IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 12, 2020, accepted March 27, 2020, date of publication April 2, 2020, date of current version April 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984886

Long-Range Dependencies and High-Order Spatial Pooling for Deep Model-Based Full-Reference Image Quality Assessment

MENGYANG LIU[1,2], LAI-MAN PO[1], (Senior Member, IEEE), XUYUAN XU[2], KWOK WAI CHEUNG[3],(Member, IEEE), YUZHI ZHAO[1], KIN WAI LAU[1], AND CHANG ZHOU[1]

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong
[2]Tencent Video, Tencent Holdings Ltd., Shenzhen 518000, China
[3]School of Communication, The Hang Seng University of Hong Kong, Hong Kong

Corresponding author: Mengyang Liu (lmyleon2014@gmail.com)'

This work was supported by MF EXT-CityU Internal Funds for External Grant Schemes under Project 9678141.

ABSTRACT Deep Learning based image quality assessment (IQA) has been shown to greatly improve the quality score prediction accuracy of images with single distortion. However, because these models lack generalizability and the accuracy of multidistortion-based image data is relatively low, designing reliable IQA systems is still an open issue. In this paper, we propose to introduce long-range dependencies between local artifacts and high-order spatial pooling into a convolutional neural network (CNN) model to improve the performance and generalizability of the full-reference IQA (FR-IQA). This long-range dependencies model is based on the hypothesis that local apparent artifacts can affect the overall image quality. The proposed network architecture adopts a non-local means algorithm to establish connections between all positions in the deep feature space and uses the Minkowski function to improve the non-linearity of the spatial pooling. Based on this architecture, a robust FR-IQA system has been constructed and evaluated on three well-known single-distortion-based IQA databases (LIVE, CSIQ, and TID2013) and a multidistortion-based IQA database (MDID). Experimental results demonstrate that, compared with the latest FR-IQA systems, the proposed long-range dependencies-boosted CNN-based FR-IQA system can achieve state-of-the-art performance. A comprehensive cross-database evaluation also shows that the proposed system is sufficiently generalized between different databases and multidistortion-based image data is more useful for training robust image quality metrics.

INDEX TERMS Full-reference image quality assessment, quantization, long-range dependencies, convolutional neural networks, spatial pooling.

I. INTRODUCTION

As an important part of digital multimedia, digital images are now used in many aspects of life. With the rapid development of network technology, photographic equipment and online social media in the past three decades, the amount of digital image data has increased dramatically. It has put tremendous pressure on data storage and transmission. Image processing techniques such as compression, super-resolution, and image generation are needed to alleviate this pressure. It is also vital to have metrics or design guidelines for computationally evaluating the perceived quality of processed images of these processing techniques. Image quality assessment (IQA) methods can be divided into three categories based on the amount of priori information (original version of the processed or distorted image) obtained from the reference image. The full-reference (FR) method utilizes all of the information in the reference image, while the no-reference (NR) method does not consider priori information. The reduce-reference (RR) method includes an intermediate IQA method because only features extracted from the reference image are used in the algorithm.

There are several well-known FR-IQA metrics, including structural similarity (SSIM) [1] index and its extension [2], [3], Gradient Magnitude Similarity Deviation (GMSD) [4], and Visual Information Fidelity (VIF) [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Yuming Fang.

They are closely related to human perception and are evaluated based on traditional test databases such as LIVE [6]. These methods attempt to model the human visual system (HVS) with handcrafted features. As these FR-IQA algorithms are based on handcrafted features, they have very limited performance on newly released test databases with more complex distortions, such as multidistortion-based IQA database (MDID) [7] and TID2013 [8]. Since 2012, Convolutional Neural Networks (CNNs) have achieved excellent performance in many different image processing tasks such as image classification [9], object detection [10], and image enhancement [11]. These demonstrate the benefits of good representation in deep models. In 2014, Kang *et al.* [12] introduced the CNN-based structure to NR-IQA and established the first deep model-based system. Nowadays, many deep-model-based approaches [7], [13]–[17] have been upgraded to provide better performance on newly released IQA databases.

Compared to conventional handcrafted IQA methods, most deep-model-based metrics follow a top-down design process. The training process aims to reduce the difference between predicted quality values and human opinions. In [12] and [15], the average value of the predicted patch quality score or the global average pooling of feature maps is used to obtain the overall image quality score while [7] and [17] use a fully connected neural network layer after the feature map to output a single image quality score. In [14], the reliability map is predefined and different weights are provided for the output of the deep model to obtain the final output, instead of learning the weights from the training process. However, these methods may not pay sufficient attention to HVS modeling. Optimization goals for HVS modeling are added in [16] and [13]. In [16], a patch-based deep IQA model is proposed to predict individual patch quality score. The final output is the weighted average of these scores. The system proposed by [13] can predict the sensitivity map of the entire test image. Element-wise multiplication is then used to combine the objective error map and sensitivity map to generate a perceptual quality map. However, the two models have a common drawback that only a small range of spatial information is considered when predicting the value of each position on the generated quality map. The reliability score in [16] is generated from a single patch while the value of a single pixel on the sensitivity map in [13] is determined by the receptive field of the convolutional kernel.

In the saliency detection study of [18], it is found that the image perception of each local area is affected by the overall image content, and highly noticeable artifacts may be far apart. To make use of this HVS characteristic, conventional CNN architectures use large kernel sizes [10] or dilated convolutions [19] to increase the receptive field sizes. However, the number of learning parameters is significantly increased and the size of the receptive field is still limited.

In order to effectively use the findings in [18], we hypothesized that, in HVS, local appearance artifacts can affect overall image quality perception. This characteristic can be modeled as the long-range dependencies in IQA. Based on this hypothesis, CNN-based Long-Range Dependencies-Boosted (LRDB) model is proposed for improving the performance and generalizability of the FR-IQA. In LRDB, the intermediate of the network provides not only quality distribution from the small receptive field of the convolutional kernels but also the long-range dependencies among the image elements [20]. Each pixel on the generated image is determined by the pattern of all pixels in the input image. The proposed network then generates an attention map based on the entire feature space. Dot product of two feature maps [21], [22] is used to generate a square-shaped attention map. Long-range dependencies can therefore be established through connections formed between all locations in the feature map. Instead of using the conventional deep structure to construct the residual block [21], the sensitivity map is applied to the objective error map so that the effect of the generated sensitivity map can be directly applied to the final output.

On the other hand, a global average pooling or fully connected layer is usually placed at the end of the CNN model to link local features together for final classification or detection result. However, the representation ability is insufficient to perform complex regressions such as quality assessment task. Therefore, traditional spatial pooling methods that either consider all elements equally or the use handcrafted weighted map may limit the overall performance of the IQA systems. In order to improve the representation ability and stabilize the training process, we also proposed a high-order spatial pooling method that incorporates Minkowski function [23] into a fully connected layer at the end of LRDB model.

Basically, the proposed LRDB model is an end-to-end fully CNN, which is very flexible for images with different aspect ratios. In this study, we evaluated LRDB model on four well-known IQA databases: three single-distortion-based databases (LIVE [6], CSIQ [24], and TID2013 [8]) and one multiple-distortion-based database (MDID [7]). LRDB-based FR-IQA can achieve state-of-the-art performance on well-known and challenging CSIQ, TID2013 and MDID. In addition, LRDB model can be listed as one of the top three methods in the LIVE IQA database. It shows a significant improvement in performance when evaluating images with multiple distortions simultaneously. Since the performance of data-driven methods depends heavily on the data used for training, we also performed a comprehensive cross-database evaluation of each possible database pair. LRDB model shows sufficient generalizability in cross-database evaluation, which indicates that multidistortion-based image data are more critical for establishing practical IQA metric for single and multiple distorted images. Ablation studies have demonstrated the effectiveness of the long-range dependencies and Minkowski pooling strategy.

