Quality Assessment for Stitched Panoramic Images via Patch Registration and Bidimensional Feature Aggregation

Yu Zhou, Weikang Gong, Yanjing Sun, *Member, IEEE*, Leida Li, *Member, IEEE*, Ke Gu, *Senior Member, IEEE*, and Jinjian Wu, *Member, IEEE*

***Abstract*—Quality assessment for stitched panoramic images (SPIQA) is of great significance for the stitching algorithm optimization. By contrast, this task is much more challenging and arduous than traditional IQA task due to the high resolution of stitched panoramic images and the particularity and complexity of stitching distortions. For this task, we propose an effective method based on patch registration and bidimensional feature aggregation (PRBFA). First, inspired by the attention mechanism of the human visual system and the limited range of human vision, a soft patch segmentation and selection method is presented to determine the key patches in panoramic images to participate in the following patch matching and feature alignment stages, achieving patch registration between the panoramic image and the corresponding constituent images. Further, to fully simulate the human visual perception process from local viewport to panorama, the feature exploration is successively performed from local to global, which is also adaptive to the complexity of the distortions in stitched panoramic images. For performance testification, extensive experiments are conducted on the publicly released SPIQA database, the results of which prove the performance superiority of the PRBFA method.**

***Index Terms*—Image quality assessment, stitched panoramic images, patch registration, bidimensional feature aggregation.**

I. Introduction

With the rapid development of virtual reality (VR) and augmented reality (AR), the immersive and interactive experience has been gaining popularity [1], [2]. Visual experience is the most basic and pivotal one of consumers' various sensory experiences, and panoramic images are the carrier of information in the 360° visual experience, so the quality of which is undoubtedly crucial. However, various distortions introduced by the series of processes from image acquisition to final presentation often cause quality degradation of panoramic images. Therefore, the panoramic image quality assessment (PIQA) is of great necessity and significance, which plays vital role in the design and optimization of all technologies involved in the entire process.

In recent years, more and more works have been proposed for the PIQA task. Similar to conventional IQA works [3], [4], [5], [6], [7], these works can be divided into the general-purpose ones and the targeted ones. For the general-purpose works [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], the knowledge of the distortion type does not need to be known in advance. This type of PIQA works has been verified to achieve impressive performance on several publicly released PIQA databases that contain common distortions, such as compression distortions, gaussian blur and gaussian noise, etc. The targeted works are those specifically designed for the panoramic images with a certain type of distortions. As a panoramic image is commonly generated by using the stitching algorithm to stitch a set of constituent images captured from different angles and existing stitching algorithms are usually imperfect, stitching distortions are no doubt introduced into the final panoramic images, resulting in the decline of visual quality. Due to the great difference between the stitching distortions and the common distortions in distribution and appearance, the existing general-purpose works that have been demonstrated excellent performance in quality evaluation of panoramic images with conventional distortions have obvious performance degradation when they deal with the stitching distortions in stitched panoramic images. Based on this, some scholars have been working on the targeted works for the stitching distortions [18], [19], [20], [21], [22], [23], [24], [25], [26]. This type of works is dubbed as the quality assessment methods for stitched panoramic images (SPIQA) to distinguish from the general-purpose PIQA task. Through deep analysis, we

Manuscript received 9 May 2023; revised 27 July 2023; accepted 21 August 2023. Date of publication 30 August 2023; date of current version 14 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62001475, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20200649, in part by the National Natural Science Foundation of China under Grants 62171340, 62322302, 62273011, and 62076013, and in part by Beijing Natural Science Foundation under Grant JQ21014. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Qianqian Xu. *(Corresponding author: Yanjing Sun.)*

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

Weikang Gong and Yanjing Sun are with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (e-mail: gongweikang1998@163.com; yjsun@cumt.edu.cn).

Leida Li and Jinjian Wu are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: ldli@xidian.edu.cn; jinjian.wu@mail.xidian.edu.cn).

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

Digital Object Identifier 10.1109/TMM.2023.3310276

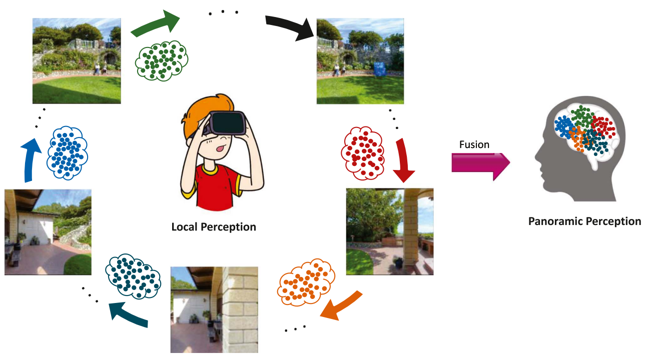


Fig. 1. Presentation of the process of human determining the quality of a panoramic image.

summarize the disadvantages of existing SPIQA metrics that hinder their performance improvement. First, most of them are based on hand-crafted features, the limited feature representation ability of which seriously restricts the performance development. Besides, for the only two deep learning based metrics [25], [26], they both evaluate the stitched panoramic images from only one aspect of the local or global feature representation, which is inconsistent with the consumers' visual perception process from local to global when they view a panoramic image. To be more specific, due to the limited scope of human vision, consumers equipped with the head mounted display (HMD) cannot perceive the entire 360° panoramic image at once. Actually, they only obtain the content in the viewport range at a moment. Then, they rotate heads to perceive the local information in different viewport images and give the overall quality evaluation by integrating all the local information they have perceived. This process is illustrated in Fig. 1.

To adapt to the above facts, we propose an effective SPIQA metric based on Patch Registration and Bidimensional Feature Aggregation (PRBFA). This is a patch-level metric to cater to the characteristic of limited human vision range. First, inspired by the process of human determining image quality by measuring the distance between the tested content and the reference prior [26], [28], and with accessibility of the raw constituent images that generate the stitched panoramic image, the patch registration module is presented to seek for the reference version of the distorted patches. Specifically, motivated by the attention mechanism of the human visual system (HVS), a soft patch segmentation and selection method is presented to determine the key patches in the stitched panoramic image that have great impact on quality perception. Then, the image patch matching each key patch is determined in the constituent images, and the feature alignment operation is designed to ensure consistency between two matching patches. Further, enlightened by the perception process from local to global, we propose to explore local and global features successively. This scheme of bidimensional feature exploration is also adaptive to the distortion complicacy of the SPIs. Extensive comparisons on the SPIQA dataset demonstrate the advantages of the PRBFA method to the state-of-the-art quality metrics.

Major contributions of the PRBFA method include the following aspects:

First, a patch registration method is proposed to generate the reference version of the distorted patches from the constituent images, which is beneficial for more accurate quality measurement.

Second, the feature extraction is implemented successively from local to global, which is demonstrated to well simulate the process of human perceiving panoramic images from local viewport contents to panorama.

Finally, the proposed method shows obvious superiority than the state-of-the-art quality metrics in terms of both the average prediction accuracy and the prediction stability.

II. Literature Review

In this part, we give a review of existing PIQA works, including the general-purpose PIQA works and the targeted SPIQA works.

A. *General-Purpose PIQA Works*

Yu et al. [8] proposed to assess the quality of panoramic contents in the spherical domain. In the work [9], the authors presented a metric based on the craster parabolic projection space, where the peak signal to noise ratio of the resampled pixels was computed, producing the CPP-PSNR metric. In [10], a weighted-to-spherically-uniform PSNR method (WS-PSNR) was proposed. In [11], Chen et al. proposed an omnidirectional video QA metric by integrating the similarities of three aspects, including luminance, contrast and structure. In [12], the original panoramic images with the equirectangular projection (ERP) format were first projected to the segmented spherical domain, generating the bipolar and equatorial regions. Then, several types of features were obtained to train the quality prediction model. Sun et al. [13] proposed to first segment the panoramic image into six viewport images and then extract features from each viewport image for quality prediction using the multi-channel convolutional neural network. In [14] and [15], the authors both proposed to first predict the scores and weights of all patches and then integrated them to produce the quality score. In [16], three metrics were proposed. For one metric, six groups of features were extracted on six cubemaps and fused for quality prediction. For the other two metrics, the attention weights and the attention distortion features were severally represented and fused with the features obtained in the first metric. Xu et al. [17] proposed to extract features from both the whole ERP format panoramic image and several selected viewport images, and fuse all features for quality score prediction.

