Blind Image Quality Assessment Using A Deep Bilinear Convolutional Neural Network

Weixia Zhang, Kede Ma, *Member, IEEE,* Jia Yan, Dexiang Deng, and Zhou Wang, *Fellow, IEEE*

***Abstract*—We propose a deep bilinear model for blind image quality assessment (BIQA) that works for both synthetically and authentically distorted images. Our model constitutes two streams of deep convolutional neural networks (CNN), specializing in the two distortion scenarios separately. For synthetic distortions, we first pre-train a CNN to classify the distortion type and level of an input image, whose ground truth label is readily available at a large scale. For authentic distortions, we make use of a pre-train CNN (VGG-16) for the image classification task. The two feature sets are bilinearly pooled into one representation for a final quality prediction. We fine-tune the whole network on target databases using a variant of stochastic gradient descent. Extensive experimental results show that the proposed model achieves state-of-the-art performance on both synthetic and authentic IQA databases. Furthermore, we verify the generalizability of our method on the large-scale Waterloo Exploration Database, and demonstrate its competitiveness using the group maximum differentiation competition methodology.**

***Index Terms*—Blind image quality assessment, convolutional neural networks, bilinear pooling, gMAD competition, perceptual image processing.**

I. Introduction

NOWADAYS, digital images are captured by various stationary and mobile cameras, compressed by traditional and novel techniques [1], [2], transmitted through diverse communication channels [3], and stored in a variety of storage devices. Each stage in the image acquisition, processing, transmission and storage pipeline could introduce unexpected distortions, and cause perceptual information loss and quality degradation. Image quality assessment (IQA), therefore, becomes increasingly important in monitoring the quality of images and assuring the reliability of image processing systems. Since the human visual system is the ultimate judge of perceptual image quality, subjective IQA is most reliable, but is also time-consuming and expensive. Hence, it is essential to design accurate and efficient objective IQA algorithms to push IQA from laboratory research to real-world applications [4]. Objective IQA is traditionally classified into three categories depending on the availability of reference information: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference or blind IQA (BIQA) [5]. Because no reference information is available (or may not even exist) in many realistic situations, BIQA attracts a significant amount of research interests in recent years [6].

Traditional BIQA models commonly adopt low-level features either hand-crafted [7] or learned [8] to characterize the level of deviations from statistical regularities of natural scenes, based on which a quality prediction function is learned [9]. Until recently, there has been limited effort towards end-to-end optimized BIQA using deep convolutional neural networks (CNN) [10], [11], primarily due to the lack of sufficient ground truth labels such as the mean opinion scores (MOS) for training. A naïve solution is to directly fine-tune a CNN pre-trained on ImageNet [12] for quality prediction [13]. The resulting CNN-based quality model achieves reasonable performance on the LIVE Challenge Database [14] (authentically distorted), but does not deliver standout performance on legacy IQA databases such as LIVE [15] and TID2013 [16] (synthetically distorted). Another commonly adopted strategy is patch-based training, where the quality score of a patch is either inherited from that of the corresponding image [10] or approximated by FR-IQA models [17]. This strategy is very effective at learning CNN models for synthetic distortions, but fails to handle authentic distortions due to the non-homogeneity of distortions and the absence of reference images for patch quality annotation. Other methods [11], [18] take advantage of the known synthetic degradation processes (*e.g.*, distortion types) to find reasonable initializations of CNN models for quality prediction, which however are not directly applicable to authentic distortions.

In this work, we aim for an end-to-end solution to BIQA of both synthetically and authentically distorted images. We first learn feature representations that are matched with the two degradation scenarios separately. For synthetic distortions, inspired by previous works [11], [18], [20], we construct a large-scale pre-training set based on the Waterloo Exploration Database [19] and PASCAL VOC 2012 [21], where the images are synthesized with nine distortion types and two to five distortion levels. Instead of rating each distorted image in the pre-training set, we take advantage of the known distortion type and level information and pre-train a CNN through a multi-class classification task. For authentic distortions, it is difficult to simulate the degradation processes due to their complexities [22]. Here, we opt to use another CNN model (VGG-16 [23] to be exact) that is pre-trained on ImageNet [12], containing many realistic natural images of different quality, and is therefore better matched to the rich content and distortion variations in authentically distorted images. We model

This work was supported in part by the National Natural Science Foundation of China under Grant 61701351.

Weixia Zhang, Jia Yan, and Dexiang Deng are with the Electronic Information School, Wuhan University, Wuhan, China (e-mail: zhangweixia@whu.edu.cn; yanjia2011@gmail.com; ddx@whu.edu.cn).

Kede Ma is with the Center for Neural Science, New York University, New York, NY 10012, USA (e-mail: k29ma@uwaterloo.ca).

Zhou Wang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: zhou.wang@uwaterloo.ca).

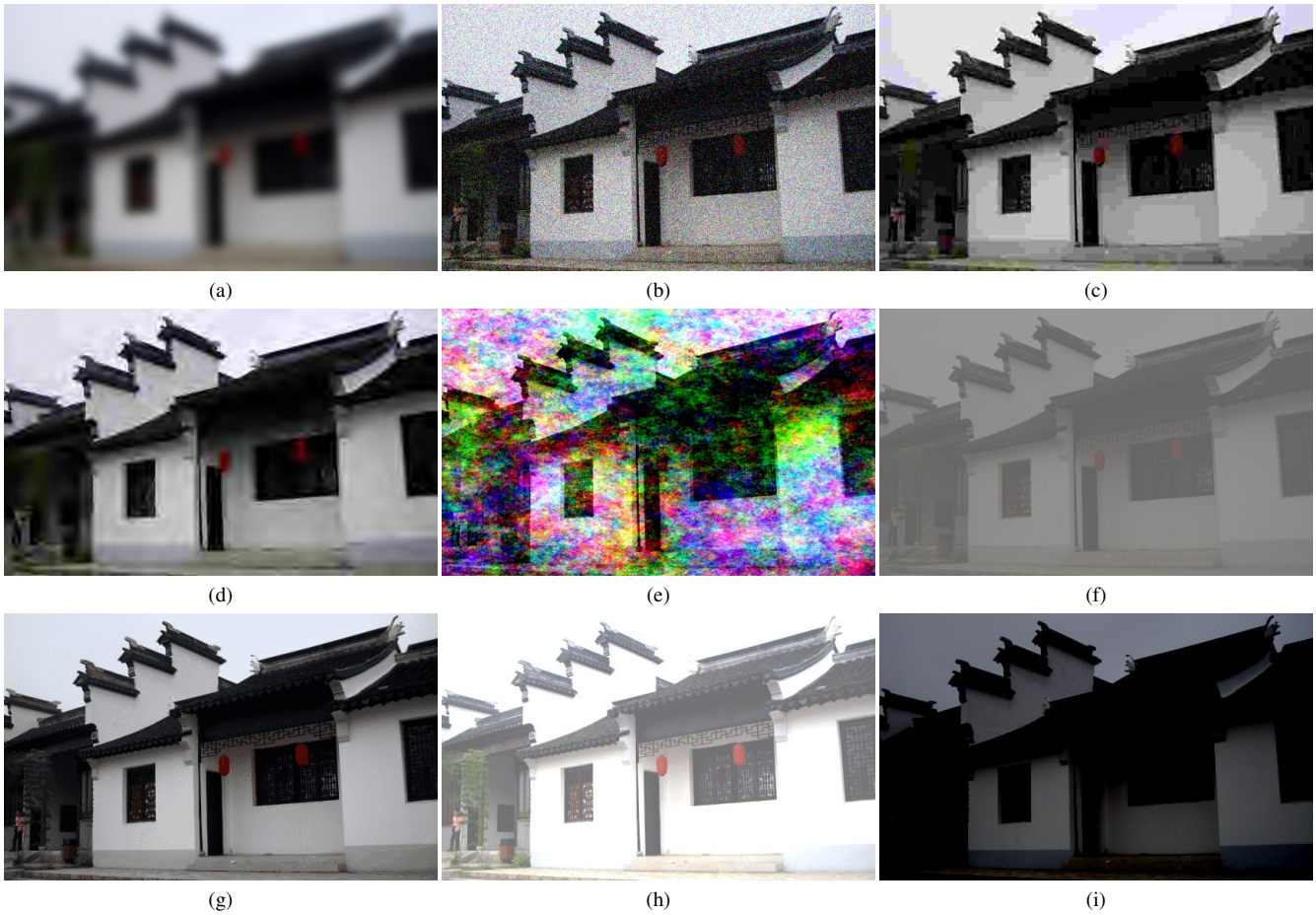

Fig. 1. Sample distorted images synthesized from a reference image in the Waterloo Exploration Database [19]. (a) Gaussian blur. (b) White Gaussian noise. (c) JPEG compression. (d) JPEG2000 compression. (e) Pink noise. (f) Contrast stretching. (g) Image color quantization with dithering. (h) Over-exposure. (i) Under-exposure.

