Multi-task Guided No-Reference Omnidirectional Image Quality Assessment with Feature Interaction

Yun Liu, *Member, IEEE,* Sifan Li, Huiyu Duan, Yu Zhou, *Member, IEEE,* Daoxin Fan, and Guangtao Zhai, *Fellow, IEEE*

***Abstract*—Omnidirectional image quality assessment (OIQA) has become an increasingly vital problem in recent years. Most previous no-reference OIQA methods only extract local features from the distorted viewports, or extract global features from the entire distorted image, lacking the interaction and fusion between local and global features. Moreover, the lack of reference information also limits their performance. Thus, we propose a no-reference OIQA model which consists of three novel modules, including a bidirectional pseudo-reference module, a Mamba-based global feature extraction module, and a multi-scale local-global feature aggregation module. Specifically, by considering the image distortion degradation process, a bidirectional pseudo-reference module capturing the error maps on viewports is first constructed to refine the multi-scale local visual features, which can supply rich quality degradation reference information without the reference image. To well complement the local features, the VMamba module is adopted to extract the representative multi-scale global visual features. Inspired by human hierarchical visual perception characteristics, a novel multi-scale aggregation module is built to strengthen the feature interaction and effective fusion which can extract deep semantic information. Finally, motivated by the multi-task managing mechanism of human brain, a multi-task learning module is introduced to assist the main quality assessment task by digging the hidden information in compression type and distortion degree. Extensive experimental results demonstrate that our proposed method achieves the state-of-the-art performance on the no-reference OIQA task compared to other models.**

***Index Terms*—Bidirectional pseudo reference, omnidirectional image quality assessment, Mamba, multi-scale aggregation, multi-task learning, no-reference (NR).**

I. Introduction

QUALITY degradation exists in various multimedia contents, which can degrade their perceptual quality and limit the applications. The quality degradation problem is more serious in omnidirectional scenarios due to the increased data volume, and may affect the quality of experience (QoE) more due to the immersive nature of virtual reality (VR) [1]. Therefore, it is significant to develop more effective quality assessment method to further help optimize the QoE in VR environment [2]–[4].

As a fundamental problem of QoE assessment, the task of image quality assessment (IQA) has attracted remarkable attention for a long time [5], [6], including two-dimensional (2D) image, omnidirectional image (OI) and so on [7]–[9]. Among them, many omnidirectional image quality assessment (OIQA) methods have been proposed in recent years with the development of VR area, including full-reference (FR) metrics [10]–[13], reduced-reference (RR) metrics [14], [15] and no-reference (NR) metrics [16]–[19] according to whether the reference information is introduced. In the early stage, many FR OIQA methods have extended previous 2D IQA models such as peak-signal-to-noise ratio (PSNR) and structural similarity (SSIM) into omnidirectional projection to perform evaluation [10], [20]. Since an omnidirectional image (OI) has a wider viewing range and more complicated perceptual features than a traditional 2D image, the above 2D IQA-based models lead to mediocre performances. In addition, for real applications, the reference image for the omnidirectional image is not always available, which makes the NR OIQA metrics have more practical significance than the other two types of OIQA methods.

Motivated by the important role of human visual system (HVS) in image quality assessment area, many NR OIQA models have been proposed by capturing representative visual features and semantic information [21]–[23], which can be categorized into three types: image-based, patch-based, and viewport-based. Image-based models generally treat an omnidirectional image as a 2D image to capture global visual information [24], [25]. However, due to the wide scene range and large image size, such methods may ignore local visual distortions, which limits the performance [26], [27]. Because the HVS is extremely sensitive to local information, many patch-based models have been proposed based on the cropped 2D image patches from an omnidirectional image, which obtains rich local information to improve the overall performance [28], [29]. Motivated by the viewing characteristics, some viewport-based models have achieved good performance based on the features extracted from each field of view (FoV). Moreover,

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

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

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

Yu Zhou is with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China, and also with Xuzhou First People's Hospital, Xuzhou 221116, China (e-mail: zhouy@cumt.edu.cn).

considering the vital role of local and global features in IQA tasks, some OIQA models have been proposed by integrating local and global information, which presents an effective way to obtain the representative visual features [30]–[32].

As the boost of deep learning, the deep neural network (DNN) based methods have started to show their ability and have become the mainstream of OIQA models [33]–[35]. Inspired by the human visual system, Jiang *et al.* [36] built an effective network by mimicking human visual perception. Xu *et al.* [22] proposed a GCN based OIQA method with an elaborated viewports choosing algorithm, which proves the importance of local features. Zhang *et al.* [37] proposed a deep-learning based joint network to model the no-reference quality assessment of omnidirectional images by considering the viewports visual features. However, few works pay attention to multi-scale local and global information, and their interactive relationship. To thoroughly learn the global and local features in OIs, Zhou *et al.* [38] proposed a perception-oriented u-shaped Transformer Network, which proves the importance of local and global features in OIQA area. Inspired by this, Tofighi *et al.* [39] introduced the local global Transformer for OIQA by integrating local and global information, which offers an effective IQA method for OIs. Later, Fan *et al.* [40] proposed an omnidirectional image assessment network, which proves the importance of multiscale feature extraction and fusion on the representation of distortion information. Although further progress has been achieved by the aforementioned OIQA models, significant efforts on digging representative local and global features and interactive relationship should be made to build more effective models.

To deal with the problems and challenges mentioned above, we propose an NR OIQA network by deeply fusing representative multi-scale local and global semantic information. Specifically, inspired by the pseudo reference conception proposed in [41], [42], we first propose a bidirectional pseudo reference module to extract multi-scale local semantic information from two directions: the restoration direction and the degradation direction, which can capture important quality changing information in distorted image without the help of reference image to augment the prediction accuracy. Inspired by VMamba's unprecedented capability in extracting features from long range images or texts [43], a global feature extraction module based on VMamba is built to obtain multi-scale global information. Then motivated by human hierarchical visual perception characteristics, a multi-scale aggregation module is adopted to extract the interactive information and refine the shared features, which can achieve a better performance than the simple fusion or concatenation way. To further optimize the learning process, we apply a multi-task module to assist the model to adaptively assign weights among different tasks, which can yield a more stable performance. Our main contributions are summarized as follows:

1) A bidirectional pseudo reference module is proposed to extract representative local differences from two opposite directions, which can well reflect the local quality degradation and improve the feature representation.
2) A VMamba-based feature extraction module is designed to catch efficient multi-scale global visual information to well complement the local features, which can reduce the data volume burden of the model and improve the overall performance.
3) A multi-scale interactive feature fusion module is introduced to the OIQA task to strengthen the feature interaction and deep fusion, which can improve the accuracy of our model.
4) A multi-task learning module is designed to guide the model to adaptively assign weights among different degradations, which can further improve the effectiveness of our model.

To illustrate the idea more structurally, we arrange the remainder of this paper as follows. The related works of this paper are briefly reviewed in section II. Our proposed method is introduced in detail in section III, and the experimental results and the analysis are reported in section IV. Finally, the conclusion of this work is presented in section V.

II. Related Work

We review the related works including the previous OIQA models and the VMamba structure in this section.

A. OIQA Models

FR OIQA metrics require all information of the reference image, which can easily obtain the quality difference between a distorted image and a reference image. Many traditional FR OIQA methods extended the previous 2D IQA models to evaluate the quality, such as PSNR and SSIM. The spherical domain-based model (S-PSNR) was proposed by calculating the PSNR value in the spherical domain. Then a CPP-PSNR metric was designed by calculating the PSNR value in the space of Craster parabolic projection (CPP) [26]. Motivated by the above works, the weighted-to-spherically-uniform PSNR and spherical domain-based SSIM models were built [44], [45]. Although the above FR OIQA models present relatively satisfactory results at the early stage, they were designed based on 2D IQA metrics, and failed to obtain specific visual features of OI, which limits further development of the omnidirectional image quality assessment.

FR OIQA model presents a way to capture the quality degradation in distorted image, but the reference information is generally missing in real applications, which makes the NR OIQA models more practical and popular [46], [47]. The existing NR OIQA methods can be categorized into three types, which are the whole image-based methods, patch-based methods and viewport-based methods [21]. The whole image-based methods took the equirectangular projection (ERP) image as the input and directly calculated the image quality [27]. The patch-based methods mainly focus on seeking for a better representation space to obtain more effective features, which digs deeply into the characteristics of the projection methods including the segmented spherical projection (SSP), cube map projection (CMP), equirectangular projection (ERP) [21], etc. For example, in [8], an NR OIQA metric were designed for OIs on the SSP space based on the local details and global features in both bipolar regions of the reprojection space. Jiang *et al.* [36] focused on the local visual features

in the CMP space and built a NR OIQA work. Kim *et al.* [48] introduced a no-reference patch based OIQA metric by segmenting the OI in ERP format into non-overlapping patches in a uniformed size and exploring the positional features of them. Liu *et al.* [49] introduced an effective quality assessment metric by fusing the local structural features and global natural features based on the OI in ERP format.

