Blind Image Quality Assessment via Deep Learning

Weilong Hou, Xinbo Gao, *Senior Member, IEEE*, Dacheng Tao, *Senior Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*— This paper investigates how to blindly evaluate the visual quality of an image by learning rules from linguistic descriptions. Extensive psychological evidence shows that humans prefer to conduct evaluations qualitatively rather than numerically. The qualitative evaluations are then converted into the numerical scores to fairly benchmark objective image quality assessment (IQA) metrics. Recently, lots of learning-based IQA models are proposed by analyzing the mapping from the images to numerical ratings. However, the learnt mapping can hardly be accurate enough because some information has been lost in such an irreversible conversion from the linguistic descriptions to numerical scores. In this paper, we propose a blind IQA model, which learns qualitative evaluations directly and outputs numerical scores for general utilization and fair comparison. Images are represented by natural scene statistics features. A discriminative deep model is trained to classify the features into five grades, corresponding to five explicit mental concepts, i.e., excellent, good, fair, poor, and bad. A newly designed quality pooling is then applied to convert the qualitative labels into scores. The classification framework is not only much more natural than the regression-based models, but also robust to the small sample size problem. Thorough experiments are conducted on popular databases to verify the model's effectiveness, efficiency, and robustness.**

***Index Terms*— Deep learning, image quality assessment (IQA), natural scene statistics (NSS), no reference.**

I. Introduction

THE aim of image quality assessment (IQA) is to devise an approach to assess the quality of perceived visual stimuli. In recent years, this topic has attracted increased attention because of its significance from both theoretical and practical perspectives. On one hand, before being presented to human consumers, images suffer from a great deal of handling and processing, e.g., restoration [1], super-resolution [2], [3], and so on, all of which introduce extra noise. As a result, the distorted images are difficult for human observers to understand or impair the performance of following processing algorithms. On the other hand, IQA, more importantly, is an attempt to solve the puzzle about how humans perceive images and try to mimic the hominine ability.

The extreme receptor of images commonly is human beings. Hence, the humans subjective judgment is always deemed as the most accurate and reliable method to assess visual quality of given images. However, the subjective judgment is not always suitable for applications because its well-known drawbacks are time- and labor-consuming. Consequently, it is very important to design a proper computational model that simulates human visual behaviors, to evaluate images accurately and automatically.

Existing IQA metrics can be classified into three categories according to the accessibility of the reference/original image: 1) full-reference IQA; 2) reduced-reference IQA; and 3) no-reference/blind IQA (BIQA). Of these approaches, BIQA does not require any reference information, which enhances its applicability remarkably and renders it significant in practice. In recent years, machine learning-based models [4]–[12] have obtained promising performance for IQA. Each of these models learns a particular mapping function from image features to perceived quality scores. To train these models, a huge training set that includes images and their corresponding subjective scores must be obtained by conducting subjective experiments. Subjects are asked to label an image with a numerical score, but the numerical scores evaluated by human observers are strongly affected by individual experience and background. For example, scoring an image with either 70 or 75 is difficult and irregular as a result of individual subjectivity. This kind of evaluation is clearly very noisy and thus unreliable.

Psychological evidence shows that humans prefer to conduct evaluations qualitatively, not quantitatively, using natural language. Qualitative description is often said to be naturalistic, that is, its goal is to understand behavior in a natural setting [13]. Hence, people are not likely to describe image quality with exact numbers in practice. Instead, qualitative adjectives are usually used, such as excellent, good, and bad. Therefore, asking subjects to qualitatively evaluate image quality is a much more natural and operable way to conduct subjective experiments, and can dramatically reduce the randomness of the scores and the burden placed on subjects.

Manuscript received January 11, 2014; revised June 22, 2014; accepted July 3, 2014. Date of publication August 6, 2014; date of current version May 15, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61125204, Grant 61125106, Grant 61172146, Grant 61101250, and Grant 61372130, in part by the Fundamental Research Funds for the Central Universities under Grant K5051202048, Grant BDZ021403, Grant JB149901, in part by the Program for Changjiang Scholars and Innovative Research Team with the University of China under Grant IRT13088, in part by the Shaanxi Innovative Research Team for Key Science and Technology under Grant 2012KCT-02, in part by the Key Research Program, Chinese Academy of Sciences, Beijing, China, under Grant KGZD-EW-T03, and in part by the Australian Research Council under Project FT-130101457, Project DP-120103730, and Project LP-140100569.

W. Hou and X. Gao are with the School of Electronic Engineering, Xidian University, 2 South Taibai Road, Xi'an, 710071, Shaanxi, P. R. China (e-mail: weilonghou@gmail.com; xbgao@ieee.org).

D. Tao is with the Centre for Quantum Computation and Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 235 Jones Street, Ultimo, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China (e-mail: xuelong_li@opt.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2336852

Contemporary subjective experiment is always based on this principle [14]. Two kinds of subjective evaluation are principally used to construct an IQA database. The absolute category rating is recommended by the international telecommunication union [14] for image/video quality assessment, and widely used in most IQA databases, such as LIVE [15], IVC [16], and MICT [17]. In addition, pairwise comparison is also a very popular method of conducting subjective evaluation [18], based on the fact that it is much easier for observers to rank the images according to their qualities than it is to score each of them in isolation. It shows strong robustness to build large-scale databases for IQA. Then, the acquired evaluations are converted into numerical ratings to provide a groundtruth for fair comparison of objective IQA metrics.

Recently, lots of machine learning-based IQA models are proposed by analyzing the mapping from the images to the numerical ratings. But, the learnt mapping can hardly be accurate enough because some information has been lost in such an irreversible conversion from the linguistic descriptions to numerical rating. Therefore, why cannot the model learn from the qualitative evaluation directly?

The major challenge, therefore, is how to learn rules from non-numerical descriptions by humans and output a numerical score for follow-up processing algorithms. In specific, this paper focuses on how to learn rules from qualitative description rather than comparison order, because, in daily life, it is important to evaluate the quality of an image without using a comparison example, e.g., humans simply use personal experience (which is amphibolous) to describe images.

To bridge the gap between the qualitatively labeled samples and numerical outputs, a novel BIQA model via deep learning is proposed in this paper. We recast the blind assessment as a five-grade classification problem, corresponding to five explicit mental concepts, i.e., excellent, good, fair, poor, and bad, to facilitate learning the qualitative descriptions given by humans. A simple yet efficient quality pooling is applied to produce numerical outputs for general utilization. Notably, input images are represented by natural scene statistics (NSS) features [19], [20]. With this representation, the images are first classified into five grades with probabilistic confidence by a deep classifier, which is pre-trained with the deep belief net (DBN) [21] and discriminatively fine-tuned by back-propagation. The labels and their corresponding probabilistic confidences are subsequently transformed to numerical scores in the quality pooling phase, as shown in Fig. 1. A simple subjective test is conducted to provide the parameter settings of the proposed model.

The main contribution of this paper is the new classification-based framework for IQA. To the best of our knowledge, this framework is original and conceptually different from the existing regression-based approaches. Four significant advantages are summarized as follows.

1) *Reasonability:* The proposed model adopts a classification framework instead of a regression framework, which has been widely used in the current IQA schemes. Since human prefers to evaluate images with linguistic labels, the proposed classification-based model is much more natural than the regression-based models.

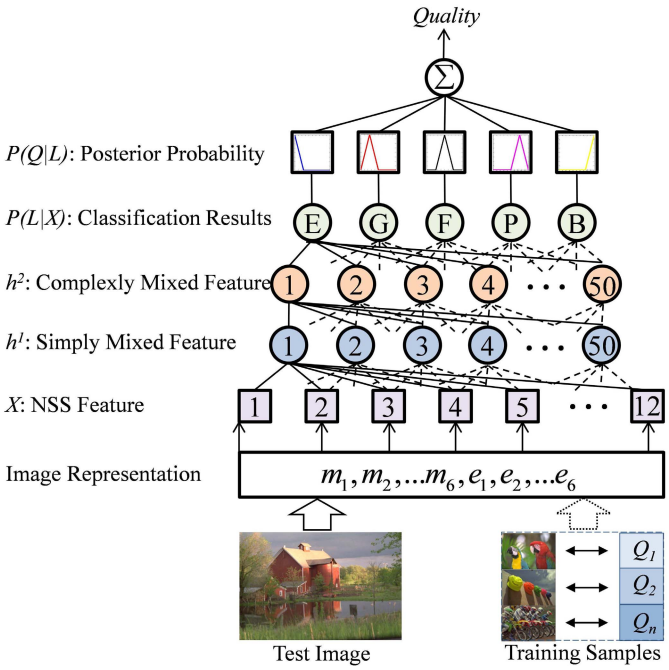

Fig. 1. Overview of the proposed BIQA framework via deep learning. Deep model is pre-trained by DBN and discriminatively fine-tuned by backpropagation. Characters in the classification results represent excellent, good, fair, poor, and bad, respectively.