synthetic and authentic distortions as two-factor variations, and bilinearly pool the two pre-trained feature sets into a unified representation, resulting in a deep bilinear CNN (DB-CNN) [24] for quality prediction. The proposed DB-CNN is fine-tuned on target databases with a variant of the stochastic gradient descent method. Extensive experimental results on four IQA databases demonstrate the effectiveness of DB-CNN for both synthetic and authentic distortions. Furthermore, through the group MAximum Differentiation (gMAD) competition [25], we observe that DB-CNN is more robust than the most recent CNN-based BIQA models [11], [26].

The remainder of this paper is organized in the following manner. Section II reviews CNN-based models for BIQA with emphasis on their limitations. Section III details the construction of the proposed DB-CNN model. We present extensive comparison and ablation experiments in Section IV. Section V concludes the paper.

II. Related Work

In this section, we provide a review of recent CNN-based BIQA models. For a more detailed treatment of BIQA, readers can refer to [6], [9], [27], [28].

Tang *et al.* [29] pre-trained a deep belief network with a radial basis function and fine-tuned it to predict image quality. Bianco *et al.* [30] investigated various design choices of CNN for BIQA. They first adopted CNN features pre-trained on the image classification task as inputs to learn a quality evaluator using support vector regression (SVR). They then fine-tuned the pre-trained features in a multi-class classification setting by quantizing MOSs 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 [30]. Kang *et al.* [10] trained a CNN using a large number of spatially normalized image patches and computed the quality score of an input image by averaging the predicted scores of all image patches cropped from it. They then simultaneously estimated image quality and distortion type via a traditional multi-task CNN [18]. While the quality scores of patches are directly inherited from the corresponding image, it may be problematic since local perceptual quality is not always consistent with global quality due to the high non-stationarity of image content across spatial locations and the intricate interactions between content and distortions [11], [13]. Taking this problem into consideration, Bosse *et al.* [26] trained CNN models using two different strategies: 1) directly averaging features from multiple patches and 2) weighted averaging quality scores of patches weighted by their relative

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2886771, IEEE Transactions on Circuits and Systems for Video Technology

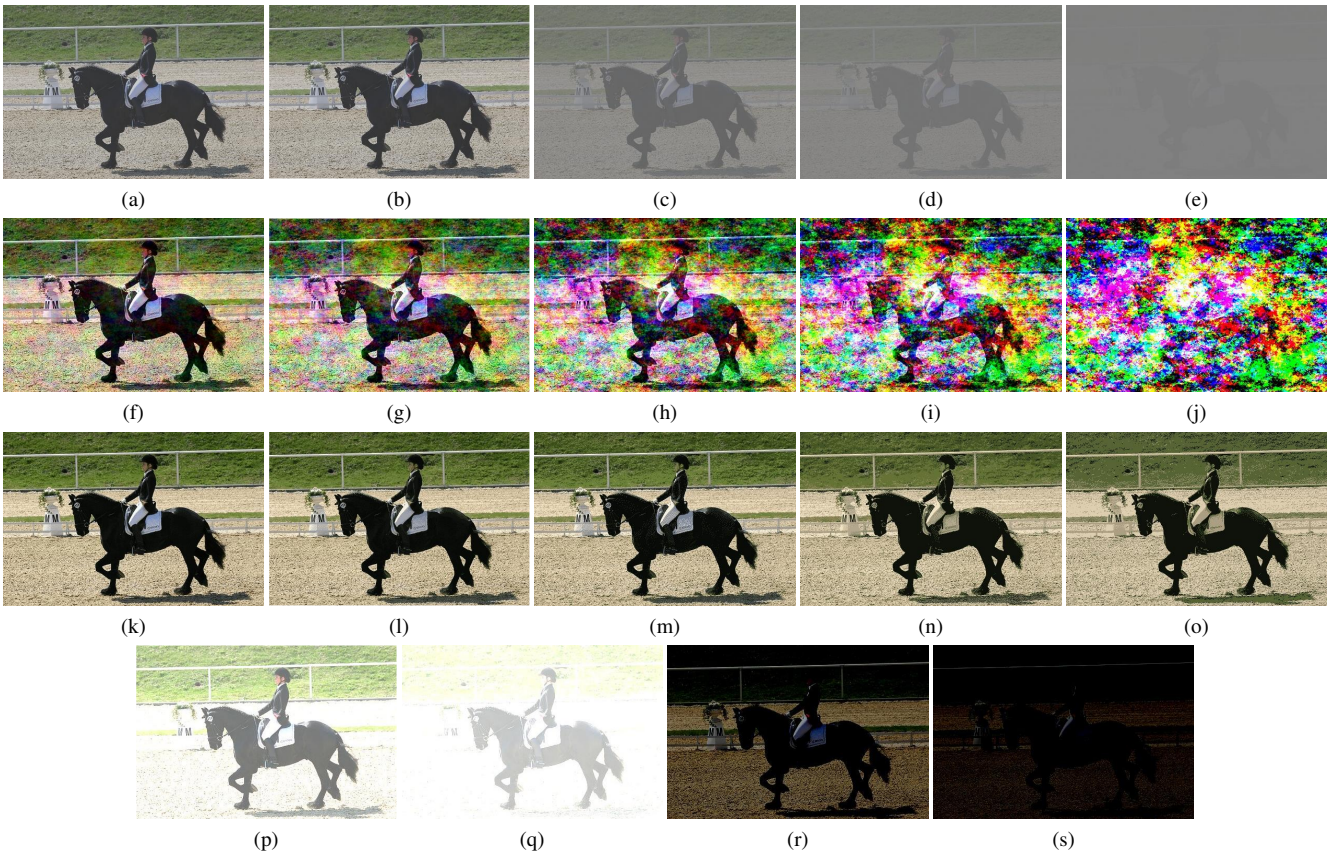

Fig. 2. Illustration of the five new distortion types with increasing degradation levels from left to right. (a)-(e) Contrast stretching. (f)-(j) Pink noise. (k)-(o) Image color quantization with dithering. (p)-(q) Over-exposure. (r)-(s) Under-exposure.

importance. Kim *et al.* [17] first pre-trained a CNN model using numerous patches with proxy quality scores acquired by an FR-IQA model [31] and then summarized the patch-level feature representations using mean and standard deviation statistics for fine tuning. A closely related work to ours is MEON [11], a cascaded multi-task framework for BIQA. A distortion type identification network is first trained, for which large-scale training samples are readily available. Then, starting from the pre-trained early layers and the outputs of the distortion type identification network, a quality prediction network is trained subsequently. The proposed DB-CNN takes a step further by taking not only distortion type but also distortion level information into account, which results in better quality-aware initializations. It is worth noting that the aforementioned three methods [11], [17], [26] only partially address the training data shortage problem in the synthetic distortion scenario. Extending them to account for authentic distortions is difficult.

III. DB-CNN FOR BIQA

In this section, we first describe the construction of the pre-training set and the architecture of the CNN for synthetically distorted images. We then present the tailored VGG-16 network for authentically distorted images. Finally, we introduce our bilinear pooling module along with the fine-tuning procedure.