The viewport-based NR OIQA metrics aim to simulate the visual mechanism when human watching the VR contents [50]. In [22], the final quality score was obtained by calculating both the global prediction quality of the entire image and the local prediction quality of the viewports. Later, considering the importance of viewports in six directions, a multi-channel viewport-based method was designed [35]. Based on scene content understanding, Zhang *et al.* [51] invented a saliency-guided NR OIQA method by fusion multi-scale features of each viewport. Considering the quality degradation in an image is related to the type and degree of the distortion, the auxiliary task for distortion type discrimination was utilized in a multi-stream network [52], which motivates us to build an effective multi-task model. Overall, the above works proves that multi-scale features from both the local and global degrees are particularly important in the OIQA task.

Although NR OIQA models perform better than FR models, there is still room for improvement to make the OIQA more effective. More accurate local and global visual features need to be obtained, and the interactive relationship between them should be considered in OIQA models. In addition, it is hard to capture quality changing information without the reference image. An effective module that can capture important quality degradation information hidden in the distorted image should be deeply dug. Overall, OIQA is a complex and challenging work, which needs to take many factors into consideration, such as human visual characteristics, quality degradation between the distorted image and the reference, etc.

B. OSIQA Models

By combining multiple images with overlapping areas, it can produce an image with wide field-of-view (FoV) and high resolution, which is called omnidirectional stitching image. Omnidirectional stitching image quality assessment (OSIQA) is a task aiming to assess the quality of omnidirectional stitching images (OSIs) that contains multiple stitching distortions. Tian *et al.* [53], [54] proposed some effective stitched omnidirectional image quality evaluators by considering the viewport-based visual features. Duan *et al.* [55] established an omnidirectional stitching image quality assessment (OSIQA) dataset, and proposed an effective IQA model. Zhou *et al.* [56] proposed a hierarchical quality prediction method for stitched panoramic images by aggregating pyramid features. Later, they introduced another method for stitched panoramic image quality assessment with patch registration and bidimensional feaeture aggregation [57]. By applying integration of spatio-temporal feature, Hu *et al.* [58] introduced an effective OSIQA method. Although many works have been proposed on the omnidirectional image quality assessment, the performance of the omnidirectional stitching quality has not been sufficiently explored. Since OSI is still an omnidirectional image, OSIQA task can be applied to validate the generalization ability of the OIQA model.

C. Mamba-based Vision Models

Mamba [59] has recently drawn considerable attention in various areas, which yields significant results in long sequence modeling tasks [60]–[62]. Mamba consists of repeated Mamba blocks with state space model (SSM) blocks [63]–[65], standard normalization layers, and residual connections, which relieves the constraints of modeling in a convolutional neural network (CNN) [66]–[68]. Compared to Transformer [69], Mamba provides us with advanced and excellent modeling capabilities, but without secondary computing complexity. The significant advantages over CNNs and Transformer demonstrate Mamba's enormous potential as a base model for vision tasks, which promotes its further development.

Inspired by Mamba, VMamba [70] was proposed as an efficient model based on down-sampling operations [71] and Visual State Space (VSS) blocks with 2D-selective-scan (SS2D) blocks [43], as shown in Fig. 5. VMamba has bidirectional selective state space model (SSM) blocks [59] along 2D axes by integrating the information from all the other four pixels in different directions around each pixel [43], which can capture rich global semantic information by combining the information of each pixel and reduce the time complexity. Many works have introduced VMamba to various visual tasks and achieved significant performance [72]–[76]. For example, Xie *et al.* [77] introduced VMamba and reformed it in dynamic feature enhancement for multi-modal image fusion, which performs well in medical imaging. Yang *et al.* [78] presented the advantages of Mamba in feature extraction by inventing a scheme with it for image segmentation, which achieves good results. Shi *et al.* [74] introduced a network with VMamba for image restoration and a state-of-the-art performance was achieved. Ma *et al.* [76] then applied VMamba on crowd counting work to solve the problems in counting specific points of a scene, and invented a new approach that inherited the merits of VMamba for global modeling and low computational costs, which achieves a remarkable performance. Considering the remarkable effectiveness of VMamba in the 2D image area, especially its significance for long-range modeling, we choose it as the backbone of the global branch of our model to capture the global semantic features.

III. Proposed Method

In this section, our proposed model is described in detail. The overall framework is shown in Fig. 1. Specifically, the distorted OI is fed into a bidirectional pseudo-reference (BPR) module to obtain two opposite pseudo-reference images, and then split into viewports to extract the multi-scale local features based on the error maps. To well complement the local features on viewports, a VMamba module [70] is adopted to extract the multi-scale global features. Then the local features and global features are fused from shallow to deep based on the Bi-Stream Multi-Scale Fusion Aggregation (BS-MSFA) module. Finally, a multi-task learning module is introduced

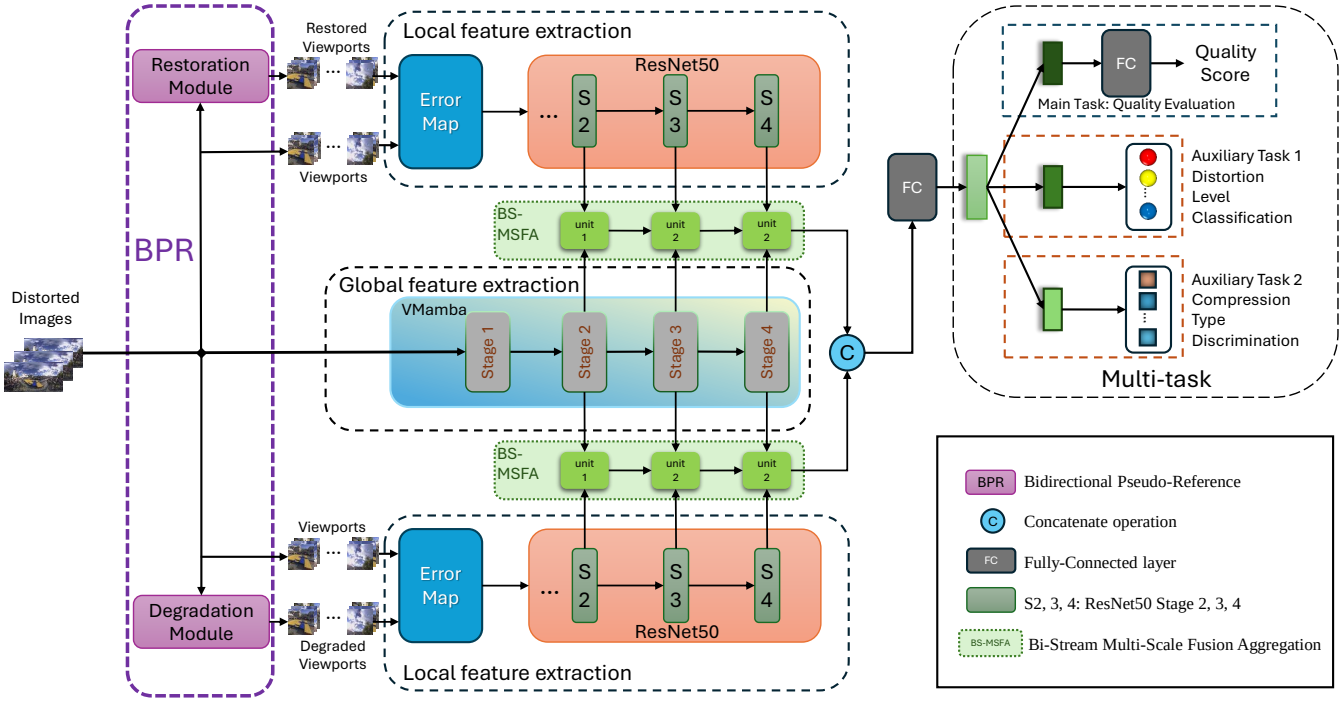

Fig. 1. Overview architecture of our proposed method.

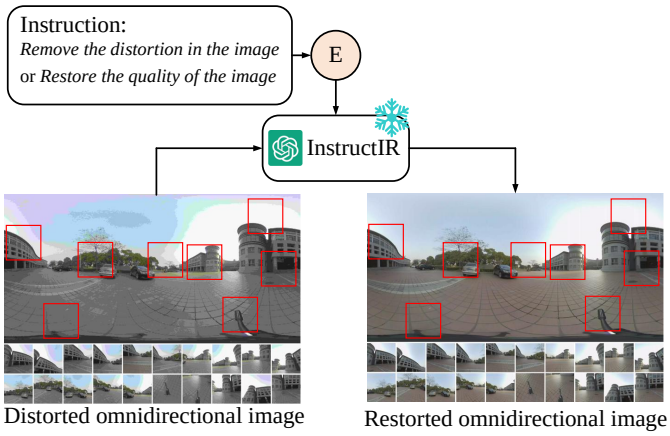

Fig. 2. The process of restoring a distorted omnidirectional image with textual prompts by using InstructIR. E represents the embedding process.

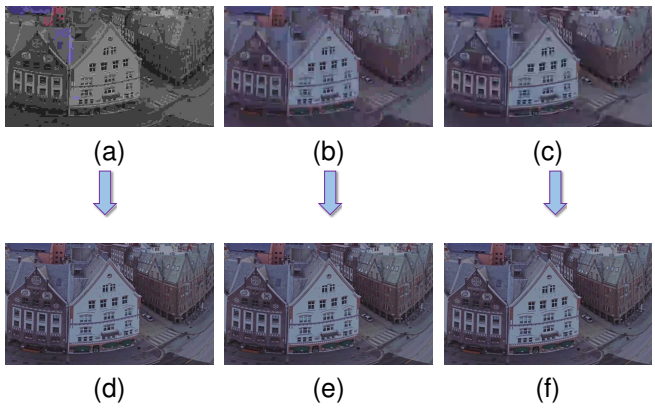

Fig. 3. The distorted viewports and their restored viewports. (a), (b) and (c) are the viewports of the OIs with different types of distortions. (d), (e) and (f) are the restored viewports of (a), (b) and (c), respectively.

to assist the main quality assessment task. All the modules in our method are present in detail in this section.

