Objective Quality Evaluation of Dehazed Images

Xiongkuo Min, Guangtao Zhai, *Member, IEEE*, Ke Gu, *Member, IEEE*, Xiaokang Yang, *Senior Member, IEEE*, and Xinping Guan, *Fellow, IEEE*

***Abstract*— Vision-based intelligent systems like automatic driving or driving assistance can be improved by enhancing the visibility of the scenes captured in bad weather conditions. In particular, many image dehazing algorithms (DHAs) have been proposed to facilitate such applications in hazy weather. Contrary to the substantial progress of DHA developing, the quality evaluation of DHAs falls behind. Generally, DHAs can be evaluated qualitatively by human subjects or quantitatively by objective quality measures. Compared with the subjective evaluation which is time consuming and difficult to apply, objective measures with quantitative results are more needed in practical systems. But in the literature, very few measures are widely utilized, and even less measures correlate well with the overall dehazing quality (DHQ). In this paper, we study the DHQ evaluation using real hazy images systematically. We first construct a DHQ database, which is the largest of its kind so far and includes 1750 dehazed images generated from 250 real hazy images of various haze densities using seven representative DHAs. A subjective quality evaluation study is subsequently conducted on the DHQ database. Then, we propose an objective DHQ index (DHQI) by extracting and fusing three groups of features, including: 1) haze-removing features; 2) structure-preserving features; and 3) over-enhancement features, which have captured the most key aspects of dehazing. DHQI can be utilized to evaluate DHAs or optimize practical dehazing systems. Validations on the constructed DHQ database and three other databases with synthetic haze have verified the effectiveness of DHQI. Finally, we give an overview of the current DHA quality evaluation strategies, discuss their merits and demerits, and give some suggestions on systematic DHA quality evaluation. The DHQ database and the code of DHQI will be released to facilitate further research.**

***Index Terms*— Single image dehazing, dehazed image, quality assessment, dehazing algorithm evaluation.**

Manuscript received February 10, 2018; revised June 22, 2018; accepted August 30, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61831015, 61521062, and 61527804, and in part by the China Postdoctoral Science Foundation under Grant BX20180197. The Associate Editor for this paper was Z. Duric. *(Corresponding author: Guangtao Zhai.)*

X. Min, G. Zhai, and X. Yang are with the Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and also with the Institute of Image Communication and Network Engineering, Shanghai Key Laboratory of Digital Media Processing and Transmissions, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: minxiongkuo@gmail.com; zhaiguangtao@sjtu.edu.cn; xkyang@sjtu.edu.cn).

K. Gu is with the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: guke@bjut.edu.cn).

X. Guan is with the Key Laboratory of System Control and Information Processing, Ministry of Education of China, Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xpguan@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2868771

I. Introduction

VISION based automatic driving and other driving assistance systems may suffer from low visibility problems in bad weather conditions such as haze, rain, and dust. Compared with the human visual systems, the relevant techniques required by the driving assistance systems such as lane, vehicle and pedestrian detection are more likely to suffer from the visibility problems. Since perfect weather conditions may not be asked for, many specific techniques have been proposed for visibility enhancement under various extreme weather conditions [1]–[5]. Among them, haze removal has been widely researched and many single image dehazing algorithms (DHAs) have been proposed due to the more and more frequent hazy weather [6]–[16]. DHAs can be utilized to enhance the visibility and restore the image details in practical image capturing systems, especially the systems used outdoors.

A. Single Image Dehazing Algorithms

Images captured in hazy weather can be described by the following model [17], which is the basis of many image dehazing studies

$$\mathbf{I}(i,j) = \mathbf{J}(i,j)e^{-\beta \mathbf{d}(i,j)} + \mathbf{A}(1 - e^{-\beta \mathbf{d}(i,j)}), \tag{1}$$

where $\mathbf{I}$ is the captured image suffering from haze, $\mathbf{J}$ is the real scene image, $\mathbf{A}$ is the global atmospheric light, $t(i,j) = e^{-\beta \mathbf{d}(i,j)}$ is the medium transmission, β is the scattering coefficient of the atmosphere, $\mathbf{d}$ indicates the scene depth, and i, j are pixel indexes. The first term $\mathbf{J}(i,j)e^{-\beta \mathbf{d}(i,j)}$ is *attenuation*, which describes the scene radiance and its decay in the atmosphere; whereas the second term $\mathbf{A}(1-e^{-\beta \mathbf{d}(i,j)})$ is *airlight*, which describes the environmental illumination. The objective of dehazing is to estimate $\mathbf{J}$, t, and $\mathbf{A}$ from $\mathbf{I}$.

A latest review of single image DHAs is given in [15]. We shortly review several representative DHAs here. Fattal [6] solved image dehazing as a non-linear inverse problem. A dark channel prior was introduced in [8] for dehazing, which is based on a phenomenon that at least one color channel has very low intensity values at some pixels. Tarel and Hautière [7] solved dehazing as a particular filtering problem. Another method based on guided joint bilateral filter was introduced in [9]. Meng *et al.* [10] proposed a regularization method to remove haze. Lai *et al.* [12] derived the optimal transmission map under some scene priors. Berman *et al.* [13] introduced a non-local scene prior which describes that color clusters in RGB space are distributed along lines. Some other methods

TABLE I
TWO QUANTITATIVE DHA EVALUATION STRATEGIES

Strategy 1 Using synthetic hazy image
Input: Haze-free image $\mathbf{J}$, corresponding depth d, target DHA
Output: Evaluation of the DHA q
1: Synthesize hazy image $\mathbf{I}_h$ via Eq.(1)
2: Generate the dehazed image $\mathbf{I}_d$ from $\mathbf{I}_h$ using the target DHA
3: Compute the quality using FR quality measure $q = \mathrm{FR}(\mathbf{J}, \mathbf{I}_d)$

Strategy 2 Using real hazy image
Input: Real hazy image $\mathbf{I}_h$, target DHA
Output: Evaluation of the DHA q
1: Generate the dehazed image $\mathbf{I}_d$ from $\mathbf{I}_h$ using the target DHA
2: Compute the quality using NR quality measure $q = \mathrm{NR}(\mathbf{I}_d)$

learned a mapping from synthetic hazy images to the haze-free images, for example a random forest based method was used in [11], and an end-to-end system using neural network was introduced in [14]. Readers can refer to [15] for more DHAs.

B. Quality Evaluation of DHAs

Besides DHAs themselves, the evaluation of DHAs is also very important. When proposing a DHA, one needs to evaluate it and compare it with the state-of-the-art, or when applying DHAs, one needs to select the best algorithm. Moreover, an effective and comprehensive evaluation criterion can promote image dehazing research forward in a right direction. In the current literature, most DHAs are evaluated from two aspects: qualitative evaluation given by human subjects, and quantitative evaluation given by objective measures. Qualitative evaluation is straightforward and accurate, since humans are often the ultimate receiver of the dehazed images. It is the most recognized evaluation strategy, and most DHAs are evaluated qualitatively when proposed [6]–[16], [18]. But it suffers from several critical disadvantages. First, it is time-consuming and expensive, which makes large scale evaluation difficult. Then, the qualitative evaluation becomes "controllable", since there is no widely utilized large scale evaluation set, and the selected limited number of hazy images only occupy a tiny subset of real practices. Moreover, subjective evaluation is difficult to be applied and embedded into practical systems, thus timely optimization for the system also becomes tough.

Considering the drawbacks of qualitative evaluation, quantitative evaluation via objective quality measures is introduced. Generally, two strategies can be adopted for quantitative evaluation: using real hazy images and using synthetic hazy images. Table I has summarized and compared these two strategies. Using real hazy images is the more straightforward way. It utilizes some measures to assess the quality of dehazed images generated from real hazy images directly [7]. It is a no-reference (NR)[1] image quality assessment (IQA) problem since the ground-truth haze-free image is not available. It is difficult due to the complexity of dehazing. During recent years, some quality measures are proposed for this objective [19]–[23]. In [23], three descriptors are proposed by comparing the gradient of the visible edges of the images before and after dehazing. But these descriptors only measure the dehazing effect, rather than the overall dehazing quality. There are also some measures proposed for quality assessment of enhanced images [19]–[22], since dehazing is an image enhancement process. But these measures are not specifically designed for dehazing and are not effective enough for DHA quality evaluation.

[1]Though some measures use the hazy image as a "reference", this "reference" is different from the traditional perfect quality reference, thus we still describe these methods as NR in this paper.

(a) (b)

Fig. 1. A comparison of synthetic haze and real haze. (a) Synthetic haze. (b) Real haze.

Another quantitative evaluation strategy is using synthetic hazy images [11], [12], [14]–[16], [24]. An overall discussion of this strategy is given in [24], and an effective quality measure for this strategy is also proposed. These methods synthesize hazy images from haze-free images and the corresponding depth using the widely utilized haze model. The haze-free images are taken as the ground-truth of dehazing, and full-reference (FR) IQA measures can be utilized as the evaluation criteria. It is easy to conduct quantitative evaluation using such strategy. The main drawback of this strategy is that real haze may be different from the synthetic haze. As illustrated in Fig. 1, synthetic haze and real haze look quite different. The practical haze may not be perfectly modeled by the ideal haze model. The synthetic haze is usually assumed to be homogenous, whereas the real haze is often far more complicated than that. Moreover, synthetic hazy and many DHAs are both based on the ideal haze model, which may reduce the difficulty of dehazing. A good synthetic haze-removing effect does not necessarily guarantee a good real haze-removing effect. Thus this strategy can be utilized to evaluate DHAs from one aspect, but quantitative evaluation on real hazy images is still needed.