These general-purpose PIQA metrics have achieved very impressive performance for common distortions, especially for the coding distortions. However, the huge property difference of common distortions from the distortions in the stitched panoramic images results in the predictable performance fading of these metrics in the SPIQA task.

B. *SPIQA Works*

Most SPIQA methods were designed based on the representation and integration of some hand-crafted features. In [18], both

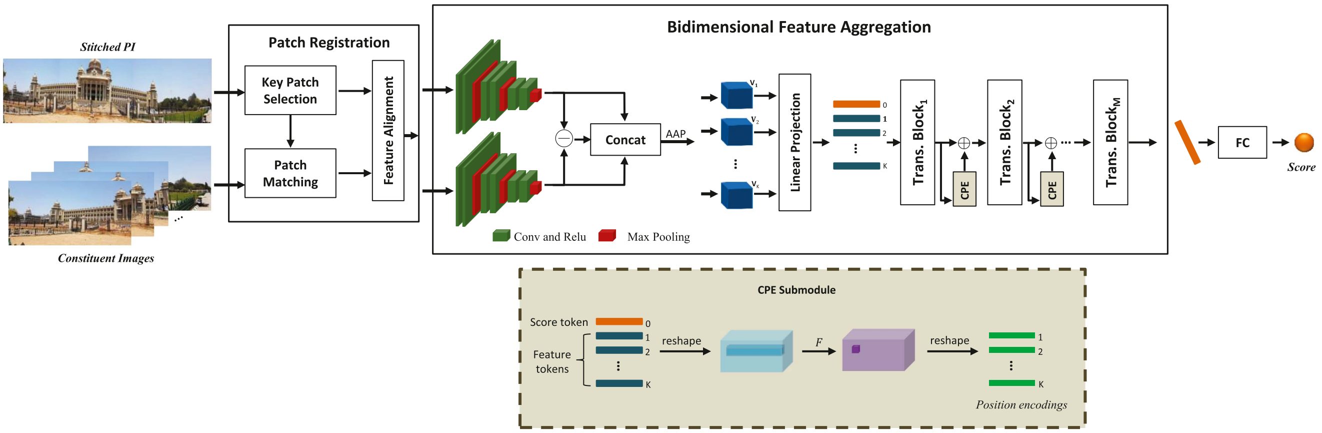


Fig. 2. Architecture of the proposed PRBFA method. It mainly includes the patch registration module and the bidimensional feature aggregation module. For the patch registration module, three submodules are included, i.e., key patch selection, patch matching and feature alignment. For the feature aggregation module, the details of its CPE submodule is presented in the dashed line area with the yellow background at the bottom of this figure.

geometric and structural errors were measured and integrated into one quality score of the stitched panoramic image. Ling et al. [19] proposed a metric by measuring two types of distortions, namely the ghosting and structural distortions, where the extracted features were used as the input of the support vector regressor to learning the quality model. In [20], an SPI database named the Indian Institute of Science Stitched IQA (ISIQA) was first built and then an SPIQA method was presented, where the feature distances between the SPI and the constituent images were calculated for model training. In [21], a stitched IQA metric was proposed. The feature descriptors were extracted by computing the local measurement errors and the global statistical properties. Wang et al. [22] presented an SPIQA method based on bi-directional matching, extraction of multiple types of features and feature fusion through SVR. In [23], an entropy based metric was proposed for the SPIQA. In [24], a quality metric was proposed by aggregating both local and global features via SVR.

Even though the above metrics are specifically targeted at the SPIQA task, the limited representation ability of hand-crafted features hinders their performance improvement. With the development of deep learning, Hou et al. [25] presented a multi-task learning network for the SPIQA task, where the quality ranking was designed as the auxiliary task of the quality prediction task. In [26], Zhou et al. also proposed a deep learning based SPIQA metric by first predicting the pure version of the distorted panoramic image using the transfer learning idea and then aggregating the pyramid features extracted from multiple convolutional layers for quality prediction. In [27], an SPIQA database with 300 SPIs was provided and an attentive multi-channel SPIQA method was proposed by improving the hyper-ResNet with spatial attention.

III. Proposed SPIQA Metric

With consideration of the characteristic that human vision is limited, namely human cannot capture the whole view of the 360-degree panorama at a moment, we propose a patch-level method for the SPIQA task instead of directly assessing the whole panoramic image. First, to alleviate the inconsistency with the HVS characteristics of the commonly used rigid patch partition method and accommodate to the visual attention mechanism, we propose a soft key patch selection module to determine the patches that have non-negligible impact on quality of panoramic images. Then, the patches matching these key patches are subsequently detected from the constituent images to form the matching patch pairs. Further, a feature alignment stage is conducted for more accurate measure of feature distance. Successively, quality-aware features are explored from local to global to more fully simulate the process of human perceiving the panoramic image. In summary, the proposed PRBFA method includes two modules, i.e. the patch registration module and the bidimensional feature aggregation module, which can be seen in Fig. 2. Among them, the patch registration module consists of three submodules, including the key patch selection, patch matching and feature alignment. The following bidimensional feature aggregation module includes the local feature representation stage and the global feature representation stage.

A. Patch Registration

For an input distorted panoramic image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the keypoints are first detected using the Scale-Invariant Feature Transform (SIFT) [29], which has been widely used for robust local feature representation in many image processing tasks, such as object detection [30] and fingerprint indexing [31], etc. Here, the descriptors of each keypoint are together generated. To avoid the inaccurate detection caused by noise interference, we further identify and screen out the outliers among the initially detected keypoints using the local outlier factor (LOF) [32], which works by assigning a degree of being an outlier to each keypoint. Namely, this factor describes the isolation degree between one detected keypoint and the surrounding neighborhood keypoints. For each initially detected keypoint, if its LOF value is larger than 1, the corresponding point is determined as the

