Omnidirectional Image Quality Assessment by Distortion Discrimination Assisted Multi-Stream Network

Yu Zhou, Yanjing Sun, Leida Li, *Member, IEEE*, Ke Gu, *Member, IEEE*, and Yuming Fang, *Senior Member, IEEE*

***Abstract*— Omnidirectional image (OI) quality assessment is crucial to facilitate the development of virtual reality (VR) related technology. In this work, a distortion discrimination assisted multi-stream network is proposed for OI quality assessment. The multi-stream architecture is constructed by generating the viewport images received by the retina at one point to simulate the characteristics of humans perceiving VR contents. Additionally, the strategy of generating several viewport image sets from one OI is proposed for data augmentation. Furthermore, the facts that the human brain has the ability for both quality assessment and distortion type distinguishment, and the process of human brain handling two tasks exists information interaction inspire us to employ an auxiliary distortion discrimination task to facilitate the quality assessment task learning. Extensive experiments conducted on two public OI databases demonstrate the superiority of the proposed method to both traditional 2D quality metrics and existing metrics specific for OIs. Moreover, utilizing the assistant task is proven to be more effective than the single task learning for OI quality evaluation. Better generalization performance is also verified to be another valuable trait of the proposed method.**

***Index Terms*— Image quality assessment, virtual reality (VR), omnidirectional image (OI), viewport generation, distortion discrimination.**

Manuscript received February 19, 2021; revised April 29, 2021; accepted May 13, 2021. Date of publication May 17, 2021; date of current version April 5, 2022. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20200649; in part by the National Natural Science Foundation of China under Grant 62001475, Grant 62071472, Grant 61771473, Grant 62076013, and Grant 62021003; and in part by the Program for the Industrial Internet of Things (IoT) and Emergency Collaboration Innovative Research Team in China University of Mining and Technology (CUMT) under Grant 2020ZY002. This article was recommended by Associate Editor H. Yue. *(Corresponding author: Yanjing Sun.)*

Yu Zhou and Yanjing Sun are with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China, and also with the Xuzhou Engineering Research Center of Intelligent Industry Safety and Emergency Collaboration, Xuzhou 221116, China (e-mail: zhouy@cumt.edu.cn; yjsun@cumt.edu.cn).

Leida Li is with the Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China, and also with the Pazhou Laboratory, Guangzhou 510330, China (e-mail: ldli@xidian.edu.cn).

Ke Gu is with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China, and also with the Beijing Laboratory of Smart Environmental Protection, the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Ministry of Education, Engineering Research Center of Intelligent Perception and Autonomous Control, Beijing Artificial Intelligence Institute, Beijing University of Technology, Beijing 100124, China (e-mail: guke.doctor@gmail.com).

Yuming Fang is with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China (e-mail: fa0001ng@e.ntu.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3081162.

Digital Object Identifier 10.1109/TCSVT.2021.3081162

I. Introduction

VIRTUAL reality (VR) has been attracting much attention due to the realistic and immersive visual experience it provides for consumers [1], [2]. This experience is achieved by resorting to the head-mounted displays (HMDs) to perceive the 360° omnidirectional information. Undoubtedly, the visual quality of omnidirectional images (OIs) plays a vital role in the users' experience. Poor quality contents inevitably cause physical and mental discomfort [3]. However, the quality degradation of OIs is frequent in practice [4], [5]. In more detail, to provide the 360° free-view experience, OIs in the VR system are originally spherical, which is quite different from traditional 2D plane images that are displayed on the flat screen. This characteristic of OIs predestines to require much higher resolution and larger storage space, which brings great challenges to the current image acquisition, encoding, transmission and display technologies. In this case, more or less OI quality degradation is caused. Therefore, designing quality metrics for OIs is of remarkable significance, which can be employed to guide the studies of VR related technology, e.g. the stitching and coding techniques.

So far, lots of quality approaches have been proposed, which can be divided into subjective and objective ones, and the latter is needed more urgently for the advantages of higher stability, time saving and economy, etc [6], [7]. Further, the objective metrics can be classified into three types, namely full-reference (FR) [8]– [12], reduced-reference (RR) [13], [14] and no-reference (NR) metrics [15]– [18]. As far as we know, existing metrics have got excellent performance for traditional 2D plane images. Nevertheless, they expose disadvantages in the OI quality assessment. Particularly, conventional quality metrics are designed for the 2D equirectangular images, while raw OIs are in the spherical form. Of course, OIs are usually transformed to 2D format in practice using the non-linear projection methods to facilitate the compression, transmission and storage stages [19], [20]. However, structures in the bipolar regions of the sphere OIs conceivably produce geometric deformations during this transformation process. Unfortunately, traditional quality metrics are sensitive to geometric deformations, which will be mistaken as the real distortions in the raw OIs, leading to misjudgement and non-ideal evaluation results. This indicates the urgent need for the quality metrics specific for OIs.

In the literature, there exist a few of OI quality approaches, including FR and NR types. Among them, the FR approaches are mainly based on existing 2D quality metrics. More specifically, Yu *et al.* [21] proposed to calculate the Peak Signal-to-Noise Ratio (PSNR) value of the uniformly sampled points in the sphere domain, producing the S-PSNR method. In [22], the craster parabolic projection PSNR (CPP-PSNR) metric was proposed by calculating the PSNR value of the resampled pixels in this projection domain. Sun *et al.* [23] presented a weighted-to-spherically-uniform PSNR (WS-PSNR) method by multiplying the pixel error in the projection plane and the weight that was computed based on the stretching ratio of area from projection to the spherical domain. Due to the inconsistency of PSNR with the human visual system, these PSNR based OI quality metrics cannot even perform as well as traditional 2D quality metrics, which has been reported in [24], [25]. To avoid this problem, the sphere structural similarity metric was proposed to assess the quality of omnidirectional videos from aspects of luminance, contrast and structural similarities [26]. The above FR metrics are only feasible with the availability of the corresponding reference OIs without distortions, which heavily limits their applications in real scenarios. As comparison, the NR approaches independent of reference information are more popular [27], [28]. For this, a blind omnidirectional image quality assessment quality metric (SSP-BOIQA) based on the segmented spherical projection (SSP) representation was proposed [5]. The SSP OI transformed from the equirectangular projection (ERP) version was first segmented into two bipolar and one equatorial regions. Then, several sets of features with heat map features as weights were separately extracted from three regions. Finally, the random forest was employed for the quality model training. In [29], a multi-channel convolutional neural network (CNN) was presented for OI quality assessment. The hyper-ResNet34 based CNN was used for feature representation and then a quality regressor was employed to integrate the features into the quality score. Kim *et al.* [1] and Lim *et al.* [30] proposed the OI quality methods. A network for the patch-based quality score calculation and weight prediction was first constructed, and subsequently the whole OI quality was measured by fusing the predicted quality scores and weights of all patches. Ling *et al*. [31] proposed a blind quality metric for stitched panoramic images by quantifying the ghosting and structure inconsistency artifacts using convolutional sparse coding. In [32], an NR stereoscopic OI quality method based on latitude and binocular perception was presented. The multi-scale monocular features and the binocular perception features were first extracted. Then the random forest regression model was utilized to train the quality model from these features. Jiang *et al*. [33] presented three schemes for blind quality evaluation of OIs in the cubemap projection domain. The first scheme was achieved by pooling the quality scores of six cubemap faces. The second scheme was produced by introducing the human attention factor to the first scheme. The last scheme was proposed based on pooling the attention distortion features. In [34], a blind quality index for OIs was proposed by predicting and aggregating the local and global quality. The global quality was evaluated based on the entire OI while the local quality was assessed based on the sampled viewport images. Even though these metrics have made encouraging progress in the OI quality prediction, great efforts are still needed for more superior performance.

