Waterloo Exploration Database: New Challenges for Image Quality Assessment Models

Kede Ma, *Student Member, IEEE*, Zhengfang Duanmu, *Student Member, IEEE*, Qingbo Wu, *Member, IEEE*, Zhou Wang, *Fellow, IEEE*, Hongwei Yong, Hongliang Li, *Senior Member, IEEE*, and Lei Zhang, *Senior Member, IEEE*

***Abstract*— The great content diversity of real-world digital images poses a grand challenge to image quality assessment (IQA) models, which are traditionally designed and validated on a handful of commonly used IQA databases with very limited content variation. To test the generalization capability and to facilitate the wide usage of IQA techniques in real-world applications, we establish a large-scale database named the Waterloo Exploration Database, which in its current state contains 4744 pristine natural images and 94 880 distorted images created from them. Instead of collecting the mean opinion score for each image via subjective testing, which is extremely difficult if not impossible, we present three alternative test criteria to evaluate the performance of IQA models, namely, the pristine/distorted image discriminability test, the listwise ranking consistency test, and the pairwise preference consistency test (P-test). We compare 20 well-known IQA models using the proposed criteria, which not only provide a stronger test in a more challenging testing environment for existing models, but also demonstrate the additional benefits of using the proposed database. For example, in the P-test, even for the best performing no-reference IQA model, more than 6 million failure cases against the model are "discovered" automatically out of over 1 billion test pairs. Furthermore, we discuss how the new database may be exploited using innovative approaches in the future, to reveal the weaknesses of existing IQA models, to provide insights on how to improve the models, and to shed light on how the next-generation IQA models may be developed. The database and codes are made publicly available at: https://ece.uwaterloo.ca/~k29ma/exploration/.**

***Index Terms*— Image quality assessment, image database, discriminable image pair, listwise ranking consistency, pairwise preference consistency, mean opinion score.**

I. Introduction

IMAGE quality assessment (IQA) aims to quantify human perception of image quality, which may be degraded during acquisition, compression, storage, transmission and reproduction [1], [2]. Subjective testing is the most straightforward and reliable IQA method and has been conducted in the construction of the most widely used IQA databases (e.g., LIVE [3] and TID2013 [4]). Despite its merits, subjective testing is cumbersome, expensive and time-consuming [5]. Developing objective IQA models that can automate this process has been attracting considerable interest in both academia and industry. Objective measures can be broadly classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) approaches based on their accessibility to the pristine reference image, which is also termed as the "source image" that is assumed to have pristine quality. FR-IQA methods assume full access to the reference image [6]. RR-IQA methods utilize features extracted from the reference to help evaluate the quality of a distorted image [7]. NR-IQA methods predict image quality without accessing the reference image, making them the most challenging among the three types of approaches.

With a variety of IQA models available [8]–[16], how to fairly evaluate their relative performance becomes pivotal. The conventional approach in the literature is to compute correlations between model predictions and the "ground truth" labels, typically the mean opinion scores (MOSs) given by human subjects, of the images on a handful of commonly used IQA databases. However, collecting MOS via subjective testing is a costly process. In practice, the largest IQA database that is publicly available contains a maximum of $3,000$ subject-rated images, many of which are generated from the same source images with different distortion types and levels. As a result, only less than 30 source images are included. By contrast, the space of digital images is of very high dimension, which is equal to the number of pixels in the images, making it extremely difficult to collect sufficient subjective opinions to adequately cover the space. Perhaps more importantly, using only a few dozens of source images is very unlikely to provide a sufficient representation of the variations of real-world image content. Moreover, most objective IQA methods are developed after the commonly used IQA databases became publicly available and often involve machine learning or manual parameter tuning steps to boost their performance. All these issues cast challenges on the generalization capability of existing IQA models in real-world applications.

We believe that a large-scale database with greater content diversity is critical to fairly compare IQA models, to test their generalization capability, and to develop the next-generation IQA algorithms. This motivates us to build the Waterloo Exploration Database, or in short the Exploration

Manuscript received July 1, 2016; revised October 25, 2016; accepted November 15, 2016. Date of publication November 21, 2016; date of current version January 5, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dacheng Tao.

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

Q. Wu and H. Li are with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: wqb.uestc@gmail.com; hlli@uestc.edu.cn).

H. Yong and L. Zhang are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: cshyong@comp.polyu.edu.hk; cslzhang@comp.polyu.edu.hk).

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

Digital Object Identifier 10.1109/TIP.2016.2631888

TABLE I
COMPARISON OF IQA DATABASES

Database	# of Pristine Images	# of Distorted Images	Subjective Testing Methodology
LIVE [3]	29	779	single-stimulus continuous scale
TID2008 [8]	25	1, 700	paired comparison
TID2013 [4]	25	3, 000	paired comparison
CSIQ [9]	30	866	multi-stimulus absolute category
LIVE MD [10]	15	405	single-stimulus continuous scale
LIVE Challenge [11]	—	1, 162	single-stimulus continuous scale with crowdsourcing
Waterloo Exploration	4, 744	94, 880	need-based

database, which in its current state contains 4, 744 pristine natural images that span a variety of real-world scenarios. We extend it by adding four distortion types, namely JPEG compression, JPEG2000 compression, white Gaussian noise contamination and Gaussian blur with five distortion levels each, resulting in 99, 624 images in total. Given the large number of sample images, it is extremely difficult (if not impossible) to collect MOSs for all images in a well controlled laboratory environment. Therefore, innovative approaches are necessary to evaluate the relative performance of IQA models. Here we propose three evaluation criteria, termed as the pristine/distorted image discriminability test (D-test), the listwise ranking consistency test (L-test) and the pairwise preference consistency test (P-test), respectively. Each of them tests the robustness and generalization capability of an IQA model from a different aspect. Specifically, the D-test exams whether an IQA model well separates the pristine from distorted images. The L-test checks whether an IQA model gives consistent ranking of images with the same distortion type and content but different distortion levels. The P-test evaluates the preference concordance of an IQA measure on quality-discriminable image pairs (DIPs), which are carefully selected image pairs whose quality is clearly discriminable. By applying the three evaluation criteria to the Exploration database, we perform a systematic comparison of 20 well-known IQA models. Furthermore, we demonstrate that innovative approaches may be developed to leverage the large-scale Exploration database in order to reveal the weaknesses of even top performing IQA models, a task that is not easily achieved using existing IQA databases. Careful investigations of the failure cases of these models also provide valuable insights on potential ways to improve them.