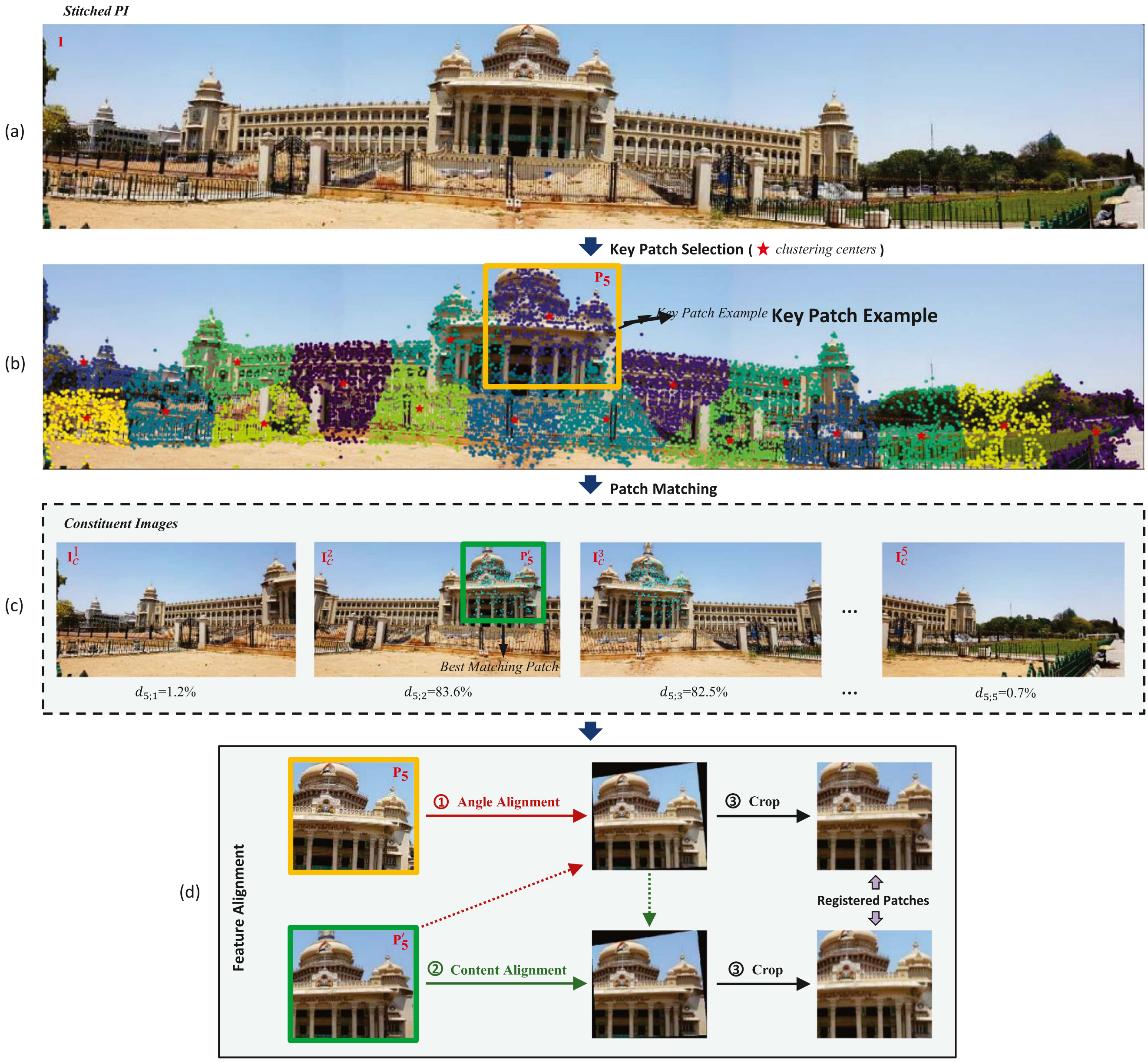


Fig. 3. Illustration of the patch registration process.

outlier and will be removed. Otherwise, it is preserved as the final key point. After removing all outliers, each remaining keypoint is further distributed to one of multiple clusters through multiple updates of the clustering centers using the K-means++ method [33]. In this work, the number of clustering centers is denoted by $K = \lfloor \frac{H}{N} \rfloor \cdot \lfloor \frac{W}{N} \rfloor$, where N denotes the patch size. So far, K clustering centers are detected. Then, the patches with size of $N \times N$ centered on each clustering center are determined as the key patches for the following feature representation stage, which are denoted by $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_K\}$. Fig. 3(b) gives an example to show the key patch selection process. The colorful scatter points in image (b) are the determined key points and those red "$\star$" represent the final clustering centers. It can be seen from image (b) that a total of 18 "$\star$" are detected, so 18 key patches can be generated. For clear presentation, only the 5th patch $\mathbf{P}_5$ is presented as the key patch example, which is marked with the yellow border in image (b).

With the generated key patches $\mathbf{P}$, the patches matching these key patches are subsequently detected from the constituent images based on pixel-level matching. Taking the i-th key patch $\mathbf{P}_i$ as an example, the keypoints and the corresponding feature descriptors in it have been obtained, so we further detect the keypoints and obtain the feature descriptors on each constituent image $\mathbf{I}_c^j (j \in [1, z])$ via SIFT, where c denotes the first letter of the word "constituent" to distinguish the constituent image $\mathbf{I}_c$ from the input panoramic image $\mathbf{I}$ and z represents the number of constituent images that stitch the panoramic image $\mathbf{I}$. For each keypoint in patch $\mathbf{P}_i$, i.e. $p_{i,k}$ $(k \in [1, n_i], i \in [1, K])$, where n_i denotes the number of keypoints in $\mathbf{P}_i$, its feature similarity with each keypoint $p_{c,j,l}$ $(l \in [1, n_j])$ in the constituent image $\mathbf{I}_c^j$ is

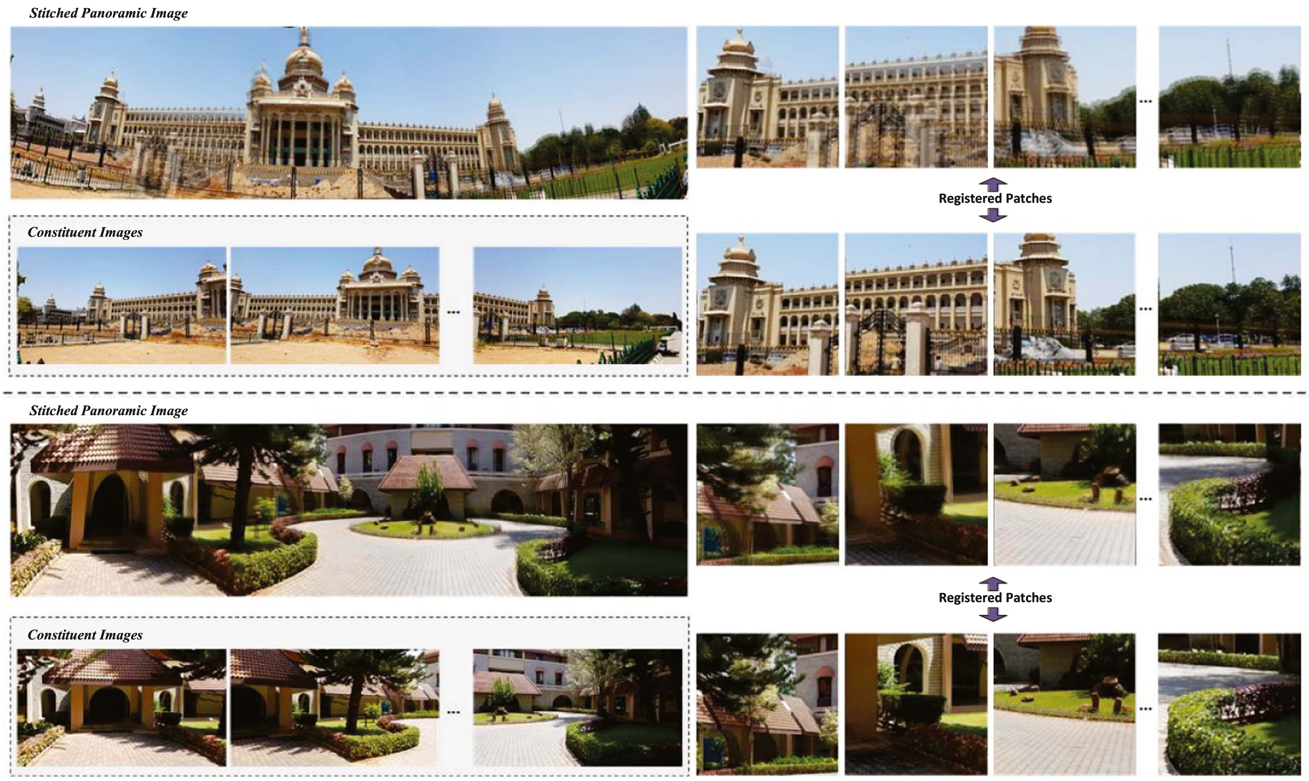


Fig. 4. Presentation of the patch registration results.

measured by computing the Euclidean distance,

$$s_{i,k;c,j,l} = \sqrt{(v_1 - v_1')^2 + \ldots + (v_m - v_m')^2}, \quad (1)$$

where m denotes the dimension of feature descriptors, which is set to the default value 128 [29], v_m and v_m' are the SIFT feature values of keypoints $p_{i,k}$ and $p_{c,j,l}$, n_j denotes the number of keypoints in $\mathbf{I}_c^j$.

