Deep CNN-Based Blind Image Quality Predictor

Jongyoo Kim, *Member, IEEE*, Anh-Duc Nguyen, and Sanghoon Lee, *Senior Member, IEEE*

***Abstract*— Image recognition based on convolutional neural networks (CNNs) has recently been shown to deliver the state-of-the-art performance in various areas of computer vision and image processing. Nevertheless, applying a deep CNN to no-reference image quality assessment (NR-IQA) remains a challenging task due to critical obstacles, i.e., the lack of a training database. In this paper, we propose a CNN-based NR-IQA framework that can effectively solve this problem. The proposed method—deep image quality assessor (DIQA)—separates the training of NR-IQA into two stages: 1) an objective distortion part and 2) a human visual system-related part. In the first stage, the CNN learns to predict the objective error map, and then the model learns to predict subjective score in the second stage. To complement the inaccuracy of the objective error map prediction on the homogeneous region, we also propose a reliability map. Two simple handcrafted features were additionally employed to further enhance the accuracy. In addition, we propose a way to visualize perceptual error maps to analyze what was learned by the deep CNN model. In the experiments, the DIQA yielded the state-of-the-art accuracy on the various databases.**

***Index Terms*— Convolutional neural network (CNN), deep learning, image quality assessment (IQA), no-reference IQA (NR-IQA).**

I. Introduction

THE goal of image quality assessment (IQA) is to predict the perceptual quality of digital images in a quantitative manner. Digital images are likely to be inevitably degraded in the process from content generation to consumption. The acquisition, transmission, storage, postprocessing, or compression of images introduces various distortions, such as Gaussian white noise, Gaussian blur (GB), or blocking artifacts. A reliable IQA algorithm can help quantify the quality of images obtained blindly from the Internet and accurately assess the performance of image processing algorithms, such as image compression and super-resolution, from the perspective of a human observer. IQA is classified in general into three categories, depending on whether a reference image (the pristine version of an image) is available: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference IQA (NR-IQA). In general, the performance of these techniques, in order of decreasing accuracy, is FR-IQA, RR-IQA, and NR-IQA. However, since reference images are not accessible in a number of practical scenarios, NR-IQA is most appropriate as the most general method.

The bit rate of computer networks has continued to increase in recent years and has enabled the provision of high-quality entertainment to end users who do not have reference images; hence, significant research efforts have been made to enhance the accuracy of NR-IQA from the perspective of the end user. Many recently proposed NR-IQA algorithms involve the use of machine learning, such as support vector machines (SVMs) and neural networks (NN), to blindly predict image quality scores. Research has shown that the accuracy of NR-IQA depends heavily on designing elaborate features. Natural scene statistics (NSS) [1], [2] is one of the most successful features under the assumption that natural images have statistical regularity that is altered when distortions are introduced. Due to the difficulties involved in obtaining reliable features, research on NR-IQA has progressed significantly since NSS. Deep learning has lately been adopted in a few NR-IQA studies as a different method from conventional approaches based on NSS [3], [4]. However, most such studies have continued to use handcrafted features, and deep models, such as deep belief networks (DBNs) and stacked autoencoders, have been used in place of conventional regression machines.

A. Problems of Applying CNNs to NR-IQA

Convolutional neural networks (CNNs) form the most popular deep learning model nowadays due to their strong representation capability and impressive performance. CNNs have been successfully applied to various computer vision and image processing problems.

The performance of deep neural networks heavily depends on the number of training data. However, the currently available IQA databases are much smaller compared to the typical computer vision data set for deep learning. For example, the LIVE IQA database [5] contains 174–233 images for each distortion type, while the widely used data set for image recognition contain more than 1.2 million pieces of labeled data [6]. Moreover, obtaining large-scale reliable human subjective labels is very difficult. Unlike classification labels, constructing an IQA database requires a complex and time-consuming psychometric experiment. To expand the training data set, one can use data augmentation techniques such as rotation, cropping, and horizontal reflection. Unfortunately, any transformation of images would affect perceptual quality scores.

Moreover, the perceptual process of the human visual system (HVS) includes multiple complex processes, which makes

Manuscript received March 1, 2017; revised October 22, 2017 and March 23, 2018; accepted April 9, 2018. This work was supported by the National Research Foundation of Korea through the Korea Government (MSIT) under Grant 2016R1A2B2014525. *(Corresponding author: Sanghoon Lee.)*

The authors are with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, South Korea (e-mail: jongky@yonsei.ac.kr; slee@yonsei.ac.kr).

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.2018.2829819

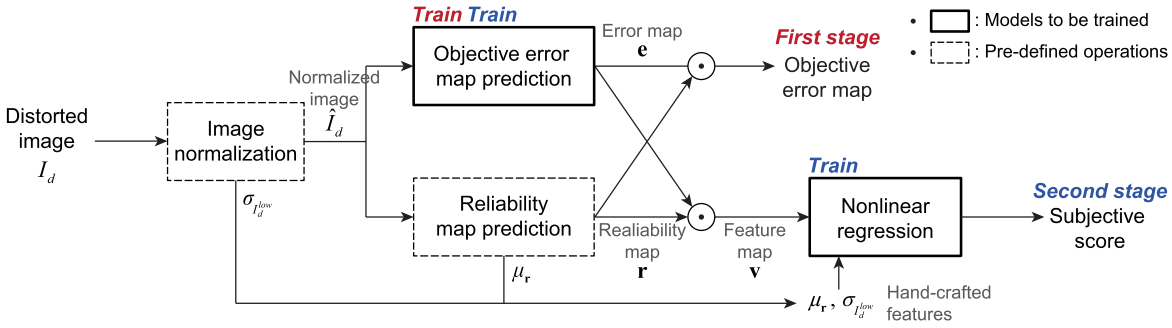


Fig. 1. Overall flowchart of DIQA. The training process consists of two stages: regression onto objective error maps and regression onto subjective scores. The squares with red "train" (blue "train") indicates that the subnetwork will be trained in the first (second) stage.

training of a deep model with limited data set even harder. For example, the visual sensitivity of the HVS varies according to spatial frequency of stimuli [7], [8], and the presence of texture hinders other spatially coincident image changes [9]. In addition, the perceived signals go through bandpass, multiscale, and directional decompositions in the visual cortex [10]. Such complex behaviors need to be embedded in the data set with human subjective labels. However, it is difficult to claim that a small data set can represent general visual stimuli, which results in an overfitting problem.

B. Proposed Framework

To tackle this problem, we propose a novel NR-IQA framework called deep blind image quality assessor (DIQA). The DIQA is trained in two separated stages as shown in Fig. 1. In the first stage, an objective error map is used as a proxy training target to expand the data set labels. The existing database provides a subjective score for each distorted image. In other words, one training data item includes a mapping from a 3-D tensor (width, height, and channel) to a scalar value. Given a distorted image I_d and a scalar subjective score S, the optimal parameter of a model θ should be sought by $\arg\min_\theta \|f(I_d;\theta)-S\|^2$, where $f(\cdot)$ is a prediction function. In contrast, the DIQA utilizes reference images during training and generates a 2-D intermediate target called the objective error map. Please note that the reference images are accessible during training as long as the database provides them, and the ground-truth objective error map can be easily derived by comparing the reference and distorted images. By expanding the training target to a 2-D error map $\mathbf{e}$, we have $\arg\min_\theta \sum_{(i,j)} \|f(I_d;\theta)(i,j) - \mathbf{e}(i,j)\|^2$, where (i,j) is a pixel index. In other words, it yields the same effect of increasing the number of training pairs up to the dimensions of the error map by giving more constraints. Once the deep neural network is trained with sufficient training data set, the model is fine tuned to predict the subjective scores. Since the objective error map is somewhat correlated with the subjective score, the second stage can be trained without great difficulty by using even a limited data set. In the end, our model can predict the subjective scores without accessing the ground-truth objective error maps during testing.

Overall, we resolve the NR-IQA problem by dividing it into the objective distortion and the HVS-related parts. In the objective distortion part, a pixelwise objective error map is predicted using the CNN model. In the HVS-related part, the model further learns the human visual perception behavior.

However, there persists another problem in the objective error map prediction phase. When severe distortion is applied to an image and its high-frequency detail is lost, its error map obtains more high-frequency components. Meanwhile the distorted image does not have high-frequency details. Therefore, without the reference image, it is difficult to predict an accurate error map from the distorted image, in particular, on homogeneous regions. To avoid this problem, we propose deriving a reliability map by measuring textural strength to compensate for the inaccuracy of the error map.

To visualize and analyze the learned human visual sensitivity, we further propose an alternative model, which we call DIQA-SENS. We use two separated CNN branches where each is dedicated to learn the objective distortion and the human visual sensitivity, respectively. In particular, the visual sensitivity branch predicts local visual weights of the objective error map by seeing the triplet of a distorted image, its objective error map, and its ground-truth subjective score. The multiplication of the objective error map and the sensitivity map results in a perceptual error map, which can explain the degree of distortion in the perspective of the HVS.

Our contributions can be summarized as follows.

1) Using the simple objective error map, the training data set can be easily augmented, and the deep CNN model can be trained without an overfitting problem.
2) DIQA is trained via end-to-end optimization so that the parameters can be thoroughly optimized to achieve state-of-the-art correlation with human subjective scores.
3) DIQA-SENS generates the objective error map and the perceptual error maps as intermediate results, which provide an intuitive analysis of local artifacts given distorted images.

The remainder of this paper is organized as follows. Section II introduces related work and relatively recent NR-IQA algorithms that use deep learning methods and pooling methods. Section III describes the DIQA framework, including the architecture of our model, reliability map prediction, and training methods in detail. Section IV provides comprehensive ablation studies of DIQA. In addition, a statistical evaluation of the proposed method is provided. Section V

introduces an alternative model of the DIQA to analyze the learned perceptual error maps. The visualization and analysis results of the trained deep model are presented. Conclusions and suggestions for future work are provided in Section VI.

