MC360IQA: A Multi-channel CNN for Blind 360-Degree Image Quality Assessment

Wei Sun, Xiongkuo Min, *Member, IEEE*, Guangtao Zhai, *Senior Member, IEEE*, Ke Gu, *Member, IEEE*, Huiyu Duan, and Siwei Ma, *Senior Member, IEEE*

***Abstract*—360-degree images/videos have been dramatically increasing in recent years. The characteristic of omnidirectional-view results in high resolution of 360-degree images/videos, which makes them difficult to be transported and stored. To deal with the problem, video coding technologies are used to compress the omnidirectional content but they will introduce the compression distortion. Therefore, it is important to study how popular coding technologies affect the quality of 360-degree images. In this paper, we present a study on both subjective and objective quality assessment of compressed virtual reality (VR) images. We first build a compressed VR image quality (CVIQ) database including 16 reference images and 528 compressed ones with three prevailing coding technologies. Then, we propose a multi-channel convolution neural network (CNN) for blind 360-degree image quality assessment (MC360IQA). To be consistent with the visual content seen in the VR device, we project each 360-degree image into six viewport images, which are adopted as inputs of the proposed model. MC360IQA consists of two parts, a multi-channel CNN and an image quality regressor. The multi-channel CNN includes six parallel hyper-ResNet34 networks, where the hyper structure is used to incorporate the features from intermediate layers. The image quality regressor fuses the features and regresses them to final scores. The experimental results show that our model achieves the best performance among the state-of-art full-reference (FR) and no-reference (NR) image quality assessment (IQA) models on the CVIQ database and other available 360-degree IQA database.**

***Index Terms*—360-degree images, image quality assessment, convolution neural network, hyper-structure.**

Manuscript received April 15, 2019; revised September 9, 2019; accepted November 8, 2019. Date of publication November 22, 2019; date of current version February 5, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61831015, Grant 61527804, Grant 61521062, Grant 61901260, Grant 61771305, and Grant 61927809, and in part by the China Postdoctoral Science Foundation under Grant BX20180197 and Grant 2019M651496. This work was presented in part at the IEEE 20th International Workshop on Multimedia Signal Processing, Vancouver, Canada, August 2018 [1] and IEEE International Symposium on Circuits and Systems, Sapporo, Japan, May 2019 [2]. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Patrick LE CALLET. *(Corresponding author: Guangtao Zhai.)*

W. Sun, X. Min, G. Zhai, and H. Duan are with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: sunguwei@sjtu.edu.cn; minxiongkuo@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn; huiyuduan@sjtu.edu.cn).

K. Gu is with the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: guke@bjut.edu.cn).

S. Ma is with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: swma@pku.edu.cn).

Digital Object Identifier 10.1109/JSTSP.2019.2955024

I. Introduction

VIRTUAL Reality (VR) can provide immersive and interactive visual experience through devices such as the Head Mounted Display (HMD), thus attracting a lot of interest from industry and research in recent years. As an important form of VR content, 360-degree videos, also known as panoramic, omnidirectional or VR videos, record views in every direction at the same time, which have accounted for 90% of VR content [3]. Users can view the content of videos in any directions through VR devices by rotating the head orientation. As a consequence, the immersive experience of real-world scenes makes 360-degree videos popular in social media, live concert events or sports events, and VR movies.

Due to the omnidirectional-view recording, 360-degree images/videos often need high resolution, e.g. 4 K, 8 K, and higher, to meet the quality of experience (QoE) of users, and thus are often compressed heavily for easy transmission and storage. However, 360-degree images at low resolution or with serious compression artifacts usually make people feel uncomfortable, sometimes even produce motion sickness [4]. Therefore, it is crucial to study the quality assessment of 360-degree images, which has significant implications in leading the development of 360-degree image compression.

Image quality assessment (IQA) has been thoroughly studied in the past twenty years and lots of IQA algorithms have been proposed to evaluate image quality [5]–[19]. Generally speaking, IQA algorithms can be classified into full-reference IQA (FR IQA), reduced-reference IQA (RR IQA), and no-reference IQA (NR IQA). FR IQA and RR IQA models need full and part reference image information respectively while NR IQA takes only the distortion image as input. FR IQA algorithms can be regarded as an image fidelity metric which measure the similarity between the distorted image and the reference image. The most widely used FR IQA model is Structural Similarity index (SSIM) [5] which calculates luminance, contrast and structure similarity between two images. Multi-SSIM (MS-SSIM) [6] and Information content weight SSIM index (IW-SSIM) [7] are extensions of SSIM, where MS-SSIM calculates different scale SSIM of the images and IW-SSIM applies the information content weighted pooling to SSIM. Feature-similarity (FSIM) [13] index uses the phase congruency and the image gradient magnitude to characterize the image local quality. These FR IQA metrics take advantages of the human visual system (HVS) and can achieve accurate quality predictions for traditional 2-dimensional

images. However, it is usually hard, or even impossible in most cases to obtain an ideal reference image, the NR IQA metirc is more realistic and receives substantial attention in recent year. According to the methodology of the measures, the NR IQA metrics can be categorized to natural scene statistics (NSS)-based, learning-based, and HVS-based measures. For example, NSS-based models include distortion identification-based image verify and integrity evaluation (DIIVINE) [20], natural image quality evaluator (NIQE) [21], pesudo reference image based measures BPRI [22] and BMPRI [23] etc. The codebook representation for NR IQA (CORNIA) [24] and NR free-energy based robust metric (NFERM) [15] are typical methods for learning-based and HVS-based measures, respectively.

However, as far as we know, limited work has been done on the quality assessment of 360-degree images. Some work [25]–[30] attempts to extend the existing IQA models such as PSNR and SSIM [5] to evaluate the quality of 360-degree images. These models mainly consider the geometric distortion occurring in the projection. Note that 360-degree images are usually mapped to the rectangular plane for easy storage and visualization. Among all the projection methods, the equirectangular projection is the simplest and most widely used projection for 360-degree images now. In order to offset the distortion, Yu *et al.* [25] proposed a sphere based PSNR (S-PSNR), which computes PSNR for the set of points uniformly distributed on a spherical surface instead of on the rectangular domain. Sun *et al.* [26] proposed the Weighted Spherical PSNR (WS-PSNR), of which the weight is determined by how much the sampled area is stretched in the representation. Zakharchenko *et al.* [27] proposed Craster Parabolic Projection PSNR (CPP-PSNR). They remapped both the distorted and reference images to the Craster parabolic projection and computed the PSNR in that domain. Xu *et al.* [28] noticed that the perceived quality of omnidirectional videos is tightly related to human attention on videos. They proposed two perceptual video quality assessment (P-VQA) methods, non-content-based P-VQA (NCP-PSNR) and content-based P-VQA (CP-PSNR), where the first one weights the distortion of pixels according to their locations in omnidirectional video and the second one assigns weights to pixel-wise distortion based on their predicted viewing direction. However, due to the inconsistency between PSNR and the experience of the HVS, the performance of these models is significantly inferior to the traditional successful IQA models for 2D natural images according to the studies of [1], [31], [32]. Therefore, the SSIM-based IQA models for 360-degree images have been proposed. For example, Chen *et al.* [29] proposed the spherical structural similarity index (S-SSIM) for omnidirectional video quality evaluation, which calculates the luminance, contrast and structural similarities of each pixel in the spherical domain. Researchers from Facebook proposed SSIM360 and 360VQM to verify the performance of 360 video pipeline on encoding and streaming [30]. SSIM360 is calculated by putting a weight on each per-sample SSIM, where the weight is determined by how much the sampled area is stretched. 360VQM replaces the weight by the new scaling factor derived from pixel density change due to view change. Different from traditional images, omnidirectional images may suffer from artifacts such as ghosting, structure inconsistency, etc, which are generated by image stitching algorithms. Ling *et al.* [33] used convolutional sparse coding to locate stitching-specific distortions and designed trained kernels to quantify the compound effects of multiple distortion types in artifact regions. Chennagiri *et al.* [34] utilized bivariate statistics of neihboring coefficients of steerable pyramid decompositions to model spatial correlation caused by stitching-specific artifacts.

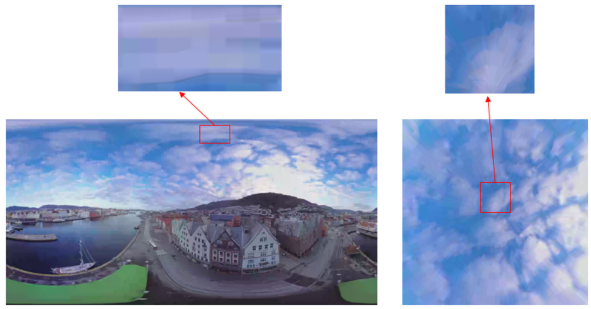


Fig. 1. Distortion comparison between the viewport image seen in VR devices and its corresponding omnidirectional image in the equirectangular format.

