Visual Quality Assessment for Super-resolved Images: Database and Method

Fei Zhou, Rongguo Yao, Bozhi Liu, and Guoping Qiu

Abstract—**Image super-resolution (SR) has been an active research problem which has recently received renewed interest due to the introduction of new technologies such as deep learning. However, the lack of suitable criteria to evaluate the SR performance has hindered technology development. In this paper, we fill a gap in the literature by providing the first publicly available database as well as a new image quality assessment (IQA) method specifically designed for assessing the visual quality of super-resolved images (SRIs). In constructing the Quality Assessment Database for SRIs (QADS), we carefully selected 20 reference images and created 980 SRIs using 21 image SR methods. Mean opinion score (MOS) for these SRIs are collected through 100 individuals participating a suitably designed psychovisual experiment. Extensive numerical and statistical analysis is performed to show that the MOS of QADS has excellent suitability and reliability. The psychovisual experiment has led to the discovery that, unlike distortions encountered in other IQA databases, artifacts of the SRIs degenerate the image structure as well as image texture. Moreover, the structural and textural degenerations have distinctive perceptual properties. Based on these insights, we propose a novel method to assess the visual quality of SRIs by separately considering the structural and textural components of images. Observing that textural degenerations are mainly attributed to dissimilar texture or checkerboard artifacts, we propose to measure the changes of textural distributions. We also observe that structural degenerations appear as blurring and jaggies artifacts in SRIs and develop separate similarity measures for different types of structural degenerations. A new pooling mechanism is then used to fuse the different similarities together to give the final quality score for an SRI. Experiments conducted on the QADS demonstrate that our method significantly outperforms classical as well as current state-of-the-art IQA methods.**

Index Terms—**Full reference, image database, image quality assessment, image super-resolution.**

I. Introduction

IMAGE super-resolution (SR) is an important research problem in the field of image processing. It is extensively used in many applications, including high definition television, security surveillance, coding and transmission [1], etc. The aim of SR is to generate images with higher spatial resolution that are free from aliasing and other artifacts. One of the key challenges

The authors are with the College of Information Engineering, Shenzhen University, and with Guangdong Key Laboratory for Intelligent Information Processing, Shenzhen 518060, P. R. China. (e-mail: flying.zhou@163.com; rongguo_yao@163.com; bozhi.liu@hotmail.com; qiu@szu.edu.cn).

G. Qiu is also with the School of Computer Science, University of Nottingham, Nottingham NG7 2RD, U.K. (guoping.qiu@nottingham.ac.uk).

to the development of SR techniques is the assessment of the visual quality of the super-resolved images (SRIs). This paper fills a gap in the image SR literature by providing a carefully designed database and a state-of-the-art visual quality assessment method for image SR research.

A. Image Super-resolution

In the literature, different terminologies have been used to refer to the process of increasing the spatial resolution of an image, for simplicity, this work loosely refers to such a process as SR regardless of the underlying technique used. According to the available inputs, image SR techniques can be divided to multi-frame SR [2] and single-image SR [3]. In this work, we only consider the image super-resolved from one single low-resolution (LR) input. Some authors use image SR as a synonym of image upscaling [4]. Some early techniques are based on interpolation [5]. Commonly-used methods include nearest neighbor, bilinear, and bicubic interpolation [6], [7]. The traditional cubic convolution algorithm [6] is improved in [8] by modelling a non-separable convolution. In spite of the mathematical beauty in [6]-[8], image properties are not explicitly considered. To well capture the orientational property of images, the interpolation is performed along local isophotes in [9]. In all interpolation methods, image upscaling is treated as the problem of signal resampling. Different sampling kernels produce different methods. The sampling kernel can even be implicitly expressed, e.g., in an iterative correction [10] or in an iterative feedback [11]. Regardless of the expression of kernels, in the context of signal resampling, it is straightforward to use mean squared error (MSE) or peak signal-to-noise-ratio (PSNR) as the evaluation criterion [8]-[10].

Instead of aiming at the design of sampling kernels, most SR methods try to learn extra information from training samples. One of the earliest learning-based SR methods employs linear neural networks and vector quantization to predict the missing information in SRIs [1] [12]. Another early method is known as example-based SR [13], where the nearest neighbors from the training samples is used in a Markov field. Locally linear embedding (LLE) is employed in SR [14], based on the assumption that LR and high-resolution (HR) patches form manifolds with similar local geometry. The methods in [13] and [14] are inefficient, since they crudely employ the raw training samples without any compact coding or transformation. In [15], SR is achieved by support vector regression in the discrete cosine transform domain. In [16], kernel ridge regression (KRR) with a sparse solution is utilized for SR. Sparse-coding based SR (SCSR) is introduced in [17] by assuming that natural images

can be represented sparsely using a specified dictionary. It has many extensions, such as adaptive sparse domain selection (ASDS) [18], semi-coupled dictionary learning (SCDL) [19], statistical prediction model (SPM) [20], compact kernel sub-dictionary learning [21], consistent coding scheme (CCS) [22], etc. These extensions have achieved success via the incorporation of more knowledge in the image priors. Efficiency SR in both the training and testing stages are also attractive. To avoid the tedious optimization process in sparse coding, the anchored neighborhood regression method [23] and its extension, known as A+ [24], make use of the collaborative representation to infer HR images. The mapping between LR and HR features in [25] is formed as very simple functions (SF), and thus it is much fast. In some interesting work, the training samples are from self-exemplars, instead of external datasets [26]. Loosely speaking, we can call the methods in [12]-[26] as dictionary-based SR. In this kind of SR, PSNR or MSE is still the preferred choice to perform their quantitative comparisons [12]-[26]. In addition, structural similarity (SSIM) index [27], starts to become popular [18], [20]-[22], [25], [26]. However, neither of them is suitable for SR tasks [28].

Recently, deep neural networks (DNN) have been applied to SR with remarkable success. A two-hidden-layer convolutional neural network (CNN) is exploited in [29] to imitate the coding and de-coding in the dictionary-based SR. In [30], sparse constraint is embedded in the CNN by inserting a sub-network to imitate the behavior of sparse coding. In [31], the deconvolution layer is moved to the end of CNN so that the parameter number is largely reduced. With a skip-connection, the works in [32] achieve a very deep SR (VDSR) network. In [33], the deep CNN is improved by multiple skip-connections and network in network (DCSCN). In [34], the deep recursive residual network (DRRN) for SR is proposed by recursively using the same module, combined with skip-connections. The motivation behind [32]-[34] is to build much deeper networks for SR. Besides, the behaviors of skip-connections or residual modules are similar to that of ensemble learning [35]. The supervision on intermediate layers is used in a Laplacian pyramid SR network (LapSRN) [36] to achieve large magnification factors. In [38], features of trained VGGNet [37], which are believed to relate to visual perception, are used as the loss function of SR. In [39], the SR using a generative adversarial network (SRGAN) is proposed by transferring the distribution of output images to that of the ground truth. The results of [38] and [39] are visually pleasing, but their PSNR and SSIM values are less competitive. In DNN-based methods, PSNR and SSIM continue to serve as the evaluation criteria in quantitative comparisons [29]-[39]. To show the superiority of SRGAN, the authors of [39] have to use mean opinion score (MOS) to complement PSNR and SSIM.

B. Image Quality Assessment

Image quality assessment (IQA) has attracted extensive research interest in recent times [40]. In this work, we focus on full-reference (FR) IQA, where the image of 'perfect' quality is available. In the context of image SR, the 'perfect' image is the ground truth. The most popular FR IQA method is PSNR, but it is well known that the correlation between PSNR and perceived

quality is low. This motivates the development of other IQA measures. In order to correlate with the human visual system (HVS) well, a straightforward idea is to incorporate the properties of the HVS. In [41], visual signal-to-noise-ratio makes use of the near-threshold and supra-threshold characteristics of human vision. In [42], a method named as the most apparent distortion (MAD) adaptively exploits Fourier transformation and log-Gabor filtering to extract visual features according to the visibility of the distortion. Moreover, in [41] and [42], different channel decompositions are employed in an attempt to exploit the discovery that there may exist multiple channels of octave spacing radial frequency in the visual pathway [43]. However, a complete understanding of the HVS is still unavailable, thus only part of its properties can be modelled [44].

