Compressed Image Quality Metric Based on Perceptually Weighted Distortion

Sudeng Hu, Lina Jin, Hanli Wang, *Senior Member, IEEE*, Yun Zhang, *Member, IEEE*, Sam Kwong, *Fellow, IEEE*, and C.-C. Jay Kuo, *Fellow, IEEE*

***Abstract*— Objective quality assessment for compressed images is critical to various image compression systems that are essential in image delivery and storage. Although the mean squared error (MSE) is computationally simple, it may not be accurate to reflect the perceptual quality of compressed images, which is also affected dramatically by the characteristics of human visual system (HVS), such as masking effect. In this paper, an image quality metric (IQM) is proposed based on perceptually weighted distortion in terms of the MSE. To capture the characteristics of HVS, a randomness map is proposed to measure the masking effect and a preprocessing scheme is proposed to simulate the processing that occurs in the initial part of HVS. Since the masking effect highly depends on the structural randomness, the prediction error from neighborhood with a statistical model is used to measure the significance of masking. Meanwhile, the imperceptible signal with high frequency could be removed by preprocessing with low-pass filters. The relation is investigated between the distortions before and after masking effect, and a masking modulation model is proposed to simulate the masking effect after preprocessing. The performance of the proposed IQM is validated on six image databases with various compression distortions. The experimental results show that the proposed algorithm outperforms other benchmark IQMs.**

***Index Terms*— Image quality assessment, compressed image, human visual system, masking effect, low-pass filter.**

I. Introduction

RELIABLE assessment of image quality is important in improving the performance of image processing systems. Due to the inconvenience of subjective image quality assessment, a large number of objective image quality metrics (IQM) have been developed. Generally, there are two different categories of objective IQMs. In the first category, the characteristics of Human Visual System (HVS) are explored and incorporated into IQM algorithms [1]–[8]. In [1], the luminance adaptation and the Contrast Sensitivity Function (CSF) of HVS are considered in human's perception to luminance difference. In [2], a wavelet CSF is employed and the distortion is analyzed in multiple channels after the wavelet transform. In [3], the Haar wavelet is used to model the space-frequency localization property of HVS responses. In [4], a model of noise detection threshold is proposed to determine the visibility of discrete wavelet transform noise in image compression, which is similar to the concept of just noticeable distortion (JND) [5]. In [6], the noise thresholds are determined on contrast via CSF, and two-stage schemes are proposed for the distortion less or larger the threshold. Recently, visual attention has been studied extensively for IQMs [7], [8]. Due to non-uniform distribution of the photo receptors on the retina and visual attention that drives the most sensitive part on interesting objects, images are not perceived with the same resolution for each region and the visual attention drive the eye and make the most sensitive region of region focus on interenting objects. Therefore the distortion is not perceived equally and should be given different weights. In the second category, rather than simulating the process of HVS, IQMs are proposed from the view of signal processing by involving image properties like structure information [9]–[11], statistical information [12], [13]. In [9], the structural similarity is computed using local mean and variance and the overall performance is measured by averaging the local structural similarity. In [12] and [13], the information fidelity criterion is proposed by quantifying the information shared between a reference and a distorted image. Recently the edge or gradient similarity have been proved effective in modeling IQMs [14]–[16]. More HVS based image quality metrics could be found in the literature such as [57] and [58].

Most of the above IQMs are aimed at handling a large range of distortion types and usually tested in databases with multiple distortion types such as the TID database [17]. However developing an universal quality metric is quite challenge. Due to the wide application of image compression in image delivery and storage, the compression distortion is one of major distortion among various distortion types. Besides, IQM plays a key role in image coding in the processes such as Rate-Distortion Optimization (RDO) [18]–[20]. Therefore, it is highly desired to have accurate IQMs for compressed images.

Manuscript received January 6, 2015; revised June 6, 2015 and September 8, 2015; accepted September 8, 2015. Date of publication September 23, 2015; date of current version October 15, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61472281 and 61471348, in part by the Shu Guang Project of Shanghai Municipal Education Commission and Shanghai Education Development Foundation under Grant 12SG23, and in part by the Shenzhen Overseas High-Caliber Personnel Innovation and Entrepreneurship Project under Grant KQCX20140520154115027. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Stefan Winkler.

S. Hu, L. Jin, and C.-C. J. Kuo are with the Ming Hish Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: sudenghu@gmail.com; linajin.j@gmail.com; cckuo@sipi.usc.edu).

