dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs

Kede Ma, *Student Member, IEEE*, Wentao Liu, *Student Member, IEEE*, Tongliang Liu, Zhou Wang, *Fellow, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Objective assessment of image quality is fundamentally important in many image processing tasks. In this paper, we focus on learning blind image quality assessment (BIQA) models, which predict the quality of a digital image with no access to its original pristine-quality counterpart as reference. One of the biggest challenges in learning BIQA models is the conflict between the gigantic image space (which is in the dimension of the number of image pixels) and the extremely limited reliable ground truth data for training. Such data are typically collected via subjective testing, which is cumbersome, slow, and expensive. Here, we first show that a vast amount of reliable training data in the form of quality-discriminable image pairs (DIPs) can be obtained automatically at low cost by exploiting large-scale databases with diverse image content. We then learn an opinion-unaware BIQA (OU-BIQA, meaning that no subjective opinions are used for training) model using RankNet, a pairwise learning-to-rank (L2R) algorithm, from millions of DIPs, each associated with a perceptual uncertainty level, leading to a DIP inferred quality (dipIQ) index. Extensive experiments on four benchmark IQA databases demonstrate that dipIQ outperforms the state-of-the-art OU-BIQA models. The robustness of dipIQ is also significantly improved as confirmed by the group MAximum Differentiation competition method. Furthermore, we extend the proposed framework by learning models with ListNet (a listwise L2R algorithm) on quality-discriminable image lists (DIL). The resulting DIL inferred quality index achieves an additional performance gain.**

***Index Terms*— Blind image quality assessment (BIQA), learning-to-rank (L2R), dipIQ, RankNet, quality-discriminable image pair (DIP), gMAD.**

I. Introduction

OBJECTIVELY assessing image quality is of fundamental importance due in part to the massive expansion of online image volume. Objective image quality assessment (IQA) has become an active research topic over the last decade, with a large variety of IQA models proposed [1], [2]. They can be categorized into full-reference models (FR, where the reference image is fully available when evaluating a distorted image) [3], reduced-reference models (RR, where only partial information about the reference image is available) [4], and blind/no-reference models (NR, where the reference image is not accessible) [5]. In many real-world applications, reference images are unavailable, making blind IQA (BIQA) models highly desirable in practice.

Many BIQA models are developed by supervised learning [6]–[14] and share a common two-stage structure: 1) perception- and/or distortion-relevant features (denoted by $\mathbf{x}$) are extracted from the test image; and 2) a quality prediction function $f(\mathbf{x})$ is learned by statistical machine learning algorithms. The performance and robustness of these approaches rely heavily on the quality and quantity of the ground truth data for training. The most common type of ground truth data is in the form of the mean opinion score (MOS), which is the average of quality ratings given by multiple subjects. Therefore, these models are often referred to as opinion-aware BIQA (OA-BIQA) models and may incur the following drawbacks. First, collecting MOS via subjective testing is slow, cumbersome, and expensive. As a result, even the largest publicly available IQA database, TID2013 [15], provides only $3,000$ images with MOSs. This limited number of training images is deemed extremely sparsely distributed in the entire image space, whose dimension equals the number of pixels and is typically in the order of millions. As such, the generalizability of BIQA models learned from small training samples is questionable on real-world images. Second, among thousands of sample images, only a few dozen source reference images can be included, considering the combinations of reference images, distortion types and levels. For example, the TID2013 database [15] includes 25 source images only. It is extremely unlikely that this limited number of reference images sufficiently represent the variations that exist in real-world images. Third, since these BIQA models are trained with individual images to make independent quality predictions, the cost function is blind to the relative perceptual order between images. As a result, the learned models are weak at ordering images with respect to their perceptual quality.

In this paper, we show that a vast amount of reliable training data in the form of so-called quality-discriminable image pairs (DIP) can be generated by exploiting large-scale databases with diverse image content. Each DIP is associated with a perceptual uncertainty measure to indicate the confidence

Manuscript received August 7, 2016; revised February 11, 2017; accepted May 16, 2017. Date of publication May 26, 2017; date of current version June 13, 2017. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, and the Australian Research Council Projects FT-130101457, DP-140102164, and LP-150100671. K. Ma was partially supported by the CSC. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kalpana Seshadrinathan. *(Corresponding author: Kede Ma.)*

K. Ma, W. Liu, and Z. Wang are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: k29ma@uwaterloo.ca; w238liu@uwaterloo.ca; zhou.wang@uwaterloo.ca).

T. Liu and D. Tao are with the UBTech Sydney Artificial Intelligence Institute, The University of Sydney, Darlington, NSW 2008, Australia, and also with the Faculty of Engineering and Information Technologies, School of Information Technologies, The University of Sydney, Darlington, NSW 2008, Australia (e-mail: tongliang.liu@sydney.edu.au; dacheng.tao@sydney.edu.au).

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/TIP.2017.2708503

level of its quality discriminability. We show that such DIPs can be generated at very low cost without resorting to subjective testing. We then employ RankNet [16], a neural network-based pairwise learning-to-rank (L2R) algorithm [17], [18], to learn an opinion-unaware BIQA (OU-BIQA, meaning that no subjective opinions are used for training) model by incorporating the uncertainty measure into the loss function. Extensive experiments on four benchmark IQA databases demonstrate that the DIP inferred quality (dipIQ) indices significantly outperform previous OU-BIQA models. We also conduct another set of experiments in which we train the dipIQ indices using different feature representations as inputs and compare them with OA-BIQA models using the same representations. The generalizability and robustness of dipIQ are improved across all four IQA databases and verified by the group MAximum Differentiation (gMAD) competition method [19], which examines image pairs optimally selected from the Waterloo Exploration Database [20]. Furthermore, we extend the proposed pairwise L2R approach for OU-BIQA to a listwise L2R one by evoking ListNet [21] (a listwise L2R extension of RankNet [16]) and transforming DIPs to quality-discriminable image lists (DIL) for training. The resulting DIL inferred quality (dilIQ) index leads to an additional performance gain.

The remainder of the paper is organized as follows. BIQA models and typical L2R algorithms are reviewed and categorized in Section II. The proposed dipIQ approach is introduced in Section III. Experimental results using dipIQ on four benchmark IQA databases compared with state-of-the-art BIQA models are presented in Section IV, followed by an extension to the dilIQ model in Section V. We conclude the paper in Section VI.

II. Related Work

We first review existing BIQA models according to their two-stage structure: feature extraction and quality prediction model learning. We then review typical L2R algorithms. Details of RankNet [16] are provided in Section III.

A. Existing BIQA Models

From the feature extraction point of view, three types of knowledge can be exploited to craft useful features for BIQA. The first is knowledge about our visual world that summarizes the statistical regularities of undistorted images. The second is knowledge about degradation, which can then be explicitly taken into account to build features for particular artifacts, such as blocking [22]–[24], blurring [25]–[27] and ringing [28]–[30]. The third is knowledge of the human visual system (HVS) [31], namely perceptual models derived from visual physiological and psychophysical studies [32]–[35]. Natural scene statistics (NSS), which seek to capture the natural statistical behavior of images, embody the three-fold modeling in a rather elegant way [5]. NSS can be extracted directly in the spatial domain or in transform domains such as DFT, DCT, and wavelets [36], [37].

In the spatial domain, edges are presumably the most important image features. The edge spread can be used to detect blurring [38], [39], and the intensity variance in smooth regions close to edges can indicate ringing artifacts [28]. Step edge detectors that operate at 8×8 block boundaries measure the severity of discontinuities caused by JPEG compression [22]. The sample entropy of intensity histograms is used to identify image anisotropy [40], [41]. The responses of image gradients and the Laplacian of Gaussian operators are jointly modeled to describe the destruction of statistical naturalness of images [12]. The singular value decomposition of local image gradient matrices may provide a quantitative measure of image content [42]. Mean-subtracted and contrast-normalized pixel value statistics have also been modeled using a generalized Gaussian distribution (GGD) [8], [43]–[45], inspired by the adaptive gain control mechanism seen in neurons [33].

Statistical modeling in the wavelet domain resembles the early visual system [32], and natural images exhibit statistical regularities in the wavelet space. Specifically, it is widely acknowledged that the marginal distribution of wavelet coefficients of a natural image (regardless of content) has a sharp peak near zero and heavier than Gaussian tails. Therefore, statistics of raw [4], [6], [46], [47] and normalized [48], [49] wavelet coefficients, and wavelet coefficient correlations in the neighborhood [10], [29], [50]–[52] can be individually or jointly modeled as image naturalness measurements. The phase information of wavelet coefficients, for example expressed as the local phase coherence, is exploited to describe the perception of blur [26] and sharpness [53].

