Pairwise Learning to Rank for Image Quality Assessment

YIQING SHI1, YUZHEN NIU1,2, (Member, IEEE), WENZHONG GUO1,2, (Member, IEEE), YIZE HUANG1, AND JIAMEI ZHAN1

1 Fujian Key Laboratory of Network Computing and Intelligent Information Processing, College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China
2 Key Laboratory of Spatial Data Mining and Information Sharing, Ministry of Education, Fuzhou University, Fuzhou 350116, China

Corresponding author: Wenzhong Guo (fzugwz@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672158, Grant 61672159, and Grant U1705262; and in part by the Natural Science Foundation of Fujian Province under Grant 2019J02006.

ABSTRACT Because the pairwise comparison is a natural and effective way to obtain subjective image quality scores, we propose an objective full-reference image quality assessment (FR-IQA) index based on pairwise learning to rank (PLR). We first compose a large number of pairs of images, extract their features, and compute their preference labels as training labels. We then obtain a pairwise preference model by training a binary classifier using the features and labels. Because image quality is affected by the masking effect, we propose extracting frequency-aware quality features by adapting state-of-the-art IQA metrics. The learned pairwise preference model is then used to predict the preference between pairs of images in the testing dataset. The quality of each image is computed as the number of preferences. Experimental results on four IQA databases validate that the proposed PLR-based IQA index achieves higher consistency with human subjective evaluation than the state-of-the-art IQA metrics.

INDEX TERMS Image quality assessment (IQA), pairwise learning to rank, preference model, masking effect.

I. INTRODUCTION

Digital images and videos are an important part of our daily lives, and there is an on-going demand for high-quality images. However, the quality of images and videos can be affected by distortions introduced during acquisition, encoding, storage, transmission, and reconstruction. To ensure that the quality of images can meet the requirements of users, image quality assessment (IQA) has been introduced and IQA metrics are used in many image processing systems.

IQA can either be subjective or objective. In a subjective assessment, people directly observe an image and assess its quality by assigning a score to it. However, such the assessment cannot be applied in real-time image processing systems. In contrast, an objective assessment involves evaluating image quality through mathematical calculation. Existing objective IQA metrics can be classified into three categories: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference/blind IQA (NR-IQA/BIIQA). FR-IQA metrics, such as peak signal-to-noise ratio (PSNR) [1] and structural similarity index (SSIM) [2], have been widely adopted as the evaluation metrics of choice for a variety of image restoration and enhancement algorithms, such as image super-resolution [3], image denoising [4], [5], image dehazing [6], and image de-raining [7], [8].

Despite recent progress of IQA research, there is still a gap between subjective and objective assessments. To better simulate people’s ability to distinguish a small quality difference between two images using the comparison strategy, we propose an objective IQA index based on pairwise learning to rank (PLR) in this article.

Pairwise comparison is a natural and effective way to obtain subjective image quality scores. When two images have similar qualities, people are not good at rating a very small difference. However, people are good at comparing two images and choosing the one with a better or worse quality. The pairwise comparison was used in [9], [10] to train visual comfort assessment (VCA) models. The VCA is proposed to predict the effects induced by stereo image content on visual health. The proposed PLR-based IQA (PLRA) index first extracts training features from many pairs of images.
and computes training labels by comparing the corresponding pairs of subjective mean opinion scores (MOSs). The PLRA index then trains a binary classifier using the features and labels assigned and obtains a pairwise preference model. Because image quality is usually affected by the masking effect, a phenomenon that the visibility of distortion is reduced [11] when the distortion signal is masked by the image signal, we propose extracting frequency-aware quality features. Furthermore, because current IQA metrics have achieved a certain degree of consistency with subjective quality perception and can complement each other, we adapt these IQA metrics to extract these frequency-aware quality features.

In the testing step, the trained pairwise preference model is used to predict the preference between the pairs of images in the testing dataset. The quality of each testing image is computed as the number of preferences over the other testing images. Experimental results using four commonly used IQA databases [12]–[15] demonstrate the superior performance of the proposed PLRA index over the state-of-the-art IQA metrics.

The contributions of the paper are as follows: 1) We propose an IQA index based on pairwise learning to rank, which better simulates people’s ability to distinguish the small quality difference between two images. 2) The frequency-aware features computed in the spatial domain are introduced to incorporate the influence of the masking effect on IQA and applied to the pairwise comparison framework. 3) The proposed PLRA index has broad and high practical application values, especially for fine-grained distorted image quality assessment.

The rest of the paper is organized as follows. Section II provides related works on objective FR-IQA metrics and learning to rank methods. In Section III, we present a detailed description of the proposed PLRA index. In Section IV, the experimental results are presented. Section V concludes the paper.

II. RELATED WORK

We describe related works on FR-IQA and learning to rank in subsections II-A and II-B, respectively.

A. FR-IQA

Traditional error quantification-based FR-IQA metrics, such as mean square error (MSE), signal to noise ratio (SNR), and PSNR, do not consider the content of an image and the characteristics of the human visual system (HVS). Therefore, these metrics usually result in a weak consistency with human perception.

Therefore, HVS-inspired and image content-aware IQA metrics have been presented. Many researchers have attempted to combine the characteristics of HVS with pure mathematical algorithms. Visual SNR (VSNR) [16] combines the mathematical convenience of root-mean-squared contrasted with both the low-level and mid-level properties of vision in order to better quantify the visual fidelity of distorted images. HVS-based peak SNR (PSNR-HVS) [17] combines PSNR with HVS characteristics by considering the contrast sensitivity function (CSF). PSNR-HA [18] introduces the CSF, between-coefficient contrast masking of discrete cosine transform (DCT) basis functions, and the mean shift into PSNR. PSNR-HMA [18] introduces contrast changing into PSNR-HA.

In 2002, Wang and Bovik claimed that the human eye obtains image information through three channels: brightness, contrast, and structure. They developed the SSIM [2] from this concept. The edge-strength-similarity-based image quality metric (ESSIM) [19] was proposed using the assumption that the edge strength of each pixel could represent the semantic information of images. Gu et al. found that the extractable information for a given visual scene considerably depended on viewing distance and image resolution; a concept, which had been mostly overlooked in IQA. They proposed a new optimal scale selection (OSS) model [20] which considered viewing distance and image resolution before using the IQA models, and their experimental results show that the OSS-SSIM model achieves better performance than SSIM itself.

Unlike most IQA metrics that adopt the single most relevant strategy used by HVS to judge image quality, the most apparent distortion (MAD) [15] index presented by Larson and Chandler is modeled using two separate strategies: detection-based and appearance-based. Zhang et al. proposed the feature similarity index (FSIM) [21], which used phase congruency and gradient magnitude as primary and secondary features.

In recent years, more HVS-inspired IQA metrics have been presented. Zhang et al. proposed a visual saliency-induced metric (VSI) [22] which introduced visual saliency into assessment. During the final pooling, the saliency value of each pixel is used as a weighting value which affects the importance of different pixels. The DCT sub-bands similarity (DSS) [23] index uses changes in the structural information of the subbands in the DCT domain and the weighted quality estimates for these subbands to predict image quality. Visual information fidelity (VIF) [24] criterion was proposed to quantify the loss of image information to the distortion process and explore the relationship between image information and visual quality. Gradient magnitude similarity deviation (GMSD) [25] was proposed based on the observation that image gradients are sensitive to image distortions. GMSD predicts image quality based on the gradient magnitude similarity between the reference and the distorted images. The improved color-image-difference (iCID) [26] metric assesses image quality by considering multiple assessment components, including SSIM, color contrast, and pixel value difference. Nafchi et al. proposed a mean deviation similarity index (MDSI) [27] to assess image quality by considering gradient similarity and chromaticity similarity to measure structural and color distortions, respectively. The superpixel-based similarity index (SPSIM) [28] evaluates the overall visual impression on local images by calculating superpixel lumi-
nance similarity and superpixel chrominance similarity, then quantifies the structural variations by gradient similarity, and finally evaluates the image quality by combining all the three types of features.

