End-to-End Blind Image Quality Prediction With Cascaded Deep Neural Network

Jinjian Wu, *Member, IEEE*, Jupo Ma, Fuhu Liang, Weisheng Dong, *Member, IEEE*, Guangming Shi, *Senior Member, IEEE*, and Weisi Lin, *Fellow, IEEE*

***Abstract*— The deep convolutional neural network (CNN) has achieved great success in image recognition. Many image quality assessment (IQA) methods directly use recognition-oriented CNN for quality prediction. However, the properties of IQA task is different from image recognition task. Image recognition should be sensitive to visual content and robust to distortion, while IQA should be sensitive to both distortion and visual content. In this paper, an IQA-oriented CNN method is developed for blind IQA (BIQA), which can efficiently represent the quality degradation. CNN is large-data driven, while the sizes of existing IQA databases are too small for CNN optimization. Thus, a large IQA dataset is firstly established, which includes more than one million distorted images (each image is assigned with a quality score as its substitute of Mean Opinion Score (MOS), abbreviated as pseudo-MOS). Next, inspired by the hierarchical perception mechanism (from local structure to global semantics) in human visual system, a novel IQA-orientated CNN method is designed, in which the hierarchical degradation is considered. Finally, by jointly optimizing the multilevel feature extraction, hierarchical degradation concatenation (HDC) and quality prediction in an end-to-end framework, the Cascaded CNN with HDC (named as CaHDC) is introduced. Experiments on the benchmark IQA databases demonstrate the superiority of CaHDC compared with existing BIQA methods. Meanwhile, the CaHDC (with about 0.73M parameters) is lightweight comparing to other CNN-based BIQA models, which can be easily realized in the microprocessing system. The dataset and source code of the proposed method are available at https://web.xidian.edu.cn/wjj/paper.html.**

***Index Terms*— Blind image quality assessment (BIQA), hierarchical degradation concatenation, end-to-end, deep convolutional neural network.**

I. Introduction

NOWADAYS, objective image quality assessment (IQA) plays an important role in image/video processing. Over the past several decades, a variety of IQA methods have been introduced and they can be divided into three categories: full-reference (FR) IQA (for which the whole reference image is required), reduced-reference (RR) IQA (which use partial reference information) and no-reference (NR) IQA (for which no reference information is required) [1]. In reality, however, the reference image is usually unavailable. Thus, NRIQA, also called blind IQA (BIQA), becomes a hot research topic.

Most early BIQAs, which belong to knowledge-driven method, need to design feature descriptors manually based on the properties of Human Visual System (HVS) or Natural Scene Statistics (NSS) [2]–[11]. However, it is hard to design handcrafted features which can efficiently represent the quality degradation for BIQA. Due to the powerful feature representation ability of convolutional neural network (CNN), some CNN-based BIQAs have been proposed recently (which belong to data-driven). These methods are mainly based on two ideas. One is to adopt existing pre-trained CNN models as feature or multilevel-feature extractor and SVR for quality predication [12]–[14], which cannot jointly optimize the whole framework as a whole. The other follows end-to-end manner for BIQA [15]–[19], like many CNN models used for image recognition task, in which only the features from the last convolutional layer are utilized. All these CNN-based BIQAs can not make full use of the perceptual properties of HVS. Although a lot of CNN frameworks have been designed and achieved great success in image recognition task, the characteristics of IQA is different from that of image recognition. Image recognition task should be sensitive to visual content and robust to distortion, while IQA task should be sensitive to both distortion and visual content. Many existing BIQA methods use recognition-oriented CNN for quality prediction, which may not be fully adapted. Thus, we need design an IQA-oriented CNN method for BIQA.

Moreover, a common problem of these CNN-based BIQAs is the lack of large quality-annotated IQA database which is needed to train a network with strong generalization ability. Thus, we firstly establish a large-scale IQA dataset which includes more than one million distorted images generated from ten thousands of high quality pristine images (with 21 distortion types under 5 levels). Since the performance of existing FRIQAs [20]–[24] is highly consistent with the HVS, the best FRIQAs for each distortion type are chosen and merged to set a quality score for each distorted image as its MOS, abbreviated as pseudo-MOS (the reliability of our proposed method assigning pseudo-MOS is experimentally verified in Section III).

In this work, in order to meet another challenge of lacking IQA-oriented CNN-based methods, we consider to make use of the hierarchical degradation during visual perception for BIQA framework designing. Research on neuroscience

Manuscript received August 12, 2019; revised January 7, 2020 and April 6, 2020; accepted June 9, 2020. Date of publication June 19, 2020; date of current version July 13, 2020. This work was supported in part by the NSF of China under Grant 61772388 and Grant 61632019 and in part by the National Key Research and Development Program of China under Grant 2018AAA0101400. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chaker Larabi. *(Corresponding author: Jinjian Wu.)*

Jinjian Wu, Jupo Ma, Fuhu Liang, Weisheng Dong, and Guangming Shi are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: jinjian.wu@mail.xidian.edu.cn).

Weisi Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798.

Digital Object Identifier 10.1109/TIP.2020.3002478

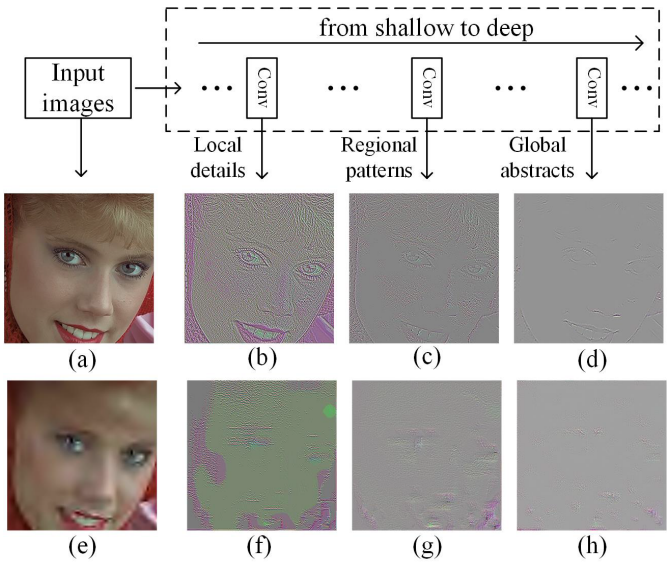


Fig. 1. Illustration of the effect of hierarchical degradation on image quality. (a) Reference image as input. (b)-(d) Visualization of features at different layers for the reference image. (e) Distorted image as input. (f)-(h) Visualization of features at different layers for the distorted image.

indicates the hierarchical process for visual perception [25], [26]. Coincidentally, CNN naturally learns hierarchical features (from low-level to high-level) with the depth of layers from shallow to deep. Low-level features concern more on local details, middle-level features mainly focus on regional patterns, and high-level features are rich in global abstracts. As shown in Fig.1, through the visualization of convolutional network[1] by the method of [27], we can observe that distortion affects different levels of features and causes hierarchical quality degradation from the perspective of IQA. For example, at low level, network focuses on local details as show in Fig.1(b), and the local details will be destroyed by distortion as show in Fig.1(f). The hierarchical degradation refers to the destruction on hierarchical features caused by distortion. It's necessary to consider the hierarchical degradation for IQA. Thus, we design an end-to-end cascaded CNN framework, in which the procedures of feature extraction, hierarchical degradation concatenation and quality prediction can be jointly optimized. Experiment results demonstrate the superiority of CaHDC compared with existing BIQA methods. It is worth mentioning that the number of parameters of CaHDC is far smaller than other CNN-based BIQAs.

The main contributions of this paper can be summarized as below:

1) A large-scale quality-annotated dataset is established to address the problem of limited training data, which spans a great diversity in visual contents and distortions. The pseudo-MOS assigned by our proposed method is reliable and comparable to subjective test.

2) Inspired by the hierarchical perception mechanism in HVS, we propose an IQA-oriented CNN method, in which the hierarchical degradations are concatenated for BIQA and the whole procedures, i.e., feature extraction, hierarchical degradation concatenation and quality regression can be optimized by an end-to-end manner. Benefiting from the hierarchical degradation concatenation and the end-to-end optimization, CaHDC can better learn the nature of quality degradation. Experimental results indicate that CaHDC achieves the state-of-the-art.