The rest of this paper is organized as follows. Section II first introduces the system architecture of the proposed FR-IQA system based on LRDB deep model and the details of each module. Section III presents extensive experiments

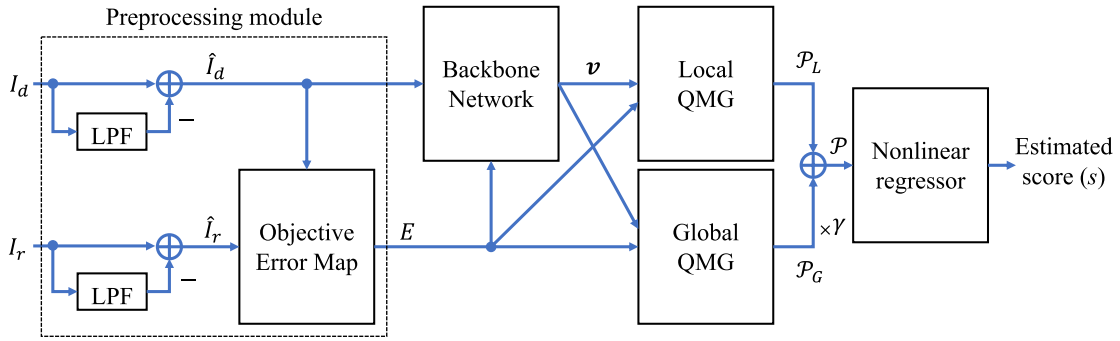


FIGURE 1. Overview basic architecture of the proposed LRDB-based FR-IQA system.

to evaluate the performance of the proposed LRDB FR-IQA and also includes computational complexity analysis and discussions. Finally, Section IV gives the conclusions with comments.

II. SYSTEM ARCHITECTURE

The overall architecture of the proposed FR-IQA system based on LRDB deep model is shown in Figure 1. The system is composed of five modules:

(1) Preprocessing module to generate a high-frequency distorted image ($\hat{I}_d$) and an objective error map (E) derived from input images;
(2) Shallow backbone network to extract preliminary features (v);
(3) Local Quality Map Generation (QMG) module to generate a quality map based on short-range dependencies;
(4) Global QMG module to generate a quality map based on long-range dependencies;
(5) Nonlinear regressor that integrates the multidegree Minkowski spatial pooling to output estimated image score using the elementwise added local and global quality maps with a trainable weight (γ).

Details of each module are provided in the following subsections.

A. PREPROCESSING MODULE

As a FR-IQA system, the proposed system requires two inputs: a distorted image (I_d) and a reference image (I_r). As HVS is more sensitive to differences in the high-frequency band [13], [25], high spatial-variance regions are more critical than low-frequency difference and uniform regions for IQA. In order to pay more attention to HVS modeling, this discovery is adopted in the preprocessing module to generate high-frequency images from the original distorted image and reference image. The high-frequency distorted image ($\hat{I}_d$) is one of the two inputs used by the backbone neural network for first stage feature map (v) extraction. To obtain the high-frequency image ($\hat{I}_d$) from the distorted image (I_d), we applied a highpass filter (HPF) to the original image. In practice, this is implemented by computing the difference between the distorted image and a low-passed distorted image (I_d^{low})

$$\hat{I}_d = I_d - H_L(I_d) = I_d - I_d^{low} \tag{1}$$

where $H_L(\cdot)$ is the transfer function of the lowpass filter implemented by $\times 16$ downsampling the input image and then upsampling back to its original size by interpolation. In addition, the high-passed image ($\hat{I}_d$) is normalized to the range of $[-1, 1]$ after subtraction computation and (1/255) scaling. Similarly, the high-frequency reference image ($\hat{I}_r$) of the reference image (I_r) is generated as follows.

$$\hat{I}_r = I_r - H_L(I_r) = I_r - I_r^{low} \tag{2}$$

Since HVS is also sensitive to differences in high-frequency distorted image and reference image, $\hat{I}_d$ and $\hat{I}_r$ are used to generate an objective error map (E) [13], which is another input of backbone convolutional neural network and is defined as

$$E = \frac{1}{\log\left(\frac{255^2}{\varepsilon}\right)} \log\left(\frac{1}{\left(\hat{I}_d - \hat{I}_r\right)^2 + \frac{\varepsilon}{255^2}}\right) \tag{3}$$

where ε is used to avoid the zero-denominator and is empirically set to 1. The main characteristic of this function is that it outputs a value of 1 for same pixel values of $\hat{I}_d$ and $\hat{I}_r$, and outputs small values for large differences. Thus, the output value of this function is ranging from a small negative value to 1. This normalized difference between $\hat{I}_d$ and $\hat{I}_r$ can stabilize the backbone neural network training process. It can also provide information related to human visual perception for image quality assessment. We can consider this objective error map E as an additional handcrafted HVS low-level feature for assisting the higher level feature extraction of the backbone network module. The high-frequency images and objective error maps of the training data can be calculated in advance to train the end-to-end LRDB network model, thereby making the training process very efficient.

B. BACKBONE NETWORK

The parallel inputs of the CNN backbone network are the low-level handcrafted features of high-frequency distorted

image ($\hat{I}_d$) and objective error map (E) from the outputs of the preprocessing module. The main goal of the backbone network is to extract intermediate-level feature (v) with assistance from the handcrafted features. These two inputs are first processed separately by CB_1 and CB_2 respectively, and then concatenated into a 32-channel feature map as inputs to cascade CB_3 and CB_4 for extracting higher level feature ($\boldsymbol{v}$). We believe that better feature can be effectively extracted with the assistance of the handcrafted feature of E. The structure of this backbone network is illustrated in Figure 2, where c is the output channel dimension of each block. This network consists of four convolutional blocks ($\{CB_i|i = 1, 2, 3, 4\}$), each contains two convolutional layers. A batch normalization layer (BN) [26] and a leaky rectified linear unit (LReLU) [27] activation function are added after each convolutional layer. The second convolution layer uses a stride of 2, which causes each CB to halve the size of the feature map. Details of the settings are provided in Table 1.

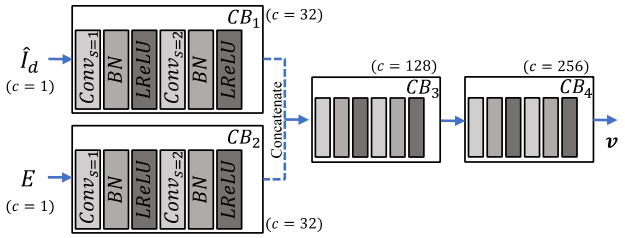


FIGURE 2. Structure of the backbone convolutional neural network.

C. LOCAL QUALITY MAP GENERATION

The backbone network is used to explore intermediate-level feature (v) from low-level handcrafted features of $\hat{I}_d$ and E. Two quality map generation (QMG) modules are used to explore short-range and long-range high-level features from the intermediate-level feature (v) as well as the handcrafted low-level feature of E. The local QMG module can be considered as a short-range QMG, which employs a QMG generation pipeline [13], [14] based on local dependencies [22] using regular convolutional layer structure. Each pixel on the output local quality map is linked to a small area of the input, which is the receptive field of the convolution kernel. The size of the receptive field is determined by the kernel size, which is usually set to 3×3. The convolutional layer in this module generates a sensitivity map (S_L) from the output of the backbone network, and then applies this sensitivity map to the objective error map using elementwise multiplication

$$\mathcal{P}_L = \hat{E} \odot S_L \tag{4}$$

where $\hat{E}$ is the downsampled objective error map by average pooling, and "$\odot$" is elementwise multiplication. The generated map is a biased quality map ($\mathcal{P}_L \in \mathbb{R}^{H \times W}$) that can effectively manipulate human perceptual sensitivity for IQA [13], [14].

TABLE 1. Configuration of the proposed network.