II. RELATED WORK

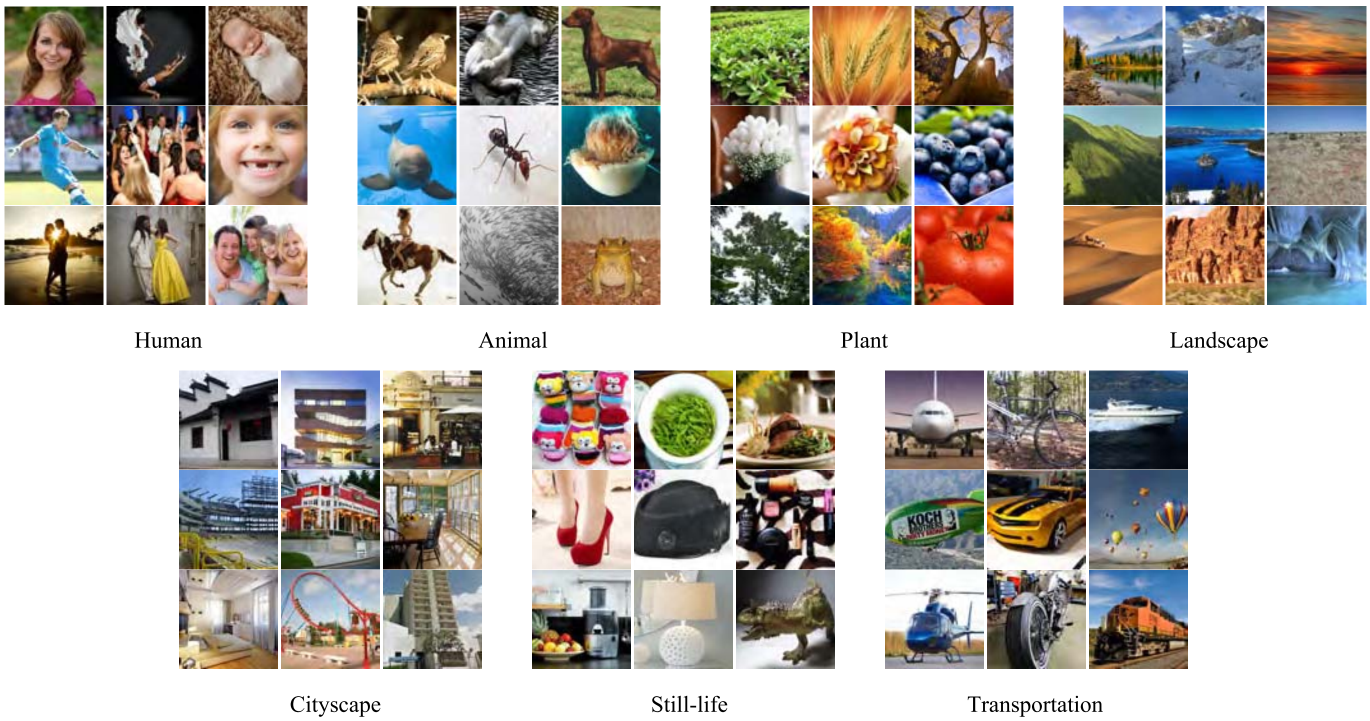
Several well-known IQA databases have been widely used in the literature. In 2005, Sheikh *et al.* [3] conducted a "large-scale" subjective image quality study and created the LIVE database that consists of 29 reference and 779 distorted images with five distortion types—JPEG2000 compression, JPEG compression, white Gaussian noise, Gaussian blur and fast fading transmission error. A single-stimulus continuous-scale method [21] is adopted for testing, where the reference images are also evaluated under the same experimental configuration [22]. MOS scaling and realignment (based on an additional double-stimulus subjective experiment) are performed to align the scores across different distortion sessions. In particular, the scaling compensates different scales used by different subjects during rating, while the realignment avoids significant bias of MOS values towards any specific distortion type and/or level.

The TID2008 [17] database contains 24 pristine natural and 1 computer generated images. 18 of them are originated from LIVE [3], differing only in size via cropping. Seventeen types of distortions with four distortion levels are added, resulting in a total of 1, 700 distorted images. The testing methodology is a paired comparison method [23], where the reference image is also shown to the subjects. A Swiss competition principle is used to reduce the number of pairs for subjective testing such that each image appears in at most nine pairs. No explicit MOS scaling and realignment are reported to refine the raw MOSs collected from multiple sessions in three countries. TID2008 was later extended to TID2013 [4] by adding seven new distortion types and one additional distortion level, making it the largest public database so far.

The CSIQ [18] database contains 30 reference images and 866 distorted images by adding six distortion types with four to five distortion levels. CSIQ uses a multi-stimulus absolute category method based on a linear displacement of the images of the same content across four calibrated LCD monitors placed side by side with equal viewing distance to the observer. MOSs of images with different content are realigned according to a separate, but identical, experiment in which observers place subsets of all the images linearly in space.

The LIVE multiply distorted (MD) database [19] and the LIVE in the wild image quality challenge database [20] (LIVE Challenge) focus on images with mixture of distortions. LIVE MD simulates two multiple distortion scenarios, one for image storage (Gaussian blur followed by JPEG compression) and the other for digital image acquisition (Gaussian blur followed by white Gaussian noise). It contains 15 pristine images and 405 distorted ones. The test methodology is the same as is used in LIVE [3]. LIVE Challenge database takes a step further and directly works with authentically distorted images captured from mobile devices. A total of 1, 162 images are included, whose MOSs are crowdsourced using the Amazon Mechanical Turk platform. Substantial efforts have been put to process the noisy raw data and to verify the reliability of the obtained human opinions from the uncontrolled test environment. A summary of the aforementioned databases are given in Table I.

Other widely known but smaller databases include IVC [24], Toyama-MICT [25], Cornell A57 [26] and WIQ [27], etc.



Human

Animal

Plant

Landscape

Cityscape

Still-life

Transportation

Fig. 1. Sample source images in the Waterloo Exploration Database.

A useful collection of IQA databases can be found at [28], [29].

A major common issue of all the existing IQA databases is the limited numbers of source images being used (as a matter of fact, none of the databases includes more than 30 source images), which creates a large gap between the diversity of real-world images and the variation of the image content that can be tested using the databases. As a result, IQA models developed and validated using such databases are inevitably questioned on their generalization capability to real-world applications. This is evidenced by the recent test results on the LIVE Challenge database, a collection of images from the real-world, where the performance of the most advanced NR-IQA models drops significantly [20]. The limitation on the number of source images is largely due to the limited capacity of the affordable subjective testing experiments. For example, testing and comparing the $1,700$ distorted images in TID2008 [17] is an expensive and highly time-consuming "large-scale" subjective testing task, but given the combinations of the distortion types and levels that are applied to each source image, eventually, only 25 source images can be included.

The above issue motivates us to build a new database for IQA research, which aims to significantly expand the diversity of image content. Meanwhile, testing all images in the database using conventional subjective testing methodologies becomes extremely difficult, if not impossible. Therefore, innovative approaches on how to use the database to test and compare IQA models have to be developed in order to meet the challenge. These are the key questions we would like to answer in this work.

III. Constructing the Waterloo Exploration Database

We construct a new image database, namely the Waterloo Exploration Database, which currently contains $4,744$ pristine natural images that span a great diversity of image content. An important consideration in selecting the images is that they need to be representative of the images we see in our daily life. Therefore, we resort to the Internet and elaborately select 196 keywords to search for images. The keywords can be broadly classified into 7 categories: human, animal, plant, landscape, cityscape, still-life and transportation. We initially obtain more than $200,000$ images. Many of these images contain significant distortions or inappropriate content, and thus a sophisticated manual process is applied to refine the selection. In particular, we first remove those images that have obvious distortions, including heavy compression artifacts, strong motion blur or out of focus blur, low contrast, underexposure or over-exposure, substantial sensor noise, visible watermarks, artificial image borders, and other distortions due to improper operations during acquisition. Next, images of too small or too large sizes, cartoon and computer generated content, and inappropriate content are excluded. After this step, about $7,000$ images remain. To make sure that the images are of pristine quality, we further carefully inspect each of the remaining images multiple times by zooming in and remove those images with visible compression distortions. Eventually, we end up with $4,744$ high quality natural images. Sample images grouped into different categories are shown in Fig. 1.

Four distortion types with five levels each are chosen to alter the source images. All distorted images are generated using MATLAB functions as follows:

- JPEG compression: The quality factor that parameterizes the DCT quantization matrix is set to be $[43, 12, 7, 4, 0]$ for five levels, respectively.
- JPEG2000 compression: The compression ratio is set to be $[52, 150, 343, 600, 1200]$ for five levels, respectively.
- Gaussian blur: 2D circularly symmetric Gaussian blur kernels with standard deviations (std) of $[1.2, 2.5, 6.5, 15.2, 33.2]$ for five levels are used to blur the source images.
- White Gaussian noise: white Gaussian noise is added to the source images, where variances are set to be $[0.001, 0.006, 0.022, 0.088, 1.000]$ for five levels, respectively.

The above four distortion types are the most common ones in existing IQA databases [22], [30], and many IQA models are claimed to excel at handling these distortions [12]–[15], [31]–[39]. Therefore, whether these models perform well on the new Waterloo Exploration Database becomes a strong test on the claims of these methods and their generalization capability in the real-world. The parameters that control the distortion levels for each distortion type are chosen so as to cover the full range of subjective quality scale, which is measured by VIF [9] calibrated on the LIVE database with a nonlinear mapping. Specifically, we select the distortion parameters for each distortion type separately so that the distorted images are roughly evenly distributed in the score range. As a result, the discriminability between two adjacent levels can be guaranteed. Once determined, the parameters are fixed for all images. Overall, the Exploration database contains a total of $99,624$ images. The numbers of pristine and distorted images are 150 times and 30 times, respectively, more than those of the largest existing databases so far.

IV. Evaluating Objective IQA Models

To make use of the Exploration database for comparing the relative performance of IQA models, we present three test criteria, namely the pristine/distorted image discriminability test (D-test), the listwise ranking consistency test (L-test), and the pairwise preference consistency test (P-test).