II. Related Work

Most previously proposed NR-IQA methods were developed based on the machine learning framework. Researchers attempted to design elaborate features that could discriminate distorted images from the pristine images. One popular feature is a family of NSS that assumes that natural scenes contain statistical regularities. Various types of NSS features have been defined in transformation and spatial domains in the literature. Moorthy and Bovik [1] extracted features in the wavelet domain, and Saad *et al.* [2] defined them in the discrete cosine transform coefficients. Recently, Mittal *et al.* [11], [12] captured NSS features using only locally normalized images without any domain transformation.

In addition to NSS features, various kinds of features have been developed for NR-IQA. Li *et al.* [13] employed a general regression neural network relative to phase congruency, entropy, and image gradients. Tang *et al.* [14] considered such multiple features as natural image statistics, distortion textures, blur, and noise statistics. Meanwhile, in [15] and [16], dictionary learning was adapted to capture effective features from the raw patches. Most of these studies were based on conventional machine learning algorithms, such as SVMs and NNs. Since such models have a limited number of parameters, the size of the data set was not a significant issue. However, they yielded lower accuracies than FR-IQA metrics.

Relatively recently, attempts have been made to adopt a deep learning technique for the NR-IQA problem to enhance prediction accuracy [42]. Hou *et al.* [3] used a DBN, where NSS-related features were extracted in the wavelet domain and fed into the deep model. Similarly, Li *et al.* [4] derived NSS-related features from Shearlet-transformed images. The extracted features were then regressed onto a subjective score using a stacked autoencoder. Lv *et al.* [17] used DoG features and the stacked autoencoder. Ghadiyaram and Bovik [18] attempted to capture a large number of NSS features using multiple transforms and then used a DBN to predict the subjective score. However, most studies have used the deep model in place of the conventional regression machine. This involved designing handcrafted features of sufficiently small size such that the neural networks were not sufficiently deep to take full advantage of deep learning. Kang *et al.* [19] applied a CNN to the NR-IQA problem without handcrafted features to conduct end-to-end optimization. To resolve the data set size, an input image was divided into multiple patches, and an equal mean opinion score (MOS) was used for all patches in an image. Strictly speaking, this approach cannot reflect properties of the HVS, the pixelwise perceptual quality of which varies over the spatial domain. Bosse *et al.* [20] adopted a deep CNN model with 12 layers. The loss function was similar to [19]; however, they suggested an additional model, which learns the individual importance of each patch. Recently, we proposed a CNN-based NR-IQA framework, where FR-IQA metrics were employed as intermediate training targets of the CNN [21], and the statistical pooling over minibatch was introduced for end-to-end optimization. On the other hand, to overcome the limited training set, other attempts have been made by generating discriminable image pairs [22], or employing multitask learning [23].

In contrast to past work, the DIQA resolves the issue of the lack of a data set by utilizing reference images in training to generate an intermediate target. Different from our previous work [21], the DIQA does not depends on complicated FR-IQA metrics. In addition, the DIQA uses only convolutional layers in the pretraining stage so that the model can be deeper and can use a larger proxy target. Our proposed framework achieves state-of-the-art prediction accuracy using the strong representation capability of CNN models, which is discussed in Section IV.

III. Deep Image Quality Assessment Predictor

The overall framework of the DIQA is shown in Fig. 1. Once an input-distorted image is normalized (Section III-B), it passes through two paths: 1) a CNN branch and 2) a reliability map prediction branch (Section III-C). In the first training stage, the CNN branch is trained to predict an objective error map $\mathbf{e}$ (Section III-D). The ground-truth error map $\mathbf{e}_{\text{gt}}$ is obtained by comparing the reference and distorted images. In the second stage, the model is further trained to predict a human subjective score S (Section III-E). In each stage, the reliability map $\mathbf{r}$ is supplemented to compensate the inaccuracy on homogeneous regions.

A. Model Architecture

The design of the proposed CNN architecture is motivated by [24]. The structure of the DIQA is shown in Fig. 2. For the error map prediction part, the model consists of only convolutional layers and zeros are padded around the border before each convolution; therefore, the output does not lose relative pixel position information. Each layer except the last one has a 3×3 filter and a rectified linear unit (ReLU) [25]. We call the output of Conv8 as a feature map (filled with yellow in Fig. 2), which is reused for the second stage of training. In the last layer of the first training stage, the feature map is reduced to a one-channel objective error map using a 1×1 filter without nonlinear activation. If we directly feed the predicted error map into the modules of the second stage, it would hinder the abundant representation of features, because there is only one channel in the error map. To avoid this problem, we employ a simple linear combination over channels in Conv9, so that we can generate a meaningful feature map closely related to the ground-truth error map, meanwhile having multiple channels for better representation. The size of the output of Conv9 is $1/4$ times the original input image. Correspondingly, the ground-truth objective error maps are downscaled by $1/4$. For the downsampling operation, convolution with a stride of 2 is used. In the second training stage, the extracted feature map is fed into the global average pooling layer followed by two fully connected layers. We additionally use two handcrafted features, which will be

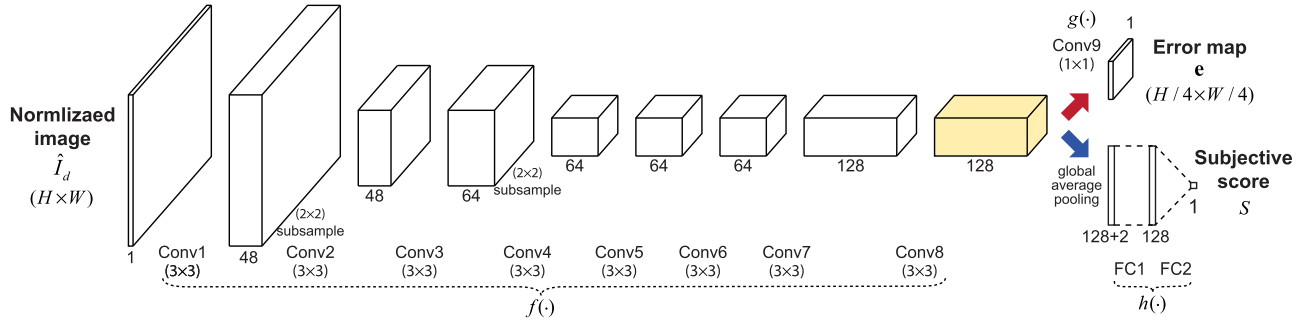


Fig. 2. Architecture of the objective pixel error map prediction subnetwork. "Conv" indicates the convolutional layers, and "FC" indicates fully connected layers. The text below "Conv" indicates its size of filter. The red (blue) arrows indicate the flows of the first (second) stage.

explained later. The handcrafted features are concatenated with the pooled features before FC1, and then regressed onto a subjective score. For convenience, we denote the procedure from Conv1 to Conv8 by $f(\cdot)$, the operation of Conv9 by $g(\cdot)$, and the procedure including FC1 and FC2 by $h(\cdot)$.

B. Image Normalization

As a preprocessing, the input images are first converted to grayscale, and they are subtracted from their low-pass filtered images. Let I_r be a reference image and I_d be the corresponding distorted image. The normalized versions are then denoted by $\hat{I}_r$ and $\hat{I}_d$, respectively. The low-frequency image is obtained by downscaling the input image to $1/4$ and upscaling it again to the original size, which is denoted by I_r^{low} and I_d^{low}. A Gaussian low-pass filter and subsampling were used to resize the images.

There are two reasons for this simple normalization. First, image distortions barely affect the low-frequency component of images. For example, white Gaussian noise (WN) adds random high-frequency components to images, GB removes high-frequency details, and blocking artifacts introduce new high-frequency edges. The distortions due to JPEG and JPEG2000 (JP2K) can be modeled by a combination of these artifacts [26]. Second, the HVS is not sensitive to a change in the low-frequency band. The CSF shows a bandpass filter shape peaking at approximately four cycles per degree, and sensitivity drops rapidly at low frequency [8]. Although there are small distortions in the low-frequency band, the HVS hardly notices them. Though there are benefits of employing this normalization scheme, there is also a drawback of losing information. To compensate this, two handcrafted features are supplemented in the second training stage.

C. Reliability Map Prediction

Many distortions, such as quantization by JP2K, or GB, make images blurry. However, unlike FR-IQA, it is difficult to determine whether the blurry region is distorted without knowing its pristine image. Furthermore, as severe distortion is applied to an image, its error map receives more high-frequency components. Meanwhile, the distorted image loses more high-frequency details, as shown in Fig. 3. Therefore, the model is likely to fail to predict the objective error map on homogeneous regions.

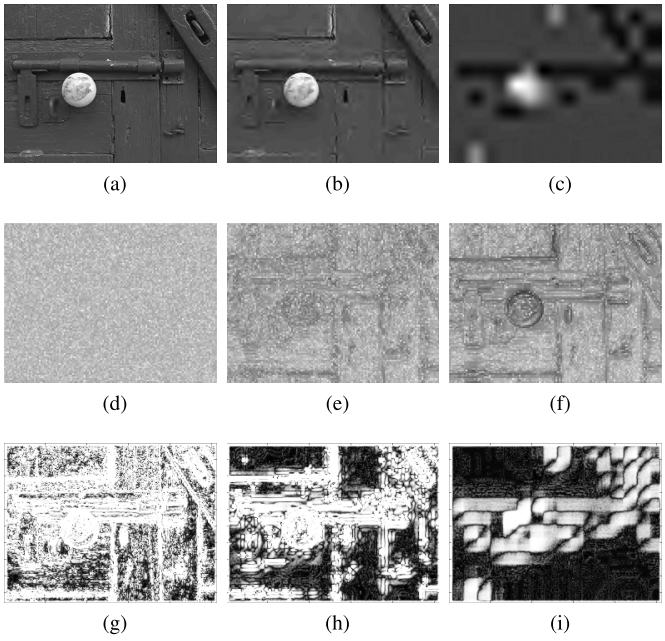


Fig. 3. Examples of estimated reliability maps. (a)–(c) JPEG2000 distorted images in the TID2013 data set at the distortion levels of 1, 3, and 5. (d)–(f) Difference maps derived by using (5). (g)–(i) Reliability maps of (a)–(c).