3) As a lightweight network with only 0.73M parameters, the proposed IQA model is easily realized in the microprocessing system (e.g., NVIDIA JETSON TX2), which can meet the requirements of accuracy and real-time.

[1]The CNN model used for visualization is the proposed CaHDC.

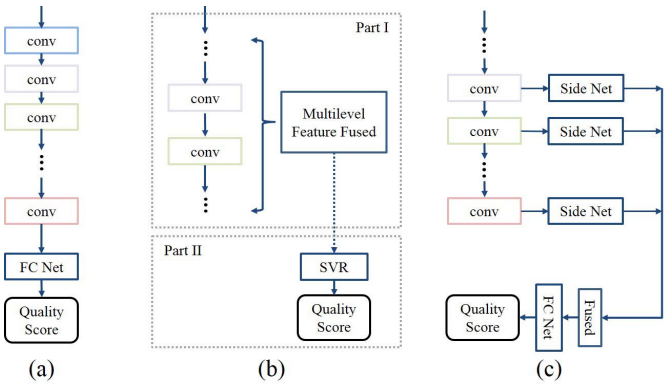


Fig. 2. Existing BIQA architectures: (a) End-to-end but not hierarchical framework. (b) Hierarchical but not end-to-end framework. (c) Our proposed end-to-end network simultaneously combining hierarchical degradation.

II. Related Work

A. Traditional Blind Image Quality Assessment

Traditional BIQA aims to design hand-crafted feature descriptors, which try to extract features that can efficiently represent the quality degradation. Then, the non-linear regression procedure (e.g., SVR) is adopted to regress the high dimension features into a quality score. The most classical approach is natural scene statistic (NSS) based BIQA. This kind of methods estimate the statistic distribution of natural image, and then capture parametric bias to evaluate image quality degradation. For example, DIIVINE [2] first distinguishes the distortion type, and then employs distortion-specific methods to acquire quality score by utilizing NSS features. BLIINDS-II [3] exploits a NSS model of discrete cosine transform (DCT) coefficients to evaluate image quality. BRISQUE [4] employs the generalized Gaussian distribution (GGD) to extract features in the spatial domain, and then SVR is employed to map feature space to quality score. NIQE [28] constructs quality aware features and matches them to the multivariate Gaussian (MVG) model. Besides, there are also other methods to design feature descriptors, such as RISE [8] imitates multiscale characteristic of HVS by learning multiscale features in both the spatial and spectral domains to evaluate the image sharpness.

B. CNN-Based Blind Image Quality Assessment

Over the past few years, with the outstanding performance of CNN in various visual tasks, some CNN-based BIQAs have been proposed. There are mainly two types of CNN-based BIQAs, whose architectures are shown in Fig.2 (a) and (b). The first type is end-to-end but no hierarchical degradation integrated as show in Fig.2 (a). For example, WaDIQaM [16] presents an end-to-end method for deep neural network-based

BIQA, in which weighted average patch aggregation is used to get the global image quality. BIECON [19], following the FRIQA behavior, uses the local quality maps as intermediate targets for convolutional neural networks, and then pooled features are regressed into quality score. RANK [17] trains a Siamese Network to rank images which are generated by adding synthetic distortions to reference images. MEON [18] is composed by two sub-networks: a distortion identification sub-network and a quality prediction sub-network. Although these methods employ end-to-end optimized framework, they only use the output of the last layer to evaluate image quality. However, different levels of distortion generate different degradations on hierarchical features. These methods described above cannot efficiently represent the hierarchical degradation.

In order to capture the hierarchical degradation, some researchers extract multilevel features from the existing pre-trained CNN models (on other tasks, e.g., object classification), then these features are regressed with SVR to predict image quality. The structure of such type is represented by Fig.2(b). For example, BLINDER [12] extracts features at each layer of VGG16 [29]. Then SVR is utilized to obtain a score at each layer, and the final quality score is computed by averaging the layer-wise scores. HFD-BIQA [13] combines low-level local structure features with high-level semantic features extracted from the ResNet [30]. Afterwards, the combined features are fed into SVR to acquire the final quality score. Although these methods combine different levels of features and measure the quality degradation from multiple scales, they extract features and predict quality score separately. Since they are not within an end-to-end optimized network structure, such kind of BIQAs cannot jointly optimize the whole procedure. Besides, the performance and generalization ability of these models are always constrained by the tasks used for pre-training.

Moreover, there exists a common problem for all of these CNN-based BIQAs: the lacking of big training data. Existing databases (their size are too small) cannot provide sufficient training images to optimize a network with high generalization ability. Therefore, data augmentation is ineluctably adopted, and the most widely used method is the patch-based method. Although this skill is valid on databases with synthetic distortions, there are many drawbacks in dividing image into small patches such as size of 32×32: 1) Assigning MOS of the original image to its sampled patches, while the quality of each patch is different due to content and authentic distortions spatial inhomogeneities. 2) Several researches utilized FRIQA to generate proxy quality label for each patch. However, such small patches may not contain enough semantic information to judge its quality. And one FRIQA method can't achieve the best performance in all distortion types. 3) The subjective perception quality of each patch is not exactly the same as the whole image.

In this work, we firstly build a large-scale quality-annotated dataset to solve the problem of lacking training data based on merging multiple FRIQAs. Indeed, some previous researches have leveraged FRIQA to generate a quality score for unlabeled image. In BIQME [31], one high-accuracy FRIQA method, i.e., colorfulness-based PCQI [32], is proposed to predict quality score for enhanced image. BLISS [33] uses unsupervised rank aggregation to combine different FRIQAs to generate a synthetic score. Besides, MMF [34] also proposes a regression approach to fuse multiple FRIQAs. Specifically, the new MMF score is set to be the nonlinear combination of scores from multiple FRIQAs. Different from previous strategies, we propose an intuitive and effective method to merge multiple FRIQAs which will be introduced in Section III.

Following, an end-to-end cascaded CNN model (called CaHDC, as shown in Fig.2(c)) is proposed, which considers hierarchical degradation and simultaneously optimizes the whole procedure jointly. It's worthy mentioning that while CaHDC has a small number of parameters, it still maintains high performance. It greatly alleviates overfitting and achieves the superior cross-database performance.

TABLE I
COMPARISON OF EXISTING IQA DATABASES AND OUR PROPOSED DATASET

Dataset	Ref.imgs	Distortions	Dist.imgs	Quality label
LIVE [35]	29	5	779	**Yes**
CSIQ [36]	30	6	886	**Yes**
TID2013 [37]	25	**24**	3,000	**Yes**
Waterloot [38]	4,744	4	94,880	No
Proposed	**10,000**	21	**1,050,000**	**Yes**

III. LARGE-SCALE DATASET WITH PSEUDO-MOS

Optimizing deep convolutional neural network with high generalization ability needs a huge amount of data, however, the most popular IQA databases, such as LIVE [35], CSIQ [36], TID2013 [37], are usually too small. The largest existing dataset, i.e., TID2013, only possesses 3000 distorted images which are derived from 25 pristine images. Limited data size can easily lead to overfitting of deep neural networks. Waterloo Exploration Database [38] contains 94 880 distorted images generated from 4 744 high quality natural images. However it contains only 4 distortion types and all the images are short of quality labels. Since it is laborious to collect MOS for images by subjective experiment, which usually needs highly controlled conditions, resulting in small data collections relative to other image analysis databases. Towards overcoming these problems, we establish a large-scale quality-annotated dataset with pseudo-MOS based on merging multiple FRIQAs.

In this work, 10,000 images with high qualities are firstly chosen from MSCOCO [39] as reference images. Next, each reference image is degraded by 21 types of distortion under 5 noise levels. As a result, 1,050,000 distorted images are collected. Then, the best FRIQA for each distortion type is selected to compute quality scores for images. Five classical FRIQA metrics are adopted. Ultimately, we normalize the quality scores of all distorted images by building a nonlinear mapping function for each distortion type to get the unified pseudo-MOS. Tab.I lists the comparison between our proposed dataset and 4 other databases. Our proposed dataset is far ahead of other databases in terms of image quantity. Consequently we can leverage sufficient labeled data to train a stable and robust deep network which can greatly alleviate the overfitting.

