SPSIM: A SuperPixel-based SIMilarity Index for Full-reference Image Quality Assessment

Wen Sun, Qingmin Liao, Jing-Hao Xue, and Fei Zhou

***Abstract*—Full-reference image quality assessment algorithms usually perform comparisons of features extracted from square patches. These patches do not have any visual meanings. On the contrary, a superpixel is a set of image pixels that share similar visual characteristics and is thus perceptually meaningful. Features from superpixels may improve the performance of image quality assessment. Inspired by this, we propose a new superpixel-based similarity index (SPSIM) by extracting perceptually meaningful features and revising similarity measures. The proposed method evaluates image quality on the basis of three measurements, namely, superpixel luminance similarity, superpixel chrominance similarity, and pixel gradient similarity. The first two measurements assess the overall visual impression on local images. The third measurement quantifies structural variations. The impact of superpixel-based regional gradient consistency on image quality is also analyzed. Distorted images showing high regional gradient consistency with the corresponding reference images are visually appreciated. Therefore, the three measurements are further revised by incorporating regional gradient consistency into their computations. A weighting function that indicates superpixel-based texture complexity is utilized in the pooling stage to obtain the final quality score. Experiments on several benchmark databases demonstrate that the proposed method is competitive with state-of-the-art metrics.**

***Index Terms*—Full-reference, image quality assessment, superpixel, regional gradient consistency, texture complexity.**

I. Introduction

IMAGE quality assessment (IQA) is widely used as a benchmark in numerous image processing tasks, such as image super-resolution [1], image compression [2], and image enhancement [3]. Subjective assessment by humans is the most accurate IQA metric because images are finally presented to human beings. However, subjective assessment is inapplicable to practical tasks because it is laborious. Objective assessment is more practical than subjective assessment in this case because the quality of an image is automatically predicted by machines. Objective IQA metrics can be categorized into full-reference (FR), reduced-reference (RR), and no-reference (NR) [4]. In FR methods, the information of the reference image is completely available, whereas NR methods do not require a reference image. RR metrics are between them and the reference image is partially accessible. This study focuses on FR IQA algorithms.

Early FR IQA methods, such as peak signal to noise ratio (PSNR) and mean squared error (MSE), evaluate image quality based on intensity differences between reference and distorted images. In these two methods, only a numerical comparison is performed while the visual mechanism of humans is ignored. To solve this problem, scholars have proposed many metrics for incorporating the characteristics of the human visual system (HVS). Visual signal to noise ratio (VSNR) exploits near-threshold and supra-threshold properties of human vision to measure image fidelity [5]. In the metric called most apparent distortion (MAD), distortion visibility is calculated, and different strategies are adopted for near-threshold and clearly visible distortions [6]. Visual information fidelity (VIF) [7] predicts image quality by using shared information between reference and distorted images. Current research on HVS is limited, and only part of its characteristics has been modeled and utilized [8].

Based on the assumption that human visual perception is highly sensitive to structural information, the structural similarity (SSIM) index is used to assess image quality from three aspects, namely, luminance comparison, contrast comparison, and structure comparison [9]. SSIM is one of the most well-known FR metrics due to its computational efficiency and satisfactory performance. Extended versions of SSIM have been presented. Wavelet-domain structural similarity (WDSSIM) performs structural similarity in the wavelet-domain, and the relative importance of edge information is considered [10]. In information content weighted structural similarity (IW-SSIM) [11], information content is measured and used as a perceptual weight for pooling. A new multivariate SSIM (MvSSIM) index is proposed in [12] to assess the quality of hyperspectral images by considering the pixel spectrum as a multivariate random vector.

Image gradient has been frequently included in IQA algorithms because it conveys important visual information [13]. Given that HVS understands images by low-level features, the feature similarity (FSIM) index uses phase congruency and gradient magnitude as primary features [14]. Moreover, phase congruency functions as a local weight in the final pooling. Liu et al. [13] proposed the gradient similarity (GSIM), in which image gradient is obtained by using four directional filters and compared by considering the masking effect and distortion visibility. In [15], a computationally efficient and highly effective method, namely, gradient magnitude similarity deviation (GMSD), is proposed. The most remarkable innovation of GMSD is a new pooling strategy that exploits the global

This work was partly supported by the National Natural Science Foundation of China under Grant 61771276. (*Corresponding author: Fei Zhou*)

Wen Sun and Qingmin Liao are with the Shenzhen Key Laboratory of Information Science and Technology, Shenzhen Engineering Laboratory of IS&DRM, Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: sunw13@mails.tsinghua.edu.cn; liaoqm@tsinghua.edu.cn).

Jing-Hao Xue is with the Department of Statistical Science, University College London, UK (e-mail: jinghao.xue@ucl.ac.uk).

Fei Zhou is with the College of Information Engineering, Shenzhen University, Shenzhen, China (e-mail: flying.zhou@163.com).

variation of gradient similarity to characterize image quality. The visual saliency-induced index (VSI) incorporates image saliency analysis into IQA, where visual saliency similarity and gradient similarity on each pixel are calculated and pooled by using visual saliency as local weights [16].

Conventional IQA metrics focus on the intensity parts of images because HVS perceives intensity changes more easily than chromatic variations. However, chromatic variations are also important and should be considered appropriately [17]. Several methods have been designed recently to measure chromatic variations. Directional statistics are used in [18], where color descriptors from three color channels, namely, hue, chroma, and lightness are extracted and compared. In [19], a method based on color contrast similarity and color value difference (CSVD) was developed to evaluate the quality of color correction images. CSVD calculates color contrast, average-based color value difference in the CMYK color space, and span-based color value difference in the HSV color space. The final score is a weighted sum of the scores of each part.

Although much progress has been achieved in FR IQA, several problems still exist. First, the features used in existing methods are generally extracted from square image patches. These patches do not have visual meanings, and thus the resulting features may not be optimal. Second, in many FR models, the quality of a given pixel is determined by the change of features on that pixel between the reference and distorted images, whereas the overall change of features in a small region is ignored. Image pixels are only meaningful when gathered as image regions, indicating that regional quality assessment should be performed. Finally, in most traditional FR methods, a large difference of local features indicates poor local quality. However, this is not always true for commonly used features. For example, the quality of contrast-enhanced images may still be acceptable, despite evident differences detected using common features [8].

To solve these problems, this study proposes a new method, namely, superpixel-based similarity index (SPSIM), to accurately predict image quality. In this method, images are segmented into visually meaningful regions, namely, superpixels. Then, the mean values of the intensity and chrominance components are extracted within each superpixel and compared to describe local characteristics precisely. This procedure is proposed to address the first problem. In addition to the two similarity measures above, gradient similarity is employed to improve the performance on structural variations. Furthermore, in each superpixel, the regional consistency of gradient magnitudes between reference and distorted images is measured. This measure focuses on the overall changes of all gradients in one superpixel and is used to improve the accuracy of the three similarities. This process aims to solve the second and third problems. Finally, texture complexity is utilized as local weights to pool the pixel-wise similarity map into a single score. The main contributions of our work can be summarized briefly as follows: 1) we use superpixels, which is perceptually more meaningful and accurate, to extract features and reflect image characteristics; 2) the regional overall variations in features are considered and utilized to revise feature similarity measures. Experiments on four databases demonstrate that SPSIM is superior to most existing methods in predicting image quality.

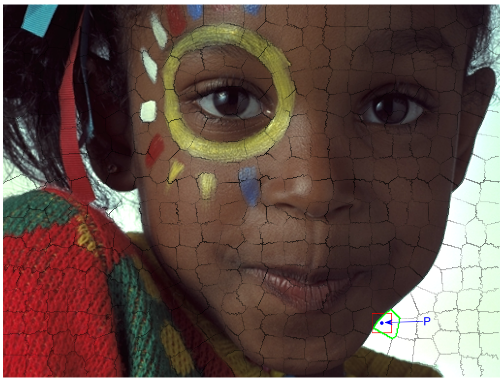


Fig. 1. Illustration of the SLIC superpixel segmentation

The remainder of this paper is organized as follows. The feature extraction and similarity measures are explained in Section II. Section III describes the proposed IQA index. In Section IV, the experimental results are presented and analyzed. Finally, Section V concludes this work.

II. Proposed Feature Extraction and Similarity Measures

In this section, we analyze the extraction of features and the similarity measures used in our IQA method. Many features, such as mean, standard deviation, and covariance in SSIM, linear correlation coefficient in the local linear model (LLM) [20], and color component difference in [19], have been utilized in IQA. A common point among them is that almost all features are extracted from a square image patch centered at a given pixel. These patches are convenient for computation but are usually meaningless for visual perceptions. We believe that extracting features from visually meaningful regions may improve the performance of IQA algorithms. Superpixels may play an important role in this regard.