A. CNN for Synthetic Distortions

To address the enormous content variations in real world images, we start with two large-scale databases, *i.e.*, Waterloo Exploration Database [19] and PASCAL VOC 2012 [21]. Waterloo Exploration Database contains $4,744$ pristine images covering various image content. It also provides source code to synthesize four common distortions, *i.e.*, JPEG compression, JPEG2000 compression, Gaussian blur and while Gaussian noise at five degradation levels from the pristine images. PASCAL VOC 2012 is a large database for object recognition, detection and semantic segmentation. It contains $17,125$ images of acceptable quality covering 20 semantic classes. We merge the two databases to a total of $21,869$ source images. In addition to the four common distortion types mentioned above, we add five more — pink noise, contrast stretching, image quantization with color dithering, over-exposure, and under-exposure. Since some source images (especially in PASCAL VOC 2012) may not have perfect quality, we only include synthesized distorted images in the pre-training set and make sure that the added distortions dominate the perceived quality. Following [19], we synthesize distorted images with five degradation levels except for over-exposure and under-exposure, for which only two levels are generated [32]. Sample distorted images are shown in Fig. 1 and the degradation levels of the five new distortion types are shown in Fig. 2. In summary, the pre-training set contains $852,891$ distorted

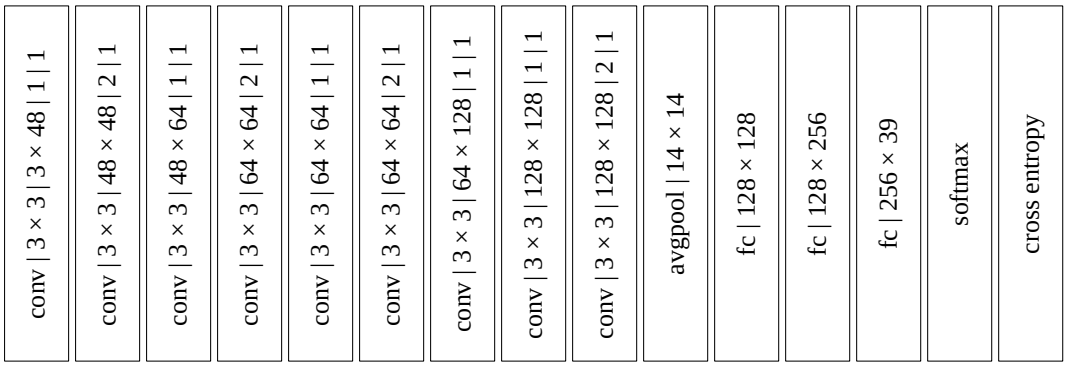

Fig. 3. The architecture of S-CNN for synthetic distortions. We follow the style and convention in [2], and denote the parameterization of the convolutional layer as "height × width | input channel × output channel | stride | padding". For brevity, we ignore all ReLU layers here.

images.

Due to the scale of the pre-training set, it is far from realistic to carry out a full subjective test to obtain a MOS for each image. We resolve this problem by taking advantage of the distortion type and level information used in the synthesis process, and pre-train the network to classify the distortion type and meanwhile identify the degradation level. Compared to previous methods that only exploit distortion type information [11], [18], [20], our pre-training strategy offers better initializations, leading to better local optimum (shown in Section IV-B5). Specifically, we form the ground truth label for pre-training as a 39-class indicator vector with only one entry activated to encode the underlying distortion type at the specific distortion level. The dimension of the ground truth vector comes from the fact that there are seven distortion types with five levels and two distortion types with two levels.

Inspired by the simple architecture design of VGG-16 network [23], we design our CNN for synthetic distortions (S-CNN) with a similar philosophy subject to some modifications. The network architecture is detailed in Fig. 3. In a nutshell, the size of the input RGB image is cropped to 224×224. All convolutions have a kernel size of 3×3. Zero padding is adopted to keep the resolution of feature activations. We adopt rectified linear unit (ReLU) as the nonlinear activation function since it delivers reliable performance in many computer vision applications [23], [33]. Although generalized divisive nomarlization (GDN) demonstrates promising performance in MEON [11] with lower depths and fewer parameters, considering that our S-CNN is a deeper network with more parameters, we opt to use ReLU for its simplicity and effectiveness in accelerating the training of deep neural networks [34]. Spatial max-pooling is replaced by the strided convolution with a step of two such that the spatial resolution is reduced by half in both directions. The feature activations at the last convolutional layer are averaged into a single feature vector followed by fully connected layers. All model parameters are collectively denoted by $\mathbf{W}$. The softmax function and the cross entropy loss are considered here for training. Specially, given N training data tuples $\{(\mathbf{X}^{(1)}, \mathbf{p}^{(1)}), ..., (\mathbf{X}^{(N)}, \mathbf{p}^{(N)})\}$, where $\mathbf{X}^{(i)}$ denotes the i-th raw input RGB image and $\mathbf{p}^{(i)}$ is the ground-truth multi-class indicator vector. By denoting the i-th activation value of the last fully connected layer of the k-th input image as $y_i^{(k)}$, the softmax function is defined as

$$\hat{p}_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}) = \frac{\exp\left(y_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W})\right)}{\sum_{j=1}^{39} \exp\left(y_j^{(k)}(\mathbf{X}^{(k)}; \mathbf{W})\right)}, \quad (1)$$

where $\hat{\mathbf{p}}^{(k)} = [\hat{p}_1^{(k)}, \cdots, \hat{p}_{39}^{(k)}]^T$ is a 39-dimensional probability vector of the k-th input in a mini-batch, which indicates the probability of each distortion type at the specific degradation level. The empirical cross entropy loss is computed by

$$\ell_s(\{\mathbf{X}^{(k)}\}; \mathbf{W}) = -\sum_{k=1}^{N} \sum_{i=1}^{39} p_i^{(k)} \log \hat{p}_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}). \quad (2)$$

B. CNN for Authentic Distortions

Unlike training S-CNN for synthetic distortions, where special strategies (such as the one used in Section III-A) may be employed to produce a large amount of training data, it is difficult to obtain sufficient ground truth data to train a CNN for authentic distortions from scratch, on the other hand, limited number of labeled training data often leads to overfitting problem. Here we opt to a CNN, namely VGG-16 [23] that has been pre-trained for the image classification task on ImageNet [12], to extract relevant features for authentically distorted image. The hypothesis is that the VGG-16 feature representations can adapt to authentic distortions because the distortions in ImageNet occur as a natural consequence of photography rather than simulations. As a result, features trained from such a data set are likely to improve the classification performance [13].

C. DB-CNN by Bilinear Pooling

We consider bilinear techniques to combine S-CNN for synthetic distortions and VGG-16 for authentic distortions into a single model. Bilinear models have been shown to be effective in modeling two-factor variations, such as style and content of images [35], location and appearance for fine-grained recognition [24], temporal and spatial aspects for video analysis [36], text and visual features for question-answering [37], and flow and image features for action recognition [38]. Here we tackle the BIQA problem with a

similar philosophy, where synthetic and authentic distortions are modeled as the two-factor variations, resulting in a DB-CNN model.

The structure of DB-CNN is presented in Fig. 4. We tailor the pre-trained S-CNN and VGG-16 by discarding all layers after the last convolution. Given an input image $\mathbf{X}$ and its activations of the last convolutional layers of the two streams, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are with size of $h_1 \times w_1 \times d_1$ and $h_2 \times w_2 \times d_2$, respectively. The bilinear pooling of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ requires $h_1 \times w_1 = h_2 \times w_2$, which holds in our case for an input image of arbitrary size because S-CNN and VGG-16 share the same padding and downsampling routines. We use VGG-16 mainly due to the fact that the design of S-CNN is inspired by VGGNet for its conciseness and effectiveness, which brings convenience to hold $h_1 \times w_1 = h_2 \times w_2$ as required by the intrinsic characteristic of bilinear pooling. Other CNNs such as ResNet [39] may also be adopted in our framework if the structure of S-CNN is adjusted appropriately. The bilinear pooling of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ is formulated as

$$\mathbf{B} = \mathbf{Y}_1^T \mathbf{Y}_2, \tag{3}$$

where the outer product $\mathbf{B}$ is of dimension $d_1 \times d_2$.