Some machine-learning based methods have also been presented. These tackle IQA in two steps: first by designing relevant features and then by performing regression or classification using those features. Narvaria and Lin proposed a singular value decomposition-based metric using support vector regression (SVR) [29]. SVDR uses singular value decomposition (SVD) to extract singular vectors as features of reference and distortion images, and then uses support vector regression (SVR) for image quality prediction. Liu et al. presented a multi-method fusion (MMF) for IQA [30] which was motivated by the observation that no single method could give the best performance in all distortion types. This metric first classifies the distortion type of distorted images into five categories and then combines multi IQA metrics using SVR to predict the image quality scores. A machine learning-based metric with distortion measured by non-negative matrix factorization (NMF) was proposed in [31] which used the extreme learning machine (ELM) method to address the limitations of existing pooling techniques.

Because convolutional neural network (CNN) is capable of learning features and performing regression based on raw image data, it has been introduced to solve the IQA problem. Bosse et al. presented a weighted average deep image quality measure for FR-IQA (WaDIQaM-FR) [32]. It is purely data-driven and does not rely on hand-crafted features or other types of prior domain knowledge about HVS or image statistics. Kim and Lee presented a deep image quality assessment (DeepQA) model [33] which learned the visual sensitivity characteristics of the HVS by using a deep CNN.

B. LEARNING TO RANK
The problem of learning to rank has recently received a large amount of attention within the machine learning literature. It enables transform a multi-class classification problem into a number of binary classification problems.

In [34], Fl"{a}wrkranz et al. considered supervised learning of a ranking function, and introduced pairwise preference learning as an extension of pairwise classification to constrain classification which is a mapping from instances to total orders over a set of labels (options). In [35], Cao et al. proposed a new approach to learning to rank, referred to as the list-wise approach which used a list of objects as instances in learning instead of using object pairs; this approach has been proven to be better than the traditional ones. In [36], H"{u}llermeier et al. presented a label ranking learning algorithm which depended on the ranking by pairwise comparison. In [37], Mai and Liu used the pairwise-based learning to rank method to compare and rank salient object detection results. When choosing the best salient object detection result, the performance of this method is better than the individual methods. Jiang et al. [10] presented a simple yet effective visual comfort assessment approach for stereoscopic images from the perspective of learning to rank. Gao et al. [38] presented an NR-IQA framework which formulated the problem of mapping difference feature vectors to preference labels as a classification problem and used a multiple kernel learning algorithm to learn a classification model. Oszust [39] introduced a hybrid FR-IQA index which used the lasso regression and pairwise score differences. Ma et al. [40] employed a neural network-based PLR algorithm to learn an opinion-unaware NR-IQA model. An NR-IQA index for the retargeted images was proposed in [41], which resorted the images to the pairwise rank learning approach to discriminate the perceptual quality between the retargeted image pairs. Bin et al. [42] proposed a deep evaluator for the retargeted images which combined the geometry and content features with two segmented stacked AutoEncoders. Liu et al. [43] proposed an NR-IQA approach that learned from ranking using the Siamese network. And then they transferred knowledge learned from ranked images to a traditional CNN fine-tuned on IQA data. Prashnani et al. [44] presented a perceptual image-error metric by leveraging the pairwise preference to create large IQA datasets and used a pairwise-learning framework to train an error-estimation function. Niu et al. [45] used a Siamese convolutional neural network that shares weights to extract features of the input two image patches to transform the IQA task into a task of quality comparison between images.

III. IQA BASED ON PAIRWISE-BASED LEARNING TO RANK
We show the framework of the proposed PLRA index in Figure 1. To train a pairwise preference model, we first extract features for each distorted image in the training dataset, compose image pairs, and then use the PLR method to train the model. In the testing process, we extract features for each distorted image in the testing dataset, then use the trained pairwise preference model to determine the preference between the two testing images, and finally get the quality score of a testing image by summing its number of preferences over other testing images.

In subsections III-A-III-C, we describe the details of feature extraction, pairwise preference model learning, and image quality prediction, respectively.

A. FEATURE EXTRACTION
1) EXISTING IQA METRICS
IQA is complex and depends on many factors. In this subsection, we describe how we design our features to represent the factors that affect image quality perception. Existing IQA metrics evaluate image quality mainly based on low-level image features, such as color, gradient, edge, and structure. Some IQA metrics integrate the HVS characteristics with image features. As shown in Figure 2, different IQA metrics usually use different image features. The performance of these IQA metrics on some public databases shows that they have achieved a certain degree of consistency with subjective
quality perception. Therefore, we adapt the existing IQA metrics to extract quality features.

In this article, we select fifteen IQA metrics to compute features: SSIM [2], PSNR [1], PSNR-HVS [17], MAD [15], PSNR-HMA [18], FSIM [21], FSIMc [21], ESSIM [19], VSI [22], DSS [23], OSS-SSIM [20], VIF [24], GMSD [25], iCID [26], and MDSI [27]. All metrics have positive correlation with subjective visual perception except for the MAD, GMSD, iCID, and MDSI metrics. The information and targeting attributes of fifteen IQA metrics are shown in Table 1. The extracted features are denoted as \( F_c = \{F_1, F_2, \ldots, F_{15}\} \), where \( F_m \) is calculated using the \( m \)th IQA index we selected, and \( m = 1, 2, \ldots, 15 \).

Figure 3 shows several input images for FR-IQA metrics with different distortion types and levels. Table 2 shows the MOSs and the quality scores calculated by the 15 IQA metrics for these distorted images. For each IQA index, if the ranking of the quality score for an image within the five images is different from the ranking of its MOS, the quality score is inconsistent with the MOS. As shown in Table 2, eight of the metrics have inconsistent scores, and 21 out of 75 scores are inconsistent with the MOSs.
2) FREQUENCY-AWARE QUALITY FEATURE

When the image signal masks the distortion signal, the visibility of distortion is reduced [11]. This phenomenon is called masking and is common for distorted images. Masking generally occurs when these two signals appear in roughly the same spatial location and share many of the same spatial frequencies. Masking is also a very complex process, and in some cases, the visibility of distortion is actually increased by the image signal. Generally, the masking effect on random backgrounds is more pronounced than it is on regular backgrounds [11]. The high frequency implies higher change of values which happens when there is sharp contrast in the image edge and contours. While the low frequency means smaller change in pixel values which corresponds to plain areas in images. It should be noted that the HVS frequency sensitivity is closely related to the masking effect [46]. For example, the high-frequency region has obvious masking effect to noise, whereas human eyes are sensitive to noise in low-frequency parts of images [46]. Meanwhile, the blur is easier to be detected in high-frequency parts than low-frequency parts of images. As shown in Figure 4(c), the distortion in the sky region is more evident than that in the cactus region. Conversely, the distortion in the cactus region is more pronounced in Figure 4(e). Therefore, image contents related to high-frequency and low-frequency should be considered separately.

To take the masking effect into consideration, we separate the image into high-frequency and low-frequency portions and assess their qualities. Based on the knowledge that high-frequency information is usually located around the image edge and contours, we first convert the input image into a gray image, and apply a fast bilateral filter [47] which is known for its edge-preserving capability during image smoothing. We then use the Canny edge detection method [48] to get an edge map and then obtain an expanded edge map by expanding the extent of the edges using a dilation operator [49].

For each expanded edge map, we treat the detected area as the high-frequency portion of the image and the remaining area as the low-frequency portion. Since the expanded edge map is a binary image, we take the edge map into consideration in the pooling stage of each IQA index to help assess the quality of frequency. Based on the expanded edge map, the pixels in a quality map can be divided into two sets. Pixels in each set follow the pooling method of the original IQA index to calculate the scores of low and high frequency parts. The quality scores are obtained using the following formulae:

\[
F^h_m \left( \frac{m}{m} \right) = \text{Pooling} \left( H \right), \quad H = \{ S(i, j) | E(i, j) = 1 \}, \quad (1)
\]

\[
F^l_m \left( \frac{m}{m} \right) = \text{Pooling} \left( L \right), \quad L = \{ S(i, j) | E(i, j) = 0 \}, \quad (2)
\]

where Pooling is the pooling function of each IQA index, \( S(i, j) \) is the quality score at pixel \( (i, j) \), \( E(i, j) \) is the edge map of the reference image, and \( H \) and \( L \) are the sets of quality scores whose pixels are in the high- and low-frequency parts of the image, respectively. In this way, we obtain 30 scores from each image. The image scores are denoted as \( F^h = \{ F^h_1, F^h_2, \ldots, F^h_{15} \} \) and \( F^l = \{ F^l_1, F^l_2, \ldots, F^l_{15} \} \), which represent the quality scores of the high- and low-frequency parts.