A. Pristine/Distorted Image Discriminability Test (D-Test)

Considering the pristine and distorted images as two distinct classes in a meaningful perceptual space, the D-test aims to test how well an IQA model is able to sperate the two classes. An illustration using the Exploration database is shown in Fig. 2, where an IQA model with strong discriminability (e.g. Wang05 [40]) is able to map pristine and distorted images onto easily separable intervals with minimal overlaps, whereas a less competitive model creates two distributions of scores with large overlaps. Here we introduce a measure to quantify this discriminability. Let q_i represent the predicted quality of the i-th image by a model, we group indices of pristine and distorted images into the sets of S_p and S_d, respectively. We then apply a threshold T on $\{q_i\}$ to classify the images such that $S'_p = \{i | q_i > T\}$ and $S'_d = \{i | q_i \leq T\}$. The average correct classification rate is given by

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

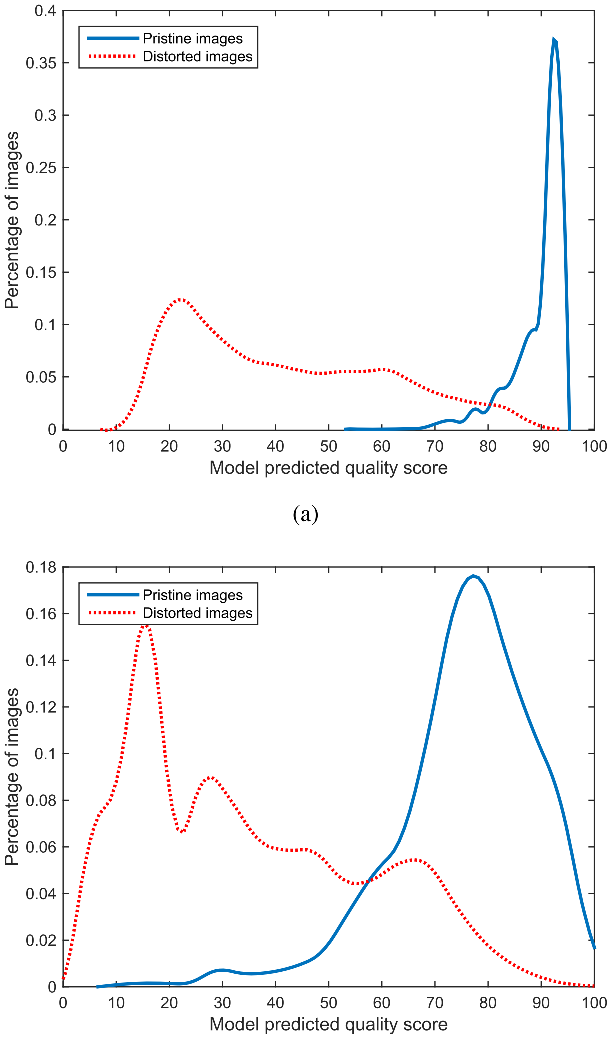


Fig. 2. Distributions of IQA model prediction scores of pristine and distorted images of the Waterloo Exploration Database. Ideal IQA models are expected to have strong discriminability of the distributions, and are expected to create small overlaps between the two distributions. (a) WANG05 [40] model; (b) DIIVINE [13] model. (a) WANG05 [40]. (b) DIIVINE [13].

It is worth noting that most existing IQA databases, including the Waterloo Exploration Database are class-imbalanced, where the collection of samples is overwhelmed by the distorted images. By normalizing the correctly classified samples, we avoid the trivial solution that all images are classified as distorted, which could also result in a not bad R. The value of T should be optimized to yield the maximum correct classification rate. Thus, we define a discriminability index as

$$D = \max_{T} R(T). \quad (2)$$

D lies in $[0, 1]$, with a larger value indicating a better separability between pristine and distorted images. The single-variable optimization problem can be solved using a line search method.

B. Listwise Ranking Consistency Test (L-Test)

The idea behind the L-test has been advocated by Xue *et al.* [29], [41]. The goal is to evaluate the robustness of

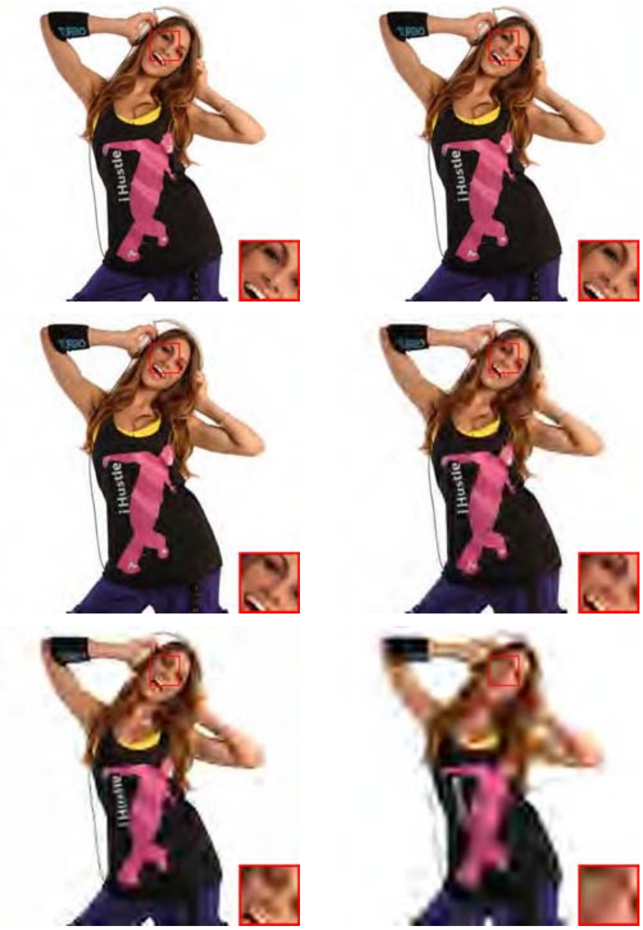


Fig. 3. L-test of "Hip-hop Girl" images under JPEG2000 compression. The image quality degrades with the distortion level from left to right and from top to bottom. A competitive IQA model (e.g., ILNIQE [14]) rank-lists the images in exactly the same order. By contrast, a less competitive model (e.g., QAC [35]) may give different rankings.

IQA models when rating images with the same content and the same distortion type but different distortion levels. The underlying assumption is that the quality of an image degrades monotonically with the increase of the distortion level for any distortion type. Therefore, a good IQA model should rank these images in the same order. An example on the Exploration database is given in Fig. 3, where different models may or may not produce the same quality rankings in consistency with the image distortion levels. Given a database with N pristine images, K distortion types and L distortion levels, we use the average Spearman's rank-order correlation coefficient (SRCC) and Kendall's rank-order correlation coefficient (KRCC) to quantify the ranking consistency between the distortion levels and the model predictions, which are defined as

$$L_s = \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathrm{SRCC}(l_{ij}, q_{ij}), \tag{3}$$

and

$$L_k = \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathrm{KRCC}(l_{ij}, q_{ij}), \tag{4}$$

where l_{ij} and q_{ij} are both length-L vectors representing the distortion levels and the corresponding distortion/quality scores given by a model to the set of images that are from the same (i-th) source image and have the same (j-th) distortion type.

C. Pairwise Preference Consistency Test (P-Test)

The P-test compares preference predictions of IQA models on pairs of images whose quality is clearly discriminable. We call such pairs of images quality-discriminable image pairs (DIPs). A good IQA model should consistently predict preferences concordant with the DIPs. Paired comparison is a widely used subjective testing methodology in IQA research, as discussed in Section II. Pairwise preference has also been exploited previously to learn rank-IQA models [16], [42]. Nevertheless, in all previous work, the DIPs that can be used for testing or developing objective models are obtained exclusively from subjective quality ratings, which largely limits the number of available DIPs, and is impractical for large-scale image databases such as the Exploration database.

Here, we propose a novel automatic DIP generation engine by leveraging the quality prediction power of several most-trusted FR-IQA measures in the literature. Specifically, we consider an image pair to be a valid DIP if the absolute differences of the predicted scores from the FR models are all larger than a pre-defined threshold, T.

