TFFN: Three-branch Feature Fusion Network for Stereoscopic Omnidirectional Image Quality Assessment

Yun Liu, *Member, IEEE,* Sifan Li, Daoxin Fan, Huiyu Duan, Peiguang Jing,
Guanghui Yue, *Member, IEEE,* and Guangtao Zhai, *Fellow, IEEE*

***Abstract*—Stereoscopic omnidirectional image (SOI) has both omnidirectional and stereoscopic perception features. Many previous models have proved the viewport characteristics and stereoscopic visual features are crucial for quality perception of SOI. However, effective monocular and binocular visual features extraction and fusion are difficult due to the size of SOI and inaccuracy of feature representation. In this paper, we proposed a three-branch feature fusion network (TFFN) by fusing two-stream binocular visual features and the important monocular features based on the viewport perspective. The hierarchical fusion module is first designed to fuse effective binocular visual features from different semantic scales, and the pseudo-difference information extraction module is built to obtain the accuracy monocular visual features to complement the binocular visual features. Finally, the above monocular and binocular visual features are fused together to measure the quality of SOI. The comparison experiments are conducted on three public datasets and the analysis of the results demonstrate the effectiveness of the proposed method.**

***Index Terms*—Stereoscopic omnidirectional image quality assessment, stereoscopic image, feature fusion, human vision.**

I. Introduction

With the evolution of information technology, omnidirectional content rises to new demands for quality evaluation when they applied to real applications, such as virtual reality (VR) field, autonomous driving and intelligent monitoring field, digitization of cultural heritage filed, and so on. To enjoy an immersive stereoscopic omnidirectional vision feast, stereoscopic omnidirectional image (SOI) is expected to become the main visual contents [1], [2]. Similar to other image formats, it is inevitable to introduce distortion during capturing, compression, transmission and so on, which lead to poor quality perception [3]. To solve the quality evaluation problem, many researchers pay much attention to design the effective image quality assessment (IQA) models [4]–[6]. Different from traditional stereoscopic image, stereoscopic omnidirectional image has wider view range and richer information, which makes stereoscopic omnidirectional image quality assessment (SOIQA) more challenge [7]–[10].

SOIQA can be divided into subjective methods, which is time-consuming and costly, and objective method, which is the convenient in real application [11]–[13]. Among three types of objective methods, full-reference (FR) [1], [14], reduced-reference (RR) and blind/no-reference (NR) methods [2], [15]–[18], since the reference information is not always available in real application, NR method has become the mainstream way [19]–[21]. Since SOIQA is the combination problem of stereoscopic image quality assessment (SIQA) [22] and omnidirectional image quality assessment (OIQA) [23], so most SOIQA models are designed based on them. Observed that the earlier objective SIQA models and OIQA models are built based on 2D IQA metrics, such as peak signal-to-noise ratio (PSNR) [24], structural similarity (SSIM) [25], visual information fidelity (VIF) [26], MS-SSIM [27] and so on, the earlier SOIQA models are follow the above idea, which cannot meet the requirements of the real application. Later, in the NR SIQA area, motivated by the psychophysical and neurophysiological studies, many works focus on stereoscopic perception theory are proposed and present satisfactory performance, which proves the effectiveness of human stereoscopic visual characteristics in SIQA [28]–[32]. Therefore, the stereoscopic perception evaluation in the SOIQA metrics is then designed based on binocular perception characteristics [12], [14], [33], which achieve better results than the models built based on 2D IQA metrics. For the NR OIQA area, with the help of head-mounted display (HMD), people can enjoy the free viewpoint as spherical content. To fit the needs of encoders and decoders, the spherical content must be converted to a 2D plane using projection methods, such as equirectangular projection (EPR) [34], which promote the existence of viewport-based methods [35]–[39]. The success of viewport-based OIQA methods proves the importance of viewport characteristics in OIQA, which further promotes the development of SOIQA [23], [40].

Manuscript received February 17, 2025. This work was supported in part by Shenyang science and technology plan project under Grant 23-407-3-32, in part by Liaoning Province Natural Science Foundation under Grant 2023-MS-139, and in part by National Natural Science Foundation of China under Grant 61901205. *(Corresponding author: Huiyu Duan.)*

Yun Liu, Sifan Li, and Daoxin Fan are with the Faculty of Information, Liaoning University, Shenyang 110036, China (e-mail: yunliu@lnu.edu.cn; sflijohn@foxmail.com; fdx_0729@163.com).

Huiyu Duan and Guangtao Zhai are with the Institute of Image Communication and Information Processing, Shanghai Key Laboratory of Digital Media Processing and Transmissions, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: huiyuduan@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn).

Peiguang Jing is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300000, China (e-mail: pgjing@tju.edu.cn).

Guanghui Yue is with the School of Biomedical Engineering, Shenzhen University, Shenzhen 518000, China (e-mail: yueguanghui@szu.edu.cn).

In all, the above works prove the effectiveness of human stereoscopic visual characteristics in SIQA and the merits of viewport characteristics in OIQA, which provide us with the way to design an effective SOIQA model.

Although many efforts have been made in SOIQA area, facing to varies scenes and new complex distortion types, there is still gap between human subjective perception and objective quality evaluation [3]. There is an urgent need for targeted SOI quality evaluation schemes based on effective feature extraction modules, such as the large language models, feature fusion modules and so on. SOI involves the characteristics of stereoscopic image and omnidirectional image [18], so multiple aspects, such as the stereoscopic visual perception [23], the view format, and feature fusion should be all considered in SOIQA task. Besides, how to combine the above aspects together and extract quality-sensitive information is still a big challenge [41], [42]. Up to now, just a few SOIQA datasets with different compression distortions are built, which makes the SOIQA task more difficult [11], [17]. To solve the above problems, in this paper, we pay much attention not only to accurate visual feature extraction [43], but also effective feature fusion to solve the SOI quality assessment, either symmetrically or asymmetrically distorted [44], [45]. For the stereoscopic visual feature extraction, conventional perspective believes that the binocular rivalry is low-level competition [46], while the predictive coding perspective takes it as the high-level competition [47], [48], which has not reached a unanimous conclusion. It is reasonable to conclude that the binocular rivalry is multi-scale competition, which is in line with the HVS [49]. Based on the above analysis, we propose a three-branch feature fusion network (TFFN) to conduct the quality assessment task. For the viewport characteristics, we not only introduce them in the binocular visual features extraction module, but also in the monocular visual features extraction module, which can help to capture rich quality sensitive visual features. Considering that binocular perception features play an indispensable role in binocular interaction and stereoscopic quality perception, we design a new multi-scale binocular feature extraction and two-stream parallel fusion (TPF) module for stereoscopic perception evaluation in SOIQA model. Besides, we propose a pseudo-difference information extraction (PDIE) module to extract monocular visual features to further augment the prediction accuracy. The main contributions are as follows:

1) The proposed asymmetric binocular difference (BD) module can well strengthen the binocular difference information, which can capture rich stereoscopic binocular difference information hidden in SOI.
2) Two-stream multi-scale aggregation module is designed to capture the binocular perception information, including binocular difference and binocular summation, respectively, which can well dig the different semantic representations and extract the stereoscopic perception interactive information instead of simply fusion.
3) By introducing a larger large language model, we build a PDIE module to accurately extract the representative monocular visual features, which can well reflect the presentative monocular quality degradation and complement the binocular visual features.
4) To strengthen the importance of binocular perception in stereoscopic perception, asymmetric feature fusion (FF) module is also built to fuse the monocular and two binocular visual features, which can reduce the data volume burden of the model and improve the overall performance.

The rest of the contents are list as follows. Section II presents the related works, and Section III is our proposed model. Section IV is the experimental results and Section V is the conclusion.

II. Related Works

The SOI has the characteristics of both SI and OI, so the SIQA and OIQA related works are also introduced in this section.

