Computer Modeling in Engineering & Sciences CMES, vol.123, no.1, pp.201-216, 2020

Towards No-Reference Image Quality Assessment Based on Multi-Scale Convolutional Neural Network

Yao Ma[1], Xibiao Cai[1, *] and Fuming Sun[2]

Abstract: Image quality assessment has become increasingly important in image quality monitoring and reliability assuring of image processing systems. Most of the existing no-reference image quality assessment methods mainly exploit the global information of image while ignoring vital local information. Actually, the introduced distortion depends on a slight difference in details between the distorted image and the non-distorted reference image. In light of this, we propose a no-reference image quality assessment method based on a multi-scale convolutional neural network, which integrates both global information and local information of an image. We first adopt the image pyramid method to generate four scale images required for network input and then provide two network models by respectively using two fusion strategies to evaluate image quality. In order to better adapt to the quality assessment of the entire image, we use two different loss functions in the training and validation phases. The superiority of the proposed method is verified by several different experiments on the LIVE datasets and TID2008 datasets.

Keywords: Image pyramid, global information, local information, image distortion.

1 Introduction

In the image transmission process, there is a variety of distortion and degradation to reduce image quality, which further affects the accuracy and adequacy of the obtained information. Therefore, it is crucial to establish an effective image quality assessment mechanism. Image quality assessment algorithms are generally classified into three categories according to whether reference images are available for the comparison: full-reference image quality assessment (FR-IQA) algorithms, reduced-reference image quality assessment (RR-IQA) algorithms, and no-reference image quality assessment (NR-IQA) algorithms. The FR-IQA algorithms use all the information of the non-distorted reference image to evaluate image quality scores. Classical FR-IQA algorithms mainly include the SSIM

[1] School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou, 121001, China.
[2] School of Information and Communication Engineering, Dalian Minzu University, Dalian, 116600, China.
* Corresponding Author: Xibiao Cai. Email: lgcaixb@163.com.
Received: 05 July 2019; Accepted: 25 September 2019.

[Wang, Bovik, Sheikh et al. (2004)], the GMSD [Xue, Zhang, Mou et al. (2014)], the FSIM [Zhang, Zhang, Mou et al. (2011)], and many others. The RR-IQA algorithms use only partial information about the reference image as a reference to evaluate image quality scores, and the NR-IQA algorithms evaluate image quality without any reference-image information. However, in practical applications, it is often difficult to obtain a non-distorted reference image. Therefore, no-reference image quality assessment methods have been attracting more attention.

Due to the usage of reference images, the FR-IQA has put forward a lot of work based on the research of human visual system (HSV), but it becomes difficult for NR-IQA. The NR-IQA [Saad, Bovik and Charrier (2012); Mittal, Moorthy and Bovik (2012); Li, Zhu and Qian (2014)] starts from the distorted image and extracts the features that can represent the distorted characteristics. Saad et al. [Saad, Bovik and Charrier (2012)] stated that the degree and type of image distortion are closely related to discrete cosine transformation (DCT) coefficients, so they extracted features in the DCT domain to predict the quality score. The aforementioned works have achieved some significant results, but the HSV is a very complex system, and it is very difficult to extract accurate and comprehensive features manually [Qiao, Xu and Yan (2020); Qiao, Wang and Xu (2020); Sun, Tang, Li et al. (2014)]. Thanks to the deep learning development, manually extracted features have been gradually replaced by the ones acquired by the feature learning methods. For instance, in Bare et al. [Bare, Li and Yan (2017)], an end-to-end method without any handcrafted features was directly used to implement the NR-IQA. The reliability map was used in Kim et al. [Kim, Nguyen and Lee (2018)] as a middle target to learning changes caused by distortion.

However, there are two main problems that need to be addressed. On the one hand, for the whole image, the distortion effect is often reflected in a change in high-frequency details, which is only a small part of the information contained in the whole image; it has a little effect on the energy-intensive low-frequency information, as demonstrated in Fig. 1. As can be seen in the Fig. 1, the more severe the image distortion is, the more severe the loss of high-frequency information is, but the low-frequency information is almost unchanged.

By using the black-box pool learning method, the problem is analyzed from the pixel value of the image without considering the process of human's observation of an image. Moreover, it is easy to focus only on the energy-intensive outline information, ignoring the lost realistic texture, which greatly increases the learning difficulty of a network. This phenomenon makes fully learning of distortion characteristics from small datasets such as IQA very difficult. Therefore, we propose a method of combining multi-scale with CNNs. The image pyramid fully simulates the HSV process, reduces the details in the down-sampling process, and shows the image at different resolutions from a coarse level to a fine level. It is well known that multi-scale input can merge the details of images with different resolutions and can combine global information with local information. The multi-scale input is very helpful for image feature extraction so that it is applied to various aspects related to image processing. By inputting four different scales of images