In the DFT domain, blur kernels can be efficiently estimated [50], [51], [54] to quantify the degree of image blurring. The regular peaks at feature frequencies can be used to identity blocking artifacts [23], [55]. Moreover, it is generally hypothesized that most perceptual information in an image is stored in the Fourier phase rather than the Fourier amplitude [56], [57]. Phase congruency [58] is such a feature that identifies perceptually significant image features at spatial locations where Fourier components are maximally in-phase [40].

In the DCT domain, blocking artifacts can be identified in a shifted 8×8 block [24]. The ratio of AC coefficients to DC components can be interpreted as a measure of local contrast [59]. The kurtosis of AC coefficients can be used to quantify the structure statistics. In addition, AC coefficients can also be jointly modeled using a GGD [7].

There is a growing interest in learning features for BIQA. Ye *et al.* learned quality filters on image patches using K-means clustering and adopted filter responses as features [9]. They then took one step further by supervised filter learning [45]. Xue *et al.* [60] proposed a quality-aware clustering scheme on the high frequencies of raw patches, guided by an FR-IQA measure [61]. Kang *et al.* investigated a convolutional neural network to jointly learn features and nonlinear mappings for BIQA [62].

From the model learning perspective, SVR [63], [64] is the most commonly used tool to learn $f(\mathbf{x})$ for BIQA [6], [9], [10], [12], [45], [52]. The capabilities of neural networks to pre-train a model without labels and to easily scale up have also been exploited for this purpose [40], [47], [51], [62]. Another typical quality regression is the example-based

method, which predicts the test image quality score using the weighted average of training image quality scores, where the weight encodes the perceptual similarity between the test and training images [14], [52], [60]. Saad *et al.* [7], [59] jointly modeled $\mathbf{x}$ and MOS using a multivariate Gaussian distribution and performed prediction by maximizing the conditional probability $P(\mathbf{x}|\text{MOS})$. Similar probabilistic modeling strategies have been investigated [43], [65]. Pairwise L2R algorithms have also been used to learn BIQA models [66], [67]. However, in these methods, DIP generation relies solely on MOS availability, which limits the number of DIPs produced. Moreover, their performance is inferior to that of existing BIQA methods. Other advanced learning algorithms include topic modeling [68], Gaussian process [51], and multi-kernel learning [67], [69].

B. Existing L2R Algorithms

Existing L2R algorithms can be broadly classified into three categories based on the training data format and loss function: pointwise, pairwise, and listwise approaches. An excellent survey of L2R algorithms can be found in [17]. Here we only provide a brief overview.

Pointwise approaches assume that each instance's importance degree is known. The loss function usually examines the prediction accuracy of each individual instance. In an early attempt on L2R, Fuhr [70] adopted a linear regression with a polynomial feature expansion to learn the score function $f(\mathbf{x})$. Cossock and Zhang [71] utilized a similar formulation with some theoretical justifications for the use of the least squares loss function. Nallapati [72] formulated L2R as a classification problem and investigated the use of maximum entropy and support vector machines (SVMs) to classify each instance into two classes—relevant or irrelevant. Ordinal regression-based pointwise L2R algorithms have also been proposed such as PRanking [73] and SVM-based large margin principles [74].

Pairwise approaches assume that the relative order between two instances is known or can be inferred from other ground truth formats. The goal is to minimize the number of misclassified instance pairs. In the extreme case, if all instance pairs are correctly classified, they will be correctly ranked [17]. In RankSVM [75], Joachims creatively generated training pairs from clickthrough data and reformulated SVM to learn the score function $f(\mathbf{x})$ from instance pairs. Proposed in 2005, RankNet [16] was probably the first L2R algorithm used by commercial search engines, which had a typical neural network with a weight-sharing scheme forming its skeleton. Tsai *et al.* [76] replaced RankNet's loss function [16] with a fidelity loss originating from quantum physics. In this paper, RankNet is adopted as the default pairwise L2R algorithm to learn OU-BIQA models for reasons that will be described later. RankBoost [77] is another well-known pairwise L2R algorithm based on AdaBoost [78] with an exponential loss.

Listwise approaches provide the opportunity to directly optimize ranking performance criteria [17]. Representative algorithms include SoftRank [79], SVM^{map} [80], and RankGP [81]. Another subset of listwise approaches choose to optimize listwise ranking losses. For example, as a direct extension of RankNet, ListNet [21] duplicates RankNet's structure to accommodate an instance list as input and optimizes a ranking loss based on the permutation probability distribution [21]. In this paper, we also employ ListNet to learn OU-BIQA models as an extension of the proposed pairwise L2R approach.

III. Proposed Pairwise L2R Approach for OU-BIQA

In this section, we elaborate the proposed pairwise L2R approach to learn OU-BIQA models. First, we propose an automatic DIP generation engine. Each DIP is associated with an uncertainty measure to quantify the confidence level of its quality discriminability. Second, we detail RankNet [16] and extend its capability to learn from the generated DIPs with uncertainty.

A. DIP Generation

Our automatic DIP generation engine is described as follows. We first choose three best-trusted FR-IQA models, namely MS-SSIM [82], VIF [83], and GSMD [84]. A logistic nonlinear function suggested in [85] is adopted to map predictions of the three models to the MOS scale of the LIVE database [86]. After that, the score range of the three models roughly spans $[0, 100]$, where higher values indicate better perceptual quality. We associate each candidate image pair with a nonnegative T, which is equal to the smallest score difference of the three FR models. Intuitively, the perceptual uncertainty level of quality discriminability should decrease monotonically with the increase of T. By varying T, we can generate DIPs with different uncertainty levels. To quantify the level of uncertainty, we employ a raised-cosine function given by

$$U(T) = \begin{cases} \frac{1}{2}\left(1 + \cos\left(\frac{\pi T}{T_c}\right)\right) & \text{if } T \le T_c \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where $U(T)$ lies in $[0, 1]$, with a higher value indicating a greater degree of uncertainty and T_c is a constant, above which the uncertainty goes to zero. In the current implementation, we set $T_c = 20$, whose legitimacy can be validated from two sources. First, the average standard deviation of MOSs on LIVE is around 9, which is approximately half of T_c, therefore guaranteeing the perceived discriminability of two images. Second, based on the subjective experiments conducted by Gao *et al.* [67] on LIVE, the consistency between subjects on the relative quality of one pair increases with the absolute difference and, when it is larger than 20, the consistency approaches 100%. Fig. 1 shows the shape of the uncertainty function as a function of T and some representative DIPs, where the left images have better quality in terms of the three chosen FR-IQA models with $T > 0$. All the shown DIPs are generated from the training image set that will be described later. It is clear that setting T close to zero produces the highest level of uncertainty of quality discriminability. Careful inspection of Fig. 1(a) and Fig. 1(b) reveals that the uncertainty manifests itself in two ways. First, the right

