International Journal of Automation and Computing
www.ijac.net

18(2), April 2021, 204-218
DOI: 10.1007/s11633-020-1270-z

Structured Computational Modeling of Human Visual System for No-reference Image Quality Assessment

Wen-Han Zhu[1,2] Wei Sun[2] Xiong-Kuo Min[2] Guang-Tao Zhai[2] Xiao-Kang Yang[1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China

[2] Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: Objective image quality assessment (IQA) plays an important role in various visual communication systems, which can automatically and efficiently predict the perceived quality of images. The human eye is the ultimate evaluator for visual experience, thus the modeling of human visual system (HVS) is a core issue for objective IQA and visual experience optimization. The traditional model based on black box fitting has low interpretability and it is difficult to guide the experience optimization effectively, while the model based on physiological simulation is hard to integrate into practical visual communication services due to its high computational complexity. For bridging the gap between signal distortion and visual experience, in this paper, we propose a novel perceptual no-reference (NR) IQA algorithm based on structural computational modeling of HVS. According to the mechanism of the human brain, we divide the visual signal processing into a low-level visual layer, a middle-level visual layer and a high-level visual layer, which conduct pixel information processing, primitive information processing and global image information processing, respectively. The natural scene statistics (NSS) based features, deep features and free-energy based features are extracted from these three layers. The support vector regression (SVR) is employed to aggregate features to the final quality prediction. Extensive experimental comparisons on three widely used benchmark IQA databases (LIVE, CSIQ and TID2013) demonstrate that our proposed metric is highly competitive with or outperforms the state-of-the-art NR IQA measures.

Keywords: Image quality assessment (IQA), no-reference (NR), structural computational modeling, human visual system, visual feature extraction.

Citation: W. H. Zhu, W. Sun, X. K. Min, G. T. Zhai, X. K. Yang. Structured computational modeling of human visual system for no-reference image quality assessment. *International Journal of Automation and Computing*, vol.18, no.2, pp.204–218, 2021. http://doi.org/10.1007/s11633-020-1270-z

1 Introduction

In the 21st century, an informative network era, the Internet has become an important way for people to acquire the latest information and entertainment. Visual information, including images and videos, has accounted for more than 80% of the total Internet traffic. High quality visual experience is the common basis of major applications such as the digital media industry and network information service. Image quality assessment (IQA), dedicated to evaluating human visual perception and predict image quality, has been a fundamental issue in image processing fields[1]. Although subjective IQA is the most accurate approach, the slowness, time-consuming, laboriousness and difficult repetition of subjective IQA immensely limit its progress. By contrast, objective IQA that resorts to mathematical metrics for predicting the perceived quality of images automatically and efficiently has been widely researched. In common IQA databases, the distorted images are usually degraded from a pristine image called the reference image. According to the available information of the reference image, objective IQA algorithms can be classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) algorithms, respectively.

FR IQA models calculate the target image quality with fully accessible reference images, and they usually measure the "distance" between the perfect original image and its degraded image[2]. The mean-squared error (MSE) and peak signal-to-noise ratio (PSNR) are two classic and widely used metrics proposed long ago. However, they have a poor correlation with subjective perceptions in some conditions[3]. For this purpose, Wang et al.[4] propose the structural similarity index (SSIM) combining luminance information, contrast information and structure information based on the assumption that the human visual system (HVS) is highly sensitive to the structures in the image. In addition, plenty of successful FR IQA algorithms are subsequently designed, such as the visual information fidelity (VIF)[5], the visual signal-

Research Article
Manuscript received July 31, 2020; accepted November 17, 2020; published online January 4, 2021
Recommended by Associate Editor Zhi-Jie Xu
Colored figures are available in the online version at https://link.springer.com/journal/11633

to-noise ratio (VSNR)[6] and the perceptual similarity (PSIM)[7]. When only partial information about the original image is available, RR IQA models take full advantage of this information to evaluate the image quality. Wang et al.[8] propose an effective method using the natural scene statistics (NSS) features in the transform domain. In the spatial domain, Liu et al.[9] report a RR model compositing the bottom-up and top-down features as reference data. Soundararajan and Bovik[10] develop the reduced reference entropic differencing (RRED) metric via the wavelet coefficients of original and distorted images to assess their qualities. Min et al.[11] measure image quality based on the saliency similarity.

However, the pristine image is nonexistent or unavailable in most of the actual scenarios, and thus NR IQA models are highly desirable which require no original references. Since there is no prior knowledge of the reference image, NR IQA is more difficult and challenging than FR and RR IQA. In fact, the development of NR IQA has been rapid and brilliant in recent years. Based on the design philosophies of the measures, the NR IQA algorithms can be divided into three types, which are NSS-based, learning-based and HVS-based measures. NSS-based NR IQA models are the earliest NR metrics and their motivation is that high quality natural images possess some kind of statistical properties, while degraded images no longer possess such properties. Typical NSS-based NR models follow three major steps: feature extraction, NSS modeling, and feature regression[12]. In the literature, Moorthy and Bovik[13] propose a distortion identification based image verity and integrity evaluation (DIIVINE) model based on the statistical properties of wavelet coefficients. Mittal et al.[14] design a natural image quality evaluator (NIQE) using the NSS of image patches with high image contrast in the spatial domain. Min et al.[15] develop a blind IQA model called blind pseudo-reference image based metric (BPRI) based on the NSS of pseudo-reference images. Liu et al.[16] propose an unsupervised method with the structure, naturalness, and the perception quality variations based on a pristine multivariate Gaussian model. An increasing number of learning-based NR measures have been proposed in recent years, which try to learn and integrate the quality features. For example, Ye et al.[17] report an unsupervised feature learning framework method CORNIA (codebook representation for no-reference image assessment) by constructing an unlabeled codebook from raw image patches via K-means clustering. Xu et al.[18] introduce a NR IQA metric based on high order statistics aggregation (HOSA). A blind image evaluator using an optimized end-to-end convolutional neural network is proposed by Kim and Lee[19].

The human eye is the final receiver of visual signals and the ultimate criterion of visual experience for human beings. Computational modeling of HVS is a key scientific problem in visual experience optimization. Thus, in addition to the above two categories of NR models, the HVS-based NR metric is also an important component of NR algorithms, which is motivated by some properties of the HVS, and extracts some perceptual-based features to predict the image quality. Zhai et al.[20] propose a psychovisual image no-reference free-energy-based quality metric (NFEQM) inspired by the free-energy principle interpreting the perception of an image as an active inference process. Gu et al.[21] put forward a NR method incorporating free-energy based features, structural information and gradient magnitude. Li et al.[22] design an NR IQA metric employing no-reference quality assessment using structural and luminance (NRSL) features inspired by human visual perception of images. Li et al.[23] report an NR IQA algorithm extracting perceptual features from first-order and second-order structural patterns of images. Saha and Wu[24] extract features from scale-space, Fourier domain and wavelet domain to compute the quality score of the target image. Although there are a lot of HVS-based NR algorithms and the effectiveness of these models has been proved, these metrics still have defects. Traditional black-box regression models have low interpretability, which can hardly guide visual experience optimization effectively, while the models based on physiological simulation are difficult to integrate into practical visual communication services due to theirs high computational complexity. How to construct a NR IQA metric with high interpretability to bridge the gap between signal distortion and visual experience still deserves to be researched.

In the literatures of neuroscience and brain theory, visual experience can be induced by external stimuli, such as the appearance of an image[25]. Localization of the neural structure is an important step in the process of comprehending the fundamental mechanisms of the visual system[26]. The human brain is limited in its ability to interpret all perceptual stimuli appearing in the visual field at any position in time and relies on a gradual cognitive process of the stimuli based on the contingencies of the moment[27]. During perception, activation of visual imagery ultimately results from bottom-up impacts from the retina[28, 29]. Therefore, we attempt to propose a bottom-up structured HVS-based approach to illustrate the information transfer and feedback mechanism of visual perception in the human brain. Combined with knowledge of image processing, we divide visual stimuli into three bottom-up layers, which are a low-level visual layer, a middle-level visual layer and a high-level visual layer. Specifically, for an image, the global image can be regarded as the high-level visual excitation, and the local primitives obtained from the decomposition of the global image can be treated as the middle-level visual stimulus, while the low-level visual layer is composed of all individual pixels in the image. Conversely, the complete global image can be acquired by synthesizing its local primitives, which are constituted by individual pixels of the image. The diagram of our proposed structural computational modeling in the human visual system is shown in Fig. 1.

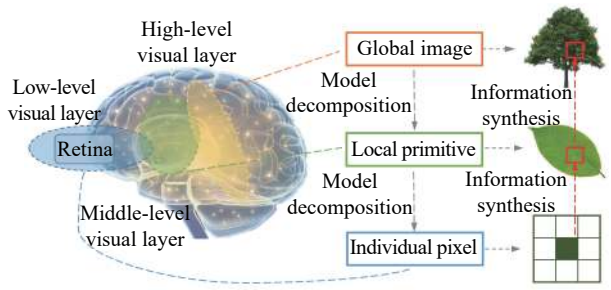


Fig. 1 Diagram of our proposed structural computational modeling of human visual system

In this paper, inspired by the above-mentioned framework, a new NR IQA algorithm based on structural computational modeling of the human visual system is proposed, called NSCHM (no-reference structural computational of human visual system metric). We deeply investigate and analyze the perception mechanism in HVS based on multi-layer representations of the image. A set of quality-aware NSS-based features are extracted as low-level visual features. Deep features in the convolution network are considered as middle-level features and free-energy based features are treated as high-level features in our proposed method. Finally, support vector regression (SVR) is used to aggregate these three layers′ features into a perception quality index that can predict the quality scores of target images. In order to demonstrate the effectiveness of our method, extensive experiments are performed on three common image quality databases (LIVE[30], CSIQ[31] and TID2013[32]) and the method is compared with eight mainstream general-purpose NR algorithms. Experimental results show that the proposed NSCHM method is effective and superior or comparable to the state-of-the-art NR models.