Based on the observation that the HVS is highly adaptive to specific structural information, the SSIM index [27] performs visual comparisons on three aspects, i.e., luminance, contrast, and structure. It has been utilized to replace PSNR in many applications of image processing, including SR. Some schemes aim at improving SSIM, e.g., [45]-[48]. In [45], multi-scale SSIM (MS-SSIM) index is achieved by performing SSIM on multiple scales of images. In [46], SSIM is performed in the wavelet domain. In [47], information content weighted SSIM (IW-SSIM) index focuses on the pooling strategy. A multivariate SSIM is proposed in [48] to assess the quality of hyperspectral images. In addition to mean, variance, and covariance, many other features or cues are adopted in IQA. Information fidelity criterion (IFC) [49], as well as its extension visual information fidelity (VIF) [50], calculates the visual quality as the mutual information between the reference and the distorted images. Besides, gradient features are adopted in IQA algorithms since it is believed that image gradients can convey important visual information. In [51], gradient magnitudes combined with phase congruency are employed to calculate feature similarity (FSIM) index. In [52], gradient similarity (GSIM) index is proposed to incorporate the gradient features with masking effect and distortion visibility. Gradient magnitude similarity deviation (GMSD) in [53] utilizes the global variation of local gradient similarity as the pooling strategy. In IQA, visual saliency of images is also utilized, as both the local features and pooling weights [54]. In some IQA methods, the predictability is used as a cue. In the method based on internal generative mechanism (IGM) [55], different strategies are adopted for the predicted and unpredicted portion of images. The difference of predicted coefficients is employed in [56] to measure the visual quality, followed by a CNN-based distortion compensation. Some work goes further in the description of image structure. The structure contrast defined in the discrete cosine transformation (DCT) domain is employed in [57]. To measure the image structure, directional anisotropy structure measurement (DASM) in [58] involves local gradient, anisotropy, and directionality. In our recent work, a super-pixel-based similarity (SPSIM) index [59] is proposed to extract simple local features within each super pixel instead of image patch.

Most of existing FR IQA methods emphasize the importance of distortions on structure, based on the consensus that image structure is dominant in the visual perception. However, for

SRIs, image details are also important. Moreover, some artifacts that may appear in the SRIs are not considered in the existing IQA methods. Therefore, these existing methods are not best suited for SRIs.

C. This Work

As discussed above, many researchers have been aware that PSNR and SSIM, the most popular criteria for performance evaluation of SR methods, are not suitable for the IQA of SRIs. In addition to SSIM and PSNR, several SR methods also turn to some other existing IQA methods, such as VIF used in [21] and IFC used in [36]. Nevertheless, in this work, we will show that although VIF and IFC are much better than PSNR and SSIM in the context of image SR, their performances are still unsatisfactory. Moreover, it is worthwhile to notice that the gradient is of great important in image SR. Hence, it seems the IQA methods using the gradient as a feature, e.g., [51]-[53], are promising in the visual quality assessment for SRIs. In this work, we have tested them on SRIs in Section IV. Unfortunately, their performance is far from satisfactory. Another strategy to evaluate the SR methods is to use MOS by subjective evaluations, e.g., [39]. However, obtaining MOS is a labor-intensive process. Furthermore, the reusability of MOS values is very low. Unlike the PSNR and SSIM values, the values of MOS obtained from one psychovisual experiment cannot be directly applied in new SR comparisons. It is because the settings to obtain MOS can hardly remain the same, and individuals involved in the evaluation are also varied. That is, the subjective evaluation would have to be re-performed from the beginning when new comparisons are required. Therefore, it is highly demanded to develop objective assessment algorithms for SRIs, in accordance with human vision.

To conduct IQA researches, an image database with MOS or differential MOS (DMOS) is necessary. Some well-known IQA databases, such as IVC [60], LIVE [61], MICT [62], TID2008 [63], CSIQ [42], and TID2013 [64], are publicly available and support the development of IQA. Recently, several new IQA databases are built for different purposes. For example, a multiply distorted image database (MDID) is exhibited in [65] to enable the research on the quality assessment for images with multiply types of distortions. To facilitate the IQA research in tone mapping of images with high dynamic range (HDR), an IQA database for tone-mapped images is presented in [66]. However, these IQA databases are unsuitable in the case of SRIs. Specifically, none of their images is a super-resolved one. In most existing FR IQA databases [42], [60]–[65], the distorted images are produced by introducing some common distortions to reference images. In [66], the images are generated by 11 HDR processing algorithms. Many distortions or artifacts in the images of these databases, e.g., impulse noise, JPEG, JPEG2000, etc., can hardly be observed in SRIs while some possible artifacts in SRIs, e.g., jaggies and checkerboard, are not included in these IQA databases. To our knowledge, there is no publicly available database for the quality assessment of SRIs up to now. Thus, establishing an IQA database for SRIs is highly necessary and fills a gap in the literature.

Fig. 1. Illustration of source images.

Motivated by the above, in this paper, we focus on the visual quality assessment for SRIs. Specifically,

Firstly, a Quality Assessment Database for SRIs (QADS) is presented to facilitate the research. The benchmark database contains 20 HR images as the reference and 980 SRIs created using 21 SR methods. Almost all the artifacts that frequently appear in SRIs can be found in QADS. Using a psychovisual experiment procedure specifically designed for subjective assessment of SRI visual quality, 100 individuals participated in the subjective evaluation to acquire reliable MOS.

Secondly, based on the observation that the visual artifacts on the structure and the texture behave differently, we propose a new method to assess the visual quality of SRIs by separately considering the structural and textural parts of images. Since textural degenerations mainly manifest as dissimilar texture or checkerboard artifacts, we propose to measure the changes of textural distributions to take into account both types of textural artifacts. For the structural component, separate similarity measures are calculated to measure the artifacts of blurring and jaggies respectively. The pooling is first performed on individual ingredients to get respective scores, and then the scores are fused to get the final single score.

Thirdly, experiments conducted on QADS show that our new visual quality assessment method significantly outperforms classical as well as state-of-the-art IQA methods.

II. DATABASE FOR SRI QUALITY ASSESSMENT

In this section, we will introduce the details of QADS, including the preparation of images, subjective evaluation and data analysis. The suitability and reliability of QADS will also be explained. This database is publicly available at [67].

A. Preparation of Images

In our QADS, the reference images, also known as source images, serve as the HR ground truth in image SR. Since image contents will have an impact on visual quality assessment, selection of reference images is nontrivial. The general principle is that the reference images should be clean, and their contents ought to be varied [65], [68]. Initially, we selected the 20 reference images in MDID [65] as our source. It has been demonstrated in [65] that these source images contain a wider range of spatial information (SI) and colorfulness [69] than other IQA databases. However, during the course of sub-

TABLE I
COMPARISONS OF QADS WITH SOME OTHER IQA DATABASES

IQA Databases	SI index d	Averaged SD†	KL divergence
IVC [60]	61.57	11.91	0.0824
LIVE [61]	101.77	N/A	0.1324
MICT [62]	88.92	14.73	0.1553
TID2008 [63]	105.83	7.28	0.3831
CSIQ [42]	105.56	7.80	0.1703
TID2013 [64]	105.83	7.21	0.3910
MDID [65]	107.78	6.48	0.0233
QADS	108.59	6.76	0.0247

† SD is normalized by SD/max(MOS)×100, since the score ranges in different databases are not the same.

TABLE II
SOME DETAILS IN GENERATING SRIS

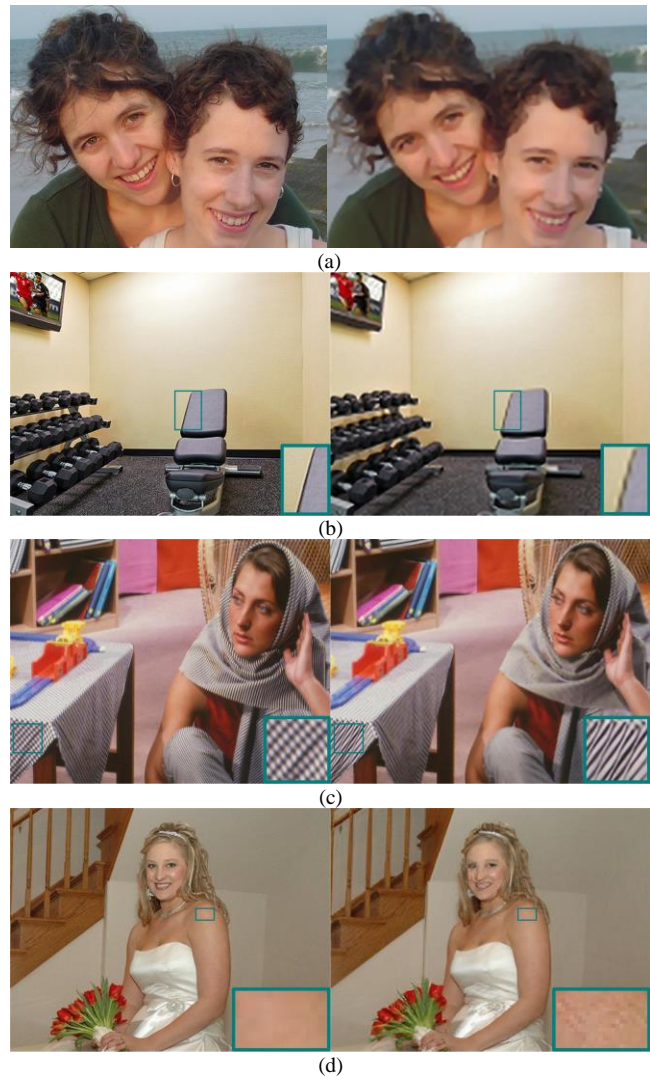| SR methods		Factors	Number
Interpolation based	Bilinear interpolation	2, 3, 4	60
	Bicubic interpolation	2, 3, 4	60
	Orientational interpolation [9]	2, 3, 4	60
	Fast up-sampling [11]	2, 3, 4	60
Dictionary based	Example-based SR [13]	3, 4	40
	LLE [14]	3, 4	40
	Sparse KRR [16]	2, 3, 4	60
	SCSR [17]	3, 4	40
	ASDS [18]	2, 3, 4	60
	SCDL [19]	2, 3, 4	60
	SF [25]	2, 3, 4	60
	SPM [20]	3	20
	A+ [24]	2, 4	40
	Self-exemplars [26]	3, 4	40
	CCS [22]	2, 4	40
DNN based	VDSR [32]	2, 3, 4	60
	VGGNet [38]	2, 4	40
	DCSCN [33]	3, 4	40
	DRRN [34]	3, 4	40
	LapSRN [36]	2, 4	40
	SRGAN [39]	4	20