Bilinear representation is usually mapped from Riemannian manifold into an Euclidean space [40] using signed square root and ℓ_2 normalization [41]:

$$\tilde{\mathbf{B}} = \frac{\text{sign}(\mathbf{B}) \odot \sqrt{|\mathbf{B}|}}{\|\text{sign}(\mathbf{B}) \odot \sqrt{|\mathbf{B}|}\|_2}, \tag{4}$$

where $\odot$ refers to element-wise multiplication. $\tilde{\mathbf{B}}$ is fed to a fully connected layer for quality prediction, which produces an overall quality score. We consider the ℓ_2 norm as the empirical loss, which is widely used in previous works [10], [13], [26] to fine-tune the whole DB-CNN on a target IQA database

$$\ell = \frac{1}{N} \sum_{i=1}^{N} \|s_i - \hat{s}_i\|_2, \tag{5}$$

where s_i is the ground truth subjective quality score of the i-th image in a mini-batch and $\hat{s}_i$ is the predicted quality score by the proposed DB-CNN.

According to the chain rule, the backward propagation of the loss ℓ through the bilinear pooling layer to $\mathbf{Y}_1$ and $\mathbf{Y}_2$ can be computed by

$$\frac{\partial \ell}{\partial \mathbf{Y}_1} = \mathbf{Y}_2 \left(\frac{\partial \ell}{\partial \mathbf{B}} \right)^T \tag{6}$$

and

$$\frac{\partial \ell}{\partial \mathbf{Y}_2} = \mathbf{Y}_1 \left(\frac{\partial \ell}{\partial \mathbf{B}} \right). \tag{7}$$

It is worth noting that bilinear pooling is a global strategy and therefore DB-CNN accepts an input image of arbitrary size. As a result, we can directly feed the whole image instead of patches cropped from it into DB-CNN during both training and testing.

IV. Experiments

In this section, we first describe the experimental setups, including IQA databases, evaluation protocols, performance criteria, and implementation details of DB-CNN. After that, we compare the performance of DB-CNN with state-of-the-art BIQA models on individual databases and cross databases. We also test the robustness of DB-CNN on the Waterloo Exploration Database using discriminability and rating consistency testing criteria. Finally, we conduct several critical ablation experiments to justify the rationality of DB-CNN.

A. Experimental Setups

1) IQA Databases: The main experiments are conducted on three legacy singly synthetic IQA databases, *i.e.*, LIVE [15], CSIQ [42] and TID2013 [16] along with a multiply distorted synthetic dataset LIVE MD [43] and the authentic LIVE Challenge database [14]. LIVE [15] contains 779 distorted images synthesized from 29 reference images covering five distortion types—JPEG compression (JPEG), JPEG2000 compression (JP2K), Gaussian blur (GB), white Gaussian noise (WN) and fast fading error (FF) at seven to eight degradation levels. Difference MOS (DMOS) is collected with a higher value indicating lower perceptual quality, roughly in the range $[0, 100]$. CSIQ [42] is composed of 866 distorted images generated from 30 reference images, including six distortion types, *i.e.*, JPEG, JP2K, GB, WN, contrast change (CG), and pink noise (PN) at three to five degradation levels. DMOS in the range $[0, 1]$ is provided as the ground truth. TID2013 [16] consists of $3,000$ distorted images from 25 reference images with 24 distortion types at five degradation levels. MOS in the range $[0, 9]$ is provided to indicate the perceptual quality. LIVE MD [43] contains 450 images generated from 15 source images with corruption under two multiple distortion scenarios, *i.e.*, blur followd by JPEG compression and blur followed by white Gaussian noise. DMOS in the range$[0, 100]$ is provideds as the subjective quality score for each image. LIVE Challenge [14] is an authentic IQA database, which contains $1,162$ images captured from diverse real-world scenes by numerous photographers with various levels of photography skills using different camera devices, and hence undergo complex realistic distortions. MOS in the range $[0, 100]$ is collected from over $8,100$ unique human evaluators via an online crowdsourcing platform.

2) Experimental Protocols and Performance Criteria: We conduct experiments by following the same protocol in [13]. Specifically, for synthetic databases LIVE, CSIQ, TID2013 and LIVE MD, distorted images are divided into two splits, 80% of which are used for fine-tuning the DB-CNN and the rest 20% for testing. The splitting is conducted according to source images to guarantee the independence of image content. The training and testing procedures are randomly repeated ten times on all databases.

We adopt two commonly used metrics to benchmark the models: Spearman rank order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC). SRCC measures the prediction monotonicity and PLCC measures

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2886771, IEEE Transactions on Circuits and Systems for Video Technology

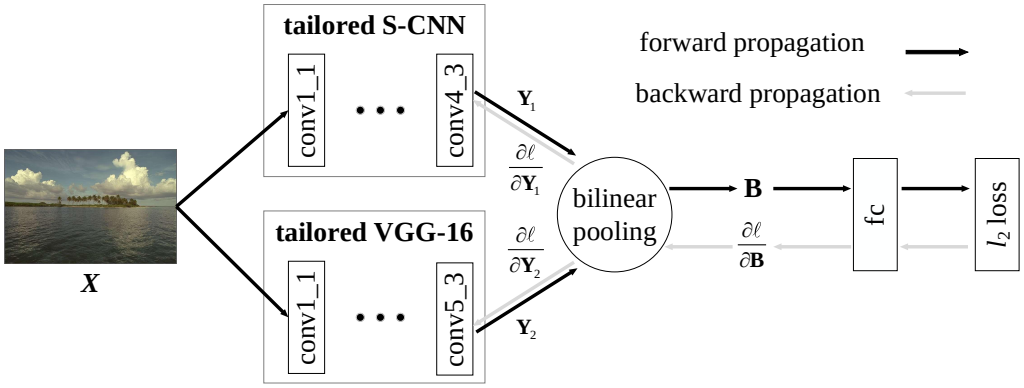

Fig. 4. The structure of the proposed DB-CNN.

prediction precision. As suggested in [44], the predicted quality scores are passed through a nonlinear logistic mapping function before computing PLCC:

$$\tilde{s} = \beta_1 \left(\frac{1}{2} - \frac{1}{\exp(\beta_2(\hat{s} - \beta_3))} \right) + \beta_4 \hat{s} + \beta_5, \qquad (8)$$

where $\{\beta_i; i = 1, 2, 3, 4, 5\}$ are regression parameters to be fitted. SRCC and PLCC results from the ten sessions are reported.

3) Implementation Details: All parameters of S-CNN are initialized with the method introduced in [33] and trained from scratch using the Adam optimization algorithm [45] with a mini-batch of 64. We run 30 epoches with a learning rate decaying logarithmically in the interval $[10^{-2}, 10^{-4}]$. Batch normalization [46] is used to assure the stability during training. Images are scaled to $256 \times 256 \times 3$ and we randomly crop $224 \times 224 \times 3$ patches as inputs.

During fine-tuning of DB-CNN, we again adopt Adam [45] with a learning rate of 10^{-6} for LIVE [15] and CSIQ [42], 10^{-5} for TID2013 [16], LIVE MD [43] and LIVE Challenge [14], respectively. The mini-batch size is set to eight. We feed images of original size to DB-CNN during both fine-tuning and testing phases.

DB-CNN is implemented using the MatConvNet toolbox [47] and will be made publicly available at github.com/zwx8981/BIQA_project.