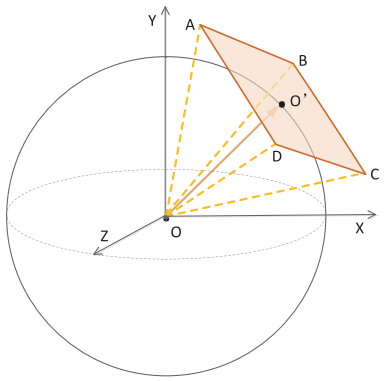


Fig. 1. Illustration of the viewport image received by the viewers equipped with the HMD at center O.

At present, the main challenges in OI quality assessment are multi-fold. First, the human has a limited field of view when they gaze in one direction instead of perceiving the entire 360-degree contents at a time. Meanwhile, the nonlinear transformation between the 2D equirectangular format OI and the original sphere version causes the gap between structures in the equirectangular image and those actually captured by human eyes in the spherical viewing environment. Therefore, one key problem is how to get the contents that are actually received by human vision. Second, as the subjective study of OI quality assessment is high-cost, the number of OI data is limited, which hinders the development of high-performance algorithms, especially for the deep learning based methods. Moreover, the experience comfort is one important indicator for the quality of VR contents. Obviously, modeling the users' experience is a much more complex and arduous task, and related works are relatively rare.

To this end, we propose a viewport images based multi-stream network for OI quality assessment, together with the distortion discrimination as an assistant task. As only contents in the viewport region can be perceived by viewers at a moment, instead of the entire spherical OI, a viewport generation algorithm is first introduced in the proposed method to face the challenge of producing image contents that are actually perceived by human eyes. Specifically, when users equipped with the HMD watch the VR content from an angle, the OI content in the corresponding viewport region will be modeled as a plane segment tangential to the sphere resorting to the head motion data [21]. Fig. 1 illustrates a viewport image example ABCD that is perceived by users at center O, where $\overrightarrow{OO'}$ is the gazing direction. Viewers rotate viewing directions to perceive multiple viewport images that cover the full OI content, and further comprehensively judge the whole quality by fusing the quality of all viewport images. Besides, distortions that actually appear to users in the viewport images are quite different from those in equirectangular OIs, which can be obtained from Fig. 2. Figs. 2(a) and 2(b) show a distorted equirectangular OI with JPEG compression and the corresponding viewport images generated by the method in

Ref. [21], together with the enlarged versions of local distorted regions placed in the bottom right of each image. By comparison, we can find that the JPEG compression presents the square block effects in the equirectangular OI but appears quite different in the viewport regions that are intuitively received by human eyes, especially for the "top" viewport image. This undermines the advantages of the quality metrics that are designed for flat images with traditional distortions. These above facts guide us to design a novel multiple viewports based quality metric, naturally forming the parallel multi-steam framework. Besides, as users usually choose different starting viewing angles during the VR experience, several sets of viewport images for each OI are generated by setting different starting viewing angles in this work. This is also an effective way to relieve the pressure of data volume. Furthermore, enlightened by the fact that using the related task as an assistant is effective in promoting the performance improvement of the individual task [35], an auxiliary task is employed to assist the viewport quality prediction. Considering the quality degradation is caused by distortions and various types of distortions affect the human perceiving results differently, we design a distortion type discrimination learning network to aid the quality prediction. In more detail, the input equirectangular OI is first projected to a set of non-overlapping viewport images, and the produced viewport images are used as the input of the proposed parallel multi-stream network. The architecture of each stream includes one shared network and two individual task learning branches. The shared part not only explores the mutual information between two tasks but also releases the data volume pressure. The experimental results demonstrate the advantages of the proposed method in the OI quality prediction task, and the cross-dataset validation also indicates the best generalization ability of the proposed approach.

Main contributions of this work are summarized as follows:

- The proposed viewport images based multi-stream network is more aligned with the human perception process compared with the framework based on entire OIs. Besides, the proposed strategy of generating multiple viewport image sets from one equirectangular OI is effective for data augmentation.
- We propose to use the distortion discrimination task as the assistant of the main quality prediction task, which is helpful to better train the quality evaluation network by delving the shared information between two tasks and ease the data volume burden.
- The proposed method is demonstrated by extensive experiments to outperform the state-of-the-art quality metrics and have the best generalization ability.

The rest of this paper is organized as follows. Section II details the proposed method. Section III presents the experimental results for performance testification. In the last section IV, the conclusion of this article is drawn.

II. Proposed Method

Fig. 3 illustrates the framework of the proposed method during training. It consists of four modules, namely the viewport generation module, shared network module and two individual task learning modules. As aforementioned, the first module is to simulate the contents that are perceived by users with HMDs in the VR experience. The second module is to learn the mutual information between two tasks, and the last two modules are for quality prediction and distortion discrimination tasks, respectively. During the training stage, the auxiliary information about the distortion discrimination is transferred to the shared network by the gradient backpropagation under the supervision of the fused loss of two tasks. In other words, the auxiliary network helps to learn the model of the main quality assessment task through optimizing the shared network. During test, only the shared network and the above quality evaluation branch are employed to predict the quality score. Subsequently, we will give more details about the proposed method.

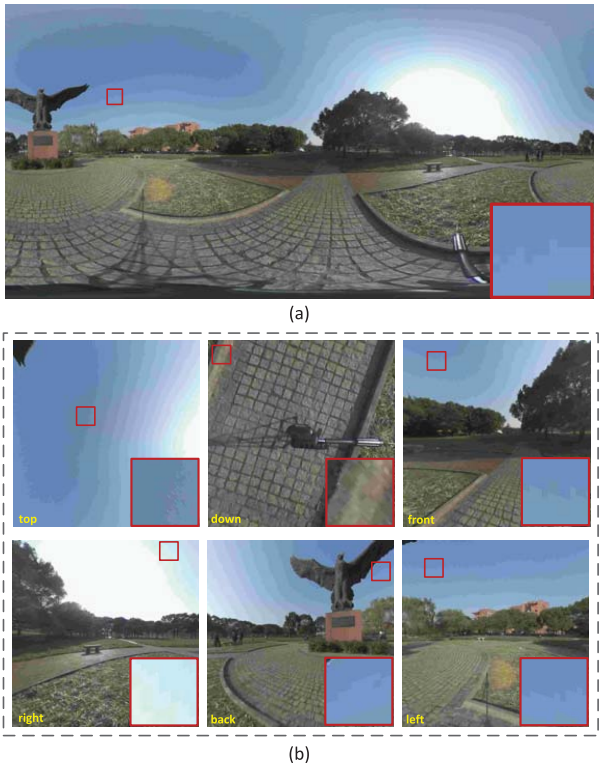


Fig. 2. Illustration of a distorted equirectangular OI with JPEG distortion and the corresponding viewport images, together with the local distorted regions magnified and placed in the bottom right. (a): A distorted equirectangular OI with JPEG compression; (b): Six non-overlapping viewport images.

