Attentive Deep Image Quality Assessment for Omnidirectional Stitching

Huiyu Duan, Xiongkuo Min, *Member, IEEE,* Wei Sun, Yucheng Zhu, Xiao-Ping Zhang, *Fellow, IEEE*, and Guangtao Zhai, *Senior Member, IEEE*

***Abstract*—Omnidirectional images or videos are commonly generated via the stitching of multiple images or videos, and the quality of omnidirectional stitching strongly influences the quality of experience (QoE) of the generated scenes. Although there were many studies research the omnidirectional image quality assessment (IQA), the evaluation of the omnidirectional stitching quality has not been sufficiently explored. In this paper, we focus on the IQA for the omnidirectional stitching of dual fisheye images. We first establish an omnidirectional stitching image quality assessment (OSIQA) database, which includes 300 distorted images and 300 corresponding reference images generated from 12 raw scenes. The database contains a variety of distortion types caused by omnidirectional stitching, including color distortion, geometric distortion, blur distortion, and ghosting distortion, *etc.* A subjective quality assessment study is conducted on the database and human opinion scores are collected for the distorted omnidirectional images. We then devise a deep learning based objective IQA metric termed Attentive Multi-channel IQA Net. In particular, we extend hyper-ResNet by developing a subnetwork for spatial attention and propose a spatial regularization item. Experimental results show that our proposed FR and NR models achieve the best performance compared with the state-of-the-art FR and NR IQA metrics on the OSIQA database. The OSIQA database as well as the proposed Attentive Multi-channel IQA Net will be released to facilitate future research.**

***Index Terms*—Image quality assessment, image stitching, omnidirectional image, virtual reality.**

I. Introduction

VIRTUAL Reality (VR) refers to technologies that aims at providing the users with simulated/expected experience of the real world. With the help of head-mounted displays (HMDs), users can experience omnidirectional contents, which provides realistic and immersive visual experience. As an important part of VR contents, omnidirectional images/videos, *a.k.a,* 360-degree images/videos, can reproduce the omnidirectional visual experience of the real world, which has attracted a lot of attention. Omnidirectional images are generally obtained by capturing images with overlapping areas and stitching them together to cover the whole field of view (FOV). Thus, omnidirectional image stitching is an important aspect of generating omnidirectional images. To cover the entire FOV, at least two fisheye lenses are needed, and dual-fisheye camera is the most popular and economical equipment to capture omnidirectional images or videos. Therefore, dual fisheye image stitching is also a very important research topic.

Many panoramic image stitching algorithms have been proposed in literature [1]–[3], and some stitching methods for dual fisheye images have also been studied [4], [5]. Nevertheless, the quality assessment of the stitched panoramic images, especially omnidirectional images, has rarely been studied. Such assessment can help benchmark, compare, and even improve various image stitching algorithms. Moreover, stitching quality assessment is an important part of omnidirectional image quality assessment (IQA). Serious local stitching distortions would have severe impact on the entire image quality. Thus it is important and significant to study the IQA of omnidirectional stitching. In this paper, we mainly focus on the subjective and objective quality assessment of omnidirectional image stitching in the context of VR applications.

There are some studies related to stitching IQA [6]–[8]. They mainly focus on the presence of specific artifacts such as photometric and geometric distortions. Qureshi *et al.* [7] proposed to measure the geometric error by calculating the structural similarity (SSIM) index between the high frequency information of the stitched and unstitched images. Bellavia *et al.* [9] extended the work of Xu *et al.* [6] and used feature similarity index (FSIM) to evaluate the color differences. However, these IQA models only study 2D image stitching and only consider one specific stitching distortion, such as geometric distortion or color distortion.

Recently, a few subjective and objective studies for omnidirectional image/video quality assessment (I/VQA) have been conducted. Sun *et al.* [10] established a compressed VR images database and conducted a subjective IQA study. Duan *et al.* [11] studied subjective and objective quality assessment of omnidirectional images with four commonly encountered distortions, including JPEG compression, JPEG2000 compression, Gaussian blur and Gaussian noise. They found that humans prefer high frequency contents and image details in

Manuscript received September 2, 2022; revised January 15, 2024; accepted February 9, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62225112, Grant 61831015, Grant 62271312, and Grant 62101326, in part by the National Key R&D Program of China 2021YFE0206700, in part by the Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102, in part by the Shanghai Pujiang Program under Grant 22PJ1407400, and in part by the China Postdoctoral Science Foundation 2022M712090. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mylène C.Q. Farias. *(Corresponding authors: Xiongkuo Min; Guangtao Zhai.)*

Huiyu Duan, Xiongkuo Min, Wei Sun, Yucheng Zhu, and Guangtao Zhai are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: huiyuduan@sjtu.edu.cn; minxiongkuo@sjtu.edu.cn; sunguwei@sjtu.edu.cn; zyc420@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn).

Xiao-Ping Zhang is with Tsinghua Berkeley Shenzhen Institute, China, and the Department of Electrical, Computer & Biomedical Engineering, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada. (e-mail: xzhang@ee.ryerson.ca).

VR HMDs. Duan *et al.* [12] also conducted subjective quality assessment experiments on omnidirectional videos with various degradations of bit rate, frame rate and resolution. Chen *et al.* [13] proposed to use structural similarity index (SSIM) on spherical space for the FR evaluation of omnidirectional video quality. Sun *et al.* [14] proposed a deep neural network based model for the NR quality assessment of omnidirectional images. All these studies considered the distortions introduced during acquisition, transmission, and display. However, they did not consider stitching distortions. As an important part of omnidirectional IQA, local stitching distortions such as color distortion, ghosting distortion [15]–[17], and blur distortion may have a huge influence on the quality of the entire omnidirectional image. In addition, stitching distortions such as geometric distortion not only affect the image quality but also aggravate the motion sickness [18], [19] in VR environment, making the quality of experience (QoE) worse.

Towards the assessment of QoE for VR, some studies conducted stitching quality assessment using HMD. Yang *et al.* [20] established a stitching quality assessment dataset and designed a ghosting and structure inconsistency based model to evaluate the quality of image. However, the stitched images they used to evaluate are 2D plane images, while they conducted the subjective experiment under the VR environment. Their proposed metric is raised for 2D stitched images. Madhusudana *et al.* [21] also conducted a subjective 2D panoramic stitching quality assessment study under VR environment and proposed a Gaussian mixture model to capture ghosting artifacts. Similarly, their study is not omnidirectional stitching IQA. Stitching dual-fisheye images for generating the omnidirectional image has more steps than common plane image stitching and faces more serious distortions in stitching area. Li *et al.* [22] proposed a cross-reference stitching quality assessment method and established a FR IQA database for dual fisheye image stitching. However, this study mainly focused on the global color distortion. The distortion types are not broad.

In this paper, we build a new omnidirectional stitching image quality assessment (OSIQA) database towards both FR and NR IQA for omnidirectional stitching, and propose a FR-IQA metric and a NR-IQA metric for evaluating omnidirectional stitching distortions. We follow the method proposed in [22] and capture 12 sets of cross-reference raw fisheye images (each set contains four fisheye images, two for the distorted omnidirectional image generation, another two for the reference omnidirectional image generation). Based on these 12 sets of raw images, 300 distorted omnidirectional images with various stitching distortions and 300 corresponding reference images are obtained using various stitching algorithms. Then we conduct a large scale subjective quality assessment study and collect 40 human opinion scores (20 for the front stitching area, 20 for the back stitching area, see Section III-C for detailed information) for each distorted image. Next, an OSIQA-FR model and an OSIQA-NR model based on Attentive Mult-channel IQA Net are proposed for better evaluating the quality of omnidirectional stitching. Specifically, we propose an Attentive Multi-channel IQA Net which uses multi-field of view (FOV) images as the input to the network, and adopt a hyper-ResNet structure combined with a subnetwork for spatial attention to extract features. Moreover, a spatial regularization loss has been designed to further restrict the training process. The experimental results show that our proposed model achieves the best performance among all state-of-the-art FR and NR IQA metrics.

Overall, the main contributions of this paper are summarized as follows. (i) We establish an OSIQA database with various stitching distortions, including color distortion, geometric distortion, blur distortion, and ghosting distortion. (ii) We propose an Attentive Multi-channel IQA Net which adopt hyper-ResNet and spatial attention subnetwork to extract features, and then regress to predict the image quality. A spatial regularization method is proposed to further restrict the training process. (iii) A OSIQA-FR model and a OSIQA-NR model are proposed based on the Attentive Multi-channel IQA Net. (iv) Extensive FR and NR IQA experiments are conducted for omnidirectional stitching, and experimental results demonstrate the superiority of the proposed models.

The rest of the paper is organized as follows. In Section II, we briefly reviewed related works. Section III describes the construction procedure of the OSIQA database, including distortion generation and subjective data collection. In Section IV, we introduce the proposed Attentive Multi-channel IQA Net in detail. The experimental validation process is presented in Section V. Section VI conclude the whole paper.

II. Related Works

A. Classical IQA Index and Learnable IQA Index.

In terms of FR IQA methods, many classical metrics have been proposed and widely used [23]–[27], including mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) index [28], feature similarity (FSIM) index [29], *etc.* Regarding NR IQA indexes, there are also many popular methods such as natural image quality evaluator (NIQE) [30], blind quality assessment based on pseudo-reference image (BPRI) [31], and NR free-energy based robust metric (NFERM) [32], *etc.*

Driven by the rapid development of deep neural networks (DNNs) recently, some learning based IQA algorithms have been proposed. Kang *et al.* [33] proposed to use multi-task convolutional neural networks to evaluate the image quality and use 32×32 patches for training. Bosse *et al.* [34] proposed both FR and NR IQA metrics by joint learning of local quality and local weights. Some studies located specific distortions by using convolutional sparse coding and then evaluated the image quality [35].

