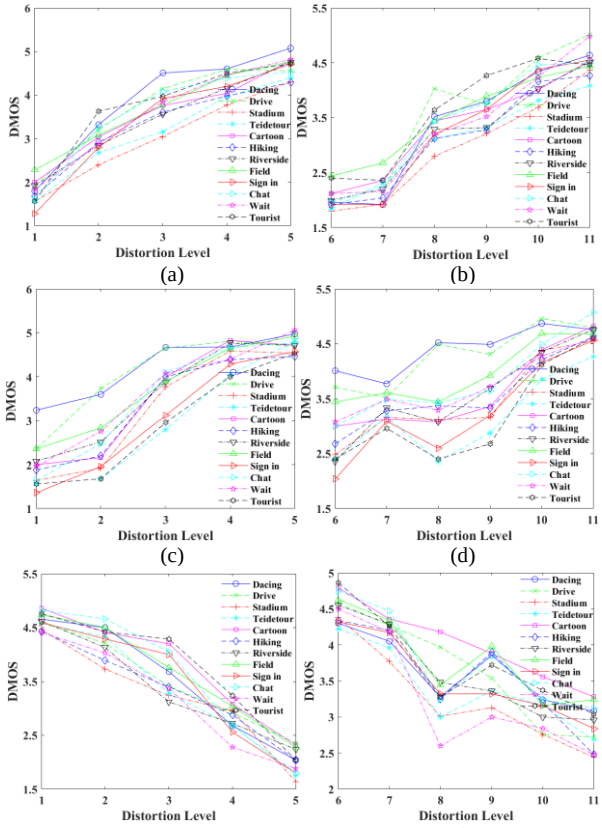


Fig. 4. The relationship between distortion levels and DMOS for the distorted SOIs. (a) JPEG symmetrical distortion. (b) JPEG asymmetrical distortion. (c) JPEG2000 symmetrical distortion. (d) JPEG2000 asymmetrical distortion. (e) HEVC symmetrical distortion. (f) HEVC asymmetrical distortion.

relationship between distortion levels and DMOS for the distorted SOIs, where the distortion levels are shown in Table II. Generally, for the symmetrically distorted SOIs, the relationship between subjective quality and distortion levels shows good monotonicity and the corresponding DMOS values are evenly distributed on the whole quality scales. These results indicate that the selected symmetrical distortion levels span a wide range of visual quality. However, there is no good monotonicity for the asymmetrically distorted SOIs, which are mainly caused by the binocular masking effect.

IV. Proposed VP-BSOIQA Method

Fig. 5 shows the framework of the proposed VP-BSOIQA method consisted of two perception models, namely BPM and OPM. The BPM aims at generating a binocular combination perception map for predicting binocular masking effect. The OPM performs the feature extraction process on the viewport images rather than ERP images, and aggregates the perception-related features of all viewport images with the joint influences of visual attention and peripheral vision sensitivity. Moreover, a new multi-orientation structural feature extraction approach is given to combine both BPM and OPM. The detail of the proposed VP-BSOIQA method is described in the following four subsections.

A. Binocular Perception Model (BPM)

Tensor decomposition of SOI is firstly used for simulating the binocular fusion and rivalry characteristics of HVS. Then, binocular energy weighting factors are designed for the perceptual characteristics combination. Finally, the left and right views of SOI and the obtained binocular combination perception map are integrated to form the binocular pattern ERP image stacks as the input of OPM.

1) *Tensor Decomposition of SOI*: Due to the multi-dimensional nature of real SOI data, traditional data representations cannot well reflect their complex structure and properties. To overcome this limitation, a distorted SOI is represented as the form of four-order tensor, *i.e.*, $\boldsymbol{I}_{\mathrm{d}} \in \Re^{h\times w\times 3\times 2}$, where h and w are the height and width of left view $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{L}}$ (or right view $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{R}}$), respectively. Furthermore, in order to preserve the influence of the interaction between brightness and color, Tucker decomposition [52] is used for the dimension reduction of SOI in this work. Specifically, for $\boldsymbol{I}_{\mathrm{d}}$, the core tensor $C \in \Re^{h\times w\times 2}$ containing the main information of original tensor can be obtained by simultaneously reducing the third and fourth dimensions which characterize the color and view information, respectively. The above process of dimension reduction is expressed as

$$C \approx \boldsymbol{I}_{\mathrm{d}} \times_3 (\boldsymbol{U}^{(3)})^{\mathrm{T}} \times_4 (\boldsymbol{U}^{(4)})^{\mathrm{T}} \tag{1}$$

where $\times_n (n=3,4)$ is the n-mode product operator of tensor and matrix, $\boldsymbol{U}^{(n)}(n=3,4)$ is the singular value matrix, and the core tensor can be expanded into the corresponding matrix sets, namely, $C=\{\boldsymbol{I}_1^{\mathrm{TS}}, \boldsymbol{I}_2^{\mathrm{TS}}\}$, expressed as the first and second tensor sub-bands, respectively.

In our previous work [53], it has been found that the first sub-band of core tensor represented the background information of video, *i.e.*, the similarities between video frames. Meanwhile, other sub-bands represented the motion information of video, *i.e.*, the differences between video frames. Similarly, $\boldsymbol{I}_1^{\mathrm{TS}}$ and $\boldsymbol{I}_2^{\mathrm{TS}}$ can be used to describe the binocular fusion and binocular rivalry characteristics of HVS, *i.e.*, the similarities and differences between the left and right views of distorted SOI, respectively.

Furthermore, although there is disparity between the left and right views of SOIs, real scene is ultimately imaged in the brain as a single stable stereoscopic image. Thus, there exists the mutual filtering effect between stereoscopic views [54], and the structure transfer property of guided image filter, $F_{\mathrm{G}}(\cdot)$, is utilized to eliminate the content differences of both views. Specifically, the filtering output $\boldsymbol{I}_{\mathrm{o}}$ of guided image filter is determined by filtering input $\boldsymbol{I}_{\mathrm{i}}$ and guided image $\boldsymbol{I}_{\mathrm{g}}$, and expressed as $\boldsymbol{I}_{\mathrm{o}} = F_{\mathrm{G}}(\boldsymbol{I}_{\mathrm{i}}, \boldsymbol{I}_{\mathrm{g}})$. When $\boldsymbol{I}_{\mathrm{g}}$ is inconsistent with $\boldsymbol{I}_{\mathrm{i}}$, $\boldsymbol{I}_{\mathrm{o}}$ will retain the strong structural information of $\boldsymbol{I}_{\mathrm{g}}$. Consequently, the left view $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{L}}$ and right view $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{R}}$ of SOI are used as the filtering input and guided image, respectively, so as to obtain the filtered left image, *i.e.*, $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{L,R}} = F_{\mathrm{G}}(\boldsymbol{I}_{\mathrm{d}}^{\mathrm{L}}, \boldsymbol{I}_{\mathrm{d}}^{\mathrm{R}})$. Similarly, when $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{L}}$ is used as guided image, the filtered right image $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{R,L}}$ can also be obtained. Then, a newly filtered four-order tensor

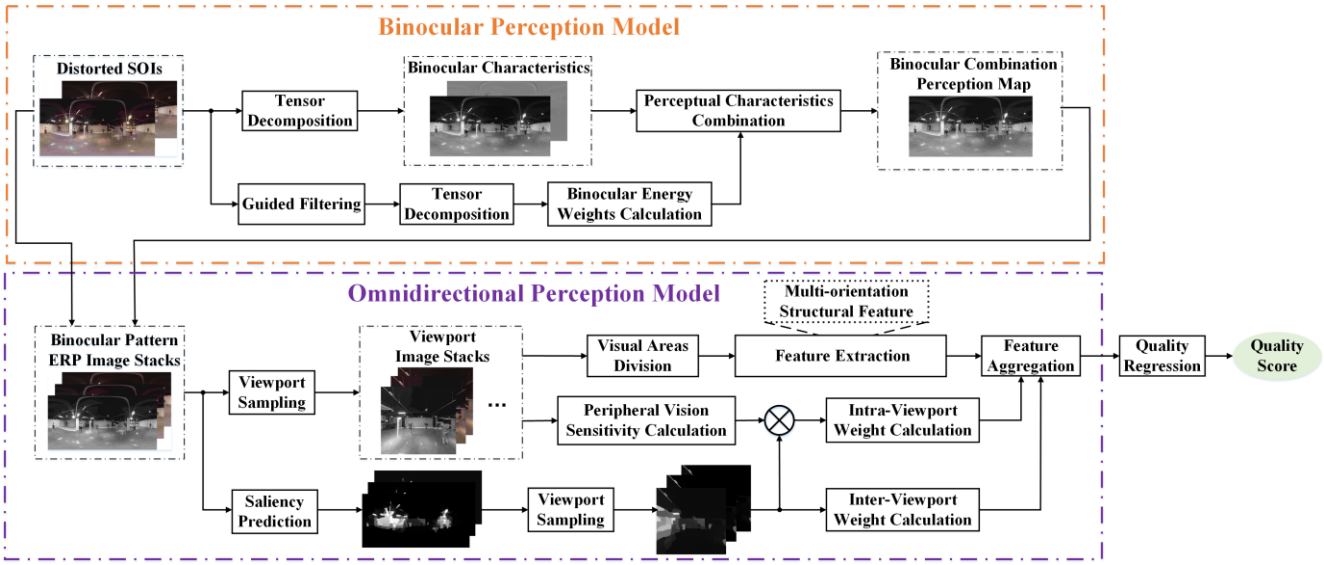


Fig. 5. The proposed viewport perception based blind stereoscopic omnidirectional image quality assessment method.