In Figure 5, we use the SSIM [2] index to show the calculation of the frequency-aware feature. The SSIM algorithm is first used to generate a quality map (as shown in Figure 5(e)). The quality map consists of the quality scores of all the pixels in the image, and the original SSIM index is used to calculate the average of all the quality scores of the pixels, providing the quality score of the image. To compute the frequency-aware SSIM feature, we use the expanded edge map (Figure 5(d)) of the reference image (Figure 5(a)) as the weighting map. We calculate the frequency-aware quality

**TABLE 2. MOSs and scores of the existing FR-IQA metrics of the distorted images in Figure 3.** (The quality scores that are inconsistent with the MOSs are shown in boldface. An up-arrow indicates that a high score denotes a good quality image.)

| Formula and MOS | Distorted image 1 | Distorted image 2 | Distorted image 3 | Distorted image 4 | Distorted image 5 |
|-----------------|------------------|------------------|------------------|------------------|------------------|
| MOS†            | 5.7275          | 5.0305          | 4.0882          | 3.7555          | 2.6667          |
| PSNR†           | 31.5241         | 29.4309         | 31.8390         | 25.7778         | 23.9587         |
| PSNR-HVS†       | 36.6545         | 24.7866         | 26.3506         | 25.5018         | 18.4648         |
| MAD†            | 27.9552         | 17.6707         | 65.0681         | 82.3334         | 105.6043        |
| PSNR-IMA†       | 43.8209         | 33.9515         | 29.2139         | 28.1029         | 22.0574         |
| FSIM†           | 0.9969          | 0.9758          | 0.9350          | 0.9409          | 0.7534          |
| FSDMC†          | 0.9955          | 0.9737          | 0.9195          | 0.9235          | 0.7413          |
| BSSIM†          | 1.0000          | 0.9998          | 0.9995          | 0.9995          | 0.9966          |
| VSI†            | 0.9945          | 0.9930          | 0.9702          | 0.9064          | 0.9007          |
| DSS†            | 0.9909          | 0.9604          | 0.7085          | 0.7582          | 0.3575          |
| OSSI-SSIM†      | 0.9994          | 0.9705          | 0.9621          | 0.9725          | 0.8187          |
| VIP†            | 0.8576          | 0.8373          | 0.4376          | 0.3869          | 0.1300          |
| GMSD_L†         | 0.0055          | 0.0150          | 0.0998          | 0.0893          | 0.2587          |
| iCIDE†          | 0.0145          | 0.1067          | 0.3079          | 0.3300          | 0.4846          |
| MDSL†           | 0.1327          | 0.1691          | 0.2665          | 0.2733          | 0.4389          |

**FIGURE 3.** Input images for IQA metrics. The associated distortion types and levels of these images are different: (a) Reference image; (b) Masked noise, L2; (c) Contrast change, L2; (d) Spatially correlated noise, L2; (e) Impulse noise, L4; (f) Gaussian blur, L4.
FIGURE 4. Reference image (a), distorted image 1 (b), two magnified distorted regions (c), distorted image 2 (d), and two magnified distorted regions (e). The distortion in image (b) is additive white Gaussian noise, and the distortion in image (d) is Bulr.

FIGURE 5. Diagram of the calculation of frequency-aware features using SSIM.

TABLE 3. MOSs and scores of the distorted image in Figure 3 computed by frequency-aware features $F^{hm}_{m}$. (The quality scores that are inconsistent with the MOSs are shown in boldface. An up-arrow indicates that a high score denotes a good quality image. A down-arrow indicates that a low score denotes a good quality image).

| MOS and MOS | Distorted image 1 | Distorted image 2 | Distorted image 3 | Distorted image 4 | Distorted image 5 |
|-------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| MOS†       | 5.7273            | 5.0303            | 4.0882            | 3.7353            | 2.6667            |
| SSIM†      | 0.9031            | 0.9160            | 0.8848            | 0.6843            | 0.5203            |
| PSNR†      | 30.3148           | 29.1106           | 31.3099           | 25.4939           | 23.0277           |
| PSNR-HVS†  | 35.5379           | 31.7593           | 28.2595           | 25.2042           | 17.6504           |
| MAD L      | 26.9981           | 35.9255           | 56.5164           | 76.4642           | 103.1248          |
| PSNR-HMA†  | 42.8240           | 33.4513           | 29.2749           | 27.9710           | 21.2950           |
| FSIM†      | 0.9982            | 0.9960            | 0.9203            | 0.8708            | 0.8482            |
| FSIMc†     | 0.9975            | 0.9973            | 0.9681            | 0.9666            | 0.9999            |
| BSIM†      | 1.0000            | 0.9997            | 0.9996            | 0.9999            | 0.9956            |
| VSIT†      | 0.9982            | 0.9903            | 0.9772            | 0.9741            | 0.8838            |
| DSS†       | 0.9907            | 0.9601            | 0.8430            | 0.7589            | 0.3559            |
| OSS-SSIM†  | 0.9993            | 0.9670            | 0.9728            | 0.9774            | 0.7780            |
| VIFP†      | 0.8442            | 0.8197            | 0.4720            | 0.3825            | 0.1064            |
| GMSSD†     | 0.0062            | 0.0129            | 0.0682            | 0.0708            | 0.2652            |
| TCDL†      | 0.0145            | 0.0171            | 0.3108            | 0.3310            | 0.4972            |
| MDSL†      | 0.1331            | 0.1710            | 0.2699            | 0.2741            | 0.4006            |

Separating images into different regions has been used similarly in work [50]. However, the proposed IQA index is different from work [50] in the following three aspects. First, the PLRA index gets the training feature vector by concatenating the training feature vectors of the pair of images and then formulates the pairwise preference learning problem as a binary classifier to get relative quality scores. However, work [50] directly applies training features into a regression problem. Second, the regions are separated using different methods. The PLRA index calculates frequency information based on edge detection as high-frequency information is usually located around image edges, while work [50] divides the image into the rough and smooth regions according to gradient magnitude. Thirdly, the PLRA index extracts frequency-aware quality features by adapting a variety of existing IQA metrics, while work [50] extracts the most preferred structure feature based on rough and smooth regions.

Tables 3 and 4 present the scores of the distorted images shown in Figure 3 computed by the frequency-aware features. Among all the $F^{hm}_{m}$ scores measured in the high-frequency parts, 6 out of 75 scores are inconsistent with the MOSs. For $F^{l}_{l}$ scores measured in the low-frequency parts, 12 out of 75 scores are inconsistent with the MOSs.

3) FEATURE REPRESENTATION

For each distorted image in the training dataset, we combine $F^{h}$ and $F^{l}$ to compose a $k$-dimensional ($k = 30$) feature vector using the following formulae:

$$ F^{h}_{1} = \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} (E(i,j) S(i,j))}{\sum_{i=1}^{p} \sum_{j=1}^{q} E(i,j)}, \quad (3) $$

$$ F^{l}_{1} = \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} ((1 - E(i,j)) S(i,j))}{\sum_{i=1}^{p} \sum_{j=1}^{q} (1 - E(i,j))}, \quad (4) $$

where $S(i,j)$ is the quality score at pixel $(i,j)$ calculated by the SSIM index. $E(i,j)$ is the frequency value at pixel $(i,j)$ of the expanded edge map, and $p$ and $q$ are the height and width of the image, respectively.

VOLUME 8, 2020
a set of feature values, \( j \) denotes the index of selected feature and \( i \) is the number of the image in the training or testing dataset, \( X \). We concatenate their feature vectors to get the training \( F \) and testing data. Specifically, we adopt the following Map-Min-Max normalization method:

\[
Y = \begin{cases} 
\frac{(X - X_{\min})}{(X_{\max} - X_{\min})}, & \text{if } X_{\max} \neq X_{\min}, \\
1, & \text{if } X_{\max} = X_{\min}, 
\end{cases}
\] (5)

where \( X = \{ M_1, M_2, \ldots, M_n \} \), \( X \) is a set of feature values, \( j \) denotes the index of the image and \( n \) is the number of the image in the training or testing dataset, \( i \) denotes the index of selected feature and \( k \) is the dimension of the feature vector. \( X_{\max} \) and \( X_{\min} \) are the minimum and maximum values of set \( X \), \( Y \) is the set of normalized feature values.