TABLE II
THE 21 DISTORTION TYPES AND CORRESPONDING BEST FRIQA

No	Distortion Type	Best FRIQA	*No*	Distortion Type	Best FRIQA
#1	Additive Gaussian noise	VSI	#12	Non eccentricity pattern noise	GMSD
#2	Additive noise in color components	PSNR	#13	Local block-wise distortions	GMSD
#3	Spatially correlated noise	VSI	#14	Mean shift (intensity shift)	GSM
#4	Masked noise	PSNR	#15	Contrast change	GMSD
#5	High frequency noise	VSI	#16	Change of color saturation	FSIMc
#6	Impulse noise	PSNR	#17	Multiplicative Gaussian noise	VSI
#7	Quantization noise	GMSD	#18	Lossy compression of noisy images	GMSD
#8	Gaussian blur	GSM	#19	Image color quantization with dither	PSNR
#9	Image denoising	GMSD	#20	Chromatic aberrations	FSIMc
#10	JPEG compression	VSI	#21	Sparse sampling and reconstruction	GMSD
#11	JPEG2000 compression	VSI			

TABLE III
SROCC OF FRIQAS FOR EACH DISTORTION TYPE

FRIQA	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
VSI [23]	**0.946**	0.872	**0.937**	0.779	**0.921**	0.874	0.875	0.961	0.948	**0.954**	**0.971**
FSIMc [21]	0.912	0.854	0.89	0.809	0.904	0.825	0.881	0.955	0.933	0.936	0.959
PSNR	0.929	**0.898**	0.92	**0.832**	0.914	**0.897**	0.881	0.914	0.948	0.919	0.884
GSM [22]	0.914	0.824	0.925	0.737	0.893	0.803	0.89	**0.969**	0.945	0.932	0.964
GMSD [24]	**0.946**	0.868	0.935	0.717	0.916	0.764	**0.905**	0.911	**0.952**	0.951	0.966
FRIQA	#12	#13	#14	#15	#16	#17	#18	#19	#20	#21	
VSI [23]	0.806	0.171	0.77	0.475	0.81	**0.912**	0.956	0.884	0.891	0.963	
FSIMc [21]	0.794	0.552	0.749	0.473	**0.836**	0.857	0.949	0.882	**0.893**	0.958	
PSNR	0.686	0.099	0.767	0.44	0.101	0.891	0.914	**0.927**	0.887	0.904	
GSM [22]	0.807	0.631	**0.779**	0.479	0.355	0.849	0.958	0.904	0.881	0.967	
GMSD [24]	**0.814**	**0.663**	0.735	**0.621**	0.295	0.889	**0.963**	0.91	0.853	**0.968**	

A. Pristine Images

The reference images of our proposed dataset come from a large-scale database MSCOCO [39], which is widely used for object detection, segmentation and caption. There are numerous images suffering from severe distortion or poor perceptual quality in MSCOCO. Therefore, a manual process to select high quality images is essential. Only pristine images with high quality and clear content are chosen as the reference images. Specifically, we first removed those low-quality images undergoing obvious distortion, such as motion blur, defocus blur, Gauss noise, impulse noise, compression artifacts, under or over exposure, low contrast, artificial borders, watermarks, and other distortions. Next, gray images and images with minor size or low resolution are also removed. Finally, only 10,000 high-quality pristine images are left as the reference images.

B. Distorted Images

Similar to TID2013 (the largest IQA database, which contains 3,000 distorted images with 24 distortion types under 5 levels), we generate 21 distortion types, which are listed as Tab. II. Afterwards, each reference image is degraded by the 21 distortion types under 5 noise levels. As a consequence, a total of 1,050,000 distorted images are collected.

C. Generation of Pseudo-MOS

For purpose of generating credible pseudo-MOS for each distorted image, five classical FRIQAs, i.e., PSNR, FSIMc [21], GSM [22], VSI [23], GMSD [24] are adopted. Tab.III lists the SROCC of these five FRIQAs on each distortion type of TID2013. Pseudo-MOS can be assigned through two methods. The simplest approach to generate pseudo-MOS is to adopt a single FRIQA for all distorted images. Nevertheless one FRIQA responds to distinct distortion types discrepantly. Although some FRIQAs can achieve good performance on most distortion types, their performance is unsatisfactory on some specific distortion types. For instance, on TID2013, VSI achieves the best SROCC (0.971) on #11 (JPEG2000 Compression distortion), but the SROCC dramatically decreases to 0.171 on #13 (local block-wise distortions). Accordingly we leverage the other comprehensive means which takes FRIQA's discrepancies on distinct distortion types into account. We pick out a best FRIQA for each distortion type according to the performance. Tab.II lists the best FRIQA for each distortion type.

Another problem emerges when combining multiple FRIQAs. Different FRIQAs produce disparate quality value scales for diverse distortion types, so we need to normalize them into a unified range. In this work, a nonlinear mapping function is adopted to map predicted quality values from different FRIQAs into a unified scale as that in TID2013. The nonlinear mapping function is formulated as [35], [40]

$$Q = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2 (Q_s - \beta_3))} \right) + \beta_4 Q_s + \beta_5 \quad (1)$$

where Q is the normalized score, and Q_s is the predicted score from FRIQA. $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ are parameters to be fitted. Intuitively, there are two ways to build mapping models to merge predicted quality values.

1) Merge by FRIQA From Tab.II we can see that one FRIQA may achieve the best performance on more than one distortion types. Thus one way to merge predicted quality

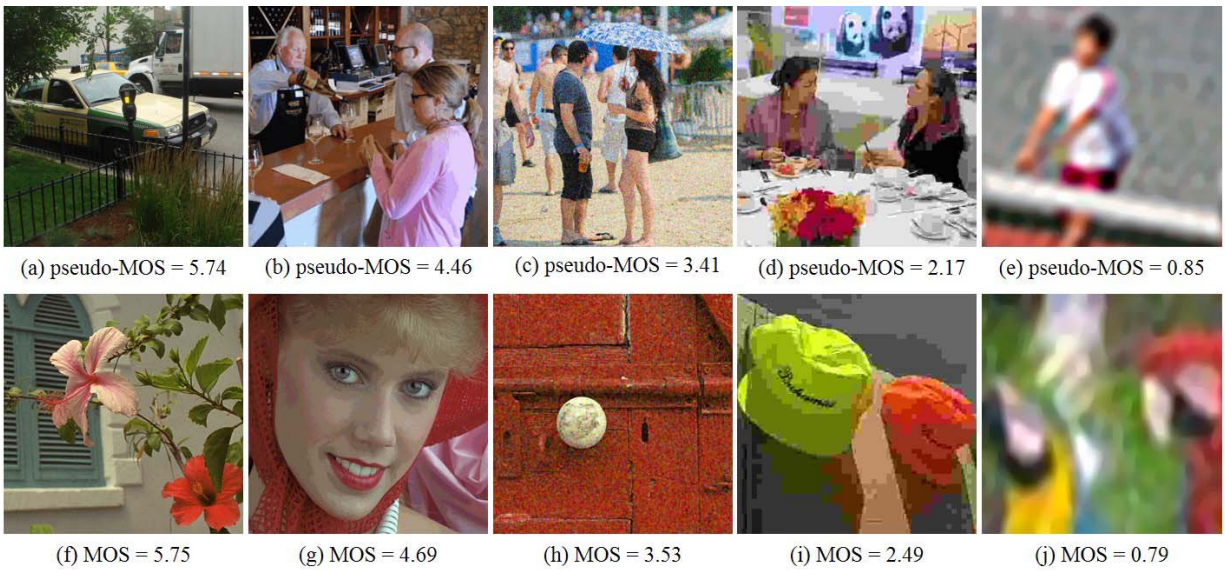


Fig. 3. Comparison between our proposed dataset and TID2013. (a)-(e) Different perceptual quality images with pseudo-MOS in our proposed dataset. (f)-(j) Different perceptual quality images with subjective MOS in TID2013.