The keypoint corresponding to the minimum distance s is determined as the final matching points. Then, the similarity degree between $\mathbf{P}_i$ and the constituent image $\mathbf{I}_c^j$ is defined as,

$$d_{i;j} = \frac{M}{n_i} \times 100\%, \quad (2)$$

where M represents the number of matching points between $\mathbf{P}_i$ and $\mathbf{I}_c^j$.

The constituent image $\mathbf{I}_c^j$ corresponding to the highest similarity degree is determined as the constituent image best matching with $\mathbf{P}_i$. Further, the best matching patch $\mathbf{P}_i'$ is produced by taking the minimum enclosing rectangle of all the matching keypoints in $\mathbf{I}_c^j$. Fig. 3(c) shows the five constituent images of the stitched PI in image (a), i.e. $\mathbf{I}_c^1$-$\mathbf{I}_c^5$ and their corresponding similarity degree d with patch $\mathbf{P}_5$. It can be observed from the results that the constituent image $\mathbf{I}_c^2$ has the highest similarity with $\mathbf{P}_5$ and the best matching patch $\mathbf{P}_5'$ is highlighted with the green border.

For more accurate distance measure between the features of $\mathbf{P}_i$ and $\mathbf{P}_i'$, a feature alignment stage is further designed, which can also be seen from Fig. 2. Specifically, the shooting angle of $\mathbf{P}_i$ is first adjusted to be same to that of $\mathbf{P}_i'$, which can be formulated as,

$$\mathbf{P}_i' = \mathbf{P}_i \cdot \mathbf{H}, \quad (3)$$

where $\mathbf{H}$ is an affine transformation matrix, more details of which can be obtained from [29]. After angle alignment, the content between $\mathbf{P}_i$ and $\mathbf{P}_i'$ is further aligned by template mapping. Finally, a cropping operation is implemented to remove the useless contents introduced by the feature alignment stage.

Fig. 3(d) shows the feature alignment between $\mathbf{P}_5$ and $\mathbf{P}_5'$. From the results, we can see that two patches are well registered and have achieved feature alignment. To more fully illustrate the effectiveness of the proposed patch registration module, Fig. 4 gives another two sets of stitched panoramic images together with the corresponding constitute images. In each set, patches in the right half are the registered patches. The top row presents some example key patches we detected and the bottom row are the corresponding registered patches. This figure presents the good registration effect of our patch registration module, which is beneficial to the following feature representation.

B. Bidimensional Feature Aggregation

1) Local Feature Representation (LFR): Due to the powerful information exploration ability of the convolutional neural network (CNN), it is employed to hierarchically capture the local information. With the ith key patch $\mathbf{P}_i$ and the corresponding matching patch $\mathbf{P}_i'$, they are first fed into the CNN we designed for local feature exploration, which is composed of several Convolution, ReLU activation and Max pooling operations. With CNN, the local feature maps of $\mathbf{P}_i$ and $\mathbf{P}_i'$ are generated, which are denoted by $\mathbf{f}_i$ and $\mathbf{f}_i'$. Then, the difference map between $\mathbf{f}_i$ and $\mathbf{f}_i'$ is calculated to describe the feature distance caused by quality degradation,

$$\mathbf{D}_i = |\mathbf{f}_i - \mathbf{f}_i'|. \tag{4}$$

Then, three feature maps $\mathbf{f}_i$, $\mathbf{f}_i'$ and $\mathbf{D}_i$ are concatenated together for local feature representation, generating the feature map $\mathbf{F}_i$,

$$\mathbf{F}_i = [\mathbf{f}_i, \mathbf{f}_i', \mathbf{D}_i] \in \mathbb{R}^{(C_1+C_2+C_3)\cdot h\cdot w}, \tag{5}$$

where C_1, C_2 and C_3 denote the channel number of three feature maps; h and w represent the height and weight, respectively. In this work, three feature maps have the same channel number.

Further, the adaptive average pooling (AAP) is implemented, converting $\mathbf{F}_i$ to $\mathbf{V}_i$.

2) Global Feature Representation (GFR): Inspired by the global feature representation ability of the transformer architecture that is initially applied to process the natural language processing task and now has achieved wide applications in various computer vision tasks, the classic vision Transformer (ViT) [34] method is improved for our work. This is the first time to introduce the transformer framework to this task. The first difference from the original ViT method is that the input of the transformer block is the feature map with shrinking size instead of the raw input image, which has the advantage of easing training difficulty. Concretely, with the feature vector of each key patch, i.e. $\mathbf{V}_i (i \in [1, K])$, a 1×1 convolution is first implemented to reduce channel dimensions, the result of which is dubbed as $\mathbf{U}_i$. Then, $\mathbf{U}_i$ are flattened and mapped to E dimensions using a learnable linear projection L, generating the patch embeddings. The linear projection stage can be described as,

$$\mathbf{T}_i = \mathbf{U}_i \times \mathbf{L}, \mathbf{L} \in R^{(N^2\times J)\times E}, i \in [1, K] \tag{6}$$

where N and J are the size and the channel number of the input feature map $\mathbf{U}_i$, and the value of E is equal to the product of N^2 and J.

Further, a score token is prepended to the sequence of embedded patches, which is denoted by $\mathbf{T}_0$. The output of this score token serves as the global feature representation. So far, the vector sequence can be represented as $\mathbf{T} = \{\mathbf{T}_0; \mathbf{T}_1; \ldots; \mathbf{T}_K\}$. It is fed into the transformer encoder, which is composed of M cascaded transformer blocks. Each transformer block consists of a series of Layer Normalization (LN), Multihead Self-Attention (MHSA) and Multi-Layer Perception (MLP) operations. Besides, different from ViT or other vision transformers that adopt fixed or learnable positional encodings [34], the conditional positional encoding (CPE) that is dynamically generated and conditioned on the local neighborhood of input tokens is added between each two transformer blocks [49], which can adapt to the change of the input size and ensure translation equivalence and is also helpful to the improvement of both generalization and performance. The above process can be expressed by the following formulas,

$$\mathbf{T}(0) = \mathbf{T}, \tag{7}$$

$$\mathbf{T}''(\xi-1) = \text{MHSA}(\text{LN}(\mathbf{T}(\xi-1))) + \mathbf{T}(\xi-1), \tag{8}$$

$$\mathbf{T}'(\xi-1) = \text{MLP}(\text{LN}(\mathbf{T}''(\xi-1))) + \mathbf{T}''(\xi-1), \tag{9}$$

$$\mathbf{T}(\xi) = \text{CPE}[\mathbf{T}'(\xi-1)] + \mathbf{T}'(\xi-1), \xi \in [1, M] \tag{10}$$

where $\mathbf{T}'(\xi-1)$ denotes the output of the $(\xi-1)$th transformer block, and the details of the CPE submodule is presented at the bottom of Fig. 2. It can be seen from this figure that the feature token sequence of $\mathbf{T}'(\xi-1)$, i.e. $\mathbf{T}'_{\{1,2\ldots,K\}}(\xi-1) \in R^{K\times J}$ is first reshaped to $\hat{\mathbf{T}}'_{\{1,2\ldots,K\}}(\xi-1) \in R^{H\times W\times J}$ in the 2-D space. Subsequently, a function F implemented with a 2-D convolution is applied to each local patch in $\hat{\mathbf{T}}'_{\{1,2\ldots,K\}}(\xi-1)$, the output of which is further reshaped to the dimension $R^{K\times J}$, generating the conditional position encodings. More details about this submodule can be found in [49].

C. Quality Prediction

The score token in the output feature vector of the global feature representation module is adopted for quality prediction, which is formulated as,

$$X = \text{LN}(\mathbf{T}_0(M)). \tag{11}$$

Then, one fully-connected (FC) layer with 2048 nodes and one FC layer with 1 node are sequentially employed to predict the final quality score Q_p. For model training, the Mean Absolute Error (MAE) between Q_p and the subjective quality score Q_s is measured as supervision,

$$l = \frac{1}{B}\sum_{i=1}^{B} |Q_p(i) - Q_s(i)|. \tag{12}$$

where B denotes the batchsize during model training.