	Layers	Kernel Size (Stride)	output channel	Input
Backbone Network	CB_1	$\begin{bmatrix} 3\times3\ (1,1) \\ BN \\ LReLU \\ 3\times3\ (2,2) \\ BN \\ LReLU \end{bmatrix}$	32	$\hat{I}_d$
	CB_2	Δ*	Δ	$\mathcal{E}$
	Concat	-	-	CB_1, CB_2
	CB_3	Δ	128	Concat
	CB_4	Δ	256	CB_3
Local QMG	CB_5	Δ	512	CB_4
	CB_6	Δ	Δ	CB_5
	Up-sampling	× 4, bilinear	Δ	CB_6
	$conv_4$	$\begin{bmatrix} 3\times3\ (1,1) \\ BN \\ ReLU \end{bmatrix}$	1	Up-sampling
Global QMG	$conv_1, conv_2$	$1\times1\ (1,1)$	32	CB_4
	$conv_3$	Δ	Δ	$\mathcal{E}$
	Ave-pooling	$8\times8\ (8,8)$	Δ	$conv_3$
	$conv_5$	$\begin{bmatrix} 3\times3\ (1,1) \\ BN \\ ReLU \end{bmatrix}$	1	Ave-pooling
Regressor	Spatial pooling	-	-	$conv_4 + \gamma conv_5$
	fc_1	LReLU	8	Spatial pooling
	Dropout	0.1	-	fc_1
	fc_2	Linear	1	Dropout

* Δ: same as the upper one

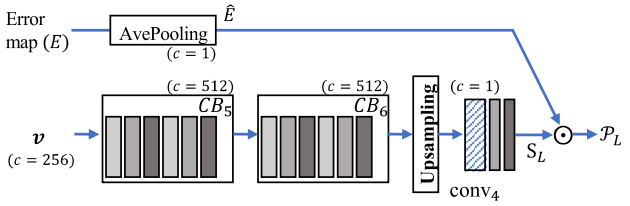


FIGURE 3. Structure of the local QMG module.

The structure of the local QMG module is shown in Figure 3, in which two more convolutional blocks (CB_5 and CB_6) are used to further shrink the size of the feature map, so that the output of CB_6 is 32 times smaller than $\hat{I}_d$ and E in terms of height and width. An upsampling layer and a single convolutional layer ($conv_4$) are applied after the CB_6. The upsampling layer enlarges the feature map four times using a bilinear interpolation function without extra parameters, and $conv_4$ outputs a single channel feature map that is considered as the sensitivity map. Average pooling reduces the dimension of the error map to fit the dimensions of the sensitivity map (S_L). The main advantage of this network structure is that it can alleviate the high variance problem of the conventional GMG architecture [13]. In addition, when the error mapping is directly reduced by 32 times, the upsampling operation can also avoid losing too much information.

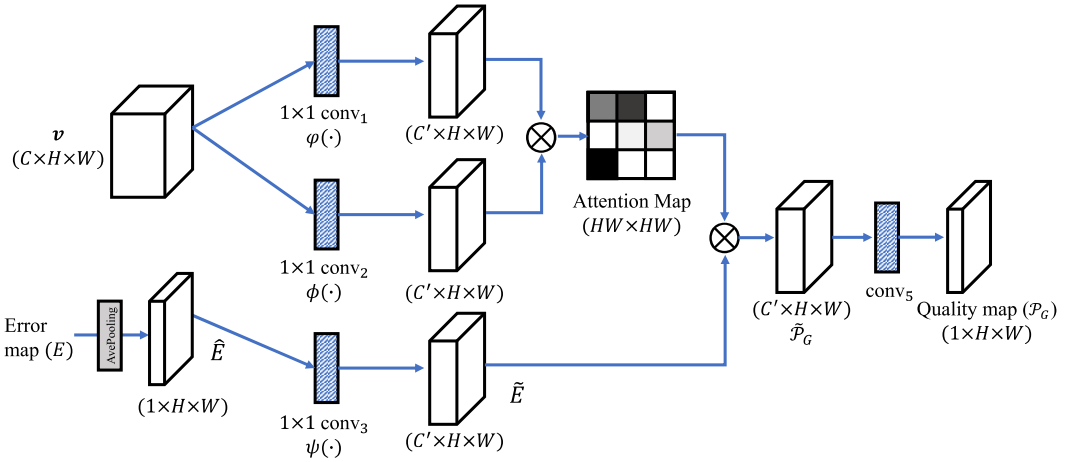


FIGURE 4. Structure of the global QMG module.

D. GLOBAL QUALITY MAP GENERATION

The local QMG can only provide quality distribution from the small receptive field of the convolutional kernels. In order to make use of the findings in [18], the proposed FR-IQA model adds a long-range branch (global QMG) to establish long-range dependencies in the network model. This is based on the hypothesis that local apparent artifacts can affect the overall image quality. The global QMG module also uses the intermediate-level feature (v) and the handcrafted low-level feature objective error map (E) to explore the high-level global quality feature map with long-range dependencies. The main idea was inspired by the NL-means algorithm for image denoising [20], which can be summarized as follows:

$$\boldsymbol{y}(i) = \frac{1}{\mathcal{C}(i)} \sum_{\forall j} w(i,j)\, \boldsymbol{s}(j) \tag{5}$$

$$\mathcal{C}(i) = \sum_{\forall j} w(i,j) \tag{6}$$

where i and j are the pixel location indexes; $\boldsymbol{y}$ is the output image, and $\boldsymbol{s}$ is the input image; $w(i, j)$ indicates the relationship (similarity) between pixels i and j; and $\mathcal{C}(i)$ is the normalization factor for each output location. According to (5), the value of each pixel on the output image is determined by the weighted average of the values of all pixels in the input image. The entire image is taken into account when calculating the output value of one pixel. In [20], Euclidean distance between these two pixels (i and j) determines the weight $w(i, j)$, but the method used to obtain the weight can be changed. Based on [21] and [22], in which this idea is applied to a deep model to generate a global (or long-range) feature map, we rewrite (5) as

$$\tilde{\mathcal{P}}_G = \tilde{E}\boldsymbol{A}^T \tag{7}$$

for the global QMG module in Figure 4, where $\tilde{\mathcal{P}}_G$ is the generated multichannel perceptual quality map; $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is an attention map generated from the neural network; and $\tilde{E}$ is the channel expanded objective error map. To implement matrix multiplication, we combine the width (W) and height (H) dimensions into column matrix with $N = H \times W$ dimensions. Thus, $\boldsymbol{v}$ belongs to $\mathbb{R}^{C\times N}$. In contrast to the nonlocal block in [21] and the self-attention block in [22], three embedding functions ($\varphi(\cdot)$, $\phi(\cdot)$, and $\psi(\cdot)$) are used in the proposed long-range QMG module with different inputs. The objective error map $\tilde{E} \in \mathbb{R}^{C'\times N}$ is generated by $\tilde{E} = \psi\left(\hat{\mathrm{E}}\right) = \boldsymbol{W}_{\psi}^{T}\hat{E}$ based on the objective error map (E) while $\varphi(\cdot)$ and $\phi(\cdot)$ use the intermediate feature ($\boldsymbol{v}$) from the backbone network as the input. The three embedding functions ($\varphi(\cdot)$, $\phi(\cdot)$, and $\psi(\cdot)$) are implemented as 1×1 convolutional layers, as shown in Figure 4, with $\boldsymbol{W}_{\varphi} \in \mathbb{R}^{C\times C'}$, $\boldsymbol{W}_{\phi} \in \mathbb{R}^{C\times C'}$, and $\boldsymbol{W}_{\psi} \in \mathbb{R}^{1\times C'}$ as learnable parameters of $\{conv_i | i = 1, 2, 3\}$ in which the bias parameters of each layer are included. We set $C' = C/8$ to reduce the computational cost of the dot product calculation.

Wang *et al.* [21] discussed various methods of generating $\boldsymbol{A}$, including simple Gaussian, embedded Gaussian, and dot product. We used a dot product in the global QMG design to generate $\boldsymbol{A}$ as

$$\boldsymbol{A} = \frac{1}{N}\varphi(\boldsymbol{v})^{T}\phi(\boldsymbol{v}) \tag{8}$$

where $\varphi(\boldsymbol{v}) = \boldsymbol{W}_{\varphi}^{T}\boldsymbol{v}$ and $\phi(\boldsymbol{v}) = \boldsymbol{W}_{\phi}^{T}\boldsymbol{v}$ are two embedding functions with the same input and N is the normalization factor used to maintain the stability of the system for inputs of different sizes. This formulation is simple and has low computation cost.

At the end of this branch, the height and width dimensions are extracted to ensure $\tilde{\mathcal{P}}_G \in \mathbb{R}^{C'\times H\times W}$. We applied another convolutional layer ($conv_5$) to merge all C' channels into one channel $\mathcal{P}_G = \theta_5(\tilde{\mathcal{P}}_G)$, where θ_5 represents the operation of $conv_5$. Therefore, $\mathcal{P}_G$ and $\mathcal{P}_L$ have the same dimensions. Table 1 provides details of the network settings.

