Neurocomputing 431 (2021) 78–89

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep blind image quality assessment based on multiple instance regression

Dong Liang [a], Xinbo Gao [a,b,*], Wen Lu [a], Jie Li [a]

[a] School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China
[b] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

ARTICLE INFO

Article history:
Received 11 June 2020
Revised 24 October 2020
Accepted 5 December 2020
Available online 16 December 2020
Communicated by Steven Hoi

Keywords:
Image quality assessment
Multiple instance regression
Convolutional neural networks
Deep learning

ABSTRACT

In recent years, the research of image quality assessment (IQA) based on deep learning, especially convolutional neural network (CNN), has made rapid development. The most widely used way to build a CNN-based IQA model is to divide image into patches and conduct a patch-based training. However, this method has a critical defect that the local ground-truth for each patch is not available. This defect leads to a sub-optimal result. To address this issue, we propose a novel deep blind IQA algorithm under the multiple instance regression (MIR) framework. Specifically, we assume each instance (patch) has a certain probability to be the prime instance of the bag (image), which is responsible for the bag label. Then the global quality score of the bag can be computed by the weighted summation of the local quality scores of the instances, where the weights are the probabilities of the instances to be the prime one. To simplify the training procedure, we propose an EM-like algorithm, called conditional EM algorithm, to train the deep MIR IQA model. Experimental results show that the proposed deep MIR IQA algorithm performs better than the traditional deep blind IQA algorithm. Moreover, the proposed algorithm can be used as a unified framework to improve the performance of any patch-based deep IQA models.

1. Introduction

Digital images may suffer different distortions at different level during the acquisition, transmission, reproduction and processing in a system [1]. Image quality assessment (IQA) aims to evaluate the visual quality of an image from the human's point of view. Benefiting from IQA, one can further evaluate the performance of the algorithm, monitor the quality of the service (QoS) for the visual data business, and improve the effect of the image processing system. According to the availability of the reference image, the IQA algorithms can be grouped into three categories: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA) and no-reference IQA (NR-IQA or blind IQA). Since the reference information is not available in many applications, the blind IQA algorithms are the most promising technique and attract more and more attention recently.

The IQA algorithms can be categoried as traditional algorithm [2–12] and deep learning based algorithm [13–27]. Convolutional neural network (CNN) based IQA has developed rapidly in recent years because of its outstanding performance [15–27]. To build a CNN-based IQA model, the most widely used way is to divide image into patches and conduct a patch-based training. In the training stage, the quality score of an image is assigned to its patches as label. In the testing stage, the predicted quality scores of the patches are pooled to obtain the quality score of the image.

There are some reasons to conduct the patch-based training. One important reason is to augment the training dataset. The performance of deep learning heavily depends on huge training data. Unfortunately, the public IQA databases are not large enough because it is time consuming to collect IQA data [22]. For instance, the TID2013 database [28] for IQA only contains 3000 images. As a contrast, the ImageNet database [23] for image recognition contains more than 14 million training samples. Moreover, the effect of data augmentation based on image transformation is very limited. Except the horizontal reflection, all the other image transformations, such as rotations and vertical reflections, will significantly change the perceived quality of the distorted image [22]. Therefore, the major data augmentation for IQA is to divide the image into patches. Another important reason to conduct patch-based training is that many CNN models require fixed size input image due to the existing of the fully connected layers, which are widely used in the architecture of the CNN models [15–27].

However, the patch-based training has a critical defect that the local ground-truth for each patch is not available. Firstly, human perceives the visual quality by observing an image as a whole.

* Corresponding author.
E-mail addresses: dl_21century@163.com (D. Liang), xbgao@mail.xidian.edu.cn (X. Gao), luwen@mail.xidian.edu.cn (W. Lu), leejie@mail.xidian.edu.cn (J. Li).

https://doi.org/10.1016/j.neucom.2020.12.009
0925-2312/

The patch only provides a local scene. Therefore, local quality of the patch does not equals to the global quality of the image. Secondly, even with same distortion type and distortion level, the perception of the visual quality for different content will be different. Therefore, it is not reasonable to assign same label to all the patches in an image. Thirdly, the visual degradation does not distribute evenly in space. For instance, the distortion additive Gaussian White Noise (WN) in the flat region of an image looks more obvious than that in the texture region. Therefore, the patches with different textures will have different visual quality. Lastly, the patches in different images will have different contributions to the quality of their associated images. For above reasons, the traditional CNN-based IQA model were trained with approximated ground-truth, which definitely leads to a sub-optimal result.

Although the local ground-truth is not available, we know the membership between the image and its patches. This motivates us to address this issue by multiple instance learning (MIL). More specifically, we build the CNN-based IQA model by multiple instance regression (MIR) since the label of the IQA problem is a real value. As illustrated in Fig. 1, traditional CNN-based IQA algorithms belong to single instance regression (SIR) algorithm, which treats the patches as individual samples and assume the local ground-truth is same to the global ground-truth. As a contrast, the MIR IQA algorithm treats each distorted image as a bag and treats its patches as the instances of the bag. It keeps the reliable ground-truth at the bag level and trains IQA model based on the relationship between the bag and the instances.

In this paper, we proposed a novel algorithm to train CNN-based blind IQA model under the MIR framework. Specifically, we assume each instance (patch) has a certain probability to be the prime instance of the bag (image). The prime instance is responsible for the bag label (quality score). The contribution of each instance to the bag label is proportional to the prime instance probability. The quality score of the bag is predicted by the summation of the quality scores of its instances weighted by their prime instance probabilities. Our contributions are summarized as follows.

(1) We propose an end-to-end solution to train CNN-based IQA model under the MIR framework. To the best of our knowledge, it is the first time to train deep blind IQA model by multiple instance learning.
(2) We propose a unified framework to improve the traditional deep IQA models. Any patch-based IQA models can be embeded into the proposed MIR framework to improve their performance furtherly.
(3) We propose a novel conditional EM algorithm to optimize the MIR IQA model. The conditional EM algorithm executes the E-step conditionally. This way simplifies the training procedure of the deep MIR IQA model.

The organization of the rest of this paper is as follows. Section 2 gives a review on related work. Section 3 describes the proposed algorithm in detail. Section 4 shows the experimental results. Section 5 concludes this paper and discusses the future work.

2. Related work

2.1. Image quality assessment

The traditional IQA algorithms built models based on well-designed features [2–12]. Wang et al. [2] utilized the natural scene

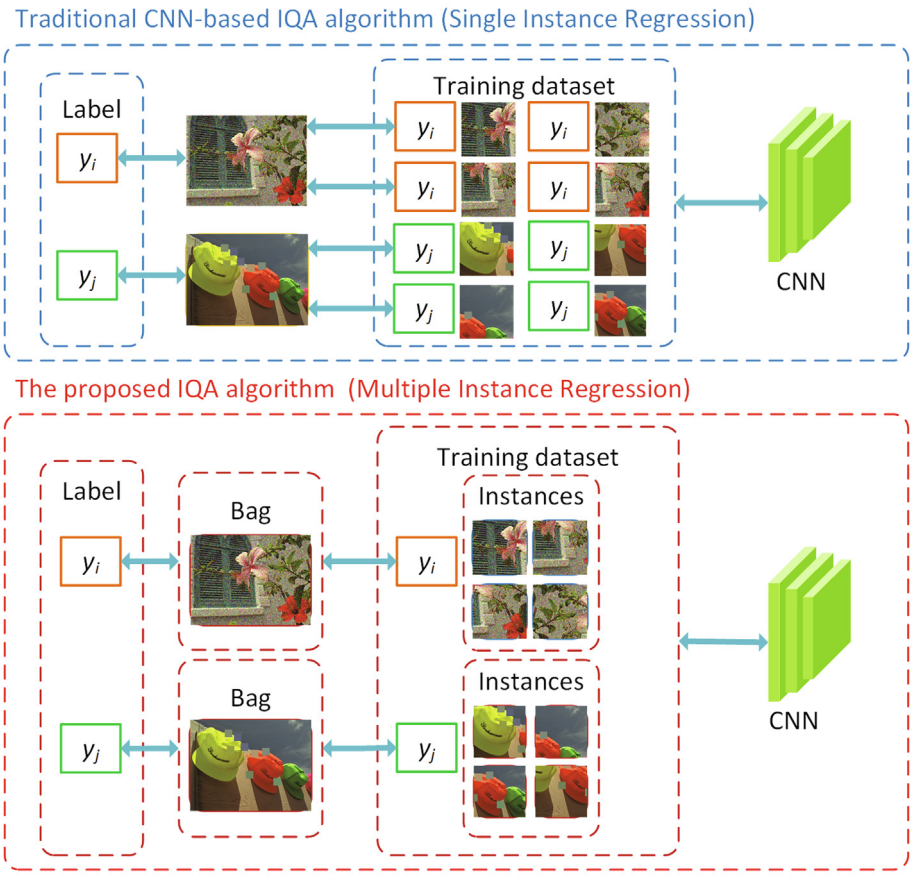


Fig. 1. Traditional CNN-based IQA algorithm and the proposed IQA algorithm.