$\boldsymbol{I}_{\mathrm{d}}^{\mathrm{F}} \in \Re^{h\times w\times 3\times 2}$ composed of $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{L,R}}$ and $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{R,L}}$ is decomposed into the filtered tensor sub-bands representing binocular fusion and binocular rivalry characteristics by Eq. (1), denoted as $\boldsymbol{I}_1^{\mathrm{FS}}$ and $\boldsymbol{I}_2^{\mathrm{FS}}$, respectively.

2) *Binocular Energy Weights Calculation*: Binocular energy responses determine the level of binocular perception by modeling the binocular features, and Log-Gabor filter can simulate the simple cells in primary visual cortex well. Hence, binocular energy is simply calculated by utilizing the responses of Log-Gabor filter. Generally, a set of responses in different directions o and scales s obtained by Log-Gabor filter can be denoted as $[\boldsymbol{\eta}_{s,o}, \boldsymbol{\xi}_{s,o}]$, and the transfer function $f_{s,o}^{\mathrm{G}}$ of Log-Gabor filter is defined as

$$f_{s,o}^{\mathrm{G}}(\omega,\psi)=\exp\left(-\frac{(\log(\omega/\omega_s))^2}{2\sigma_s^2}\right)\cdot\exp\left(-\frac{(\psi-\psi_o)^2}{2\sigma_o^2}\right) \quad (2)$$

where ω and ψ are the radial frequency and orientation angle of the filter, respectively. ω_s and ψ_o are the corresponding center frequency and center orientation angle of the filter, respectively.

Then, according to the response summation in all directions and scales, the local energy $\boldsymbol{E}$ is computed by

$$\boldsymbol{E}=\sqrt{\left(\sum_s\sum_o \boldsymbol{\eta}_{s,o}\right)^2+\left(\sum_s\sum_o \boldsymbol{\xi}_{s,o}\right)^2} \quad (3)$$

Finally, the corresponding binocular energy weights with respect to two binocular perception characteristic maps are calculated by

$$\boldsymbol{w}_1^{\mathrm{FS}}=\boldsymbol{E}_1^{\mathrm{FS}}/(\boldsymbol{E}_1^{\mathrm{FS}}+\boldsymbol{E}_2^{\mathrm{FS}}+\varepsilon) \quad (4)$$

$$\boldsymbol{w}_2^{\mathrm{FS}}=\boldsymbol{E}_2^{\mathrm{FS}}/(\boldsymbol{E}_1^{\mathrm{FS}}+\boldsymbol{E}_2^{\mathrm{FS}}+\varepsilon) \quad (5)$$

where $\boldsymbol{E}_1^{\mathrm{FS}}$ and $\boldsymbol{E}_2^{\mathrm{FS}}$ are the energy response of $\boldsymbol{I}_1^{\mathrm{FS}}$ and $\boldsymbol{I}_2^{\mathrm{FS}}$ calculated by Eq. (3), respectively, and ε is a constant to avoid the denominator close to zero.

3) *Perceptual Characteristics Combination*: HVS can perceive the external binocular visual stimuli and integrate them into a single combination perception. When similar contents exist in both views, the combination perception of this area is easily dominated by binocular fusion. If different contents exist in both views, the combination perception is dominated by binocular rivalry. Therefore, based on the binocular energy weights, the binocular combination perception map $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{P}}$ of SOI can be expressed as

$$\boldsymbol{I}_{\mathrm{d}}^{\mathrm{P}}=\boldsymbol{w}_1^{\mathrm{FS}}\cdot\boldsymbol{I}_1^{\mathrm{TS}}+\boldsymbol{w}_2^{\mathrm{FS}}\cdot\boldsymbol{I}_2^{\mathrm{TS}} \quad (6)$$

where $\boldsymbol{I}_1^{\mathrm{TS}}$ and $\boldsymbol{I}_2^{\mathrm{TS}}$ are the binocular perceptual characteristic maps obtained by Eq. (1). Finally, distorted left and right views of SOI and their binocular combination perception map are integrated to form the distorted binocular pattern ERP image stacks as the input of OPM, denoted as $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{E}}=[\boldsymbol{I}_{\mathrm{d}}^{\mathrm{P}}, \boldsymbol{I}_{\mathrm{d}}^{\mathrm{L}}, \boldsymbol{I}_{\mathrm{d}}^{\mathrm{R}}]$.

To illustrate the effectiveness of the obtained binocular combination perception map more intuitively, a demonstration test for asymmetrical distortion is depicted in Fig. 6. Specifically, six binocular combination perception maps and their local enlarged maps (corresponding to red rectangular boxes) are given at different distortion levels, where each column from the left to right in Figs. 6(c) and 6(d) corresponds to JPEG distortion level 6-11, respectively. By observing the local enlarged maps, it can be found that the images in the first two columns have more obvious blocking artifacts than those in the last four columns, and the texture or smooth areas of the images in the last two columns are clearer than those in the middle two columns. Moreover, the quality degradation between the images in the first two columns (or the middle and last two columns) is very similar subjectively. According to the subjective experiment, the average DMOS values of the SOIs corresponding to Figs. 6(a) and 6(b) are 1.96, 2.08, 3.48, 3.78, 4.40 and 4.57, respectively, which are consistent with the

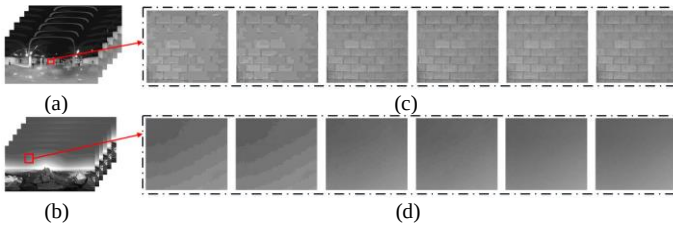


Fig. 6. Binocular combination perception maps at JPEG asymmetrical distortion levels. (a)-(b) Binocular combination perception maps. (c)-(d) Corresponding local enlarged areas of images in the first column.

subjective feelings of binocular combination perception maps. Therefore, the calculated binocular combination perception maps can accurately measure the binocular masking effect of HVS when viewing SOIs.

B. Omnidirectional Perception Model (OPM)

Firstly, the special viewport sampling approach is applied to the distorted binocular pattern ERP image stacks generated by BPM. Then, the joint influences of visual attention and peripheral visual sensitivity are used to simulate the user behavior. Finally, the intra-viewport and inter-viewport weighting factors are designed to aggregate the features extracted from all potential viewports.

1) *Viewport Sampling*: Omnidirectional images are usually stored in ERP format, which will cause stretching distortion in the polar regions. Therefore, viewport rendering is an essential part to restore the original spherical shape of an SOI. Fig. 7 depicts three examples of viewport sampling approaches [41, 55], named as uniform, tropical and proposed, respectively. From Fig. 7, it can be found that the tropical sampling only collects multiple viewports in the tropical areas (45 degrees off equator). The uniform sampling has a better coverage of sphere, but it usually contains a large number of viewports. Considering that HVS tends to focus on the equator from user behavior analysis [56], we design a special viewport sampling approach (*i.e.*, Fig. 7(c)) to cover all viewing information in theoretically possible directions. Specifically, let M be the total number of viewports, then, a total of M-2 viewports are circularly distributed at equal intervals, *i.e.*, the angle between two adjacent viewports can be calculated as $\psi=2\pi/(M\text{-}2)$, while the other two viewports are toward the polar regions, respectively.

2) *User Behavior Simulation*: On the basis of the presented viewport sampling approach, the relevant distorted binocular pattern viewport image stacks can be denoted as $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{V}_m}=[\boldsymbol{I}_{\mathrm{d}}^{\mathrm{P,V}_m},\boldsymbol{I}_{\mathrm{d}}^{\mathrm{L,V}_m},\boldsymbol{I}_{\mathrm{d}}^{\mathrm{R,V}_m}]$ (m=1,2,...,M), where m is the index of viewport. Moreover, it is well-known that different areas of the current viewport will appear varying degrees of visual perception, which is relevant to the user's viewing behavior, namely, human eyes tend to pay more attention to the areas with salient objects.

According to the related visual and neuroscientific research work [57], the binocular FoV of human is about 220° in the horizontal direction, which mainly includes three different visual areas, *i.e.*, central vision area (CVA) with one-side 9° eccentrically, near peripheral area (NPA) with one-side 30° eccentrically and far peripheral area (FPA) covering the remaining regions. To study the influence of peripheral vision on the immersive image quality, a subjective experiment based on the just noticeable difference (JND) criteria was conducted in [58]. The results indicate that the relationship between JND threshold q_{p} and one-side eccentricity angle θ_{p} can be modeled by the generalized Gaussian function, which is expressed as

$$q_{\mathrm{p}}=\frac{1}{c\sqrt{2\pi}}\cdot\exp\left(-\frac{|(b\cdot\theta_{\mathrm{p}})^{a}|}{2c^{2}}\right)+d \tag{7}$$

where a, b, c and d are fitting parameters derived by least square principle, respectively.

To quantify the influence of peripheral visual sensitivity on the current viewport, we first divide each viewport image into 100 blocks with the same resolution and each block's FoV is 11°. As depicted in Fig. 8(a), the distribution of these blocks is roughly similar to the distribution of CVA, NPA and FPA in human vision. The blocks with different colors constitute five areas with different visual sensitivities, where θ_{p} corresponding to each visual area is 11°, 22°, 33°, 44° and 55°, respectively. Then, the JND thresholds $\{q_{\mathrm{p}}\}$ of five areas can be calculated by Eq. (7), and smaller threshold means higher visual sensitivity for the distortion in the viewport image, *i.e.*, a larger weight allocated. Finally, the normalized visual sensitivity weights with respect to five visual areas can be determined as $\{w_{\mathrm{s}}\}$={0.4490, 0.2601, 0.1272, 0.0851, 0.0786}, and the associated peripheral vision sensitivity map $\boldsymbol{S}^{\mathrm{V}}$ for each viewport image is depicted in Fig. 8(b). Evidently, human eyes are more sensitive to the quality degradation of CVA, which is probably related to the highly uneven distribution of photoreceptors on the human retina.