A. Viewport Generation

As reported in Ref. [21], users with HMDs can only perceive the visual information in a limited field of view at a time, which is dubbed as viewport. As aforementioned, the plane viewport image is tangential to the sphere, which is modeled by resorting to the head motion data, including the viewing angle and the field of view [21]. Fig. 1 presents an illustration. To adapt to this perceiving property, the proposed method is designed based on the quality prediction of viewport images. First, the viewport images are generated based on Ref. [21], where the field of view in this work is set to 90°. Further, with

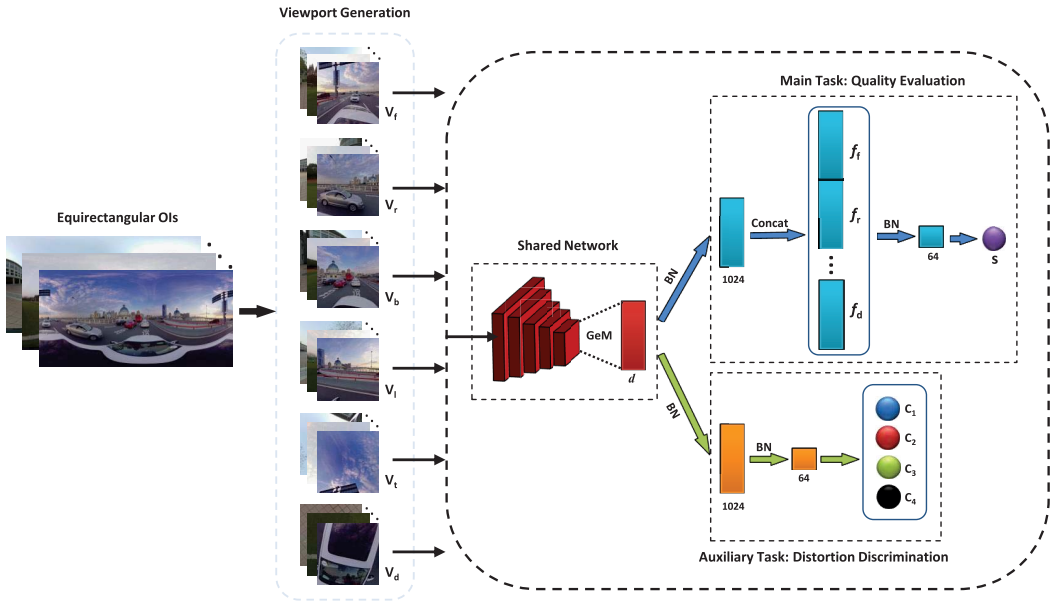


Fig. 3. Framework of the proposed method during training, where GeM and BN denote generalized-mean pooling and batch normalization, respectively.

consideration of the fact that users quantify the OI quality by rotating the head position to perceive contents within 360°, six non-overlapping viewport images at six directions, including top, bottom, front, back, left, and right, are modeled to cover the full OI content, ensuring information integrity. Specifically, the top and bottom viewport images are two ones centered at the North and South poles, denoted by $\mathbf{V}_t$ and $\mathbf{V}_d$, respectively. The other four viewport images are the ones covering the entire equator fields, namely the front, back, left and right viewports, which are denoted by $\mathbf{V}_f$, $\mathbf{V}_b$, $\mathbf{V}_l$ and $\mathbf{V}_r$. Furthermore, due to the fact that users usually choose different starting viewing angles during the VR experience in practice, we propose to produce many sets of viewport images for each OI by setting different starting viewing angles in implementation. The viewport images generated from the same OI and with the same starting viewing angle are grouped into one set. This viewport generation scheme not only is consistent with the actual viewing case but also realizes effective data augmentation to avoid the over-fitting problem. Particularly, the starting view angle is selected every θ degree from 0° to 360° by rotating the longitude, generating n groups for each OI, where $n = \lfloor 360/\theta \rfloor$, $\lfloor \cdot \rfloor$ means rounding in the negative direction. Therefore, for a specific OI $\mathbf{V}$, the viewport image sets are denoted as $\mathbf{V} = \{\mathbf{V}_t^i, \mathbf{V}_d^i, \mathbf{V}_f^i, \mathbf{V}_b^i, \mathbf{V}_l^i, \mathbf{V}_r^i\}$, where $i \in [1, n]$. In this work, θ is empirically set to 5°.

Fig. 4 presents two reference equirectangular OIs and one set of viewport images generated from the starting viewing angle of 0° using the above method. From this figure, we can see that the geometric distortions are introduced to the equirectangular OIs during the projection from the spherical OI. However, these distortions vanish in the viewport images. In other words, contents in the viewport images are distinctly more consistent with the real scenes viewed by users with the HMDs, which is undoubtedly a good beginning for the following multi-stream network learning.

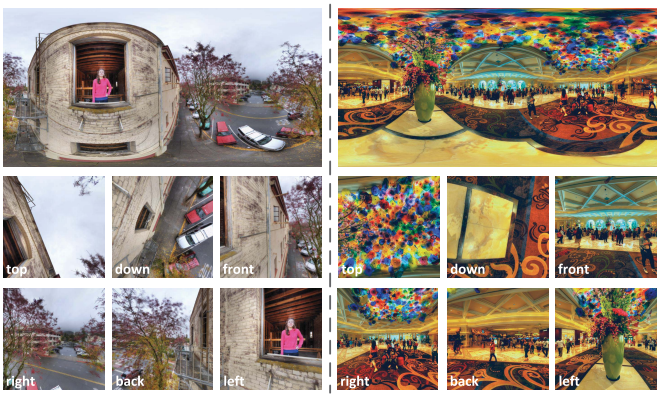


Fig. 4. Two sets of reference equirectangular OIs and the corresponding viewport images generated with the starting view angle of 0°.

B. Shared Network

We denote the input training data by $\{(\mathbf{V}(k), q_k, c_k)\}_{k=1}^N$, where k denotes the k-th set of input viewport images, q_k is the subjective quality score, c_k is the distortion classification tag, and N represents the total number of viewport image sets for training. One set of images is attached with one subjective quality score and one distortion type tag. For network training, the training data $\mathbf{V}(k)$ is fed into the shared network, which is built upon the pre-trained CNN model, such as Inception-v3 [36] or ResNet50 [37]. The fully-connected (FC) layers are removed and the Generalized-Mean (GeM) pooling layer [38] is appended to the last convolutional layer to translate the 3D feature map into 1D feature vector d. Suppose the output of the last convolutional layer as 3D feature map $E \in R^{C \times H \times W}$, the GeM operation can be formulated as,

$$d = \left(\frac{1}{|E|} \sum_{e_i \in E} e_i^p\right)^{\frac{1}{p}}, \tag{1}$$

where p denotes the pooling hyperparameter, which can be manually set or learnt by the back-propagation operation, and

$d \in R^{C\times 1\times 1}$. The GeM pooling method approaches to max pooling when $p \to \infty$ and approaches to average pooling when $p \to 1$.