B. Omnidirectional Image Quality Assessment

Omnidirectional IQA has been broadly studied recently. A few databases including commonly encountered distortions (such as coding, compression, etc) for omnidirectional IQA have been established, and corresponding subjective studies have been conducted [10]–[12], [37]. There are some objective metrics for omnidirectional IQA have been porposed. Yu *et al.* [38] proposed to calculate PSNR on the points uniformly distributed on the sphere and named the method sphere

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2023.3250956

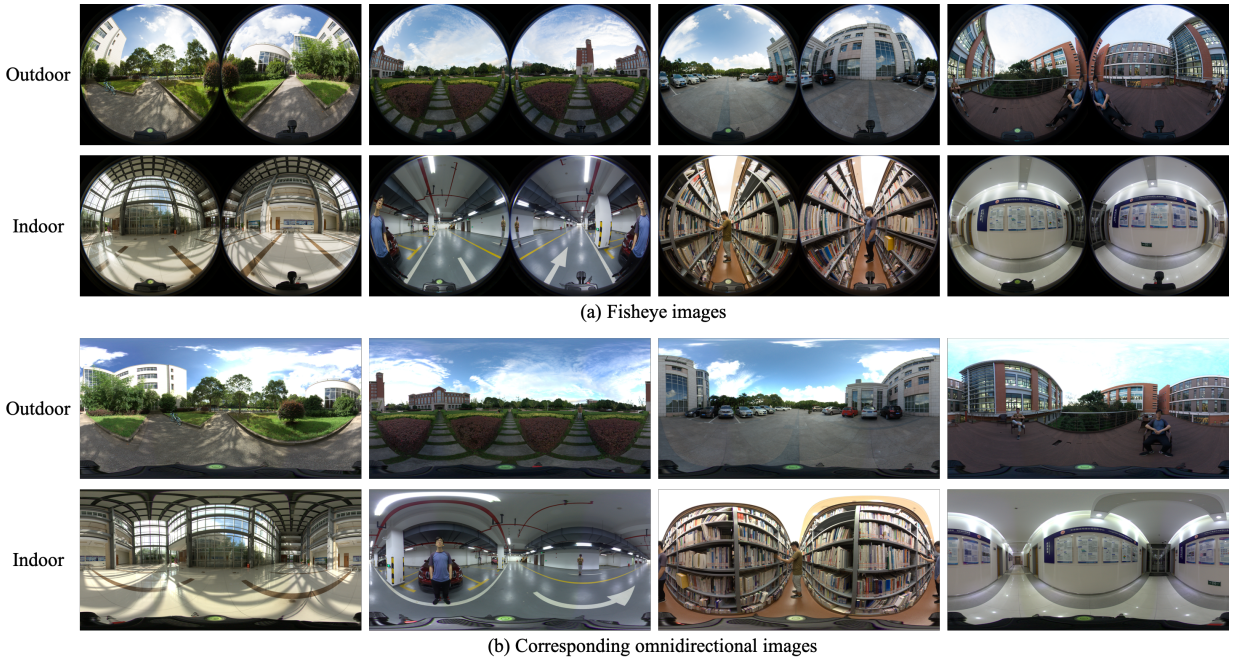


Fig. 1. Examples of raw fisheye images and corresponding optimal stitched omnidirectional images in our OSIQA dataset. The first row in each sub-figure shows outdoor scenes, and the second row in each sub-figure shows indoor scenes. (a) Raw fisheye images captured by the fisheye lens. (b) Corresponding stitched omnidirectional images with the best perceptual quality in our dataset (obtained from the stitching method provided by Insta360 stitching software [36]).

based PSNR (S-PSNR). A Weighted Spherical PSNR (WS-PSNR) metric has been proposed by Sun *et al.* [39]. The weights were represents by how much the sampled area is stretched. Xu *et al.* [40] presented to use attention in the process of omnidirectional VQA, and proposed a non-content-based P-VQA (NCP-PSNR) index and a content-based P-VQA (CP-PSNR) index. Moreover, a spherical SSIM method for omnidirectional VQA has been proposed by Chen *et al.* [13]. Kim *et al.* [41] split the omnidirectional image into small patches and used adversarial network to estimat the local quality and the weight of each pathc. Sun *et al.* [14] proposed a mulit-channel IQA metric based on deep learning and got great performance for omnidirectional IQA. However, omnidirectional image stitching quality assessment has rarely been studied in previous literature.

C. Stitching Image Quality Assessment

Previous stitching IQA mainly focus on some specific artifacts, such as color distortion and ghosting, and concentrated on plane panoramic images. Yang *et al.* [20] proposed a ghosting and structure inconsistency based model to evaluate the quality of stitched images. Ling *et al.* [42] presented a method using convolutional sparse coding and compound feature selection to capture and evaluate stitching distortions. Madhusudana *et al.* [21] conducted a 2D panoramic stitching quality assessment study under VR environment and proposed a Gaussian mixture model to capture ghosting artifacts. All these studies are conducted towards 2D panoramic stitching quality assessment rather than omnidirectional stitching quality assessment. Zhu *et al.* [43] created a video benchmark for various blending algorithms during stitching. They mainly focused on evaluating different blending algorithms. Li et al. [22] established a cross-reference dual fisheye stitching database and proposed two omnidirectional stitching image quality assessment (OS-IQA) metrics for stitched regions and the whole stitched images respectively. However, their database mainly focused on the color distortion, other distortions such as geometric distortion and ghosting are not considered.

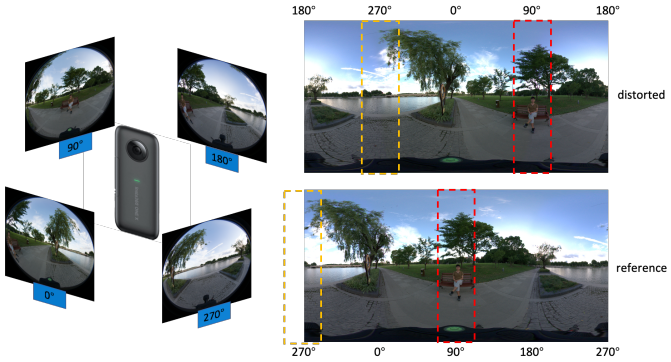


Fig. 2. Cross-reference stitching method. For each scene, we capture 4 fisheye images by 2 shoots. The fisheye images in 0° and 180° directions are stitched to generate the distorted omnidirectional images. The fisheye images in 90° and 270° directions are stitched to generate the reference omnidirectional images. The dashed boxes in the bottom-right reference image can provide high-quality ground-truth information for the dashed stitched areas in the top-right distorted image.

III. OSIQA Database

To address the absence of omnidirectional stitching quality assessment studies, we first establish an omnidirectional stitching IQA (OSIQA) database. In this section, we introduce

our OSIQA database construction methodology. First, we introduce our raw cross-reference image collection procedure. Then, omnidirectional stitching quality degradation method as well as distortion types are presented. Next, the experimental methodology to perform quality assessment is introduced. At last, we process these subjective quality scores to obtain the mean opinion scores (MOSs) and analyze their distribution. To the best of our knowledge, the OSIQA database is the largest image quality assessment database related to omnidirectional stitching distortions. Our OSIQA database will be released to facilitate future studies.

A. Image Collection

Since the most common way to acquire omnidirectional images is by stitching two fisheye images, in this research, we use a consumer panoramic camera (Insta360 ONE X [36]) with dual fisheye lens to collect raw images. Each fisheye lens can take a fisheye image with the highest resolution of 3040×3040, and the resolution of the combined omnidirectional image can be up to 6080×3040. Since in this work, our main focus is towards modeling distortions during dual fisheye stitching, and we do not address artifacts introduced during capturing. Therefore, the highest resolution was taken during image collection procedure. We captured 12 sets of raw images using Insta360 ONE X in various scenarios to enhance the robustness of our dataset. These 12 various scenarios can be further classified into two categories: 1) outdoor: including street, park, residential area, *etc.*; and 2) indoor: including hall, underground parking, libraries, *etc.* Fig. 1 demonstrates examples of our raw fisheye images and stitched omnidirectional images with indoor and outdoor scenes, respectively.

To obtain the reference images of distorted images, here we applied a similar image collection method as described in [22]. On account of the dual fisheye lens of the camera, each shot can get two fisheye images in opposite directions. Thus, as shown in Fig. 2, for each scene, we captured this scene twice, and got 4 fisheye images with scenes in $0°$, $90°$, $180°$, $270°$ direction, respectively. Specifically, for each scene, firstly, we captured this scene in $0°$ and $180°$ directions. These two fisheye images are used to generate the distorted omnidirectional image. Then, we rotated the camera by 90 degrees, and captured the scene in $90°$ and $270°$ directions. And these two fisheye images are used to generate the reference omnidirectional image later. In this way, we can get the reference omnidirectional images of the distorted omnidirectional images. Thus, the stitched distortion areas (as shown in the rectangle dashed line areas of the top-right figure in Fig. 2) can find their corresponding captured reference areas with high-quality (as shown in the rectangle dashed line areas of the bottom-right figure in Fig. 2) as the ground-truth.

B. Omnidirectional Stitching Distortions

There are many studies related to dual-fisheye image stitching [4], [5]. Omnidirectional stitching involves a series of operation from dual-fisheye images with overlapping fields of view to the stitched equirectangular image. Generally, at least four steps are needed during this procedure, as shown in Fig. 3. First, captured dual-fisheye images are transformed into equirectangular format. In this step, fisheye lens parameters usually need to be estimated or provided. If the estimated or provided parameters are not explicit, it may affect the following stitching steps and cause geometric distortion. Then, features are extracted and matched from the acquired dual images after transforming. Next, according to the matched features acquired from the second step, these two images are warped to match each other. If the extracted features are not matched well, some distortions may be introduced during the image warping step. Finally, dual images are blended and output as the final equirectangular images. There are many sub-steps during this blending process, including exposure compensation, color correction, and blending strategies, *etc.* If the previous warped images are not matched well, blending algorithms may fail to perform well, and the output equirectangular images may have server stitching distortions. Each step in the stitching pipeline may introduce distortions and degrade the quality of stitched images, and distortions introduced in the early steps of stitching may accumulate and make the following process perform worse.