A. *Superpixel versus Image Patch*

A superpixel is a perceptually meaningful region comprised of spatial neighboring pixels. These pixels usually share many common characteristics, such as similar colors, intensities or structures, aside from spatial adjacency. These points make superpixels a convenient and effective tool to compute image features in image processing tasks [21]. In [22], superpixel segmentation is used to compress images more efficiently than traditional techniques. Liu et al. proposed to calculate the inter-superpixel similarity, global contrast, and spatial sparsity to generate a superpixel-level saliency map [23]. Superpixels are also applicable to image decomposition [24], multisensory video fusion [25], and image synthesis [26]. Many superpixel algorithms have been presented. In [27], superpixels are generated by using a geodesic distance, which produces small superpixels in structure-dense regions and large ones in structure-sparse regions. Simple linear iterative clustering

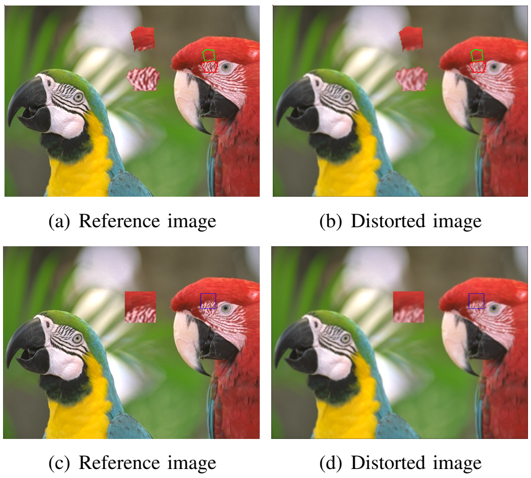


Fig. 2. Comparison of superpixel (the upper row) and image patch (the lower row) in separating regions with different sensitivities to image blur. The distorted image is obtained by degrading the reference image with Gaussian blur. It can be observed that the superpixel well separates textured regions (sensitive to Gaussian blur) and flat regions (insensitive to Gaussian blur) while the image patch contains both textured and flat areas.

(SLIC) [21] firstly initializes a number of cluster centers and each pixel is assigned to its nearest center according to a predefined distance measure. Then each cluster center is updated by the mean attributes of its corresponding elements. Superpixels are produced by repeating these steps. Giraud et al. proposed a fast method to compute superpixels by considering the linear path to the superpixel barycenter in designing distance measures [28]. In our work, we adopt the SLIC method, which is computationally efficient and shows leading adherence to image boundaries. Moreover, SLIC can be easily implemented by simply setting the number of cluster centers (Nc). We present an example of SLIC segmentation in Fig. 1, where $Nc = 400$.

Extracting features from superpixels is beneficial. Since image pixels in a superpixel are similar to each other in colors and intensities, obtaining the low-level features such as mean luminance is more accurate. We take the luminance comparison of SSIM as an example. In SSIM, the luminance of pixel P is calculated by the mean intensity of the pixels inside the red square, as illustrated in Fig. 1. However, in superpixels, luminance computation is performed on the pixels encircled by the green line. The mathematical expressions of these two methods are as follows:

$$L_P = \frac{1}{|C_r|} \sum_{j \in C_r} Intensity(j) \tag{1}$$

$$L_P = \frac{1}{|C_g|} \sum_{j \in C_g} Intensity(j) \tag{2}$$

where C_r is the set of pixels in the red square, C_g is the set of pixels inside the green line, $|C_r|$ denotes the number of pixels in C_r, and $|C_g|$ stands for the number of pixels in C_g. Apparently, the second equation is more precise in describing pixel luminance.

It is also convenient to analyze the peculiarities of image areas with superpixels. In FR IQA, two peculiarities of HVS are usually highlighted. First, textured areas are sensitive to image blur but insensitive to Gaussian noise. Second, Gaussian noise in flat areas is easy to be perceived, whereas image blur is not. Based on above two observations, various strategies have been employed to predict image patch quality [29-31]. For this type of FR methods, superpixels are superior to square patches because they can separate image regions of different styles with higher boundary consistency. Examples of textured and flat regions using superpixels are presented in Figs. 2(a) and 2(b), where the region encircled by the red line is a textured area and the region encircled by the green line is a flat part. This segmentation effectively separates two regions with different sensitivities to image blur. For comparison, the patch-based results are also provided in Figs. 2(c) and 2(d). The figures show that a square patch may contain both textured and flat areas, which are common especially when the patch is close to their boundaries. In this case, it is difficult to evaluate the quality degradation. These examples verify the good performance of superpixels in describing image peculiarities. For this reason, the regional gradient consistency and weights in Section III are computed with superpixels.

B. Superpixel Segmentation of Reference and Distorted Images

As introduced in Section I, FR IQA predicts the quality of a distorted image with the reference image being available. For these two images, superpixel generation can be performed in three modes: segmenting the reference image and applying this segmentation to the distorted image, segmenting the distorted image and applying this segmentation to the reference image, and segmenting the reference and distorted images separately. In our work, we select the first mode, i.e., the distorted image is segmented directly following the reference image. The reason for this selection is that reference images are high-quality images with invisible distortions. Their segmentations are consistent with visual perceptions. On the contrary, distorted images are degraded in various modes; hence, existing superpixel algorithms (e.g., SLIC) cannot obtain a widely accepted segmentation result, especially when these images are seriously distorted by noise [32]. Therefore, we segment the distorted image similarly to that of the reference image.

It is necessary to segment images into a reasonable number of superpixels. A number of superpixels larger than 200 is generally sufficient for edge preservation [23]. Excessive superpixels would lead to a high computational cost. Therefore, we set $Nc = 400$, in which the number of generated superpixels is between 250 and 400.

C. Superpixel Similarity: Luminance and Chrominance

Image luminance represents the brightness perceived by HVS, and it is an important feature in predicting image quality. Color, which is ignored in many conventional metrics, also influences human perception about image quality and has been increasingly emphasized in recent research [18] [19]. In this section, we present the luminance similarity and the chrominance similarity between the reference image **r** and the distorted image **d** from the viewpoint of superpixels.

The process is as follows. Images **r** and **d** are firstly segmented into many superpixels as described in Section II. B. The intensity and chromatic components are then derived by the YUV composition [33]. Using the Y component, the luminance of the i-th pixel is estimated by the mean intensity as follows:

$$L_i = \frac{1}{|s_i|} \sum_{j \in s_i} Y(j)$$

where s_i is the superpixel that encloses the i-th pixel and $|s_i|$ is the number of elements in s_i. Then, we can compute the pixel-wise luminance similarity as follows:

$$M_L(i) = \frac{2L_r(i)L_d(i) + T_1}{L_r^2(i) + L_d^2(i) + T_1} \tag{3}$$

where $L_r(i)$ and $L_d(i)$ represent the luminance of the i-th pixel in **r** and **d**, respectively, and T_1 is a positive variable to avoid instability when $L_r^2(i) + L_d^2(i)$ is extremely small. Similarly, we can derive $M_U(i)$ and $M_V(i)$. The chrominance similarity is the product of $M_U(i)$ and $M_V(i)$ as follows:

$$M_C(i) = M_U(i)M_V(i) \tag{4}$$

D. Pixel Similarity: Gradient

Luminance similarity and chrominance similarity can appropriately characterize low-level features. In other words, they measure the overall impression when an image is perceived by humans. As shown in Fig. 1, a superpixel is usually a homogeneous area and structures or variations are widely distributed in the boundaries of superpixels. The similarity measures in Section II. C are powerless to reflect the impact of structures. Gradient similarity can overcome this shortcoming. Image gradient is calculated by differentiating pixel intensities, and thus it can appropriately describe local structural changes. Image gradient has been employed in many image processing tasks, including FR IQA. For example, gradient serves as a primary feature in FSIM, GSIM, GMSD, and VSI. In our work, we exploit gradient information likewise to effectively measure structural degradations. As suggested in [15], the Prewitt operators are adopted to extract vertical and horizontal image gradients, denoted by $G_v(i)$ and $G_h(i)$, where i stands for the i-th pixel. Then, gradient magnitude is calculated as $G(i) = \sqrt{G_h^2(i) + G_v^2(i)}$.

Gradient similarity is defined as the similarity of gradient magnitudes on each pixel between **r** and **d** as follows:

$$M_G(i) = \frac{2G_r(i)G_d(i) + T_2}{G_r^2(i) + G_d^2(i) + T_2} \tag{5}$$

where $G_r(i)$ and $G_d(i)$ represent the gradient magnitudes of the i-th pixel in **r** and **d**, respectively. The role of T_2 is similar to that of T_1. It is worthwhile to notice that the values of T_1 and T_2 greatly influence FR IQA. The selection of T_1 and T_2 will be discussed in the next section.

III. Proposed IQA Metric

In this section, we explore the impacts of regional gradient consistency (RGC) on quality assessment and revise the similarity measures presented in Section II using RGC. Further, a weighted pooling strategy is employed to process pixel-wise similarity into a global quality score. The framework of the proposed IQA metric is illustrated in Fig. 3, where superpixel-based calculations are highlighted in red while pixel-wise operations are in gray. The inputs **r** and **d** are initially partitioned into many superpixels using the SLIC segmentation of **r** and then decomposed into YUV components. With the Y components, gradient magnitude is computed and RGC is measured in each superpixel. Luminance, chrominance, and gradient similarities are calculated subsequently in consideration of RGC. Finally, the integration of these similarities is obtained and a pooling operation is conducted to derive the final quality score.