In addition, users usually focus on the area of interest by freely rotating their heads. Here, an advanced saliency detection algorithm [59] designed for stereoscopic images is used to estimate the head movement of user, so that the left and right visual attention maps with certain parallax can be obtained and denoted as $\boldsymbol{A}^{\mathrm{L}}$ and $\boldsymbol{A}^{\mathrm{R}}$, respectively. In the same way, the

Fig. 7. Comparison of different viewport sampling approaches [41]. (a) Uniform. (b) Tropical. (c) Proposed.

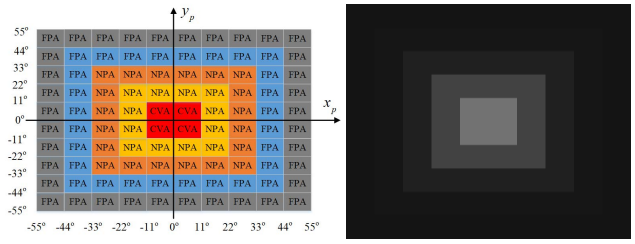


Fig. 8. The perceptual impact of peripheral vision. (a) Visual areas division. (b) Peripheral vision sensitivity map.

binocular combination attention map $\boldsymbol{A}^{\mathrm{P}}$ is calculated according to Eqs. (1)-(6), and combined with $\boldsymbol{A}^{\mathrm{L}}$ and $\boldsymbol{A}^{\mathrm{R}}$ to form the visual attention ERP image stacks, denoted as $\boldsymbol{A}^{\mathrm{E}}=[\boldsymbol{A}^{\mathrm{P}}, \boldsymbol{A}^{\mathrm{L}}, \boldsymbol{A}^{\mathrm{R}}]$. Then, according to the proposed viewport sampling approach, the visual attention viewport image stacks can be expressed as $\boldsymbol{A}^{\mathrm{V}_m}=[\boldsymbol{A}^{\mathrm{P,V}_m}, \boldsymbol{A}^{\mathrm{L,V}_m}, \boldsymbol{A}^{\mathrm{R,V}_m}]$ (m=1,2,...,M). Finally, considering both of visual attention and peripheral vision sensitivity, the intra-viewport perceptual weight matrix $\boldsymbol{w}_{\mathrm{in}}^{\mathrm{V}_m} \in \boldsymbol{R}^{H_{\mathrm{v}} \times W_{\mathrm{v}}}$ is constructed to simulate user's visual perception for different areas of the current viewport, which is computed by

$$\boldsymbol{w}_{\mathrm{in}}^{\mathrm{V}_m} = \boldsymbol{A}^{\mathrm{V}_m} \cdot \boldsymbol{S}^{\mathrm{V}}, \ m = 1,2,...,M \tag{8}$$

where H_{v} and W_{v} are the height and width of viewport image, respectively. Since the total visual attention areas contained in different viewports are usually various, the inter-viewport perceptual weight $w_{\mathrm{out}}^{\mathrm{V}_m}$ is constructed to indicate user's degree of interest in different viewports, which is calculated by

$$w_{\mathrm{out}}^{\mathrm{V}_m} = \sum_{p} \boldsymbol{A}^{\mathrm{V}_m}(p) / \sum_{m} \sum_{p} \boldsymbol{A}^{\mathrm{V}_m}(p), \ m = 1,2,...,M \tag{9}$$

where $\boldsymbol{A}^{\mathrm{V}_m}(p)$ is the visual attention value at the position of p on the m-th viewport image.

3) *Feature Extraction and Aggregation*: Given a feature matrix $\boldsymbol{F}^{\mathrm{V}_m}$ measuring the distortion of viewport images, the final viewport perception weighted feature value F_{w} can be calculated by the linear pooling operation according to $\boldsymbol{w}_{\mathrm{in}}^{\mathrm{V}_m}$ and $w_{\mathrm{out}}^{\mathrm{V}_m}$, where $\boldsymbol{F}^{\mathrm{V}_m}$ is calculated by traditional IQA metrics. Remarkably, for the FR-IQA metrics, $\boldsymbol{F}^{\mathrm{V}_m}$ is usually obtained by measuring the difference of brightness or features between the pixels of reference image and distorted image, and is also known as the local quality map. Therefore, $\boldsymbol{F}^{\mathrm{V}_m}$ can be directly weighted by $\boldsymbol{w}_{\mathrm{in}}^{\mathrm{V}_m}$ to conduct the feature aggregation. However, for the BIQA metrics, the pixel-based $\boldsymbol{F}^{\mathrm{V}_m}$ is unable to be obtained because the metrics usually adopt the global feature extraction ways, that is, only a corresponding feature value can be extracted for the whole image or local area. To bridge this gap, $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{V}_m}$ and $\boldsymbol{w}_{\mathrm{in}}^{\mathrm{V}_m}$ are first divided into multiple blocks as described in Fig. 8(a), and the global feature values for 100 local blocks are extracted by BIQA metrics to form the block-based feature matrix $\boldsymbol{F}_{\mathrm{b}}^{\mathrm{V}_m} \in \boldsymbol{R}^{10\times10}$. Then, the block-based intra-viewport perceptual weight matrix $\boldsymbol{w}_{\mathrm{in,b}}^{\mathrm{V}_m} \in \boldsymbol{R}^{10\times10}$ is also formed by summing the total visual attention areas contained in each block of $\boldsymbol{w}_{\mathrm{in}}^{\mathrm{V}_m}$. Finally, according to two perceptual weights, $\boldsymbol{w}_{\mathrm{in,b}}^{\mathrm{V}_m}$ and $w_{\mathrm{out}}^{\mathrm{V}_m}$, F_{w} can be computed by

$$F_{\mathrm{w}} = \sum_{m=1}^{M} \left(\sum_{p} \left(\boldsymbol{F}_{\mathrm{b}}^{\mathrm{V}_m}(p) \cdot \boldsymbol{w}_{\mathrm{in,b}}^{\mathrm{V}_m}(p) \right) / \sum_{p} \boldsymbol{w}_{\mathrm{in,b}}^{\mathrm{V}_m}(p) \right) \cdot w_{\mathrm{out}}^{\mathrm{V}_m} \tag{10}$$

where $\boldsymbol{F}_{\mathrm{b}}^{\mathrm{V}_m}(p)$ and $\boldsymbol{w}_{\mathrm{in,b}}^{\mathrm{V}_m}(p)$ are the feature value and the intra-viewport perceptual weight value for the image block at the position of p on the m-th viewport image, respectively.

C. Multi-orientation Structural Feature Extraction

In this section, an efficient feature extraction approach based on the multi-orientation gradient maps' local binary pattern (MOGLBP) is given to measure the distortion in each viewport image, thereby combined with the above presented BPM and OPM. It can be divided into two stages, *i.e.*, MOGLBP construction and MOGLBP histogram calculation.

1) *MOGLBP Construction*: Given a viewport image $\boldsymbol{I}_{\mathrm{v}}$, its horizontal and vertical gradients are computed by convolving images with Prewitt operator, thus the gradient magnitude $\boldsymbol{G}_{\mathrm{m}}$ and gradient orientation $\boldsymbol{G}_{\mathrm{o}}$ can be obtained. Due to the simplicity of calculation, eight common orientation angles are selected to make full use of the image gradient information, denoted as ϕ={0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°}, and the i-th orientation channel map $\boldsymbol{O}_{\mathrm{c}}^{i}$ can be calculated by

$$\boldsymbol{O}_{\mathrm{c}}^{i} = \cos(\boldsymbol{G}_{\mathrm{o}}) \cdot \cos(\phi(i)) + \sin(\boldsymbol{G}_{\mathrm{o}}) \cdot \sin(\phi(i)), \ i=1,2,...,8 \tag{11}$$

Then, the overall gradient maps $\boldsymbol{G}_{\mathrm{om}}^{i}$ are computed as

$$\boldsymbol{G}_{\mathrm{om}}^{i} = \boldsymbol{G}_{\mathrm{m}} \cdot \boldsymbol{O}_{\mathrm{c}}^{i}, \ i = 1,2,...,8 \tag{12}$$

Finally, local binary pattern (LBP) [60] is applied to $\boldsymbol{G}_{\mathrm{om}}^{i}$ for finely encoding the associated primitive structures of images, thereby generating the MOGLBP operator $\boldsymbol{L}_{P,R}^{\mathrm{MO}}$ as follows

$$\boldsymbol{L}_{P,R}^{\mathrm{MO}} = [\boldsymbol{L}_{P,R}^{1}, \boldsymbol{L}_{P,R}^{2}, ..., \boldsymbol{L}_{P,R}^{8}] \tag{13}$$

where $\boldsymbol{L}_{P,R}^{i}$ (i=1,2,...,8) is the gradient LBP map along the i-th orientation. P is the number of neighbors and R is the size of neighborhood in LBP operation.

Generally, $\boldsymbol{L}_{P,R}^{i}$ may have P+2 patterns, and it is accurate and reliable to discriminate structural information in different orientations. When images are destroyed by the distortions with various intensities, the MOGLBP patterns would be transformed from one type to another. In Fig. 9, the first column depicts a reference viewport image patch and its distorted versions generated from three codecs, and the other columns show the corresponding MOGLBP maps. From Fig. 9, it can be observed that the MOGLBP patterns change as distortion types vary. So, MOGLBP can be utilized to identify some complex distortions.