TABLE I
AVERAGE SRCC AND PLCC RESULTS ACROSS TEN SESSIONS. THE TOP TWO RESULTS ARE HIGHLIGHTED IN BOLDFACE

SRCC	LIVE [15]	CSIQ [42]	TID2013 [16]	LIVE MD [43]	LIVE Challenge [14]
BRISQUE [7]	0.939	0.746	0.604	0.886	0.608
M3 [48]	0.951	0.795	0.689	0.892	0.607
FRIQUEE [22]	0.940	0.835	0.680	**0.923**	0.682
CORNIA [8]	0.947	0.678	0.678	0.899	0.629
HOSA [49]	0.946	0.741	0.735	0.913	0.640
Le-CNN [10]	0.956	—	—	—	—
BIECON [17]	0.961	0.815	0.717	0.909	0.595
DIQaM-NR [26]	0.960	—	**0.835**	—	0.606
WaDIQaM-NR [26]	0.954	—	0.761	—	0.671
ResNet-ft [13]	0.950	**0.876**	0.712	0.909	**0.819**
IW-CNN [13]	**0.963**	0.812	0.800	0.914	0.663
DB-CNN	**0.968**	**0.946**	**0.816**	**0.927**	**0.851**
PLCC	LIVE	CSIQ	TID2013	LIVE MD	LIVE Challenge
BRISQUE [7]	0.935	0.829	0.694	0.917	0.629
M3 [48]	0.950	0.839	0.771	0.919	0.630
FRIQUEE [22]	0.944	0.874	0.753	**0.934**	0.705
CORNIA [8]	0.950	0.776	0.768	0.921	0.671
HOSA [49]	0.947	0.823	0.815	0.926	0.678
Le-CNN [10]	0.953	—	—	—	—
BIECON [17]	0.962	0.823	0.762	0.933	0.613
DIQaM-NR [26]	**0.972**	—	**0.855**	—	0.601
WaDIQaM-NR [26]	0.963	—	0.787	—	0.680
ResNet-ft [13]	0.954	**0.905**	0.756	0.920	**0.849**
IW-CNN [13]	0.964	0.791	0.802	0.929	0.705
DB-CNN	**0.971**	**0.959**	**0.865**	**0.934**	**0.869**

B. Experimental Results

1) Performance on Individual Databases: We compare the proposed model against several state-of-the-art BIQA methods: BRISQUE [7], M3 [48], FRIQUEE [22], CORNIA [8], HOSA [49], and dipIQ [27], whose source codes are provided by the respective authors. We re-train and/or validate using the same randomly generated training-testing splits. For deep learning-based counterparts, we directly report the performance in the corresponding papers due to the unavailability of the training codes. SRCC and PLCC results on the four databases are listed in Table I, from which we obtain several interesting observations. First, while all competing models achieve comparable performance on LIVE [15], their performance on CSIQ [42] and TID2013 [16] are rather diverse. Compared with classical domain knowledge-based models, CNN-based models deliver better performance on CSIQ and TID2013, which we believe arises from the end-to-end feature learning in replacement of hand-crafted feature engineering. Second, as for the multiply distorted image dataset LIVE MD, DB-CNN also delivers better performance against other methods although it dose not incorporate any multiply distorted image in the pre-training set. This suggests that DB-CNN generalizes well to slightly different distortion scenarios. Last, as for the authentic database LIVE Challenge, FRIQUEE [22] that combines a set of quality-aware features extracted from multiple color spaces outperforms other classical BIQA models and all CNN-based models except for ResNet-ft [13] and the proposed DB-CNN. It manifests that the intrinsic characteristics of authentic distortions cannot be fully captured by low-level features learned from synthetically

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2886771, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II
AVERAGE SRCC AND PLCC RESULTS OF INDIVIDUAL DISTORTION TYPES ACROSS TEN SESSIONS ON LIVE [15]

SRCC	JPEG	JP2K	WN	GB	FF
BRISQUE [7]	0.965	0.929	0.982	**0.964**	0.828
M3 [48]	0.966	0.930	**0.986**	0.935	0.902
FRIQUEE [22]	0.947	0.919	**0.983**	0.937	0.884
CORNIA [8]	0.947	0.924	0.958	0.951	0.921
HOSA [49]	0.954	0.935	0.975	**0.954**	**0.954**
dipIQ [27]	**0.969**	**0.956**	0.975	0.940	—
DB-CNN	**0.972**	**0.955**	0.980	0.935	**0.930**
PLCC	JPEG	JP2K	WN	GB	FF
BRISQUE [7]	0.971	0.940	0.989	**0.965**	0.894
M3 [48]	0.977	0.945	**0.992**	0.947	0.920
FRIQUEE [22]	0.955	0.935	**0.991**	0.949	0.936
CORNIA [8]	0.962	0.944	0.974	0.961	0.943
HOSA [49]	0.967	0.949	0.983	**0.967**	**0.967**
dipIQ [27]	**0.980**	**0.964**	0.983	0.948	—
DB-CNN	**0.986**	**0.967**	0.988	0.956	**0.961**

TABLE III
AVERAGE SRCC AND PLCC RESULTS OF INDIVIDUAL DISTORTION TYPES ACROSS TEN SESSIONS ON CSIQ [42]

SRCC	JPEG	JP2K	WN	GB	PN	CC
BRISQUE [7]	0.806	0.840	0.723	0.820	0.378	0.804
M3 [48]	0.740	0.911	0.741	0.868	0.663	0.770
FRIQUEE [22]	0.869	0.846	0.748	0.870	**0.753**	**0.838**
CORNIA [8]	0.513	0.831	0.664	0.836	0.493	0.462
HOSA [49]	0.733	0.818	0.604	0.841	0.500	0.716
dipIQ [27]	0.936	**0.944**	0.904	**0.932**	—	—
MEON [11]	**0.948**	0.898	**0.951**	0.918	—	—
DB-CNN	**0.940**	**0.953**	**0.948**	**0.947**	**0.940**	**0.870**
PLCC	JPEG	JP2K	WN	GB	PN	CC
BRISQUE [7]	0.828	0.887	0.742	0.891	0.496	0.835
M3 [48]	0.768	0.928	0.728	0.917	0.717	0.787
FRIQUEE [22]	0.885	0.883	0.778	0.905	**0.769**	**0.864**
CORNIA [8]	0.563	0.883	0.687	0.904	0.632	0.543
HOSA [49]	0.759	0.899	0.656	0.912	0.601	0.744
dipIQ [27]	0.975	**0.959**	0.927	**0.958**	—	—
MEON [11]	**0.979**	0.925	**0.958**	0.946	—	—
DB-CNN	**0.982**	**0.971**	**0.956**	**0.969**	**0.950**	**0.895**

distorted images. The success of DB-CNN on LIVE Challenge verifies the effectiveness of employing more relevant features from VGG-16 to measure the severity of authentic distortions. In summary, the proposed DB-CNN model achieves state-of-the-art performance on both synthetic and authentic IQA databases.

2) Performance on Individual Distortion Types: To take a closer look at the behaviors of DB-CNN on individual distortion types along with several competing BIQA models, we train models using images with all kinds of distortion types and test them on a specific distortion type. Table II, III, and IV show the results on LIVE [15], CSIQ [42], and TID2013 [16], respectively, where we can observe that DB-CNN is among the top two performing models 34 out of 46 times, showing a significant advantage. Specifically, on LIVE, DB-CNN does not perform well on FF, which we believe is caused by its absence during the construction of the pre-training set. As for CSIQ, DB-CNN outperforms other counterparts by a large margin especially on pink noise and contrast change, which validates the effectiveness of pre-training S-CNN, a stream of DB-CNN. On the most challenging synthetic database TID2013, all BIQA models fail to deliver satisfactory performance on three distortion types, *i.e.*, non-eccentricity pattern noise, local block-wise distortions, and mean shift. DB-CNN performs relatively better on contrast change, which is consistent with the results on CSIQ and change of color saturation, which is attributed to its feature extraction from color images. Although we do not synthesize as many distortion types as in TID2013, an interesting finding is that DB-CNN still performs well on distortion types with similar artifacts that have been contained in our pre-training set. To be specific, as shown in Fig. 5, grainy noise ubiquitously exists in images distorted by additive Gaussian noise, additive noise in color components, and high frequency noise; Gaussian blur, image denoising, and sparse sampling and reconstruction mainly introduce blur; image color quantization with dither and quantization noise also share similar appearances. Trained by synthesized images with distortions of additive Gaussian noise, Gaussian blur, and image color quantization with dither, DB-CNN well generalizes to unseen distortions with similar perceived artifacts.