statistics (NSS) to extract features for the IQA regression. Gao et al. [3] extracted features to mimic the multi-channel structure of human visual system for IQA. Wee et al. [4] constructed the image quality metric with the discrete orthogonal moments by spatial information. Moorthy et al. [5] identified the distortion types with the extracted features and then used distortion specific IQA algorithm to evaluate the visual quality. Lu et al. [6] applied contourlet transform to the NSS of image and combined the extracted features in each sub-band to assess the quality. Saad et al. [7] extracted features from DCT domain to enhance the IQA model. Mittal et al. [8] extracted NSS features from locally normalized luminance coefficients. Ye et al. [9] conducted dictionary learning to extracted features from raw image patches for IQA. Gao et al. [10] collected three types of NSS features together and utilized multiple kernel learning to construct IQA metrics. Zhang et al. [11] learned multivariate Gaussian model from the NSS features to conduct an opinion-unaware IQA model. Yang et al. [12] selected key-frame sequence and extracted motion texture features to evaluate the visuality quality.

Recently, deep learning have been widely used to improve the accuracy of IQA [13–27]. Li et al. [13] trained stacked auto-encoder IQA model on the features extracted from Shearlet transformed images. Hou et al. [14] trained a DBN IQA model with NSS features. Yang et al. [21] evaluated the quality of stereoscopic image with a segmented stacked auto-encoder. Kang et al. [15] firstly introduced CNN architecture to the study of IQA. They also put forward a multi-task CNN model to evaluate visual quality and identify distortion type at same time [16]. Ma et al. [18] designed a multi-task CNN IQA model to improve the performance with assistant from the sub-task of distortion identification. Gao et al. [26] extrated intermediate-level and high-level representations from CNN and used SVR to build IQA model. Kim et al. [17,20] proposed to pre-train the CNN model by referring to the reference images and finetune the CNN model in the train stage. Wu et al. [27] proposed a cascade CNN framework to conduct feature extraction, hierarchicaldegradation concatenation and quality prediction. Most of the above algorithms adopted the average pooling stragegy to obtain the image quality score, which took the assumption that each patch has same contribution to image quality as the prior knowledge.

Some CNN-based algorithms adopted different pooling strategies. He et al. [25] proposed the saliency pooling strategy that the image score is weighted by the significance of the patches. Liu et al. [24] focused on the research on ground truth assignment and pooling strategy of the CNN-based model. They proposed to adjust the local ground-truth by visual saliency and pooled patch scores by the gradient features. All of the above algorithms designed the pooling strategy with specific prior knowledge. Bosse et al. [19] trained a multi-task CNN IQA model to predict the local quality score and relative importance of patch simultaneously. The relative importance are used as weights to fuse the patch scores in test stage. This algorithm designed the pooling strategy with the posterior knowledge and did not consider any prior knowledge. Morover, the multi-task architecture lead the algorithm cannot utilize exist powerful CNN model directly. Its training was also not flexible enough since it requires that patches in same image must be distributed in same mini-batch.

Different from above CNN-based IQA algorithm, the proposed algorithm jointly considers the prior knowledge and the posterior knowledge about the contribution of patches to the image quality. By setting the prior probabilities, the proposed algorithm refers to the user defined prior knowledge. By predicting the quality score for training samples, the proposed algorithm refers to the posterior knowledge. Moreover, the proposed algorithm can embed any patch-based deep IQA model into the MIR framework easily without any change of the network architecture. The proposed algorithm allows to randomly shuffle the patches in different images into different mini-batches, which makes the training more flexible.

2.2. Multiple instance learning

The most commonly studied MIL problems are multiple instance classification (MIC), which learns to offer the classification for the bag with the instances of it [29–34]. Another group of the MIL problems is multiple instance regression (MIR), which handles the case that the bag label is a real value. Unfortunately, the research on the MIR problems are very limited due to the less applications and databases [30]. Ray et al. [35] carried out the original study on the MIR problem. They proposed the prime instance assumption, which assumes the bag label is determined by only one prime instance in the bag. They used the EM algorithm to determine the prime instances and trained linear regression model with the prime instances. However, they did not describe how to determine the prime instance for the unseen bag. Amar et al. [36] used the extended k-nearest neighbor (kNN) and the citation-KNN to predict the real-valued label of the bag. Wang et al. [37] proposed to use mean or median averaging algorithm to predict the unseen bag. Wagstaff et al. [38] proposed the collective assumption, which assumes that a set of instances contribute to bag label and their contributions correspond to their relevance. However, it is a NP-hard problem to determine the relevance of all the instances. Wagstaff et al. [39] also proposed a clustering-based algorithm which selected the instances in the most relevant clusters as exemplar samples and performed regression on them. Wang et al. [40] proposed a probabilistic method based on the mixed model, where the prediction of the bag was computed by the weighted summation of the predictions of the instances. Inspired by this paper, we embed the CNN model into the MIR framework. The major differences between Wang's algorithm and our algorithm are as follows. Firstly, they applied MIR to the traditional regression models. We apply MIR to the CNN models. Secondly, they used a standard EM algorithm to optimize the model, which executed the E-step and the M-step alternatively in each iteration. We propose a conditional EM algorithm that executes the E-step only if the specified condition is met. Thirdly, they executed the EM algorithm with a standard manner, which completely trained a regression model in each M-step. We simplify the process of the EM algorithm, which only trains the CNN model with one epoch in each M-step.

3. The proposed algorithm

3.1. Definition of MIR IQA

The MIR IQA is to learn a regression function $h : 2^X \rightarrow Y$ from a dataset$\{(X_1, y_1), \cdots, (X_m, y_m)\}$, where each image $X_i \subseteq X$ is treated as a bag, the patches of it are treated as a set of instances $\{x_{i_1}, \cdots, x_{i_{n_i}}\}$ $(x_{ij} \in \chi, j = 1, \cdots, n_i)$ of the bag, the visual quality score $y_i \subseteq R$ is the real-valued label of the bag, the function h denotes the CNN-based IQA model. For an test bag (image) X_t, the function h gives the prediction $h(X_t) \subseteq R$ as the quality score of the test image.

3.2. The framework

Fig. 2 shows the framework of the proposed algorithm. In the training stage, we use an EM-like algorithm, called conditional EM algorithm, to optimize the MIR model. The conditional EM algorithm iterates over E-step and M-step. In the M-step, the posterior probabilities of the instances (patches) to be the prime one are used as sample weights to train the CNN model. The model is

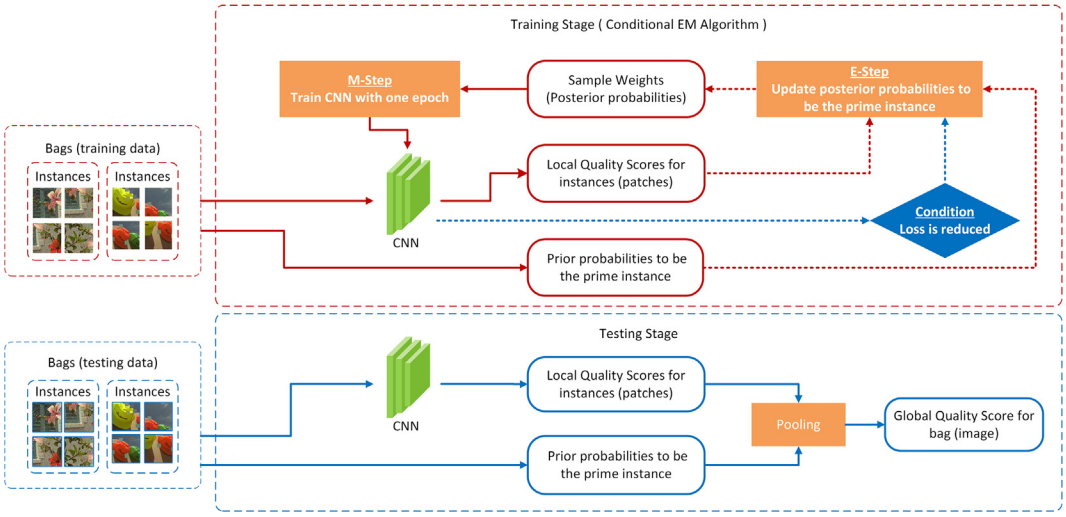


Fig. 2. The framework of the deep MIR IQA algorithm.

trained with one epoch in each M-step. In the E-step, the posterior probabilities of the instances to be the prime one are updated based on the prior probabilities of the instances to be the prime one and the local quality scores of the instances. The CNN model predicts the local quality scores. The dotted lines indicate that the E-step is executed conditionally. The execution condition is that the loss is reduced in current M-step. In the testing stage, the CNN model predicts the local quality scores of the instances (patches) in the test bag (image). Then we use the prior probabilities of the instances as the weights and pool the local quality scores of the instances to the global quality score of the bag.