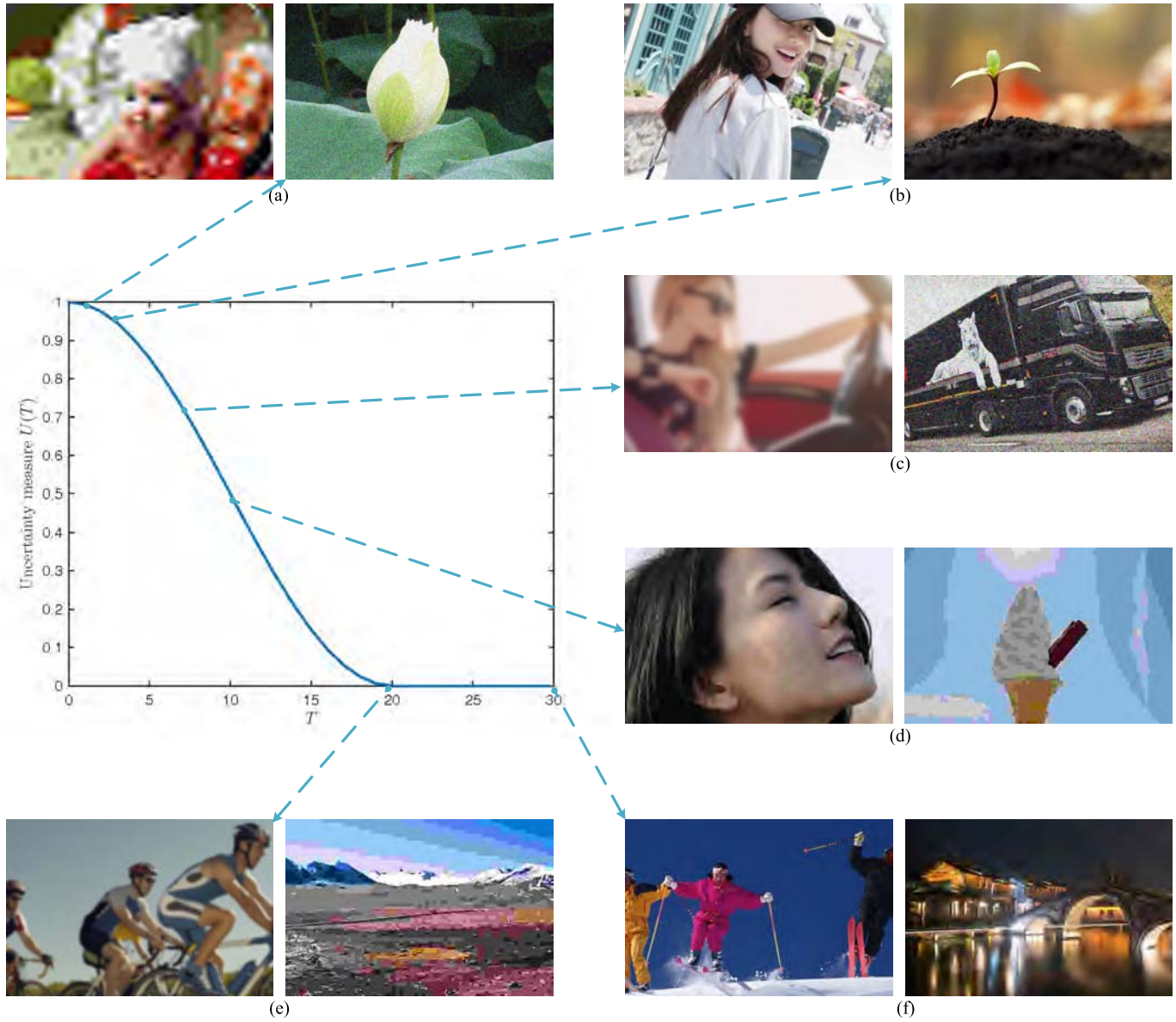


Fig. 1. Illustration of the perceptual uncertainty of quality discriminability of DIPs as a function of T. The left images of all DIPs have better quality in terms of the three FR-IQA models with $T > 0$. However, the quality discriminability differs significantly. All images are originated from the 700 training images and cropped for better visibility. (a)–(f) DIPs with different levels of uncertainty.

image in Fig. 1(a) has better perceived quality to many human observers compared with the left one, which disagrees with the three FR-IQA models. Second, both images in Fig. 1(b) have distortions that are barely perceived by the human eye. In other words, they have very similar perceptual quality. The perceptual uncertainty generally decreases if T increases and when $T > 20$, the DIP is clearly discriminable, further justifying the selection of $T_c = 20$.

B. RankNet [16]

Given a number of DIPs, a pairwise L2R algorithm would make use of their perceptual order to learn quality models while taking the inherent perceptual uncertainty into account. Here, we revisit RankNet [16], a pairwise L2R algorithm that was the first of its kind used by commercial search engines [17]. We extend it to learn from DIPs associated with uncertainty. Fig. 2 shows RankNet's architecture, which is based on classical neural networks and has two parallel streams to accommodate a pair of inputs. The two-stream weights are shared, which is achieved by using the same initializations and the same gradients during backpropagation [16]. The quality prediction function $f(\mathbf{x})$, namely the dipIQ index, is implemented by one of the streams, and the loss function is defined on a pair of images with the help of f. Specifically, let $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ be the output of the first and second streams, whose difference is converted to a probability using

$$P_{ij}(f) = \frac{\exp\big(f(\mathbf{x}_i) - f(\mathbf{x}_j)\big)}{1 + \exp\big(f(\mathbf{x}_i) - f(\mathbf{x}_j)\big)}, \tag{2}$$

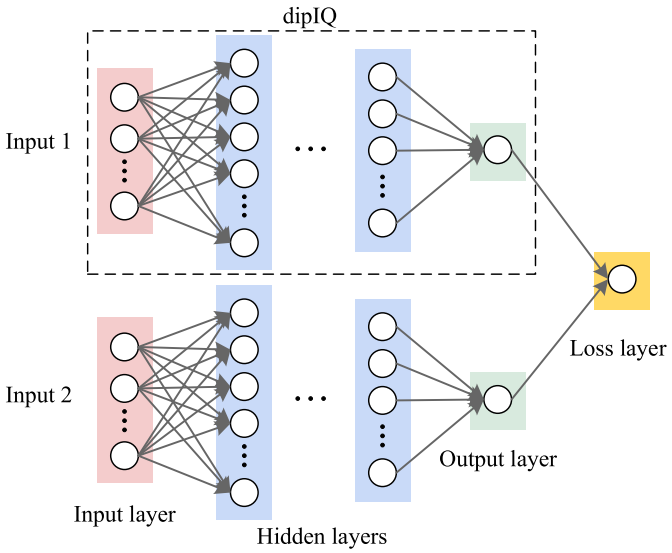


Fig. 2. The architecture of dipIQ based on RankNet [16]. (a)–(f) DIPs with different levels of uncertainty.

based on which we define the cross entropy loss as

$$\begin{aligned} L(f;\mathbf{x}_i,\mathbf{x}_j,\bar{P}_{ij}) &= -\bar{P}_{ij}\log P_{ij} - (1-\bar{P}_{ij})\log(1-P_{ij}) \\ &= -\bar{P}_{ij}\left(f(\mathbf{x}_i)-f(\mathbf{x}_j)\right) \\ &\quad +\log\left(1+\exp\left(f(\mathbf{x}_i)-f(\mathbf{x}_j)\right)\right), \end{aligned} \tag{3}$$

where $\bar{P}_{ij}$ is the ground truth label associated with the training pair, consisting of the i-th and j-th images. In the case of DIPs described in the Section III-A, $\bar{P}_{ij}$ is always 0 or 1, indicating that the quality of the i-th image is worse or better than the j-th one. Within the mini-batch stochastic gradient minimization framework, we define the batch-level loss function using the perceptual uncertainty of each DIP as a weighting factor

$$L_b(f) = \sum_{\langle i,j\rangle\in\mathcal{B}} (1-U_{ij})L(f;\mathbf{x}_i,\mathbf{x}_j,\bar{P}_{ij}), \tag{4}$$

where $\mathcal{B}$ is the batch containing the DIP indices currently being trained. As Eq. (4) makes clear, DIPs with higher uncertainty contribute less to the overall loss. With some derivations, we obtain the gradient of L_b with respect to the model parameters collectively denoted by $\mathbf{w}$ as follows

$$\begin{aligned} \frac{\partial L_b(f)}{\partial \mathbf{w}} &= \sum_{\langle i,j\rangle\in\mathcal{B}} \left(-\bar{P}_{ij} + \frac{\exp\left(f(\mathbf{x}_i)-f(\mathbf{x}_j)\right)}{1+\exp\left(f(\mathbf{x}_i)-f(\mathbf{x}_j)\right)}\right) \\ &\quad \times \left(1-U_{ij}\right)\left(\frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}} - \frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}}\right). \end{aligned} \tag{5}$$

In the case of a linear dipIQ containing no hidden layers and no nonlinear activations, Eq. (3) is reduced to

$$\begin{aligned} L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j,\bar{P}_{ij}) &= -\bar{P}_{ij}\left(\mathbf{w}^T(\mathbf{x}_i-\mathbf{x}_j)\right) \\ &\quad +\log(1+\exp(\mathbf{w}^T(\mathbf{x}_i-\mathbf{x}_j)), \end{aligned} \tag{6}$$

which is easily recognized as logistic regression. The convexity of Eq. (6) ensures the global optimality of the solution. We investigate both linear and nonlinear dipIQ cases with the cross entropy as loss. In fact, any probability distribution measures can be adopted as alternatives. For example, Tsai *et al.* [76] proposed a fidelity loss measure from quantum physics. We find in our experiments that the fidelity loss impairs performance, so we use the cross entropy loss throughout the paper.

We select RankNet [16] as our first choice of pairwise L2R algorithm for two reasons. First, it is capable of handling a large number (millions) of training samples using stochastic or mini-batch gradient descent algorithms. By contrast, the training of other pairwise L2R methods such as RankSVM [75], even with a linear kernel, is painfully slow. Second, since RankNet [16] embodies classical neural network architectures, we embrace the latest advances in training deep neural networks [87], [88] and can easily upscale the network by adding more hidden layers to learn powerful nonlinear quality prediction functions.

IV. Experiments

In this section, we first provide thorough implementation details of RankNet [16] to learn OU-BIQA models. We then describe the experimental protocol based on which a fair comparison is conducted between dipIQ and state-of-the-art BIQA models. After that, we discuss how to extend the proposed pairwise L2R approach for OU-BIQA to a listwise one that could possibly boost the performance.