2) *MOGLBP Histogram Calculation*: The MOGLBP operator is computed from the difference relationship between center pixel and its surrounding neighbors, so it is invariant to the gradient magnitude value at the position of center pixel. However, HVS is highly sensitive to the local contrast change of images that has a significant influence on perceptual quality. To simultaneously capture the structure and contrast information of images, the gradient-weighted MOGLBP histogram h_{MO} is calculated by simply integrating gradient information in each orientation into different pattern descriptions, which is defined as

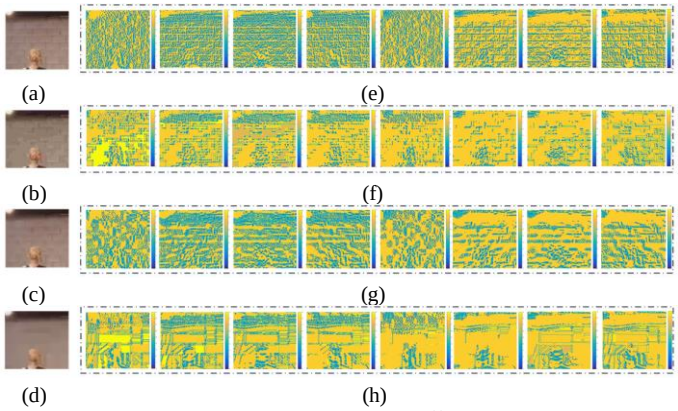


Fig. 9. Viewport image patches under different distortions and the corresponding MOGLBP maps. (a) Reference image patch. (b) Image patch with JPEG distortion. (c) Image patch with JPEG2000 distortion. (d) Image patch with HEVC distortion. (e)-(h) Corresponding MOGLBP maps of image patches in the first column.

$$h_{\mathrm{MO}} = [h_{\mathrm{L}}^{1}, h_{\mathrm{L}}^{2}, ..., h_{\mathrm{L}}^{8}] \tag{14}$$

where h_{L}^{i} (i=1,2,...,8) is the gradient-weighted LBP histogram along the i-th orientation, and deduced as

$$h_{\mathrm{L}}^{i}(j) = \sum_{p=1}^{N_{\mathrm{p}}} \boldsymbol{G}_{\mathrm{om}}^{i}(p') \cdot f_{\mathrm{L}}(\boldsymbol{L}_{P,R}^{i}(p'), j), \quad i = 1,2,...,8 \tag{15}$$

$$f_{\mathrm{L}}(x, y) = \begin{cases} 1, & x = y \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

where N_{p} is the total number of pixels in a LBP map and j is the possible LBP patterns index. In this way, the structural degradation with strong contrast change in the distorted image is emphasized.

Furthermore, P and R in the above equations are empirically set to 8 and 1 for MOGLBP calculation, respectively. Therefore, 80-dimensional features can be obtained to measure the distortion in each viewport image in terms of the MOGLBP histogram. However, in order to robustly combine OPM with global feature extraction ways commonly used in traditional BIQA metrics, the presented MOGLBP operator is applied to each viewport image block divided from $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{V}_m}$ according to the block-dividing scheme in Fig. 8(a). Then, for the feature at each dimension, the corresponding viewport perception weighted feature value F_{w} can be calculated by Eq. (10). Finally, the formed quality-aware feature vector is 240 dimensions in total, denoted as $\boldsymbol{F}_{\mathrm{f}} = [\boldsymbol{F}_{\mathrm{w}}^{\mathrm{P}}, \boldsymbol{F}_{\mathrm{w}}^{\mathrm{L}}, \boldsymbol{F}_{\mathrm{w}}^{\mathrm{R}}]$, where $\boldsymbol{F}_{\mathrm{w}}^{\mathrm{P}}$, $\boldsymbol{F}_{\mathrm{w}}^{\mathrm{L}}$ and $\boldsymbol{F}_{\mathrm{w}}^{\mathrm{R}}$ are the aggregated 80-dimensional feature vectors extracted from $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{P},\mathrm{V}_m}$, $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{L},\mathrm{V}_m}$ and $\boldsymbol{I}_{\mathrm{d}}^{\mathrm{R},\mathrm{V}_m}$, respectively.

D. Quality Regression

After feature extraction, the feature space is mapped to predict the objective quality Q_{f} of SOIs, which is expressed as

$$Q_{\mathrm{f}} = f_{\mathrm{m}}(\boldsymbol{F}_{\mathrm{f}}) \tag{17}$$

where $f_{\mathrm{m}}(\cdot)$ is a non-linear mapping function that can be achieved by machine learning, and $\boldsymbol{F}_{\mathrm{f}}$ is the extracted feature vector. Specifically, a quality prediction model is first learned using a set of training images. Then the trained model is used to evaluate the quality of testing images. During the prediction process, we have tested both support vector regression (SVR) [61] and random forest regression (RFR) [62], from which we have chosen RFR because it significantly outperformed SVR in the following experiments. The setting of hyper-parameters in the process of RFR, including the number of trees N_{T} and the maximum number of features N_{M}, will be stated in the subsection V-A.

V. Experimental Results and Discussions

In this section, the proposed VP-BSOIQA method is tested and compared with the state-of-the-art IQA metrics on the individual database and distortion type. Then, the impacts of the viewport sampling approaches, the number of viewports, and two perception models are analyzed. Finally, the discussions about strengths and weaknesses of the method are given.

A. Experimental Settings

1) *Databases*: The comparative experiments were conducted on two available SOI databases (*i.e.*, NBU-SOID [46] and SOLID [45]). The self-built NBU-SOID database contains 396 distorted images with three common distortion types (*i.e.*, JPEG, JPEG2000 and HEVC). The SOLID database includes 138 distorted images with three depth levels degraded by BPG compression and three associated mean opinion score (MOS) values (*i.e.*, image quality, depth perception and overall quality of experience). It is noteworthy that only image quality is considered in our experiment.

2) *Evaluation Criteria*: Three commonly-used evaluation criteria are used to reflect the relationship between the predicted objective quality and human subjective scores, *i.e.*, Pearson linear correlation coefficient (PLCC), Spearman rank-order correlation coefficient (SRCC) and root mean squared error (RMSE). The PLCC, SRCC and RMSE measure the accuracy, monotonicity and error in the process of prediction, respectively. The closer the absolute values of PLCC and SRCC are to 1, and the closer RMSE is to 0, the better the performance of objective IQA methods is. According to the report from video quality expert group (VQEG) [63], a 5-parametric logistic regression process is additionally applied to make the relationship between the subjective and objective scores more linear before calculating the values of PLCC and RMSE.

3) *Experimental Parameters*: In terms of the proposed VP-BSOIQA method, there are several parameters to be determined in the implementation process, *i.e.*, the height, weight and FoV of viewport image (H_{v}, W_{v} and F_{v}), the number of viewport (M), the constant in Eqs. (4)-(5) (ε), the fitting parameters in Eq. (7) (a, b, c, d), the number and size of neighbors in MOGLBP operation (P and R), and two hyper-parameters in the process of RFR (N_{T} and N_{M}). Firstly, considering that the objective evaluation object is required to be consistent with the image viewed by user as far as possible, viewport rendering process in this work is strictly conducted by using official software 360Lib [64], and H_{v}, W_{v} and F_{v} are the same as the internal parameters of HMD used in the subjective experiment, *i.e.*, H_{v}=1600, W_{v}=1440, F_{v}=110.

TABLE III
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND TWENTY-ONE IQA METRICS ON TWO DATABASES.

Types	Metrics	NBU-SOID Database [46]			SOLID Database [45]		
		PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
2D FR-IQA	PSNR	0.781	0.791	0.596	0.710	0.673	0.653
	SSIM [7]	0.830	0.833	0.531	0.857	0.878	0.477
	VIF [8]	0.828	0.837	0.535	0.875	0.794	0.449
	RFSIM [9]	0.861	**0.867**	0.485	0.892	**0.890**	0.420
	GMSD [10]	0.862	0.864	0.484	0.827	0.779	0.521
FR-OIQA	S-PSNR [26]	0.809	0.827	0.561	0.720	0.663	0.643
	CPP-PSNR [27]	0.803	0.820	0.568	0.708	0.658	0.654
	WS-PSNR [28]	0.795	0.811	0.579	0.702	0.657	0.660
	WS-SSIM [29]	0.832	0.835	0.529	0.877	0.889	0.446
	VA-PSNR [31]	0.797	0.820	0.575	0.761	0.720	0.601
2D BIQA	BRISQUE [12]	0.747	0.592	0.573	0.797	0.583	0.471
	OG [13]	0.811	0.753	0.509	0.827	0.691	0.507
	NIQE [14]	0.650	0.628	0.725	0.604	0.604	0.739
	dipIQ [15]	0.743	0.735	0.638	0.595	0.538	0.745
	BPRI [16]	0.647	0.602	0.727	0.815	0.801	0.538
	HOSA [17]	0.741	0.710	0.641	0.564	0.539	0.766
	UCA [18]	0.649	0.347	0.726	0.895	0.890	0.414
BSIQA	SINQ [25]	0.803	0.762	0.496	0.808	0.728	0.451
BOIQA	ASY-OIQA [34]	0.843	0.830	0.456	0.749	0.630	0.526
	SSP-OIQA [37]	0.740	0.689	0.553	0.765	0.695	0.541
BSOIQA	Yang *et al.* [35]	0.861	0.856	0.455	0.855	0.801	0.478
	Proposed (SVR)	0.834	0.828	0.501	0.914	0.843	0.542
	Proposed (RFR)	**0.883**	0.860	**0.410**	**0.946**	0.878	**0.278**

Secondly, the constant in Eqs. (4)-(5) is used to avoid the denominator close to zero, so we set ε as 10^{-8}. Thirdly, the fitting parameters in Eq. (7) is derived by least square principle, so we set a, b, c, d as 2.2, 0.08, 1.38, 0.05, respectively, which are in accordance with the advice in [58]. Fourthly, the parameters, P and R, are set as 8 and 1, respectively, which is in line with the recommendation in [60]. Fifthly, N_T is the number of trees before taking the maximum voting or averages of predictions, and N_M is the maximum number of features in individual tree. Generally, higher number of trees gives a better performance and slower efficiency, and N_M is not greater than the square root of total number of features in individual run. Hence, N_T and N_M are set as 1500 and 15, respectively. Finally, the number of viewport, M, will probably affect the performance of method. Therefore, we select the optimal value (*i.e.*, M=10) to achieve a trade-off between complexity and accuracy, and the specific details about analyzing the impact of number of viewport on performance will be discussed in the following sections. Our source code and NBU-SOID database will be released soon at https://github.com/qyb123/.