Fig. 2. Examples of various artifacts in QADS. (a) Blurring. (b) Jaggies. (c) Dissimilar texture. (d) Checkerboard. In (a)-(d), the left is the reference image (HR ground truth), while the right is the distorted image, i.e., SRI.

jectively evaluating these images, we found that for two of the source images, many SR methods will generate visually indistinguishable SRIs. It turns out that these two source images contain small SI values resulting in different SR methods producing visually similar results. This would make subjective scores less meaningful and even unreliable. To make the subjective evaluation more meaningful and reliable, we replaced these two by another two images that frequently serve as the testing images in SR. The 20 reference images finally adopted in QADS are shown in Fig. 1. All the reference images were initially cropped into size 504×384 without scaling or rotation.

For IQA databases, the ranges of SI and colorfulness, defined in [69], can be used to analyze the suitability of the selections of reference images. In the context of image SR, SI plays a more important role, since the aim of SR is to increase the spatial resolution. From the standpoint of building IQA databases with various image contents, larger range of SI is preferred. From the view of evaluating SRIs, the images with richer SI are more meaningful. Hence, in this work, we comprehensively analyze the range and value of SI by using the SI index defined as

$$d = \sqrt{\text{mean}(v) \cdot \text{range}(v)}$$

where d denotes SI index and v represents the SI values defined in [69]. The functions mean($\cdot$) and range($\cdot$) respectively return the mean value and the value range of their arguments, over all the reference images. Generally speaking, larger mean value implies richer SI. As can be seen in Table I, compared with other databases, QADS has the largest SI index, indicating the suitability of its reference images for SRI quality assessment.

SRIs can be treated as distorted images from the view of IQA databases. To obtain the distorted images, we first use bicubic down-sampling to reduce the size of the reference images by a factor of k (k = 2, 3, 4), and then use 21 SR methods to super-resolve the reduced image back to their original sizes. Note that the size of the original reference images is 504×384, the size of the SRI is cut to 500×380 to avoid the visual impacts from image borders. The reason is that the pixels from the borders of SRIs are often abnormal due to the padding operation in many SR methods, and we do not want these abnormal pixels to influence the opinions of subjects.

The 21 methods include 4 interpolation-based methods, 11 dictionary-based SR, and 6 DNN-based SR. The selected SR methods are representative, i.e., they are either widely-accepted or state-of-the-art. In Table II, the details of the 21 SR methods

Fig. 3. Photograph of the evaluation environment.

and associated down-scaling/up-scaling factors are provided. The bilinear and bicubic interpolations were implemented using the built-in function of MATLAB. The method in [9] was implemented by ourselves. All the other methods were implemented by the codes provided by their authors or downloaded from the homepages of their authors. For all the codes, we directly used the default settings of parameters and the dictionaries or DNNs that had already been trained. This implies that the training data for different SR methods might be varied. However, it does not matter because our aim is just to produce real SRIs, instead of performing comparisons among these 21 methods. Several methods use different down-scaling methods rather than bicubic to create their input images. For example, in the default setting, the input of SCDL is expected to be the down-sampled version of HR images using delta sampling kernel. Nevertheless, the same inputs are employed for all the methods in creating the database, since we believe that the artifacts caused by the inconsistency between testing inputs and training inputs should be taken into consideration. The number of SRIs produced by each method is also provided in the last column of Table II. By summing the last column of Table II, it can be shown that the total number of SRIs is 980. For each reference image, the number of SRIs is 49.

In QADS, the distorted images, i.e., the SRIs, contains typical artifacts that frequently appear in image SR. Aliasing in high-frequency areas would appear due to the low sampling rate when generating digital images. The key of image SR is to retrieve the aliased high-frequency information. However, if the high-frequency information of SRIs fails to be recovered, the images would look blurry. A visual example of a reference image and the blurring artifact appears in its SR version is displayed in Fig. 2(a). Another type of commonly encountered artifact in SRIs is jaggies, which is also known as zigzags. The emergence of jaggies is derived from the aliasing as well. When anti-aliasing operations in the SR methods create incorrect high-frequency components, the jaggies would appear. An example of jaggies is provided in Fig. 2(b), where the part with obvious jaggies is highlighted. Some DNN-based methods may generate very sharp images but with dissimilar texture to the original, as shown in Fig. 2(c). Checkerboard artifacts, mainly caused by the overlapping pattern of convolution kernels in the deconvolution layer [70], also appear in QADS. Fig. 2(d) shows an example of checkerboard artifacts. In addition to the 4 kinds of artifacts shown in Fig.2, other artifacts which often

Fig. 4. Screenshot of the software interface used in subjective evaluation. The resolution of the interface is 1440×900, which is the same as that of the monitor. With this resolution, the images in the interface can be exhibited completely without scaling.

appear in SR also can be found in QADS, e.g., ringing artifacts in some over-sharpened SRIs. Obviously, the artifacts in QADS are essentially different from the distortions in existing IQA databases.

B. Subjective Evaluation and Data Analysis

Subjective evaluation is an important procedure in building IQA database, although it is laborious. A total of 100 subjects participated in this procedure. The subjects were postgraduates from different disciplines, and all of them were with normal eyesight. The environment as well as the devices to perform the subjective evaluation were fixed. Specifically, all the subjects were required to accomplish their evaluations in an indoor environment without any background light. The device to show the evaluation interface was a 23.8-inch liquid crystal display monitor with spatial resolution of 1440×900. The other configurations of the monitor, such as the brightness and color temperature, remained default and unchanged during the whole subjective evaluation. A photograph of the evaluation environment is provided in Fig. 3.

The software interface used in subjective evaluation is illustrated in Fig. 4, where there are four image windows shown simultaneously. Therein, the top row is two SRIs to be evaluated, they share the same reference image shown in the bottom-right window. Furthermore, we propose to use the bottom-left window to show an image that can be controlled by the subjects during the evaluation. The participant can control which of the 3 other images will be shown in this window by pressing the key "1", "2", or "3" on the keyboard. In the psychovisual evaluation, we find that the subjects can make their decisions much more quickly and more precisely by flipping the three images at exactly the same position, i.e., the bottom-left window, than only observing them in a side-by-side display. During the evaluation, subjects were required to make their decisions as soon as possible and were instructed to click the button ">", "<", or "=" on the interface to indicate their judgements. The initial distance between the subjects and the monitor was approximately twice the screen height, as suggested in [68]. After starting a round of evaluation, the subject was able to slightly adjust the viewing distance to make more precise decisions. For more details about the psychovisual
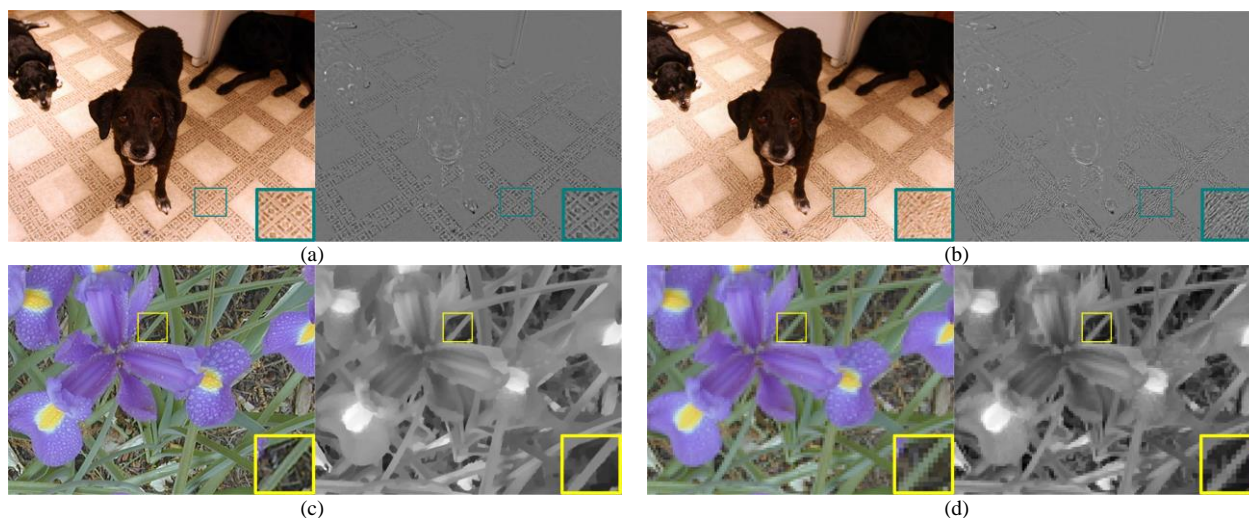
Fig. 5. Impacts of artifacts on textural and structural components. (a) A reference image and its textural component. (b) An SRI and its textural component. (c) A reference image and its structural component. (d) An SRI with its structural component.

evaluation, please refer to the online materials in [67]. To record and process judgements by the subjects, we employed the pair comparison sorting (PCS) algorithm [65], which enables the option of "=" in the sorting of subjective evaluation. After sorting the subject scores, for each SRI, we can obtain a number that indicates the index of its quality. Since there are 49 SRIs for each reference image in QADS, the numbers range from 1 to 49. A smaller number means worse quality.