3.3. Single instance regression

While learning a regression function h, we assume it is disturbed by random noise. Therefore, in the SIR framework, the label y_i is the function of the sample x_i plus a random noise item:

$$y_i = h(x_i, w) + e_i \tag{1}$$

where $h(x_i, w)$ denotes the CNN model, w denotes the network weights to be learned, e_i denotes the random variable of the noise. Normally, we assume e_i obeys Gaussian distribution, i.e. $e_i \sim \mathcal{N}(0, \sigma^2)$. Then the label y_i also obeys Gaussian distribution. Therefore, the conditional probability of the label y_i given the sample x_i can be written as:

$$\begin{aligned} p(y_i|x_i) &= \mathcal{N}(h(x_i, w), \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - h(x_i,w))^2}{2\sigma^2}\right) \end{aligned} \tag{2}$$

Then the log maximum likelihood of model h is:

$$\begin{aligned} h_{ML} &= \arg\max_h \ln P(Y|X) \\ &= \arg\max_h \ln \prod_{i=1}^{m} p(y_i|x_i) \\ &= \arg\max_h \sum_{i=1}^{m} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - h(x_i,w))^2}{2\sigma^2}\right)\right) \end{aligned} \tag{3}$$

It is equivalent to optimize the network weights w with the minimum square error (MSE) loss:

$$h_{ML} = \arg\min_w \sum_{i=1}^{m} (y_i - h(x_i, w))^2 \tag{4}$$

3.4. Multiple instance regression

In the MIR framework, we follow the prime instance assumption, which assumes there is a single instance being responsible for the bag label. Similar to the SIR framework, the MIR regression model is also disturbed by random noise. Therefore, the label y_i of the bag X_i is the function of the prime instance x_{ij} plus a random noise item:

$$y_i = h(x_{ij}, w) + e_{ij} \tag{5}$$

where j denotes the index of the prime instance of the bag X_i, $h(x_{ij}, w)$ denotes the CNN model, w denotes the network weights to be learned, e_{ij} denotes the random variable of the noise. Both e_{ij} and y_i obey Gaussian distribution. Therefore, the conditional probability of the bag label y_i given the prime instance x_{ij} can be written as:

$$\begin{aligned} p(y_i|x_{ij}) &= \mathcal{N}(h(x_{ij}, w), \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - h(x_{ij},w))^2}{2\sigma^2}\right) \end{aligned} \tag{6}$$

We introduce a random latent variable z_i=[z_{i_1},$\cdots$,$z_{i_{n_i}}$] for bag X_i, where $z_{ij} = 1$ if the j-th instance is the prime instance and $z_{ij} = 0$ otherwise. According to the prime instance assumption, there is only one element in z_i equals to one. We can conclude that $\sum_{j=1}^{n_i} z_{ij} = 1$. The data $X \bigcup Y \bigcup Z$ is called the completed data, the data $X \bigcup Y$ is called the observed data, where X=(X_1,$\cdots$,X_m), Y=(y_1,$\cdots$,y_m), Z=(z_1,$\cdots$,z_m).

Since there is only one element in z_i=[z_{i_1},$\cdots$,$z_{i_{n_i}}$] equals to one and all the other elements in z_i equal to zero, the conditional probability of the bag label given the bag, i.e. $p(y_i|X_i)$, can be written as:

$$\begin{aligned} p(y_i|X_i) &= p(y_i, z_{i_1}, z_{i_2}, \ldots, z_{i_{n_i}}|X_i) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} z_{ij} (y_i - h(x_{ij}, w))^2\right) \end{aligned} \tag{7}$$

Then the log likelihood of the completed data $X \bigcup Y \bigcup Z$ will be:

$$\begin{aligned} \ln P(Y, Z|X) &= \ln \prod_{i=1}^{m} p(y_i, z_{i_1}, z_{i_2}, \ldots, z_{i_{n_i}}|X_i) \\ &= \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_{ij} (y_i - h(x_{ij}, w))^2 \end{aligned} \tag{8}$$

According to Eq. 8, we will train a CNN model with the latent variable z_{ij}. Since the latent variables cannot be observed directly, we propose an EM-like algorithm to optimize the network weights w of the CNN model and determines the expectation of the latent variables z_{ij} alternatively. As the EM algorithm usually does, we define a *Q-function* as following:

$$\begin{aligned} Q(h'|h) &= E_{Z|Y,X}[\ln P(Y,Z|X)] \\ &= E\left[\sum_{i=1}^{m} \ln \tfrac{1}{\sqrt{2\pi\sigma^2}} - \tfrac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} z_{ij}(y_i - h(x_{ij},w))^2\right] \\ &= \sum_{i=1}^{m} \ln \tfrac{1}{\sqrt{2\pi\sigma^2}} - \tfrac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} E[z_{ij}](y_i - h(x_{ij},w))^2 \end{aligned} \tag{9}$$

where h indicates the model with old network weights, h' denotes the model with updated network weights.

Therefore, the maximum likelihood of the model h can be written as:

$$\begin{aligned} h_{ML} &= \arg\max_h Q(h'|h) \\ &= \arg\max_h \left(\sum_{i=1}^{m} \ln \tfrac{1}{\sqrt{2\pi\sigma^2}} - \tfrac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} E[z_{ij}](y_i - h(x_{ij},w))^2\right) \end{aligned} \tag{10}$$

With fixed value of the variance σ, it is equivalent to optimize the network weights w with a weighted MSE loss:

$$h_{ML} = \arg\min_w \sum_{i=1}^{m}\sum_{j=1}^{n_i} E[z_{ij}](y_i - h(x_{ij},w))^2 \tag{11}$$

According to Eq. 11, after introducing the latent variable z_{ij}, the patch-based CNN model h can be optimized by reducing the sample weighted MSE loss. The expectation $E[z_{ij}]$ can be treated as the sample weights.

Now we will deduce the expectation $E[z_{ij}]$. It is known that the conditional probability of the bag label given the bag, i.e. $p(y_i|X_i)$, is the marginal of the joint distribution $p(y_i, z_i|X_i)$ after we introduce the latent variable z_i.

$$\begin{aligned} p(y_i|X_i) &= \sum_{z_i} p(y_i, z_i|X_i) \\ &= \sum_{z_i} p(z_i|X_i)p(y_i|z_i, X_i) \\ &= \sum_{j=1}^{n_i} p(z_i = 1|X_i)p(y_i|z_{ij} = 1, X_i) \end{aligned} \tag{12}$$

where $p(z_{ij} = 1|X_i)$ is the probability that $z_{ij} = 1$ when we observe the bag X_i, which can be considered as the prior probability of the *j-th* instance to be the prime instance of the *i-th* bag.

Based on the prime instance assumption, the prime instance x_{ij} is responsible for the bag label. Then we have

$$p(y_i|z_{ij} = 1, X_i) = p(y_i|x_{ij}) \tag{13}$$

The expectation $E[z_{ij}]$ can be computed as following:

$$\begin{aligned} E[z_{ij}] &= E_{Z|Y,X}[z_{ij}] \\ &= 0 \cdot p(z_{ij} = 0|y_i, X_i) + 1 \cdot p(z_{ij} = 1|y_i, X_i) \\ &= p(z_{ij} = 1|y_i, X_i) \end{aligned} \tag{14}$$

where $p(z_{ij} = 1|y_i, X_i)$ is the probability that $z_{ij} = 1$ when we observe the bag label y_i and the bag X_i, which can be considered as the posterior probability of the *j-th* instance to be the prime instance of the *i-th* bag.

