No-reference image quality assessment with multi-scale weighted residuals and channel attention mechanism

Changzhong Wang · Xiang Lv · Weiping Ding · Xiaodong Fan

Accepted: 22 September 2022 / Published online: 5 October 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

With the rapid development of deep learning, no-reference image quality assessment (NR-IQA) based on convolutional neural network (CNN) plays an important role in image processing. Currently, most CNN-based NR-IQA methods focus primarily on the global features of images while ignoring detail-rich local features and channel dependencies. In fact, there are subtle differences in detail between distorted and reference images, as well as differences in the contribution of different channels to IQA. Furthermore, multi-scale feature extraction can be used to fuse the detailed information from images with different resolutions, and the combination of global and local features is critical in extracting image features. As a result, in this paper, a multi-scale residual CNN with an attention mechanism (MsRCANet) is proposed for NR-IQA. Specifically, a multi-scale residual block is first used to extract features from distorted images. Then, the residual learning with active weighted mapping strategy and channel attention mechanism is used to further process image features to obtain more abundant information. Finally, the fusion strategy and full connection layer are used to evaluate image quality. The experimental results on four synthetic databases and three in-the-wild IQA databases, as well as cross-database validation results, show that the proposed method has good generalization ability and can be compared with the most advanced methods.

Keywords Multi-scale · No-reference image quality assessment · Channel attention · Active weighted mapping strategy

1 Introduction

With the rapid development of computer technology and mobile devices, image information has become an information source for perceptual content in daily life and learning. However, image distortions may occur in storage devices as a result of compression, copying, transmission, and other operations. Therefore, image quality assessment (IQA) has emerged as a critical technology in computer vision and image processing (Hong et al. 2016; Hao et al. 2019).

In general, IQA is divided into two categories: subjective and objective evaluations. Subjective evaluation is to evaluate images with people as observers to reflect people’s visual perception. Objective evaluation calculates image quality by designing models and algorithms. It reflects the subjective perception of human eyes and provides evaluation results based on digital calculation. The objective IQA technology can be divided into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). For many IQA methods, reference images are difficult to obtain due to factors like environmental impacts on real-world practical applications, etc. As a result, the FR-IQA and RR-IQA methods are severely limited in their application of IQA. Because the NR-IQA does not require many reference images, this technology has piqued the interest of many researchers.

Due to the rapid development of deep learning, convolutional neural network (CNN) has been successfully applied in the field of computer vision (Holzinger 2018, 2021), such as image super-resolution and recognition tasks. It has achieved numerous remarkable results (Jin et al. 2016). Despite CNN’s impressive performance in many visual tasks, NR-IQA remains challenging due to a lack of real-rich sam-
In existing synthetic IQA databases, such as LIVE (Sheikh et al. 2006), CSIQ (Larson and Chandler 2010), TID2008 (Ponomarenko et al. 2009), and TID2013 (Ponomarenko et al. 2015) for NR-IQA is relatively small with only about 20 images with different contents, respectively; however, there are NR images for them. Second, image distortion is typically reflected in the variation of high-frequency information details across an entire image. In general, the more severe the image distortion, the more the high-frequency information is lost, whereas the low-frequency information hardly changes. People will ignore some real details and only pay attention to the dense contour information if they do not consider the observation process of objects and only think about the problem from the pixel value of images. This will increase the difficulty of learning.

To address the issue of a limited number of training samples, Kang et al. (2014) proposed a very shallow network. They obtained the global image quality by learning and predicting the local image quality using 32 × 32 image blocks as input. This method used image enhancement techniques, which are necessary for image denoising or super-resolution problems. The network structure, however, is relatively simple, and the number of network layers is relatively shallow. To extract image features, only one layer of the network structure with a convolution kernel of 7 × 7 size is used, and local features cannot describe global features. As a result, this method is incapable of meeting the requirement of sufficient accuracy.

Many researchers have devoted themselves to the task of NR-IQA based on CNNs and proposed many effective algorithms. Li et al. (2016); Dash et al. (2017) trained a network using a full image of 224 × 224 pixels. Pan et al. (2016); Zuo et al. (2016) proposed saliency maps to adjust the weight of each patch. Kim and Lee (2017) predicted subjective scores by learning human visual perception behaviors. Lin and Wang (2018) introduced GAN into a quality assessment network to provide fake reference images for NR-IQA. Yang et al. (2019) proposed a NR-IQA network based on a two-stream CNN structure, with the network receiving both the original and gradient images. To extract different features, Li et al. (2018) adopted multi-scale convolution to process images for super-resolution, resulting in more effective high-frequency details and good results. Inspired by this, Chen et al. (2019) also applied multi-scale feature extraction in the field of image quality evaluation. Experiments showed that multi-scale feature extraction can improve IQA performance.

However, most existing methods focused on extracting more contour features from images, ignoring detail-rich local features and feature map channel dependencies. In fact, fusion of features with different resolutions can help to reduce the loss of excessive details in severely distorted images. Channel attention mechanism and active weighted mapping strategy have achieved good results in the super-resolution direction of image processing. They mainly simulate human vision system for weighted image processing. As far as we know, no one has combined multi-scale (three-scale) feature extraction, channel attention mechanism, and active weighted mapping for IQA.

To address the above issues, this study proposed a novel multi-scale residual CNN with an attention mechanism for extracting contour and detail features and processing channel correlation. The proposed network employs a multi-scale residual block (MSRB) with three parallel branches to extract distorted images, which differs significantly from the structures of Refs. Li et al. (2018); Chen et al. (2019). MSRB can fuse the global and local information from images with different resolutions. In addition, the network employs residual learning of active weighted mapping strategy (AWMS) and channel attention mechanism (CAM) to process global and local features for obtaining richer information. The experiments showed that the proposed network model can improve quality assessment performance for no-reference images.

The contributions of this study are as follows:

1. A novel IQA algorithm based on convolution neural network is proposed, which used several effective modules to extract image details and contour features. Experimental results show that it can better learn image quality features and evaluate image quality more accurately.

2. A MSRB with three parallel branches is proposed to extract distorted images. Furthermore, the active weighted mapping mechanism is used in MSRB for residual learning, with its main role being to redesign input mappings rather than directly adding two residual learning results.

3. CAM is incorporated into the IQA task. This mechanism determines the weight values of various channels. This mechanism can effectively improve cross-layer information flow through residual connections and channel relationships.

The rest of the paper is organized as follows. Section 2 provides a brief overview of some methods related to the proposed network. Section 3 describes the proposed network model. Section 4 describes the experimental details and conducts comparative and cross-validation experiments. Section 5 presents the conclusion and direction for future research.
2 Related works

In this section, we review some related methods and algorithms, and briefly discuss their advantages and disadvantages.

2.1 No-reference IQA image quality assessment based on CNN

Several years ago, Kang et al. (2014) first introduced CNN into the field of NR-IQA and proposed a network with only five layers, which has crucial research significance. The method adopted the image block training strategy. The original image is first divided into 32 × 32 non-overlapping image blocks, with the quality of each image block assumed to be roughly the same as that of the original image. The image blocks are then sequentially fed into the network for training parameters. The proposed network pushed NR-IQA to a high level despite the lack of samples in the training database. Many scholars have since devoted themselves to researching the application of CNN to NR-IQA. However, the network proposed by Kang et al. has only one convolutional layer, which is relatively shallow. As a result, the network has some limitations and fails to meet accuracy requirements.