To explore this idea, we first experiment with the LIVE database [3], from which we extract all possible image pairs whose absolute MOS differences are larger than $T_l = 20$ and consider them as the "ground truths" DIPs. The legitimacy of $T_l = 20$ on LIVE can be validated from two sources. First, the average std of MOSs on LIVE is around 9 and $T_l = 20$ is right outside the ± 1 std range, which guarantees the perceptual quality discriminability of the pair of images. Second, from the subjective experiment conducted by Horita *et al.* [16], it is observed that the consistency between subjects on the relative quality of one pair from LIVE increases with T_l, and when T_l is larger than 20, the consistency approaches 100%. Using the available MOS values in LIVE [3], we are able to generate 206, 717 "ground truth" DIPs, termed as the "ground truth" set. After that, we use our DIP generation engine to generate DIPs on LIVE and observe whether the generated pairs are in the "ground truth" set. Fig. 4 shows the percentage p of generated DIPs in the "ground truth" DIP set as a function of T for different combinations of FR-IQA measures, where three base FR-IQA measures, namely MS-SSIM [43], VIF [9] and GMSD [10] are selected. It can be seen that p increases when more FR-IQA models are involved, and is maximized when all the 3 FR-IQA models are used. Using all the three models together with $T = 40$, we achieve $p = 99.81\%$ accuracy, which verifies the reliability of the DIP generation engine. This configuration is used as the default setting. Note that the model predictions of the three FR-IQA models should be mapped to the same perceptual scale before DIP generation. Fig. 5 shows 3 DIPs generated by the proposed engine on the Exploration database. One can see that the left images of the 3 DIPs have superior perceived quality compared to the right ones.

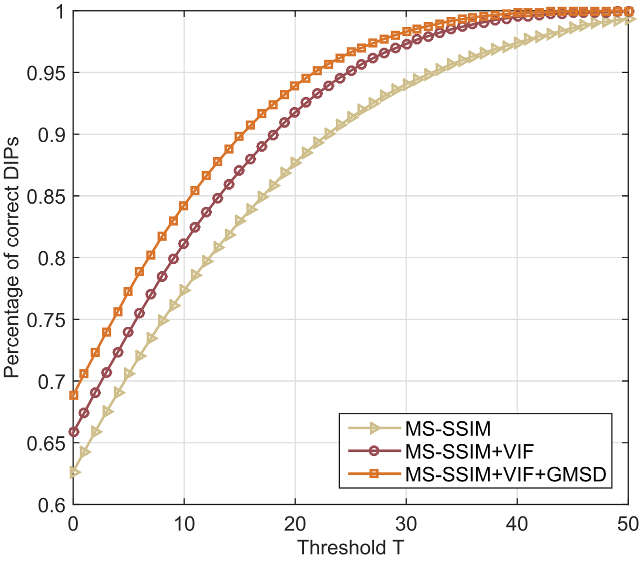


Fig. 4. The percentage of generated DIPs in the "ground truth" set on the LIVE [3] database as a function of T for different combinations of base FR-IQA models.

Given an image database $\mathbf{D}$, the DIP generation goes through all possible pairs of images to create the full DIP set from $\mathbf{D}$. Suppose that the total number of DIPs in the set is M and the number of concordant pairs of an IQA model (meaning that the model predicts the correct preference) is M_c, a pairwise preference consistency ratio is defined as

$$P = \frac{M_c}{M}. \tag{5}$$

P lies in [0, 1] with a higher value indicating better performance of the IQA model being tested.

D. Discussion

The above test criteria are defined independent of any particular database, regardless of their size or content. Each of them challenges an IQA model from a different perspective. One would not be surprised to see that one model is superb under one criterion but subpar under another (as we will see in Section V). Meanwhile, all of them benefit from larger databases, where the weaknesses and failure cases of the test models have more chances to be detected. These failure cases may provide insights on how to improve IQA models.

V. Experimental Results

We apply the aforementioned test criteria on the Waterloo Exploration Database and compare the performance of 20 well-known IQA models, which are selected to cover a wide variety of design methodologies with an emphasis on NR-IQA methods. The models include FR-IQA measures 1) PSNR, 2) SSIM [8], 3) MS-SSIM [43], 4) FSIM [44], 5)VIF [9], 6) GMSD [10], RR-IQA measures 7) WANG05 [40], 8) RRED [45], and NR-IQA methods 9) BIQI [31], 10) BLINDS_II [32], 11) BRISQUE [33], 12) CORNIA [12], 13) DIIVINE [13], 14) IL-NIQE [14], 15) LPSI [38], 16) M3 [36], 17) NFERM [39], 18) NIQE [34], 19) QAC [35] and 20) TCLT [15]. The implementations of all algorithms are obtained from the original authors or their public websites. For training based IQA methods, we use the whole LIVE database [3] to learn the models. Furthermore, we adopt a 4-parameter logistic nonlinear function as suggested in [21] to map the predicted scores of candidate models to the MOS scale of LIVE [22]. The nonlinear mapping compensates the nonlinearity of model predictions on the human perception of image quality and make the results more interpretable. As a result, the score range of all algorithms spans between [0, 100], where a higher value indicates better perceptual quality.

A. D-Test

Fig. 7 shows the D-test results on the Exploration database of 20 IQA measures. It can be observed that FR-IQA and RR-IQA models perform the best and often give nearly perfect performance. This is not surprising because they have full or partial access to the pristine images. Second, TCLT [15], CORNIA [12], QAC [35] and BRISQUE [33] are among the top performing NR-IQA models. Despite their superior performance, by looking into their common failure cases, we are able to identify their weaknesses. Some examples are shown in Fig. 6. In general, the pristine images that are misclassified as distorted ones often exhibit low illumination or low intensity variations. There are also exceptions. For example, complex textures as those in Fig. 6(c) resemble noise structures and may fool NR-IQA models. On the other hand, the distorted images that are misclassified as pristine ones are often induced by white Gaussian noise and JPEG compression at mild distortion levels. Since slightly distorted images may not be visually differentiable from pristine images, we are not expecting an ideal NR-IQA model to have a perfect or nearly perfect D value on the Exploration database.

We also run the D-test on LIVE [3] which has less than $1,000$ test images. The top performing NR-IQA models TCLT [15] and CORNIA [12] on the Exploration database perform perfectly on LIVE (achieving $D = 1$), which means that no failure case can be found. This manifests the benefits of using the Exploration database which contains substantially more images to better distinguish between IQA models by more easily identifying their failure cases.

B. L-Test

We perform the L-test on the Exploration database that includes $4,744 \times 4 = 18,976$ sets of images, each of which contains a list of images generated from the same source with the same distortion type but at different distortion levels. Fig. 8 shows the L_s and L_k results of 20 IQA models, from which we have several observations. First, it is not surprising that FR- and RR-IQA models generally perform better than NR-IQA approaches because they are fidelity measures that quantify how far a distorted image departs from the source image, and such fidelity typically decreases monotonically with the increase of distortion levels. Second, the NR model NIQE [34] and its feature enriched extension ILNIQE [14] outperform all other NR-IQA models. It is worth mentioning

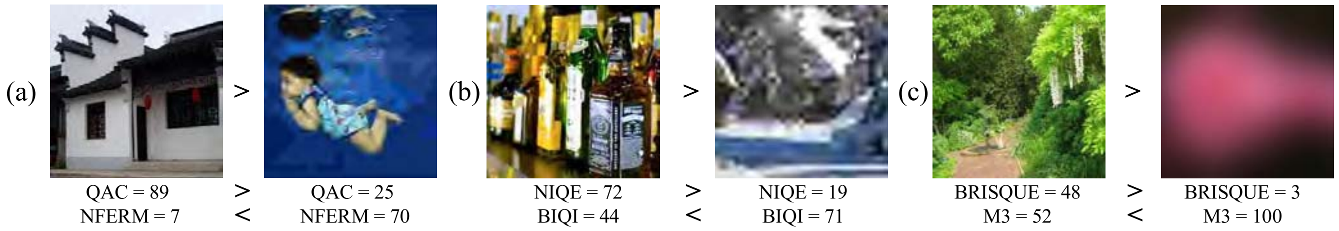


Fig. 5. Sample DIPs from the Exploration database. (a), (b) and (c) show 3 DIPs, where the left images have clearly better quality than the right images. A good model is able to give concordant opinions, whereas a less competitive model tends to perform randomly or provide discordant opinions.