C. Individual Branch Network

1) Quality Evaluation Network: On top of the shared network, an FC layer containing 1024 nodes is added for feature representation of the quality evaluation task. For each viewport image, a feature vector $\boldsymbol{f}$ with dimension of 1024×1 can be obtained. Therefore, for one set of viewport images of six directions, six feature vectors can be got, denoted by $\{\boldsymbol{f}_{\mathrm{f}},\boldsymbol{f}_{\mathrm{r}},\boldsymbol{f}_{\mathrm{b}},\boldsymbol{f}_{\mathrm{l}},\boldsymbol{f}_{\mathrm{t}},\boldsymbol{f}_{\mathrm{d}}\}$. Then, six vectors are concatenated to produce one vector. Afterwards, two FC layers with 64 nodes and 1 node are adopted in succession to produce the final quality score of the whole OI. Besides, the Batch Normalization (BN) operation [39] is implemented following both the GeM layer and each FC layer to accelerate network training and improve the generalization ability.

2) Distortion Discrimination Network: Two FC layers respectively containing 1024 and 64 nodes are appended to the shared network to extract features for the distortion classification task. The BN layers are also added following the GeM layer and each FC layer. After that, the Softmax operator [40], [41] is employed to implement the nonlinear transformation, which is defined as,

$$\hat{p}_k^i\left(\mathbf{V}(k);\mathbf{W}\right)=\frac{e^{p_k^i(\mathbf{V}(k);\mathbf{W})}}{\sum_{j=1}^{C}e^{p_k^j(\mathbf{V}(k);\mathbf{W})}},\quad i\in[1,C] \tag{2}$$

where p_k^i denotes the i-th activation value of the last FC layer of the k-th input image, $\mathbf{W}$ is the collection of the parameters in the shared network and the distortion discrimination branch network, C is the total number of the distortion types. Moreover, since the distortion type in a set of viewport images is the same, only one viewport image among a set of six viewport images is randomly selected for training this distortion discrimination branch.

D. Network Training

In this work, the end-to-end training mode is adopted for model learning. The Euclidean and cross entropy loss functions are severally used to optimize the quality prediction network and the distortion discrimination network, which are defined as follows,

$$L_q\left(\mathbf{V}(k);\mathbf{W}_0\right)=\frac{1}{N}\sum_{k=1}^{N}\left\|s_k\left(\mathbf{V}(k);\mathbf{W}_0\right)-q_k\right\|_2^2, \tag{3}$$

where s_k and q_k are the predicted quality score and the subjective score of the k-th training sample in a mini-batch, $\mathbf{W}_0$ denotes the collection of the parameters in the shared network and the quality prediction network.

$$L_d\left(\mathbf{V}(k);\mathbf{W}\right)=-\sum_{k=1}^{N}\sum_{i=1}^{C}c_k^i\log\hat{p}_k^i\left(\mathbf{V}(k);\mathbf{W}\right), \tag{4}$$

where $\mathbf{c}_k$ is the ground-truth multi-class indicator vector with only one item activated to represent the distortion type, $\hat{p}_k^i$ indicates the predicted probability of the distortions in the k-th image belonging to the i-th distortion type.

In implementation, the Stochastic Gradient Descent (SGD) optimization algorithm is utilized to push the network to learn the superior parameters for both tasks. The final target of training the proposed network is to minimize the L value,

$$L=L_q+\lambda L_d, \tag{5}$$

where λ reflects the relative importance of two losses and is empirically set to 0.1 in this work. In this manner, the quality score of a test image can be generated using the trained quality prediction branch network.

III. Experimental Results

A. Experimental Settings

1) Datasets: In this work, the performance of the proposed method is validated on the public OIQA [42] and CVIQ [29] databases. Details of these two databases are shown as follows.

OIQA database: It includes 320 distorted OIs degraded by four types of distortions from 16 reference images with different scenes. The distortion types are JPEG compression, JPEG2000 compression, Gaussian blur (GB), and Gaussian noise (GN). Besides, the resolutions are various, ranging from 11332×5666 to 13320×6660. Similar to [29], the SS method is adopted for the subjective test and also the Mean Opinion Score (MOS) value is provided.

CVIQ database: It consists of 528 distorted OIs compressed by three coding technologies (JPEG [43], H.264/AVC [44] and H.265/HEVC [45]) from 16 reference images with different scenes but with the same resolution of 4096×2048. The MOS value of each image is provided as the benchmark by conducting the single-stimulus (SS) subjective experiment.

2) Evaluation Criterion: For performance comparison of the proposed method with the state-of-the-arts, three commonly adopted and widely acknowledged criteria in the quality assessment field are employed, including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC) and Root Mean Squared Error (RMSE). Among them, PLCC and RMSE are for the prediction accuracy evaluation, while SRCC is for the monotonicity evaluation [46], [47]. Higher PLCC and SRCC values, and lower RMSE value indicate the higher prediction accuracy and consistency. Except SRCC, both PLCC and RMSE are computed following the five-parameter nonlinear mapping [48], [49],

$$F(\theta)=\rho_1\left(\frac{1}{2}-\frac{1}{1+e^{\rho_2(\theta-\rho_3)}}\right)+\rho_4\theta+\rho_5, \tag{6}$$

where θ and $F(\theta)$ denote the predicted quality score and the mapped objective score, respectively, and $\rho_1,\rho_2,\cdots,\rho_5$ represent the fitting parameters. The aim of this function is to unify the objective scores calculated by different quality metrics into the same range.

In implementation, the 8-fold cross validation method is adopted for performance evaluation. Specifically, the database is first randomly segmented into eight folds according to the reference images, ensuring images with the same scene are allocated to the same fold. Then seven folds are chosen for

TABLE I
IMPACT OF DIFFERENT λ VALUES ON THE SRCC VALUE OF THE PROPOSED METHOD

λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
OIQA	0.880	**0.923**	0.902	0.888	0.857	0.842	0.820	0.834	0.817	0.805	0.789
CVIQ	0.866	**0.911**	0.894	0.887	0.864	0.853	0.825	0.834	0.801	0.791	0.770

model training and the remaining one is for test. This iteration process is conducted eight times until all folds have been used for test for one time. For each iteration, one set of criterion values can be generated, which is called PLCC(t), SRCC(t) and RMSE(t), where $t \in [1, 8]$ denotes the iteration time. Then, eight criterion values are averaged to produce the final performance,

$$P = \frac{1}{8}\sum_{t=1}^{8} P(t), \tag{7}$$

where P is the pronoun of PLCC, SRCC and RMSE.

3) Comparison Metrics: To validate the performance of the proposed approach, both FR and NR state-of-the-art quality metrics are employed for comparison. Meanwhile, not only traditional quality metrics, but also the methods specific for OIs are adopted. Particularly, the FR metrics include PSNR, SSIM [8], S-PSNR [21], CPP-PSNR [22] and WS-PSNR [23], and the NR metrics include BIQI [15], BLIINDS [18], DESQUE [50], DIIVINE [51], IL-NIQE [52], BMPRI [53], MEON [54], dipIQ [55], CSC [31], SSP-BOIQA [5] and MC360IQA [29]. Among them, S-PSNR [21], CPP-PSNR [22], WS-PSNR [23], CSC [31], SSP-BOIQA [5] and MC360IQA [29] approaches are specifically designed for OI quality assessment, while the remaining approaches are traditional ones. The S-PSNR, CPP-PSNR and WS-PSNR approaches are implemented using the 360*tools* package,[1] and the source codes of the rest approaches are provided by the authors.