### B. PAIRWISE PREFERENCE MODEL LEARNING

#### 1) PAIRWISE COMPARISON

The performance of IQA metrics is evaluated from three aspects, namely correlation, accuracy, and monotonicity. Monotonicity compares the rank order of objective assessment results to that of subjective assessment results, and it is regarded as the most critical aspect of metric performance.

In this article, we propose to evaluate the quality of a given image by comparing it to a certain number of other images to get a relative quality score. The comparison strategy adopted in this article is different from existing IQA metrics which directly evaluate the quality of a given distorted image.

We adopt the pairwise-based learning to rank methodology [34]–[37] to train a pairwise preference model. In the training process, we compare each pair of images in the training dataset. The feature vector of each comparison is composed using the feature vectors of the pair of images. Its corresponding label is obtained by comparing the subjective scores of the pair of images.

For each pair of images, we first obtain their feature vectors, and their corresponding MOSs or difference mean opinion scores (DMOSs). For example, take images \( A \) and \( B \), and their feature vectors \( F_A \) and \( F_B \). We concatenate their feature vectors to get the training feature vector \( CF_{A,B} = (F_A, F_B) \) and compute the training label \( CL_{A,B} = \{MOS_A > MOS_B:1 \mid 0 \} \) or \( CL_{A,B} = \{DMOS_A < DMOS_B:1 \mid 0 \} \). In this way, we formulate the pairwise preference model learning problem as a binary classifier and get a model \( M \).

#### 2) BALANCED COMPARISON STRATEGY

The comparison strategy that determines the composed pairs of images used in training can affect the performance and efficiency of the training of the preference model \( M \). To achieve high performance and efficiency, the comparison strategy needs to satisfy the following three constraints:

1) An image can not be compared with itself, which means there is no \( C(A, A) \).

2) Once an image \( A \) is compared with an image \( B \), image \( B \) cannot be compared with image \( A \) in other words, \( C(A, B) \) and \( C(B, A) \) cannot exist for the same training or testing event.

3) Each image is compared with the same number of other images. The numbers of each image appearing in the first position of the pair is the same and the numbers of each image appearing in the second position of the pair is also the same.

Inspired by the Latin square [51] method, we propose a balanced comparison strategy. We first divide all reference images into two groups. All distorted images derived from images in the first and second groups are used in training and testing, respectively. For a group of \( r \) reference images, we get a Latin square as shown in Figure 6(a).

\[
\begin{align*}
\text{CL} & = \begin{bmatrix} G_1 & G_2 & \cdots & G_4 & G_5 \\
G_6 & G_7 & \cdots & G_9 & G_{10} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
G_{11} & G_{12} & \cdots & G_{14} & G_{15} \\
G_{16} & G_{17} & \cdots & G_{18} & G_{19} \\
\end{bmatrix} \\
\text{Row number matrix} & = \begin{bmatrix} G_1 & G_2 & \cdots & G_4 & G_5 \\
G_6 & G_7 & \cdots & G_9 & G_{10} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
G_{11} & G_{12} & \cdots & G_{14} & G_{15} \\
G_{16} & G_{17} & \cdots & G_{18} & G_{19} \\
\end{bmatrix} \\
\text{Full comparison matrix} & = \begin{bmatrix} G_1 & G_2 & \cdots & G_4 & G_5 \\
G_6 & G_7 & \cdots & G_9 & G_{10} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
G_{11} & G_{12} & \cdots & G_{14} & G_{15} \\
G_{16} & G_{17} & \cdots & G_{18} & G_{19} \\
\end{bmatrix} \\
\text{Balanced comparison matrix} & = \begin{bmatrix} G_1 & G_2 & \cdots & G_4 & G_5 \\
G_6 & G_7 & \cdots & G_9 & G_{10} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
G_{11} & G_{12} & \cdots & G_{14} & G_{15} \\
G_{16} & G_{17} & \cdots & G_{18} & G_{19} \\
\end{bmatrix}
\end{align*}
\]

FIGURE 6. An example of composing a balanced comparison matrix.
In a Latin square of $r \times r$, each of $n$ different symbols occurs exactly once in each row and exactly once in each column. Then we assign $r$ symbols to $r$ rows in order and obtain an $r \times r$ array as shown in Figure 6(b). Without considering efficiency, a full comparison strategy shown in Figure 6(c) can be obtained by composing each element in the Latin square with its corresponding element in the row number matrix. For example, the last elements are composed as $(G_r, G_{r-1})$.

To satisfy constraints 1 and 2 for efficiency, we remove the first column and last $\left\lceil \frac{r-1}{2} \right\rceil$ columns. The resulting balanced comparison matrix satisfies all constraints and is shown in Figure 6(d).

Each symbol in the balanced comparison matrix indicates a reference image from which some distorted images are derived. For each composed pair, for example $(G_i, G_j)$, we compare all distorted images derived from reference image $G_i$ with all distorted images derived from reference image $G_j$.

As described in subsection II-B, the pairwise comparison has been gradually applied into objective image quality assessment metrics. The proposed method is different from the previous ones in terms of comparison strategies and feature processing. The main differences between the proposed PLRA index and the previous methods lie in the following three aspects. First, the proposed PLRA index selects the metrics based on different types of image features and human visual characteristics and extracts frequency-aware quality features based on the masking effects by adapting these metrics. Second, it gets the training feature vector by concatenating the training feature vectors of the pair of images, while most existing methods calculate the corresponding difference vector as the training features. Third, a balanced comparison strategy is proposed to improve the performance and efficiency of quality prediction.

**C. IMAGE QUALITY PREDICTION**

We use $n$-fold cross validation to carry out the training/learning and testing/prediction procedures. We first randomly divide the reference images into $n$ folds. All folds have the same or at least similar numbers of reference images. We then put each distorted image into the fold into which its reference image belongs. A pairwise preference model $M$ is first learned using $n-1$ folds of training data, and then model $M$ is used to compute the preference between any two images in the last fold of testing data.

In this article, we use the random forest classification (RFC) machine learning algorithm to learn the pairwise preference model $M$. We use the RFC implementation for MATLAB from Jaiantilal et al. proposed in [52].

The relative quality score of image $A$ is calculated based on the binary labels as follows:

$$ S_A = \sum_{C=1}^{T} P_{A,C}, $$

where $T$ donates the number of comparisons that image $A$ appears at the first position of the comparison, $T$ is same for all testing images. $P_{A,C}$ is the predicted binary label for image pairs $A$ and $C$ which can essentially reflect the quality of image $A$ relative to image $C$. $P_{A,C} = 1$ indicates that the quality of image $A$ is better than image $C$, $P_{A,C} = 0$ indicates that the quality of image $A$ is worse than image $C$. $S_A$ is the quality score of image $A$ and computed as the number of compared images whose quality are predicted worse than image $A$.

IV. EXPERIMENTS

In this section, we present the experimental results of the proposed PLRA index and comparisons with some state-of-the-art IQA metrics on four widely used IQA databases.

A. IQA DATABASES AND PERFORMANCE METRICS

In the experiments, we used the following four IQA databases.

The TID2008 database [13] shares the same reference images with the TID2013 database. It includes 17 types of distortions, and each distortion type has four intensity levels. The MOS is obtained from 838 ratings conducted by observers. The subjective evaluation score MOS of each image is within the range of $[0, 9]$.

The TID2013 database [14] is a widely used IQA database. It contains 25 reference images and 3000 distorted images. Among the 25 reference images, 24 are obtained from the Kodak database, and the remaining one is a synthesized image. The database has 24 types of distortion. The subjective evaluation experiment on the TID2013 database involves 971 participants from five different countries, namely, Finland, France, Italy, Ukraine, and the United States. The subjective evaluation score MOS of each image is within the range of $[0, 9]$.

The LIVE2005 database [12] contains 39 reference images and 866 distorted images. There are a total of six types of distortions at four to five different levels. The distortion types in LIVE2005 are JPEG compression, JPEG2000 compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blurring. A total of 16 non-expert observers participated in this subjective experiment. The database contains 5000 subjective ratings reported in the form of DMOS.

The LIVE2005 database [12] contains 92 reference images and 779 distorted images. A total of five types of distortion are included, namely, JPEG2000, JPEG, white noise in the RGB components, Gaussian blur in the RGB components, and fast fading Rayleigh. Approximately 20-29 observers have rated each image. Each distortion type is evaluated by various subjects using the same equipment and viewing conditions. The raw scores are converted to DMOS which are ranged $[1,100]$.