C. Contributions of This Paper

In this paper, we study dehazing quality evaluation using real hazy images systematically. To facilitate the research, we first construct a large dehazing quality (DHQ) database which includes 1,750 dehazed images created from 250 hazy images using 7 representative DHAs. We select 250 hazy images of various haze densities from [25], and 7 representative DHAs are selected for quality evaluation. Then a subjective quality evaluation study is conducted on the DHQ database. The performance of all compared DHAs are

analyzed and the mean opinion scores (MOSs) of all dehazed images are collected for the following DHA evaluation study. To the best of our knowledge, the DHQ database is the largest dehazing quality database so far which includes human labeled ground-truth quality scores of the dehazed images.

To evaluate DHAs on real hazy images, we propose a dehazing quality index (DHQI) for dehazed images. A good DHA should be able to remove haze as much as possible, preserve image structures from damage, and avoid side-effects such as over-enhancement. Thus we extract 3 groups of features, including 1) haze-removing features, 2) structure-preserving features, and 3) over-enhancement features, to describe the above 3 key objectives of dehazing. The extracted features are integrated to the overall DHQI via a regression module, which is trained using the collected subjective evaluation data. The effectiveness of DHQI is verified on the DHQ database and 3 other databases with synthetic haze. DHQI can be used to evaluate DHAs or optimize practical dehazing systems.

One major use of dehazing quality measure is to evaluate DHAs, thus we give an overview of the current DHA quality evaluation methods and discuss their merits and demerits. We use the subjective quality evaluation data collected in this study and another database to compare the two typical strategies of quantitative DHA evaluation, and based on the analyses, we give some suggestions on the overall and systematic DHA quality evaluation. The overview, discussions, and suggestions is the third contribution of this study besides the constructed DHQ database and the proposed DHQI.

The rest of this paper is organized as follows. Section II presents the subjective quality evaluation study. In Section III, we describe the details of the proposed DHQI. Experimental verification is given in Section IV. In Section V, we give some discussions and suggestions on systematic quality evaluation of DHAs. Section VI concludes this paper.

II. Subjective Quality Evaluation of Dehazed Images

For further objective dehazing quality evaluation study, we construct a large scale dehazing quality (DHQ) database, which includes 1,750 dehazed images, and conduct a subjective quality evaluation study on the DHQ database. To the best of our knowledge, the DHQ database is the largest dehazing quality database which includes human labeled ground-truth quality scores of the dehazed images.

A. Hazy and Dehazed Images

We select 250 hazy images from a total of 500 hazy images used in [25]. The images are labeled with the human-rated haze densities. We select 250 of them which have various haze densities to test the DHAs' effectiveness under different haze conditions. Seven representative DHAs, including Fattal08 [6], Tarel09 [7], He09 [8], Xiao12 [9], Meng13 [10], Tang14 [11], and Lai15 [12], are selected to process the hazy images. A total of 1,750 dehazed images are generated. All dehazed images and the corresponding hazy images constitute the DHQ database.

TABLE II
Subjective Experiment Settings

Category	Item	Detail
Monitor	Model	SONY KD-85X8500D
	Resolution	3840×2160
Methodology	Method	Double-stimulus
	Quality-scale	5-grade categorical
	Presentation order	Random
Test settings	Sessions	5
	Subjects number	51 valid / 3 outliers 32 male / 22 female
	Viewing distance	3 times screen height
	Environment	Laboratory

B. Subjective Quality Evaluation

We conduct a subjective quality evaluation study on the DHQ database. A double-stimulus strategy is adopted, and both the hazy and dehazed images are shown side-by-side. Subjects are asked to rate the overall dehazing quality using a five-grade categorical rating scale. Subjects are suggested to give an overall quality rating mainly from 2 aspects: if the haze is totally removed, and if the DHA introduces any artifacts. Each pair of images is shown for 2 seconds, with a 1 second gray image shown in between. All 1,750 images are randomly and evenly divided into 5 sessions. A total of 54 subjects participate in the tests, and most of them take part in 3 sessions, with a few subjects only take part in 1 session. Each image is rated by 30 subjects. A 5 minutes break is given between sessions to avoid fatigue. All test images are randomly shown on a LED monitor, which is calibrated according to the recommendations given by ITU-R BT.500-13 [26]. The subjects are seated at a viewing distance of around 3 times the screen height in a laboratory environment with normal indoor illumination conditions. A summary of the experiment settings is given in Table II.

C. Subjective Data Processing

We follow the practices in [24] and [27] to process the raw subjective ratings. If the raw quality rating for an image is far from the average (2 or $\sqrt{20}$ standard deviations (stds) for the Gaussian or non-Gaussian case), it is detected as outlier, and a subject with more than 5% outlier ratings is detected as outlier subject. Both outlier ratings and outlier subjects are excluded from following processes. The ratings for every subject is normalized, and the normalized ratings for an image are averaged over all valid subjects to the mean opinion score (MOS). The MOSs are taken as the ground-truth quality of the dehazing, and are included in the DHQ database. Fig. 2 illustrates a histogram of the MOSs of the DHQ database. It is observed that the perceptual quality spreads over the whole quality range.

III. Objective Dehazing Quality Evaluation

To evaluate DHA quantitatively, we propose a dehazing quality index (DHQI) for dehazed images. We measure

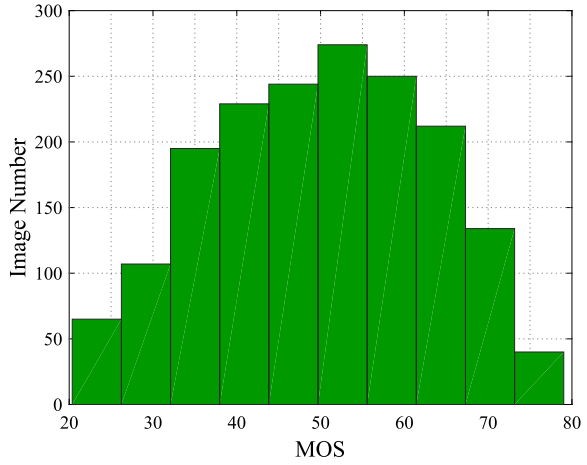


Fig. 2. Histogram of the MOSs of the DHQ database.

the quality from the following 3 key aspects of dehazing: haze-removing, structure-preserving, and over-enhancement. We extract 3 groups of features to measure the quality of these 3 aspects, and integrate these features into an overall dehazing quality.

A. Pre-Processing

The extracted 3 groups of features share some common processes, which are described here as the pre-processing. Given an image $\mathbf{I}$, we first compute the local mean and variance [27]–[29]

$$\boldsymbol{\mu}(i,j) = \sum_{k,l} \mathbf{w}(k,l)\mathbf{I}(i+k, j+l), \tag{2}$$

$$\boldsymbol{\sigma}(i,j) = \Big[\sum_{k,l} \mathbf{w}(k,l)\big(\mathbf{I}(i+k, j+l) - \boldsymbol{\mu}(i,j)\big)^2\Big]^{\frac{1}{2}}, \tag{3}$$

where i, j are pixel indexes, $\boldsymbol{\mu}$ is the local mean, and $\mathbf{w}$ is a local Gaussian weighting window. Then we normalize the image

$$\hat{\mathbf{I}}(i,j) = \frac{\mathbf{I}(i,j) - \boldsymbol{\mu}(i,j)}{\boldsymbol{\sigma}(i,j) + 1}. \tag{4}$$

For the dehazed image $\mathbf{I}_d$ and hazy image $\mathbf{I}_h$, we follow the same processes and compute their local mean, variance, and normalized image as $\boldsymbol{\mu}_d$, $\boldsymbol{\mu}_h$, $\boldsymbol{\sigma}_d$, $\boldsymbol{\sigma}_h$, $\hat{\mathbf{I}}_d$, $\hat{\mathbf{I}}_h$, where the subscripts d and h indicate the dehazed and hazy images, respectively. We only utilize the hazy image when measuring the image structure-preserving and over-enhancement, thus if without subscript, we generally indicate the dehazed image. Similarly, we only utilize color information when measuring the haze-removing, thus if without specific explanation, we are handling the converted gray-scale image.

B. Haze-Removing

Haze mainly introduces visibility problems such as contrast reduction and loss of image details, while dehazing tries to recover these lost contrast and image details. We use several haze-aware descriptors to detect the haze left in the dehazed image to evaluate the haze-removing effect. More haze left generally indicates worse dehazing quality.

A dark channel prior (DCP) is found in [8], which describes that at least one color channel has very low intensity in the haze-free images. The existence of haze can break such prior, and heavier haze generally leads to brighter dark channel. We use a pixel-wise DCP to measure the haze left in the dehazed image

$$\mathbf{I}_{dark}(i,j) = \min_{c\in\{R,G,B\}} \mathbf{I}_c(i,j), \tag{5}$$

where $c \in \{R, G, B\}$ indicates the RGB channels of the dehazed image.

Images with less haze generally contain more image details, thus we use image entropy to measure the image details of the dehazed image

$$H = -\sum_i p_i \log(p_i), \tag{6}$$

where $\mathbf{p} = [p_1, \ldots, p_{256}]$ denotes the histogram probability of the luminance of the dehazed image.