A. Implementation Details

1) Training Set Construction: We collect 840 high quality and high resolution natural images to represent scenes we see in the real world. They can be roughly clustered into seven groups: human, animal, plant, landscape, cityscape, still-life, and transportation. Sample source images are shown in Fig. 3. We preprocess each source image by down-sampling it using a bicubic kernel so that the maximum height or width is 768. Following the procedures described in [19], we add four distortion types, namely JPEG and JPEG2000 (JP2K) compression, white Gaussian noise contamination (WN), and Gaussian blur (BLUR), each with five distortion levels. As a result, our training set consists of 17,640 test images, with 840 source and 16,800 distorted images. We randomly hold out 140 source images and their corresponding distorted images and use them as the validation set. For the rest 14,700 images, we adopt the proposed DIP generation engine to produce more than 80 million DIPs, which constitute our training set.

2) Base Feature: We adopt CORNIA features [9] to represent test images because they appear to be highly competitive in a recent gMAD competition on the Waterloo Exploration Database [19]. In addition, a top performing OU-BIQA model, BLISS [89], also chooses CORNIA features as input and trains on synthetic scores. As such, we offer a fair testing bed to compare dipIQ learned by a pairwise L2R approach (RankNet [16]) against BLISS [89] learned by a regression method (SVR).

3) RankNet Instantiation: We investigate both linear and nonlinear dipIQ models, denoted by dipIQ* and dipIQ, respectively. The input dimension to RankNet is 20,000, equaling the feature dimension in CORNIA [9]. The loss layer is

Fig. 3. Sample source images in the training set. (a) Human. (b) Animal. (c) Plant. (d) Landscape. (e) Cityscape. (f) Still-life. (g) Transportation. All images are cropped for better visibility.

implemented by the cross entropy function in Eq. (3). For dipIQ*, the input layer is directly connected to the output layer without adding hidden layers or going through nonlinear transforms. The use of the cross entropy loss ensures the convexity of the optimization problem. For dipIQ, we add 3 hidden layers, which have a 256 - 128 - 3 structure. All layers are fully connected, followed by rectified linear units (ReLU) [90] as nonlinearity activations. We choose the node number of the third hidden layer to be 3 so that we can visualize the 3D embedding of test images. Other choices are somewhat ad-hoc, and a more careful exploration of alternative architectures could potentially lead to significant performance improvements.

The RankNet training procedure generally follows Simonyan and Zisserman [91]. Specifically, the training is carried out by optimizing the cross entropy function using mini-batch gradient descent with momentum. The weights of the two streams in RankNet are shared. The batch size is set to 512, and momentum to 0.9. The training is regularized by weight decay (the L_2 penalty multiplier set to 5×10^{-4}). The learning rate is fixed to 10^{-4}. Since we have a plenty of DIPs (more than 80 million) for training, each DIP is exposed to the learning algorithm once and only once. The learning stops when the entire set of DIPs have been swept. The weights that achieve the lowest validation set loss are used for testing.

B. Experimental Protocol

1) Databases: Four IQA databases are used to compare dipIQ with state-of-the-art BIQA measures. They are LIVE [86], CSIQ [92], TID2013 [15] and Waterloo Exploration Database [20]. The first three are small subject-rated IQA databases that are widely adopted to benchmark objective IQA models. Each test image is associated with an MOS to represent its perceptual quality. In our experiments, we only consider distortion types that are shared by all four databases, namely JP2K, JPEG, WN, and BLUR. As a result, LIVE [86], CSIQ [92], and TID2013 [15] contain 634, 600, and 500 test images, respectively. The Exploration database contains $4,744$ reference and $94,880$ distorted images. Although the MOS of each test image is not available in the Exploration database, innovative evaluation criteria are employed to compare BIQA measures as will be specified next.

2) Evaluation Criteria: We use five evaluation criteria to compare the performance of BIQA measures. The first two are included in previous tests carried out by the video quality experts group (VQEG) [93]. Others are introduced in [20] to take into account image databases without MOS. Details are given as follows.

- Spearman's rank-order correlation coefficient (SRCC) is defined as
$$\text{SRCC} = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)}, \tag{7}$$
where N is the number of images in a database and d_i is the difference between the i-th image's ranks in the MOS and model prediction.
- Pearson linear correlation coefficient (PLCC) is computed by
$$\text{PLCC} = \frac{\sum_i (s_i - \bar{s})(q_i - \bar{q})}{\sqrt{\sum_i (s_i - \bar{s})^2}\sqrt{\sum_i (q_i - \bar{q})^2}}, \tag{8}$$
where s_i and q_i stand for the MOS and model prediction of the i-th image, respectively.
- Pristine/distorted image discriminability test (D-test) considers pristine and distorted images as two distinct classes, and aims to measure how well an IQA model is able to separate the two classes. More specifically, indices of pristine and distorted images are grouped into

sets S_p and S_d, respectively. A threshold T is adopted to classify images such that $S'_p = \{i|q_i > T\}$ and $S'_d = \{i|q_i \leq T\}$. The average correct classification rate is defined as

$$R = \frac{1}{2}\left(\frac{|S_p \cap S'_p|}{|S_p|} + \frac{|S_d \cap S'_d|}{|S_d|}\right). \tag{9}$$

The value of T should be optimized to yield the maximum correct classification rate, which results in a discriminability index

$$D = \max_T R(T). \tag{10}$$

D lies in $[0, 1]$ with a larger value indicating a better separability between pristine and distorted images.

- Listwise ranking consistency test (L-test) evaluates the robustness of IQA models when rating images with the same content and the same distortion type but different distortion levels. The assumption is that the quality of an image degrades monotonically with the increase of the distortion level for any distortion type. Given a database with S source images, K distortion types and Q distortion levels, the average SRCC is used to quantify the ranking consistency between distortion levels and model predictions

$$L_s = \frac{1}{SK}\sum_{i=1}^{S}\sum_{j=1}^{K}\text{SRCC}(\mathbf{l}_{ij}, \mathbf{q}_{ij}), \tag{11}$$

where $\mathbf{l}_{ij}$ and $\mathbf{q}_{ij}$ represent the distortion levels and the corresponding distortion/quality scores given by a model to the set of images that are from the same (i-th) source image and have the same (j-th) distortion type.

- Pairwise preference consistency test (P-test) compares the performance of IQA models on a number of DIPs, whose generation is similar to what is described Section III-A but with a stricter rule [20]. A good IQA model should give concordant preferences with respect to DIPs. Assuming that an image database contains M DIPs and that the number of concordant pairs of an IQA model (meaning that the model predicts the correct preference) is M_c, the pairwise preference consistency ratio is defined as

$$P = \frac{M_c}{M}. \tag{12}$$

P lies in $[0, 1]$ with a higher value indicating better performance. We also denote the number of incorrect preference predictions as $M_i = M - M_c$.

SRCC and PLCC are applied to LIVE [86], CSIQ [92], and TID2013 [15], while the D-test, L-test, and P-test are applied to the Waterloo Exploration Database. Note that the use of PLCC requires a nonlinear function $\hat{q} = (\beta_1 - \beta_2)/(1 + \exp(-(q - \beta_3)/|\beta_4|)) + \beta_2$ to map raw model predictions to the MOS scale. Following Mittal *et al.* [8] and Ye *et al.* [89], in our experiments we randomly choose 80% reference images along with their corresponding distorted versions to estimate $\{\beta_i|i = 1, 2, 3, 4\}$, and use the rest 20% images for testing. This procedure is repeated 1,000 times and the median SRCC and PLCC values are reported.