IV. Experimental Analysis

A. Experimental Settings

1) Dataset: The performance of the PRBFA method is testified on the most widely recognized dataset for the SPIQA task, i.e. the publicly released ISIQA dataset [20]. A total of 264 SPIs are provided in this database, which is generated from four or five constituent images using various stitching algorithms.

2) Implementation Details: Our work is implemented in PyTorch framework. To ease the pressure from data amount, our model is first pretrained on the KADID-10 K dataset [50]. Then, the pretrained model further goes through the fine-tune stage on the ISIQA dataset to more better adapt to the SPIQA task. The learning rates of the pretraining and the fine-tuning stages are 1×10^{-4} and 1×10^{-5}, respectively. Also, the batch size is set to 12. Adam [51] with the default parameters is employed as the optimizer.

For performance presentation, three acknowledged criteria are adopted, i.e. PLCC, SRCC and RMSE [37], [38], [39]. More specifically, 80% samples are randomly chosen from the ISIQA database to learn the quality module, and all remaining samples are employed for model validation [52], [53]. The above model learning and validation process based on random selection is performed 10 times. For each time, one set of PLCC, SRCC and

TABLE I
PERFORMANCE SUMMARIZATION OF THE PRBFA METHOD AND FOURTEEN STATE-OF-THE-ARTS

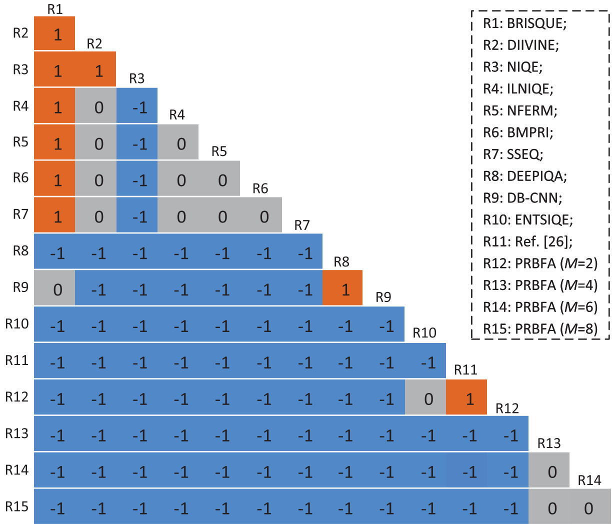
Metric	Type	PLCC	SRCC	RMSE
BRISQUE [40]	GP	0.559	0.533	0.935
DIIVINE [41]	GP	0.303	0.501	1.177
NIQE [42]	GP	0.179	0.163	1.536
ILNIQE [43]	GP	0.338	0.285	1.137
NFERM [44]	GP	0.321	0.373	1.155
BMPRI [45]	GP	0.392	0.404	1.098
SSEQ [46]	GP	0.317	0.347	1.156
DEEPIQA [47]	GP	0.633	0.596	0.856
DB-CNN [48]	GP	0.512	0.508	0.988
SIQE [20]	S	0.840	0.832	-
ENTSIQE [23]	S	0.834	0.834	0.643
Ref. [24]	S	0.853	0.841	-
BSPIQA [25]	S	0.802	0.759	-
Ref. [26]	S	0.861	0.868	0.562
PRBFA (M=2)	S	0.828	0.828	0.682
PRBFA (M=4)	S	**0.887**	**0.878**	**0.481**
PRBFA (M=6)	S	0.865	0.851	0.516
PRBFA (M=8)	S	0.869	0.853	0.509

RMSE values is calculated. The mean value of each criterion is computed as the final performance value.

B. Performance Analysis

In this part, we carry out extensive experiments to investigate the performance of the PRBFA method, nine general-purpose IQA methods and five methods specially designed for the SPIQA task. The nine general-purpose methods include BRISQUE [40], DIIVINE [41], NIQE [42], ILNIQE [43], NFERM [44], BMPRI [45], SSEQ [46], DEEPIQA [47] and DB-CNN [48], etc. The five targeted SPIQA metrics include SIQE [20], ENTSIQE [23], Ref. [24], BSPIQA [25], and Ref. [26], etc.

Table I summarizes the performance of all metrics on the ISIQA dataset, where "GP" and "S" in the second column severally denote general-purpose IQA methods and the methods targeted at the SPIQA task. For intuition, the best results have been highlighted boldfaced. It can be known from this table that the SPIQA metrics obviously outperform the general-purpose metrics. To be specific, the DEEPIQA method [47] obtains the optimal performance among all general-purpose metrics. However, both its PLCC and SRCC values are only about 0.6. By contrast, existing SPIQA metrics perform much better in terms of each performance criterion. For instance, the Ref. [26] method has the maximum PLCC and SRCC and the minimum RMSE among all existing works, which are 0.861, 0.868 and 0.562, respectively. Even so, our proposed PRBFA method (M=4) is still superior to the optimal Ref. [26] method, which demonstrates the best prediction accuracy. Moreover, it can be seen that when M is set to 2, the PRBFA method has a significant performance degradation, but it is still better than most compared metrics. When M is set to 6 or 8, the PRBFA method and the Ref. [26] method have comparable performance.



	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14
R2	1													
R3	1	1												
R4	1	0	-1											
R5	1	0	-1	0										
R6	1	0	-1	0	0									
R7	1	0	-1	0	0	0								
R8	-1	-1	-1	-1	-1	-1	-1							
R9	0	-1	-1	-1	-1	-1	-1	1						
R10	-1	-1	-1	-1	-1	-1	-1	-1	-1					
R11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1				
R12	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1			
R13	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		
R14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	
R15	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0

R1: BRISQUE;
R2: DIIVINE;
R3: NIQE;
R4: ILNIQE;
R5: NFERM;
R6: BMPRI;
R7: SSEQ;
R8: DEEPIQA;
R9: DB-CNN;
R10: ENTSIQE;
R11: Ref. [26];
R12: PRBFA (M=2)
R13: PRBFA (M=4)
R14: PRBFA (M=6)
R15: PRBFA (M=8)