Inspired by Kang’s idea, Li et al. (2016); Dash et al. (2017) proposed a CNN with multiple layers of structures. They trained the network with a full image of size 224 × 224 pixels. Although the small patch block segmentation technique was not used, the deep network structure produced relatively good results on some datasets with fewer reference images. Sun et al. (2016) and Bosse et al. (2016) achieved remarkable results by fine-tuning parameters in the deep CNN network (Li et al. 2016; Dash et al. 2017). Pan et al. (2016) and Zuo et al. (2016) also designed a deep network for the small patches. A CNN-based significance map detection method was proposed in this network. The significance map was used to adjust the weight of each patch, and the overall image quality score was output at last. Cheng et al. (2017) proposed a NR-IQA network based on a pre-significance map and proved through experiments that the prediction error of significance information was lower than that of a uniform region. Based on this fact, they proposed a pre-SM algorithm that can assign relatively large weights to small blocks of significance graphs to reduce the error.

Recently, Kim and Lee (2017) have proposed a new distorted IQA (DIQA) model. The training of this model is divided into two parts. The first part is to determine the information associated with the distorted and reference images. The second part primarily employs human subjective scores to fine-tune the proposed CNN network to predict subjective opinion scores by learning human visual perception behavior. However, when the network randomly clips the input image, the patch overlaps will be caused, resulting in double calculation and waste time and resources. Lin and Wang (2018) proposed a NR-IQA network based on a generative adversarial network. In this model, GAN was introduced into the quality assessment network to provide pseudo-reference graphs for NR. Furthermore, the loss of quality perception is added to the GAN generator and the discriminator is used to discriminate the pseudo-reference graph. Although the above methods have effectively improved and enhanced the NR-IQA, the problem of gradient propagation weakening or disappearing may occur in some databases due to the deep network.

To address this issue, Yang et al. (2019) proposed a NR-IQA network based on a two-stream CNN structure, which inputs original and gradient images into the network, enabling the network to extract different features representing distortion. Because the network extracted features using a single-size convolution kernel, the feature information extracted from Yang’s network is insufficient, and detailed features may be lost. Chen et al. (2019) applied multi-scale feature extraction in the field of image quality evaluation. In the field of super-resolution, Li et al. (2018) first used a multi-scale residual convolution kernel to process images, thereby achieving a good zoom performance effect. They did not, however, attempt to use an active weighted strategy or a CAM to enhance the effectiveness of feature representation. Sun et al. (2021) used the structure and texture similarity of feature maps extracted by intermediate layers to represent NR model features, and employed the global average and standard deviation of final feature maps fused by intermediate feature maps to represent NR model features. Sun et al. (2021) proposed a ladder structure to learn quality perception features more effectively. This structure integrates the features of the middle layers hierarchically into the final feature representation, allowing the model to fully utilize low-level to high level vision information. They could not pay more attention to the special channel of distorted images because they did not consider the correlation between channels.

2.2 Different features of extraction blocks

He et al. (2016) proposed a residual block (Fig. 1a) to avoid the disappearance or explosion of gradient during network training, which can retain the original input features. To extract richer features from images, some researchers proposed inception blocks (Fig. 1b) that use convolution kernels of different sizes. Zhang et al. (2018) designed a residual dense block (Fig. 1c) by stacking residual and dense blocks. The residual dense block is capable of fully extracting both local and global features.
2.3 Channel attention

Visual attention is the unique brain signal processing mechanism of human vision. We can obtain the important target areas that require attention by quickly scanning a global image, and then pay more attention to this part to obtain more detailed features and suppress some relatively useless features. The attention mechanism in deep learning draws lessons from human attention thinking mode. Many researchers have recently devoted themselves to investigating the mechanisms of attention in deep learning. Hu et al. (2018) proposed a squeeze and excitation (SE) block that adaptively adjusts channel weights by learning channel dependence. Zhang et al. (2018) introduced a deep residual channel attention network with CAM. To process input images, the network combined residual learning with a SE block and adjusted different feature maps based on the degree of dependence between input images. These feature maps were then given appropriate weights. The results of the experiments showed that CAM can improve image processing. Woo et al. (2018) proposed a convolutional attention module by combining channel attention and spatial attention. This module can automatically calculate the importance of feature channels as well as the different regions of the feature space. Gao et al. (2020) proposed a global second-order average pooling module that used covariance to represent the relationship between channels. Because the original input can be scaled along channel dimension, this module played a role in channel attention.

2.4 Active weighted mapping

In general residual learning, the original input information is added to the information obtained after some convolution layers, and the final information is obtained. The simplest residual learning is depicted in Fig. 2a as follows:

\[ y = x + F(x) \]  
(1)

where \( x \) and \( y \) represent the input and output of the residual block, and \( F(\cdot) \) is the residual learning convolution function.

Hyoungho et al. (2018) proposed an active weighted mapping (AWM) strategy in 2018 based on residual learning. The mechanism of AWM differs from that of residual learning. It is the result of redefining the mapping of residual learning rather than directly adding two paths. The AWM mechanism, in particular, believes that the weights of \( x \) and \( F(x) \) are different. Different weights (Fig. 2 (b)) can be derived and redistributed following the AWM mechanism. The AWM mechanism can be expressed as follows:

\[ y = \lambda x + \mu F(x) \]  
(2)

where \( \lambda \) and \( \mu \) are the weight values calculated by the AWM mechanism.

3 Proposed method

In this section, we provided a detailed description of the proposed network and its modules, as well as a discussion of the network’s structure, the function of each module, and the calculation process.
3.1 A new multi-scale residual block (MSRB)

Before we present our MsRCANet model, a MSRB structure is shown in Fig. 3, where $S^0$ and $S^{n-1}$ are the input and output of the MSRB, respectively. This module consists of three parallel branches of various convolution kernels and an AWMS mechanism. The MSRB structure is described in detail below.

In the first part, three convolution kernels of different sizes, $3 \times 3, 5 \times 5$ and $7 \times 7$, are used to extract multi-scale features from input features. Following the convolution operation, the LReLU function is used to obtain more detailed context information. The three output features are then concatenated into a tensor. The procedure is repeated twice. Finally, the $1 \times 1$ convolution kernel is used to adjust the channel number of output features. This part can be described as follows:

$$A_{11} = L(W_{11}^{3x3} \ast S^{n-1}+b_{11}^{3x3})$$  \hspace{1cm} (3)
$$A_{12} = L(W_{12}^{5x5} \ast S^{n-1}+b_{12}^{5x5})$$  \hspace{1cm} (4)
$$A_{13} = L(W_{13}^{7x7} \ast S^{n-1}+b_{13}^{7x7})$$  \hspace{1cm} (5)
$$C_1 = [A_{11}, A_{12}, A_{13}]$$  \hspace{1cm} (6)
$$B_{21} = L(W_{21}^{3x3} \ast C_1+b_{21}^{3x3})$$  \hspace{1cm} (7)
$$B_{22} = L(W_{22}^{5x5} \ast C_1+b_{22}^{5x5})$$  \hspace{1cm} (8)

where $W_{ij}^{k \times k}$ and $b_{ij}^{k \times k}$ denote the weight and bias of the branch in the layer, respectively; and $k \times k$ denotes the size of a convolution kernel, $k = 3, 5, 7$, $i = 1, 2, j = 1, 2, 3$. $L(\cdot)$ represents the LReLU function, $A_{ij}$ and $B_{ij}$ represent the results obtained using LReLU functions. concat $[\cdot]$ represents the fusion operation, and $D$ represents the result obtained after applying the $1 \times 1$ convolution kernel.