The MOS for each SRI is calculated as follows: First, individual scores are normalized to the range of (0, 1). Subsequently, outlier removal and subject rejection are conducted based on a method described in LIVE [61] to exclude the unreliable individual scores. Finally, MOS for each SRI can be calculated as the average of the remaining valid scores.

Although much previous experience on the subjective evaluation can be followed, it is still necessary to check the reliability and suitability of the final scores. The reliability can be measured by using the standard deviation (SD) of individual scores per upscaled image while the suitability can be measured by the uniformity of MOS [69]. A small SD means the consistency on the visual quality among different subjects is high, thus the score is believed to be reliable. Table I provides the averaged SD over all the SRIs of QADS and some other FR IQA databases. The uniformity of MOS expects that they should be uniformly distributed so that the full range of the rating scale can be fully utilized. We calculate the 30-bin histograms of MOS or DMOS for different databases, and the Kullback–Leibler (KL) divergences between them and uniform distributions are given in Table I. A small value of KL divergence indicates that the uniformity of MOS or DMOS is high. In Table I, comparisons with other well-known IQA databases demonstrate the reliability and suitability of MOS in QADS.

III. IQA FOR SRIs USING STRUCTURE-TEXTURE DECOMPOSITION

During the construction of QADS, we observed that some artifacts, e.g., those shown as Figs. 2(c) and (d), do not change the image structure much but are noticeable in the psychovisual

evaluation. This enlightens us to develop quality assessment for SRIs by considering the structural and textural parts of the images separately. Therefore, we make use of structure-texture decomposition (STD). As a widely known image processing technique, STD decomposes an image into two parts, i.e., structural component and textural component [71]. It has been successfully used to solve many problems, e.g., road detection [72] and defocus estimation [73]. In this work, we propose an FR IQA method for SRIs using STD, which enables us to design specialized IQA measurements for different types of artifacts. In fact, some artifacts in SRIs are conspicuous in the textural part while the other artifacts mainly exist in the structural component. Two visual examples are provided in Fig. 5. By comparing Figs. 5(a) and (b), we can readily observe the dissimilar texture in the textural component while the jaggies remains in structural part, as shown in Figs. 5(c) and (d).

The proposed method is further motivated by the fact that humans perceive the textural and structural components differently. Specifically, if the textures of two image regions belong to the same textural type, their differences can be difficult to perceive, even the difference at a fixed image location is large. Conversely, a large structural disparity at a fixed location can be easily perceived. Thus, the artifacts in the texture component will be clustered together while the artifacts in the structure component will appear at specific structural locations. Since the HVS is more sensitive to the changes on structures than on textures, the focus of most FR IQA methods is on the description of structural distortions. Nevertheless, for SRIs, it is essential to investigate the artifacts on texture, rather than ignoring them.

A. Textural Similarity

As mentioned above, for textures, the HVS concentrates on the textural type mainly. The difference between two completely different textures can be easily perceived, whereas the textures sharing similar distributions provide similar visual perception. Therefore, we use a statistical descriptor to capture the textural distribution, instead of structural features. Meanwhile, as suggested in [27], it is better to get a spatially varying
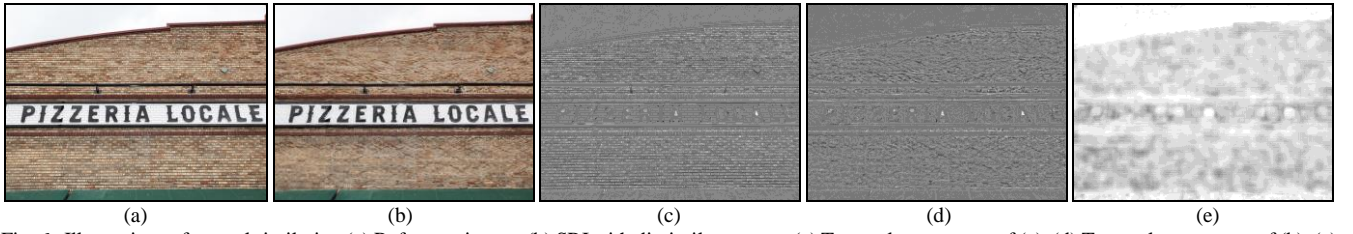
Fig. 6. Illustrations of textual similarity. (a) Reference image. (b) SRI with dissimilar texture. (c) Textural component of (a). (d) Textural component of (b). (e) Quality map of textural similarity. In (e), the smaller intensity means lower textural similarity.

Fig. 7. Illustrations of structural similarity. (a) Reference image. (b) SRI with jaggies. (c) Structural component of (a). (d) Structural component of (b). (e) Quality map of structural similarity. In (e), the smaller intensity means lower structural similarity.

quality map so that the spatial information of the artifacts can be made available. To achieve these, in this work, we employ the well-known descriptor of scale-invariant feature transform (SIFT) [74] in a dense way. Specifically, the SIFT feature with one scale and no rotation is calculated for each pixel of the textural component. The reasons to exclude the multiscale operations and rotations are twofold. Firstly, it is more computationally efficient to consider the feature in only one scale and without rotation. Secondly and crucially, our purpose is to describe textural distributions, instead of robust key points. In other words, the textures with varied scales or rotations should be distinguished in the context of image SR. The dense SIFT feature, essentially, is a concatenated histogram to describe the distribution of gradients in an image region. With the histogram-based feature, the textural similarity for the i-th pixel $M_t(i)$ is measured as

$$M_t(i) = \frac{\left\langle \mathbf{f}_r(i)/\|\mathbf{f}_r(i)\|_2, \mathbf{f}_u(i)/\|\mathbf{f}_u(i)\|_2 \right\rangle + K_t(i)}{1 + K_t(i)}, \quad (1)$$

where $\|\cdot\|_2$ denotes L2-norm, $\langle\cdot, \cdot\rangle$ denotes inner product, and $\mathbf{f}_r(i)$ and $\mathbf{f}_u(i)$ are the feature vectors of histograms at the i-th pixel in the textural components of the reference image and the SRI, respectively, and K_t is an adaptive variable, defined as

$$K_t(i) = \frac{C_t}{\max\left(\mathrm{var}\left(\mathbf{t}_r(i)\right), \mathrm{var}\left(\mathbf{t}_u(i)\right)\right)}, \quad (2)$$

where $\mathbf{t}_r(i)$ and $\mathbf{t}_u(i)$ represent the patches centered at the i-th pixel in the textural components of the reference image and the SRI, C_t is a positive constant to adjust the range of K_t, $\mathrm{var}(\cdot)$ calculates the variance of elements in its argument, and $\max(\cdot, \cdot)$ returns the maximum value of its arguments. The function of K_t is somewhat similar to that of the masking parameter in [52].

Through (1) and (2), we can find several characteristics of the textural similarity M_t. First, it is easy to prove that M_t ranges from 0 to 1. And a high correlation between normalized $\mathbf{f}_r$ and $\mathbf{f}_u$,

which indicates similar distributions of textures, would produce a large value of M_t. Second, K_t has little impact on M_t if either the reference or the super-resolved patch is with rich texture, i.e., large variance. In this case, the value of M_t is only determined by $\mathbf{f}_r$ and $\mathbf{f}_u$. Third, if both the reference and the super-resolved patches are with little texture, e.g., invisible texture, the value of K_t would be very large and force M_t to 1. A more sensible way from the view of psychophysics is to conduct a frequency analysis, incorporating the contrast sensitivity function, to judge the visibility of texture. However, we find that using the simple variance in (2) can already achieve satisfactory results in this work. After calculating M_t for each pixel, we can get a map of textural similarity.