Another main objective of dehazing is to recover the contrast, we utilize 3 descriptors to measure the contrast of the dehazed image. First, the local variance calculated via Eq.(3) is used. Considering that the local variance $\boldsymbol{\sigma}$ generally varies with the local mean $\boldsymbol{\mu}$, we derive the normalized local variance as the second contrast feature

$$\boldsymbol{\eta} = \frac{\boldsymbol{\sigma}}{\boldsymbol{\mu} + 1}. \tag{7}$$

This feature has been previously used as haze-aware features in [24] and [25]. We extract the contrast energy (CE) [30] as the third contrast feature, which estimates perceived image local contrast. The CE has been previously used in [25] for haze density prediction, and it has been proved to be an effective haze-aware contrast feature. Specifically, CE is computed as

$$\mathbf{CE} = \frac{\rho \cdot Z(\mathbf{I})}{Z(\mathbf{I}) + \rho \cdot \kappa} - \tau, \tag{8}$$

where $Z(\mathbf{I}) = \sqrt{(\mathbf{I} \otimes \mathbf{g}_x)^2 + (\mathbf{I} \otimes \mathbf{g}_y)^2}$, $\otimes$ indicates convolution, $\mathbf{g}_x$ and $\mathbf{g}_y$ are the horizontal and vertical second-order derivatives of the Gaussian function, respectively, ρ is the maximum value of $Z(\mathbf{I})$, κ controls the contrast gain, and τ is the noise threshold used to constrain the noise. The readers can refer to [25] and [30] for more details of CE. Fig. 3 has illustrated examples of the related haze-removing feature maps utilized by the proposed method.

C. Structure-Preserving

Many IQA measures utilize structural features due to its effectiveness of capturing image degradations [28], [29], [31], [32]. Image structure is also an important cue for dehazing quality prediction, since DHAs sometimes can introduce structural artifacts. Fig. 4 illustrates two typical examples of structural artifacts introduced by dehazing. One typical failure occurs when some DHAs utilize an aggressive strategy trying to remove the haze completely, but they may damage the intrinsic image structures. Another typical structural

Fig. 3. An illustration of haze-removing feature maps. From left to right are $\mathbf{I}$, $\mathbf{I}_{dark}$, $\boldsymbol{\sigma}$, $\boldsymbol{\eta}$, $\mathbf{CE}$, respectively.

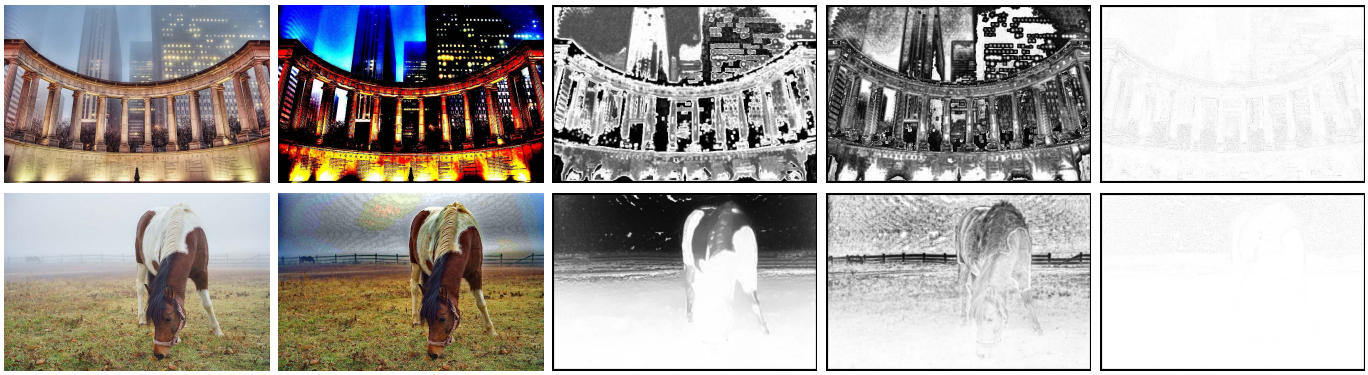


Fig. 4. Two typical structural artifacts introduced by DHAs: intrinsic structure damage (top) and over-enhancement (bottom). Related image structural feature maps are also illustrated, including $\mathbf{I}_h$, $\mathbf{I}_d$, $\mathbf{s}_\sigma$, $\mathbf{s}_\eta$, $\mathbf{s}_{\hat{I}}$ (from left to right).

artifact is the over-enhancement, which is generally observed in the low contrast areas. Some hardly-observed image details are taken as the image structures and enhanced out. We extract structure-preserving and over-enhancement features to describe such structural damage and over-enhancement, respectively.

In FR IQA, it is easy to capture structural damage by comparing the structures of the reference and distorted images. But we do not have a perfect quality reference when evaluating DHAs on real hazy images. Considering that we only utilize the structural features to describe the structural damage which significantly changes the image structures, we use the hazy image as a reference and measure the structural similarity between the hazy and dehazed images. For the purpose of measuring structural damage which occurs either in the texture-rich or very flat regions, the structure of the hazy image can roughly provide an approximate reference.

Specifically, we measure the image structure damage by computing the structural similarity between $\mathbf{I}_d$ and $\mathbf{I}_h$ using 3 structural features. First, variance similarity is derived

$$\mathbf{s}_\sigma = \frac{2\boldsymbol{\sigma}_d \cdot \boldsymbol{\sigma}_h + \epsilon_1}{\boldsymbol{\sigma}_d^2 + \boldsymbol{\sigma}_h^2 + \epsilon_1}, \tag{9}$$

where ϵ_1 is a constant used to avoid instability. Then normalized variance similarity is derived

$$\mathbf{s}_\eta = \frac{2\boldsymbol{\eta}_d \cdot \boldsymbol{\eta}_h + \epsilon_2}{\boldsymbol{\eta}_d^2 + \boldsymbol{\eta}_h^2 + \epsilon_2}, \tag{10}$$

where ϵ_2 has the same function as ϵ_1, $\boldsymbol{\eta}_d$ and $\boldsymbol{\eta}_h$ are the dehazed and hazy images' normalized variance calculated via Eq.(7). Finally is the normalized image similarity

$$\mathbf{s}_{\hat{I}} = \frac{2\hat{\mathbf{I}}'_d \cdot \hat{\mathbf{I}}'_h}{\hat{\mathbf{I}}'^2_d + \hat{\mathbf{I}}'^2_h}, \tag{11}$$

where $\hat{\mathbf{I}}'_d = \hat{\mathbf{I}}_d + 3$ and $\hat{\mathbf{I}}'_h = \hat{\mathbf{I}}_h + 3$ are the normalized dehazed and hazy images. We add a constant 3 to scale the normalized image to a positive range.

D. Over-Enhancement

As described in Section III-C and illustrated in Fig. 4, over-enhancement is another typical structural artifact. It is one kind of side-effect introduced by dehazing, and some hardly-observed image details are enhanced as image structures. As demonstrated in [24], such over-enhancement in the low contrast areas which include no obvious image structure is extremely harmful to the perceptual quality. We still utilize structural features to describe the over-enhancement. Specifically, the over-enhancement is described by the pooling of the structural similarity maps in the low contrast areas

$$o_\phi = \frac{1}{N} \sum_{(i,j)\in\Theta} \mathbf{s}_\phi(i,j), \tag{12}$$

where the subscript $\phi \in \{\sigma, \eta, \hat{I}\}$ indicates the structure features described above, N is a normalization factor representing the number of pixels in set Θ which indicates the low contrast areas and is defined as

$$\Theta = \big\{(i,j) | \boldsymbol{\sigma}_h(i,j) < E(\boldsymbol{\sigma}_h), \\ \boldsymbol{\sigma}_d(i,j) - \boldsymbol{\sigma}_h(i,j) > E(\boldsymbol{\sigma}_d - \boldsymbol{\sigma}_h)\big\}, \tag{13}$$

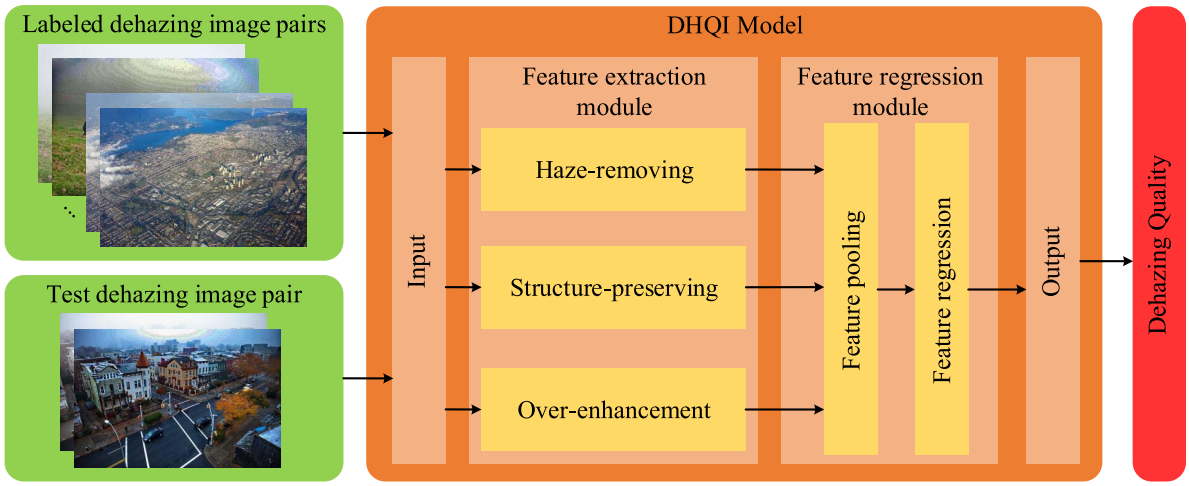


Fig. 5. A framework of the proposed DHQI method.