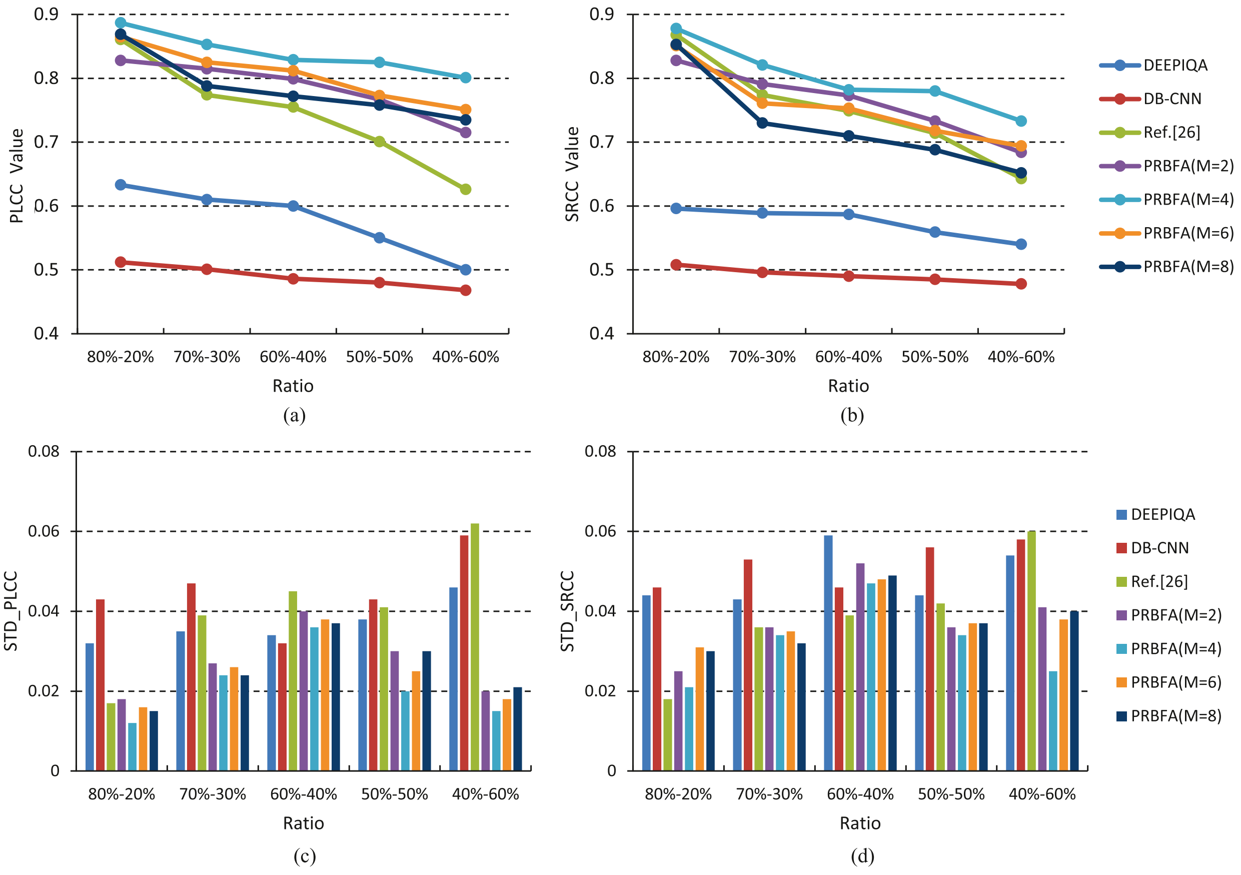
Fig. 5. Presentation of the statistical performance between each two quality metrics.

Further, we compare the statistical performance between each two quality metrics by the F-test experiment [54], [55]. First, the threshold F_c is calculated based on the residual number. In our work, the threshold F_c is 1.171 with the 90% confidence level. If two metrics to be compared are denoted as R_x and R_y, the F-test score is computed through their RMSE results, which are formulated as,

$$F_s = \left(\frac{R_x}{R_y}\right)^2. \tag{13}$$

The statistical performance is determined by the magnitudes of F_c and F_s. If the value of F_s is larger than that of F_c, it means the metric R_y outperforms R_x in terms of statistical performance. If the value of F_s is between $\frac{1}{F_c}$ and F_c, it means the competitive performance of two metrics. Otherwise, R_x outperforms R_y. For intuition, the results are presented in Fig. 5, where "1" represents the method above the square lattice is superior to the method on the left, while "0" represents the similar performance between two metrics and "-1" indicates the superior statistical performance of the method on the left side of the square lattice. It can be summarized from Fig. 5 that the proposed PRBFA method performs the best in terms of statistical performance. Concretely, when M is set to 4, 6 or 8, the PRBFA method (R_{13} R_{14} or R_{15}) has the most superior statistical performance. When M is set to 2, the PRBFA method (R_{12}) also has significantly superior statistical performance to all existing methods except the ENTSIQE and Ref. [26] methods (R_{10}, R_{11}).

Moreover, we study performance of the PRBFA method and other three deep learning based IQA metrics when different ratios of samples are chosen for model training. In this work, five ratios are tested, which are 80%, 70%, 60%, 50% and 40%. For each ratio, the random selection process is conducted for ten times and the average performance is reported, which can be seen from Fig. 6. images (a) and (b) show the PLCC and SRCC values of each metric at each ratio, and images (c) and (d) show



(a) (b) (c) (d)

Fig. 6. Performance and stability of the PRBFA metric with different M settings and other three deep learning based IQA metrics when different percentages of samples participate in model learning. Images (a) and (b): PLCC and SRCC values at each ratio; Images (c) and (d): STD_{PLCC} and STD_{SRCC} values at each ratio.

the STD_{PLCC} and STD_{SRCC} values. We can observe from images (a) and (b) that the performance of each metric has some decrease with the reduction of training images. When 80% images participate in model training, each metric reaches the performance peak. Besides, even if only 40% image samples are employed to learn the quality score prediction model, the performance of the PRBFA method still has very impressive performance. Especially when M is set to 4, the performance is still better than most compared metrics in Table I. Next, it can be found from images (c) and (d) that the proposed PRBFA method has relatively smaller STD_{PLCC} and STD_{SRCC} values in most cases, and when $M = 4$, the smallest STD values are obtained among all parameter settings. These experimental results present the less dependency of the PRBFA method on the amount of samples used for model learning, implying the promising practical application ability. Besides, compared with the SRCC criterion of the PRBFA method, the PLCC criterion always has higher average value but lower STD value, which indicates that our method performs better in prediction accuracy than in prediction monotonicity.

C. Ablation Experiment

The purpose of this experiment is to testify the contribution of one module by comparing the performance change before and after this module is removed from the whole framework. First, the contributions of three main modules are studied, including the Patch Registration (PR) module, the LFR module and the GFR module. The performance results of the PRBFA method with different M values are shown in Table II, from which we can observe that no matter for each parameter setting, the removal of any module results in the significant performance drop, especially when the LFR module is removed. This presents the effectiveness of the LFR module in simulating the local visual perception process of the high-resolution panoramic images. Besides, the performance drop caused by the removal of the GFR module further indicates the effectiveness of this module in simulating the local information integration process. In conclusion, each module in the proposed PRBFA method is demonstrated to be effective and necessary.

As stated in Section III-B, three sets of features ($\mathbf{f}$, $\mathbf{f}'$ and $\mathbf{D}$) are concatenated and aggregated for quality prediction in the LFR module. So we further test the contribution of each set of features by preserving different sets of features. The performance results of the PRBFA method with different M values are reported in Table III. For intuition, the best experimental results under each parameter setting are highlighted in bold. We can see that no matter what value M is set to, the performance reaches the peak by aggregating three sets of features together. In other words, discarding any set of features causes performance

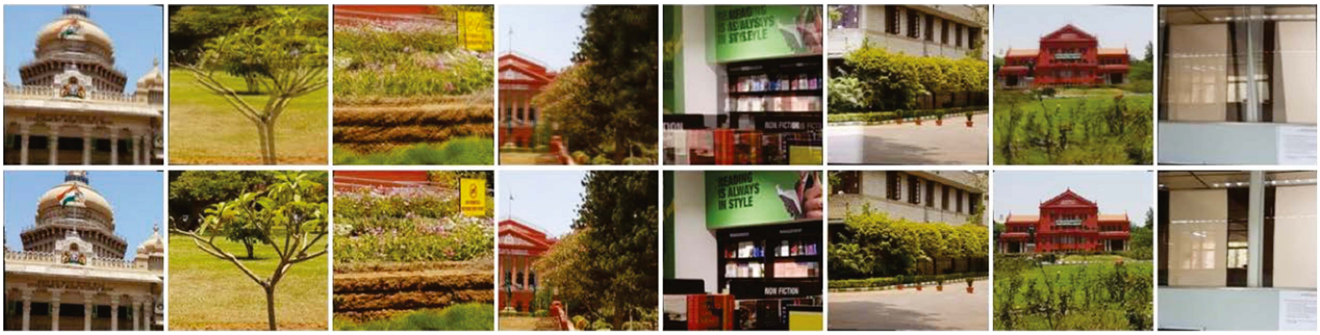


Fig. 7. Visualization of the registration effect. Top row: Key patches detected from the distorted panoramic images; Bottom row: Reference patches generated by the proposed patch registration method.