In addition, we added a learnable parameter γ as a scaling parameter to $\mathcal{P}_G$ as shown in Figure 1. The final output

perceptual quality map is an elementwise addition of $\mathcal{P}_L$ and $\mathcal{P}_G$, where γ is always initialized as "0" to allow the parameter in the local QMG module to be updated first. This process can be represented as:

$$\mathcal{P} = \mathcal{P}_L + \gamma \mathcal{P}_G \tag{9}$$

E. NONLINEAR REGRESSION MODULE

The final module of the proposed LRDB-based FR-IQA system is a nonlinear regressor that integrates a multidegree Minkowski spatial pooling into the output estimated image score. The local (short-range) and global (long-range) quality maps are merged by trainable weight (γ) elementwise addition. The design is mainly based on the spatial pooling for combining a set of local quality scores into one quality score, which plays a vital role of IQA algorithm. The Minkowski function, which is commonly used in conventional image quality assessment methods of spatial pooling, is expressed as

$$\tilde{q}_\beta = \left(\frac{1}{n \times c} \sum_{\forall i} \sum_{\forall j} \mathcal{P}_i (j)^\beta \right)^{1/\beta} \tag{10}$$

where i and j represent the channel index and the pixel location on the feature map, respectively; β is a constant exponent, usually between 1 and 4; n is the number of pixels on each $\mathcal{P}_i$; and c is the channel number.

Alternatively, global average pooling (GAP) has been widely adopted in CNN-based IQA algorithms because it makes the network a fully convolutional structure that can flexibly adapt to various input dimensions. Most of the existing deep IQA indicators [13], [14] use GAP as the spatial pooling function. When β is set to 1 and $c = 1$, the GAP of a single channel is a special case of Minkowski pooling. The max-pooling is another common method of combining local scores into one output quality score, which is a special case of Minkowski pooling when the exponent is infinite. Compared to GAP, only the largest value in max-pooling contributes to the final output. This eliminates the effect of all other values that appear on the quality map. If the goal of the system is to generate a perceptual quality map for an image, we tend to set the channel of $\mathcal{P}$ to 1. This also saves parameters compared to the case when $\mathcal{P}$ is multichannel. In this case, GAP will output a single score, and all values on the quality map will make the same contribution to this score. Although a fully connected layer is applied with multiple nodes in the hidden layer after GAP, the nonlinearity of the regressor is insufficient. In this paper, we propose a two-layer regressor that combines Minkowski pooling with various β values at the end of the system. The structure of the proposed regressor is shown in Figure 5.

To isolate the gradient of each location on the perceptual quality map, which reduces the computational complexity of the gradient calculation, we consider the β^{th} power of the output of Minkowski pooling as

$$q_\beta = \tilde{q}_\beta^\beta = f_\beta (\boldsymbol{X}) = \frac{1}{N} \sum_{\forall i} x_i^\beta \tag{11}$$

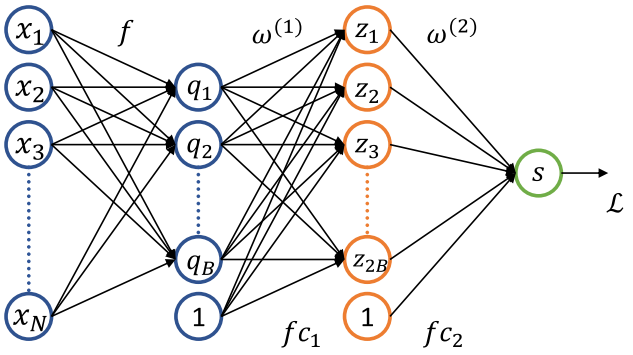


FIGURE 5. Structure of the proposed regressor.

where $\boldsymbol{X} = \{x_1, x_2, \ldots, x_N\}$ represents all possible locations on the perceptual quality map that have only one channel. The superscript (β) denotes the exponent, and the subscript (β) represents the selection of the Minkowski function. In this way, we can formulate the feedforward regressor as

$$s = \sum_{j=1}^{2B} \omega_{1,j}^{(2)} h \left(\sum_{i=1}^{B} \omega_{j,i}^{(1)} f_i (\boldsymbol{X}) + \omega_{j,0}^{(1)} \right) + \omega_{1,0}^{(2)} \tag{12}$$

where B is the total number of Minkowski exponents involved; the superscripts (1) and (2) indicate the corresponding parameters of the first and second "layer" in the regressor; ω represents the learnable parameters; and $h(\cdot)$ is the LReLU nonlinear activation function. s is the final predicted value. During the training process, the gradient of each layer is crucial for backpropagation. Therefore, to reveal the different effects of the built-in spatial pooling and traditional GAP, the gradient of $f_\beta (\boldsymbol{X})$ is defined as

$$\frac{\partial f_\beta (\boldsymbol{X})}{\partial x_i} = \frac{\partial}{\partial x_i} \left(\frac{1}{N} \sum_{\forall j} x_j^\beta \right) = \frac{1}{N} \beta x_j^{\beta - 1} \delta \text{ (i-j)} \tag{13}$$

$$\delta (x) = \begin{cases} 1, & x = 0 \\ 0, & other \end{cases} \tag{14}$$

When $\beta = 1$, Eq. (13) becomes a gradient of GAP, which gives the same gradient value (1/N) for each element on the feature map. However, when $\beta > 1$, for the gradient value, the actual value of the element (x_j) is considered. For gradient-based parameter updates, the parameters corresponding to the higher output should change drastically when the activation value is greater than 1 and modestly when the value is less than 1. In this way, each element of the feature map contributes unequally to the final output, which can be considered a regularization term for the deep model and improves the nonlinearity of the regression. Thus, the overall performance is boosted. Mean-square-error (MSE) is chosen as the loss function, which is represented as

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left\| s_{(i)} - g_{(i)} \right\|_2^2 + \frac{\lambda}{2} \| \boldsymbol{W} \|^2 \tag{15}$$

where g is the corresponding ground truth label, N denotes the number of samples in a batch, and $\boldsymbol{W}$ represents all trainable parameters of the network.

F. TRAINING PROCESS

The detailed configuration of each module in the proposed LRDB-based FR-IQA network is listed in Table 1. The structures of the four modules (Backbone Network, Local QMG, Global QMG and Regressor) are given in Figure 2 to 5. Based on these configurations, the proposed network requires approximately 7×10^6 trainable parameters. This is a fully convolutional network and is flexible for various input sizes. In this study, we avoided the patch generation process to significantly improve the assessment speed. However, we still used the patch of the image to train the network. Large patch size was chosen because the artifact distribution in the distorted image was not uniform and we intended to retain most of the artifacts; training samples have been enhanced using this tactic. During each epoch, each image in the training set was used to generate one patch, but the patch location was randomly selected. All training image samples were used once in one epoch. Additionally, we added a random horizontal flip to further increase the amount of training data. Accordingly, the training patches for different epochs were different, which led to rapid epoch iterations but more epochs of convergence. In our experiments, the training process usually stopped at approximately 3000 epochs. When we sampled each image to eight patches and expanded the training set twice by horizontal flipping, the iteration time was equal to approximately 200 epochs. After training, we used a variety of raw image sizes to test the trained model.

We chose the ADAM optimizer [28] to update the trainable weights in the proposed network in every iteration. We set $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$ following the recommended settings in [28]. The learning rate was 5×10^{-4}. To prevent overfitting, we added L2-norm to the weights as the regularization term whose weight (λ) was set to 5×10^{-3}. Following the standard protocol, the reference images were split into two subsets. One subset which contains 80% of the reference images was used for training. Another subset which contains the remaining 20% of the reference images was used for testing. The corresponding distorted images were then divided into non-overlapping training and testing sets. This procedure was conducted 10 times for the ablation study and 20 times for overall performance comparison. The average value was used to eliminate the performance bias caused by the reference image. All experiments were implemented using PyTorch [29] on a GTX 1080Ti GPU card.