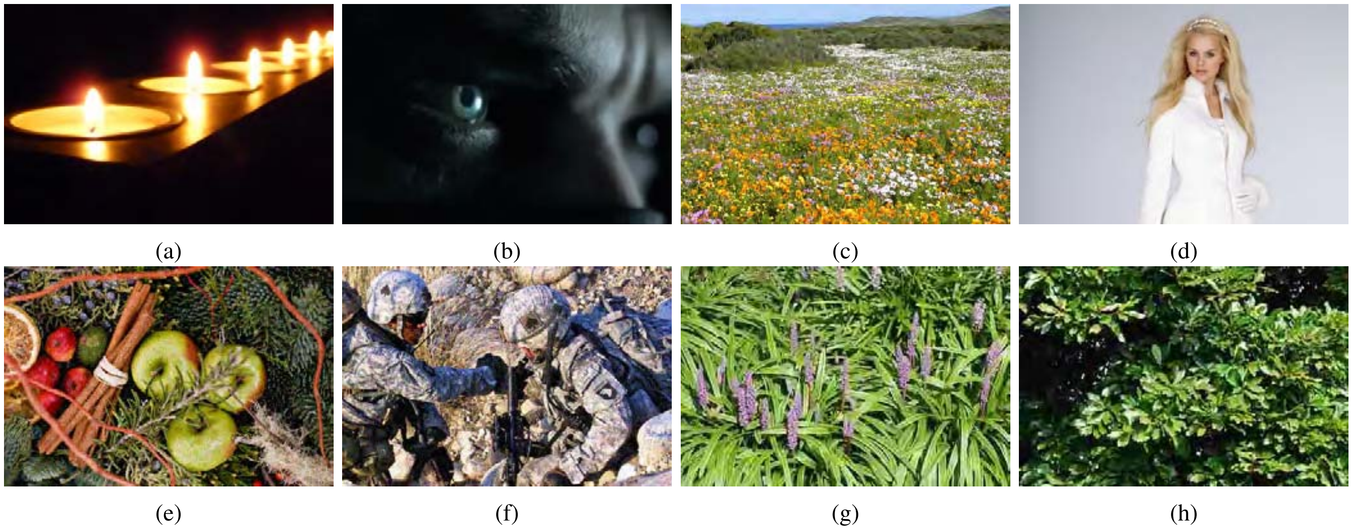


Fig. 6. Failure cases of the top four NR-IQA models (TCLT [15], CORNIA [12], QAC [35] and BRISQUE [33]) in the D-test on the Exploration database. (a)-(d): pristine images misclassified as distorted ones by the four models; (e)-(h): distorted images misclassified as pristine ones by the four models.

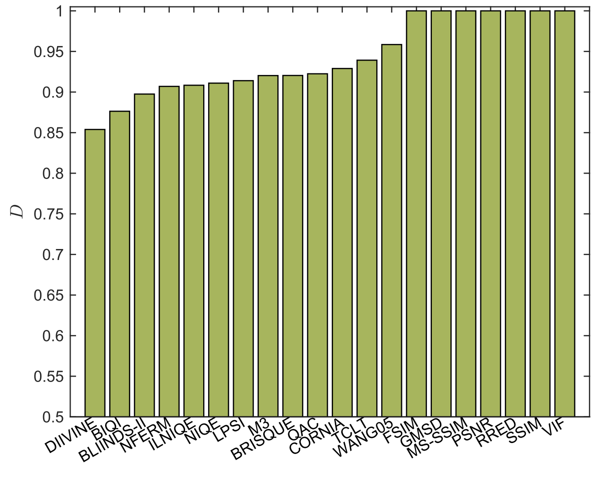


Fig. 7. D-test results of IQA models on the Exploration database.

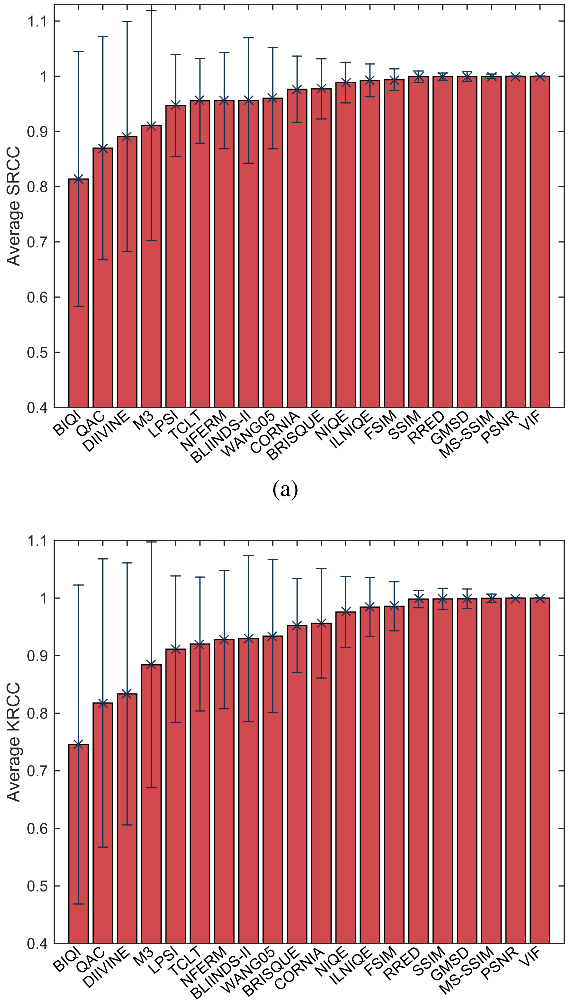
that NIQE and ILNIQE are based on perception-and distortion-relevant natural scene statistics (NSS) features, without MOS for training. This reveals the power of NSS, which map images into a perceptually meaningful space for comparison. Third, although TCLT [15] performs the best in the D-test, it is not outstanding in the L-test. Fourth, training based models, such as BIQI [31] and DIIVINE [13] generally have lower overall consistency values and larger error bars (stds), implying potential overfitting problems.

Furthermore, to demonstrate the additional benefits of the L-test, we focus on NIQE [34], one of the best performing models, observing its main failure cases and discussing how it can be improved. Fig. 9 shows sample failure cases which occur when JPEG2000 compression is applied. A common characteristic of these images is that they are a combination of strong edges and large smooth regions, which results in abundant ringing artifacts after JPEG2000 compression. The patch selection mechanism in NIQE [34] may mistakenly group such distorted structures to build the multi-variant Gaussian model (MVG), which can be close to the MVG computed from a number of natural image patches. This results in a reverse order of quality ranking. Potential ways of improving NIQE [34] include pre-screening ringing artifacts and training the MVG using natural image patches of more diverse content.

To investigate the impact of the size of the image databases on the L-test, we also run it on the LIVE [3] database. The average L_s value over 20 IQA models is 0.964 with an std of 0.025, which is only half of the std obtained using the Exploration database. This indicates that running L-test on larger databases is desirable to better differentiate IQA models.

C. P-Test

We apply the proposed DIP generation engine on the Exploration database, resulting in more than 1 billion DIPs. Fig. 10 shows the pairwise preference consistency ratios



(a)

(b)

Fig. 8. L-test results of IQA models on the Exploration database. (a) L_s results. (b) L_k results.

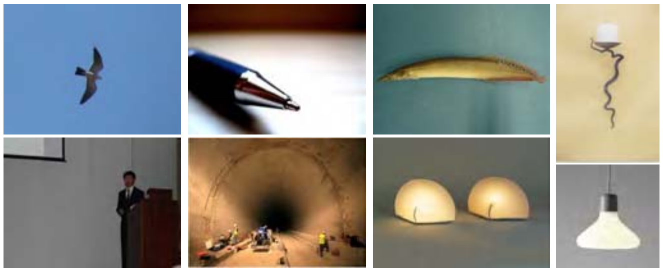


Fig. 9. Failure cases of NIQE [34] in the L-test induced by JPEG2000 compression on the Exploration database, where L_k is less than 0.5.

of 12 NR-IQA, 2 RR-IQA and 3 FR-IQA measures. MS-SSIM [43], VIF [9] and GMSD [10] are not tested here because they are used in the DIP generation process and thus are not independent of the test. Several useful observations can be made. First, all algorithms under test achieve $P \geq 90\%$, which verifies the success of these algorithms in predicting image quality to a certain extent. Second, as one of the first attempts towards RR-IQA, WANG05 [40], a top performer in the D-test, does not perform very well in the P-test compared to many NR-IQA methods. This may be because the statistical features on marginal wavelet coefficients are insufficient to fully capture the variations in image content and distortion. The performance may be further compromised due to quantization of extracted features. Third, ILNIQE [14], NIQE [33] and CORNIA [12] are among top performing NR models, which conforms to the results in the L-test.