2) *Effectiveness and Efficiency:* The proposed model is universal blind. Experimental results show that its prediction is highly correlated with the human evaluation. In addition, after learning stage, the proposed model has very low-time complexity.
3) *Robustness:* The model is robust to the small sample size problem. With the aid of classification-based framework, the new model only requires a relatively small size training set to achieve remarkable performance by comparing with the state-of-the-art approaches.
4) *Comprehensiveness:* With the aid of new designed quality pooling, the model can provide three-level quality descriptions, i.e., the qualitative labels, the quality distribution for a given population, and the numerical score, which is more informative and comprehensive than the regression-based IQA methods.

Thorough experiments are conducted on LIVE II [15], TID2008 [18], CSIQ [22], IVC [16], and MICT [17] databases to verify the effectiveness and efficiency of the new BIQA framework, and to demonstrate its robustness to small training data sets.

A. Related Work

Previous BIQA models relied on strong hypotheses. The distortion type, in particular, was given or predefined in advance, which is not applicable in real applications. For example, Wang *et al.* [23] proposed a computational and memory efficient quality assessment model for JPEG images to deal with blocking artifacts. Sheikh *et al.* [5] explored NSS by a learning-based model to measure the quality of

a JPEG2000 image. They claimed that natural scenes contain nonlinear dependencies that are disturbed by the compression process, and this disturbance can be quantified and related to human perception quality. Zhong *et al.* [24] presented a semantic no-reference image sharpness metric, using image tags from the Internet to explore human intention. Applying top–down and bottom–up saliency map to reweight the image quality, Varadarajan and Karam [25] proposed an improved perception-based no-reference image sharpness metric. They used iterative edge refinement to increase the correlation between the perceived sharpness and the sharpness metric. However, these methods perform well only when distortions are known and precisely modeled.

Many universal machine learning-based methods have since been proposed. BIQI [6] and LBIQ [7] first determine the distortion type of a given image and then employ an associated distortion-specific metric to predict its quality. Recently, techniques have been developed to directly map image features to subjective quality without distinguishing different types of distortion. For example, Saad *et al.* [8] and [9] devised BLIINDS and BLIINDS-II using NSS. Li *et al.* [11] used a general regression network to regress the image features to quality scores. Mittal *et al.* [26] presented BRISQUE to predict image quality in the spatial domain, Ye *et al.* [12] proposed CORNIA to extract features using K-means clustering and applied SVM to map the feature to quality. He *et al.* [10] proposed integrated sparse representation with NSS features in the frame of sparse coding. By weighting subjective scores, the final visual quality values are obtained; it is a simple yet effective algorithm.

These approaches nevertheless have certain drawbacks. First, many models adopt machine learning algorithms to find the correlation between images and scores, but conventional machine learning methods have insufficient depth to ascertain the highly structured representation in extremely noisy samples [27]. Second, all of these methods exploit the numerical labeled sample, which is an unnatural way to describe image quality and is not informative. Third, most of them need a large set of images associated with subjective scores to achieve a relatively good performance, which is expensive and time consuming. Recently, Mittal *et al.* [28] introduced probabilistic latent semantic analysis, which is totally free of subjective scores, to learn the latent quality factors. However, its performance is not good enough to use in practice. Considering the problems faced by the approaches discussed above, we propose using qualitative labels to train a deep learning network. The experimental results show that the classification framework is highly correlated with human perception.

II. BIQA via Deep Learning

The diagram of the proposed BIQA model via deep learning is shown in Fig. 1. After training a DBN, a test image is input to the image representation phase and then forwarded hierarchically to the discriminative deep model. The predicted quality score is ultimately obtained from the quality pooling phase.

A. Image Representation

In conventional applications of deep learning, e.g., recognition [21], image patches are always directly imported into the deep architecture without the extraction of statistical features. A huge number of labeled samples and relatively low-dimensional patches guarantee that the deep architecture can be tuned properly. For IQA, however, we assess a whole image instead of patches in the image, because visual quality is a holistic concept of an image. If we use the high-dimensional image data as the input of the deep architecture, an extremely large labeled data set would be required to train a valid model, which is tedious and impractical. Therefore, it is necessary to devise a relatively low-dimensional representation that can comprehensively encode image quality and facilitate the training of the deep model with comparatively small data sets.

Humans can easily perceive the distortions or artifacts in natural images and there must therefore be particular structures that distinguish the unnatural from the natural. Such structures are called NSS [19], [20]. Many researchers find that the NSS in the wavelet domain can be grouped into three levels of properties: 1) primary; 2) secondary; and 3) tertiary [29].

Primary properties give the wavelet coefficients of natural images significant statistical structure, such as locality and multiresolution. Secondary properties, which consist of non-Gaussianity and persistency, give rise to joint wavelet statistics. The literature shows that these properties alter when the image is noised or distorted. As a result of the tendency to change, many BIQA algorithms have been devised [5], [8]. However, these properties change irregularly with different kinds of visual content or distortion. The corresponding methods might perform well on distortion-specific tasks, but they fail to assess universal image degeneration. Fortunately, the tertiary properties show the self-similar property of scenes, of which the exponential decay across scales is the most significant property. It reflects that the magnitudes of the wavelet coefficients of real-world images decay exponentially across scale. Furthermore, the exponential decay is less dependent on particular image content and is therefore suitable for constructing a universal BIQA method. Following previous works [10], the exponential decay property is used in this paper as the image representation.

An image is initially decomposed into three scales using wavelet transform. Nine sub-bands are obtained (there are three sub-bands per scale). Because the low–high (LH) sub-band has a very similar statistical property to the high–low (HL) sub-band, we do not distinguish between LH and HL sub-bands in the same scale. Therefore, six sub-bands are used in total to calculate features. In each sub-band, the magnitude m_k and the entropy e_k are calculated according to

$$m_k = \frac{1}{N_k \times M_k} \sum_{j=1}^{N_k} \sum_{i=1}^{M_k} \log_2 |C_k(i, j)| \tag{1}$$

$$e_k = \sum_{j=1}^{N_k} \sum_{i=1}^{M_k} p[C_k(i, j)] \ln p[C_k(i, j)] \tag{2}$$

where N_k and M_k are the length and width of the kth sub-band, respectively, $C_k(i, j)$ stands for the (i, j) coefficient of the kth sub-band, and $p[\cdot]$ is the probability density function of the sub-band.

The image representation is obtained by combing six sub-bands into a single vector

$$X = [m_1, m_2, \ldots m_6, e_1, e_2, \ldots e_6]^T \tag{3}$$

where the magnitude m_k encodes the sub-band energy and the entropy e_k represents the information content.

B. Discriminative Deep Model

Deep learning models have thrived over the years [30], not only in the field of computer vision, but also in many others, such as audio [31], natural language processing [32], and so on [21], [33]–[35]. It has long been proved that deep networks are much more representative and efficient than shallow ones [27], [36]. However, traditional learning algorithms have failed to train such a deep network because they always return poor local solutions due to the extreme nonlinearity. Hinton *et al.* [21] made a breakthrough when they employed a pre-training strategy to regulate the weight space of deep networks followed by a supervised fine tuning. Since then, deep learning has been a great success.

Generally speaking, the deep network facilitates the proposed method in two ways. On one hand, deep network is an efficient way to represent highly varying functions. It can mine the inherent structure of data without labels, which is inspired by the fact that humans heavily use unsupervised learning. Therefore, it would be an excellent model for learning the highly varied mapping between visual stimuli and quality, because the human perception of quality has an extremely strong nonlinearity, and researchers still have inadequate insight into its mechanism. On the other hand, the deep network has a stronger power of generalization than shallow methods, especially when training samples are limited [31]. In this case, depth and pre-training act as a smart regularization choice to help the model prevent overfitting. When the training set is small, even shallow machine learning methods can fit the training set perfectly, but they generalize poorly.