To evaluate the consistency between IQA metrics with subjective visual perception (MOS or DMOS), three commonly used performance metrics are adopted, namely, the Spearman rank correlation coefficient (SRCC), the Pearson linear
TABLE 5. Performance comparison among 22 IQA metrics and the proposed PLRA index on four databases. (The best three results are highlighted in red, green, blue and are formatted in boldface. The symbol ‘−’ indicates that the corresponding value and the source codes of these metrics were not available).

| Metric | LIVE2005 | CSIQ | TID2008 | TID2013 |
|--------|----------|------|---------|---------|
|        | SRCC     | PLCC | RMSE    | SRCC    | PLCC | RMSE    | SRCC | PLCC | RMSE |
| SSIM   | 0.908    | 0.865 | 16.132  | 0.876   | 0.861 | 0.133   | 0.775 | 0.773 | 0.851 |
| PSNR   | 0.876    | 0.865 | 13.730  | 0.806   | 0.781 | 0.164   | 0.553 | 0.533 | 1.135 |
| PSNR-HVS | 0.919   | 0.903 | 12.540  | 0.830   | 0.804 | 0.156   | 0.594 | 0.608 | 1.065 |
| MAD    | 0.967    | 0.968 | 6.907   | 0.947   | 0.950 | 0.082   | 0.834 | 0.831 | 0.747 |
| PSNR-HMA | 0.934   | 0.918 | 10.857  | 0.912   | 0.891 | 0.119   | 0.845 | 0.832 | 0.745 |
| FSIM   | 0.963    | 0.960 | 7.678   | 0.924   | 0.912 | 0.108   | 0.881 | 0.874 | 0.653 |
| FSIMc  | 0.965    | 0.890 | 12.036  | 0.931   | 0.902 | 0.112   | 0.884 | 0.860 | 0.683 |
| ESSIM  | 0.962    | 0.799 | 15.828  | 0.933   | 0.642 | 0.167   | 0.884 | 0.841 | 0.777 |
| VSI    | 0.952    | 0.944 | 8.682   | 0.942   | 0.928 | 0.098   | 0.898 | 0.876 | 0.647 |
| DSS    | 0.962    | 0.931 | 9.961   | 0.961   | 0.957 | 0.076   | 0.873 | 0.877 | 0.644 |
| OSS-SSIM | 0.933  | 0.752 | 17.509  | 0.886   | 0.822 | 0.147   | 0.772 | 0.758 | 0.874 |
| VIF    | 0.964    | 0.960 | 7.614   | 0.920   | 0.928 | 0.098   | 0.749 | 0.808 | 0.790 |
| GMSD   | 0.959    | 0.938 | 9.442   | 0.930   | 0.947 | 0.084   | 0.889 | 0.875 | 0.758 |
| iCID   | 0.938    | 0.933 | 9.799   | 0.922   | 0.902 | 0.113   | 0.757 | 0.770 | 0.857 |
| MDSI   | 0.967    | 0.966 | 7.079   | 0.957   | 0.953 | 0.080   | 0.921 | 0.916 | 0.538 |
| SPSIM  | 0.962    | 0.960 | 7.629   | 0.944   | 0.934 | 0.093   | 0.910 | 0.893 | 0.605 |
| SVDR   | 0.879    | 0.879 | 7.686   | 0.862   | 0.888 | 0.121   | 0.777 | 0.789 | 0.825 |
| CD-MMF | 0.981    | 0.980 | 5.413   | 0.967   | 0.968 | 0.067   | 0.942 | 0.948 | 0.429 |
| NMF    | 0.976    | 0.976 | 5.866   | 0.973   | 0.976 | 0.057   | 0.947 | 0.951 | 0.419 |
| WaDIQaM-FR | 0.970 | 0.980 | -      | 0.961   | 0.965 | -       | 0.947 | 0.951 | -    |
| DeepQA | 0.981    | 0.982 | -      | 0.961   | 0.965 | -       | 0.947 | 0.951 | -    |
| P2D-IT  | 0.973    | 0.966 | 10.048  | 0.968   | 0.974 | 0.082   | 0.919 | 0.908 | 0.799 |
| PLRA   | 0.978    | 0.978 | 5.676   | 0.976   | 0.979 | 0.053   | 0.955 | 0.958 | 0.386 |

correlation coefficient (PLCC), and the root mean square error (RMSE). SRCC is used to measure the prediction monotonicity of an objective IQA index. PLCC is used to measure the relevance between the subjective evaluation and objective scores after nonlinear regression [53]. Larger PLCC and SRCC values indicate a closer relation with human subjective evaluation. RMSE is used to assess the accuracy of the predictions after nonlinear regression. A smaller RMSE value indicates a superior correlation with human perception. For a fair comparison, PLCC and RMSE are computed using the objective scores after fitting through Equation 7 suggested by Sheikh et al. [53], as follows,

\[ f(x) = \beta_1 \times (1/2 - 1/(1 + e^{\beta_2 x + \beta_3})) + \beta_4 x + \beta_5, \]

where \( \beta_i, i = 1, 2, 3, 4, 5 \), are the parameters to be fitted. \( x \) is the objective score before fitting, and \( f(x) \) is the new score after fitting.

In the experiments, the 3-fold cross-validation method was used. Moreover, to achieve more stable and convincing performance, the cross-validation experiments were repeated 100 times, and the average values of each fold in terms of SRCC, PLCC, and RMSE are reported in the paper.

B. OVERALL PERFORMANCE OF THE PROPOSED PLRA INDEX

To validate the performance of the proposed PLRA index, we compared it with 22 state-of-the-art IQA metrics. Table 5 shows the comparisons of these 22 state-of-the-art IQA metrics on four databases: four of them are machine learning-based metrics, namely SVDR [29], CD-MMF [30], NMF [31], and IrSIM [39]; two of them are CNN-based metrics, namely, WaDIQaM-FR [32] and DeepQA [33]. The three best assessment results for each database are highlighted in red, green, and blue and are formatted in boldface. The scores of MAD, GMSD, iCID, and MDSI are negatively correlated with MOS, whereas those of the other metrics are negatively correlated with DMOS. Therefore, we reported the absolute values of SRCC. As shown in Table 5, the PLRA index outperformed the other IQA metrics. Specifically, the results of our PLRA index are all better than the other 22 metrics in terms of all performance indicators except for SRCC, PLCC, and RMSE on the LIVE2005 database.

C. PERFORMANCE ON 24 DISTORTION TYPES IN THE TID2013 DATABASE

To fully verify the proposed PLRA index for the prediction ability of each distortion type, we reorganized the distorted images in the TID2013 database into 24 datasets according to the 24 types of distortion. Each dataset contains 125 images with the same type of distortion. For these datasets, we performed 3-fold cross-validation experiments.

We experimented with PLRA index and 15 state-of-the-art IQA metrics on each distortion type in the TID2013 database. The SPSIM [28], four machine learning-based IQA metrics, and two CNN-based metrics were not compared in this subsection since the quality score of each image and the source codes of these metrics were not available. In Table 6, we show...
the SRCC value of each index for each distortion type. The best three values for each distortion type are formatted in boldface. We chose the performance metric SRCC because SRCC is not influenced by fitting and because prediction monotonicity is most important for IQA. As shown in Table 6, the proposed PLRA index showed more stable performance across different distortion types than 15 state-of-the-art IQA metrics. In order to further compare our PLRA index with the state-of-the-art IQA metrics, we also computed the average ranking score. For each subset, we ranked the 16 IQA metrics, including 15 representative metrics and our PLRA index, from 1 to 16. We then averaged the ranking score of each IQA metric on the 24 subset. The average ranking score is described as follows:

\[
R(m) = \frac{\sum_{d=1}^{D} r(d, m)}{D}
\]

where \(d\) donates the distortion type, \(D (= 24)\) donates the number of distortion types, \(m\) represents the IQA metric, and \(r(d, m)\) represents the rank of IQA metric \(m\) among the 16 IQA metrics in distortion type \(d\) according to performance metric SRCC. A small \(R(m)\) indicates good performance. The ranking score results of 16 IQA metrics are shown in Table 7. As shown in Table 7, the proposed PLRA index achieved the best average ranking score.