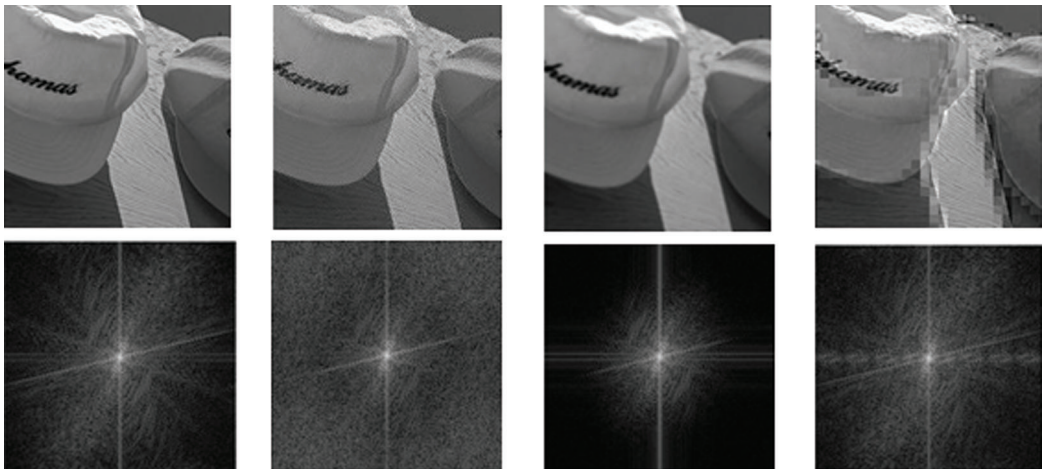

Figure 1: Images with different distortion types and their corresponding spectrograms. The first column shows the non-distorted reference image and its corresponding spectrogram. From left to right different distortion types are represented and distortion degree is gradually increased

into a network simultaneously, the reference object, i.e., the low-frequency information, can be acquired in the learning process because it is almost unchanged under the influence of distortion. Therefore, it will be easier for the network to learn the changes in high-frequency information caused by distortion, which greatly reduces the difficulty of learning and improves the learning accuracy. At the same time, in order to utilize the multi-scale input better, we designed two different fusion methods in the network, the results showed good competitiveness.

On the other hand, because the prediction ability of the CNN models heavily depends on the image datasets used, it is difficult to obtain stable and excellent results without the training with a sufficiently large amount of data. However, the existing available IQA datasets are not sufficient to support the training of the network models with a large number of parameters. In order to overcome this problem, researchers [Kang, Ye, Li et al. (2014); Dash, Wong and Mishra (2017); Dendi, Dev and Kothari (2018)] have used various methods. For instance, Kang et al. [Kang, Ye, Li et al. (2014)] split the entire image into 32×32 image patches without overlapping to expand the datasets. Dash et al. [Dash, Wong and Mishra (2017)] used a pre-trained network which enabled certain recognition capabilities to reduce the need for learning space.

To solve this problem, we adopt a method similar to the method presented in Kang et al. [Kang, Ye, Li et al. (2014)]. However, the method used in Kang et al. [Kang, Ye, Li et al. (2014)] has a problem that a network trained with the expanded datasets is difficult to adapt to the label in the original datasets. Different from the previous works, in this work, we use two different loss functions in the training and validation training and

validation. In the training, we use the quality score of small image patches as a target, that is, each image patch corresponds to a label, and the extended dataset for supervised training. In the validation, we use the quality score of the entire image as a target, that is, each minibatch corresponds to a label. After the patches obtained by dividing an image are grouped into a minibatch, we average the predicted scores for each group as an output to fine-tune the parameters. By alternating the two mentioned stages, the network can meet the requirements for image quality assessment better.

The advantages of the proposed method are three-fold.

1. A multi-scale-input method is adopted. Both local and global information of an image is extracted as an input of the network with multi-scale characteristics.
2. Two fusion strategies are provided and compared. One strategy fuses the predicted results of each scale, and the other fuses the features extracted from the convolution layer and pooling layer. Accordingly, two network models are given, Multiple Scales Concat (MS-C) and Multiple Scales Fusion (MS-F), and experiments show that each network has own advantages in different situations.
3. Different loss functions are used. Owing to the form of label in train dataset and validation dataset are different, we use two different loss functions to make the evaluation of the entire image quality score better by migrating the evaluation ability of an image patch.

The rest of this paper is organized as follows. Section 2 reviews the related works. Section 3 introduces the proposed method, including the image pyramid, normalized processing, network structure, fusion method, and loss functions. Section 4 conducts different experiments. Finally, Section 5 concludes this work and discusses our future work.

2 Related work

The NR-IQA methods can be roughly categorized into two groups, namely statistical characteristics methods and feature learning methods. The former methods unfold the statistical properties of distorted images and non-distorted images. For instance, the BRISQUE proposed in Moorthy et al. [Moorthy and Bovik (2011)] shows that the statistical characteristics of natural images are similar to the generalized Gaussian distribution (GGD), which is warped by distortion. Therefore, the statistical characteristics are used to distinguish distortion types and evaluate distorted-image quality. Jiang et al. [Jiang and Zhou (2016)] performed structural similarity calculations on multiple scales and calculated the structural similarity of the two inputs according to the certain weights. Xu et al. [Xu, Lu and Ren (2015)] extracted mutual information between the images and pixels to capture pixel changes caused by distortion. All these algorithms are based on a hypothesis that distortion changes some statistical characteristics of natural images, but it is difficult to determine these characteristics comprehensively and accurately manually.