Deep learning has been widely used for image classification over the years [21], [34]. In the proposed framework, a four-layer discriminative deep model is used to assign image representation to five grades corresponding to the five adjective labels in the LIVE II database [15], i.e., excellent, good, fair, poor, and bad. The discriminative deep model is pre-trained by DBN [21] and fine-tuned by backpropagation. As shown in Fig. 1, the first layer is filled by NSS feature X. The second and third layers form the simply and complexly mixed feature, respectively. The L layer is the classification results with corresponding probabilistic confidence $P(L|X)$. Specially, there are 12 and 5 nodes in the input and output layer, respectively, and 50 nodes in each hidden layer. The joint distribution between image representation X and the three hidden layers is as follows:

$$P(X, h^1, h^2, L) = P(X|h^1)P(h^1|h^2)P(h^2, L). \tag{4}$$

The training algorithm learns the relationship between the image representation and labels in two phases.

1) *Phase 1:* Pre-training parameters for the two adjacent layers using restricted Boltzmann machine (RBM).
2) *Phase 2:* Fine-tuning all the parameters by back propagation.

During this first phase, the DBN is pre-trained in an unsupervised greedy layer-wise manner. Each layer is initialized as an RBM, which is restricted to a single visible layer and a single hidden layer. Taking the first two layers as an example, the RBM defines a probability distribution as

$$P(X, h^1) \propto \exp(X^T W h^1 + c^T h^1 + b^T X) \tag{5}$$

where X forms the visible layer and h^1 forms the hidden layer. There are symmetric connections W between the visible layer and the hidden layer, but no connection for variables within the same layer. c and b are the bias of two layers, respectively.

This particular configuration makes it easy to compute the conditional probability distributions $P(X|h^1)$ and $P(h^1|X)$. The contrastive divergence is used for fast learning of the parameters. In particular, the input X is presented to the visible layer and values are forwarded to the h^1 layer. In reverse, the feature layer is stochastically reconstructed by the h^1 layer. Performed iteratively, the difference in the correlation of the h^1 layer and the feature forms the basis for a weight update, which is a process known as Gibbs sampling. The detailed RBM training process can be found in [21].

In Phase 2, we use a discriminative version of DBN to model $P(L|X)$. We fine-tune the network by maximizing the conditional distribution $P(L|X)$ instead of the joint distribution $P(L, X)$, because there is no need to estimate the feature-given-label conditional distribution $P(X|L)$. Notably, the parameters trained by the RBM are reconfigured as a back-propagation network. All the feature-label pairs are repeatedly presented to the network, and through backpropagation fine-tuning, a classifier is obtained.

C. Quality Pooling

Following classification, the input image is assigned to five grades with corresponding probabilistic confidence $P(L|X)$. Because the grades correspond to five adjectives containing intrinsic semantic information, the classification results can be directly used to qualitatively describe the image quality. For example, if the $P(L = \text{Excellent}|X)$ is higher than others for a given image, the quality of this image can be described as excellent. The adjective provides a natural fashion of describing image quality that is analogous to human evaluation; however, it cannot be used by other applications and fails to compare favorably with existing IQA methods.

To solve this problem, the model needs to know how to relate the labels to scores. In practice, as a result of personal experience and background, different people may have different opinions about the same image. Therefore, we assume the following.

1) Each image has an intrinsic quality Q.
2) Each well-trained individual gives constant labels when assessing images with the same intrinsic quality.

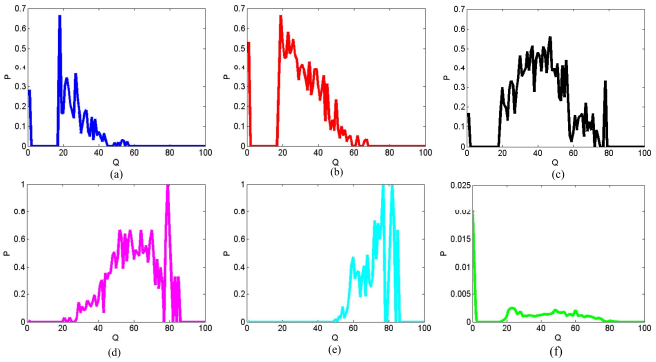

Fig. 2. Likelihood and prior distribution recorded by subjective test. (a) $P(L = \text{Excellent}|Q)$. (b) $P(L = \text{Good}|Q)$. (c) $P(L = \text{Fair}|Q)$. (d) $P(L = \text{Poor}|Q)$. (e) $P(L = \text{Bad}|Q)$. (f) $P(Q)$.

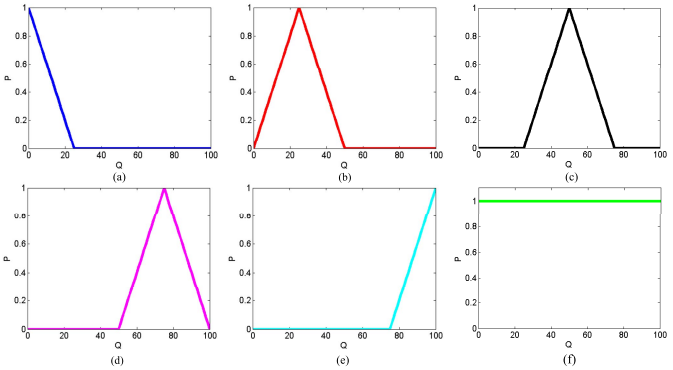

Fig. 3. Hypothetical likelihood and prior distribution. (a) $P(L = \text{Excellent}|Q)$. (b) $P(L = \text{Good}|Q)$. (c) $P(L = \text{Fair}|Q)$. (d) $P(L = \text{Poor}|Q)$. (e) $P(L = \text{Bad}|Q)$. (f) $P(Q)$.

Therefore, for a certain population, $P(L|Q)$ is invariable. Based on Bayes' rule, the posterior probability $P(Q|L)$ can be expressed as

$$P(Q|L) \sim P(L|Q)P(Q) \tag{6}$$

where $P(Q)$ indicates the prior probability distribution of images with the intrinsic quality Q. Given the input image representation X, the distribution of the intrinsic quality can be obtained by marginal distribution

$$P(Q|X) = \int P(Q|L)P(L|X)dL. \tag{7}$$

The quality distribution $P(Q|X)$ represents the evaluations by a population. By computing the mean of the quality distribution, the numerical measurement of image quality is given as follows:

$$\text{Quality} = \mathbb{E}[P(Q|X)]. \tag{8}$$

D. Parameter Setting

To compute the quality distribution, the likelihood $P(L|Q)$ and prior probability $P(Q)$ must be obtained in advance. However, the qualitative evaluation is absent in most published databases. Instead, the mean opinion score (MOS) is provided. To address this problem, we conducted a subjective evaluation on the LIVE II database, asking nine naïve subjects to classify the images into five classes, i.e., excellent, good, fair, poor, and bad. By assuming that the intrinsic quality Q is roughly linear to the DMOS (provided in LIVE database), the likelihood probability $P(L|Q)$, and prior distribution $P(Q)$ can be deduced.

In the test, each subject was individually briefed about the goal of the experiment and given a demonstration of the experimental procedure. Most of the images in this database are 768×512 pixels in size. The display monitors are 21.6-in LCD monitors at a resolution of 1024×768 pixels. Subjects viewed the monitors from a distance of approximately twice screen height. The images were shown in random order and the randomization was different for each subject. The subjects reported their judgments of quality by clicking the buttons on a graphical user interface.

The recorded categorical distribution is shown in Fig. 2(a)–(e). The horizontal axis represents the intrinsic quality Q approximated by DMOS and the vertical axis represents the $P(L|Q)$. For example, if $P(L = \text{Excellent}|Q = 40) = 0.1$, it means that an image with an intrinsic quality of 40 may have a 10% chance of being labeled as excellent by the population. Fig. 2(f) shows the quality distribution in the LIVE II database.