### D. CROSS-DATABASE TEST

In this section, we present the evaluation of the generalization capability and robustness of the proposed PLRA index using a cross-database test. Specifically, we trained the proposed PLRA using the images in one database and tested the trained PLRA model using the images in the other databases. Performance results in terms of SRCC of the cross-database test are presented in Table 8. We compared the proposed PLRA with four training-based IQA metrics, namely SVDR [29], CD-MMF [30], NMF [31], and WaDIQaM-FR [32]. Because the codes and cross-database test performance values of the other training-based metrics (IrSIM [39] and DeepQA [33]) that were compared in Table 5 were not available, we did not compare these in Table 8.

Owing to the distortion-type discrepancy between the testing and training sets, the cross-database test performance was generally inferior to the performance when training and testing used images in the same database as reported in Subsection IV-B. As indicated in Table 8, the proposed PLRA index performed better than the other existing competitive IQA metrics.

### E. VALIDATION OF THE PROPOSED FREQUENCY-AWARE FEATURES

In this section, we evaluate the effectiveness of the proposed frequency-aware features. As described in subsection III-A2, the SRCC value of each index for each distortion type. The best three values for each distortion type are formatted in boldface. The results of cross-database test. Best performance value of each test is indicated in boldface.

### TABLE 7. Average ranking score \(R\) across the 24 distortions of the proposed PLRA index and 15 IQA metrics.

| SRN | PSNR | PSNR-HVS | MAD | PSNR-HMA | PSIM | PRIMc | ESSIM | VSI | DSS | OSS-SSIM | VIP | GMSD | rCD | MDSI | PLRA |
|-----|------|---------|-----|---------|------|-------|-------|-----|-----|---------|-----|------|-----|------|-----|
| R(m) | 12.833 | 9.667 | 6.542 | 13.042 | 6.042 | 10.542 | 8.542 | 7.875 | 5.875 | 11.083 | 10.333 | 11.042 | 8.042 | 5.917 | 5.542 | 1.542 |

### TABLE 8. SRCC results of cross-database test. Best performance value of each test is indicated in boldface.

| Training database | Testing database | SVDR | CD-MMF | NMF | WaDIQaM-FR | PLRA |
|-------------------|------------------|------|--------|-----|-------------|-----|
| TID2013          | TID2008          | -    | -      | -   | 0.925       | 0.945 |
| CSIQ              | CSIQ             | -    | -      | -   | 0.914       | 0.927 |
| LIVE2005         | LIVE2005        | 0.843 | 0.912 | 0.926 | -            | 0.938 |
| TID2013          | TID2013         | 0.886 | 0.946 | 0.959 | -            | 0.965 |
| CSIQ              | CSIQ             | 0.755 | 0.853 | 0.861 | -            | 0.913 |
| LIVE2005         | LIVE2005        | 0.908 | 0.911 | 0.911 | -            | 0.964 |
| TID2013          | TID2013         | 0.742 | 0.898 | 0.899 | -            | 0.911 |
| CSIQ              | CSIQ             | 0.858 | 0.971 | 0.977 | 0.909        | 0.959 |
the proposed PLRA index extracts the frequency-aware features of distorted images, then uses the pairwise-based learning to rank method to train the model. We used different combinations of $F^c$, $F^h$, and $F^l$ to train some variations of the proposed PLRA index. The experimental results on the TID2008 database are presented in Table 9. The performance of the PLRA model with frequency-aware features $F^h$ and $F^l$ is better than the other variations.

| Dataset | Features | SRCC | PLCC | RMSE |
|---------|----------|------|------|------|
| TID2008 | $F^c$    | 0.944| 0.947| 0.430|
|         | $F^h$    | 0.951| 0.953| 0.402|
|         | $F^l$    | 0.950| 0.952| 0.408|
|         | $F^c, F^h$ | 0.953| 0.954| 0.398|
|         | $F^c, F^l$ | 0.953| 0.955| 0.395|
|         | $F^h, F^l$ (PLRA) | 0.955| 0.958| 0.386|
|         | $F^c, F^h, F^l$ | 0.954| 0.956| 0.393|

As the dimension of the feature vector increased from 30 to 45, the performance of the proposed PLRA model is slightly decreased. This is mainly due to two factors. First, we used the expanded edge map, a binary image to detect the high- and low-frequency parts of the image. This means that the high- and low-frequency features are complementary to the feature $F^c$ which is obtained using the whole image. Second, as the dimension of the feature vector increases, the proposed PLRA model becomes more difficult to train. The reason that the performance of $F^c$ was lower than both $F^h$ and $F^l$ is that different types of image distortions have different frequency sensitivity performances and then it often has an adverse effect on IQA when considering all frequency bands. The performance differences between the PLRA model and its variations demonstrate the effectiveness of the proposed frequency-aware features.

### F. VALIDATION OF FEATURE PROCESSING AND COMPARISON STRATEGY

As described in subsection III-B2, the proposed PLRA index is different from the previous ones in terms of comparison strategies and feature processing. The PLRA index gets the training feature vector by concatenating the training feature vectors of the pair of images. We replaced the training feature vector with the corresponding difference vector and obtained a variation of the PLRA index. The performance difference of the PLRA index and its variation demonstrates the effectiveness of the proposed feature processing. As indicated in Table 10, the performance of the proposed PLRA index is better than that of the variation with the corresponding difference vector.

Furthermore, to validate the effectiveness of the balanced comparison strategy, we also conducted comparison experiments using a random selection strategy. To make the experiment fair, we chose $N = 500$ as the number of compared images. The reason is that the number of compared images and the training set are close to the balanced comparison strategy. The performance of the balanced comparison strategy is better than that of the random selection strategy. From the experimental results shown in Table 10, we can conclude that the balanced comparison strategy can improve the performance and efficiency of quality prediction.

Inspire by Latin square, the full comparison matrix was proposed, then we removed the first column and last $\left\lfloor \frac{r-1}{2} \right\rfloor$ columns to obtain the balanced comparison matrix. It is also effective to use the triangular matrix without the diagonal elements (denoted as triangle matrix) in the composition of the image pairs. The proposed balanced comparison matrix is more feasible because image sampling is more uniform which can prevent the interference caused by the interactions between perceptual quality and high-level semantics [54]. To demonstrate the feasibility of the proposed balanced comparison matrix, we conducted the comparison experiment using the triangle matrix. As shown in Table 10, the performance of the proposed balanced comparison matrix is better than that of the triangle matrix.

| Dataset | Comparison strategy       | SRCC | PLCC | RMSE |
|---------|---------------------------|------|------|------|
| TID2008 | Difference vector         | 0.949| 0.949| 0.421|
|         | Random selection strategy | 0.945| 0.946| 0.448|
|         | Triangle matrix           | 0.953| 0.954| 0.401|
|         | Balanced comparison matrix| 0.955| 0.958| 0.386|

### G. INFLUENCE OF THE EDGE DETECTOR METHOD ON THE PROPOSED PLRA INDEX

In this article, we computed the frequency-aware features to incorporate the influence of the masking effect on IQA. Frequency information was calculated based on edge detection as high-frequency information is usually located around image edges and contours. To investigate the influence of the edge detector method on the proposed PLRA index, we experimented with the proposed PLRA model and a variation of edge detection methods. As described in Subsection III-A2, the frequency-aware quality features of the PLRA model is computed using the Canny edge detector method [48]. We replaced the edge detector with six other edge detectors, including Sobel [55], Robert [56], Prewitt [57], LoG [58], zero-cross [59], and EMBM [60]. We first computed the expanded edge map using Canny edge detector and six other edge detectors. As indicated in Figure 7, the results of seven edge detectors are very similar, and the Canny edge detector (Figure 7(b)) can detect the edges of clouds. The performance difference between the PLRA model and its variations demonstrates the effectiveness of the Canny edge detector. The experimental results on the TID2008 database are presented in Table 11. The performance of the PLRA model with the Canny edge detector is better than that of the variations with other edge detectors. We can conclude that
H. DISCUSSION

In this part, we further discuss the following problems:

1) FEATURE ABLATION STUDY

As described in subsection III-A1, we selected fifteen IQA metrics to construct image features, and these metrics can achieve a certain degree of consistency with subjective perception and complement each other. To justify the necessity of fifteen metrics that were chosen for feature construction, we conducted the feature ablation test. We gradually increased the number of fifteen selected metrics to train 15 variations of the proposed PLRA index. A simple greedy search algorithm [61] was used to select the features. Starting from an empty set, a feature is gradually added into the set and a new set obtained in each iteration until there are no more features that can be selected. The principle of feature selection in each iteration is the largest PLCC, since it represents the prediction accuracy of evaluation.

