IEEE Transactions on Image Processing, accepted for publication, 2017.

End-to-End Blind Image Quality Assessment Using Deep Neural Networks

Kede Ma, *Student Member, IEEE,* Wentao Liu, *Student Member, IEEE,* Kai Zhang, Zhengfang Duanmu, *Student Member, IEEE,* Zhou Wang, *Fellow, IEEE,* and Wangmeng Zuo, *Senior Member, IEEE*

***Abstract*—We propose a Multi-task End-to-end Optimized deep neural Network (MEON) for blind image quality assessment (BIQA). MEON consists of two sub-networks—a distortion identification network and a quality prediction network—sharing the early layers. Unlike traditional methods used for training multi-task networks, our training process is performed in two steps. In the first step, we train a distortion type identification sub-network, for which large-scale training samples are readily available. In the second step, starting from the pre-trained early layers and the outputs of the first sub-network, we train a quality prediction sub-network using a variant of the stochastic gradient descent method. Different from most deep neural networks (DNN), we choose biologically inspired generalized divisive normalization (GDN) instead of rectified linear unit (ReLU) as the activation function. We empirically demonstrate that GDN is effective at reducing model parameters/layers while achieving similar quality prediction performance. With modest model complexity, the proposed MEON index achieves state-of-the-art performance on four publicly available benchmarks. Moreover, we demonstrate the strong competitiveness of MEON against state-of-the-art BIQA models using the group MAximum Differentiation (gMAD) competition methodology.**

***Index Terms*—Blind image quality assessment, deep neural networks, multi-task learning, generalized divisive normalization, gMAD competition.**

I. Introduction

BLIND image quality assessment (BIQA) aims to predict the perceptual quality of a digital image with no access to its pristine counterpart [1]. It is a fundamental problem in image processing that has not been fully resolved [2]. Early BIQA models are mainly based on hand-crafted features [3]–[6], which rely heavily on knowledge of the probabilistic structures of our visual world, the mechanisms of image degradations, and the functionalities of the human visual system (HVS) [7], [8]. Built upon feature representations, a quality prediction function is learned using the ground truth data in the form of subject-rated images. Typically, the knowledge-driven feature extraction and data-driven quality prediction stages are designed separately. With the recent exciting development of deep neural network (DNN) methodologies [9], a fully data-driven end-to-end BIQA solution becomes possible.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, and the NSFC grant (61671182). K. Ma was partially supported by the CSC.

Kede Ma, Wentao Liu, Zhengfang Duanmu, and Zhou Wang are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: {k29ma, w238liu, zduanmu, zhou.wang}@uwaterloo.ca).

Kai Zhang and Wangmeng Zuo are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: {cskaizhang, cswmzuo}@gmail.com).

Although DNN has shown great promises in many vision tasks [9]–[11], end-to-end optimization of BIQA is challenging due to the lack of sufficient ground truth samples for training. Note that the largest subject-rated image quality assessment (IQA) database contains only $3,000$ annotations [12], while digital images live in a space of millions of dimensions. Previous DNN-based BIQA methods tackle this challenge in three ways. Methods of the first kind [13] directly inherit the architectures and weights from pre-trained networks for general image classification tasks [14] followed by fine-tuning. The performance and efficiency of such networks depend highly on the generalizability and relevance of the tasks used for pre-training. The second kind of methods [15]–[17] work with image patches by assigning the subjective mean opinion score (MOS) of an image to all patches within it. This approach suffers from three limitations. First, the concept of quality without context (*e.g.*, the quality of a single 32×32 patch) is not well defined [7], [18]. Second, local image quality within context (*e.g.*, the quality of a 32×32 patch within a large image) varies across spatial locations even when the distortion is homogeneously applied [19]. Third, patches with similar statistical behaviors (*e.g.*, smooth and blurred regions) may have substantially different quality [20]. Methods of the third kind [21] make use of full-reference IQA (FR-IQA) models for quality annotation. Their performance is directly affected by that of FR-IQA models, which may be inaccurate across distortion levels [2] and distortion types [12]. Other methods for generating training data involve the creation of synthetic scores [22] and discriminable image pairs (DIP) [23], both of which rely on FR-IQA models and may suffer from similar problems.

In this work, we describe a framework for end-to-end BIQA based on multi-task learning. Motivated by previous works [16], [24], we decompose the BIQA problem into two subtasks. Subtask I classifies an image into a specific distortion type from a set of pre-defined categories. Subtask II predicts the perceptual quality of the same image, taking advantage of distortion information obtained from Subtask I. On the one hand, the two subtasks are related because quality degradation arises from distortion and the quality level is also affected by the distortion amount. On the other hand, they are different because images with different distortion types may exhibit similar quality while images with the same distortion type may have drastically different quality, as shown in Fig. 1. The subtasks are accomplished by two sub-networks of linear convolutions and nonlinearities with shared features at early layers. Feature sharing not only greatly reduces the computa-

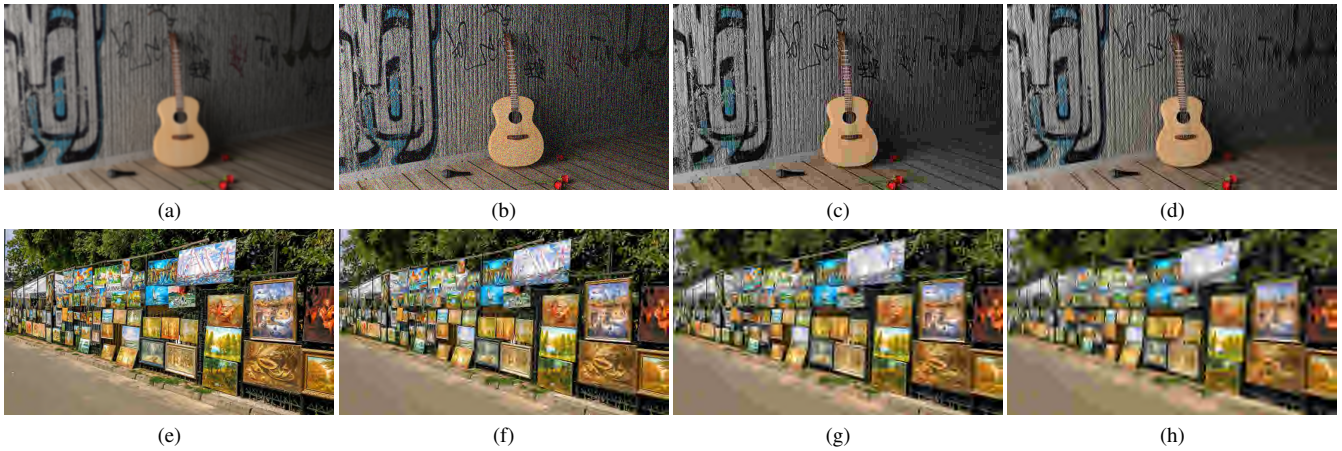


Fig. 1. Images (a)-(d) with different distortion types have similar quality while images (e)-(h) of the same distortion type have different quality, according to our subjective testing. (a) Gaussian blurring. (b) Gaussian noise contamination. (c) JPEG compression. (d) JPEG2000 compression. (e)-(h) JPEG2000 compression with increasing compression ratios from left to right.

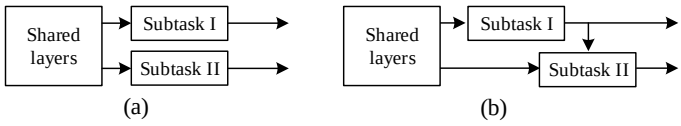


Fig. 2. (a) Traditional multi-task learning [16]. (b) Proposed multi-task learning structure.

tion, but also enables the network to pre-train the shared layers via Subtask I, for which large-scale training data (distortion type) can be automatically generated at low cost. Unlike traditional multi-task learning, Subtask II of our method depends on the outputs of Subtask I, as shown in Fig. 2. As such, the distortion information is transparent to Subtask II for better quality prediction. We define a layer that is differential with respect to both convolutional activations and outputs of Subtask I to guarantee the feasibility of backpropagation. After pre-training, the entire network is end-to-end optimized using a variant of the stochastic gradient descent method. In addition, instead of using rectified linear unit (ReLU) [25], we adopt generalized divisive normalization (GDN) joint nonlinearity as the activation function that is biologically inspired and has proven effective in assessing image quality [26], Gaussianizing image densities [27], and compressing digital images [28]. We empirically show that GDN has the capability of reducing model parameters/layers and meanwhile maintaining similar quality prediction performance. We evaluate the resulting Multi-task End-to-end Optimized Network (MEON) based image quality index on four publicly available IQA databases and demonstrate that it achieves state-of-the-art performance compared with existing BIQA models. Finally, we investigate the generalizability and robustness of MEON using the group MAximum Differentiation (gMAD) competition methodology [2] on the Waterloo Exploration Database [29]. We observe that MEON significantly outperforms the most recent DNN-based BIQA model [17] and is highly competitive against MS-SSIM [30], a well-known FR-IQA model.