B. Performance Comparison on Individual Database

To testify the performance of the proposed VP-BSOIQA method, we compare it on two individual databases with twenty-one state-of-the-art IQA metrics, including five 2D FR-IQA metrics (PSNR, SSIM [7], VIF [8], RFSIM [9] and GMSD [10]), five FR-OIQA metrics (S-PSNR [26], CPP-PSNR [27], WS-PSNR [28], WS-SSIM [29] and VA-PSNR [31]), seven 2D BIQA metrics (BRISQUE [12], OG [13], NIQE [14], dipIQ [15], BPRI [16], HOSA [17] and UCA [18]), one BSIQA metric (SINQ [25]), two BOIQA metrics (ASY-OIQA [34] and SSP-OIQA [37]) and one BSOIQA metric (Yang *et al.* [35]). In terms of BIQA metrics, seven metrics (including BRISQUE, OG, SINQ, ASY-OIQA, SSP-OIQA, Yang *et al.* and the proposed VP-BSOIQA metrics) require subjective scores for training, while other BIQA metrics do not. For these metrics based on supervised learning, the database needs to be divided into training and testing sets, and K-fold cross validation is used to evaluate the performance of IQA models. For each database, reference images along with their corresponding distorted versions are first separated into K subsets. Then, each subset is used for testing separately and the remaining K-1 subsets are used for training so that the scenes for training will not appear in the testing set. Finally, the mean performance is reported. Thus, when comparing existing learning based BIQA methods, we extract the relevant features with these models and re-train them on each database. Furthermore, for metrics other than our method and SINQ, the predicted quality scores of left and right images are averaged as the final quality of SOIs. To ensure reliability of the final results, only one subset is tested at each turn, so the value of K is consistent with the total number of reference images in each database, *i.e.*, K is set to 12 and 6 for NBU-SOID database and SOLID database, respectively. The results of PLCC, SRCC and RMSE on two databases for the proposed VP-BSOIQA method and twenty-one objective IQA metrics are tabulated in Table III, and the best performance is highlighted in boldface. It is noting that all indicators are obtained by running the source codes provided by the authors of the published literatures.

From Table III, we can have the following observations.

1) Compared with PSNR or SSIM metric, four FR-OIQA metrics specifically serving for omnidirectional images coding (other than VA-PSNR) have achieved a slight performance improvement respectively, which indicates that calculating the quality on the plane closer to spherical surface can eliminate the impact of projection transformation, and the shape change of distortion caused by projection transformation may have significant influence on human visual perception. It can be

TABLE IV
PERFORMANCE (PLCC/SRCC) OF ALL MODELS FOR DIFFERENT DISTORTION TYPES ON TWO DATABASES. THE ABBREVIATIONS 'SYM' AND 'ASYM' DENOTE SYMMETRICAL AND ASYMMETRICAL DISTORTION, RESPECTIVELY.

Types	Metrics	NBU-SOID Database [46]					SOLID Database [45]	
		JPEG	JPEG2000	HEVC	Sym	Asym	Sym	Asym
2D FR-IQA	PSNR	0.732/0.744	0.851/0.853	0.798/0.785	0.854/0.870	0.673/0.690	0.804/0.843	0.503/0.423
	SSIM [7]	0.885/0.879	0.887/0.888	0.860/0.844	0.893/0.865	0.773/0.778	0.965/0.927	0.751/0.756
	VIF [8]	0.850/0.853	0.880/0.878	0.879/0.880	0.908/0.895	0.744/0.746	0.939/**0.943**	0.792/0.668
	RFSIM [9]	0.959/0.952	0.934/**0.933**	0.883/0.872	0.926/0.907	0.797/0.788	0.964/0.912	0.809/0.802
	GMSD [10]	0.960/0.939	0.926/0.925	0.897/0.888	0.921/0.895	0.794/0.794	0.883/0.889	0.710/0.588
FR-OIQA	S-PSNR [26]	0.771/0.788	0.875/0.882	0.838/0.831	0.879/0.898	0.718/0.729	0.812/0.829	0.591/0.416
	CPP-PSNR [27]	0.808/0.816	0.871/0.877	0.828/0.825	0.898/0.893	0.713/0.723	0.804/0.824	0.554/0.407
	WS-PSNR [28]	0.739/0.750	0.869/0.874	0.819/0.815	0.865/0.890	0.688/0.710	0.799/0.824	0.576/0.403
	WS-SSIM [29]	0.894/0.886	0.885/0.890	0.854/0.846	0.894/0.870	0.775/0.777	0.973/0.934	0.785/0.784
	VA-PSNR [31]	0.696/0.729	0.897/0.907	0.817/0.811	0.867/0.899	0.692/0.716	0.831/0.861	0.592/0.504
2D BIQA	BRISQUE [12]	0.919/0.879	0.877/0.688	0.683/0.604	0.837/0.693	0.613/0.471	0.875/0.759	0.745/0.473
	OG [13]	0.925/0.907	0.896/0.834	0.858/0.771	0.818/0.782	0.815/0.697	0.871/0.819	0.718/0.566
	NIQE [14]	0.778/0.743	0.807/0.793	0.579/0.510	0.741/0.698	0.550/0.525	0.800/0.657	0.669/0.536
	dipIQ [15]	0.795/0.764	0.888/0.881	0.622/0.590	0.821/0.784	0.636/0.667	0.667/0.660	0.394/0.335
	BPRI [16]	0.752/0.652	0.851/0.843	0.562/0.508	0.687/0.613	0.633/0.593	0.859/0.778	0.794/0.752
	HOSA [17]	0.790/0.766	0.849/0.849	0.602/0.564	0.811/0.779	0.663/0.601	0.658/0.738	0.413/0.414
	UCA [18]	0.869/0.863	0.749/0.670	0.877/0.866	0.662/0.309	0.675/0.378	0.962/0.910	0.817/**0.820**
BSIQA	SINQ [25]	0.964/0.949	0.876/0.793	0.863/0.777	0.839/0.791	0.796/0.723	0.870/0.794	0.805/0.682
BOIQA	ASY-OIQA [34]	0.889/0.859	0.939/0.897	0.820/0.791	0.889/0.841	0.829/0.794	0.834/0.792	0.636/0.498
	SSP-OIQA [37]	0.927/0.881	0.887/0.848	0.865/0.801	0.831/0.715	0.747/0.627	0.839/0.779	0.761/0.693
BSOIQA	Yang *et al.* [35]	0.934/0.935	0.914/0.886	0.908/**0.899**	**0.938**/**0.909**	0.829/0.753	0.948/0.937	0.864/0.644
	Proposed	**0.973**/**0.957**	**0.949**/0.889	**0.948**/0.875	0.893/0.862	**0.873**/**0.833**	**0.982**/0.927	**0.907**/0.777

found that the VA-PSNR metric significantly outperforms PSNR metric on the SOLID database, which indicates the importance of visual attention in the immersive environment. Moreover, in terms of PLCC, RFSIM and GMSD metrics designed for 2D images reach 0.861 and 0.862 on the NBU-SOID database, respectively. We attribute the performance difference between objective IQA metrics to the appropriateness of extracted features, and the structure or texture features in RFSIM or GMSD metrics can discriminate the perceptible image degradation well.

2) For the 2D BIQA and BSIQA metrics based on supervised learning, BRISQUE and SINQ metrics are based on the natural scene statistic features extracted from spatial domain, while the relative statistic features in gradient domain are utilized in OG metric. Compared with BRISQUE metric, SINQ metric considering stereoscopic perception characteristics has higher consistency with HVS, namely increased by 0.056 and 0.170 (or 0.011 and 0.145) on PLCC and SRCC for the NBU-SOID database (or SOLID database), respectively. This result demonstrates that there exists the binocular masking effect when users view SOIs. However, due to feature extraction and distortion identification conducted on gradient maps, OG metric performs slightly better than SINQ metric in terms of PLCC. Its results prove once again that images' structural information play a significant role in SOIQA. Furthermore, in terms of five 2D BIQA metrics without MOS for training, they have not achieved satisfactory results on all databases. For instance, PLCC of dipIQ metric on SOLID database is only 0.595 and the value of BPRI metrics on NBU-SOID database is 0.647, respectively, which is probably related to the deficiency of robustness of IQA models. Similarly, PLCC of HOSA metric on SOLID database is 0.564, but it is 0.741 on NBU-SOID database. We attribute the performance difference to the particularity of distortion in database, and HOSA metric cannot identify the BPG distortion compared with traditional distortion types, *i.e.*, JPEG, JPEG2000 and HEVC distortion. Moreover, PLCC of UCA metric on NBU-SOID database is 0.649, but it is 0.895 on SOLID database. Its result is probably related to multiple distortion types in NBU-SOID database, making the performance of UCA metric reduced.