The remainder of this paper is organized as follows. In Section 2, we describe details of the NSCHM metric. Validation is given in Section 3, which demonstrates that NSCHM is comparable to or outperforms the state-of-the-art NR IQA models. Finally, we draw some general conclusions in Section 4.

2 The proposed algorithm

For characterizing the quality of images using structural computational modeling, we investigate three layers of perception mechanism in HVS. In this section, three levels of features including low-level visual features, middle-level visual features and high-level visual features are analyzed and devised to characterize the quality of distorted images effectively. After feature extraction, we adopt SVR to regress those features into the final index to represent the quality of target images. The overall diagram of the NSCHM method is illustrated in Fig. 2.

2.1 Feature extraction in the low-level visual layer

The features extracted from NSS have been widely accepted in the NR IQA field because of their stability and efficiency. The NSS-based features in the spatial domain can judge the degree of image degradation by the characteristics at pixel level since high-quality original scene images satisfy some certain statistical characteristics, while quality degradation may alter these characteristics. This is consistent with the low-level visual features we expected. Therefore, in this section we will introduce the selection of low-level visual features based on NSS in the spatial domain.

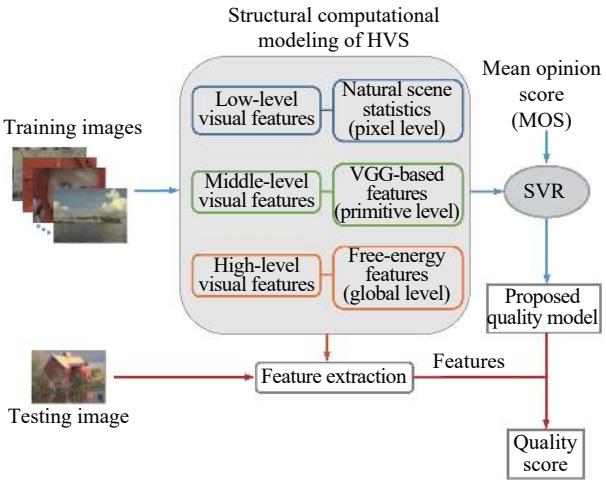


Fig. 2 Overall diagram of the NSCHM algorithm

Specifically, inspired by some previous studies[33, 34], the locally mean subtracted and contrast normalized (MSCN) coefficients of the intensity image of a target image can denote the luminance effectively. Given an image I, we first transform I to the intensity image H, and then the MSCN coefficients of H can be calculated as

$$H'(x,y) = \frac{H(x,y) - \mu(x,y)}{\sigma(x,y) + 1}$$

where $H(x,y)$ and $H'(x,y)$ represent the pristine and normalized values of the intensity image at position (x,y), $x \in \{1,2,\cdots,L_W\}$ and $y \in \{1,2,\cdots,L_H\}$ are spatial indices, L_W and L_H mean the width and height of the image respectively. $\mu(x,y)$ and $\sigma(x,y)$ denote the mean and standard deviation of the local patch with the center (x,y), which can be computed as

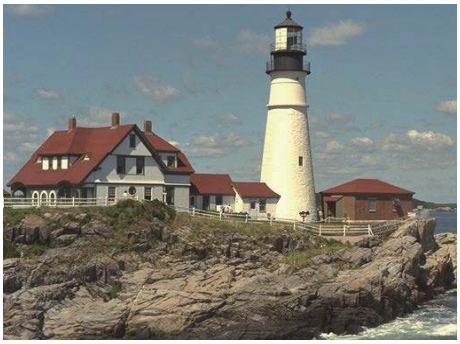
$$\mu(x,y) = \sum_{u=-U}^{U}\sum_{v=-V}^{V} \omega_{u,v} H(x+u, y+v)$$

$$\sigma(x,y) = \sqrt{\sum_{u=-U}^{U}\sum_{v=-V}^{V} \omega_{u,v}(H(x+u,y+v) - \mu(x,y))^2}$$

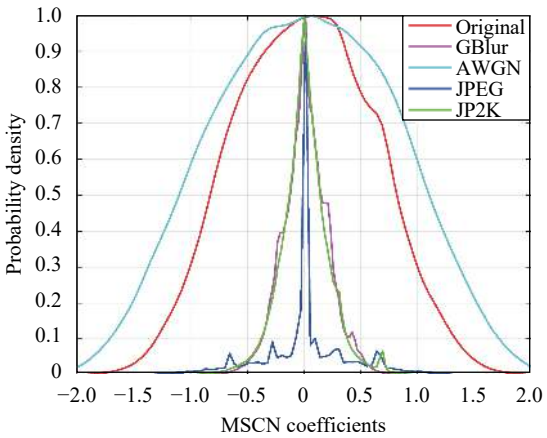
where $\omega = \{\omega_{u,v} | u = -U,\cdots,U, v = -V,\cdots,V\}$ stands for the 2D circularly-symmetric Gaussian weighting function and $U = V = 3$ in this implementation. Ruderman[34] observes that these normalized luminance

values of natural images have a great correlation with the unit normal Gaussian characteristic. These properties of MSCN coefficients can be used to describe the distortion level of the target image. To demonstrate this fact, the MSCN coefficients′ distributions of a reference image selected from TID2013 IQA database[32] and its corresponding degraded versions with different distortion types are shown in Fig. 3. It is obvious that the distributions of the reference image and its various distorted versions are different, which indicates that the statistical properties of MSCN coefficients can be changed by various distortions. In addition, as reported by [33], the distribution of the original image presents a Gaussian like appearance and each distortion deviates from such kind of properties in its own way. For describing the MSCN coefficients′ distribution specifically, a generalized Gaussian distribution (GGD) is employed which can effectively depict the broader spectrum of the distorted image statistics. The zero-mean GGD is expressed as

$$G(x;\alpha,\sigma^2)=\frac{\alpha}{2\beta(1/\alpha)}\exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right)$$



(a)



(b)

Fig. 3 MSCN distributions of a reference image and its corresponding degraded versions with different distortion types including additive white Gaussian noise (AWGN), Gaussian blur (GBlur), JPEG compression (JPEG) and JPEG2000 (JP2K): (a) Original image extracted from TID2013 database; (b) MSCN distributions.

where $\beta=\sigma\sqrt{\Gamma(1/\alpha)/\Gamma(3/\alpha)}$ and gamma function $\Gamma(\cdot)$ is defined as

$$\Gamma(\varphi)=\int_0^{\infty}\phi^{\varphi-1}e^{-\phi}\mathrm{d}\phi,\ \varphi>0$$

where α and σ are the parameters, which control the magnitude and the variance of the distribution, respectively. Then, we employ this GGD model to fit the above-mentioned MSCN distributions from the target images and extract α and σ as the quality-aware features for our low-level visual feature group.

In addition to the statistical distribution of each pixel, we also consider the statistical law of adjacent pixels, which exhibits a regular structure and is sensitive to the presence of distortion[33]. Thus, we compute the pairwise products of adjacent MSCN coefficients in four orientations including horizontal, vertical, main-diagonal and secondary-diagonal. The distributions of the pairwise products of the adjacent MSCN coefficients of the reference and its various degraded versions along the horizontal direction are illustrated in Fig. 4. The difference between the distribution of the original image and that of its distorted version can be clearly distinguished. Similarly, we adopt a zero mode asymmetric generalized Gaussian distribution (AGGD) model to fit these distributions of the adjacent coefficients:

$$G(x;\gamma,\sigma_l^2,\sigma_r^2)=\begin{cases}\dfrac{\gamma}{(\beta_l+\beta_r)\Gamma(1/\gamma)}\exp\left(-\left(\dfrac{-x}{\beta_l}\right)^{\gamma}\right), & \text{if } x<0\\ \dfrac{\gamma}{(\beta_l+\beta_r)\Gamma(1/\gamma)}\exp\left(-\left(\dfrac{-x}{\beta_r}\right)^{\gamma}\right), & \text{if } x\le 0\end{cases}$$

where $\beta_l{=}\sigma_l\sqrt{\Gamma(1/\gamma)/\Gamma(3/\gamma)}$ and $\beta_r{=}\sigma_r\sqrt{\Gamma(1/\gamma)/\Gamma(3/\gamma)}$ control the expansion of each side respectively, while γ is the parameter controlling the magnitude of the mode.