An example of the textural similarity is provided in Fig. 6. From Fig. 6(e), it can be seen that M_t captures the regions with artifacts on the texture well. It is worth to note that although SIFT-based feature is used in this work, other texture descriptors may also be used. More analysis on the selection of texture descriptors can be found in the online supplement materials in [67].

B. Structural Similarity

Although dissimilar texture and checkerboard mainly manifest in the textural component, other artifacts, e.g., jaggies, would mainly appear in the structural component of images. Jaggies, a kind of commonly encountered artifacts in SRIs, generally cause directional distortions on the structure, e.g., Fig.2 (b) and Fig.5 (d). To measure the jaggies, we choose to compare the dominant directions of the gradients in the structural components. The dominant direction of a patch can be derived from the following positive semi-definite matrix $\mathbf{J}$ [75]:

$$\mathbf{J}(i) = \begin{bmatrix} \mathbf{g}_x^T(i)\mathbf{g}_x(i) & \mathbf{g}_x^T(i)\mathbf{g}_y(i) \\ \mathbf{g}_y^T(i)\mathbf{g}_x(i) & \mathbf{g}_y^T(i)\mathbf{g}_y(i) \end{bmatrix}, \quad (3)$$

where i is the location index of the patch center, $\mathbf{g}_x$ and $\mathbf{g}_y$ are the vectors containing gradients in lexicographic order along the abscissa and the ordinate, respectively. The matrix $\mathbf{J}$ has
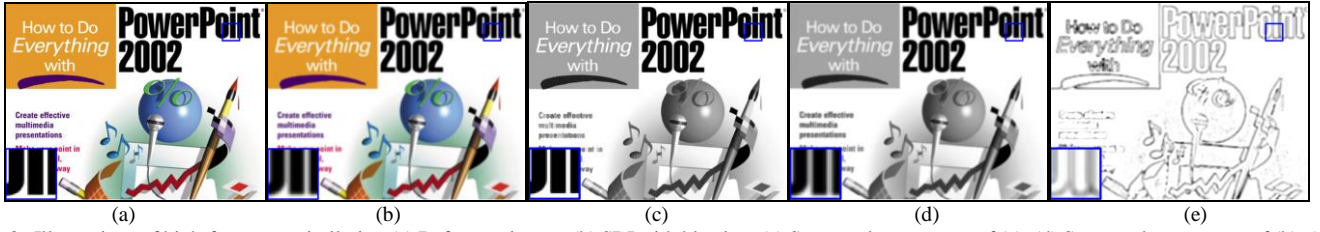
Fig. 8. Illustrations of high-frequency similarity. (a) Reference image. (b) SRI with blurring. (c) Structural component of (a). (d) Structural component of (b). (e) Quality map of high-frequency similarity. In (e), the smaller intensity means lower high-frequency similarity.

two eigenvalues, and the dominant direction can be represented by the eigenvector corresponding to the lower one. Similar to (1), the measurement of structural similarity for the i-th pixel $M_s(i)$ is designed as

$$M_s(i) = \frac{\left| \langle \mathbf{n}_r(i), \ \mathbf{n}_u(i) \rangle \right| + K_s(i)}{1 + K_s(i)}, \qquad (4)$$

where $|\cdot|$ returns the absolute value, and $\mathbf{n}_r(i)$ and $\mathbf{n}_u(i)$ are the normalized eigenvectors, indicating the dominant directions at the i-th pixel in the structural components of the reference image and the SRI, respectively. And K_s is defined as

$$K_s(i) = \frac{C_s}{\max\left(g_{mr}(i), g_{ms}(i)\right)}, \qquad (5)$$

where $g_{mr}(i)$ and $g_{ms}(i)$ represents the normalized gradient magnitude at the i-th pixel in the structural components of the reference image and the SRI respectively, and C_s is a positive constant to adjust the range of K_s. It should be noted that if $\mathbf{n}$ is a normalized eigenvector corresponding to a given eigenvalue, $-\mathbf{n}$ is also an eigenvector corresponding to the same eigenvalue. Therefore, the absolute value in (4) is nontrivial.

Obviously, M_s in (4) has the similar characters as M_t in (1), since they share similar mathematical formulas. For instance, the range of M_s is also from 0 to 1. If the vectors $\mathbf{n}_r(i)$ and $\mathbf{n}_u(i)$ point to the same direction, the structural similarity reaches 1, i.e., the maximum value. The role of K_s in (4) is also similar to that of K_t in (1). Specifically, if either the reference image or the SRI has strong gradients in the structure, the impact of K_s on (4) can be ignored. On the other hand, if both the reference image and the SRI are smooth in the structure, the estimations on the dominant directions would be susceptible to noises and thus unreliable. In this case, K_s would play a vital role in obtaining reasonable results by pushing M_s to 1. The characteristics of M_s and M_t show that the design of textural and structural similarity is in accordance with visual perception qualitatively.

To show the validity of the proposed structural similarity, we show an example in Fig. 7, where a region with severe jaggies is highlighted. From Fig. 7, it can be observed that jaggies with directional distortions can be captured.

C. High-frequency Similarity

The loss of high-frequency details in SRIs would make them look blurry. Although both the textural and structural parts of images can suffer from blurring, the textural similarity defined

in (1) has the capacity to distinguish blur textures from sharp ones. Thus, in this work, the similarity of high frequency is only calculated on the structural component of images. Since this similarity only depends on the structural component, it can also be regarded as one kind of structural similarity. Here, we name it as high-frequency similarity to distinguish it from (4). It is worthwhile to note that structural component is not a synonym of low-frequency part, although some high-frequency details exist in the textural component. Actually, the structural component of the reference images can have very sharp edges, e.g., Fig. 8(c). Losing their energy in high frequency, sharp edges in the structural component would become blur, such as Fig. 8(d). To measure the high-frequency energy h at the position of the i-th pixel, we propose to use a simple but effective expression as follows

$$h(i) = \frac{1}{N_\mathcal{N}} \sum_{j \in \mathcal{N}(i)} \left(\mathbf{s}(j) - \mathbf{s}_\sigma(j) \right)^2, \qquad (6)$$

where j is a location index, $\mathcal{N}(i)$ is the neighborhood of i, $N_\mathcal{N}$ is the number of neighbors, $\mathbf{s}$ is the structural component, and $\mathbf{s}_\sigma$ is obtained via convolving $\mathbf{s}$ by the Gaussian filter with an SD of σ. In (6), $\mathbf{s}_\sigma$ represents the low-frequency part of $\mathbf{s}$. Comparing the high-frequency energy in the reference image and the SRI, the high-frequency similarity for the i-th pixel $M_h(i)$ is given by

$$M_h(i) = \frac{2h_r(i)h_u(i) + C_h}{h_r^2(i) + h_u^2(i) + C_h}, \qquad (7)$$

where h_r and h_u are calculated using (6) in the reference image and SRI, respectively, C_h is a positive constant to avoid the instability caused by a small denominator. The mathematical form of (7) has been demonstrated to be consistent with the masking effect in many previous works [27], [52], [57], [59].

Similar to Figs. 6 and 7, we also provide a visual example of high-frequency similarity in Fig. 8. From the examples in Figs. 6-8, it can be observed that the artifacts in the textural part cluster in an image region, while the artifacts in the structural component lie in sparse locations. These results support the rationale of measuring the artifacts based on STD.

D. Pooling

Given a reference image and its corresponding SRI, we need to pool the above quality maps into a single score, which indicates the final quality of the SRI. The traditional strategy is first fusing the multiple quality maps into one map, and then

TABLE III
PERFORMANCE TESTING OF DIFFERENT METHODS ON QADS

Criteria	IQA Methods													SIS
	PSNR	SSIM [27]	MS-SSIM [45]	IFC [49]	VIF [50]	MAD [42]	IW-SSIM [47]	FSIM [51]	GSIM [52]	IGM [55]	GMSD [53]	DASM [58]	SPSIM [59]	
SROCC	0.3544	0.5290	0.7172	0.8609	0.8152	0.7234	0.8195	0.6885	0.5538	0.7145	0.7650	0.7512	0.5751	**0.9232**
KROCC	0.2441	0.3689	0.5299	0.6816	0.6249	0.5293	0.6283	0.5020	0.3908	0.5231	0.5689	0.5622	0.4071	**0.7541**
PLCC	0.3897	0.5327	0.7240	0.8657	0.8210	0.7311	0.8234	0.6902	0.5684	0.7192	0.7749	0.7585	0.5822	**0.9230**
RMSE	0.2530	0.2325	0.1895	0.1375	0.1568	0.1874	0.1559	0.1988	0.2260	0.1907	0.1736	0.1790	0.2233	**0.1057**

pooling the pixel-wise scores into a single one. However, in our method, the three quality maps describe quite different aspects of the properties during the process of image SR. It would be more meaningful to investigate the individual similarities to discover the possible weakness of a given SR method. Therefore, we first pool the three quality maps into three scores, and then fuse the three scores into one. The pooling for each map is achieved by weighted mean, i.e.,

$$p_q = \frac{1}{N}\sum_i w_q(i)M_q(i), \qquad (8)$$

where N is the number of pixels in the images, the subscript $q \in \{t, s, h\}$ is an index of the three similarities, p_q is the score of each similarity, and w_q is the weight for each pixel. In (8), the weights are calculated by considering the contents in each map:

$$w_t(i) = \frac{\max\left(\mathrm{var}\left(\mathbf{t}_r(i)\right), \mathrm{var}\left(\mathbf{t}_u(i)\right)\right)}{\sum_i \max\left(\mathrm{var}\left(\mathbf{t}_r(i)\right), \mathrm{var}\left(\mathbf{t}_u(i)\right)\right)},$$

$$w_s(i) = \frac{\max\left(g_{mr}(i), g_{mu}(i)\right)}{\sum_i \max\left(g_{mr}(i), g_{mu}(i)\right)}, \qquad (9)$$

$$w_h(i) = \frac{\max\left(h_r(i), h_u(i)\right)}{\sum_i \max\left(h_r(i), h_u(i)\right)}.$$

The denominators in (9) are used for normalization. The design of the weights is straightforward, since the individual score for each map is calculated at first. As an example, for the textural similarity, it makes senses that the patches with rich textures in either the reference image or the SRI have higher weights. But if multiple quality maps are first fused to one map, the design of weights would be much more complicated.