TABLE III
OVERVIEW OF FEATURES FOR DEHAZING QUALITY

Category	Feature ID	Feature description	Symbol	Computation
Haze-removing	f_1	Pixel-wise DCP	$\mathbf{I}_{dark}$	Eq.(5)
	f_2	Image entropy	H	Eq.(6)
	f_3	Local variance	$\boldsymbol{\sigma}$	Eq.(3)
	f_4	Normalized local variance	$\boldsymbol{\eta}$	Eq.(7)
	f_5	Contrast energy	$\mathbf{CE}$	Eq.(8)
Structure-preserving	f_6	Variance similarity	$\mathbf{s}_\sigma$	Eq.(9)
	f_7	Normalized variance similarity	$\mathbf{s}_\eta$	Eq.(10)
	f_8	Normalized image similarity	$\mathbf{s}_{\hat{I}}$	Eq.(11)
Over-enhancement	f_9-f_{11}	Over-enhancement in low contrast areas	o_σ, o_η, $o_{\hat{I}}$	Eq.(12)-Eq.(13)
	f_{12}	Blockiness	b	Eq.(14)

where $E(\cdot)$ calculates the average of all values in the matrix, Θ represents all pixels whose variance is lower than the average and variance enhancement is larger than the average.

Prior to dehazing, some hazy images undergo some compression. The compression artifacts are hardly observed in the hazy image due to proper control of the compression degree, but they may be taken as the image structures and enhanced out during dehazing. Considering that JPEG is the most widely used image compression method and blockiness is also easily enhanced during dehazing, we estimate blockiness as one kind of over-enhancement. We detect the corners and edges in the dehazed image, and calculate their regularity as the blockiness

$$b = \frac{\sum_{i,j} \mathbf{c}'(i,j)}{\sum_{i,j} \mathbf{c}(i,j))} \cdot \frac{\sum_{i,j} \mathbf{e}'(i,j)}{\sum_{i,j} \mathbf{e}(i,j))}, \tag{14}$$

where $\mathbf{c}$ is the corner map, in which $\mathbf{c}(i,j) = 1$ or 0 indicates that a corner is or is not detected at (i,j), $\mathbf{c}'$ is similar to $\mathbf{c}$, but only corners at the 8×8 block boundaries are detected, $\mathbf{e}$, $\mathbf{e}'$ are edge maps similar to $\mathbf{c}$, $\mathbf{c}'$, respectively. This feature has been proven effective for predicting the quality of block-based compressed images and videos in [27] and [31].

E. Feature Pooling and Regression

As illustrated in Fig. 5, DHQI mainly consists of two modules: feature extraction described above, and feature regression described in this section. The extracted 3 groups of features capture different aspects of the dehazing, including haze-removing, structure-preserving, and over-enhancement. These features include both single feature values, e.g., image entropy, over-enhancement, and 2D feature maps, e.g., DCP, (normalized) local variance, contrast energy, structural similarity. To predict a single value quality score from the extracted features, we first conduct feature pooling for the 2D feature maps. Though content-based or visual attention-based pooling has been proven effective in IQA [33]–[35], mean pooling is adopted in our method for simplicity. All features are pooled into single values, and are catenated to a feature vector $\mathbf{f} = [f_1, f_2, \ldots, f_{12}]$. An overview of the extracted features is given in Table III.

The last step is feature regression. Considering the successes of support vector regression (SVR) and random forest (RF), we select SVR and RF for regression. As illustrated in Fig. 5, the model training and testing share the same framework. We use labeled dehazing image pairs to train the regressor,

which then can be utilized to predict the quality of any dehazed image pair. Given the features $\mathbf{f}_i = [f_1, \ldots, f_{12}]$, the corresponding quality label q_i (MOS) and the training image set Ψ, we can train the $regressor$ using either SVR or RF

$$regressor = \text{TRAIN}(\mathbf{f}_i, q_i), \quad i \in \Psi, \tag{15}$$

where i is the image index. Then given the quality feature $\mathbf{f} = [f_1, \ldots, f_{12}]$ of any test image, we can predict the quality using the pre-trained $regressor$

$$q = \text{PREDICT}(\mathbf{f}, regressor). \tag{16}$$

We mainly use SVR as the regressor in this paper, but we will also test the RF regressor in the experiments. LIBSVM [36] is adopted to implement SVR with a radial basis function (RBF) kernel. We follow the common SVR parameter settings used in the training of mainstream IQA measures. The RF implementation for MATLAB is utilized [37], and we use the default parameters. DHQI needs labeled data for training, but after training it can predict the quality of dehazing using any dehazing image pairs.

IV. Experimental Results

A. Experimental Protocol

1) Test Databases: The proposed DHQI is validated on the following two categories of databases:

- Database with real haze, i.e., the DHQ database constructed in this paper. It includes 1,750 dehazed images generated from 250 real hazy images of various haze densities using 7 representative DHAs, and the corresponding subjective rating data. An overview of this database has been given in Section II. The proposed DHQI is mainly validated on this database.
- Databases with synthetic haze, including the SHRQ [24] database, and the reprocessed D-HAZY [38] and FRIDA [39] databases. Though DHQI is mainly designed for dehazing quality evaluation using real hazy images, we test DHQI on dehazed images generated from synthetic hazy images as a complementary. The SHRQ database (regular image subset) consists of 360 dehazed images which are generated from 45 hazy images synthesized from 45 reference haze-free images, and the corresponding subjective rating data. The evaluated DHAs include Fattal08 [6], Tarel09 [7], He09 [8], Xiao12 [9], Meng13 [10], Lai15 [12], Berman16 [13], and Cai16 [14]. In the D-HAZY and FRIDA databases, only synthetic hazy images and the reference haze-free images are available. We reprocess them by generating the dehazed images from the synthetic hazy images using the same 8 DHAs used in the SHRQ database. A total of 184, 576 dehazed images generated from 23, 72 synthetic hazy images are available in the reprocessed D-HAZY and FRIDA databases, respectively. Note that only the Middelbury subset of the original D-HAZY database is used to reduce the computation. Since no subjective rating rata is available in the reprocessed D-HAZY and FRIDA databases, we use the quality scores computed by the specifically designed FR dehazing quality measure Min18 [24] as the ground-truth quality scores.

All the above databases including the DHQ, SHRQ, and reprocessed D-HAZY and FRIDA databases will be publicly available to facilitate further research.

2) Competitors: Besides DHQI, we also test some quality measures which may be effective for dehazing quality evaluation. Specifically, the following two types of quality measures are tested:

- Blind evaluators related to haze removal and contrast-enhancement, including quality measures for contrast-enhanced images, e.g., BIQME [19], Fang15 [20], NIQMC [22], the three evaluators e, r and NS introduced in [23], and the haze density estimator FADE [25].

General-purpose

- blind IQA measures, including BRISQUE [40], CORNIA [41], IL-NIQE [42], BPRI [43], and BMPRI [44] which are assumed to be able to handle general IQA problems.

We believe the above two types of measures have included the possible quality measures which may be effective for dehazing quality evaluation. For all competitors, we use the original implementations released by the authors.

3) Evaluation Criteria: We use the following five-parameter logistic function which is frequently used in IQA model evaluation to map the predicted quality scores [27], [43], [44]

$$q' = \lambda_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\lambda_2 (q - \lambda_3)}} \right) + \lambda_4 q + \lambda_5, \tag{17}$$

where q, q' are the original and mapped quality scores, respectively, $\{\lambda_i | i = 1, 2, \ldots, 5\}$ are five parameters determined through curve fitting using q and MOSs. Then the consistency between q' and MOSs is measured as the performance of the IQA model. We choose the following 3 commonly used consistency evaluation criteria:

- Spearman rank-order correlation coefficient (SRCC), which measures the monotonicity of the IQA model.
- Pearson linear correlation coefficient (PLCC), which measures the IQA model's prediction linearity.
- Root-mean-square error (RMSE), which is a prediction accuracy measure.

B. Performance Evaluation With Real Haze

Since DHQI is designed for dehazing quality evaluation using real hazy images, we mainly test DHQI on the DHQ database constructed in this paper.