A. SIQA Models

Many SIQA models have been proposed during the past decades. Considering the two views exist in 3D image, some 2D metric-based models are directly applied on each view and take the average score as the final predicted score. Unfortunately, the above models present poor results. The reason is that 3D image has depth information, which is neglect in previous models. Besides, some distorted stereoscopic image is distorted asymmetrically, and the average calculating way fails to get the accurate result [45]. Considering the existence of depth perception in 3D image, some models focus on extracting the accurate hand-crafted binocular perception features to evaluate the objective quality scores [20], [50]. However, the above models highly rely on the handcraft extracted features, which cannot meet the demand of real application. According to physiological research of human visual system (HVS) [51], monocular visual pathway and binocular visual pathway are found in human brain, which promote the development of SIQA. Shao *et al.* [52] have proved the effectiveness of monocular and binocular visual information in SIQA. Chen *et al.* [53] proposed a cyclopean map to simulate binocular rivalry problems and build an effective SIQA model. Inspired by the binocular fusion and competition in human visual perception, Zhou *et al.* [16] then proposed a multi-layer interactive network model that can predict the weight and quality of left and right view blocks and achieved accurate prediction results. Motivated by the above works, many researchers then focus on building more effective SIQA networks to extract representative and important visual features, which yield a promising performance. From the above, we can find that the existing state-of-the-art SIQA algorithms are both considered monocular and binocular visual information, which presents a clear path to obtaining the stereoscopic perception on OSIQA.

B. OIQA Models

At the earlier stage, 2D FR IQA metrics are directly extended to the OIQA area, such as S-PSNR, CPP-PSNR, the weighted-to-spherically uniform PSNR and spherical domain

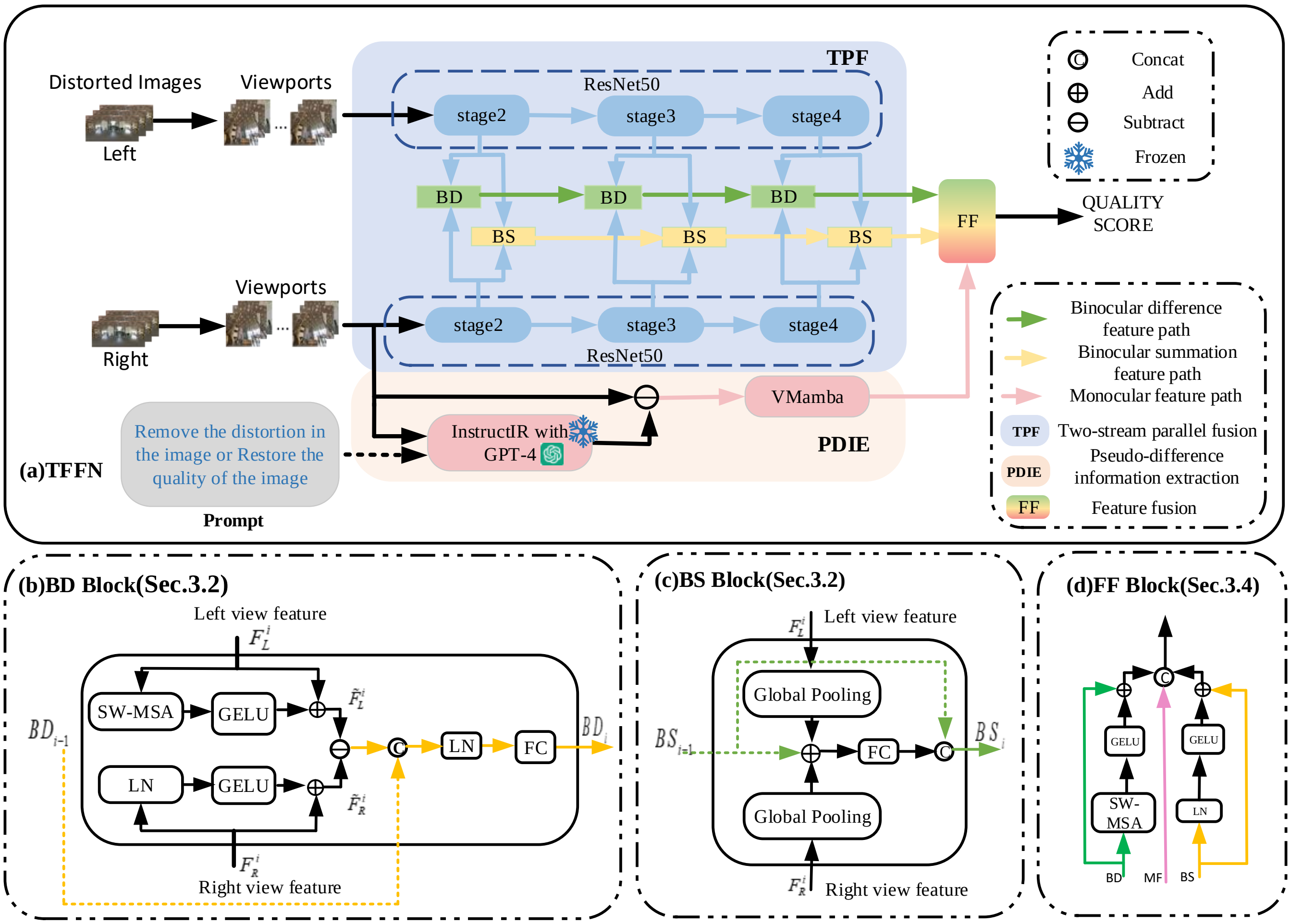

Fig. 1. The overview architecture of our proposed method. We first obtain the viewports of the distorted SOI to extract the binocular difference and binocular summation features. Then a PDIE block is designed to extract monocular features from the monocular visual difference information based on the InstructIR [65]. Finally, three-branch visual features are fused together using the FF block to obtain the quality result.

based SSIM models. Since the above models built based on 2D metrics, they neglect the specific visual features of OI, which limits their performance. Considering the reference is not always available, many NR OIQA models are proposed, which can be categories into three types: whole image-based methods, patch-based methods and viewport-based methods. The whole image-based methods took the whole OI as the input of models to design the OIQA model without considering human visual characteristics, which limits the overall performance [54]. Pay attention to the characteristics of projection methods, the patch-based methods are designed [48], [55]. Liu *et al.* [56] designed a NR OIQA model in equirectangular projection (ERP) format by fusing local and global visual features, which presents a good result. However, the above patch-based methods mainly focus on seeking for a better representation space to obtain more effective features based on the rigid segmented patches, which disturbs the integrity of the semantic meaning in OI that human cares. Then the viewport-based NR OIQA metrics, splitting viewports by simulating human viewing mechanism, are proposed, which presents a better performance than the former two types of models [51], [57], [58]. Jiang *et al.* [48] and Liu *et al.* [59] pay attention to human multi-scale visual characteristics and proposed effective OIQA models, which proves the effectiveness of multi-scale features in OIQA tasks. Later, Sun *et al.* [60] proposed a multi-channel viewport-based model based on the information from six directions' viewports, which motivated us to build a viewport-based feature extraction module to obtain the important visual information. Inspired by human hierarchical perception mechanism, Zhang *et al.* [61] proposed a saliency-guided no-reference method based on saliency-guided multiscale feature fusion, and content perception-based quality regression, which also indicate the importance of multiscale feature fusion on OIQA field.

C. SOIQA Models

Up to now, SOIQA is still in its early stages [17], and the existing SOIQA metrics are designed based on binocular perception characteristics, visual features, or the combination of them. Based on a predictive coding theory, Chen *et al.* [17] took both 3D visual perception and viewports features into consideration and conducted the quality evaluation. Based on the latitude characteristics of omnidirectional image and binocular perception of human visual system, Yang *et al.* [14] propose a blind SOIQA method, which further proves the important of viewports features and binocular perception. Considering the demanding of immersive stereoscopic perception, Yang *et al.* [30] proposed an SOIQA model by extraction the binocular subtraction information and the spherical image