To avoid this problem, the reliability of the predicted error map is estimated by measuring the texture strength of the distorted image. Our assumption is that blurry regions have lower reliability than textured regions. Preprocessed images that are bandpassed are used to measure the reliability map as

$$\mathbf{r} = \frac{2}{1 + \exp(-\alpha(|\hat{I}_d|))} - 1 \tag{1}$$

where α controls the saturation property of the reliability map. To normalize the reliability map, the positive half of the sigmoid function is used in (1), so that pixels with small values are assigned sufficiently large reliability values.

The images shown in Fig. 3(a)–(c) are distorted by JPEG2000 at different levels, and the corresponding reliability maps with $\alpha = 1$ are shown in Fig. 3(d)–(f). It can be easily checked that it is difficult to derive an accurate error map (f) from severely distorted images (c). The estimated reliability maps are shown in Fig. 3(g)–(i). As shown in Fig. 3(i),

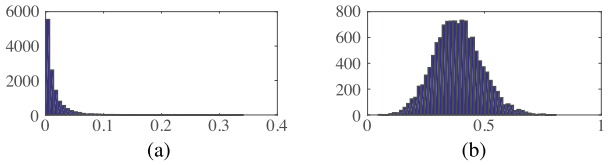


Fig. 4. Histograms of error maps with different values of p. (a) $p = 1$. (b) $p = 0.2$.

the reliability map has zero values, where there is no meaningful spatial information in Fig. 3(c).

To prevent the reliability map from directly affecting the predicted score, it is divided by its average as

$$\hat{\mathbf{r}} = \frac{1}{\frac{1}{H_{\mathbf{r}} \cdot W_{\mathbf{r}}} \sum_{(i,j)} \mathbf{r}(i, j)} \mathbf{r} \tag{2}$$

where $H_{\mathbf{r}}$ and $W_{\mathbf{r}}$ are the height and width of $\mathbf{r}$.

D. Learning Objective Error Map

In the first stage of training, the objective error maps are used as proxy regression targets to get the effect of increasing data. The loss function is defined by the mean squared error between the predicted and ground-truth error maps

$$\mathcal{L}_1(\hat{I}_d; \theta_f, \theta_g) = \| g(f(\hat{I}_d; \theta_f) - \mathbf{e}_{\text{gt}}; \theta_g) \odot \hat{\mathbf{r}} \|_2^2 \tag{3}$$

where $f(\cdot)$ and $g(\cdot)$ are defined in Fig. 2, θ represents the CNN's parameters, and $\mathbf{e}_{\text{gt}}$ is defined by

$$\mathbf{e}_{\text{gt}} = \text{err}(\hat{I}_r, \hat{I}_d). \tag{4}$$

Here, any error metric function can be used for err$(\cdot)$. In our experiment, we chose the exponent difference function

$$\mathbf{e}_{\text{gt}} = |\hat{I}_r - \hat{I}_d|^p \tag{5}$$

where p is the exponent number. When an absolute difference ($p = 1$) is used for the error metric function, most values in the error maps are small numbers close to zero. In this case, the model tends to fail to predict an accurate error map. When the training process converges, most values were zero in the experiment. Therefore, we chose $p = 0.2$ to spread the distribution of the difference map over the higher values. Fig. 4 shows a comparison of histograms for the two exponent numbers, where the histogram of $p = 0.2$ has a broader distribution between 0 and 1.

E. Learning Subjective Opinion

Once the model is trained to predict the objective error maps, we move to the next training stage, where DIQA is trained to predict subjective scores. To achieve this, the trained subnetwork $f(\cdot)$ is connected to a global average pooling layer followed by the fully connected layers as shown in Fig. 2. The feature map is averaged over spatial domain leading to a 128-D feature vector.

Here, to compensate the lost information, we consider two additional handcrafted features: the mean of the nonnormalized reliability map $\mu_{\mathbf{r}}$ and the standard deviation of the low-frequency of distorted image $\sigma_{I_d^{\text{low}}}$. If the distorted image is too blurred, the reliable area becomes too small. In this case, the overall textural strength of the distorted image becomes an important feature, which can be captured by $\mu_{\mathbf{r}}$. Therefore, the loss function is defined as

$$\mathcal{L}_2(I_d; \theta_f, \theta_h) = \| (h(\mathbf{v}, \mu_{\mathbf{r}}, \sigma_{I_d^{\text{low}}}; \theta_h) - S) \|_2^2 \tag{6}$$

where $f(\cdot)$ is a nonlinear regression function, S is the ground-truth subjective score of the input-distorted image, and $\mathbf{v}$ is the pooled feature vector. $\mathbf{v}$ is defined by:

$$\mathbf{v} = GAP(f(\hat{I}_d; \theta_f)) \tag{7}$$

where GAP indicates the global average pooling operation.

F. Training

In this section, we describe the training details of the DIQA. The layers for error map prediction are first trained by minimizing (3), where the ground-truth error map is derived from (5). When the first stage converges to a sufficient extent, (6) is then minimized in the second stage.

Since zeros are padded before each convolution, the feature maps near the borders tend to be zeros. Therefore, during the minimization of the loss functions in (3) and (6), we ignored pixels near borders around the error and the perceptual error maps. Each of four rows or columns for each border was excluded in the experiment, which compensated for information loss in the last two convolutional layers.

For better convergence of optimization, the adaptive moment estimation optimizer (ADAM) [27] with Nesterov momentum [28] was employed to alter the regular stochastic gradient descent method. The default hyperparameters suggested in the literature [27] were used for the ADAM, and the momentum parameter was set to 0.9. The learning rate was set differently for each data set from 2×10^{-4} to 5×10^{-4}. We chose the optimal value empirically. In addition, during training of the second stage, the learning rates for the pretrained layers were multiplied by 0.1. For weight decay, L_2 regularization was applied to all the layers (L_2 penalty multiplied by 5×10^{-4}).

G. Patch-Based Training

In the DIQA framework, the sizes of input images must be fixed to train the model on a GPU. Therefore, to train the DIQA using images of various sizes, such as in the LIVE IQA database [5], each input image should be divided into multiple patches of the same size. Here, the step of the sliding window is determined by the patch size and the number of ignored pixels around the borders to avoid overlapping regions when the perceptual error map is reconstructed. When the ignored pixels around the borders are four, the step should be 4, where $\text{step}_{\text{patch}} = \text{size}_{\text{patch}} - 32$ is determined by 4×2 (both sides of the border) $\times 4$ (upscaling by 4). In the experiment with the LIVE IQA database, the patch size was 112×112 and each step was 80×80.

In addition, during the training of the second stage, all patches composing an image should be in the same mini-batch [21], so that $\mathbf{v}$, $\mu_{\mathbf{r}}$, and $\sigma_{I_d^{\text{low}}}$ can be derived from the reconstructed perceptual error and reliability maps.

TABLE I

COMPARISON OF IQA DATABASES IN TERMS OF NUMBERS OF REFERENCE (REF.) IMAGES, DISTORTED (DIST.) IMAGES, DISTORTION TYPES, AND TYPE OF SUBJECTIVE SCORES

Database	Ref.	Dist.	Dist. Types	Score Type
LIVE IQA [5]	29	779	5	DMOS
CSIQ [29]	30	866	6	DMOS
TID2013 [30]	25	3,000	24	MOS
LIVE MD [31]	15	405	2	DMOS
LIVE challenge [32]	N/A	1,162	N/A	MOS
WED [33]	4,744	94,880	4	N/A

IV. EXPERIMENTS AND ANALYSIS

A. Database

Six different IQA databases were used to evaluate the proposed algorithm: LIVE IQA database [5], CSIQ [29], TID2013 [30], LIVE multiply distorted (LIVE MD) database [31], LIVE in the wild challenge (LIVE challenge) database [32], and Waterloo exploration database (WED) [33]. The summary of each database is tabulated in Table I. The LIVE IQA database contains five distortion types: JP2K compression, WN, GB, and Rayleigh fast-fading (FF) channel distortion. The CSIQ database includes six distortion types: JPEG, JP2K, WN, GB, pink Gaussian noise (PGN), and global contrast decrements [contrast distortion (CTD)]. TID2013 contains the largest number of distortion types at five levels of degradation. The LIVE MD database includes images distorted by two multiple types of distortion. One is associated with images corrupted by GB followed by JPEG (GB+JPEG) and the other one is associated with images corrupted by GB followed by WN (GB+WN). The LIVE in the wild challenge database covers the widest variety of contents. In contrast to the other ones, each distortion is caused by a regular capturing process using mobile cameras, which includes low-light blur and noise, motion blur, overexposure, underexposure, compression errors and their combination. Therefore, there are no reference images in the LIVE challenge database. Recently, the Waterloo exploration database [33] was built in response to the need for large-scale databases. Four synthetic distortions were used: JPEG, JP2K, GB, and WN. WED does not include any subjective quality scores. Instead, new evaluation strategies that do not require human opinions are proposed.

Regarding the subjective scores, in the case of differential MOS (DMOS), the lower values indicate higher perceptual quality, whereas the higher values indicate higher visual quality in MOS. All the subjective scores were rescaled to [0, 1], and DMOSs were reversed to match with MOSs.