TABLE IV
PERFORMANCES OF SINGLE FRIQA AND MULTI-FRIQAS ON TID2013

FRIQA	SROCC	PLCC
single-VSI [23]	0.897	0.899
single-FSIMc [21]	0.851	0.877
single-PSNR	0.703	0.702
single-GSM [22]	0.797	0.833
single-GMSD [24]	0.804	0.859
multi-FRIQAs (merge by FRIQA)	0.895	0.908
multi-FRIQAs (merge by distortion)	**0.947**	**0.954**

values is to build mapping model by FRIQA,

$$Q = f^n(Q_s^n) \quad \mathrm{n} \in [1, N] \tag{2}$$

where N is the number of FRIQAs we used in multi-FRIQAs, Q_s^n is the predicted scores of images belonging to the distortion types on which the n-th FRIQA achieves the best performance, and $f^n(\cdot)$ is the mapping model built for the n-th FRIQA. In this way, there are 5 mapping models built to merge predicted quality values for our proposed dataset.

2) Merge by distortion Another way to merge predicted quality value is to build mapping model by distortion type,

$$Q = f^m(Q_s^m) \quad \mathrm{m} \in [1, M] \tag{3}$$

where M is the number of distortion types in our dataset. Q_s^m is the predicted scores of images belonging to the m-th distortion type, and $f^m(\cdot)$ is the mapping model built for the m-th distortion type. There are 21 mapping models built by distortion for our proposed dataset.

We compare the performance of these methods mentioned above on TID2013. Tab.IV lists the SROCC and PLCC achieved by different single FRIQAs and multi-FRIQAs. As can be seen, because of considering properties of distinct distortion types, the SROCC and PLCC of multi-FRIQAs (merge by distortion) achieve the best performance (0.947 and 0.954), which is much higher than other methods. This indicates that the multi-FRIQAs (merge by distortion) approach is effective and reliable for quality prediction. And thus we apply it on our proposed large dataset to assign pseudo-MOS value for each distorted image.

Fig.3 presents some specimens of distorted images with pseudo-MOS in our dataset and several distorted images with subjective MOS in TID2013. The perceptual qualities of the presented images are in the range: {bad, poor, fair, good, excellent}. It's observed that when the degree of contamination is approximate, the pseudo-MOS in our proposed dataset and the subjective MOS in TID2013 is approximate too (which verifies the dependability of our proposed large quality-annotated dataset).

IV. THE QUALITY PREDICTION FRAMEWORK

An end-to-end BIQA framework (i.e., CaHDC) considering hierarchical quality degradation is proposed in this section. The proposed CaHDC can not only integrate hierarchical degradation to predict image quality (in which features are hierarchical analyzed and refined for concatenation), but also optimize the feature extraction, hierarchical degradation concatenation and quality prediction jointly in an end-to-end framework. Detailed information will be given in the following.

A. Architecture

We denote the input image with size $300 \times 300 \times 3$ by X and the pseudo-MOS/MOS of the input image by $\bar{Q}$. As depicted in Fig.4, our proposed model consists of three parts: the Hierarchical Net for feature extraction, the Side Pooling Nets (SiPNets) for hierarchical degradation fusion/concatenation, and the Regression Net for quality prediction. Their parameters are denoted as W^{α}, W^{β}, and W^{φ}, respectively. ReLU is selected as activation function and used after all convolutional layers and fully connected layers expect for special statement.

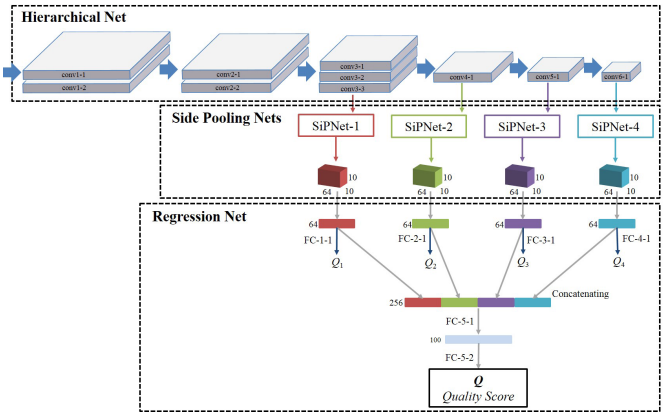


Fig. 4. Our proposed network architecture. Q_1 - Q_4 are intermediate scores at different single levels. Q is the final predicted quality score.

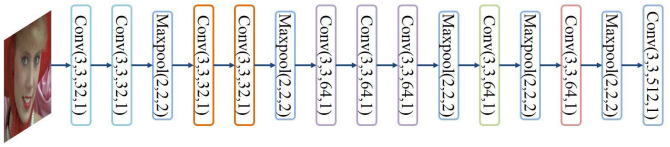


Fig. 5. The parameterization of Hierarchical Net. The presentation formats are Conv(heights, widths, output channels, strides), Maxpool(heights, widths, strides).

Hierarchical Net: The Hierarchical Net is composed of a series of convolutional layers including 6 levels, conv1-x, …, conv6-x, to extract hierarchical features (shallow-to-deep/low-to-high level). The parameterization details of the Hierarchical Net are shown in Fig.5, in which all convolutional layers apply 3×3 kernels, 1×1 stride and zero padding in order to obtain the output as the same size as the input. Meanwhile, 2×2 max pooling with stride 2 and zero padding are used on the output of each level for downsampling.

SiPNets: The branches of the SiPNets come from different levels of the Hierarchical Net except for the first and second levels. The reason why we omit the two levels is that the receptive field size is too small and too many parameters will increase the complexity of our network. As shown in Fig.6, the SiPNet first adopts a convolutional layer with 1×1 kernel and 1×1 stride. After that, a series of repeated convolutional layers with 3×3 kernel, zero padding and 2×2 stride are adopted for downsampling (until to the same size 10×10 as conv6-1). The last SiPNet separated from conv6-1 only has one convolutional layer with 1×1 kernel and 1×1 stride. The number of repeated convolutional layers for downsampling from SiPNet-1 to SiPNet-3 are $\{3, 2, 1\}$ respectively.

Regression Net: The Regression Net includes $K + 1$ fully connected layers, FC-1-x, …, FC-5-x, to map features extracted from SiPNets into quality scores, where $K = 4$ is the number of levels to be integrated. As shown in Fig.4, max pooling with the same size 10×10 as the feature map is applied to each output of SiPNet to extract the most apparent features, which are denoted as:

$$S_i = \Phi_i(X; W^{\alpha}, W_i^{\beta}) \quad i = 1, \ldots, K \tag{4}$$

where $S_i \in R^{64}$ represents the feature vector after max pooling on the output of SiPNet, W^{α} is the parameters of the Hierarchical Net, W_i^{β} is the parameters of the i-th SiPNet, and $\Phi_i(\cdot)$ denotes the process to get the feature vector S_i. Predicted scores by each single level can be described as:

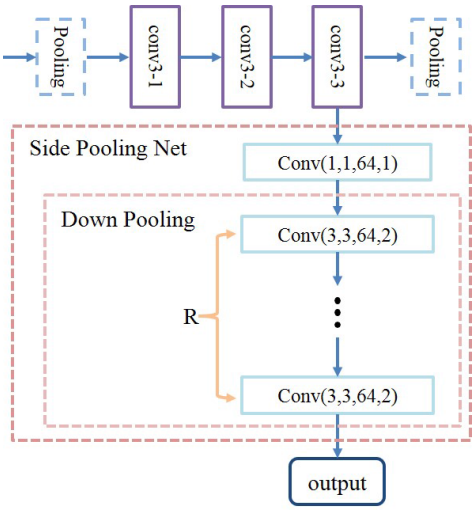


Fig. 6. Illustration of Side Pooling Net. R denotes the number of repeated convolutional layers.

$$Q_i = F_i(S_i; W_i^{\varphi}) \quad i = 1, \ldots, K \tag{5}$$

where $F_i(\cdot)$ denotes fully connected layers FC-i, and W_i^{φ} represents the parameters of FC-i. The final predicted score of the input image can be described as:

$$Q = F_{K+1}(S; W_{K+1}^{\varphi}) \tag{6}$$

$$S = S_1 \oplus S_2 \oplus S_3 \oplus S_4 \tag{7}$$

where $\oplus$ means the concatenation operation, S indicates the fused feature vector from distinct single levels. $F_{K+1}(\cdot)$ denotes the model of FC-5, and W_{K+1}^{φ} represents the parameters of FC-5.