Table 12 shows how the performance of the proposed PLRA method changes with the increase of the number of the fifteen selected metrics in the TID2008 database. It can be observed that the speed of performance improvement is roughly divided into three stages. In the first stage, when the number of metrics goes from 1 to 5, there is a rapid performance improvement. Then, the performance improves moderately when the number of metrics goes from 5 to 11. Finally, when the number goes from 11 to 15, the performance of our method improves slightly and smoothly. Therefore, the performance of the proposed PLRA index improves with the increase of the number of metrics. We can conclude that the proposed PLRA method does not share limitations with the metrics used for feature extraction.

2) VALIDATION ON HIGH-QUALITY IMAGES

Since regression in high-quality images where the distortion distributes in high-frequency regions does not show consistent performance due to the contrast masking, it is hard to quantify the quality score on the high-quality image using current IQA metrics. In order to investigate the performance of the proposed PLRA index on the high-quality image, we first selected 500 images with the highest MOS values in the TID2008 database, and then we regarded these images as high-quality images. We experimented with the PLRA index and 15 state-of-the-art IQA metrics on high-quality images. The SPSIM [28] index, the four machine learning-based IQA metrics [29]–[31], [39], and the two CNN-based metrics [32], [33] were not compared in this subsection since the source codes of these metrics were not available. As indicated in Table 13, all IQA metrics achieved decreased performance on high-quality images which indicated the difficulty in evaluating the quality of high-quality images. However, the PLRA index is designed based on pairwise learning to rank which can simulate people’s ability to distinguish the small quality difference between two images, and the quality differences between high-quality images are usually small. We can conclude that the proposed PLRA index has better performance for high-quality images than the other state-of-the-art IQA metrics.

3) INFLUENCE OF N-FOLD CROSS-VALIDATION

To investigate the stability of the proposed PLRA model, we conducted experiments with the PLRA model using different n-fold cross-validations. Specifically, n − 1 folds of data in the TID2008 database were used for training and the corresponding remaining last fold of data was used for testing. As indicated in Table 14, four different n-fold cross-validations (2-fold, 3-fold, 4-fold, and 5-fold) were compared, the differences in terms of SRCC and PLCC are less than 0.008, and the differences in terms of RMSE are less than 0.029. Although the differences in terms of RMSE are larger than those in terms of SRCC and PLCC, the RMSE values achieved using 2-fold cross-validation are superior to those of all the comparison metrics reported in Table 5. Furthermore, the performance differences among 3-fold, 4-fold, and 5-fold cross-validations are tiny. From the experimental results in Table 14, we can conclude that the proposed PLRA model performs stably when using different combinations of images for training.

4) COMPARISON WITH LEARNING TO RANK METHODS

Because the pairwise comparison is a natural and effective way to obtain subjective image quality scores, it has recently received a large amount of attention within the
To further validate the effectiveness of the proposed PLRA index, we compared the proposed PLRA index with other ranking based methods described in subsection II-B. We listed the performance results in Table 15. The symbol “−” indicates that the corresponding value and the source codes of corresponding work were not available. The works [34]–[37] were not designed for IQA problem and the works [41], [42] were designed for retargeting images. The results of works [40], [45], were not included because the corresponding experiments were not carried out on the whole database. As indicated in Table 15, the proposed PLRA index performed better than other ranking based IQA metrics except for RankIQA [43] on the LIVE2005 database. Due to the few and common types of distortions on the LIVE2005 database, it is relatively easier to train the effective models for CNN-based methods, as a result, the performance of RankIQA is slightly better than the proposed PLRA index, the differences in performance metrics are less than 0.4%.

### 5) COMPUTATIONAL COMPLEXITY

In this article, we use pairwise learning to rank method to simulate people’s ability to distinguish a small quality difference between two images. Compared to mathematical-based objective IQA algorithms, computational complexity would be increased as a consequence of a large quantity of training feature vectors. However, the PLRA index is designed to achieve better consistency with subjective visual perception. From the experimental results shown in Table 5, we can conclude that the proposed PLRA index is more consistent with subjective scores than other state-of-the-art IQA metrics. Especially, as shown in Table 8, the proposed PLRA index shows a superior generalization capability and robustness, so that the proposed PLRA index has high practical value. Finally, subjective IQA is criticized for being time-consuming and tedious, and subjective scores are sometimes difficult to obtain. At this point, the time complexity of training the PLRA model is much less than subjective experiments, thus the proposed PLRA index can serve as an alternative to subjective evaluations.

### 6) APPLICATION FIELDS

We think the proposed PLRA index can be applied in the following three fields:

1. The proposed PLRA index can be used as an alternative to subjective IQA. Subjective IQA experiments are complex and the subjective scores are time- and labor-consuming to obtain. In contrast, the image scores of the PLRA index are much easier and faster to obtain. The PLRA index not only has high consistency with subjective scores, but also has a great generalization capability and robustness. When the subjective experiments are not available, the relative score obtained by the trained PLRA model can be used as an alternative to the absolute quality score.

2. The proposed PLRA index can be used as training labels in opinion-unaware NR-IQA models. Most existing NR-IQA methods are opinion-aware models that rely on examples of distorted images and corresponding human opinion scores, but obtaining human opinion scores can be time-consuming and expensive. To overcome this limitation, there has been an increasing trend towards constructing opinion-unaware NR-IQA methods without using human opinion scores. In [62]–[65], instead of training on human opinion scores, the authors used the predicted scores computed from high-performance FR-IQA methods as training labels. In this case, the proposed PLRA index is a good choice because it combines the advantages of current IQA metrics and has better consistency with subjective visual perception.
(3) The proposed PLRA index is effective for fine-grained distorted image quality assessment, while most existing IQA metrics show inferior performance on fine-grained IQA task. The distorted images in the existing databases are usually generated by corrupting pristine images with various distortions in coarse levels. Therefore, the IQA metrics may be inefficient on images with fine-grained quality differences. The FG-IQA database [66] is constructed to provide a benchmark for compressed image quality assessment with fine-grained distortion differences. The FG-IQA database contains 100 reference images with different real-world contents and 1200 distorted images for three target bitrates. At each target bitrate, each reference image is compressed by JPEG encoders with four optimized quantization tables, namely, JPEG default quantization table, uniform quantization table, optimized quantization table with the optimization goal of PSNR, and the optimized quantization table with the optimization goal of MSSSIM. Therefore, each reference image corresponds to four distorted images at the given bitrate, and there are 1200 distorted images in total for all three target bitrates.

We experimented the proposed PLRA index on the FG-IQA database. The FG-IQA database was divided into two randomly chosen subsets. Specifically, the distorted images corresponding to 80% of the reference images were used as the training set and distorted images corresponding to the remaining 20% of the reference images were used as the testing set. We trained the PLRA model using the images in the training set for all three target bitrates and tested the trained PLRA model using the images in the testing set. The experiments were repeated 100 times. Experimental results are presented in Table 16. We compared the proposed PLRA index with 15 state-of-the-art FR-IQA metrics, including PSNR [1], PSNR-HVS [17], VSI [22], SSIM [2], IW-SSIM [67], MSSSIM [68], FSIM [21], RFSIM [69], SR-SIM [70], UQI [71], GSM [72], GSMD [25], VIF [24], IFC [73], MAD [15], and WaDIAqM-FR [32]. The performance values of 15 IQA metrics were provided in the FG-IQA database paper [66]. The proposed PLRA index achieved better performance for most cases and performs stably with different bitrates.

V. CONCLUSION

In this article, we proposed a full-reference IQA metric based on the machine learning method, which exploring and combing the advantages of frequency-aware image features computed by adapting a variety of existing IQA metrics, to enhance the consistency between subjective and objective IQA results. Specifically, we adopted the pairwise-based learning to rank method to learn a pairwise preference model which could predict the relative image quality preference between any two input distorted images. We used the pairwise preference model to predict the relative quality among the images in the testing dataset and calculated the image quality based on the prediction results. The experimental results on some IQA databases showed that the proposed PLRA index achieved higher consistency with human subjective perception than the state-of-the-art IQA metrics.