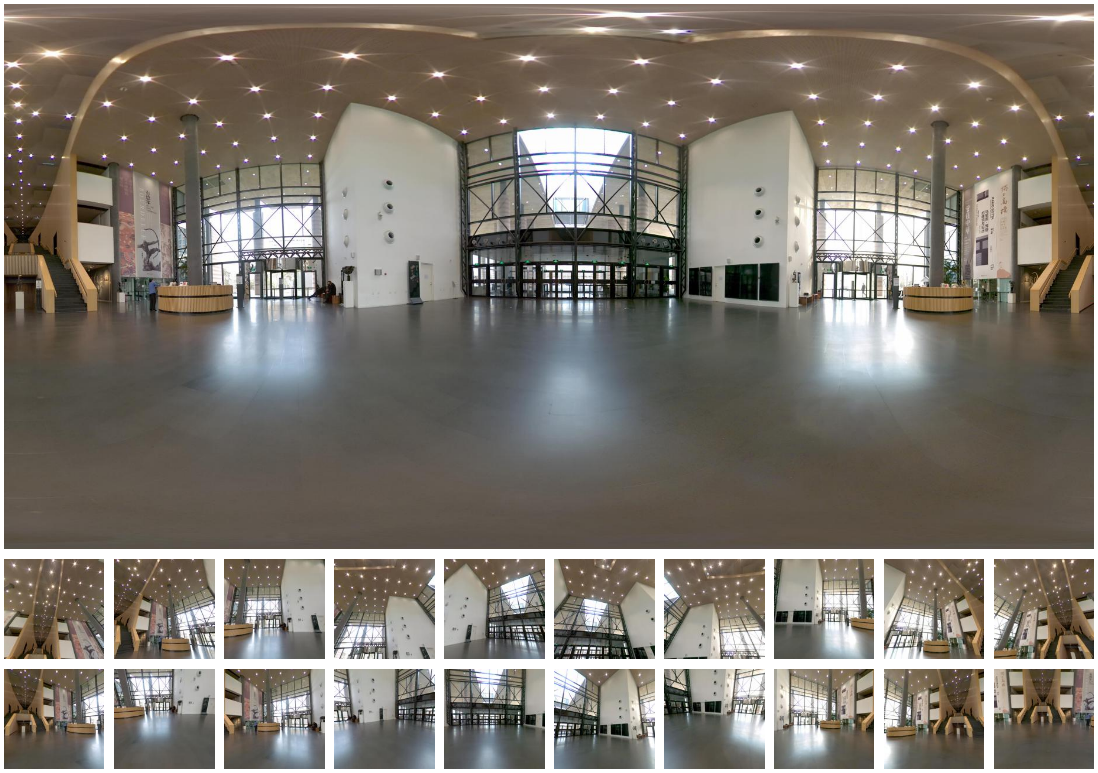

Fig. 2. The left view of one SOI and its viewports.

features. To deal with user view behavior and stereopsis, Qi *et al.* [62] proposed a SOIQA model by fusing the binocular visual feature and viewports features, which yields a superior performance. Focus on the most common format, the EPR format image, Chai *et al.* [63] introduce a deformable convolution to solve the sampling problem and adopt three channel networks to deal with the stereoscopic visual features. Wan *et al.* [64] proposed a hypergraph convolutional network to deal with the binocular visual features. In short, view behavior, viewport format and stereopsis should be all considered in SOIQA tasks, which provide us with a clear way to design effective SOIQA models.

III. Proposed Method

In this section, the proposed TFFN model is introduced in detail, and the architecture of our proposed method is presented in Fig. 1. We build a parallel aggregation structure, the TPF block, to extract and fuse the binocular difference and binocular summation features from the left viewport and right viewport and use a PDIE block to extract monocular features from the monocular visual difference information based on the InstructIR [65] which utilizes GPT-4. Finally, three-branch visual features are fused together using the FF block to obtain the quality result. In the following sections, we first introduce the TPF block, then the PDIE block and finally introduce the quality regression module through the FF block.

A. Viewport Splitting

Considering that the viewports is the basic area while human watches a distorted stereoscopic omnidirectional image (SOI) we first split the left view and the right view of the SOI into several viewports by using the work in [66], respectively, which can also provide rich local visual information than an entire image. Take the left view of one SOI as an example, the left view is first conducted the keypoint selection based on the SURF operation, and the keypoint map is captured by annotating each keypoint in the empty left view image. Then the heatmap is obtained by convoluting the keypoint map with a 2D Gaussian Filter. Finally, the viewpoint selection algorithm is adopted to calculate the central points of each viewport based on the heatmap, and 20 viewports are sampled. More specific details can be found in work [66]. Fig. 2 presents an example of an entire left view of one SOI and its viewports.

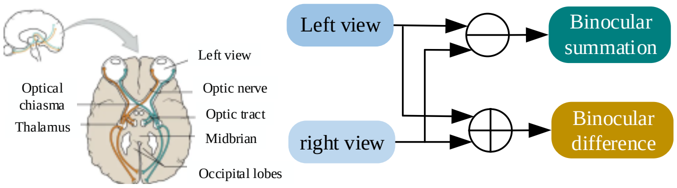

Fig. 3. Human binocular perception mechanism simulation diagram.

B. TPF Block

The mature mechanism research on human binocular perception indicated that there exists a double channel in human brain to adjust the visual parallax by neurons, which are named as binocular summation channel and difference channel [47], [67], shown in Fig. 3. The signal in summation channel indicates the fusion information, while the signal in difference channel reflects the disparity information. Many SIQA works proposed based on the above visual mechanism present a better performance than neglecting the double channel mechanism, which indicates that the double channel mechanism can well

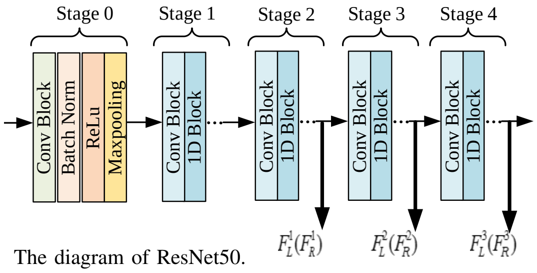

Fig. 4. The diagram of ResNet50.

simulate human binocular perception while watching the 3D content. Inspired by the above works, we apply the double channel model to extract binocular visual features based on the TPF block, in which ResNet50 [68], consists of five stages (stage 0 to 4), is the backbone for the feature extraction. The hierarchical feature fusion block can adaptively fuse binocular visual features from different layers to capture rich global representations and semantic information. Specifically, the left and right viewports are fed into the ResNet50 to obtain the multi-scale visual features from the left and right viewports, which lays the foundation for different scales binocular feature fusion. As shown in Fig. 4, the outputs of stage 2, 3, 4 are fed to the binocular difference fusion module, namely BD shown in Fig. 1(b), and binocular summation fusion module, namely BS shown in Fig. 1(c), to capture the binocular visual features and the binocular interactive relationship, and achieve effectiveness multi-scale interactive fusion from shallow to deep level instead of simply fusion the last stage of ResNet50. For the binocular difference fusion module, to strengthen the difference between the left view and right view and capture rich stereoscopic binocular difference information hidden in SOI, we only introduce the Shifted-Windows Multi-Head Self-Attention (SW-MSA) [69], [70] to the left view feature with the GELU activation function to reduce the amount of computation. The right view feature goes through Layer Normalization (LayerNorm) operation and the GELU activation function. A residual connection is applied to each view feature to dig the semantic information hidden in the viewport, and the binocular difference features can be obtained based on the subtraction operation. Then, the binocular difference features are fused with the previous hierarchical binocular difference features with LayerNorm operation and fully connected layer to increase the representative capability of the binocular difference features. Among them, F_L^i and F_R^i are the feature matrices of the left viewport features and right viewport features produced by ResNet50, respectively. BD_{i-1} denotes the features generated by the previous stage of BD, and BS_{i-1} denotes the features generated by the previous stage of BS. The multi-scale binocular difference feature extraction operation is depicted as follows:

$$\widetilde{F}_L^i = F_L^i \oplus \text{GELU}(\text{SWMSA}(F_L^i)), \tag{1}$$

$$\widetilde{F}_R^i = F_R^i \oplus \text{GELU}(\text{LN}(F_R^i)), \tag{2}$$

$$BD_i = \text{FC}(\text{LN}(Concat(BD_{i-1}, (\widetilde{F}_L^i \ominus \widetilde{F}_R^i)), \tag{3}$$

where GELU$(\cdot)$ means the GELU function, and SWMSA$(\cdot)$ means the Shifted-Windows Multi-Head Self-Attention. LN$(\cdot)$ is the LayerNorm operation. $\oplus$ denotes element-wise addition, and $\ominus$ means the element-wise subtraction operation. $Concat(\cdot)$ is the concatenation operation.

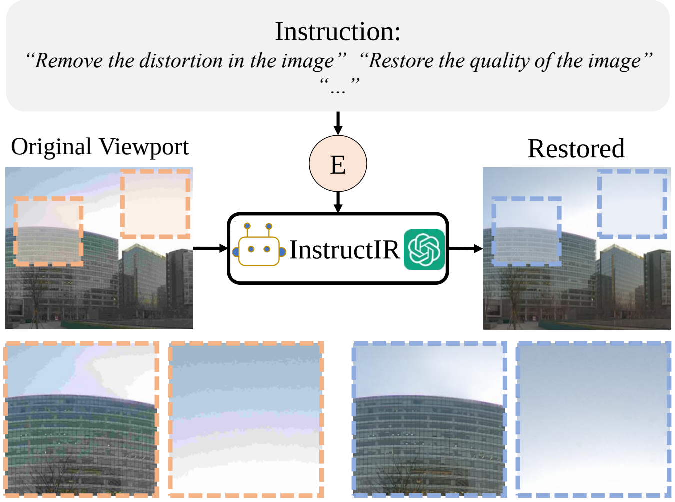
Instruction:
"Remove the distortion in the image" "Restore the quality of the image"
"..."
E
Original Viewport
InstructIR
Restored

Fig. 5. The process of restoring a distorted right viewport with textual prompt by using InstructIR [65]. E represents the embedding process.