As discussed by Madhusudana *et al.* in [21], they only varied algorithms associated with the warping stage and the blending stage, and they found that the quality of stitched image was most sensitive to these two modules. Nevertheless, the difference between dual fisheye stitching and traditional panoramic stitching is that the images need to be transformed into equirectangular format before the subsequent stitching steps. Therefore, we additionally adjusted the input fisheye lens parameters during the equirectangular transformation process as the estimated or provided parameters may be different from the actual parameters of the camera. Specifically, we varied parameters associated with field of view (FOV) and projection method during the equirectangular transformation steps and varied blending strategies during the blending step. For the variation of FOV, since the FOV provided by Insta360 ONE X [36] is $200°$, we varied the input FOV around this value and chose it from $190°, 195°, 200°$, and $205°$. For the projection method from fisheye to equirectangular format, we tried four fisheye mapping method, including stereographic projection, equidistant projection, equisolid angle projection, and orthographic projection. Moreover, we also changed blending strategies as discussed in [43]. In this work, six blending algorithms are used to generate equirectangular images from warped images, including copy-and-paste [44], feature blending [43], multi-band blending [43], poisson blending [45], convolution pyramid blending [43], modified poisson blending [46], and multi-spline blending [47]. The backbone of dual fisheye images stitching method we used is transferred from an open source code of dual fisheye video stitching repository[1]. Then we varied the relevant methods in corresponding steps of the stitching pipeline as discussed above. Furthermore, we used three stitching software to generate equirectantular images including one built-in stitching software from Insta360 (Insta360 STUDIO [36]) and two commercial stitching sofwares (PTGui [48] and Easypano [49]).

[1]https://github.com/cynricfu/dual-fisheye-video-stitching

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2023.3250956

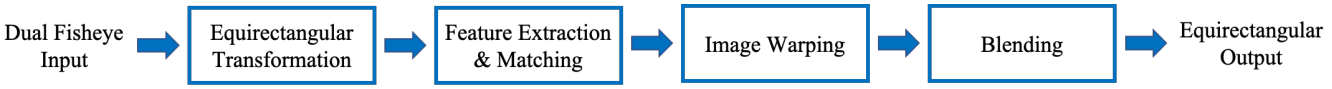


Fig. 3. General dual fisheye stitching pipeline. To generate omnidirectional images from dual fisheye inputs, four steps are usually needed.

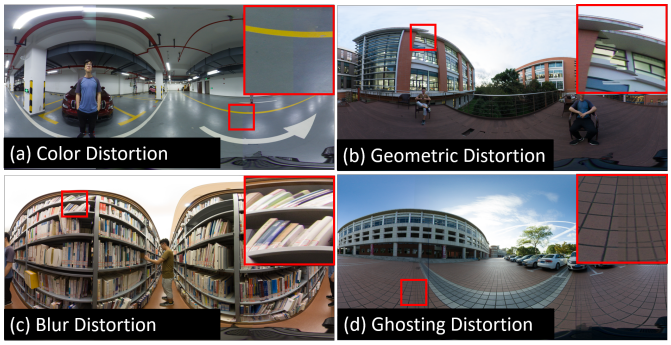


Fig. 4. Example images with various distortions due to omnidirectional stitching. The red rectangle in the upper right corner of each sub-figure shows the zoom-in view.

Finally, we obtained 25 stitched equirectangular images as distorted images and 25 stitched equirectangular images as reference images for each scene. As mentioned before, we captured 12 scenes during the image collection process. Thus, we eventually obtained 300 distorted images and 300 reference images in total. As shown in Fig. 4, four types of distortions are commonly introduced during the stitching process, including color distortion, blur distortion, geometric distortion, and ghosting distortion. Under certain circumstances, the distortion in the stitched equirectangular images may even be a combination of these four distortions. Fig. 4 (a) shows the case of color distortion. Color distortion occurs when the input dual-fisheye images have different exposure levels and the color correction step during blending procedure may not perform well. Fig. 4 (b) shows the geometric distortion. Geometric distortion occurs due to seriously inaccurate matching of feature points so that geometric distortion is introduced during image warping stage. Then the blending algorithm fails to be carried out. Fig. 4 (c) and Fig. 4 (d) show the cases of blur distortion and ghosting distortion, respectively. Blur distortion and ghosting distortion occur due to slightly inaccurate matching of extracted feature points in the overlapping regions of two input images. These misalignments are reflected in the blending stage, and then cause the blur or ghosting distortion in stitched equirectangular images. These distortions in the equirectangular image may even be magnified when viewed as an omnidirectional image in VR-HMD and make the quality of experience worse.

C. Subjective Experiment Methodology

After acquiring equirectangular images with distortions, we conducted subjective quality assessment experiment to obtain subjective quality rating scores of these images. Several subjective quality assessment methodologies have been recommended by ITU-R BT500-11 [50], including single-stimulus (SS), double-stimulus impairment scale (DSIS) and paired comparison (PC). Since in the VR-HMD, only one omnidirectional image can be seen at one time, single-stimulus continuous quality evaluation (SSCQE) procedure were employed to obtain subjective quality ratings for the distorted images in our OSIQA dataset.

We used HTC VIVE Pro [51] as the HMD to display omnidirectional images on account of its excellent graphic display and high precision tracking ability. We designed an interaction system using Unity3D software to display omnidirectional images and collect subjective quality scores automatically. All 300 distorted equirectangular images were first converted to omnidirectional images in Unity3D and then displayed in HMD. The subjects can use the controller to switch images and select quality scores inside HMD during the subjective experiment. A 10 point numerical categorical rating method was adopted to obtain subjective ratings. The higher value means the better quality. Unity3D as well as HMD were run on a computer with 4.00GHz Intel Core i7 processor, 16GB main memory, and an Nvidia GeForce GTX 1080 graphics card.

Xu *et al.* [40] suggest that at least 15 subjects are required to conduct subjective quality assessment for VR contents. In this work, 20 subjects were recruited to participant in our subjective experiment. The ages of participants ranged from 18 to 30. All subjects had normal or correct-to-normal visual acuity. Since our subjective assessment experiment was conducted under HMD environment, unlike other subjective experiments conducted on the traditional displays, we do not need to consider the viewing conditions, such as viewing distance [52], ambient luminance [53], *etc.* The experiment was conducted in an empty room without noise to avoid interference.

Before starting the experiment, each subject was individually explained the goal of this experiment and a short training session including 20 images was conducted to familiarize the subjects with the rating procedure and the distortions. The subjects were instructed to provide subjective ratings based on the perceptual quality of the images rather than the aesthetics of the images. We adjusted the focus of lens before the training session and kept it during the whole experiment to provide the best QoE. As shown in Fig. 2, dual fisheye stitching would generate two areas with apparent stitching distortions. Therefore, we divided the experiment into two sessions. For the first session, subjects were seated facing the stitched area in the 90° direction (red rectangle in Fig. 2) and evaluate all 300 images. For the second session, subjects were seated facing the stitched area in the 270° direction (yellow rectangle in Fig. 2) and evaluate all 300 images again. During the experiment, subjects were allowed to rotate their head but not their body. In this case, subjects can only view and evaluate one stitched area during each session. Using this method, we would obtain 2 subjective quality scores for each omnidirectional image.

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2023.3250956

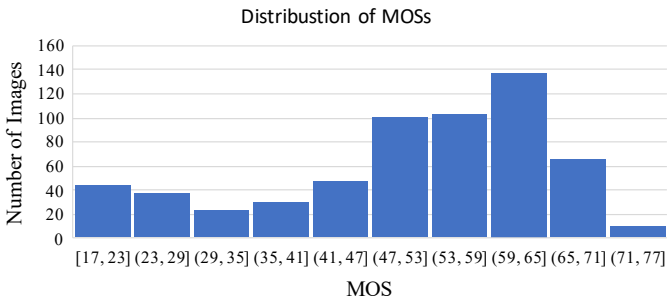


Fig. 5. Histogram of MOSs.

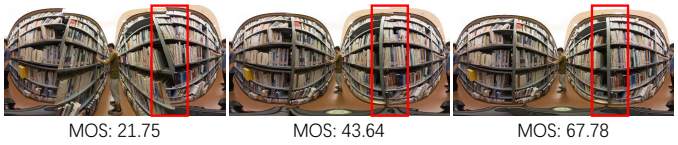


Fig. 6. Illustration of stitched omnidirectional images with different MOS values.

The images were shown in a random order for each subject. For each session, subjects were allowed to pause and resume the session whenever they felt fatigued and had enough rest. And they were required to rest for at least half an hour after evaluating every 100 images. Finally, we collected 12000 ($300 \times 2 \times 20$) subjective quality assessment scores in total for further analysis.