B. Evaluation Metrics

To evaluate the performances of the IQA algorithms, we used two standard measures, i.e., Spearman's rank-order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) by following [34]. First, the SRCC is defined by

$$\text{SRCC} = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \tag{8}$$

where d_i is the difference between the predicted score and ground-truth score of the ith image, and n is the number of images. In addition, the PLCC can be derived by

$$\text{PLCC} = \frac{\sum_i (\hat{S}_i - \mu_{\hat{S}})(S_i - \mu_S)}{\sqrt{\sum_i (\hat{S}_i - \mu_{\hat{S}})^2}\sqrt{\sum_i (S_i - \mu_S)^2}} \tag{9}$$

where $\hat{S}_i$ and S_i are the predicted and ground-truth subjective scores of the ith image, and $\mu_{\hat{S}}$ and μ_S indicate the average of each.

In addition, we use three evaluation criteria: the pristine/distorted image discriminability test (D-test), the listwise ranking consistency test (L-test), and the pairwise preference consistency test (P-test) [33]. The D-test examines whether an IQA algorithm is able to discriminate the pristine images from distorted ones. The L-test evaluates if an IQA model consistently ranks images whose distortion type and content are same, but distortion level varies. The P-test checks the preference concordance of an IQA measure on quality-discriminable image pairs.

TABLE II

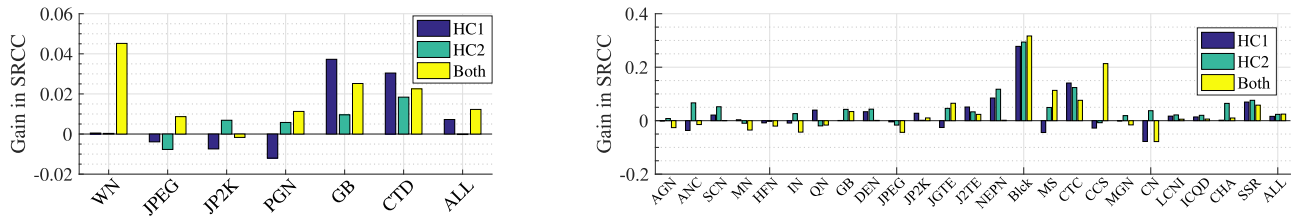
SRCC AND PLCC COMPARISON FOR DIFFERENT EPOCHS IN THE FIRST STAGE ON THE LIVE IQA AND CSIQ DATABASES

Configuration	Epochs	LIVE IQA		CSIQ	
		SRCC	PLCC	SRCC	PLCC
DIQA-BASE	0	0.963	0.964	0.812	0.791
DIQA	20	0.975	0.977	0.868	0.903
	40	0.972	0.974	0.871	0.910
	60	0.964	0.963	0.884	0.915

C. Ablation Studies

To investigate the contribution of each module and training scheme, we utilized different combinations of them. In other words, each ablated model needs to be repeatedly trained and tested while dividing training and testing sets randomly, which leads to a huge training time of deep CNN models. Thus, learning and testing were conducted on LIVE IQA and CSIQ, which are reasonable with respect to the data size.

1) Pretraining With Objective Error Map: We first studied the effects of training epochs in the first stage. In Table II, the models trained with different numbers of epochs in the first stage are compared. We tested the pretrained model with four different epochs and conducted the second stage to obtain the SRCC and PLCC on LIVE IQA and CSIQ, respectively. For all the epochs, the same neural network structure was used. The model named DIQA-BASE was directly trained to predict the subjective score with the same learning rate for all layers. Generally, the pretrained models with any epochs outperformed the DIQA-BASE. The best epochs for the first stage were different for each database. On LIVE IQA, pretraining with 20 epochs showed the best results, while 60 epochs do for CSIQ. For TID2013 and LIVE MD, we found that 80 and 100 epochs yielded the best performances, respectively.



(a) (b)

Fig. 5. Gains in SRCC by using handcrafted features on the CSIQ and TID2013 databases. HC1: using only $\sigma_{I_d^{\text{low}}}$. HC2: using only $\mu_{\mathbf{r}}$. Both: using both the features.

TABLE III
SRCC AND PLCC COMPARISON WITH AND WITHOUT RELIABILITY MAPS ON THE LIVE IQA AND CSIQ DATABASES

Reliability map	α	LIVE IQA		CSIQ	
		SRCC	PLCC	SRCC	PLCC
w/o	N/A	0.955	0.957	0.845	0.860
w/	$\alpha = 0.6$	0.973	0.969	0.844	0.852
	$\alpha = 1.0$	0.975	0.977	0.884	0.915
	$\alpha = 1.4$	0.973	0.975	0.882	0.916

TABLE IV
SRCC AND PLCC COMPARISON WITH AND WITHOUT IMAGE NORMALIZATION ON THE LIVE IQA AND CSIQ DATABASES

Input normalization	LIVE IQA		CSIQ	
	SRCC	PLCC	SRCC	PLCC
w/o	0.965	0.966	0.829	0.859
w/	0.975	0.977	0.884	0.915

2) Reliability Map: To test the contribution of the reliability map, the results of different settings are compared in Table III. In most of the cases, there was a performance gain when the reliability map is used, which confirmed the importance of the reliability map. We further analyzed the effect of hyperparameter of the reliability maps. As α of the reliability map increases, the reliable region increases. In other words, with larger α, the model considers more homogeneous pixels. To examine the effect of α, we compared three different values ($\alpha = 0.6$, 1.0, and 1.4). In general, there was no big difference between the three settings, and $\alpha = 1.0$ showed nice accuracies in common on all the databases.

3) Image Normalization: The experimental results with and without image normalization on the LIVE IQA and CSIQ databases are shown in Table IV. Using input data normalization significantly increased both the SRCC and PLCC. In particular, there was a large gap on the CSIQ database.

4) Handcrafted Features: We further investigated the effects of the adopted handcrafted features, $\sigma_{I_d^{\text{low}}}$ and $\mu_{\mathbf{r}}$, on the performance. Since they were designed to capture specific distortion statistics which could not be detected due to the normalization and reliability map processing, we investigated their effects on each individual distortion type in CSIQ and TID2013. We omitted the results on the LIVE IQA and LIVE MD databases, since there were almost no gains. Fig. 5 shows the gains in the SRCC according to each individual distortion type when using the handcrafted features. In general, using both handcrafted features led to a performance gain when all the distortion types were considered on the both data sets (0.0123 on CSIQ and 0.0246 on TID2013). On the CSIQ database, using $\sigma_{I_d^{\text{low}}}$ resulted in a large gain on CTD, and using both also yielded significant gains on WN, GB, and CTD. For the TID2013 database, there was a large gain on block (local blockwise distortions) and contrast change (CTC) by using either handcrafted feature. In particular, using both features resulted in gains on change of color saturation (CCS) and mean shift (MS). Unfortunately, the achieved SRCC on the CCS distortion was still a negative value.

TABLE V
SRCC AND PLCC COMPARISON OF DIFFERENT PROXY TARGETS IN THE FIRST STAGE ON THE LIVE IQA AND CSIQ DATABASES

Proxy target	LIVE IQA		CSIQ	
	SRCC	PLCC	SRCC	PLCC
Error map	0.975	0.977	0.884	0.915
SSIM	0.960	0.973	0.856	0.899
FSIM	0.912	0.930	0.706	0.752

In Fig. 5, using the handcrafted features sometimes lead to a negative effect. One possible reason of this phenomenon is that DIQA was trained to minimize the summed errors of all the distortion types. As a result, when the model is adapted to a specific type, it could be less accurate on the other types. In addition, the objective function of a deep CNN has lots of local minima and saddle points. Since massive number of parameters are initialized randomly, the final optimal point in the parameter space can be very different, which may lead to inconsistent results on some distortion types.

5) Proxy Training Targets: We tested the alternative training targets in the first stage. As proposed in [21], SSIM [35] and FSIM [36] were used to generate the proxy targets. To train the models, we used the same hyper parameters. The experimental results on the LIVE IQA and CSIQ databases are compared in Table V. Interestingly, the model using the objective error map ranked first and was followed by using SSIM and FSIM. One possible reason is due to usage of insufficiently optimized hyper parameters of using the SSIM and FSIM. Another reason would be the size of the proxy targets. Compared with [21], the DIQA uses eight times larger proxy targets by using the fully convolutional layers. Since FSIM highlights edges especially, using an over detailed FSIM map can cause a serious overfitting problem.

D. Effect of Architecture Depth

To examine the effect of architecture depth, we tested six different numbers of convolutional layers of DIQA (from 5 to 10). In the shortest setting, the layers

TABLE VI
SRCC AND PLCC COMPARISON ON THE FIVE DATABASES. ITALICS INDICATE DEEP LEARNING-BASED METHODS

Type	Methods	LIVE IQA		CSIQ		TID2013		LIVE MD		LIVE challenge		Weighted Average	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR	PSNR	0.876	0.872	0.806	0.800	0.636	0.706	0.725	0.815	–	–	0.709	0.756
	SSIM	0.948	0.945	0.876	0.861	0.775	0.691	0.845	0.882	–	–	0.825	0.775
	FSIMc	0.963	0.960	0.931	0.919	0.851	0.877	0.863	0.818	–	–	0.883	0.892
	DeepQA	0.981	0.982	0.961	0.956	0.939	0.947	0.938	0.942	–	–	0.949	0.954
NR	BLIINDSII	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	0.746
	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
	GMLOG	0.950	0.954	0.803	0.812	0.675	0.683	0.824	0.863	0.543	0.571	0.751	0.761
	HOSA	0.948	0.949	0.781	**0.841**	0.688	0.764	0.902	0.926	**0.659**	**0.678**	0.761	0.819
	NRSL	0.952	0.956	**0.851**	**–**	0.661	–	**0.932**	**0.946**	0.631	0.654	0.760	–
	SESANIA	0.934	0.948	–	–	–	–	–	–	–	–	–	–
	CNN	0.956	0.953	–	–	–	–	–	–	–	–	–	–
	MGDNN	0.951	0.949	–	–	–	–	–	–	–	–	–	–
	deepIQA	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	**0.790**	**0.821**
	DIQA–BASE	**0.963**	**0.964**	0.812	0.791	**0.800**	**0.803**	0.910	**0.934**	**0.663**	**0.705**	**0.836**	**0.836**
	DIQA	**0.975**	**0.977**	**0.884**	**0.915**	**0.825**	**0.850**	**0.939**	**0.942**	**0.703**	**0.704**	**0.867**	**0.888**

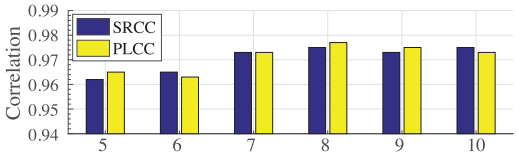


Fig. 6. Comparison of SRCC and PLCC according to model depth.