B. Loss Function and Optimization

In view of the pseudo-MOS values of our proposed dataset are generated by FRIQAs, it's possible to produce some abnormal samples (inaccurate quality labels). Hence, Huber loss is adopted to improve robustness of the network. To ensure derivatives are continuous for all degrees, the Pseudo-Huber loss (as one of the smooth approximations of the Huber loss) is adopted,

$$L_{\delta}(q) = \delta^2 \left(\sqrt{1 + \left(\frac{q - \bar{q}}{\delta} \right)^2} - 1 \right) \tag{8}$$

where δ is a super parameter, q is the predicted quality score, and $\bar{q}$ is the ground truth (i.e., MOS or pseudo-MOS). This loss function approximates quadratic for small residuals between q and $\bar{q}$, and linear for large residuals to reduce the penalty to outliers. The gradient of $L_{\delta}(q)$ with respect to q is formulated as

$$\frac{\partial L_{\delta}(q)}{\partial q} = \frac{q - \bar{q}}{\sqrt{1 + \left(\frac{q - \bar{q}}{\delta} \right)^2}}. \tag{9}$$

We train our model by end-to-end and hierarchical-joint optimization. The error between predicted score by each

single level and MOS/pseudo-MOS is adopted as auxiliary loss to train our proposed model for hierarchical degradation measurement. Therefore, the overall loss function can be expressed as

$$L_t\left(X; W^{\alpha}, W^{\beta}, W^{\varphi}\right) = L_s + L_f = \sum_{i=1}^{K} a_i L_{\delta}\left(Q_i\right) + L_{\delta}\left(Q\right) \quad (10)$$

where L_s stands for the sum of losses under each single level, L_f stand for the loss under multilevel features integration, a_i is a weight to impose relative emphasis on one over the other levels, and it can be defined as,

$$a_i = \left(S(Q_i) - \frac{1}{2}\right)^2 \quad (11)$$

where $S(\cdot)$ is sigmoid function,

$$S(Q_i) = \frac{1}{1 + exp\left(-(Q_i - \bar{Q})\right)}. \quad (12)$$

The smaller the difference between Q_i and $\bar{Q}$, the smaller the weight a_i should be. The gradient of a_i with respect to Q_i is

$$\frac{\partial a_i}{\partial Q_i} = S\left(Q_i\right)\left(2S\left(Q_i\right) - 1\right)\left(1 - S\left(Q_i\right)\right). \quad (13)$$

During the back propagating, the gradient of the overall loss L_t with respect to Q_i and Q is

$$\frac{\partial L_t}{\partial Q_i} = \frac{\partial a_i}{\partial Q_i} L_{\delta}\left(Q_i\right) + \frac{\partial L_{\delta}\left(Q_i\right)}{\partial Q_i} a_i \quad (14)$$

$$\frac{\partial L_t}{\partial Q} = \frac{\partial L_{\delta}\left(Q\right)}{\partial Q}. \quad (15)$$

Finally, we can obtain the optimal parameters by minimizing the overall loss function,

$$\left(W^{\alpha}, W^{\beta}, W^{\varphi}\right)^* = \operatorname{argmin}\left(L_t\left(X; W^{\alpha}, W^{\beta}, W^{\varphi}\right)\right) \quad (16)$$

The proposed CaHDC has the following characteristics:

1) Multiple branches separated from different layers of the trunk net allow our network to evaluate the hierarchical quality degradation.

2) The first fully connected layer with only 100 dimensions tremendously reduces the number of network parameters, which can vastly speed up the optimization of the network, and simultaneously alleviates overfitting.

3) We do not pull the features at different layers into column vectors and then integrate them, but downsample them with convolutional operations to the same scale for integrating. As a result, the number of features is reduced and the spatial information of the features is reserved.

4) The proposed CaHDC is a lightweight network with high performance and generalization ability. The network parameters of WaDIQaM [16] and BIECON [19] are 5M and 7M, whereas the total number of CaHDC is only 0.73M.

C. Strategy for Training and Testing

According to RANK [17], the size of input sub-images ought to no less than 1/3 of the original images in order to capture context information. When training in our experiments, the pseudo-MOS/MOS of the original image is assigned as ground truth to its random sampled patches with size $300 \times 300 \times 3$ (such an input size is large enough to capture the context information). For testing, the image is cropped into 4 patches evenly with size of $300 \times 300 \times 3$, and the final quality score is acquired by averaging these predicted scores.

In the training process, a mini-batch of random sampled image patches are fed into CaHDC and Adam optimization algorithm is adopted for training. Parameters of Adam are set as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. The learning rate is set as

$$\alpha = \begin{cases} \alpha_0 * d^{s/s_0} & \alpha > \alpha_m \\ \alpha_m & \alpha \leq \alpha_m \end{cases} \quad (17)$$

where α_0 is the initial learning rate, d is the decay factor, s_0 is the decay rate, s is the number of trained steps, and α_m is the minimum learning rate. When $\alpha < \alpha_m$, it is equivalent to exponential decay function. This function applies a big learning rate to train the model at the begin and lower the learning rate as the training progresses until to the minimum learning rate. For pre-training, $\alpha_0 = 10^{-4}$, $\alpha_m = 10^{-5}$, s_0 is the number of steps required for training one epoch. For fine tuning, the global learning rates is fixed to 10^{-5}, and the learning rate of the Hierarchical Net is multiplied by 0.01.

V. Experimental Results

A. Databases and Protocols

Five public IQA databases are chosen for experiment, i.e., LIVE [35], CSIQ [36], TID2013 [37], LIVE-MD [41], and LIVE-CH [42]. The information of LIVE, CSIQ, and TID2013 can be found in Tab.I. The LIVE-MD contains 450 multiply distorted images under two multiple distortion scenarios: 1) first blurred and then compressed by JPEG encoder to simulated the scenario of image storage; 2) first blurred and then corrupted by white Gaussian noise to simulated camera image acquisition process. The LIVE-CH contains 1162 images captured under highly diverse conditions by a large amount of camera devices. It involves a variety of authentic and real-world distortion.

To measure the performance of the proposed CaHDC and other BIQAs, two widely used correlation criterion are adopted.

1) Pearson Linear Correlation (PLCC)

$$PLCC = \frac{\sum_i (p_i - p_m)(\hat{p}_i - \hat{p}_m)}{\sqrt{\sum_i (p_i - p_m)^2}\sqrt{\sum_i (\hat{p}_i - \hat{p}_m)^2}} \quad (18)$$

where p_i and $\hat{p}_i$ are the MOS and the framework prediction, respectively. PLCC measures the correlation between predicted scores and MOS values.

2) Spearman Rank Order Correlation Coefficient (SROCC)

$$SROCC = 1 - \frac{6\sum_{i=1}^{L} (m_i - n_i)^2}{L(L^2 - 1)} \quad (19)$$

where L is the number of distorted images, m_i is the rank of p_i in the MOS values, n_i is the rank of $\hat{p}_i$ in the predicted scores. SROCC assesses the monotony between MOS values and predicted scores. In this paper, media SROCC and PLCC are reported across 100 sessions.