With deep learning development, feature learning gradually replaced manual feature extraction. The latter method obtains learning features instead of hand-crafted feature extraction via deep learning. Bianco et al. [Bianco, Celona and Napoletano (2018)] used different methods to

analyze and verify each component of the network to make the network learn distortion features better. Hou et al. [Hou, Gao, Tao et al. (2017)] proposed a blind IQA model that classifies distortion into five levels for qualitative analysis and predicts numerical scores by regression. Kim et al. [Kim and Lee (2017)] used four classical FR-IQA algorithms to get the quality score of image patches, which was further used as ground truth for supervised learning. Bosse et al. [Bosse, Maniry, Müller et al. (2018)] also determined the image quality score directly from the image patches but combined different patches weights to predict the image quality score. Similarly, Oh et al. [Oh, Ahn, Kim et al. (2017)] used deep learning on 3D images, and the difference was that the left image and the right image were fused by the CNNs. In Fan et al. [Fan, Zhang, Feng et al. (2018)], the authors input images into multiple networks at the same time. Since each network related to a specific distortion type, the output of each network represented the correlation between the image and specific distortion type. All of these methods adopt some data enhancement methods to expand the dataset to meet the requirement for data amount needed for adequate CNN training, but two problems about image-wise and patch-wise need to be overcome.

Hitherto, some researchers have adopted unsupervised learning methods. Because artificial subjective scores are no longer used, datasets can be extended from specific datasets to a wide variety of readily available datasets, or even to image datasets obtained by generated and distorted. In view of the sequential learning algorithm (L2R), Ma et al. [Ma, Liu, Liu et al. (2017)] collected distorted images from the real world and used unsupervised learning to realize the NR-IQA. Liu et al. [Liu, Weijer and Bagdanov (2017)] used a series of image pairs with known rank order to train the network so that the network had the ability to distinguish image quality. Due to the lack of reference images, unsupervised learning can difficultly quantify the degree of learning distortion as a score, resulting in a low final measurement score. Recently, the Generative Adversarial Networks (GAN) network have been used in the NR-IQA. Some researchers [Lin and Wang (2018); Pan, Shi, Hou et al. (2018)] used the GAN network to generate the available image to compensate for miss reference image in the dataset; for instance, the Hallucinated-IQA [Lin and Wang (2018)]. Although the generated reference image is difficult to be false, it also provides a new idea for the research in the NR-IQA field.

3 Multi-scale convolutional neural network for NR-IQA

3.1 Image pyramid

Image pyramids are a way of expressing images with multiple scales. For an image, the shorter the distance between the observer and the image is, the clearer the image is, so the more detailed content of the image the observer see. Conversely, the farther the observer is, the more blurred the image is, so the observer an only see the outline of the image. Because of the mentioned, the image pyramid expresses the distance between the observer and the image as an image scale. Due to the existence of this scale, we can convert a two-dimensional image into three-dimensional space to analysis and extract its intrinsic characteristics. For the original non-distorted image, when it is looked carefully

from the whole to the part, it can be seen that the new details basically coincide with the rough ones [Li, Zhu and Qian (2014)]. However, distorted images tend to produce different results in content when looking roughly and closely. The comparison of image local information between the clean image and noisy image is shown in Fig. 2.

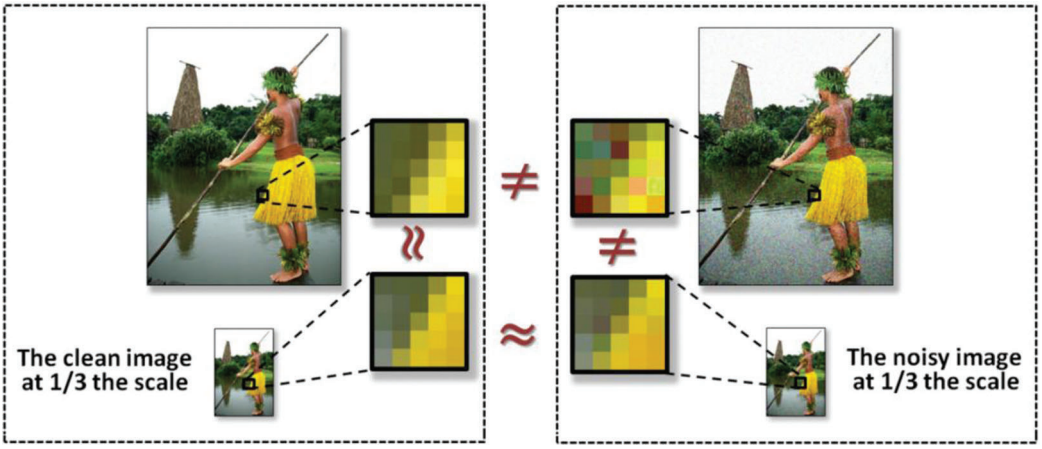

Figure 2: Figure from Li et al. [Li, Zhu and Qian (2014)]. Comparison of image local information at 1/3 the scale. (a) Clean image. (b) Noisy image