As shown by the test results, an image associated with a lower DMOS would be more possible to be labeled as excellent by a certain group of assessors and vice versa, which is intuitively straightforward. It is reasonable to assume that people share the similar assessment standard, and thus the different $P(L|Q)$ should share a similar trend that how to relate the label to quality. For comparison with the subjective results, we also produce hypothetical categorical distributions that have the similar trend to the recorded distributions, as shown in Fig. 3(a)–(e). We use triangular distribution and uniform distribution to represent such an intuitive assumption.

Both the real and hypothetical likelihood and prior distribution are tested Section III, and the results show the trend, which indicates how the labels relate to the qualities, is much more important than the distributions themselves.

III. Experiments

In this section, five experiments are conducted to test the performance of the proposed method. The consistency experiment is used to validate how the objective assessment corresponds to human evaluation. The extensibility experiment is employed to prove whether the proposed method is applicable for various images and distortions without extra training. The rationality experiment justifies the proposed method. The sensitivity experiment is conducted to demonstrate the sensitivity of the proposed method. Last, but not least, the complexity experiment tests the computational efficiency of the proposed method.

Five public IQA databases are used in our experiments, including LIVE II [15], TID2008 [18], CSIQ [22], IVC [16], and MICT [17]. The LIVE II database contains 29 high-resolution 24 bits/pixel RGB color images and 175 corresponding JPEG and 169 JPEG2000 compressed images, as well as 145 white noisy, 145 Gaussian blurred, and 145 fast-fading (FF) Rayleigh channel noisy images at a range of quality levels. The TID2008 database contains

25 reference images and 17 types of distortion are generated for each reference image. The CSIQ consists of 30 original images and each is distorted using six different types of distortion at four to five different levels of distortion. The distortion includes JPEG, JPEG2000, global contrast decrements, Gaussian blurring, additive Gaussian white noise, and additive Gaussian pink noise. In the IVC database, 10 reference images are used, and 235 distorted images are generated from four different processes: 1) JPEG2000; 2) JPEG; 3) blurring; and 4) LAR coding. In the MICT database, distorted images are generated by JPEG and JPEG2000 compression from 14 reference images. Of the comparison criteria recommended by the video quality experts group (VQEG) [37], the Pearson linear correlation coefficient (LCC) and Spearman rank-order correlation coefficient (SROCC) are the most significant for testing the performance of an IQA method. The LCC provides an evaluation of prediction accuracy and SROCC measures the prediction monotonicity. Due to the nonlinearity raised by the subjective rating process, VQEG suggests applying variance-weighted regression analysis to provide a nonlinear mapping between the objective and subjective MOS to facilitate fair comparison. In this case, the LCC and SROCC are computed between the objective and subjective scores after nonlinear regression. We also compute the root mean square error (RMSE) and mean absolute error (MAE) of the fitting procedure after nonlinear mapping as auxiliary comparison criteria.

A. Consistency Experiment

In this section, a thorough experiment on the LIVE II database is conducted to validate how the objective assessment corresponds to human evaluation. Since the machine learning-based models need samples for training, we group the reference images and their corresponding distorted versions, and randomly select several groups for training, retaining the remainder for testing. To remove the influence of the selection of the training set, the proposed method is run 100 times in this way. The performance metrics of LCC, SROCC, RMSE, and MAE are obtained by averaging 100 results. The other learning-based BIQA methods are all achieved the same way.

Four traditional full-reference IQA methods, PSNR, SSIM [38], IFC [39], and VIF [40], are used as the benchmark. In addition, eight BIQA methods are employed to compare: 1) NSS [5]; 2) BIQI [6]; 3) BLIINDS [8]; 4) BLIINDS-II [9]; 5) DIIVINE [41]; 6) SRNSS [10]; 7) BRISQUE [26]; and 8) CORNIA [42]. All of these methods are based on machine learning. The proposed method is abbreviated to DLIQA-R and DLIQA-I, representing the model based on real subjective evaluation and hypothetical data, respectively.

Tables I–IV show the experimental results of all the comparison methods on the LIVE II database with different comparison criteria. The subscript of the name of the method indicates how many groups it uses for training. Larger training sets always enhance method performance as a result of having more information.

Of all the comparison methods, VIF has the best performance as a full-reference IQA method, but it needs the reference images to conduct assessment. Of the blind methods with 23-group training, it is noted that DLIQA-R has the highest LCC on the entire database followed by DLIQA-I, although neither is superior for every distortion. The performance of SRNSS on an entire database is second to the proposed model. The BLIINDS-II and DIIVINE perform better for some individual distortions. They extract 24 and 88 features, respectively, many more than are extracted by the proposed model, but the experimental results show that the features extracted by our model are more applicable for describing image quality, while avoiding the influence of distortion type. In addition, the proposed model outperforms SRNSS for each distortion type except FF, which suggests that the framework more effectively simulates the quality perception of human beings.

TABLE I
LCC OF DIFFERENT METHODS ON LIVE II DATABASE

Method	JP2K	JPEG	WN	GBlur	FF	ALL
PSNR	0.896	0.860	0.986	0.783	0.890	0.824
SSIM	0.937	0.928	0.970	0.874	0.943	0.863
IFC	0.903	0.905	0.958	0.961	0.961	0.911
VIF	0.962	0.943	0.984	0.974	0.962	0.950
NSS_{23}	0.929	0.427	0.835	0.597	0.895	0.504
$BIQI_{23}$	0.942	0.922	0.945	0.941	0.856	0.902
$BLIINDS\text{-}II_{23}$	**0.963**	**0.979**	0.985	0.948	0.944	0.923
$DIIVINE_{23}$	0.922	0.921	**0.988**	0.923	0.888	0.917
$SRNSS_{23}$	0.936	0.939	0.940	0.936	**0.947**	0.932
$BRISQUE_{23}$	0.936	0.937	0.958	0.935	0.898	0.917
$CORNIA_{23}$	0.915	0.902	0.952	0.940	0.913	0.903
$DLIQA\text{-}I_{23}$	0.951	0.941	0.959	0.949	0.889	0.933
$DLIQA\text{-}R_{23}$	0.953	0.948	0.961	**0.950**	0.892	**0.934**
NSS_{15}	0.921	0.366	0.822	0.701	0.722	0.495
$BIQI_{15}$	0.809	0.901	0.954	0.829	0.733	0.821
$BLIINDS_{15}$	*	*	*	*	*	*
$BLIINDS\text{-}II_{15}$	0.934	0.915	0.950	0.926	0.852	0.908
$DIIVINE_{15}$	0.869	0.876	0.951	0.911	0.846	0.864
$SRNSS_{15}$	0.886	0.890	0.880	0.865	0.873	0.886
$BRISQUE_{15}$	0.939	0.916	0.941	0.938	0.869	0.909
$CORNIA_{15}$	0.891	0.893	0.936	0.917	0.881	0.881
$DLIQA\text{-}I_{15}$	0.942	0.935	0.940	0.939	0.887	0.927
$DLIQA\text{-}R_{15}$	**0.947**	**0.940**	**0.955**	**0.944**	**0.890**	**0.930**

TABLE II
SROCC OF DIFFERENT METHODS ON LIVE II DATABASE

Method	JP2K	JPEG	WN	GBlur	FF	ALL
PSNR	0.890	0.841	0.985	0.782	0.890	0.820
SSIM	0.932	0.903	0.963	0.894	0.941	0.851
IFC	0.892	0.866	0.938	0.959	0.963	0.913
VIF	0.953	0.913	0.986	0.973	0.965	0.953
NSS_{23}	0.882	0.247	0.852	0.644	0.859	0.339
$BIQI_{23}$	0.940	0.915	0.971	0.947	0.831	0.903
$BLIINDS\text{-}II_{23}$	**0.951**	**0.942**	0.978	0.944	0.927	0.920
$DIIVINE_{23}$	0.913	0.910	**0.984**	0.921	0.863	0.916
$SRNSS_{23}$	0.928	0.931	0.938	0.933	**0.941**	**0.930**
$BRISQUE_{23}$	0.910	0.919	0.955	0.941	0.874	0.920
$CORNIA_{23}$	0.903	0.889	0.958	0.946	0.915	0.906
$DLIQA\text{-}I_{23}$	0.929	0.910	0.959	0.941	0.849	0.923
$DLIQA\text{-}R_{23}$	0.933	0.914	0.968	**0.947**	0.857	0.929
NSS_{15}	0.908	0.180	0.877	0.737	0.738	0.333
$BIQI_{15}$	0.800	0.891	0.951	0.846	0.707	0.820
$BLIINDS_{15}$	0.922	0.839	**0.974**	**0.957**	0.750	0.800
$BLIINDS\text{-}II_{15}$	0.926	0.883	0.956	0.935	0.859	0.911
$DIIVINE_{15}$	0.862	0.850	0.961	0.938	0.846	0.874
$SRNSS_{15}$	0.863	0.871	0.861	0.860	0.865	0.876
$BRISQUE_{15}$	0.908	0.889	0.941	0.945	0.852	0.911
$CORNIA_{15}$	0.899	0.870	0.938	0.924	**0.901**	0.890
$DLIQA\text{-}I_{15}$	0.924	0.909	0.959	0.941	0.853	0.919
$DLIQA\text{-}R_{15}$	**0.928**	**0.912**	0.968	0.946	0.861	**0.927**