A. Regional Gradient Consistency

Digital images are composed of large numbers of pixels [34]. A single pixel has no visual meaning, but a group of pixels may present various textures and structures. When an image is perceived by HVS, the information conveyed by these pixels as a whole (namely, an image region) is more crucial than that conveyed by individual pixels. However, in most existing FR IQA models, image quality is commonly predicted by the change of features on each pixel between reference and distorted images, whereas the overall change of features in an image region is ignored. For example, image gradient is the primary feature in FSIM and GSIM, and the quality is obtained by comparing pixel-wise gradients without any consideration for regional gradient comparison. Incorporating regional gradient comparison into aforementioned models may improve their performance, given that image regions as a whole are important for human visual perception. The pooling strategy of GMSD can be considered as a special measure to estimate the global gradient relationship of all pixels (the entire image region). In [15], GMSD is obtained as the standard deviation of gradient similarities on all pixels. The motivation behind it is that if gradient similarities on all pixels are almost the same, then GMSD is very small, which means high quality for the test image. In other words, if image gradient changes in a similar trend, then the predicted quality tends to be good. In this sense, GMSD can be considered as a measure of the global relationship between the gradients of **r** and **d**. The excellent performance of GMSD demonstrates that incorporating the regional feature relationship into an IQA scheme can improve performance. Superpixels are used as image regions in our work.

Since local structures are perceived by comparing the differences of visual signals, similar variations may correspond to similar structures. These variations can be captured by the relationship of relative magnitude (RRM) of visual signals, such as gradient or intensity. Here, RRM indicates the intensity or gradient ranking of all pixels in a perceived image region. With the gradient ranking as an example, two image regions with similar RRMs generally share dominant structured and flat parts analogously, which is of great importance in IQA [35]. Due to the various patterns of image areas, we prefer using the regional gradient consistency to compare the RRMs of two regions rather than using the global relationship in

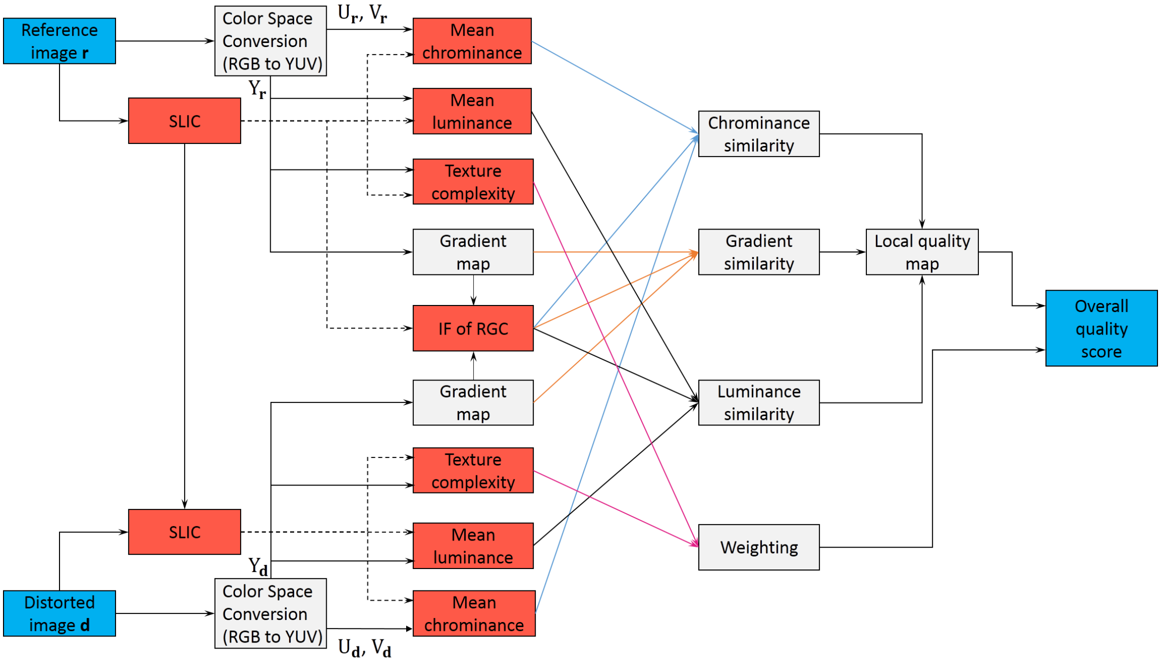


Fig. 3. Framework of the proposed method. Blue boxes indicate inputs and output, red boxes correspond to computations performed in superpixels and the others represent pixel-wise operations.

GMSD. For two image superpixels S'_r and S'_d with gradient maps g_r and g_d, we propose to use the Spearman's rank order correlation coefficient (SROCC) of g_r and g_d to measure RGC as follows:

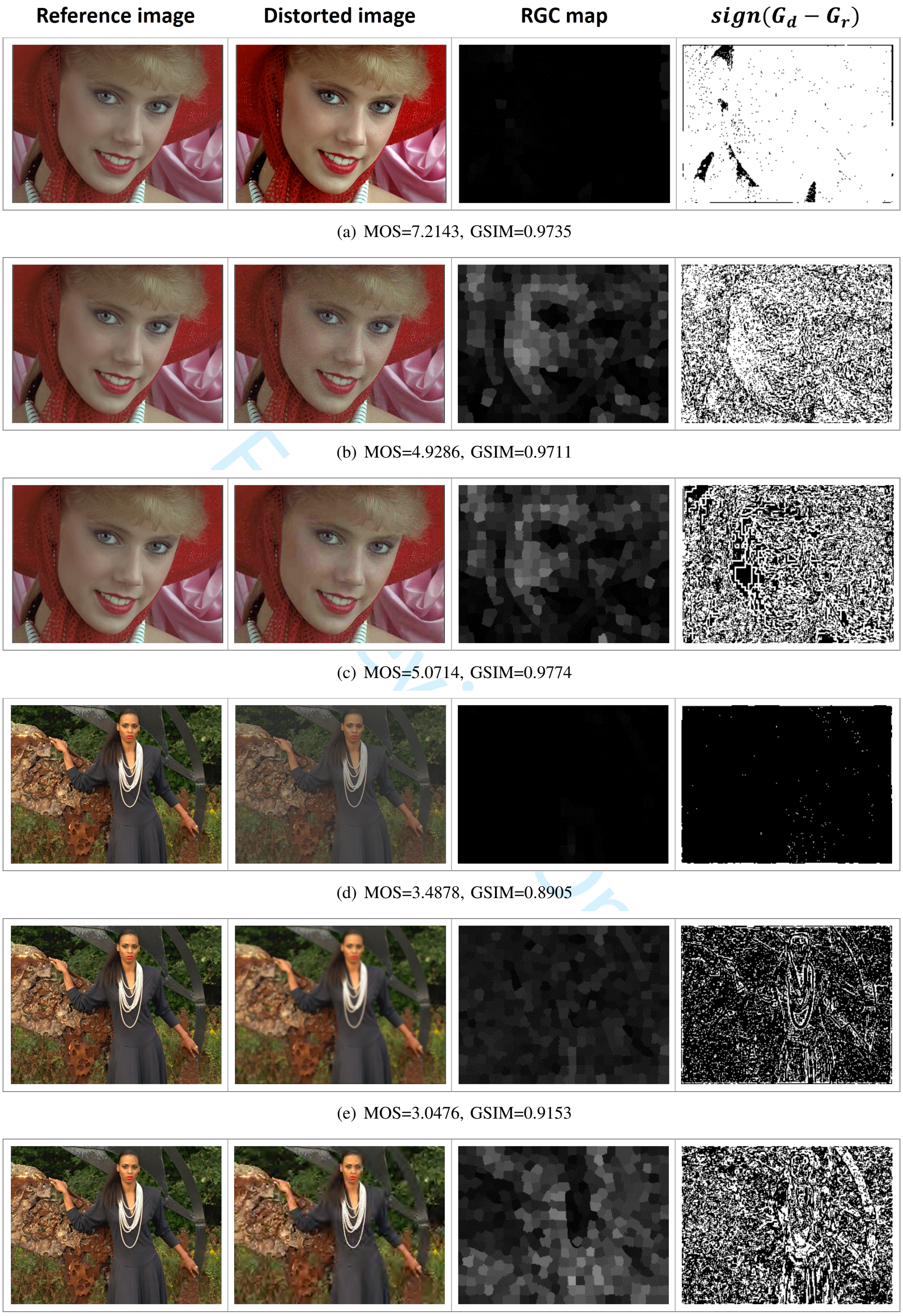
$$\begin{aligned} RGC(S'_r, S'_d) &= SROCC(g_r, g_d) \\ &= 1 - \frac{6\sum_{i=1}^{K} d_i^2}{K(K^2-1)} \end{aligned} \qquad (6)$$