TABLE I

MEDIAN SRCC AND PLCC RESULTS ACROSS 1,000 SESSIONS ON LIVE [86]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.908	0.894	0.984	0.814	0.883
SSIM [94]	0.961	0.974	0.970	0.952	0.947
QAC [60]	0.876	0.951	0.925	0.911	0.869
NIQE [43]	0.924	0.945	0.972	**0.941**	0.920
ILNIQE [65]	0.901	0.944	**0.979**	0.927	0.918
BLISS [89]	0.925	**0.956**	0.967	0.936	0.945
dipIQ*	**0.946**	**0.956**	**0.976**	**0.962**	**0.952**
dipIQ	**0.956**	**0.969**	0.975	0.940	**0.958**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.912	0.896	0.987	0.812	0.874
SSIM [94]	0.968	0.980	0.972	0.951	0.937
QAC [60]	0.876	0.960	0.895	0.912	0.855
NIQE [43]	0.932	0.956	**0.979**	**0.951**	0.912
ILNIQE [65]	0.912	0.966	0.976	0.936	0.913
BLISS [89]	0.933	**0.972**	0.978	0.948	0.945
dipIQ*	**0.958**	0.953	0.951	**0.950**	**0.948**
dipIQ	**0.964**	**0.980**	**0.983**	0.948	**0.957**

TABLE II

MEDIAN SRCC AND PLCC RESULTS ACROSS 1,000 SESSIONS ON CSIQ [86]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.941	0.901	0.943	0.936	0.928
SSIM [94]	0.962	0.956	0.912	0.965	0.935
QAC [60]	0.884	0.913	0.850	0.839	0.840
NIQE [43]	0.926	0.882	0.836	0.908	0.883
ILNIQE [65]	0.924	0.905	0.867	0.867	0.887
BLISS [89]	0.932	**0.927**	0.879	0.922	0.920
dipIQ*	**0.938**	0.926	**0.887**	**0.925**	**0.924**
dipIQ	**0.944**	**0.936**	**0.904**	**0.932**	**0.930**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.954	0.908	0.961	0.937	0.918
SSIM [94]	0.973	0.983	0.908	0.956	0.930
QAC [60]	0.898	0.942	0.865	0.855	0.847
NIQE [43]	0.944	0.946	0.824	0.935	0.900
ILNIQE [65]	0.942	0.956	0.880	0.903	0.914
BLISS [89]	0.954	0.970	0.895	0.947	0.939
dipIQ*	**0.955**	**0.971**	**0.903**	**0.951**	**0.946**
dipIQ	**0.959**	**0.975**	**0.927**	**0.958**	**0.949**

C. Experimental Results

1) Comparison With FR and OU-BIQA Models: We compare dipIQ with two well-known FR-IQA models: PSNR (whose largest value is clipped at 60 dB in order to perform a reasonable parameter estimation) and SSIM [94] (whose implementation used in the paper involves a down-sampling process [95]) and previous OU-BIQA models, including QAC [60], NIQE [43], ILNIQE [65], and BLISS [89]. The implementations of QAC [60], NIQE [43], and ILNIQE [65] are obtained from the original authors. To the best of our knowledge, the complete implementation of BLISS [89] is not publicly available. Therefore, to make a fair comparison we train BLISS [89] on the same 700 reference images and their distorted versions, which have been used to train dipIQ. The labels are synthesized using the method in [89]. The training toolbox and parameter settings are inherited from the original paper [89].

Tables I, II, and III list comparison results between dipIQ and existing OU-BIQA models in terms of median SRCC and

(a) (b) (c)

Fig. 4. The noisiness of the synthetic score [89]. (a) Synthetic score = 10. (b) Synthetic score = 10. (c) Synthetic score = 40. (a) has worse perceptual quality than (b), which in turn has approximately the same quality compared with (c). Both two cases are in disagreement with the synthetic score [89]. Images are selected from the training set.

TABLE III

MEDIAN SRCC AND PLCC RESULTS ACROSS 1,000 SESSIONS ON TID2013 [15]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.898	0.929	0.942	0.965	0.924
SSIM [94]	0.950	0.935	0.896	0.969	0.924
QAC [60]	0.883	0.885	0.668	0.879	0.837
NIQE [43]	0.901	0.873	0.854	0.821	0.812
ILNIQE [65]	**0.912**	0.873	**0.890**	0.815	**0.881**
BLISS [89]	0.906	0.893	0.856	0.872	0.836
dipIQ*	0.909	**0.903**	0.854	**0.884**	0.857
dipIQ	**0.926**	**0.932**	**0.905**	**0.922**	**0.877**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.933	0.925	0.963	0.958	0.911
SSIM [94]	0.970	0.968	0.902	0.958	0.927
QAC [60]	0.892	0.929	0.719	0.877	0.829
NIQE [43]	0.912	0.928	0.859	0.848	0.819
ILNIQE [65]	0.929	0.944	**0.899**	0.816	0.890
BLISS [89]	0.930	**0.963**	0.863	0.872	0.862
dipIQ*	**0.937**	**0.963**	0.851	**0.892**	**0.894**
dipIQ	**0.948**	**0.973**	**0.906**	**0.928**	**0.894**

TABLE IV

THE D-TEST, L-TEST AND P-TEST RESULTS ON THE WATERLOO EXPLORATION DATABASE [20].

	D	L_s	P	M_i
PSNR	1.0000	1.0000	0.9995	620,071
SSIM [94]	1.0000	0.9992	0.9991	1,131,457
QAC [60]	0.9226	0.8699	0.9779	28,447,590
NIQE [43]	0.9109	**0.9885**	0.9937	8,127,941
ILNIQE [65]	0.9084	**0.9926**	0.9927	9,435,319
BLISS [89]	0.9080	0.9801	0.9996	562,925
dipIQ*	**0.9209**	0.9863	**0.9996**	**465,069**
dipIQ	**0.9346**	0.9846	**0.9999**	**129,668**

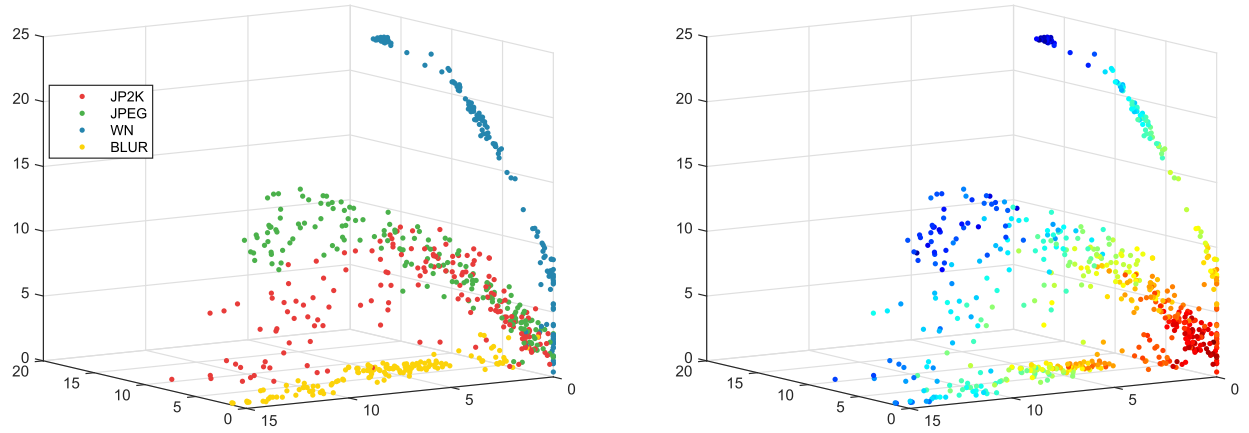
PLCC values on LIVE [86], CSIQ [92], and TID2013 [15], respectively. Both dipIQ* and dipIQ outperform all previous OU-BIQA models on LIVE [86] and CSIQ [92], and are comparable to ILNIQE [65] on TID2013 [15]. Although both dipIQ* and BLISS [89] learn a linear prediction function using CORNIA features as inputs [9], we observe consistent performance gains of dipIQ* across all three databases over BLISS [89]. This may be because dipIQ* learns from more reliable data (DIPs) with uncertainty weighting, whereas the training labels (synthetic scores) for BLISS are noisier, as exemplified in Fig. 4. It is not hard to observe that Fig. 4(a) has clearly worse perceptual quality than Fig. 4(b), which in turn has approximately the same quality compared with Fig. 4(c). Both two cases are in disagreement with the synthetic score [89].

To ascertain that the improvement of dipIQ is statistically significant, we carry out a two sample T-test (with a 95% confidence) between PLCC values obtained by different models on LIVE [86]. After comparing every possible pairs of OU-BIQA models, the results are summarized in Table V, where a symbol "1" means the row model performs significantly better than the column model, a symbol "0" means the opposite, and a symbol "-" indicates that the row and column models are statistically indistinguishable. It can be observed that dipIQ is statistically better than dipIQ*, which is better than all previous OU-BIQA models.

Table IV shows the results on the Waterloo Exploration Database. dipIQ* and dipIQ outperform all previous OU-BIQA models in the D-test and P-test, and are competitive in the L-test, whose performance is slightly inferior to NIQE [43] and ILNIQE [65]. By learning from examples with a variety of image content, dipIQ is able to crush the number of incorrect preference predictions in the P-test down to around 130,000 out of more than 1 billion candidate DIPs.

In order to gain intuitions on why the generalizability of dipIQ is excellent even without MOS for training, we visualize the 3D embedding of the LIVE database [86] in Fig 5, using the learned 3D features from the third hidden layer of dipIQ. We can see that the learned representation is able to cluster test images according to the distortion type, and meanwhile align them with respect to their perceptual quality in a meaningful way, where high quality images are clamped together regardless of image content.