TABLE III
RMSE OF DIFFERENT METHODS ON LIVE II DATABASE

Method	JP2K	JPEG	WN	GBlur	FF	ALL
PSNR	7.187	8.170	2.680	9.772	7.516	9.124
SSIM	5.671	5.947	3.916	7.639	5.485	8.126
IFC	6.972	6.813	4.574	4.360	4.528	6.656
VIF	4.449	5.321	2.851	3.533	4.502	5.024
NSS_{23}	8.911	21.25	12.13	17.22	9.821	19.69
$BIQI_{23}$	8.213	9.233	7.005	**6.566**	11.38	9.849
$BLIINDS\text{-}II_{23}$	7.257	9.103	6.825	7.894	9.709	8.800
$DIIVINE_{23}$	9.660	12.25	**5.310**	7.070	12.93	10.90
$SRNSS_{23}$	7.892	7.948	7.971	7.591	**7.157**	**7.618**
$BRISQUE_{23}$	8.150	9.230	7.273	7.516	9.536	9.538
$CORNIA_{23}$	9.666	10.32	6.541	7.689	8.917	9.935
$DLIQA\text{-}I_{23}$	7.540	7.640	5.942	6.875	9.641	8.254
$DLIQA\text{-}R_{23}$	**7.250**	**7.596**	5.881	6.570	9.540	8.149
NSS_{15}	9.506	22.53	12.53	15.52	15.28	20.09
$BIQI_{15}$	14.84	13.76	8.409	10.23	19.29	15.62
$BLIINDS_{15}$	*	*	*	*	*	*
$BLIINDS\text{-}II_{15}$	8.596	9.593	6.915	8.281	11.36	9.633
$DIIVINE_{15}$	11.72	11.58	6.982	7.176	11.69	11.55
$SRNSS_{15}$	10.88	10.91	10.27	10.86	10.33	10.73
$BRISQUE_{15}$	8.262	9.411	7.294	7.354	10.55	9.609
$CORNIA_{15}$	11.07	11.89	8.445	9.576	10.00	11.78
$DLIQA\text{-}I_{15}$	7.787	8.874	6.912	7.715	10.12	8.854
$DLIQA\text{-}R_{15}$	**7.720**	**8.184**	**6.432**	**7.136**	**9.952**	**8.445**

TABLE IV
MAE OF DIFFERENT METHODS ON LIVE II DATABASE

Method	JP2K	JPEG	WN	GBlur	FF	ALL
PSNR	5.528	6.380	2.164	7.743	5.800	7.325
SSIM	5.461	4.792	3.257	5.760	4.297	6.275
IFC	3.445	3.807	3.816	3.410	3.620	5.182
VIF	3.445	3.807	2.304	2.818	3.547	3.887
NSS_{23}	7.620	17.44	10.14	11.72	7.739	15.45
$BIQI_{23}$	6.881	7.456	5.418	5.128	9.417	7.987
$BLIINDS\text{-}II_{23}$	5.860	7.136	5.377	6.553	7.162	6.930
$DIIVINE_{23}$	8.380	8.172	4.489	7.771	8.653	8.123
$SRNSS_{23}$	6.015	5.968	6.106	6.086	**5.268**	**5.873**
$BRISQUE_{23}$	6.300	7.016	5.335	5.954	7.093	7.327
$CORNIA_{23}$	7.865	8.389	5.102	5.806	6.837	8.072
$DLIQA\text{-}I_{23}$	5.845	5.991	4.752	5.412	7.105	6.348
$DLIQA\text{-}R_{23}$	**5.658**	**5.803**	**4.406**	**5.119**	6.888	6.076
NSS_{15}	8.331	18.84	10.49	10.88	10.79	15.85
$BIQI_{15}$	9.950	8.429	5.826	5.898	10.774	9.661
$BLIINDS_{15}$	*	*	*	*	*	*
$BLIINDS\text{-}II_{15}$	6.776	7.954	5.418	6.729	7.675	7.573
$DIIVINE_{15}$	9.698	9.441	5.476	5.638	8.761	9.221
$SRNSS_{15}$	7.965	7.865	8.015	8.312	7.625	8.106
$BRISQUE_{15}$	6.544	7.460	5.356	5.782	7.749	7.385
$CORNIA_{15}$	9.280	9.783	6.963	8.783	8.954	10.88
$DLIQA\text{-}I_{15}$	6.051	6.415	4.981	5.715	7.451	6.716
$DLIQA\text{-}R_{15}$	**5.932**	**6.112**	**4.831**	**5.406**	**7.201**	**6.325**

Of the methods with 15-group training, both DLIQA-R and DLIQA-I outperform all the comparison methods for each distortion. It is remarkable that the performance of the proposed method does not drop sharply along with the reduction of the training set. The experimental results demonstrate that the framework via deep learning is robust against the small sample problem.

Figs. 4 and 5 show the scatter plots of MOS versus the quality predicted by NSS, BIQI, BLIINDS-II, DIIVINE, SRNSS, and DLIQA-R. In Fig. 4, the methods are trained on 23 groups and in Fig. 5, 15 groups are selected for training. The scatter plots give a visual expression of their performance. Each plus sign represents a test image. The closer these plus signs are to the fitting curve, the better performance the model has. It can be shown that the scatter plot of the proposed method

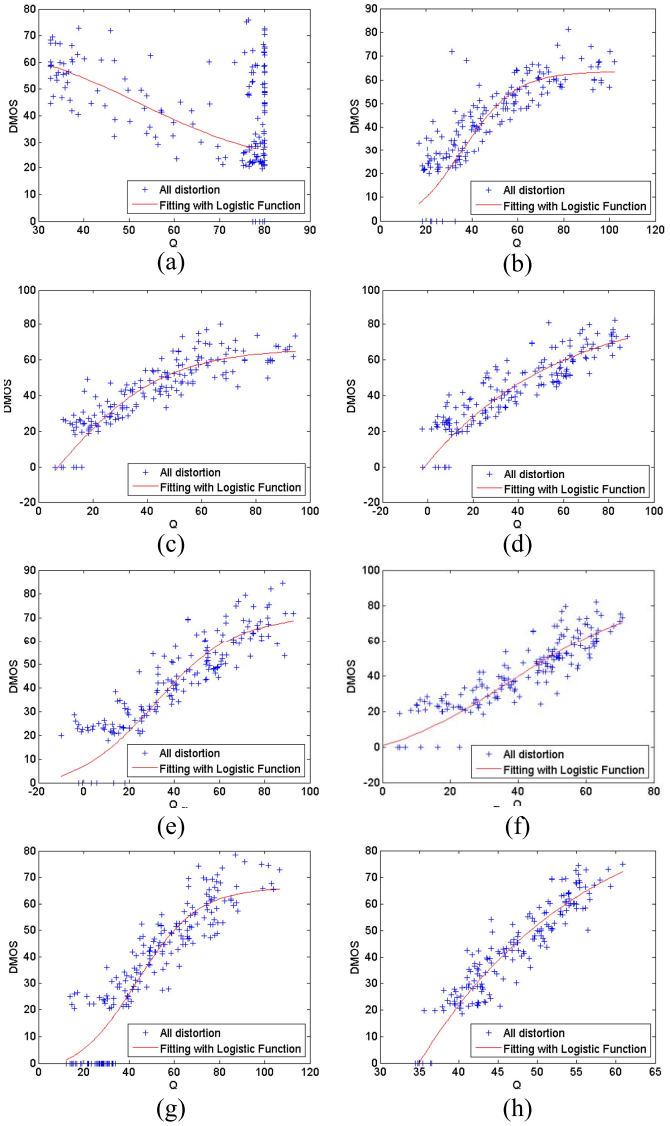

Fig. 4. Scatter plots of MOS versus prediction of the test methods with 23-group training. (a) NSS. (b) BIQI. (c) BLIINDS-II. (d) DIIVINE. (e) SRNSS. (f) BRISQUE. (g) CORNIA. (h) DLIQA-R.

is clearly more compact than others in Fig. 5, i.e., those with 15-group training, which proves that the proposed method is robust against the small sample problem.