D. Data Processing and Analysis

After collecting subjective quality scores, we process these data to obtain MOSs. We follow the recommendations as detailed in [50] to exclude outliers and reject subjects. Rating for an image is considered as outlier if it is outside 2 (if normal) or $\sqrt{20}$ (if non-normal) standard deviations (stds) about the mean rating of that image. Subjects with more than 5% outlier ratings are rejected. In our experiments, two subjects are rejected. Since for each omnidirectional image, we get two ratings (one for the left half image, one for the right half image) from each subject. Mean opinion score is calculated for each half image (left or right). Let m_{ij} denote the subjective scores provided by subject i to image j and N_i denote the number of the images evaluated by subject i. We first normalize the ratings of each subject by using:

$$z_{ij} = \frac{m_{ij} - \mu_i}{\sigma_i}, \tag{1}$$

where

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} m_{ij}, \quad \sigma_i = \sqrt{\frac{1}{N_i - 1}\sum_{j=1}^{N_i}(m_{ij} - \mu_i)^2}. \tag{2}$$

Then the ratings for each image are averaged:

$$z_j = \frac{1}{N_j}\sum_{i=1}^{N_j} z_{ij}, \tag{3}$$

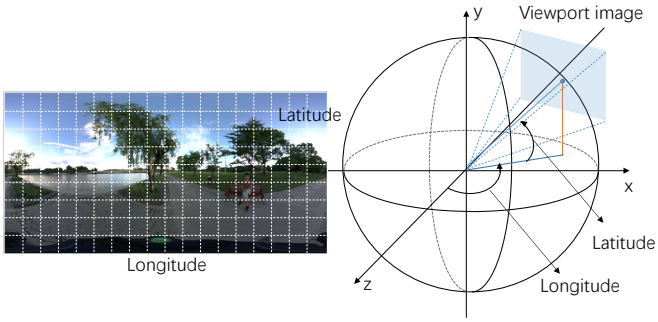


Fig. 7. Projection relationship between omnidirectional (spherical) image and equirectangular image. Left: an equirectangular image is projected on a sphere by equirectangular projection and displayed as omnidirectional image. Right: users can only see a viewport image at a certain head pose when viewing an omnidirectional image.

where N_j is the number of valid subjective ratings for image j (after outlier removing). Finally, MOS value is derived by linear rescaling to lie in the range of [0,100]:

$$MOS_j = \frac{100(z_j + 3)}{6}. \tag{4}$$

From the previous step, we obtain 600 MOSs for 300 distorted omnidirectional images (for each omnidirectional image, we get one MOS for the left half part, and one MOS for the right half part). Fig. 5 shows the histogram of MOSs indicating a reasonably wide distribution of MOS values. The MOSs lie in the range of [17, 77], which shows that the constructed database covers a wide range of perceptual quality scores. Moreover, The distribution of quality scores is relatively uniform, which further demonstrates the usefulness of the database. Fig. 6 illustrates three stitched omnidirectional images with different MOS values, which further intuitively shows the difference of stitched images with different quality. To get the quality of the entire omnidirectional image, we average the two MOSs for each omnidirectional image. Finally, we obtain 300 subjective quality ratings for 300 distorted omnidirectional images and further conduct objective quality assessment.

IV. Proposed Method

In this section, we propose our objective FR and NR quality assessment algorithms based on the subjective ratings obtained above. The overall flowchart of the proposed method is shown in Fig. 9. We first project the omnidirectional image to cubic multi-FOV images. Then data refinement and data augmentation method are employed on these cubic viewport images. Finally, through the proposed Attentive Multi-Channel IQA Net, we can get the predicted image quality. Moreover, during training stage, spatial regularization loss is adopted to further restrict training process. Detailed methods are introduced as follows.

A. Projection Method

Omnidirectional images are generally stored with equirectangular format on the computer. When displaying an equirectangular image in the HMD, it is first projected on a sphere

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2023.3250956

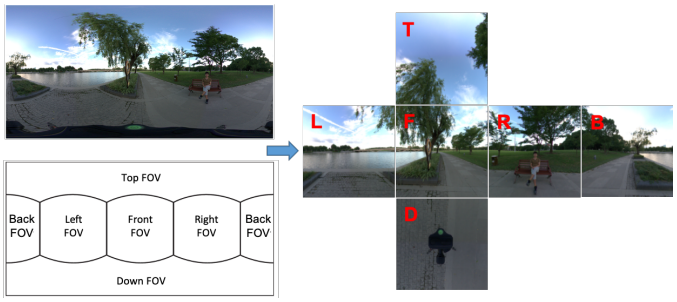


Fig. 8. The cubic viewport images and their corresponding parts in the omnidirectional image. L: Left View, F: Front View, R: Right View, B: Back View, T: Top View, D: Down View.

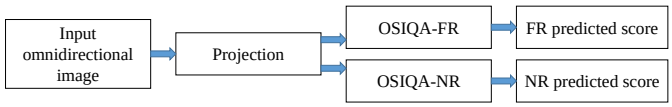


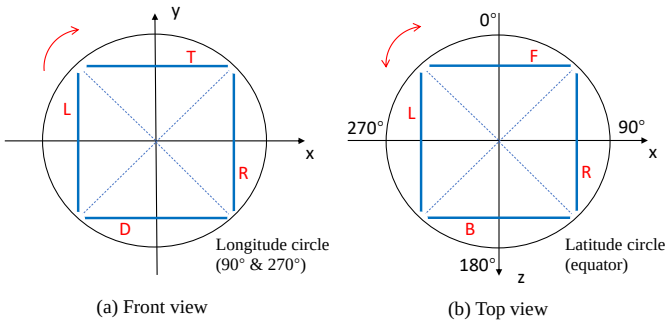
Fig. 9. The flowchart of the proposed method.

using equirectangular projection and shown as the omnidirectional image. Then the HMD can track the head movement of the user and display corresponding field of view (FOV). Fig. 7 shows the projection relationship between the equirectangular image and the omnidirectional image as well as the viewport image. Our target is evaluating the quality of omnidirectional image. However, it is hard to design image processing algorithm based on spherical image. And due to the geometric distortion of equirectangular image itself, it is not reasonable to directly evaluate the quality of the equirectangular image. Since the viewport images are the images viewed and assessed by the subjects, we propose to assess the quality of an omnidirectional image based on its corresponding viewport images.

We calculate each pixel of the viewport image through mapping it back to the omnidirectional (spherical) coordinate and then mapping back to the plane equirectangular image to find its best estimated pixel. The detailed procedure can be found in [38]. In this paper, we set the FOV as 90 degree, which is consistent with the FOV of most VR devices such as Oculus, HTC VIVE, *etc.* Moreover, six viewport images with 90 degree FOV in the left, right, front, back, top, and down orientations, respectively, can cover the full visual content of the omnidirectional image without overlapping. As shown in Fig. 8, an omnidirectional image can be converted to six cubic viewport images. In the following the paper, we use the symbols I_L, I_R, I_F, I_B, I_T, and I_D to represent the left, right, front, back, top, and down viewport images, respectively.

B. Data Refinement and Augmentation

As shown in Fig. 2, omnidirectional stitching distortions generally appear in directions near $90°$ and $270°$. And we have considered this in subjective experiments as discussed in Section III-C. As shown in Fig. 8, the FOVs of left, right, top, and down are in the $90°$ and $270°$ directions. Since the regions near $90°$ and $270°$ directions are the most important for the perceptual quality assessment, here we refine the viewport



(a) Front view (b) Top view

Fig. 10. Illustration of our data augmentation method. L: Left FOV, R: Right FOV, F: Front FOV, B: Back FOV, T: Top FOV, D: Down FOV. (a) Front view of the sphere in Fig. 7. (b) Top view of the sphere in Fig. 7. $0°, 90°, 180°, 270°$ represent the longitude angles.

images by choosing the viewports of the left, right, top, and down, and discarding the front and back FOVs. In other words, in this paper, only I_L, I_R, I_T, and I_D are used to assess the quality of omnidirectional images.

Our IQA model is based on deep neural networks. To avoid overfitting, we augment the data by rotating the viewing angle along the longitude and the latitude. This inspiration comes from the observation that users usually see multiple views from different directions and then give the final opinion score for an omnidirectional image. Specifically, we first rotate the FOV along the longitude of $90°$ and $270°$ ($90°$ longitude and $270°$ longitude are at the same longitude circle) from $0°$ to $90°$ with an interval of φ, as shown in Fig 10 (a). We only rotate it from $0°$ to $90°$ because that the viewport images of rotating $\alpha+90°$ are repeated with viewport images of rotating α degree, where α can be any angle. Then we rotate the FOV along the latitude from $-15°$ to $15°$ with an interval of γ (as shown in Fig. 10 (b) and then rotate along the longitude again to obtain another several sets of augmented data. Finally, we can get $M \times N$ groups of viewport images derived from one omnidirectional image. We denote them as $I_{FOV}^{i,j}$, where $FOV \in \{L\ (left),\ R\ (right),\ T\ (top),\ D\ (down)\}$, and $i \in \{1, 2, ..., M\}$, $M = 30/\gamma + 1$, as well as $j \in \{1, 2, ..., N\}$, $N = 90/\varphi$. Note that this data augmentation method is only used during the training stage.

C. Attentive Multi-Channel IQA Net

Deep convolution neural networks (CNN) have achieved impressive results in lots of computer vision tasks. Successful CNN models such as VGG [54], GoogleNet [55], ResNet [56] have been widely used to solve image recognition, detection, segmentation problems, *etc.* Deep neural networks have strong ability to extract high-level semantic features. Nevertheless, the quality assessment of the stitching distortions is related to both low-level and high-level features. Therefore, in this paper, we adopt ResNet as backbone CNN, and fuse the features from inter-mediate layers to accumulate features from low-level to high-level as discussed in [14], [57]. Moreover, we design a sub-network for spatial attention to further extract the features related to stitching distortions. Finally, we fuse these features extracted from different FOVs and get the estimated MOS. The detailed structure of each part is given as follows.

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2023.3250956

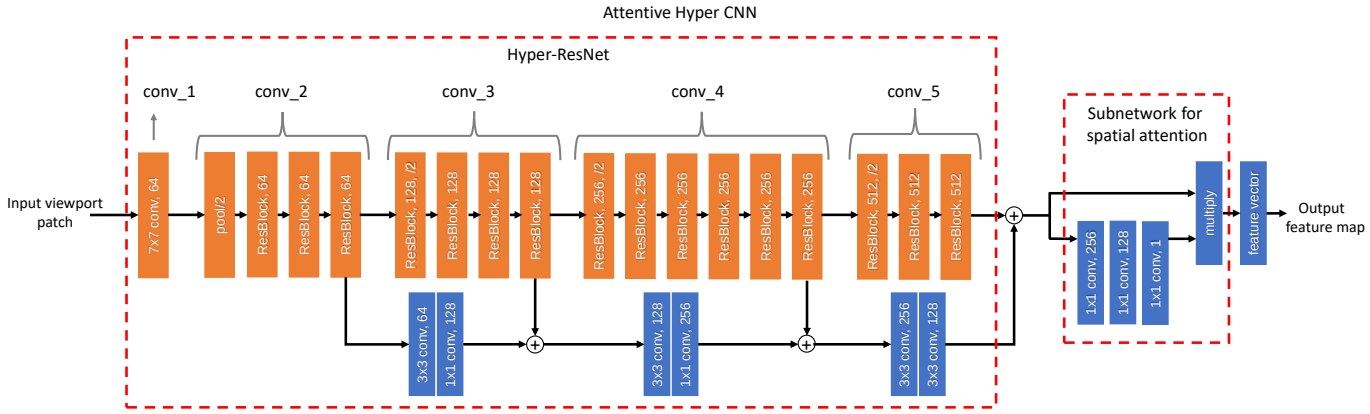


Fig. 11. Illustration of the structure of the Attentive Hyper CNN, which includes a hyper-resnet and a subnetwork for spatial attention.