where d_i denotes the difference between the ranks of corresponding gradient pair in g_r and g_d, and K is the number of pixels in the superpixel. SROCC is a widely used standard in IQA to measure the monotonicity of two datasets [36]. It can accurately describe the similarity of RRMs of two gradient maps. We present examples of RGC on images with different distortions in Fig. 4, where the first two columns are the reference and distorted images, the third column shows the RGC maps computed with superpixels, and the last column presents the signs of gradient differences (G_r and G_d denote the gradient maps of **r** and **d**, respectively). The sub-caption provides the mean opinion score (MOS) rated by humans and the predicted score obtained by GSIM [13]. In the column of RGC map, a darker pixel means a higher RGC. In the last column, positive and negative signs are displayed by white and black pixels, respectively. From Fig. 4, we can find that the distorted image in Fig. 4(a) shares similar GSIM with those in Figs. 4(b)-(c), whereas its MOS is much larger than those of the other two. Therefore, the objective quality of the distorted image in Fig. 4(a) is underestimated by GSIM, and a larger MOS difference indicates a higher degree of quality underestimation. A same observation can be obtained from Fig. 4(d) and Figs. 4(e)-(f), and a similar conclusion is achieved for the distorted image in Fig. 4(d). At the same time, it is clear that RGC maps of Figs. 4(a) and 4(d) are mostly covered with dark regions, which indicate high RGCs. These observations and analyses imply that the quality scores of images with high RGCs predicted by GSIM are usually underestimated, i.e., the distortions are overestimated. The reason is that large RGCs generally indicate similar structures from the view of image region, which exert a great impact on visual perception but is ignored in most cases. In addition, the difference of MOS between Fig. 4(a) and Figs. 4(b)-(c) significantly exceeds that between Fig. 4(d) and Figs. 4(e)-(f), which indicates that the distortion overestimation of Fig. 4(a) is more severe than that of Fig. 4(d). This result may be ascribed to the gradient difference signs in the last column, where Fig. 4(a) presents most positive signs (increased gradients) while Fig. 4(d) shows the opposite. Specifically, images with increased gradients and high RGCs are usually enhanced images, which may present good visual impressions but differ evidently from their original versions in terms of features [37]. Therefore, the increase or decrease of gradients (IDG), which can be calculated as

$$IDG(g_r, g_d) = \frac{1}{K}\sum_{i=1}^{K} psgn(g_d(i) - g_r(i)) \qquad (7)$$

where $psgn(x)$ returns 1 when $x \geq 0$ and 0 otherwise, is another important factor that influences quality assessment. If IDG is close to 1, gradients are mostly increased; if IDG is close to -1, gradients are mostly decreased. Other cases do not indicate a strong variation trend. On the basis of this analysis, we can classify distorted image regions into three types:

- type A: large RGC, large IDG, satisfying $RGC \geq \tau_0$,



(a) MOS=7.2143, GSIM=0.9735

(b) MOS=4.9286, GSIM=0.9711

(c) MOS=5.0714, GSIM=0.9774

(d) MOS=3.4878, GSIM=0.8905

(e) MOS=3.0476, GSIM=0.9153

(f) MOS=3.0714, GSIM=0.9102

Fig. 4. Comparison of RGC maps of images distorted by (a) contrast change (increment), (b) Gaussian noise, (c) JPEG, (d) contrast change (decrement), (e) Gaussian blur, and (f) JPEG2000. In RGC maps, a darker region indicates a higher RGC; In the last column, a white pixel denotes the positive sign while a black pixel stands for the negative sign.

$IDG \geq \tau_1$;

- type B: large RGC, small IDG, satisfying $RGC \geq \tau_0$,

$IDG \le -\tau_1$;

- type C: other cases;

For simplicity, we combine RGC in (6) and IDG in (7) to calculate the indicator function (IF) of RGC as:

$$\begin{array}{l} u_0 = psgn(RGC(S'_r, S'_d) - \tau_0) \\ u_1 = psgn(IDG(g_r, g_d) - \tau_1) \\ u_2 = psgn(-\tau_1 - IDG(g_r, g_d)) \\ IF_A(S'_r, S'_d) = u_0 u_1 \\ IF_B(S'_r, S'_d) = u_0 u_2 \end{array} \tag{8}$$

where τ_0 and τ_1 are thresholds, and u_0, u_1, and u_2 are temporary variables. Images constituted mostly by patches from types A ($IF_A = 1$) and B ($IF_B = 1$) are usually overestimated in terms of degradations, especially for type A. Some modifications or compensations should be performed on these images.

With the definition of feature similarity in Section II, we can make T_1 (and similarly T_2) change adaptively with IF_A and IF_B. When T_1 is increased, the degradation indicated by similarity measures declines, which can solve the problem of distortion overestimation. Therefore, T_1 and T_2 can be modified as:

$$\begin{array}{l} T_1(S'_r, S'_d) = C_1 + \lambda_1 IF_A + \lambda_2 IF_B \\ T_2(S'_r, S'_d) = C_2 + \lambda_1 IF_A + \lambda_2 IF_B \end{array} \tag{9}$$

where C_1 and C_2 are positive constants to avoid instability in (3)-(5), and λ_1 and λ_2 are positive constants to avoid the overestimation of distortions in types A and B, respectively. As we have analyzed above, the overestimation of distortions is more severe in type A than in type B. Thus, λ_1 is much larger than λ_2 for a stronger capability to avoid this overestimation.

B. SPSIM Index

The IQA index can be calculated using the similarity measures in Section II and RGC-modified parameters in Section III. A. With the superpixels that enclose the i-th pixel in **r** and **d** denoted by $S_r(i)$ and $S_d(i)$, respectively, we can obtain $T_1(S_r(i), S_d(i))$ and $T_2(S_r(i), S_d(i))$ with (9). Then, $M_L(i)$, $M_C(i)$, and $M_G(i)$ are computed by substituting T_1 and T_2 to (3)-(5). Finally, the overall comparison is expressed as:

$$M(i) = M_G(i)[M_L(i)]^{\alpha} e^{\beta(M_C(i)-1)} \tag{10}$$

where α and β are parameters to adjust the weights of luminance and chrominance similarities. The exponential form helps limit the influence of chromatic components because HVS is more sensitive to achromatic variations than to chromatic variations [18]. Specifically, with $0 < x \le 1$ and $0 < \beta < 1$, we have $0 < x^{\beta} \le e^{\beta(x-1)} \le 1$. Therefore, for the same α and β, $M(i)$ in (10) is less sensitive to chromatic variations $M_C(i)$ because the exponential form makes it closer to 1 than the power form does.

With pixel-wise measurement $M(i)$, the global quality score SPSIM can be calculated as:

$$SPSIM = \frac{\sum_{i=1}^{N} M(i)w(i)}{\sum_{i=1}^{N} w(i)} \tag{11}$$

where N is the number of pixels, and $w(i)$ denotes the weight of the i-th pixel. In our work, we employ the difference of texture complexity (TC) [38] as a local weight, which is a just noticeable difference (JND) index incorporating the contrast sensitivity function (CSF) and the contrast masking (CM) effect. This index is effective in IQA tasks [30]. In [38], TC is defined as the ratio of contrast intensity (CI) to structureness (ST), where CI can be approximated by the standard deviation [39] and ST can be computed as the kurtosis of pixel intensities [40]. To be specific, we can obtain the TC of the i-th pixel in **r** as:

$$\begin{array}{rcl} CI(i) & = & std(S_r(i)) \\ ST(i) & = & kurtosis(S_r(i)) + 3 \\ TC_r(i) & = & \dfrac{CI(i)}{ST(i)} \end{array}$$

where $std(\cdot)$ and $kurtosis(\cdot)$ mean calculating the standard deviation and kurtosis, respectively. Different from [38], the computations of CI, ST, and TC are based on superpixels in this work. When TC is achieved in both the reference and distorted images, $w(i)$ can be given by

$$w(i) = e^{0.05 fabs(TC_d(i) - TC_r(i))} \tag{12}$$

where $TC_d(i)$ stands for the TC of the i-th pixel in **d** and $fabs(\cdot)$ is the absolute operator.

IV. Experimental Results And Analysis

In this section, the proposed IQA metric is tested on four commonly used databases and compared with several well-known FR IQA methods.

A. Evaluation Databases and Criteria

Benchmark databases are necessary to evaluate the performance of IQA methods. In general, Laboratory for Image and Video Engineering (LIVE) [41], Categorical Subjective Image Quality (CSIQ) [6], Tampere Image Database 2008 (TID2008) [42], and Tampere Image Database 2013 (TID2013) [43] are most widely used databases. In LIVE, five types of distortions, namely, JPEG2000 compression, JPEG compression, white noise, Gaussian blur, and fast fading channel bit errors, are introduced to 29 reference images to obtain 779 distorted images. A total of 116 subjects rate the image quality. CSIQ contains 30 reference images and 886 distorted images with six types of distortions, namely, additive white Gaussian noise, additive pink Gaussian noise, contrast decrements, Gaussian blur, JPEG2000, and JPEG. TID2008 is a large database with 25 reference images and 1700 distorted images, and the number of distortion types is 17. TID2013 is an extended version of TID2008, in which 24 distortion types are applied, and the number of distorted images is 3000. TID2013 is currently one of the largest FR IQA databases with the most types of synthetic distortions.