II. Related Work

In this section, we provide a brief review of feature engineering in BIQA and previous studies closely related to our work. For a more comprehensive treatment of general IQA and BIQA, please refer to [23], [31]–[34].

Assuming the distortion affecting an image is known, early BIQA research focused on extracting distortion-specific features that can handle only one distortion type, *e.g.*, JPEG/JPEG2000 compression [35], [36] and blurring artifacts [37]. Only in the past decade has general BIQA become an active research topic, for which spatially normalized coefficients [4] and codebook-based features [38] are popular. In BRISQUE [4], inspired by earlier work on reduced-reference (RR) IQA using local gain control based divisive normalization [26], natural scene statistics (NSS) are extracted from locally normalized luminance coefficients. Such a normalization approach has been used in many BIQA models [38]–[40] as a starting point of feature extraction or a preprocessing step for DNN-based BIQA models [15], [16], [21]. In CORNIA [38], a codebook is constructed by clustering spatially normalized patches with k-means, based on which soft-assignment encoding and feature pooling are performed. Despite its high dimension, CORNIA features have been frequently adopted in later BIQA models such as BLISS [22] and dipIQ [23]. The feature set has been improved to HOSA [41] by incorporating higher order statistics.

Kang *et al.* [15] implemented a DNN with one convolutional and two fully connected layers for BIQA as an end-to-end version of CORNIA [38]. In order to perform both maximum and minimum pooling, ReLU nonlinearity [25] is omitted right after convolution. Bianco *et al.* [13] investigated various design choices of DNN for BIQA. They first adopted DNN features pre-trained on the image classification task as inputs to learn a quality evaluator using support vector regression (SVR) [42]. They then fine-tuned the pre-trained features in a multi-class classification setting by quantizing the MOS into five categories, and fed the fine-tuned features to SVR. Nevertheless, their proposal is not end-to-end optimized and involves heavy manual parameter adjustments [13]. Bosse *et*

TABLE I
MODEL SIZE COMPARISON OF DNN-BASED BIQA MODELS

BIQA model	Kang14 [15]	Kang15 [16]	DeepBIQ [13]	deepIQA [17]	Kim17 [21]	MEON
Model size ($\times 10^4$)	72	7.9	5,687	523	739	10.6

al. [17] significantly increased the depth of DNN by stacking ten convolutional and two fully connected layers, whose architecture was inspired by the VGG16 network [10] for image classification. They also adapted their network to handle FR-IQA. Kim and Lee [21] first utilized the local score of an FR-IQA algorithm as the ground truth to pre-train the model and then fine-tuned it using MOSs. They observed that pre-training with adequate epochs is necessary for the fine-tuning step to converge. All of the above methods either work with image patches, which may suffer from noisy training labels, or inherit network structures from other tasks with low relevance and unnecessary complexity. We summarize model complexities of DNN-based models in Table I.

Our work is motivated by two previous methods. In BIQI [24], Moorthy and Bovik proposed a two-step framework for BIQA, where an image is first classified into a particular distortion category, and then the distortion-specific quality prediction is performed [24]. The two steps of BIQI are optimized separately. Unlike BIQI, we are aiming at an end-to-end solution, meaning that feature representation, distortion type identification, and quality prediction are optimized jointly. In [16], Kang *et al.* simultaneously estimated image quality and distortion type via a traditional multi-task DNN. However, simultaneous multi-task training requires ground truths of distortion type and subjective quality to be both available, which largely limits the total number of valid training samples. In addition, the quality prediction subtask is ignorant of the output from the distortion identification subtask. As a result, the performance is less competitive.

III. MEON FOR BIQA

In the proposed MEON index, we take a raw image of $256 \times 256 \times 3$ as input and predict its perceptual quality score. How larger images are handled will be explained in Section III-C. MEON consists of two subtasks accomplished by two sub-networks. Sub-network I aims to identify the distortion type in the form of a probability vector, which indicates the likelihood of each distortion and is fed as partial input to Sub-network II, whose goal is to predict the image quality. Each subtask involves a loss function. Since Sub-network II relies on the output of Sub-network I, the two loss terms are not independent. We pre-train the shared layers in MEON via Subtask I and then jointly optimize the entire network with a unified loss function.

In this section, we first describe GDN as our nonlinear activation function used in MEON and then present in detail the construction of the two subtasks in Fig. 3. Finally, we introduce our end-to-end training and testing procedures.

A. GDN as Activation Function

Since Nair and Hinton revealed the importance of the ReLU nonlinearity in accelerating the training of DNN [25], ReLU and its variants [43], [44] have become the dominant activation functions in DNN literature. However, the joint statistics of linear filter responses after ReLU exhibit strong higher-order dependencies [27], [28]. As a result, ReLU generally requires a substantially large number of model parameters to achieve good performance for a particular task. These higher-order statistics may be significantly decorrelated through the use of a joint nonlinear gain control mechanism [45], [46] inspired by models of visual neurons [47], [48]. Previous studies also showed that incorporating the local gain control operation in DNN improves the generalizability in image classification [9] and object recognition [49], where the parameters are predetermined empirically and fixed during training. Here, we adopt a GDN transform that has been previously demonstrated to work well in density estimation [27] and image compression [28]. Specifically, given an S-dimensional linear convolutional activation $\mathbf{x}(m,n) = [x_1(m,n), \cdots, x_S(m,n)]^T$ at spatial location (m,n), the GDN transform is defined as

$$y_i(m,n) = \frac{x_i(m,n)}{\left(\beta_i + \sum_{j=1}^{S} \gamma_{ij} x_j(m,n)^2\right)^{\frac{1}{2}}}, \tag{1}$$

where $\mathbf{y}(m,n) = [y_1(m,n), \cdots, y_S(m,n)]^T$ is the normalized activation vector at spatial location (m,n). The weight matrix $\boldsymbol{\gamma}$ and the bias vector $\boldsymbol{\beta}$ are parameters in GDN to be optimized. Both of them are confined to $[0, +\infty)$ so as to ensure the legitimacy of the square root operation in the denominator and are shared across spatial locations. GDN is a differentiable transform that can be trained with any preceding or subsequent layers. In addition, GDN is proven to be iteratively invertible under mild assumptions [27], which preserves better information than ReLU.

During training, we need to backpropagate the gradient of the loss ℓ through the GDN transform and compute the gradients with respect to its inputs and parameters. According to the chain rule

$$\frac{\partial \ell}{\partial x_j(m,n)} = \sum_{i=1}^{S} \frac{\partial \ell}{\partial y_i(m,n)} \frac{\partial y_i(m,n)}{\partial x_j(m,n)}, \tag{2}$$

$$\frac{\partial \ell}{\partial \beta_i} = \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial \ell}{\partial y_i(m,n)} \frac{\partial y_i(m,n)}{\partial \beta_i}, \tag{3}$$

$$\frac{\partial \ell}{\partial \gamma_{ij}} = \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial \ell}{\partial y_i(m,n)} \frac{\partial y_i(m,n)}{\partial \gamma_{ij}}, \tag{4}$$

where M and N denote the spatial sizes of the GDN trans-

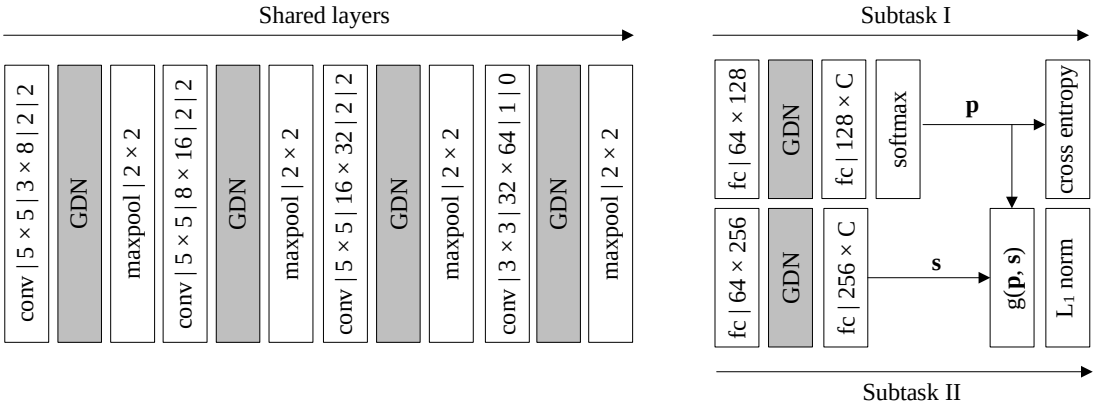


Fig. 3. Illustration of MEON configurations for BIQA, highlighting the GDN nonlinearity. We follow the style and convention in [28], and denote the parameterization of the convolutional layer as "height × width | input channel × output channel | stride | padding".