A. The Bidirectional Pseudo-Reference (BPR) Module

The BPR module is designed to obtain the pseudo-reference information from two opposite directions. Considering the image distortion degradation process, we introduce a bidirectional pseudo-reference module to obtain two pseudo-reference images by restoring and degrading the distorted image, which is presented as the restoration and degradation modules, respectively, in Fig. 1. For the restoration module, we adopt a novel generative model InstructIR [79], which is designed based on the famous large language model GPT-4 and the image restoration technique [83]. The method achieves excellent performance in image restoration. Fig. 2 shows how it works to restore a distorted image. We utilize this model with the textual instructions to generate the restored image, such as *"Remove the distortion in the image"* or *"Restore the quality of the image"*. Since InstructIR relies on the semantic content within the textual prompt, similar textual prompt also works. Fig. 3 presents the viewports of the distorted OI with different degrees of distortions and their corresponding restored viewports. It can be seen that the restored viewport has a better quality than the distorted viewport, which proves that the restored viewport can well reflect the quality difference.

For the degradation module, we obtain the degraded images by adding shuffled blur, down-sampling, and noise following

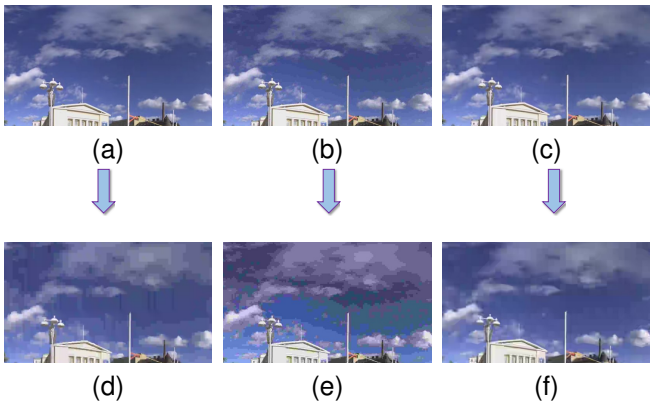

Fig. 4. The distorted viewports and their degraded viewports. (a), (b) and (c) are the viewports of the distorted OIs with different types of distortions (JPEG [80], H.264/AVC [81] and H.265/HEVC [82], respectively). (d), (e) and (f) are the degraded viewports of (a), (b) and (c), mainly adding JPEG compression in level 3, Gaussian noise with $\sigma = 0.008$ and camera sensor noise in level 3, respectively.

the work [84]. Particularly, the down-sampling is randomly used from the nearest, bilinear or bicubic interpolations, and the noise is synthesized by adding Gaussian noise in different levels, JPEG compression or camera sensor noise, which generates the degraded images with random types and random levels of distortions to describe the distorted images in reality. Specifically, JPEG compression and camera sensor noise level range from 0 to 5. The higher the level, the worse the quality. The standard deviation of Gaussian noise range from 0 to 12 ($\times 10^{-3}$). The image gets blurrier with the standard deviation increases. It needs to be mentioned that the overall performance of our model is similar by training the degradation module with a specific type of distortion to random distortions. Since the type and level of the distortion is random in real applications, we train the degradation module by adding random types and levels of distortions and test the model also with the random degradations. Fig. 4 presents the viewports of the distorted OI with different degrees of distortions and their corresponding degraded viewports. Three types of distortions with different levels are randomly aggravated to the inputs to generate the degraded images that has worse qualities. Other than restored images that present the reference score from a positive degree, degraded PR images can help the prediction model compare with the worst quality score, which can reflect the quality difference from an opposite direction.

B. Local Feature Extraction

With the help of the two pseudo-reference images, the local features extraction can be conducted based on the quality degradation information between the distorted image and its pseudo-reference image. Considering that only one local viewport of an omnidirectional image is watched at a time for a specific user, 20 viewports are first generated following the work in [54]. To effectively capture the presentative quality changing between the PR viewports and the distorted viewports, error maps are calculated to capture the rich semantic difference information from the two directions. Here, in order to obtain a better correlation with the quality perceived by viewers, the normalized log difference function [85] is adopted, which is defined as in (1):

$$E = \log_{\alpha}\left(\alpha + (I_r - I_d)^2\right), \tag{1}$$

where $\alpha = \epsilon/255^2$ is a constant with $\epsilon = 0.1$, I_r is the value of each pixel of the PR viewport, and I_d is the value of each pixel of the distorted viewport. With the log normalization processing, the pixel values can be adjusted to a narrow range, which can assist to capture important changes in an error map [86]. After the necessary procedure, the error map is then fed into the local feature extraction module. In this paper, we elaborately take ResNet50 as the backbone to obtain the local features. The ResNet50 consists of five stages: stage 0, 1, 2, 3 and 4, respectively, in which stage 1 includes three bottlenecks, and stage 2, 3 and 4 include 4, 6 and 3 bottlenecks, respectively [87]. We take the features from the last three stages as local features in our method.

C. Global Feature Extraction

To well complement the above local feature, the global feature is extracted based on the whole omnidirectional image instead of the viewports. Since VMamba recently presents a remarkable performance in feature extraction in deep learning [70], we take it as the backbone of the global feature extraction module, shown in Fig. 5. For an OI, it is firstly partitioned into patches with a stem module, and a 2D feature map $\boldsymbol{M}_d \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}}$ with the spatial dimension of $\frac{H}{4} \times \frac{W}{4}$ [70] is consequently obtained to feed into the VMamba, where H and W are the height and width of the image. Denoting the output of stage n as $\boldsymbol{v}_n$, the output of the first stage can be obtained as in (2) and (3):

$$\boldsymbol{T}_1 = \text{SS2DB}(\text{LN}(\boldsymbol{M}_d)) \oplus \boldsymbol{M}_d, \tag{2}$$

$$\boldsymbol{v}_1 = \text{FFN}(\text{LN}(\boldsymbol{T}_1)) \oplus \boldsymbol{T}_1. \tag{3}$$

For the output of stage $n(n = 2, 3, 4)$ can be obtained as in (4) and (5):

$$\boldsymbol{T}_n = \text{SS2DB}(\text{LN}(\text{DS}(\boldsymbol{v}_{n-1}))) \oplus \text{DS}(\boldsymbol{v}_{n-1}), \tag{4}$$

$$\boldsymbol{v}_n = \text{FFN}(\text{LN}(\boldsymbol{T}_n)) \oplus \boldsymbol{T}_n, \tag{5}$$

where SS2DB$(\cdot)$ denotes the SS2D block [70], LN$(\cdot)$ means the layer normalization, FFN$(\cdot)$ is the feedforward neural network, and DS$(\cdot)$ is down-sampling operation. The layer normalization helps accelerate the training process to converge, and down-sampling operation reduces the size of the matrix in the procedure. The SS2D block is defined as in (6) and (7):

$$F_{SS2D}(\cdot) = \text{SS2D}(\text{SiLU}(\text{Dwc}(\text{Linear}(\cdot)))), \tag{6}$$

$$\text{SS2DB}(\cdot) = \text{Linear}(\text{LN}(F_{SS2D}(\cdot))), \tag{7}$$

where SS2D$(\cdot)$ is the 2D-Selective-Scan operation [88], and Dwc$(\cdot)$ is a 3×3 depth-wise convolution layer. SiLU function

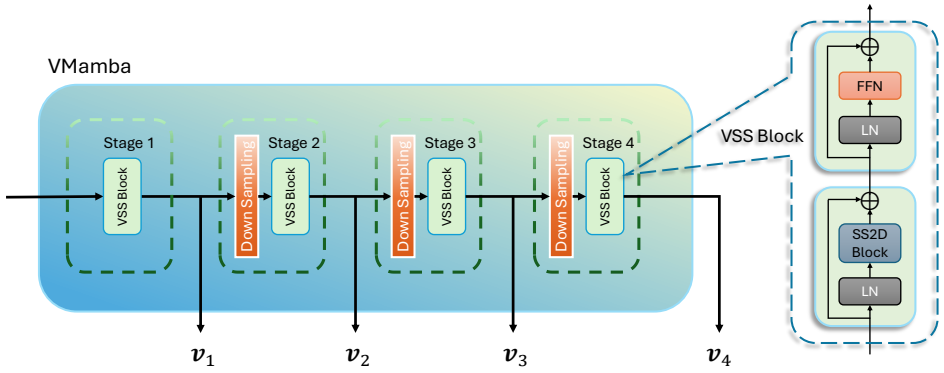

Fig. 5. The structure of VMamba [70] and VSS block.