REFERENCES

[1] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, “New full-reference quality metrics based on HVS,” in Proc. Int. Workshop Video Process. Qual. Metrics, vol. 4, 2006, pp. 1–4.
[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.
[3] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, Feb. 2016.
[4] Y. Niu, Y. Yang, W. Guo, and L. Lin, “Region-aware image denoising by exploring parameter preference,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 9, pp. 2433–2438, Sep. 2018.

[5] F. Chen, X. Zeng, and M. Wang, “Image denoising via local and non-local circulant similarity,” J. Vis. Commun. Image Represent., vol. 30, pp. 117–124, Jul. 2015.

[6] H. Zhang and V. M. Patel, “Densely connected pyramid dehazing network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3194–3203.

[7] H. Zhang and V. M. Patel, “Density-aware single image de-raining using a multi-stream dense network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 695–704.

[8] J. Park, J.-Y. Lee, D. Yoo, and I. S. Kweon, “Distort-and-recover: Color enhancement using deep reinforcement learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5928–5936.

[9] Q. Jiang, F. Shao, G. Jiang, M. Yu, and Z. Peng, “Three-dimensional visual comfort assessment via preference learning,” J. Electron. Imag., vol. 24, no. 4, pp. 1–12, 2015.

[10] Q. Jiang, F. Shao, W. Lin, and G. Jiang, “On predicting visual comfort of stereoscopic images: A learning to rank based approach,” IEEE Signal Process. Lett., vol. 23, no. 2, pp. 302–306, Feb. 2016.

[11] A. B. Watson, R. Borthwick, and M. Taylor, “Image quality and entropy masking,” Proc. SPIE, vol. 3016, pp. 70–72, Jun. 1997.

[12] H. Sheikh, (2005). LIVE Image Quality Assessment Database Release 2, [Online]. Available: http://live.ece.utexas.edu/research/quality

[13] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “TID2008—A database for evaluation of full-reference visual quality assessment metrics,” Adv. Mod. Radioelectron., vol. 10, no. 4, pp. 30–45, 2009.

[14] N. Ponomarenko, I. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, and F. Battisti, “Color image database TID2013: Peculiarities and preliminary results,” in Proc. Eur. Workshop Vis. Inf. Process., 2013, pp. 106–111.

[15] C. E. Larson and M. D. Chandler, “Most apparent distortion: Full-reference image quality assessment and the role of strategy,” J. Electron. Imag., vol. 19, no. 1, pp. 143–153, 2010.

[16] D. M. Chandler and S. S. Hemami, “VSNR: A wavelet-based visual signal-to-noise ratio for natural images,” IEEE Trans. Image Process., vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[17] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, “On between-coefficient contrast masking of DCT basis functions,” in Proc. 3rd Int. Workshop Video Process. Qual. Metrics Consum. Electron., vol. 4, 2007, pp. 1–4.

[18] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, M. Carli, “Modified image visual quality metrics for contrast change and mean shift accounting,” in Proc. CAD Symp. Microelectron., 2011, pp. 305–311.

[19] X. Zhang, X. Feng, W. Wang, and W. Xue, “Edge strength similarity for image quality assessment,” IEEE Signal Process. Lett., vol. 20, no. 4, pp. 319–322, Apr. 2013.

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

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

[22] L. Zhang, Y. Shen, and H. Li, “VSE: A visual saliency-induced index for perceptual image quality assessment,” IEEE Trans. Image Process., vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[23] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, “Image quality assessment based on DCT subband similarity,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2015, pp. 2105–2109.

[24] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” IEEE Trans. Image Process., vol. 15, no. 2, pp. 430–444, Feb. 2006.

[25] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” IEEE Trans. Image Process., vol. 23, no. 2, pp. 684–695, Feb. 2014.

[26] J. Preiss, F. Fernandes, and P. Urban, “Color-image quality assessment: From prediction to optimization,” IEEE Trans. Image Process., vol. 23, no. 3, pp. 78–1366, Mar. 2014.

[27] H. Z. Nafchi, A. Shahkoobai, R. Heidjam, and M. Cheriet, “Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator,” IEEE Access, vol. 4, pp. 5579–5590, 2016.
YUZHENG NIU (Member, IEEE) received the Ph.D. degree in computer science from the Shandong University of China, in 2010. She was a Postdoctoral Researcher with the Department of Computer Science, Portland State University, Portland, OR, USA. She is currently a Professor with the College of Mathematics and Computer Science, Fuzhou University, China. Her current research interests include computer vision, artificial intelligence, and multimedia.

WENZHONG GUO (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in communication and information systems from Fuzhou University, Fuzhou, China, in 2000, 2003, and 2010, respectively. He is currently a Full Professor with the College of Mathematics and Computer Science, Fuzhou University. He also leads the Network Computing and Intelligent Information Processing Laboratory, which is a key Laboratory of Fujian province, China. His research interests include mobile computing and evolutionary computation.

YIZE HUANG received the B.S. and M.S. degrees from the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China, in 2014 and 2017, respectively. His current research interests include image processing and computer vision.

YIQING SHI received the B.S. degree in electronic information engineering from Fuzhou University, Fuzhou, China, in 2016, where she is currently pursuing the Ph.D. degree in communication and information system. Her current research interests include pattern recognition, image processing, and computer vision.

JIAMEI ZHAN received the B.S. and M.S. degrees from the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China, in 2015 and 2018, respectively. His current research interests include image processing and computer vision.

* * *

[55] J. Matthews. (2002). An Introduction to Edge Detection: The Sobel Edge Detector. [Online]. Available: http://www.generation5.org/content/2002/int01.asp

[56] L. Roberts, “Machine perception of 3dimensional solids,” in Optical and Electro-Optical Information Processing, Cambridge, MA: USA: MIT Press, 1965.

[57] C. R. Gonzalez and E. R. Woods, Digital Image Processing, vol. 28, no. 4. Upper Saddle River, NJ, USA: Prentice-Hall, 2008, pp. 484–486.

[58] V. Torre and T. A. Poggio, “On edge detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-8, no. 2, pp. 147–163, Mar. 1986.

[59] S. Dellepiane, D. D. Giusto, S. Serpico, and G. Vernazza, “Automatic parameter computation for edge detection by the zero-crossing method,” in Proc. 12th Colloque sur le traitement du signal et des images, Jan. 1989, pp. 617–620.

[60] J. Guan, W. Zhang, J. Gu, and H. Ren, “No-reference blur assessment based on edge modeling,” J. Vis. Commun. Image Represent., vol. 29, pp. 1–7, May 2015.

[61] P. E. Black, “Greedy algorithm,” Dictionary Algorithms Data Struct., vol. 2, p. 62, Feb. 2005.

[62] P. Ye, J. Kumar, and D. Doermann, “Beyond human opinion scores: Blind image quality assessment based on synthetic scores,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 4241–4248.

[63] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, “Learning a blind quality evaluation engine of screen content images,” Neurocomputing, vol. 196, pp. 140–149, Jul. 2016.

[64] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, “Learning a no-reference quality assessment model of enhanced images with big data,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 4, pp. 1301–1313, Apr. 2018.

[65] J. Kim and S. Lee, “Fully deep blind image quality predictor,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 1, pp. 206–220, Feb. 2017.

[66] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, and W. Gao, “Fine-grained quality assessment for compressed images,” IEEE Trans. Image Process., vol. 28, no. 3, pp. 1163–1175, Mar. 2019.

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

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

[69] L. Zhang, L. Zhang, and X. Mou, “RFSIM: A feature based image quality assessment metric using riesz transforms,” in Proc. IEEE Int. Conf. Image Process., Sep. 2010, pp. 321–324.

[70] L. Zhang and H. Li, “SR SIM: A fast and high performance IQA index based on spectral residual,” in Proc. 19th IEEE Int. Conf. Image Process., Sep. 2012, pp. 1473–1476.

[71] Z. Wang and A. C. Bovik, “A universal image quality index,” IEEE Signal Process. Lett., vol. 9, no. 3, pp. 81–84, Mar. 2002.

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

[73] H. R. Sheikh, A. C. Bovik, and G. de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” IEEE Trans. Image Process., vol. 14, no. 12, pp. 2117–2128, Dec. 2005.