For the binocular summation fusion module, the left viewport features and right viewport features go through the global pooling to further reduce the amount of computation, and then add with the previous hierarchical binocular summation features together to obtain the binocular summation features with the fully connected layer. To strengthen the hierarchical binocular summation features fusion, the previous hierarchical binocular summation features are fused with the output of the fully connected layer to increase the representative capability of the binocular summation features. The multi-scale binocular summation feature extraction operation is depicted as follows:

$$BS_i = Concat(BS_{i-1}, \text{FC}(\text{g}(\widetilde{F}_L^i) \oplus \text{g}(\widetilde{F}_R^i))), \tag{4}$$

where g$(\cdot)$ means global pooling operation. After the hierarchical processing, the two-parallel binocular visual features of all viewports are concatenated together as the final binocular visual feature of the distorted SOI, including binocular difference features and binocular summation features.

C. PDIE Block

Rather than binocular perception, monocular visual information focuses on the content perception instead of stereoscopic perception, which provides more information of content degradation and can well complement the binocular visual information [71]. According to work [72], the overall perception quality of stereoscopic image degraded when the single view introduces distortions, which has been proved in many SIQA works. Herein, we design a PDIE block to dig the representative monocular information while people view a stereoscopic image. It needs to mention that the most precise way to extract quality-sensitive monocular information is to capture the quality degradation information from the difference between the distorted image and its reference image. However, the reference image is not always available in real applications. To solve the above challenge, we introduce a novel restoration model, InstructIR [65], to generate the pseudo-reference

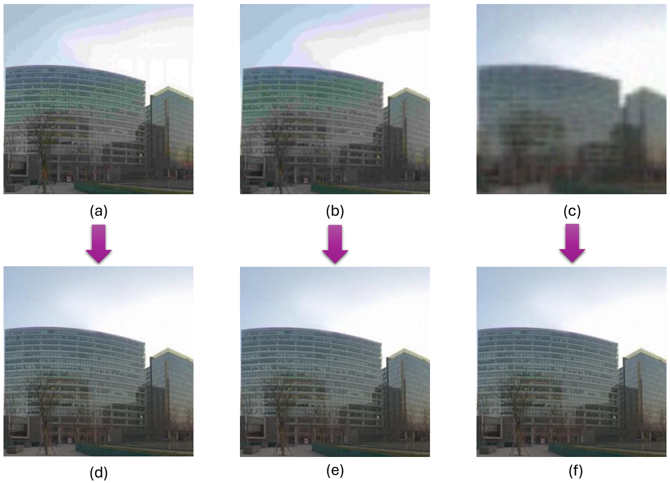

Fig. 6. The distorted viewports and the paired restored viewports. (a), (b) and (c) are the viewports of the SOI with different distortions of JPEG compression in two levels and gaussian blur, respectively. (d), (e) and (f) are the restored viewports of (a), (b) and (c), respectively.

image, which is designed based on the large language model GPT-4, shown in Fig. 5. InstructIR can restore the quality degradation occurred by different distortions based on the simple text prompts, such as “Remove the distortion in the image” or “Restore the quality of the image”. It needs to mention that InstructIR relies on the semantic content within the textual prompt, so similar textual prompt also works. The image restoring performance has been proved in its original paper, and more details can be found in work [65].

By feeding the distorted viewport and the text prompt into InstructIR, the paired pseudo-reference viewport can be got. Fig. 6 presents an example of the distorted viewport and its paired pseudo-reference viewport. As we can see, the quality of the pseudo-reference viewport is improved compared to the distorted viewport, which can be applied to dig the monocular quality degradation information.

Given a set of N viewports, the difference map between the i-th distorted viewport and the i-th pseudo-reference viewport can be obtained based on the element-wise subtraction, as follows:

$$MD_i = I_{D_i} \ominus I_{PR_j}, (i \in [1, 2, ..., N]), \tag{5}$$

where MD_i means the difference map of the i-th viewport. I_{D_i} and I_{PR_j} are the i-th distorted viewport and the i-th pseudo-reference viewport, respectively. Inspired by the effectiveness of VMamba [73] in various visual tasks, we take VMamba as the backbone to extract the monocular feature. So here, the difference maps of all viewports are concatenated together and then fed into VMamba module to extract the monocular features MF:

$$MF = \text{VMamba}(Concat(MD_i)). \tag{6}$$

It needs to mention that considering that the left view and the right view are similar, only the right view undergoes the PDIE block, which is enough to capture the monocular quality degradation information and also can reduce the computational complexity.

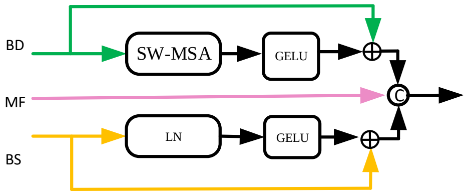

Fig. 7. The architecture of the feature fusion (FF) block.

D. FF Block

To effectively adapt and fuse the above three-branch features, we propose the FF block, shown in Fig. 7. Considering the importance of binocular perception in stereoscopic perception, binocular summation features and binocular difference features go through the activation function, while the monocular features are simply concatenated with the binocular features. Moreover, the binocular difference feature is the key information that humans perceive the stereopsis perception, so we introduce the SW-MSA to further strengthen the binocular difference feature. While the binocular summation features go through the LayerNorm operation to reduce the amount of computation, which is compute as follows:

$$\widetilde{BD} = BD \oplus \text{GELU}(\text{SWMSA}(BD)), \tag{7}$$

$$\widetilde{BS} = BS \oplus \text{GELU}(\text{LN}(BS)), \tag{8}$$

then all the three branches of features mentioned above are fused together to obtain the overall visual feature F, which is formulated as follows:

$$F = Concat(\widetilde{BD}, MF, \widetilde{BS}). \tag{9}$$

E. Quality Score Regression

Finally, the overall visual features are used to map the final quality score Q through the LayerNorm operation and the fully connected layer:

$$Q = \text{FC}(\text{LN}(F)). \tag{10}$$

Here, we employ Euclidean loss to optimize the parameters of our model, which is defined as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} ||s_i - q_i||_2^2, \tag{11}$$

where N represents the total number of samples, s_i is the ground-truth score of the i-th sample, and q_i represents the quality score predicted by the model for the i-th sample.

IV. Experimental Results and Analysis

In this section, we conduct experiments on one publicly available stereoscopic omnidirectional image quality assessment dataset, SOLID dataset [1], and two stereoscopic image quality assessment datasets, LIVE Phase I [74] and II [53], [75], to present the performance of our proposed method. Moreover, ablation experiments are conducted to verify the contributions of important components in our proposed method and demonstrate its reasonableness.

A. Experimental Settings

1) Datasets: SOLID dataset [1]: It consists of 276 distorted images, which are derived from six high-quality reference images with JPEG and BPG compression. The distorted images include 84 symmetrically distorted ones and 192 asymmetrically distorted ones. Each of them is paired with a human subjective evaluation score, Mean Opinion Score (MOS) value, ranging from 1 to 5, and a higher value means a better quality.

LIVE 3D VR dataset [76], [77]: It consists of 450 distorted images along with DMOS values, ranges from 1 to 100, and lower DMOS means better quality. The distorted images are generated based on 15 reference images with 6 different distortions and 5 levels of distortion: Gaussian blur, Gaussian noise, downsampling, stitching, VP9 compression, and H.265 compression.

LIVE Phase I dataset [74]: This dataset consists of 365 distorted images from 20 scenes, which are distorted with five different distortion types: Gaussian blurring, white noise, JPEG compression, JPEG2000 compression and fast fading for simulating packet loss of transmitted JPEG2000-compressed images. The associated Differential Mean Opinion Score (DMOS) ranged from 0 to 80 is provided with the dataset, representing the human subjective judgements. Higher DMOS means worse quality.