The final single score p is obtained by fusing the above three scores as follows,

$$p = p_t^\alpha \cdot \left(p_s \cdot p_h\right)^\beta, \qquad (10)$$

where $\alpha > 0$ and $\beta > 0$ are used to adjust the impact of different similarities. The scores of the structural and high-frequency similarities share the same parameter β, since both of them are estimated in the structural component. Empirically, β should be larger than α, due to the importance of structure in the HVS. Without loss of generality, we can simply set α to 1. To obtain the value of β, we heuristically use the ratio between the mean intensities of structural and textural parts. Specifically,

$$\beta = \frac{\beta}{\alpha} = \frac{\log\left(\mathrm{mean}\left(|\mathbf{s}|\right)\right)}{\log\left(\mathrm{mean}\left(|\mathbf{t}|\right)\right)}, \qquad (11)$$

where $\mathbf{s}$ and $\mathbf{t}$ are the intensities of the structural and textural components of the images, respectively. In (11), the $\log(\cdot)$ function is adopted to follow the Weber-Fechner law. Since structural intensities are generally lager than textural intensities, β is larger than 1. In this work, we make use of external images to estimate the parameter β instead of the images in QADS.

IV. EXPERIMENTS

In this section, the STD-based IQA method for SRIs, or SIS for short, is tested on the newly established database, i.e., QADS, and compared with some other representative FR IQA methods, including PSNR, SSIM [27], MS-SSIM [45], IFC [49], VIF [50], MAD [42], IW-SSIM [47], FSIM[51], GSIM [52], IGM [55], GMSD [53], DASM [58], and SPSIM [59].

The settings of SIS are as follows. The positive constants in (2), (5), and (7) are simply set to 1, i.e., $C_t=C_s=C_h=1$. The gradient operators used in (3) and (5) are the widely-used Sobel operators. The STD is performed via the default implementation of our previous work in [73]. The dense feature in (1) is also implemented in the default configuration, resulting in a 128-bin histogram for each location. And σ in (6) is set to 5 empirically. To obtain the value of β in (10), we employ the reference images from LIVE [61] as the external images in (11). Based on this, the estimated value of β is 3.9709. For the other IQA methods, their default settings are adopted.

Four criteria are utilized to evaluate the performance of SIS and compared IQA methods on QADS. They are Spearman rank order correlation coefficient (SROCC), Kendall rank order correlation coefficient (KROCC), Pearson's linear correlation coefficient (PLCC), and root mean squared error (RMSE) between the predicted scores of IQA methods and the MOS obtained in Section II. Among them, SROCC and KROCC indicate the prediction monotonicity, while PLCC and RMSE are used to measure the prediction accuracy. To calculate the PLCC and RMSE, a non-linear regression is required to relate objective scores and subjective ones. Based on the suggestion of [68], we use the following logistic function for the regression

$$y = \eta_1\left(\frac{1}{2} - \frac{1}{1+e^{\eta_2(x-\eta_3)}}\right) + \eta_4 x + \eta_5, \qquad (12)$$

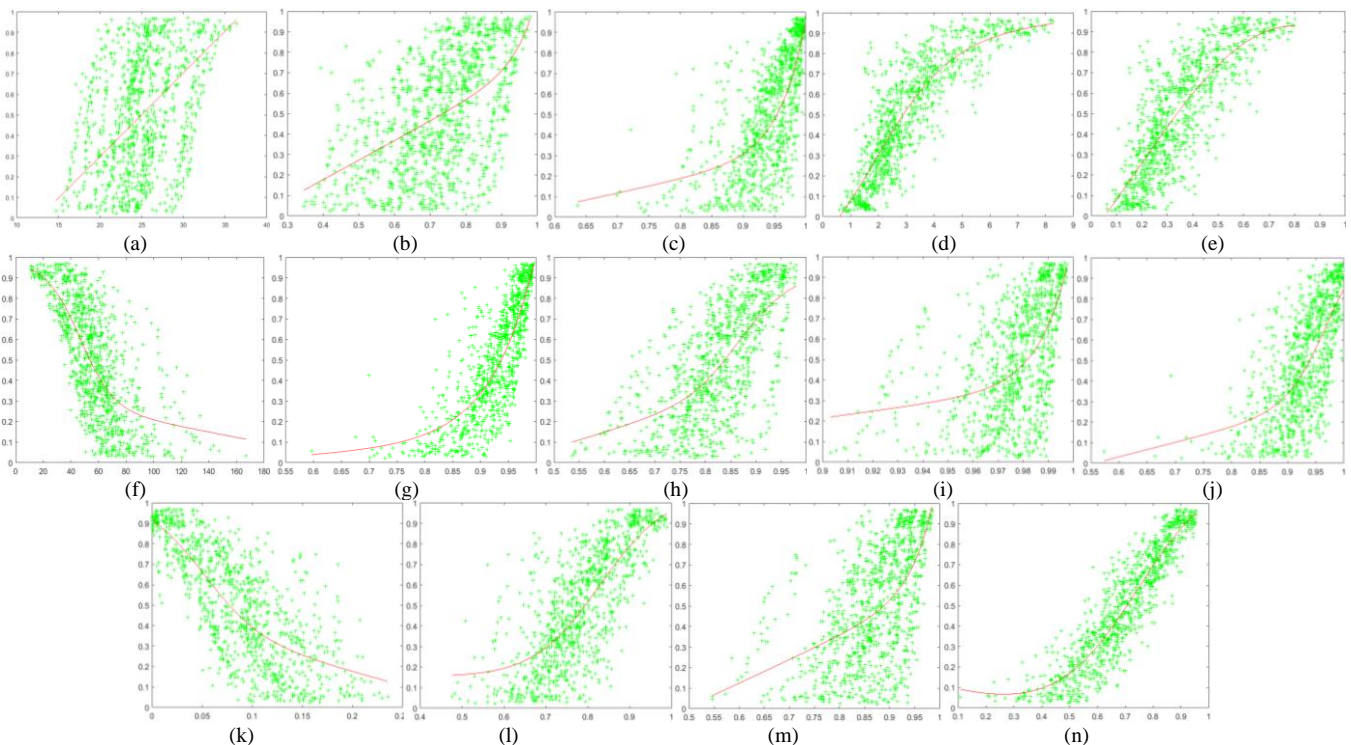where x denotes the output of IQA methods, y represents the

Fig. 9. Scatter plots of FR IQA methods. (a) PSNR, (b) SSIM, (c) MS-SSIM, (d) IFC, (e) VIF, (f) MAD, (g) IW-SSIM, (h) FSIM, (i) GSIM, (j) IGM, (k) GMSD, (l) DASM, (m) SPSIM, (n) SIS. In all the sub-figures, the abscissa is the predicted scores of IQA methods, and the ordinate is the MOS. The points marked by the green "+" represent the images in QADS, and the red curves are fitted with the logistic function in (12).

TABLE IV
SROCC VALUES OF DIFFERENT METHODS ON INDIVIDUAL SR CATEGORY

SR Category	IQA Methods													
	PSNR	SSIM [27]	MS-SSIM [45]	IFC [49]	VIF [50]	MAD [42]	IW-SSIM [47]	FSIM [51]	GSIM [52]	IGM [55]	GMSD [53]	DASM [58]	SPSIM [59]	SIS
(I)	0.2849	0.5050	0.7239	0.8662	0.8177	0.6125	0.8357	0.6805	0.5068	0.6858	0.7432	0.7315	0.5677	**0.8936**
(II)	0.3807	0.5338	0.7306	0.9000	0.8368	0.7445	0.8340	0.6570	0.5186	0.7399	0.7737	0.7894	0.5331	**0.9202**
(III)	0.2655	0.5120	0.6972	0.7792	0.7281	0.6397	0.7229	0.6637	0.5661	0.6625	0.7168	0.7065	0.5870	**0.8552**
Overall	0.3544	0.5290	0.7171	0.8609	0.8152	0.7234	0.8195	0.6885	0.5538	0.7145	0.7649	0.7512	0.5751	**0.9232**

regression values of x, and $\eta_1-\eta_5$ are the parameters to be fitted. After the non-linear regression, the values of PLCC and RMSE can be calculated by using y and MOS, rather than x and MOS. A method is a good one if it has large SROCC, KROCC, and PLCC, as well as a small RMSE.