Recently, deep learning technologies promote the development of IQA for 360-degree images [35]–[37]. Kim *et al.* [35] split the omnidirectional image into a set of patches and estimated the local quality and weight of each patch through an adversarial network. Then the final quality score is obtained by summing the local scores with their weights. Li *et al.* [36] developed a deep learning model for evaluating the quality of omnidirectional videos which integrates the weight maps of head movement and eye movement. VR motion sickness is also a quality factor of omnidirectional videos which reflects the human's physiological response to low-quality omnidirectional videos. Several work [37], [38] has studied how the factors such as the frequency of visual oscillations, the time users immersed and the scene's contents influence VR motion sickness. Kim *et al.* [37] proposed an objective VR sickness assessment (VRSA) network based on the deep generative model to predict the VR motion sickness score.

In this paper, in order to promote the development of omnidirectional image quality assessment, we build a new VR image quality assessment database and propose a NR IQA model for predicting the quality of VR images. We first construct a compressed VR image quality (CVIQ) database, which consists of 16 source 360-degree images and 528 corresponding compressed images derived from three popular coding technologies, JPEG, H.264/AVC [39] and H.265/HEVC [40]. The single stimulus (SS) method is adopted for gathering subject ratings because observers can only see one 360-degree image in the head-mounted display. Then we propose a multi-channel convolution neural network (CNN) model for NR 360-degree image quality assessment (MC360IQA). Different from the methods mentioned above [25]–[30], we use viewport-based images projected by equirectangular images instead of equirectangular images. We argue that equirectangular images suffer great structure distortion which seriously affects human's perception of omnidirectional image quality. Fig. 1 shows the distortion comparison between the viewport image seen in the VR device and its

Fig. 2. The source 360-degree spherical images in CVIQ database. (a) teaching building. (b) Playground. (c) Square. (d) Lake. (e) Sculpture. (f) Street lamp. (g) Gate. (h) Bicycles. (j) Bridge. (k) Road. (l) Classroom. (m) Multimedia room. (n) Rally. (o) Expressway. (p) Town. (q) Valley.

corresponding omnidirectional image in the equirectangular format. It shows that compression artifacts in the equirectangular projection are block-based while the distortion seen in the viewport is totally different. This illustrates why many NR IQA algorithms work poorly in this field. Therefore, we use viewport-based images as the input of MC360IQA. In the pre-processing stage, we project the equirectangular image into six equally-sized viewport images, represented each cube face with a field of view of 90 degree. We also alter the longitude of viewing angle to project many different groups of viewport images from one omnidirectional to avoid overfitting. MC360IQA model consists of two part. The first part includes six parallel CNN channels which are used to extract features of the six viewport images. The second part is the image quality regressor, which concatenates features of the six viewport images and regresses them to final quality scores. For the base CNN channel, we use the hyper-CNN architecture, which can make use of the intermediate layers of a deep CNN. It has been shown that features contained in lower layers respond to edges and corners [41], [42], which are also proved to be very useful for image quality assessment [43]. In [44], Gao *et al.* proved that the mid-layers and deep-layers are both useful for image quality prediction. Therefore, fusing information from all intermediate layers is beneficial to extract effective features for IQA. The experimental results show that the proposed model achieves the best performance among the state-of-the-art NR and FR IQA models.

In summary, this paper has made the following contributions.

1) We establish a large compressed VR image quality database, which includes 16 reference images and 528 corresponding compressed images. Three popular codec methods, JPEG, H.264/AVC, and H.265/HEVC are implemented. The built database will have significant implication to promote the development of objective IQA for compressed 360-degree images.
2) We propose a multi-channel CNN model for NR 360-degree image quality assessment. The proposed model adopt viewport images as input, which can effectively avoid geometric distortion and is more consistent with the human vision system.
3) We propose to use the hyper-CNN architecture as the base CNN channel for the MC360IQA, which can incorporate the features from the intermediate layers. The experimental results demonstrate that hyper-CNN architecture can promote the performance of the MC360IQA.

The rest of this paper is organized as follows. Section II introduces the construction of the CVIQ database and the subjective user study. In Section III, we describe the implementation of the MC360IQA in detail. In Section IV, we give the results of MC360IQA and compare the performance of MC360IQA with other popular IQA models on the available 360-degree IQA databases. Section V gives the concluding remarks.

II. Subjective Quality Assessment

In this section, we introduce how to build the CVIQ database. First, the images in the database are described. Next, subjective evaluation is applied to collect the mean opinion scores (MOSs) from subjects. Finally, the MOSs are presented and analyzed.

TABLE I
SPATIAL INFORMATION FOR CVIQ AND OTHER PREVAILING IQA DATABASES

SI	LIVE [47]	CSIQ [12]	TID203 [48]	CVIQ
mean	108.64	94.99	87.74	92.08
std	35.96	28.91	25.86	25.51

A. Compressed VR Image Quality Database

The database consists of sixteen source images, where twelve images are shot by Insta360 4 K spherical VR video camera and the other four images are obtained from the test video of the JVET. The source images contain diverse scenes such as towns, landscapes, persons, and objects, as shown in Fig. 2. Spatial information (SI) [45] is an indicator of edge energy, which reflects the scene complexity. Here, we list the mean and standard deviation of SI of reference images in the CVIQ and three most prevailing IQA databases in Table I. We can observe that the CVIQ database has a moderate mean value and a low standard deviation of SI compared to the other three IQA databases, indicating that the images in the CVIQ database have sufficient scene complexity. For the fair subjective evaluation, the source images are resized to the same resolution of 4096×2048.

We deploy three coding technologies in the CVIQ database. The first one is the Joint Photographic Experts Group (JPEG) [46], which is a commonly used method of lossy compression for digital images. Typically, JPEG can achieve 10:1 compression with little perceptible loss in image quality, which makes it one of the most commonly employed compressed formats for photographic images on the World Wide Web. The second and third coding technologies are H.264/AVC (Advanced Video Coding) [39] and H.265/HEVC (High Efficiency Video Coding) [40] respectively, which were developed for video compression. As compared with H.264/AVC, the H.265/HEVC can lead to more than 50% performance gains in most cases. According to this, these three coding technologies are introduced in this work to establish the VR image quality database.

To be more specific, we use the JPEG to compress each reference image with quality factors ranging from 50 to 0 with an interval of -5, and use the H.264/AVC and H.265/HEVC with factors from 30 to 50 with an interval of 2. On this basis, we generate 33 compressed images from each source 360-degree image. Overall, a database including 16 reference images and 528 compressed images is built.

B. Subjective Experiment Methodology

In the following, we present the general methodology and configuration of the subjective test.

- **Method**: Several subjective testing methodologies for assessing image quality have been defined by ITU-R BT500-11 [49], including single-stimulus (SS), double-stimulus impairment scale (DSIS) and paired comparison (PC). Since the viewers only see a part of the 360-degree image that falls into the field of view (FoV) of the HMD, we adopt the SS method in our test.

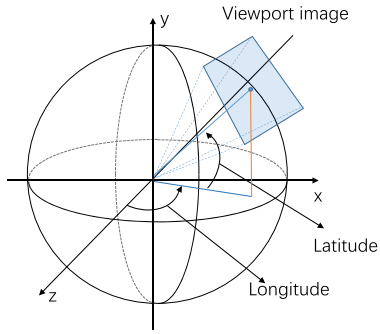


Fig. 3. Illustration for viewport images when the user sees the omnidirectional image in VR devices at a certain head pose.

- **Participants**: The study conducted by [28] suggests that at least 15 subjects are required in the subjective quality assessment for VR images. Here, **20 subjects** including 14 males and 6 females participated in the subjective test. Their ages range from 21 to 25. All participants have a normal or corrected-to-normal vision.
- **Test Condition**: Unlike other subjective experiments conducted on the traditional displays, we do not need to consider the environmental factors, e.g. viewing distance [50], ambient luminance [51], etc. The experiment was conducted in an empty room with no noise. The subjects sat on a swivel chair so they could turn their viewing direction freely.
- **Test Device**: We used the HTC VIVE as the HMD because of its excellent graphic display and high precision tracking ability. For easy operation, we designed an interaction system to automatically display the test images and collect the subjective quality scores using Unity3D software. The subjects used the controller to switch images and select the perceptual scores. Unity3D was run on a computer with 4.00 GHz Intel Core i7 processor, 32 GB main memory, and Nvidia GeForce GTX 1080 graphics.
- **Quality Rating**: The scales ranging from the lowest to highest perceptual quality are divided into 10 levels. The higher value means the better quality.