3) To illustrate the impact caused by different machine learning algorithms, we test the performance of proposed VP-BSOIQA method using SVR and RFR, respectively. It is obvious that RFR has a more excellent performance than SVR, so RFR is chosen to conduct the quality prediction in our experiments. In general, the performance of most BIQA methods on SOLID is superior to their performance on NBU-SOID. It seems obvious because three kinds of individual distortions are included in NBU-SOID and only one distortion type is contained in SOLID, making the IQA task on NBU-SOID more challenging. For the BOIQA and BSOIQA metrics, SSP-OIQA metric conducts the feature extraction process on the SSP plane rather than ERP plane, and ASY-OIQA and Yang *et al.*'s metrics adopt the weight in the WS-PSNR or WS-SSIM metrics to solve the stretching in polar region of ERP image. On this basis, Yang *et al.*'s metric also considers the perceptual impact caused by binocular masking effect. In terms of three indicators, Yang *et al.*'s metric achieve more excellent performance than ASY-OIQA and SSP-OIQA metrics, which further indicates the necessity of binocular vision in SOIQA. In addition, the proposed VP-BSOIQA method achieves 0.883 and 0.946 on PLCC on two databases, respectively. Apparently, it outperforms all competing BIQA and FR-IQA metrics, because it considers three key factors, namely viewport, user behavior and stereoscopic perception.

C. Performance Comparison on Individual Distortion Type

To further explore the adaptability and robustness of the proposed VP-BSOIQA method, we have also made the additional performance measurement on individual distortion type (*i.e.*, JPEG, JPEG2000, HEVC distortions in NBU-SOID, symmetrical and asymmetrical distortions in two databases). More concretely, for each individual distortion type, the images belonging to each distortion type in the testing subset are selected to calculate the quality scores by using the model trained on entire training subset including all distortion types. The results of performance comparison are presented in Table IV, and the best result among all IQA metrics is highlighted in boldface. It is noting that only the values of PLCC and SRCC are reported in Table IV for brevity.

From Table IV, we can have the following observations.

1) To distinguish the performance difference of each model more intuitively, the corresponding hit count (*i.e.*, the number of times ranked in the top one for each individual distortion type) of performance is computed for each IQA metric. It can be observed that the proposed method have the highest hit count (8 times), followed by Yang *et al*. [35] (3 times), then VIF [8], RFSIM [9] and UCA [18] (1 times), and finally other metrics (0 times). Evidently, the proposed method performs more stable than all competing IQA metrics over two databases. In addition, the correlations between predicted quality and subjective scores of symmetrical distortion are generally higher than those of asymmetrical distortion, indicating that binocular perception characteristics play a significant role in evaluating SOIs and the IQA tasks executed on asymmetrical distortion type seem more challenging.

2) Compared with the state-of-the-art SIQA metric (SINQ [25]) and SOIQA metric (Yang *et al*. [35]), the proposed VP-BSOIQA method shows outstanding performance when processing the asymmetrical distortion, and achieves at least 12.7% and 5.1% improvements on PLCC, respectively. We attribute the performance superiority to the role of BPM in the proposed VP-BSOIQA method. According to the previous introduction in Section IV.A, BPM aims at generating a binocular combination perception map for predicting the binocular masking effect, and asymmetrical distortion can better reflect the masking effect of human eyes compared with symmetrical distortion. Thus, the PLCC and SRCC of the proposed VP-BSOIQA method are higher than Yang *et al*.'s metric and SINQ metric on entire database, especially for asymmetrical distortion type. Furthermore, Yang *et al*.'s metric is superior to the proposed VP-BSOIQA method on NBU-SOID database when processing the symmetrical distortion, which is understandable because many kinds of features are extracted in Yang *et al*.'s metric for solving symmetrical distortion, such as nature statistics, histogram statistics and tensor distance.

3) The PLCC value of the proposed VP-BSOIQA method on NBU-SOID database when processing the symmetrical distortion is 0.893, but it reaches 0.982 on SOLID database. We attribute the performance difference to the number of distortion types contained in two databases. Obviously, three kinds of individual distortions (*i.e.*, JPEG, JPEG2000 and HEVC) are evenly distributed in SOIs impaired by symmetrical distortion on NBU-SOID database, but only one distortion type (*i.e.*, BPG) is contained in SOIs impaired by symmetrical distortion on SOLID database, making the IQA task on NBU-SOID more challenging. More surprisingly, the proposed VP-BSOIQA method outperforms all 2D FR-IQA or FR-OIQA metrics on the NBU-SOID database in terms of asymmetrical distortion type, which further demonstrates the superiority of the proposed method.

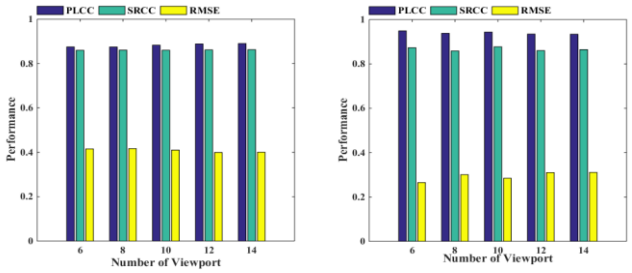


Fig. 10. Performance comparison results for different numbers of viewport on each database. (a) NBU-SOID database [46]. (b) SOLID database [45].

D. Impacts of the Number of Viewports

In terms of the proposed VP-BSOIQA method, SOIs with ERP format are first rendered into the corresponding viewport images, where the number of viewport, M, is likely to affect the performance. Therefore, a comparative experiment is designed to select the optimal number of viewport in this section. In specific, we assign M={6, 8, 10, 12, 14} and compute the values of three performance indicators on two databases for each situation, which are depicted in Fig. 10. Obviously, the variation of M will not have a significant impact on the final performance, which shows the robustness of the proposed method. Finally, we choose the median value in set (*i.e.*, M=10) as the optimal parameter to have a good tradeoff between computation complexity and accuracy.

E. Impacts of the Viewport Sampling Approaches

In terms of the proposed viewport sampling approach, most of the viewports being sampled are in the equator, which is related to the real user behavior in the immersive environment. However, different viewport sampling approaches, including uniform sampling and tropical sampling in Fig. 7, may affect the performance. Therefore, a comparative experiment is designed to explore the impact in this section. First, we collect a total of 20 viewports in the uniform sampling approach to form the uniform-based metric, *i.e.*, 8 viewports in the equator, 10 viewports in the tropic (45 degrees off equator), and 2 viewports in the polar regions. For the tropical sampling approach, we collect 10 viewports in the tropic (45 degrees off equator) to form the tropic-based metric, and the proposed method can also be named as equator-based metric. Then, three performance indicators are computed on two databases for each kind of metrics, which are tabulated in Table V, and the best performance is highlighted in boldface. From Table V, it can be found that the proposed sampling approach outperforms the other two approaches, which further verifies the fact that HVS tends to focus on the equator with less stretching distortion.

TABLE VI
PERFORMANCE (PLCC/SRCC) COMPARISON BETWEEN THE 2D-IQA METRICS AND OPTIMIZED IQA METRICS ON TWO DATABASES.

NBU-SOID Database [46]								
Types	Metrics	2D-IQA	BPM-IQA	$Gain_{BPM}$ (%)	OPM-IQA	$Gain_{OPM}$ (%)	VP-IQA	$Gain_{VP}$ (%)
2D FR-IQA	PSNR	0.781/0.791	0.813/0.836	+ 4.1/5.7	0.813/0.837	+ 4.1/5.8	0.856/0.866	+ 9.6/9.5
	SSIM [7]	0.830/0.833	0.848/0.853	+ 2.2/2.4	0.868/0.876	+ 4.6/5.2	0.886/0.893	+ 6.7/7.2
	VIF [8]	0.828/0.837	0.852/0.860	+ 2.9/2.7	0.877/0.887	+ 5.9/6.0	0.896/0.906	+ 8.2/8.2
	RFSIM [9]	0.861/0.867	0.888/0.899	+ 3.1/3.7	0.867/0.878	+ 0.7/1.3	0.896/0.907	+ 4.1/4.6
	GMSD [10]	0.862/0.864	0.884/0.886	+ 2.6/2.5	0.883/0.893	+ 2.4/3.4	0.903/0.910	+ 4.8/5.3
OA-BIQA	BRISQUE [12]	0.747/0.592	0.768/0.677	+ 2.8/14.4	0.807/0.723	+ 8.0/22.1	0.825/0.738	+ 10.4/24.7
	OG [13]	0.811/0.753	0.827/0.816	+ 2.0/8.4	0.823/0.766	+ 1.5/1.7	0.875/0.840	+ 7.9/11.6
-	Average	-	-	+ 2.8/5.7	-	+ 3.9/6.5	-	+ 7.4/10.2
SOLID Database [45]								
Types	Metrics	2D-IQA	BPM-IQA	$Gain_{BPM}$ (%)	OPM-IQA	$Gain_{OPM}$ (%)	VP-IQA	$Gain_{VP}$ (%)
2D FR-IQA	PSNR	0.710/0.673	0.734/0.694	+ 3.4/3.1	0.786/0.743	+ 10.7/10.4	0.793/0.760	+ 11.7/12.9
	SSIM [7]	0.857/0.878	0.861/0.881	+ 0.5/0.3	0.889/0.901	+ 3.7/2.6	0.891/0.903	+ 4.0/2.8
	VIF [8]	0.875/0.794	0.884/0.797	+ 1.0/0.4	0.904/0.839	+ 3.3/5.7	0.904/0.842	+ 3.3/6.0
	RFSIM [9]	0.892/0.890	0.893/0.892	+ 0.1/0.2	0.898/0.897	+ 0.7/0.8	0.900/0.897	+ 0.9/0.8
	GMSD [10]	0.827/0.779	0.835/0.787	+ 1.0/1.0	0.901/0.886	+ 8.9/13.7	0.901/0.889	+ 8.9/14.1
OA-BIQA	BRISQUE [12]	0.797/0.583	0.824/0.716	+ 3.4/22.8	0.847/0.705	+ 6.3/20.9	0.834/0.785	+ 4.6/34.6
	OG [13]	0.827/0.691	0.834/0.731	+ 0.8/5.8	0.858/0.782	+ 3.7/13.2	0.862/0.789	+ 4.2/14.2
-	Average	-	-	+ 1.5/4.8	-	+ 5.3/9.6	-	+ 5.4/12.2

TABLE V
PERFORMANCE COMPARISON AMONG THREE KINDS OF METRICS FORMED BY DIFFERENT VIEWPORT SAMPLING APPROACHES ON TWO DATABASES.