1) Performance Comparison: As described in Section III-E, DHQI includes a regression module which requires training. We follow the common practices of opinion-aware IQA model training [27], [40], [45], [46], and split the whole DHQ database into a training set and a testing set, which are completely separated. The training set includes a percentage of $ratio$ dehazed images which are randomly selected, and the rest $1 - ratio$ dehazed images are left to the testing set. The dehazed images corresponding to the same hazy image are divided into the same set to ensure a complete

TABLE IV
PERFORMANCE EVALUATION WITH REAL HAZE (ON THE DHQ DATABASE)

Ratio	Criteria	A	B	C	D	E	F	G	H	I	J	K	L	M
		BIQME	Fang15	NIQMC	e	r	NS	FADE	BRISQUE	CORNIA	IL-NIQE	BPRI	BMPRI	DHQI
80%	SRCC	0.2596	0.4340	0.2773	0.1400	0.1461	0.4114	0.2517	0.6829	0.2467	0.5950	0.2486	0.7054	**0.8622**
	PLCC	0.3301	0.5265	0.4028	0.2617	0.5315	0.5080	0.2680	0.7229	0.3189	0.6595	0.3602	0.7437	**0.8737**
	RMSE	12.353	11.126	11.982	12.582	11.071	11.268	12.588	9.0227	12.367	9.8177	12.168	8.7325	**6.3744**
50%	SRCC	0.2617	0.4261	0.2778	0.1425	0.1463	0.4139	0.2517	0.6568	0.2461	0.5961	0.2520	0.6957	**0.8558**
	PLCC	0.3156	0.5106	0.3907	0.2701	0.5276	0.5037	0.2552	0.6927	0.3035	0.6557	0.3397	0.7299	**0.8647**
	RMSE	12.439	11.274	12.072	12.644	11.141	11.350	12.676	9.4421	12.499	9.9138	12.329	8.9743	**6.5966**
20%	SRCC	0.2601	0.3999	0.2770	0.1418	0.1453	0.4147	0.2526	0.5897	0.2489	0.5964	0.2521	0.6680	**0.8380**
	PLCC	0.3096	0.4814	0.3871	0.2737	0.5261	0.5024	0.2554	0.6264	0.3009	0.6551	0.3334	0.7023	**0.8457**
	RMSE	12.488	11.507	12.109	12.639	11.174	11.357	12.701	10.234	12.525	9.9261	12.375	9.3420	**7.0082**
	Time	1.0550	**0.0291**	1.3412	5.1335	5.1335	5.1335	0.6052	0.4271	2.6430	3.2436	0.6022	1.0646	0.5401

separation of the training and testing data. We repeat this training-testing process for 1,000 times and report the median SRCC, PLCC, and RMSE performance. For the training-free methods, we conduct the same splitting and only test the performance on the testing set for fair comparison. Normally, a 80% train – 20% test split is adopted in the IQA literature. We also adopt this split strategy, but we add two more strategies: 50% train – 50% test, 20% train – 80% test to test the models' dependency on the amount of training data.

The performance is listed in Table IV, from which we have several observations. First, DHQI performs the best among all models, which verifies the effectiveness of the proposed method. Second, some general-purpose blind IQA measures have certain ability to predict the quality of dehazed images after retraining. It is because a lot of measures are based on natural scene statistics (NSS), while the distortions of dehazing, for example, the existence of haze and the structural damage, can violate the NSS and be captured by these models. But these measures have not considered the characteristics of dehazing, and they are not effective enough for dehazing quality evaluation. Third, the models which require MOSs for training perform better, which is not surprising since these models can adapt themselves to the dehazing distortions. Fourth, though some measures are designed for contrast enhancement quality assessment, e.g., BIQME, Fang15, and NIQMC, their performance is not impressive. It is mostly due to that the dehazing quality is more complicated than contrast enhancement quality as described in Section I. Last, DHQI shows quite high performance (a SRCC of 0.8380) even only 20% of data is used for training. Moreover, compared with other training-based measures, DHQI shows the least performance drop when reducing the training data, which suggests good model generalizability.

2) Statistical Significance Test: We conduct statistical test to verify if the performance differences between models are significant. Specifically, t-test [47] is conducted on the SRCC values obtained from the 1,000 80% train – 20% test splits. We compare every pairs of models, and list the results in Table V. A 1/0/- symbol indicates that the row model is statistically better than/worse than/indistinguishable from the column model (with a confidence of 95%). Similar results can be obtained using other split strategies and performance evaluation criteria. The significant superiority of DHQI is evident. Most observations described in the previous paragraph are also proved to be significant.

TABLE V
STATISTICAL SIGNIFICANCE TEST RESULTS ON THE DHQ DATABASE. A 1/0/- SYMBOL INDICATES THAT THE ROW MODEL IS STATISTICALLY BETTER THAN/WORSE THAN/ INDISTINGUISHABLE FROM THE COLUMN MODEL (WITH 95% CONFIDENCE), RESPECTIVELY. A-M ARE MODEL INDEXES GIVEN IN TABLE IV

	A	B	C	D	E	F	G	H	I	J	K	L	M
A	-	0	1	1	1	0	1	0	1	0	1	0	0
B	1	-	1	1	1	1	1	0	1	0	1	0	0
C	0	0	-	1	1	0	1	0	1	0	1	0	0
D	0	0	0	-	-	0	0	0	0	0	0	0	0
E	0	0	0	-	-	0	0	0	0	0	0	0	0
F	1	0	1	1	1	-	1	0	1	0	1	0	0
G	0	0	0	1	1	0	-	0	-	0	-	0	0
H	1	1	1	1	1	1	1	-	1	1	1	0	0
I	0	0	0	1	1	0	-	0	-	0	0	0	0
J	1	1	1	1	1	1	1	0	1	-	1	0	0
K	0	0	0	1	1	0	-	0	1	0	-	0	0
L	1	1	1	1	1	1	1	1	1	1	1	-	0
M	1	1	1	1	1	1	1	1	1	1	1	1	-

3) Feature Analysis: To have an intuitive understanding of how the DHQI features correlate with the human rated dehazing quality, we illustrate the correlation between single features and the MOSs in Fig. 6. Note that no training is utilized here, we directly test each feature's SRCC and PLCC performance on the overall DHQ database. It is observed that some single features show quite competitive performance, which is comparable to current best-performing algorithms even though they are trained on this database. Another observation is that several best-performing single features are structure-relevant features. It suggests that avoiding structural artifacts is the most important cue for dehazing quality. More detailed verification is given in the ablation experiment.

4) Contribution of Different Components: We conduct several ablation experiments to test the contributions of different

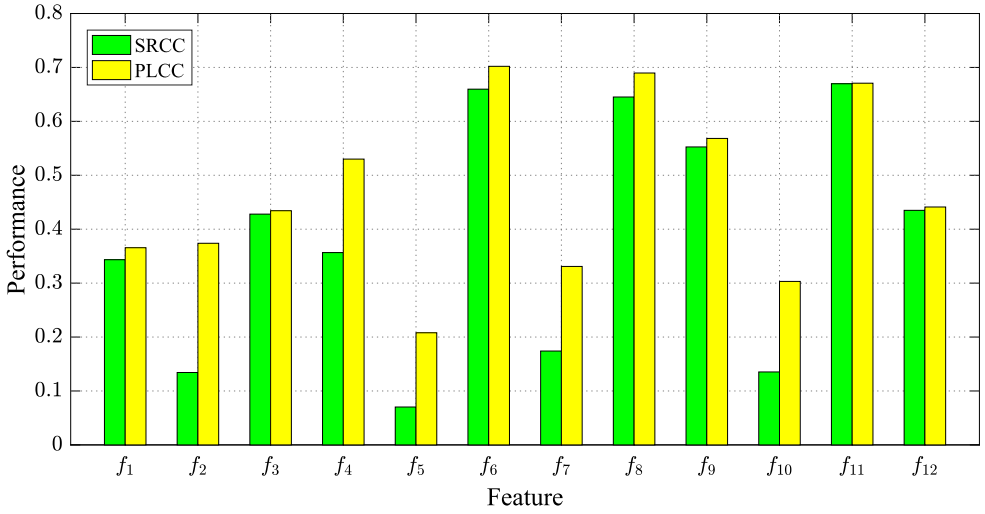


Fig. 6. Performance (SRCC and PLCC) of single features on the DHQ database. f_1-f_{12} are feature IDs given in Table III.

TABLE VI
PERFORMANCE OF DIFFERENT FEATURE GROUPS

Ratio	Criteria	*G1*	*G2*	*G3*	*G4*	*G5*	*G6*
80%	SRCC	0.6234	0.7893	0.7353	0.8346	0.8054	0.8408
	PLCC	0.6865	0.8080	0.7460	0.8489	0.8263	0.8570
	RMSE	9.4947	7.7011	8.6955	6.9043	7.3700	6.7412
50%	SRCC	0.6213	0.7883	0.7370	0.8301	0.8003	0.8374
	PLCC	0.6783	0.8040	0.7432	0.8420	0.8173	0.8503
	RMSE	9.6401	7.8031	8.7666	7.0830	7.5575	6.9073
20%	SRCC	0.6092	0.7839	0.7310	0.8177	0.7850	0.8251
	PLCC	0.6652	0.7991	0.7342	0.8293	0.8001	0.8382
	RMSE	9.7930	7.8882	8.9084	7.3463	7.8816	7.1669

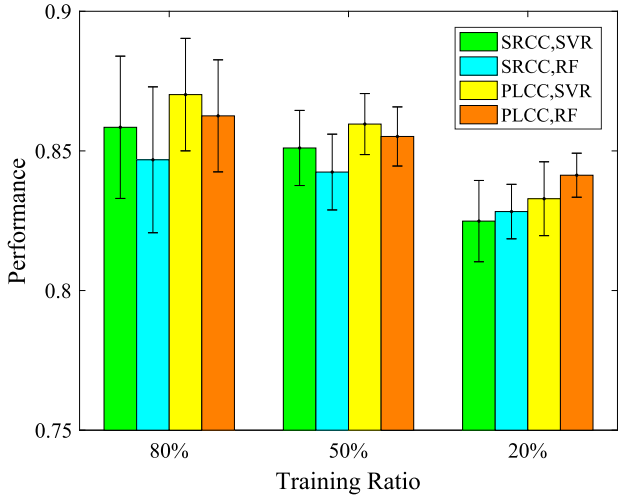


Fig. 7. Performance of DHQI using different regressors.

feature components of DHQI. Specifically, features listed in Table III are grouped according to the category, and we test the contributions of different feature groups. Specifically, we test the performance of the following feature groups:

- *G1*: Only haze-removing features.
- *G2*: Only structure-preserving features.
- *G3*: Only over-enhancement features.
- *G4*: Structure-preserving and over-enhancement features.
- *G5*: Haze-removing and over-enhancement features.
- *G6*: Haze-removing and structure-preserving features.