A. Performance Testing

The experimental results of different methods on QADS are given in Table III, where the best result for each criterion is highlighted in boldface. From Table III, we can see that, for the SRIs, the compared methods do not correlate well with subjective perception. To name just a few, the performance of PSNR for the SRIs is rather poor. SSIM also performs poorly, although it is better than PSNR. It means that the widely-used evaluation criteria in the research of SR are not appropriate. The performance of the methods based on image gradients, e.g., FSIM, GSIM, and GSMD, is also unsatisfactory, although image gradients are important in SR. Among the compared methods, IFC has the best performance. This observation is consistent with the experiments in [3], where it can be found that IFC has the highest correlation with the perceptual scores in the context of SR evaluation. It is worth to notice that this

consistency can also support the reliability of our QADS. Nevertheless, the performance of IFC is far from satisfactory as well. Their poor performance on QADS may be caused by not including the texture. In visual comparison of SRIs, the textural details should not be ignored. Furthermore, some possible artifacts in the SRIs are not taken into consideration in the compared methods. From Table III, it can be observed that SIS significantly outperforms all the competitors.

In Fig. 9, the scatter plots of subjective scores and objective predictions by the above methods are provided. The points obtained by SIS distribute more tightly along the fitted curve, in contrast to the compared methods. This also demonstrates the superiority of SIS.

B. Group-wise Analysis

In the IQA research, it is meaningful to perform comparisons on individual distortion types [47], [51]-[56]. However, for SRIs, multiple types of artifacts may simultaneously appear in one image. For example, some SRIs suffer from jaggies as well as blurring. Thus, it is difficult to examine the behaviors of IQA methods on each type of SR artifacts. Instead, we group the 21 methods into three categories and perform a group-wise anal-

Fig. 10. Results of statistical significance tests based on (a) *F*-test, and (b) the Pitman test.

ysis on each kind of SR methods. In accordance with Table II, the three categories are named as (I) interpolation-based SR, (II) dictionary-based SR, and (III) DNN-based SR. Generally, the methods in one SR category share some common characteristics in their artifacts. For instance, the checkerboard artifacts would only appear in DNN-based SR methods, and cannot be found in interpolation-based or dictionary-based methods. The group-wise analysis based on SROCC are provided in Table IV. The quantitative results show that SIS still outperforms the compared IQA methods on every SR category. Besides, from Table IV, it can be observed that, for most IQA methods, the performance on the dictionary-based SR is better than that on the DNN-based SR. This indicates the difficulty in assessing the visual quality of DNN-based SR results. Specifically, some competitors, e.g., IFC, achieve comparable results with SIS for dictionary-based SRIs. However, for DNN-based SRIs, the advantage of the proposed method is obvious in comparison with other IQA methods.

C. Statistical Significance

The testing of statistical significance is recommended in [61] to determine whether one method is statistically distinguishable from another one. In IQA research, the *F*-test is commonly performed for this purpose, e.g.,[53], [56], and [59]. Recently, the Pitman test is suggested in [76] to relax the assumption of the independence between residuals. In this work, we conduct both the hypothesis tests to show the statistical significance of the above IQA methods on QADS. The results are given in Fig. 10, where competitors are sorted according to their performance. An element of the array in Fig. 10 is filled with "1", if the method in its row statistically surpasses the one in its column. Otherwise, the element is filled with "0". From Fig. 10, it can be observed that, based on both the *F*-test and the Pitman test, SIS is significantly better than all the competitors.

D. Performance of Individual Similarities

In this sub-section, we investigate the performance of each similarity of SIS, i.e., the textural similarity p_t, structural similarity p_s, and high-frequency similarity p_h. The quantitative results are shown in Table V. It can be observed that either the textural or high-frequency similarity has already achieved similar performance as IFC. It is unsurprising that the structural

TABLE V
COMPARISONS ON INDIVIDUAL SIMILARITIES

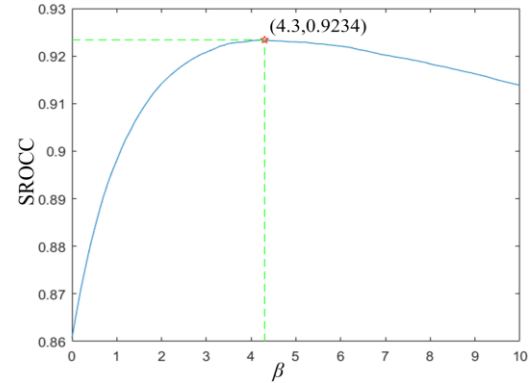| Criteria	p_t	p_s	p_h	p
SROCC	0.8608	0.5612	0.8335	0.9232
KROCC	0.6741	0.3891	0.6403	0.7541
PLCC	0.8627	0.5603	0.8368	0.9230
RMSE	0.1389	0.2274	0.1504	0.1057

Fig. 11. Performance of SIS in term of SROCC versus the parameter β in (10).

TABLE VI
RESULTS ON NON-DECOMPOSED SRIS

Criteria	p_t'	p_s'	p_h'	p'
SROCC	0.8528	0.5365	0.7156	0.6167
KROCC	0.6683	0.4554	0.5229	0.4370
PLCC	0.8507	0.5387	0.6904	0.6205
RMSE	0.1444	0.2314	0.1988	0.2154

similarity alone does not perform well, since only the directional artifacts are measured in p_s. However, the three similarities describe SRIs from the different views of visual defects, which are complementary. Actually, the integrated performance is much better than the individual performance, demonstrating the benefit to incorporate the three similarities.

It is interesting to note that the textural similarity has the best performance among the individual similarities, but it has the lowest impact, i.e., $\alpha < \beta$, on the final result. An intuitive idea is to emphasize the impact of the textural similarity, i.e., reduce the value of β, due to its relatively good performance. Here, a series of β, ranged from 0 to 10 with a step of 0.1, is adopted to test SIS. The quantitative results in term of SROCC are illustrated in Fig. 11. If $\beta = 0$, only the textural similarity is used and the value of SROCC is 0.8608. As β increases, the impact of p_s and p_h becomes greater and the SROCC value also increases. The maximum value of SROCC is 0.9234 when $\beta = 4.3$ that is larger than $\alpha = 1$. The change of SROCC is slow near the maximum value. And the default value of β, which is estimated by (11) using the reference images in LIVE, is also in the area with slow change. As β continues to increase, the SROCC gradually reduces. Hence, the following two conclusions can be drawn. On the one hand, for SRIs, the measures on the image structure still play a more important role in visual perception. On the other hand, the measure on image texture is also essential to assess the visual quality of SRIs.

E. Benefit of STD

To demonstrate the advantage of using STD in SIS, we further directly perform each similarity on non-decomposed SRIs.

The quantitative results are provided in Table VI, where p_t', p_s', and p_h' represent the texture similarity, structure similarity, and high-frequency similarity on the original distorted images. The final single score without STD is denoted as p' in Table V. By comparing the results in Table V (using STD) and those in Table VI (without STD), it can be found that our method greatly benefits from the STD. Without STD, the performance becomes poorer. This result is easy to understand. The behaviors of HVS on structural and textural components are totally different. Measuring the same similarity on different components cannot well capture their respective properties. Taking textual similarity for example, p_t has better performance than p_t', although non-decomposed images are more informative than their textual components. More importantly, without STD, the complementarity among the three similarities is undermined. Consequently, the fused score p' is with much worse performance.

Therefore, the usage of STD is beneficial and important.

V. CONCLUSION AND FUTURE WORK

In this paper, we focus on FR IQA for SRIs. A benchmark database, named as QADS, with 980 SRIs and their MOS is presented. The database contains commonly encountered artifacts in SRIs. 21 SR methods and 100 subjects are involved in building QADS. The suitability of the selection of reference images is demonstrated by the SI index, while the suitability and reliability of subjective scores are also confirmed. Besides, an FR IQA method based on the STD is proposed for SRIs. Different measurements are employed for textural and structural components, since some artifacts appear in the textural part while others exist in the structural component. Furthermore, the perceptual properties of the textural and structural degenerations are distinctive. For textural components, we measure the similarity of feature distributions. For structural components, the similarities of the dominant directions and high-frequency energy are taken into consideration. To obtain the final visual quality score, three individual similarity maps are first pooled to three scores, which are then fused into one.