formed coefficients and

$$\frac{\partial y_i(m,n)}{\partial x_j(m,n)} = \begin{cases} \frac{\beta_i + \sum_{k\neq i}\gamma_{ik}x_k(m,n)^2}{\left(\beta_i + \sum_{k=1}^{S}\gamma_{ik}x_k(m,n)^2\right)^{\frac{3}{2}}} & i = j \\ \frac{-\gamma_{ij}x_i(m,n)x_j(m,n)}{\left(\beta_i + \sum_{k=1}^{S}\gamma_{ik}x_k(m,n)^2\right)^{\frac{3}{2}}} & i \neq j \end{cases}, \quad (5)$$

$$\frac{\partial y_i(m,n)}{\partial \beta_i} = \frac{-x_i(m,n)}{2\left(\beta_i + \sum_{j=1}^{S}\gamma_{ij}x_j(m,n)^2\right)^{\frac{3}{2}}}, \quad (6)$$

$$\frac{\partial y_i(m,n)}{\partial \gamma_{ij}} = \frac{-x_i(m,n)x_j(m,n)^2}{2\left(\beta_i + \sum_{k=1}^{S}\gamma_{ik}x_k(m,n)^2\right)^{\frac{3}{2}}}. \quad (7)$$

Some DNNs incorporate the batch normalization (BN) transform [50] that whitens the responses of linear filters to reduce the internal covariate shift and to rescale them in a reasonable operating range. GDN is different from BN in many ways. First, during testing, the mean and variance parameters are fixed and BN is simply an affine transform applied to the input. By contrast, GDN offers high nonlinearities especially when it is cascaded in multiple stages. Second, BN jointly normalizes all the activations across the mini-batch and over all spatial locations, which makes it an element-wise operation. Although the parameters in GDN are shared across the space similar to BN, the normalization of one activation at one location involves all activations across the channel, making it spatially adaptive. Another transform that is closely related to GDN is the local response normalization (LRN) [9], which has a form of

$$y_i(m,n) = \frac{x_i(m,n)}{\left(\beta' + \gamma' \sum_{j=\max(1,i-S'/2)}^{\min(S,i+S'/2)} x_j(m,n)^2\right)^{\alpha'}}, \quad (8)$$

where α', β', γ', and S' are scalar parameters predetermined using a validation set. The sum in the denominator runs over S' adjacent activations at the same spatial location. LRN has been used to boost the performance of image classification [9] and object recognition [49]. Both GDN and LRN are inspired by models of biological neurons. When the fixed exponent of $\frac{1}{2}$ in the denominator is generalized to a scalar parameter, LRN becomes a special case of GDN. We experiment with such a generalized version of Eq. (1), but do not observe noticeable performance gains. Therefore, we choose to use Eq. (1) throughout the paper.

B. Network Architecture

We denote our input mini-batch training data set by $\left\{\left(\mathbf{X}^{(k)}, \mathbf{p}^{(k)}, q^{(k)}\right)\right\}_{k=1}^{K}$, where $\mathbf{X}^{(k)}$ is the k-th raw input image, $\mathbf{p}^{(k)}$ is a multi-class indicator vector with only one entry activated to encode the ground truth distortion type, and $q^{(k)}$ is the MOS of the k-th input image. As depicted in Fig. 3, we first feed $\mathbf{X}^{(k)}$ to the shared layers, which are responsible for transforming raw image pixels into perceptually meaningful and distortion relevant feature representations. It consists of four stages of convolution, GDN, and maxpooling, whose model parameters are collectively denoted by $\mathbf{W}$. The parameterizations of convolution, maxpooling, and connectivity from layer to layer are detailed in Fig. 3. We reduce the spatial size by a factor of 4 after each stage via convolution with a stride of 2 (or without padding), and 2×2 maxpooling. As a result, we represent a $256 \times 256 \times 3$ raw image by a 64-dimensional feature vector. On top of the shared layers, Sub-network I appends two fully connected layers with an intermediate GDN transform to increase nonlinearity, whose parameters are denoted by $\mathbf{w}_1$. We adopt the softmax function to encode the range to $[0, 1]$

$$\hat{p}_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1) = \frac{\exp\left(y_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1)\right)}{\sum_{j=1}^{C}\exp\left(y_j^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1)\right)}, \quad (9)$$

where $\hat{\mathbf{p}}^{(k)} = [\hat{p}_1^{(k)}, \cdots, \hat{p}_C^{(k)}]^T$ is a C-dimensional probability vector of the k-th input in a mini-batch, which indicates the probability of each distortion type. We take pristine images into account and use one entry to represent the "pristine" category. $\hat{\mathbf{p}}^{(k)}$ is the quantity fed to sub-network II and creates the dependent structure. For Subtask I, we consider the empirical cross entropy loss

$$\ell_1(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1) = -\sum_{k=1}^{K}\sum_{i=1}^{C} p_i^{(k)} \log \hat{p}_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1). \quad (10)$$

Since we feed pristine images into Sub-network I by adding the "pristine" category, our training set is mildly unbalanced. Specifically, the number of images suffering from a particular distortion is L times as many as pristine images, where L is the number of distortion levels. It is straightforward to offset such class imbalance by adding weights in Eq. (10) according to the proportion of each distortion type. In our experiments, instead of over-weighting pristine images in the loss function, we over-sample them L times during training, which is beneficial for learning strong discriminative features to handle mild distortion cases.

Sub-network II takes the shared convolutional features and the estimated probability vector $\hat{\mathbf{p}}^{(k)}$ from Sub-network I as inputs. It predicts the perceptual quality of $\mathbf{X}^{(k)}$ in the form of a scalar value $\hat{q}^{(k)}$, where a lower score indicates worse perceptual quality. As in Sub-network I, to increase nonlinearity, we append two fully connected layers with an intermediate GDN layer, whose parameters are collectively denoted by $\mathbf{w}_2$. We double the node number of the first fully connected layer compared with that of Sub-network I, because predicting image quality is expected to be more difficult than identifying the distortion type. After the second fully connected layer, the network produces a score vector $\mathbf{s}^{(k)}$, whose i-th entry represents the perceptual quality score corresponding to the i-th distortion type. We define a fusion layer that combines $\hat{\mathbf{p}}^{(k)}$ and $\mathbf{s}^{(k)}$ to yield an overall quality score

$$\hat{q}^{(k)} = g(\hat{\mathbf{p}}^{(k)}, \mathbf{s}^{(k)}) . \tag{11}$$

We continue by completing the definition of $g(\cdot)$. First, in order to achieve theoretically valid backpropagation, g should be differentiable with respect to both $\hat{\mathbf{p}}^{(k)}$ and $\mathbf{s}^{(k)}$. Second, pairs $(\hat{p}_i^{(k)}, s_i^{(k)})$ and $(\hat{p}_j^{(k)}, s_j^{(k)})$ should be interchangeable in g to reflect the equal treatment of each distortion type under no privileged information. Third, g needs to be intuitively reasonable. For example, more emphasis should be given to $s_i^{(k)}$ if $\hat{p}_i^{(k)}$ is larger; $\hat{q}^{(k)}$ should be monotonically non-decreasing with respect to each entry of $\mathbf{s}^{(k)}$. Here, we adopt a probability-weighted summation [24] as a simple implementation of g

$$\hat{q}^{(k)} = \hat{\mathbf{p}}^{(k)T}\mathbf{s}^{(k)} = \sum_{i=1}^{C} \hat{p}_i^{(k)} \cdot s_i^{(k)} , \tag{12}$$

which is easily seen to obey all the properties listed above. We have also tried the outer product implementation with non-negative weights learned during training and obtained similar results. For subtask II, we use the ℓ_1-norm as the empirical loss function

$$\ell_2(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_2) = \|\mathbf{q} - \hat{\mathbf{q}}\|_1 = \sum_{k=1}^{K} |q^{(k)} - \hat{q}^{(k)}| . \tag{13}$$

We have also tried the ℓ_2-norm as the loss and observed similar performance. This is different from patch-based DNN methods [17], which show a clear preference to the ℓ_1-norm due to a high degree of label noise in the training data.