Four criteria calculated between prediction results and human-rated scores, namely, Pearson's linear correlation coefficient (PLCC), root mean squared error (RMSE), SROCC, and Kendall's rank order correlation coefficient (KROCC), are utilized to compare the performance of different IQA metrics [36]. Among them, PLCC and RMSE indicate the prediction accuracy, and SROCC and KROCC show the prediction

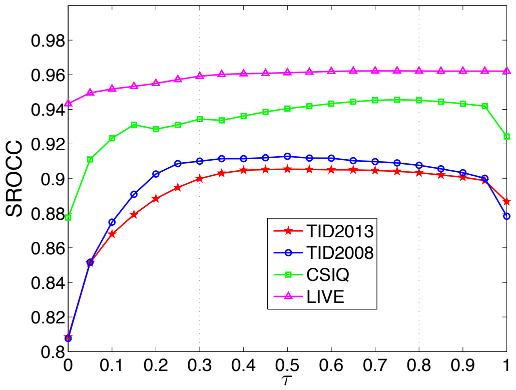


Fig. 5. Performance of SPSIM in terms of SROCC against τ on four databases.

monotonicity. In most cases, the relationship between objective results and subjective scores is nonlinear and a regression between them is necessary to reduce this nonlinearity. In our work, we adopt the logistic function [14] for nonlinear regression as follows:

$$p = \eta_1(\frac{1}{2} - \frac{1}{1 + e^{\eta_2(q-\eta_3)}}) + \eta_4 q + \eta_5$$

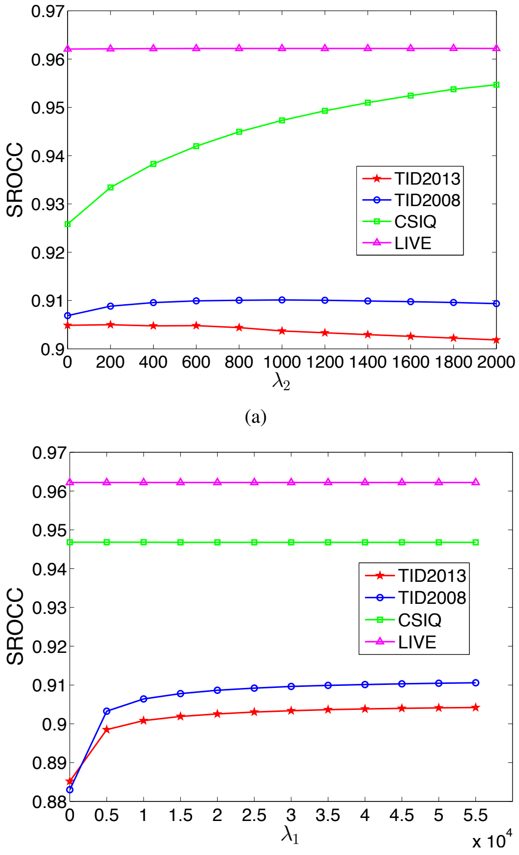
where q represents the results of an IQA method, p denotes the regression values of q, and $\eta_i (i = 1, 2, 3, 4, 5)$ are parameters to be fitted. The method in [44] is helpful for computing these parameters. After nonlinear regression, PLCC and RMSE can be calculated using p and subjective scores. With regard to SROCC and KROCC, they can be computed directly with subjective scores and objective results of the IQA method. Generally, an attractive IQA method usually presents high PLCC, SROCC, and KROCC with a small RMSE.

B. Parameter Setting

In this work, four pairs of parameters are involved, including τ_0 and τ_1 in (8), C_1 and C_2 in (9), λ_1 and λ_2 in (9), α and β in (10). Among them, C_1 and C_2 are commonly used in existing IQA models [14, 16] and are empirically set as $C_1 = 600$ and $C_2 = 210$. The other parameters are investigated in the following part. It is noteworthy that when we study the current parameters, the others remain unchanged.

Parameters τ_0 and τ_1 serve as the thresholds of RGC and IDG. They are set as $\tau = \tau_0 = \tau_1$ in our experiments for simplicity. In Fig. 5, SROCC curves against τ on the four databases are presented. We can find that for all databases, the performance is stable when τ is in the interval [0.3, 0.8]. In our experiments, we set $\tau = 0.6$.

The most important parameters are λ_1 and λ_2, which influence the degree of reducing distortion overestimation. In Fig. 6, we show the SROCC curves against λ_1 and λ_2 on the four databases. In Fig. 6(a), λ_1 is fixed as 40000 and λ_2 ranges from 0 to 2000. SROCC increases initially and then decreases on databases TID2008 and TID2013. The overall performance is optimal for these databases when λ_2 is within the interval [600, 1000]. In CSIQ, SROCC increases with λ_2, but the rate of increase is progressively small especially when $\lambda_2 \geq 1000$. For better comprehensive performance on different databases, we regulate the optimal λ_2 in the interval [900,1000] (e.g., 950). With regard to Fig. 6(b), the curve is obtained by varying λ_1 when $\lambda_2 = 950$. It can be observed that a larger λ_1 leads to a higher SROCC on TID2008 and TID2013, but this trend is not evident when $\lambda_1 \geq 40000$. Also, an overlarge λ_1 may result in failure in predicting the quality of over-enhanced images, although this condition is usually ignored and not reflected in current research and databases. Therefore, we set $\lambda_1 = 40000$ in our experiment. In summary, parameters λ_1 and λ_2 are fixed as 40000 and 950, respectively. It can be observed that the value of λ_1 significantly exceeds that of λ_2, which is consistent with our analysis in the end of Section III. A.



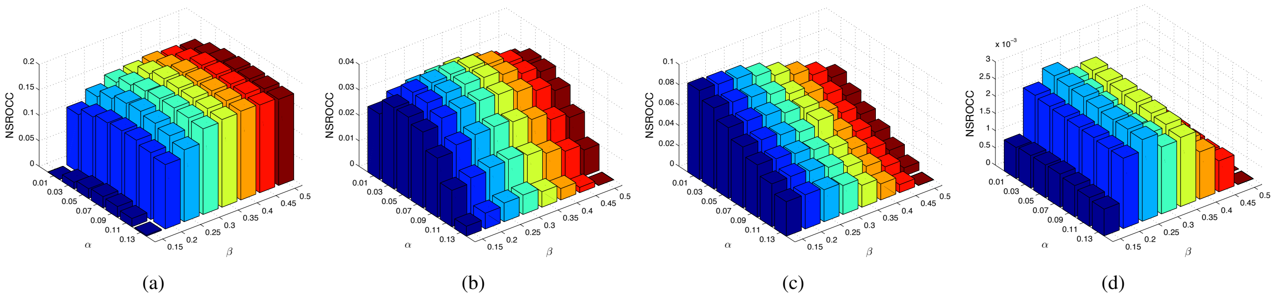
(a)

(b)

Fig. 6. Performance of SPSIM in terms of SROCC against (a) λ_2 and (b) λ_1 on four databases.

The last two parameters, namely, α and β, adjust the weights of luminance and color similarity measurements. Similar to parameters mentioned above, we further discuss their influences on the performance of SPSIM. The results are shown in Fig. 7, where $\textbf{NSROCC} = 10 \times (\textbf{SROCC} - min(\textbf{SROCC}))^1$. It can be observed that the optimal α and β of TID2013, TID2008, CSIQ, and LIVE are in intervals [0.05, 0.09]×[0.35, 0.45], [0.03, 0.07]×[0.30, 0.40], [0.01, 0.05]×[0.15, 0.30], and [0.03,

[1] SROCC is transformed into NSROCC on each database to make the changes visible and distinguishable in the figures.



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

Fig. 7. Performance of SPSIM in terms of SROCC against α and β on (a) TID2013, (b) TID2008, (c) CSIQ, and (d) LIVE.