Before starting the experiment, the goal of this subjective test and instruction were introduced to each subject. The whole experiment involves two stages. The first stage is the pilot experiment. Subjects previewed some example images which would not appear in the formal experiment so they would have an idea on how to provide their scores on the image quality. The second stage is the formal experiment. 20 subjects participated in the test. They were asked to provide their perceptual opinions. The presentation order of the images was randomized for each subject. After the subjective experiment, we collected the scores of all the images rated by all the subjects and did further analysis.

C. Data Processing and Analysis

From the subjective test, we have collected all the subjects' scores. We follow the MOS calculation method as detailed in [49]. Let m_{ij} denote the raw subjective scores assigned by

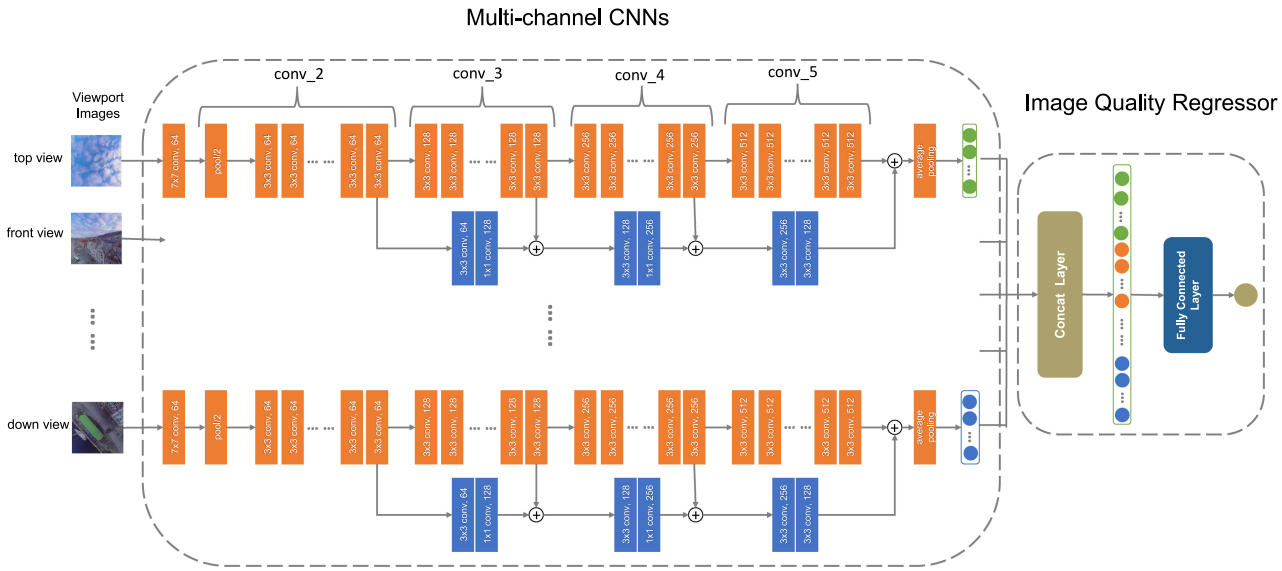


Fig. 4. The network architecture of the MC360IQA. The multi-channel CNN includes six parallel ResNet34. We omit the four of them which are sent front, left, right and back view images for simplcity.

subject i to image j. First, the score m_{ij} needs to be converted to a Z-score Z_{ij} using

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} m_{ij},\ \sigma_i = \sqrt{\frac{1}{M_i - 1}\sum\nolimits_{j=1}^{M_i}(m_{ij} - \mu_i)}, \quad (1)$$

$$Z_{ij} = \frac{m_{ij} - \mu_i}{\sigma_i}, \quad (2)$$

where N_i denotes the number of test images viewed by subject i.

After that, we discard scores from unreliable subjects by using the subject rejection procedure specified in the ITU-R BT500-11 [49]. Then Z-score Z_{ij} needs to be linearly rescaled to lie in the range of [0, 100]:

$$Z'_{ij} = \frac{100(Z_{ij} + 3)}{6}, \quad (3)$$

Finally, the MOS of the image j is calculated by averaging the Z'_{ij} from M_j subjects:

$$\text{MOS}_j = \frac{1}{M_j}\sum_{i=1}^{M_j} Z'_{ij}, \quad (4)$$

Note that the MOSs in the CVIQ database are mainly centralized from scores "30" to "70" and the number of MOSs which are more than "80" is none. This means that the visual effect is still barely satisfied with those compressed 360 degree spherical images with a resolutions of 4 K.

III. Proposed Method

In this section, we detail the pipeline of MC360IQA for evaluating the 360-degree image quality. A diagram of the proposed MC360IQA is illustrated in Fig. 4. The proposed model takes a 360-degree image as input and projects it onto six viewport images using the method described in Section III-A. Then six viewport images are sent to the multi-channel CNN, the details of which we depict in Section III-B. The features extracted by the multi-channel CNN are fused and finally regressed to the objective quality score.

A. Projection Method

When users view the visual content of the 360-degree image in the VR device, the equirectangular image is first represented by a sphere in 3D spherical coordinates and then the visual content is rendered as a plane segment tangential to the sphere decided by the viewing angle and the FoV of the VR device. We show this process in Fig. 3. Users can view all contents of the 360-degree image by rotating the head to change the viewing angle. When assessing the quality of a 360-degree image, the viewer should look around the 360-degree image from several viewing angles to cover the entire 360-degree image.

Inspired by this behavior, we propose to use the viewport-based images to evaluate the omnidirectional image quality. The pixel in the viewport image can be calculated through mapping it backward to find the best estimate pixel in the spherical image. The detailed procedure can be found in [25]. We set the field of view (FoV) as 90 degree, which is consistent with the FoV of most popular VR devices such as HTC VIVE, Oculus, Gear VR, etc. To cover the full visual content of the omnidirectional image, six viewport images are rendered by one omnidirectional image. Two of these views are oriented towards the nadir and zenith, and the other four are pointed towards the horizon but rotated horizontally to cover the entire band at the sphere's equator, which is shown in Fig. 5. We use the symbols $VP_{\text{front}}, VP_{\text{back}}, VP_{\text{right}}, VP_{\text{left}}, VP_{\text{top}}, VP_{\text{down}}$ to represent the six viewport images in the front, back, right, left, top and down views, respectively. On the other side, users usually view the image from different starting viewing angles, which enlightens us that training samples should include different

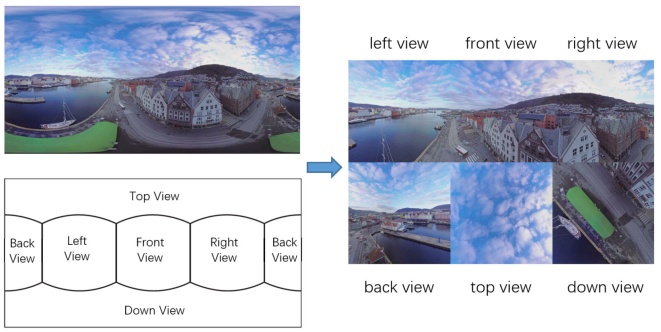


Fig. 5. The viewport images and their corresponding parts in the omnidirectional image.

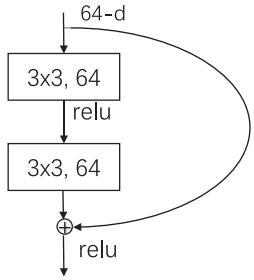


Fig. 6. A building block (on 56 × 56 feature maps) for ResNet34.

TABLE II
THE ARCHITECTURE OF RESNET34

Layer Name	Output Size	Convolution Layer
conv1	112×112	7×7, 64, stride2
conv2_x	56×56	3×3 max pool, stride2 $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$
	10×1	average pool

starting viewing angles. Therefore, we rotate the longitude of the viewing angle of the front view from 0 to 360 degree with an interval of φ degree and then project omnidirectional images to six viewport images at each front viewing angle respectively. Finally, we can get N groups of viewport images derived from one omnidirectional image. We denote them as VP^i_{view}, where $view \in \{front, back, right, left, top, down\}$ and $i \in [1, 2, \ldots, N]$, $N = 360/\varphi$. Also, it is an effective method for the CNN model to avoid overfitting.