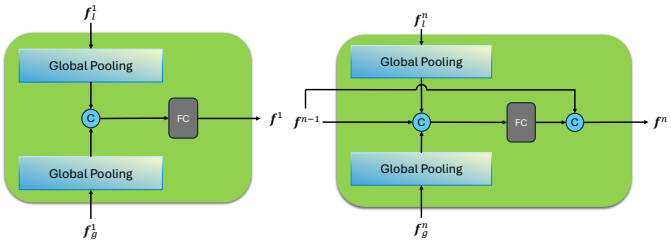

Fig. 6. BS-MSFA module: (a) is BS-MSFA UNIT 1, and (b) is BS-MSFA UNIT 2. © represents concatenate operation, and FC represents a fully-connected layer.

is an activation function used commonly in deep models. The linear operation helps VMamba to optimize the training process. The outputs $\boldsymbol{v}_2$, $\boldsymbol{v}_3$, and $\boldsymbol{v}_4$ are then prepared for the multi-scale fusion.

D. Bi-Stream Multi-Scale Feature Aggregation (BS-MSFA)

Bi-stream multi-scale feature aggregation (BS-MSFA) module is designed to fuse the extracted local and global features. To fully apply the interactive relationship between local and global features, we adopt an interactive fusion module to extract the interactive information and refine the shared features. Due to the task insensitivity, the regression learning module based on the one layer output fails to give a satisfactory performance. The multi-scale representation can provide us with a new way to capture the crucial features [40], [89]. Here, we apply the outputs of last three stages of ResNet50 as the multi-scale local features and take the outputs of the last three VMamba stages as the multi-scale global features. To avoid the rigid connection between multi-scale features, the interactive fusion module with residual structure, namely Bi-Stream Multi-Scale Feature Aggregation (BS-MSFA) module, is proposed to reduce the shallow feature dimension and achieve efficient fusion by digging the interactive relationship between local and global features, shown in Fig. 6.

Specifically, BS-MSFA consists of one BS-MSFA unit 1 and two BS-MSFA unit 2. For the unit 1, as shown in Fig. 6a, two features from both branches are fed into it to capture the first level fusion features $\boldsymbol{f}^1$, which is defined as in (8):

$$\boldsymbol{f}^1 = \mathrm{FC}(g(\boldsymbol{f}_l^1)©g(\boldsymbol{f}_g^1)), \tag{8}$$

where $\mathrm{FC}(\cdot)$ denotes a fully-connected layer, $g(\cdot)$ denotes a global pooling operation, and © means the concatenate operation. $\boldsymbol{f}_l^1$ and $\boldsymbol{f}_g^1$ are the first level of local features and global features, respectively.

Then the first level fusion features $\boldsymbol{f}^1$, the second level local features and the second level global features are then fed into a BS-MSFA unit 2, as shown in Fig. 6b, to capture the second level fusion features $\boldsymbol{f}^2$. Before feeding $\boldsymbol{f}^1$ into the BS-MSFA unit 2, $\boldsymbol{f}^1$ is processed with a fully-connected layer. The second level fusion feature $\boldsymbol{f}^2$ is shown as in (9):

$$\boldsymbol{f}^2 = \boldsymbol{f}^1©\mathrm{FC}(\boldsymbol{f}^1©g(\boldsymbol{f}_l^2)©g(\boldsymbol{f}_g^2)). \tag{9}$$

The same as the above procedure, the third level fusion features are obtained. Then the multi-scale fusion features of the 20 viewports from the restoration viewports and degradation viewports are concatenated together and processed by a fully-connected layer for the multi-task module outputing the final quality score.

E. The Multi-Task Module

The multi-task module can motivate the model to focus on the representative information. Considering that the compression types and distortion degrees have different impacts on quality perception, we design a multi-task module to refine the shared information and optimize the model's performance by the assistance of the distortion level classification task and compression type discrimination task. For the main task, a fully-connected layer is adopted to adjust the size and get the final quality score. For the two auxiliary tasks, two fully-connected layers containing 1024 nodes and 64 nodes, respectively, are utilized.

Specifically, in our multi-task procedure, Euclidean loss function and cross-entropy loss function are adopted to optimize the network. The Euclidean loss is defined as in (10):

$$L_q = \frac{1}{N}\sum_{k=1}^{N}\left\|s_k - q_k\right\|_2^2, \tag{10}$$

where k denotes the k-th sample. s_k is the subjective quality score, and q_k is the predicted quality score.

The cross-entropy loss based on the distortion level classification task is defined as in (11):

$$L_d = -\sum_{k=1}^{N}\sum_{i=1}^{M} \boldsymbol{m}_k^i \log \hat{p}_k^i, \tag{11}$$

where $\boldsymbol{m}_k^i$ is the ground-truth multi-class indicator vector. If the k-th sample is in the i-th distortion level, then $\boldsymbol{m}_k^i$ will be one, otherwise $\boldsymbol{m}_k^i$ will be zero. $\hat{p}_k^i$ denotes the predicted probability of whether the distortion in the k-th sample is in the i-th distortion level.

Similar to the loss for the distortion level classification, the cross-entropy loss of the compression type discrimination task is defined as in (12):

$$L_c = -\sum_{k=1}^{N}\sum_{i=1}^{C} \boldsymbol{c}_k^i \log \hat{r}_k^i, \tag{12}$$

where $\boldsymbol{c}_k^i$ is the ground-truth multi-class indicator vector for the compression type discrimination. If the distortion type of the k-th sample is the i-th compression type, the $\boldsymbol{c}_k^i$ will be one, otherwise it will be zero. $\hat{r}_k^i$ denotes the predicted probability of whether the compression type of the k-th sample is the i-th compression type.

To learn the superior parameters for all the three tasks, the total loss is defined as in (13):

$$L = L_q + 0.1L_d + 0.1L_c, \tag{13}$$

where the coefficients of L_q, L_d and L_c represent the importance of the three losses and are empirically set to 1, 0.1 and 0.1, respectively, in this work.

IV. Experimental Results

A. Experimental Settings

1) Datasets: **OIQA dataset** [90]: It consists of 320 distorted OIs in equireclangular format based on two types of compressions (JPEG and JPEG2000) and two types of degradations (Gaussian blur (GB), and Gaussian noise (GN)), and 16 reference images. The resolutions of the OIs vary in a range from 11332×5666 to 13320×6660. The Mean Opinion Score (MOS) value of each image is provided with the dataset by conducting the subjective experiment in which the single-stimulus (SS) method [91] is adopted.

CVIQ dataset [35]: It provides 544 images in total, which include 528 distorted omnidirectional images based on three different types of coding compressions (JPEG [80], H.264/AVC [81] and H.265/HEVC [82]). The rest 16 OIs are the reference images. All the images are in the same resolution of 4096×2048. Like [90], the SS method is also adopted in the subjective experiment to get the MOS value.

OSIQA dataset [55]: It is a dataset designed for omnidirectional stitching image quality assessment, which is used to validate the generalization ability of our model. It provides 700 omnidirectional stitching images generated by stitching the packs from different views. It includes 350 distorted OIs based on 14 scenes with different stitching distortions. MOS values are provided based on the extensive experiments in [55].

2) Evaluation Criteria: Three prevalent criteria which are Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC) and Root Mean Squared Error (RMSE) are adopted to make the monotonicity comparison and accuracy prediction. The three criteria are formulated as in (14), (15) and (16):

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

where N denotes the number of the samples. s_i is the MOS of the i-th sample, and p_i is the prediction score. $\bar{s}$ is the mean value of the MOS's, and $\bar{p}$ is the mean value of the score that the model predicted for each sample.

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{15}$$

where d_i denotes the distance between the rank of the MOS and the rank of the prediction score given by the model for the i-th sample.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - p_i)^2}, \tag{16}$$

where s_i is the MOS of the i-th sample and p_i is the prediction score given by the model.

Among the three criteria, PLCC and RMSE are calculated by the five-parameter nonlinear mapping [92], [93], which aims to unify the prediction scores given by different metrics into the same range [52].

3) Implementation Details: In implementation, the dataset is split into a training set and a testing set following the commonly used standard method in [94]–[97]. PyTorch framework [98] is adopted to implement the proposed method and the fine-tuning operation is implemented on both the OIQA and the CVIQ datasets. The SGD optimization [99]–[101] is employed with the momentum parameter set to 0.9, while the batch size and the weight decay parameter are set to 16 and 10^{-4}, respectively. For the two branches of the network, we set the initial learning rates to 10^{-3}, and the learning rate drops with a factor of 0.9 for each epoch with the total number of epochs is 300. The entire experiment processes are implemented on a device with Intel(R) Core (TM) i7-10870H CPU with 16 GB RAM, and one NVIDIA GeForce RTX 2060 graphic card, which shows the advantage of the proposed method on lower-level devices.