LIVE Phase II dataset [53], [75]: This dataset consists of 360 stereoscopic images from 8 scenes. Each scene contains five types of distortions which are the same as those in LIVE Phase I [74]. Each type of distortion contains 3 symmetrically distorted and 6 asymmetrically distorted images. The DMOS of each image is provided with the dataset.

2) Criteria: Based on the common methods in the image quality assessment area, we apply three prevalent criteria to evaluate the performance of our proposed method and other state-of-the-art methods. The three criteria are Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC) and Root Mean Squared Error (RMSE). The formulae are depicted as follows:

$$\text{PLCC} = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2 \sum_{i=1}^{N}(p_i - \bar{p})^2}}, \tag{12}$$

where N denotes the number of the images. s_i is the MOS of the i-th image, and p_i is the prediction score. $\bar{s}$ is the mean value of the MOS's, and $\bar{p}$ is the mean value of the score that the model predicted for each image. PLCC ranges from 0 to 1, and the higher the value is, the better performance is.

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{13}$$

where d_i denotes the distance between the rank of the MOS and the rank of the prediction score given by the model for the i-th image. SRCC value ranges from -1 to 1. The higher the value is, the better performance is.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - p_i)^2}, \tag{14}$$

where s_i is the MOS of the i-th image and p_i is the prediction score given by the model. The lower the value is, the better performance is.

3) Implementation Details: The dataset is split into a training set and a testing set by a ratio of 8:2, following the commonly used standard [78]. The SGD with momentum optimizer is used with the momentum parameter initially set to 0.9, and the weight decay parameter is set to 10^{-4}. The batch size is set to 32 while the initial learning rate is set to 10^{-3}. The entire experimental process is implemented on a machine configured with one NVIDIA GeForce RTX 4090 graphic card with 24GB VRAM.

B. Experimental Results and Analysis

1) Performance on Benchmark Datasets: To verify the advanced performance of our proposed method, many most cited and advanced state-of-the-art metrics of FR and NR methods are applied to make the comparison. The FR metrics for comparing include six 2D IQA-based metrics (PSNR, IW-PSNR [79], SSIM [25], MS-SSIM [27], FSIM [80] and VSI [81]), three 2D OIQA-based metrics (S-PSNR [82], WS-PSNR [83] and CPP-PSNR [84]), four 3D IQA metrics (Chen *et al.* [53], W-SSIM [44], W-FSIM [44] and SINQ [85]), and five 3D OIQA metric (BSOIQA [86], VP-BSOIQA [62], Chai *et al.* [63] , SOIQE [17] and Wan *et al.* [64]).

The results on two SOI datasets, SOLID Dataset and LIVE 3D VR Dataset, are listed in Table I, and the best metrics are highlighted in boldface. It can be seen that our proposed method performs the best among all the compared methods, which demonstrates that our method is reasonable and effective. Specifically, 2D OIQA models yield the worst performance among all types of IQA models on SOLID Dataset, and 2D IQA models present the worst performance on LIVE 3D VR Dataset. Since conventional 2D OIQA models are designed specifically for 2D OI and not for 3D OI, and conventional 2D IQA models are designed specifically for 2D image not for OI, 2D OIQA models fail to capture the special visual stereoscopic perception information in 3D OI, and 2D IQA models fail to capture the special visual perception information in OI. The aforementioned results indicate the importance of combing panoramic information and binocular visual features in SOIQA area, which is one of our contributions. Considering Fovea visual characteristics and binocular rivalry visual characteristics, SOIQE proposed an effective predictive coding model, which further proves that combining binocular features and panoramic information is reasonable. Although SOIQE provides promising performance on SOLID Dataset, it fails to predict the SOI quality on LIVE 3D VR Dataset. This may be because the interactive information hidden in binocular visual perception procedure is neglected in their model, which proves the importance of designing effective feature fusion modules. For the Wan et al. [64] model, it presents a competition performance on 3D VR Dataset, it fails to effectively predict the SOI quality on SOLID Dataset. In contrast, our method ranks the first among all the models on both two datasets, which indicates the reasonable and effectiveness of our model. Besides, it can be seen that all

TABLE I
PERFORMANCE COMPARISON ON SOLID DATASET AND LIVE 3D VR DATASET

Type	Method	SOLID			LIVE 3D VR		
		PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓
2D IQA	PSNR [24]	0.629	0.603	0.789	0.526	0.536	9.701
	IW-PSNR [79]	0.773	0.751	0.643	0.535	0.532	9.683
	SSIM [25]	0.882	0.888	0.478	0.633	0.658	8.592
	MS-SSIM [27]	0.773	0.755	0.643	0.607	0.596	9.165
	FSIM [80]	0.889	0.883	0.465	0.613	0.587	9.362
	VSI [81]	0.881	0.873	0.479	0.626	0.599	9.357
2D OIQA	S-PSNR [82]	0.593	0.567	0.816	0.662	0.625	8.342
	WS-PSNR [83]	0.585	0.559	0.823	0.636	0.625	8.603
	CPP-PSNR [84]	0.593	0.566	0.817	0.627	0.630	8.702
3D IQA	Chen *et al.* [53]	0.853	0.827	0.530	0.775	0.753	6.806
	W-SSIM [44]	0.893	0.891	0.457	0.696	0.633	8.965
	W-FSIM [44]	0.889	0.885	0.464	0.703	0.684	8.301
	SINQ [85]	0.810	0.779	0.460	0.802	0.799	5.925
3D OIQA	BSOIQA [86]	0.790	0.762	0.480	0.740	0.738	7.024
	VP-BSOIQA [62]	0.853	0.842	0.411	0.796	0.801	6.848
	Chai-SOIQE [63]	0.879	0.872	0.372	0.862	0.863	5.688
	SOIQE [17]	0.927	0.924	0.383	0.682	0.660	8.229
	Wan *et al.* [64]	0.884	0.890	0.382	0.878	0.870	4.915
	TFFN (**ours**)	**0.965**	**0.963**	**0.269**	**0.886**	**0.891**	**4.347**

TABLE II
PERFORMANCE COMPARISON ON 3D IQA DATABASE LIVE PHASE I AND II

Type	Method	LIVE Phase I			LIVE Phase II		
		PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓
2D IQA	PSNR [24]	0.864	0.855	8.242	0.658	0.637	8.495
	SSIM [25]	0.869	0.860	8.087	0.684	0.679	8.229
	MS-SSIM [27]	0.882	0.894	7.710	0.727	0.724	7.740
3D IQA	You *et al.* [87]	0.830	0.814	7.746	0.800	0.786	6.772
	Benoit *et al.* [28]	0.881	0.878	7.061	0.748	0.728	7.490
	Hewage *et al.* [88]	0.902	0.899	9.139	0.558	0.501	9.364
	Bensalma *et al.* [89]	0.887	0.875	7.559	0.770	0.751	7.204
	Chen *et al.* [53]	0.895	0.891	7.247	0.895	0.880	5.102
3D OIQA	SOIQE [17]	0.920	0.917	6.266	0.915	0.907	4.544
	TFFN (**ours**)	**0.938**	**0.939**	**5.094**	**0.932**	**0.933**	**4.536**

3D OIQA models have better performance on SOLID Dataset than that on LIVE 3D VR Dataset, the probable reason is that LIVE 3D VR Dataset has more complexity distortions and more varied scenes than SOLID Dataset, which makes quality assessment on SOLID Dataset challenge. Our method solves the above challenges by not only considering the binocular summation features and binocular difference features but also digging the interactive relationship between them through an effective aggregation module, which makes our model more suitable for real application with complexity and wide range of distortions.

To validate the ability of the model on evaluating the stereoscopic perception quality of traditional 3D images, we verify the validity of our model on two public 3D image datasets, LIVE Phase I [74] and II [53], [75]. The whole data processing, training and testing procedure on 3D image are the same as that on SOI, and the results are presented in Table II, where the best performing metric result is highlighted in boldface. It can be seen that 2D-based models have limited performance, and You's model [87] presents the worst performance among all of the models, which proves that simply applying 2D IQA algorithms to each view directly to get the final quality score is unfeasible. Although Benoit *et al.* [28] took the depth information into consideration, they fail to achieve a promising result without considering the features of human binocular visual process. While Chen *et al.* [53] proposed FR and NR SIQA models by considering binocular perception characteristics, the performance of each of the models was further improved, which proves the importance of binocular visual features in the quality assessment area. The SOIQE model, focusing on binocular features and content information, presents potential performance in the SIQA area. However, they neglect the two binocular channels in human brains while watching a stereoscopic image. In contrast, our method not only focus on two binocular visual features, binocular summation and binocular difference features, but also the monocular content degradation information, and achieves the best performance on all criteria on LIVE Phase I and II, which demonstrates the superiority and the potential practical application of our model. Overall, the results presented above prove the effectiveness and generalization ability of our model, which can be feasibly applied to evaluate the quality of SOIs.