H. Wang is with the Key Laboratory of Embedded System and Service Computing, Department of Computer Science and Technology, Ministry of Education, Tongji University, Shanghai 200092, China (e-mail: hanliwang@tongji.edu.cn).

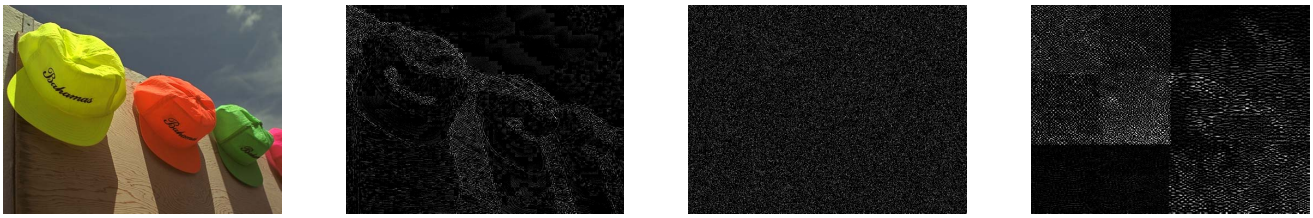
Y. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: yun.zhang@siat.ac.cn).

S. Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: cssamk@cityu.edu.hk).

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

Digital Object Identifier 10.1109/TIP.2015.2481319

1057-7149 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.



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

Fig. 1. Compression distortion is content dependent. (a) Original image. (b) Compression distortion. (c) Additive distortion. (d) Transmission error distortion.

Compression distortion could include various types of visual artifacts, which mainly are blurriness, blocking and ringing artifacts. In fact, compression distortion has its unique characteristics comparing to other distortion types. Masking effect is widely exploited in the image codecs, and that makes compression distortion content dependent. In codecs, high frequency components usually are quantized with larger quantizers than low frequency components. Moreover, for prediction based codecs, larger prediction residual in complicated area could also result in larger distortions. In addition, most perceptual image codecs try to hide distortion in the area that has large masking effect. Therefore as shown in Fig. 1, the distortion relates to original image that it is larger in complex content than in smooth content. On the other hand, masking effect from complex content could significantly prevent the distortion being perceived. Therefore the masking effect become critical to the compression distortion and it is important to make a quantitative analysis of the masking effect on MSE.

Masking effect refers to human's reduced ability to detect a stimulus on a spatially or temporally complex background. The traditional way to measure the masking effect is using a divisive gain control method, which decomposes the image into multiple channels and analyzes the masking effect among the channels by divisive gain normalization [21], [22], [54]–[56]. However, the mechanism of gain control mostly remains unknown. Additionally, since only simple masker such as sinusoidal gratings or white noise is used in the experiments to search for optimal parameters to fit the gain control model, there is no guarantee that these models are applicable to natural images [23]. In [24] and [25], it is pointed out that masking effect highly depends on the level of randomness created by the background. Usually the regular background contains predictable content and the stimulus will become distinct from neighborhood when it is different from human's expectation of its position. While in the random background, the content is unpredictable, and thus any change on it will be less noticed. Therefore, there is higher masking in the random background than the regular background. In [24], a concept of entropy masking is proposed to measure masking effect of background using zero order entropy. However, it fails to consider the spatial relation of pixel values. In addition, a single value might not be enough to indicate randomness of the whole background, because the content in the background may vary significantly. Furthermore, only with masking measurement is insufficient to predict the perceptual distortion, because it is unclear how the proposed masking measurement affects the perceived distortion.

In this paper, we first propose a method to measure the randomness of the background with a spatial statistics model. Since a regular structure has strong spatial correlation among their neighborhood, which makes it easier to predict the background from the neighborhood. Therefore, the prediction error actually reflects the randomness of background. The random background is less spatially predictable, resulting in larger prediction error. Thus the spatial prediction error is used as the measurement of randomness, indicating how much the background could mask the noise. With this method, we have a randomness map, rather than a single value, to indicate the randomness of the structure at each pixel. Then we investigate the model of masking modulation, which mathematically analyzes how distortion is reduced with the proposed randomness measurement based on the observation of perceptual qualities in terms of MOS in different databases. Meanwhile, we propose a simple but effective preprocessing scheme, which removes the imperceivable error signals.

The rest of this paper is organized as follows. In Section II, the scheme of randomness measurement is proposed. The masking modulation model is introduced in Section III. In Section IV, the experimental results are given to compare the performance of the proposed IQM with other benchmarks. Finally, Section V concludes this paper.