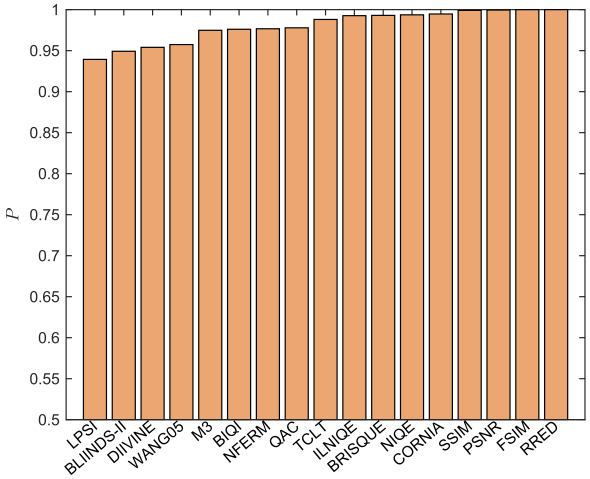


Fig. 10. P-test results of IQA models on the Exploration database.

Note that the size of the Exploration database is fairly large and therefore a small difference of the P-test may indicate significant space for improvement. For example, although CORNIA [12] outperforms all the other NR-IQA methods and achieve $P = 0.995$, it still makes $6,808,400$ wrong predictions. Representative failure cases are shown in Fig. 11. Careful investigations show its weaknesses and provide potential ways to improve it. Specifically, CORNIA tends to favor artificial structures introduced in smooth regions, for example blocking structures in the sky in Fig. 11(a1), and ringing around sharp edges in Fig. 11(c1). This may be a consequence of its unsupervised feature learning mechanism that may not be capable of reliably differentiating real structures from artificially created distortions in smooth areas.

We run the P-test on LIVE [3] for comparison. Only $90,870$ DIPs can be generated, which is less than 0.01% of the DIPs generated from the Exploration database. All 14 algorithms perform perfectly on LIVE, achieving $P = 1$. No failure case is found of any IQA model. This result manifests the value of the Exploration database, and meanwhile shows the capability of the P-test at exploiting large databases.

We also conduct experiments using the P-test on the LIVE Challenge database [20]. We generate DIPs based on the MOS provided by the database. Specifically, we consider an image pair to be a valid DIP if their absolute MOS difference is larger than one std of the MOS.[1] As such, a total number of $330,752$ DIPs are generated. Note that the reference images are not available in the Challenge database and therefore only NR-IQA models are tested. Fig. 12 shows the P values of 12 NR-IQA models. It can be observed that the top performing

[1] Every image in the LIVE Challenge database has a MOS and an std associated with it, computed from all valid subjective scores. Here we use the average std of all images.

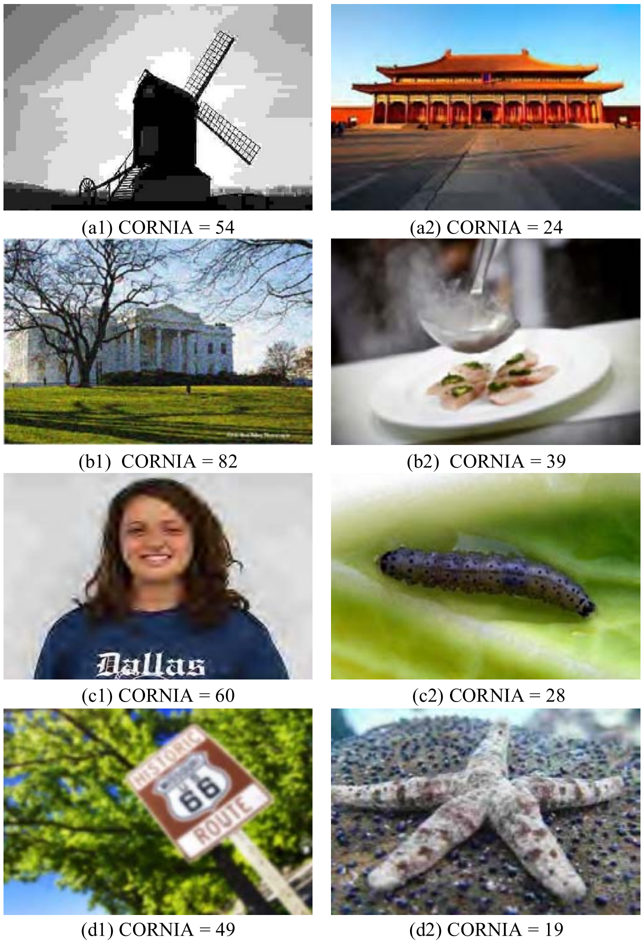

(a1) CORNIA = 54
(a2) CORNIA = 24
(b1) CORNIA = 82
(b2) CORNIA = 39
(c1) CORNIA = 60
(c2) CORNIA = 28
(d1) CORNIA = 49
(d2) CORNIA = 19

Fig. 11. Failure cases of CORNIA [12] in the P-test on the Exploration database. The left images have inferior quality compared with the right ones, but CORNIA [12] gives incorrect preference predictions. (a1) CORNIA = 54. (a2) CORNIA = 24. (b1) CORNIA = 82. (b2) CORNIA = 39. (c1) CORNIA = 60. (d1) CORNIA = 49. (c2) CORNIA = 28. (d2) CORNIA = 19.

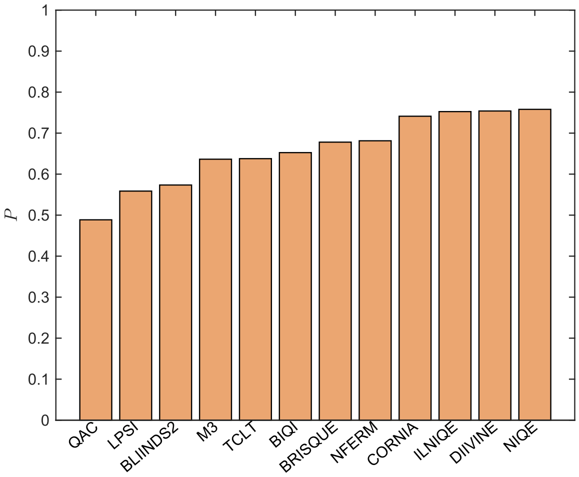


Fig. 12. The P-test results of 12 NR-IQA models on the LIVE Challenge database [20].

NR-IQA models CORNIA [12], NIQE [34], ILNIQE [14] on the Waterloo Exploration Database are also ranked high on the LIVE Challenge database. However, DIIVINE [13], a less competitive model on the Exploration database, performs the second best on the Challenge database. One possible reason might be the relatively large noise levels of the MOS in the LIVE Challenge database, whose samples were collected via crowdsourcing from an uncontrolled environment. The differences in the ranks of the IQA models may also result from a combination of the nature of the image distortions in different databases and the properties of the features employed in different models. Further investigations are needed to better explain the observations.

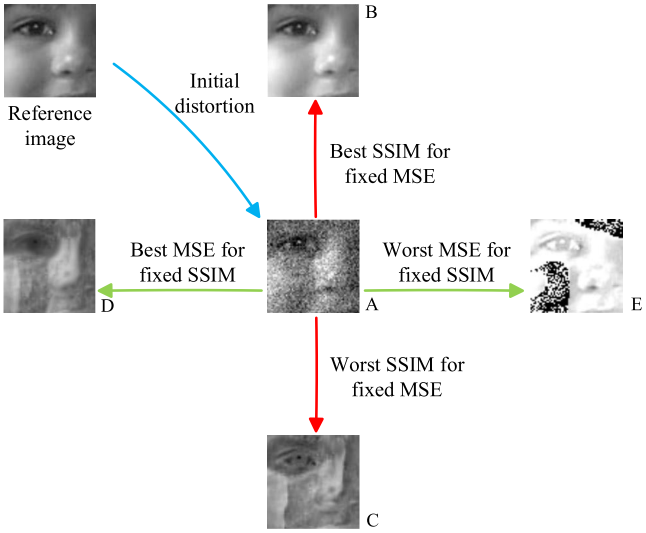


Fig. 13. Illustration of the MAD competition method [46] with synthesized image pairs.