2) Performance on Different Distortion Types: To further prove the practical effectiveness of our method, we also conduct experiments on each type of distortion on SOLID [1] database. Table III presents the results of our model on JPEG images and BPG images, and Table IV concludes the results on symmetrically or asymmetrically distorted images. The best performing metric is highlighted in boldface. From Table III, we can be seen that the performances of all the models

TABLE III
PERFORMANCE COMPARISON FOR DIFFERENT DISTORTION TYPES ON SOLID DATABASE

Method	PLCC ↑		SRCC ↑		RMSE ↓	
	JPEG	BPG	JPEG	BPG	JPEG	BPG
PSNR [24]	0.564	0.740	0.538	0.673	0.901	0.624
SSIM [25]	0.907	0.857	0.893	0.879	0.460	0.477
MS-SSIM [27]	0.841	0.730	0.833	0.687	0.591	0.633
FSIM [80]	0.894	0.896	0.880	0.902	0.490	0.411
VSI [81]	0.898	0.888	0.885	0.886	0.480	0.426
S-PSNR [82]	0.515	0.736	0.477	0.660	0.936	0.627
WS-PSNR [83]	0.505	0.732	0.464	0.658	0.949	0.631
CPP-PSNR [84]	0.517	0.735	0.475	0.660	0.934	0.628
Chen *et al.* [53]	0.909	0.797	0.904	0.736	0.454	0.559
W-SSIM [44]	0.905	0.887	0.888	0.879	0.464	0.428
W-FSIM [44]	0.893	0.933	0.885	0.933	0.492	0.333
SINQ [85]	0.819	0.821	0.790	0.789	0.499	0.377
BSOIQA [86]	0.880	0.723	0.889	0.774	0.437	0.429
VP-BSOIQA [62]	0.883	0.861	0.852	0.856	0.551	0.373
Chai-SOIQE [63]	0.898	0.891	0.882	0.885	0.310	0.378
SOIQE [17]	0.933	0.955	0.928	0.939	0.393	**0.275**
Wan *et al.* [64]	0.872	0.848	0.865	0.844	0.446	0.447
TFFN (**ours**)	**0.956**	**0.960**	**0.952**	**0.950**	**0.299**	0.287

TABLE IV
PERFORMANCE COMPARISON FOR SYMMETRICALLY AND ASYMMETRICALLY DISTORTED IMAGES ON SOLID DATABASE

Method	PLCC ↑		SRCC ↑		RMSE ↓	
	Sym	Asym	Sym	Asym	Sym	Asym
PSNR [24]	0.791	0.394	0.789	0.354	0.758	0.756
SSIM [25]	0.944	0.821	0.902	0.814	0.409	0.470
MS-SSIM [27]	0.869	0.631	0.836	0.615	0.613	0.638
FSIM [80]	0.930	0.853	0.890	0.847	0.456	0.430
VSI [81]	0.931	0.834	0.887	0.807	0.454	0.454
S-PSNR [82]	0.805	0.364	0.766	0.313	0.735	0.766
WS-PSNR [83]	0.807	0.325	0.762	0.302	0.732	0.778
CPP-PSNR [84]	0.806	0.334	0.766	0.310	0.734	0.775
Chen *et al.* [53]	0.944	0.767	0.890	0.700	0.411	0.528
W-SSIM [44]	0.944	0.834	0.902	0.832	0.409	0.454
W-FSIM [44]	0.930	0.845	0.890	0.842	0.456	0.440
SINQ [85]	0.847	0.766	0.792	0.745	0.365	0.493
BSOIQA [86]	0.887	0.720	0.893	0.692	0.455	0.437
VP-BSOIQA [62]	0.894	0.831	0.853	0.776	0.396	0.414
Chai-SOIQE [63]	0.887	0.848	0.885	0.832	0.337	0.433
SOIQE [17]	0.970	0.867	0.931	0.866	0.301	0.411
Wan *et al.* [64]	0.885	0.873	0.872	0.859	0.423	0.412
TFFN (**ours**)	**0.973**	**0.902**	**0.952**	**0.903**	**0.285**	**0.384**

present a similar trend with the performance on the overall datasets in Table I, which further proves the effectiveness of our method. Table IV indicates that the performance on symmetrically distorted images is better than that on asymmetrically distorted images. It is harder to extract the difference between binocular views, which makes the quality assessment of asymmetrically distorted images difficult. Considering the asymmetrically distorted image is common in real application, recent methods have made progress in tackling this problem. Among them, SOIQE indicates an impressive performance, but they ignore the interaction between binocular visual features and monocular features. Our model takes both comprehensive stereoscopic visual features and the feature aggregation into consideration and yields the best overall performance on different distortion types of distorted images, which further proves that our method is reasonable and effective.

3) Ablation Experiments: To verify the contribution of each component in our method, including the monocular visual features, binocular summation features, binocular difference features, the hierarchical feature fusion module, and the restoration pseudo-reference module, the ablation experiments are conducted. The results are presented in Table V, and the best performing metric is highlighted in boldface. Experiment (1), named "TFFN (ours)", is our proposed model. Experiment (2), named "w/o TPF-BD", denotes the model deleting the binocular difference features, Experiment (3), named "w/o TPF-BS", denotes the model deleting the binocular summation features, and Experiment (4), named "w/o PDIE", denotes the model deleting the monocular visual features. Experiment (5), named "FF replaced by simple Concat", denotes the model using concatenate operation to fuse all the extracted features instead of using the proposed FF feature fusion model. Experiment (6), named "Double-Side SW-MSA", denotes that the model adopts a double-side BD module, which not only introduces the Shifted-Windows Multi-Head Self-Attention to the left view features, but also to the right view features.

As we can see that Experiment (2) and Experiment (3) both present worse performance than our model on two SOIQA datasets, which proves that the binocular difference information and binocular summation information both play an important role in SOIQA area. Experiment (5) ("FF replaced by simple Concat") presents the worst performance than our model, which proves that simply using the concatenation operation fails to capture representative quality-sensitive visual information. Experiment (6) ("Double-Side SW-MSA") is set to prove the effectiveness of the single-side SW-MSA in BD module. It can be seen that the "Double-Side SW-MSA" model has worse performance than our proposed model, which proves the effectiveness of the single-side SW-MSA module proposed in our work. Although Experiment (4) presents a promising performance, our model yields the best performance among all the models, which proves that the monocular visual features can well complement the binocular visual features and further improve the accuracy of the model. All the results of the ablation experiments verify the contributions of the important components in our proposed method and demonstrate our proposed method is effective and reasonable. Based on the analysis above, it can be concluded that our method can be applied to SOIQA tasks effectively and achieve an advanced performance.

V. CONCLUSION

In this paper, we propose a new blind stereoscopic omnidirectional image quality assessment method considering the visual features of the binocular viewport and the visual features of the monocular viewport. Binocular difference information and binocular summation information are hierarchically extracted and fused to capture rich binocular semantic information, which can delve into representative stereoscopic perception. The monocular visual features are extracted on the basis of pseudo-difference information, which can well complement the binocular visual feature and provide accuracy quality-degradation information. Extensive experiments prove the effectiveness and reasonability of our method. In the future, we will pay attention to design effective quality assessment

TABLE V
THE RESULTS OF THE ABLATION EXPERIMENTS

No.	Method	SOLID			LIVE 3D VR		
		PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓
(1)	TFFN (**ours**)	**0.965**	**0.963**	**0.269**	**0.886**	**0.891**	**4.347**
(2)	w/o TPF-BD	0.921	0.926	0.401	0.792	0.803	5.668
(3)	w/o TPF-BS	0.930	0.929	0.395	0.802	0.798	5.728
(4)	w/o PDIE	0.942	0.947	0.329	0.845	0.850	4.982
(5)	FF replaced by simple $Concat$	0.922	0.916	0.417	0.804	0.795	5.123
(6)	Double-Side SW-MSA	0.923	0.918	0.376	0.798	0.794	4.981

methods for stereoscopic omnidirectional videos based on extracting more essential features with a lower computational budget and hierarchical multi-scale feature fusion manner.