B. MC360IQA

Convolution neural networks have shown great performance in solving visual signal problems in recent years. Many successful CNN models such as VGG [52], GoogleNet [53], ResNet [54] have been proposed for solving image recognition, detection, segmentation problems, etc. These models usually have strong ability in extracting high-level semantic features. We adopt ResNet as the base CNN-channel since ResNet has an excellent generalization ability in lots of visual tasks and has a relatively small memory consumption. We also fuse the features from intermediate layers because they are also useful features for image quality assessment. We detail the architecture of MC360IQA as follows.

The MC360IQA model consists of two parts, multi-channel CNN and image quality regressor. We illustrate the framework in Fig. 4. The multi-channel CNN includes six parallel ResNet34 s which are used to extract features of corresponding six viewport images. ResNet utilizes residual learning to further deepen the CNN network, which can be generally represented by several deeper building blocks. According to the number of layers, ResNet has several architectures such as ResNet18, ResNet34, ResNet50, ResNet101, etc. Here, considering the efficiency and accuracy of the model, we choose the ResNet34 as the base CNN-channel. For ResNet34, each building block includes two layers convolutions where the dimension of kernels is both 3 × 3. The identity shortcut connection is inserted from the input of the building block to the output of the building block. We show the building block in Fig. 6. The ResNet34 can be represented by five parts, which are denoted as $conv1$, $conv2\_x$, $conv3\_x$, $conv4\_x$, and $conv5\_x$, respectively. We illustrate the architecture of ResNet34 in Table II. In $conv1$, convolution kernels with the dimension of 7 × 7 and 64 channels are performed with a stride 2. The $conv2\_x$ includes a 3 × 3 max pooling with a stride 2 and three repeated building blocks which are named as $conv2\_1$, $conv2\_2$, $conv2\_3$ ($x = 1, 2,$ and 3), respectively. Each building block has the same structure illustrated in Fig. 6. The $conv3\_x$ to $conv5\_x$ have 4, 6, and 3 repeated building blocks, respectively. The difference in each building block is that the kernel channels in $conv3\_x$ to $conv5\_x$ are 128, 256, and 512 respectively. We replace the last layer of each baseline ResNet34 with 10 output features by average pooling. The more detail of ResNet network structure can be found in [54].

As mentioned above, feature maps extracted from low layers are corresponding to low-level visual stimulus such as edges and corners, which is important for human to perceive the image quality. To take advantages of these features, we use the hyper-ResNet structure to fuse the features from intermediate layers. Similar to [41], we use hierarchical element-wise addition to fuse features. Specifically, we fuse feature maps extracted by $conv2\_x$, $conv3\_x$, $conv4\_x$, $conv5\_x$. Due to that the channel and dimension of feature maps in each stage are not the same, we use two convolution operation to downscale the dimension and add the channels. As illustrated in Fig. 6, the features from $conv2\_x$ are first reduced its resolution through a

3×3 convolution kernel with a stride of 2 and then are passed through a 1×1 convolution layer to increase the number of channels. So, the number of channels and the size of dimension can match the features from the next stage $conv3\_x$. Then, element-wise addition is implemented between the feature maps from $conv3\_x$ and $conv2\_x$ to produce the merged feature maps. The merged feature maps will continue to downscale its resolution and add its channels and then are element-wise added by feature maps from $conv4\_x$. The same operation is also implemented to fuse the new merged feature maps and the feature maps from $conv5\_x$. Finally, average pooling is applied to produce a feature vector with a dimension of 10×1.

The six hyper-ResNet34 channels share the same weights and are trained to extract the unified features for different compression artifacts. The image quality regressor first fuses the features by concatenating the outputs of multi-channel CNNs. According to [55]–[58], users focus more on the equator area and seldom view the nadir and zenith areas, which indicates that the importance of each viewport image is different for the final quality score. Therefore, another function of the image quality regressor is to assign weights for different viewport images. Finally, the quality score can be calculated by using a fully connected layer in the image quality regressor.

For the end-to-end training, the loss function is set as:

$$L = (q_{\text{predict}} - q_{\text{label}})^2, \tag{5}$$

where q_{predict} is the predicted score calculated by the MC360IQA and q_{label} is the MOS derived from subjective experiments.

Two metrics are proposed to measure the quality of 360-degree images. The first metric uses the score calculated by the MC360IQA using the viewport images without longitude rotating, denoted by $MC360IQA_{\text{origin}}$. The second metric uses the mean score of N groups of viewport images calculated by the MC360IQA, denoted by $MC360IQA_{\text{mean}}$. We formulate two metrics as:

$$\begin{aligned} MC360IQA_{\text{origin}} &= MC360IQA(VP^1_{\text{view}}), \\ MC360IQA_{\text{mean}} &= \sum_{i=1}^{N} MC360IQA(VP^i_{\text{view}}), \end{aligned} \tag{6}$$

where $view \in \{front, back, right, left, top, down\}$ and $i \in [1, 2, \ldots, N]$, $N = 360/\varphi$.

IV. Experimental Validation

In this section, we first present the experimental protocol in detail and then evaluate the prediction performance of the proposed model on the two 360-degree image quality databases, CVIQ and OIQA [59] databases. After that, the sensitivity analysis and cross-database evaluation are conducted to prove the robustness and effectiveness of the MC360IQA model.

A. Experiment Protocol

1) Implementation Details: The MC360IQA model is implemented in PyTorch [60]. The six hyper-ResNet34 channels of the MC360IQA share the same weights, which are initialized by training on ImageNet [61]. Other weights are randomly initialized. The interval angle φ is set as 2 degree, which means 180 groups of six viewport images are rendered from one omnidirectional image. The resolution of each viewport image is resized to 224×224. We trained and tested our model on a server with Intel Xeon Silver 4114 CPU @ 2.20 GHz, 64 GB RAM and NVIDIA GTX 1080Ti. The batch size is set as 20. We choose the RMSprop algorithm [62] to speed up mini-batch learning. The learning rate and smoothing constant are set as 0.0001 and 0.9, respectively. We stop the training after 20 epochs. For fair evaluation, we use 5-fold cross validation. So, the database is split into the training set with 80% distorted images and the testing set with 20% distorted images. The distorted images corresponding to the same original image are assigned to the same set to ensure complete separation of the training and testing content.

2) Test Databases: The MC360IQA model is validated on the two 360-degree image quality databases:

- CVIQ: The CVIQ database is introduced in Section II. It includes 524 compressed images generated from 16 original images. Three popular coding technologies are deployed in the database, which are JPEG, H.264/AVC, and H.265/HEVC, respectively.
- OIQA: The OIQA database [59] is an omnidirectional image quality database which also provides the head and eye movement data. The OIQA consists of 16 original images and 320 distorted images degraded by JPEG compression, JPEG2000 compression, Gaussian blur, and Gaussian noise. Since the proposed MC360IQA can automatically learn the weight of each viewport image related to the final quality score, we do not use the head and eye movement data here.

3) Comparing Algorithms: To demonstrate the effectiveness of the proposed model, we compare the MC360IQA model with several state-of-the-art IQA models, including:

- FR IQA models for omnidirectional images: WS-PSNR [26], CPP-PSNR [27], and S-PSNR [25]. They are PSNR-based IQA models specially designed for evaluating the 360-degree image quality.
- FR IQA models for traditional 2-D images: PSNR, SSIM [5], and MS-SSIM [6]. They are most commonly used FR IQA models in practical applications such as video coding, image enhancement, image denoising, image super-resolution, etc.
- NR IQA models for traditional 2-D images: BRISQUE [63], GMLF [64], NIQE [21], QAC [65], and SISBLIM [66]. They are general-purpose NR IQA models which are not limited by distortion types.