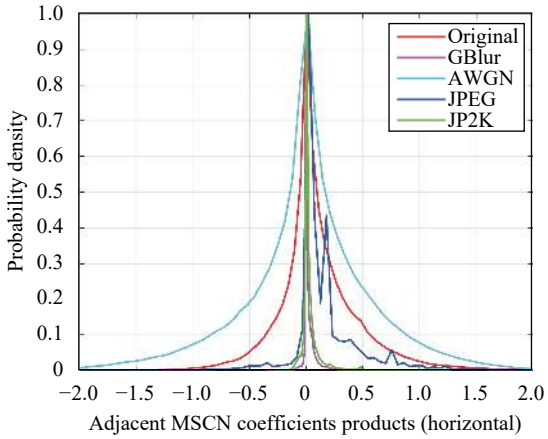


Fig. 4 Distributions of the products of the adjacent MSCN coefficients along the horizontal orientation of an original image and its corresponding degraded versions distorted by AWGN, GBlur, JPEG and JP2K.

Then the mean of the above distribution can be calculated as

$$\eta = (\beta_r - \beta_l)\frac{\Gamma(2/\gamma)}{\Gamma(1/\gamma)}.$$

The means η in the informative model parameters $(\gamma, \eta, \beta_l, \beta_r)$ of this AGGD model are extracted as our low-level visual features considering its high sensitivity to various degradation of images proved by extensive experiments. Since multi-scale processing contributes to improve the correlation between predicting scores of QA models and the human perception, we extract all features at two scales including the original scale and a reduced resolution downsampled with a factor of two. Finally, a total of twelve features, six at each scale, are employed as the low-level visual features L_i to measure the quality of the target image.

2.2 Feature extraction in the middle-level visual layer

Following the low-level visual feature extraction, in this section, we will discuss the middle-level visual feature extraction. As mentioned above, we consider that the middle-level visual feature is more advanced than the low-level visual features, which is no longer the information at the pixel level, but the characteristic of some local primitives in the images. It is known that the convolutional neural network (CNN) can extract local features of images by calculating the cross-correlation between convolution kernels and feature maps. With the development of deep learning in recent years, deep CNNs show great performance in solving various visual signal problems, such as image recognition[35, 36], detection[37, 38], tracking[39, 40], etc. Also, many studies indicate that local features extracted by CNNs response to edge, texture, etc., which is consistent with the reaction of neurons in the human visual system. The core of deep learning is passing the kernel through continuous convolution iteration between layers to realize the final goal, which accords with the properties of the middle visual layer conceived by us. How to extract suitable deep features as the middle-level visual features is the target of this section.

As a novel concept, the pseudo-reference image using the worst image to act as a reverse reference image is proved to be effective in NR IQA models[15]. Inspired by this concept[41], we combine a deep convolutional neural network with this framework to extract middle-level visual features. The framework of the proposed middle-level feature extraction is illustrated in Fig. 5. First, we need to confirm the distortion types for the distortion aggravation to produce the pseudo-reference images. Since different categories of distortion bring in different artifacts, the pseudo-reference image associated with a specific distortion needs to be defined to comply with the properties of the given distortion. Generally, in most of widely used

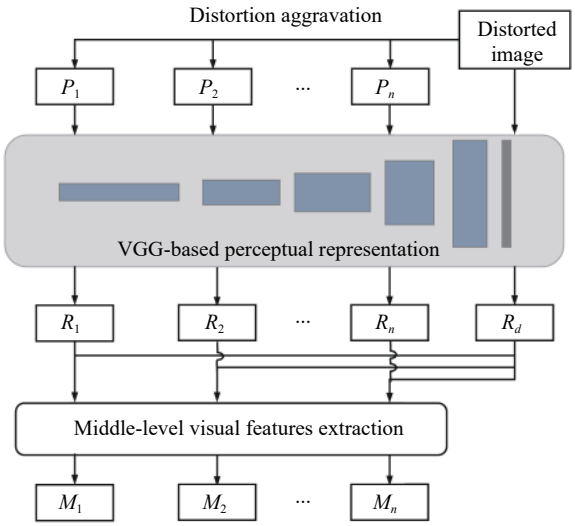


Fig. 5 Framework of the proposed middle-level visual features extraction. P_1 to P_n are multiple pseudo reference images, R_d and R_1 to R_n denote the representation maps in VGG. The VGG-based perceptual representation includes six stages, in which the zeroth stage is the raw pixels. M_1 to M_n indicate the distorted image′ s middle-level visual features combining the texture and structure features of target distorted image and its corresponding multiple pseudo reference images at different convolution layers.

subjective IQA databases, AWGN, GBlur, JPEG and JP2K are the four most mainstream encountered distortion types. Thus, these four distortion types are used to further measure the noising, blurring, blocking and ringing artifacts via degrading the distorted image. For different categories of distortion aggravation, VGG-based representation maps are generated and the middle-level visual features are extracted based on the features in these maps. Since the visual geometry group (VGG) network has great power in representing image local features, we calculate the VGG-based representation maps for different categories of distortion aggravation images and extract the middle-level visual features from these maps. The details are introduced as follows.

Firstly, we introduce the methods of distortion aggravation for each distortion type. To achieve noising effects, we add white noise to the distorted image D to obtain the multiple pseudo reference images (MPRIs) P_{ni}:

$$P_{ni} = D + \mathcal{N}(0, v_i)$$

where i represents the i-th degree of distortion aggravation, $\mathcal{N}(0, v_i)$ indicates a random normal distribution with zero mean and v_i variance. For the blurring effect, we blur the distorted image D to MPRIs P_{bi} by employing Gaussian kernels:

$$P_{ni} = \mathcal{G}_i * D$$

where $\mathcal{G}_i$ is a Gaussian kernel with determinate standard deviation and $*$ denotes a convolution operator. To

realize the blocking effect, the JPEG encoder is used to compress the distorted image D to the MPRIs P_{ji}:

$$P_{ji} = \mathrm{JPEG}(D, J_i)$$

where JPEG indicates the JPEG encoder and J_i adjusts the compression quality. For producing the ringing effect, we compress the distorted image D to the MPRIs P_{qi} by adopting the JP2K encoder:

$$P_{qi} = \mathrm{JP2K}(D, Q_i)$$

where JP2K means the JP2K encoder and Q_i is used to change the compression ratio. In total, the subscripts n, b, j, q denote noising, blurring, blocking and ringing effects, respectively. In addition, the degrees of distortion aggravation are divided into five levels for each distortion type, which means $i = 1, 2, \cdots, 5$ in this work.

After distortion aggravation, we carry out the process of extracting middle-level features based on the target distorted image and its corresponding MPRIs. Ding et al.[42] find that only calculating the spatial means and variances of feature maps in the convolutional neural network receive an efficient parametric model towards visual quality. Thus, in this work, we employ a VGG network in the target distorted images and their corresponding MPRIs and calculate the mean and variance in each feature map of the VGG network as well as the input image. Specifically, the MPRI connected to the convolution responses of five corresponding VGG layers is composed of the representation:

$$R(x) = \left\{ \tilde{x}_j^{(i)}; i = 0, \cdots, m; j = 1, \cdots, k_i; x = 1, \cdots, n \right\}$$

where $m = 5$ in this work, which means the number of convolution layers of R and k_i denote the number of feature maps in the i-th convolution layer. $R(x)$ is the representation for the x-th MPRI. Similarly, we can also derive the representation for the target distorted image:

$$R(d) = \left\{ \tilde{d}_j^{(i)}; i = 0, \cdots, m; j = 1, \cdots, k_i \right\}.$$

After that, the quality features extracted from $R(x)$ and $R(d)$ are required to be specified. Inspired by the features in SSIM[4], we calculate the quality features of the texture and structure of each pair of the feature maps of the target image and its corresponding MPRI based on the global means and variances:

$$t(\tilde{x}_j^{(i)}, \tilde{d}_j^{(i)}) = \frac{2\mu_{\tilde{x}_j}^{(i)} \mu_{\tilde{d}_j}^{(i)} + c_1}{\left(\mu_{\tilde{x}_j}^{(i)}\right)^2 + \left(\mu_{\tilde{d}_j}^{(i)}\right)^2 + c_1}$$

$$y(\tilde{x}_j^{(i)}, \tilde{d}_j^{(i)}) = \frac{2\sigma_{\tilde{x}_j \tilde{d}_j}^{(i)} + c_2}{\left(\sigma_{\tilde{x}_j}^{(i)}\right)^2 + \left(\sigma_{\tilde{d}_j}^{(i)}\right)^2 + c_2}$$

where t and y denote the similarities of the global means (texture features) and global correlation (structure features), respectively. $\mu_{\tilde{x}_j}^{(i)}$, $\mu_{\tilde{d}_j}^{(i)}$, $\sigma_{\tilde{x}_j}^{(i)}$, $\sigma_{\tilde{d}_j}^{(i)}$ and $\sigma_{\tilde{x}_j \tilde{d}_j}^{(i)}$ indicate the global means and variances of $\tilde{x}_j^{(i)}$ and $\tilde{d}_j^{(i)}$, as well as the global covariance between $\tilde{x}_j^{(i)}$ and $\tilde{d}_j^{(i)}$, respectively. c_1 and c_2 are two small constants to prevent instabilities when the denominators are close to zero.