TABLE V

PERFORMANCE COMPARISON OF DIFFERENT CNN MODELS (TRAINING ON TID2013 AND TESTING ON LIVE-MD)

	SROCC	PLCC
AlexNet [43]	0.644	0.667
VGGNet [29]	0.599	0.623
DenseNet [44]	0.714	0.741
CaHDC	**0.773**	**0.826**

TABLE VI

PERFORMANCE COMPARISON OF ABLATION EXPERIMENTS (TRAINING ON TID2013 AND TESTING ON LIVE-MD)

	SROCC	PLCC
CaHDC w/o SiPNet-1&2&3	0.509	0.558
CaHDC w/o SiPNet-1&2	0.698	0.762
CaHDC w/o SiPNet-1	0.745	0.801
No Single Level Loss L_s	0.753	0.817
CaHDC	**0.773**	**0.826**

B. Comparison With Traditional CNN Models

We firstly compare the proposed CaHDC against traditional CNN models used for image recognition task to demonstrate the efficiency of the proposed model. Image recognition task mainly focuses on high-level global abstracts, and features from the last layer are sufficient at most cases for classification. However, degradation from low-level to high-level will degrade the perceptual quality, and degradations on multilevels should be considered for quality predictation. We compare the performance of CaHDC with traditional CNN models (i.e., AlexNet [43], VGGNet-16 [29], DenseNet[2] [44]) by cross-database evaluation. All the models are first pre-trained on our proposed dataset for better initialization. After that, they are trained on TID2013 and tested on LIVE-MD. Median SROCC and PLCC results of different models across 100 sessions are reported in Tab.V. Thanks to hierarchical degradation integration, the proposed IQA-orientated CaHDC achieves a remarkable improvement.

C. Analysis of Hierarchical Degradation

In this subsection, we will comprehensively analyze the effect of hierarchical degradation. CaHDC is an IQA-orientated CNN method designed by integrating hierarchical degradation, which is inspired by the hierarchical process mechanism of HVS. We firstly conduct the ablation experiment to compare the performance under the conditions of hierarchical integration or not. Tab.VI lists the performance when removing the SiPNets step-wisely. **CaHDC w/o SiPNet-i&j** means that the SiPNet-i and SiPNet-j in Fig.4 are removed. All the models in Tab.VI are firstly pre-trained on our proposed dataset, then trained on TID2013 and tested on LIVE-MD. We can see that if there is no hierarchical integration, the performance of **CaHDC w/o SipNet-1&2&3** is really poor. Considering more hierarchical degradation step-wisely can improve the performance of quality prediction gradually. Our proposed complete model CaHDC considers all the hierarchical degradation and obtains the best results, about 51.9% improvement on SROCC and 48% improvement on PLCC compared to **CaHDC w/o SipNet-1&2&3**. Besides, our proposed loss function L_t is also conducive to better learning hierarchical degradation for BIQA. By embedding the Q_i into the overall loss function L_t, CaHDC can better measure the quality degradation at different levels, which will be beneficial to the final quality prediction. From Tab.VI, we can see that under **No Single Level Loss** L_s, i.e., $L_t = L_f$, the SROCC and PLCC are 0.753 and 0.817, which are lower than our proposed CaHDC.

[2]In our experiments, the DenseNet is set to be the same depth as CaHDC for fair comparison.

Fig. 7. An example from TID2013 to analysis the impact of distortion on hierarchical features. (a) Reference image. (b)-(d) Distorted images by JPEG compression under different levels, their subjective MOS values in TID2013 are 5.9, 5.0, 2.2 respectively. The higher the MOS value, the better the image quality.

Next, the impact of distortion on hierarchical features are deeply analyzed. A pristine image and its distorted images (with JPEG compression under different degradation level) are shown in Fig.7. Intuitively, with the increase of distortion level from Fig.7 (b) to (d), their quality is gradually degraded. As shown in Fig.7 (b), weakly distortion mainly degrades the low-level edge, and has slightly affect on the high level semantics (we can still clearly understand the window, the flower and the leaf of this image). However, severely distortion will directly destroy the high level semantics, as shown in Fig.7 (d).

In order to further clearly demonstrate the hierarchical degradation on multilevel features quantitatively, we define a metric which is called Error Ratio (ER) to analyze the impact of distortion on multilevel features. ER is formulated as

$$ER = \frac{RMSE_i}{\sum_{i=1}^{K} RMSE_i} \tag{20}$$

where $K = 4$ is the number of levels to be integrated, $RMSE_i$ is Root Mean Square Error of features (i.e., the output of SiPNet-i) between the reference image and distorted image.

TABLE VII
RMSE VALUES OF DIFFERENT FEATURE LEVELS

Fig.7	SiPNet-1	SiPNet-2	SiPNet-3	SiPNet-4
(b)	7.833	5.338	4.996	3.255
(c)	9.514	7.342	7.024	4.560
(d)	16.239	13.810	15.282	11.242

TABLE VIII
ER VALUES OF DIFFERENT FEATURES LEVELS

Fig.7	SiPNet-1	SiPNet-2	SiPNet-3	SiPNet-4
(b)	36.6%	24.9%	23.3%	15.2%
(c)	33.5%	25.8%	24.7%	16.0%
(d)	28.7%	24.4%	27.0%	19.9%

Before computing RMSE, Min-Max Normalization is applied to the output of SiPNet-i. Tab.VII and Tab.VIII respectively list the RMSE and ER of different levels of features between distorted image and reference image. It can be seen that with the aggravation of distortion, the degradation of features (i.e., RMSE) becomes larger and larger. When the distortion is weak, low level features are mainly degraded while high level semantics are slightly affected, e.g., the ER of SiPNet-1 between Fig.7(b) and Fig.7(a) is largest among all the four hierarchical levels. When distortion is severe, high level semantics are destroyed apparently, so the ER of SiPNet-4 between Fig.7(d) and Fig.7(a) is larger than the other distortion levels. As the distortion increases, ER in higher levels become larger.

In conclusion, distortion causes hierarchical quality degradation, and it is essential to integrate hierarchical features for image quality assessment. The proposed CaHDC is IQA oriented, which can efficiently extract hierarchical features and integrate hierarchical degradation for BIQA.

D. Performance Comparison With Existing BIQAs

We compare the proposed CaHDC with 15 classical BIQAs (8 traditional and 7 CNN-based) on the five benchmark databases. We randomly split each database into 80% images for training and 20% images for testing by reference images for no overlapping in context. We tried our best to collect the results of these existing BIQAs, which are listed at Tab.IX. For the 8 traditional BIQAs, all the results are calculated by the source code released by authors. For the 7 CNN-based BIQAs, all the results come from existing papers.

As can be seen, the proposed CaHDC performs better than all of these traditional BIQAs on LIVE, CSIQ, TID2013, LIVE-MD and LIVE-CH (especially for CSIQ and TID2013). When comparing with these CNN-based BIQAs, the proposed CaHDC also achieves the state-of-the-art performance on the five benchmark databases. On LIVE, most of the CNN-based BIQAs achieve high performance. Although the performance is slightly lower than some other algorithms, our proposed CaHDC achieves 0.965 SROCC and 0.964 PLCC, which is acceptable in most cases. On CSIQ, CaHDC gets the best SROCC, which is about 2.1% improvement compared to DIQA [45]. On TID2013, CaHDC achieves the best results among all the compared BIQAs except for BPSQM [46], which is about 0.8% higher than our proposed on PLCC. Different from LIVE, CSIQ and TID2013, LIVE-MD focus on multiply distorted images, and CaHDC still achieves the best PLCC result and the second SROCC result, which verifies that CaHDC can also measure the quality degradation caused by multiple distortions well. LIVE-CH is the most challenging because it is composed of authentic distorted images. Our proposed CaHDC still achieves the best performance among the compared BIQAs, nevertheless there is still a lot of room for improvement. In summary, by considering the hierarchical degradation for BIQA and the end-to-end manner for optimization, the proposed CaHDC performs highly consistently with the subjective perception.

E. Cross Database Evaluations

In order to verify that the proposed CaHDC is not limited by the database that it has been trained, the cross-database evaluations are given. For the five benchmark databases, LIVE, CSIQ and TID2013 are legacy databases which distorted by synthetic distortion; LIVE-MD is distorted by synthetic multiple distortions; LIVE-CH is distorted by a wide variety of authentic distortions. Therefore, it is a big challenge for cross database evaluation.