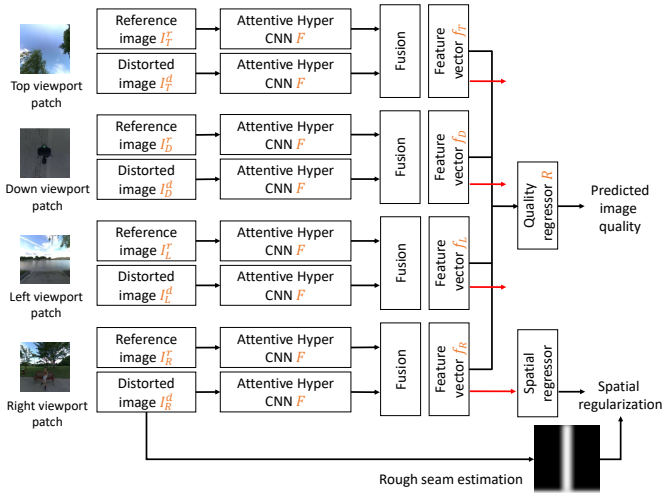


Fig. 12. Illustration of the proposed Attentive Multi-channel CNN framework for FR omnidirectional stitching image quality assessment, which includes 8 parallel Attentive Hyper CNNs with shared weights to extract features. The quality regressor and spatial regressor are demonstrated in Fig. 14.

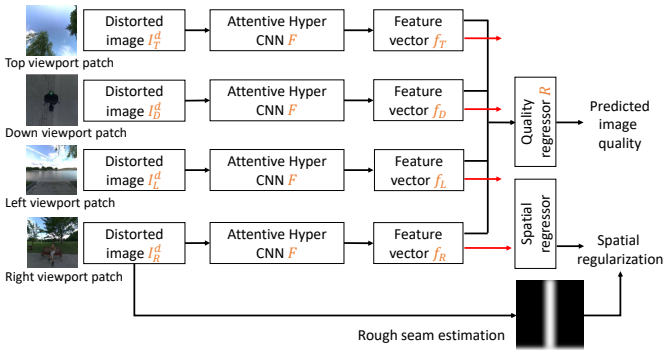


Fig. 13. Illustration of the proposed Attentive Multi-channel CNN framework for NR omnidirectional stitching image quality assessment, which includes 4 parallel Attentive Hyper CNNs with shared weights to extract features. The quality regressor and spatial regressor are shown in Fig. 14.

Fig. 11 illustrates the Attentive Hyper CNN for feature extraction. Each Attentive Multi-channel CNN consists two parts, one hyper-ResNet and one subnetwork for spatial attention. The backbone of hyper-Resnet is Resnet, which using residual learning to further deepen the CNN network [56]. ResNet has several architectures such as ResNet18, ResNet34, ResNet50, ResNet101, *etc.*, in accordance with the number of layers. In this work, we adopt ResNet34 as the backbone considering the efficiency and accuracy of the model. The ResNet34 includes five parts which are denoted as conv_1, conv_2, conv_3, conv_4, and conv_5, respectively in Fig. 11. The first conv_1 is a convolutional layer with 7×7 kernel size and 64 channels, and the stride is 2. The rest parts conv_2, conv_3, conv_4, and conv_5, consist of residual blocks [56]. As mentioned above, the evaluation of stitching distortions is not only related to high-level semantic features, but also related to low-level features such as edges, corners, *etc.* To take advantage of these low-level features, we use hyper-ResNet [57] to fuse the features from inter-mediate layers. The hyper-ResNet is followed by a spatial attention subnetwork [58], which is used to further refine the extracted features. The inspiration comes from the observation that the stitching distortions mainly occur in part of the image.

Fig. 12 and Fig. 13 illustrates the proposed Attentive Multi-channel CNN frameworks for FR and NR omnidirectional stitching image quality assessment, respectively. For each viewport patch I_{FOV}^l, where $FOV \in \{L\ (left),\ R\ (right),\ T\ (top),\ D\ (down)\}$ and $l \in \{d\ (distorted), r\ (reference)\}$, we first feed the patch into a Attentive Hyper CNN F to extract the feature vector f_{FOV}^l for this patch, which is formulated as:

$$f_{FOV}^l = F(I_{FOV}^l). \tag{5}$$

Then for the FR-IQA model, we fuse the two feature vectors for each viewport as follows:

$$f_{FOV} = f_{FOV}^d - f_{FOV}^r, \tag{6}$$

where d denotes the distorted patch, r denotes the reference patch, and $FOV \in \{L,\ R,\ T,\ D\}$. And for the NR-IQA model, we only use the distorted viewport feature vector as the viewport feature vector, which is denoted as:

$$f_{FOV} = f_{FOV}^d. \tag{7}$$

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2023.3250956

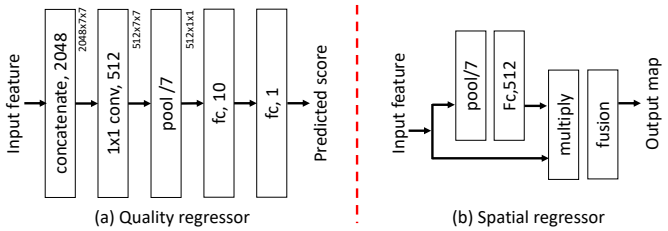


Fig. 14. Illustration of the quality regressor and spatial regressor.

Thus, we get the extracted feature vectors f_T, f_D, f_L, f_R for viewport images I_T, I_D, I_L, I_R, respectively. Then we feed these feature vectors to the regressor network which consists of one concatenate layer, one convolutional layer, one average pooling layer and two fully connected layers as shown in Fig. 14 (a). The convolutional layer here is used to aggregate the features. We analyze the importance of this layer in Section V-B3 and TABLE III and denote related results as "aggregation" item. The process can be represented by:

$$q_{predicted} = \mathbf{R}(f_T, f_D, f_L, f_R), \tag{8}$$

where $q_{predicted}$ denotes the predicted image quality, f_T, f_D, f_L, f_R denote the feature vectors extracted from the top, down, left, right FOV images, respectively.

For the end-to-end training, we follow the paper [14] which uses mean squared error (MSE) as the loss function. This loss function can be denoted as:

$$\mathcal{L}_{MSE} = (q_{predicted} - q_{labeled})^2, \tag{9}$$

where $\mathcal{L}_{MSE}$ denotes the MSE loss function, $q_{predicted}$ is the predicted score calculated by the Attentive Multi-channel IQA Net, and $q_{labeled}$ is the MOS derived from subjective experiments.

D. Spatial Regularization

To further restrict the training process, a spatial regularization term is designed. We follow the work in [59] and design a channel-wise subnet for feature fusion. Detailed network structure is demonstrated in Fig. 14 (b). Specifically, for each extracted feature vector, we first feed it to an average pooling layer then pass it to a fully connected layer to learn channel-wise weights. Next, we multiply the feature vector with the channel-wise weights. After fusing, we get the Class Activation Map (CAM). Then a spatial regularization term is designed based on common pixel-wise cross-entropy loss between the activation map (CAM) and the rough seam estimation map. Since in our database, stitching seam generally appears in a column of areas in the center of the FOV, we simply use this area as a rough seam estimation area. Let $p_{i,j}$ and $e_{i,j}$ denote the activation map and rough seam estimation map, respectively. Spatial regularization loss can be written as:

$$\mathcal{L}_{SR} = -\left[\frac{\sum_{i,j} e_{i,j} \log p_{i,j}}{\sum_{i,j} e_{i,j}} + \frac{\sum_{i,j}(1-e_{i,j})\log(1-p_{i,j})}{\sum_{i,j}(1-e_{i,j})}\right]. \tag{10}$$

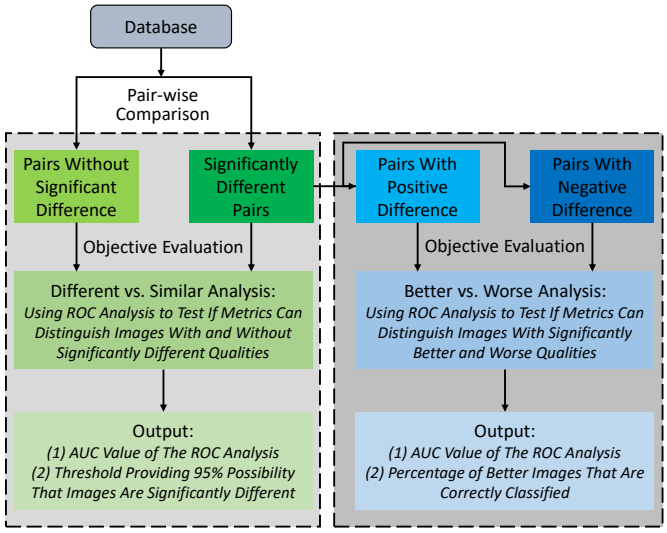


Fig. 15. A framework of the new metric evaluation methodology.

Our final objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_{SR}, \tag{11}$$

where the balance factor λ is empirically set as 0.025 to control the contributions of the two terms.