B. Performance Evaluation

In this section, three mainstream quantified experiments are conducted to validate our method's performance: performance on benchmark dataset, performance on different training-testing proportions and performance on the generalization ability, which can give a comprehensive conclusion about the advantage of our model. For the performance on benchamrk dataset, we evaluate the performance on each distortion type and overall performances on two OIQA datasets, which can well prove the effectiveness of our model. For the performance

TABLE I
THE PERFORMANCE COMPARISON ON THE OIQA DATASET [90]

	Metrics	JPEG			JP2K			GN			GB			Overall		
		PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓
FR	PSNR	0.758	0.731	10.245	0.781	0.768	9.379	0.958	0.931	3.654	0.529	0.506	11.268	0.492	0.497	12.528
FR	SSIM [10]	0.803	0.934	9.355	0.802	0.936	8.985	0.904	0.886	5.467	0.768	0.925	8.500	0.856	0.880	7.436
FR	S-PSNR [20]	0.87	0.829	7.738	0.816	0.849	8.686	0.919	0.885	5.033	0.699	0.692	9.501	0.716	0.712	10.030
FR	CPP-PSNR [26]	0.865	0.829	7.873	0.849	0.837	7.943	0.920	0.885	5.001	0.672	0.667	9.830	0.707	0.703	10.167
FR	WS-PSNR [44]	0.861	0.828	7.994	0.844	0.832	8.070	0.922	0.885	4.942	0.661	0.658	9.966	0.689	0.693	10.428
NR	BRISQUE [102]	0.935	0.921	8.689	0.725	0.733	11.355	0.968	0.979	4.551	0.844	0.857	9.161	0.823	0.831	9.262
NR	DESQUE [27]	0.897	0.868	6.952	0.739	0.732	10.120	0.953	0.937	3.882	0.749	0.663	8.799	0.725	0.712	9.903
NR	dipIQ [33]	0.829	0.789	8.783	0.916	0.918	6.030	0.955	0.943	3.772	0.932	0.898	4.816	0.701	0.691	10.259
NR	MEON [34]	0.823	0.779	8.935	0.680	0.601	11.017	0.952	0.930	3.895	0.764	0.716	8.572	0.749	0.717	9.536
NR	BMPRI [41]	0.918	0.909	6.210	0.185	0.166	14.768	0.961	0.949	3.534	0.356	0.354	12.248	0.431	0.338	12.984
NR	SSP-BOIQA [8]	0.877	0.834	7.620	0.853	0.852	7.501	0.905	0.843	5.451	0.854	0.862	6.834	0.860	0.865	7.313
NR	MC360IQA [35]	0.912	0.901	6.535	0.896	0.882	6.573	0.913	0.926	5.240	0.893	0.918	6.072	0.890	0.909	6.697
NR	Zhou *et al.* [52]	0.936	0.94	5.691	0.920	0.934	5.886	0.968	0.957	3.330	0.925	0.920	4.972	0.899	0.923	6.396
NR	MUSIQ [103]	0.963	0.928	5.105	0.962	0.937	5.302	0.945	0.931	3.588	0.976	0.953	3.208	0.962	0.948	4.522
NR	CVRKD-IQA [104]	0.952	0.932	5.027	0.954	0.947	4.313	0.952	0.927	3.262	0.984	0.962	3.855	0.967	0.947	4.123
NR	VGCN [22]	0.954	0.929	4.288	0.977	0.946	4.313	0.981	0.975	3.617	0.985	0.965	4.213	0.958	0.952	4.385
NR	PICS (Pro.) [21]	0.968	0.946	3.988	0.980	0.972	4.047	0.989	0.983	3.575	0.990	0.974	3.827	0.970	0.964	3.991
NR	**Ours**	**0.979**	**0.971**	**3.078**	**0.990**	**0.982**	**3.558**	**0.995**	**0.987**	**1.998**	**0.992**	**0.985**	**2.961**	**0.992**	**0.982**	**2.641**

TABLE II
THE PERFORMANCE COMPARISON ON THE CVIQ DATASET [35]

	Metrics	JPEG			H.264/AVC			H.265/HEVC			Overall		
		PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓
FR	PSNR	0.889	0.766	7.824	0.784	0.783	7.674	0.746	0.745	8.000	0.786	0.757	8.692
FR	SSIM [10]	0.852	0.929	8.946	0.941	0.940	4.177	0.918	0.917	4.763	0.897	0.885	6.230
FR	S-PSNR [20]	0.892	0.778	7.727	0.789	0.786	7.589	0.762	0.758	7.785	0.785	0.761	8.714
FR	CPP-PSNR [26]	0.884	0.765	7.996	0.779	0.777	7.751	0.751	0.748	7.936	0.779	0.754	8.822
FR	WS-PSNR [44]	0.880	0.756	8.101	0.775	0.773	7.814	0.747	0.744	7.993	0.777	0.751	8.850
NR	BRISQUE [102]	0.913	0.938	5.144	0.780	0.779	7.715	0.771	0.758	8.340	0.826	0.828	7.572
NR	DESQUE [27]	0.912	0.870	7.003	0.385	0.173	11.410	0.328	0.152	11.362	0.566	0.417	11.603
NR	dipIQ [33]	0.928	0.793	6.353	0.620	0.635	9.695	0.361	0.326	11.216	0.706	0.623	9.960
NR	MEON [34]	0.808	0.566	10.057	0.599	0.574	9.900	0.783	0.782	7.484	0.665	0.567	10.510
NR	BMPRI [41]	0.776	0.498	10.767	0.533	0.520	10.459	0.846	0.840	6.412	0.627	0.621	10.962
NR	SSP-BOIQA [8]	0.915	0.853	6.847	0.885	0.861	7.042	0.854	0.841	6.302	0.890	0.856	6.941
NR	MC360IQA [35]	0.941	0.923	5.804	0.932	0.941	5.357	0.914	0.899	4.801	0.939	0.904	4.606
NR	Zhou *et al.* [52]	0.957	0.961	5.601	0.953	0.949	3.873	0.929	0.914	4.525	0.902	0.911	6.117
NR	MUSIQ [103]	0.975	0.968	3.156	0.965	0.945	3.982	0.920	0.933	4.566	0.960	0.952	3.992
NR	CVRKD-IQA [104]	0.982	0.955	2.858	0.968	0.954	3.657	0.942	0.939	4.103	0.963	0.955	3.527
NR	VGCN [22]	0.989	0.976	2.359	0.972	0.966	3.149	0.940	0.943	4.026	0.965	0.964	3.657
NR	PICS (Pro.) [21]	0.990	0.983	2.136	0.976	0.972	2.967	0.959	0.962	3.577	0.976	0.973	3.290
NR	**Ours**	**0.992**	**0.993**	**1.953**	**0.983**	**0.974**	**2.556**	**0.995**	**0.987**	**3.053**	**0.987**	**0.991**	**2.734**

on different training-testing proportions, it is designed to indicate the effect of the training-testing proportion on our model. For the performance on the generalization ability, it is applied to validate the generalization ability and the robustness of our model using the cross-dataset validation and omnidirectional stitching image quality assessment (OSIQA) task.

1) Performance on Benchmark Dataset: To prove the advanced performance of our model, several classical FR IQA models cited in most previous NR IQA works are adopted to conduct the performance comparison, including PSNR, SSIM [10], S-PSNR [20], CPP-PSNR [26] and WS-PSNR [44]. Considering the reference image is not always available, several classical and state-of-the-art NR IQA models are also cited to make the comparison, including BRISQUE [102], DESQUE [27], dipIQ [33], MEON [34], BMPRI [41], SSP-BOIQA [8], MC360IQA [35], Zhou *et al.* [52], MUSIQ [103], CVRKD-IQA [104], VGCN [22] and PICS (Pro.) [21], which are also the most cited works in OIQA area. Among them, S-PSNR [20], CPP-PSNR [26], WS-PSNR [44], SSP-BOIQA [8], MC360IQA [35], Zhou *et al.* [52], VGCN [22] and PICS (Pro.) [21] are specifically designed for OIQA. Based on the comparison results with FR IQA models, NR IQA models and NR OIQA models, the comprehensive performance analysis of our model can be concluded. The experimental results of all the metrics on the OIQA dataset and CVIQ dataset are summarized in Table I and II, respectively. "↑" indicates that the higher the value is, the better the performance is, while "↓" indicates that the lower the value is, the better the performance is. And the top performances are emphasized with boldface. The results demonstrated that the latest deep learning-based NR OIQA models without any reference information, such as MC360IQA [35], SSP-BOIQA [8], Zhou *et al.* [52], VGCN [22], and PICS [21], achieve a better overall performance than all the FR quality metrics and some early NR OIQA models. One of the reasons could be that FR methods mainly rely on handcrafted features, and deep learning technology matures gradually, which can boost the improvement of OIQA models. Although SSIM model [10] presents a promising performance on the whole dataset and on each distortion type, it highly relies on the reference information without considering human visual characteristics on OI, which limits its application.