Finally, based on the structure features of the target distorted image and its corresponding MPRI at different convolution layers, the middle-level visual features M are extracted:

$$M(x, d) = 1 - \sum_{i=0}^{m} \sum_{j=1}^{k_i} \left(\xi_{ij} t(\tilde{x}_j^{(i)}, \tilde{d}_j^{(i)}) + \delta_{ij} y(\tilde{x}_j^{(i)}, \tilde{d}_j^{(i)}) \right)$$

where $\{\xi_{ij}, \delta_{ij}\}$ represents the positive learnable weights, which satisfy $\sum_{i=0}^{m} \sum_{j=1}^{k_i} (\xi_{ij} + \delta_{ij}) = 1$.

2.3 Feature extraction in the high-level visual layer

After discussing the extraction of low-level visual features and middle-level visual features, in this section we will explore and analyze the high-level visual feature extraction. Since the high-level visual features take the global image as a whole, we need to seek a model aiming at the whole image to extract the features. We thoroughly investigate the visual perception models of the human brain and attempt to characterize the quality of image from the high-level visual perception in HVS.

Specifically, we employ the free-energy principle method, which unifies several findings in brain theory and neuroscience, to simulate the process of human action, perception and learning[43]. A fundamental theory of the free-energy principle is that the process of cognition or comprehending is an active inference behavior managed by an internal generative model (IGM) in the human brain[44]. When a "surprise", such as an image signal, transmits to the human brain via the retina, the brain will spontaneously produce the useful part of the information and ignore the redundant uncertain components for explaining sensations using this IGM[20]. The perceptual quality of the input thus has high correlation with the discrepancy between the input image and its corresponding representation generated by IGM[21]. Since IGM yields the perception of the visual signals based on the integrated input image, free-energy based features are regarded as high-level visual features in this work.

For mathematical formulation, we adopt $\mathcal{K}$ to represent the internal generative model. Also, we assume that the process of visual perception is parametric, which adjusts the parameter vector θ to explain visual scenes. Given the input image I, the joint distribution $P(I, \theta|\mathcal{K})$ with the model parameters vector θ can measure the value of free-energy:

$$-\log P(I|\mathcal{K}) = -\log \int P(I,\theta|\mathcal{K})\mathrm{d}\theta.$$

To simplify this mathematical expression, an auxiliary term $Q(\theta|I)$ is introduced to both the numerator and denominator of the above equation. Using Jensen's inequality and dropping the generative model $\mathcal{K}$ in order to make the formula clear, we can alter this equation to:

$$-\log P(I|\mathcal{K}) \leq -\int Q(\theta|I)\log\frac{P(I,\theta)}{Q(\theta|I)}\mathrm{d}\theta.$$

Afterwards, we can regard the right side of the above equation as the free energy according to the knowledge of statistical physics and thermodynamics[45]:

$$F(\theta) = -\int Q(\theta|I)\log\frac{P(I,\theta)}{Q(\theta|I)}\mathrm{d}\theta.$$

Notice that $P(I,\theta) = P(\theta|I)P(I)$, we can further infer the above equation as

$$F(\theta) = -\log P(I) + KL(Q(\theta|I)||P(\theta|I))$$

where $KL(\cdot)$ denotes the Kullback-Leibler divergence between the approximate posterior and the true posterior distributions. A more detailed derivation of free energy can be found in [20].

Since the human brain is extremely complicated and far beyond our current knowledge, the explicit expression model of the free-energy has not yet been developed. To solve this problem, some research attempts to approximate the free-energy calculating model using existing models for simulating image perception of the human brain. In some earlier works[20, 21], the linear auto-regressive (AR) model is employed to acquire the approximation $F(\hat{\theta})$ of the free energy $F(\theta)$. However, the calculation process of the AR model is too complex, which leads to a relatively long time for feature extraction. Based on the neurobiological findings, sparse representation is suitable for denoting natural images that agree with some properties, such as spatial localization, orientation and bandpass in the mammalian primary visual cortex of the brain [46]. Thus, Liu et al.[47] and Zhu et al.[48] use a sparse representation method to approximately express the free energy. The performance of the sparse representation method is demonstrated to be more efficient and effective than the linear AR model for predicting the quality of images. Therefore, in this paper, we employ the sparse representation model to approximate the IGM.

Specifically, given the input image I, we first select a patch $x_s \in \mathbf{R}^B$ from it with an extraction operator $O_s(\cdot)$, where B denotes the size of the patch. Then, the sparse representation of x_s over an over-complete dictionary $D \in \mathbf{R}^{B\times U}$ is equal to compute a vector $a_s \in \mathbf{R}^U$ to represent x_s, which can be indicated as

$$a_s^* = \arg\min_{a_s} \frac{1}{2}||x_s - Da_s||_2 + \lambda||a_s||_p$$

where D is the dictionary that can be expressed as $[d_1,\cdots,d_U]$, $a_s \in \mathbf{R}^U$ is the representation coefficient vector of the extracted patch and U represents the number of atoms in the sparse representation model. λ is a positive constant used to balance the weight of the reconstruction fidelity constraint term and the sparse punishment term. Moreover, $||\cdot||_p$ represents the l^p norm. From the above formula, the sparse vector a_s^* for representing x_s can be obtained. After that, the sparse representation of the whole input image I can be expressed as

$$\hat{I} = \sum_{s=1}^{n_p} O_s^{\mathrm{T}}(Da_s^*)./\sum_{s=1}^{n_p} O_s^{\mathrm{T}}(1_B)$$

where $\hat{I}$ is the sparse representation of the entire image I, which is regarded as the representation of I in human brain. "./" means the element-wise division of two matrices and n_p refers to the number of patches. $O_s^{\mathrm{T}}(\cdot)$ represents the transpose operation of $O_s(\cdot)$ and 1_B denotes the vector whose values are all 1 with the size of B.

According to the above-mentioned analysis, the free energy indicates a discrepancy between the input image and its best prediction image by the IGM. Thus, free energy can be considered as a natural proxy for the quality of perceptions. Based on the expression of free energy, the prediction residual of input image I is defined as

$$RE = \left|I - \hat{I}\right|$$

where RE refers to the prediction residual of input image I and $|\cdot|$ is the magnitude operation. After that, the uncertainty of RE can be obtained by measuring its entropy:

$$H = -\sum_{i=0}^{255} p_i \log_2 p_i$$

where

$$\widehat{\theta} = \arg\min_{\theta} H(\theta|\mathcal{K}, I)$$

and H shows the entropy of RE, which is also regarded as the value of free energy. p_i refers to the probability density of the i-th gray scale in RE.

For illustrating the effectiveness of the free energy feature on describing image quality intuitively, the distorted images generated from two reference images are selected from the TID2013 database[32]. As shown in Figs. 6(a) and 6(b), these two images have different image complexity in that Fig. 6(a) possesses simple image content and Fig. 6(b) has complicated texture information. Two com-

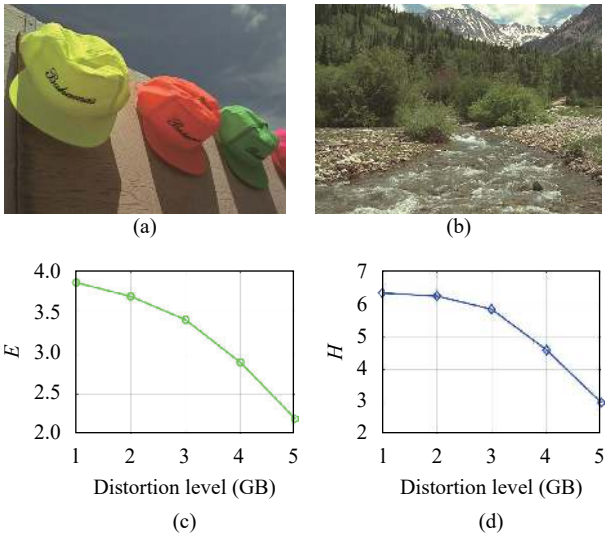


Fig. 6 Relationship between high-level visual feature H and distortion levels with different types. (a) and (b) are two reference images selected from the TID2013 database. (c) shows the visual feature H of (a) over different distortion levels distorted by GBlur. (d) shows the visual feature H of (b) over different distortion levels degraded by JPEG.

mon distortion types, GBlur and JPEG compression and five distortion levels are employed. The relationship between high-level visual feature H and distortion levels with different types are illustrated in Fig. 6. As exhibited, it can be observed that the values of H reduce gradually with the deepening of degradation. Based on the great capacity of free energy features to measure degradations of image quality effectively and its high-level visual properties, we select the free-energy feature H to be the high-level visual feature in this work.

2.4 Quality evaluation

After extracting the quality-aware features from low-level, middle-level and high-level visual layers, we need to seek an appropriate mapping to learn the subjective MOS values from the feature space using the regression module, and then employ it to produce objective quality scores. A total of 33 features are extracted from the three visual layers, as shown in Table 1. Based on the number of features and the effectiveness of regressors, we adopt SVR[49] to aggregate the quality-aware features, which has been widely used in the NR IQA field[21, 33].