V. Experimental Validation

In this section, we introduce the experimental validation of the proposed method. We first present the detailed settings of the experiments including the implementation of the training and testing process and the evaluation criteria. Then the experimental results including the comparison with the state-of-the-art FR-IQA metrics and the ablation studies are presented.

A. Experimental Settings

1) Implementation Details: Our Attentive Multi-channel IQA Net is implemented based on PyTorch [60]. The eight Attentive Multi-channel CNNs share the same weights, and the ResNet backbone is initialized by training on ImageNet [61]. Other weights of the neural network are randomly initialized. We train and test the proposed model on a computer with 4.00GHz Intel Core i7 processor, 16GB RAM, and an Nvidia GeForce RTX 1080 graphics card. We set the batch size to 10. The learning rate and the smoothing constant are set as 0.0001 and 0.9, respectively. RMSprop [62] is used as the optimizer to speed up minibatch learning. The training process is stopped after 50 epochs. As discussed in Section IV-B, we augment the training data by rotating the omnidirectional image along the latitude and the longitude. During the training process, we set the interval ϕ as 2 degrees and the interval γ as 15 degrees. Therefore, for each omnidirectional image, we get 45×3 sets of viewport images, with each set of viewport images containing 4 FOVs. As mentioned before, we have 300 distorted omnidirectional images and 300 corresponding reference images, respectively, for 12 raw scenes. For fair and robust evaluation, we use 6-fold cross validation to inspect the model. Specifically, 12 raw scenes are randomly split into 6

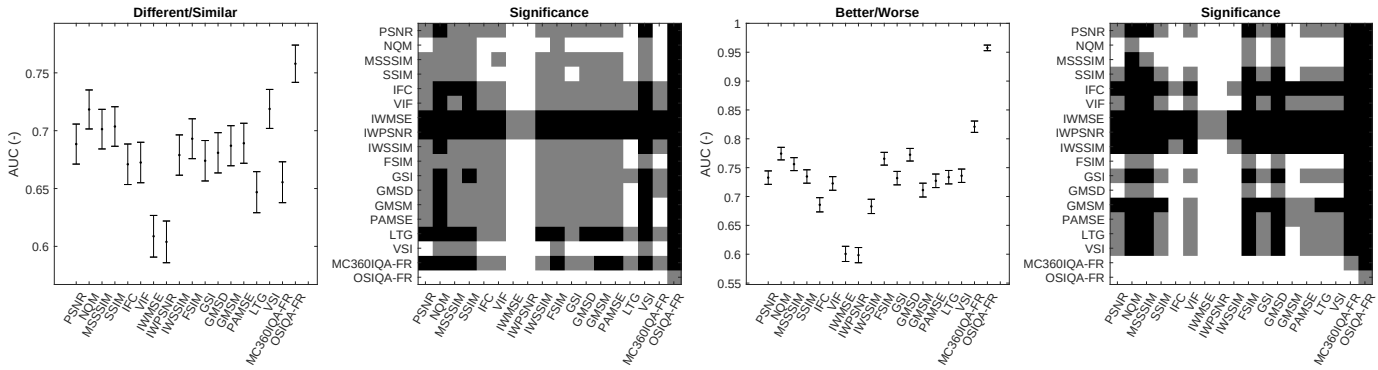


Fig. 16. New criteria performance of 17 state-of-art FR IQA models and the proposed metric on the OSIQA database. Left two figures are the different vs. similar ROC analysis results. Right two figures are the better vs. worse analysis results. Note that a white/black square in the significance figures means the row metric is statistically better/worse than the column one. A gray square means the row method and the column method are statistically indistinguishable.

TABLE I
PERFORMANCE COMPARISON OF THE STATE-OF-THE-ART FR-IQA MODELS ON THE CONSTRUCTED OSIQA DATABASE. THE BEST PERFORMING METRIC UNDER EACH CRITERION IS HIGHLIGHTED WITH BOLD FONT.

Model \ Criteria	SRCC	KRCC	PLCC	RMSE
PSNR	0.3215	0.2202	0.6403	9.8125
NQM [63]	0.3812	0.2786	0.5994	9.5568
SSIM [28]	0.4220	0.2984	0.6866	8.3829
IFC [64]	0.1939	0.1194	0.4709	10.643
VIF [65]	0.2884	0.1862	0.3628	11.358
IW-MSE [66]	0.1796	0.1319	0.3263	11.721
IW-PSNR [66]	0.1825	0.1327	0.1853	12.611
IW-SSIM [66]	0.1996	0.1371	0.5761	10.277
FSIM [29]	0.3879	0.2903	0.6465	9.8387
GSI [67]	0.3651	0.2745	0.6372	9.7378
GMSD [68]	0.4010	0.3001	0.5186	10.266
GMSM [68]	0.3836	0.2807	0.6959	8.8666
PAMSE [69]	0.3132	0.2170	0.7271	8.4045
LTG [70]	0.3438	0.2554	0.5252	10.682
VSI [71]	0.3377	0.2442	0.6212	9.8267
MC360IQA [14]	0.6434	0.4819	0.7983	7.5870
OSIQA-FR (proposed)	**0.8101**	**0.6380**	**0.8813**	**6.1004**

groups, where each group contains two scenes. The distorted and reference omnidirectional images corresponding to the same original image are assigned to the same group to ensure complete separation of the training and testing content. We use 5 groups as training set, and the remaining one group as testing set. The ratio of the training set to the testing set is 5:1. Thus, through 6-fold cross validation, we would traverse the whole group and validate the model.

2) Evaluation Criteria: Two kinds of evaluation criteria are utilized to evaluate the performance of IQA models, which includes "traditional evaluation criteria" and a "new evaluation methodology". Traditional evaluation criteria commonly calculate correlation or deviation between predicted scores and MOSs as metrics. Four evaluation criteria including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SRCC), Kendall Rank Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE) are used to measure the performance of the model. To calculate the performance, the scores predicted by the IQA models are first mapped to subjective quality ratings through a five parameter logistic function [72]:

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}}\right) + \beta_4 x + \beta_5, \qquad (12)$$

in which $\beta_i (i = 1, 2, 3, 4, 5)$ are the parameters to be fitted, x denotes the predicted score, and $f(x)$ represents the corresponding mapped score. Then these mapped scores are compared with the MOSs (through four evaluation criteria) to measure the performance of the model. Different statistical indexes demonstrate different aspects of the performance of the IQA model. Specifically, PLCC reflects the prediction linearity of the IQA metric, SRCC and KRCC demonstrate the prediction monotonicity, and RMSE indicates the prediction accuracy. The larger PLCC, SRCC, KRCC values (closer to 1) and the smaller RMSE value (closer to 0) mean better performance.

As a complementary, a new evaluation methodology based on receiver operating characteristic (ROC) analysis [73], [74] is also adopted for metric evaluation, which is based on two aspects in real application scenarios, *i.e.*, *whether two stimuli are qualitatively different and if they are, which of them is of higher quality*. Fig. 15 illustrates the framework of this evaluation methodology. We first conduct pair-wise comparison for all possible image pairs, and then classify them into pairs with and without significant quality differences. Then the ROC analysis is used to determine whether various objective metrics can discriminate images with and without significant differences, termed "*Different vs. Similar ROC Analysis*". Next, the image pairs with significant differences are classified into pairs with positive and negative differences, and the ROC analysis is used to test if various objective metrics can distinguish images with positive and negative differences, termed "*Better vs. Worse ROC Analysis*". The area under the ROC curve (AUC) values of two analysis are mainly reported in this paper, of which the higher values indicate better performance.

B. Experimental Results

1) Performance Comparison with State-of-the-art FR-IQA models: We first compare the proposed OSIQA-FR model with 16 state-of-the-art FR-IQA models on our constructed

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2023.3250956

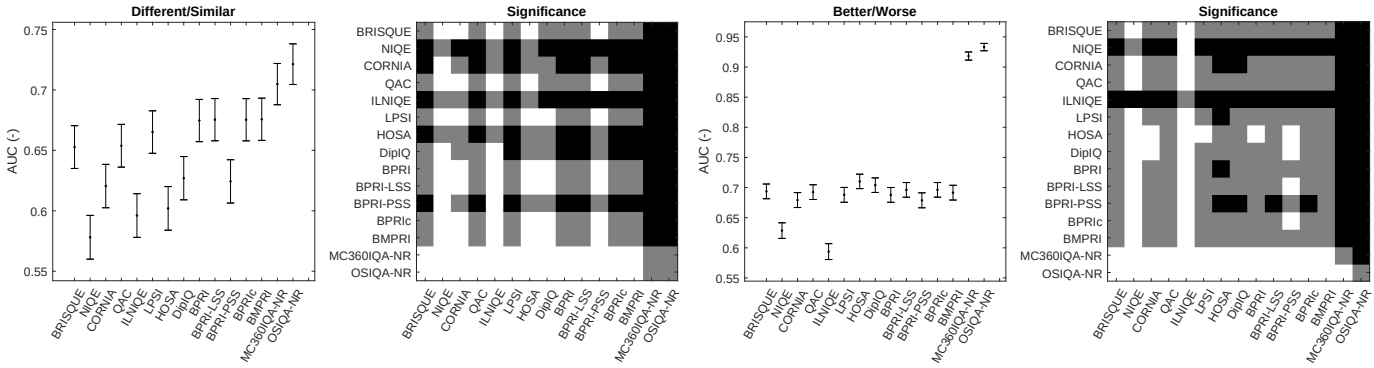


Fig. 17. New criteria performance of 15 state-of-art NR IQA models and the proposed metric on the OSIQA database. Left two figures are the different vs. similar ROC analysis results. Right two figures are the better vs. worse analysis results. The black/white/gray squares in the significance figures have the same meaning with that in Fig. 16

TABLE II
PERFORMANCE COMPARISON OF THE STATE-OF-THE-ART NR-IQA MODELS ON THE CONSTRUCTED OSIQA DATABASE. THE BEST PERFORMING METRIC UNDER EACH CRITERION IS HIGHLIGHTED WITH BOLD FONT.