We now define the overall loss function of MEON as

$$\ell(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1, \mathbf{w}_2) = \ell_1 + \lambda \ell_2 , \tag{14}$$

where λ is the balance weight to account for the scale difference between the two terms or to impose relative emphasis on one over the other.

We finish this subsection by highlighting another special treatment of MEON in addition to Eq. (11) and Eq. (12). The gradient of ℓ with respect to $\hat{p}_i^{(k)}$ in Sub-network I

$$\frac{\partial \ell}{\partial \hat{p}_i^{(k)}} = \frac{\partial \ell_1}{\partial \hat{p}_i^{(k)}} + \lambda \frac{\partial \ell_2}{\partial \hat{p}_i^{(k)}} \tag{15}$$

$$= -\frac{p_i^{(k)}}{\hat{p}_i^{(k)}} - \lambda \text{sign}\left(q^{(k)} - \hat{q}^{(k)}\right) s_i^{(k)} \tag{16}$$

depends on the gradient backpropagated from Sub-network II.

C. Training and Testing

The success of DNN is largely owing to the availability of large-scale labeled training data. However, in BIQA, it is difficult to source accurate MOSs at a large scale. MEON tackles this problem by dividing the training into two steps: pre-training and joint optimization. At the pre-training step, we minimize the loss function in Subtask I

$$(\hat{\mathbf{W}}, \hat{\mathbf{w}}_1) = \text{argmin}\, \ell_1(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1) . \tag{17}$$

The training set used for pre-training can be efficiently generated without subjective testing. Details will be discussed in Section IV. At the joint optimization step, we initialize $(\mathbf{W}, \mathbf{w}_1)$ with $(\hat{\mathbf{W}}, \hat{\mathbf{w}}_1)$ and minimize the overall loss function

$$(\mathbf{W}^\star, \mathbf{w}_1^\star, \mathbf{w}_2^\star) = \text{argmin}\, \ell(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1, \mathbf{w}_2) . \tag{18}$$

During testing, given an image, we extract $256 \times 256 \times 3$ sub-images with a stride of U. The final distortion type is computed by majority vote among all predicted distortion types of the extracted sub-images. Similarly, the final quality score is obtained by simply averaging all predicted scores.

IV. Experiments

In this section, we first describe the experimental setups including implementation details of MEON, IQA databases, and evaluation criteria. We then compare MEON with classic and state-of-the-art BIQA models. Finally, we conduct a series of ablation experiments to identify the contributions of the core factors in MEON.

A. Experimental Setups

1) Implementation Details: Both pre-training and joint optimization steps adopt the Adam optimization algorithm [51] with a mini-batch of 40. For pre-training, we start with the learning rate $\alpha = 10^{-2}$ and subsequently lower it by a factor of 10 when the loss plateaus, until $\alpha = 10^{-4}$. For joint optimization, α is fixed to 10^{-4}. Other parameters in Adam are set by default [51]. The learning rates for biases are doubled. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in GDN are projected to nonnegative values after each update. Additionally, we enforce $\boldsymbol{\gamma}$ to be symmetric by averaging it with its transpose as recommended in [28]. The balance weight in Eq. (14) is set to account for the scale difference between the two terms (0.2 for LIVE [52] and

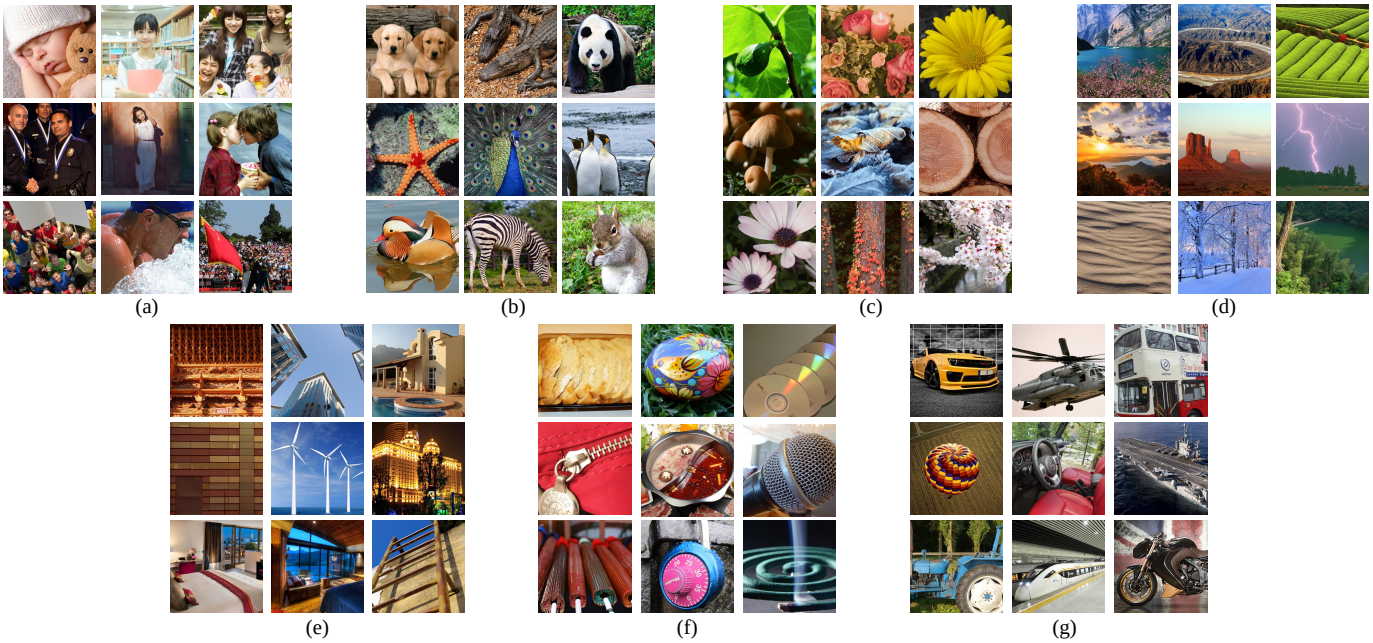


Fig. 4. Sample source images used for pre-training. (a) Human. (b) Animal. (c) Plant. (d) Landscape. (e) Cityscape. (f) Still-life. (g) Transportation. All images are cropped for better visibility.

1 for TID2013 [12]). During testing, the stride U is set to 128. We augment the training data by randomly horizontal flipping and changing their contrast and saturation within the range that is indiscernible to human eyes. Since quality changes with scales which correspond to different viewing distances, we do not augment training data across scales.

We select 840 high-resolution natural images with nearly pristine quality as the basis to construct the dataset for pre-training. They can be loosely categorized into seven classes: human, animal, plant, landscape, cityscape, still-life, and transportation, with representative images shown in Fig. 4. We down-sample each image to further reduce possible compression artifacts, keeping a maximum height or width of 768. $C-1$ distortion types (excluding the "pristine" category) are applied to those images, each with 5 distortion levels. As previously described, we over-sample pristine images to balance the class labels during pre-training. Therefore, our dataset contains a total of $C \times 840 \times 5$ images with ground truth labels automatically generated.

2) IQA Databases: We compare MEON with classic and state-of-the-art BIQA models on four standard IQA databases. They are LIVE [52], CSIQ [53], TID2013 [12], and the Waterloo Exploration Database [29]. The first three databases are subject-rated while MOSs are not available in the Exploration database (which calls for innovative evaluation criteria as will be introduced in Section IV-A3). In the first set of experiments, we consider four distortion types that are common in the four databases: JPEG2000 compression (JP2K), JPEG compression (JPEG), white Gaussian noise contamination (WN), and Gaussian blur (BLUR). This leaves us 634, 600, 500, and 94880 test images in LIVE [52], CSIQ [53], TID2013 [12], and the Exploration database, respectively. In the second set of experiments, we investigate the effectiveness of MEON on handling more distortion types (24 to be specific) by considering all $3,000$ test images in TID2013 [12].