TABLE II
PERFORMANCE OF OUR PRBFA METHOD BEFORE AND AFTER ONE MODULE IS REMOVED

M value	Model	PLCC	SRCC	RMSE
M=2	Without PR	0.748	0.769	0.762
	Without LFR	0.330	0.363	1.150
	Without GFR	0.805	0.782	0.719
	PRBFA	**0.828**	**0.828**	**0.682**
M=4	Without PR	0.836	0.828	0.569
	Without LFR	0.359	0.425	1.119
	Without GFR	0.805	0.782	0.719
	PRBFA	**0.887**	**0.878**	**0.481**
M=6	Without PR	0.860	0.846	0.563
	Without LFR	0.298	0.320	1.183
	Without GFR	0.805	0.782	0.719
	PRBFA	**0.865**	**0.851**	**0.516**
M=8	Without PR	0.836	0.749	0.568
	Without LFR	0.296	0.341	1.185
	Without GFR	0.805	0.782	0.719
	PRBFA	**0.869**	**0.853**	**0.509**

The bold values represents the better results.

TABLE III
PERFORMANCE WHEN DIFFERENT FEATURE SETS ARE EMPLOYED FOR MODEL TRAINING

M value	Features	PLCC	SRCC	RMSE
M=2	$\{\mathbf{f}\}$	0.748	0.769	0.762
	$\{\mathbf{D}\}$	0.792	0.779	0.721
	$\{\mathbf{f}, \mathbf{D}\}$	0.807	0.767	0.713
	$\{\mathbf{f}, \mathbf{f}'\}$	0.822	0.748	0.697
	$\{\mathbf{f}, \mathbf{f}', \mathbf{D}\}$ (Pro.)	**0.828**	**0.828**	**0.682**
M=4	$\{\mathbf{f}\}$	0.836	0.828	0.569
	$\{\mathbf{D}\}$	0.823	0.791	0.696
	$\{\mathbf{f}, \mathbf{D}\}$	0.859	0.849	0.563
	$\{\mathbf{f}, \mathbf{f}'\}$	0.865	0.854	0.518
	$\{\mathbf{f}, \mathbf{f}', \mathbf{D}\}$ (Pro.)	**0.887**	**0.878**	**0.481**
M=6	$\{\mathbf{f}\}$	0.860	0.846	0.563
	$\{\mathbf{D}\}$	0.858	0.841	0.564
	$\{\mathbf{f}, \mathbf{D}\}$	0.827	0.813	0.687
	$\{\mathbf{f}, \mathbf{f}'\}$	0.836	0.827	0.569
	$\{\mathbf{f}, \mathbf{f}', \mathbf{D}\}$ (Pro.)	**0.865**	**0.851**	**0.516**
M=8	$\{\mathbf{f}\}$	0.836	0.749	0.568
	$\{\mathbf{D}\}$	0.836	0.809	0.569
	$\{\mathbf{f}, \mathbf{D}\}$	0.838	0.820	0.565
	$\{\mathbf{f}, \mathbf{f}'\}$	0.859	0.836	0.563
	$\{\mathbf{f}, \mathbf{f}', \mathbf{D}\}$ (Pro.)	**0.869**	**0.853**	**0.509**

The bold values represents the better results.

degradation. Besides, it can be observed that when M is set to 4, the PLCC and SRCC values both reach the highest level, which are 0.887 and 0.878. Therefore, all three sets of features are employed and M is set to 4 for the final PRBFA method.

We further study the advantages of the proposed LFR module, which is achieved by replacing it with existing CNN networks. Table IV reports the experimental results. By contrast, the proposed network is more effective than the existing classic CNNs in the local feature representation task in terms of both the average prediction performance and the prediction stability.

D. Visualization

To more intuitively display the effectiveness of the proposed patch registration method, several pairs of registered patches are presented in Fig. 7. The top row presents some example key patches detected from the distorted panoramic images using the proposed patch selection method and the bottom row shows the corresponding reference patches obtained by the proposed patch matching and feature alignment steps. By comparison, we can observe that the registered patches in the bottom row indeed provide the reference information of the distorted patches in the top row and the features between each pair of images are aligned well, which proves the excellent performance of the patch registration module.

Further, the feature visualization results of five images are given in Fig. 8 to intuitively display the effectiveness of the PRBFA method. The first row are the test images and the following rows are the attention maps obtained from the DEEP-IQA [47], DB-CNN [48], Ref. [26] and PRBFA metrics, respectively. The first two test images are the ones without obvious stitching distortions while the following three test images are the ones with obvious distortions. Besides, five test images

TABLE IV
Performance of the PRBFA Method With Various LFR Networks

Network	PLCC	SRCC	STD_{PLCC}	STD_{SRCC}
Resnet18	0.810	0.811	0.044	0.050
Resnet34	0.849	0.844	0.042	0.040
VGG16	0.770	0.754	0.056	0.058
VGG19	0.850	0.825	0.163	0.228
Proposed (M=4)	**0.887**	**0.878**	**0.012**	**0.027**

The bold values represents the better results.

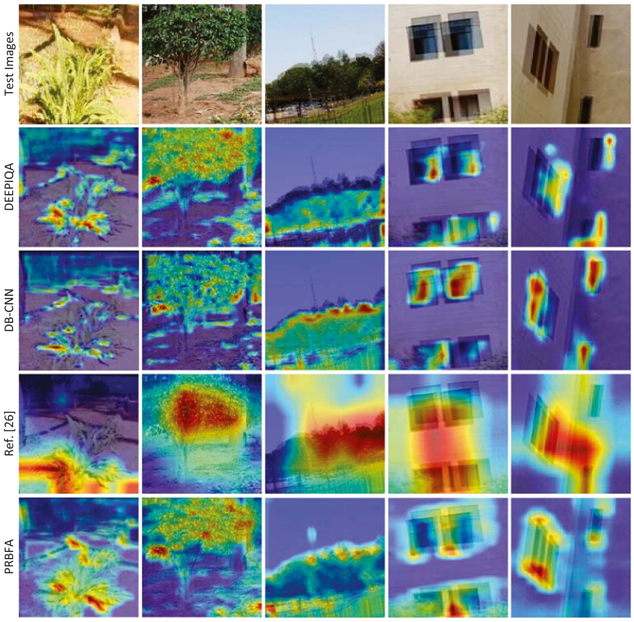


Fig. 8. Feature visualization results.

are the key patches detected during the patch registration process, instead of the whole SPIs with high resolutions. The aim is to ensure consistency with the input of the feature aggregation module and also ensure clarity. By comparison, it can be found from the first two columns that the results obtained from our PRBFA method are more consistent with the results of the human visual perception, where the critical regions are more accurately focused. Moreover, we can see from the following three columns that our PRBFA method can more accurately capture those conspicuous distortion regions, especially for the ghosting distortions at edge regions. This experiment more intuitively proves the effectiveness of the PRBFA metric.

V. Conclusion

In this article, we have proposed an SPIQA metric by more comprehensively simulating the process of human perceiving the panoramic images. First, it is a patch-level metric to accommodate the limited range of human vision. Second, a soft key patch segmentation and selection module is designed to conform to the visual attention mechanism. Furthermore, the local features are explored followed by the global feature exploration. This successive feature representation process is designed to simulate the human perception process of panoramic images from local to global. The experimental results have presented that the PRBFA method performs favorably against state-of-the-art in terms of both prediction accuracy and stability.

References

[1] Q. Zhang, J. Wei, S. Wang, S. Ma, and W. Gao, "RealVR: Efficient, economical, and quality-of-experience-driven VR video system based on MPEG OMAF," *IEEE Trans. Multimedia*, 2022, early access, Jul. 14, 2022, doi: 10.1109/TMM.2022.3190697.

[2] Y. Zhu, G. Zhai, X. Min, and J. Zhou, "The prediction of saliency map for head and eye movements in 360 degree images," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2331–2344, Sep. 2020.

[3] H. W. Chen et al., "Perceptual quality assessment of cartoon images," *IEEE Trans. Multimedia*, vol. 25, pp. 140–153, 2023.

[4] C. Meng et al., "Objective quality assessment of lenslet light field image based on focus stack," *IEEE Trans. Multimedia*, vol. 24, pp. 3193–3207, 2022.