Fig. 6 shows the relationship between the number of groups for training and the performance of DLIQA-R. LIVE II database contains 29 group images. We train our method using one to 28 groups and DLIQA-R is tested on each of the other images 100 times. The average LCC, SROCC, RMSE, and MAE, along with their 95% confidence interval, are obtained. As the figure shows, the curve decreases slowly with the reduction of the training set. With only nine groups for training, the LCC is still bigger than 0.9. The RMSE and MAE curve shows the same trend, which demonstrates that the proposed method performs better when the training images are fewer.

B. Extensibility Experiment

To verify the extensibility and generalization of the blind methods, we train them on the LIVE II database and test them

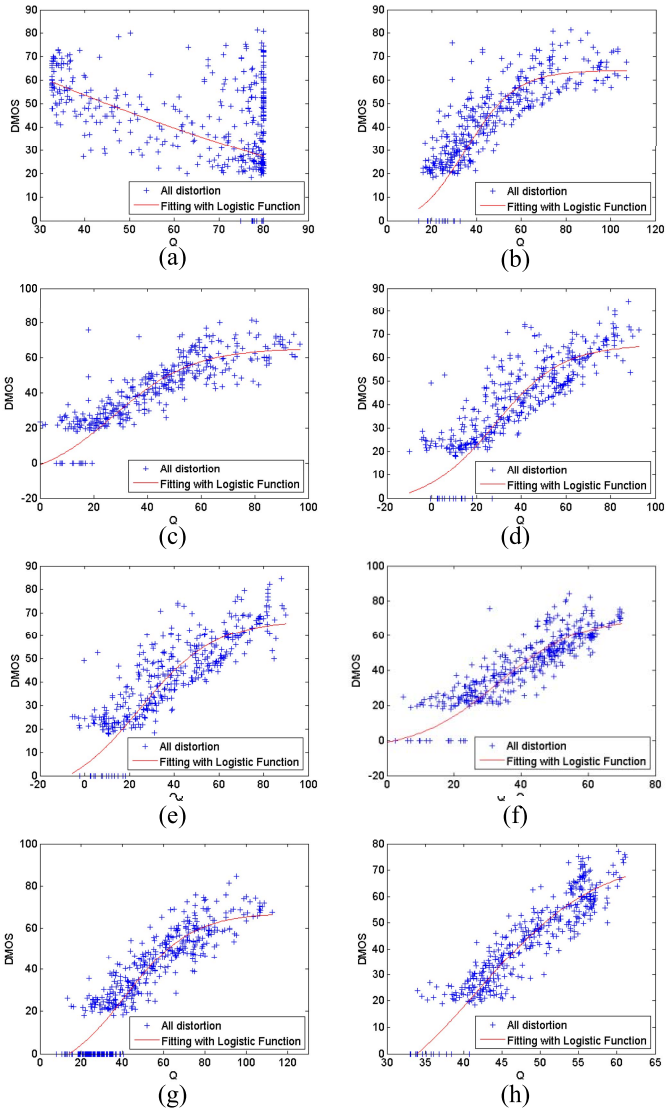

Fig. 5. Scatter plots of MOS versus prediction of the test methods with 15-group training. (a) NSS. (b) BIQI. (c) BLIINDS-II. (d) DIIVINE. (e) SRNSS. (f) BRISQUE. (g) CORNIA. (h) DLIQA-R.

on other publicly available databases, including TID, CSIQ, IVC, and MICT. The machine learning-based methods are highly affected by the training data. In the experiment, the test methods are trained using the LIVE database and tested on others. Many distortion types in test datasets do not appear in the training dataset. Therefore, for better illustration, the test images are separated into two groups according to whether their distortion types appear in the training dataset. Group 1 includes the distortion appearing in the LIVE II database and group 2includes the rest.

Figs. 7 and 8 show the LCC metrics of NSS, BIQI, BLIINDS-II, DIIVINE, SRNSS, BRISQUE, CORNIA, and the proposed model in the two groups. In group 1, all the distortion types appear in the training data set. It is observed that most of the test methods show relatively good performance across databases (Fig. 7). Among the methods, the proposed DLIQA-R outperforms the others for most distortion types. For 6 of 13 distortion types, the LCC metric of DLIQA-R reaches over 0.9, and for 11 of 13 distortion types, it reaches over 0.8. The results demonstrate that the proposed framework is robust against different image databases. The extensibility of the proposed method is verified.

In group 2, the test methods try their best to predict the images with unknown distortion types. In most cases, the majority fails. This is because the training data set does not contain these kinds of distortion. However, in some cases, such as high-frequency noise in the TID database, the test methods perform well. This might be because these kinds of distortion have similar visual appearance to the distortions in the training set. For example, high-frequency noise looks very similar to the white noise, which obtains the highest performance in group 2. By contrast, the local block-wise distortion in TID has a distinct visual appearance with the learnt distortion types, which results in the failure of the method.

To further illustrate the influence of the training data set, we train the proposed method on part of the TID database and use the rest of the database for testing. The TID database is divided into 25 groups according to the reference images. The 20 groups are randomly selected for training and the rest for testing. After 100 runs, the average LCC metric is shown along with the performance trained on LIVE II database in Fig. 9. It is observed that using part of the TID database for training remarkably enhances the LCC metrics for almost all distortion types. However, the performance is not improved for impulse noise and JPEG compression. We use 20 out of 25 groups for training, i.e., 80 images of each distortion, and 17 distortion types are mixed. We believe that the training samples are insufficient compared to the distortion types, which results in the deterioration. Thus, the TID database is more challenging than the LIVE II database.

C. Rationality Experiment

A rationality experiment [10], [43] is conducted in this section to demonstrate that the proposed framework produces rational evaluations according to different degrees of distortion. Four types of distortion, JPEG2000 compression, JPEG compression, Gaussian blurring, and Gaussian white noise, are adopted in this section. When the images are compressed by JPEG2000, the quantification processing sets many small wavelet coefficients at zero, which results in ringing and blurring distortion. Blurring loses high-frequency information and ringing introduces many artifacts. For the JPEG compression, the images are degenerated with blocking artifacts and blurring within blocks. We use the original image Monarch from the LIVE II database and generate its six distorted versions for each distortion type. The prediction trends of the proposed BIQA framework are shown in Fig. 10. In these figures, the vertical axis indicates the predicted quality by the proposed method and the horizontal axis represents JPEG2000/JPEG compression rate R. Gaussian blurring window size is W and Gaussian noise variance is V. For ease of viewing, only part of the images is shown in the figures. As the figures show, the proposed method produces rational quality prediction, which is consistent with the variations of the degree of distortion. When the degree of distortion increases, the predicted quality scores rises. It is noted that the proposed method is tuned to produce a DMOS-

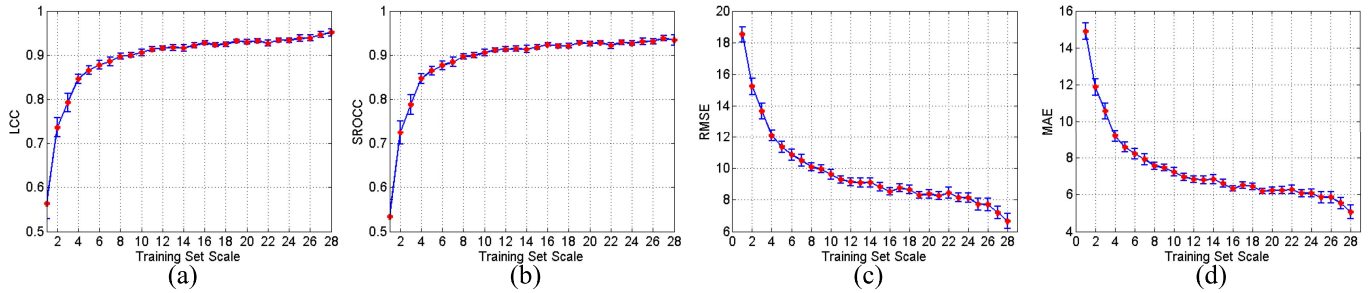

Fig. 6. Performance of the proposed method versus groups of training images. Results are averaged on 100 times run. (a) LCC metric. (b) SROCC metric. (c) RMSE metric. (d) MAE metric.