For the NR metrics in the comparison experiment, it can be found that BMPRI [41] with multiple pseudo reference images (MPRIs) takes the worst overall performance among all the NR metrics. One probable reason is the big intervals between two degradation degrees, which may make the network puzzled in building a clear reference standard and lead to a deficient performance. In addition, this model yields the worst results

TABLE III
PERFORMANCE OF OUR PROPOSED METHOD AND TWO ADVANCED METHODS (VGCN [22] AND PICS (PRO.) [21]) WITH DIFFERENT QUANTITIES OF TRAINING SAMPLES ON OIQA DATASET AND CVIQ DATASET

Dataset	Proportion	VGCN			PICS (Pro.)			Ours		
		PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓
OIQA	0.8/0.2	**0.958**	**0.952**	**4.385**	**0.970**	**0.964**	**3.991**	**0.992**	**0.982**	**2.641**
	0.7/0.3	0.868	0.870	7.254	0.896	0.893	6.483	0.909	0.910	5.755
	0.6/0.4	0.802	0.793	8.739	0.817	0.806	8.664	0.844	0.842	7.864
	0.5/0.5	0.725	0.721	9.918	0.753	0.737	9.315	0.792	0.779	7.839
	0.4/0.6	0.673	0.652	10.826	0.682	0.654	10.634	0.721	0.699	9.497
CVIQ	0.8/0.2	**0.965**	**0.964**	**3.657**	**0.976**	**0.973**	**3.290**	**0.992**	**0.985**	**2.961**
	0.7/0.3	0.903	0.805	5.941	0.915	0.911	5.307	0.947	0.941	4.909
	0.6/0.4	0.831	0.819	7.796	0.844	0.835	7.685	0.900	0.880	6.908
	0.5/0.5	0.755	0.730	9.381	0.795	0.782	8.241	0.821	0.816	6.918
	0.4/0.6	0.694	0.658	10.024	0.713	0.707	9.846	0.799	0.776	8.572

on JP2K and GB distortion among all the models, since it is impossible to design one pseudo generation model to cover all types of distortions, and it is hard to obtain an accurate result only relying on the pseudo quality changing information. So, a more reasonable way to build an effective OIQA model should be capturing the pseudo reference information from different directions and training the network with them, which is one of our contributions. In general, the NR models designed for OIQA present a better performance than other NR models, while MUSIQ [103] and CVRKD-IQA [104] yield a competitive results. Although MUSIQ [103] and CVRKD-IQA [104] are designed for the general IQA task rather than the OIQA task, they also achieves the good performance on the OIQA task after finetuning but still perform worse than some OIQA methods, which indicates the importance of designing specific architectures for the OIQA task. Among the NR models designed for OIQA, PICS [21] achieves a better performance, which may benefit from the rich semantic information from the generative complementary images. However, it ignores the importance of multi-scale fusion of the features. Our model not only captures the quality changing information from two opposite directions, but also interactively fuses multi-scale local and global features, which achieves the best results not only on the whole dataset, but also on each distortion type of the dataset. Overall, the results above prove that our model is reasonable, and able to be effectively applied to evaluate the quality of OIs.

To further prove the practical effectiveness of our model, Table II gives a performance comparison on each compression type and the overall CVIQ dataset. It can be observed that all the models present a better performance, and the trend of the performances in Table II present a similar trend in Table I. It needs to be mentioned that CVIQ dataset focuses on the types of compression distortion, which is more useful for image transcoding or transmission. Our model achieves the best performance on the overall CVIQ dataset and each distortion type, which further demonstrates the superiority and the potential practical application of our model.

We also visualize the prediction results of samples on the CVIQ dataset, as shown in Fig. 7. We take the promising model, PICS (Pro.) [21], as the comparison model, and ($*\%$) is the difference between MOS and the predicted scores. The smaller the value of the difference, the better the performance.

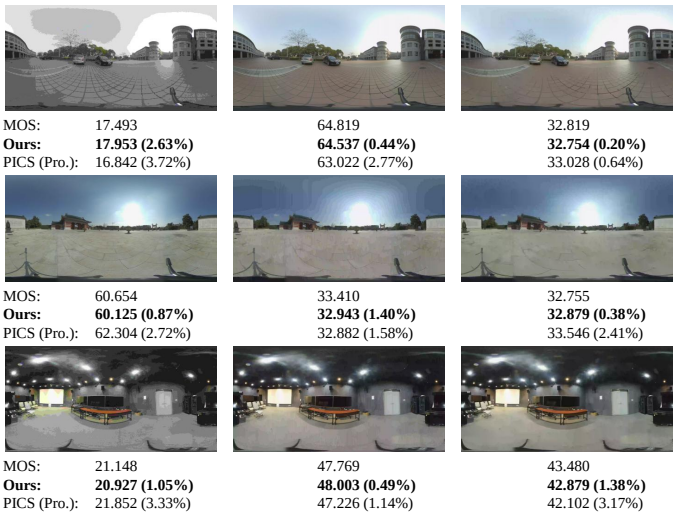

Fig. 7. The qualitative comparison on CVIQ dataset. MOS is the subjective score. Ours and PICS (Pro.) [21] present the predicted scores by our model and PICS (Pro.), respectively. The best results are presented in **boldface**.

As can be seen, the quality predicted by our model coincide better with the subjective score. This in turn confirms that the proposed model can better capture representative features for quality assessment.

2) Performance on Different Training-Testing Proportions: To study the effect of the training-testing proportion, experiments on how the performance varies with the different dataset proportions are conducted. The procedure is repeated five times, and the average results are presented in Table III, and the best performances are emphasized with boldface. It can be seen that as the number of training samples increases, the performances of all three metrics are improved. It needs to be mentioned that the performance of our proposed method is superior to VGCN [22] and PICS [21] methods on all training-testing splits. Moreover, our method still achieves a remarkable performance only training with half of the images. Specifically, both the PLCC and SRCC are over 0.8, which performs even better than some of the existing metrics with the best proportion in Table I and II. Furthermore, it can be concluded that our method is relatively dependent on the quantity of the images for model training, which means an

TABLE IV
THE RESULTS OF THE CROSS-DATASET VALIDATION

Train	Test	Criterion	BRISQUE	dipIQ	MEON	BMPRI	MC360IQA	Zhou *et al.*	MUSIQ	CVRKD-IQA	VGCN	PICS (Pro.)	**Ours**
CVIQ	OIQA	PLCC ↑	0.682	0.583	0.604	0.331	0.705	0.735	0.762	0.803	0.787	0.827	**0.905**
		SRCC ↑	0.524	0.502	0.551	0.192	0.684	0.684	0.792	0.801	0.778	0.815	**0.903**
		RMSE ↓	10.870	11.747	11.399	13.576	10.178	10.178	6.447	6.325	5.437	5.124	**4.440**
OIQA	CVIQ	PLCC ↑	0.754	0.630	0.688	0.586	0.823	0.823	0.898	0.933	0.924	0.935	**0.970**
		SRCC ↑	0.689	0.587	0.624	0.548	0.814	0.814	0.901	0.902	0.928	0.931	**0.972**
		RMSE ↓	9.381	10.904	10.145	11.403	7.811	7.811	5.657	4.892	5.462	4.887	**3.857**

TABLE V
PERFORMANCE COMPARISON OF THE STATE-OF-THE-ART NR-IQA MODELS ON OSIQA DATASET

Metrics	PLCC ↑	SRCC ↑	RMSE ↓
BRISQUE [102]	0.3072	0.2450	12.296
NIQE [105]	0.3167	0.2288	12.053
CORNIA [106]	0.3404	0.2271	12.008
QAC [107]	0.5747	0.2635	9.9765
ILNIQE [108]	0.3957	0.1658	11.707
LPSI [109]	0.5789	0.2127	10.599
HOSA [110]	0.3270	0.2457	11.859
dipIQ [33]	0.2394	0.1994	499.01
BPRI [111]	0.5993	0.2656	9.9980
BPRI-LSS [111]	0.4889	0.3200	10.994
BPRI-PSS [111]	0.5085	0.2356	10.957
BPRIc [111]	0.5685	0.3171	10.270
BMPRI [41]	0.3703	0.2666	11.320
MC360IQA [35]	0.7943	0.6807	6.9597
OSIQA-NR [55]	0.8214	0.7236	**6.2442**
Ours	**0.8670**	**0.7541**	6.8241

application potential.

3) Performance on Generalization Ability: To validate the generalization ability and the robustness of our method, the cross-dataset validation is conducted. We first train the model on one of the two datasets and then test on the other. The results of the cross-dataset validation are presented in Table IV, and the top performances are emphasized with boldface. It can be observed that all the results are lower than the results of training and testing on the same datasets. The PICS (Pro.) [21] metric presents a potential competition but is inferior to our method. Our method achieves the highest PLCC and SRCC and the lowest RMSE and is the only method performing with both the PLCC and SRCC over 0.9 when the model is trained on CVIQ dataset and tested on OIQA dataset. In addition, it can be observed that the results of models trained on OIQA dataset are all better than the models trained on CVIQ dataset. One probable reason is that the OIQA dataset includes four types of compressions and degradations, while the CVIQ dataset includes images with only the compression types of distortions.

To further prove the generalization ability of our method, the omnidirectional stitching image quality assessment (OSIQA) [55] task is conducted. Although OSIQA task is focused on stitching OIs, it is still an OIQA task, which can be applied to prove the generalization of our model. It needs to be mentioned that since the OSIQA dataset does not include the reference image, the FR results cannot be obtained. Herein, in this section, we only provide NR methods for the experimental comparison. The results are shown in Table V, and the top performances are emphasized with boldface. It can be seen that, even compared to the state-of-the-art in the OSIQA area, our method achieves a remarkable performance, which proves the potential performance in the OSIQA area. Our model achieves the best performance on PLCC and SRCC criteria and ranks second on RMSE. The probable reason is that the geometry distortion in OSIQA tasks is not considered, but the RMSE value of our model is close to the top one, which proves the generalization ability of our method.

Overall, the results above indicate the generalization ability of our proposed method.