4) Implementation Details: Pytorch [56] is employed to implement the proposed method. The viewport images are first resized to $299 \times 299 \times 3$ before feeding into the proposed network. The Inception-v3 [36] network pre-trained on ImageNet [57] is adopted as the backbone. For the shared network and the branch networks, the initial learning rates are severally set to $1e$-4 and $1e$-3. The learning rate drops to a factor of 0.9 each epoch, and total epoch is 50. For the hyper-parameters momentum, the value is 0.9. The training and test processes are implemented on a server with Intel Xeon(R) Platinum(R) 8163 CPU@2.5 GHz, 32 GB RAM and NVIDIA Tesla V100-SXM2.

B. Parameter Selection

In this subsection, we compare the performance of the proposed method with different values of λ in Eq. (5). As the quality evaluation task is the main task, the λ value is tested from 0 to 1 with the step of 0.1. This parameter is optimized based on the SRCC value, which is a commonly used criterion for parameter optimization in the field of image quality assessment [58], [59]. The optimization process is conducted on both the OIQA and CVIQ databases. Table I shows the experimental results. It can be observed from this table that the proposed method performs the best on both databases when λ is set to 0.1. This indicates the relative importance of L_q for model training, which highlights the leading role of the quality evaluation task.

[1] [Online]. Available: https://github.com/Samsung/360tools

C. Performance Evaluation

1) Comparison With State-of-the-Arts: In this part, we first compare the performance of the proposed method with the aforementioned state-of-the-arts on the OIQA and CVIQ databases, respectively. Table II presents the experimental results on the OIQA database, where the best two results are marked in boldface. From this table, we can see that the proposed method is among the top two best for images with each type of distortions and the whole database. Especially on images with JPEG and GN distortions, the proposed method achieves the highest PLCC and SRCC values, and the lowest RMSE value. On images with JP2K and GB distortions, the SSIM [8] and dipIQ [55] methods obtain the highest PLCC or SRCC value. However, for images with other two types of distortions, both SSIM [8] and dipIQ [55] methods have poorer performance than the proposed method. Furthermore, it can be observed from the results shown in the last three columns of Table II that the proposed method has more superior performance than all the other metrics on the whole OIQA database, which indicates that the proposed method possesses the best ability to evaluate images across different types of distortions.

Table III further summarizes the experimental results on the CVIQ database, including results on each distortion type and on the whole database. From this table, we can see that the performance of the proposed method ranks the top two on all the distortion types and the entire database, which indicates the superiority of the proposed method in the OI quality prediction. In detail, on the images with JPEG and H.264/AVC distortions, the proposed method achieves both the highest PLCC and SRCC values and the lowest RMSE values among all the FR and NR metrics. On the images with H.265/HEVC distortion, the proposed method also obtains the highest PLCC and lowest RMSE values among either FR or NR metrics. Even though the SSIM method [8] achieves a slightly larger SRCC value than the proposed method, the dependency of the reference OI leads to serious application limitation in the real scenario. On the entire database, the proposed method has the best monotonicity prediction ability. Despite the higher PLCC value of MC360IQA [29], it performs poorly than our method in all other cases. In conclusion, the proposed method has the best overall performance on the CVIQ database.

Through the above two sets of experiments, it can be found that the proposed method obtains the overall optimal performance on both databases. Moreover, we can observe

TABLE II

PERFORMANCE OF THE PROPOSED METHOD AND THE STATE-OF-THE-ARTS ON THE OIQA DATABASE

Method	JPEG			JP2K			GN			GB			Overall		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
PSNR	0.758	0.731	10.245	0.781	0.768	9.379	0.958	0.931	3.654	0.529	0.506	11.268	0.492	0.497	12.528
SSIM [8]	0.803	**0.934**	9.355	0.802	**0.936**	8.985	0.904	0.886	5.467	0.768	**0.925**	8.500	0.856	0.880	7.436
S-PSNR [21]	0.870	0.829	7.738	0.816	0.849	8.686	0.919	0.885	5.033	0.699	0.692	9.501	0.716	0.712	10.030
CPP-PSNR [22]	0.865	0.829	7.873	0.849	0.837	7.943	0.920	0.885	5.001	0.672	0.667	9.830	0.707	0.703	10.167
WS-PSNR [23]	0.861	0.828	7.994	0.844	0.832	8.070	0.922	0.885	4.942	0.661	0.658	9.966	0.689	0.693	10.428
BIQI [15]	0.817	0.839	9.044	0.058	0.133	15.001	0.805	0.803	7.571	0.519	0.424	11.350	0.473	0.414	12.678
BLIINDS [18]	0.757	0.737	10.250	**0.918**	0.925	**5.950**	0.925	0.913	4.860	0.729	0.691	9.089	0.772	0.774	9.137
DESQUE [50]	0.897	0.868	6.952	0.739	0.732	10.120	0.953	0.937	3.882	0.749	0.663	8.799	0.725	0.712	9.903
DIIVINE [51]	0.373	0.230	14.564	0.774	0.748	9.511	0.960	**0.954**	3.560	0.826	0.775	7.476	0.616	0.581	11.338
IL-NIQE [52]	0.783	0.755	9.756	0.300	0.249	14.336	0.695	0.675	9.178	0.661	0.663	9.965	0.597	0.573	11.540
BMPRI [53]	**0.918**	0.909	**6.210**	0.185	0.166	14.768	**0.961**	0.949	**3.534**	0.386	0.354	12.248	0.431	0.338	12.984
MEON [54]	0.823	0.779	8.935	0.680	0.601	11.017	0.952	0.930	3.895	0.764	0.716	8.572	0.749	0.717	9.536
dipIQ [55]	0.829	0.789	8.783	0.916	0.918	6.030	0.955	0.943	3.772	**0.932**	0.898	**4.816**	0.701	0.691	10.259
CSC [31]	0.779	0.765	9.831	0.791	0.812	9.107	0.886	0.812	6.031	0.767	0.689	8.515	0.764	0.707	9.294
SSP-BOIQA [5]	0.877	0.834	7.620	0.853	0.852	7.501	0.905	0.843	5.451	0.854	0.862	6.834	0.860	0.865	7.313
MC360IQA [29]	0.912	0.901	6.535	0.896	0.882	6.573	0.913	0.926	5.240	0.893	0.918	6.072	**0.890**	**0.909**	**6.697**
Proposed	**0.936**	**0.940**	**5.691**	**0.920**	**0.934**	**5.886**	**0.968**	**0.957**	**3.330**	**0.925**	**0.920**	**4.972**	**0.899**	**0.923**	**6.396**