TABLE I
Performance Comparison of FR IQA Methods on Four Databases

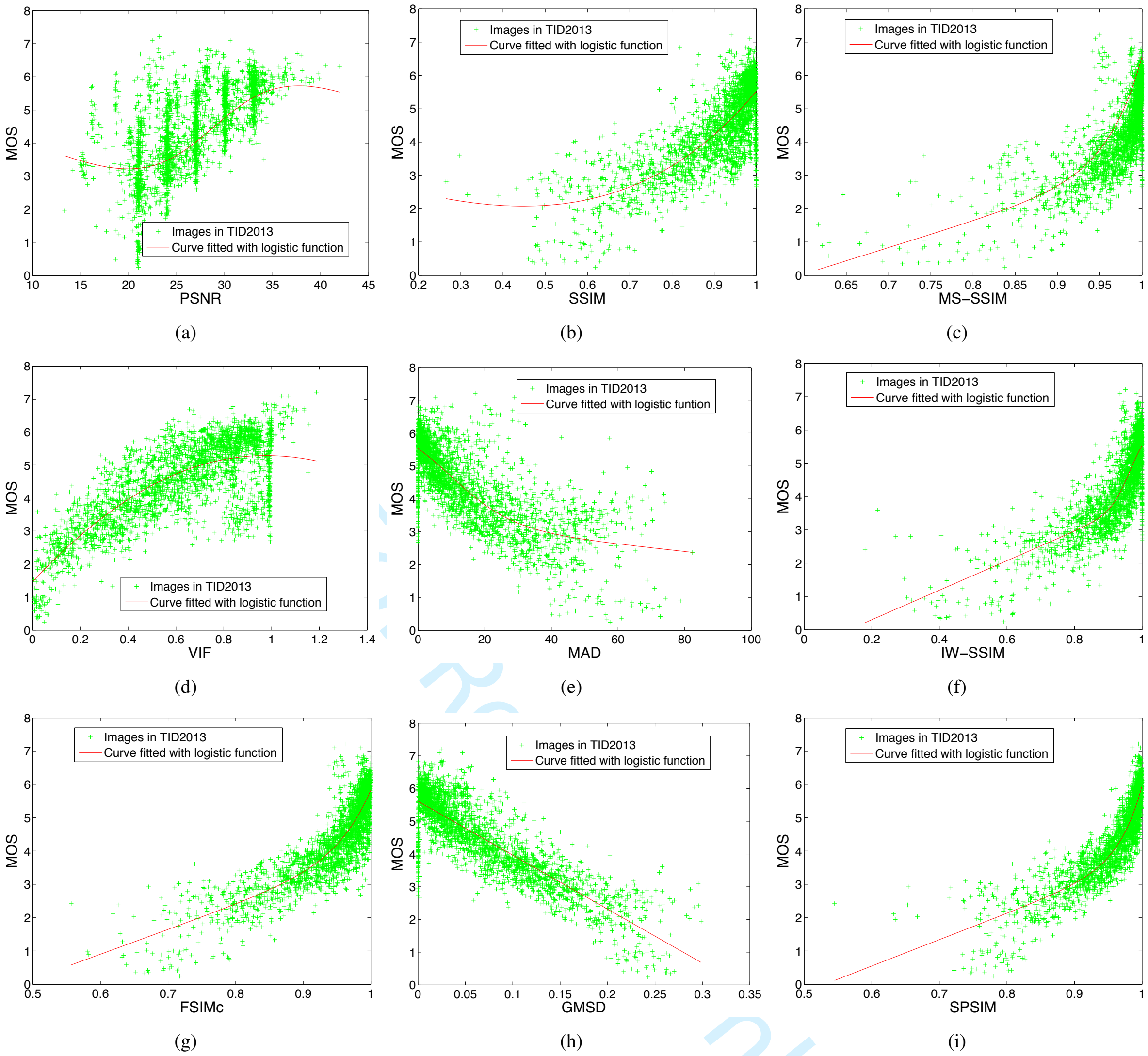
		PSNR	SSIM	MS-SSIM	VIF	MAD	IW-SSIM	FSIMc	GMSD	LLM	SPSIM
LIVE	SROCC	0.8756	0.9479	0.9513	**0.9636**	**0.9669**	0.9567	**0.9645**	0.9603	0.9608	0.9622
	KROCC	0.6865	0.7963	0.8045	**0.8282**	**0.8421**	0.8175	**0.8363**	0.8269	0.8230	0.8271
	PLCC	0.8723	0.9449	0.9489	**0.9604**	**0.9675**	0.9522	**0.9613**	0.9602	0.9578	0.9599
	RMSE	13.360	8.9455	8.6188	**7.6137**	**6.9073**	8.3473	**7.5296**	7.6211	7.7678	7.6288
CSIQ	SROCC	0.8005	0.8756	0.9133	0.9193	**0.9466**	0.9213	0.9310	**0.9572**	0.9050	**0.9440**
	KROCC	0.5984	0.6907	0.7393	0.7534	**0.7970**	0.7529	0.7690	**0.8134**	0.7238	**0.7880**
	PLCC	0.7998	0.8613	0.8991	0.9277	**0.9502**	0.9144	0.9192	**0.9542**	0.9000	**0.9344**
	RMSE	0.1576	0.1334	0.1149	0.0980	**0.0818**	0.1063	0.1034	**0.0786**	0.1232	**0.0934**
TID2008	SROCC	0.5245	0.7749	0.8542	0.7491	0.8340	0.8559	0.8840	**0.8906**	**0.9077**	**0.9104**
	KROCC	0.3696	0.5768	0.6568	0.5860	0.6445	0.6636	0.6991	**0.7090**	**0.7368**	**0.7374**
	PLCC	0.5309	0.7732	0.8451	0.8084	0.8308	0.8579	**0.8762**	0.8717	**0.8971**	**0.8927**
	RMSE	1.1372	0.8511	0.7173	0.7899	0.7468	0.6875	**0.6468**	0.6565	**0.5982**	**0.6046**
TID2013	SROCC	0.6394	0.7417	0.7859	0.6769	0.7807	0.7779	**0.8510**	0.8045	**0.9037**	**0.9044**
	KROCC	0.4696	0.5588	0.6047	0.5147	0.6035	0.5977	**0.6665**	0.6331	**0.7209**	**0.7251**
	PLCC	0.7017	0.7895	0.8329	0.7720	0.8267	0.8319	**0.8769**	0.8542	**0.9068**	**0.9091**
	RMSE	0.8832	0.7608	0.6861	0.7880	0.6975	0.6880	**0.5959**	0.6444	**0.5277**	**0.5165**
Weighted Average	SROCC	0.6596	0.7942	0.8419	0.7645	0.8405	0.8403	**0.8847**	0.8675	**0.9120**	**0.9186**
	KROCC	0.4870	0.6108	0.6616	0.6049	0.6702	0.6635	**0.7101**	0.7018	**0.7381**	**0.7496**
	PLCC	0.6903	0.8140	0.8594	0.8261	0.8619	0.8649	**0.8928**	0.8856	**0.9095**	**0.9145**

0.07]×[0.25, 0.35]. In all databases, the optimal α and β are in similar intervals, which is consistent with our knowledge that luminance and chrominance changes are visually perceived with certain weights in image quality perception. In this work, α and β are fixed as 0.05 and 0.35, respectively. In our future work, we would like to discuss the selections of α and β from the perspective of psychovisual experiments.

C. Performance Comparison

We compare the proposed method with nine well-known FR IQA approaches, namely, PSNR, SSIM, multi-scale SSIM (MS-SSIM) [45], VIF, MAD, IW-SSIM, FSIMc, GMSD, and LLM. The experimental results on the four benchmark databases are shown in Table I, where the top three results in each row are highlighted in boldface. Generally, the research on FR IQA has made great progress. Many methods provide accurate predictions about image quality on these four databases, especially the newest models FSIMc, GMSD, LLM, and SPSIM. Moreover, it can be observed that the performance of the same FR method diminishes on databases from LIVE to TID2013, which may be attributed to increasing numbers of distortion types in these four databases. Meanwhile, the distribution of boldfaced figures in Table I shows that no method performs best on all databases. VIF works effectively on LIVE. MAD provides precise results on LIVE and CSIQ. FSIMc is efficient on LIVE, TID2008, and TID2013. GMSD evaluates image quality consistently with subjective scores on CSIQ and TID2008. LLM performs effectively on TID2008 and TID2013. The proposed method achieves the best outcomes on TID2008 and TID2013 as well as the top three result on CSIQ. To compare these models comprehensively, we present the weighted average criteria at the bottom of Table I, where the weight is defined as the number of distorted images in each database. The weighted results show that our method obtains the best overall performance. In Fig. 8, the scatter plots of subjective scores and objective predictions by above methods on TID2013 are shown. It can be observed that data points of SPSIM are distributed more tightly along the fitted curve, which verifies the capability of SPSIM to assess image quality more consistently with human ratings.