TABLE VI
THE RESULTS OF THE ABLATION EXPERIMENTS

Module	Status	OIQA PLCC ↑	OIQA SRCC ↑	OIQA RMSE ↓	CVIQ PLCC ↑	CVIQ SRCC ↑	CVIQ RMSE ↓
Local features	✗	0.915	0.921	12.669	0.934	0.923	9.600
Global features	✗	0.944	0.948	4.722	0.951	0.954	4.591
BS-MSFA	✗	0.969	0.975	3.227	0.976	0.972	3.141
Auxiliary tasks	✗	0.982	0.968	3.193	0.983	0.974	3.175
Global backbone	ViT	0.981	0.979	4.499	0.975	0.974	3.133
All	✓	**0.992**	**0.982**	**2.641**	**0.987**	**0.991**	**2.734**

C. Ablation Experiments

To verify the contribution of the local features, the global features, the BS-MSFA module, the auxiliary tasks and the backbone of VMamba, respectively, the ablation experiments are conducted, and the results are presented in Table VI. The top performances are emphasized with boldface. It can be concluded that each component has contribution to the final performance, while the model without the local branch yields the worst performance, which proves that the local branch is the most essential information for OIQA. The model without the global features ranks second to the last, which proves that it is reasonable to combine the local and global features. The model without BS-MSFA module designed by using the traditional concatenating fusion way presents a worse performance than the proposed model, which presents the significant role of the bi-stream multi-scale feature aggregation. And the model without the assistance of multi-task module present a lower results than our model, which proves the positive effect of two auxiliary tasks. We can see that the model using the ViT as the backbone to capture the global features presents a worse performance than our model, which means the effectiveness of VMamba in OIQA tasks. Compared to ViT, VMamba has an efficient zigzags method to scan an image and captures rich

information [70], and makes VMamba a better choice as the backbone, which is validated by the results of our ablation experiments. All the above results prove all our contributions and the reasonableness of our model. Overall, our proposed model achieves a high consistency with the human perception and can be effectively applied to evaluate the quality of OIs.

V. Conclusion

In this paper, a multi-task framework based on multi-scale local and global features is proposed for OIQA. Considering the quality degradation varying in distorted images, a bidirectional pseudo reference module is utilized to capture the rich local features from two opposite directions. Based on the state-of-art performance of VMamba in features extraction, we adopt it as the global feature extractor to obtain multi-scale global information, which can complement the above local features well. To utilize the interactive relationship between the local and global information, the multi-scale feature aggregation module is constructed to make a hierarchically deep fusion. Furthermore, a multi-task learning is applied to optimize the entire model for the quality prediction. The experimental results demonstrate that our method can effectively and accurately predict the quality of an omnidirectional image. In the future, we will extend our method to more omnidirectional quality assessment tasks, and will develop more effective models with advanced technology to evaluate the quality of omnidirectional images by digging into the relationship between human visual characteristics and high-level semantic information.

References

[1] X. Min, H. Duan, W. Sun, Y. Zhu, and G. Zhai, "Perceptual video quality assessment: a survey," *Science China Information Sciences*, vol. 67, no. 11, pp. 211301, 2024.

[2] H. Duan, X. Zhu, Y. Zhu, X. Min, and G. Zhai, "A quick review of human perception in immersive media," *IEEE Open Journal on Immersive Displays*, vol. 1, pp. 41–50, 2024.

[3] C. Li, Z. Zhang, H. Wu, K. Zhang, L. Bai, X. Liu, G. Zhai, and W. Lin, "Paps-ovqa: Projection-aware patch sampling for omnidirectional video quality assessment," in *Proceedings of the 2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2024, pp. 1–5.

[4] X. Zhu, H. Duan, Y. Cao, Y. Zhu, Y. Zhu, J. Liu, L. Chen, X. Min, and G. Zhai, "Perceptual quality assessment of omnidirectional audio-visual signals," in *Proceedings of the Artificial Intelligence: Third CAAI International Conference, CICAI 2023, Fuzhou, China, July 22–23, 2023, Revised Selected Papers, Part II*, 2024, pp. 512–525.

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

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

[7] 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, 2019.

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

[9] G. Yue, C. Hou, T. Zhou, and X. Zhang, "Effective and efficient blind quality evaluator for contrast distorted images," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 8, pp. 2733–2741, 2019.

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

[11] 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.

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

[13] X. Sui, K. Ma, Y. Yao, and Y. Fang, "Perceptual quality assessment of omnidirectional images as moving camera videos," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 3022–3034, 2022.

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

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

[16] Q. Wu, H. Li, K. N. Ngan, and K. Ma, "Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2078–2089, 2018.

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

[18] Q. Jiang, W. Gao, S. Wang, G. Yue, F. Shao, Y.-S. Ho, and S. Kwong, "Blind image quality measurement by exploiting high-order statistics with deep dictionary encoding network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7398–7410, 2020.

[19] L. Li, Y. Zhou, J. Wu, F. Li, and G. Shi, "Quality index for view synthesis by measuring instance degradation and global appearance," *IEEE Transactions on Multimedia*, vol. 23, pp. 320–332, 2021.

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

[21] Y. Zhou, Y. Ding, Y. Sun, L. Li, J. Wu, and X. Gao, "Perceptual information completion-based siamese omnidirectional image quality assessment network," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–10, 2024.

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

[23] Y. Liu, X. Yin, Y. Wang, Z. Yin, and Z. Zheng, "Hvs-based perception-driven no-reference omnidirectional image quality assessment," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[24] J. Xu, W. Zhou, Z. Chen, S. Ling, and P. L. Callet, "Binocular rivalry oriented predictive auto-encoding network for blind stereoscopic image quality measurement," *arXiv preprint arXiv:1909.01738*, 2020. [Online]. Available: https://arxiv.org/abs/1909.01738

[25] L. Yang, M. Xu, D. Xin, and B. Feng, "Spatial attention-based non-reference perceptual quality prediction network for omnidirectional images," *arXiv preprint arXiv:2103.06116*, 2021. [Online]. Available: https://arxiv.org/abs/2103.06116

[26] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," *Optics and Photonics for Information Processing X*, vol. 9970, pp. 99700C, 2016.

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

[28] 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.

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

[30] A. K. R. Poreddy, R. B. C. Ganeswaram, B. Appina, P. Kokil, and R. B. Pachori, "No-reference virtual reality image quality evaluator using global and local natural scene statistics," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–16, 2023.

[31] Y. Liu, X. Yin, Z. Wan, G. Yue, and Z. Zheng, "Toward a no-reference omnidirectional image quality evaluation by using multi-perceptual features," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 19, no. 2, 2023.

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

[33] 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*, vol. 26, no. 8, pp. 3951–3964, 2017.

[34] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.

[35] W. Sun, W. Luo, X. Min, G. Zhai, X. Yang, K. Gu, and S. Ma, "Mc360iqa: The multi-channel cnn for blind 360-degree image quality assessment," in *Proceedings of 2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2019, pp. 1–5.

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

[37] C. Zhang and S. Liu, "No-reference omnidirectional image quality assessment based on joint network," in *Proceedings of the ACM International Conference on Multimedia, ser. MM '22. New York, NY, USA: Association for Computing Machinery*, 2022, pp. 943–951.

[38] M. Zhou, L. Chen, X. Wei, X. Liao, Q. Mao, H. Wang, H. Pu, J. Luo, T. Xiang, and B. Fang, "Perception-oriented u-shaped transformer network for 360-degree no-reference image quality assessment," *IEEE Transactions on Broadcasting*, vol. 69, no. 2, pp. 396–405, 2023.

[39] N. J. Tofighi, M. H. Elfkir, N. Imamoglu, C. Ozcinar, A. Erdem, and E. Erdem, "Omnidirectional image quality assessment with local–global vision transformers," *Image and Vision Computing*, vol. 148, no. C, 2024.

[40] Y. Fan and C. Chen, "Omiqnet: Multiscale feature aggregation convolutional neural network for omnidirectional image assessment," *Applied Intelligence*, vol. 54, no. 7, pp. 5711–5727, 2024.

[41] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.

[42] J. Hu, X. Wang, F. Shao, and Q. Jiang, "Tspr: Deep network-based blind image quality assessment using two-side pseudo reference images," *Digital Signal Processing*, vol. 106, pp. 102849, 2020.

[43] R. Xu, S. Yang, Y. Wang, Y. Cai, B. Du, and H. Chen, "Visual mamba: A survey and new outlooks," *arXiv preprint arXiv:2404.18861*, 2024. [Online]. Available: https://arxiv.org/abs/2404.18861

[44] 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.

[45] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, "Spherical structural similarity index for objective omnidirectional video quality assessment," in *Proceedings of 2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[46] L. Liu, P. Ma, C. Wang, and D. Xu, "Omnidirectional image quality assessment with knowledge distillation," *IEEE Signal Processing Letters*, vol. 30, pp. 1562–1566, 2023.

[47] X. Liu, J. Yan, L. Huang, Y. Fang, Z. Wan, and Y. Liu, "Perceptual quality assessment of omnidirectional images: A benchmark and computational model," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 20, no. 6, 2024.

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

[49] Y. Liu, H. Yu, B. Huang, G. Yue, and B. Song, "Blind omnidirectional image quality assessment based on structure and natural features," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[50] W. Zhou, J. Xu, Q. Jiang, and Z. Chen, "No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1778–1791, 2022.