3) Evaluation Criteria: Five evaluation criteria are adopted as follows.

- Spearman's rank-order correlation coefficient (SRCC): It is a nonparametric measure and is defined as

$$\mathrm{SRCC} = 1 - \frac{6\sum_i d_i^2}{I(I^2-1)}, \tag{19}$$

  where I is the test image number and d_i is the rank difference between the MOS and the model prediction of the i-th image. SRCC is independent of monotonic mappings.
- Pearson linear correlation coefficient (PLCC): It is a nonparametric measure of the linear correlation

$$\mathrm{PLCC} = \frac{\sum_i (q_i - q_m)(\hat{q}_i - \hat{q}_m)}{\sqrt{\sum_i (q_i - q_m)^2}\sqrt{\sum_i (\hat{q}_i - \hat{q}_m)^2}}, \tag{20}$$

  where q_i and $\hat{q}_i$ stand for the MOS and the model prediction of the i-th image, respectively.
- Pristine/distorted image discriminability test (D-test) [29]: It quantifies the ability of a BIQA model to discriminate pristine from distorted images. Given a database, we group the indices of pristine and distorted images into sets S_p and S_d, respectively. Based on model predictions, an optimal threshold T^* can be found to maximize the correct classification rate

$$D = R(T^*) = \frac{1}{2}\left(\frac{|S_p \cap S_p'|}{|S_p|} + \frac{|S_d \cap S_d'|}{|S_d|}\right), \tag{21}$$

  where $S_p' = \{i|\hat{q}_i > T^*\}$ and $S_d' = \{i|\hat{q}_i \leq T^*\}$. D ranges from $[0, 1]$ with a larger value indicating a better separability induced by the BIQA model.

TABLE II
MEDIAN SRCC AND PLCC RESULTS ACROSS 1,000 SESSIONS ON CSIQ [53]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
DIIVINE [3]	0.844	0.819	0.881	0.884	0.835
BRISQUE [4]	0.894	0.916	**0.934**	0.915	0.909
CORNIA [38]	0.916	0.919	0.787	**0.928**	0.914
ILNIQE [40]	0.924	0.905	0.867	0.867	0.887
BLISS [22]	**0.932**	0.927	0.879	0.922	0.920
HOSA [41]	0.920	0.918	0.895	0.915	0.918
dipIQ [23]	**0.944**	**0.936**	0.904	**0.932**	**0.930**
deepIQA [17]	0.907	0.929	0.933	0.890	0.871
MEON	0.898	**0.948**	**0.951**	0.918	**0.932**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
DIIVINE [3]	0.898	0.818	0.903	0.909	0.855
BRISQUE [4]	0.937	0.960	**0.947**	0.936	0.937
CORNIA [38]	0.947	0.960	0.777	**0.953**	0.934
ILNIQE [40]	0.942	0.956	0.880	0.903	0.914
BLISS [22]	**0.954**	0.970	0.895	0.947	0.939
HOSA [41]	0.946	0.958	0.912	0.940	0.942
dipIQ [23]	**0.959**	**0.975**	0.927	**0.958**	**0.949**
deepIQA [17]	0.931	0.951	0.933	0.906	0.891
MEON	0.925	**0.979**	**0.958**	0.946	**0.944**

TABLE III
MEDIAN SRCC AND PLCC RESULTS ACROSS 1,000 SESSIONS ON TID2013 [12]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
DIIVINE [3]	0.857	0.680	0.879	0.859	0.795
BRISQUE [4]	0.906	0.894	0.889	0.886	0.883
CORNIA [38]	0.907	0.912	0.798	**0.934**	0.893
ILNIQE [40]	0.912	0.873	0.890	0.815	0.881
BLISS [22]	0.906	0.893	0.856	0.872	0.836
HOSA [41]	**0.933**	0.917	0.843	0.921	**0.904**
dipIQ [23]	0.926	**0.932**	0.905	**0.922**	0.877
deepIQA [17]	**0.948**	**0.921**	**0.938**	0.910	0.885
MEON	0.911	0.919	**0.908**	0.891	**0.912**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
DIIVINE [3]	0.901	0.696	0.882	0.860	0.794
BRISQUE [4]	0.919	0.950	0.886	0.884	0.900
CORNIA [38]	0.928	0.960	0.778	**0.934**	0.904
ILNIQE [40]	0.929	0.944	0.899	0.816	0.890
BLISS [22]	0.930	0.963	0.863	0.872	0.862
HOSA [41]	**0.952**	0.949	0.842	0.921	**0.918**
dipIQ [23]	0.948	**0.973**	0.906	**0.928**	0.894
deepIQA [17]	**0.963**	0.960	**0.943**	0.897	**0.913**
MEON	0.924	**0.969**	**0.911**	0.899	0.912

- Listwise ranking consistency test (L-test) [29]: It exams the consistency of a BIQA model under test images differing only in distortion levels. The assumption here is that image quality degrades monotonically with the increase of the distortion level for any distortion type. Given a database with J source images, C distortion types, and L distortion levels, the average SRCC is adopted to quantify the ranking consistency

$$L_s = \frac{1}{JC}\sum_{i=1}^{J}\sum_{j=1}^{C}\mathrm{SRCC}(\mathbf{l}_{ij}, \mathbf{s}_{ij})\,, \tag{22}$$

  where $\mathbf{l}_{ij}$ and $\mathbf{s}_{ij}$ indicate distortion levels and model predictions to images that are generated from the i-th source image by applying the j-th distortion type.
- Pairwise preference consistency test (P-test) [29]: It builds upon the notion of DIP, which consists of two images whose perceptual quality is discriminable. Given a database with Q DIPs, where a BIQA model correctly predicts the concordance of Q_c DIPs, the pairwise preference consistency ratio is computed by

$$P = \frac{Q_c}{Q}\,. \tag{23}$$

  P lies in $[0, 1]$ with a higher value indicating better performance.

SRCC and PLCC are standard evaluation criteria adopted by the video quality experts group (VQEG) [54]. We apply them to LIVE [52], CSIQ [53], and TID2013 [12]. The other three tests are introduced by Ma *et al.* [29] to account for large-scale image databases without MOSs, such as the Waterloo Exploration Database [29] used in the paper.

B. Experimental Results

1) Results on Four Distortions: We compare MEON with classic and state-of-the-art BIQA models on four common distortion types in LIVE [52], CSIQ [53], TID2013 [12], and the Waterloo Exploration Database [29]. The competing algorithms are chosen to cover a diversity of design philosophies, including three classic ones: DIIVINE [3], BRISQUE [4] and CORNIA [38], and five state-of-the-art ones: ILNIQE [40], BLISS [22], HOSA [41], dipIQ [23] and deepIQA [17]. All implementations except BLISS [22] are obtained from the authors. We implement our own version of BLISS and train it on the dataset used for pre-training MEON. In order to make a fair comparison, all models are re-trained/validated on the full LIVE database and tested on CSIQ, TID2013, and the Exploration database. As for MEON, we randomly select 23 reference and their corresponding distorted images in LIVE for training, and leave the rest 6 reference and their distorted images for validation. The model parameters with the lowest validation loss are chosen. When testing, we follow the common practice of Mittal *et al.* [4] and Ye *et al.* [22], and randomly choose 80% reference images along with their corresponding distorted images to estimate the parameters $\{\eta_i | i = 1, 2, 3, 4\}$ of a nonlinear function $\tilde{q} = (\eta_1 - \eta_2)/(1 + \exp(-(\hat{q} - \eta_3)/|\eta_4|)) + \eta_2$, which is used to map model predictions to the MOS scale. The rest 20% images are left out for testing. This procedure is repeated $1,000$ times and the median SRCC and PLCC values are reported.