In the second part, AWMS is used to deal with the output feature maps and the initial input feature maps $S^{n-1}$ obtained in the first part (see Fig. 4). Assume $S^{n-1} = \{S_1, S_2, \ldots, S_k\}$ has $k$ different feature maps, and the size of the feature maps is $M \times N$. The global average channel information $Z = \{z_1, z_2, \ldots, z_k\}$ is obtained after global average pooling. It means that the two-dimensional feature map on each channel will become a statistic after global average pooling, and the $c$th element of $Z$ can be expressed as follows:

$$z_c = Ave(S_c) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} S_c(i, j)$$  \hspace{1cm} (12)

Fig. 3 Multi-scale residual block (MSRB)

Fig. 4 Active weighted mapping strategy (AWMS)
where \( Ave(\cdot) \) is the global average pool operation and \( S_c(i, j) \) is the value for the position \((i, j)\) in the \( cth \) channel feature map. As shown in Fig. 4, \( Z^1 \) and \( Z^2 \) represent two statistics after pooling the global average.

We concatenate \( Z^1 \) and \( Z^2 \) into \( Z \) to get the weights of \( D \) and \( S^{n-1} \). The weights \( \lambda, \mu \) of \( S^{n-1} \) and \( D \) are calculated using two nonlinear fully connected layers. The procedure is as follows:

\[
Z^1 = Ave(S^{n-1}) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} S^{n-1}(i, j) \tag{13}
\]

\[
Z^2 = Ave(D) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} D(i, j) \tag{14}
\]

\[
Z = \begin{bmatrix} Z^1, Z^2 \end{bmatrix} \tag{15}
\]

\[
y = \sigma(W_2\delta(W_1Z + b_1) + b_2), \gamma = \{\lambda, \mu\} \tag{16}
\]

\[
S^n = \lambda S^{n-1} + \mu D \tag{17}
\]

where \( W_1 \in \mathbb{R}^{h \times 2k} \) and \( W_2 \in \mathbb{R}^{2 \times h} \) denote the weight matrices of the fully connected layers and \( h \) (2 < \( h \) < 2\( k \)) represents the number of dimension nodes in the hidden layer. \( \sigma \) and \( \delta \) represent Sigmoid and ReLu functions, respectively. \( \gamma = \{\lambda, \mu\} \) is the weight vector of \( D \) and \( S^{n-1} \), and \( S^n \) is the MRSB output.

### 3.2 Channel attention mechanism (CAM)

In general, it is considered that the feature map of each channel input is equally important. In fact, the importance of feature information input of each channel is different; thus, each channel should be given a different weight. This study employs CAM similar to Hu et al. (2018) to assign weights to each channel. As shown in Fig. 5, CAM consists of two parts: squeeze and excitation. These two parts are used to learn the weight of each channel to generate the attention of channel domain. Next, the structure and operation process of squeeze will be given in detail.

The squeezing function is similar to the AWMS mechanism. If \( X = \{X_1, X_2, \ldots, X_C\} \) is the input of CAM and its space size is \( M \times N \), and the global average channel information is \( Z = \{z_1, z_2, \ldots, z_C\} \) after global average pooling, the \( kth \) element of \( Z \) can be expressed as follows:

\[
z_k = Ave(X_k) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} X_k(i, j) \tag{18}
\]

The excitation function consists of two fully connected layers. It is mainly used to generate channel weights for input feature maps; the learned weights represent the importance of feature channels. Here, the Relu and Sigmoid functions are selected. After obtaining the weight of each feature channel, the weight is multiplied by the corresponding original feature channel. The weighted feature maps are created using these specific steps. The specifics are as follows:

\[
\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_C\} = \sigma(W_2\delta(W_1Z)) \tag{19}
\]

\[
\hat{X} = \lambda \odot X = \begin{bmatrix} \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_C \end{bmatrix} \tag{20}
\]

where \( W_1 \in \mathbb{R}^{C \times C} \) and \( W_2 \in \mathbb{R}^{C \times C} \) represent the weights of the two fully connected layers, \( r \) is a scaling ratio, and \( \sigma \) and \( \delta \) represent the Sigmoid and ReLu functions, respectively. \( \lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_C\} \) is the learned weight vector of feature channels. \( \odot \) is the result of channel-wise multiplication, and \( \hat{X} \) is the output of CAM. In particular, the \( kth \) element in \( \hat{X} \) can be expressed as \( \hat{X}_k = \lambda_k \cdot X_k \).

### 3.3 Network model architecture

Figure 6 depicts the proposed network’s architecture. \textit{Input} and \textit{Output} represent the network’s input and output, respectively, where \textit{Input} is a locally normalized 32 × 32 image patch. To begin, the network employs a 3 × 3 convolution kernel to extract the detailed features of an input image. The network then employs a 9 × 9 convolution kernel to extract contour features. During the training process, the quality score of each patch is used as the true label score of its initial image.
In the testing stage, the score estimation of real quality for each image is obtained by averaging its patch prediction scores. As shown in Fig. 6, the network is divided into three sections—the initial feature extraction, advanced feature extraction, and fully connected network parts. The local normalization method and three main components of the network are described in detail below.

Local normalization employs a straightforward normalization method similar to that described in Zhang et al. (2016). Assuming that the pixel value at the position \((x, y)\) is \(I(x, y)\), the normalized value \(\tilde{I}(x, y)\) is as follows:

\[
\begin{align*}
\tilde{I}(x, y) &= \frac{I(x, y) - \mu(x, y)}{\sigma(x, y) + \theta} \quad (21) \\
\mu(x, y) &= \frac{1}{PQ} \sum_{p=-P}^{p=P} \sum_{q=-Q}^{q=Q} I(x + p, y + q) \quad (22) \\
\sigma(x, y) &= \sqrt{\frac{1}{PQ} \sum_{p=-P}^{p=P} \sum_{q=-Q}^{q=Q} (I(x + p, y + q) - \mu(x, y))^2} \quad (23)
\end{align*}
\]

where \(\theta\) is a small positive number (ensuring that the denominator is not 0), and \(PQ\) is the total number of image pixels in the local window. In the experiment, this study uses a window with \(P = Q = 3\), which has a relatively good effect. We found that the effect of contrast and brightness of image patches on the learning of network structure can be alleviated by local normalization.

The initial feature extraction part employs the \(3 \times 3\) convolution kernel to extract features \(p_1\) from Input. The \(9 \times 9\) convolution kernel then extracts contour features and obtains feature information \(p_2\). The two operations can be described as follows:

\[
\begin{align*}
p_1 &= W_1^{3 \times 3} \ast input + b_1 \quad (24) \\
p_2 &= W_2^{9 \times 9} \ast p_1 + b_2 \quad (25)
\end{align*}
\]

where \(\ast\) denotes the convolution operation, and \(W_1^{3 \times 3}\) and \(b_1\) denote the weight matrix and bias of the \(3 \times 3\) convolution kernel, respectively. \(W_2^{9 \times 9}\) and \(b_2\) represent the weight matrix and bias of the \(9 \times 9\) convolution kernel, respectively. The output \(p_2\) is the input of the next phase.