Conv1–Conv4 and Conv8 were used. In the longest setting, two 3×3 convolutional layers with 64 filters were added after the Conv6 layer. Fig. 6 shows the accuracy comparison between the six models on the LIVE IQA database. When the depth was five, the SRCC and PLCC were the lowest. However, as the depth increased, the correlations scores almost saturated around 0.97. It can be concluded that the eight convolutional layers are rational for our framework.

E. Benchmark

1) Performance on Individual Databases: To train and test the DIQA, we randomly divided the reference images into two subsets, 80% for training and 20% for testing. Then, the corresponding distorted images were divided into training and testing sets so that there was no overlap between the two. Two training stages shared the same training and testing sets. To increase the number of training data items, horizontally flipped images were supplemented. In the second stage, an early stopping method was used to avoid overfitting. Following the recommendation in [34], we evaluated the performance of the IQA algorithms using two standard measures: SRCC and PLCC.

We compared DIQA with four FR-IQA methods (PSNR, SSIM [35], FSIMc [36], and DeepQA [37]) and seven NR-IQA methods (BLIINDS II [2], BRISQUE [11], CORNIA [15], ILNIQE [38], GMLOG [39], HOSA [16], and NRSL [40]). In addition, five deep learning-based NR-IQA methods were benchmarked: SESANIA [4], CNN [19], MGDNN [17], deepIQA [20], and BIECON [21]. For the D-, L-, and P-tests, dipIQ [22] and MEON [23] are additionally compared with.

First, we evaluated all the methods on each individual database as shown in Table VI. Since there is no reference image in the LIVE challenge database, the FR-IQA methods were not tested on this. To train the DIQA on LIVE challenge, the model was pretrained on TID2013, and then the second training stage was conducted on LIVE challenge. The correlation coefficients of the DIQA were averaged after the procedure was repeated 20 times while dividing the training and testing sets randomly in order to eliminate performance bias. The best three models among the NR-IQA methods for each evaluation criterion are shown in bold. The weighted average of the SRCC and PLCC over the four databases (LIVE, CSIQ, TID2013, and LIVE MD) is also shown in the last column. The weight to each database is proportional to the number of distorted images in that database.

Among NR-IQA methods, deep learning-based methods were generally superior to the previous methods. In particular, the DIQA achieved the highest correlation on most databases as shown in the weighted average column. Regarding the LIVE MD database, it contains only 15 reference images, which would hamper the training of the DIQA and DIQA-BASE.

2) Performance on Individual Distortion Types: In Table VII, the SRCC and PLCC of the FR- and NR-IQA algorithms are compared according to the individual distortion type. The DIQA models were trained with all the distortion types from the training set (80%), and tested on the individual distortion type from the testing set (20%). The best three models among the NR-IQA methods are shown in bold.

Even when each distortion type was tested separately, the DIQA was the best model on most distortion types. The DIQA in particular showed high accuracy of prediction in white noise. For FF in the LIVE IQA database, it outperformed the other NR-IQA methods by a wide margin. Most previous methods failed on several distortion types in TID2013, such as JGTE, NEPN, Block, MS, CTF, and CN. However, the DIQA achieved remarkably enhanced accuracies on those distortion types except for CCS. Since the DIQA methods used grayscale

TABLE VII
SRCC COMPARISON ON INDIVIDUAL DISTORTION TYPES. ITALICS INDICATE DEEP LEARNING-BASED METHODS

	Dist.type	PSNR	SSIM	FSIMc	*DeepQA*	BLIINDS2	BRISQUE	CORNIA	IL-NIQE	GMLOG	NRSL	*DIQA-BASE*	*DIQA*
LIVE IQA	JP2K	0.895	0.961	0.972	0.972	**0.930**	0.914	0.921	0.902	0.926	0.903	**0.935**	**0.961**
	JPEG	0.881	0.972	0.979	0.980	0.950	**0.965**	0.938	0.931	0.963	0.891	**0.963**	**0.976**
	WN	0.985	0.969	0.971	0.986	0.947	0.977	0.957	0.975	0.983	**0.984**	**0.987**	**0.988**
	GB	0.782	0.952	0.968	0.982	0.915	0.951	**0.957**	0.911	0.929	0.808	**0.979**	**0.962**
	FF	0.891	0.956	0.950	0.963	0.874	0.877	**0.906**	0.827	0.899	0.895	**0.931**	**0.912**
CSIQ	WN	0.963	0.897	0.936	0.904	0.702	0.682	0.763	**0.850**	0.804	**0.810**	0.766	**0.835**
	JPEG	0.888	0.956	0.966	0.948	0.846	0.846	0.842	0.899	0.864	**0.903**	**0.931**	**0.931**
	JP2K	0.936	0.961	0.970	0.972	0.850	0.817	0.869	0.906	0.890	**0.912**	**0.908**	**0.927**
	PGN	0.934	0.892	0.937	0.930	0.812	0.743	0.567	**0.874**	0.774	0.836	**0.843**	**0.893**
	GB	0.929	0.961	0.973	0.970	**0.880**	0.808	0.854	0.858	0.857	**0.896**	0.827	**0.870**
	CTD	0.862	0.792	0.944	0.956	0.336	0.396	0.533	0.501	0.562	**0.659**	**0.614**	**0.718**
TID2013	AGN	0.934	0.867	0.910	0.920	0.714	0.706	0.550	**0.890**	0.748	**0.813**	0.730	**0.915**
	ANC	0.867	0.773	0.854	0.816	**0.728**	0.523	0.209	**0.823**	0.591	0.457	0.524	**0.755**
	SCN	0.916	0.852	0.890	0.884	0.825	0.776	0.717	**0.929**	0.769	**0.867**	0.784	**0.878**
	MN	0.836	0.777	0.801	0.839	0.358	0.295	0.360	**0.649**	0.491	0.393	**0.520**	**0.734**
	HFN	0.913	0.863	0.904	0.935	0.852	0.836	0.797	**0.881**	0.875	**0.902**	0.842	**0.939**
	IN	0.900	0.750	0.825	0.835	0.664	**0.802**	0.585	**0.802**	0.693	0.787	**0.836**	**0.843**
	QN	0.875	0.866	0.880	0.865	0.780	0.682	0.727	**0.881**	**0.833**	0.700	0.781	**0.858**
	GB	0.910	0.967	0.955	0.969	0.852	0.861	0.840	0.845	**0.878**	**0.886**	0.858	**0.920**
	DEN	0.953	0.925	0.933	0.899	0.754	0.500	0.721	**0.778**	0.721	**0.795**	0.774	**0.788**
	JPEG	0.922	0.920	0.934	0.896	0.808	0.790	0.806	**0.875**	**0.823**	0.818	0.782	**0.892**
	JP2K	0.886	0.947	0.959	0.919	0.862	0.779	0.800	**0.911**	0.872	0.891	**0.901**	**0.912**
	JGTE	0.806	0.845	0.861	0.830	0.251	0.254	**0.595**	0.310	0.400	0.345	**0.582**	**0.861**
	J2TE	0.891	0.883	0.912	0.908	0.755	0.723	0.654	0.627	0.731	**0.805**	**0.803**	**0.812**
	NEPN	0.679	0.782	0.794	0.661	0.081	**0.213**	0.157	-0.117	0.190	0.117	**0.449**	**0.659**
	Block	0.330	0.572	0.553	0.324	**0.371**	0.197	0.016	-0.051	0.318	**0.323**	0.308	**0.407**
	MS	0.757	0.775	0.749	0.596	0.159	0.217	0.177	**0.222**	0.119	0.136	**0.272**	**0.299**
	CTC	0.447	0.378	0.468	0.671	-0.082	0.079	**0.262**	0.026	0.224	0.194	**0.532**	**0.687**
	CCS	0.634	0.414	0.836	0.351	**0.109**	**0.113**	**0.170**	-0.101	-0.121	-0.110	-0.335	-0.151
	MGN	0.883	0.780	0.857	0.939	0.699	0.674	0.407	**0.736**	0.701	**0.753**	0.676	**0.904**
	CN	0.841	0.857	0.914	0.885	0.222	0.198	**0.541**	0.388	0.202	0.434	**0.539**	**0.655**
	LCNI	0.916	0.806	0.949	0.934	0.451	0.627	0.696	**0.869**	0.664	0.751	**0.802**	**0.930**
	ICQD	0.920	0.854	0.882	0.893	0.815	0.849	0.649	0.793	**0.886**	0.866	**0.895**	**0.936**
	CHA	0.880	0.878	0.893	0.859	0.568	0.724	0.689	**0.789**	0.648	0.694	**0.773**	**0.756**
	SSR	0.911	0.946	0.958	0.920	0.856	0.811	0.874	**0.893**	**0.915**	0.887	0.868	**0.909**
LIVE MD	GB+JPEG	0.736	0.898	0.885	0.956	0.899	0.903	0.900	**0.911**	0.824	**0.928**	**0.929**	0.896
	GB+WN	0.743	0.912	0.899	0.952	0.892	0.902	0.899	0.924	0.863	**0.937**	**0.941**	**0.941**

TABLE VIII
D-TEST, L-TEST, AND P-TEST RESULTS ON WED

Type	Methods	D-test	L-test	P-test
FR	PSNR	1.0000	1.0000	0.9995
	SSIM	1.0000	0.9992	0.9991
NR	BRISQUE	0.9204	0.9772	0.9930
	CORNIA	0.9290	0.9764	0.9947
	ILNIQE	0.9084	**0.9926**	0.9927
	HOSA	0.9175	0.9647	**0.9983**
	dipIQ	**0.9346**	**0.9846**	**0.9999**
	MEON	**0.9384**	0.9669	**0.9984**
	deepIQA	0.9074	0.9467	0.9628
	DIQA (CSIQ)	0.8982	0.9594	0.9431
	DIQA (LIVE IQA)	**0.9530**	**0.9790**	0.9970

images, they failed on CCS where color information is the major cue of distortion.