In Fig. 2, it can be seen that there are obvious differences in details between the distorted and non-distorted images. Therefore, the NR-IQA method cannot achieve satisfactory results by using only global information. To address this issue, we use both local and global information of images. In order to extract scale invariant features of distorted images, we first use multi-scale images as an input of a neural network. In this way, global information is obtained without losing local information. The method of obtaining a multi-scale image adopted in this paper is similar to the Gaussian pyramid. This is mainly due to two reasons. On the one hand, it does not introduce other noise in the convolution process because Gaussian kernels are linear. On the other hand, the statistical characteristics of the distorted and non-distorted images keep unchanged using the Gaussian kernels since the data distribution of natural images is similar to the generalized Gaussian distribution [Mittal, Moorthy and Bovik (2012)]. The Gaussian convolution function we use in this work is given by:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e \frac{(x-x_0)^2 + (y-y_0)^2}{2\sigma^2} \tag{1}$$

As for the parameter σ, the experimental results showed that the result is the best at a fixed value of 1.6. At first, we regard the original image as an input of the first layer. Through down-sampling the convolution result of the original image and Gaussian kernels, a new image that is 1/4 of the original image is obtained. Then, the new image is reshaped to

the original size and used as an input of the second layer. Comparing the two layers' images at the same size at the pixel level, we can find that there are obvious differences between them. We can denote this difference as a result of image distortion. In this way, we have four scale images as a network input.

3.2 Normalized processing

Before performing the convolution operation on the original image and Gaussian kernels, we first perform a simple normalization of the image similar to that in Kang et al. [Kang, Ye, Li et al. (2014)] and use the mean and variance to decorrelation. Consequently, the effect of image redundancy features that are weakly related to the image quality can be eliminated. The local normalization method we use is as follows:

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{2}$$

$$\mu(i,j) = \sum_{p=-P}^{\mathrm{p=P}} \sum_{q=-Q}^{q=Q} I(i+p, j+q) \tag{3}$$

$$\sigma(i,j) = \sqrt{\sum_{p=-P}^{p=P} \sum_{q=-Q}^{q=Q} (I(i+p, j+q) - \mu(i,j))^2} \tag{4}$$

where C is a positive constant that prevents dividing by zero, μ and σ are the mean and variance, respectively; P and Q are the normalization window size lengths and widths. In practical application, we take P and Q as fixed values of 3. The window size should be smaller than the image size, and the smaller the window size is, the better the normalization effect will be. The experimental results in Kang et al. [Kang, Ye, Li et al. (2014)] showed that the effect is the best when P and Q are fixed at 3.

3.3 Network structure and fusion method

The proposed network structure, including two fusion strategies, Multiple Scale Concat (MS-C) and Multiple Scale Fusion (MS-F), is shown in Fig. 3.

We also provide two fusion strategies. In one, the predicted results of all the scale are fused, and in the other, the features from the convolution layer are fused. The MS-C and MS-F networks share the CONV (convolution) layer. The CONV consists of one convolution layer and two pooling layers. There are 50 convolution kernels with the size of 7×7 and the stride of 1 in the convolution layer. A maximum pooling layer and a minimum pooling layer are used to reduce the feature dimensions. After implementing the Gaussian pyramid on the original image, four multi-scale images, showing different distortion degrees are obtained. Further, four feature vectors are generated after the convolution, i.e., *f1*, *f2*, *f3*, and *f4*. The convolution method we use is as follows:

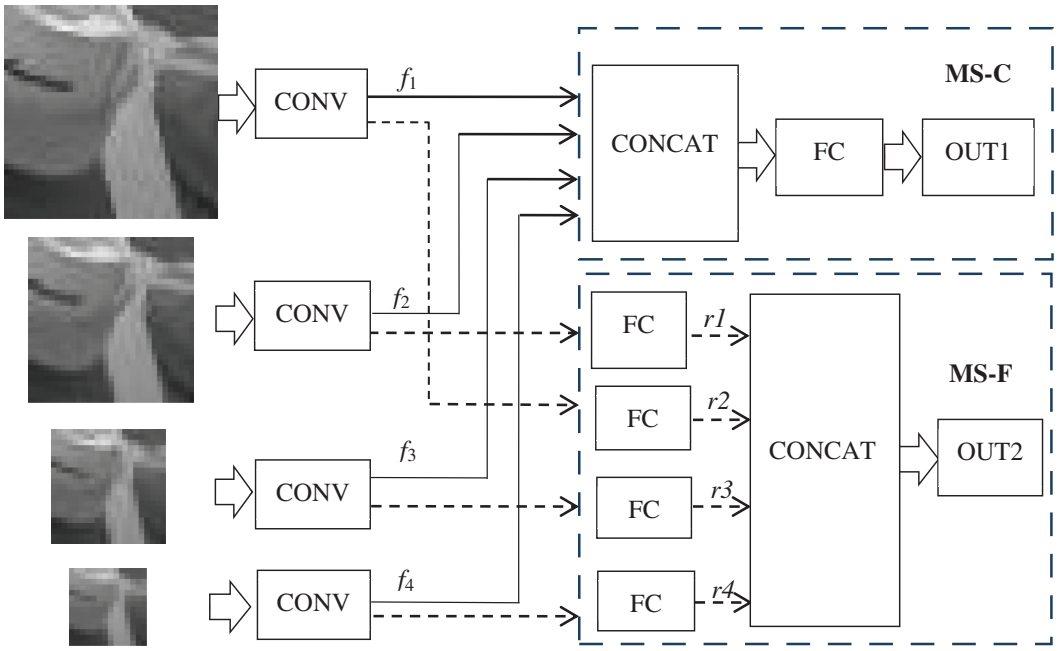

Figure 3: The proposed network structure

$$f_i = \max_{a,b}(b_k + \sum_k w_k^i * x_{a,b}^i) + \min_{a,b}(b_k + \sum_k w_k^i * x_{a,b}^i) \tag{5}$$

where * denotes convolution operations, f_i denotes feature vectors extracted by convolution and pooling at the scale of i, a and b denotes max and min values of location (a, b) of the feature map obtained by the i-th scales convolution, $x_{a,b}^i$ denotes location (a, b) of input images at i-th scales, w_k^i denotes weights of k convolutions on i-th scales and b_k denotes bias.

In the MS-C network, the four features are fused before the fully connected layer (FC), that is, *f=concat (f1, f2, f3, f4)*. Then, *f* is used as an input of the fully connected layer to obtain the image quality score. So, the MS-C connects four networks to form output with the size of 8×50; then follows the FC, which contains two fully connected layers consisted of 800 nodes. In the MS-F network, the four features are considered as four independent features, and the full connection layer four outputs, *r1*, *r2*, *r3*, and *r4* are obtained respectively. The four results are taken as an input of the penultimate layer, which makes the neural network optimize the weights of each layer by learning to predict the image quality score. So, the MS-F has four FC after the pooling layer. Finally, the predicted score is obtained by the last linear regression layer. The two networks types are compared by the experiments.

The proposed NR-IQA algorithm based on the multi-scale convolutional neural network is given in Algorithm 1.

Algorithm 1: Multi-scale convolutional neural network based NR-IQA algorithm

Input: Distortion image **I**

Output: The predicted quality score of a distorted image **q**

Procedure:

1: For input **I**, normalized according to Eq. (2);

2: Get four scale images by the image pyramid;

3: Input four scale images to the CONV layers to obtain four feature vectors f_1, f_2, f_3, and f_4 on Eq. (5)

MS-C:

1. f =concat(f_1,f_2,f_3,f_4), fusion of the features f_1, f_2, f_3, and f_4
2. Pass f through the FC layers
3. Last fully connected layer regression output q

MS-F:

1. Input f_1, f_2, f_3, f_4 into the FC layers, respectively, and obtain four results r_1, r_2, r_3, r_4.
2. Combine r_1, r_2, r_3, r_4 to the final output layer
3. Last fully connected layer regression output q

3.4 Loss functions

When the small IQA datasets are extended into new datasets by the non-overlapping segmentation, in this work, it is assumed that the quality score of each patch is equal to that of its source. When the network is trained with image-patch datasets, it will finally learn to evaluate the quality score of patches. But when obtained the label of patches, the distortion degree of some small patches is not completely equal to the entire image, so there will be some deviation. Therefore we adopt two different loss functions to correct the error caused by this assumption.

In the training dataset, all the image patches correspond to a label of their own, after disorder were used, and the quality score of each image patch was evaluated. In the validation dataset, the quality scores of image patches were not our target, the entire image correspond to a label. Since all the image patches were obtained by the non-overlapping segmentation of the entire image, the quality score of the entire image was evaluated when all the patches from that image were taken as an input. Therefore, the loss functions for training and validation were different. All the image patches were used during the training. In the validation, the patches obtained from one image were grouped into a minibatch, and the average value of minibatch was used as a label of that image. Therefore, the loss function $L1$ during the training, and the loss function L_2 during the validation were respectively defined as follows:

$$L_1 = \frac{1}{N}\sum_{n=1}^{N} ||f(x_n; w) - y_n||_{l1} \tag{6}$$

$$L_2 = \frac{1}{N}\sum_{n=1}^{N} ||\frac{1}{M}\sum_{i=1}^{M} f(x_n;w) - y_n||_{l1} \tag{7}$$

where x_n and y_n denoted the input patch image and its ground truth score, respectively; $f(x_n; w)$ denoted the predicted score of x_n with network weight w, N was the batch size of 128, and M denoted the non-overlapping patches number of the image.

4 Experimental results and analysis

4.1 Datasets

The following two datasets were used in experiments.