4) Evaluation Criteria: Two kinds of evaluation criteria are utilized to evaluate the performance of IQA models. The first is recommended by video quality experts group (VQEG) [67]–[69], which calculates a series of correlation values between predicted scores and MOSs. The second is developed by Krasula *et al.* [70]–[73], which evaluates classification abilities of IQA models to distinguish which of the two images is better or of the

TABLE III
VQEG PERFORMANCE OF 11 STATE-OF-ART FR AND NR IQA MODELS AND TWO PROPOSED METRICS ON THE CVIQ DATABASE. THE METRIC $MC360IQA_{origin}$ AND $MC360IQA_{mean}$ ARE DENOTED AS PRO. AND PRO.+ RESPECTIVELY, WHICH ARE ALSO USED IN FOLLOWING FIGURES AND TABLES. THE BEST PERFORMING MODELS IN FR AND NR IQA CATEGORIES ARE HIGHLIGHTED IN EACH ROW

Metrics		Full Reference						No Reference						
		PSNR	S-PSNR	WS-PSNR	CPP-PSNR	SSIM	MS-SSIM	QAC	GMLF	NIQE	BRISQUE	SISBLIM	Pro.	Pro.+
JPEG	SRCC	0.8643	0.8772	0.8802	0.8886	**0.9749**	0.9628	0.9537	0.7801	0.8525	0.9091	0.9186	**0.9332**	0.9316
	PLCC	0.7342	0.7520	0.7604	0.7729	**0.9334**	0.9140	0.8680	-0.4484	-0.8585	-0.8489	-0.8433	0.9724	**0.9746**
	RMSE	8.5866	8.1974	8.1019	7.8302	**3.7986**	4.6101	5.1324	10.6822	8.9237	7.1137	6.7479	3.7884	**2.6388**
AVC	SRCC	0.7592	0.7708	0.7748	0.7854	**0.9457**	0.8805	0.8681	0.4864	0.8467	0.7294	0.8547	0.9196	**0.9244**
	PLCC	0.7572	0.7690	0.7726	0.7815	**0.9451**	0.8794	0.8681	-0.1748	-0.8358	-0.7193	-0.8122	**0.9558**	0.9461
	RMSE	8.0448	7.8743	7.8143	7.6506	**4.0165**	5.8583	6.1348	10.8000	6.5773	8.4558	6.4159	3.7884	**2.6983**
HEVC	SRCC	0.7215	0.7428	0.7469	0.7578	**0.9232**	0.8610	0.8749	0.1491	0.8649	0.7104	0.5620	**0.8986**	0.8985
	PLCC	0.7169	0.7389	0.7430	0.7540	**0.9220**	0.8604	0.8764	-0.0232	-0.8681	-0.7151	-0.5041	0.9118	**0.9126**
	RMSE	8.3279	8.0515	7.9974	7.8471	**4.6219**	6.1165	5.8249	11.8923	6.0370	8.4646	9.9474	4.3467	**3.2935**
Overall	SRCC	0.7662	0.7741	0.7755	0.7819	**0.8972**	0.8875	0.8681	0.6134	0.5329	0.7641	0.7439	0.9113	**0.9139**
	PLCC	0.7320	0.7467	0.7498	0.7574	**0.8857**	0.8762	0.8299	-0.2246	-0.5126	-0.7448	-0.6554	0.9503	**0.9506**
	RMSE	9.0397	8.9066	8.8816	8.7695	**6.2140**	6.4836	6.9820	11.1101	11.9038	9.0751	9.4014	4.1484	**3.0935**

same quality. We denote the first one as VQEG criteria and the second one as New criteria.

For VQEG criteria, the scores predicted by IQA models are first mapped using the following five-parameter logistic function.

$$q(s) = \epsilon_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\epsilon_2 (s - \epsilon_3)}} \right) + \epsilon_4 s + \epsilon_5, \tag{7}$$

where $\{\epsilon_i | i = 1, 2, \ldots, 5\}$ are parameters to be fitted, s and $q(s)$ denote the predicted scores and mapped scores respectively. Then three correlation values are calculated between the mapped scores and MOSs, which are Spearman Rank-Order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC) and Root Mean Squared Error (RMSE), respectively. These three statistical indexes have different meaning for demonstrating the performance of IQA models. To be more specific, PLCC reflects the prediction linearity of the IQA algorithm, SRCC indicates the prediction monotonicity, and RMSE represents the prediction accuracy. An excellent IQA model should obtain the value of SRCC and PLCC close to 1, yet the value near 0 for RMSE.

The new criteria is based on two real-world scenarios: *whether two stimuli are qualitatively different* and *if they are, which of them is of higher quality*. Following the procedures in [70], we first use the statistical method suggested by [74] to analyze the subjective rating data and then determine whether each pair of stimuli is significantly different with a 95% confidence. So, the pairs in the dataset can be divided into two groups, significantly different and similar. For significantly different pairs, the higher MOS means the better quality. We further divide them into groups with positive and negative MOS difference. Next, we calculate the differences of scores predicted by each IQA model for all stimuli pairs. An excellent IQA model should have enough abilities to distinguish different/similarity pairs and better/worse pairs. Receiver Operating Characteristic (ROC) Analysis [75] is frequently used to determine the abilities of binary classifiers. We use the Area Under the ROC curve (AUC) to evaluate the classification performance of IQA models. In addition, AUC values derived from different IQA models are compared [76] to determine if the performance differences between IQA models are statistically significant. Note that we combine the stimuli pairs (different/similar, better/worse) directly in each cross-validation and analyze all the pairs together, while the correlation values in VQEG criteria are averaged across five cross-validations.

B. Performance Comparison With State-of-the-Art IQA Models

1) VQEG Criteria: We list the VQEG performance on the CVIQ database in Table III and the performance on the OIQA database in Table IV in terms of single distortion and overall database. The best performing models in FR and NR IQA categories are highlighted in each row in Table III and Table IV. From the performance listed on Table III and Table IV, we have several observations. We first focus on overall performance. It is shown that most of NR IQA models perform poorly on the CVIQ database and OIQA database, which indicates the current NR IQA models do not work on 360-degree images. Fortunately, our model makes up this gap. For FR IQA models, it is obvious that the performance of PSNR-based IQA models is remarkably inferior to the traditional successful IQA models like SSIM, MS-SSIM, though they are specially designed for 360-degree images. The reason is that PSNR performs not well on content-independent distortions and does not agree with the experience of the HVS. SSIM-based methods are still suitable for 360-degree images since they can effectively measure the

TABLE IV
VQEG PERFORMANCE OF 11 STATE-OF-ART FR AND NR IQA MODELS AND TWO PROPOSED METRICS ON OIQA DATABASE. THE BEST PERFORMING MODELS IN FR AND NR IQA CATEGORIES ARE HIGHLIGHTED IN EACH ROW

Metrics		Full Reference						No Reference						
		PSNR	S-PSNR	WS-PSNR	CPP-PSNR	SSIM	MS-SSIM	QAC	GMLF	NIQE	BRISQUE	SISBLIM	Pro	Pro.+
JPEG	SRCC	0.8291	0.8285	0.8278	0.8282	**0.9346**	0.9188	0.7755	0.4150	-0.6084	-0.8677	-0.0876	0.8988	**0.9190**
	PLCC	0.8658	0.8703	0.8607	0.8654	**0.9409**	0.9312	0.8176	0.5932	0.6498	0.8986	0.1828	0.9025	**0.9279**
	RMSE	7.8570	7.7319	7.9919	7.8678	**5.3193**	5.7214	9.0392	12.6390	11.9332	6.8890	15.4352	4.8394	**4.5058**
JP2K	SRCC	0.8421	0.8489	0.8322	0.8375	**0.9357**	0.9267	0.8820	0.4715	-0.6346	-0.9179	-0.3965	0.8790	**0.9252**
	PLCC	0.8492	0.8555	0.8435	0.8488	**0.9336**	0.9265	0.8863	0.5081	0.7214	0.9199	0.4885	0.9082	**0.9324**
	RMSE	7.9357	7.7811	8.0719	7.9449	**5.3829**	5.6560	6.9593	12.9423	10.4060	5.8936	13.1112	4.8006	**4.5825**
GN	SRCC	0.9008	0.8846	0.8853	0.8851	0.8846	**0.9484**	0.7329	**-0.9528**	-0.9258	-0.9283	-0.9104	0.9312	0.9345
	PLCC	0.9317	0.9190	0.9221	0.9201	0.9026	**0.9672**	0.8866	**0.9609**	0.9337	0.9417	0.9210	0.9079	0.9344
	RMSE	4.6392	5.0329	4.9415	5.0010	5.4965	**3.2460**	5.9058	**3.5339**	4.5716	4.2957	4.9759	4.1595	3.7908
BLUR	SRCC	0.6374	0.6917	0.6583	0.6666	**0.9238**	0.8624	0.9289	-0.3523	-0.6420	**-0.9459**	-0.9042	0.9336	0.9353
	PLCC	0.6357	0.6929	0.6609	0.6734	**0.9188**	0.8623	0.9490	0.3853	0.8195	**0.9572**	0.9118	0.9075	0.9220
	RMSE	10.2503	9.5736	9.9652	9.8162	**5.2404**	6.7250	4.1860	12.2528	7.6088	**3.8447**	5.4517	4.8172	4.5256
Overall	SRCC	0.6802	0.7115	0.6932	0.7028	**0.8798**	0.8332	0.7665	0.1185	-0.3306	-0.7794	-0.4539	0.9020	**0.9187**
	PLCC	0.6918	0.7153	0.6985	0.7057	**0.8892**	0.8427	0.7725	0.4427	0.4701	0.7778	0.5326	0.9045	**0.9247**
	RMSE	10.3882	10.0524	10.2944	10.1920	**6.5814**	7.7442	9.1350	12.8994	12.6970	9.0409	12.1757	5.2704	**4.6247**