3) Performance across Different Databases: Robust BIQA models are expected to not only perform well on the training database, but also generalize well to other IQA databases. In this subsection, we conduct cross database validations to compare the generalizability of DB-CNN against BRISQUE [7], M3 [48], FRIQUEE [22], CORNIA [8], and HOSA [49]. The results of CNN-based counterparts are reported if available from the original papers. All experiments are conducted by training models on one entire database and test them on the other databases. SRCC results are reported in Table V. It is expected that models trained on LIVE are much easier to generalize to CSIQ and vice versa than other cross database pairs. As for training on TID2013 and testing on the other two synthetic databases, the proposed DB-CNN performs superior to other models. Unfortunately, it is evident that models trained on synthetic databases are difficult to generalize to the LIVE Challenge authentic database or vice versa. This shows different intrinsic characteristics between synthetic and authentic distortions. Despite this, DB-CNN still achieves higher prediction accuracies than all other models under such a challenging experimental setup, which justifies the effectiveness of the proposed method.

4) Results on the Waterloo Exploration Database: Although SRCC and PLCC have been widely used as the performance criteria in IQA research, they cannot be applied to arbitrarily large-scale databases due to the absence of ground truth MOS labels of all images. Three testing criteria are introduced along with the large-scale Waterloo Exploration Database in [19], *i.e.*, Pristine/Distorted Image Discriminability Test (D-Test), Listwise Ranking Consistency Test (L-Test), and Pairwise Preference Consistency Test (P-Test), which measure the ability of BIQA models in discriminating distorted from pristine images, rating images with the same content and the same distortion type but different degradation levels in a consistent rank, and predicting concordance with pairs of images whose quality is clearly discriminable, respectively. More details of these criteria can be found in [19]. Here we examine the robustness of the proposed DB-CNN model using these criteria on the Waterloo Exploration Database. We first

TABLE IV
AVERAGE SRCC RESULTS OF INDIVIDUAL DISTORTION TYPES ACROSS TEN SESSIONS ON TID2013 [16]. WE OBTAIN SIMILAR RESULTS USING PLCC AS THE PERFORMANCE METRIC

SRCC	BRISQUE [7]	M3 [48]	FRIQUEE [22]	CORNIA [8]	HOSA [49]	MEON [11]	DB-CNN
Additive Gaussian noise	0.711	0.766	0.730	0.692	**0.833**	**0.813**	0.790
Additive noise in color components	0.432	0.560	0.573	0.137	0.551	**0.722**	**0.700**
Spatially correlated noise	0.746	0.782	**0.866**	0.741	0.842	**0.926**	0.826
Masked noise	0.252	0.577	0.345	0.451	0.468	**0.728**	**0.646**
High frequency noise	0.842	**0.900**	0.847	0.815	0.897	**0.911**	0.879
Impulse noise	0.765	0.738	0.730	0.616	**0.809**	**0.901**	0.708
Quantization noise	0.662	**0.832**	0.764	0.661	0.815	**0.888**	0.825
Gaussian blur	0.871	**0.896**	0.881	0.850	0.883	**0.887**	0.859
Image denoising	0.612	0.709	0.839	0.764	**0.854**	0.797	**0.865**
JPEG compression	0.764	0.844	0.813	0.797	**0.891**	0.850	**0.894**
JPEG2000 compression	0.745	0.885	0.831	0.846	**0.919**	0.891	**0.916**
JPEG transmission errors	0.301	0.375	0.498	0.694	0.730	**0.746**	**0.772**
JPEG2000 transmission errors	**0.748**	0.718	0.660	0.686	0.710	0.716	**0.773**
Non-eccentricity pattern noise	**0.269**	0.173	0.076	0.200	0.242	0.116	**0.270**
Local bock-wise distortions	0.207	0.379	0.032	0.027	0.268	**0.500**	**0.444**
Mean shift	0.219	0.119	**0.254**	**0.232**	0.211	0.177	-0.009
Contrast change	-0.001	0.155	**0.585**	0.254	0.362	0.252	**0.548**
Change of color saturation	0.003	-0.199	0.589	0.169	0.045	**0.684**	**0.631**
Multiplicative Gaussian noise	0.717	0.738	0.704	0.593	**0.768**	**0.849**	0.711
Comfort noise	0.196	0.353	0.318	0.617	**0.622**	0.406	**0.752**
Lossy compression of noisy images	0.609	0.692	0.641	0.712	**0.838**	0.772	**0.860**
Color quantization with dither	0.831	**0.908**	0.768	0.683	**0.896**	0.857	0.833
Chromatic aberrations	0.615	0.570	0.737	0.696	**0.753**	**0.779**	0.732
Sparse sampling and reconstruction	0.807	0.893	0.891	0.865	**0.909**	0.855	**0.902**

TABLE V
SRCC RESULTS IN A CROSS DATABASE SETTING

Training	LIVE [15]			CSIQ [42]		
Testing	CSIQ	TID2013	LIVE Challenge	LIVE	TID2013	LIVE Challenge
BRISQUE [7]	0.562	0.358	0.337	0.847	0.454	0.131
M3 [48]	0.621	0.344	0.226	0.797	0.328	0.183
FRIQUEE [22]	**0.722**	0.461	0.411	**0.879**	**0.463**	0.264
CORNIA [8]	0.649	0.360	0.443	0.853	0.312	**0.393**
HOSA [49]	0.594	0.361	**0.463**	0.773	0.329	0.291
DIQaM-NR [26]	0.681	0.392	—	—	—	—
WaDIQaM-NR [26]	0.704	**0.462**	—	—	—	—
DB-CNN	**0.758**	**0.524**	**0.567**	**0.877**	**0.540**	**0.452**
Training	TID2013 [16]			LIVE Challenge [14]		
Testing	LIVE	CSIQ	LIVE Challenge	LIVE	CSIQ	TID2013
BRISQUE [7]	0.790	0.590	0.254	0.238	0.241	0.280
M3 [48]	**0.873**	0.605	0.112	0.059	0.109	0.058
FRIQUEE [22]	0.755	0.635	0.181	**0.644**	**0.592**	**0.424**
CORNIA [8]	0.846	0.672	0.293	0.588	0.446	0.403
HOSA [49]	0.846	0.612	**0.319**	0.537	0.336	0.399
DIQaM-NR [26]	—	0.717	—	—	—	—
WaDIQaM-NR [26]	—	**0.733**	—	—	—	—
DB-CNN	**0.891**	**0.807**	**0.457**	**0.746**	**0.697**	**0.424**

retrain the S-CNN stream using distorted images generated from PASCAL VOC 2012 only to ensure the independence of image content between training and testing. Experimental results are tabulated in Table VI, where we observe that DB-CNN achieves the best two results in D-Test and P-Test, and is competitive in L-Test.

We also conduct gMAD competition games [25] on Waterloo Exploration Database [19]. Evaluating in a large-scale dataset is more credible in real-world application to overcome the contradiction between the high-dimensional inner charateristic of digital images and the extremely limited sample space of tradional IQA datasets, which only contain at most a few thousands images covering very limited content variations. gMAD competition is a preferable way to evaluate an IQA model since it can automatically and most efficiently select the optimal test image pairs from a large-scale image dataset such as Waterloo Exploration Database [19] and let the model competes against other opponents. gMAD extends the idea of MAximum Differentiation (MAD) competition [50] that one counter-example is sufficient to disprove a model by allowing a group of models for competition and by finding the optimal stimuli in a large database [25]. Image pairs are automatically generated by searching for the maximum quality difference by an aggressive model (attacker), while keeping predictions of another resistant model (defender). To be specific, DB-CNN first plays the role of attacker while deepIQA [26] a defender. Then the procedure is repeated with the roles of two models exchanged. As shown in Fig. 6 (a)-(d), deepIQA [26] considers