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

[52] Y. Zhou, Y. Sun, L. Li, K. Gu, and Y. Fang, "Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1767–1777, 2022.

[53] C. Tian, X. Chai, G. Chen, F. Shao, Q. Jiang, X. Meng, L. Xu, and Y.-S. Ho, "Vsoiqe: A novel viewport-based stitched 360° omnidirectional image quality evaluator," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6557–6572, 2022.

[54] C. Tian, F. Shao, X. Chai, Q. Jiang, L. Xu, and Y.-S. Ho, "Viewport-sphere-branch network for blind quality assessment of stitched 360° omnidirectional images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2546–2560, 2023.

[55] H. Duan, X. Min, W. Sun, Y. Zhu, X.-P. Zhang, and G. Zhai, "Attentive deep image quality assessment for omnidirectional stitching," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 6, pp. 1150–1164, 2023.

[56] Y. Zhou, W. Gong, Y. Sun, L. Li, J. Wu, and X. Gao, "Pyramid feature aggregation for hierarchical quality prediction of stitched panoramic images," *IEEE Transactions on Multimedia*, vol. 25, pp. 4177–4186, 2023.

[57] Y. Zhou, W. Gong, Y. Sun, L. Li, K. Gu, and J. Wu, "Quality assessment for stitched panoramic images via patch registration and bidimensional feature aggregation," *IEEE Transactions on Multimedia*, vol. 26, pp. 3354–3365, 2024.

[58] H. Hu, F. Shao, H. Wang, B. Mu, H. Chen, Q. Jiang, and W. Chen, "Spatio-temporal feature integration for quality assessment of stitched omnidirectional images," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 2, pp. 1484–1499, 2024.

[59] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2024. [Online]. Available: https://arxiv.org/abs/2312.00752

[60] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2022. [Online]. Available: https://arxiv.org/abs/2111.00396

[61] A. Gu, A. Gupta, K. Goel, and C. Ré, "On the parameterization and initialization of diagonal state space models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc.*, 2022.

[62] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," in *Proceedings of the 36th International Conference on Neural Information Processing Systems, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc.*, 2022.

[63] A. Orvieto, S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu, and S. De, "Resurrecting recurrent neural networks for long sequences," in *Proceedings of the 40th International Conference on Machine Learning, ser. ICML'23. JMLR.org*, 2023.

[64] J. T. H. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," *arXiv preprint arXiv:2208.04933*, 2023. [Online]. Available: https://arxiv.org/abs/2208.04933

[65] R. Hasani, M. Lechner, T.-H. Wang, M. Chahine, A. Amini, and D. Rus, "Liquid structural state-space models," *arXiv preprint arXiv:2209.12951*, 2022. [Online]. Available: https://arxiv.org/abs/2209.12951

[66] A. Gu, I. Johnson, A. Timalsina, A. Rudra, and C. Ré, "How to train your hippo: State space models with generalized orthogonal basis projections," *arXiv preprint arXiv:2206.12037*, 2022. [Online]. Available: https://arxiv.org/abs/2206.12037

[67] J. Wang, W. Zhu, P. Wang, X. Yu, L. Liu, M. Omar, and R. Hamid, "Selective structured state-spaces for long-form video understanding," in *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6387–6397.

[68] H. Mehta, A. Gupta, A. Cutkosky, and B. Neyshabur, "Long range language modeling via gated state spaces," *arXiv preprint arXiv:2206.13947*, 2022. [Online]. Available: https://arxiv.org/abs/2206.13947

[69] A. Chubarau and J. Clark, "VTAMIQ: Transformers for attention modulated image quality assessment," *arXiv preprint arXiv:2110.01655*, 2021. [Online]. Available: https://arxiv.org/abs/2110.01655

[70] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "VMamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024. [Online]. Available: https://arxiv.org/abs/2401.10166

[71] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11999–12009.

[72] C.-S. Chen, G.-Y. Chen, D. Zhou, D. Jiang, and D.-S. Chen, "Res-VMamba: Fine-grained food category visual classification using selective state space models with deep residual learning," *arXiv preprint arXiv:2402.15761*, 2024. [Online]. Available: https://arxiv.org/abs/2402.15761

[73] C. Du, Y. Li, and C. Xu, "Understanding robustness of visual state space models for image classification," *arXiv preprint arXiv:2403.10935*, 2024. [Online]. Available: https://arxiv.org/abs/2403.10935

[74] Y. Shi, B. Xia, X. Jin, X. Wang, T. Zhao, X. Xia, X. Xiao, and W. Yang, "VmambaIR: Visual state space model for image restoration," *arXiv preprint arXiv:2403.11423*, 2024. [Online]. Available: https://arxiv.org/abs/2403.11423

[75] Z. Wang, J.-Q. Zheng, C. Ma, and T. Guo, "VMambaMorph: a multi-modality deformable image registration framework based on visual state space model with cross-scan module," *arXiv preprint arXiv:2404.05105*, 2024. [Online]. Available: https://arxiv.org/abs/2404.05105

[76] H.-Y. Ma, L. Zhang, and S. Shi, "VMambaCC: A visual state space model for crowd counting," *arXiv preprint arXiv:2405.03978*, 2024. [Online]. Available: https://arxiv.org/abs/2405.03978

[77] X. Xie, Y. Cui, C.-I. Ieong, T. Tan, X. Zhang, X. Zheng, and Z. Yu, "FusionMamba: Dynamic feature enhancement for multimodal image fusion with mamba," *arXiv preprint arXiv:2404.09498*, 2024. [Online]. Available: https://arxiv.org/abs/2404.09498

[78] Y. Yang, C. Ma, J. Yao, Z. Zhong, Y. Zhang, and Y. Wang, "ReMamber: Referring image segmentation with mamba twister," *arXiv preprint arXiv:2403.17839*, 2024. [Online]. Available: https://arxiv.org/abs/2403.17839

[79] M. V. Conde, G. Geigle, and R. Timofte, "InstructIR: High-quality image restoration following human instructions," *arXiv preprint arXiv:2401.16468*, 2024. [Online]. Available: https://arxiv.org/abs/2401.16468

[80] G. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[81] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[82] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[83] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18392–18402.

[84] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4771–4780.

[85] T. T. Huong, D. T. Ha, H. T. T. Tran, N. D. Viet, B. D. Tien, N. H. Thanh, T. C. Thang, and P. N. Nam, "An effective foveated 360° image assessment based on graph convolution network," *IEEE Access*, vol. 10, pp. 98165–98178, 2022.

[86] R. Zhang and A. C. S. Chung, "A fine-grain error map prediction and segmentation quality assessment framework for whole-heart segmentation," *arXiv preprint arXiv:1907.12244*, 2019. [Online]. Available: https://arxiv.org/abs/1907.12244

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

[88] X. Pei, T. Huang, and C. Xu, "EfficientVMamba: Atrous selective scan for light weight visual mamba," *arXiv preprint arXiv:2403.09977*, 2024. [Online]. Available: https://arxiv.org/abs/2403.09977

[89] Y. Liu, J. Wu, L. Li, W. Dong, and G. Shi, "Quality assessment of ugc videos based on decomposition and recomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1043–1054, 2023.

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

[91] G. M. Pace, M. T. Ivancic, G. L. Edwards, B. A. Iwata, and T. J. Page, "Assessment of stimulus preference and reinforcer value with profoundly retarded individuals," *Journal Of Applied Behavior Analysis*, vol. 18, no. 3, pp. 249–255, 1985.

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

[93] 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.

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

[95] L. Li, Y. Zhou, K. Gu, Y. Yang, and Y. Fang, "Blind realistic blur assessment based on discrepancy learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3859–3869, 2020.

[96] 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 *Proceedings of the 26th ACM International Conference on Multimedia, ser. MM '18. New York, NY, USA: Association for Computing Machinery*, 2018, pp. 932–940.

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

[98] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: an imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[99] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[100] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, 1952.

[101] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2017. [Online]. Available: https://arxiv.org/abs/1609.04747

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

[103] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.

[104] G. Yin, W. Wang, Z. Yuan, C. Han, W. Ji, S. Sun, and C. Wang, "Content-variant reference image quality assessment via knowledge distillation," *arXiv preprint arXiv:2202.13123*, 2022. [Online]. Available: https://arxiv.org/abs/2202.13123

[105] 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, 2013.

[106] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[107] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 995–1002.

[108] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[109] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proceedings of 2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 339–343.

[110] 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*, vol. 25, no. 9, pp. 4444–4457, 2016.

[111] 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*, vol. 20, no. 8, pp. 2049–2062, 2018.

Yun Liu (Member, IEEE) received the Ph.D. degree in communication and information engineering from Tianjin University, China, in 2016. From 2014 to 2015, she was a visiting Ph.D. student at the Visual Space Perception Laboratory, University of California, Berkeley, United States. She is currently an associate professor at the Faculty of Information, Liaoning University, Shenyang, China. Her research interests include multimedia quality assessment, image processing, computer vision, and pattern recognition.

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

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

Yu Zhou (Member, IEEE) received the B.S. and Ph.D. degrees from the China University of Mining and Technology, Xuzhou, China, in 2014 and 2019, respectively. She is currently an Associate Professor with the School of Information and Control Engineering, China University of Mining and Technology. Her research interests include computer vision, multimedia image processing, and artificial intelligence.

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

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