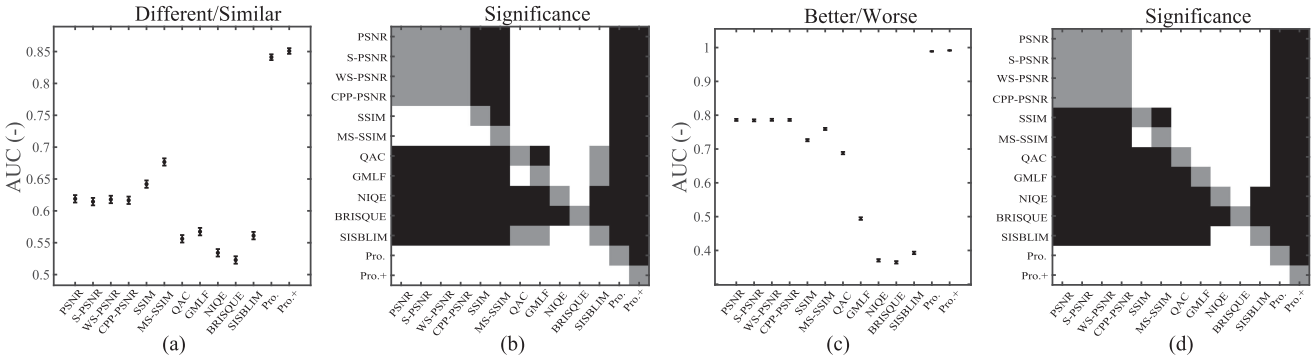


Fig. 7. New criteria performance of 11 state-of-art FR and NR IQA models and two proposed metrics on the CVIQ database. (a) and (b) are the different vs. similar ROC analysis results. (c) and (d) are the better vs. worse analysis results. Note that a white/black square in (b) and (d) means the row metric is statistically better/worse than the column one. A gray square means the row method and the column method are statistically indistinguishable.

structural similarity between two images. The proposed two metrics $MC360IQA_{origin}$ and $MC360IQA_{mean}$ outperform the state-of-the-art FR and NR IQA models on the two databases. The metric $MC360IQA_{mean}$ is slightly better than the metric $MC360IQA_{origin}$. The reason is that the mean score of the N groups of viewport images is more stable and less susceptible to abnormal predictive scores. But $MC360IQA_{mean}$ needs to consume N times of computing resources than that of $MC360IQA_{origin}$. In comparison, $MC360IQA_{origin}$ is more efficient.

Then we observe the performance of IQA models on each single distortion. The proposed two metrics still achieve the best performance among these state-of-the-art IQA models on the CVIQ database and have the comparative performance with the best model SSIM. For the OIQA database, our metrics also achieve the best performance on JPEG compression and blur distortion and is just slightly inferior to the best model on the JPEG2000 compression and Gaussian noise. This is because traditional IQA models like SSIM is easier to deal with the single distortion since the distortion rule of images are the same. However, our model tries to learn the uniform features of multiple distortion types for omnidirectional quality assessment, which cases our model to perform better on the overall database and perform relatively poorly on each distortion type.

2) New Criteria: We illustrate the performance evaluated by the new criteria on the CVIQ database in Fig. 7 and on the OIQA

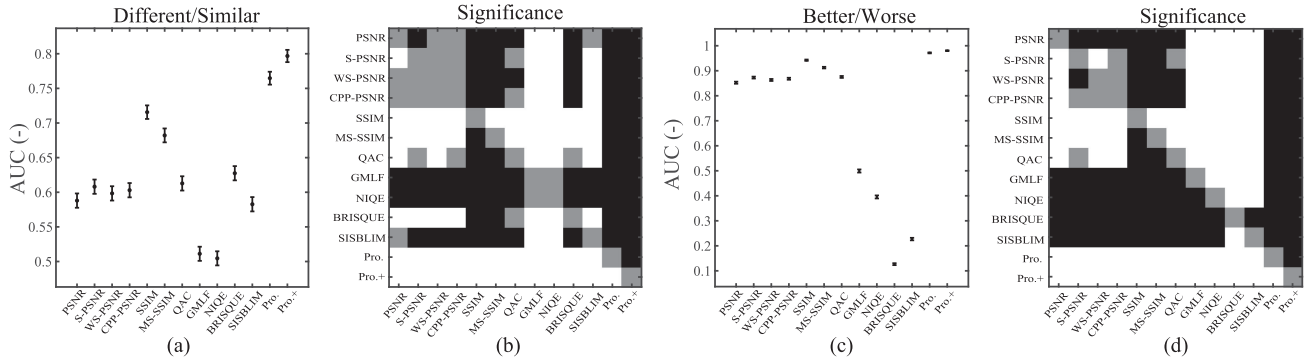


Fig. 8. New criteria performance of 11 state-of-art FR and NR IQA models and two proposed metrics on the OIQA database. (a) and (b) are the different vs. similar ROC analysis results. (c) and (d) are the better vs. worse analysis results. The black/white/gray squares in (b) and (d) have the same meaning in Fig. 7.

database in Fig. 8. From these two figures, we can obtain similar conclusions derived from the VQEG performance. First, the proposed two metrics $MC360IQA_{origin}$ and $MC360IQA_{mean}$ outperform other FR and NR IQA models on *Different vs. Similar* and *Better vs. Worse* classification tasks by a large margin. The statistics analysis also show that the proposed two metrics are significantly superior to others on both two databases. We notice the AUC values of the proposed metrics on *Better vs. Worse* classification task are higher than *Different vs. Similar* classification task, which indicates that *Different vs. Similar* classification is a more hard task and there is still some room for improvement in this classification task. Second, we observe that the metric $MC360IQA_{mean}$ is also significantly better than $MC360IQA_{origin}$ on two classification tasks, which means $MC360IQA_{mean}$ can provide stronger quality prediction ability to distinguish qualities of different images.

The experimental results of VQEG performance and new criteria performance both show that the proposed two metrics achieve the best performance among the state-of-the-art IQA models, which proves the effectiveness of the MC360IQA model from many aspects. In the following sections, we only report the VQEG performance since we can derive the same conclusions from both VQEG criteria and new criteria.

C. The Effect of Hyper-Structure

The hyper-structure can effectively incorporate features from multiple intermediate layers, where these features are sensitive to perceiving the image quality. With the help of the hyper-structure, the MC360IQA achieves excellent performance on the 360-degree image quality databases. In this section, we investigate the contribution of this hyper-structure by comparing the performance of the MC360IQA with and without hyper-structure. The MC360IQA without hyper-structure adopts the ResNet34 as baseline channel directly. The two models are tested on the CVIQ and OIQA databases with the same parameters configuration in Section IV-A. We list the performance in Table V and Table VI.

From the Table V and Table VI, we observe that the performance of the MC360IQA with hyper-structure is significantly improved at each statistical index when compared with the model without hyper-structure, and only the value of SRCC of $MC360IQA_{mean}$ is slightly inferior to the model without hyper-structure. It proves that the features from the intermediate layers are beneficial to image quality evaluation. Compared with $MC360IQA_{mean}$, the performance of $MC360IQA_{first}$ is improved more obviously, which indicates that the hyper-structure can make the $MC360IQA$ more stable.