The advanced feature extraction part is the core part of the network. First, the MSRB is used to extract multi-scale features \(p_3\) with the AWMS mechanism. Second, CAM is used to learn the correlation between feature channels. Then, the network uses a \(5 \times 5\) convolution layer for feature extraction and employs max pooling and min pooling for the output features, following by three convolution layers and a parallel pooling layer. Finally, the pooling results are fed into a fully connected network component. The advanced feature extraction part is described as follows:

\[
\begin{align*}
p_3 &= f(p_2) \quad (26) \\
p_4 &= g(p_3) \quad (27) \\
p_5 &= W_5^{5 \times 5} \ast p_4 + b_5 \quad (28) \\
p_6^1 &= \max - \text{pool}(p_5) \quad (29) \\
p_6^2 &= \min - \text{pool}(p_5) \quad (30) \\
p_6 &= \left[ p_6^1, p_6^2 \right] \quad (31) \\
p_7 &= W_7^{7 \times 7} \ast p_6 + b_7 \quad (32) \\
p_8 &= W_8^{3 \times 3} \ast p_7 + b_8 \quad (33) \\
p_9 &= W_9^{1 \times 1} \ast p_8 + b_9 \quad (34) \\
p_{10}^1 &= \max - \text{pool}(p_9) \quad (35) \\
p_{10}^2 &= \min - \text{pool}(p_9) \quad (36) \\
p_{10} &= \left[ p_{10}^1, p_{10}^2 \right] \quad (37)
\end{align*}
\]

where \(f(\cdot)\) and \(g(\cdot)\) represent the MSRB and CAM processes, respectively; \(\max - \text{pool}(\cdot)\) and \(\min - \text{pool}(\cdot)\) represent the \(2 \times 2\) convolution kernel process. Similarly, the output \(p_{10}\) of this stage is the input of the next stage.

The fully connected network part consists of three fully connected layers. To reduce the complex co-adaptive rela-
tionship between neurons, the dropout method is used after the first full connection layer, and the ratio is set to 0.5. This part can be described as follows:

\[ p_{11} = \max(0, W_1 p_{10}) \]  
\[ p_{12} = \max(0, W_2 p_{11}) \]  
\[ output = p_{13} = W_3 p_{12} \]

where \( W_1, W_2, \) and \( W_3 \) represent the weight matrices of the three full connection layers with the node numbers 100, 800, and 1, respectively. The output represents the final predicted score of a test image.

### 3.4 Loss function

We can use mean square error or mean absolute error as the objective function in regression problems. Although the mean square error is a simple calculation, the mean absolute error has greater robustness to outliers and is better suited for the IQA model regression calculation. As a result, the mean absolute error serves as the optimization objective function in this study. Assume that \( x_n, y_n \) are an input patch and its true score and \( \theta \) is the parameter of network, then the optimization objective function is defined as follows:

\[ LOSS = \frac{1}{N} \sum_{i=1}^{N} \| (f(x_n; \theta), y_n) \|_1 \]  

Our primary goal is to solve the following optimization problem:

\[ \hat{\theta} = \arg \min_{\theta} LOSS \]
\[ = \arg \min_{\theta} \left( \frac{1}{N} \sum_{i=1}^{N} \| (f(x_n; w), y_n) \|_1 \right) \]  

### 4 Experimental results and analysis

In this section, we describe experimental settings, perform comparative experiments on several databases, and analyze the experimental results in detail.

#### 4.1 Database

When training the network, patches of 32 × 32 size (non-overlapping) are obtained from the initial image and used as the network’s input. Compared with taking the initial image as the original input, more training samples can be obtained by partitioning the initial image into patches. Kang et al. (2014) pointed out in experiments that the quality score of each patch could be used as the ground score of the initial image. Therefore, the average value of the predicted scores for input patches is used in the test stage as the estimate of the quality score of the initial image. This study uses four synthetic databases: LIVE (Holzinger 2021), CSIQ (Larson and Chandler 2010), TID2008 (Ponomarenko et al. 2009), and TID2013 (Ponomarenko et al. 2015). Besides the synthetic IQA databases, we also validate the proposed model on three in-the-wild IQA databases, including LIVE Challenge (Deepti and Alan 2015), KonIQ-10K (Vlad et al. 2020), and SPAQ (Fang et al. 2020). Detailed information about the four synthetic databases and three in-the-wild databases is described as follows.

LIVE database consists of 29 reference images and 779 distorted images. It has five types of distortion: JP2K compression (JP2K), JPEG compression (JPEG), white Gaussian (WN), Gaussian blur (GB), and fast fading (FF). LIVE database provides differential mean subjective scores (DMOS) for all distorted images, with values in the range \([0,100]\). The closer the DMOS is to 100, the worse the image quality and the more severe the distortion. Figure 7 shows the DOMS comparison of the same image with different distortion types in the LIVE dataset.

CSIQ database consists of 30 reference images and 866 distorted images, which is a shared database developed by the Computational Perception and Image Lab of Oklahoma State University. There are six types of distortion in the CSIQ database. CSIQ database provides DMOS for all distorted images, with DMOS values in the range \([0,1]\). The closer the DMOS value is to 1, the worse the image quality and the more severe the distortion. In this paper, we use a subset of this database with the same distortion types as the LIVE database.

TID2008 was established by the N504 Department of Signal Reception, Transmission, and Processing of the National Aeronautical and Aerospace University of Ukraine. TID2008 consists of 25 reference images and 1700 distorted images. There are 17 types of distortion, such as GB, image noise, JPEG compression, JP2K compression, JPEG transmission error, and JP2K transmission error. TID2008 provides mean subjective scores (MOS) for all distorted images. The value of MOS is in the range \([0,9]\). Contrary to DMOS, the closer the value is to 0, the worse the image quality and the more distorted the image.

TID2013 was established by the National Aeronautical and Aerospace University of Ukraine. TID2013 consists of 25 reference images and 3000 distorted images. There are 24 types of distortion, such as GB, JPEG compression, and JP2K compression. TID2013 provides mean subjective scores (MOS) for all distorted images. The value of MOS is in the range \([0,9]\). In this paper, we use a subset of this database with the same distortion types as the LIVE database.
LIVE Challenge was established by LIVE Lab of University of Austin, Texas, USA, in 2016, containing 1,162 true distorted images. Subjective scores were collected through crowdsourcing with more than 8,100 subject’s people. The collected subjective scores exceeded 350,000, and the value range of MOS was [3.42, 92.43].

KONIQ-10K (Konstanz Authentic Image Quality Database) was established by University of Konstanz in Germany in 2020, containing 10,073 distorted images. The subjective score was obtained through crowdsourcing with 1,459 annotators and 1.2 million subjective data. At the same time, the database also provides image attributes and EXIF (Exchangeable Image File Format) information.

SPAQ (Smartphone Photography Attribute and Quality) was established by Jiangxi University of Finance and Economics in 2020, mainly for the evaluation of mobile phone imaging quality, including 11,125 images collected from 66 different imaging devices. The subjective experiment was carried out in a standard laboratory environment, and the value range of MOS is [0, 100]. At the same time, the database provides image attribute information, category information, and EXIF information.

4.2 Evaluation metrics

To effectively evaluate the proposed NR-IQA algorithm, it is necessary to evaluate the algorithm’s performance. This study uses the Spearman rank correlation coefficient (SROCC), Kendall rank-order correlation coefficient (KROCC), Pearson linear correlation coefficient (PLCC), and root mean squared error (RMSE) to evaluate the performance of the proposed IQA algorithm. SROCC, PLCC, KROCC, and RMSE are used to measure the monotonicity and accuracy of the NR-IQA results, respectively.

\[ SROCC = 1 - \frac{6}{N(N^2 - 1)} \sum_{i=1}^{N} D_i^2 \]  
\[ KROCC = \frac{n_c - n_d}{0.5n(n - 1)} \]  
\[ PLCC = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}} \]  
\[ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2} \]

where \( x_i \) and \( y_i \) represent the subjective and objective quality score of the \( i \)th image, respectively, and \( \bar{x} \) and \( \bar{y} \) represent the average values of the subjective and objective quality scores, respectively. The range of PLCC is in [0,1]. PLCC is closer to 1 and RMSE is smaller, indicating that the image quality evaluation algorithm performs better.