LIVE: This dataset consisted of a total of 779 distorted images, including 29 reference images and five types of distortion (gblur, jp2k, jpeg, wn, and fast fading).

TID2008: This dataset consisted of a total of 1700 distorted images, including 25 reference images and 17 types of distortion. Four of these 17 distortion types were common to the LIVE dataset, namely gblur, jp2k, jpeg, wn. The experiments were mainly performed for the four common types of distortion.

We divided the datasets into the training set, validation set, and test set, which were used in the experiment. The ratio of the training set was 0.6, and the ratio of the validation set and test set was the same and equal to 0.2. We set the batch size to 128, and the learning rate to 0.0001.

4.2 Gauss convolution

In the process of generating the multi-scale input images, the Gauss convolution was indispensable. In order to demonstrate the effect of parameters of the Gaussian convolution function, the results for different settings are shown in Tab. 1. As presented in Tab. 1, the larger the value of σ was, the larger the range of data that could affect it was. We compared six different values, σ=1.0, 1.2, 1.4, 1.6, 1.8, and 2.0. The experimental results show that adding the Gauss convolution to the image pyramid was effective, but different values did not result in a large difference. Overall, the result for σ of 1.6 was slightly better than others, so we fixed σ at 1.6.

Table 1: The results for different σ values

	N/A	**σ**					
		1.0	**1.2**	**1.4**	**1.6**	**1.8**	**2.0**
SROCC	0.9414	0.9530	0.9537	0.9553	0.9580	0.9528	0.9529
PLCC	0.9443	0.9676	0.9673	0.9688	0.9690	0.9578	0.9575

4.3 Evaluation on LIVE

The performance of the proposed method on the LIVE dataset was validated, and obtained results are given in Tab. 2. Tab. 2 shows the comparison results of the proposed method and the state-of-the-art NR-IQA methods.

Table 2: The SROCC and PLCC on LIVE dataset

SROCC	**JP2K**	**JPEG**	**WN**	**BLUR**	**FF**	**ALL**
BRISOUE	0.910	0.919	0.955	0.941	0.874	0.920
CORNIA	0.903	0.889	0.958	0.946	0.915	0.906
DLIQA	0.933	0.914	0.968	0.947	0.857	0.929
CNN	0.952	0.977	0.978	0.962	0.908	0.956
MS-F	0.957	0.961	0.991	0.979	0.866	0.940
MS-C	0.973	0.974	0.986	0.977	0.879	0.958
PLCC	**JP2K**	**JPEG**	**WN**	**BLUR**	**FF**	**ALL**
BRISOUE	0.936	0.937	0.958	0.935	0.898	0.917
CORNIA	0.915	0.902	0.952	0.940	0.913	0.903
DLIQA	0.953	0.948	0.961	0.950	0.892	0.934
CNN	0.953	0.981	0.984	0.953	0.933	0.953
MS-F	0.948	0.973	0.992	0.972	0.852	0.933
MS-C	0.963	0.980	0.968	0.975	0.870	0.969

We compared the proposed method with the four representative NR-IQA methods: the BRISOUE, the CORNIA, the DLIQA, and the CNN. In Tab. 2, it can be seen that the proposed method performed well for each distortion type. The CNN is the result of the method presented in Kang et al. [Kang, Ye, Li et al. (2014)]. It can be seen that in most cases, the results of our proposed method were superior to those of the CNN. Especially when the distortion types were blur and wn, the effects of both proposed networks were higher than of the others, and the accuracy of MS-F network achieved the highest. For the fast fading distortion type, the result was worse than the CNN. For the remaining two types of distortion, the result was almost the same as that of the other types. The full name of the BLUR was Gaussian Blur, that is, R, G, and B components were filtered using a circular-symmetric 2-D Gaussian kernel with the standard deviation sigma. The experimental results show that the input image obtained by the Gaussian pyramid method enhance the recognition ability of the distortion caused by the Gaussian noise. When the five types of distortion were used simultaneously in the training process, the MS-C performed better than the MS-F, and they both outperformed all the state-of-the-art NR-IQA methods.

4.4 Evaluation on TID2008

We conducted the experiments on the TID2008 dataset to further evaluate the proposed method. We divided the TID2008 dataset (containing 17 distortion types) into two parts; one included the same distortion types as those in the LIVE dataset, the other included the thirteen remaining distortion types. Tabs. 3 and 4 show the results of the MS-C and the MS-F, respectively.