In the future, we would like to extend this work in the following two aspects. The first is to increase the size of the database. More reference images and SRIs can be included. For reference images, in addition to nature images, we can also add some synthetic ones that are sensitive to the changes of spatial resolution, e.g., the EIA-1956 Resolution Chart. For SRIs, results from multi-frame SR can be involved. The multi-frame SR would introduce new artifacts, such as the artifacts caused by inexact registrations. The second is to develop IQA methods without any reference images. In many real applications, the reference image, i.e., HR ground truth, is not available, while the LR input is known definitely. Since the LR input only contains reduced information of the ground truth instead of full information, it can be regarded as reduced-reference IQA.

REFERENCES

[1] G. Qiu, "A progressively predictive image pyramid for efficient lossless coding," *IEEE Trans. Image Process.*, vol. 8, no.1, pp. 109 – 115, 1999.

[2] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technique overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.

[3] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 372–386.

[4] J. D. van Ouwerkerk, "Image super-resolution survey," *Image Vision Comput.*, vol. 24, no. 10, pp. 1039–1052, Oct. 2006.

[5] E. Meijering, "A chronology of interpolation: From ancient astronomy to modern signal and image processing," *Proc. IEEE*, vol. 90, no. 3, pp. 319–342, Mar. 2002.

[6] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.

[7] H. S. Hou and H. C. Andrews, "Cubic splines for image interpolation and digital filtering," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 6, pp. 508–517, Dec. 1978.

[8] S. E. Reichenbach and F. Geng, "Two-dimensional cubic convolution," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 857–865, Aug. 2003.

[9] Q. Wang and R. K. Ward, "A new orientation-adaptive interpolation method," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 889–900, 2007.

[10] A. Giachetti and N. Asuni, "Real-time artifact-free image upscaling," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2760–2768, Oct. 2011.

[11] Q. Shan, Z. Li, J. Jia, and C.-K. Tang, "Fast image/video upsampling," *ACM Trans. Graph.*, vol. 27, no. 5, pp.153:1-153:7, Dec. 2008.

[12] G. Qiu, "Inter-resolution look-up table for improved spatial magnification of images," *J. Vis. Commun. Image Represent.*, vol. 11, no. 4, pp. 360-373, 2000.

[13] W. T. Freeman, T. R. Jones, and E. Z. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, 2002.

[14] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2004, pp. I-275–I-282.

[15] K. S. Ni and T. Q. Nguyen, "Image superresolution using support vector regression," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1596–1610, Jun. 2007.

[16] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Patt. Anal. Mach.Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[17] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[18] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.

[19] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2216-2223.

[20] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2569–2582, Jun. 2014.

[21] F. Zhou, T. Yuan, W. Yang, and Q. Liao, "Single-image super-resolution based on compact KPCA coding and kernel regression," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 336–340. , Mar. 2015.

[22] W. Yang, Y. Tian, F. Zhou, Q. Liao, and H Chen, "Consistent coding scheme for single-image super-resolution via independent dictionaries," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 313–325, Mar. 2016.

[23] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[24] R. Timofte, V. De, and L. V. Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. IEEE Asian Conf. Comput. Vis.*, 2014, pp. 111–126.

[25] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 561–568.

[26] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5197–5206.

[27] Z.Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2898638, IEEE Transactions on Image Processing

13

assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[28] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1059–1066.

[29] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal.Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[30] D. Liu, Z. Wang, J. Yang, W. Han, and T. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3194–3207, Jul. 2016.

[31] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.

[32] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1646-1654.

[33] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep CNN with skip connection and network in network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 217–225.

[34] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network", in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3147–3155.

[35] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2016, pp. 550–558.

[36] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 624–632.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[38] J. Justin, A. Alexandre, and F.-F. Li "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[39] C. Ledig, *et al*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 105 – 114.

[40] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, San Rafael, CA, USA: Morgan and Claypool, 2006, p. 156.

[41] 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.

[42] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006-1–011006-21, 2010.

[43] R. Snowden, R. J. Snowden, P. Thompson and T. Troscianko, *Basic vision: An introduction to visual perception*. Oxford, UK: Oxford University Press, 2012, p. 424.

[44] D. M. Chandler, "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Process.*, vol. 2013, no. 8, 2013, Art. no. 905685.

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

[46] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2385–2401, Nov. 2009.

[47] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[48] R. Zhu, F. Zhou, and J.-H. Xue, "MvSSIM: A quality assessment index for hyperspectral images," *Neurocomputing*, vol. 272, pp. 250–257, 2017.

[49] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[50] 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.

[51] L. Zhang, D. 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.

[52] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[53] 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.

[54] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assesssment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Mar. 2014.

[55] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 43–54, Jan. 2013.

[56] H. Wang, *et al.*, "Image quality assessment based on local linear information and distortion specific compensation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 915–926, Feb. 2017.

[57] S.-H. Bae and M. Kim, "A novel image quality assessment with globally and locally consilient visual quality perception," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2392–2406, May 2016.

[58] L. Ding, H. Huang, and Y. Zang, "Image quality assessment using directional anisotropy structure measurement," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1799–1809, Apr. 2017.

[59] W. Sun, Q. Liao, J.-H. Xue, and F. Zhou, "SPSIM: A superpixel-based similarity index for full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4232–4244, Sep. 2018.

[60] A. Ninassi, P. Le Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," in *Proc. SPIE: Human Vis. Electron. Imag.*, San Jose, CA, Jan. 2006, vol. 6057.

[61] 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.

[62] Y. Horita, K. Shibata, Y. Kawayoke, and Z. M. P. Sazzad, "MICT image quality evaluation database," 2000. [Online]. Available: http://mict.eng. toyama.ac.jp/mict/index2.html

[63] N. Ponomarenko, *et al.*, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Mod. Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2004.

[64] N. Ponomarenko, *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *Proc. Eur. Workshop Vis. Inf. Process.*, 2013, pp. 106–111.

[65] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, Jan. 2017.

[66] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-scale crowdsourced study for tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4725–4740, Oct. 2017.

[67] F. Zhou *et al.*, "Quality assessment database for super-resolved images: QADS", 2018. [Online]. Available: http://www.vista.ac.cn/projects/vista-h/super-resolution/

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

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

[70] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://doi.org/10.23915/distill.00003

[71] J.-F. Aujol, G. Gilboa, T. Chan, S. Osher, "Structure-texture image decomposition—Modeling, algorithms, and parameter selection," *Int. J. Comput. Vis.*, vol. 67, no. 1, pp. 111–136, 2006.

[72] J. Gilles and Y. Meyer, "Properties of BV − G structures + textures decomposition models. Application to road detection in satellite images," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2793–2800, Nov. 2010.

[73] S. Liu, F. Zhou, and Q. Liao, "Defocus map estimation from a single image," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5943–5956, Dec. 2016.

[74] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[75] T. Aach, *et al.*, "Analysis of superimposed oriented patterns," *IEEE Trans.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2898638, IEEE Transactions on Image Processing

14

Image Process., vol. 15, no. 12, pp. 3690–3700, Dec. 2006.

[76] R. Zhu, F. Zhou, W.Yang, and J.-H. Xue, "On hypothesis testing for comparing image quality assessment metrics," *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 133–136, Jul. 2018.

Fei Zhou received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, in 2007, and the Ph.D. degree in electronic engineering, Tsinghua University, in 2013. From 2013 to 2016, he was a post-doctoral fellow with the Graduate School at Shenzhen, Tsinghua University. From 2017 to 2018, he was a Visiting Scholar with the Department of Statistical Science, University College London. He is currently an Assistant Professor with the College of Information Engineering, Shenzhen University. He has authored over 40 papers internationally. His research interests include image super-resolution, image interpolation, and image quality assessment. He is a reviewer of many well-known journals, including IEEE Transactions on Image processing, IEEE Signal Processing Letters, Information Sciences, etc. He also serves as a guest editor of Neurocomputing and Signal Processing: Image Communication.

Rongguo Yao was born in Shaoyang city, Hunan province, China, in 1995. He is now a Master in the College of Information Engineering, Shenzhen University, China. His researches focus on the topic of image quality assessment and image saliency detection.

Bozhi Liu received the B.S. degree in 2008 from Nanjing University of Science and Technology, China, and the Master and PhD degrees in 2010 and 2018 from the University of Nottingham, UK. He is currently a Research Fellow with the College of Information Engineering, Shenzhen University, China. His research interests include color constancy, image processing, and video enhancement.

Guoping Qiu is a Distinguished Professor of Information Engineering, Director of Shenzhen University Intelligent Robotics Centre at Shenzhen University, China, and a Chair Professor of Visual Information Processing at the University of Nottingham, Nottingham, UK. He has taught in universities in the UK and Hong Kong and also consulted for multinational companies in Europe, Hong Kong and China. His research interests include image processing, pattern recognition, and machine learning. He is particularly known for his pioneering research in high dynamic range imaging and machine learning based image processing technologies. He has published widely and holds several European and US patents. Technologies developed in his lab have laid the cornerstone for successful spinout companies who are developing advanced digital photography software enjoyed by tens of millions of global users.