The same training-testing processes described in Section IV-B1 are conducted, and the results are summarized in Table VI.

Agreeing with the analyses given in Section IV-B3, it is observed that structure-preserving features contribute the most to DHQI. Besides, using only structure-preserving or over-enhancement features can achieve considerable performances (SRCC of 0.7893 or 0.7353 when 80% of data is used for training), which suggests that avoiding structural artifacts as described in Section III-C and Section III-D is very important for dehazing from a perceptual quality perspective of view. Haze-removing features also contribute to DHQI, which is not surprising since removing haze is the primary objective of dehazing and more haze left in the dehazed image indicates lower dehazing quality.

5) Performance of Using Different Regressors: As described in Section III-E, different regressors can be utilized to predict the final quality. We test the performance of DHQI using different regressors, i.e., SVR and RF. The same training-testing processes described in Section IV-B1 are conducted, and the performance is illustrated in Fig. 7. It is observed that DHQI achieves considerable performances when using SVR or RF, which suggests that what contributes most to DHQI is feature extraction rather than feature regression.

6) Computational Complexity: It is important for algorithms to have low computational complexity in practical use. We analyze the computational complexity of all compared measures by comparing the average running cost (seconds/image). The experiments are conducted with MATLAB R2016a operating on a computer with Intel Core i7-6700K CPU @4.00 GHz and 32 GB RAM. We select 100 images with a fixed resolution of 512×512 as the test set. The results

TABLE VII
PERFORMANCE EVALUATION WITH SYNTHETIC HAZE (ON THE SHRQ DATABASE)

Ratio	Criteria	BIQME	Fang15	NIQMC	e	r	NS	FADE	BRISQUE	CORNIA	IL-NIQE	BPRI	BMPRI	DHQI
80%	SRCC	0.2862	0.4812	0.4145	0.2580	0.0071	0.0112	0.2977	0.4844	0.1137	0.3437	0.0130	0.5063	**0.6794**
	PLCC	0.3397	0.6775	0.6119	0.4496	0.4853	0.2618	0.3238	0.7194	0.3290	0.6380	0.2952	0.7142	**0.8124**
	RMSE	12.842	9.9933	10.808	12.044	11.904	13.095	12.867	9.4756	12.802	10.534	12.963	9.5439	**7.9382**
50%	SRCC	0.2761	0.4589	0.4048	0.2384	0.0164	0.0225	0.2950	0.4219	0.1189	0.3437	0.0092	0.4749	**0.6556**
	PLCC	0.2813	0.6370	0.5731	0.3442	0.4486	0.2088	0.2792	0.6245	0.2161	0.5955	0.2209	0.6625	**0.7825**
	RMSE	13.248	10.670	11.304	12.807	12.356	13.509	13.277	10.773	13.437	11.089	13.463	10.377	**8.6042**
20%	SRCC	0.2736	0.4272	0.4003	0.2311	0.0162	0.0198	0.2942	0.2765	0.1229	0.3358	0.0138	0.3930	**0.5999**
	PLCC	0.2710	0.5760	0.5567	0.3004	0.4358	0.1861	0.2705	0.4559	0.1725	0.5882	0.1933	0.5692	**0.7291**
	RMSE	13.336	11.320	11.482	13.148	12.470	13.605	13.343	12.335	13.629	11.178	13.587	11.358	**9.4674**

are listed in Table IV. For all competitors, we use the original implementations released by the authors. It is observed that DHQI has considerable low computational complexity and it is one of the fast measures.

C. Performance Evaluation With Synthetic Haze

As described in Section I, DHAs in the current literature are evaluated via two strategies: using synthetic hazy images and using real hazy images. An overall discussion of the strategy of using synthetic hazy images is given in [24]. Under this strategy, the haze-free images are available, and they can be used as the reference. Under this circumstance, FR IQA measures generally perform better, and it is less meaningful to develop and use blind measures. The proposed DHQI follows another strategy and it is designed for dehazing quality evaluation using real hazy images. Since the reference is not available, DHQI has to perform quality evaluation in a NR manner. Though DHQI is designed for real haze, we still try it with synthetic haze to test its generalizability.

1) Evaluation on the SHRQ Database (With Subjective Data): The SHRQ database is constructed in [24], and subjective quality evaluation data of all dehazed images is also available. A brief introduction of the SHRQ database has been given in Section IV-A1. Readers can also refer to [24] for more details. We test the DHQI and competitors on the regular image subset of this database. Specifically, we follow the same evaluation protocols described in Section IV-B1, and compare DHQI with the same competitors on the SHRQ database. The experimental results are summarized in Table VII. Though DHQI is not so effective as evaluating real dehazed images, it still shows considerable performance, and it performs the best among all competitors. It suggests that DHQI is also effective for synthetic haze removing quality evaluation, though FR measures are often better choices under such circumstances.

Besides NR measures, FR IQA measures can be utilized when using synthetic hazy images. According to the performance on the SHRQ database, we select the 5 best-performing FR measures, including VIF [48], FSIM [49], GMSD [50], PSIM [51], as well as the specifically designed FR dehazing quality measure Min18 [24], and then compare them with DHQI. Since DHQI requires training, we adopt the same 80% train – 20% test split used in previous experiments. In Fig. 8,

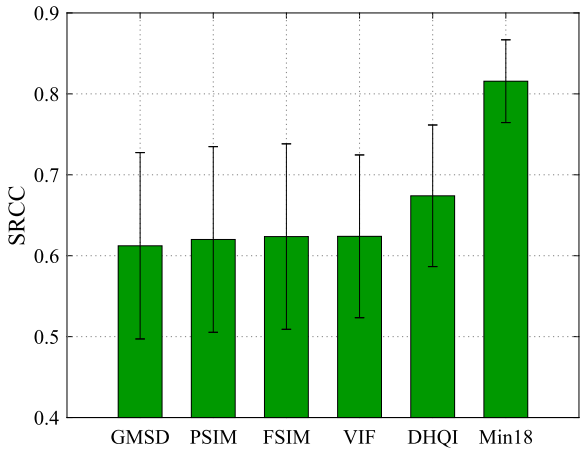


Fig. 8. Mean and standard error bar of the SRCC values obtained from the 1,000 train-test trials for the best-performing FR IQA measures on the SHRQ database.

we illustrate the mean and std of the SRCC values obtained from the 1,000 train-test splits. It is not surprising that the specifically designed FR dehazing quality measure Min18 [24] performs the best. Except for Min18 [24], DHQI performs the best, though it does not utilize the haze-free image as a reference while the rest are FR measures.

2) Evaluation on the Reprocessed D-HAZY and FRIDA Databases (Without Subjective Data): Besides the SHRQ database which includes subjective rating data, there are also some databases without subjective data, for example the D-HAZY [38] and FRIDA [39] databases. These databases cionsist of synthetic hazy images and the corresponding reference haze-free images. We reprocess the D-HAZY and FRIDA databases by generating dehazed images from the synthetic hazy images using 8 representative DHAs and labeling the dehazed images using the specifically designed FR dehazing quality measure Min18 [24]. A brief introduction of the reprocessed D-HAZY and FRIDA databases has been given in Section IV-A1. We follow the same evaluation protocols described in Section IV-C1 and Section IV-B1. The only difference is that we use the objective scores labeled by Min18 [24] to replace the subjective rating scores. The performance comparison results are listed in Table VIII. Only SRCC

TABLE VIII
SRCC PERFORMANCE EVALUATION WITH SYNTHETIC HAZE (ON THE REPROCESSED D-HAZY AND FRIDA DATABASES)

Database	Ratio	BIQME	Fang15	NIQMC	e	r	NS	FADE	BRISQUE	CORNIA	IL-NIQE	BPRI	BMPRI	DHQI
D-HAZY	80%	0.4669	0.4106	0.4782	0.3337	0.1040	0.0697	0.2376	0.6098	0.0846	0.3406	0.0761	0.4734	**0.6493**
	50%	0.4522	0.3900	0.4844	0.3142	0.0868	0.0816	0.2318	0.5743	0.0480	0.3291	0.0531	0.4986	**0.6196**
	20%	0.4495	0.2935	0.4803	0.3176	0.0840	0.0707	0.2365	0.5800	0.0477	0.3236	0.0486	0.5635	**0.6186**
FRIDA	80%	0.0059	0.8690	0.2290	0.0700	0.0485	0.1445	0.2633	0.9043	0.0723	0.2324	0.1330	0.8605	**0.9165**
	50%	0.0059	0.8610	0.2346	0.0707	0.0475	0.1398	0.2662	0.8881	0.0647	0.2208	0.1341	0.8274	**0.9128**
	20%	0.0055	0.8311	0.2347	0.0696	0.0478	0.1389	0.2661	0.8609	0.0633	0.2210	0.1370	0.7350	**0.9024**

is reported for simplicity, but similar results can be obtained using other evaluation criteria. It is observed that DHQI shows the best performance among all competitors, which agrees with the previous validations on the DHQ and SHRQ databases.