Metrics	NBU-SOID [46]			SOLID [45]		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
Uniform	0.864	0.841	0.425	0.935	0.877	0.301
Tropical	0.865	0.825	0.425	0.931	0.878	0.387
Proposed	**0.883**	**0.860**	**0.410**	**0.946**	**0.878**	**0.278**

F. Gains of BPM and OPM

To fully validate the effectiveness of BPM and OPM, several comparative experiments were conducted on the NBU-SOID and SOLID databases. Five 2D FR-IQA metrics (*i.e.*, PSNR, SSIM [7], VIF [8], RFSIM [9] and GMSD [10]) and two 2D BIQA metrics (*i.e.*, BRISQUE [12] and OG [13]) based on supervised learning were optimized by BPM and OPM to develop the novel BPM-IQA, OPM-IQA and VP-IQA metrics, respectively. Firstly, the BPM-IQA metrics extra consider the quality of binocular combination perception map on the basis of 2D-IQA metrics that simply average the left and right ERP image quality. Secondly, in order to eliminate the impact of BPM as much as possible, local features are only extracted from the left and right viewport images and the final quality is obtained by averaging the calculated viewport perception weighted quality of left and right views of SOI in the OPM-IQA metrics. Thirdly, similar to the proposed VP-BSOIQA method, all perceptual factors are taken into account in the VP-IQA metrics. The results of performance (*i.e.*, PLCC and SRCC) comparison among the 2D-IQA and the optimized IQA metrics are demonstrated in Table VI. Moreover, we also calculate the performance gains of BPM-IQA, OPM-IQA and VP-IQA metrics relative to the corresponding 2D-IQA metrics, denoted as $Gain_{BPM}$, $Gain_{OPM}$ and $Gain_{VP}$, respectively. Due to the similar calculation ways for three kinds of gains, we only give the definition of $Gain_{VP}$, and it can be expressed as $Gain_{VP}=(P_{VP}-P_{2D})/P_{2D}$, where P_{VP} and P_{2D} are the performance of VP-IQA and 2D-IQA metrics, respectively. As shown in Table VI, it can be found that the IQA metrics optimized by the perception models outperform the original 2D-IQA metrics, which proves the proposed BPM and OPM are effective in evaluating the SOIs.

G. Discussions

Due to the particularity of SOIs in storage form, FoV and rendering device, three key factors should be considered in the SOIQA metric, *i.e.*, viewport, user behavior and stereoscopic perception. In this paper, we have constructed two perception models (*i.e.*, BPM and OPM) to simulate the unique perceptual characteristics in the immersive environments, and proposed a new approach to designing the VP-BSOIQA metric. As an example, a simple and efficient VP-BSOIQA metric is given based on the novel multi-orientation structural feature, whose performance on two available databases is better than other competing IQA methods. Moreover, the presented two perception models can be robustly combined with some existing 2D-IQA metrics to improve the performance, thus making them suitable for SOIs, which is meaningful to explore several potential applications (*e.g.*, quality monitoring, perceptual coding and image enhancement).

Although the proposed VP-BSOIQA method achieves better results in evaluating images degraded with coding distortion, there are still some limitations in some respects. For instance, when the OPM is combined with BIQA metrics, it is necessary to perform the additional block-dividing operation due to the property of global feature extraction, where the size of block is corresponding to the visual areas partition of human eyes. However, in terms of the existing OU-BIQA metrics (*e.g.*, NIQE [14], dipIQ [15], BPRI [16], HOSA [17] and UCA [18]), the final predicted quality score is obtained by averaging the quality of all image blocks with tiny size, and the size of these pre-divided blocks is closely relevant to the pre-learned multivariate Gaussian model or codebook. Thus, it is difficult to change their original size of block to match the OPM.

VI. Conclusion and Future Work

In this paper, a viewport perception based blind stereoscopic omnidirectional images (SOIs) quality assessment (VP-BSOIQA) method has been proposed by designing the binocular perception model (BPM) and omnidirectional perception model (OPM). For the BPM, the dimensionality reduction of stereoscopic views by tensor decomposition is used to simulate the binocular fusion and binocular rivalry characteristics, and the binocular combination perception map for measuring binocular masking effect is generated by additionally considering the impact of binocular energy weights. For the OPM, viewport images are evaluated instead of ERP images to ensure the consistency of evaluation objects, and the intra-viewport and inter-viewport weighting factors are constructed by simulating the user behavior in the immersive environments to aggregate the novel multi-orientation structural features of all potential viewports. Moreover, a large-scale and diverse subjective SOIs database (named as NBU-SOID) is constructed and released for further research demand. Experimental results on two available databases have demonstrated the effectiveness of the proposed VP-BSOIQA method in predicting the quality of SOIs. In the future work, we are about to explore the perceptual characteristics in SOIs further to improve the adaptability and accuracy of SOIQA metrics.

Acknowledgment

The authors would like to thank Dr. Z. Chen and Dr. J. Yang for providing their subjective assessment databases.

References

[1] Y. Ye, J. M. Boyce, and P. Hanhart, "Omnidirectional 360º video coding technology in responses to the joint call for proposals on video compression with capability beyond HEVC," *IEEE Trans. Circuits Syst. Video Technol*., vol. 30, no. 5, pp. 1241-1252, May 2020.

[2] R. Ghaznavi-Youvalari, A. Zare, A. Aminlou, M. M. Hannuksela and M. Gabbouj, "Shared coded picture technique for tile-based viewport-adaptive streaming of omnidirectional video," *IEEE Trans. Circuits Syst. Video Technol*., vol. 29, no. 10, pp. 3106-3120, Oct. 2019.

[3] M. Xu, C. Li, S. Zhang, and P. L. Callet, "State-of-the-art in 360º video/image processing: perception, assessment and compression," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 5-26, Jan. 2020.

[4] Y. Liu, K. Gu, S. Wang, D. Zhao, and W. Gao, "Blind quality assessment of camera images based on low-level and high-level statistical features," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 135-146, Jan. 2019.

[5] Q.Yan, D. Gong, and Y. Zhang, "Two-Stream convolutional networks for blind image quality assessment," *IEEE Trans. Image Process*., vol. 28, no. 5, pp. 2200-2211, May 2019.

[6] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, *et al*., "Group maximum differentiation competition: model comparison with few samples," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 42, no. 4, pp. 851-864, Apr. 2020.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process*., vol. 13, no. 4, pp. 600-612, Apr. 2004.

[8] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process*., vol. 15, no. 2, pp. 430-444, Feb. 2006.

[9] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *proc. IEEE Int. Conf. Image Process*., Sept. 2010, pp. 321-324.

[10] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process*., vol. 23, no. 2, pp. 684-695, Feb. 2014.

[11] Y. Niu, Y. Zhong, W. Guo, Y. Shi, and P. Chen, "2D and 3D image quality assessment: a survey of metrics and challenges," *IEEE Access*, vol. 7, pp. 782-801, Jan. 2019.

[12] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference image quality assessment in the spatial domain," *IEEE Trans. Image Process*., vol. 21, no. 12, pp. 4695-4708, Dec. 2012.

[13] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, "Blind image quality assessment by relative gradient statistics and adaboosting neural network," *Signal Process., Image Commun*., vol. 40, pp. 1-15, Jan. 2016.

[14] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett*., vol. 20, no. 3, pp. 209-212, Mar. 2013.

[15] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process*., vol. 26, no. 8, pp. 3951-3964, Aug. 2017.

[16] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. Chen, "Blind quality assessment based on pseudo reference image," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2049-2062, Aug. 2018.

[17] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process*., vol. 25, no. 9, pp. 4444-4457, Sept. 2016.

[18] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Trans. Image Process*., vol. 26, no. 11, pp. 5462-5474, Nov. 2017.

[19] Z. Chen, Y. Li, and Y. Zhang, "Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation," *Signal Process*., vol. 146, pp. 66-78, May 2018.

[20] S. Khan and S. S. Channappayya, "Estimating depth-salient edges and its application to stereoscopic image quality assessment," *IEEE Trans. Image Process*., vol. 27, no. 12, pp. 5892-5903, Dec. 2018.

[21] P. Gorley and N. Holliman, "Stereoscopic image quality metrics and compression," in *proc. Stereoscopic Displays and Applications XIX*, vol. 6803, Feb. 2008.

[22] W. Zhou, L. Yu, Y. Zhou, W. Qiu, M. Wu, and T. Luo, "Blind quality estimator for 3D images based on binocular combination and extreme learning machine," *Pattern Recognit*., vol. 71, pp. 207-217, Nov. 2017.

[23] M. J. Chen, C. C. Su, D. K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereoscopic images by modeling binocular rivalry," in *proc. 46th Asilomar Conf. Signals, Syst. Comput. (ACSSC)*, Nov. 2012, pp. 721-725.

[24] G. Jiang, H. Xu, M. Yu, T. Luo, and Y. Zhang, "Stereoscopic image quality assessment by learning non-negative matrix factorization-based color visual characteristics and considering binocular interactions," *J. Vis. Commun. Image Represent*., vol. 46, pp. 269-279, Jul. 2017.