2) Comparison With OA-BIQA Models: In the second set of experiments, we train dipIQ using different feature representations as inputs and compare with OA-BIQA models using the same representations and MOS for training. BRISQUE [8] and DIIVINE [10] are selected as representative features extracted from the spatial and wavelet domain, respectively. We also compare dipIQ with CORNIA [9], whose features are adopted as the default input to dipIQ. We re-train BRISQUE [8], DIIVINE [10], and CORNIA [9] on the LIVE database, whose learning tools and parameter settings



(a) (b)

Fig. 5. 3D embedding of the LIVE database [86]. (a) Color encodes distortion type. (b) Color encodes quality; the warmer, the better. The learned features from the third hidden layer of dipIQ are able to cluster images based on distortion types and align them in a perceptually meaningful way.

TABLE V

STATISTICAL SIGNIFICANCE MATRIX BASED ON THE HYPOTHESIS TESTING. A SYMBOL "1" MEANS THAT THE PERFORMANCE OF THE ROW ALGORITHM IS STATISTICALLY BETTER THAN THAT OF THE COLUMN ALGORITHM, A SYMBOL "0" MEANS THAT THE ROW ALGORITHM IS STATISTICALLY WORSE, AND A SYMBOL "–" MEANS THAT THE ROW AND COLUMN ALGORITHMS ARE STATISTICALLY INDISTINGUISHABLE

PLCC	PSNR	SSIM	QAC	NIQE	ILNIQE	BLISS	dipIQ*	dipIQ
PSNR	-	0	1	0	0	0	0	0
SSIM [94]	1	-	1	1	1	0	0	0
QAC [60]	0	0	-	0	0	0	0	0
NIQE [43]	1	0	1	-	-	0	0	0
ILNIQE [65]	1	0	1	-	-	0	0	0
BLISS [89]	1	1	1	1	1	-	0	0
dipIQ*	1	1	1	1	1	1	-	0
dipIQ	1	1	1	1	1	1	1	-

TABLE VI

MEDIAN SRCC AND PLCC RESULTS ACROSS 1,000 SESSIONS, TRAINING ON LIVE [86] AND TESTING ON CSIQ [92]. THE SUPERSCRIPTS *B* AND *D* INDICATE THAT THE INPUT FEATURES OF DIPIQ ARE FROM BRISQUE [8] AND DIIVINE [10], RESPECTIVELY

SRCC	JP2K	JPEG	WN	BLUR	ALL4
BRISQUE [8]	0.894	0.916	**0.934**	0.915	0.909
dipIQ^B	**0.938**	**0.938**	**0.934**	**0.943**	**0.926**
DIIVINE [10]	0.844	0.819	0.881	0.884	0.835
dipIQ^D	**0.930**	**0.939**	**0.904**	**0.920**	**0.912**
CORNIA [9]	0.916	0.919	0.787	0.928	0.915
dipIQ	**0.944**	**0.936**	**0.904**	**0.932**	**0.930**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
BRISQUE [8]	0.937	0.960	**0.947**	0.936	0.937
dipIQ^B	**0.956**	**0.974**	0.945	**0.959**	**0.943**
DIIVINE [10]	0.898	0.818	0.903	0.909	0.855
dipIQ^D	**0.949**	**0.973**	**0.924**	**0.944**	**0.942**
CORNIA [9]	0.947	0.960	0.777	0.953	0.934
dipIQ	**0.959**	**0.975**	**0.927**	**0.958**	**0.949**

TABLE VII

MEDIAN SRCC AND PLCC RESULTS ACROSS 1,000 SESSIONS, TRAINING ON LIVE [86] AND TESTING ON TID2013 [15]

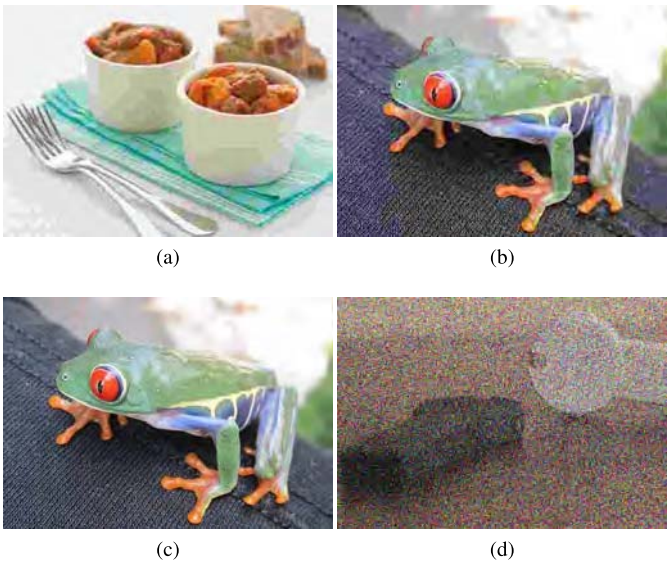
SRCC	JP2K	JPEG	WN	BLUR	ALL4
BRISQUE [8]	0.906	0.894	0.889	0.886	**0.883**
dipIQ^B	**0.927**	**0.921**	**0.921**	**0.917**	**0.883**
DIIVINE [10]	0.857	0.680	0.879	0.859	0.795
dipIQ^D	**0.912**	**0.889**	**0.887**	**0.905**	**0.872**
CORNIA [9]	0.907	0.912	0.798	**0.934**	**0.893**
dipIQ	**0.926**	**0.932**	**0.905**	0.922	0.877
PLCC	JP2K	JPEG	WN	BLUR	ALL4
BRISQUE [8]	0.919	0.950	0.886	0.884	**0.901**
dipIQ^B	**0.942**	**0.957**	**0.923**	**0.906**	0.883
DIIVINE [10]	0.901	0.696	**0.882**	0.860	0.794
dipIQ^D	**0.945**	**0.947**	0.881	**0.896**	**0.892**
CORNIA [9]	0.923	0.960	0.778	**0.934**	**0.904**
dipIQ	**0.948**	**0.973**	**0.906**	0.928	0.894

follow their respective papers. We adjust the dimension of the input layer of dipIQ to accommodate features of different dimensions and train them on the 700 reference images and their distorted versions, as described in IV-A. All models are tested on CSIQ [92], TID2013 [15] and the Exportation database [20]. From Tables VI, VII, and VIII, we observe that dipIQ consistently performs better than the corresponding OA-BIQA model on CSIQ [92] and the Exploration database, and is comparable on TID2013 [15]. The reason we do not obtain noticeable performance gains on TID2013 [15] may be that TID2013 [15] has 18 references images originated from LIVE [86], based on which the OA-BIQA models have been trained. This creates dependencies between training and testing sets. We may also draw conclusions about the effectiveness of the feature representations based on their performance

TABLE VIII

THE D-TEST, L-TEST AND P-TEST RESULTS ON THE EXPLORATION DATABASE [20], TRAINING ON LIVE [86]

	D	L_s	P	M_i
BRISQUE [8]	0.9204	**0.9772**	0.9930	9,004,685
dipIQ^B	**0.9265**	0.9753	**0.9996**	**503,911**
DIIVINE [10]	0.8538	0.8908	0.9540	59,053,011
dipIQ^D	**0.9191**	**0.9588**	**0.9983**	**2,124,199**
CORNIA [9]	0.9290	0.9764	0.9947	6,808,400
dipIQ	**0.9346**	**0.9846**	**0.9999**	**129,668**



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

Fig. 6. gMAD competition between dipIQ^B and BRISQUE [8]. (a) best BRISQUE for fixed dipIQ^B. (b) worst BRISQUE for fixed dipIQ^B. (c) best dipIQ^B for fixed BRISQUE. (d) worst dipIQ^B for fixed BRISQUE.

under the same pairwise L2R framework: generally speaking, CORNIA [9] features > BRISQUE [8] features > DIIVINE [10] features.

We further compare dipIQ^B and BRISQUE [8] using the gMAD competition methodology on the Waterloo Exploration Database. Specifically, we first find a pair of images that have the maximum and minimum dipIQ^B values from a subset of images in the Exploration database, where BRISQUE [8] rates them to have the same quality. We then repeat this procedure, but with the roles of dipIQ^B and BRISQUE [8] exchanged. The two image pairs are shown in Fig. 6, from which we conclude that images in the first row exhibits approximately the same perceptual quality (in agreement with dipIQ^B) and those in the second row has drastically different perceptual quality (in disagreement with BRISQUE [8]). This verifies that the robustness of dipIQ^B is significantly improved over BRISQUE [8] using the same feature representations and MOS for training. Similar gMAD competition results are obtained across all quality levels, and for dipIQ^D versus DIIVINE [10] and dipIQ versus CORNIA [9].