III. EXPERIMENTS AND RESULTS

In this section, we describe the experiments used to evaluate the proposed LRDB model on four well-known databases. We carefully compared our method with other well-known IQA metrics in terms of overall performance, distortion-specific performance, and cross-database performance. We also conducted the ablation study to show that each module of the proposed deep model can positively contribute to overall performance. We extracted the $\mathcal{P}_L$, $\mathcal{P}_G$, and $\mathbf{A}$ from the middle layer of the proposed network to demonstrate the model's ability to predict perceptual quality maps and to generate the attention map for quality assessment. Pearson's Linear Correlation Coefficient (LCC) and the Spearman Rank-Order Correlation Coefficient (SROCC) were used to numerically assess performance. They are defined as

$$LCC\,(\boldsymbol{s},\boldsymbol{g}) = \frac{\bar{\boldsymbol{g}} \cdot \bar{\boldsymbol{s}}^T}{\sqrt{\bar{\boldsymbol{s}} \cdot \bar{\boldsymbol{s}}^T \cdot \bar{\boldsymbol{g}} \cdot \bar{\boldsymbol{g}}^T}} \tag{16}$$

$$SROCC\,(\boldsymbol{s},\boldsymbol{g}) = 1 - \frac{6\sum_{i=1}^{n}\left(rg\,(s_i) - rg\,(g_i)\right)^2}{n\left(n^2-1\right)} \tag{17}$$

where $\boldsymbol{s} = [s_1, s_2, s_3, \ldots, s_n]$ is the predicted score vector, and $\boldsymbol{g} = [g_1, g_2, g_3, \ldots, g_n]$ is the corresponding ground truth vector with n evaluation samples. $\bar{\boldsymbol{s}} = \boldsymbol{s} - \frac{1}{n}\sum_{i=1}^{n} s_i$ and $\bar{\boldsymbol{g}} = \boldsymbol{g} - \frac{1}{n}\sum_{i=1}^{n} g_i$. The operator $\mathrm{rg}(\cdot)$ obtains the rank of the observations.

A. DATASETS

Three widely used IQA databases, namely LIVE [6], TID2013 [8], and CSIQ [24], as well as the well-known multiple distorted image database of MDID [7], were used for our evaluation. The LIVE [6] IQA database consists of 29 reference images and 982 distorted images with five types of distortion: white Gaussian noise (WHIT), Gaussian blur (BLUR), Rayleigh fast-fading channel distortion (FAST), JPEG, and JPEG2000 (JP2K). This database uses the differential mean opinion score (DMOS) as the quality score. The size of the image varies. The range of DMOS is [0, 100], where a higher value represents lower perceptual image quality. To maintain the same label range in all databases, we normalized the DMOS to [0, 10] by dividing the original label by 10.

The TID2013 [8] IQA database is an extended version of the TID2008 [30] IQA database. The TID2013 database contains 25 reference images and 3000 distorted images with 24 types of distortion at 5 levels of distortion. These types are divided into six overlapping subsets: ''Noise'' contains a conventional noise type; ''Actual'' contains the common type of distortion in practice; ''Simple'' is a standard type of distortion; ''Exotic'' contains difficult types of distortion for quality metrics; ''New'' refers to new types introduced by this database; and ''Color'' is color-based distortion. The detailed content of each subset can be found in the original paper [8]. Various and complete distortion type causes TID2013 to be a very challenging IQA database. The dimension of all images in this database is 512×384. The range of the label is [0, 10].

The CSIQ [24] IQA database contains 30 reference images and 866 distorted images of six types: contrast change (CONT), Gaussian white noise (AWGN), Gaussian blur (BLUR), JPEG compression, JPEG2000 (JP2K) compression, and additive pink Gaussian noise (APGN). A total

TABLE 2. Performance comparison with other IQA metrics on different databases.

		LIVE IQA		CSIQ		TID2013		MDID	
		LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC
FR	PSNR	0.872	0.876	0.800	0.806	0.706	0.636	0.661	0.639
	SSIM	0.945	0.948	0.861	0.876	0.691	0.637	0.743	0.715
	IW-SSIM	0.952	0.957	0.914	0.921	-	-	0.899	0.892
	DOG-SSIM	0.963	0.961	0.954	0.943	0.919	0.907	-	-
	VIF	0.960	0.963	0.928	0.920	0.772	0.677	0.934	0.928
	VSI	0.948	0.952	0.928	0.942	0.900	0.897	-	-
	MAD	0.968	0.967	0.950	0.947	0.827	0.780	-	-
	GMSD	0.960	0.960	0.954	0.957	0.859	0.804	0.886	0.873
	FSIM(c)	0.961	0.960	0.919	0.931	0.877	0.851	0.899	0.896
	FR-DCNN	0.966	0.963	0.943	0.954	0.934	0.926	-	-
	WaDIQaM	0.980	0.970	-	-	0.946	0.940	-	-
	DeepQA	0.982	**0.981**	0.965	0.961	0.947	0.939	-	-
	LRDB (proposed)	0.982	0.981	**0.972**	**0.972**	**0.949**	**0.947**	**0.956**	**0.956**
NR	DIQA	0.977	0.975	0.915	0.884	0.850	0.825	-	-
	PS-NRIQA	**0.987**	0.976	-	-	0.890	0.878	-	-
	BLIINDS-II	0.916	0.912	0.832	0.780	0.628	0.536	0.603	0.497
	DIIVINE	0.922	0.912	0.882	0.864	0.768	0.715	0.591	0.489
	dipIQ	0.957	0.958	0.949	0.930	0.877	0.894	-	-

of 5000 subjects provided the quality score, and across-image ratings were obtained to "realign" the ratings. The DMOS score is in the range [0, 1]. We linearly enlarged the DMOS 10 times in our experiments.

The MDID [7] IQA database is a recently proposed database and represents a new challenge for IQA. The distorted images in this database have more than one type of distortion. In total, the images in the database have five types of distortion: Gaussian noise, Gaussian blur, contrast change, JPEG, and JPEG2000. Each type has four levels of distortion. Each distorted image has a random combination f distortion types and levels. The database contains 20 reference images and 1600 distorted images. It uses Pair Comparison Sorting (PCS) to label image quality. The dimension of all images is 512×512, and the range of the label is [0, 10].

B. PERFORMANCE COMPARISON

The evaluations described in this subsection were based on the full proposed model and the aforementioned training protocol. All presented values were based on the standard experiment protocol. All result provided in this section is the average performance of the 20-times reference image-based training-testing split. We compared the performance of the proposed model with that of other latest FR- and NR-IQA methods on four commonly used IQA databases in terms of LCC and SROCC. The results are presented in Table 2. For FR-IQA, PSNR, SSIM and its successors (IW-SSIM [3] and DOG-SSIM [31]), handcrafted feature-based methods (VIF [5], VSI [32], MAD [24], GMSD [4], and FSIM(c) [11]) and recently proposed deep model-based IQA metrics (FR-DCNN [33], WaDIQaM [16], and DeepQA [13]) are included. The compared NR-IQA metrics contain DIQA [14], BLIINDS-II [34], DIIVINE [35], PS-NRIQA [36], and dipIQ [7].

We found that the performance of the proposed system is superior to that of other latest methods on the CSIQ, TID2013, and MDID IQA databases, except for DeepQA [13] for SROCC and PS-NRIQA [36] for LCC on the LIVE IQA database. On LIVE, LRDB yields the second highest performance for both LCC and SROCC. The proposed method clearly shows improved performance on both the CSIQ and TID2013 databases. On the MDID database, a decrease in performance of many handcrafted feature-based methods is noticeable due to the complex combination of multiple types of image distortion. However, the proposed system can maintain high-level performance on other single distortion-based databases.

To provide an in-depth evaluation of the proposed method, we assessed the performance of the proposed method with specific types of distortion on the LIVE IQA database and with different subsets of the TID2013 IQA database. Table 3 shows that the proposed method yields the optimal performance for all types of distortion on LIVE, except for JPEG compression loss. The block effect of JPEG compression ignores the high frequency information provided by the image and generates new high frequency noise at the border of the block. Consequently, the generated error map contains a confusing pattern created by the image normalization process. When training a nondistortion-specific method, this confusing pattern degrades performance. Table 4 shows that the proposed method can deal more effectively with difficult, newly proposed, and color-based distortion than other methods, but performance degrades moderately for the conventional type of distortion. Notably, the proposed method utilizes the grayscale image as the input but performs well on the color-based distortion subset of the TID2013 IQA database. We found that the objective error map can still reveal differences in color for common color-based distortion, even if based on the grayscale image. The generation

TABLE 3. Performance comparison for individual types of distortion on the LIVE IQA database.