Specifically, given the training set Φ, the quality-aware features $\mathcal{F}_i$ and the corresponding subjective quality labels q_i (MOS) of the images are employed to train the model:

$$\text{model} = \text{SVR\_TRAIN}([\mathcal{F}_i], [q_i], I_i \in \Phi)$$

where $\mathcal{F}_i$ is composed of the low-level visual features L_i, the middle-level visual features M_i, and the high-level visual features H_i of the training image I_i in the training set Φ. Then, we can utilize this regressor to predict the quality score of any target image with its corresponding feature $\mathcal{F}$:

$$V = \text{SVR\_PERDICT}([\mathcal{F}], \text{model})$$

where V stands for the predicted objective quality score of the target image. In this work, the LIBSVM package[50] is utilized with a radial basis function (RBF) kernel to teach our proposed model.

Table 1 Summary of the quality-aware features extracted from three visual layers

Layer	Category	Number
Low-level	NSS-based features	12
Middle-level	VGG-based features	20
High-level	Free-energy feature	1

3 Experimental results and analysis

In this section, we first compare the performance of our proposed method with the performance of the popular NR IQA models on three common large-scale image databases: LIVE[30], CSIQ[31] and TID2013[32] for validating the proposed NSCHM quality metric. The four most mainstream distortion types that we mentioned above: AWGN, GBlur, JPEG and JP2K are employed in the experiment and distortion type of the Rayleigh fast-fading channel simulation (FF) in the LIVE database[30] is also included. The performance on single distortion types is also discussed. In addition, we analyze the robustness of our proposed method through cross-validation under mismatched conditions. Finally, the ablation experiment is employed to demonstrate the effect of features in different visual layers.

3.1 Parameter settings and training procedure

In the process of exacerbating the distortion in the middle-level visual feature extraction, the distorted image is degraded by AWGN, GBblur, JPEG and JP2K distortions with five degradation levels for each type. We employ the Matlab to apply these four distortions and the specific parameters are as follows. The five Gaussian kernels of AWGN with standard deviations are from 0.5 to 2.5 with a step of 0.5, the five variances of GBlur are from 0.3 to 0.7 with a step of 0.1 and the five quality parameters of the JPEG encoder are from 0 to 8 with a step of 2 as well as the five compression ratios of JPEG2000 encoder are from 150 to 250 with a step of 25. In addition, since the perceptual weights ξ and δ are undetermined, we train the VGG-based representation model on the KADID dataset[51] to learn ξ and δ.

In the part of the sparse representation in the high-level visual feature extraction, the predefined dictionary

adopts an overcomplete discrete cosine transform (DCT) dictionary with the size of 64×144 which includes 144 atoms for sparse representation. The size B of each patch vector is set to 64. The orthogonal matching pursuit (OMP) algorithm[52] is used to work out the optimization problem of sparse representation.

Since the model we proposed requires training, we randomly divided the distorted images in each testing databases into two parts: a training set and a testing set, which respectively include 80% and 20% of the images. We train our proposed algorithm using the training set and measure its performance with the testing set. This 80% train − 20% test process is repeated one thousand times to guarantee the robustness of our metric[53]. The median results over these one thousand iterations are selected as the final performance to avoid the performance bias as much as possible.

3.2 Experimental protocol

1) Databases: For examining the performance of the proposed model, three widely used IQA databases are employed as testbeds, including LIVE[30], CSIQ[31] and TID2013[32]. A brief introduction of these three databases is presented below:

The LIVE database[30] is released by the University of Texas at Austin, and is the most famous IQA database. It contains 770 lossy images generated from 29 pristine images by degrading them with five different types of distortions: AWGN, GBlur, JPEG, JP2K and FF.

The CSIQ database[31] is provided at Oklahoma State University including 886 images created from 30 original images. Six types of distortions are considered in the CSIQ database, which are GBlur, AWGN, JPEG, JP2K, global contrast decrements (CC) and additive pink Gaussian noise (APGN) at four or five distortion levels respectively.

The TID2013 database[32] is the updated version of the TID2008 database, which is developed with a joint international cooperation among Finland, Italy and Ukraine. This database consists of 3 000 distorted images generated by corrupting 25 reference ones with 24 distortion types at five distinct distortion levels.

2) Comparing algorithms: Eight popular IQA algorithms are compared with our proposed NSCHM metric, which are DIIVINE[13], BLINDS2[54], BRISQUE[33], NIQE[14], QAC[55], IL-NIQE[56], LPSI[57] and BPRI[15]. In these NR models, DIIVINE, BLINDS2 and BRISQUE are opinion-aware models which need to be trained to integrate the NSS features extracted from the wavelet domain, DCT domain and spatial domain, respectively. The rest of them are opinion-unaware models, where NIQE and IL-NIQE are based on spatial domain NSS, QAC learns a codebook to achieve quality-aware clustering, LPSI uses local image structure statistics and BPRI utilizes a local binary pattern.

3) Evaluation criteria: Four commonly used evaluation criteria are applied to measure the performance of the compared IQA metrics, including spearman rank-order correlation coefficient (SRCC), Kendall's rank-order correlation coefficient (KRCC), Pearson linear correlation coefficient (PLCC) and root mean squared error (RMSE)[58, 59]. The mathematical expressions of these four measurements are as follows:

$$\mathrm{SRCC} = 1 - \frac{6\sum_{i=1}^{Z} d_i^2}{Z(Z^2-1)}$$

$$\mathrm{KRCC} = \frac{Z_c - Z_d}{\frac{1}{2}Z(Z-1)}$$

$$\mathrm{PLCC} = 1 - \frac{\sum_{i=1}^{Z}(p_i-\bar{p})(q_i-\bar{q})}{\sqrt{\sum_{i=1}^{Z}(p_i-\bar{p})^2(q_i-\bar{q})^2}}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{Z}\sum_{i=1}^{Z}(p_i-q_i)^2}$$

where d_i represents the difference between the ranks of the i-th images in subjective and objective assessments, and Z denotes the number of images in testing data set. Z_c and Z_d mean the numbers of concordant and discordant pairs in the testing database. p_i and q_i indicate the converted objective score and subjective score of the i-th image after the nonlinear regression. $\bar{p}$ and $\bar{q}$ are the means of all p_i and q_i. Specifically, SRCC represents the prediction monotonicity by only considering the relative orders between the inputs, and KRCC is another monotonicity index employed to evaluate the association between the data. PLCC describes the prediction linearity of an IQA metric and RMSE indicates the prediction accuracy. A good IQA measure is expected to acquire high values, which close to 1, in SRCC, KRCC and PLCC, yet the low values, which near 0, in RMSE.

Furthermore, following the suggestions of the video quality experts group (VQEG)[60], PLCC and RMSE cannot calculate performance by using the subjective scores and the corresponding objective ratings directly. According to the guidance of VQEG, we adopt a regression analysis to conduct a nonlinear mapping between the subjective MOSs and the corresponding objective ratings predicted by target IQA metrics. For the nonlinear regression, a monotonic logistic function of five parameters $\{\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5\}$ is employed:

$$f(x) = \zeta_1\left(0.5 - \frac{1}{1+\mathrm{e}^{\zeta_2(x-\zeta_3)}}\right) + \zeta_4 x + \zeta_5$$

where x and $f(x)$ represent the raw input ratings and mapped scores, and $\zeta_j, j = 1, 2, 3, 4, 5$ stand for five parameters to be ascertained during the process of the nonlinear fitting.

3.3 Overall performance comparison

First, we compare the overall performance of our proposed algorithm with the above-mentioned eight state-of-the-art NR IQA models on three widely used databases: LIVE, CSIQ and TID2013. For a fair comparison, we retrain the opinion-aware algorithms: DIIVINE, BLIINDS2 and BRISQUE, as well as our proposed method on the same training set and measure them on the testing set of each database. For the remaining models, we employ the same testing set to test their performance. The overall performance in terms of SRCC, KRCC, PLCC and RMSE are tabulated in Table 2, where the three top-performing models are highlighted.

It is observed that our proposed algorithm shows great comprehensive performance and achieves the top three positions on all databases in terms of various criteria. By comparison, DIIVINE, NIQE and LPSI show relatively moderate performance on three databases. BRISQUE demonstrates good performance in LIVE and BLIINDS2 has high correlation with the subjective scores on LIVE and CSIQ. Another observation is that IL-NIQE and BPRI achieve great prediction performance in TID2013 and CSIQ, respectively. These experiments clearly demonstrate that our proposed method has high stability and superiority in assessing the perceived quality of images.

3.4 Performance on different distortion types

In addition to testing the overall performance of algorithms on individual databases, we also examine the prediction performance of all NR IQA metrics on individual distortions. The same training-testing process described in Section 3.3 is implemented. The 80% degraded images in the training set are all employed to train the models, while only images with the target distortion type in the testing set selected from the rest 20% distorted images are applied to test. The mean results of our proposed method and the compared blind IQA models on single distortion types are summarized in Table 3. The three best performances of each distortion type on different databases is highlighted with boldface. For simplicity, we only list SRCC values in Table 3, but we can acquire similar evaluation results with other evaluation criteria.

From Table 3, it can be clearly observed that the competition between each NR IQA algorithm is more intense, and each metric has its own advantages. Specifically, our proposed NSCHM is also comparable to these popular metrics when performed on individual distortions, which is consistent with the results of the overall performance evaluation introduced in Section 3.3. In addition, we can find that BRISQUE obtains the best results on the LIVE and has relatively mediocre performance on the TID2013, while BPRI and LPSI perform much better on CSIQ and TID2013. Furthermore, our proposed model shows more stable performance than other NR measures, and NSCHM has no SRCC value lower than 0.88 for a single distortion type. Our NSCHM metric has no obvious weakness in these four common distortion types on three popular databases.