3) Performance on WED: To evaluate the DIQA on WED, we used two models trained on the full sets of the LIVE IQA and CSIQ databases, respectively. For both the models, we used only four common distortion types (white noise, GB, JPEG, and JPEG2000). In Table VIII, the results of the D-test, L-test, and P-test on WED are reported. The DIQA models trained on LIVE IQA and CSIQ are named DIQA (LIVE IQA) and DIQA (CSIQ), respectively. The three best models among the NR-IQA methods are shown in bold. The DIQA (LIVE IQA) outperformed all previous NR-IQA models in the D-test, which indicates that our model figure out the existence of distortions well. In the L-test and P-test, dipIQ yielded the best, since it was trained on a large number of natural images. It also achieved competitive scores in both the L-test and P-test. However, the DIQA (CSIQ) was not competitive in the three tests on WED. Comparing with the CNN-based models, the DIQA and MEON showed the similar performances in the P-test, while the scores of the DIQA were the best in the D-test and L-test.

F. Cross Data Set Test

To evaluate the generalizability of the DIQA, the model was trained using all the images from one database, and then tested on another database. In the CSIQ and TID2013 databases, four overlapping distortion types (white noise, GB, JPEG, and JPEG2000) were used. The results of the cross-data set test are shown in Table IX. For each test, the best two models are shown in bold. Even though the DIQA can easily be over adapted to the specific data set due to a large number of parameters, the DIQA generally yielded a competitive SRCC. In most tests, the DIQA was one of the best two models. It can be concluded that the DIQA performs well in terms of subjective score prediction, and its performance does not depend on the database.

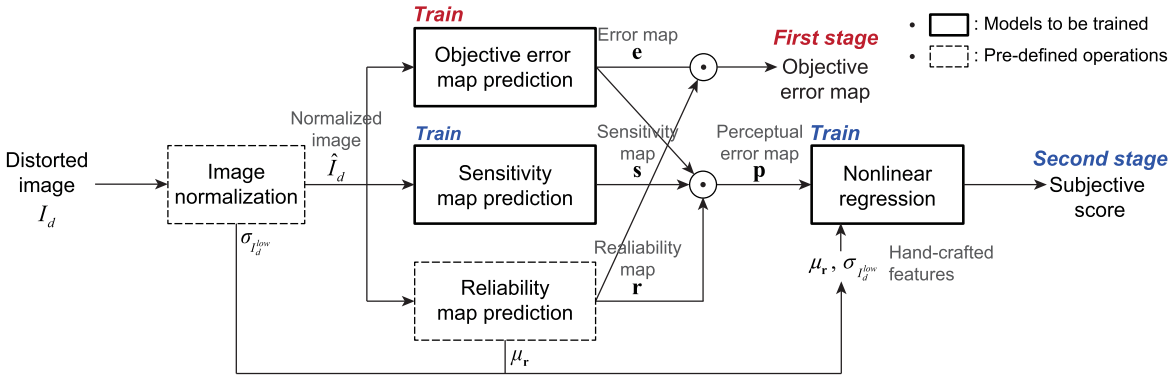


Fig. 7. Overall structure of DIQA-SENS. An input distorted image is normalized and goes through three paths: 1) objective error map prediction; 2) sensitivity map prediction; and 3) reliability map prediction. In the first stage, the subnetwork of objective error map prediction is trained, and then the subnetworks of sensitivity map prediction and nonlinear regression are trained in the second stage.

TABLE IX

SRCC Comparison of the Cross Data Set Test

Train	Test	BLIINDS2	BRISQUE	IL-NIQE	CORNIA	GMLOG	NRSL	*DIQA-BASE*	*DIQA*
LIVE IQA	CSIQ subset	0.901	0.890	0.880	0.898	0.897	**0.917**	0.906	**0.915**
	TID2013 subset	0.855	0.878	0.877	0.879	0.907	0.916	**0.918**	**0.922**
CSIQ subset	LIVE IQA	0.894	0.919	0.913	0.920	0.903	0.921	**0.923**	**0.926**
	TID2013 subset	0.765	0.874	**0.945**	0.852	0.879	0.921	0.915	**0.923**
TID2013 subset	LIVE IQA	0.894	0.877	**0.906**	**0.907**	0.889	0.896	0.905	0.904
	CSIQsubset	0.864	0.861	0.861	0.859	0.794	**0.875**	0.871	**0.877**

V. Learning to Visualize Perceptual Error Map

To study and analyze what was learned by a deep model, we additionally propose a variant version of the DIQA, named DIQA-SENS. Different from the normal model, the DIQA-SENS contains three paths: 1) objective error map prediction, 2) sensitivity map prediction, and 3) reliability map prediction. The overview of DIQA-SENS is shown in Fig. 7. The same architecture of the DIQA (from Conv1 to Conv9 in Fig. 2) is used for the objective error map prediction and the sensitivity map prediction subnetworks, respectively. The hidden layer of the fully connected layer has 20 perceptrons. Its training scheme is similar to that of the DIQA. However, the objective error map prediction subnetwork is frozen while the sensitivity map prediction subnetwork is updated. In addition, the perceptual error map is obtained by multiplying the predicted error and sensitivity maps, and then directly regressed onto the subjective score.

A. Learning Visual Weight of Error Map

Because there is no ground-truth sensitivity map available, the model cannot be trained to directly minimize the pixelwise difference. Instead, we show a triplet of a distorted image, its objective error map, and its corresponding ground-truth subjective score to the model. Then, the model seeks the optimal weights of the pixels in the error map such that the predicted score approaches the subjective score. The visual sensitivity map $\mathbf{s}$ is first derived from the CNN model. The perceptual error map $\mathbf{p}$ is then defined by

$$\mathbf{p} = \mathbf{s} \odot \mathbf{e} \odot \hat{\mathbf{r}} \tag{10}$$

$$\mathbf{s} = f_{\text{sens}}(\hat{I}_d; \theta_{\text{sens}}) \tag{11}$$

where $\odot$ is the Hadamard product, $f_{\text{sens}}(\cdot)$ indicates the second subnetwork with parameter θ_2, and $\mathbf{e}$ is the predicted error map, $\mathbf{e} = f_{\text{err}}(\hat{I}_d; \theta_{\text{err}})$, which is obtained from the error map prediction subnetwork $f_{\text{err}}(\cdot)$.

For the subjective score regression, the average of $\mathbf{p}$ and two handcrafted features $\mu_{\mathbf{r}}$ and $\sigma_{I_d^{\text{low}}}$ are used as follows:

$$\mathcal{L}_{\text{SENS}}(I_d; \theta_{f_{\text{sens}}}, \theta_{h'}) = \|(h'(\mu_{\mathbf{p}}, \mu_{\mathbf{r}}, \sigma_{I_d^{\text{low}}}; \theta_{h'}) - S)\|_2^2 \tag{12}$$

where $h'(\cdot)$ is a regression function of the DIQA-SENS.

When the model is optimized to minimize (12) without any constraints, it generates a noisy sensitivity map, which is not desirable. Therefore, we apply a smoothing constraint to penalize the high frequency in the sensitivity map using the total variation (TV) L_2 norm.

$$\text{TV}(\mathbf{s}) = \frac{1}{H_{\mathbf{s}} \cdot W_{\mathbf{s}}} \sum_{(i,j)} (\mathbf{s}_{\text{horz}}(i,j)^2 + \mathbf{s}_{\text{vert}}(i,j)^2) \tag{13}$$

where $H_{\mathbf{s}}$ and $W_{\mathbf{s}}$ indicate the height and width of $\mathbf{s}$, and $\mathbf{s}_{\text{horz}}$ and $\mathbf{s}_{\text{vert}}$ are Sobel-filtered sensitivity maps in the horizontal and vertical directions, respectively.

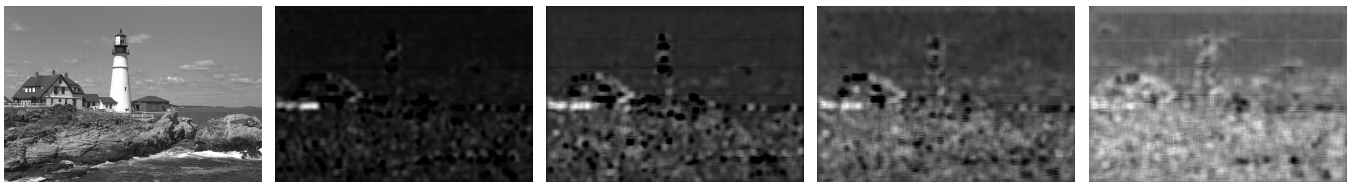
The final loss function of the DIQA-SENS is

$$\mathcal{L}_{\text{DIQA-SENS}}(\hat{I}_d; \theta_1, \theta_2) = w_s \mathcal{L}_s + w_{\text{TV}} \text{TV} \tag{14}$$

where w_s and w_{TV} are weights of two losses. In the experiment, we set $w_s = 10$ and $w_{\text{TV}} = 10^{-4}$.