TABLE IV
THE D-TEST, L-TEST, AND P-TEST RESULTS ON THE WATERLOO EXPLORATION DATABASE [29]

	D-test	L-test	P-test
DIIVINE [3]	0.8538	0.8908	0.9540
BRISQUE [4]	0.9204	0.9772	0.9930
CORNIA [38]	0.9290	0.9764	0.9947
ILNIQE [40]	0.9084	**0.9926**	0.9927
BLISS [22]	0.9080	0.9801	**0.9996**
HOSA [41]	0.9175	0.9647	0.9983
dipIQ [23]	**0.9346**	**0.9846**	**0.9999**
deepIQA [17]	0.9074	0.9467	0.9628
MEON	**0.9384**	0.9669	0.9984

Tables II, III, and IV show the results on CSIQ [53],

TABLE V
THE CONFUSION MATRICES PRODUCED BY MEON ON CSIQ, TID2013, AND THE EXPLORATION DATABASE. THE COLUMN AND THE RAW CONTAIN GROUND TRUTHS AND PREDICTED DISTORTION TYPES, RESPECTIVELY

Accuracy		JP2K	JPEG	WN	BLUR	Pristine
CSIQ	JP2K	**0.847**	0.007	0.000	0.093	0.053
	JPEG	0.040	**0.820**	0.000	0.027	0.113
	WN	0.000	0.000	**0.947**	0.013	0.040
	BLUR	0.067	0.006	0.000	**0.827**	0.100
	Pristine	0.067	0.000	0.100	0.166	**0.667**
TID2013	JP2K	**0.944**	0.016	0.000	0.040	0.000
	JPEG	0.032	**0.968**	0.000	0.000	0.000
	WN	0.000	0.000	**1.000**	0.000	0.000
	BLUR	0.088	0.008	0.000	**0.848**	0.056
	Pristine	0.160	0.000	0.040	0.000	**0.800**
Exploration	JP2K	**0.985**	0.000	0.000	0.015	0.000
	JPEG	0.006	**0.994**	0.000	0.000	0.000
	WN	0.000	0.000	**1.000**	0.000	0.000
	BLUR	0.003	0.000	0.000	**0.997**	0.000
	Pristine	0.213	0.050	0.067	0.234	**0.436**

TID2013 [12], and the Exploration database [29], respectively, from which the key observations are as follows. First, MEON achieves state-of-the-art performance on all three databases. Although there is slight performance bias towards JPEG and WN, MEON aligns all distortions pretty well across the perceptual space. Second, MEON significantly outperforms DIIVINE [3], an improved version of BIQI [24] with more advanced NSS. The performance improvement is largely due to the jointly end-to-end optimization for feature and multi-task learning. Third, MEON performs the best in D-test on the Exploration database, which is no surprise because we are optimizing a finer-grained version of D-test through Subtask I. More specifically, the network learns not only to classify the image into pristine and distorted classes but also to identify the specific distortion type when distorted. Fourth, we observe stronger generalizability of MEON on the Exploration database compared with another DNN-based method, deepIQA [17]. We believe the performance improvement arises because 1) the proposed novel learning framework has the quality prediction subtask regularized by the distortion identification subtask; 2) images instead of patches are used as inputs to reduce the label noise; 3) the pre-training step enables the network to start from a more task-relevant initialization, resulting in a better local optimum.

As a by-product, MEON outputs the distortion information of a test image, whose accuracy on CSIQ [53], TID2013 [12], and the Exploration database [29] is shown in Table V. Empirical justifications for the correlation of the two subtasks can be easily seen, where a lower classification error of a particular distortion generally leads to better quality prediction performance on that distortion and vice versa (*e.g.*, WN and BLUR). Since the statistical behaviors of WN have obvious distinctions with the other three distortions, MEON predicts WN nearly perfectly. On the other hand, it confounds JP2K with BLUR sometimes because JP2K often introduces significant blur at low bit rates. When the distortion level is mild, MEON occasionally labels distorted images as pristine, which is not surprising because the HVS is also easily fooled by such cases. Finally, there is still much room for improvement of correctly classifying pristine images. We conjecture that adding more training data in the pre-training step may help improve the results.

Moreover, we let MEON play the gMAD competition game [2] with deepIQA [17]. Instead of attesting a computational model for a perceptual quantity, the MAximum Differentiation (MAD) competition [55] method works by falsifying it, which has the capability to minimize the number of testing stimuli because essentially even one counter-example is sufficient to disprove a model. gMAD extends the idea by allowing a group of models for competition and by finding the optimal stimuli in a large database [2]. We choose the Exploration database [29] as the playground. An image pair is automatically searched for the maximum quality difference in terms of MEON, while keeping deepIQA [17] predictions at the same quality level. The procedure is then repeated with the roles of the two models exchanged. Four such image pairs are shown in Fig. 5 (a)-(d), where MEON considers pairs (a) and (b) of the same quality at low- and high-quality levels, respectively, which is in close agreement with our visual observations. By contrast, deepIQA incorrectly predicts the top images of (a) and (b) to have much better quality than that of the bottom images. Similar conclusions can be drawn by examining pairs (c) and (d), where the roles of the two models are reversed. The results of gMAD provide strong evidence that the generalizability of MEON is significantly improved over deepIQA [17]. We further compare MEON through gMAD with MS-SSIM [30], an FR-IQA model that performs the best among 16 IQA models according to a recent subjective experiment [2]. Fig. 6 (a)-(d) show the results, from which we observe that MEON is highly competitive against MS-SSIM [30] in the sense that both methods are able to fail each other by successfully finding strong counter-examples. Specifically, MS-SSIM [30] tends to over-penalize WN but under-penalize BLUR. MEON is able to reveal such weaknesses of MS-SSIM, which can be easily discerned in the bottom images of Fig. 6 (c) and (d). On the other hand, MS-SSIM takes advantage of the fact that MEON does not handle BLUR and JP2K well enough and finds counter-examples from those distortions.

2) Results on More Distortion Types: We investigate the scalability of our multi-task learning framework to handle more distortion types by training and testing on the full TID2013 database [12]. For pre-training, we make our best effort to reproduce 15 out of the 24 distortions in TID2013 and apply them to the 840 high-quality images. As a result, only parameters of the shared layers $\mathbf{W}$ are provided with meaningful initializations. Since BLISS [22] and dipIQ [23] cannot be trained without all distorted images available, we exclude them from the comparison. For joint optimization, we follow Bosse *et al.* [17] and use 15, 5, and 5 reference and their corresponding distorted images for training, validation, and testing, respectively. Median SRCC results are reported based on 10 random splits in Table VI. All other competing BIQA models except deepIQA [17] are re-trained, validated, and tested in exactly the same way. Since the training codes of deepIQA are not available, we copy the results from the

TABLE VI
MEDIAN SRCC RESULTS ACROSS 10 SESSIONS ON THE FULL TID2013 DATABASE

SRCC	#01	#02	#03	#04	#05	#06	#07	#08	#09	#10	#11	#12	#13
DIIVINE [3]	0.756	0.464	0.869	0.374	0.794	0.704	0.650	**0.900**	0.814	0.795	0.804	0.514	**0.892**
BRISQUE [4]	0.674	0.550	0.804	0.222	0.824	0.749	0.677	0.855	0.492	0.751	0.696	0.285	0.719
CORNIA [38]	0.496	0.130	0.655	0.373	0.715	0.647	0.632	0.844	0.688	0.758	0.866	0.587	0.603
ILNIQE [40]	**0.924**	**0.847**	**0.947**	**0.786**	**0.908**	**0.847**	**0.933**	0.869	**0.846**	**0.901**	**0.930**	0.400	0.708
HOSA [41]	**0.833**	0.575	0.808	0.432	0.906	0.817	0.783	**0.903**	**0.873**	**0.903**	**0.920**	**0.712**	**0.743**
deepIQA [17]	—	—	—	—	—	—	—	—	—	—	—	—	—
MEON	0.813	**0.722**	**0.926**	**0.728**	**0.911**	**0.901**	**0.888**	0.887	0.797	0.850	0.891	**0.746**	0.716

SRCC	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24	All
DIIVINE [3]	**0.215**	**0.389**	0.124	0.189	0.280	0.691	0.340	0.690	0.769	0.700	0.795	0.632
BRISQUE [4]	0.158	0.362	0.253	0.102	0.200	0.587	0.211	0.546	**0.842**	0.770	0.764	0.572
CORNIA [38]	**0.282**	-0.025	0.194	0.145	-0.006	0.461	**0.560**	0.648	0.646	0.672	0.867	0.611
ILNIQE [40]	-0.173	0.000	**0.328**	0.080	0.103	**0.773**	0.507	**0.911**	0.822	**0.801**	**0.878**	0.534
HOSA [41]	0.143	0.330	**0.279**	**0.307**	**0.414**	0.711	**0.537**	0.756	0.840	**0.821**	**0.903**	0.707
deepIQA [17]	—	—	—	—	—	—	—	—	—	—	—	**0.761**
MEON	0.116	**0.500**	0.177	**0.252**	**0.684**	**0.849**	0.406	**0.772**	**0.857**	0.779	0.855	**0.808**