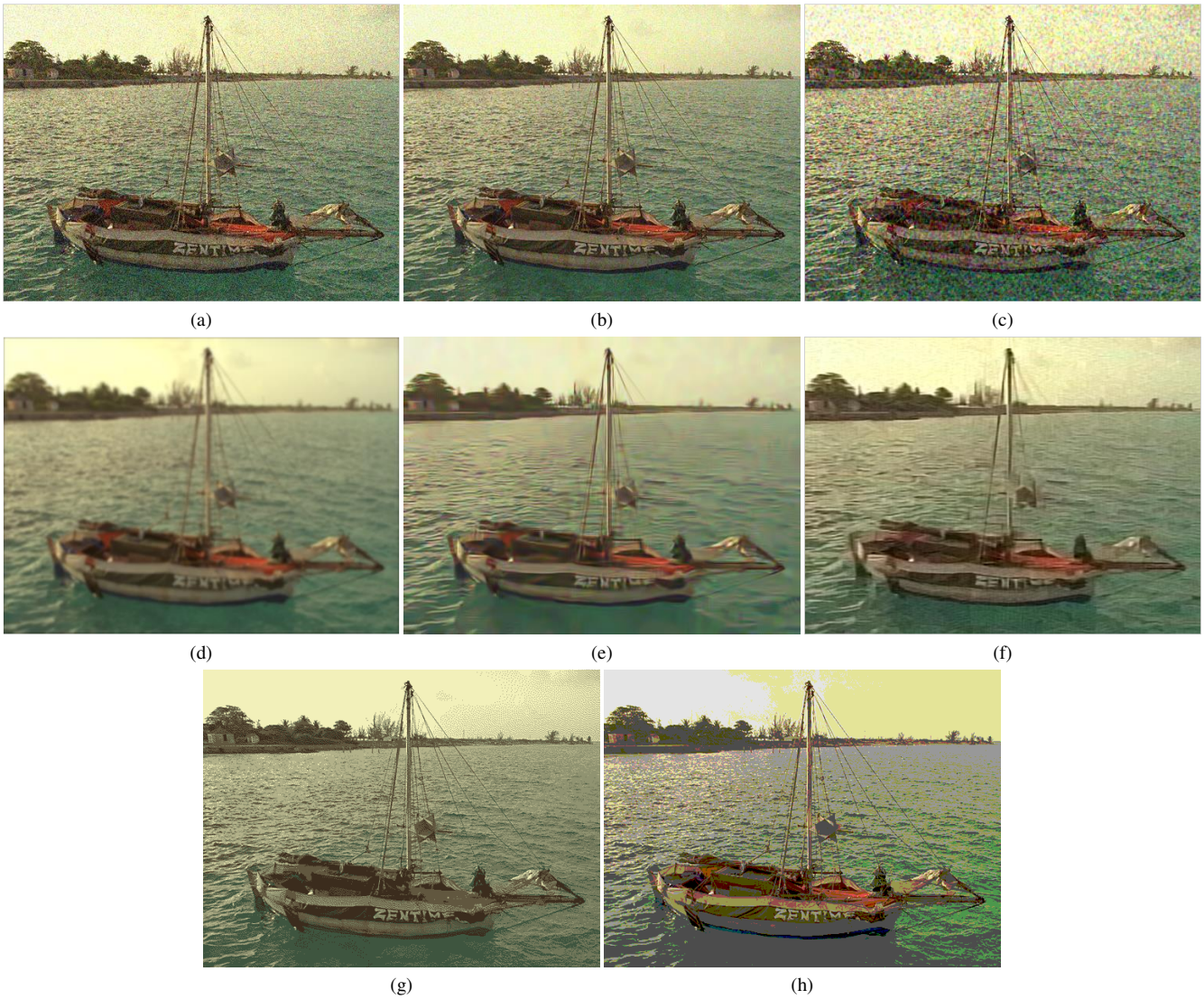

Fig. 5. Images with different distortion types may share similar distorted appearances. (a) Additive Gaussian noise. (b) Additive noise in color components. (c) High frequency noise. (d) Gaussian blur. (e) Image denoising. (f) Sparse sampling and reconstruction. (i) Image color quantization with dither. (j) Quantization noise.

pairs (a) and (b) of the same quality at low- and high-quality level respectively, which are obviously not in agreement with the perceptual quality. On the contrary, DB-CNN predicts much better quality of top images in pairs (a) and (b), which is closer to human subjective opinions. As for (c) and (d), with the roles exchanged, deepIQA [26] fails to falsify DB-CNN, which shows a better resistance of DB-CNN. We then let DB-CNN fights against MEON [11], pairs of which are shown in Fig. 7 (a)-(d). We can observe from (a) and (c) that both DB-CNN and MEON are able to fail each other at low-quality level by finding strong counter-examples. Specificially, DB-CNN fails to disprove MEON [11] in (a), which reveal its weakness in BLUR and conversely, MEON [11] does not handle JP2K well enough, which leads to the successful defend of DB-CNN in pair (c). As for high quality pair of (b), DB-CNN fails MEON [11] by finding the bottom image of (b) to have apparently lower quality than the top one. On the other hand, DB-CNN also successfully defends attack from MEON [11] in pair (d), which has two images with similar perceptual quality.

5) Ablation Experiments: In order to evaluate the design rationality of DB-CNN, we conduct several ablation experiments with setups and protocols following Section IV-A. We first work with a baseline version, where only one stream (either S-CNN and VGG-16) is included. The bilinear pooling is kept, which turns out to be the outer-product of the activations of the last convolutional layer with themselves. We then replace the bilinear pooling module with a simple feature concatenation and ensure that the number of parameters of the subsequent fully connected layer is approximately the same as in DB-CNN. From Table VII, we observe that S-CNN and VGG-16 can only deliver promising performance on synthetic and authentic databases, respectively. By contrast, DB-CNN is capable of simultaneously handling synthetic and authentic distortions. We also train two DB-CNN models, one from scratch and the other using the distortion type

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2886771, IEEE Transactions on Circuits and Systems for Video Technology

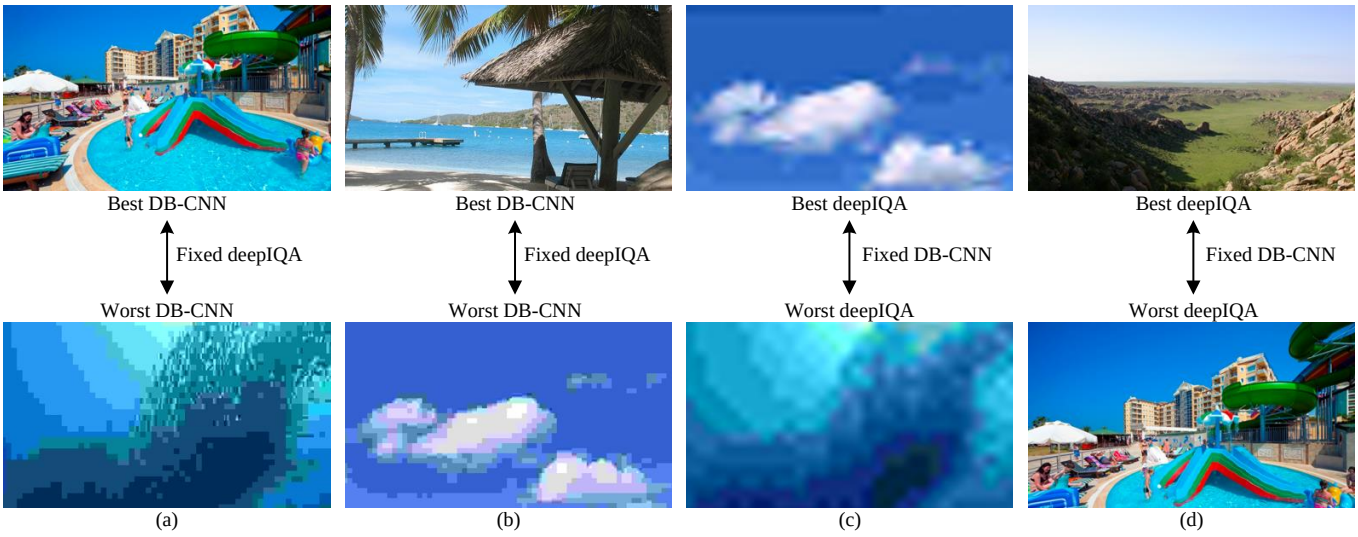

Fig. 6. gMAD competition results between DB-CNN and deepIQA [26]. (a) Fixed deepIQA at the low-quality level. (b) Fixed deepIQA at the high-quality level. (c) Fixed DB-CNN at the low-quality level. (d) Fixed DB-CNN at the high-quality level.