VI. Discussions

The D-test, L-test and P-test presented in this paper are by no means the only ways we could use the Exploration database to test, compare and improve existing IQA models. The rich diversity of the database allows for many innovative and advanced approaches for testing and new model development.

A concept that is worth deeper investigation is the MAximum Differentiation (MAD) competition method, introduced by Wang and Simoncelli [46]. The fundamental idea behind MAD, which is substantially different from standard approaches of model evaluation, is to disprove a model by visually inspecting automatically generated "counter-examples", instead of trying to prove a model using pre-selected and subject-rated stimuli. This could largely reduce the required number of samples for subjective testing because conceptually even one "counter-example" is sufficient to disprove a model. In the context of IQA, image pairs are automatically synthesized to optimally distinguish two IQA models in comparison. An illustration is shown in Fig. 13, where we first synthesize a pair of images that maximize/minimize SSIM [8] while holding MSE fixed. We then repeat this procedure, but with the roles of SSIM [8] and MSE exchanged. An implementation issue that impedes the wide applicability of MAD competition is that the image synthesis process relies on gradient computations to perform an iterative constrained optimization process, which is not plausible for many IQA models whose gradients are difficult to compute, if not

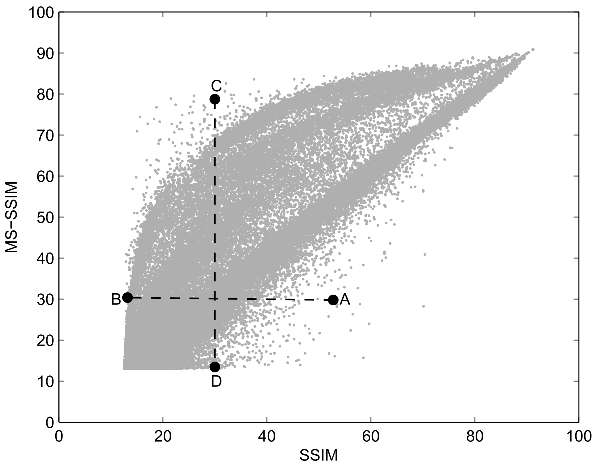
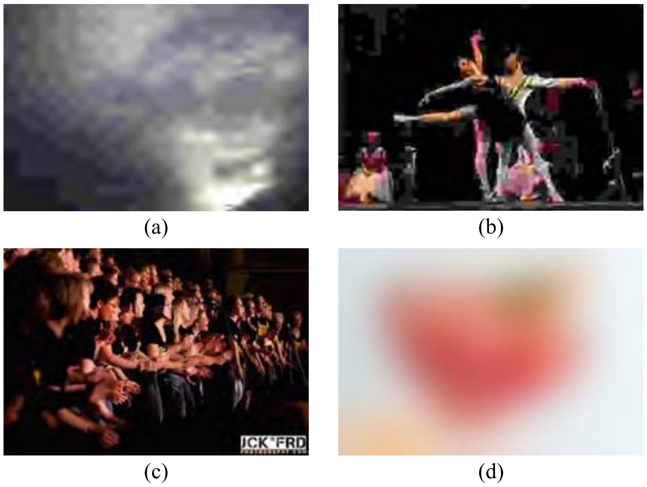


Fig. 14. MAD-motivated image pairs selection using the Exploration database. A pair of images (A, B) is selected by maximizing/minimizing SSIM but holding MS-SSIM fixed. Similarly, a pair of images (C, D) is selected by maximizing/minimizing MS-SSIM but holding SSIM fixed.



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

Fig. 15. Image pairs found by MAD competition between SSIM [8] and MS-SSIM [43] on the Exploration database. (a) MS-SSIM = 3. SSIM = 53. (b) MS-SSIM = 30. SSIM = 13. (c) SSIM = 30. MS-SSIM = 78. (d) SSIM = 30. MS-SSIM = 13.

impossible. The rich diversity of the Exploration database allows us to bypass this difficult step by replacing the image synthesis process with a search step for pairs of images with one model fixed but the other maximally differentiated. This corresponds to finding image pairs on the scatter plots of two models that have the longest distance in a given row or column (where we assume that the quality predictions of two models are mapped to the same perceptual scale), as exemplified in Fig. 14, where SSIM competes with MS-SSIM. The corresponding image pairs are shown in Fig. 15, from which we can see that images in the first row exhibits approximately the same perceptual quality (in agreement with MS-SSIM [8]) and those in the second row have drastically different perceptual quality (in disagreement with SSIM [43]). This suggests that MS-SSIM may be a significant improvement over SSIM.

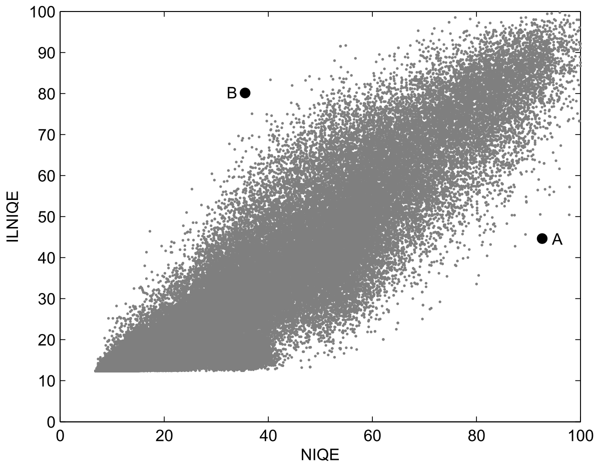
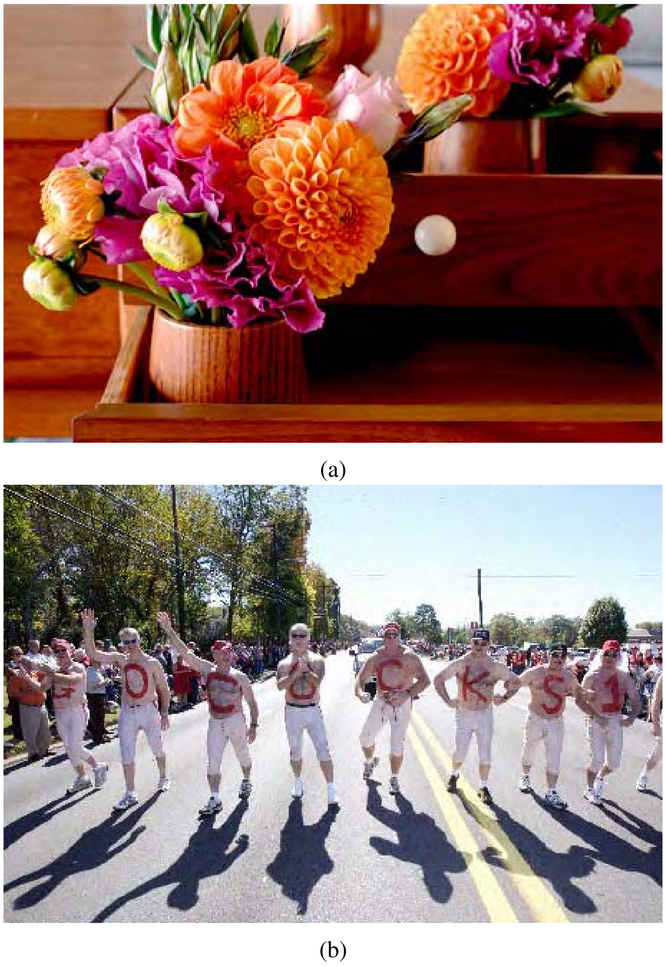


Fig. 16. Selection of a pair of images that two models have the strongest opposite opinions. (A, B) corresponds to the images for which the quality predictions by NIQE [34] and ILNIQE [14] are maximized/minimized.



(a)

(b)

Fig. 17. The pair of images in the Exploration database for which NIQE [34] and ILNIQE [14] have the strongest opposite opinions. (a) NIQE = 93. ILNIQE = 45. (b) NIQE = 36. ILNIQE = 81.

Inspired by the spirit of MAD, we may explore the idea even further by looking for image pairs that two models have exactly opposite opinions. An extreme case is to find the outmost outlier image pair in the scatter plot of two models, as exemplified in Fig. 16, where we pick two images corresponding to $\max_i (q_i - q_i')$ and $\min_j (q_j - q_j')$, respectively. Using this strategy, we find the outmost outlier image pair of NIQE [34] and ILNIQE [14] on the Exploration database, as shown in Fig. 17. Surprisingly, although ILNIQE [14] is claimed to improve upon NIQE [34], NIQE [34] is in closer agreement with human perception in this test. This suggests that the evolvement from NIQE [34] to ILNIQE [14] may have lost certain merits originally in NIQE [34].