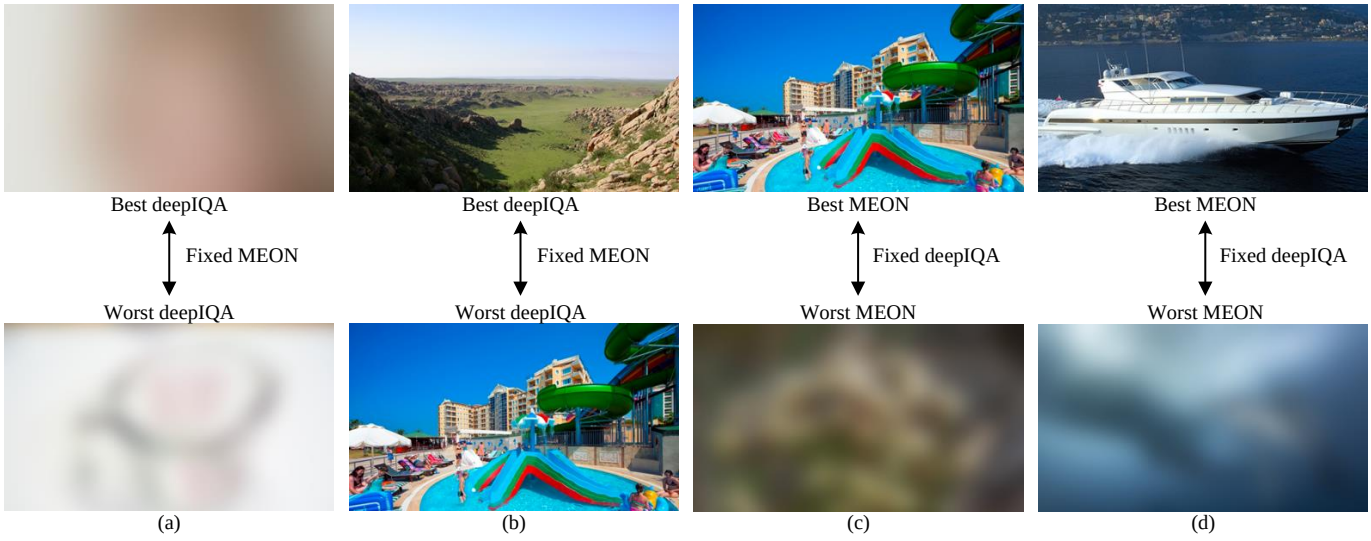


Fig. 5. gMAD competition results between MEON and deepIQA [17]. (a) Fixed MEON at the low-quality level. (b) Fixed MEON at the high-quality level. (c) Fixed deepIQA at the low-quality level. (d) Fixed deepIQA at the high-quality level.

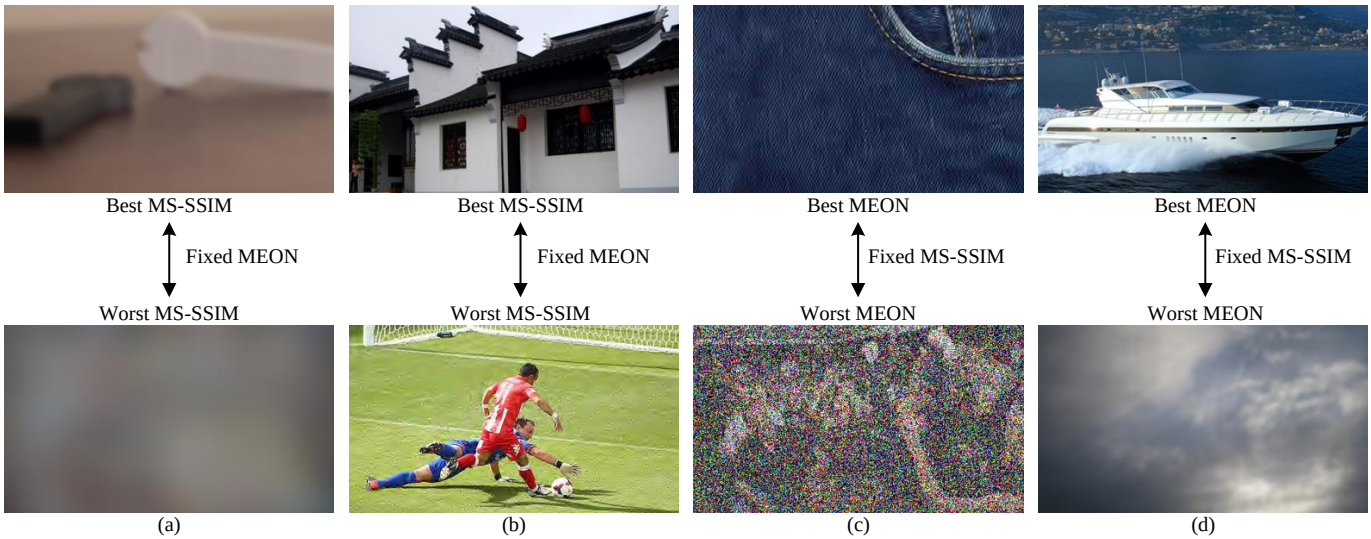


Fig. 6. gMAD competition results between MEON and MS-SSIM [30]. (a) Fixed MEON at the low-quality level. (b) Fixed MEON at the high-quality level. (c) Fixed MS-SSIM at the low-quality level. (d) Fixed MS-SSIM at the high-quality level.

TABLE VII
MEDIAN SRCC RESULTS OF ABLATION EXPERIMENTS ACROSS $1,000$ SESSIONS ON CSIQ [53] AND TID2013 [12]

	CSIQ	TID2013
Single-task	0.844	0.850
Traditional multi-task	0.885	0.871
MEON w/o pre-training	0.894	0.880
MEON with pre-training	**0.932**	**0.912**

TABLE VIII
SRCC RESULTS OF CONFIGURATIONS WITH DIFFERENT ACTIVATION FUNCTIONS AND MODEL COMPLEXITIES

	CSIQ	TID2013
ReLU + single layer	0.922	0.891
ReLU + double layers	0.924	0.900
ReLU + double layers + BN	0.930	**0.918**
MEON (GDN + single layer)	**0.932**	0.912

original paper for reference (note that the random seeds for the 10 data splits may be different).

From Table VI, we observe that MEON outperforms previous BIQA models by a clear margin, aligning 24 distortions in the perceptual space remarkably well. By contrast, although ILNIQE [40] does an excellent job in predicting image quality under the same distortion type, which is also reflected in its superior performance in L-test on the Exploration database, it fails to align distortion types correctly. Moreover, all competing BIQA models including MEON do not perform well on mean shift (#16) and contrast change (#17) cases. This is not surprising for methods that adopt spatial normalization as preprocessing, such as BRISQUE [4], CORNIA [38], ILNIQE [40], and HOSA [41] because the mean and contrast information has been removed at the very beginning. Moreover, mean shift and contrast change may not be considered as distortions at all because modest mean shift may not affect perceptual quality and contrast change (*e.g.*, contrast enhancement) often improves image quality.

3) Ablation Experiments: We conduct a series of ablation experiments to single out the core contributors of MEON. We first train Sub-network II with random initializations as a simple single-task baseline. We also experiment with the traditional multi-task learning framework by directly producing an overall quality score. From Table VII, we observe that without pre-training, MEON achieves the best performance. Moreover, pre-training brings the prediction accuracy to the next level. We conclude that the proposed multi-task learning framework and the pre-training mechanism are keys to the success of MEON.

Next, we analyze the impact of the GDN transform on model complexity and quality prediction performance. We start from a baseline by replacing all GDN layers with ReLU. We then double all convolutional and fully connected layers in both Sub-networks I and II with ReLU nonlinearity to see whether a deeper network improves the performance. Last, we introduce the BN transform right before each ReLU layer. The results are listed in Table VIII, from which we see that simply replacing GDN with ReLU leads to inferior performance. The network with a deeper architecture slightly improves the performance. When combined with BN, it achieves competitive performance against MEON. This suggests that GDN may be an effective way to reduce model complexity without sacrificing performance. Specifically in our case, GDN is able to half the layers and parameters of the network while achieving similar performance when compared with ReLU.