V. DISCUSSIONS AND SUGGESTIONS ON SYSTEMATIC QUALITY EVALUATION OF DHAS

One major use of dehazing quality measure is to evaluate DHAs. In this section, we first summarize the current DHA evaluation methods and give some discussions. Then we compare the two typical strategies of quantitative DHA evaluation, and after that, we give some suggestions on conducting overall and systematic DHA quality evaluation.

A. Summary of Current DHA Quality Evaluation Methods

Compared with the substantial progress of DHA developing [6]–[15], the quality evaluation of DHAs falls short and needs more work. Though there are some DHA evaluation methods introduced and used in the literature, reliable and quantitative measure which correlates well with the overall dehazing quality still lacks. Generally, DHAs can be evaluated using the following methods:

- *Method 1*: Qualitative evaluation performed by human subjects. DHAs are directly tested on real hazy images.
- *Method 2*: Quantitative evaluation using synthetic hazy images. FR quality measures are utilized in this method.
- *Method 3*: Quantitative evaluation using real hazy images. NR quality measures are needed in this method.

Method 1 and *Method 3* follow the same strategy of using real hazy images, while *Method 2* follows another strategy of using synthetic hazy images.

An overview of the above evaluation methods has been given in Section I. The advantages and drawbacks of these methods are also discussed. *Method 1* is reliable and accurate, but it is difficult to conduct large scale evaluation. *Method 2* is easy to follow, but DHA evaluation using this method may not be an exact FR IQA problem [24]. Moreover, synthetic haze may be different from real haze. *Method 3* is desirable, but effective NR quality measure is needed, which motivates us to construct the DHQ database and propose the DHQI. More comparison of *Method 2* and *Method 3* is given in the following section.

Fig. 9. Problem of using the haze-free image as the ground-truth: the dehazed image may not be so close to the reference haze-free image, but it still has high perceptual quality. Two examples are shown in this figure.

B. Comparison of Quantitative DHA Evaluation Using Real and Synthetic Hazy Images

Both *Method 2* and *Method 3* can be utilized for quantitative dehazing quality evaluation. Current literature generally selects one of them for quantitative evaluation. Due to the convenience of conducting evaluation and comparison, *Method 2* has been widely utilized [11], [12], [14], [15], [24]. Some basic FR IQA measures like PSNR and SSIM [28] are directly used in these papers. A comprehensive study of this method is conducted in [24]. As illustrated in Fig. 8, state-of-the-art FR IQA measures are not effective enough. The main reason is that closing to the haze-free image can guarantee a pretty good dehazing quality, but some dehazed image still has good perceptual quality though it is not so close to the reference. This problem is easily observed in Fig. 9. We have considered this problem, incorporated some haze-aware features, and proposed an effective measure for this strategy in [24]. As described in Section I and illustrated in Fig. 1, another problem of *Method 2* lies in that real haze may be different from the synthesized haze. Few work has discussed this problem before. In this section, we will test the effectiveness of using synthetic hazy images through subjective evaluation data.

Method 3 is a more straightforward way, but it is not easy to conduct quantitative evaluation due to its NR nature and the complexity of dehazing. Due to the lack of reliable measures, it is less used in [7]. As described in Section I, though some measures are proposed for this objective [20], [21], [23], they do not correlate well with the overall dehazing quality. The proposed DHQI evaluates the dehazed image from an overall

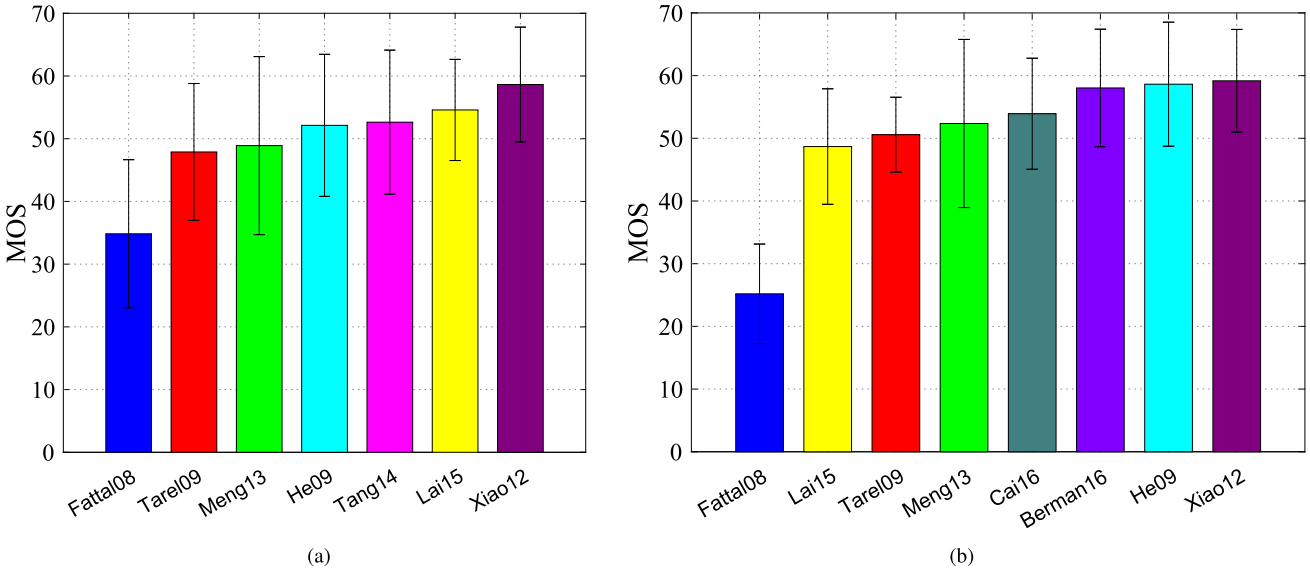


Fig. 10. A comparison of subjective DHA evaluation results using two different strategies. The two databases share 6 DHAs, and the same color indicates the same DHA. (a) The DHQ database (using real hazy images). (b) The SHRQ database (using synthetic hazy images).

quality perspective of view, and provides an effective measure for this strategy.

We use the subjective evaluation data collected in the DHQ and SHRQ (regular image subset) databases to analyze the effectiveness of the strategy of using synthetic hazy images. Specifically, we calculate the mean and std of the MOSs of all dehazed images generated from the same DHA, and illustrates the results in Fig. 10. The mean MOS is used as the ground-truth quality of the DHA. Berman16 [13] and Cai16 [14] are two DHAs not considered in the DHQ database. The two databases share 6 DHAs, i.e., Fattal08, Tarel09, Meng13, He09, Lai15, and Xiao12. We compare the evaluation results of the shared DHAs, and analyze the consistency of the two databases. On synthetic hazy images, the relative performance rank of the shared DHAs is

Fattal08 < Lai15 < Tarel09 < Meng13 < He09 < Xiao12.

While on real hazy images, the rank becomes

Fattal08 < Tarel09 < Meng13 < He09 < Lai15 < Xiao12.

It is observed that on one hand, 5 of 6 DHAs share the same rank using these two different evaluation strategies, which indicates that using synthetic hazy images is reliable to a certain degree. On the other hand, the last DHA Lai15 performs not well on synthetic hazy images, but it shows pretty good performance on real hazy images, which suggests that using synthetic hazy images may not be so accurate due to the differences between real and synthetic haze.

C. Suggestions on Systematic DHA Quality Evaluation

Current papers generally perform qualitative evaluation using several examples, and then select one from *Method 2* and *Method 3* to conduct quantitative evaluation. We suggest keeping the qualitative evaluation, and using both quantitative evaluation methods for an overall and systematic DHA evaluation. Owing to the availability of the reference, FR DHA evaluation is usually more stable than the NR one. But considering that different DHAs have different abilities to transfer from synthetic haze to real haze, evaluation on real hazy images is still needed. We think that the qualitative evaluation and the quantitative evaluation using both synthetic and real hazy images together can give a more systematic evaluation, while the proposed DHQI provides an effective measure for the strategy of using real hazy images.

VI. Conclusion

Though DHAs can be evaluated using synthetic hazy images, this strategy may not be reliable due to the differences between synthetic haze and real haze. In this paper, we evaluate DHAs by assessing the quality of real dehazed images directly and study this strategy systematically. As a major contribution, we first construct a dehazing quality (DHQ) database, which is the largest of its kind and includes 1,750 dehazed images generated from 250 hazy images of various haze densities. A subjective quality evaluation study is then conducted. Considering that the key objectives of dehazing is to remove the haze, preserve the intrinsic image structures, and prevent over-enhancement, we propose an objective dehazing quality index (DHQI) by extracting and integrating 3 groups of features which are responsible for the above 3 objectives. DHQI is validated on the constructed DHQ database, and besides that, DHQI shows certain prediction ability for dehazed images generated from synthetic hazy images. The proposed DHQI is another key contribution of this paper, and it can be utilized to evaluate DHAs quantitatively and optimize practical dehazing systems. The last contribution lies in that we give an overview and discussion of DHA quality evaluation methods. Based on subjective data, we suggest that the qualitative evaluation and the quantitative evaluation using both synthetic and real hazy images together can give an overall and systematic evaluation of the DHAs.

REFERENCES

[1] J. P. Tarel, N. Hautière, L. Caraffa, A. Cord, H. Halmaoui, and D. Gruyer, "Vision enhancement in homogeneous and heterogeneous fog," *IEEE Intell. Transp. Syst. Mag.*, vol. 4, no. 2, pp. 6–20, 2012.