3.5 Cross-validation under mismatched conditions

In Sections 3.3 and 3.4, the performance of the NR algorithms is based on the training-testing procedure on the same database. Thus, in this section, we attempt to carry out cross-validation experiments to test the robustness of our proposed method under mismatched conditions. We use LIVE, CSIQ and TID2013 databases as the training set respectively, and then employ the corresponding remaining two databases as the testing set. The results are

Table 2 Overall performance comparison of the ten popular IQA methods and our proposed metric on LIVE, CSIQ and TID2013 databases. We highlight the three top-performing models in each row.

Database	Metric	DIIVINE[13]	BLIINDS2[54]	BRISQUE[33]	NIQE[14]	QAC[55]	IL-NIQE[56]	LPSI[57]	BPRI[15]	NSCHM(pro.)
LIVE	SRCC	0.869 7	**0.9187**	**0.943 6**	0.908 8	0.872 3	0.902 1	0.819 9	0.908 2	**0.948 3**
	PLCC	0.879 9	**0.926 8**	**0.947 2**	0.649 5	0.868 2	0.711 1	0.826 1	0.896 6	**0.953 1**
	KRCC	0.689 6	0.761 5	**0.800 0**	0.734 2	0.680 2	0.724 5	0.625 1	**0.743 5**	**0.806 0**
	RMSE	12.746 1	**10.401 0**	**8.786 6**	20.565 6	13.429 3	19.163 8	15.289 5	12.119 4	**8.303 3**
CSIQ	SRCC	0.863 4	**0.897 7**	0.866 9	0.887 6	0.841 0	0.888 5	0.780 8	**0.902 8**	**0.906 1**
	PLCC	0.897 5	**0.922 5**	0.896 1	0.907 2	0.874 5	0.920 6	0.872 9	**0.924 2**	**0.932 8**
	KRCC	0.683 4	**0.728 2**	0.698 8	0.705 5	0.651 3	0.710 9	0.598 5	**0.735 1**	**0.738 3**
	RMSE	0.125 2	0.108 2	0.128 4	0.118 7	0.135 5	**0.107 3**	0.137 1	**0.105 7**	**0.099 2**
TID2013	SRCC	0.751 3	0.839 5	0.863 4	0.799 5	0.859 1	**0.875 7**	0.716 9	**0.899 5**	**0.915 4**
	PLCC	0.793 9	0.880 6	**0.893 1**	0.812 4	0.869 8	**0.893 4**	0.814 7	0.893 0	**0.928 8**
	KRCC	0.588 0	0.654 7	0.682 6	0.598 0	0.660 9	**0.686 2**	0.512 7	**0.719 3**	**0.751 9**
	RMSE	8.379 1	6.621 5	6.306 7	8.090 1	7.001 2	**6.275 8**	8.153 2	**6.295 9**	**5.328 5**

Table 3 SRCC values of our NSCHM and other IQA metrics in various individual distortion types on LIVE, CSIQ and TID2013 databases. We highlight the three top-performing models with boldface.

Database	Metric	DIIVINE[13]	BLIINDS2[54]	BRISQUE[33]	NIQE[14]	QAC[55]	IL-NIQE[56]	LPSI[57]	BPRI[15]	NSCHM(pro.)
LIVE	AWGN	0.960 8	0.944 2	**0.984 0**	0.972 4	0.948 8	**0.980 4**	0.957 5	**0.982 9**	0.977 8
	GBlur	0.857 6	0.909 7	**0.953 3**	**0.937 5**	0.923 5	0.929 5	0.930 1	0.937 5	**0.945 9**
	JPEG	0.881 2	0.949 0	**0.966 2**	0.943 1	0.947 5	0.941 3	**0.970 6**	**0.968 5**	0.951 3
	JP2K	0.815 1	**0.933 6**	0.911 2	**0.925 4**	0.890 1	0.905 7	**0.938 6**	0.923 7	0.924 8
	FF	0.791 1	08456	**0.877 2**	**0.861 6**	0.829 6	0.823 4	0.785 5	0.840 9	**0.905 5**
CSIQ	AWGN	0.805 1	0.868 9	**0.902 7**	0.837 3	0.822 5	0.867 9	0.734 8	**0.943 6**	**0.885 5**
	GBlur	0.882 9	**0.917 2**	0.894 9	**0.9113**	0.840 5	0.870 5	**0.915 9**	0.909 5	0.906 3
	JPEG	0.883 4	0.904 1	0.903 7	0.890 9	0.908 5	0.908 4	**0.954 1**	**0.933 3**	**0.909 1**
	JP2K	0.853 8	0.901 4	08274	**0.926 3**	0.875 8	**0.922 5**	**0.928 8**	0.879 7	0.904 8
TID2013	AWGN	0.664 7	0.702 3	0.848 0	0.859 4	0.754 7	**0.888 5**	0.833 1	**0.930 4**	**0.888 1**
	GBlur	0.847 7	0.845 4	0.873 8	0.796 1	**0.8835**	0.841 2	**0.896 5**	0.878 5	**0.895 4**
	JPEG	0.669 9	0.818 8	0.851 1	0.857 6	0.876 5	0.861 5	**0.928 4**	**0.9223**	**0.901 7**
	JP2K	0.796 3	0.875 9	0.862 4	0.888 7	0.891 2	**0.907 7**	**0.902 2**	0.890 0	**0.898 9**

shown in Table 4. It can be observed that although the performance declines compared with the performance under mismatched conditions, it still maintains moderate results without serious deviation, which is within the acceptable range. Therefore, the independence of our proposed algorithm is favourable.

To demonstrate that our algorithm also has acceptable performance under the mismatched conditions, we compare NSCHM with other competitive algorithms in this section. For the fairness of this experiment, we select the opinion-aware algorithms, which are DIIVINE, BLIINDS2 and BRISQUE as well as a state-of-the-art training algorithm NFERM[21] to compare with our proposed method. We employ the TID2013 database as the training set and measure the performance of these metrics on LIVE and CSIQ databases. The performance results are demonstrated in Table 5. It is obvious that our proposed algorithm has advantages compared with other opinion-aware algorithms. The results of PLCC, KRCC and RMSE in LIVE as well as the results of PLCC and RMSE achieve the best performance among these models. In addition, there are no relatively poor results for each sub-item indicating that the robustness of our algorithm is good.

3.6 Statistical significance analysis

For computing the statistical significance of our proposed NSCHM with these compared algorithms, we employ a t-test to measure prediction residuals between the converted objective ratings after the nonlinear fitting of different NR IQA models and subjective scores. The pairwise t-test evaluations are performed on LIVE, CSIQ and TID2013, respectively. The statistical significance results are listed in Table 6, where the symbols "1", "0" and "−1" indicate that the proposed measure is statistically

Table 4 Cross-validation experiments under mismatched conditions using LIVE, CSIQ and TID2013 databases

Database	Metric	LIVE	CSIQ	TID2013
LIVE	SRCC	–	0.843 5	0.707 7
	PLCC	–	0.870 5	0.777 3
	KRCC	–	0.676 6	0.586 7
	RMSE	–	0.122 7	8.008 7
CSIQ	SRCC	0.698 1	–	0.765 6
	PLCC	0.702 6	–	0.896 0
	KRCC	0.519 1	–	0.586 7
	RMSE	15.788 1	–	5.651 9
TID2013	SRCC	0.744 4	0.858 4	–
	PLCC	0.773 0	0.912 4	–
	KRCC	0.567 4	0.681 2	–
	RMSE	14.075 6	0.102 0	–

Table 5 Performance results of our NSCHM and other NR metrics in cross-validation experiments under mismatched conditions. TID2013 database is employed as the training set and LIVE and CSIQ databases are applied to test the models.

Database	Metric	SRCC	PLCC	KRCC	RMSE
LIVE	DIIVINE[13]	0.540 7	0.617 8	0.404 4	17.447 0
	BLIINDS2[54]	0.700 9	0.718 6	0.537 5	15.428 5
	BRISQUE[33]	0.468 1	0.573 5	0.367 8	18.175 9
	NFERM[21]	**0.753 7**	0.737 4	0.555 6	14.985 0
	NSCHM(pro.)	0.744 4	**0.773 0**	**0.567 4**	**14.075 6**
CSIQ	DIIVINE[13]	**0.890 4**	0.910 8	**0.718 1**	0.102 9
	BLIINDS2[54]	0.702 2	0.735 6	0.520 1	0.168 9
	BRISQUE[33]	0.878 4	0.886 6	0.713 5	0.115 3
	NFERM[21]	0.723 8	0.847 8	0.547 8	0.132 2
	NSCHM(pro.)	0.859 4	**0.912 4**	0.681 2	**0.102 0**

(with 95% confidence) better, imperceptible and worse than the corresponding NR IQA metrics in each column. From Table 6, it is easy to find that NSCHM is superior to all competitive NR IQA models on the LIVE database and has great advantages compared with other competitors on the TID2013 database, where only IL-NIQE and BPRI are comparable to our method. In addition, although the performance on the CSIQ database is not as outstanding as that on the other two databases, no competitor algorithm is superior to NSCHM. Thus, this experiment demonstrates the advantage of NSCHM in evaluating the image quality statistically.