Table 3: Results of the MS-C on TID2008 dataset

MS-C	WN	GBLUR	JPEG	JP2K	4-all	13-all	ALL
SROCC	0.941	0.953	0.947	0.946	0.950	0.753	0.782
PLCC	0.957	0.979	0.991	0.986	0.970	0.750	0.811

Table 4: Results of the MS-F on TID2008 dataset

MS-F	WN	GBLUR	JPEG	JP2K	4-all	13-all	ALL
SROCC	0.955	0.950	0.927	0.932	0.945	0.727	0.756
PLCC	0.966	0.970	0.986	0.978	0.959	0.719	0.798

We first separately trained the four distortion types shared by the TID2008 and LIVE datasets. Compared with the LIVE datasets, in the TID2008 datasets, the number of distortion types was larger, but the number of images contained in each distortion type was smaller. This reason makes the training of models more difficult so that the performance in TID2008 datasets was relatively reduced. However, compared with the other networks, we proposed method was still competitive. After the training with the four types together, the test results on the TID2008 datasets were very good. In Tabs. 3 and 4, "13-all" denotes the thirteen distortion types that LIVE did not include but which were included in the TID2008 dataset; "ALL" refers to the seventeen distortion types in the TID2008 dataset. In Tabs. 3 and 4, it can be seen that the recognition ability of distortion types that were not included in the LIVE dataset was the weakest. When one distortion type was tested, the accuracy of the two networks was almost the same, while multiple distortion types were tested together, such as "13-all" and "ALL", the accuracy of MS-C network was much higher than that of MS-F network. It can be seen that MS-C network has stronger ability to capture common characteristics of different distortion types. As the number of distortion types increased, the performance of the both networks decreased significantly. However, compared with the state-of-the-art NR-IQA methods, the experimental results of the proposed method were much competitive.

According to several different experiments, it can be seen that the predict ability of networks for each distortion type varies greatly, which shows that the characteristics of each distortion type are different. With the increase of distortion types, it becomes more and more difficult

to design a general method that is applicable to all distortion types. How to find a balance among different distortion types and find more representative common characteristics among them are the problem that we need to further study.

4.5 Cross-dataset evaluation

To verify the generalization ability of the proposed method, we conducted the cross-dataset experiment. The model was trained with the LIVE dataset and then tested on the TID2008 dataset. Tab. 5 shows the comparison results of different NR-IQA methods.

Table 5: The SROCC and PLCC obtained by the network trained on the LIVE dataset and tested on the TID2008 dataset

SROCC	**JP2K**	**JPEG**	**WN**	**BLUR**	**ALL**
BRISOUE	0.902	0.875	0.821	0.857	0.865
CORNIA	0.920	0.899	0.647	0.901	0.866
DVRM	0.943	0.930	0.909	0.733	0.894
MS-F	0.946	0.937	0.961	0.941	0.919
MS-C	0.945	0.941	0.958	0.938	0.920
PLCC	**JP2K**	**JPEG**	**WN**	**BLUR**	**ALL**
BRISOUE	0.908	0.909	0.812	0.855	0.873
CORNIA	0.904	0.928	0.642	0.887	0.878
DVRM	0.945	0.942	0.801	0.919	0.911
MS-F	0.925	0.982	0.962	0.978	0.913
MS-C	0.937	0.985	0.969	0.981	0.920

This experiment was designed such that to analysis the generalization capability of the proposed method. We compared our method with three representative NR-IQA methods: the BRISOUE, the CORNIA, and the DLIQA. Only the four types of distortions that were common for the LIVE and TID2008 datasets were examined in this experiment. At this test, the superiority of the proposed method was even more apparent. In this experiment, all the methods except for the proposed one were difficult to maintain the good performance as test on LIVE dataset, while the proposed method performed well. The experimental results show that on the TID2008 dataset test, the results did not fluctuate much compared with the results on the LIVE dataset, and it was even not inferior compared with the results in Tab. 2. And it can be seen that each network had own advantages in different situations. Therefore, it can be concluded that the proposed method showed a strong generalization ability. Compared with the other methods, both MS-F and MS-C networks achieved better results in the cross-dataset experiments.

5 Conclusion

In this paper, a strong multi-scale convolutional neural network for the no-reference image quality assessment is presented. The proposed method can study the unique properties of distorted images, and make the research work of NR-IQA even further. The experimental results show that the proposed method has very good performances both in terms of consistency with human subjective scores and generalization. However, as the characteristics of each distortion type are different, the recognition ability of the general network is still difficult to get rid of the distortion type constraint, so the results vary widely on different distortion types. This phenomenon is widespread in the related research work on the NR-IQA.

Unfortunately, it is expected that there will be even more distortion types in the future, so general methods will difficultly perform well for all distortion types. Moreover, in the real world, distorted images often contain many types of distortion, which makes it difficult for the existing methods to show excellent results on specific distortion types. Accordingly, the current NR-IQA methods cannot achieve high performances on real-life distortion pictures, and this will be the focus of our future work on the no-reference image quality assessment.

Acknowledgement: This work has been supported by the National Natural Science Foundation of China (Grant No. 61772171) and the Major Science and Technology Platform Project of the Normal Universities in Liaoning (Grant No. JP2017005). These supports are gratefully acknowledged.

Funding Statement: Xibiao Cai received these funds (Grant No. 61772171 and JP2017005).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Bare, B.; Li, K.; Yan, B. (2017). An accurate deep convolutional neural networks model for no-reference image quality assessment. *IEEE Conference on Multimedia and Expo*, *Hong Kong, China*. pp. 1356-1361.