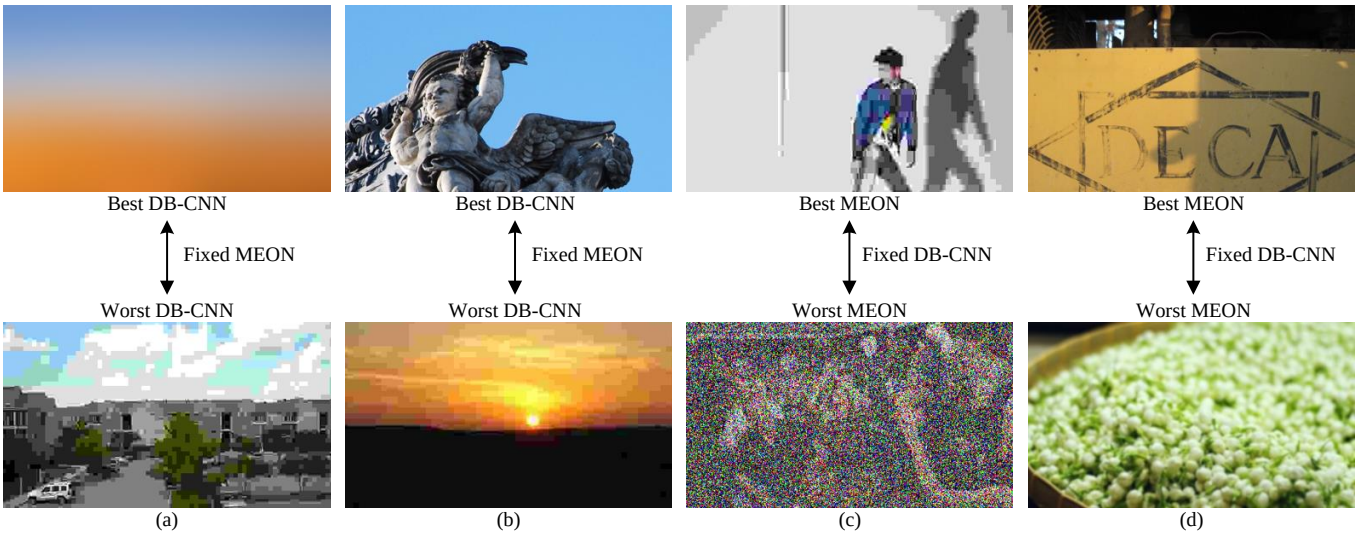

Fig. 7. gMAD competition results between DB-CNN and MEON [11]. (a) Fixed MEON at the low-quality level. (b) Fixed MEON at the high-quality level. (c) Fixed DB-CNN at the low-quality level. (d) Fixed DB-CNN at the high-quality level.

information only during pre-training S-CNN, to validate the necessity of pre-trained stages. From the table, we observe that with more meaningful initializations, DB-CNN achieves better performance.

V. Conclusion

We propose a deep bilinear CNN-based BIQA model for both synthetic and authentic distortions by conceptually modeling them as two-factor variations followed by bilinear pooling. DB-CNN demonstrates state-of-the-art performance on both synthetic and authentic IQA databases, which we believe arises from the two-steam architecture for variation modeling, pre-training for better initializations, and bilinear pooling for meaningful feature blending. In addition, through validations across different databases, experiments on the Waterloo Exploration Database, and results from the gMAD competition, we have shown the scalability, generalizability, and robustness of the proposed DB-CNN model.

DB-CNN is versatile and extensible. For example, more

TABLE VI
Results on the Waterloo Exploration Database [19]

Model	D-Test	L-Test	P-Test
BRISQUE [7]	0.9204	**0.9772**	0.9930
M3 [48]	0.9203	0.9106	0.9748
CORNIA [8]	0.9290	0.9764	0.9947
HOSA [49]	0.9175	0.9647	0.9983
dipIQ [27]	0.9346	**0.9846**	**0.9999**
deepIQA [26]	0.9074	0.9467	0.9628
MEON [11]	**0.9384**	0.9669	0.9984
DB-CNN	**0.9616**	0.9614	**0.9992**

TABLE VII
AVERAGE SRCC RESULTS OF ABLATION EXPERIMENTS ACROSS TEN SESSIONS. "SCRATCH" MEANS DB-CNN IS TRAINED FROM SCRATCH WITH RANDOM INITIALIZATIONS. "DISTYPE" MEANS THE S-CNN STREAM IS PRE-TRAINED TO CLASSIFY DISTORTION TYPES ONLY, IGNORING THE DISTORTION LEVEL INFORMATION

SRCC	LIVE [15]	CSIQ [42]	TID2013 [16]	LIVE Challenge [14]
S-CNN	**0.963**	**0.950**	**0.810**	0.680
VGG-16	0.943	0.824	0.758	**0.848**
Concatenation	0.951	0.856	0.701	0.811
DB-CNN scratch	0.875	0.541	0.488	0.625
DB-CNN distype	**0.963**	0.928	0.761	—
DB-CNN	**0.968**	**0.946**	**0.816**	**0.851**

distortion types and levels can be added into the pre-training set; more sophisticated designs of S-CNN and more powerful CNNs such as ResNet [39] can be utilized. One may also improve DB-CNN by considering other variants of bilinear pooling [51].

The current work deals with synthetic and authentic distortions separately by fine-tuning DB-CNN on either synthetic or authentic databases. How to extend DB-CNN toward a more unified BIQA model, especially in the early feature extraction stage, is an interesting direction yet to be explored.

REFERENCES

[1] A. C. Bovik, *Handbook of Image and Video Processing*. Academic Press, 2010.

[2] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *CoRR*, vol. abs/1611.01704, 2016. [Online]. Available: http://arxiv.org/abs/1611.01704

[3] Z. Duanmu, K. Ma, and Z. Wang, "Quality-of-experience of adaptive video streaming: Exploring the space of adaptations," in *ACM Multimedia*, 2017, pp. 1752–1760.

[4] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," in *Human Vision and Electronic Imaging*, 2015, pp. 1–11.

[5] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[6] ——, "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.

[7] 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.

[8] 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.

[9] K. Ma, "Blind image quality assessment: Exploiting new evaluation and design methodologies," Ph.D. dissertation, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, 2017.

[10] 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.

[11] K. Ma, K. Zeng, and Z. Wang, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[13] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[14] 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.

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

[16] 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.

[17] 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.

[18] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *IEEE International Conference on Image Processing*, 2015, pp. 2791–2795.

[19] 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. 26, no. 2, pp. 1004–1016, Feb. 2017.

[20] 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.

[21] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[22] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32–32, Jan. 2017.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[24] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[25] 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 Vision and Pattern Recognition*, 2016, pp. 1664–1673.

[26] S. Bosse, D. Maniry, K. R. Mller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[27] 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.

[28] P. Ye, "Feature learning and active learning for image quality assessment," Ph.D. dissertation, Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA, 2014.

[29] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2877–2884.

[30] 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.

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

[32] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[33] 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.

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

[35] J. B. Tenenbaum and W. T. Freeman, "Separating style and content," in *Advances in Neural Information Processing Systems*, 1997, pp. 662–668.

[36] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[37] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *CoRR*, vol. abs/1606.01847, 2016.
[38] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.
[39] 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.
[40] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, Jan. 2006.
[41] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, 2010, pp. 143–156.
[42] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 1–21, Jan. 2010.
[43] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Signals, Systems and Computers*, 2013, pp. 1693–1697.
[44] 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
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
[46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.
[47] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for Matlab," in *ACM International Conference on Multimedia*, 2015, pp. 689–692.
[48] 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 Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.
[49] 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.
[50] 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. 8–8, Sep. 2008.
[51] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 317–326.