According to Eq. 12, Eq. 13 and the bayesian formulation, the expectation $E[z_{ij}]$ can be computed as following:

$$\begin{aligned} E[z_{ij}] &= p(z_{ij} = 1|y_i, X_i) \\ &= \tfrac{p(z_{ij}=1, y_i|X_i)}{p(y_i|X_i)} \\ &= \tfrac{p(z_{ij}=1|X_i)p(y_i|z_{ij}=1, X_i)}{p(y_i|X_i)} \\ &= \tfrac{p(z_{ij}=1|X_i)p(y_i|z_{ij}=1, X_i)}{\sum_{j=1}^{n_i} p(z_{ij}=1|X_i)p(y_i|z_{ij}=1, X_i)} \\ &= \tfrac{p(z_{ij}=1|X_i)p(y_i|x_{ij})}{\sum_{j=1}^{n_i} p(z_{ij}=1|X_i)p(y_i|x_{ij})} \end{aligned} \tag{15}$$

Eq. 15 demonstrates that the expectation $E[z_{ij}]$, i.e. the posterior probability to be the prime instance, is jointly determined by the prior probability to be the prime instance and the conditional probability of the bag label given the prime instance. The prior probability $p(z_{ij} = 1|X_i)$ is determined by the prior knowledge of the IQA problem. According to Eq. 6, the conditional probability $p(y_i|x_{ij})$ can be computed with the bag label y_i, the local quality score $h(x_{ij}, w)$ and the variance σ. The bag label y_i is the ground-truth of the training data. The local quality score $h(x_{ij}, w)$ is predicted by the CNN model h. The variance σ is a super parameter of the MIR IQA model. By using $E[z_{ij}]$ as the sample weights, the CNN model are optimized by jointly referring to the prior knowledge and posterior knowledge at same time.

3.5. Conditional EM algorithm

Based on Eq. 11, we propose an EM-like algorithm to optimize the MIR IQA model. We call this algorithm as conditional EM algorithm because the E-step will be executed only if a condition has been met in current M-step. The conditional EM algorithm is shown in Algorithm 1.

In the M-step, the expectation $E[z_{ij}]$ are used as the sample weights to train the CNN model. The algorithm only trains the CNN model with one epoch in each M-step.

In the E-step, the algorithm performs three sub-steps. Firstly, it predicts the local quality score $h(x_{ij}, w)$ with the updated CNN model h. Secondly, it computes the conditional probability of the bag label $p(y_i|x_{ij})$ with the label y_i, the local quality score $h(x_{ij}, w)$ and the variance σ by Eq. 6. Thirdly, it updates the expectation $E[z_{ij}]$ with the prior probability $p(z_{ij} = 1|X_i)$ and conditional probability $p(y_i|x_{ij})$ by Eq. 15.

Algorithm 1: The Conditional EM algorithm

Input: $X \bigcup Y \bigcup Z, p(z_{ij} = 1|X_i), \sigma, K$.
Output: the trained model h.
1: $e \leftarrow 0$.
2: **while** h is not convergent **and** $e < K$ **do**
3: **(M-step)** Trains h with one epoch, where $E[z_{ij}]$ are the sample weights.
4: **if** the loss is reduced in current M-step (epoch) **then**
5: **(Conditional E-step)** Exexutes following sub-steps:
1) Predicts $h(x_{ij}, w)$.
2) Computes $p(y_i|x_{ij})$ with $y_i, h(x_{ij}, w)$, and σ.
3) Updates $E[z_{ij}]$ with $p(z_{ij} = 1|X_i)$ and $p(y_i|x_{ij})$.
6: **endif**
7: $e \leftarrow e + 1$.
8: **end while**
9: **return** h.

There are two different points between the traditional EM algorithm and the conditional EM algorithm. Firstly, the traditional EM algorithm executes the E-step and the M-step alternatively in each iteration. The conditional EM algorithm executes the E-step only if

the specified condition is met. By setting the loss reduction as the execution condition, the sample weights will be updated only when the CNN model is improved in current epoch. Secondly, the traditional EM algorithm completes the whole training process in each M-step. The conditional EM algorithm trains the CNN model with one epoch in each M-step. These measures simplify the training procedure of the MIR IQA model.

The proposed algorithm requires to predict the local quality scores for the training data in the train stage. The time of the prediction is much less than that of one training epoch. Moreover, the prediction will only be conducted when the loss has been reduced. Therefore, the deep MIR IQA algorithm improves the performance of the deep SIR IQA algorithm with an acceptable increase in the training time.

3.6. prior probability

The prior probability $p(z_{ij} = 1|X_i)$ responses the prior knowledge about the constributions of patches to the image quality. It can be determined by the expert knowledge. We define three types of prior probability.

The first kind of prior probability assumes each patch contributes to the image quality equally, which can be written as:

$$p(z_{ij} = 1|X_i) = 1/|X_i| \tag{16}$$

where $|X_i|$ is the number of instances (patches) in bag X_i.

The second kind of prior probability determines the contributions of patches according to the saliency information of the patches, which can be written as:

$$p(z_{ij} = 1|X_i) = S_{ij}/\sum_{j=1}^{n_i} S_{ij} \tag{17}$$

where S_{ij} is the summation of the pixel values of the croresponding patch in the salency map. The saliency map is generated by the GBVS algorithm [41].

The third kind of prior probability determines the contributions of patches based on the quality predictions of the training samples, which can be written as:

$$p(z_{ij} = 1|X_i) = \frac{\exp(h(x_{ij}, w))}{\sum_{j=1}^{n_i} \exp(h(x_{ij}, w))} \tag{18}$$

where $h(x_{ij}, w)$ is the predicted quality score for the instance (patch) x_{ij}. According to [40], the optimization of CNN makes the prior probability approximate to the posterior probability.

In this paper, the first kind of prior probability will be used to build the MIR model as base model if there is no special statement.

3.7. Global quality score

According to Eq. 12 and Eq. 13, the conditional probability of the bag label given the bag is the weighted summation of the conditional probabilities of the bag label given different prime instances, where the weights are the prior probabilities of different instances to be the prime one. Therefore, the global quality score y_t of the test bag (image) X_t can be predicted by:

$$y_t = \sum_{j=1}^{n_t} p(z_{tj} = 1|X_t) \cdot h(x_{tj}, w) \tag{19}$$

where $p(z_{tj} = 1|X_t)$ is the prior probability of the *j-th* instance to be the prime instance of the *t-th* bag, which can be determined by Eq. 16, Eq. 17 or Eq. 18. $h(x_{tj}, w)$ is the predicted local quality score for the instance (patch) x_{tj}.

3.8. Training details

According to Eq. 6, Eq. 11 and Eq. 15, the MIR IQA model involves a super parameter σ, i.e. the variance of the Gaussian distribution for the noise. The value of σ is set according to the experimental experience. As for the number of instances in a bag, we set same value both for training bag and for testing bag.

Since the model h cannot give meaningful predictions with randomly initialized network weights, we initialize the sample weights by the prior probabilities in the first M-step (epoch). From the second M-step, we begin to use the posterior probabilities as the sample weights.

To train the CNN model, the MSE is used as the loss function. The optimizer ADAM [42] is employed in the training process. The learning rate is set as 1×10^{-4}. The maximum number of epochs is set as 30. The patience of the early stopping is set as five.

4. Experimental results

4.1. Experimental protocols

4.1.1. Databases

We evaluate the proposed algorithm on five IQA databases: LIVE [43], CSIQ [44], TID2013 [28], LIVE MD [45], and LIVE in the wild challenge (LIVE Challenge) [46]. The LIVE database contains 779 singly distorted images and 5 distortion types. The CSIQ database contains 866 singly distorted images and 6 distortion types. The TID2013 database contains 3000 singly distorted images and 24 distortion types. The LIVE MD database contains 480 multiply distorted images and five distortion types: Gaussian Blur (GB), JPEG compression (JPEG), Gaussian White Noise (WN), GB + JPEG and GB + WN. The LIVE Challenge database contains 1169 authentic distorted images.

For the LIVE, CSIQ, TID2013 and LIVE MD databases, we randomly partition the reference images into two subsets. Then we partition the distorted images into training dataset and testing dataset according to the partition of the reference images. This makes the contents of the training images and testing images have no overlap. For the LIVE Challenge database, we directly partition the distorted images into training dataset and testing dataset since the content of each image is unique. The proportion of the training dataset and testing dataset are 80% and 20%. To remove the bias, we repeat the training and testing twenty times and report the averaged results.

4.1.2. Backbones

To evaluate the effects of the propose algorithm for different CNN models, we select two public models and one self-created model as the backbones.