Firstly, we compare the generalization ability of CaHDC with other BIQAs when training on the TID2013 (the largest legacy database) and testing on the other four databases. For the 5 traditional BIQAs, all the results are calculated by the source code released by authors. For the two CNN-based BIQAs (i.e., WaDIQaM [16] and RANK [17]), the results are computed by the model released by the authors. As shown in Tab. X, the proposed CaHDC performs the best on three legacy databases (i.e., LIVE, CSIQ, and LIVD-MD). Especially for the LIVE-MD, the proposed CaHDC achieves a remarkable improvement against the other BIQAs, about 51.9% improvement on SROCC and 61% improvement on PLCC compared to BLIINDS-II [3]. It verifies that our proposed CaHDC can better learn the nature of quality degradation caused by distortion, and has good generalization ability from single distortion to multiple distortion. However, probably because the distortion in LIVE-CH is quite different from that in the other synthetic databases, all of these BIQAs performs poorly on the wild LIVE-CH (the best one with SROCC 0.419), and the proposed CaHDC performs the second place (a slightly worse than the best one). Meanwhile,the two CNN-based BIQAs (WaDIQaM [16] and RANK [17]) perform extremely poorly.

Tab.XI lists the SROCC and PLCC for models when trained on the wild LIVE-CH and tested on the other four legacy databases. LIVE-CH is composed of authentic distorted images and the other four databases are composed of synthetic distorted images. It's a big challenge to achieve good results when training on LIVE-CH and testing on the other four databases. Fortunately, benefiting from our proposed large-scale quality-annotated dataset and the hierarchical degradation concatenation in CaHDC, our method can still get perfect results. Specifically, benefiting from pre-training on the proposed dataset, CaHDC can effectively alleviate overfitting. Benefiting from hierarchical degradation concatenation and end-to-end optimization, CaHDC can better learn the nature

TABLE IX
PERFORMANCE COMPARISON ON FIVE BENCHMARK IQA DATABASES

	LIVE [35]		CSIQ [36]		TID2013 [37]		LIVE-MD [41]		LIVE-CH [42]	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
BLIINDS-II [3]	0.919	0.920	0.570	0.534	0.536	0.628	0.827	0.845	0.405	0.450
DIIVINE [2]	0.925	0.923	0.784	0.836	0.654	0.549	0.874	0.894	0.546	0.568
BRISQUE [4]	0.939	0.942	0.750	0.829	0.573	0.651	0.897	0.921	0.607	0.585
NIQE [28]	0.915	0.919	0.630	0.718	0.299	0.415	0.745	0.815	0.430	0.480
CORNIA [47]	0.942	0.943	0.714	0.781	0.549	0.613	0.900	0.915	0.618	0.662
HOSA [48]	0.948	0.949	0.781	0.842	0.688	0.764	0.902	0.926	0.660	0.680
ILNIQE [49]	0.902	0.865	0.807	0.808	0.519	0.640	0.878	0.892	0.430	0.510
FRIQUEE [50]	0.948	0.962	0.839	0.863	0.669	0.704	0.925	0.940	**0.720**	**0.720**
MEON [18]	-	-	-	-	0.808	-	-	-	-	-
WaDIQaM [16]	0.954	0.963	-	-	0.761	0.787	-	-	0.671	0.680
RANK [17]	**0.981**	**0.982**	-	-	0.780	0.799	0.921	0.936	-	-
VIDGIQA [51]	0.969	0.973	-	-	-	-	-	-	-	-
DIQA [45]	**0.975**	**0.977**	**0.884**	**0.915**	0.825	0.850	**0.939**	**0.942**	0.703	0.704
BIECON [19]	0.958	0.960	0.815	0.823	0.717	0.762	0.909	0.933	0.595	0.613
BPSQM [46]	0.973	0.963	0.874	**0.915**	**0.862**	**0.885**	-	-	-	-
CaHDC	0.965	0.964	**0.903**	0.914	**0.862**	**0.878**	**0.927**	**0.950**	**0.738**	**0.744**

TABLE X
CROSS DATABASE EVALUATIONS WHEN TRAINING ON TID2013

Train	TID2013 [37]							
Test	LIVE [35]		CSIQ [36]		LIVE-MD [41]		LIVE-CH [42]	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
BLIINDS-II [3]	0.836	0.840	0.568	0.661	0.509	0.513	0.142	0.207
DIIVINE [2]	0.687	0.688	0.590	0.661	0.479	0.499	**0.419**	**0.445**
BRISQUE [4]	0.681	0.687	0.491	0.494	0.314	0.356	0.110	0.166
HOSA [48]	0.842	0.836	0.622	0.698	0.469	0.509	0.279	0.319
FRIQUEE [50]	0.847	0.849	0.637	0.722	0.421	0.550	0.241	0.294
WaDIQaM [16]	0.817	0.807	0.690	0.750	0.318	0.334	0.107	0.162
RANK [17]	0.641	0.594	0.475	0.399	0.376	0.415	0.146	0.141
CaHDC	**0.930**	**0.921**	**0.736**	**0.808**	**0.773**	**0.826**	0.396	0.438

TABLE XI
CROSS DATABASE EVALUATIONS WHEN TRAINING ON LIVE-CH

Train	LIVE-CH [42]							
Test	LIVE [35]		CSIQ [36]		TID2013 [37]		LIVE-MD [41]	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
BLIINDS-II [3]	0.516	0.465	0.409	0.575	0.282	0.441	0.693	0.700
DIIVINE [2]	0.461	0.414	0.436	0.462	0.346	0.468	0.342	0.405
BRISQUE [4]	0.455	0.452	0.181	0.311	0.296	0.465	0.270	0.362
HOSA [48]	0.450	0.469	0.337	0.475	0.317	0.482	0.339	0.331
FRIQUEE [50]	0.553	0.530	0.550	0.598	0.368	0.508	0.494	0.657
CaHDC	**0.922**	**0.911**	**0.756**	**0.822**	**0.706**	**0.727**	**0.723**	**0.773**

of quality degradation. As shown in Tab.XI, the proposed CaHDC performs much better than the other existing BIQAs on all of these databases, which demonstrates that the proposed CaHDC has a very strong generalization ability and performs state-of-the-art.

F. Efficiency of the Proposed Dataset

The proposed dataset can also benefit the existing CNN-based IQA models. Two existing CNN-based IQA models (i.e., WaDIQaM and VIDGIQA) are adopted to verify the improvements of pre-training on the proposed dataset. We implement these two models as setups in their original papers. Both WaDIQaM and VIDGIQA are first pre-trained on our proposed dataset, then trained on LIVE and tested on TID2013. Tab.XII lists the cross-database evaluation results. The SROCC and PLCC results of WaDIQaM [16] and VIDGIQA [51] come from their original papers. As can be seen, the generalization performances of both models are greatly improved. WaDIQaM achieves up to about 34.6% improvement on SROCC. And VIDGIQA achieves up to 35.4% improvement both on SROCC and PLCC.

G. Analysis on the Number of Parameters

Finally, the number of parameters for different CNN-based BIQAs are analyzed. In real applications, limited by the memory space of microprocessing system, it is critical to

TABLE XII

IMPROVEMENTS OF OUR PROPOSED DATASET FOR EXISTING BIQA MODELS (TRAINING ON LIVE AND TESTING ON TID2013)

	SROCC	PLCC
WaDIQaM [16]	0.462	-
VIDGIQA [51]	0.415	0.477
WaDIQaM+Our dataset	0.622	0.667
VIDGIQA+Our dataset	0.562	0.646

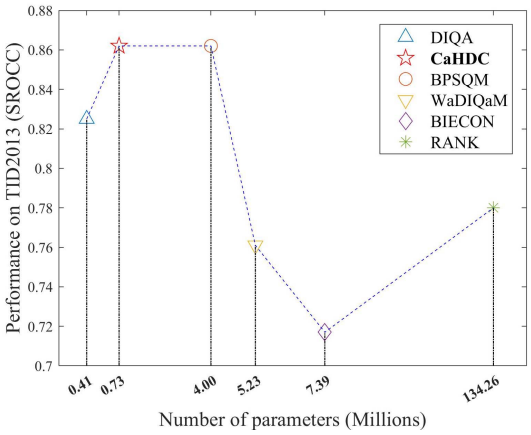


Fig. 8. Number of parameters vs. Performance. The cross-validation results (SROCC) on TID2013 are adopted to illustrate the performance of CNN-based BIQAs. For better visual effect, the scale of the X-axis is not linearly distributed.

design a lightweight model with small number of parameters while ensuring higher performance. Fig.8 illustrates the parameters-performance curve for representative CNN-based BIQAs. Among the compared methods, CaHDC achieves the state-of-the-art SROCC with only 0.73M parameters, which is at least one order of magnitude smaller than other models expect for DIQA [45]. Although DIQA has fewer parameters, its performance is poorer than CaHDC. In summary, CaHDC is lightweight while maintaining high performance. It can be easily applied to microprocessing systems. We have implemented the proposed CaHDC on NVIDIA JETSON TX2 (GPU: Nvidia Pascal^{TM}; CPU: HPM Dual Denver 2/2MB L2 + Quad ARM A57/2MB L2). It costs 110 seconds to run the entire TID2013 database (about 27 images per second, satisfying the real-time requirement).