II. Randomness Measurement

The visual signal is affected by masking effect and the visibility of compression distortion significantly depends on the background of the images. Usually the distortion is easy to be observed in the regular region and hard to be perceived in disordered regions. To measure the masking effect of the image content, the spatial randomness of image structure should be measured. In this section, the randomness is measured quantitatively using the spatial estimation error. Meanwhile proper selection of prediction neighborhood is discussed as well.

A. Randomness Measured With Spatial Statistics

For regular structure, the pixels always have strong correlation with the neighboring pixels and the presence of particular combinations of neighboring pixels will increase the possibility of certain values of the current pixel. On the other hand, for a disordered structure, the neighboring pixels will provide less useful information to estimate the current pixel.

Let $Y(u)$ and $\mathbf{X}(u)$ be jointly distributed random variable and random vector standing for the current pixel and neighboring pixels, respectively. At a particular position,

$y(i,j)$ is an example of $Y(u)$ and similarly $\mathbf{x}(i,j)$ is an example of $\mathbf{X}(u)$ representing the neighboring pixels. The reasonable estimation of $y(i,j)$ is $E(y(i,j)|\mathbf{X}(u)=\mathbf{x}) = \sum_{y(i,j)\in\mathcal{S}} y(i,j)P_{Y|X}(y|\mathbf{x})$ where $P_{Y|X}(y|\mathbf{x})$ is conditional probability of y given $\mathbf{X}(u)=\mathbf{x}$ and $\mathcal{S}$ is the set of all possible y. However the estimation of $P_{Y|X}$ is not easy and thus we assume a linear estimation that

$$\hat{Y}(u) = \mathbf{H}\mathbf{X}(u), \tag{1}$$

where $\mathbf{H}$ is an $1\times n$ matrix. The optimal $\mathbf{H}^*$ is determined by achieving the minimum mean of the error $|(Y(u)-\hat{Y}(u)|$ over all possible combination of $Y(u)$ and $\mathbf{X}(u)$, which is expressed as

$$\mathbf{H}^* = \underset{\mathbf{H}\in\mathcal{R}^{1\times n}}{\operatorname{argmin}}\ E[(Y(u)-\mathbf{H}\mathbf{X}(u))^2], \tag{2}$$

where $E[\cdot]$ is the expected value operator. To achieve the optimal value, the following equation must be satisfied as

$$\frac{\partial E[(Y(u)-\mathbf{H}\mathbf{X}(\mathbf{u}))^2]}{\partial \mathbf{H}} = 2\mathbf{H}^*\cdot E[\mathbf{X}(u)\mathbf{X}(u)^T] - 2E[Y(u)\mathbf{X}(u)^T] = \mathbf{0}, \tag{3}$$

where T is the transpose operator. From Eq. (3), we could have $H^* = E[Y\mathbf{X}(u)^T]E[\mathbf{X}(u)\mathbf{X}(u)^T]^{-1}$ and hence the optimal estimation of $y(i,j)$ given the neighboring pixels $\mathbf{x}$ is

$$\hat{y}(\mathbf{x})(i,j) = \mathbf{R}_{YX}\mathbf{R}_X^{-1}\mathbf{x}(i,j), \tag{4}$$

where $\mathbf{R}_{YX} = E[Y\mathbf{X}(u)^T]$ is the cross-correlation matrix between $\mathbf{X}(u)$ and $Y(u)$ and $\mathbf{R}_X = E[\mathbf{X}(u)\mathbf{X}(u)^T]$ is the correlation matrix of $\mathbf{X}$. $\mathbf{R}_{YX}$ and $\mathbf{R}_X$ carry the structure information of image content and vary as the image structure changes.

If the neighboring pixels x_i, (*i.e.,* the components in $\mathbf{X}$) are linear dependent, $\mathbf{R}_X$ is not full rank and thus it is not invertible in Eq. (4). For example, in exactly plain regions, the structural information is so limited that the rank of $\mathbf{R}_X$ is actually one. In such a case, $\mathbf{R}_X^{-1}$ in Eq. (4) could be replaced by pesudo-inverse $\tilde{\mathbf{R}}_X^+$, which is expressed as

$$\tilde{\mathbf{R}}_X^+ = \mathbf{U}_m\Lambda_m^{-1}\mathbf{U}_m^T, \tag{5}$$

where Λ_m is the eigenvalue matrix of all non-zero eigenvalues of matrix $\mathbf{R}_X$ and $\mathbf{U}_m$ is the corresponding eigenvector matrix. As proved in appendix, the pesudo-inverse operation also provides the best estimation. Actually $\tilde{\mathbf{R}}_X^+$ is a generalized form of $\mathbf{R}_X^{-1}$. When $\mathbf{R}_X$ is full rank, they are equivalent.