The two public models are Resnet50 [47] and VGG16 [48], which are pre-trained on the ImageNet database [23]. We fine-tune the two models by adding one output layer on top of the existing layers. A dropout layer is added before the output layer and the drop rate is set as 0.5 [49]. The input patch size is 224×224 for the two models. For the LIVE, CSIQ, LIVE MD and LIVE Challenge databases, we randomly crop 50 patches from each distorted image. For the TID2013 database, we randomly crop 32 patches from each distorted image since the number of distorted images are much more than other databases.

Table 1 shows the architecture of the self-created model. ReLU is applied as activation functions for each convolution layer and fully connected layer. We call the self-created model as ConvL10 since there are ten convolutional layers in the architecture. The input patch size is 128×128. For the LIVE, CSIQ, LIVEMD and LIVE Challenge databases, we randomly crop 100 patches from each

Table 1
Network architecture of the self-created CNN model.

Layer	Output Size	Parameter		
		Kernel	Stride	Padding
Input	$128 \times 128 \times 3$			
Conv2D	$128 \times 128 \times 32$	3×3	1	0
Conv2D	$128 \times 128 \times 32$	3×3	1	0
MaxPooling2D	$64 \times 64 \times 32$	2×2	2	
Conv2D	$64 \times 64 \times 64$	3×3	1	0
Conv2D	$64 \times 64 \times 64$	3×3	1	0
MaxPooling2D	$32 \times 32 \times 64$	2×2	2	
Conv2D	$32 \times 32 \times 128$	3×3	1	0
Conv2D	$32 \times 32 \times 128$	3×3	1	0
MaxPooling2D	$16 \times 16 \times 128$	2×2	2	
Conv2D	$16 \times 16 \times 256$	3×3	1	0
Conv2D	$16 \times 16 \times 256$	3×3	1	0
MaxPooling2D	$8 \times 8 \times 256$	2×2	2	
Conv2D	$8 \times 8 \times 512$	3×3	1	0
Conv2D	$8 \times 8 \times 512$	3×3	1	0
GlobalAveragePooling2D	512			
Dense[1]	512			
Dense	512			
Output	1			

[1] Dense denotes the fully connected layer.

distorted image. For the TID2013 database, we randomly crop 64 patches from each distorted image.

4.1.3. Evaluation metrics

To evaluate the effect of the proposed algorithm, we use two widely used metrics: Spearman's rank-order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC).

The metric SRCC is computed by:

$$SRCC = 1 - \frac{6 \cdot \sum_i (ry_i - r\hat{y}_i)^2}{n(n-1)} \tag{20}$$

where ry_i is the *i-th* image's rank in the ground-truth scores, $r\hat{y}_i$ is the *i-th* image's rank in the predicted scores, n is the number of images.

The metric PLCC is computed by:

$$PLCC = \frac{\sum_i (y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_i (y_i - \mu_y)^2 \sum_i (\hat{y}_i - \mu_{\hat{y}})^2}} \tag{21}$$

where y_i is the ground-truth score of the *i-th* image, $\hat{y}_i$ is the predicted score of the *i-th* image, μ_y is the averaged value of the ground-truth scores, $\mu_{\hat{y}}$ is the averaged value the predicted scores.

4.2. Effects of the proposed algorithm

4.2.1. SIR vs. MIR

We compare the performance between the traditional deep SIR IQA model and the proposed deep MIR IQA model. We use the Resnet50 model as the backbone. We train three MIR models with different variances σ: 0.1, 0.5 and 1.0. The experiment is conducted on the TID2013 database, the LIVE MD database and the LIVE Challenge database.

The comparison of the overall performance are shown in Table 2. It can be observed that nearly all the MIR models obtain better results than their corresponding SIR models. The best result on every database is achieved by a MIR model. Even the SIR model and the MIR model adopt the same prior knowledge that each patch has same contribution to the image quality, The MIR model optimize the CNN model by referring to the posterior probabilities as sample weights. The additional posterior knowledge improves the performance of the CNN-based IQA model.

It also can be observed that the MIR models with smaller variance can obtain better performance in the three databases. The model MIR(σ=0.1) and MIR(σ=0.5) improves the SIR model on all the databases. Only the model MIR(σ=1.0) performs a little worse than the SIR model. We guess this is because that its convergence velocity is slow and thus cannot obtain optimal model within 30 epochs. In general, one can obtain the best MIR model by tuning the super parameter.

We compare the performance on individual artificial distortion types on the LIVE MD database and the TID2013 database. Table 3 shows the comparison on the LIVE MD database. The MIR models achieve best SRCC on nearly all distortion types except distortion WN. Table 4 shows the comparison on the TID2013 database. The MIR models achieve best SRCC on all distortion types. The model MIR(σ=0.5) performs best, which gets the best SRCC on 15 distortion types. The model MIR(σ=1.0) and model MIR(σ=0.1) gets the best SRCC on seven and two distortion types respectively. This is consistent with their overall performance.

We also compare the gains in SRCC and PLCC for the three MIR models by comparing them with the SIR model. Fig. 3 shows the gains on individual distortion types. It can be observed that the each MIR model obtains positive gains on most of the distortion types. On one hand, it intuitively demonstrates the effect of the proposed algorithm. The model MIR(σ=0.5) has the largest number of positive gains, this leads to best performance on the databases. On the other hand, the MIR model still obtain negative gains on some distortion types. It is a normal case since the MIR model try to improve the overall performance by adapting itself to most of the distortion types rather than specific one.

4.2.2. Consistency of IQA

To analyze the consistency of IQA for different distortion types between the SIR model and the MIR models, we compare their tendency of the metrics on individual distortion types. The results are shown in Fig. 4.

Fig. 4(a) and Fig. 4(b) show the SRCC and PLCC for the LIVE MD database respectively. All the three MIR models and the SIR model have similar tendency of the metrics on individual distortion types. All of them obtain relative worse performance on distortion WN and relative better performance on other four distortions. In other words, the SIR model and the MIR models have similar sensitivity on individual distortion types. Fig. 4(c) and Fig. 4(d) show the SRCC

Table 2
Comparison of SRCC and PLCC between the SIR IQA model and the MIR IQA model. The bolded digits denotes the best metrics.

Model	TID2013		LIVE MD		LIVE Challenge	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
SIR	0.782	0.816	0.925	0.936	0.786	0.822
MIR($\sigma = 0.1$)	0.784	0.820	0.928	0.937	**0.793**	**0.831**
MIR($\sigma = 0.5$)	**0.796**	**0.821**	**0.931**	**0.940**	0.785	0.826
MIR($\sigma = 1.0$)	0.787	0.817	0.926	0.937	0.782	0.821

Table 3
Comparison of SRCC on individual distortion types for the LIVE MD database. The bolded digits denotes the best metrics.

Model	GB	JPEG	WN	GB + JPEG	GB + WN
SIR	0.871	0.820	**0.712**	0.856	0.856
MIR($\sigma = 0.1$)	**0.876**	0.845	0.694	0.864	**0.874**
MIR($\sigma = 0.5$)	0.873	**0.850**	0.711	**0.871**	0.866
MIR($\sigma = 1.0$)	0.874	0.841	0.669	0.865	0.866

Table 4
Comparison of SRCC on individual distortion types for the TID2013 database. The bolded digits denotes the best metrics.

Model	AGN	ANC	SCN	MN	HFN	IN	QN	GB	DEN	JPEG	JP2K	JGTE
SIR	0.685	0.515	0.822	0.338	0.779	0.782	0.817	0.873	0.819	0.842	0.879	0.742
MIR($\sigma = 0.1$)	0.696	**0.537**	0.820	0.315	0.781	0.788	**0.832**	0.866	0.816	0.832	0.871	0.740
MIR($\sigma = 0.5$)	**0.703**	0.534	**0.824**	0.311	0.786	**0.834**	0.818	**0.883**	**0.851**	**0.854**	**0.884**	0.752
MIR($\sigma = 1.0$)	0.697	0.530	0.810	**0.369**	**0.787**	0.810	0.816	0.862	0.810	0.847	0.881	**0.755**
Model	J2TE	NEPN	Block	MS	CTC	CCS	MGN	CN	LCNI	ICQD	CHA	SSR
SIR	0.833	0.421	0.537	0.239	0.109	0.607	0.657	0.710	0.838	0.727	0.692	0.891
MIR($\sigma = 0.1$)	0.824	0.428	0.542	0.207	0.071	0.653	0.667	0.716	0.836	0.703	0.704	0.881
MIR($\sigma = 0.5$)	**0.846**	**0.457**	0.551	**0.273**	0.071	**0.690**	0.669	0.733	**0.855**	**0.750**	**0.727**	**0.900**
MIR($\sigma = 1.0$)	0.839	0.447	**0.587**	0.270	**0.146**	0.637	**0.680**	**0.736**	0.846	0.746	0.685	0.879