[5] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Online hodgeRank on random graphs for crowdsourceable QoE evaluation," *IEEE Trans. Multimedia*, vol. 16, pp. 373–386, 2014.

[6] Q. Xu, J. C. Xiong, Q. M. Huang,, and Y. Yao, "Robust evaluation for quality of experience in crowdsourcing," in *Proc. ACM Conf. Multimedia*, 2013, pp. 43–52.

[7] Q. Xu et al., "Hodgerank on random graphs for subjective video quality assessment," *IEEE Trans. Multimedia*, vol. 14, pp. 844–857, 2012.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] W. Sun et al., "MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 64–77, Jan. 2020.

[14] H. G. Kim, H. T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 917–928, Apr. 2020.

[15] H.-T. Lim, H. G. Kim, and Y. M. Ra, "VR IQA NET: Deep virtual reality image quality assessment using adversarial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6737–6741.

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

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

[18] L. Y. Yang, Z. G. Tan, Z. Huang, and G. Cheung, "A content-aware metric for stitched panoramic image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2487–2494.

[19] 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, pp. 1–6.

[20] 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.

[21] C. Z. Tian, X. L. Chai, and F. Shao, "Stitched image quality assessment based on local measurement errors and global statistical properties," *J. Vis. Commun. ImageRepresentation*, vol. 81, pp. 1–13, 2021.

[22] X. J. Wang, X. L. Chai, and F. Shao, "Quality assessment for color correction-based stitched images via bi-directional matching," *J. Vis. Commun. Image Representation*, vol. 75, pp. 1–8, 2021.

[23] K. Okarma et al., "Entropy-based combined metric for automatic objective quality assessment of stitched panoramic images," *Entropy*, vol. 23, pp. 1–13, 2021.

[24] C. Z. Tian, X. L. Chai, and F. Shao, "Stitched image quality assessment based on local measurement errors and global statistical properties," *J. Vis. Commun. Image Representations*, vol. 81, 2021, Art. no. 103324, doi: 10.1016/j.jvcir.2021.103324.

[25] J. Hou, W. Lin, and B. Zhao, "Content-dependency reduction with multi-task learning in blind stitched panoramic image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 3463–3467.

[26] Y. Zhou et al., "Pyramid feature aggregation for hierarchical quality prediction of stitched panoramic images," *IEEE Trans. Multimedia*, early access, May 03, 2022, doi: 10.1109/TMM.2022.3171684.

[27] H. Duan et al., "Attentive deep image quality assessment for omnidirectional stitching," *IEEE J. Sel. Topics Signal Process.*, early access, Mar. 01, 2023, doi: 10.1109/JSTSP.2023.3250956.

[28] K.-Y. Lin and G. X. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 732–741.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[30] W. L. Zhao and C. W. Ngo, "Flip-invariant SIFT for copy and object detection," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 980–991, Mar. 2013.

[31] X. Shuai, C. Zhang, and P. Hao, "Fingerprint indexing based on composite set of reduced SIFT features," in *Proc. IEEE 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[32] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM Sigmod Int. Conf. Manage. Data. ACM*, 2000, pp. 93–104.

[33] D. Arthur and S. Vassilvitskii, "K-Means++:The advantages of careful seeding," in *Proc. 8th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035.

[34] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[35] P. Shaw, J. Uszkoreit,, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 464 468.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[37] Y. Zhou et al., "No-reference quality assessment for view synthesis using DoG-based edge statistics and texture naturalness," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4566–4579, Sep. 2019.

[38] Y. Zhou, Y. Sun, L. Li, K. Gu, and Y. Fang, "Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1767–1777, Apr. 2022, doi: 10.1109/TCSVT.2021.3081162.

[39] Y. Zhou et al., "Blind quality index for multiply distorted images using biorder structure degradation and nonlocal statistics," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3019–3032, Nov. 2018.

[40] 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.

[41] 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.

[42] 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.

[43] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[44] 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.

[45] 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.

[46] L. X. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process., Image Commun.*, vol. 29, pp. 856–863, 2014.

[47] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[48] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[49] X. X. Chu, Z. Tian, B. Zhang, X. L. Wang, and C. H. Shen, "Conditional positional encodings for vision transformers," in *Proc. IEEE Int. Conf. Learn. Representations*, 2023, pp. 517–522.

[50] H. Lin, V. Hosu, and D. Saupe, "KADID-10 k: A large-scale artificially distorted IQA database," in *Proc. IEEE Int. Conf. Qual. Multimedia Experience*, 2019, pp. 1–3.

[51] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[52] X. Wang et al., "Exploiting local degradation characteristics and global statistical properties for blind quality assessment of tone-mapped HDR images," *IEEE Trans. Multimedia*, vol. 23, pp. 692–705, 2021.

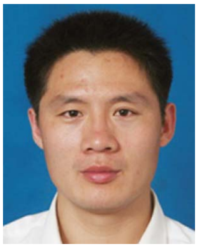
[53] S. Ling et al., "Re-visiting discriminator for blind free-viewpoint image quality assessment," *IEEE Trans. Multimedia*, vol. 23, pp. 4245–4258, 2021.

[54] L. Li et al., "Image sharpness assessment by sparse representation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1085–1097, Jun. 2016.

[55] L. Li, Y. Li, J. Wu, L. Ma, and Y. Fang, "Quality evaluation for image retargeting with instance semantics," *IEEE Trans. Multimedia*, vol. 23, pp. 2757–2769, 2021.

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

Weikang Gong received the B.S. degree in communication engineering from the North China Institute of Aerospace Engineering, Langfang, China, in 2020 and the M.S. degree from the China University of Mining and Technology, Xuzhou, China. His research interests include image quality assessment, computer vision, and deep learning.



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

Leida Li (Member, IEEE) received the B.E. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid Rich Object Search Lab, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. From 2009 to 2019, he was an Assistant Professor, Associate Professor, and Professor with the China University of Mining and Technology, Xuzhou, China. He is currently a Professor with the School of Artificial Intelligence, Xidian University. His research interests include multimedia quality assessment, computational aesthetics, and affective computing. He was the Area Chair for IJCAI 2019–2020 and ICME 2023, the Session Chair for ICMR 2019 and PCM 2015, and a TPC for AAAI 2019, ACM MM 2019–2020, ACM MM-Asia 2019, ACII 2019, and PCM 2016. He is an Associate Editor for *Journal of Visual Communication and Image Representation*.

Ke Gu (Senior Member, IEEE) 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. Dr. Gu was the recipient of the Best Paper Award from IEEE TRANSACTIONS ON MULTIMEDIA and Best Student Paper Award at IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO (ICME) in 2016. He was the Leading Special Session Organizer in IEEE INTERNATIONAL CONFERENCE ON VISUAL COMMUNICATIONS AND IMAGE PROCESSING 2016 and IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING 2017. He is also an Associate Editor for *Computer Animation and Virtual Worlds* and *IET Image Processing*, the Area Editor of *Signal Processing: Image Communication*, and the Editor of *Applied Sciences*, *Displays*, and *Entropy*. He is also a Reviewer for 20 top SCI journals.

Jinjian Wu (Member, IEEE) received the B.Sc. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. From 2013 to 2014, he was a Postdoctoral Research Fellow with Nanyang Technological University, Singapore. From 2015 to 2019, he was an Associate Professor with Xidian University, where he had been a Full Professor, since 2019. His research interests include visual perceptual modeling, biomimetic imaging, quality evaluation, and object detection. He was the recipient of the Best Student Paper Award at the ISCAS 2013. He was an Associate Editor for *Circuits, Systems and Signal Processing*, the Special Section Chair of IEEE Visual Communications and Image Processing 2017, and the Section Chair/Organizer/TPC Member of ICME 2014–2015, PCM 2015–2016, ICIP 2015, VCIP 2018, ACMMM Asia 2019, CICAI 2021, PRCV 2021, AAAI 2019–2021, and ACMMM 2021–2022.