[25] L. Liu, B. Liu, C. C. Su, H. Huang, and A. C. Bovik, "Binocular spatial activity and reverse saliency driven no-reference stereopair quality assessment," *Signal Process., Image Commun*., vol. 58, pp. 287-299, Oct. 2017.

[26] M. Yu, H. Lakshman and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *proc. IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2015, pp. 31-36.

[27] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *proc. Optics and Photonics for Information Processing X*, vol. 9970, Sept. 2016.

[28] Y. Sun, A. Lu and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Process. Lett*., vol. 24, no. 9, pp. 1408-1412, Sept. 2017.

[29] Y. Zhou, M. Yu, H. Ma, H. Shao, and G. Jiang, "Weighted-to-spherically-uniform SSIM objective quality evaluation for panoramic video," in *proc. IEEE Int. Conf. Signal Process*., Aug. 2018, pp. 54-57.

[30] E. Upenik, M. Rerabek, and T. Ebrahimi, "On the performance of objective metrics for omnidirectional visual content," in *proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2017, pp. 1-6.

[31] E. Upenik, and T. Ebrahimi, "Saliency driven perceptual quality metric for omnidirectional visual content," in *proc. IEEE Int. Conf. Image Process*., Sept. 2019, pp. 4335-4339.

[32] H. G. Kim, H. Lim and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Trans. Circuits Syst. Video Technol*., vol. 30, no. 4, pp. 917-928, Apr. 2020.

[33] H. Lim, H. G. Kim, and Y. M. Ra, "VR IQA NET: deep virtual reality image quality assessment using adversarial learning," in *proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Apr. 2018, pp. 6737-6741.

[34] Y. Xia, Y. Wang, and Y. Peng, "Blind panoramic image quality assessment via the asymmetric mechanism of human brain," in *proc. IEEE Int. Conf. Vis. Commun. Image Process.*, Dec. 2019, pp. 1-4.

[35] Y. Yang, G. Jiang, M. Yu, and Y. Qi, "Latitude and binocular perception based blind stereoscopic omnidirectional image quality assessment for VR system," *Signal Process.*, vol. 173, pp. 1-18, Mar. 2020.

[36] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, "MC360IQA: a multi-channel CNN for blind 360-degree image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 64-77, Jan. 2020.

[37] X. Zheng, G. Jiang, M. Yu, and H. Jiang, "Segmented spherical projection-based blind omnidirectional image quality assessment," *IEEE Access*, vol. 8, pp. 31647-31659, 2020.

[38] S. Croci, C. Ozcinar, E. Zerman, J. Cabrera, and A. Smolic, "Voronoi-based objective quality metrics for omnidirectional video," in *proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1-6.

[39] Z. Chen, J. Xu, C. Lin, and W. Zhou, "Stereoscopic omnidirectional image quality assessment based on predictive coding theory," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 103-117, Jan. 2020.

[40] J. Xu, Z. Luo, W. Zhou, and W. Zhang, "Quality assessment of stereoscopic 360-degree images from multi-viewports," in *proc. Picture Coding Symposium*, Nov. 2019, pp. 1-5.

[41] R. G. de A. Azevedo, N. Birkbeck, I. Janatra, B. Adsumilli, and P. Frossard, "A viewport-driven multi-metric fusion approach for 360-degree video quality assessment," in *proc. Int. Conf. Multimedia and Expo (ICME)*, Jul. 2020, pp. 1-6.

[42] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *arXiv preprint*, arXiv: 2002.09140, 2020.

[43] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, "Viewport proposal CNN for 360° video quality assessment," in *proc. IEEE Int. Conf. Computer Vision and Pattern Recognit.*, Jun. 2019, pp. 10169-10178.

[44] R. G. de A. Azevedo, N. Birkbeck, F. De Simone, I. Janatra, B. Adsumilli, and P. Frossard, "Visual distortions in 360° videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2524-2537, Aug. 2020.

[45] J. Xu, C. Lin, W. Zhou, and Z. Chen, "Subjective quality assessment of stereoscopic omnidirectional image," in *proc. Pacific Rim Conference on Multimedia*, Sept. 2018, pp. 589-599.

[46] *Stereoscopic omnidirectional image database from Ningbo University*. [Online]. Available: https://github.com/qyb123/NBU-SOID/

[47] *Stereoscopic omnidirectional videos from MPEG*. [Online]. Available: https://157.159.160.118/mpegcontent/ws-mpegcontent/content/Explorations/360VR/

[48] *Stereoscopic omnidirectional videos from VRQ-TJU*. [Online]. Available: ftp://eeec.tju.edu.cn/VR/

[49] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, 2008.

[50] ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, 2012.

[51] Y. Zhang, Y. Wang, F. Liu, Z. Liu, Y. Li, D. Yang, *et al.*, "Subjective panoramic video quality assessment database for coding applications," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 461-473, Jun. 2018.

[52] N. D. Sidiropoulos, L. D. Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551-3582, Jul. 2017.

[53] G. Jiang, S. Liu, M. Yu, F. Shao, Z. Peng, and F. Chen, "No reference stereo video quality assessment based on motion feature in tensor decomposition domain," *J. Vis. Commun. Image Represent.*, vol. 50, pp. 247-262, Jan. 2018.

[54] J. Yang, K. Sim, B. Jiang, and W. Lu, "No-reference stereoscopic image quality assessment based on hue summation-difference mapping image and binocular joint mutual filtering," *Appl. Optics*, vol. 57, no. 14, pp. 3915-3926, 2018.

[55] R. G. de A. Azevedo, N. Birkbeck, I. Janatra, B. Adsumilli, and P. Frossard, "Subjective and viewport-based objective quality assessment of 360-degree videos," *Proceedings of Image Quality and System Performance XVII*, vol. 17, pp. 1-6, Jan. 2020.

[56] F. Duanmu, Y. Mao, S. Liu, S. Srinivasan, and Y. Wang, "A subjective study of viewer navigation behaviors when watching 360-degree videos on computers," in *proc. Int. Conf. Multimedia and Expo (ICME)*, Jul. 2018, pp. 1-6.

[57] H. Strasburger, I. Rentschler, and M. Juettner, "Peripheral vision and pattern recognition: A review," *Journal of Vision*, vol. 11, no. 5, pp. 1-82, May. 2011.

[58] P. Guo, Q. Shen, Z. Ma, D. J. Brady, and Y. Wang, "Perceptual quality assessment of immersive images considering peripheral vision impact," *arXiv preprint*, arXiv:1802.09065, 2018.

[59] W. Wang, J. Shen, Y. Yu, and K. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 8, pp. 2014-2027, Aug. 2017.

[60] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971-987, Jul. 2002.

[61] C. C. Chang, and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1-27, Apr. 2011.

[62] L. Breiman, "Random forests," *Machine Language*, vol. 45, no. 1, pp. 1-33, Oct. 2001.

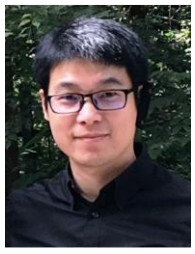
[63] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," Jun. 2000. [Online]. Available: http://www.vqeg.org/.

[64] Y. Ye, E. Alshina, and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360Lib," *in proc. Conf. for Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, Jan. 2017, pp. 1-22.

Yubin Qi received the B.S. degree and M.S. degree from Ningbo University, Ningbo, China, in 2017 and 2020. His current research interest lies in perceptual image processing, high dynamic range imaging, and image quality assessment.

Gangyi Jiang (M'04-SM'19) received the M.S. degree from Hangzhou University, China, in 1992 and the Ph.D. degree from Ajou University, South Korea, in 2000. He is currently a Professor with the Faculty of Information Science and Engineering, Ningbo University, China. He is a senior member of the IEEE. He has authored over 100 technical articles in refereed journals. His research interests mainly include visual communication and image/video processing, 3D video coding, omnidirectional video processing, light field imaging, visual perception and quality assessment.

Mei Yu received the B.S. and M. S. degrees from Hangzhou Institute of Electronics Engineering, China, in 1990 and 1993, and the Ph.D. degree from Ajou University, South Korea, in 2000. She is currently a Professor with the Faculty of Information Science and Engineering, Ningbo University, China. Her research interests mainly include image/video coding and visual perception.



Yun Zhang (M'12-SM'16) received the B.S. and M.S. degrees in electrical engineering from Ningbo University, Ningbo, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2010. From 2009 to 2014, he was a Postdoc Researcher with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. From 2010 to 2017, he was an Assistant Professor and an Associate Professor in Shenzhen Institutes of Advanced Technology (SIAT), CAS, where he is currently a Professor in SIAT, CAS. His research interests are video compression, 3D video processing and visual perception.

Yo-Sung Ho (SM'06–F'16) received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, Korea, in 1981 and 1983, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1990. He joined ETRI (Electronics and Telecommunications Research Institute), Daejeon, Korea, in 1983. From 1990 to 1993, he was with North America Philips Laboratories, Briarcliff Manor, New York, where he was involved in development of the Advanced Digital High-Definition Television (AD-HDTV) system. In 1993, he rejoined the technical staff of ETRI and was involved in development of the Korean DBS Digital Television and High-Definition Television systems. Since 1995, he has been with Gwangju Institute of Science and Technology (GIST), where he is currently Professor of School of Electrical Engineering and Computer Science. Since August 2003, he has been Director of Realistic Broadcasting Research Center at GIST in Korea. He has served as Associate Editors of IEEE Transactions on Multimedia (T-MM) and IEEE Transactions on Circuits and Systems Video Technology (T-CSVT). His research interests include Digital Image and Video Coding, Image Analysis and Image Restoration, Three-dimensional Image Modeling and Representation, Advanced Source Coding Techniques, Augmented Reality (AR) and Virtual Reality (VR), Three dimensional Television (3DTV), and Realistic Broadcasting Technologies. He is a Fellow of IEEE.