V. CONCLUSION AND DISCUSSION

We propose a novel multi-task learning framework for BIQA, namely MEON, by decomposing the BIQA task into two subtasks with dependent loss functions. We optimize MEON for both distortion identification and quality prediction in an end-to-end fashion. The resulting MEON index demonstrates state-of-the-art performance, which we believe arises from pre-training for better initializations, multi-task learning for mutual regularization, and GDN for biologically inspired feature representations. In addition, we show the scalability of MEON to handle more distortion types and its strong competitiveness against state-of-the-art BIQA approaches in the gMAD competition.

The general idea behind the proposed approach does not limit its application scope to BIQA only. With proper modifications of the MEON network architecture, we may learn end-to-end FR- and RR-IQA networks. Furthermore, such deep learning based IQA networks may be incorporated into other image processing applications. For example, through backpropagation, a DNN-based IQA model may be directly used to drive DNN-based image compression and restoration algorithms.

Another promising future direction is to extend the current work to other problems that involve perceptual attributes of images. For example, in the fields of authentic [56] and aesthetic [57] IQA, we are faced with the same problem of limited training data, which casts great challenges to train DNN without over-fitting. How to extend the idea of the current work to these problems is an interesting direction yet to be explored.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Mingming Gong for fruitful discussions on multi-task learning and Mu Li for insightful comments on efficiently implementing GDN. We thank the NVIDIA Corporation for donating a GPU for this research.

REFERENCES

[1] Z. Wang and A. C. Bovik, "Reduced-and no-reference image quality assessment: The natural scene statistic model approach," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, Nov. 2011.

[2] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, "Group MAD competition − a new methodology to compare objective image quality models," in *IEEE Conference on Computer Vsion and Pattern Recognition*, 2016, pp. 1664–1673.

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

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

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

[6] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 838–842, Jul. 2015.

[7] B. A. Wandell, *Foundations of Vision.* Sinauer Associates, 1995.

[8] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, Mar. 2001.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[12] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, Jan. 2015.

[13] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *CoRR*, vol. abs/1602.05531, 2016.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[15] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[16] ——, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *IEEE International Conference on Image Processing*, 2015, pp. 2791–2795.

[17] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *CoRR*, vol. abs/1612.01697, 2016.

[18] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 3313–3316.

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

[20] K. Ma, H. Fu, T. Liu, Z. Wang, and D. Tao, "Local blur mapping: Exploiting high-level semantics by deep neural networks," *CoRR*, vol. abs/1612.01227, 2016.

[21] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[22] P. Ye, J. Kumar, and D. Doermann, "Beyond human opinion scores: Blind image quality assessment based on synthetic scores," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4241–4248.

[23] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

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

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *IEEE International Conference on Machine Learning*, 2010, pp. 807–814.

[26] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[27] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in *International Conference on Learning Representations*, 2016.

[28] ——, "End-to-end optimized image compression," in *International Conference on Learning Representations*, 2017.

[29] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo Exploration Database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 1004–1016, Feb. 2017.

[30] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, 2003, pp. 1398–1402.

[31] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[32] ——, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[33] Z. Wang, "Objective image quality assessment: Facing the real-world challenges," in *IS&T Electronic Imaging: Image Quality and System Performance*, 2016.

[34] P. Ye, "Feature learning and active learning for image quality assessment," Ph.D. dissertation, University of Maryland, 2014.

[35] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *IEEE International Conference on Image Processing*, 2002, pp. 477–480.

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

[37] Z. Wang and E. P. Simoncelli, "Local phase coherence and the perception of blur," in *Advances in Neural Information Processing Systems*, 2003.

[38] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

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

[40] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

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

[42] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[44] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *CoRR*, vol. abs/1511.07289, 2015.

[45] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, no. 8, pp. 819–825, Aug. 2001.

[46] S. Lyu, "Divisive normalization: Justification and effectiveness as efficient coding transform," in *Advances in Neural Information Processing Systems*, 2010, pp. 1522–1530.

[47] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Visual Neuroscience*, vol. 9, no. 02, pp. 181–197, Aug. 1992.

[48] M. Carandini and D. J. Heeger, "Normalization as a canonical neural computation," *Nature Reviews Neuroscience*, vol. 13, no. 1, pp. 51–62, Jan. 2012.

[49] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *IEEE International Conference on Computer Vision*, 2009, pp. 2146–2153.

[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

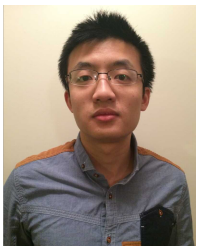
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[52] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, Image and Video Quality Assessment Research at LIVE [Online]. Available: http://live.ece.utexas.edu/research/quality/.

[53] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *SPIE Journal of Electronic Imaging*, vol. 19, no. 1, pp. 1–21, Jan. 2010.

[54] VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment 2000 [Online]. Available: http://www.vqeg.org.

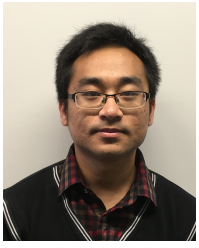
[55] Z. Wang and E. P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *Journal of Vision*, vol. 8, no. 12, pp. 1–13, Sep. 2008.
[56] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, Jan. 2016.
[57] W. Liu and Z. Wang, "A database for perceptual evaluation of image aesthetics," in *IEEE International Conference on Image Processing*, 2017.



Kede Ma (S'13) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2012. He then received the M.S. and Ph.D. degrees in electrical and computer engineering from University of Waterloo, ON, Canada, in 2014 and 2017, respectively. He will do a posdoc in the Center for Neural Science at New York University. His research interests lie in perceptual image processing, computational vision, and computational photography.

Wentao Liu (S'15) received the B.E. and the M.E. degrees from Tsinghua University, Beijing, China in 2011 and 2014, respectively. He is currently working toward the Ph.D. degree in the Electrical & Computer Engineering Department, University of Waterloo, ON, Canada. His research interests include perceptual quality assessment of images and videos.

Kai Zhang received the M.Sc. degree in applied mathematics from China Jiliang University, Hangzhou, China, in 2014. He is currently pursuing the Ph.D. degree in computer science and technology at Harbin Institute of Technology, Harbin, China, under the supervision of Prof. Wangmeng Zuo and Prof. Lei Zhang. His research interests include machine learning and image processing.



Zhengfang Duanmu (S'15) received the B.A.Sc. and the M.A.Sc. degrees in electrical and computer engineering from the University of Waterloo in 2015 and 2017, respectively, where he is currently working toward the Ph.D. degree in electrical and computer engineering. His research interests lie in perceptual image processing and quality of experience.

Zhou Wang (S'99-M'02-SM'12-F'14) received the Ph.D. degree from The University of Texas at Austin in 2001. He is currently a Professor in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include image processing, coding, and quality assessment; computational vision and pattern analysis; multimedia communications; and biomedical signal processing. He has more than 100 publications in these fields with over 30,000 citations (Google Scholar).

Dr. Wang serves as a Senior Area Editor of IEEE Transactions on Image Processing (2015-present), and an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology (2016-present). Previously, he served as a member of IEEE Multimedia Signal Processing Technical Committee (2013-2015), an Associate Editor of IEEE Transactions on Image Processing (2009-2014), Pattern Recognition (2006-present) and IEEE Signal Processing Letters (2006-2010), and a Guest Editor of IEEE Journal of Selected Topics in Signal Processing (2013-2014 and 2007-2009). He is a Fellow of Canadian Academy of Engineering, and a recipient of 2016 IEEE Signal Processing Society Sustained Impact Paper Award, 2015 Primetime Engineering Emmy Award, 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, 2013 IEEE Signal Processing Magazine Best Paper Award, 2009 IEEE Signal Processing Society Best Paper Award, and 2009 Ontario Early Researcher Award.

Wangmeng Zuo received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. He is currently a Professor in the School of Computer Science and Technology, Harbin Institute of Technology. His current research interests include image enhancement and restoration, object detection, visual tracking, and image classification. He has published over 60 papers in top-tier academic journals and conferences. He has served as a Tutorial Organizer in ECCV 2016, an Associate Editor of the *IET Biometrics* and *Journal of Electronic Imaging*, and the Guest Editor of *Neurocomputing*, *Pattern Recognition*, *IEEE Transactions on Circuits and Systems for Video Technology*, and *IEEE Transactions on Neural Networks and Learning Systems*.