VI. CONCLUSION

In this work, inspired by the hierarchical perception mechanism of HVS, we have introduced a novel IQA-orientated CNN-based method for BIQA. To meet the demand of big data for CNN optimization, a large quality-annotated IQA dataset has been primarily established, which contains 10,000 reference images and 1,050,000 distorted images. Benefiting from the pre-training on the proposed dataset, our model can effectively alleviate overfitting. Afterwards, a cascaded hierarchical net with hierarchical degradation concatenation is proposed, which can effectively measure the effect of hierarchical degradation on the overall image quality. Eventually, by jointly optimizing the feature extraction, hierarchical degradation concatenation, and quality prediction in an end-to-end manner, CaHDC can better learn the nature of quality degradation. Experimental results on five benchmark IQA databases have demonstrated the efficiency of the proposed CaHDC. Moreover, experiments of cross-database verification have further proved the high generalization ability of the proposed CaHDC.

REFERENCES

[1] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, May 2011.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2394–2402.

[7] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4559–4565, Dec. 2017.

[8] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, "No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1030–1040, May 2017.

[9] Q. Wu, H. Li, K. N. Ngan, and K. Ma, "Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2078–2089, Sep. 2018.

[10] G. Zhai, "On blind quality assessment of JPEG images," in *Proc. 7th Int. Conf. Cloud Comput. Big Data (CCBD)*, Nov. 2016, pp. 1–7.

[11] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.

[12] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognit.*, vol. 81, pp. 432–442, Sep. 2018.

[13] J. Wu, J. Zeng, Y. Liu, G. Shi, and W. Lin, "Hierarchical feature degradation based blind image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 510–517.

[14] C. Sun, H. Li, and W. Li, "No-reference image quality assessment based on global and local content perception," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.

[15] 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.

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

[17] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1040–1049.

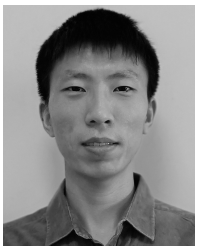
[18] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-End blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[19] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[20] 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.

[21] 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.
[22] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.
[23] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.
[24] 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.
[25] S. Hochstein and M. Ahissar, "View from the top: Hierarchies and reverse hierarchies in the visual system," *Neuron*, vol. 36, no. 5, pp. 791–804, 2002.
[26] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, no. 11, pp. 1019–1025, Nov. 1999.
[27] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 818–833.
[28] 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.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: https://arxiv.org/abs/1512.03385
[31] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1301–1313, Apr. 2018.
[32] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2387–2390, Dec. 2015.
[33] 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.
[34] T.-J. Liu, W. Lin, and C.-C.-J. Kuo, "Image quality assessment using multi-method fusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1793–1807, May 2013.
[35] 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.
[36] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 19–21, 2010.
[37] N. Ponomarenko *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *Proc. Eur. Workshop Vis. Inf. Process. (EUVIP)*, Jun. 2013, pp. 106–111.
[38] 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.
[39] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision (ECCV)*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 740–755.
[40] VQEG. (Mar. 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
[41] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Conf. Rec. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 1693–1697.
[42] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.
[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[45] J. Kim, A. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.
[46] D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, and Y. Zhang, "Blind predicting similar quality map for image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6373–6382.
[47] 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.
[48] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.
[49] 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.
[50] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, Jan. 2017.
[51] J. Guan, S. Yi, X. Zeng, W.-K. Cham, and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2505–2520, Nov. 2017.

Jinjian Wu (Member, IEEE) received the B.Sc. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. From 2011 to 2013, he was a Research Assistant with Nanyang Technological University, Singapore, where he was a Postdoctoral Research Fellow from 2013 to 2014. From 2015 to 2019, he was an Associate Professor with Xidian University, where he has been a Professor since 2019. His research interests include visual perceptual modeling, biomimetic imaging, quality evaluation, and object detection. He received the Best Student Paper Award at the ISCAS 2013. He has served as the Special Section Chair of the IEEE Visual Communications and Image Processing (VCIP) 2017, the Section Chair/Organizer/TPC Member of the ICME2014-2015, the PCM2015-2016, the ICIP2015, the VCIP2018, and the AAAI2019 Quality Assessment, and an Associate Editor for the *Journal of Circuits, Systems and Signal Processing* (CSSP).



Jupo Ma received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Xidian University, Xi'an, China. His research interests include image quality assessment, deep learning, and dynamic vision sensor (DVS).

Fuhu Liang received the B.S. degree from Xidian University, Xi'an, China, in 2017. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Xidian University. His research interests include machine learning, weak target detection, and image quality assessment (IQA).

Weisheng Dong (Member, IEEE) received the B.S. degree in electronic engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2004, and the Ph.D. degree in circuits and system from Xidian University, Xi'an, China, in 2010. He was a Visiting Student with Microsoft Research Asia, Beijing, China, in 2006. From 2009 to 2010, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. In 2010, he joined Xidian University, as a Lecturer, where has been a Professor since 2016. He is currently with the School of Artificial Intelligence, Xidian University. His research interests include inverse problems in image processing, sparse signal representation, and image compression. He was a recipient of the Best Paper Award at the SPIE Visual Communication and Image Processing (VCIP) in 2010. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and the *SIAM Journal on Imaging Sciences*.

Guangming Shi (Senior Member, IEEE) received the B.S. degree in automatic control, the M.S. degree in computer control, and the Ph.D. degree in electronic information technology from Xidian University, Xi'an, China, in 1985, 1988, and 2002, respectively. He had studied at the University of Illinois at Urbana–Champaign and The University of Hong Kong. Since 2003, he has been a Professor with the School of Electronic Engineering, Xidian University, where he is currently the Academic Leader on circuits and systems. He has authored or coauthored over 200 papers in journals and conferences. His research interests include compressed sensing, brain cognition theory, multirate filter banks, image denoising, low-bitrate image and video coding, and implementation of algorithms for intelligent signal processing. He was awarded as the Cheung Kong Scholar Chair Professor by the Ministry of Education in 2012. He served as the Chair of the 90th MPEG and 50th JPEG of the International Standards Organization (ISO) and the Technical Program Chair of the FSKD06, the VSPC 2009, the IEEE PCM 2009, the SPIE VCIP 2010, and the IEEE ISCAS 2013.

Weisi Lin (Fellow, IEEE) received the Ph.D. degree from King's College London, U.K. He served as the Laboratory Head of visual processing with Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the School of Computer Engineering. His research interests include image processing, perceptual signal modeling, video compression, and multimedia communication, in which he has published 170 journal articles, more than 230 conference papers, filed seven patents, and authored two books. He is a fellow of IET and an Honorary Fellow of the Singapore Institute of Engineering Technologists. He has been the Technical Program Chair of the IEEE ICME 2013, the PCM 2012, the QoMEX 2014, and the IEEE VCIP 2017. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He has been an invited/panelist/keynote/tutorial speaker for more than 20 international conferences. He was a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2016 to 2017 and the Asia-Pacific Signal and Information Processing Association (APSIPA) from 2012 to 2013.