[2] M. Negru, S. Nedevschi, and R. I. Peter, "Exponential contrast restoration in fog conditions for driving assistance," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2257–2268, Aug. 2015.

[3] R. Gallen, A. Cord, N. Hautière, É. Dumont, and D. Aubert, "Nighttime visibility analysis and estimation method in the presence of dense fog," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 310–320, Feb. 2015.

[4] H. Kuang, X. Zhang, Y.-J. Li, L. L. H. Chan, and H. Yan, "Nighttime vehicle detection based on bio-inspired image enhancement and weighted score-level feature fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 927–936, Apr. 2017.

[5] M. Rezaei, M. Terauchi, and R. Klette, "Robust vehicle detection and distance estimation under challenging lighting conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2723–2743, Oct. 2015.

[6] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, p. 72, Aug. 2008.

[7] J.-P. Tarel and N. Hautière, "Fast visibility restoration from a single color or gray level image," in *Proc. IEEE Conf. Comput. Vis.*, Sep. 2009, pp. 2201–2208.

[8] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1956–1963.

[9] C. Xiao and J. Gan, "Fast image dehazing using guided joint bilateral filter," *Vis. Comput.*, vol. 28, nos. 6–8, pp. 713–721, Jun. 2012.

[10] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proc. IEEE Conf. Comput. Vis.*, Dec. 2013, pp. 617–624.

[11] K. Tang, J. Yang, and J. Wang, "Investigating haze-relevant features in a learning framework for image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2995–3000.

[12] Y.-H. Lai, Y.-L. Chen, C.-J. Chiou, and C.-T. Hsu, "Single-image dehazing via optimal transmission map under scene priors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 1, pp. 1–14, Jan. 2015.

[13] D. Berman *et al.*, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1674–1682.

[14] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[15] Y. Li, S. You, M. S. Brown, and R. T. Tan, "Haze visibility enhancement: A survey and quantitative benchmarking," *Comput. Vis. Image Understand.*, vol. 165, pp. 1–16, Dec. 2017.

[16] B. Li *et al.* (2017). "RESIDE: A benchmark for single image dehazing." [Online]. Available: https://arxiv.org/abs/1712.04143

[17] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Trans. Pattern Anal. Mach. Learn.*, vol. 25, no. 6, pp. 713–724, Jun. 2003.

[18] K. Ma, W. Liu, and Z. Wang, "Perceptual evaluation of single image dehazing algorithms," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 3600–3604.

[19] 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.

[20] 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.

[21] Z. Chen, T. Jiang, and Y. Tian, "Quality assessment for comparing image enhancement algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3003–3010.

[22] 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.

[23] N. Hautière, J.-P. Tarel, D. Aubert, and É. Dumont, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Anal. Stereol. J.*, vol. 27, no. 2, pp. 87–95, Jun. 2008.

[24] X. Min, G. Zhai, K. Gu, J. Zhou, X. Yang, and X. Guan, "Quality evaluation of image dehazing methods using synthetic hazy images," *IEEE Trans. Multimedia*, to be published.

[25] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3888–3901, Nov. 2015.

[26] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500-13, Jan. 2012.

[27] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5462–5474, Nov. 2017.

[28] 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.

[29] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, Nov. 2003, pp. 1398–1402.

[30] I. I. Groen, S. Ghebreab, H. Prins, V. A. F. Lamme, and H. S. Scholte, "From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category," *J. Neurosci.*, vol. 33, no. 48, pp. 18814–18824, 2013.

[31] X. Min *et al.*, "Blind quality assessment of compressed images via pseudo structural similarity," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2016, pp. 1–6.

[32] K. Gu, J. Qiao, X. Min, G. Yue, L. Weisi, and D. Thalmann, "Evaluating quality of screen content images via structural variation analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 10, pp. 2689–2701, Oct. 2018.

[33] X. Min, K. Gu, G. Zhai, M. Hu, and X. Yang, "Saliency-induced reduced-reference quality index for natural scene and screen content images," *Signal Process.*, vol. 145, pp. 127–136, Apr. 2018.

[34] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 1, pp. 6:1–6:23, 2016.

[35] X. Min *et al.*, "Visual attention analysis and prediction on human faces," *Inf. Sci.*, vol. 420, pp. 417–430, Dec. 2017.

[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[37] A. Jaiantilal. *Random Forest Implementation for MATLAB*. Accessed: Jun. 16, 2018. [Online]. Available: https://code.google.com/archive/p/randomforest-matlab/

[38] C. Ancuti, C. O. Ancuti, and C. De Vleeschouwer, "D-HAZY: A dataset to evaluate quantitatively dehazing algorithms," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2016, pp. 2226–2230.

[39] J.-P. Tarel, N. Hautière, A. Cord, D. Gruyer, and H. Halmaoui, "Improved visibility of road scene images under heterogeneous fog," in *Proc. IEEE Intell. Veh. Symp.*, 2010, pp. 478–485.

[40] 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.

[41] 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.

[42] 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.

[43] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2049–2062, Aug. 2018.

[44] 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.

[45] 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.

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

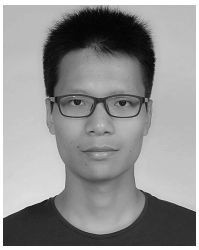
[47] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL, USA: CRC Press, 2003.

[48] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[49] 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.

[50] 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.

[51] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro- and macro-structures," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 3903–3912, May 2017.



Xiongkuo Min received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018. From 2016 to 2017, he was a Visiting Student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Post-Doctoral Fellow with Shanghai Jiao Tong University. His research interests include image quality assessment, visual attention modeling, and perceptual signal processing. He received the Best Student Paper Award from the IEEE ICME 2016.

Guangtao Zhai (M'10) received the B.E. and M.E. degrees from Shandong University, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He is currently a Research Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. His research interests include multimedia signal processing and perceptual signal processing. He received the National Excellent Ph.D. Thesis Award from the Ministry of Education of China in 2012.

Ke Gu (M'13) received the B.S. and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015, respectively. He is currently a Professor with the Beijing University of Technology, Beijing, China. His research interests include image analysis, environmental perception, quality assessment, and machine learning. He received the Best Paper Award from the IEEE TRANSACTIONS ON MULTIMEDIA, the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2016, and the Excellent Ph.D. Thesis Award from the Chinese Institute of Electronics in 2016. He was the Leading Special Session Organizer in the VCIP 2016 and the ICIP 2017, and serves as a Guest Editor for the *Digital Signal Processing Journal*. He is currently an Associate Editor of the IEEE ACCESS and the *IET Image Processing*. He is a Reviewer for 20 top SCI journals.

Xiaokang Yang (M'00–SM'04) received the B.S. degree from Xiamen University, Xiamen, China, in 1994, the M.S. degree from the Chinese Academy of Sciences, Shanghai, China, in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, in 2000.

From 2000 to 2002, he was a Research Fellow with the Centre for Signal Processing, Nanyang Technological University, Singapore. From 2002 to 2004, he was a Research Scientist with the Institute for Infocomm Research, Singapore. From 2007 to 2008, he visited the Institute for Computer Science, University of Freiburg, Freiburg im Breisgau, Germany, as an Alexander von Humboldt Research Fellow. He is currently a Distinguished Professor with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, where he is also the Deputy Director of the Institute of Image Communication and Information Processing. He has authored over 200 refereed papers and has filed 60 patents. His current research interests include image processing and communication, computer vision, and machine learning.

Dr. Yang is a member of the Asia-Pacific Signal and Information Processing Association, the VSPC Technical Committee of the IEEE Circuits and Systems Society, and the MMSP Technical Committee of the IEEE Signal Processing Society. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and a Senior Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. He was a Series Editor of the Springer CCIS and an Editorial Board Member of *Digital Signal Processing*. He is also the Chair of the Multimedia Big Data Interest Group of the MMTC Technical Committee, IEEE Communication Society.

Xinping Guan (F'18) received the Ph.D. degree in control and systems from the Harbin Institute of Technology, Harbin, China, in 1999. From 1998 to 2008, he was a Professor and the Dean of the School of Electrical Engineering, Yanshan University, China. He is currently a Chair Professor with Shanghai Jiao Tong University, Shanghai, China, where he is also the Deputy Director of the University Research Management Office and the Director of the Key Laboratory of Systems Control and Information Processing, Ministry of Education of China.

He has authored or co-authored four research monographs, over 270 papers in the IEEE TRANSACTIONS and other peer-reviewed journals, and numerous conference papers. His current research interests include industrial cyber-physical systems, wireless networking and applications in smart city and smart factory, and underwater sensor networks. As a Principal Investigator, he has finished/been working on many national key projects. He is the Leader of the prestigious Innovative Research Team, National Natural Science Foundation of China. He is an Executive Committee Member of the Chinese Automation Association Council and the Chinese Artificial Intelligence Association Council. He was a recipient of the IEEE Transactions on Fuzzy Systems Outstanding Paper Award in 2008. He received the First Prize of the Natural Science Award from the Ministry of Education of China in 2006 and 2016 and the Second Prize of the National Natural Science Award of China in 2008. He is a "National Outstanding Youth" honored by the NSF of China, a "Changjiang Scholar" by the Ministry of Education of China, and a "State-Level Scholar" of the "New Century Bai Qianwan Talent Program" of China.