Model \ Criteria	SRCC	KRCC	PLCC	RMSE
BRISQUE [75]	0.2450	0.1758	0.3072	12.296
NIQE [30]	0.2288	0.1524	0.3167	12.053
CORNIA [76]	0.2271	0.1494	0.3404	12.008
QAC [77]	0.2635	0.1736	0.5747	9.9765
ILNIQE [78]	0.1658	0.1087	0.3957	11.707
LPSI [79]	0.2127	0.1475	0.5789	10.599
HOSA [80]	0.2457	0.1840	0.3270	11.859
DipIQ [81]	0.1994	0.1363	0.2394	499.01
BPRI [31]	0.2656	0.1821	0.5993	9.9980
BPRI-LSS [31]	0.3200	0.2216	0.4889	10.994
BPRI-PSS [31]	0.2356	0.1627	0.5085	10.957
BPRIc [31]	0.3171	0.2197	0.5685	10.270
BMPRI [82]	0.2666	0.1802	0.3703	11.320
MC360IQA [14]	0.6807	0.5070	0.7943	6.9597
OSIQA-NR (proposed)	**0.7236**	**0.5512**	**0.8214**	**6.2442**

TABLE III
IMPACT OF DIFFERENT COMPONENTS.

Criteria \ Model	base	hyper	attentive	aggregation	SR
SRCC	0.6310	0.6434	0.7440	0.7857	**0.8101**
KRCC	0.4710	0.4819	0.5672	0.6105	**0.6380**
PLCC	0.7561	0.7983	0.8776	**0.8869**	0.8813
RMSE	8.2987	7.5870	6.2349	**6.0018**	6.1004

TABLE IV
IMPACT OF DIFFERENT AUGMENTATION METHODS.

Criteria \ Model	w/o multi-channel	common augmentation	OSIQA-FR w/o SR
SRCC	0.6635	0.7520	**0.7857**
KRCC	0.5031	0.5741	**0.6105**
PLCC	0.7718	0.8663	**0.8869**
RMSE	7.5536	6.3329	**6.0018**

OSIQA database, including PSNR, NQM [63], SSIM [28], IFC [64], VIF [65], IW-MSE [66], IW-PSNR [66], IW-SSIM [66], FSIM [29], GSI [67], GMSD [68], GMSM [68], PAMSE [69], LTG [70], VSI [71], and MC360IQA [14]. Table I shows the performance comparison of 16 state-of-the-art FR-IQA metrics and the proposed OSIQA-FR metric. The model with the best performance under each criterion is highlighted with bold font. The results demonstrate that the proposed model achieves the best performance compared with other state-of-the-art FR-IQA metrics, which validates the effectiveness of the proposed Attentive Multi-channel IQA Net from many aspects.

Fig. 16 illustrates the performance evaluated by the new criteria on the OSIQA database. First, we observe that the proposed OSIQA-FR model outperforms other state-of-the-art FR-IQA models on *Different vs. Similar Analysis* and *Better vs. Worse Analysis* tasks by a large margin. The significance matrices also illustrate the superiority is statistical significance. Furthermore, we notice the AUC values of the OSIQA-FR model on the *Better vs. Worse* classification task are higher than the *Different vs. Similar* classification task, which indicates that the *Different vs. Similar* classification is a more hard task and there is still room for improvement in this classification task.

2) Performance Comparison with State-of-the-art NR-IQA models: We then compare the proposed OSIQA-NR model with 14 state-of-the-art NR-IQA models on our constructed OSIQA database, including BRISQUE [75], NIQE [30], CORNIA [76], QAC [77], ILNIQE [78], LPSI [79], HOSA [80], DipIQ [81], BPRI [31], BPRI-LSS [31], BPRI-PSS [31], BPRIc [31], BMPRI [82], MC360IQA [14]. Table II shows the performance comparison between 14 state-of-the-art NR-IQA metrics and the proposed OSIQA-NR metric. The model with the best performance under each criterion is highlighted with bold font. The results demonstrate that the proposed model achieves the best performance compared with other state-of-the-art NR-IQA metrics. Moreover, as shown in Fig. 17, the proposed OSIQA-NR model significantly outperforms other state-of-the-art NR-IQA models on *Different vs. Similar*

TABLE V
PERFORMANCE COMPARISON OF THE STATE-OF-THE-ART IQA MODELS ON THE CROSS DATASET [22].

Criteria \ Model	MSE	PSNR	SSIM [28]	BRISQUE [75]	NIQE [30]	PIQE [83]	CNN [84]	OS-IQA [22]	OSIQA-FR (proposed)
SRCC	0.012	0.158	0.089	0.336	0.354	0.476	0.158	0.737	**0.857**
KRCC	0.024	0.143	0.071	0.238	0.262	0.365	0.103	0.602	**0.710**

Analysis and *Better vs. Worse Analysis* by a large margin, which further validates the effectiveness of the proposed model.

3) Ablation Studies: In this section, we conduct corresponding ablation studies to further validate the robustness of the model. Table III shows the results of the ablation studies. We first test the performance of the base Multi-channel ResNet model, which means the model mainly consists of the ResNet. The results after 6-fold cross validation are shown in the second column and denoted as "base". Then we test the performance of Multi-channel hyper-ResNet model, which we add hyper layers to the "base" ResNet model. This model is the same with the MC360IQA [14]. The 6-fold cross validation performance here (denoted as "hyper" in the third column) is the same with the column "MC360IQA" in Table I. Then we validate the importance of the subnetwork for spatial attention, of which the performance is represented by "attentive" in the fourth column Table III. Comparing the third column and the fourth column in Table III, it can be observed that the performance is boosted a lot due to the contribution of spatial attention subnetwork. As shown in the fifth column denoted as "aggregation", the performance is further improved with the help of the aggregation layer as discussed in Section IV-C. The "SR" in the sixth column in Table III denotes the spatial regularization. By comparing the fifth column and the sixth column, it can be concluded that the spatial regularization loss function can improve the prediction monotonicity of the model but will slightly decrease the prediction accuracy.

Table IV further shows the effectiveness of the proposed data augmentation method. We conduct ablation experiments based on the "aggregation" model discussed in Table III. First of all, our proposed multi-channel method can be regarded as a simple augmentation method. As shown in Table IV, without (w/o) using multi-channel method (*i.e.,* just using raw equirectangular images), the performance drops a lot. Moreover, we further compare the proposed augmentation method with the common augmentation method (*i.e.,* flip, rotate, *etc.*). It can be observed that our proposed augmentation method is better than common augmentation method on this task.

4) Generalization ability validation: To further validate the generalization ability of the proposed model. We further conduct experiments on the CROSS dataset [22]. Table V demonstrates the results of finetuning our OSIQA-FR model on the CROSS dataset. It can be observed that our proposed method achieves the best performance compared to other methods.

5) Limitation: The limitation of the proposed methods is that our devised model need to convert the omnidirectional images from equirectangular format to cubic format before feeding into the deep neural network, which introduces additional calculation and may affect the running speed.

VI. CONCLUSION

In this paper, we conduct a comprehensive IQA study for dual-fisheye omnidirectional stitching. An omnidirectional stitching image quality assessment (OSIQA) database is established first, which contains 300 distorted images with various stitching distortions (such as color distortion, geometric distortion, blur distortion, and ghosting distortion), as well as 300 corresponding reference images. We further conduct subjective quality assessment experiments and collect quality ratings from 20 subjects. These data are also included in the OSIQA database. A Attentive Hyper CNN Net is proposed based on a novel spatial attention subnetwork and a novel spatial regularization method. Then we devise an Attentive Multi-channel IQA Net for the objective FR-IQA and NR-IQA evaluation of omnidirectional stitching. Experimental results on our OSIQA database show that the proposed method achieves the best performance compared to the state-of-the-art FR IQA metrics.

REFERENCES

[1] R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2006.

[2] J. Zaragoza, T.-J. Chin, Q.-H. Tran, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving dlt," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1285–1298, 2014.

[3] C.-H. Chang, Y. Sato, and Y.-Y. Chuang, "Shape-preserving half-projective warps for image stitching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3254–3261.

[4] T. Ho and M. Budagavi, "Dual-fisheye lens stitching for 360-degree imaging," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2172–2176.

[5] I.-c. Lo, K.-t. Shih, and H. H. Chen, "Image stitching for dual fisheye cameras," in *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3164–3168.

[6] W. Xu and J. Mulligan, "Performance evaluation of color correction approaches for automatic multi-view image and video stitching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 263–270.

[7] H. Qureshi, M. Khan, R. Hafiz, Y. Cho, and J. Cha, "Quantitative quality assessment of stitched panoramic images," *IET Image Processing*, vol. 6, no. 9, pp. 1348–1358, 2012.

[8] P. Paalanen, J.-K. Kämäräinen, and H. Kälviäinen, "Image based quantitative mosaic evaluation with artificial video," in *Scandinavian Conference on Image Analysis*. Springer, 2009, pp. 470–479.

[9] F. Bellavia and C. Colombo, "Dissecting and reassembling color correction algorithms for image stitching," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 735–748, 2017.

[10] W. Sun, K. Gu, G. Zhai, S. Ma, W. Lin, and P. Le Calle, "Cviqd: Subjective quality evaluation of compressed virtual reality images," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3450–3454.

[11] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5.

[12] H. Duan, G. Zhai, X. Yang, D. Li, and W. Zhu, "Ivqad 2017: An immersive video quality assessment database," in *Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2017, pp. 1–5.

[13] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, "Spherical structural similarity index for objective omnidirectional video quality assessment," in *Proceedings of the IEEE international conference on multimedia and expo (ICME)*, 2018, pp. 1–6.

[14] 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*, 2019.

[15] H. Duan, W. Shen, X. Min, Y. Tian, J.-H. Jung, X. Yang, and G. Zhai, "Develop then rival: A human vision-inspired framework for superimposed image decomposition," *IEEE Transactions on Multimedia (TMM)*, 2022.