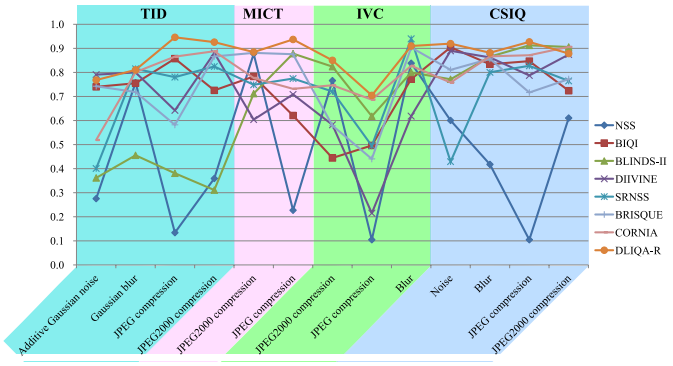

Fig. 7. LCC metrics of test methods in group 1. The horizontal axis denotes the distortion types in each database and the vertical axis denotes the LCC.

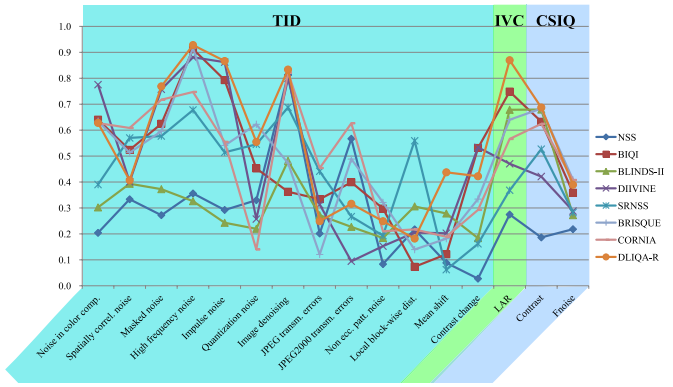

Fig. 8. LCC metrics of test methods in group 2. The horizontal axis denotes the distortion types in each database and the vertical axis denotes the LCC.

like score, which has a higher score value with lower visual quality.

Fig. 10(d) also illustrates that the quality trend does not increase monotonously. When the variance of Gaussian white noise is large, the numerical scores are slightly inconsistent with the deviation in variance. Regarding the experiments, results are tolerable because of the disagreement in the evaluations scores of low-quality images provided by subjects.

D. Sensitivity Experiment

A sensitivity experiment is conducted to test whether a blind method produces reasonable assessment. It is well known that, although PSNR is widely used, it has a fatal drawback: it cannot perceive image content and only computes the mathematical difference between two images. Therefore, the images of obviously different visual quality might have the same PSNR scores. We use the original image Monarch from the LIVE II database and generate its three distorted versions: 1) the mean shift image; 2) the contrast-stretching image; and 3) the JPEG compressed image. The three distorted images have very similar PSNR scores (shown in Fig. 11). Table V reports the quality of their PSNR, VIF, NSS, BIQI, BLIINDS-II, DIIVINE, SRNSS, BRISQUE, CORNIA, and DLIQA-R scores. Since the contrast-stretching is, indeed, an image enhancement, the contrast-stretching images should have higher quality. However, the LIVE II database does not contain that distortion. The reference image is regarded as perfect and therefore has the highest quality. All the comparison methods are trained on LIVE II database. Except NSS, they all produce reasonable results, in which the reference get the highest quality and the JPEG compressed image get the lowest quality. In addition, we train the proposed model on TID database as well (denoted by DLIQA/TID). The results show that the contrast-stretching image get lower score (higher quality) than mean-shift and JPEG compressed one, but still slightly higher than reference. It might be because the training data set contain no reference images. On some level, all test methods are consistent with human visual perception except PSNR and NSS. The sensitivity of the proposed framework can be verified.

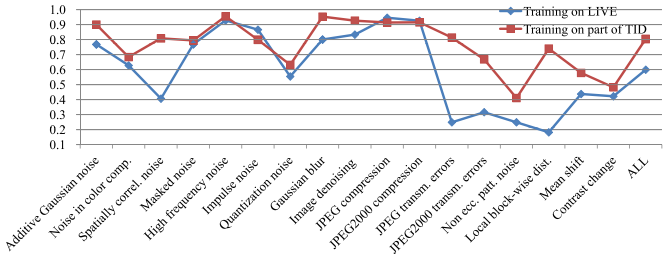

Fig. 9. LCC metrics on TID database of the proposed method.

E. Complexity Experiment

We also conduct an experiment to demonstrate the time complexity of the proposed method. Because DLIQA-R and DLIQA-I only use different likelihood and prior distribution, their time and space complexity are exactly the same. In this section, only DLIQA-R has been tested. In addition, NSS, BIQI, BLIINDS-II, DIIVINE, SRNSS, BRISQUE, and

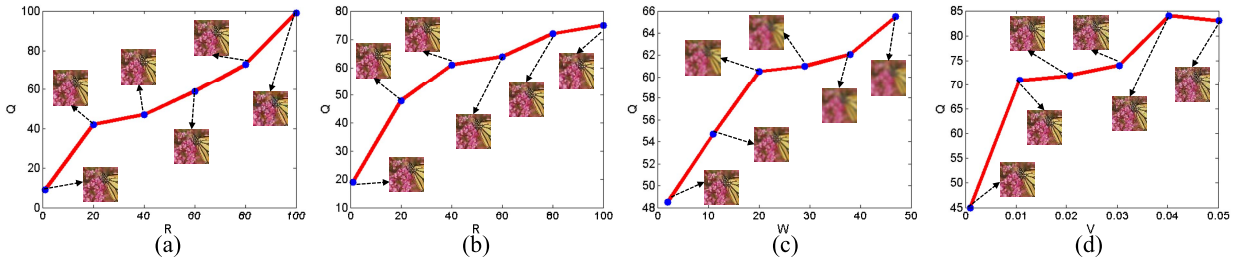

Fig. 10. Quality trend predicted by the proposed method for different distortion types. (a) JPEG. (b) JPEG2000. (c) Gaussian blurring. (d) Gaussian white noise.

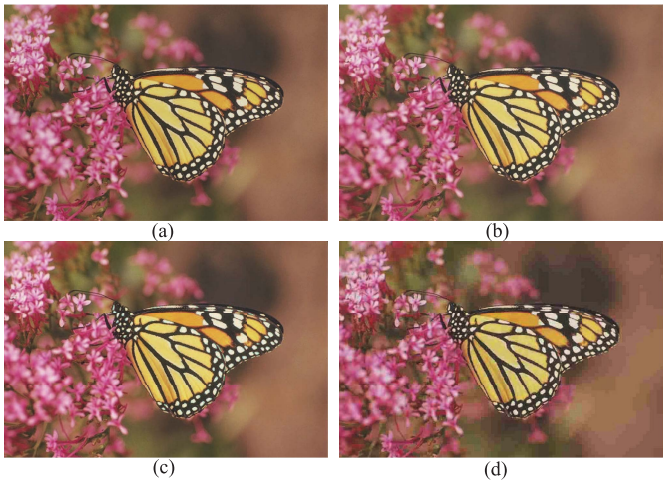

Fig. 11. Images with similar PSNR, but different visual appearances. (a) Original image. (b) Mean shift, PSNR = 30.0692. (c) Contrast stretching, PSNR = 30.6863. (d) JPEG compression, PSNR = 30.2582.