In summary, the proposed pairwise L2R approach is proved to learn OU-BIQA models with improved generalizability and robustness compared with OA-BIQA models using the same feature representations and MOS for training.

V. LISTWISE L2R APPROACH FOR OU-BIQA

In this section, we extend the proposed pairwise L2R approach for OU-BIQA to a listwise L2R one. Specifically, we first construct three-element DILs by concatenating DIPs. For example, given two DIPs $\langle i, j\rangle$ and $\langle j, k\rangle$ with the same level of uncertainty, we create a list $\langle i, j, k\rangle$ with the ground truth label $\bar{P}_{ijk} = 1$, indicating that the quality of the i-th image is better than the j-th image, whose quality is better than the k-th image. The uncertainty level is transferred as well. We then employ ListNet [21], a listwise L2R extension of RankNet [16] to learn OU-BIQA models. The major differences between ListNet and RankNet are twofold. First, ListNet can have multiple streams with the same weights to accommodate a list of inputs, where each stream is implemented by a classical neural network architecture similar to RankNet, as shown in Fig. 2. In this paper, we instantiate a three-stream ListNet to fit three-element DILs. Second, the loss function of ListNet is defined using the concept of permutation probability. More specifically, we define a permutation $\pi = \langle \pi(1), \pi(2), \ldots, \pi(n)\rangle$ on a list of n instances as a bijection from $\{1, 2, .., n\}$ to itself, where $\pi(j)$ denotes the instance at position j in the permutation. The set of all possible permutations of n instances is termed as Π. We define the probability of permutation π given the list of predicted scores $\{f(\mathbf{x}_i)\}$ as

$$P_\pi(f) = \prod_{j=1}^{n} \frac{\exp\left(f(\mathbf{x}_{\pi(j)})\right)}{\sum_{k=j}^{n} \exp\left(f(\mathbf{x}_{\pi(k)})\right)}, \tag{13}$$

which satisfies $P_\pi(f) > 0$ and $\sum_{\pi\in\Pi} P_\pi(f) = 1$ as proved in [21]. The loss function can then be defined as the cross entropy function between the ground truth and permutation probabilities

$$L(f; \{\mathbf{x}_i\}, \{\bar{P}_\pi\}) = -\sum_{\pi\in\Pi} \bar{P}_\pi \log(P_\pi). \tag{14}$$

When $n = 2$, the loss function of ListNet [21] in Eq. (14) becomes equivalent to that of RankNet [16] in Eq. (3). In the case of three-element DILs, we have $\bar{P}_\pi = 1$, if $\pi = \langle i, j, k\rangle$ and $\bar{P}_\pi = 0$ otherwise. Therefore, the loss function in Eq. (14) can be simplified as

$$\begin{aligned} &L(f; \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \bar{P}_{ijk}) \\ &\quad = -f(\mathbf{x}_i) - f(\mathbf{x}_j) + \log\left(\sum_{l\in\{i,j,k\}} \exp\left(f(\mathbf{x}_l)\right)\right) \\ &\qquad + \log\left(\sum_{l\in\{j,k\}} \exp\left(f(\mathbf{x}_l)\right)\right), \end{aligned} \tag{15}$$

based on which we define the batch-level loss as

$$L_b(f) = \sum_{\langle i,j,k\rangle\in\mathcal{B}} (1 - U_{ijk}) L(f; \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \bar{P}_{ijk}), \tag{16}$$

where U_{ijk} is the uncertainty level of the list, transferred from the corresponding DIPs. The gradient of Eq. (16) w.r.t. the parameters $\mathbf{w}$ can be easily derived. Note that ListNet [21] does not add new parameters.

TABLE IX

MEDIAN SRCC AND PLCC RESULTS ACROSS 1, 000 SESSIONS ON LIVE [86], USING LISTNET [21] FOR TRAINING

SRCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.956	0.969	0.975	0.940	0.958
dilIQ	0.956	0.966	**0.976**	**0.953**	0.958
PLCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.964	0.980	0.983	0.948	0.957
dilIQ	0.964	0.978	**0.985**	**0.956**	0.954

TABLE X

MEDIAN SRCC AND PLCC RESULTS ACROSS 1, 000 SESSIONS ON CSIQ [92], USING LISTNET [21] FOR TRAINING

SRCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.944	0.936	0.904	0.932	0.930
dilIQ	0.930	0.925	0.893	**0.939**	**0.936**
PLCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.959	0.975	0.927	0.958	0.949
dilIQ	0.954	0.968	0.920	**0.960**	**0.954**

TABLE XI

MEDIAN SRCC AND PLCC RESULTS ACROSS 1, 000 SESSIONS ON TID2013 [15], USING LISTNET [21] FOR TRAINING

SRCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.926	0.932	0.905	0.922	0.877
dilIQ	0.918	0.849	0.905	**0.925**	**0.891**
PLCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.948	0.973	0.906	0.928	0.894
dilIQ	0.948	0.923	0.903	**0.929**	**0.915**

TABLE XII

THE D-TEST, L-TEST AND P-TEST RESULTS ON THE EXPLORATION DATABASE [20], USING LISTNET [21] FOR TRAINING

	D	L_s	P	M_i
dipIQ	0.9346	0.9846	0.9999	129,668
dilIQ	0.9346	**0.9893**	0.9998	198,650

We generate 50 million DILs from the available DIPs as the training data for ListNet [21]. The training procedure is exactly the same as training RankNet [16]. The training stops when the entire set of image lists have been swept once. The weights that achieve the lowest validation set loss are used for testing.

We list the comparison results between dilIQ trained by ListNet [21] and the baseline dipIQ on LIVE [86], CSIQ [92], TID2013 [15], and the Exploration database in Tables IX, X, XI, and XII, respectively. Remarkable performance improvements have been achieved on CSIQ and TID2013. This may be because the ranking position information is made explicit to the learning process. dilIQ is comparable to dipIQ on LIVE and the Exploration database.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an OU-BIQA model, namely dipIQ, using RankNet [16]. The input to the dipIQ training model are an enormous number of DIPs, not obtained by expensive subjective testing but automatically generated with the help of most trusted FR-IQA models at low cost. Extensive experimental results demonstrate the effectiveness of the proposed dipIQ indices with higher accuracy and improved robustness in content variations. We also learn an OU-BIQA model, namely dilIQ, using a listwise L2R approach, which achieves an additional performance gain.

The current work opens the door to a new class of OU-BIQA models and can be extended in many ways. First, novel image pair and list generation engines may be developed to account for situations that reference images are not available (or do not ever exist). Second, advanced L2R algorithms are worth exploring to improve the quality prediction performance. Third, in practice, a pair of images may be regarded as having indiscriminable quality. Such knowledge could be obtained either from subjective testing (*e.g.*, paired comparison between images) or from the image source (*e.g.*, two pristine images acquired from the same source), and is informative in constraining the behavior of an objective quality model. The current learning framework needs to be improved in order to learn from such quality-indiscriminable image pairs. Fourth, given the powerful DIP generation engine developed in the current work and the remarkable success of recent deep convolutional neural networks, it may become feasible to develop end-to-end BIQA models that bypass the feature extraction process and achieve even stronger robustness and generalizability.

ACKNOWLEDGMENT

The authors would like to thank Zhengfang Duanmu for suggestions on the efficient implementation of RankNet, and the anonymous reviewers for constructive comments.

REFERENCES

[1] H. R. Wu and K. R. Rao, *Digital Video Image Quality and Perceptual Coding*. Boca Raton, FL, USA: CRC Press, 2005.
[2] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. San Rafael, CA, USA: Morgan & Claypool, 2006.
[3] S. J. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," *Proc. SPIE*, vol. 1666, pp. 2–15, Aug. 1992.
[4] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.
[5] Z. Wang and A. C. Bovik, "Reduced- and no-reference image quality assessment: The natural scene statistic model approach," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 29–40, Nov. 2011.
[6] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.
[7] 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.
[8] 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.
[9] 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.*, Jun. 2012, pp. 1098–1105.
[10] 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.
[11] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 339–343.

[12] 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.

[13] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

[14] Q. Wu *et al.*, "Blind image quality assessment based on multi-channel features fusion and label transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 425–440, Mar. 2016.

[15] 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

[16] C. Burges *et al.*, "Learning to rank using gradient descent," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 89–96.

[17] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[18] L. Hang, "A short introduction to learning to rank," *IEICE Trans. Inf. Syst.*, vol. 94, no. 10, pp. 1854–1862, Oct. 2011.

[19] K. Ma *et al.*, "Group MAD competition? A new methodology to compare objective image quality models," in *Proc. IEEE Conf. Comput. Vsion Pattern Recognit.*, Jun. 2016, pp. 1664–1673.

[20] 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.

[21] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 129–136.