	JP2K	JPEG	WN	BLUR	FF
PSNR	0.895	0.881	0.985	0.782	0.891
SSIM	0.961	0.972	0.969	0.952	0.955
VIF	0.969	**0.984**	0.985	0.972	0.965
GMSD	0.968	0.973	0.974	0.957	0.942
FSIM(c)	0.972	0.979	0.971	0.968	0.950
DeepQA	0.970	0.978	0.988	0.971	0.968
LRDB (proposed)	**0.976**	0.966	**0.989**	**0.980**	**0.973**

TABLE 4. Performance comparison for new individual types of distortion on the LIVE IQA database.

	Noise	Actual	Simple	Exotic	New	Color
PSNR	0.822	0.825	0.913	0.597	0.618	0.535
SSIM	0.757	0.788	0.837	0.632	0.579	0.505
FSIM(c)	0.902	0.915	0.947	0.841	0.788	0.775
DOG-SSIM	0.922	0.933	0.959	0.889	0.908	0.911
WaDIQaM	**0.969**	**0.970**	**0.971**	0.925	0.941	0.934
LRDB	0.961	0.966	0.962	**0.936**	**0.947**	**0.946**

of the proposed method by the attention map is dependent on both the error map and distorted image, which are robust to color-based distortion. However, a well-defined distortion, which can leave the grayscale version of the image unchanged, may cause the proposed metric to fail.

C. CROSS-DATASET EVALUATION

We evaluated the proposed system in every train-test combination of the datasets used in this work (a total of 12 combinations). The full database was used for both training and testing. Table 5 provides a subset of evaluation that includes the results for models trained on the LIVE database and tested on the TID2013 (LIVE-TID2013) and CSIQ (LIVE-CSIQ) databases, on the TID2013 database and tested on the LIVE database (TID2013-LIVE), and the CSIQ database (TID2013-CSIQ) for comparison with the other two methods. Because a large scope remains for comparing existing NR-IQA methods with the FR-IQA methods in cross-database evaluation [16], we did not compare NR-IQA methods with our proposed method. LRDB-IQA shows superior generalization capability than the other two methods when it was trained on a large and complex database and tested on a smaller database. It also cannot perform well for unseen types of distortion, which led to the poor results in the LIVE-TID2013 evaluation. By contrast, the proposed model can obtain sufficient generalizability from a large database, which yielded superior results in the TID2013-LIVE and TID2013-CSIQ evaluations.

TABLE 5. Performance comparison in cross-database evaluation (in SROCC).

Trained on:	LIVE		TID2013	
Tested on:	TID2013	CSIQ	CSIQ	LIVE
DOG-SSIM	0.751	**0.914**	0.925	0.948
WaDIQaM-FR	**0.751**	0.909	0.931	0.936
LRDB (Proposed)	0.735	0.907	**0.938**	**0.952**

TABLE 6. Complete cross-database evaluation (in SROCC).

Test\Train	LIVE	TID2013	CSIQ	MDID
LIVE	-	0.952	0.965	0.966
TID2013	0.735	-	0.764	0.753
CSIQ	0.907	0.938	-	0.924
MDID	0.789	0.860	0.870	-

Table 6 presents the complete evaluation results, and Figure 6 provides the corresponding scatter plot. Taking the LIVE and CSIQ databases as the test databases, we provided the SROCC value for each specific type of distortion. The meaning of the abbreviations for types of distortion can be found in section III.A. However, when the TID2013 database is the test database, we provide the SROCC values of four common types of distortion (AGN, BLUR, JPEG, and JP2K) and samples of other types of distortion (Others). As the samples in the MDID database do not have a specific type of distortion, all samples provide the same color and shape. Although the types of distortion in the MDID database are included in the other three databases, the model trained on a single distortion-based database cannot effectively assess the quality of an image with multiple distortions. This is apparent in the performance for models trained on all other databases and tested on the MDID database. The single distortion-based IQA database cannot therefore be an effective database for assessing the quality of multiple-distorted images, which are more common in practice. Conversely, models trained on the MDID database perform similarly to models trained on the TID2013 database on LIVE and CSIQ. However, the scale of MDID is only half that of TID2013. We therefore conclude that the multiple-distortion-based database is effective for improving the learning-based IQA method in practice.

D. ABLATION STUDY

1) EFFECTS OF THE ERROR MAP INPUT

In this section, the effects of including the error map (E) in the proposed LRDB-IQA method were evaluated. We built a new network that only uses $\hat{I}_d$ as input; thus, CB_2 was removed in this high-frequency distorted image input network structure. This network was trained 10 times using the LIVE and the TID2013 IQA databases. The other experimental settings were the same as those presented in section II-F. Table 7 provides the average and the standard deviation of LCC and SROCC values for the trained model on test sets. We observed that performance degraded on both databases compared with that on the proposed LRDB-IQA using both the error map and the high-frequency distorted image as

TABLE 7. Effects of the error map input.

	LIVE		TID2013	
	LCC	SROCC	LCC	SROCC
$\hat{I}_d$ only	0.969±.013	0.969±.008	0.827±.035	0.810±.035
E and $\hat{I}_d$	**0.985**±.000	**0.978**±.000	**0.952**±.007	**0.948**±.006

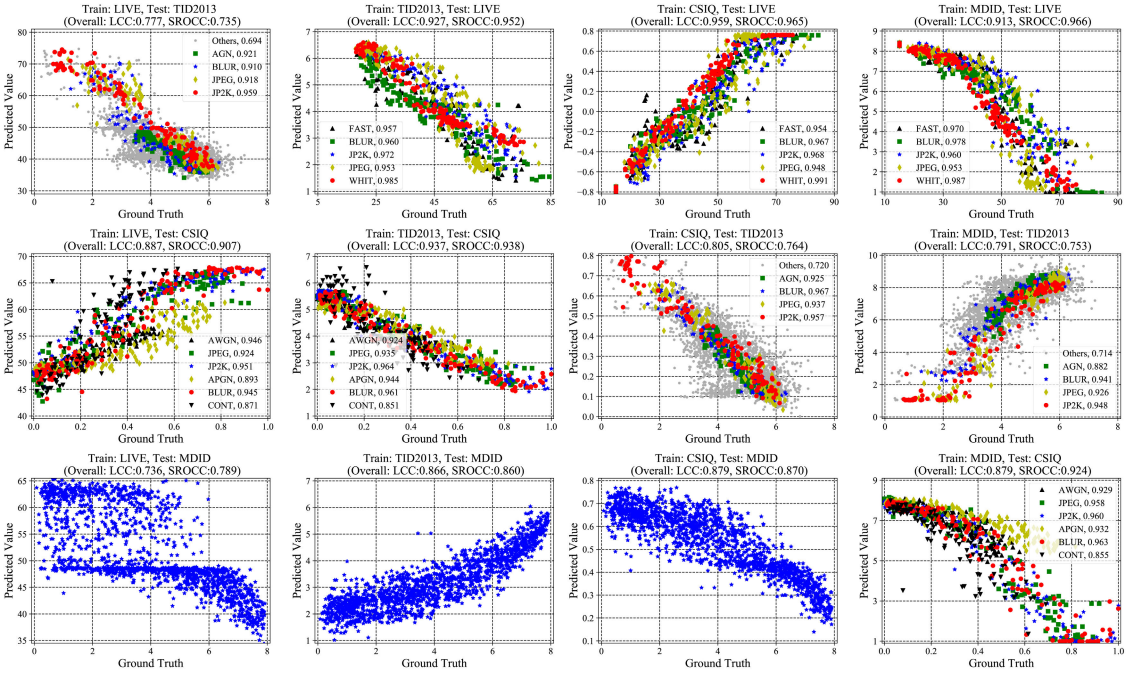


FIGURE 6. Scatter plots of predicted quality scores from the model trained on one database against the ground truth scores of anther database. Different types of distortion are represented by different color and shapes.

TABLE 8. Evaluation of training patch size effectiveness.