The randomness of the structure could be measured by the estimation error from the neighborhood with structural correlation as

$$S(i,j) = |y(i,j) - \mathbf{R}_{YX}\tilde{\mathbf{R}}_X^+\mathbf{x}(i,j)|. \tag{6}$$

The large value of $S(i,j)$ means the structure is more disordered and thus contains more randomness. On the other hand, for the regular structure, $S(i,j)$ will be close to zero.

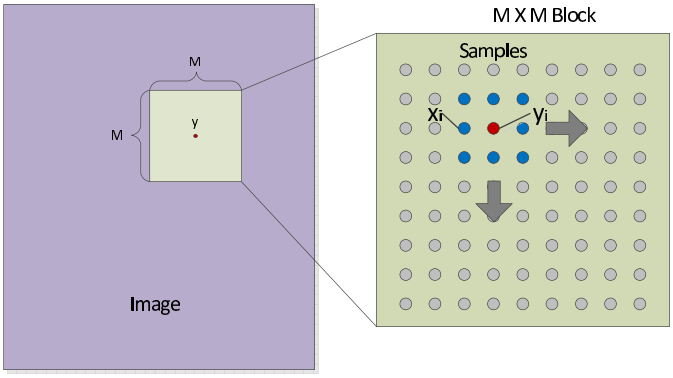


Fig. 2. Demonstration of sample extraction for $y(i,j)$ and $\mathbf{x}(i,j)$.

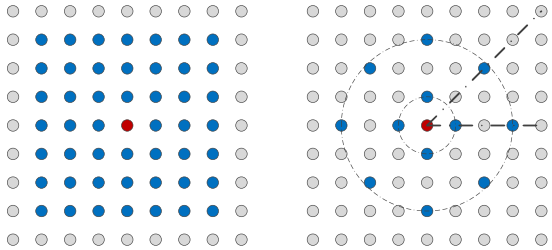


Fig. 3. Different neighborhood sampling. (a) Dense sampling. (b) Sparse sampling.

B. Estimation of Local Statistics

$\mathbf{R}_{YX}$ and $\mathbf{R}_X$ are the local properties of image content patterns, and change with image content. They could be estimated from pairs of y and $\mathbf{x}$ within local regions. A block with the size of $M\times M$ centered at y is used to extract the samples as shown in Fig. 2. The extracted samples are $\mathbf{X_S} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_N}]^T$, and $\mathbf{Y_S} = [y_1, y_2, \cdots, y_N]^T$, where N is the number of samples depending on the size of local block M and $\mathbf{x_i}$ and y_i are sample pairs in a particular position. The unbiased estimations of $\mathbf{R}_X$ and $\mathbf{R}_{YX}$ could be calculated from the sample correlation matrix and the sample cross-correlation matrix as

$$\hat{R}_X = \frac{1}{N-1}X_S X_S^T, \quad \hat{R}_{YX} = \frac{1}{N-1}Y_S X_S^T, \tag{7}$$

By replacing $\mathbf{R}_{YX}$ and $\mathbf{R}_X$ in Eq. (6) with their estimation in Eq. (7), we could estimate the randomness with local structure information.

C. Sparse Sampling of Neighborhood

The choice of neighboring pixels is not limited to the adjacent pixels. Only the closest neighboring pixels are not enough to capture the structure information of the patterns with large size. Thus more neighboring pixels within reasonable distance should be included as shown in Fig. 3 (a). A large size of neighborhood will increase the number of neighboring pixels and consequently will increase the computational complexity to estimate the randomness. Usually the dense neighboring pixels as shown in Fig. 3 (a) may contain significant redundancy. In order to achieve a proper size of neighborhood