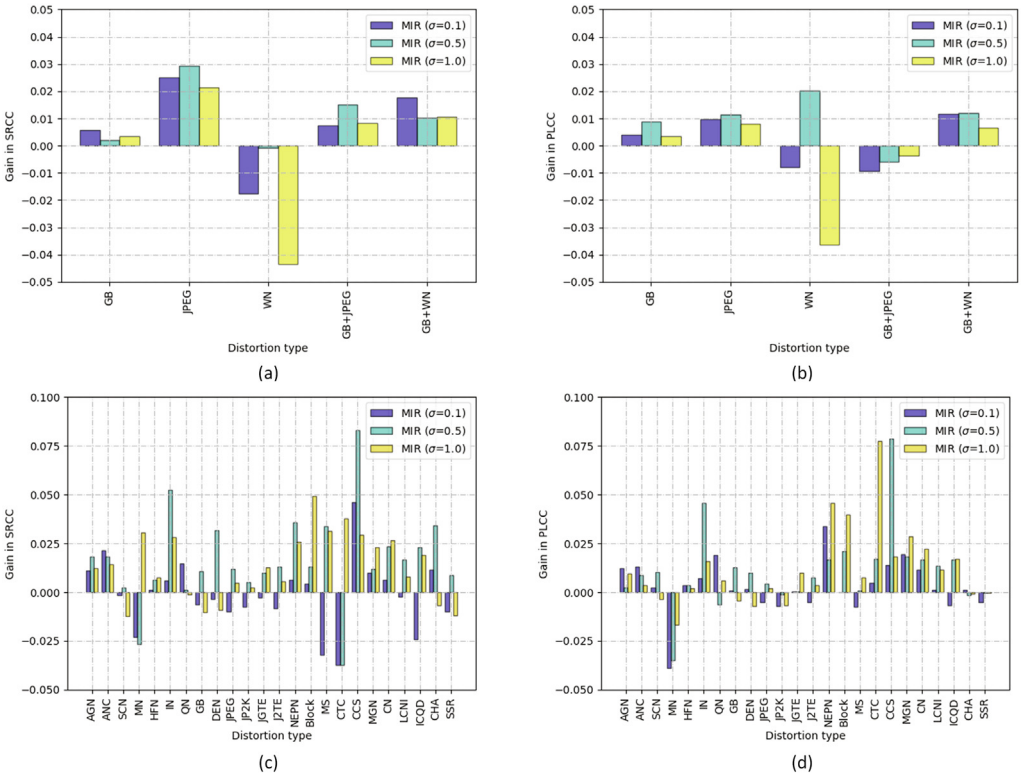


Fig. 3. The gains in SRCC and PLCC on individual distortion types. (a) The gains in SRCC for the LIVE MD database. (b) The gains in PLCC for the LIVE MD database. (c) The gains in SRCC for the TID2013 database. (d) The gains in PLCC for the TID2013 database.

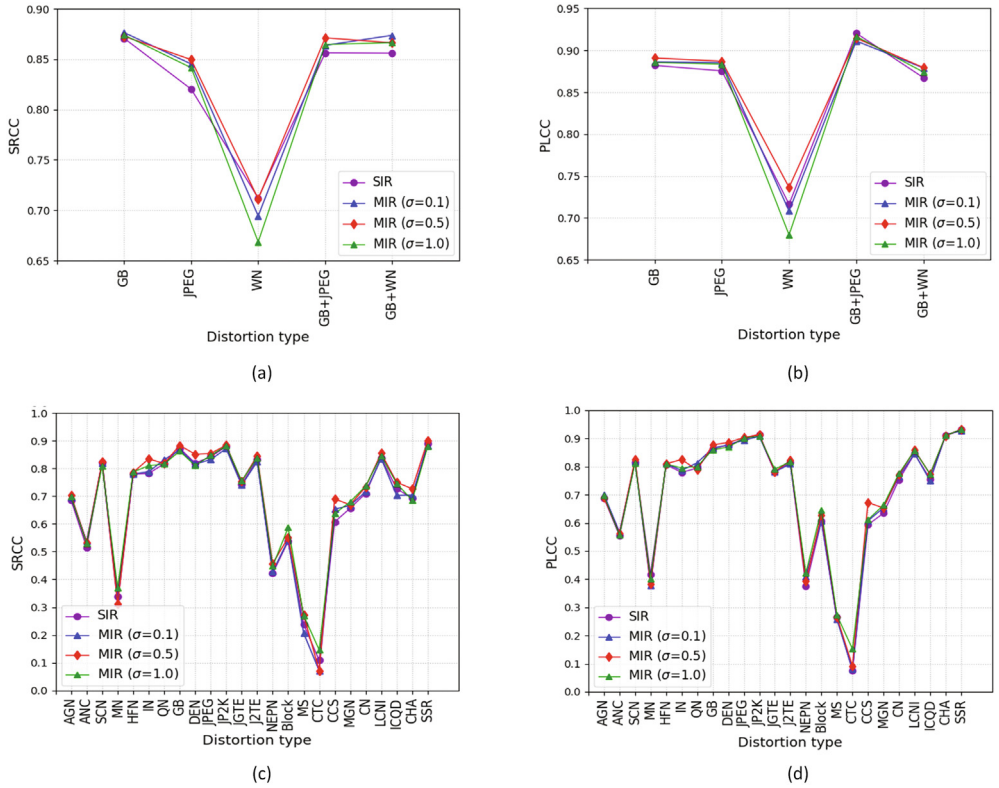


Fig. 4. The tendency of SRCC and PLCC on individual distortion types. (a) The SRCC for the LIVE MD database. (b) The PLCC for the LIVE MD database. (c) The SRCC for the TID2013 database. (d) The PLCC for the TID2013 database.

and PLCC for the TID2013 database respectively. We can also find that all the three MIR models and the SIR model have similar sensitivity on individual distortion types.

From above experiments, it can be concluded that the MIR model can keep the consistency of IQA result with the SIR model. This is a good property since the MIR model can improve the overall performance of the SIR model without changing the sensitivity on individual distortion types.

4.2.3. Generality

To test the generality of the proposed algorithm, we build the SIR model and the MIR model for three backbones and compare their performance. We conduct the experiment on the TID2013, LIVEMD and LIVE Challenge databases. We set the variance σ as 0.1 for Resnet50 and VGG16, and set the variance σ as 0.5 for ConvL10.

Table 5 shows the experiment results, the MIR models perform better than the SIR model for all the three backbones. For Resnet50, even though the model MIR(σ=0.1) is not the best model for the TID2013 database and the LIVE MD database (refer to Table 2), it still improves the performance of the SIR model. This experiment demonstrates the good generality of the proposed algorithm. The deep MIR IQA model stands on the shoulders of the deep SIR IQA model. In the future, one can concentrate on designing more powerful CNN-based IQA algorithms and utilize the proposed MIR framework to improve their performance.

4.2.4. Effects of different prior probabilities

According to the prior probabilities defined in Eq. 16, Eq. 17 or Eq. 18, we train three MIR models and call them MIR-avg, MIR-sal and MIR-pred respectively. We conduct the experiment on the LIVE, TID2013 and LIVE Challenge databases.

Table 6 shows the experiment results. For the LIVE database, the model MIR-avg achieves best results both on SRCC and PLCC. For the TID2013 database, all the MIR models performs better than the SIR model. The model MIR-pred achieves the best results among them. The model MIR-pred can update the prior probabilities according to the feedback of the quality predictions. Therefore, the larger data in TID2013 helps it to find the optimal values of the prior probabilities. For the LIVE Challenge database, the model MIR-sal achieves the best results. It looks the prior knowledge

Table 5
Comparison of SRCC and PLCC for different backbones. The bolded digits denotes the better metrics.

Backbone	Framework	TID2013		LIVE MD		LIVE Challenge	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Resnet50	SIR	0.782	0.816	0.925	0.936	0.786	0.822
	MIR(σ=0.1)	**0.784**	**0.820**	**0.928**	**0.937**	**0.793**	**0.831**
VGG16	SIR	0.792	0.826	0.916	0.925	0.758	0.798
	MIR(σ=0.1)	**0.795**	**0.829**	**0.919**	**0.926**	**0.762**	**0.803**
ConvL10	SIR	0.724	0.770	0.893	0.903	0.674	0.705
	MIR(σ=0.5)	**0.736**	**0.780**	**0.896**	**0.909**	**0.675**	**0.706**

Table 6
Comparison of SRCC and PLCC among the MIR models with various prior porbabilities. The bolded digits denotes the best metrics.