3.7 Ablation experiment

As described in Section 2, our proposed NSCHM consists of three groups of features, namely low-level visual features, middle-level visual features and high-level visual features. Therefore, it is interesting to analyze the contribution of each part to the overall algorithm. We conduct the ablation study on the LIVE, CSIQ and TID2013 databases. For quantitative analysis, we compute the median values of SRCC, PLCC, KRCC and RMSE via the same 80% train − 20% test process described above for each group of features. In addition, in order to make a more detailed division, we divide the low-level visual characteristics into the MSCN coefficient features and adjacent MSCN coefficients features. The performance of each feature group on different databases is demonstrated in Table 7. In Table 7, LOW1 and LOW2 stand for the MSCN coefficient features and adjacent MSCN coefficients features in low-level visual features, respectively. MIDDLE and HIGH denote the features extracted from the middle-level and high-level visual layers. It is observed that each set of features has favourable performance, with LOW1 and MIDDLE performing better and LOW2 and HIGH performing relatively worse. Another observation is that the performance of each set of features is inferior to the final proposed algorithm, which means that each set of groups has its own impact on improving the predicted accuracy of our proposed metric in evaluating the perceived quality of images.

4 Conclusions

In this paper, a novel perceptual NR IQA metric named NSCHM is proposed based on structural computational modeling of HVS. The proposed metric is inspired by the fact that the human brain processes visual stimuli in a hierarchical manner. We first analyze the process of the human brain to handle the images and introduce the framework of structured computing model. After that, three groups of features are extracted, which are the low-level visual features at the pixel level, the middle-level visual features at the primitive level and the high-level visual features at the global image level, respectively. Then, we employ SVR to integrate these three feature groups and predict the image quality ratings. Validation experiments are conducted on three widely used IQA databases, i.e., LIVE, CSIQ and TID2013, demonstrating that NSCHM has outstanding performance with state-of-the-art NR methods in overall performance comparison. For individual distortion types, our metric still maintains favourable performance. The cross validation experiments testify the stable performance of NSCHM under mismatched conditions.

Acknowledgements

This work was supported by National Natural Science

Table 6 Statistical significance comparison between NSCHM and other competing NR IQA models. The prediction residuals between NSCHM and subjective scores are compared with the prediction residuals between the NR IQA algorithms and subjective ratings employing the t-test. The symbols "1", "0" and "−1" indicate that the proposed measure is statistically (with 95% confidence) better, imperceptible and worse than the corresponding NR IQA metrics.

NR IQA model Database	DIIVINE	BLIINDS2	BRISQUE	NIQE	QAC	IL-NIQE	LPSI	BPRI
LIVE	1	1	1	1	1	1	1	1
CSIQ	1	0	0	0	1	0	1	1
TID2013	1	1	1	1	1	0	1	0

Table 7 Performance of the ablation study measured by SRCC, PLCC, KRCC and RMSE on LIVE, CSIQ and TID2013 databases. LOW1 and LOW2 mean the MSCN coefficient features and adjacent MSCN coefficients features in low-level visual features, respectively. MIDDLE and HIGH denote the features extracted from middle-level and high-level visual layers.

Metric	LIVE				CSIQ				TID2013			
	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE
LOW1	0.822 9	0.833 7	0.640 7	14.923 6	0.907 8	0.932 0	0.736 5	0.099 7	0.863 5	0.875 5	0.679 2	6.695 4
LOW2	0.747 1	0.779 0	0.560 4	16.980 2	0.742 9	0.806 1	0.555 5	0.168 9	0.711 5	0.774 5	0.520 6	8.858 2
MIDDLE	0.907 3	0.921 2	0.738 6	10.739 8	0.892 3	0.921 3	0.714 9	0.109 1	0.848 3	0.881 5	0.663 0	6.578 2
HIGH	0.643 1	0.760 9	0.469 2	17.655 4	0.737 7	0.831 6	0.551 0	0.154 9	0.651 7	0.762 6	0.467 9	8.988 7

Foundation of China (Nos. 61831015 and 61901260), Key Research and Development Program of China (No. 2019YFB1405902).

Open Access

References

[1] G. T. Zhai, X. K. Min. Perceptual image quality assessment: A survey. *Science China Information Sciences*, vol. 63, no. 11, Article number 211301, 2020. DOI: 10.1007/s11432-019-2757-1.

[2] W. S. Lin, C. C. J. Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011. DOI: 10.1016/j.jvcir.2011.01.005.

[3] Z. Wang, A. C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009. DOI: 10.1109/MSP.2008.930649.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, 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. DOI: 10.1109/TIP.2003.819861.

[5] H. R. Sheikh, A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006. DOI: 10.1109/TIP.2005.859378.

[6] D. M. Chandler, 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. DOI: 10.1109/TIP.2007.901820.

[7] K. Gu, L. D. Li, H. Lu, X. K. Min, W. S. Lin. A fast reliable image quality predictor by fusing micro- and macro-structures. *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 3903–3912, 2017. DOI: 10.1109/TIE.2017.2652339.

[8] Z. Wang, G. X. Wu, H. R. Sheikh, E. P. Simoncelli, E. H. Yang, A. C. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1680–1689, 2006. DOI: 10.1109/TIP.2005.864165.

[9] M. Liu, K. Gu, G. T. Zhai, P. Le Callet, W. J. Zhang. Perceptual reduced-reference visual quality assessment for contrast alteration. *IEEE Transactions on Broadcasting*, vol. 63, no. 1, pp. 71–81, 2017. DOI: 10.1109/TBC.2016.2597545.

[10] R. Soundararajan, A. C. Bovik. RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 517–526, 2012. DOI: 10.1109/TIP.2011.2166082.

[11] X. K. Min, K. Gu, G. T. Zhai, M. H. Hu, X. K. Yang. Saliency-induced reduced-reference quality index for natural scene and screen content images. *Signal Processing*, vol. 145, pp. 127–136, 2018. DOI: 10.1016/j.sigpro.2017.10.025.

[12] Y. T. Liu, K. Gu, S. Q. Wang, D. B. Zhao, W. Gao. Blind quality assessment of camera images based on low-level and high-level statistical features. *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 135–146, 2019. DOI: 10.1109/TMM.2018.2849602.

[13] A. K. Moorthy, A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011. DOI: 10.1109/TIP.2011.2147325.

[14] Mittal, R. Soundararajan, A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013. DOI: 10.1109/LSP.2012.2227726.

[15] X. K. Min, K. Gu, G. T. Zhai, J. Liu, X. K. Yang, C. W. Chen. Blind quality assessment based on pseudo-reference image. *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018. DOI: 10.1109/TMM.2017.2788206.

[16] Y. T. Liu, K. Gu, Y. B. Zhang, X. Li, G. T. Zhai, D. B. Zhao, W. Gao. Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 929–943, 2020. DOI: 10.1109/TCSVT.2019.2900472.

[17] P. Ye, J. Kumar, L. Kang, D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Providence, USA, pp. 1098−1105, 2012. DOI: 10.1109/CVPR.2012.6247789.

[18] J. T. Xu, P. Ye, Q. H. Li, H. Q. Du, Y. Liu, D. Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016. DOI: 10.1109/TIP.2016.2585880.

[19] J. Kim, S. Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, 2017. DOI: 10.1109/JSTSP.2016.2639328.

[20] G. T. Zhai, X. L. Wu, X. K. Yang, W. S. Lin, W. J. Zhang. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 41–52, 2012. DOI: 10.1109/TIP.2011.2161092.

[21] K. Gu, G. T. Zhai, X. K. Yang, W. J. Zhang. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2015. DOI: 10.1109/TMM.2014.2373812.

[22] Q. H. Li, W. S. Lin, J. T. Xu, Y. M. Fang. Blind image quality assessment using statistical structural and luminance features. *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2457–2469, 2016. DOI: 10.1109/TMM.2016.2601028.

[23] Q. H. Li, W. S. Lin, Y. M. Fang. BSD: Blind image quality

assessment based on structural degradation. *Neurocomputing*, vol. 236, pp. 93–103, 2017. DOI: 10.1016/j.neucom.2016.09.105.

[24] A. Saha, Q. M. J. Wu. Utilizing image scales towards totally training free blind image quality assessment. *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1879–1892, 2015. DOI: 10.1109/TIP.2015.2411436.

[25] N. Dijkstra, P. Zeidman, S. Ondobaka, M. A. J. Van Gerven, K. Friston. Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, vol. 7, no. 1, Article number 5677, 2017. DOI: 10.1038/s41598-017-05888-8.

[26] S. M. Kosslyn, G. Ganis, W. L. Thompson. Neural foundations of imagery. *Nature Reviews Neuroscience*, vol. 2, no. 9, pp. 635–642, 2001. DOI: 10.1038/35090055.

[27] F. Katsuki, C. Constantinidis. Bottom-up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist*, vol. 20, no. 5, pp. 509–521, 2014. DOI: 10.1177/1073858413514136.

[28] H. J. Park, K. Friston. Structural and functional brain networks: From connections to cognition. *Science*, vol. 342, no. 6158, Article number 1238411, 2013. DOI: 10.1126/science.1238411.