Patch size	LIVE		TID2013		CSIQ		MDID	
(H = W)	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC
96	0.986	0.977	0.914	0.912	0.977	0.976	0.973	0.974
128	0.982	0.978	0.925	0.928	0.979	0.972	**0.980**	**0.979**
160	**0.986**	**0.982**	**0.941**	**0.942**	**0.984**	**0.981**	0.973	0.975
192	0.975	0.975	0.930	0.925	0.975	0.975	0.972	0.973
224	0.972	0.969	0.934	0.934	0.972	0.969	0.954	0.950

inputs. These results showed that the information contained in the error map plays a crucial role when predicting the quality of an image, especially for complex types of distortion. Performance on the TID2013 IQA database degraded markedly because this database contains more complex types of distortion such as local block-wise distortions, comfort noise, and chromatic aberrations. However, $\hat{I}_d$ as input does not sufficiently emphasize the distortion. It is also the main reason why NR-IQA cannot achieve adequate performance on the TID2013 IQA database.

2) EFFECTS OF PATCH SIZE

In this section, we describe the variation in performance with different training patch sizes. To obtain the reliable label for each patch, the patches size was larger than that for conventional patch based deep IQA metrics, which is normally 32×32 pixels. We tried five patch sizes from 96×96 to 224×224 pixels. Patches that are too small cannot represent the overall distortion on an image for the proposed LRDB-IQA method, whereas patches that are too large sharply increase computational complexity. Table 8 provides the results of one-time training and we selected 160×160 pixels as the patch size for all experiments in the following sections.

3) EFFECTS OF TWO BRANCHES

In this section, the effects of the local QMG module and the global QMG module were evaluated. First, we recorded the changing γ value along with the training iteration (Figure 7). In these experiments, we fixed the epoch iteration to 6000 to provide a clearer illustration. The γ value was initialized to zero which ignores the effects of the long-range dependencies. The local QMG module was trained at the beginning. Figure 7 shows that the γ value dramatically increased and reached a stable value at approximately 2000 epochs. In this period, the training loss also rapidly decreased. This finding showed the importance of the long-range dependencies in making a positive contribution to the overall performance.

To further evaluate the effects of the long-range dependencies, we tried the proposed system with $\mathcal{P}_L$ only (short-range dependencies only) and $\mathcal{P}_G$ only (long-range dependencies only). All other settings remained the same.

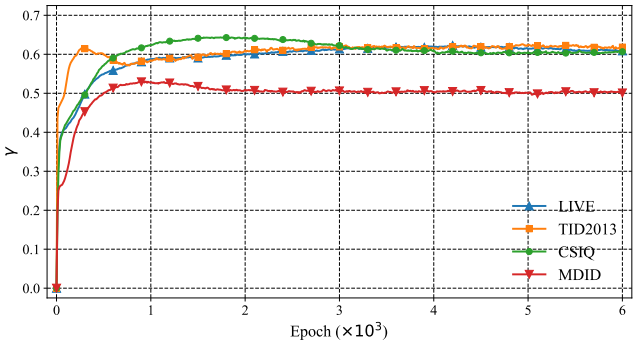


FIGURE 7. Changes in the γ value during training.

The average values of LCC and SROCC of the 10-times reference image-based training-testing split are provided in Table 9. These results indicate that each branch can attain acceptable performance alone, but more favorable performance can be achieved when two modules of short-range and long-range dependencies are combined.

TABLE 9. Effects of two branches.

	LIVE		TID2013	
	LCC	SROCC	LCC	SROCC
$\mathcal{P}_s$ only	0.983±.006	0.977±.004	0.948±.013	0.945±.014
$\mathcal{P}_a$ only	0.972±.016	0.974±.006	0.939±.013	0.942±.011
$\mathcal{P}_s + \gamma\mathcal{P}_a$	**0.985**±.000	**0.978**±.000	**0.952**±.007	**0.948**±.006

4) EFFECTS OF POOLING METHODS

The proposed regressor combines many Minkowski pooling functions with different exponents. Table 10 presents the comparison of performance when choosing different B values (the total number of Minkowski exponents involved)

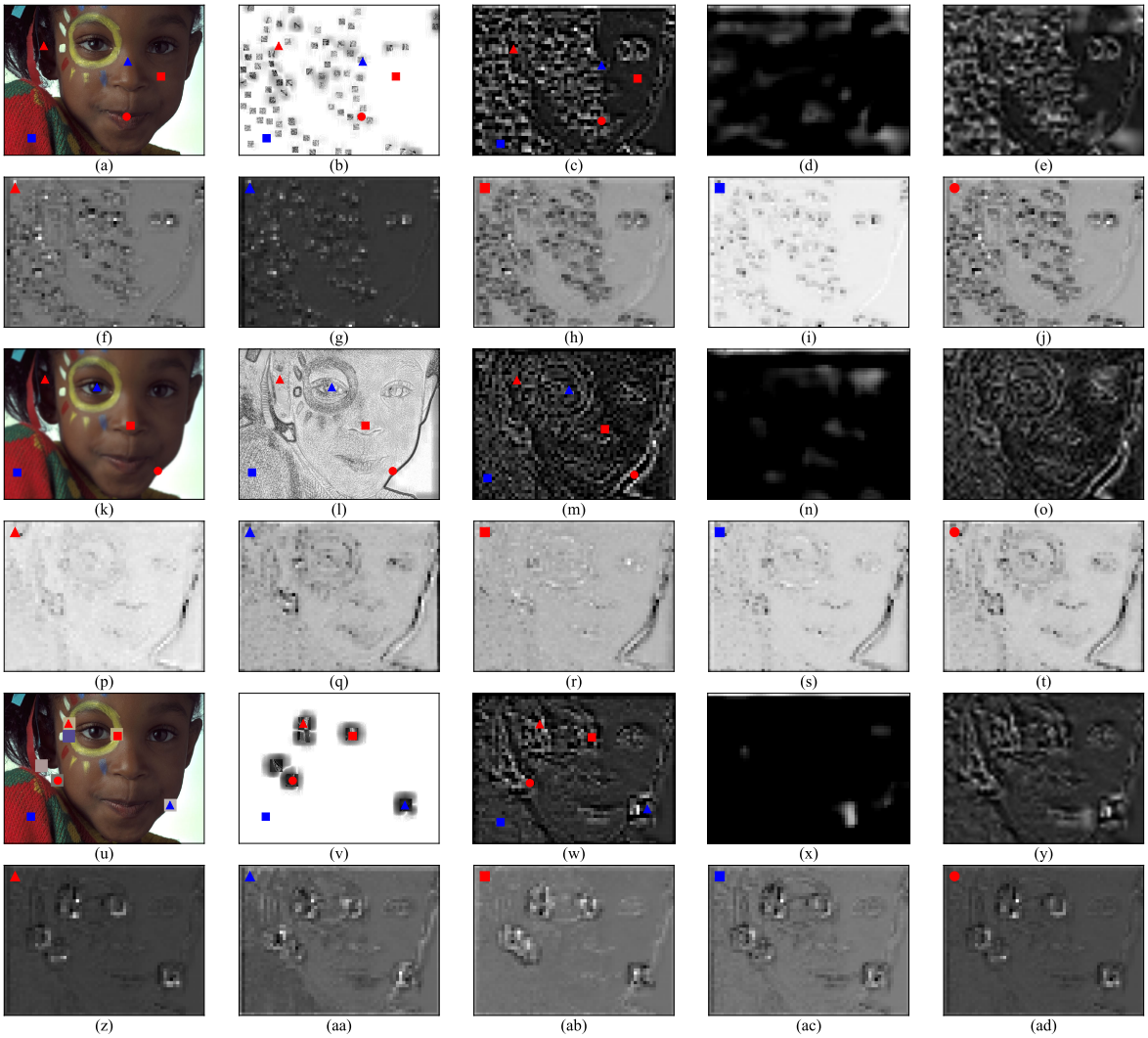


FIGURE 8. Examples of the predicted perceptual quality maps: (a), (k), and (u) are the distorted images; (b), (l), and (v) are the objective error maps (E); (c), (m), and (w) are the predicted quality maps from the attention branch ($\mathcal{P}_G$); (d), (n), and (x) are the predicted quality maps from the main path ($\mathcal{P}_L$); (e), (o), and (y) are the perceptual quality maps after elementwise addition (P); (f)-(j), (p)-(t), and (z)-(ad) are the predicted attention weights from A at different locations.