REFERENCES

[1] J. Xu, C. Lin, W. Zhou, and Z. Chen, "Subjective quality assessment of stereoscopic omnidirectional image," in *Advances in Multimedia Information Processing, PCM 2018*, Springer International Publishing, 2018, pp. 589–599.

[2] W. Zhou and Z. Wang, "Perceptual depth quality assessment of stereoscopic omnidirectional images," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[3] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[4] Z. Zhou, F. Zhou, and G. Qiu, "Blind Image Quality Assessment Based on Separate Representations and Adaptive Interaction of Content and Distortion," *IEEE Transactions on Circuits and Systems for Video Technology*. vol. 34, pp. 2484-2497, 2024.

[5] Q. Qu, H. Liang, X. Chen, Y. Chung, and Y. Shen, "NeRF-NQA: No-Reference Quality Assessment for Scenes Generated by NeRF and Neural View Synthesis Methods," *IEEE Transactions on Visualization and Computer Graphics*. vol. 30, pp. 2129-2139, 2024.

[6] Z. Chen, X. Zhang, W. Li, R. Pei, F. Song, X. Min, X. Liu, X. Yuan, Y. Guo, and Y. Zhang, "Grounding-iqa: Multimodal language grounding model for image quality assessment," 2024.

[7] L. Cao, G. Jiang, Z. Jiang, M. Yu, Y. Qi, and Y.-S. Ho, "Quality measurement for high dynamic range omnidirectional image systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2021.

[8] H. Ai, Z. Cao, J. Zhu, H. Bai, Y. Chen, and L. Wang, "Deep learning for omnidirectional vision: A survey and new perspectives,", 2022.

[9] G. Pintore, A. Jaspe-Villanueva, M. Hadwiger, J. Schneider, M. Agus, F. Marton, F. Bettio, and E. Gobbetti, "Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image," *Computers & Graphics*, vol. 119, pp. 103907, 2024.

[10] W. Zhou and Z. Wang, "Perceptual depth quality assessment of stereoscopic omnidirectional images," 2024.

[11] J. Xu, Z. Luo, W. Zhou, W. Zhang, and Z. Chen, "Quality assessment of stereoscopic 360-degree images from multi-viewports,", 2019.

[12] X. Zhou, Y. Zhang, N. Li, X. Wang, Y. Zhou, and Y.-S. Ho, "Projection invariant feature and visual saliency-based stereoscopic omnidirectional image quality assessment," *IEEE Transactions on Broadcasting*, vol. 67, no. 2, pp. 512–523, 2021.

[13] S. Biswas, B. Appina, P. Kokil, and S. S. Channappayya, "Subjective and objective quality assessment methods of stereoscopic videos with visibility affecting distortions," 2024.

[14] Y. Yang, G. Jiang, M. Yu, and Y. Qi, "Latitude and binocular perception based blind stereoscopic omnidirectional image quality assessment for vr system," *Signal Processing*, vol. 173, pp. 107586, 2020.

[15] R. Akhter, Z. M. P. Sazzad, Y. Horita, and J. Baltes, "No-reference stereoscopic image quality assessment," in *Stereoscopic Displays and Applications XXI*, 2010, pp. 75240T.

[16] W. Zhou, Z. Chen, and W. Li, "Dual-stream interactive networks for no-reference stereoscopic image quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3946–3958, 2019.

[17] Z. Chen, J. Xu, C. Lin, and W. Zhou, "Stereoscopic omnidirectional image quality assessment based on predictive coding theory," 2019.

[18] Z. Wan, Q. Yang, Z. Li, X. Fan, W. Zuo, and D. Zhao, "Dual-stream perception-driven blind quality assessment for stereoscopic omnidirectional images," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10431–10439.

[19] X. Chai and F. Shao, "Blind quality assessment of omnidirectional videos using spatio-temporal convolutional neural networks," *Optik*, vol. 226, pp. 165887, 2021.

[20] L. Shen, X. Chen, Z. Pan, K. Fan, F. Li, and J. Lei, "No-reference stereoscopic image quality assessment based on global and local content characteristics," *Neurocomputing*, vol. 424, pp. 132–142, 2021.

[21] Y. Liu, X. Yin, G. Yue, Z. Zheng, J. Jiang, Q. He, and X. Li, "Blind omnidirectional image quality assessment with representative features and viewport oriented statistical features," *Journal of Visual Communication and Image Representation*, vol. 91, pp. 103770, 2023.

[22] H. Zhang, S. Li, H. Chang, and P. Lin, "Towards top-down stereo image quality assessment via stereo attention," 2023.

[23] W. Zhou and Z. Wang, "Blind omnidirectional image quality assessment: Integrating local statistics and global semantics," in *2023 IEEE International Conference on Image Processing*, 2023, pp. 1405–1409.

[24] A. Horé and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *20th International Conference on Pattern Recognition*, 2010, pp. 2366–2369.

[25] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[26] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[27] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402 Vol.2.

[28] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 659024, 2009.

[29] G. Jiang, M. He, M. Yu, F. Shao, and Z. Peng, "Perceptual stereoscopic image quality assessment method with tensor decomposition and manifold learning," *IET Image Processing*, vol. 12, no. 5, pp. 810–818, 2018.

[30] J. Yang, H. Xu, Y. Zhao, H. Liu, and W. Lu, "Stereoscopic image quality assessment combining statistical features and binocular theory," *Pattern Recognition Letters*, vol. 127, pp. 48–55, 2019.

[31] Y. Chang, S. Li, J. Jin, A. Liu, and W. Xiang, "Stereo image quality assessment considering the difference of statistical feature in early visual pathway," *Journal of Visual Communication and Image Representation*, vol. 89, pp. 103643, 2022.

[32] D. Bandhu, M. M. Mohan, N. A. P. Nittala, P. Jadhav, A. Bhadauria, and K. K. Saxena, "Theories of motivation: A comprehensive analysis of human behavior drivers," *Acta Psychologica*, vol. 244, pp. 104177, 2024.

[33] J. You, G. Jiang, H. Jiang, H. Xug, Z. Jiang, Z. Zhu, and M. Yu, "Visual perception-oriented quality assessment for high dynamic range stereoscopic omnidirectional video system," *Displays*, vol. 80, pp. 102515, 2023.

[34] Y. Wang, Z. Chen, and S. Liu, "Equirectangular projection oriented intra prediction for 360-degree video coding," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2020, pp. 483–486.

[35] Y. Zhang, H. Zhang, D. Li, L. Liu, H. Yi, W. Wang, H. Suitoh, and M. Odamaki, "Toward real-world panoramic image enhancement," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 2675–2684.

[36] O.-J. Kwon, J. Cho, and S. Choi, "A center-to-edge progression for equirectangular projected 360° jpeg images," *IEEE Access*, vol. 9, pp. 6921–6929, 2021.

[37] S. S. Drisya, A. Mahapatra, and S. Priyadharshini, "360-degree image classification and viewport prediction using deep neural networks," in *Advances in Distributed Computing and Machine Learning*, Singapore: Springer Singapore, 2022, pp. 483–492.

[38] F. A. Brunetti, "Equirectangular pictures and surrounding visual experience. spherical immersive photographic projections at: Boito architetto archivio digitale, historical exhibition at politecnico di milano," in *ICGG 2022 - Proceedings of the 20th International Conference on Geometry and Graphics*, L.-Y. Cheng, Ed. Cham: Springer International Publishing, 2023, pp. 541–553.

[39] T. Qiu, I. K. Jain, R. Wu, D. Bharadia, and P. Cosman, "Delivering 360-degree video with viewport-adaptive truncation," in *2022 25th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2022, pp. 7–11.

[40] Y. Zhang, L. Wan, D. Liu, X. Zhou, P. An, and C. Shan, "Saliency-guided no-reference omnidirectional image quality assessment via scene content perceiving," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2024.

[41] J. Yang, Y. Zhao, C. Yi, and J. C.-W. Chan, "No-reference hyperspectral image quality assessment via quality-sensitive features learning," *Remote Sensing*, vol. 9, no. 4, 2017.

[42] X. Zhang, Y. Zhang, W. Yu, L. Nie, N. Jiang, and J. Gong, "Qs-hyper: A quality-sensitive hyper network for the no-reference image quality assessment," in *Neural Information Processing*, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds. Cham: Springer International Publishing, 2021, pp. 311–322.