[29] K. E. Stephan. On the role of general system theory for functional neuroimaging. *Journal of Anatomy*, vol. 205, no. 6, pp. 443–470, 2004. DOI: 10.1111/j.0021-8782.2004.00359.x.

[30] H. R. Sheikh, Z. Wang, L. Cormack, A. C. Bovik. Live image quality assessment database release 2, [Online], Available: http://live.ece.utexas.edu/research/quality/, 2020.

[31] E. C. Larson, D. M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, vol. 19, no. 1, Article number 011006, 2010. DOI: 10.1117/1.3267105.

[32] N. Ponomarenko, L. N. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C. C. J. Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing*: *Image Communication*, vol. 30, pp. 57–77, 2015. DOI: 10.1016/j.image.2014.10.009.

[33] A. Mittal, A. K. Moorthy, A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012. DOI: 10.1109/TIP.2012.2214050.

[34] D. L. Ruderman. The statistics of natural images. *Network*: *Computation in Neural Systems*, vol. 5, no. 4, pp. 517–548, 1994. DOI: 10.1088/0954-898X_5_4_006.

[35] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[36] M. El Mallahi, A. Zouhri, H. Qjidaa. Radial meixner moment invariants for 2D and 3D image recognition. *Pattern Recognition and Image Analysis*, vol. 28, no. 2, pp. 207–216, 2018. DOI: 10.1134/S1054661818020128.

[37] T. Y. Lin, P. Dollár, R. Girshick, K. M. He, B. Hariharan, S. Belongie. Feature pyramid networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 936–944, 2017. DOI: 10.1109/CVPR.2017.106.

[38] X. Y. Gong, H. Su, D. Xu, Z. T. Zhang, F. Shen, H. B. Yang. An overview of contour detection approaches. *International Journal of Automation and Computing*, vol. 15, no. 6, pp. 656–672, 2018. DOI: 10.1007/s11633-018-1117-z.

[39] C. Ma, J. B. Huang, X. K. Yang, M. H. Yang. Hierarchical convolutional features for visual tracking. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Santiago, Chile, pp. 3074–3082, 2015. DOI: 10.1109/ICCV.2015.352.

[40] C. Ma, X. K. Yang, C. Y. Zhang, M. H. Yang. Long-term correlation tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 5388–5396, 2015. DOI: 10.1109/CVPR.2015.7299177.

[41] X. K. Min, G. T. Zhai, K. Gu, Y. T. Liu, X. K. Yang. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018. DOI: 10.1109/TBC.2018.2816783.

[42] K. Y. Ding, K. D. Ma, S. Q. Wang, E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. [Online], Available: //https://arxiv/abs/2004.07728, 2020.

[43] K. Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010. DOI: 10.1038/nrn2787.

[44] K. Friston, J. Kilner, L. Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, vol. 100, no. 1–3, pp. 70–87, 2006. DOI: 10.1016/j.jphysparis.2006.10.001.

[45] R. P. Feynman. *Statistical Mechanics: A Set of Lectures (Advanced Books Classics)*, Reading, USA: Westview Press, 1998.

[46] B. A. Olshausen, D. J. Field. How close are we to understanding v1? *Neural Computation*, vol. 17, no. 8, pp. 1665–1699, 2005. DOI: 10.1162/0899766054026639.

[47] Y. T. Liu, G. T. Zhai, X. M. Liu, D. B. Zhao. Perceptual image quality assessment combining free-energy principle and sparse representation. In *Proceedings of IEEE International Symposium on Circuits and Systems*, IEEE, Montreal, Canada, 2016, pp. 1586–1589. DOI: 10.1109/ISCAS.2016.7538867.

[48] W. H. Zhu, G. T. Zhai, X. K. Min, M. H. Hu, J. Liu, G. D. Guo, X. K. Yang. Multi-channel decomposition in tandem with free-energy principle for reduced-reference image quality assessment. *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2334–2346, 2019. DOI: 10.1109/TMM.2019.2902484.

[49] B. SchöLkopf, A. J. Smola, R. C. Williamson, P. L. Bartlett. New support vector algorithms. *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000. DOI: 10.1162/089976600300015565.

[50] C. C. Chang, C. J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, Article number 27, 2011. DOI: 10.1145/1961189.1961199.

[51] H. H. Lin, V. Hosu, D. Saupe. KADID-10K: A large-scale artificially distorted IQA database. In *Proceedings of the 11th International Conference on Quality of Multimedia Experience*, IEEE, Berlin, Germany, 2019, DOI: 10.1109/QoMEX.2019.8743252.

[52] J. A. Tropp, A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007. DOI: 10.1109/TIT.2007.909108.

[53] J. Demšar. Statistical comparisons of classifiers over mul-

tiple data sets. *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. DOI: 10.5555/1248547.1248548.

[54] M. A. Saad, A. C. Bovik, C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012. DOI: 10.1109/TIP.2012.2191563.

[55] W. F. Xue, L. Zhang, X. Q. Mou. Learning without human scores for blind image quality assessment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Portland, USA, pp. 995–1002, 2013. DOI: 10.1109/CVPR.2013.133.

[56] L. Zhang, L. Zhang, A. C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015. DOI: 10.1109/TIP.2015.2426416.

[57] Q. B. Wu, Z. Wang, H. L. Li. A highly efficient method for blind image quality assessment. In *Proceedings of IEEE International Conference on Image Processing*, IEEE, Quebec City, Canada, pp. 339–343, 2015. DOI: 10.1109/ICIP.2015.7350816.

[58] X. K. Min, G. T. Zhai, K. Gu, X. K. Yang, X. P. Guan. Objective quality evaluation of dehazed images. *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 2879–2892, 2019. DOI: 10.1109/TITS.2018.2868771.

[59] X. K. Min, K. D. Ma, K. Gu, G. T. Zhai, Z. Wang, W. S. Lin. Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5462–5474, 2017. DOI: 10.1109/TIP.2017.2735192.

[60] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. [Online], Available: https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phasi-i.aspx, 2000.

Wen-Han Zhu received the B. Eng. degree in electronic information engineering from Huazhong University of Science and Technology, China in 2015. He is currently a Ph. D. degree candidate in information and communication engineering with Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China.

His research interests include image quality assessment and image processing.

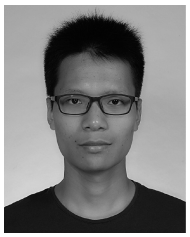
E-mail: zhuwenhan823@sjtu.edu.cn

ORCID iD: 0000-0001-8781-1110

Wei Sun received the B. Eng. degree in automation from East China University of Science and Technology, China in 2016. He is currently a Ph. D. degree candidate in control science and engineering at Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China.

His research interests include image quality assessment, perceptual signal processing and mobile video processing.

E-mail: sunguwei@sjtu.edu.cn



Xiong-Kuo Min received the B. Eng. degree in electronic and information engineering from Wuhan University, China in 2013, and the Ph. D. degree in information and communication engineering from Shanghai Jiao Tong University, China in 2018. From January 2016 to January 2017, he was a visiting student at Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a post-doctoral fellow with Shanghai Jiao Tong University. He received the Best Student Paper Award at IEEE ICME 2016.

His research interests include visual quality assessment, visual attention modeling and perceptual signal processing.

E-mail: minxiongkuo@sjtu.edu.cn

Guang-Tao Zhai received the B. Eng. degree in information science and engineering and M. Eng. degree in information science and engineering from Shandong University, China in 2001 and 2004, respectively, and the Ph. D. degree in communication and information system from Shanghai Jiao Tong University, China in 2009, where he is currently a research professor with Institute of Image Communication and Information Processing. From 2008 to 2009, he was a visiting student with Department of Electrical and Computer Engineering, McMaster University, Canada, where he was a post-doctoral fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012.

His research interests include multimedia signal processing and perceptual signal processing.

E-mail: zhaiguangtao@sjtu.edu.cn (Corresponding author)

ORCID iD: 0000-0001-8165-9322

Xiao-Kang Yang received the B. Sc. degree in physics from Xiamen University, China in 1994, the M. Sc. degree in physics from Chinese Academy of Sciences, China in 1997, and the Ph. D. degree in pattern recognition and intelligent system from Shanghai Jiao Tong University, China in 2000. He is currently a Distinguished Professor with the School of Electronic Information and Electrical Engineering, and the Deputy Director of the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, China. From 2000 to 2002, he was a Research Fellow with the Centre for Signal Processing, Nanyang Technological University, Singapore. From 2002 to 2004, he was a Research Scientist with the Institute for Infocomm Research, Singapore. From 2007 to 2008, he visited Institute for Computer Science, University of Freiburg, Germany, as an Alexander von Humboldt Research Fellow. He has published over 200 refereed papers, and has filed 60 patents. He is an Associate Editor of *IEEE Transactions on Multimedia* and a Senior Associate Editor of *IEEE Signal Processing Letters*. He was a Series Editor of Springer CCIS, and an Editorial Board Member of *Digital Signal Processing*. He is a member of Asia-Pacific Signal and Information Processing Association, the VSPC Technical Committee of the IEEE Circuits and Systems Society, and the MMSP Technical Committee of the IEEE Signal Processing Society. He is also Chair of the Multimedia Big Data Interest Group of MMTC Technical Committee, IEEE Communication Society.

His research interests include image processing and communication, computer vision, and machine learning.

E-mail: xkyang@sjtu.edu.cn