[16] H. Duan, X. Min, Y. Zhu, G. Zhai, X. Yang, and P. Le Callet, "Confusing image quality assessment: Towards better augmented reality experience," *IEEE Transactions on Image Processing (TIP)*, 2022.

[17] H. Duan, W. Shen, X. Min, D. Tu, J. Li, and G. Zhai, "Saliency in augmented reality," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022.

[18] H. Duan, G. Zhai, X. Min, Y. Zhu, W. Sun, and X. Yang, "Assessment of visually induced motion sickness in immersive videos," in *Pacific Rim Conference on Multimedia*. Springer, 2017, pp. 662–672.

[19] J. Yang, G. Zhai, and H. Duan, "Predicting the visual saliency of the people with vims," in *Proceedings of the IEEE Visual Communications and Image Processing (VCIP)*, 2019, pp. 1–4.

[20] L. Yang, Z. Tan, Z. Huang, and G. Cheung, "A content-aware metric for stitched panoramic image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2487–2494.

[21] P. C. Madhusudana and R. Soundararajan, "Subjective and objective quality assessment of stitched images for virtual reality," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5620–5635, 2019.

[22] J. Li, K. Yu, Y. Zhao, Y. Zhang, and L. Xu, "Cross-reference stitching quality assessment for 360 omnidirectional images," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2360–2368.

[23] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, vol. 26, no. 6, pp. 1275–1286, 2014.

[24] L. He, D. Tao, X. Li, and X. Gao, "Sparse representation for blind image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1146–1153.

[25] B. Hu, L. Li, H. Liu, W. Lin, and J. Qian, "Pairwise-comparison-based rank learning for benchmarking image restoration algorithms," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 8, pp. 2042–2056, 2019.

[26] B. Hu, L. Li, J. Wu, and J. Qian, "Subjective and objective quality assessment for image restoration: A critical survey," *Signal Processing: Image Communication*, vol. 85, p. 115839, 2020.

[27] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "Esim: Edge similarity for screen content image quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 10, pp. 4818–4831, 2017.

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

[29] 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.

[30] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[31] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Transactions on Multimedia (TMM)*, vol. 20, no. 8, pp. 2049–2062, 2017.

[32] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2014.

[33] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Proceedings of the IEEE international conference on image processing (ICIP)*, 2015, pp. 2791–2795.

[34] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[35] Y. Yuan, Q. Guo, and X. Lu, "Image quality assessment: a sparse learning way," *Neurocomputing*, vol. 159, pp. 227–241, 2015.

[36] Insta360. (2020, Jun.) Insta360 one x. [Online]. Available: https://www.insta360.com/product/insta360-onex

[37] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images: Subjective experiment and objective model evaluation," *ZTE Communications*, vol. 17, no. 1, pp. 38–47, 2019.

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

[39] 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.

[40] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3516–3530, 2018.

[41] 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, 2019.

[42] 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 *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[43] Z. Zhu, J. Lu, M. Wang, S. Zhang, R. R. Martin, H. Liu, and S.-M. Hu, "A comparative study of algorithms for realtime panoramic video blending," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2952–2965, 2018.

[44] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Gp-gan: Towards realistic high-resolution image blending," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2487–2495.

[45] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 313–318.

[46] M. Tanaka, R. Kamio, and M. Okutomi, "Seamless image cloning by a closed form solution of a modified poisson problem," in *SIGGRAPH Asia 2012 Posters*, 2012, pp. 1–1.

[47] R. Szeliski, M. Uyttendaele, and D. Steedly, "Fast poisson blending using multi-splines," in *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*, 2011, pp. 1–8.

[48] PTGui. (2020, Jun.) Ptgui. [Online]. Available: https://www.ptgui.com

[49] Easypano. (2020, Jun.) Easypano. [Online]. Available: https://www.easypano.com

[50] R. I.-R. BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2002.

[51] HTC. (2020, Jun.) Htc vive pro. [Online]. Available: https://www.vive.com

[52] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 520–531, 2015.

[53] W. Sun, G. Zhai, X. Min, Y. Liu, S. Ma, J. Liu, J. Zhou, and X. Liu, "Dynamic backlight scaling considering ambient luminance for mobile energy saving," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 25–30.

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

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

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

[57] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose

estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2017.
[58] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5659–5667.
[59] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.
[60] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
[62] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, 2013.
[63] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, 2000.
[64] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005.
[65] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
[66] 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, 2010.
[67] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2011.
[68] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2013.
[69] W. Xue, X. Mou, L. Zhang, and X. Feng, "Perceptual fidelity aware mean squared error," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 705–712.
[70] K. Gu, G. Zhai, X. Yang, and W. Zhang, "An efficient color image quality metric with local-tuned-global model," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 506–510.
[71] 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.
[72] 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 Transactions on Image Processing*, vol. 26, no. 11, pp. 5462–5474, 2017.
[73] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *Proceedings of the IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.
[74] L. Krasula, P. Le Callet, K. Fliegel, and M. Klíma, "Quality assessment of sharpened images: Challenges, methodology, and objective metrics," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 3, pp. 1496–1508, 2017.
[75] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing (TIP)*, vol. 21, no. 12, pp. 4695–4708, 2012.
[76] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1098–1105.
[77] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 995–1002.
[78] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 8, pp. 2579–2591, 2015.
[79] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 339–343.
[80] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing (TIP)*, vol. 25, no. 9, pp. 4444–4457, 2016.
[81] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 8, pp. 3951–3964, 2017.
[82] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting (TBC)*, vol. 64, no. 2, pp. 508–517, 2018.
[83] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 Twenty First National Conference on Communications (NCC)*. IEEE, 2015, pp. 1–6.
[84] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1733–1740.

Huiyu Duan received the B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. 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.

Xiongkuo Min (Member, IEEE) received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018, where he is currently a tenure-track Associate Professor with the Institute of Image Communication and Network Engineering. From Jan. 2016 to Jan. 2017, he was a visiting student at University of Waterloo. From Jun. 2018 to Sept. 2021, he was a Postdoc at Shanghai Jiao Tong University. From Jan. 2019 to Jan. 2021, he was a visiting Postdoc at The University of Texas at Austin. He received the Best Paper Runner-up Award of IEEE Transactions on Multimedia in 2021, the Best Student Paper Award of IEEE International Conference on Multimedia and Expo (ICME) in 2016, and the excellent Ph.D. thesis award from the Chinese Institute of Electronics (CIE) in 2020. His research interests include image/video/audio quality assessment, quality of experience, visual attention modeling, extended reality, and multimodal signal processing.

Wei Sun received the B.E. degree from the East China University of Science and Technology, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. His research interests include image quality assessment, perceptual signal processing and mobile video processing.

Yucheng Zhu received the B.E. degree from the Shanghai Jiao Tong University, Shanghai, China, in 2015, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2021. He is currently a Post-Doctoral Fellow with Shanghai Jiao Tong University. His research interests include visual quality assessment, visual attention modeling and perceptual signal processing.

Xiao-Ping Zhang (Fellow, IEEE) received B.S. and Ph.D. degrees from Tsinghua University, in 1992 and 1996, respectively, both in Electronic Engineering. He holds an MBA in Finance, Economics and Entrepreneurship with Honors from the University of Chicago Booth School of Business, Chicago, IL.

He is Chair Professor at Tsinghua-Berkeley Shenzhen Institute (TBSI). He has also been with the Department of Electrical, Computer and Biomedical Engineering, Toronto Metropolitan University (Formerly Ryerson University), Toronto, ON, Canada, as a Professor and the Director of the Communication and Signal Processing Applications Laboratory, and has served as the Program Director of Graduate Studies. He is cross-appointed to the Finance Department at the Ted Rogers School of Management, Toronto Metropolitan University. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015 and 2017. He is a frequent consultant for biotech companies and investment firms. His research interests include image and multimedia content analysis, sensor networks and IoT, machine learning, statistical signal processing, and applications in big data, finance, and marketing.

Dr. Zhang is Fellow of the Canadian Academy of Engineering, Fellow of the Engineering Institute of Canada, Fellow of the IEEE, a registered Professional Engineer in Ontario, Canada, and a member of Beta Gamma Sigma Honor Society. He is the general Co-Chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2021. He is the general co-chair for 2017 GlobalSIP Symposium on Signal and Information Processing for Finance and Business, and the general co-chair for 2019 GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business. He was an elected Member of the ICME steering committee. He is the General Chair for the IEEE International Workshop on Multimedia Signal Processing, 2015. He is Editor-in-Chief for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He is Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He served as Senior Area Editor the IEEE TRANSACTIONS ON SIGNAL PROCESSING and Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE SIGNAL PROCESSING LETTERS. He is selected as IEEE Distinguished Lecturer by the IEEE Signal Processing Society and by the IEEE Circuits and Systems Society.

Guangtao Zhai (Senior Member, IEEE) is a professor at Department of Electronics Engineering, Shanghai Jiao Tong University. His research interests are in the fields of multimedia and perceptual signal processing. He has received the Humboldt fellowship in 2011, national PhD thesis awards of China in 2012, best student paper award of Picture Coding Symposium (PCS) 2015, best student paper award of IEEE International Conference of Multimedia and Expo (ICME) 2016, best paper award of IEEE Trans. Multimedia 2018, Saliency360! Grand Challenge of ICME 2018, best paper award of IEEE Mobile Multimedia Computing Workshop (MMC) 2019, best paper award of IEEE CVPR Dyna-Vis Workshop 2020, best paper runner-up of IEEE Trans. Multimedia 2021, 1st place of UGC VQA Contest FR Track in IEEE ICME 2021. He also received the "Eastern Scholar" and "Dawn" program professorship of Shanghai, China, NSFC excellent young researcher award and national top young researcher award in China. He is a member of IEEE CAS MSA TC and SPS IVMSP TC. He serves as Editor-in-Chief of Displays (Elsevier), he is also on the editorial board of Digital Signal Processing (Elsevier) and Science China: information Science (Springer).