To train the network, 100 iterations were conducted in the experiments. In each training iteration, the training, test, and verification sets were divided according to the ratio of 6:2:2. The proposed network was trained in the Pytorch deep.
4.3 Evaluation of synthetic databases and in-the-wild databases

Evaluation of synthetic databases. The distorted images were trained on four synthetic databases. Figures 8 and 9 show SROCC and PLCC variation and loss charts obtained by using 100 epochs of training on LIVE database, respectively. As shown in the figures, the network converged after about 60 epochs. Tables 1 and 2 compare the most advanced IQA algorithms in four synthetic databases. The best performance results for these IQA models in the tables are shown in bold, the second-best performance model is shown in italics.

Table 1 Performance comparison of state-of-the-art IQA methods based on LIVE, TID2008, and TID2013 databases

| Databases | LIVE | TID2008 | TID2013 |
|-----------|------|---------|---------|
|           | SROCC | KROCC | PLCC | RMSE | SROCC | KROCC | PLCC | RMSE | SROCC | KROCC | PLCC | RMSE |
| FR-IQA    |       |        |      |      |       |        |      |      |       |        |      |      |      |
| PSNR Wang et al. (2004) | 0.872 | 0.690 | 0.903 | 10.145 | 0.579 | 0.422 | 0.592 | 1.080 | 0.646 | 0.471 | 0.006 | 1.230 |
| SSIM Wang and Bovik (2002) | 0.912 | 0.753 | 0.941 | 7.860 | 0.771 | 0.581 | 0.778 | 0.833 | 0.734 | 0.549 | 0.787 | 0.752 |
| FSIM Zhang et al. (2011) | 0.966 | 0.842 | 0.957 | 8.106 | 0.885 | 0.708 | 0.878 | 0.645 | 0.850 | 0.656 | 0.874 | 0.608 |
| VSI Zhang et al. (2014) | 0.954 | 0.813 | 0.938 | 9.579 | 0.899 | 0.717 | 0.883 | 0.622 | 0.899 | 0.723 | 0.906 | 0.520 |
| NR-IQA    |       |        |      |      |       |        |      |      |       |        |      |      |      |
| CORNIA Ye et al. (2012) | 0.934 | 0.780 | 0.922 | 9.009 | 0.598 | 0.436 | 0.639 | 1.180 | 0.546 | 0.392 | 0.600 | 1.095 |
| BRISQUE Anish et al. (2012) | 0.942 | 0.768 | 0.947 | 8.721 | 0.640 | 0.463 | 0.669 | 1.021 | 0.581 | 0.415 | 0.640 | 0.995 |
| CNN Kang et al. (2014) | 0.953 | 0.818 | 0.945 | 13.720 | 0.671 | 0.484 | 0.710 | 1.003 | 0.618 | 0.447 | 0.708 | 0.928 |
| CNN++ Kang et al. (2015) | 0.962 | 0.839 | 0.964 | 6.760 | 0.726 | 0.533 | 0.770 | 0.909 | 0.672 | 0.488 | 0.735 | 0.883 |
| HOSA Xu et al. (2016) | 0.946 | 0.808 | 0.939 | 9.253 | 0.743 | 0.545 | 0.759 | 0.950 | 0.712 | 0.526 | 0.756 | 0.880 |
| RAN4IQA Ren et al. (2018) | 0.971 | 0.859 | 0.976 | 6.533 | 0.834 | 0.654 | 0.841 | 0.767 | 0.788 | 0.596 | 0.831 | 0.717 |
| RankIQA Liu and Bagdanov (2017) | 0.975 | 0.853 | 0.976 | 11.32 | 0.846 | 0.662 | 0.851 | 1.358 | 0.798 | 0.603 | 0.820 | 1.084 |
| dynamic Chen et al. (2020) | 0.968 | 0.853 | 0.973 | 6.415 | 0.865 | 0.679 | 0.869 | 0.682 | 0.812 | 0.631 | 0.845 | 0.677 |
| static Chen et al. (2020) | 0.977 | 0.875 | 0.981 | 6.130 | 0.862 | 0.681 | 0.864 | 0.663 | 0.820 | 0.640 | 0.842 | 0.699 |
| MsRCANet | 0.981 | 0.892 | 0.984 | 5.462 | 0.884 | 0.723 | 0.892 | 0.763 | 0.820 | 0.652 | 0.845 | 0.712 |
| Methods                        | Source     | LIVE SROCC | LIVE PLCC | CSIQ SROCC | CSIQ PLCC | TID2013 SROCC | TID2013 PLCC | Weighted average SROCC | Weighted average PLCC |
|-------------------------------|------------|------------|-----------|------------|-----------|---------------|---------------|------------------------|------------------------|
| NIQE Mittal et al. (2012)     | SPL 2012   | 0.912      | 0.901     | 0.883      | 0.901     | 0.811         | 0.824         | 0.869                  | 0.875                  |
| IL-NIQE Zhang et al. (2015)   | TIP 2015   | 0.902      | 0.908     | 0.821      | 0.865     | 0.521         | 0.648         | 0.748                  | 0.807                  |
| FRIQUEE Ghadiyaram and Bovik (2017) | JoV 2017   | 0.940      | 0.944     | 0.835      | 0.874     | 0.680         | 0.753         | 0.818                  | 0.857                  |
| BIECON Kim and Lee (2017)     | JSTSP 2017 | 0.961      | 0.960     | 0.815      | 0.823     | 0.717         | 0.762         | 0.831                  | 0.848                  |
| ResNet-ft Kim et al. (2017)   | SPM 2017   | 0.950      | 0.954     | 0.876      | 0.905     | 0.712         | 0.756         | 0.846                  | 0.872                  |
| WaDIQaM-NR Bosse et al. (2017) | TIP 2018   | 0.954      | 0.954     | 0.963      | 0.852     | 0.844         | 0.787         | 0.761                  | 0.864                  |
| MFIQA Kim et al. (2018)       | ICIP 2018  | 0.970      | 0.972     | 0.844      | 0.880     | 0.825         | 0.850         | 0.880                  | 0.901                  |
| BMPRI Min et al. (2018)       | TBC 2018   | 0.931      | 0.933     | 0.909      | 0.933     | 0.908         | 0.910         | 0.913                  | 0.928                  |
| SFA Li et al. (2018)          | TMM 2018   | 0.883      | 0.895     | 0.796      | 0.818     | 0.818         | 0.854         | 0.832                  | 0.867                  |
| NRVDPY Wu et al. (2019)       | Inf 2019   | 0.943      | 0.947     | 0.840      | 0.889     | 0.683         | 0.768         | 0.822                  | 0.868                  |
| ENIQA Chen et al. (2019)      | JVIP 2019  | 0.938      | 0.944     | 0.851      | 0.873     | 0.702         | 0.750         | 0.830                  | 0.856                  |
| SGDNet Yang et al. (2019)     | MM 2019    | 0.965      | 0.969     | 0.883      | 0.903     | 0.843         | 0.861         | 0.897                  | 0.911                  |
| DisNet Dendi et al. (2019)    | SPL 2019   | 0.950      | 0.950     | 0.880      | 0.910     | 0.790         | 0.820         | 0.873                  | 0.893                  |
| QUEADI Rajchel and Oszust (2020) | SIVIP 2020 | 0.844      | 0.831     | 0.641      | 0.698     | 0.533         | 0.615         | 0.673                  | 0.715                  |
| Two-Stream Yang et al. (2019) | IET IP 2020| 0.942      | 0.944     | 0.862      | 0.886     | 0.736         | 0.786         | 0.847                  | 0.872                  |
| CaHDIC Wu et al. (2020)       | TIP 2020   | 0.965      | 0.964     | 0.903      | 0.914     | 0.862         | 0.885         | 0.910                  | 0.921                  |
| DB-CNN Zhang et al. (2020)    | TCSV 2020  | 0.968      | 0.971     | 0.946      | 0.959     | 0.816         | 0.865         | 0.910                  | 0.931                  |
| HyperIQA Su et al. (2020)     | CVPR 2020  | 0.962      | 0.966     | 0.923      | 0.942     | –             | –             | –                      | –                      |
| UNIQUE Zhang et al. (2021)    | TIP 2021   | 0.969      | 0.968     | 0.902      | 0.927     | –             | –             | –                      | –                      |
| HFF Sun et al. (2021)         | TMM 2021   | 0.964      | –         | 0.919      | –         | –             | –             | –                      | –                      |
| MMMNet Li et al. (2021)       | TCSV 2021  | 0.970      | 0.970     | 0.924      | 0.937     | 0.832         | 0.853         | 0.909                  | 0.920                  |
| MsRCANet                      |            | **0.981**  | **0.984** | **0.955**  | **0.966** | **0.830**     | **0.855**     | **0.919**              | **0.932**              |
As shown in Table 1, four FR-IQA methods and nine NR-IQA methods were compared with our proposed MsRCANet on four indexes: SROCC, KROCC, PLCC, and RMSE. The four FR-IQA methods included SSIM (Wang et al. 2004), PSNR (Wang and Bovik 2002), FSIM (Zhang et al. 2011), and VSI (Zhang et al. 2014) and nine NR-IQA methods consisted of CORNIA (Ye et al. 2012), BRISQUE (Anish et al. 2012), CNN (Kang et al. 2014), CNN++ (Kang et al. 2015), HOSA (Xu et al. 2016), RAN4IQA (Ren et al. 2018), RankIQA (Liu and Bagdanov 2017), dynamic (Chen et al. 2020), and static (Chen et al. 2020).