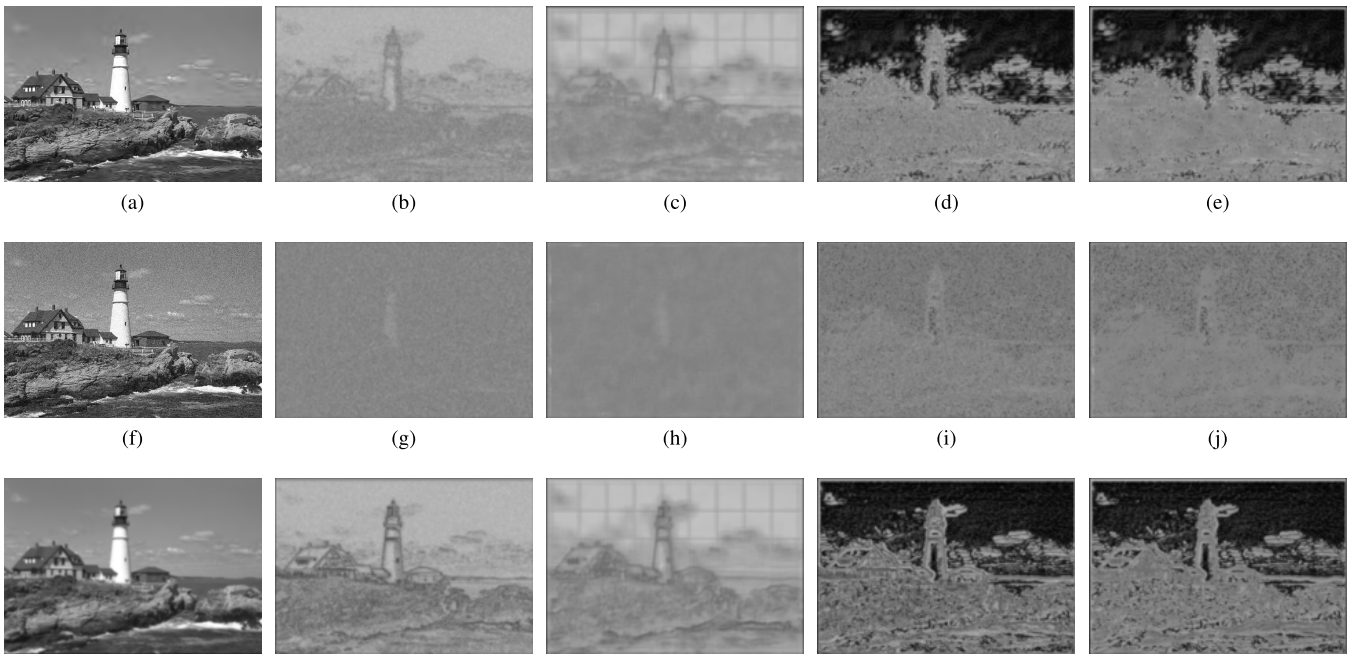
B. Experimental Results

The experimental results of the DIQA-SENS are compared with the DIQA and DIQA-BASE in Table X. The SRCC and PLCC of the DIQA-SENS are marginally lower than the DIQA and higher than the DIQA-BASE. The decreased performance



(a) (b) (c) (d) (e)

Fig. 8. Examples of predicted sensitivity maps with various TV regularization weights. (a) Distorted images. (b)–(e) Predicted sensitivity maps with different TV regularization weights.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

Fig. 9. Examples of predicted error maps. (a), (f), and (k) Distorted images with JPEG2000, white noise, and GB. (b), (g), and (l) Ground-truth error maps. (c), (h), and (r) Predicted error maps. (d), (i), and (n) Ground-truth error maps multiplied by reliability maps. (e), (j), and (t) Predicted error maps multiplied by reliability maps.

TABLE X

SRCC AND PLCC COMPARISON OF DIFFERENT COMBINATIONS OF FEATURE MAPS ON THE LIVE IQA AND CSIQ DATABASES

	LIVE IQA		CSIQ	
Configuration	SRCC	PLCC	SRCC	PLCC
DIQA-BASE	0.955	0.957	0.845	0.860
DIQA	0.975	0.977	0.884	0.915
DIQA-SENS	0.970	0.972	0.872	0.874

might be caused by one-channel feature map, which is fed into the global average pooling layer. However, the DIQA-SENS can generate the perceptual error maps, which can explain the local importance of the objective error map.

To investigate the effects of TV regularization on prediction accuracy, five weights ($w_{\mathrm{TV}} = 0$, 10^{-4}, 10^{-3}, 10^{-2}, and 10^{-1}) were tested. Fig. 8 shows the predicted sensitivity maps according to the TV regularization weight. When the weight was very small ($w_{\mathrm{TV}} = 10^{-4}$), the overall sensitivity map tended to be zeros, and only the small regions had high values. This indicates that most regions were regarded as distorted pixels; however, this did not provide a clear interpretation from the perspective of the HVS. As TV weight increased, the distribution of the sensitivity maps tended to be more uniform, as shown in Fig. 8(e) and (j).

The weight of TV also affected the prediction accuracy of the trained model. The models with sufficient magnitudes of weights ($w_{\mathrm{TV}} = 10^{-2}$ and 10^{-1}) for the TV regularization term showed higher SRCC, as shown in Table XI. Too sparse a sensitivity map, as in Fig. 8(b) and (g), could not generalize over the various database and distortion types well. It can be concluded that the TV term actually works as regularization during training and enhanced testing accuracy.

C. Error and Perceptual Error Maps Visualization

Following the training of the error map prediction stage, the model can generate an objective error map without using reference images. In Fig. 9, the predicted error maps and their ground-truth versions are compared. The "Lighthouse2" image from the LIVE database was used to show the results. This image was not included in the training set. Fig. 9 shows the prediction results of the error map. The images in the first column are distorted images with JEPG2000, white noise, and GB, respectively. In the second and third columns, the darker regions indicate more distorted pixels. For white noise, distortion was distributed uniformly over the image as shown

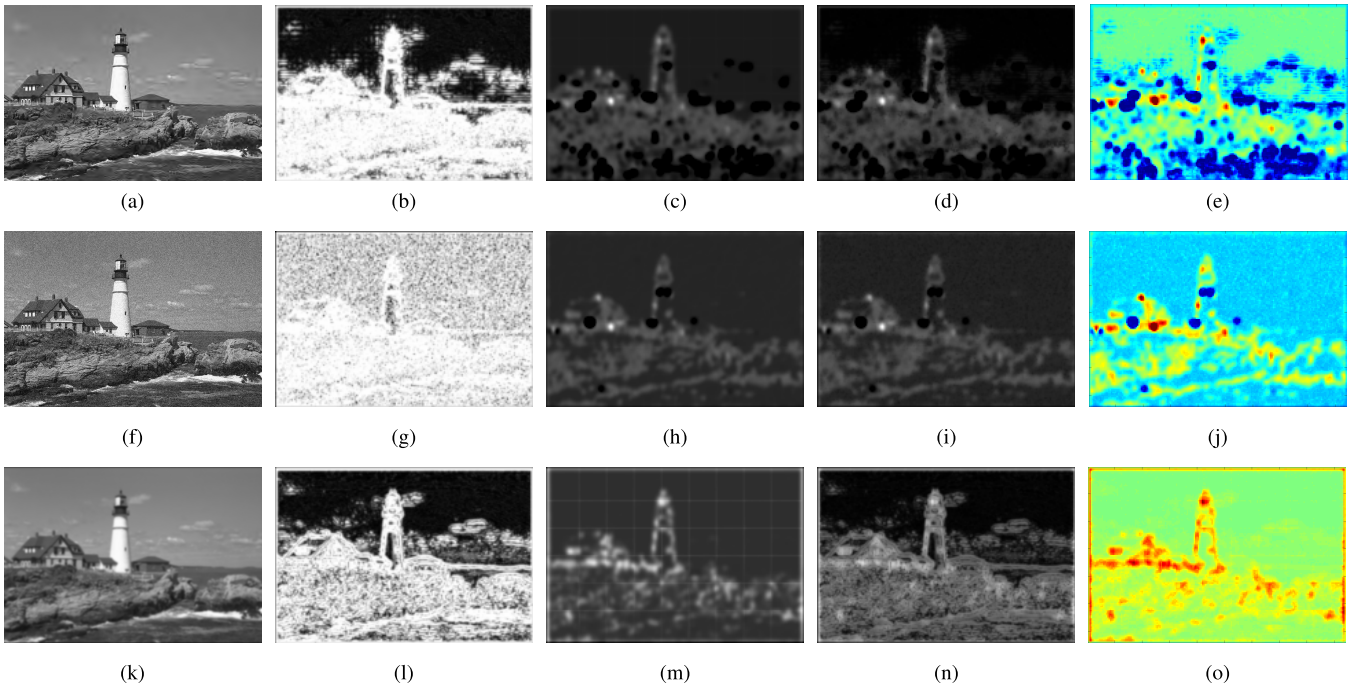


Fig. 10. Examples of perceptual error maps. (c), (h), and (r) Distorted images with JPEG2000, white noise, and GB. (b), (g), and (l) Reliability maps. (c), (h), and (r) Predicted sensitivity maps. (d), (i), and (n) Perceptual error maps. (e), (j), and (t) Comparison maps between objective and perceptual error maps.

TABLE XI
SRCC AND PLCC COMPARISON FOR EACH TV REGULARIZATION WEIGHT ON THE LIVE DATABASE

	TV regularization weight (w_{TV} =)				
Metric	0	10^{-4}	10^{-3}	10^{-2}	10^{-1}
SRCC	0.961	0.965	0.966	0.971	0.969
PLCC	0.963	0.966	0.967	0.972	0.970

Fig. 11. Example of JND map. (a) Reference image. (b) JND map [41].

in Fig. 9(g), and this was predicted well in Fig. 9(h). Since regions with low reliability were ignored during the training of the error map, the prediction was inaccurate on homogeneous regions. In the fourth and fifth columns, the ground-truth and predicted error maps are multiplied by the reliability maps as in (4), such that the inaccurate regions were ignored. The images in the last column are images actually used for NR-IQA.