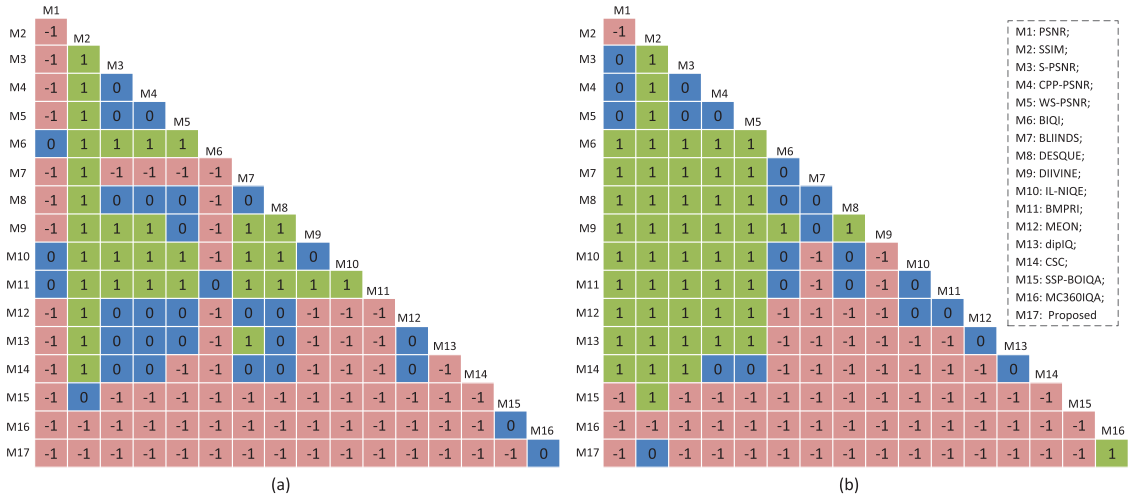


Fig. 5. Statistical performance between any two objective quality metrics on OIQA and CVIQ databases. (a): OIQA database; (b): CVIQ database.

that the PSNR-based metrics [21]–[23] that are specific for OIs even perform poorer than some traditional methods. This is mainly because the PSNR model is inconsistent with the human visual system [24], [25].

2) Statistical Performance Comparison: We further conduct the F-test [60] to compare the statistical performance between any two objective quality metrics. Suppose M_α and M_β are two metrics for comparison. The F-test score is first calculated using the RMSE values of M_α and M_β,

$$F_t = \left(\frac{\text{RMSE}_{M_\alpha}}{\text{RMSE}_{M_\beta}}\right)^2. \tag{8}$$

Subsequently, the threshold $F_{critical}$ is computed using the MATLAB function *finv* based on the number of residuals and the confidence level. In this work, the confidence level is set to 95%. If the F_t value is larger than the $F_{critical}$ value, it indicates that M_β has significantly better statistical performance than M_α. If the F_t value is between $1/F_{critical}$ and $F_{critical}$, two methods are believed to have the competitive performance. Otherwise, M_β performs significantly worse than M_α in the statistical performance. For the OIQA and CVIQ databases, the $F_{critical}$ values are 1.2022 and 1.1541, respectively.

Fig. 5 illustrates the statistical performance comparison results on two databases. In this figure, "1" denotes the metric above the square lattice has significantly better statistical performance than the metric to the left; "0" indicates the competitive statistical performance while "−1" represents the rather superior performance of the metric to the left. Three different colors are for distinguishment. From this figure, we can observe that the "M_2" (SSIM) method performs significantly better than all the other metrics except the proposed method

TABLE III
PERFORMANCE OF THE PROPOSED METHOD AND THE STATE-OF-THE-ARTS ON THE CVIQ DATABASE

Method	Type	JPEG			H.264/AVC			H.265/HEVC			Overall		
		PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
PSNR	FR	0.889	0.766	7.824	0.784	0.783	7.674	0.746	0.745	8.000	0.786	0.757	8.692
SSIM [8]	FR	0.852	**0.929**	8.946	**0.941**	0.940	**4.177**	**0.918**	**0.917**	**4.763**	0.897	0.885	6.230
S-PSNR [21]	FR	0.892	0.778	7.727	0.789	0.786	7.589	0.762	0.758	7.785	0.785	0.761	8.714
CPP-PSNR [22]	FR	0.884	0.765	7.996	0.779	0.777	7.751	0.751	0.748	7.936	0.779	0.754	8.822
WS-PSNR [23]	FR	0.880	0.756	8.101	0.775	0.773	7.814	0.747	0.744	7.993	0.777	0.751	8.850
BIQI [15]	NR	0.794	0.695	10.382	0.572	0.593	10.139	0.095	0.029	11.973	0.548	0.442	11.766
BLIINDS [18]	NR	0.104	0.241	16.980	0.724	0.714	8.527	0.703	0.708	8.548	0.465	0.529	12.458
DESQUE [50]	NR	0.912	0.870	7.003	0.385	0.173	11.410	0.328	0.152	11.362	0.566	0.417	11.603
DIIVINE [51]	NR	0.693	0.546	12.316	0.639	0.546	9.511	0.441	0.392	10.795	0.438	0.413	12.650
IL-NIQE [52]	NR	0.626	0.534	12.968	0.832	0.743	7.478	0.372	0.358	11.173	0.626	0.534	10.968
BMPRI [53]	NR	0.776	0.498	10.767	0.533	0.520	10.459	0.846	0.840	6.412	0.627	0.621	10.962
MEON [54]	NR	0.808	0.566	10.057	0.599	0.574	9.900	0.783	0.782	7.487	0.665	0.567	10.510
dipIQ [55]	NR	0.928	0.793	6.353	0.620	0.635	9.695	0.361	0.326	11.216	0.706	0.623	9.960
CSC [31]	NR	0.752	0.684	11.145	0.776	0.709	7.782	0.750	0.714	7.938	0.732	0.756	9.387
SSP-BOIQA [5]	NR	0.915	0.853	6.847	0.885	0.861	7.042	0.854	0.841	6.302	0.890	0.856	6.941
MC360IQA [29]	NR	**0.941**	0.923	**5.804**	0.932	**0.941**	5.357	0.914	0.899	4.801	**0.939**	**0.904**	**4.606**
Proposed	NR	**0.957**	**0.961**	**5.601**	**0.953**	**0.949**	**3.873**	**0.929**	**0.914**	**4.525**	**0.902**	**0.911**	**6.117**

(M_{17}) and the "M_{16}" (MC360IQA) method on both databases. On the OIQA database, only the "M_{16}" metric has competitive statistical performance to ours. On the CVIQ database, no method has significantly better statistical performance than the proposed method except the "M_{16}" method, which can be observed from Fig. 5(b). These demonstrate the obvious advantages of the proposed method in terms of statistical performance than most metrics.

D. Ablation Ability

In this subsection, several ablation experiments are conducted to verify the positive role of adding the auxiliary distortion discrimination task, the GeM pooling method compared with traditional max pooling and average pooling methods, and the SGD optimization compared with Adam and GCD optimization in this task. Particularly, we first compare the performance of the proposed method with and without the auxiliary task. In Fig. 6(a) and Fig. 6(b), the green bars illustrate the SRCC values of the method discarding the auxiliary branch and the red bars on the green ones represent the SRCC increase by adding the auxiliary task. From the results, we can observe that appending the auxiliary branch indeed contributes to improving the quality prediction consistency.

Further, we respectively replace the GeM pooling method by the max pooling method and the average pooling method to test the performance of the proposed method with different pooling ways. The cyan bars and orange bars in image (a) present the results of the proposed method with max pooling and the performance increase with the alternative GeM pooling on the OIQA database, from which we can see that the proposed method using the GeM pooling performs much better than the method with max pooling regardless of the distortion type. Besides, the blue and orange bars in the right part of image (a) also indicate the superiority of the GeM pooling to the average pooling approach. Similar conclusions can be drawn from the results on the CVIQ database, which can be obtained from Fig. 6(b). Therefore, we employ the GeM pooling method in our metric.