As shown in Table 2, twenty-one methods were compared with our proposed algorithm on two indexes: SROCC and PLCC. These methods included NIQE (Mittal et al. 2012), IL-NIQE (Zhang et al. 2015), FRIQUEE (Ghadiyaram and Bovik 2017), BIECON (Kim and Lee 2017), ResNet-ft (Kim et al. 2017), WaDIQA-M-NR (Bosse et al. 2017), MFIQA (Kim et al. 2018), BMPRI (Min et al. 2018), SFA (Li et al. 2018), NRVPD (Wu et al. 2019), ENIQA (Chen et al. 2019), SGDNet (Yang et al. 2019), DisNet (Dendi et al. 2019), QUEADI (Rajchel and Oszust 2020), Two-Stream (Yang et al. 2019), CaHDC (Wu et al. 2020), DB-CNN (Zhang et al. 2020), HyperIQA (Su et al. 2020), UNIQUE (Zhang et al. 2021), HFF (Sun et al. 2021), and MMMNet (Li et al. 2021).

Table 1 lists the results of SROCC, PLCC, KROCC, and RMSE on the synthetic databases LIVE, TID2008, and TID2013. It can be seen that our algorithm is superior to all of the compared FR-IQA and NR-IQA methods. Table 2 lists the results of SROCC and PLCC on the synthetic databases LIVE, CSIQ, and TID2013 and average values of the three databases. As shown in Table 2, the performance of the proposed network is the best and most competitive when compared to other advanced models, except for the TID2013 database. The proposed model has the best performance after averaging the SROCC and PLCC on the three synthetic databases, indicating that it has strong generalization ability and is superior to other NR-IQA algorithms.

To demonstrate that the proposed network has good generalization ability, we tested it and other algorithms (PSNR (Wang and Bovik 2002), FSIM (Zhang et al. 2011), BRISQUE (Anish et al. 2012), GM-log (Xue et al. 2014), CNN (Kang et al. 2014), BIECON Kim and Lee (2017)) in each distortion of the LIVE, CSIQ, and TID2008 databases. The experimental results are shown in Table 3. We can see that the obtained results of our network on SROCC are better than other networks in all distortion types for these databases. It demonstrates that the network is relatively effective in evaluating these distorted images, indicating that our algorithm has good generalization ability.

**Evaluation of in-the-wild databases.** The proposed algorithm and six popular IQA algorithms were trained on distorted images from three in-the-wild IQA databases. These algorithms include QAC (Xue et al. 2013), NIQE (Mittal et al. 2012), IL-NIQE (Zhang et al. 2015), FRIQUEE (Ghadiyaram and Bovik 2017), BIECON (Kim and Lee 2017), ResNet-ft (Kim et al. 2017).
tal et al. 2012), BRISQUE (Anish et al. 2012) and BMPRI (Min et al. 2018), CNN (Kang et al. 2014), and WaDIQaM-NR (Bosse et al. 2017). Table 4 shows the comparative results on three in-the-wild databases. It can be seen that although the performance of our proposed algorithm on the three in-the-wild IQA databases was not always the best, it did achieve better performance results, which indicates that our proposed model has good generalization ability on in-the-wild IQA databases. The best performance results for the IQA models in the table are shown in bold, while the second-best performance model is shown in italics.

### 4.4 Cross-database evaluations

To test the generalization ability of the proposed network model on different databases, this section trained the proposed network model on three synthetic databases LIVE, CSIQ, and TID2013 and one in-the-wild IQA database LIVE Challenge, and tested network performance on the other databases. The proposed mode was compared with BRISQUE (Anish et al. 2012), GM-LOG (Xue et al. 2014), FRIQUEE (Ghadiyaram and Bovik 2017), CORNIA (Ye et al. 2012), HOSA (Xu et al. 2016), WaDIQaM-NR (Bosse et al. 2017), DB-CNN (Zhang et al. 2020), and MMMNet (Li et al. 2021). Table 5 displays the SROCC results. To make the comparison more intuitive, we drew a visualization result diagram, as shown in Figs. 10, 11, 12, and 13. As given in Table 5 and Figs. 10, 11, 12, and 13, the generalization ability of the proposed model is better than the most advanced NR-IQA models in such challenging experiments, demonstrating the effectiveness of the proposed method.

### 4.5 Ablation study

In this section, we performed ablation experiments on the proposed network modules to verify whether they were helpful to network performance.

The size of the input patch was first compared. We used patches of $16 \times 16$, $32 \times 32$, and $64 \times 64$ sizes for the experiments while ensuring that other modules remained unchanged, and the results are shown in Table 6 and Fig. 14. It was seen that the evaluation effect was best when the patch size was $32 \times 32$. As a result, in our experiments, the input patch size was set to $32 \times 32$.