TABLE V
PERFORMANCE COMPARISON BETWEEN MC360IQA MODELS WITH AND WITHOUT HYPER-STRCUTURE ON THE CVIQ DATABASE USING VQEG CRITERIA. THE BETTER MODEL IS HIGHLIGHTED AT EACH EVALUATION CRITERION

Metrics		SRCC	PLCC	RMSE
Pro.	no hyper	0.9069	0.9361	4.2349
	hyper	**0.9113**	**0.9503**	**4.1484**
Pro.+	no hyper	**0.9153**	0.9480	3.7621
	hyper	0.9139	**0.9506**	**3.0935**

TABLE VI
PERFORMANCE COMPARISON BETWEEN MC360IQA MODELS WITH AND WITHOUT HYPER-STRCUTURE ON THE OIQA DATABASE USING VQEG CRITERIA. THE BETTER MODEL IS HIGHLIGHTED AT EACH EVALUATION CRITERION

Metrics		SRCC	PLCC	RMSE
Pro.	no hyper	0.8872	0.8693	5.7543
	hyper	**0.9020**	**0.9045**	**5.2704**
Pro.+	no hyper	0.9134	0.8971	5.0319
	hyper	**0.9187**	**0.9247**	**4.6247**

D. Model Sensitivity for the Longitude of the Front View

In Section III-B, we propose to rotate the longitude of the front view to project one omnidirectional image onto several groups of viewport images to argument the database. Theoretically, we can obtain a quality score from each group of viewport images using the MC360IQA model and the quality scores should be the same when their corresponding group of images are projected by one omnidirectional image. Therefore, we study the sensitivity

TABLE VII
THE PERFORMANCE OF THE MC360IQA MODEL TRAINED ON THE OIQA DATABASE AND TESTED ON THE CVIQ DATABASE

Metrics	JPEG			AVC			HEVC			ALL		
	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE
Pro.	0.8396	0.8969	4.1802	0.8822	0.8657	3.9854	0.8261	0.8206	3.6823	0.8560	0.8596	3.9854
Pro.+	0.8234	0.8772	4.6072	0.8728	0.8349	3.8295	0.8240	0.7785	4.3618	0.8442	0.8249	4.5423

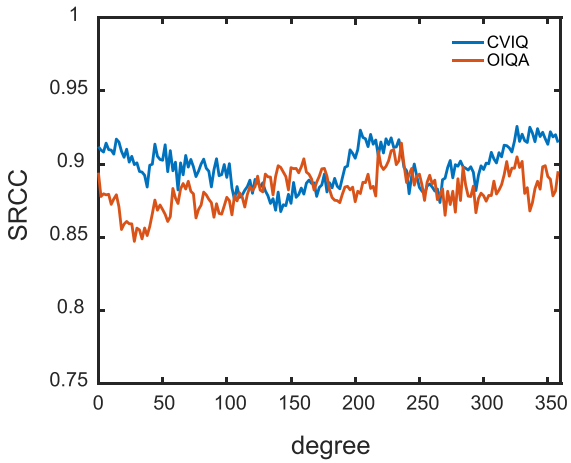


Fig. 9. The SRCC of the $MC360IQA_i$ on the CVIQ and OIQA databases.

of the MC360IQA model to the different groups of viewport images projected by the same image with different longitudes of the front view in this section.

We have defined two MC360IQA metrics $MC360IQA_{origin}$ and $MC360IQA_{mean}$. Here, we calculate the performance of $MC360IQA_i$ on the CVIQ and OIQA databases to show the stability of the MC360IQA. The metric $MC360IQA_i$ is defined as:

$$MC360IQA_i = MC360IQA(VP^i_{view}), \tag{8}$$

where $view \in \{front, back, right, left, top, down\}$ and $i \in [1, 2, \ldots, N]$, $N = 360/\varphi$.

We use the SRCC to evaluate the performance of MC360IQA since it is an important criterion which demonstrates the convergence and monotonicity between the subjective perception and objective quality metric. The variation tendency of SRCC is shown in Fig. 9. We observe that almost all values of SRCC exceed 0.85, which indicates that any of them can effectively evaluate the quality of omnidirectional images. What's more, the maximum and minimum values of SRCC differ by 0.058 in the CVIQ database, and by 0.0672 in the OIQA database, which shows the variation of the curves is small and demonstrates that our model is not sensitive to different sets of images when projected by the same image with different longitudes of the front view.

E. Cross-Dataset Evaluation

In this section, we test the generalization capability of the MC360IQA via cross-dataset evaluation. Specifically, we use the CVIQ and OIQA database to train the MC360IQA model and then use the OIQA and CVIQ database to test the corresponding model, respectively. We consider the OIQA database contains more distortion types such as Gaussian noise and Gaussian blur while the CVIQ database does not include, so we just test JPEG compression and JP2K compression on the model trained by CVIQ database. The parameters configuration is also as the same as in Section IV-A. We list the corresponding results in Table VII and Table VIII.

Since JPEG compression is the only distortion type which appears in both databases, we find that the performance of the $MC360IQA$ on JPEG compression is pretty good on both databases from the Table VII and Table VIII. However, we observe that the performance on JPEG compression does not reach to the performance in Section IV-B. The reason may be that distortion levels for each image in the two databases are not the same. It is found that each reference image in the OIQA database is compressed by JPEG with 5 different quality factors while in the CVIQ database, each reference image is compressed by JPEG with 11 different quality factors. Also, we see that the performance on JPEG compression in Table VIII is better than the performance in Table VII, which is due to the more distortion levels in the CVIQ database. Apart from JPEG compression, we notice that the MC360IQA model also achieves good performance on ACV and HEVC compression in Table VII, though these two compression types do not appear in the training database. The performance of MC360IQA on JPEG2000 compression drops a little in Table VII, but it is still satisfied since the training database does not contain the JPEG2000 compression images. Overall, we can conclude that the proposed MC360IQA model has a strong generalization capability, which demonstrates that the MC360IQA can be applied to multiple circumstances.

F. Computational Complexity

To compare the computational complexity of the MC360IQA model with other IQA models, we report the running time for 100 images with a fixed resolution of 4096 × 2048 on a computer with 4.00 GHz Intel Core i7-6700 K CPU, 32 GB RAM, and NVIDIA GTX 1080. S-PSNR, WS-PSNR, and CPP-PSNR are implemented using the code.[1] The codes of other IQA models are provided by authors. The running time of each model is listed in Table IX. It is observed that the metric $MC360IQA_{origin}$ has the considerably low running time with the support of the Graphics Processing Units (GPU). We also test the running time of the MC360IQA model without GPU, and

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

TABLE VIII
THE PERFORMANCE OF THE MC360IQA MODEL TRAINED ON THE CVIQ DATABASE AND TESTED ON THE OIQA DATABASE

Metrics	JPEG			JP2K			ALL		
	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE
Pro.	0.8407	0.8905	4.5669	0.6298	0.6634	5.5108	0.7012	0.7509	6.1373
Pro.+	0.8412	0.8898	4.3950	0.6221	0.6211	5.1294	0.6981	0.7443	5.9184

TABLE IX
COMPARISON OF COMPUTATIONAL COMPLEXITY FOR 11 STATE-OF-ART FR AND NR IQA MODELS AND TWO PROPOSED METRICS. THE TWO PROPOSED METRICS ARE IMPLEMENTED USING THE GRAPHIC PROCESSING UNITS (GPU). TIME: SECONDS/IMAGE

Metrics	PSNR	S-PSNR	WS-PSNR	CPP-PSNR	SSIM	MS-SSIM	QAC	GMLF	NIQE	BRISQUE	SISBLIM	Pro.	Pro.+
Time	0.0197	0.7116	0.1056	7.2035	0.1584	0.5835	1.3549	0.9638	3.3419	1.4401	33.9489	0.0359	6.4257

the running time of $MC360IQA_{origin}$ and $MC360IQA_{mean}$ is 0.8179 and 147.2276 seconds/image respectively. We can find the $MC360IQA_{origin}$ implemented without GPU also has competitive running time compared with other NR IQA models. For $MC360IQA_{mean}$, it needs to consume 180 times of running time than that of $MC360IQA_{origin}$. So, it is a bit slow if without the support of GPU. Overall, $MC360IQA_{mean}$ is more suitable for applications which need higher precision but are not sensitivity to running time. Note that if omnidirectional images are not in the cubemap format, we need to convert them into the cubemap format first. It consumes 2.4781 seconds/image using the Matlab code,[2] but it will be faster if use the specialized transform code.

V. CONCLUSION

In this paper, we comprehensively investigate an emerging quality assessment problem of compressed 360-degree spherical images in VR display systems. We first build a compressed VR image quality (CVIQ) database, including 16 sources images and 528 compressed ones under three coding technologies, i.e. JPEG, H.264/AVC, and H.265/HEVC. Then we propose a blind IQA model for 360-degree images, named MC360IQA. The MC360IQA uses the multi-channel CNN architecture to extract the features of viewport images projected by omnidirectional images and then regresses features to objective scores. The hyper-structure is used in each CNN channel to incorporate the features of intermediate layers. According to the using of different groups of viewport images, we propose two IQA metrics. The first one uses the scores calculated by viewport images without data augmentation and the second one uses the mean scores of lots of groups of viewport images with data augmentation via altering the longitude of view direction. The experimental results show that the proposed two metrics achieve the best performance among the state-of-art NR and FR IQA models and also demonstrate the robustness and effectiveness of the MC360IQA model. The link to the CVIQ database is https://github.com/sunwei925/CVIQDatabase.git and we think this study will have great significance to promote the development of objective IQA methods for compressed 360-degree images.