To further compare the performance of those FR IQA metrics, we conducted statistical significance tests, which are commonly performed in the IQA research [15] [20]. The outcomes are presented in Fig. 9, where a value of ‘1’ indicates that the method in the row is statistically better than that in the column and ‘0’ otherwise. It can be observed that on the TID2008 and TID2013 databases, the proposed method significantly surpasses all other approaches, except for LLM. On the two other databases, only one model is significantly better than the proposed method, that is, MAD on LIVE and GMSD on CSIQ. In total, SPSIM achieves the value of ‘1’ 27 times, followed by GMSD (24 times), LLM (22 times),

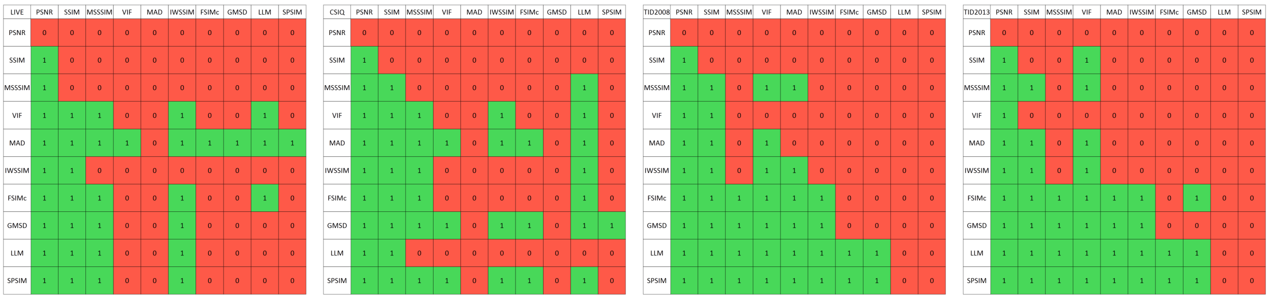


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 8. Scatter plots of FR IQA algorithms (a) PSNR, (b) SSIM, (c) MS-SSIM, (d) VIF, (e) MAD, (f) IW-SSIM, (g) FSIMc, (h) GMSD, and (i) SPSIM on TID2013.



LIVE	PSNR	SSIM	MSSSIM	VIF	MAD	IWSSIM	FSIMc	GMSD	LLM	SPSIM
PSNR	0	0	0	0	0	0	0	0	0	0
SSIM	1	0	0	0	0	0	0	0	0	0
MSSSIM	1	0	0	0	0	0	0	0	0	0
VIF	1	1	1	0	0	1	0	0	1	0
MAD	1	1	1	1	0	1	1	1	1	1
IWSSIM	1	1	0	0	0	0	0	0	0	0
FSIMc	1	1	1	0	0	1	0	0	1	0
GMSD	1	1	1	0	0	1	0	0	0	0
LLM	1	1	1	0	0	1	0	0	0	0
SPSIM	1	1	1	0	0	1	0	0	0	0

(a)

CSIQ	PSNR	SSIM	MSSSIM	VIF	MAD	IWSSIM	FSIMc	GMSD	LLM	SPSIM
PSNR	0	0	0	0	0	0	0	0	0	0
SSIM	1	0	0	0	0	0	0	0	0	0
MSSSIM	1	1	0	0	0	0	0	0	1	0
VIF	1	1	1	0	0	1	0	0	1	0
MAD	1	1	1	1	0	1	1	0	1	0
IWSSIM	1	1	1	0	0	0	0	0	1	0
FSIMc	1	1	1	0	0	0	0	0	1	0
GMSD	1	1	1	1	0	1	1	0	1	1
LLM	1	1	0	0	0	0	0	0	0	0
SPSIM	1	1	1	1	0	1	1	0	1	0

(b)

TID2008	PSNR	SSIM	MSSSIM	VIF	MAD	IWSSIM	FSIMc	GMSD	LLM	SPSIM
PSNR	0	0	0	0	0	0	0	0	0	0
SSIM	1	0	0	0	0	0	0	0	0	0
MSSSIM	1	1	0	1	1	0	0	0	0	0
VIF	1	1	0	0	0	0	0	0	0	0
MAD	1	1	0	1	0	0	0	0	0	0
IWSSIM	1	1	0	1	1	0	0	0	0	0
FSIMc	1	1	1	1	1	1	0	0	0	0
GMSD	1	1	1	1	1	1	0	0	0	0
LLM	1	1	1	1	1	1	1	1	0	0
SPSIM	1	1	1	1	1	1	1	1	0	0

(c)

TID2013	PSNR	SSIM	MSSSIM	VIF	MAD	IWSSIM	FSIMc	GMSD	LLM	SPSIM
PSNR	0	0	0	0	0	0	0	0	0	0
SSIM	1	0	0	1	0	0	0	0	0	0
MSSSIM	1	1	0	1	0	0	0	0	0	0
VIF	1	0	0	0	0	0	0	0	0	0
MAD	1	1	0	1	0	0	0	0	0	0
IWSSIM	1	1	0	1	0	0	0	0	0	0
FSIMc	1	1	1	1	1	1	0	1	0	0
GMSD	1	1	1	1	1	1	0	0	0	0
LLM	1	1	1	1	1	1	1	1	0	0
SPSIM	1	1	1	1	1	1	1	1	0	0

(d)

Fig. 9. Results of statistical significance tests of the competing IQA approaches on the databases of (a) LIVE, (b) CSIQ, (c) TID2008, and (d) TID2013. A value of ‘1’ (highlighted in green) indicates that the approach in the row is significantly better than the approach in the column, while a value of ‘0’ (highlighted in red) represents that the first approach is not significantly better than the second approach.

FSIMc (22 times), and MAD (22 times). This demonstrates that the proposed method is superior to other models. Except for SPSIM, FSIMc, GMSD, and LLM also obtain leading performance. In some specific cases, they are competitive with SPSIM, such as FSIMc on LIVE, GMSD on CSIQ, and LLM on TID2008 and TID2013. However, this is only

TABLE II
RESULTS OF SIGNIFICANCE TESTS OF FOUR METRICS

Database	Significance tests results			Rank of Metrics
LIVE	SPSIM∼FSIMc	SPSIM∼GMSD	SPSIM∼LLM	FSIMc∼**SPSIM**∼LLM∼GMSD
CSIQ	SPSIM>FSIMc	SPSIM<GMSD	SPSIM>LLM	GMSD>**SPSIM**>FSIMc>LLM
TID2008	SPSIM>FSIMc	SPSIM>GMSD	SPSIM∼LLM	**SPSIM**∼LLM>GMSD∼FSIMc
TID2013	SPSIM>FSIMc	SPSIM>GMSD	SPSIM∼LLM	**SPSIM**∼LLM>FSIMc>GMSD

TABLE III
SROCC OF SPSIM WITH FIXED T_1 AND T_2 ON TID2013

SROCC on TID2013		T_2						
		100	250	400	550	700	850	1000
T_1	100	0.8778	0.8808	0.8810	0.8803	0.8792	0.8780	0.8768
	200	0.8787	0.8843	0.8860	0.8864	0.8864	0.8860	0.8854
	300	0.8758	0.8824	0.8850	0.8862	**0.8867**	**0.8867**	0.8866
	400	0.8719	0.8795	0.8827	0.8844	0.8853	0.8857	0.8858
	500	0.8678	0.8762	0.8799	0.8820	0.8831	0.8838	0.8840
	600	0.8638	0.8728	0.8770	0.8793	0.8807	0.8816	0.8818

valid in a few cases while SPSIM performs excellently in the vast majority of instances. From the significance tests we can find: 1) SPSIM is superior to FSIMc on CSIQ, TID2008, and TID2013; 2) SPSIM achieves better performance than GMSD on TID2008 and TID2013; 3) SPSIM predicts image quality more accurately than LLM does on CSIQ. We show the significance tests of the four metrics mentioned above in a more conspicuous way in Table II, where '>' means that SPSIM is statistically better than the right-hand one, '<' means that SPSIM is worse than the right-hand algorithm, and '∼' indicates no significant difference. In the column of 'Rank of Metrics', if two metrics are not significantly distinguishable, the one with higher SROCC is placed in front. It can be easily found that SPSIM is superior to FSIMc and GMSD in most cases. As for LLM, it employs a deep learning based classification method, which depends heavily on training data. Therefore, LLM performs excellently on TID2008 and TID2013 (part as a training set) but produces unsatisfactory results on CSIQ (4-5 percent lower in SROCC). On the contrary, our method achieves top three results on almost all databases. In all, the proposed method is superior to others due to its excellent performance and universality. In addition to the F-test, the significance tests using the Pitman test [46] can also verify the superiority of SPSIM over other IQA approaches.

D. Discussion about the Adaptive Selection of T_1 and T_2

In our work, T_1 and T_2 are adaptive to the regional overall change of features. Additional experiments are conducted to compare the performance difference between adaptive T_1, T_2 and fixed T_1, T_2. Results of the proposed method using fixed values are shown in Table III, where experiments are conducted on TID2013 with other parameters unchanged. From Table III, we can find that SROCC increases initially when $T_1 < 300$ and then decreases when $T_1 > 300$. Moreover, an upward trend of SROCC is observed with the growth of T_2, but this trend is not evident when $T_2 > 850$. The optimal SROCC is approximately 0.8867, which is nearly 2 percent lower than that of adaptive T_1 and T_2 (0.9044). This performance gap demonstrates that using adaptive T_1 and T_2 is effective for predicting image quality. Further, we test a special pair of fixed T_1 and T_2, which is obtained by averaging all adaptive T_1 and T_2 on TID2013. The outcomes are $T_1 = 1097$, $T_2 = 1487$, and $SROCC = 0.8737$. This pair of T_1 and T_2 shares the same mean values with the adaptive ones, but its performance is inferior. This fact further verifies the superiority of adaptive selection of parameters. Similar comparisons and conclusions can be obtained on the three other databases.