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH DIFFERENT OPTIMIZATION ALGORITHMS

Database	Optimizer	PLCC	SRCC	RMSE
OIQA	Adam	0.878	0.884	6.846
	GCD	0.849	0.847	7.581
	SGD	**0.899**	**0.923**	**6.396**
CVIQ	Adam	0.891	0.887	6.932
	GCD	0.856	0.859	7.583
	SGD	**0.902**	**0.911**	**6.117**

Moreover, as optimizer impacts the performance of the trained model, we compare the performance of the proposed method when different optimization algorithms are used, including the commonly used Adam, GCD and SGD. Table IV presents the comparison results. From this table, we can see that performance of various optimizers is differently. Besides, the proposed quality model optimized using the SGD algorithm performs the best on both databases. Therefore, it is employed for model training in our work.

E. Generalization Ability

Further, to verify the generalization ability of the proposed method, the cross-database experiments are implemented for all the training-based metrics in Tables II and III. Particularly,

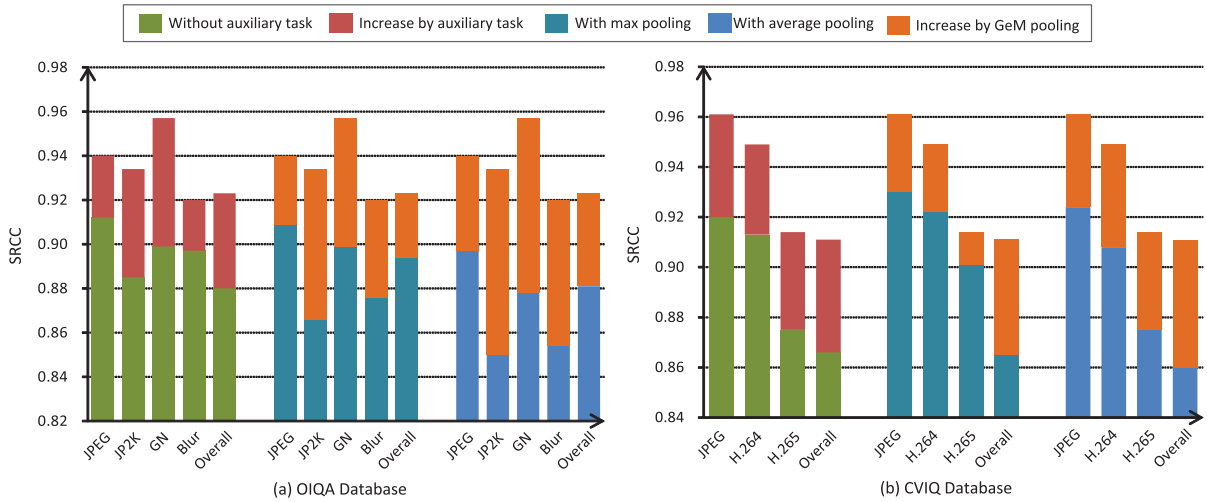


Fig. 6. SRCC values of the proposed method with and without the auxiliary task, and with max pooling and average pooling methods replacing the GeM pooling method, respectively.

TABLE V
CROSS-DATABASE PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH ALL THE TRAINING-BASED METHODS

Train	Test	Criterion	Metric										
			[15]	[18]	[50]	[51]	[53]	[54]	[55]	[31]	[5]	[29]	**Proposed**
CVIQ	OIQA	PLCC	0.385	0.423	0.410	0.405	0.331	0.604	0.583	0.601	0.627	0.705	**0.735**
		SRCC	0.205	0.308	0.273	0.260	0.192	0.551	0.502	0.543	0.601	0.684	**0.741**
		RMSE	13.274	13.087	13.118	13.186	13.576	11.399	11.747	11.502	11.038	10.178	**9.712**
OIQA	CVIQ	PLCC	0.636	0.752	0.810	0.693	0.586	0.688	0.630	0.705	0.726	0.823	**0.847**
		SRCC	0.604	0.701	0.749	0.616	0.548	0.624	0.587	0.657	0.705	0.814	**0.825**
		RMSE	10.862	9.352	8.254	10.025	11.403	10.145	10.904	9.980	9.588	7.811	**7.721**

the CVIQ database is first used for model training and the OIQA database is for model test. Then, two databases are swapped for model training and test. Table V shows the experimental results. From this table, we can see that the proposed model trained on the CVIQ database achieves the highest PLCC and SRCC values but the lowest RMSE value when it is tested on the OIQA database. For the other cross-database approach, the same conclusion can be obtained. These results demonstrate the best generalization ability of the proposed method. Besides, it can be found that no matter for the proposed method or the state-of-the-arts, the models trained on the OIQA database all outperform the models trained on the CVIQ database. This case is reasonable and explicable, because the CVIQ database only contains images with compression distortions, while the OIQA database also includes images with blur and noise distortions in addition to the images with compression distortions. Therefore, the models trained on the CVIQ database cannot accurately evaluate the alien distortions in the OIQA database, resulting in the inferior performance. This inspires that more diverse samples will help to learn a more generalized model.

IV. CONCLUSION

In this paper, we have proposed a multi-stream framework for OI quality assessment by evaluating the quality of viewport images that are truly received by human eyes in the VR experience. Specifically, with the generated viewport images, a parallel multi-stream network is constructed. Furthermore, the multi-task learning idea is applied to better train the quality evaluation model by simultaneously mining the interactive information with the distortion discrimination task. Both the viewport images based multi-stream framework and the multi-task learning idea are consistent with the human perception process and also contribute to solving the problem of the limited data amount. For performance verification, thirteen state-of-the-art quality metrics have been compared on two publicly released OI databases. The experimental results have demonstrated the advantages of the proposed method over existing traditional quality metrics and those specific for OIs regardless of the distortion type. Furthermore, the proposed method has also been proved to have the best generalization ability.

REFERENCES

[1] H. G. Kim, H.-T. 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.

[2] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. Conrad Bovik, "Study of 3D virtual reality picture quality," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 89–102, Jan. 2020.

[3] *Summary Surv. Virtual reality*, document ISO/IEC JTC 1/SC 29/WG 11 N16542, The MPEG Virtual Reality Ad-hoc Group, 2016.

[4] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3516–3530, Dec. 2019.

[5] 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.

[6] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang, and X. Gao, "No-reference quality assessment for view synthesis using DoG-based edge statistics and texture naturalness," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4566–4579, Sep. 2019.

[7] Q. Wu, H. Li, K. N. Ngan, and K. Ma, "Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2078–2089, Sep. 2018.

[8] 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.

[9] Q. Jiang, W. Zhou, X. Chai, G. Yue, F. Shao, and Z. Chen, "A full-reference stereoscopic image quality measurement via hierarchical deep feature degradation fusion," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9784–9796, Dec. 2020.

[10] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[11] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[12] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4280, Aug. 2014.

[13] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Fourier transform-based scalable image quality measure," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3364–3377, Aug. 2012.