TABLE 10. Effects of Minkowski pooling.

	LIVE		TID2013	
	LCC	SROCC	LCC	SROCC
B = 1	0.985±.002	0.978±.003	0.944±.015	0.945±.011
B = 2	0.983±.003	0.977±.001	0.949±.007	0.944±.009
B = 3	0.978±.006	0.975±.004	0.948±.009	0.946±.009
B = 4	0.985±.000	0.978±.000	**0.952**±.007	**0.948**±.006

of Eq. (12). All values in this table are the average values and standard deviation of a 10-times training-testing split. On the LIVE database, the effect of B is not obvious. However, performance clearly improves on the TID2013 database when B values increases.

E. PREDICTED PERCEPTUAL QUALITY MAP

In this section, we present a few examples of the predicted perceptual quality map from the middle layer of the proposed system (Figure 8), which includes $\mathcal{P}_G$, $\mathcal{P}_L$, and $\mathcal{P}$. In Figure 8, darker regions indicate lower quality. Here, the model was trained on the TID2013 IQA database, and the images provided were included in the test set. The objective error map (Figure 8-(b), (l), and (v)) shows that the types of distortion in these examples are different. As shown in Figure 8-(c), (m) and (w), $\mathcal{P}_G$ maintains the detailed outline of the image and the error map. $\mathcal{P}_L$ (Figure 8-(d), (n), and (x)) ignores most areas and focuses on a few areas. $\mathcal{P}$ is primarily affected by $\mathcal{P}_G$, which aligns with the results of the ablation study (3). This also supports our assumption that global dependencies are critical for quality assessment.

To further evaluate the generated attention map, Figure 8 (f)-(j), (p)-(t), and (z)-(ad) illustrates the attention weights that correspond to five locations and are extracted from the attention map ($\mathbf{A}$). The attention map aims to link a single point on the quality map to every pixel on the objective error map with different weights. The value of each pixel on $\mathcal{P}_G$ is determined by the weighted combination of the entire objective error map Eq. (7). Thus, the proposed structure considers the long-range dependencies on the error map to boost the quality assessment performance. From these subplots, different locations on $\mathcal{P}_G$ have different patterns for giving weights to each pixel of the error map. For example, the block artifacts in Figure 8-(u) are more visible than those in Figure 8-(a), but both artifacts are obvious on the error map. The attention weights for the visible artifacts (Figure 8 (z)-(ad)) have larger values than their neighbors, and the weights for the nonvisible artifacts (Figure 8 (f)-(j)) yield smaller values than other areas. The attention weights emphasize the area on the error map where the visible artifacts appear and depreciate the nonvisible artifacts on the error map. Global dependencies have thus been built in this way.

IV. CONCLUSION

In this study, we found that the FR-IQA system can benefit from a CNN-based deep model that can take into account the long-range dependencies between local pixels and has a higher-order spatial pooling module. A global QMG module, which connects all positions in the input deep feature map through a series of matrix operations, is proposed. In order to improve the representation ability, we also proposed a high-order spatial pooling method that combines the Minkowski function into a fully connected layer at the end of the model. Comprehensive experimental results show that the proposed FR-IQA system with long-range dependencies-boosted path can achieve good prediction accuracy of image quality scores with high generalization. In addition, compared with a system that only predicts a single quality score, the system can also output a perceptually weighted quality map, thereby expanding the application scope of the IQA system. However, further study is needed on the generalizability of unknown types of distortion and on establishing long-range dependencies for NR-IQA.

REFERENCES

[1] 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.
[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.
[3] 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.
[4] 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.
[5] H. Sheikh and A. C. Bovik, "Image information and visual quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jan. 2004, p. III-709.
[6] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. *LIVE Image Quality Assessment Database Release 2*. Accessed: Apr. 20, 2020. [Online]. Available: http://live.ece.utexas.edu/research/quality
[7] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "DipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.
[8] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.
[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2016, pp. 21–37.
[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[11] 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.
[12] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1733–1740.
[13] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1676–1684.
[14] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.
[15] 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.

[16] S. Bosse, D. Maniry, K.-R. Muller, 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.

[17] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1040–1049.

[18] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[19] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[20] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 60–65.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[22] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318

[23] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 533–546, Apr. 2015.

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

[25] R. Hassen, Z. Wang, and M. Salama, "No-reference image sharpness assessment based on local phase coherence measurement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 2434–2437.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–11.

[27] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML Workshop Deep Learn. Audio, Speech Lang. Process.*, vol. 30, 2013, p. 3.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 1–4.

[30] 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. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[31] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual DOG model fused with random forest," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3282–3292, Nov. 2015.

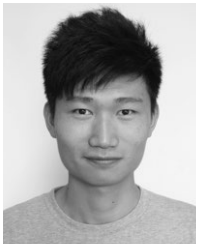
[32] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[33] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng, "Image quality assessment using similar scene as reference," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2016, pp. 3–18.

[34] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[35] 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.

[36] L.-M. Po, M. Liu, W. Y. F. Yuen, Y. Li, X. Xu, C. Zhou, P. H. W. Wong, K. W. Lau, and H.-T. Luk, "A novel patch variance biased convolutional neural network for no-reference image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1223–1229, Apr. 2019.



MENGYANG LIU received the B.E. degree in optoelectronic engineering from the Shanghai University of Electric Power, Shanghai, China, in 2014, and the M.Sc. degree in electronic and information engineering and the Ph.D. degree from the City University of Hong Kong, in 2015 and 2019, respectively. He is currently an Engineer with the Tencent Video, Tencent Holdings Ltd. His research interests include image and video processing, video indexing, computer vision, and machine learning.

LAI-MAN PO (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Hong Kong, in 1988 and 1991, respectively.

Since 1991, he has been with the Department of Electronic Engineering, City University of Hong Kong. He is currently an Associate Professor with the Department of Electrical Engineering, City University of Hong Kong. He has authored over 150 technical journal articles and conference papers. His research interests include image and video coding with an emphasis on deep learning-based computer vision algorithms. He is a member of the Technical Committee on Multimedia Systems and Applications and the IEEE Circuits and Systems Society. He was the Chairman of the IEEE Signal Processing Hong Kong Chapter, in 2012 and 2013. He served on the Organizing Committee of the IEEE International Conference on Acoustics, Speech, and Signal Processing, in 2003, and the IEEE International Conference on Image Processing, in 2010. From 2011 to 2013, he was an Associate Editor of *HKIE Transactions*.

XUYUAN XU received the B.E. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, in 2010 and 2014, respectively. He is currently a Senior Engineer with the Tencent Video, Tencent Holdings Ltd. His research interests include 3D video coding, 3D view synthesis, audio/video fingerprint, and object detection.

KWOK WAI CHEUNG (Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees from the City University of Hong Kong, in 1990, 1994, and 2001, respectively, all in electronic engineering. From 1990 to 1995, he worked as an Engineer with Hong Kong Telecom. From 1996 to 2002, he was a Research Assistant with the Department of Electronic Engineering, City University of Hong Kong. From 2002 to 2016, he was an Assistant Professor with the Chu Hai College of Higher Education, Hong Kong. Since 2016, he has been an Associate Professor with the School of Communication, The Hang Seng University of Hong Kong. His research interests are in the areas of image processing, machine learning, and social computing.

YUZHI ZHAO received the B.Eng. degree in electronic information from the Huazhong University of Science and Technology, Wuhan, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, City University of Hong Kong. His research interests include image processing, computer vision, deep learning, and machine learning.

KIN WAI LAU received the B.E. degree (Hons.) in information engineering from the City University of Hong Kong, in 2017, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. He is also working as a Software Engineer with TFI Digital Media Ltd. His research interests include image and video processing, video/image retrieval, computer vision, and deep learning.

CHANG ZHOU received the B.Sc. degree from the Donghua University of China, Shanghai, in 2016, and the master's degree from the City University of Hong Kong, Hong Kong, in 2017, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His research interests are in computer vision and deep learning.