Model	LIVE		TID2013		LIVE Challenge	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
SIR	**0.967**	0.969	0.782	0.816	0.786	0.822
MIR-avg[1]	**0.967**	**0.971**	0.796	0.821	0.785	0.826
MIR-sal[1]	0.962	0.964	0.783	0.817	**0.793**	**0.828**
MIR-pred[1]	0.965	0.966	**0.799**	**0.829**	0.783	0.821

[1] The model uses resnet50 as the backbone and sets the super parameter σ as 0.5.

based on saliency map plays bigger role on evaluating the authentic distorted images.

4.2.5. Performance comparison

We compare the performance of the proposed algorithm with other classic blind IQA algorithms. The compared algorithms includes BLIINDS II [7], BRISQUE [8], CORNIA [9], ILNIQE [11], CNN-IQA [14], deep-IQA [18], BIECON [16], DIQA [19] and CaHDC [27]. We build the SIR model and the MIR model by fine-tuning the Resnet50 model.

Table 7 shows the comparison result. First of all, the MIR IQA model performs better than the SIR IQA model for every database. Besides, the MIR model embedded with Resnet50 ranks in the top three models on most of the databases. The last column in Table 7 shows the weighted average of the metrics for the five databases, where the weight is proportional to the number of distorted images in the database. For SRCC, the MIR Resnet50 model gets the third best result. For PLCC, the MIR Resnet50 model gets the second best result.

The model BIECON and DIQA performed worse for the LIVE Challenge database. The two algorithms required the reference images in the train stage and this limits their applications. There is no reference images in the LIVE Challenge database and thus lead to their worse performance. The model CaHDC also performed worse for the LIVE Challenge database. On the contrary, the MIR Resnet50 model has better generalization capability both on the first four databases with artificial distortions and on the LIVE Challenge database with authentic distortions.

Generally speaking, the proposed MIR framework has no special requirements in application. One can select suitable CNN-based model for specific application and use the MIR framework to improve its performance furtherly. Moreover, the deep IQA model can be embeded into the MIR framework easily without changing the network architecture. All of these advantages make the MIR framework have wide applications.

4.2.6. Cross data set test

We conduct the cross data set test with the LIVE, CSIQ and TID2013 databases. Four common distortion types, including GB, WN, JPEG and JPEG 2000 compression (JPEG2000), are used in the experiment.

Table 8 shows the experiment results. It can be observed that the MIR IQA model has better generalization capability than the SIR IQA model on the cross data set test. Moreover, compared with other IQA models, the MIR IQA model offers competitive performance. In the cross data set test between the LIVE and CSIQ database, the MIR IQA model obtains the best result. This experiment

Table 7
Comparison of SRCC and PLCC for different blind IQA algorithms. The CNN-based algorithms are shown in italics. The bolded digits denotes the top three metrics.

Algorithm	LIVE		CSIQ		TID2013		LIVE MD		LIVE Challenge		Weighted Averag	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BLIINDS II	0.912	0.916	0.780	0.832	0.536	0.628	0.887	0.902	0.463	0.507	0.664	0.729
BRISQUE	0.939	0.942	0.775	0.817	0.572	0.651	0.897	0.921	0.607	0.645	0.689	9.746
CORNIA	0.942	0.943	0.714	0.781	0.549	0.613	0.900	0.915	0.618	0.662	0.666	0.717
ILNIQE	0.902	0.908	0.821	0.865	0.521	0.648	0.902	0.914	0.594	0.589	0.662	0.747
CNN-IQA	0.956	0.953	–	–	–	–	–	–	–	–	–	–
deep-IQA	0.960	**0.972**	–	–	**0.835**	**0.855**	–	–	–	–	–	–
BIECON	0.958	0.962	**0.825**	0.838	0.721	0.765	0.912	0.928	0.595	0.613	0.755	0.783
DIQA	**0.975**	**0.977**	**0.884**	**0.915**	**0.825**	**0.850**	**0.939**	**0.942**	0.703	0.704	**0.838**	**0.855**
CaHDC	0.965	0.964	**0.903**	**0.914**	**0.862**	**0.878**	**0.927**	**0.950**	**0.738**	**0.744**	**0.863**	**0.874**
DSIR-IQA[1]	**0.967**	0.969	0.820	0.878	0.782	0.816	0.925	0.936	**0.786**	**0.822**	0.822	0.854
DMIR-IQA[1]	**0.967**	**0.971**	0.823	**0.881**	0.796	0.821	**0.931**	**0.940**	**0.785**	**0.826**	**0.830**	**0.858**

[1] The model uses resnet50 as the backbone and sets the super parameter σ as 0.5.

Table 8
Comparison of SRCC in cross data set test. The bolded digits denotes the top three metrics.

Train	Test	BLINDS II	BRISQUE	CORNIA	ILNIQE	DIQA	CaHDC	DSIR-IQA[1]	DMIR-IQA[1]
LIVE	CSIQ	0.901	0.890	0.898	0.880	**0.915**	-	**0.903**	**0.922**
	TID2013	0.855	0.878	**0.879**	0.877	**0.922**	-	0.878	**0.891**
CSIQ	LIVE	0.894	0.919	0.920	0.913	**0.926**	-	**0.939**	**0.947**
	TID2013	0.765	0.874	0.852	**0.945**	**0.923**	-	0.906	**0.907**
TID2013	LIVE	0.894	0.877	**0.907**	**0.906**	0.904	**0.930**	0.879	0.880
	CSIQ	**0.864**	0.861	0.859	0.861	**0.877**	0.736	0.856	**0.867**

[1] The model uses resnet50 as the backbone and sets the super parameter σ as 0.5.

demonstrates that the proposed algorithm has good generality on different databases.

5. Conclusion

In this paper, we propose a novel deep blind IQA algorithm based on multiple instance regression to overcome the critical defect that the local ground-truth is not available in the traditional CNN-based IQA models. The proposed MIR framework can jointly consider the prior knowledge and the posterior knowledge about the contributions of patches to image quality. As a result, it improves the performance of the traditional CNN-based IQA model. To increase the training efficiency, we propose the conditional EM algorithm, which only executes the E-step when the model is improved in the M-step. Moreover, the proposed algorithm can be used as a unified framework to improve any CNN-based IQA models. The experimental results demonstrate the effect and the generality of the proposed algorithm.

Some further study can be conducted base on this work. In this paper, we test three types of prior probabilities. In the future, we can try to design new prior probabilities that better simulate the visual perception of human to the image quality. This may be a potential way to improve the proposed MIR framework.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

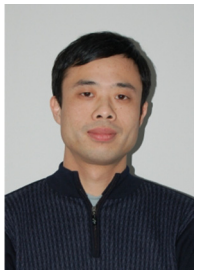
Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62036007, 61772402, 61871311 and 61876146, in part by the National Key Research and Development Program of China under Grants 2018AAA0102702 and 2018AAA0103202, in part by the Key Industrial Innovation Chain Project in Industrial Domain of Shaanxi Province under Grant 2020ZDLGY05-01.