[43] Z. Xiao and S. Li, "A real-time, robust and versatile visual-slam framework based on deep learning networks," 2024.

[44] J. Wang, A. Rehman, K. Zeng, S. Wang, and Z. Wang, "Quality prediction of asymmetrically distorted stereoscopic 3d images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3400–3414, 2015.

[45] Y. Zhang, X. Liu, H. Liu, and C. Fan, "Depth perceptual quality assessment for symmetrically and asymmetrically distorted stereoscopic 3d videos," *Signal Processing: Image Communication*, vol. 78, pp. 293–305, 2019.

[46] H.-H. Li, J. Rankin, J. Rinzel, M. Carrasco, and D. J. Heeger, "Attention model of binocular rivalry," *Proceedings of the National Academy of Sciences*, vol. 114, no. 30, pp. E6192–E6201, 2017.

[47] H. R. De Silva and S. H. Bartley, "Summation and subtraction of brightness in binocular perception," *British Journal of Psychology. General Section*, vol. 20, no. 3, pp. 241–250, 1930.

[48] H. Jiang, G. Jiang, M. Yu, Y. Zhang, Y. Yang, Z. Peng, F. Chen, and Q. Zhang, "Cubemap-based perception-driven blind quality assessment for 360-degree images," *IEEE Transactions on Image Processing*, vol. 30, pp. 2364–2377, 2021.

[49] P. Kora, C. P. Ooi, O. Faust, U. Raghavendra, A. Gudigar, W. Y. Chan, K. Meenakshi, K. Swaraja, P. Plawiak, and U. R. Acharya, "Transfer learning techniques for medical image analysis: A review," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 79–107, 2022.

[50] C. Chen, H. Zhao, H. Yang, T. Yu, C. Peng, and H. Qin, "Full-reference screen content image quality assessment by fusing multilevel structure similarity," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 3, 2021.

[51] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.

[52] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1940–1953, 2013.

[53] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143–1155, 2013.

[54] Y. Zhang and D. M. Chandler, "An algorithm for no-reference image quality assessment based on log-derivative statistics of natural scenes," *Image Quality and System Performance X*, vol. 8653, pp. 86530J, 2013.

[55] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917–928, 2020.

[56] Y. Liu, X. Yin, C. Tang, G. Yue, and Y. Wang, "A no-reference panoramic image quality assessment with hierarchical perception and color features," *Journal of Visual Communication and Image Representation*, vol. 95, pp. 103885, 2023.

[57] P. Ma, L. Liu, C. Xiao and D. Xu, "Omnidirectional image quality assessment with mutual distillation," *IEEE Transactions on Broadcasting*, vol. 71, no. 1, pp. 264–276, 2025.

[58] Y. Liu, S. Li, H. Duan, Y. Zhou, D. Fan and G. Zhai, "Multi-task guided blind omnidirectional image quality assessment with feature interaction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[59] Y. Liu, B. Huang, G. Yue, J. Wu, X. Wang, and Z. Zheng, "Two-stream interactive network based on local and global information for no-reference stereoscopic image quality assessment," *Journal of Visual Communication and Image Representation*, vol. 87, pp. 103586, 2022.

[60] 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 Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 64–77, 2020.

[61] Y. Zhang, L. Wan, D. Liu, X. Zhou, P. An and C. Shan, "Saliency-guided no-reference omnidirectional image quality assessment via scene content perceiving," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2024.

[62] Y. Qi, G. Jiang, M. Yu, Y. Zhang, and Y.-S. Ho, "Viewport perception based blind stereoscopic omnidirectional image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3926–3941, 2021.

[63] X. Chai, F. Shao, Q. Jiang, X. Meng, and Y.-S. Ho, "Monocular and binocular interactions oriented deformable convolutional networks for blind quality assessment of stereoscopic omnidirectional images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3407–3421, 2022.

[64] Z. Wan, X. Yan, Z. Li, X. Fan, W. Zuo and D. Zhao, "Blind stereoscopic omnidirectional image quality assessment via a binocular viewport hypergraph convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, pp.1–1, 2025.

[65] M. V. Conde, G. Geigle, and R. Timofte, "Instructir: High-quality image restoration following human instructions," 2024.

[66] J. Xu, W. Zhou and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.

[67] Z. Li and J. Atick, "Efficient stereo coding in the multiscale representation," *Network: Computation in Neural Systems*, vol. 5, no. 2, pp. 157–174, 1994.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[69] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[70] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, and A. Li, "Hifuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomedical Signal Processing and Control*, vol. 87, pp. 105534, 2024.

[71] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as bayesian inference," *Annual Review of Psychology*, vol. 55, no. Volume 55, 2004, pp. 271–304, 2004.

[72] K. Friston, "Functional integration and inference in the brain," *Progress in Neurobiology*, vol. 68, no. 2, pp. 113–143, 2002.

[73] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," 2024.

[74] A. K. Moorthy, C.-C. Su, A. Mittal, and A. C. Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 870–883, 2013.

[75] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-reference quality assessment of natural stereopairs," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3379–3391, 2013.

[76] M. Chen, Y. Jin, T. Goodall, X. Yu and A. C. Bovik, "LIVE 3D VR IQA Database," Online: http://live.ece.utexas.edu/research/VR3D/index.html, 2019.

[77] M. Chen, Y. Jin, T. Goodall, X. Yu and A. C. Bovik, "Study of 3D Virtual Reality Picture Quality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 89–102, 2020.

[78] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, "Viewport proposal cnn for 360° video quality assessment," in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10169—10178.

[79] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[80] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[81] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[82] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*, 2015, pp. 31–36.

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

[84] V. Zakharchenko, K. P. Choi, and J. Park, "Quality metric for spherical panoramic video," in *Optical Engineering + Applications*, 2016.

[85] 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 Processing: Image Communication*, vol. 58, pp. 287–299, 2017.

[86] Y. Yang, G. Jiang, M. Yu, and Y. Qi, "Latitude and binocular perception based blind stereoscopic omnidirectional image quality assessment for VR system," *Signal Processing*, vol. 173, 2020.

[87] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2 d image quality metrics and disparity analysis," 2010.

[88] C. T. E. R. Hewage and M. G. Martini, "Reduced-reference quality metric for 3d depth map transmission," in *2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 2010, pp. 1–4.

[89] R. Bensalma and M.-C. Larabi, "A perceptual metric for stereoscopic image quality assessment based on the binocular energy," *Multidimensional Syst. Signal Process.*, vol. 24, no. 2, pp. 281–316, 2013.

Yun Liu (Member, IEEE) received the Ph.D. degree in communication and information engineering from Tianjin University, China, in 2016. From 2014 to 2015, she was a visiting Ph.D. student at the Visual Space Perception Laboratory, University of California, Berkeley, United States.

She is currently an associate professor at the Faculty of Information, Liaoning University, Shenyang, China. Her research interests include multimedia quality assessment, image processing, computer vision, and pattern recognition.

Sifan Li is currently pursuing the M.S. degree in computer science and technology at the Faculty of Information, Liaoning University, Shenyang, China. His research interests include multimedia quality assessment, image processing, efficient training and inference, and computer vision.

Daoxin Fan is currently pursuing the M.S. degree in computer science and technology at the Faculty of Information, Liaoning University, Shenyang, China. His research interests include multimedia quality assessment, image processing, and computer vision.

Huiyu Duan received the B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2024. He is currently a Postdoctoral Fellow at Shanghai Jiao Tong University. From Sept. 2019 to Sept. 2020, he was a visiting Ph.D. student at the Schepens Eye Research Institute, Harvard Medical School, Boston, USA. He received the Best Paper Award of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) in 2022. His research interests include perceptual quality assessment, quality of experience, visual attention modeling, extended reality (XR), and multimedia signal processing.

Peiguang Jing received the M.S. degree and Ph.D. degree from Tianjin University, in 2013 and 2018, respectively. He is currently an associate professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. From 2014 to 2015, he was a visiting student with the National University of Singapore. His current research interests include multimedia computing, signal processing, and machine learning.

Guanghui Yue (Member, IEEE) received the B.S. degree in communication engineering and the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2014 and 2019, respectively. He was a joint Ph.D. student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore, from 2017 to 2019.

He is currently an Assistant Professor with the School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China. His research interests include signal processing, pattern recognition, medical image analysis, and multimedia quality assessment.

Guangtao Zhai (Fellow, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012. His research interests include multimedia signal processing and perceptual signal processing.