To verify the effectiveness of the proposed MSRB and CAM, a comparative experiment was then carried out on the LIVE database to determine whether the use of MSRB and CAM was effective for our network. The results are shown in Table 7 and Fig. 15. From the results, we can conclude that the models with feature extraction blocks MSRB and CAM are better than those without MSRB and CAM. That is, multi-scale feature extraction and channel attention mechanism indeed improved the evaluation results of the proposed model.

| Databases | NIQE Mittal et al. (2012) | BRISQUE Anish et al. (2012) | BMPRI Min et al. (2018) | CNN Kang et al. (2014) | WaDIQaM-NR Bosse et al. (2017) | MsRCANet | LIVE Challenge SROCC |
|-----------|--------------------------|-----------------------------|--------------------------|--------------------------|-----------------------------|----------|-----------------------|
| KonIQ10k  | 0.601                    | 0.523                       | 0.658                    | 0.684                    | 0.702                       | 0.697    | 0.706                 |
| SROCC     | 0.487                    | 0.523                       | 0.658                    | 0.684                    | 0.702                       | 0.697    | 0.706                 |
| PLCC      | 0.523                    | 0.685                       | 0.796                    | 0.857                    | 0.806                       | 0.857    | 0.857                 |
| SROCC     | 0.523                    | 0.685                       | 0.796                    | 0.857                    | 0.806                       | 0.857    | 0.857                 |
| PLCC      | 0.523                    | 0.685                       | 0.796                    | 0.857                    | 0.806                       | 0.857    | 0.857                 |
| SPQA      | 0.523                    | 0.685                       | 0.796                    | 0.857                    | 0.806                       | 0.857    | 0.857                 |
| Weighted average | 0.523 | 0.685 | 0.796 | 0.857 | 0.806 | 0.857 | 0.857 |
Table 5 SROCC results in the cross-database setting

| Training          | Testing          | LIVE challenge | CSIQ challenge | TID2013 challenge | LIVE challenge | CSIQ challenge | TID2013 challenge | LIVE challenge | CSIQ challenge | TID2013 challenge |
|-------------------|------------------|----------------|----------------|-------------------|----------------|----------------|-------------------|----------------|----------------|-------------------|
| BRISQUE-Andiaz et al. (2012) | 0.562 | 0.358 | 0.337 | 0.847 | 0.562 | 0.454 | – | – | – | – |
| GM-LOG-Xue et al. (2014) | 0.601 | 0.344 | 0.183 | 0.879 | 0.601 | 0.344 | – | – | – | – |
| FRQUEE-Ghadiyaram and Bovik (2017) | 0.722 | 0.461 | 0.264 | 0.853 | 0.722 | 0.461 | – | – | – | – |
| CORNIA-Ye et al. (2012) | 0.649 | 0.360 | 0.443 | 0.833 | 0.649 | 0.360 | – | – | – | – |
| HOSA-Xue et al. (2016) | 0.594 | 0.361 | 0.463 | 0.831 | 0.594 | 0.361 | – | – | – | – |
| WADQM-NR-Bosse et al. (2017) | 0.704 | 0.462 | 0.392 | 0.846 | 0.704 | 0.462 | – | – | – | – |
| MMNet-Li et al. (2021) | 0.758 | 0.524 | 0.567 | 0.853 | 0.758 | 0.524 | – | – | – | – |
| MsRCANet            | 0.921 | 0.663 | 0.952 | 0.921 | 0.663 | 0.952 | – | – | – | – |

Fig. 10 Results of training on the LIVE database and testing on the other three databases

Fig. 11 Results of training on the CSIQ database and testing on the other three databases

Fig. 12 Results of training on the TID2013 database and testing on the other three databases

neural network and are very effective for image quality evaluation.

5 Conclusion

In this paper, a novel no-reference quality assessment model MsRCANet based on CNN is proposed. This proposed model can extract contour and detail features, predict image quality scores better, and improve IQA accuracy. To better extract
multi-scale features and deal with the information flow between channels, we designed a multi-scale feature residual extraction block, in which active weighted residual learning mechanism is used to process the relevant information between channels. Specifically, parallel branches with three different convolution kernel sizes were first used to extract distorted images, and then, active weighted mechanism strategy (AWMS) was used to process input information. As different channels have different contributions to image features, we used the channel attention mechanism (CAM) to process the information of feature mapping in each channel and further applied the backpropagation algorithm to learn the dependence between channels in CAM. In this way, the mechanism can be used to deduce the corresponding weights of different channels, and then, the final output results through different channels can be obtained.

The performance of the proposed method was analyzed by experiments. The MsRCANet outperforms the most advanced NR-IQA methods in terms of image quality evaluation performance. Although MsRCANet performed well in IQA, there are still some issues worthy of further study. For example, how do we visualize extracted feature maps in the IQA process? How can we improve the performance of in-the-wild IQA databases better? In future research, we will concentrate on these issues to find more effective ways to implement IQA.

**Table 6** SROCC and PLCC of LIVE database in different patch sizes

| Size   | SROCC  | PLCC  |
|--------|--------|-------|
| 16×16  | 0.956  | 0.961 |
| 32×32  | 0.981  | 0.984 |
| 64×64  | 0.951  | 0.947 |

**Table 7** SROCC and PLCC of LIVE database in different components

| MSRB | CAM | SROCC | PLCC |
|------|-----|-------|------|
| ×    | ×   | 0.952 | 0.948 |
| ✓    | ×   | 0.960 | 0.963 |
| ×    | ✓   | 0.959 | 0.955 |
| ✓    | ✓   | 0.981 | 0.984 |

**Fig. 13** Results of training on the LIVE Challenge database and testing on the other three databases

**Fig. 14** SROCC and PLCC of LIVE database in different patch sizes

**Fig. 15** SROCC and PLCC of LIVE database in different components

**Funding** Funding was provided by the National Natural Science Foundation of China (Grant Numbers 61976027, 61572082) and the Liaoning Revitalization Talents Program (Grant Number XLYC2008002).

**Data availability** Enquiries about data availability should be directed to the authors.

**Declaration**

**Conflict of interest** The authors have not disclosed any competing interests.

**References**

Anish M, Anush K, Alan Bovik C (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 12:4695–4708

Bosse S, Maniry D, Miller K et al (2017) Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans Image Process 27(1):206–219
Bosse S, Maniry D, Wiegand T, Samek W (2016) A deep neural network for image quality assessment. In: Proceedings of IEEE international conference on image processing, Phoenix, AZ, USA, Sep. pp 3773–3777

Chen D, Wang Y, Gao W (2020) No-reference image quality assessment: an attention driven approach. IEEE Trans Image Process 29(99):6496–6506

Cheng Z, Takeuchi M, Katto J (2017) A pre-saliency map based blind image quality assessment via convolutional neural networks. In: Proceedings of IEEE international symposium on multimedia (ISM), pp 77–82

Chen P, Niu Y, Huang D (2019) No-reference image quality assessment based on multi-scale convolutional neural networks. In: Intelligent computing-proceedings of the computing conference. Springer, Cham, pp 1202–1216

Chen X, Zhang Q, Lin M et al (2019) No-reference color image quality assessment: from entropy to perceptual quality. J Image Video Proc 77:1–14

Dash PP, Wong A, Mishra A (2017) VeNICE: A very deep neural network approach to no-reference image assessment. In: Proceedings of the IEEE international conference on industrial technology (ICIT), pp 1091–1096

Deepti G, Alan C (2015) Massive online crowdsourced study of subjective and objective picture quality. IEEE Trans Image Process 25(1):372–387

Dendi S, Dev C et al (2019) Generating image distortion maps using convolutional autoencoders with application to no reference image quality assessment. IEEE Signal Process Lett 26(1):89–93

Fang Y, Zhu H, Zeng Y et al (2020) Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE conference on computer vision and pattern Recognition, pp 3677–3686