References

[1] Z. Wang, A.C. Bovik, Modern Image Quality Assessment, Morgan & Claypool, New York, 2006.
[2] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Processing 13 (4) (2016) 600–612.
[3] X.-B. Gao, W. Lu, D.-C. Tao, X.-L. Li, Image Quality Assessment Based on Multiscale Geometric Analysis, IEEE Trans. Image Processing 18 (7) (2009) 1409–1423.
[4] C.-Y. Wee, R. Paramesran, R. Mukundan, X.-D. Jiang, Image quality assessment by discrete orthogonal moment, Pattern Recogn. 43 (2010) 4055–4068.
[5] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Trans. Image Processing 20 (12) (2011) 3350–3364.
[6] W. Lu, K. Zeng, D.-C. Tao, Y. Yuan, X.-B. Gao, No-reference image quality assessment in contourlet domain, Neurocomputing 73 (4) (2010) 784–794.
[7] M.A. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Trans. Image Processing 21 (8) (2012) 3339–3352.
[8] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Processing 21 (12) (2012) 4695–4708.
[9] P. Ye, J. Kumar, L. Kang, D. Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: Proceedings of the IEEE Conference on Comput. Vis. Pattern Recognit. (2012) 1098–1105..
[10] X.-B. Gao, F. Gao, D.-C. Tao, X.-L. Li, Universal Blind Image Quality Assessment Metrics via Natural Scene Statistics and Multiple Kernel Learning, IEEE Trans. Neural Networks Learning System 24 (12) (2013) 2013–2026.
[11] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Processing 24 (8) (2015) 2579–2591.
[12] J.-C. Yang, Y. Zhao, B. Jiang, W. Lu, X.-B. Gao, No-Reference Quality Evaluation of Stereoscopic Video Based on Spatio-Temporal Texture, IEEE Trans. Multimedia 22 (10) (2020) 2635–2644.
[13] Y. Li, L.-M. Po, X.-Y. Xu, L. Feng, F. Yuan, No-reference image quality assessment with shearlet transform and deep neural networks, Neurocomputing 154 (2015) 94–109.
[14] W.-L. Hou, X.-B. Gao, D.-C. Tao, L.X.-L. Blind Image Quality Assessment via Deep Learning, IEEE Trans. Neural Networks Learning Systems 26 (2015) 1275–1286.
[15] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: IEEE Conf. on Comput. Vis. Pattern Recognit. (2014) 1733–1740..
[16] L. Kang, P. Ye, Y. Li, D. Doermann, Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks, IEEE Int. Conf. Image Process (2015) 2791–2795.
[17] J. Kim, S. Lee, Fully deep blind image quality predictor, IEEE J. Sel. Topics Signal Process 1 (1) (2017) 206–220.
[18] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-end blind image quality assessment using deep neural networks, IEEE Trans. Image Processing 27 (3) (2018) 1202–1203.
[19] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Trans. Image Processing 27 (1) (2018) 206–219.
[20] J. Kim, A.-D. Nguyen, S. Lee, Deep CNN-based blind image quality predictor, IEEE Trans, Neural Networks Learning System 30 (1) (2019) 11–24.
[21] J.-C. Yang, K.-H. Sim, X.-B. Gao, W. Lu, Q.-G. Meng, B.-H. Li, A Blind Stereoscopic Image Quality Evaluator With Segmented Stacked Autoencoders Considering the Whole Visual Perception Route, IEEE Trans. Image Processing 28 (3) (2018) 1314–1328.
[22] J. Kim et al., A.C.B., Deep Convolutional Neural Models for Picture-Quality Prediction: Challenges and Solutions to Data-Driven Image Quality Assessment, IEEE Signal Processing Mag. 34 (6) (2017) 130–141.
[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A largescale hierarchical image database, IEEE Conf. Computer Vision Pattern Recognition (2009) 248–255.
[24] H. Liu, H. Yang, Z.-K. Pan, B.-X. Huang, J. Wang, A Learning Based Image Quality Assessment Model Assisted with Visual Saliency and Gradient Features, Int. Conf. Signal and Image Processing (ICSIP) (2019) 836–840.
[25] L. He, Y. Zhong, W. Lu, X.-B. Gao, A Visual Residual Perception Optimized Network for Blind Image Quality Assessment, IEEE Access 7 (2019) 176087–176098.
[26] F. Gao, J. Yu, S.-G. Zhu, Q.-M. Huang, Q. Tian, Blind image quality prediction by exploiting multi-level deep representations, Pattern Recogn. 8 (2019) 432–442.
[27] J.-J. Wu, J.-B. Ma, F.-H. Liang, W.-S. Dong, G.-M. Shi, W.-S. Lin, End-to-End Blind Image Quality Prediction With Cascaded Deep Neural Network, IEEE Trans. Image Process. 29 (2020) 7414–7426.
[28] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.C.J. Kuo, Image database TID2013: Peculiarities, results and perspectives, Signal Processing: Image Commun. 30 (2015) 57–77.
[29] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, S. Vluymans, Multiple Instance Learning: Foundations and Algorithms, 1st ed,., Springer, New York, 2016.
[30] M.-A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: A survey of problem characteristics and applications, Pattern Recogn. 77 (2018) 329–353.
[31] Z.-H. Zhou, Y.-Y. Sun, Y.-F. Li, Multi-instance learning by treating instances as non-I.I.D. samples, in: Proceedings of International Conference on Machine Learning, (ICML)..
[32] A. Zafra, M. Pechenizkiy, S. Ventura, ReliefF-MI: an extension of relieff to multiple instance learning, Neurocomputing 75 (1) (2012) 210–218.
[33] Y. Li, D.M.J. Tax, R.P.W. Duin, M. Loog, Multiple-instance learning as a classifier combining problem, Pattern Recogn. 46 (2013) 865–874.
[34] Y.-Q. Zhang, H.-B. Zhang, Y.-J. Tian, Sparse multiple instance learning with non-convex penalty, Neurocomputing 391 (2013) 142–156.
[35] S. Ray, D. Page, Multiple-instance regression, in: Int. Conf. on Mach. Learn. (2001) 425–432..
[36] R.A. Amar, D.R. Dooly, S.A. Goldman, Q. Zhang, Multiple instance learning of real-valued data, J. Mach. Learn. Res. (JMLR) 3 (4) (2001) 3–10.
[37] Z. Wang, V. Radosavljevic, Z.O.B. Han, S. Vucetic, Aerosol optical depth prediction from satellite observations by multiple instance regression, in: Proceeding of SIAM Int. Conf. on Data Mining (2008) 165–176..
[38] K.L. Wagstaff, T. Lane, Salience assignment for multiple-instance regression, presented at the Int. Conf. Mach. Learn. Workshop Constrained Optim. Structured Output Spaces..
[39] K.L. Wagstaff, T. Lane, A. Roper, Multiple-instance regression with structured data, in: Int. Conf. Data Mining. Workshop (2008) 291–300..
[40] Z. Wang, L. Lan, S. Vucetic, Mixture Model for Multiple Instance Regression and Applications in Remote Sensing, IEEE Trans. Geosci. Remote Sens. 50 (6) (2012) 2226–2237.
[41] J. Harel, C. Koch, P. Perona, Graph-Based Visual Saliency, in: Int. Conf. Neural Information Processing Systems (2006) 545–552.
[42] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceeding of Int. Conf. Learn. Represent. (ICLR) (2015) 1–15..
[43] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Processing 15 (11) (2006) 3440–3451.

[44] E.C. Larson, D.M. Chandler, Most apparent distortion: Full- reference image quality assessment and the role of strategy, J. Electron. Imag. 10 (1) (2010) 011006.
[45] D. Jayaraman, A. Mittal, A.K. Moorthy, A.C. Bovik, Objective quality assessment of multiply distorted images, in: 46th Asilomar Conf. on Signals, Syst. Comput. (ASILOMAR) (2012) 1693–1697..
[46] D. Ghadiyaram, A. Bovik, Massive online crowd-sourced study of subjective and objective picture quality, IEEE Trans. Image Processing 25 (1) (2016) 372–387.
[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: in Proceeding of IEEE Conf. Computer Vision and Pattern Recognition (2016) 770–778..
[48] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014..
[49] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learning Res. (JMLR) 15 (2014) 1929–1958.



Dong Liang received the B.Eng. degree in mechanical design and theory from Wuhan University of Technology, Wuhan, China, in 1997, and the M.Sc. degree in computer software and theory from Xidian University, Xi'an, China, in 2004. He is currently pursuing the Ph.D. degree in intelligent information processing at Visual Information Processing Lab, School of Electronic Engineering, Xidian University. His current research interests include image quality assessment, computational vision, and machine learning.

Xinbo Gao received the B.Eng., M.Sc. and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a post-doctoral research fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education of P. R. China, a Professor of Pattern Recognition and Intelligent System of Xidian University and a Professor of Computer Science and Technology of Chongqing University of Posts and Telecommunications. His current research interests include Image processing, computer vision, multimedia analysis, machine learning and pattern recognition. He has published six books and around 300 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology and a Fellow of the Chinese Institute of Electronics.

Wen Lu received the BSc, MSc and PhD degrees insignal and information processing from Xidian University, China, in 2002, 2006 and 2009 respectively. He is currently the professor at Xidian University and postdoctoral research in the department of electrical engineering at Stanford University, USA. His research interests include image & video quality metric, human vision system, computational vision. He has published 2 books and around 30 technical articles in refereed journals and proceedings including IEEE TIP, TSMC, Neurocomputing, Signal processing etc.

Jie Li received the B.Sc. degree in electronic engineering, the M.Sc. degree in signal and information processing, and the Ph.D. degree in circuit and systems, from Xidian University, Xi'an, China, in 1995, 1998, and 2004, respectively. She is currently a Professor in the school of electronic engineering, Xidian University, China. Her research interests include video processing. Pattern recognition and machine learning. She has published over 50 technical articles in refereed journals and proceedings including IEEE TIP, TSMC, Neurocomputing, Signal processing etc.