TABLE IV
COMPARISON OF RUNNING TIME

IQA metric	Running time (s)
PSNR	0.0023
SSIM	0.0155
MS-SSIM	0.0878
VIF	0.7329
MAD	20.5872
IW-SSIM	0.3812
FSIMc	0.2665
GMSD	0.0116
LLM	-
SPSIM	0.2174

E. Computational Complexity

It is necessary to analyze the computational complexity of an algorithm because the running time is crucial in many real-time applications and systems. In Table IV, the running time of several IQA metrics on a 384×512 image is listed. Experiments are performed on a computer with Intel Core i7-870 CPU@2.8 GHz and 8G RAM. The software platform is Matlab R2013a. As shown in Table IV, PSNR and GMSD are the fastest IQA approaches. The proposed SPSIM shows a moderate running speed among all compared approaches. Our method mainly involves three steps of operations, namely, superpixel segmentation, feature extraction, and quality pooling. Compared with other models, the only added part is the operation of superpixel segmentation. In experiments, we find that the time cost of superpixel segmentation is about 0.0621s and this time can be reduced by off-line segmentation. In many systems (e.g., image transmission and image compression), the superpixel segmentation of the original image can be conducted in advance. Since we segment a distorted image following its reference image, applying our method to predict the quality of output images will save a lot of time, making this approach highly practical in real applications.

V. Conclusion

This study proposes a new FR IQA method from the viewpoint of superpixels. Based on the observation that visual meaningful regions are beneficial for image quality assessment, we segment reference and distorted images into many superpixels. Then, mean values of luminance and chromatic components are computed and compared in superpixels instead of square patches to effectively reflect local characteristics. Furthermore, we employ image gradients to characterize structural degradations. The comparisons of these three features are further revised by superpixel-based RGC, which shows that the quality of a distorted image is usually underestimated if the RGCs between this image and its reference are generally large. Finally, in order to obtain an overall quality score, a weighting strategy utilizing texture complexity is adopted. The experimental results on four databases demonstrate that our method predicts image quality more consistently with human assessment than most existing models do.

VI. Acknowledgement

The authors would like to thank all the reviewers and editors for their constructive comments which largely improve this work.

References

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

[2] H. Noda, S. Haraguchi, and M. Niimi, "Local map estimation for quality improvement of compressed color images," in *TENCON 2010 - 2010 IEEE Region 10 Conference*, pp. 1657–1662, 2011.

[3] C. T. Vu, T. D. Phan, P. S. Banga, and D. M. Chandler, "On the quality assessment of enhanced images: A database, analysis, and strategies for augmenting existing methods," in *Image Analysis and Interpretation*, pp. 181–184, 2012.

[4] Z. Wang and A. Bovik, "Modern image quality assessment," *Synthesis Lectures on Image Video and Multimedia Processing*, vol. 2, no. 1, p. 156, 2006.

[5] D. M. Chandler and S. S. Hemami, "VSNR: a wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[6] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.

[7] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

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

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] L. Liu, Y. Wang, and Y. Wu, "A wavelet-domain structure similarity for image quality assessment," in *International Congress on Image and Signal Processing*, pp. 1–5, 2009.

[11] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[12] R. Zhu, F. Zhou, and J.-H. Xue, "MvSSIM: A quality assessment index for hyperspectral images," *Neurocomputing*, 2017.

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

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

[15] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index.," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[16] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[17] D. Liu, F. Li, and H. Song, "Image quality assessment using regularity of color distribution," *IEEE Access*, vol. 4, pp. 4478–4483, 2016.

[18] D. Lee and K. N. Plataniotis, "Towards a full-reference quality assessment for color images using directional statistics," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3950–3965, 2015.

[19] Y. Niu, H. Zhang, W. Guo, and R. Ji, "Image quality assessment for color correction based on color contrast similarity and color value difference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2016.

[20] H. Wang, F. Jie, W. Lin, S. Hu, C. C. J. Kuo, and L. Zuo, "Image quality assessment based on local linear information and distortion-specific compensation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 915–926, 2017.

[21] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "S-LIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2281, 2012.

[22] G. Fracastoro, F. Verdoja, M. Grangetto, and E. Magli, "Superpixel-driven graph transform for image compression," in *IEEE International Conference on Image Processing*, pp. 2631–2635, 2015.

[23] Z. Liu, M. Le, and S. Luo, "Superpixel-based saliency detection," in *International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 1–4, 2013.

[24] X. Jin and Y. Gu, "Superpixel-based intrinsic image decomposition of hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4285–4295, 2017.

[25] V. N. Gangapure, S. Nanda, and A. S. Chowdhury, "Superpixel based causal multisensor video fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[26] C. Peng, X. Gao, N. Wang, and J. Li, "Superpixel-based face sketchcphoto synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 2, pp. 288–299, 2017.

[27] G. Zeng, P. Wang, J. Wang, and R. Gan, "Structure-sensitive superpixels via geodesic distance," in *International Conference on Computer Vision*, pp. 447–454, 2012.

[28] R. Giraud, V. T. Ta, and N. Papadakis, "SCALP: Superpixels with contour adherence using linear path," in *International Conference on Pattern Recognition*, pp. 2374–2379, 2016.

[29] K. H. Thung, R. Paramesran, and C. L. Lim, "Content-based image quality metric using similarity measure of moment vectors," *Pattern Recognition*, vol. 45, no. 6, pp. 2193–2204, 2012.

[30] S. H. Bae and M. Kim, "A novel image quality assessment with globally and locally consilient visual quality perception," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2392–2406, 2016.

[31] M. Narwaria, W. Lin, and A. E. Cetin, "Scalable image quality assessment with 2d mel-cepstrum and machine learning approach," *Pattern Recognition*, vol. 45, no. 1, pp. 299–313, 2012.

[32] X. Liu, H. Jia, L. Cao, C. Wang, J. Li, and M. Cheng, "Superpixel-based coastline extraction in sar images with speckle noise removal," in *Geoscience and Remote Sensing Symposium*, pp. 1034–1037, 2016.

[33] M. Wang and T. Blu, "Generalized yuv interpolation of cfa images," in *IEEE International Conference on Image Processing*, pp. 1909–1912, 2010.

[34] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Publishing House of Electronics Industry, 2010.

[35] D. Li, H. Huang, and Z. Yu, "Image quality assessment using directional anisotropy structure measurement," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1799–1809, 2017.

[36] T. V. Q. E. Group, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II," 2003.

[37] K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu, "Subjective and objective quality assessment for images with contrast change," in *IEEE International Conference on Image Processing*, pp. 383–387, 2014.

[38] S. H. Bae and M. Kim, "A novel generalized DCT-based JND profile based on an elaborate CM-JND model for variable block-sized transforms in monochrome images.," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3227–3240, 2014.

[39] B. Moulden and L. F. Gatley, "The standard deviation of luminance as a metric for contrast in random-dot images," *Perception*, vol. 19, no. 1, pp. 79–101, 1990.

[40] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: From quality assessment to automatic enhancement.," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 284–297, 2016.

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

[42] 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, pp. 30–45, 2004.

[43] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, and F. Battisti, "Color image database TID2013: Peculiarities and preliminary results," in *European Workshop on Visual Information Processing*, pp. 106–111, 2013.

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

[45] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, pp. 1398–1402, 2004.

[46] R. Zhu, F. Zhou, W. Yang, and J.-H. Xue, "On hypothesis testing for comparing image quality assessment metrics," *IEEE Signal Processing Magazine,minor revision*.

Wen Sun received the B.Eng. degree from Department of Information Science and Engineering, Southeast University, China, in 2013. He is currently working towards the Ph.D. degree in electronics engineering at Tsinghua University, China. His research interests include image quality assessment and image understanding.

Qingmin Liao received the Ph.D. degree in signal processing and telecommunications from the University of Rennes 1, France, in 1994. He became a professor in the Department of Electronic Engineering of Tsinghua University, in 2002. Since 2010, he has been the Director of the Division of Information Science and Technology in the Graduate School at Shenzhen, Tsinghua University. His research interests include image/video processing, transmission and analysis; biometrics; and their applications to teledetection, medicine, industry, and sports. He has published over 100 papers internationally.

Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is a senior lecturer in the Department of Statistical Science, University College London. His research interests include statistical machine learning, high-dimensional data analysis, pattern recognition and image analysis.

Fei Zhou received the B.Eng. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology in 2007, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University in 2013. From 2013 to 2016, he worked as a postdoctoral fellow with the Shenzhen Graduate School, Tsinghua University. From 2017 to 2018, he worked as a visiting scholar of the Department of Statistical Science, University College London. He is currently an assistant professor in the College of Information Engineering, Shenzhen University. His research interests include applications of image processing and pattern recognition in video surveillance, image super-resolution, image interpolation, image quality assessment and object tracking.