TABLE V
QUALITY OF DIFFERENT METHODS FOR IMAGE IN FIG. 11

Method	Original	Mean shift	Contrast stretching	JPEG
PSNR	*	30.07	30.69	30.26
SSIM	*	0.91	0.65	0.41
IFC	*	0.88	0.71	0.45
VIF	*	0.83	0.75	0.39
NSS	79.01	79.99	79.97	79.89
BIQI	18.20	22.34	28.59	34.03
BLIINDS-II	19.50	33.00	31.00	61.50
DIIVINE	12.01	16.87	17.08	23.63
SRNSS	35.51	38.36	38.47	60.79
BRISQUE	3.137	17.13	17.66	44.14
CORNIA	20.72	35.17	31.28	48.39
DLIQA/LIVE	25.91	37.88	37.90	53.72
DLIQA/TID	42.99	45.42	43.21	56.84

CORNIA are tested in the experiment. The MATLAB codes all come from their official websites or authors. The runtime environment is MATLAB R2010b on 64 bit Windows7. For each method, the same image is evaluated 10 times and total time consumed is recorded by the MATLAB functions tic and toc. Divided by 10, the CPU time is obtained. The results are shown in Table VI. We do not run optimization for our MATLAB code. As shown in Table IV, the proposed DLIQA-R has the lowest time complexity of all the comparison methods. In addition, our method extracts only 12 features to express image quality, making it one of the methods using the fewest features.

TABLE VI
TIME COMPLEXITIES OF BIQA METHODS

Method	CPU time/s
NSS	1.0098
BIQI	1.2968
BLIINDS-II	151.9483
DIIVINE	42.6760
SRNSS	0.5057
BRISQUE	0.3035
CORNIA	0.8642
DLIQA-R	0.1942

IV. CONCLUSION

This paper proposes a novel classification-based framework for universal BIQA via deep learning. Aiming to learn from the linguistic description of image quality and output numerical scores for general utilization, the proposed model recasts the BIQA as a classification problem associated with a newly designed quality pooling. A deep learning network is designed to classify (actually score) an image into five grades and then convert the labels to numerical scores. Thorough experimental results verify its effectiveness, efficiency, and robustness to small training sets, demonstrating that the proposed model corresponds well to human evaluation. It has the potential to perform even better. The NSS feature is strong in representing image quality, but it is hand-crafted, which requires time-consuming hand-tuning. The extensibility experiment shows that the features fail to express certain eccentric distortions. Using deep learning to learn more powerful image representation for describing image quality remains a great challenge that has yet to be resolved. Developing semisupervised or unsupervised convolutional neural networks could obtain effective solutions in the future work.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

[1] Y. Xia, C. Sun, and W. X. Zheng, "Discrete-time neural network for fast solving large linear L_1 estimation problems and its application to image restoration," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 812–820, May 2012.

[2] J. Yu, X. Gao, D. Tao, X. Li, and K. Zhang, "A unified learning framework for single image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 4, pp. 780–792, Apr. 2014.

[3] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with multiscale similarity learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1648–1659, Oct. 2013.

[4] X. Gao, F. Gao, D. Tao, and X. Li, "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 2013–2026, Dec. 2013.

[5] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[6] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[7] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 305–312.

[8] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, Jun. 2010.

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

[10] L. He, D. Tao, X. Li, and X. Gao, "Sparse representation for blind image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1146–1153.

[11] C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Trans. Neural Netw.*, vol. 22, no. 5, pp. 793–799, May 2011.

[12] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1098–1105.

[13] E. Hutchins and G. Lintern, *Cognition in the Wild*. Cambridge, MA, USA: MIT Press, 1995.

[14] *BT.500: Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. BT.500-11 (06/02), 2002.

[15] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[16] P. Callet and F. Autrusseau. (2006). *Subjective Quality Assessment-IVC Database*. [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/

[17] Y. Horita, K. Shibata, Y. Kawayoke, and Z. Sazzad. (2000). *MICT Image Quality Evaluation Database*. [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html

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

[19] Y. Weiss and W. T. Freeman, "What makes a good model of natural images?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[20] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, 2001.

[21] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

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

[23] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. Int. Conf. Image Process.*, vol. 1. 2002, pp. I-477–I-480.

[24] S.-H. Zhong, Y. Liu, Y. Liu, and F.-L. Chung, "A semantic no-reference image sharpness metric based on top-down and bottom-up saliency map modeling," in *Proc. 17th IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 1553–1556.

[25] S. Varadarajan and L. J. Karam, "An improved perception-based no-reference objective image sharpness metric using iterative edge refinement," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 401–404.

[26] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[27] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[28] A. Mittal, G. S. Muralidhar, J. Ghosh, and A. C. Bovik, "Blind image quality assessment without human training using latent quality factors," *IEEE Signal Process. Lett.*, vol. 19, no. 2, pp. 75–78, Feb. 2012.

[29] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1056–1068, Jul. 2001.

[30] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep machine learning—A new frontier in artificial intelligence research [research frontier]," *IEEE Comput. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, Nov. 2010.

[31] A. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 5060–5063.

[32] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[33] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[34] S.-H. Zhong, Y. Liu, and Y. Liu, "Bilinear deep learning for image classification," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 343–352.

[35] L. Shao, D. Wu, and X. Li, "Learning deep and wide: A spectral method for learning deep networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. PP, no. 99, p. 1, 2014, doi: 10.1109/TNNLS.2014.2308519.

[36] M. Bianchini and F. Scarselli, "On the complexity of neural network classifiers: A comparison between shallow and deep architectures," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1553–1565, 2014.

[37] VQEG. (2009). *Validation of Reduced-Reference and No-Reference Objective Models for Standard Definition Television, Phase I*. [Online]. Available: http://www.vqeg.org/

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

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

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

[41] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[42] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1098–1105.

[43] X. Gao, W. Lu, D. Tao, and X. Li, "Image quality assessment based on multiscale geometric analysis," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1409–1423, Jul. 2009.

Weilong Hou received the B.Eng. degree in electronic information engineering from Xidian University, Xi'an, China, in 2010, where he is currently pursuing the Ph.D. degree in intelligence information processing.

He has been a visiting Ph.D. student with the University of Technology, Sydney, NSW, Australia, since 2013. His current research interests include visual quality assessment, visual attention, and computationally modeling of human visual system.

Xinbo Gao (M'02–SM'07) received the B.Eng., M.Sc. and Ph.D. degrees in signal and information processing from Xidian University, China, in 1994, 1997 and 1999 respectively. From 1997 to 1998, he was a research fellow in the Department of Computer Science at Shizuoka University, Japan. From 2000 to 2001, he was a postdoctoral research fellow in the Department of Information Engineering at the Chinese University of Hong Kong. Since 2001, he joined the School of Electronic Engineering at Xidian University. Currently, he is a Cheung Kong Professor of Ministry of Education, China, a Professor of Pattern Recognition and Intelligent System, and Director of the State Key Laboratory of Integrated Services Networks, Xidian University. His research interests are computational intelligence, machine learning, computer vision, pattern recognition and wireless communications. In these areas, he has published 5 books and around 200 technical articles in refereed journals and proceedings including IEEE TIP, TCSVT, TNN, TSMC, IJCV, Pattern Recognition etc.. He is on the editorial boards of several journals including Signal Processing (Elsevier), and Neurocomputing (Elsevier). He served as general chair/co-chair or program committee chair/co-chair or PC member for around 30 major international conferences. Now, he is a Fellow of IET and Senior Member of IEEE.

Dacheng Tao (M'07–SM'12) is Professor of Computer Science with the Centre for Quantum Computation and Intelligent Systems and the Faculty of Engineering and Information Technology in the University of Technology, Sydney. He mainly applies statistics and mathematics for data analysis problems in data mining, computer vision, machine learning, multimedia, and video surveillance. He has authored and co-authored 100+ scientific articles at top venues including IEEE T-PAMI, T-NNLS, T-IP, NIPS, ICML, AISTATS, ICDM, CVPR, ICCV, ECCV; ACM T-KDD, Multimedia and KDD, with the best theory/algorithm paper runner up award in IEEE ICDM'07 and the best student paper award in IEEE ICDM'13.

Xuelong Li (M'02–SM'07–F'12) is a Full Professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.