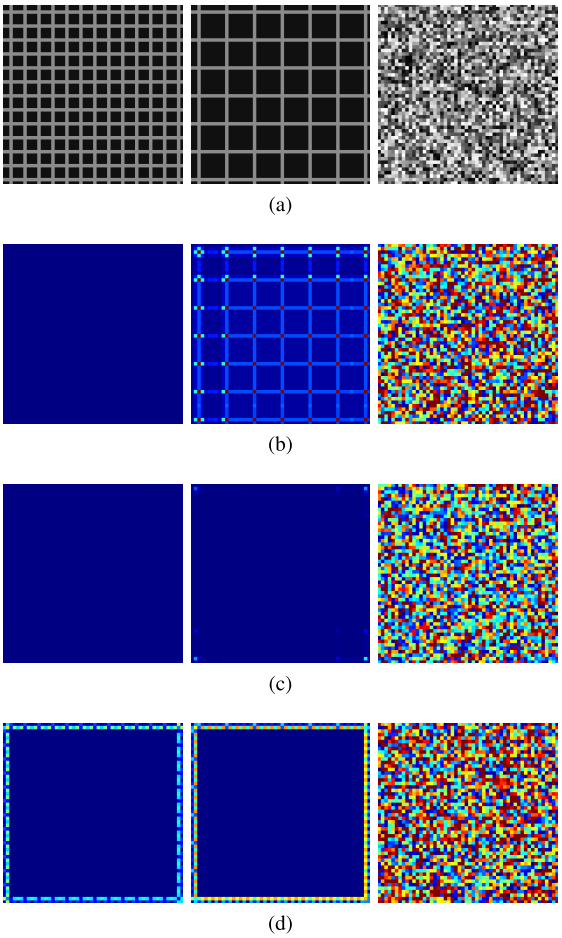


Fig. 4. Different patterns and the heat maps of randomness with different size of neighborhood. The images in each column are original images and the corresponding randomness maps with different methods. (a) Regular patterns with the size of 16 and 32 pixel respectively and a random pattern. (b) Dense sampling within a block of 9×9 size. (c) Dense sampling within a block of 17×17 size. (d) Sparse sampling within 17×17 block.

while maintaining a small number of neighboring pixels, the neighboring pixels are evenly sampled from the neighborhood as shown in Fig. 3 (b), and the sampled neighboring pixel set could be expressed in a polar coordinate system as

$$V = \left\{(\theta, r)|\theta = \frac{k\pi}{2}; r = 2l + 1 \leq L\right\} \bigcup \left\{(\theta, r)|\theta = \frac{(2k+1)\pi}{4}; r = 2\sqrt{2}l \leq L\right\}, \quad (8)$$

where $k = 0, 1, 2, 3$, and $l = 1, 2, \cdots,$ N; L is the size of neighborhood. Please note that the sampling method is not unique and the sampling method as illustrated in Fig. 3 (b) is adopted due to its simplicity and effectiveness.

To investigate the effect of neighboring pixels on the randomness calculation, different neighborhood sizes and different sampling methods are tested on simple patterns and the results are shown in Fig. 4. Fig. 4 (a) shows a regular pattern with a small size and a large size and a random pattern where the pixel values are independently uniform distributed. In Fig. 4 (b), the neighboring pixels are dense sampled within a small neighborhood size. We could see that the proposed randomness measure could correctly estimate the randomness of the pattern with small size, but fails for large size. That is because the small size of neighborhood only covers information of limited area. A large size of neighborhood with dense sampling is used in Fig. 4 (c), where the randomness is correctly estimated for both small and large size of pattern. While in Fig. 4 (d), large neighboring size is used and neighboring pixels are sampled sparsely as shown in Fig. 3 (b). We could see that the calculated randomness correctly captures the characteristics of images and achieves similar performance with dense sampling except for some errors due to the boundary effects. For the random pattern in Fig. 3, since its structure is random and neighboring pixels are independent with each others, all estimations give high randomness.

Usually a larger neighborhood could provide better estimation. However the scope of visual attention is limited, the optimal size of neighborhood L in Eq. (8) varies according to the pixel density and the viewing distance. Since in this paper we assume these parameters are fixed, a constant size of neighborhood is adopted. The randomness estimation on natural images are shown in Fig. 5, where the left half of image is more disordered while the right half is more regular and the corresponding calculated randomness with consistent with human perception.