Gao Z, Xie J, Wang Q et al (2020) Global second-order pooling convolutional networks[C]. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE

Ghadiyaram D, Bovik AC (2017) Perceptual quality prediction on authentically distorted images using a bag of features approach. J Vis 17:32

Hao S, Guo Y, Wei Z et al (2019) Lightness-aware contrast enhancement for images with different illumination conditions. Multimed Tools Appl 78:3817–3830

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Holzinger A (2018) From machine learning to explainable AI. In: World symposium on digital intelligence for systems and machines (DISA), pp 55–66

Holzinger A (2021) Explainable AI and multi-modal causability in medicine. i-com 19(3):171–179

Hong H et al (2016) Image detail enhancement with spatially guided filters. Signal Process Official Publ Eur Assoc Signal Process 120:789–796

Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision, pp 7132–7141

Hyoungho J, Ryong L, Sanghwan L et al (2018) Residual convolutional neural network revisited with active weighted mapping

Jin X, Wu L, Li X, et al (2016) ILGNet: inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation. IET computer vision, pp 1–6

Kang L, Ye P, Li Y, Doermann D (2014) Convolutional neural networks for no-reference image quality assessment. Proceedings of IEEE conference on computer vision and pattern recognition, Jun. 2014:1733–1740

Kang L, Ye P, Li Y, Doermann D (2015) Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: Proceedings of IEEE International Conference on Image Processing, pp 2791–2795

Kim J, Lee S (2017) Fully deep blind image quality predictor selected topics in signal processing. IEEE J 11(1):206–220

Kim J, Lee S (2017) Fully deep blind image quality predictor. IEEE J Select Topics Signal Process 11(1):206–220

Kim J, Zeng H, Ghadiyaram D, Lee S, Zhang L, Bovik AC (2017) Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. IEEE Signal Proc Mag 34(6):130–141

Kim J, Nguyen A, Ahn S, Luo C, Lee S (2018) Multiple level feature-based universal blind image quality assessment model. In: Proceedings ICIP, pp 291–295

Larson C, Chandler M (2010) Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electron Imag 19(1):011006

Li D, Jiang T, Lin W, Jiang M (2018) Which has better visual quality: The clear blue sky or a blurry animal? IEEE Trans Multimed 21(5):1221–1234

Li F, Fang K, Mei G, Zhang (2018) Multi-scale residual network for image super-resolution, In: Proceedings of European conference on computer vision, pp 527–542

Lin K, Wang G (2018) Hallucinated-IQA: no-reference image quality assessment via adversarial learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Li Y, Po L-M, Feng L, Yuan F (2016) No-reference image quality assessment with deep convolutional neural networks. In: Proceedings of IEEE international conference on digital signal processing (DSP), pp 685–689

Liu X, Bagdanov A (2017) RankIQA: Learning from rankings for no-reference image quality assessment

Li F, Zhang Y et al (2021) MMMNet: an end-to-end multi-task deep convolution neural network with multi-scale and multi-hierarchy fusion for blind image quality assessment. IEEE Trans Circuit Syst Video Technol

Min X, Zhai G, Gu K, Liu Y et al (2018) Blind image quality estimation via distortion aggravation. IEEE Trans Broadcast 64(2):508–517

Mittal A, Soundararajan R, Bovik AC (2012) Making a completely blind image quality analyzer. IEEE Signal Process Lett 20(3):209–212

Pan C, Xu Y, Yan Y, Gu K, Yang X (2016) Exploiting neural models for no-reference image quality assessment. In: Proceedings of visual communications and image processing, pp 1–4

Ponomarenko, N et al (2015) Image database TID2013: Peculiarities, results and perspectives. Signal Process, Image Commun, pp 57–77

Ponomarenko N, Lukin V, Zelensky A et al (2009) TID2008-a database for evaluation of full-reference visual quality assessment metrics. Adv Modern Radio Electron 10(4):30–45

Rajchel M, Oszust M (2020) No-reference image quality assessment of authentically distorted images with global and local statistics, Signal, Image and Video Processing, (SVIP)

Ren H, Chen D, Wang Y (2018) RAN4IQA: Restorative adversarial nets for no-reference image quality assessment. In: Proceedings of the AAAI conference on artificial intelligence, pp 7308–7314

Sheikh H, Sabir M, Bovik A (2006) A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process 15(11):3440–3451

Sun C, Li H, Li W (2016) No-reference image quality assessment based on global and local content perception. In: Proceedings of visual communications and image processing, pp 1–6

Sun W, Min X, Zhai G, Ma S (2021) Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training

Sun W, Wang T, Min X et al (2021) Deep learning based full-reference and no-reference quality assessment models for compressed UGC
videos. In: IEEE international conference on multimedia expo workshops (ICMEW) IEEE

Su S, Yan Q, Zhu Y et al (2020) Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3667–3676

Vlad H, Hanhe L et al (2020) KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment. IEEE Trans Image Process 29:4041–4056

Wang Z, Bovik AC (2002) A universal image quality index. IEEE Sign Process Lett 9(3):81–84

Wang Z, Bovik A, Sheikh H et al (2004) Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612

Woo S, Park J, Lee J et al (2018) CBAM: convolutional block attention module[J]. Springer, Cham

Wu J, Ma J, Liang F, Dong W, Shi G, Lin W (2020) End-to-end blind image quality prediction with cascaded deep neural network. IEEE Trans Image Process 99:7414–7426

Wu J, Zhang M, Li L, Dong W, Lin GW (2019) No-reference image quality assessment with visual pattern degradation. Inf Sci 504:487–500

Xue W, Mou X, Zhang L et al (2014) Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. IEEE Trans Image Process 23(11):4850–4862

Xue W, Zhang L, Mou X (2013) Learning without human scores for blind image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern Recognition, pp 995–1002

Xu J, Ye P, Li Q, Du H, Liu Y, Doermann D (2016) Blind image quality assessment based on high order statistics aggregation. IEEE Trans Image Process 25(9):4444–4457

Yang Q, Gong D, Zhang Y (2019) Two-stream convolutional networks for blind image quality assessment. IEEE Trans Image Process 28(5):2200

Yang S, Jiang Q, Lin W, Wang Y (2019) SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In: Proceedings of ACM international conference on multimedia association for computing machinery, pp 1383–1391

Ye P, Kumar J, Kang L, Doermann D (2012) Unsupervised feature-learning framework for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1098–1105

Zhang L, Mou X et al (2011) FSIM: a feature similarity index for image quality assessment. IEEE Trans Image Process 20(8):2378–2386

Zhang L, Zhang L, Bovik AC (2015) A feature-enriched completely blind image quality evaluator. IEEE Trans Image Process 24(8):2579–2591

Zhang W, Ma K, Yan J et al (2020) Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans Circuits Syst Video Technol 30(1):36–47

Zhang Y, Li K et al (2018) Image super-resolution using very deep residual channel attention networks

Zhang W, Ma K, Zhai G et al (2021) Uncertainty-aware blind image quality assessment in the laboratory and wild. IEEE Trans Image Process 30:3474–3486

Zhang W, Qu C, Ma L et al (2016) Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. Pattern Recogn 59:176–187

Zhang L, Shen Y, Li H (2014) VSI: A visual saliency-induced index for perceptual image quality assessment. IEEE Trans Image Process 23(10):4270–4281

Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2472–2481

Zuo L, Wang H, Fu J (2016) Screen content image quality assessment via convolutional neural network. In: Proceedings of the IEEE international conference on image processing (ICIP), pp 2082–2086

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.