[14] J. Wu, W. Lin, G. Shi, and A. Liu, "Reduced-reference image quality assessment with visual information fidelity," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1700–1705, Nov. 2013.

[15] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[16] Q. Jiang *et al.*, "Blind image quality measurement by exploiting high-order statistics with deep dictionary encoding network," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7398–7410, Oct. 2020.

[17] Q. Jiang, F. Shao, W. Gao, Z. Chen, G. Jiang, and Y.-S. Ho, "Unified no-reference quality assessment of singly and multiply distorted stereoscopic images," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1866–1881, Apr. 2019.

[18] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[19] X. Y. Xiu, Y. W. He, Y. Ye, and B. Vishwanath, "An evaluation framework for 360-degree video compression," in *Proc. VCIP*, 2017, pp. 1–4.

[20] 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.

[21] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, Fukuoka, Japan, Oct. 2015, pp. 31–36.

[22] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," *Proc. SPIE Optics Photon. Inf. Process.*, vol. 9970, Oct. 2016, Art. no. 99700C.

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

[24] W. Sun, K. Gu, S. W. Ma, W. H. Zhu, N. Liu, and G. T. Zhai, "A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process.*, Mar. 2018, pp. 1–6.

[25] W. Sun, K. Gu, G. Zhai, S. Ma, W. Lin, and P. Le Calle, "CVIQD: Subjective quality evaluation of compressed virtual reality images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3450–3454.

[26] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, "Spherical structural similarity index for objective omnidirectional video quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2018, pp. 1–6.

[27] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "BLIQUE-TMI: Blind quality evaluator for tone-mapped images based on local and global feature analyses," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 323–335, Feb. 2019.

[28] Q. Wu *et al.*, "Blind image quality assessment based on rank-order regularized regression," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2490–2504, Nov. 2017.

[29] 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.

[30] H.-T. 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. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6737–6741.

[31] S. Ling, G. Cheung, and P. Le Callet, "No-reference quality assessment for stitched panoramic images using convolutional sparse coding and compound feature selection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

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

[33] H. Jiang *et al.*, "Cubemap-based perception-driven blind quality assessment for 360-degree images," *IEEE Trans. Image Process.*, vol. 30, pp. 2364–2377, 2021.

[34] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1724–1737, May 2021, doi: 10.1109/TCSVT.2020.3015186.

[35] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3741–6737.

[38] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn,* 2015, pp. 448–456.

[40] G. A. Anastassiou, "Univariate hyperbolic tangent neural network approximation," *Math. Comput. Model.*, vol. 53, nos. 5–6, pp. 1111–1132, Mar. 2011.

[41] L. Jiao and J. Zhao, "A survey on the new generation of deep learning in image processing," *IEEE Access*, vol. 7, pp. 172231–172263, 2019.

[42] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images," in *Proc. IEEE Int. Symp. Circuits Syst.*, Oct. 2018, pp. 1–5.

[43] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consum. Electron.*, vol. 38, no. 1, pp. 33–34, Feb. 1992.

[44] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[45] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[46] L. Li, Y. Zhou, K. Gu, W. Lin, and S. Wang, "Quality assessment of DIBR-synthesized images by measuring local geometric distortions and global sharpness," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 914–926, Apr. 2018.

[47] L. Li, Y. Zhou, K. Gu, Y. Yang, and Y. Fang, "Blind realistic blur assessment based on discrepancy learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3859–3869, Nov. 2020.

[48] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang, and B. Chen, "No-reference quality assessment of deblocked images," *Neurocomputing*, vol. 177, pp. 572–584, Feb. 2016.

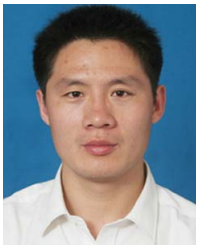
[49] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5462–5474, Nov. 2017.

[50] Y. Zhang and D. M. Chandler, "An algorithm for no-reference image quality assessment based on log-derivative statistics of natural scenes," in *Proc. Image Qual. Syst. Perform.*, Feb. 2013, pp. 16–21.

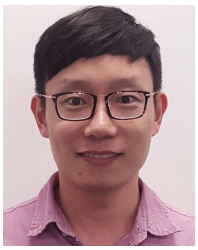
[51] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[52] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.
[53] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.
[54] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-End blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.
[55] 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.
[56] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. Int. Conf. Neural Inf. Process. Syst. Workshop*, 2017, pp. 1–6.
[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.
[58] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2049–2062, Aug. 2018.
[59] Y. Zhan and R. Zhang, "No-reference JPEG image quality assessment based on blockiness and luminance change," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 760–764, Jun. 2017.
[60] Y. Zhou, L. Li, J. Wu, K. Gu, W. Dong, and G. Shi, "Blind quality index for multiply distorted images using biorder structure degradation and nonlocal statistics," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3019–3032, Nov. 2018.

Yu Zhou received the B.S. and Ph.D. degrees from the China University of Mining and Technology, Xuzhou, China, in 2014 and 2019, respectively. She is currently an Assistant Professor with the School of Information and Control Engineering, China University of Mining and Technology. Her research interests include multimedia quality assessment and perceptual image processing.



Yanjing Sun received the Ph.D. degree in information and communication engineering from the China University of Mining and Technology in 2008. Since 2012, he has been a Professor with the School of Information and Control Engineering, China University of Mining and Technology, where he is currently the Director of the Network and Information Center. His current research interests include wireless communication, the Internet of Things, embedded real-time systems, wireless sensor networks, and cyberphysical systems. He is also a Council Member of the Jiangsu Institute of Electronics and a member of the Information Technology Working Committee of the China Safety Production Association.



Leida Li (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. In 2008, he was a Research Assistant with the Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. He is currently a Professor with the School of Information and Control Engineering, China University of Mining and Technology, China, and also with the School of Artificial Intelligence, Xidian University. His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics. He has served as an SPC for IJCAI 2019-2020, the Session Chair for ICMR in 2019 and PCM in 2015, and on the TPC for AAAI in 2019, ACM MM 2019-2020, ACM MM-Asia in 2019, ACII in 2019, and PCM in 2016. He is also an Associate Editor of the *Journal of Visual Communication and Image Representation* and the *EURASIP Journal on Image and Video Processing*.

Ke Gu (Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015, respectively. He is currently a Professor with the Beijing University of Technology, Beijing, China. His current research interests include environmental perception, image processing, quality assessment, and machine learning. He was a recipient of the Best Paper Award from the IEEE TRANSACTIONS ON MULTIMEDIA (T-MM), the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo (ICME), in 2016, and the Excellent Ph.D. Thesis Award from the Chinese Institute of Electronics, in 2016. He was the Leading Special Session Organizer in the VCIP 2016 and the ICIP 2017, and serves as a Guest Editor for the *Digital Signal Processing* Journal. He is currently an Associate Editor for IEEE ACCESS and *IET Image Processing (IET-IPR)*, and an Area Editor for the *Signal Processing: Image Communication (SPIC)*. He is a reviewer for 20 top SCI journals.

Yuming Fang (Senior Member, IEEE) received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree from Nanyang Technological University, Singapore. He is currently a Professor with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include visual attention modeling, visual quality assessment, computer vision, and 3D image/video processing. He serves on the Editorial Board for *Signal Processing: Image Communication*.