VII. Conclusion and Future Work

We introduced the Waterloo Exploration Database, currently the largest database for IQA research. We presented three evaluation criteria, the D-test, L-test and P-test, and applied them to the Exploration database to assess 20 well-known IQA models, resulting in many useful findings. In addition, innovative approaches for comparing IQA models were also discussed. Both the Exploration database and the proposed testing tools are made publicly available to facilitate future IQA research.

The current work can be extended in many ways. First, other existing and future IQA models may be tested and compared by making use of the database. Second, the database is readily extended by adding more pristine images, more distortion types and/or more distortion levels. Third, the failure cases discovered in the database using the proposed testing methodologies may be exploited to improve existing IQA models or to combine the merits of multiple models. Fourth, new machine learning based approaches may be developed using the database, aiming for IQA models with stronger generalization capability.

Acknowledgment

The authors would like to thank Dr. Ke Gu for providing the codes of [39], Dr. Tongliang Liu, Wentao Liu and the anonymous reviewers for insightful comments.

References

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. San Rafael, CA, USA: Morgan Claypool Publishers, 2006.

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

[3] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack. *Image and Video Quality Assessment Research at LIVE*, accessed on Jan. 2016. [Online]. Available: http://live.ece.utexas.edu/research/quality/

[4] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015. [Online]. Available: http://ponomarenko.info/tid2013.htm

[5] T. Hoßfeld *et al.*, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.

[6] S. J. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," *Proc. SPIE*, vol. 1666, pp. 2–15, Aug. 1992.

[7] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.

[8] 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.

[9] 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.

[10] 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.

[11] F. Gao and J. Yu, "Biologically inspired image quality assessment," *Signal Process.*, vol. 124, pp. 210–219, Jul. 2016.

[12] 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.

[13] 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.

[14] 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.

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

[16] F. Gao, D. Tao, X. Gao, and X. Li, "Learning to rank for blind image quality assessment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2275–2290, Oct. 2015.

[17] N. Ponomarenko and K. Egiazarian. (2008). *Tampere Image Database TID2008*. [Online]. Available: http://www.ponomarenko.info/tid2008

[18] 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, p. 011006, 2010.

[19] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. IEEE 46th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2012, pp. 1693–1697.

[20] 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.

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

[22] 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.

[23] H. A. David, "The method of paired comparisons," DTIC Document, Tech. Rep., 1963, vol. 12.

[24] P. L. Callet and F. Autrusseau. (2005). *Subjective Quality Assessment IRCCyN/IVC Database*. [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/

[25] Y. Horita, K. Shibata, Y. Kawayoke, and Z. M. Parvez. (2010). *Toyama-MICT Image Quality Evaluation Database*. [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb

[26] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[27] U. Engelke, M. Kusuma, H.-J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Process., Image Commun.*, vol. 24, no. 7, pp. 525–547, 2009.

[28] S. Winkler. (2016.) *Image and Video Quality Resources*. [Online]. Available: http://stefan.winkler.net/resources.html/

[29] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[30] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

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

[32] 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.

[33] 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.

[34] 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.

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

[36] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[37] A. Saha and Q. M. J. Wu, "Utilizing image scales towards totally training free blind image quality assessment," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1879–1892, Jun. 2015.

[38] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 339–343.

[39] 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.

[40] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Proc. SPIE*, vol. 5666, pp. 149–159, Mar. 2005.

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

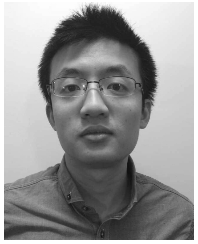
[42] L. Xu, W. Lin, J. Li, X. Wang, Y. Yan, and Y. Fang, "Rank learning on training set selection and image quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.

[43] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2. Nov. 2003, pp. 1398–1402.

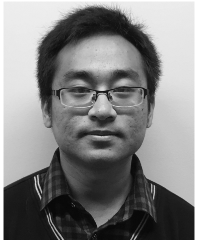
[44] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[45] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, Feb. 2012.

[46] Z. Wang and E. P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *J. Vis.*, vol. 8, no. 12, p. 8, 2008.



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



Zhengfang Duanmu (S'15) received the B.A.Sc. degree in electrical and computer engineering from the University of Waterloo in 2015, where he is currently pursuing the M.A.Sc. degree in electrical and computer engineering. His research interests include perceptual image processing and quality of experience.

Qingbo Wu (S'12–M'13) received the B.E. degree from the Education of Applied Electronic Technology, Hebei Normal University, in 2009, and the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China in 2015. Since 2014, he has been a Research Assistant with the Image and Video Processing Laboratory, The Chinese University of Hong Kong. From 2014 to 2015, he was a Visiting Scholar with the Image & Vision Computing Laboratory, University of Waterloo. He is currently a Lecturer with the School of Electronic Engineering, University of Electronic Science and Technology of China. His research interests include image/video coding, quality evaluation, and perceptual modeling and processing.

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

Dr. Wang was a Senior Area Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (2015-present), and an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2016-present). He was a member of the IEEE Multimedia Signal Processing Technical Committee (2013–2015), an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (2009–2014), the *Pattern Recognition* (2006-present), and the IEEE SIGNAL PROCESSING LETTERS (2006–2010), and a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (2013–2014 and 2007–2009). He is a fellow of the Canadian Academy of Engineering, and a recipient of the 2015 Primetime Engineering Emmy Award, the 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, the 2013 IEEE Signal Processing Magazine Best Paper Award, the 2009 IEEE Signal Processing Society Best Paper Award, the 2009 Ontario Early Researcher Award, and the ICIP 2008 IBM Best Student Paper Award (as senior author).

Hongwei Yong received the B.Sc. and M.Sc. degrees from Xi'an Jiaotong University, Xi'an, China, in 2013 and 2016, respectively. His current research interests include low-rank modeling, background subtraction, and video analysis.

Hongliang Li (SM'12) received the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, China, in 2005. From 2005 to 2006, he was with the Visual Signal Processing and Communication Laboratory, The Chinese University of Hong Kong, as a Research Associate, where he was a Post-Doctoral Fellow from 2006 to 2008. He was involved in many professional activities. He is currently a Professor with the School of Electronic Engineering, University of Electronic Science and Technology of China. He has authored or co-authored numerous technical articles in well-known international journals and conferences. He is a Co-Editor of a book titled *Video Segmentation and its Applications* (Springer). His research interests include image segmentation, object detection, image and video coding, visual attention, and multimedia communication system. He is a member of the Editorial Board of the *Journal on Visual Communications and Image Representation*, and the Area Editor of the *Signal Processing: Image Communication* (Elsevier Science). He served as a Technical Program Co-Chair of the ISPACS 2009, the General Co-Chair of the ISPACS 2010, the Publicity Co-Chair of the IEEE VCIP 2013, the Local Chair of the IEEE ICME 2014, and the TPC members in a number of international conferences, e.g., ICME 2013, ICME 2012, ISCAS 2013, PCM 2007, PCM 2009, and VCIP 2010. He also serves as a Technical Program Co-Chair for the IEEE VCIP2016.

Lei Zhang (M'04–SM'14) received the B.Sc. degree from the Shenyang Institute of Aeronautical Engineering, Shenyang, China, in 1995, and the M.Sc. and Ph.D. degrees in control theory and engineering from Northwestern Polytechnical University, Xi'an, China, in 1998 and 2001, respectively. From 2001 to 2002, he was a Research Associate with the Department of Computing, The Hong Kong Polytechnic University. From 2003 to 2006 he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, McMaster University, Canada. In 2006, he was with the Department of Computing, The Hong Kong Polytechnic University, as an Assistant Professor, where he has been a Full Professor since 2015. He has authored over 200 papers in his research areas. Since 2016, his publications have been cited over 20,000 times in the literature. His research interests include computer vision, pattern recognition, image and video processing, and biometrics. He is an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING and *SIAM Journal of Imaging Sciences and Image and Vision Computing*. He is a Web of Science Highly Cited Researcher selected by Thomson Reuters.