Bianco, S.; Celona, L.; Napoletano, P. (2018): On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355-362. DOI 10.1007/s11760-017-1166-8.

Bosse, S.; Maniry, D.; Müller, K. R.; Wiegand, T.; Samek, W. (2018): Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206-219. DOI 10.1109/TIP.2017.2760518.

Dash, P. P.; Wong, A.; Mishra, A. (2017). VeNICE: a very deep neural network approach to no-reference image assessment. *IEEE International Conference on Industrial Technology*, *Toronto,Canada*. pp. 1091-1096.

Dendi, S. V. R.; Dev, C.; Kothari, N. (2018): Generating image distortion maps using convolutional autoencoders with application to no reference image quality assessment. *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 89-93. DOI 10.1109/LSP.2018.2879518.

Fan, C.; Zhang, Y.; Feng, L.; Jiang, Q. (2018): No reference image quality assessment based on multi-expert convolutional neural networks. *IEEE Access*, vol. 6, pp. 8934-8943. DOI 10.1109/ACCESS.2018.2802498.

Hou, W.; Gao, X.; Tao, D.; Liu, W. (2017): Blind image quality assessment via deep learning. *IEEE Transactions on Neural Networks & Learning Systems*, vol. 26, no. 6, pp. 1275-1286.

Jiang, Z.; Zhou, Y. (2016): Research on quality evaluation of printed images based on multi-scale structural similarity. *Packaging Engineering*, vol. 37, no. 9, pp. 134-137.

Kang, L.; Ye, P.; Li, Y.; Doermann, D. (2014). Convolutional neural networks for no-reference image quality assessment. *IEEE Conference on Computer Vision and Pattern Recognition*, *NW Washington, DC, USA*. pp. 1733-1740.

Kim, J.; Lee, S. (2017): Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206-220. DOI 10.1109/JSTSP.2016.2639328.

Kim, J.; Nguyen, A. D.; Lee, S. (2018): Deep CNN-based blind image quality predictor. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 11-24. DOI 10.1109/TNNLS.2018.2829819.

Li, L.; Zhu, H.; Qian, J. (2014): No-reference quality metric of blocking artifacts based on color discontinuity analysis. *IEICE Transactions on Information and Systems*, vol. E97.D, no. 4, pp. 993-997. DOI 10.1587/transinf.E97.D.993.

Lin, K. Y.; Wang, G. (2018). Hallucinated-IQA: no-reference image quality assessment via adversarial learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City, USA*. pp. 732-741.

Liu, X.; Weijer, J.; Bagdanov, A. D. (2017). RankIQA: learning from rankings for no-reference image quality assessment. *IEEE International Conference on Computer Vision*, *Venice, Italy.* pp. 1040-1049.

Ma, K.; Liu, W.; Liu, T.; Wang, Z.; Tao, D. (2017): dipIQ: blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951-3964. DOI 10.1109/TIP.2017.2708503.

Mittal, A.; Moorthy, A. K.; Bovik, A. C. (2012): No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695-4708. DOI 10.1109/TIP.2012.2214050.

Moorthy, A. K.; Bovik, A. C. (2011): Blind image quality assessment: from natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350-33641. DOI 10.1109/TIP.2011.2147325.

Oh, H.; Ahn, S.; Kim, J.; Lee, S. (2017): Blind deep s3d image quality evaluation via local to global feature aggregation. *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4923-4936. DOI 10.1109/TIP.2017.2725584.

Pan, D.; Shi, P.; Hou, M.; Ying, Z.; Fu, S. et al. (2018). Blind predicting similar quality map for image quality assessment. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City, USA*. pp. 6373-6382.

Qiao, L.; Xu, D.; Yan, Y. (2020): High-order ADI orthogonal spline collocation method for a new 2D fractional integro-differential problem. *Mathematical Methods in the Applied Sciences*. DOI 10.1002/mma.6258.

Qiao, L.; Wang, Z.; Xu, D. (2020): An alternating direction implicit orthogonal spline collocation method for the two dimensional multi-term time fractional integro-differential equation. *Applied Numerical Mathematics*, vol. 151, pp. 199-212. DOI 10.1016/j.apnum.2020.01.003.

Sun, F.; Tang, J.; Li, H.; Qi, G.; Huang, TS. (2014): Multi-label image categorization with sparse factor representation. *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1028-1037. DOI 10.1109/TIP.2014.2298978.

Saad, M. A.; Bovik, A. C.; Charrier, C. (2012): Blind image quality assessment: a natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339-3352. DOI 10.1109/TIP.2012.2191563.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. (2004): Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612. DOI 10.1109/TIP.2003.819861.

Xu, H.; Lu, W.; Ren, Y. (2015). Image quality assessment based on local pixel correlation. *CCF Chinese Conference on Computer Vision*, Berlin, Heidelberg: Springer, pp. 266-275.

Xue, W.; Zhang, L.; Mou, X.; Bovik, A. C. (2014): Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684-695. DOI 10.1109/TIP.2013.2293423.

Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. (2011): FSIM: a feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386. DOI 10.1109/TIP.2011.2109730.