The predicted perceptual error maps are shown in Fig. 10. The reliability maps [Fig. 10(b), (g), and (l)] emphasized high-frequency components, such as edges and complex textures. To analyze human visual sensitivity, we observed the perceptual error map rather than the sensitivity map. The role of the sensitivity map is tuning the objective error map by weighting. It is clear that low values in the perceptual error map can be regarded as perceptually distorted regions. However, it is difficult to ensure that low values in the sensitivity map indicate less important pixels, since the SNES subnetwork was trained based on the prepredicted error map.

For better comparison, the difference between the objective error and the perceptual error maps are shown in the last column [Fig. 10(e), (j), and (o)]. The blue (red) regions indicate that the perceptual error maps have lower (higher) values.

For JPEG2000, as shown in Fig. 10(e), the distortions around the clouds and the sea were emphasized, which agrees with the human perception. On the other hand, when the image was distorted by white noise, there was much decrease in sky regions than the rock regions, as shown in Fig. 10(j), which indicates that noise on the homogeneous regions was more noticeable than in the textural regions. For blurred images, overall tendency was similar between the objective and perceptual errors as shown in Fig. 10(o).

To further analyze the predicted perceptual error maps, we utilized the just noticeable difference (JND) models [41] and an FR-IQA metric, SSIM [35]. Fig. 11(b) shows the estimated JND map of the reference image (a), which was not included in the training set. The JND describes the maximum perceptual distortion that the typical HVS does not perceive. The JND models were designed based on perceptual computational models, such as luminance adaption and contrast masking. The higher values in the JND maps indicate the higher visibility threshold.

Fig. 12(a), (d), and (g) are objective error maps (JPEG2000, white noise, and GB) where the darker regions represent more distorted pixels. The corresponding distorted images are Fig. 10(a), (f), and (p). Due to the reliability maps, it is hard

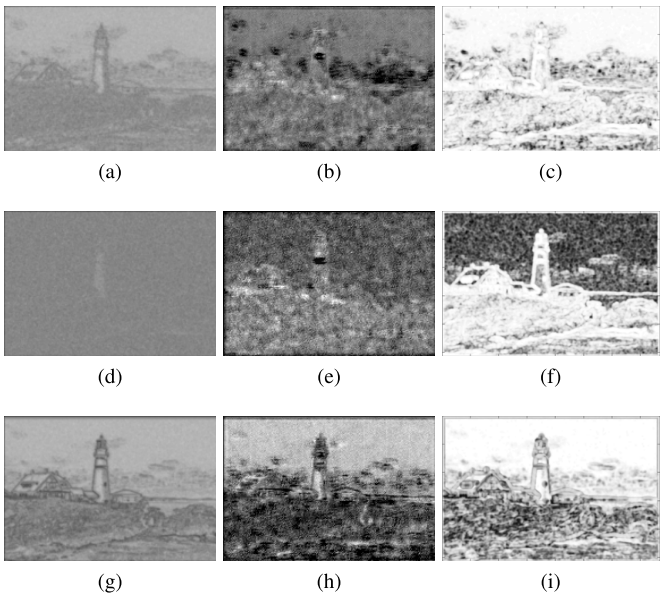


Fig. 12. Examples of perceptual error maps and SSIM maps. (a), (d), and (g) Ground-truth error maps caused by JPEG2000, white noise, and GB. (b), (e) and (h) Predicted perceptual error maps without reliability maps. (c), (f), and (i) Corresponding SSIM maps.

to analyze the perceptual error maps in the perspective of spatial frequency of pixels in Fig. 10. Therefore, we visualized the perceptual error maps when the reliability maps are not considered while using the same structure in an FR-IQA framework [43] in Fig. 12(b), (e), and (h).

By comparing the error maps with the JND map, it is clear that the distortions on the dark regions in the JND map are more visible to human observers. In Fig. 12(a), the distortions around the houses and cloud were more noticeable than those on the rocks, as shown in Fig. 12(b). For white noise, the objective error was uniformly distributed over the image. However, in Fig. 12(e), the distortions on the homogeneous regions were more noticeable than those on the textural regions, which agrees with contrast masking. When the image was distorted by GB, strong edges were especially distorted as Fig. 12(g), and the perceptual error map also had similar tendency, as shown in Fig. 12(h). We additionally showed the SSIM maps in Fig. 12(c), (f), and (i). The SSIM assumes that the HVS is highly sensitive to structural information. In general, the emphasized distortions coincide with the perceptual error maps. Though the predicted perceptual error maps were learned without using any prior knowledge of the HVS, it can be concluded that the results generally agree with the HVS indeed.

VI. Conclusion

We described a deep CNN-based NR-IQA framework. Applying a CNN to NR-IQA is a challenging issue, because there are critical obstacles. In the DIQA, an objective error map was used as an intermediate regression target to avoid overfitting with the limited database. When the first training stage is not run enough, the DIQA suffers from the overfitting problem leading to a degradation of performance. The input normalization and the reliability map increased the accuracy significantly as well. The final DIQA model outperformed all the benchmarked full-reference methods as well as no-reference methods. We further showed that the performance of the DIQA is independent of the selection of the database. We additionally proposed the DIQA-SENS to visualize and analyze the learned perceptual error maps. The perceptual error maps followed the behavior of the HVS. In the future, we will investigate a new way to obtain more meaningful sensitivity maps that can provide a more interpretable analysis with respect to the HVS.

References

[1] 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.

[2] 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.

[3] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015.

[4] Y. Li *et al.*, "No-reference image quality assessment with shearlet transform and deep neural networks," *Neurocomputing*, vol. 154, pp. 94–109, Apr. 2015.

[5] 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.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[7] S. J. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," *Proc. SPIE*, vol. 1666, pp. 179–206, Jan. 1992.

[8] A. B. Watson and A. J. Ahumada, "A standard model for foveal detection of spatial contrast," *J. Vis.*, vol. 5, no. 9, p. 6, 2005.

[9] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Amer.*, vol. 70, no. 12, pp. 1458–1471, Dec. 1980.

[10] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 4, no. 12, pp. 2379–2394, 1987.

[11] 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.

[12] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[13] 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.

[14] 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.

[15] 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.

[16] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.

[17] Y. Lv, G. Jiang, M. Yu, H. Xu, F. Shao, and S. Liu, "Difference of Gaussian statistical features based blind image quality assessment: A deep learning approach," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2344–2348.

[18] D. Ghadiyaram and A. C. Bovik, "Feature maps driven no-reference image quality prediction of authentically distorted images," *Proc. SPIE*, vol. 9394, p. 93940J, Mar. 2015.

[19] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1733–1740.

[20] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3773–3777.

[21] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.
[22] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.
[23] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.
[24] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[26] 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.
[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[28] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Proc. ICLR Workshop*, 2016.
[29] 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.
[30] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015. [Online]. Available: http://ponomarenko.info/tid2013.htm
[31] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 1693–1697.
[32] D. Ghadiyaram and A. C. Bovik, "Massive Online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.
[33] K. Ma *et al.*, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.
[34] *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II (FR-TV2)*, document ITU-T SG09, VQEG, 2003.
[35] 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.
[36] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[37] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1969–1977.
[38] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.
[39] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.
[40] Q. Li, W. Lin, J. Xu, and Y. Fang, "Blind image quality assessment using statistical structural and luminance features," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2457–2469, Dec. 2016.
[41] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, "Just noticeable difference for images with decomposition model for separating edge and textured regions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1648–1652, Nov. 2010.
[42] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, 2017.
[43] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1676–1684.

Jongyoo Kim (M'15) received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2011, 2013, and 2018, respectively.

He joined Microsoft Research Asia in 2018. His current research interests include 2-D/3-D image and video processing based on human visual system, quality assessment of 2-D/3-D image and video, 3-D computer vision, and deep learning.

Dr. Kim was a recipient of the Global Ph.D. Fellowship at the National Research Foundation of Korea from 2011 to 2016.

Anh-Duc Nguyen received the B.Eng. degree in automatic control from the Hanoi University of Science and Technology, Hanoi, Vietnam, in 2015. He is currently pursuing the Ph.D. degree with Yonsei University, Seoul, South Korea.

His current research interests include computer vision, image/video analysis, and machine learning.

Sanghoon Lee (M'05–SM'12) received the B.S. degree from Yonsei University, Seoul, South Korea, in 1989, the M.S. degree from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1991, and the Ph.D. degree from The University of Texas at Austin, Austin, TX, USA, in 2000, all in E.E.

From 1991 to 1996, he was with Korea Telecom, Seongnam, South Korea. From 1999 to 2002, he was with Lucent Technologies, Murray Hill, NJ, USA, on 3G wireless and multimedia networks. In 2003, he joined the Department of Electrical and Electronics Engineering, Yonsei University, as a Faculty Member, where he is currently a Full Professor. His current research interests include image/video quality assessment, computer vision, graphics, cloud computing, multimedia communications, and wireless networks.

Dr. Lee received the 2015 Yonsei Academic Award from Yonsei University, the 2012 Special Service Award from the IEEE Broadcast Technology Society, and the 2013 Special Service Award from the IEEE Signal Processing Society. He was the Technical Program Co-Chair of the International Conference on Information Networking 2014 and the Global 3D Forum 2012 and 2013, respectively, and the General Chair of the 2013 IEEE IVMSP Workshop. He has been serving as the Chair of the IEEE P3333.1 Quality Assessment Working Group since 2011. He currently serves on the Technical Committee of the IEEE IVMSP Technical Committee since 2014 and the IEEE Multimedia Signal Processing since 2016. He also served as an Editor of the JOURNAL OF COMMUNICATIONS AND NETWORKS from 2009 to 2015 and a special issue Guest Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING in 2013. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2010 to 2014. He has been an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS since 2014 and the *Journal of Electronic Imaging* since 2015.