[2][Online]. Available: https://github.com/rayryeng/equi2cubic

REFERENCES

[1] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, and G. 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.*, 2018, pp. 1–6.

[2] W. Sun *et al.*, "MC360IQA: The multi-channel cnn for blind 360-degree image quality assessment," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2019, pp. 1–5.

[3] HUAWEI iLab, "VR Data Report," Shenzhen, China, Huawei, Accessed March 18, 2019. [Online]. Available: http://www-file.huawei.com/-/media/CORPORATE/PDF/ilab/vr-ar-cn.pdf

[4] J.-W. Lin, H. B.-L. Duh, D. E. Parker, H. Abi-Rached, and T. A. Furness, "Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment," in *Proc. IEEE, Virtual Reality*, 2002, pp. 164–171.

[5] 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.

[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. & Comput.*, 2003, vol. 2, pp. 1398–1402.

[7] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[8] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[9] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 43–54, Jan. 2013.

[10] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *IEEE Trans. Broadcast.*, vol. 62, no. 2, pp. 446–456, Jun. 2016.

[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] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, 2010, Art. no. 011006.

[13] 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.

[14] 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. 4270–4281, Oct. 2014.

[15] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

[16] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41–52, Jan. 2012.

[17] K. Gu *et al.*, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

[18] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Trans. Broadcast.*, vol. 60, no. 3, pp. 555–567, Sep. 2014.

[19] G. Zhai, X. Min, and N. Liu, "Free-energy principle inspired visual quality assessment: An overview," *Digital Signal Process.*, vol. 91, pp. 11–20, 2019.

[20] 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.

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

[22] 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.

[23] 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.

[24] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3129–3138, Jul. 2012.

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

[26] S. Yule, A. Lu, and Y. Lu, "WS-PSNR for 360 video objective quality evaluation," MPEG Joint Video Exploration Team, 116, 2016.

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

[28] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[29] 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.*, 2018, pp. 1–6.

[30] Facebook, "Quality assessment of 360 video view sessions," Accessed: Apr. 12, 2019. [Online]. Available: https://code.fb.com/video-engineering/quality-assessment-of-360-video-view-sessions/

[31] E. Upenik, M. Rerabek, and T. Ebrahimi, "On the performance of objective metrics for omnidirectional visual content," in *Proc. IEEE, 9th Int. Conf. Quality Multimedia Experience*, 2017, pp. 1–6.

[32] W. Sun, K. Gu, G. Zhai, S. Ma, W. Lin, and P. L. Callet, "CVIQD: Subjective quality evaluation of compressed virtual reality images," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3450–3454.

[33] 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.*, 2018.

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

[35] 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.*, to be published.

[36] C. Li, M. Xu, X. Du, and Z. Wang, "Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *Proc. ACM Multimedia Conf. Multimedia Conf.*, 2018, pp. 932–940.

[37] H. G. Kim, H.-T. Lim, S. Lee, and Y. M. Ro, "VRSA net: VR sickness assessment considering exceptional motion for 360 VR video," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1646–1660, Apr. 2019.

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

[39] 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.

[40] G. J. Sullivan, J. 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.

[41] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[42] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 818–833.

[43] 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.

[44] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, 2017.

[45] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

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

[47] H. Sheikh, "Live image quality assessment database release 2," 2005. [Online]. Available:http://live.ece.utexas.edu/research/quality

[48] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, 2015.

[49] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Recommendation ITU-R BT. 500-11, ITU Telecommunication Standardization Sector, Geneva, Switzerland, vol. 7, 2002.

[50] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and ima resolution," *IEEE Trans. Broadcast.*, vol. 61, no. 3, pp. 520–531, Sep. 2015.

[51] W. Sun *et al.*, "Dynamic backlight scaling considering ambient luminance for mobile energy saving," in *Proc. IEEE Int. Conf. on Multimedia Expo.*, pp. 25–30.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, vol. abs/1409.1556, May 2015.

[53] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1–9.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in in *Proc. Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[55] Y. Zhu, G. Zhai, and X. Min, "The prediction of head and eye movement for 360 degree images," *Signal Process.: Image Commun.*, vol. 69, pp. 15–25, 2018.

[56] J. Ling, K. Zhang, Y. Zhang, D. Yang, and Z. Chen, "A saliency prediction model on 360 degree images using color dictionary based sparse representation," *Signal Process.: Image Commun.*, vol. 69, pp. 60–68, 2018.

[57] Y. Fang, X. Zhang, and N. Imamoglu, "A novel superpixel-based saliency detection model for 360-degree images," *Signal Process.: Image Commun.*, vol. 69, pp. 1–7, 2018.

[58] M. Startsev and M. Dorr, "360-aware saliency estimation with conventional image saliency predictors," *Signal Process.: Image Commun.*, vol. 69, pp. 43–52, 2018.

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

[60] A. Paszke *et al.*, "Automatic differentiation in pytorch," 2017.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," PP*Proceedings of the 25th International Conference on Neural Information Processing SystemsProc. Advances Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.

[62] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013.

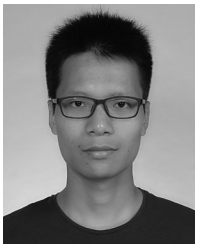
[63] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[64] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[65] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 995–1002.

[66] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Trans. Broadcast.*, vol. 60, no. 3, pp. 555–567, Sep. 2014.
[67] W. Zhu *et al.*, "Multi-channel decomposition in tandem with free-energy principle for reduced-reference image quality assessment," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2334–2346, Sep. 2019.
[68] X. Min, G. Zhai, K. Gu, X. Yang, and X. Guan, "Objective quality evaluation of dehazed images," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2879–2892, Aug. 2019.
[69] X. Min *et al.*, "Quality evaluation of image dehazing methods using synthetic hazy images," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2319–2333, Sep. 2019.
[70] 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 *Proc. IEEE, 8th Int. Conf. Quality Multimedia Experience*, 2016, pp. 1–6.
[71] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi, "How to benchmark objective quality metrics from paired comparison data?" in *Proc. IEEE, 8th Int. Conf. Quality Multimedia Experience*, 2016, pp. 1–6.
[72] L. Krasula, P. Le Callet, K. Fliegel, and M. Klíma, "Quality assessment of sharpened images: challenges, methodology, and objective metrics," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1496–1508, Mar. 2017.
[73] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Preference of experience in image tone-mapping: Dataset and framework for objective measures comparison," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 64–74, Feb. 2017.
[74] M. H. Brill, J. L, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1AI1," *Signal Process.: Image Commun.*, vol. 19, no. 2, pp. 101–107, 2004.
[75] J. A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Hove, U.K.: Psychology Press, 2014.
[76] J. Hanley and B. McNeil, "A method of comparing the area under two roc curves derived from the same cases," *Radiology*, vol. 148, pp. 839–843, 1983.

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



Xiongkuo Min (M'19) 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. From Jan. 2016 to Jan. 2017, he was a Visiting Student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Postdoctoral Fellow with Shanghai Jiao Tong University, Shanghai, China. His research interests include visual quality assessment, visual attention modeling, and perceptual signal processing. He was the recipient of the Best Student Paper Award at IEEE ICME, 2016.

Guangtao Zhai (M'10) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a Research Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Postdoctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. His research interests include multimedia signal processing and perceptual signal processing. He was the recipient of the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012.

Ke Gu (M'19) 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 research interests include environmental perception, image processing, quality assessment, and machine learning. 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 (DSP). He is currently an Associate Editor for the IEEE ACCESS and *IET Image Processing (IET-IPR)*, and an Area Editor for *Signal Processing Image Communication (SPIC)*. He is a Reviewer for 20 top SCI journals. He was the 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.

Huiyu Duan received the B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. He is currently working toward the Ph.D. degree with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. He is currently a Visiting Ph.D. Student with the Schepens Eye Research Institute, Harvard Medical School, Boston, USA. His research interests include visual quality assessment, visual attention modeling, perceptual signal processing, and psychovisual modulation in virtual reality.

Siwei Ma (M'03–SM'12) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He held a postdoctoral position with the University of Southern California, Los Angeles, CA, USA, from 2005 to 2007. He joined the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing, China, where he is currently a Professor. He has authored over 200 technical articles in refereed journals and proceedings in image and video coding, video processing, video streaming, and transmission. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *Journal of Visual Communication and Image Representation*.