[22] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Process. Lett.*, vol. 4, no. 11, pp. 317–320, Nov. 1997.

[23] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Process.*, Jun. 2000, pp. 981–984.

[24] S. Liu and A. C. Bovik, "Efficient DCT-domain blind measurement and reduction of blocking artifacts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 12, pp. 1139–1149, Dec. 2002.

[25] H. Tong, M. Li, H. Zhang, and C. Zhang, "Blur detection for digital images using wavelet transform," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2004, pp. 17–20.

[26] Z. Wang and E. P. Simoncelli, "Local phase coherence and the perception of blur," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 1435–1442.

[27] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *Proc. Int. Workshop Quality Multimedia Exper.*, 2009, pp. 64–69.

[28] S. Oğuz, Y. Hu, and T. Q. Nguyen, "Image coding ringing artifact reduction using morphological post-filtering," in *Proc. IEEE Workshop Multimedia Signal Process.*, Jun. 1998, pp. 628–633.

[29] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, no. 1, pp. 1918–1927, Nov. 2005.

[30] H. Tao, N. Klomp, and I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 529–539, Apr. 2010.

[31] B. A. Wandell, *Foundations of Vision*. Sunderland, MA, USA: Sinauer Associates, 1995.

[32] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.

[33] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *J. Neurosci.*, vol. 9, no. 2, pp. 181–197, 1992.

[34] D. J. Field, "What is the goal of sensory coding?" *Neural Comput.*, vol. 6, pp. 559–601, Jul. 1994. [Online]. Available: http://portal.acm.org/citation.cfm?id=188132.188136.

[35] W. S. Geisler and R. L. Diehl, "Bayesian natural selection and the evolution of perceptual systems," *Philos. Trans. Roy. Soc. London B, Biologic. Sci.*, vol. 357, no. 1420, pp. 419–448, Apr. 2002.

[36] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[37] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[38] X. Li, "Blind image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Jun. 2002, pp. 449–452.

[39] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 163–172, Feb. 2004.

[40] 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.

[41] 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 Process. Lett.*, vol. 22, no. 7, pp. 838–842, Jul. 2015.

[42] X. Zhu and P. Milanfar, "Automatic parameter selection for denoising algorithms using a no-reference measure of image content," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3116–3132, Dec. 2010.

[43] 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.

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

[45] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 987–994.

[46] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Human Vis. Electron. Imag.*, 2005, pp. 149–159.

[47] 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.

[48] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[49] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012.

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

[51] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2877–2884.

[52] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3129–3138, Jul. 2012.

[53] R. Hassen, Z. Wang, and M. M. A. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2798–2810, Jul. 2013.

[54] L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 157–170.

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

[56] T. Huang, J. Burnett, and A. Deczky, "The importance of phase in image processing filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 6, pp. 529–542, Dec. 1975.

[57] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[58] P. Kovesi, "Image features from phase congruency," *J. Comput. Vis. Res.*, vol. 1, no. 3, pp. 1–26, Jun. 1999.

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

[60] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 995–1002.

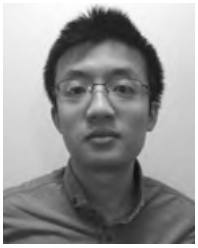
[61] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[62] 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.*, Jun. 2014, pp. 1733–1740.

[63] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

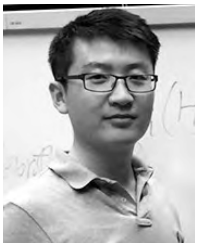
[64] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, May 2000.

[65] 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.
[66] L. Xu, W. Lin, J. Li, X. Wang, Y. Yan, and Y. Fang, "Rank learning on training set selection and image quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.
[67] F. Gao, D. Tao, X. Gao, and X. Li, "Learning to rank for blind image quality assessment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2275–2290, Oct. 2015.
[68] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1, pp. 177–196, Jan. 2001.
[69] X. Gao, F. Gao, D. Tao, and X. Li, "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 2013–2026, Dec. 2013.
[70] N. Fuhr, "Optimum polynomial retrieval functions based on the probability ranking principle," *ACM Trans. Inf. Syst.*, vol. 7, no. 3, pp. 183–204, Jul. 1989.
[71] D. Cossock and T. Zhang, "Subset ranking using regression," in *Proc. Conf. Learn. Theory*, 2006, pp. 605–619.
[72] R. Nallapati, "Discriminative models for information retrieval," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2004, pp. 64–71.
[73] K. Crammer and Y. Singer, "Pranking with ranking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 641–647.
[74] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 937–944.
[75] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 133–142.
[76] M. F. Tsai, T. Y. Liu, T. Qin, H. H. Chen, and W. Y. Ma, "FRank: A ranking method with fidelity loss," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2007, pp. 383–390.
[77] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *J. Mach. Learn. Res.*, vol. 4, no. 6, pp. 170–178, Nov. 2003.
[78] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," in *Proc. Eur. Conf. Comput. Learn. Theory*, 1995, pp. 23–37.
[79] M. Taylor, J. Guiver, S. Robertson, and T. Minka, "SoftRank: Optimizing non-smooth rank metrics," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2008, pp. 77–86.
[80] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2007, pp. 271–278.
[81] J.-Y. Yeh, J.-Y. Lin, H.-R. Ke, and W.-P. Yang, "Learning to rank for information retrieval using genetic programming," in *Proc. SIGIR Workshop Learn. Rank Inf. Retr.*, 2007, pp. 1–8.
[82] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, Jun. 2003, pp. 1398–1402.
[83] 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.
[84] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.
[85] 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.
[86] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack. *Image and Video Quality Assessment Research*, LIVE, accessed on Apr. 18, 2016. [Online]. Available: http://live.ece.utexas.edu/research/quality/
[87] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[88] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[89] P. Ye, J. Kumar, and D. Doermann, "Beyond human opinion scores: Blind image quality assessment based on synthetic scores," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4241–4248.
[90] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. IEEE Int. Conf. Mach. Learn.*, Jun. 2010, pp. 807–814.
[91] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[92] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *SPIE J. Electron. Imag.*, vol. 19, no. 1, pp. 1–21, Jan. 2010.
[93] VQEG. (2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*. [Online]. Available: http://www.vqeg.org
[94] 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.
[95] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. *The SSIM Index for Image Quality Assessment*, accessed on Apr. 18, 2016. [Online]. Available: https://ece.uwaterloo.ca/ z70wang/research/ssim/



Kede Ma (S'13) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2012, and the M.A.Sc. degree from the University of Waterloo, ON, Canada, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests lie in perceptual image processing and computational photography.

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



Tongliang Liu received the B.Eng. degree in electronic engineering and information science from the University of Science and Technology of China, and the Ph.D. degree from the University of Technology Sydney. He is currently a Lecturer with the School of Information Technologies, Faculty of Engineering and Information Technologies, The University of Sydney, and a Core Member with the UBTech Sydney Artificial Intelligence Institute, The University of Sydney. His current research interests include statistical learning theory, computer vision, and optimization. He has authored and co-authored over 20 research papers, including IEEE T-PAMI, T-NNLS, T-IP, ICML, and KDD.

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 with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His current research interests include image processing, coding, and quality assessment; computational vision and pattern analysis; multimedia communications; and biomedical signal processing. He has over 100 publications in these fields with over 30 000 citations (Google Scholar).

Dr. Wang is currently a Senior Area Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (since 2015), and an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (since 2016). He was a member of the IEEE Multimedia Signal Processing Technical Committee (2013–2015), an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (2009–2014), Pattern Recognition (since 2006), and the IEEE SIGNAL PROCESSING LETTERS (2006–2010), and a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (2013–2014 and 2007–2009). He is a fellow of the Canadian Academy of Engineering, and a recipient of the 2016 IEEE Signal Processing Society Sustained Impact Paper Award, the 2015 Primetime Engineering Emmy Award, the 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, the 2013 IEEE Signal Processing Magazine Best Paper Award, the 2009 IEEE Signal Processing Society Best Paper Award, and the 2009 Ontario Early Researcher Award.

Dacheng Tao (F'15) is a Professor of Computer Science and an ARC Future Fellow with the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTech Sydney Artificial Intelligence Institute, The University of Sydney. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and over 500 publications at prestigious journals and prominent conferences, such as the IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, CIKM, ICML, CVPR, ICCV, ECCV, AISTATS, and ICDM; and ACM SIGKDD, with several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award in IEEE ICDM'07, the Best Student Paper Award in IEEE ICDM'13, the 2014 ICDM 10-Year Highest-Impact Paper Award, and the 2017 IEEE Signal Processing Society Best Paper Award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellor's Medal for Exceptional Research. He is a Fellow of the OSA, IAPR and SPIE.