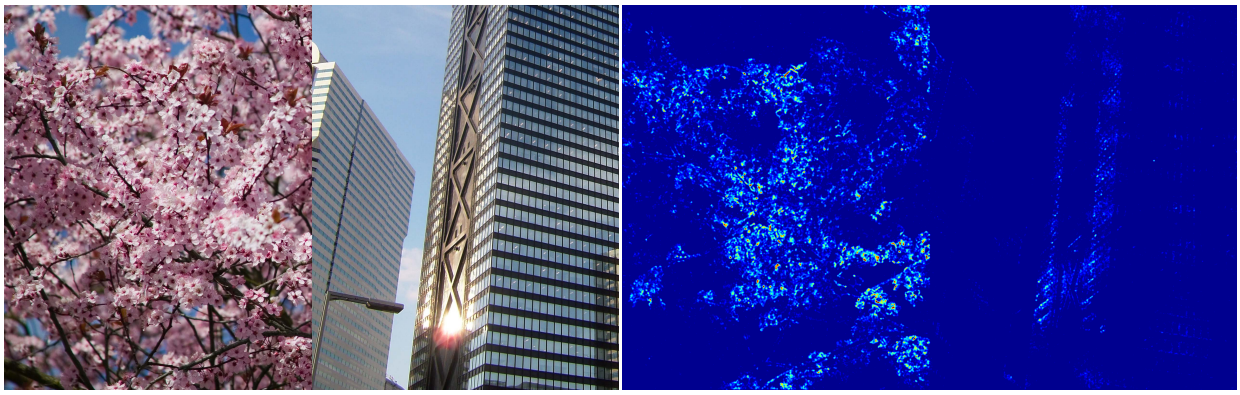
III. Masking Modulation With Randomness

After estimating the masking effect with proposed randomness quantitatively, it is critical to investigate the relation of the perceptual distortion and the randomness. Intuitively, the distortion at the pixel with high randomness should be reduced more than with low randomness. However, the exact model of how randomness modulates the actual distortion is not clear. Besides, different coding methods and image content could result in distortion with very different properties. Some distortion may contain more imperceivable distortion and some may contain less. That makes MSE inconsistent among various coding methods. Therefore, to simulate the processing occurred in the initial parts of HVS, proper preprocessing that removes imperceivable distortion is required. In this section, we first preprocess the error with a low-pass filter. Then we investigate the masking modulation at image level and later extend the developed modulation relation to pixel level.

A. Preprocessing With Low-Pass Filtering

The initial visual signal processing in HVS includes two steps. In the first step, the visual signal goes through eye's optics, forming an image on the retina. Because of the diffraction and other imperfections in the eye, such processing would blur the passed image. In the second step, the image will be filtered by neural filter as it is received by photoreceptor cells on retina and then passed on to lateral geniculate nucleus (LGN) and the primary visual cortex. These processes are more like low-pass filtering and will hide parts of signal from perception.

We assume the initial vision processing could be characterized by a linear transfer function and the magnitude of input



(a) (b)

Fig. 5. Illustration of randomness. (a) Original image. (b) Heat map of randomness.

and output signal in frequency domain is modeled as

$$I_F(\Omega) = G(\Omega) \cdot I(\Omega), \tag{9}$$

where $I(\Omega)$ and $I_F(\Omega)$ are the input image and output image in frequency Ω; $G(\Omega)$ is a modulation transfer function (MTF), reflecting the gain of the initial visual processing to various spatial frequencies. $G(\Omega)$ is the concatenation of the two MTFs at each step in the initial visual processing. In the first step, the eye's optics could be modeled as a simplified pinhole imaging system and its optical MTF could be expressed as a Gaussian blur function [29]. However the neural MTF in the second step that occurs in the neural system is hard to measure and model.

The CSF, which is defined as the inverse of contrast threshold of detectable contrast at a given frequency, provides a comprehensive measure of spatial vision. Although it is not exactly equivalent to MTF, it reflects the same trend as the modulation gain. For instance, a higher sensitivity at particular frequencies always means a higher modulation gain at the corresponding frequencies and *vice versa*. Therefore, many researchers have treated the CSF as the spatial MTF, and used it to define characteristics of initial processing in HVS [26]–[28]. In this paper, we adopt CSF as the MTF of initial visual processing. There are various CSF models proposed in past [30]–[37], and a generalized model is proposed in [34] and [35] as

$$G(\Omega) = (a + b\Omega)e^{-c\Omega}, \tag{10}$$

where Ω is the spatial frequency and a, b, c are constant model parameters and according to [34], they are set to 0.31, 0.69, and 0.29, respectively. The CSF is a low-pass filter which peaks at a certain frequency and then drops significantly. The CSF indicates that the human eye is less sensitive to higher frequency distortion. Therefore, the perceived distortion could be expressed as

$$\begin{aligned} \Delta\mathbf{I}_F &= \mathbf{g} * \mathbf{I} - \mathbf{g} * \mathbf{I_C} \\ &= \mathbf{g} * \Delta\mathbf{I}, \end{aligned} \tag{11}$$

where $\mathbf{I}$ and $\mathbf{I}_C$ are the original and compressed images; the operator $*$ means the convolution; $\Delta\mathbf{I}$ is the actual distortion

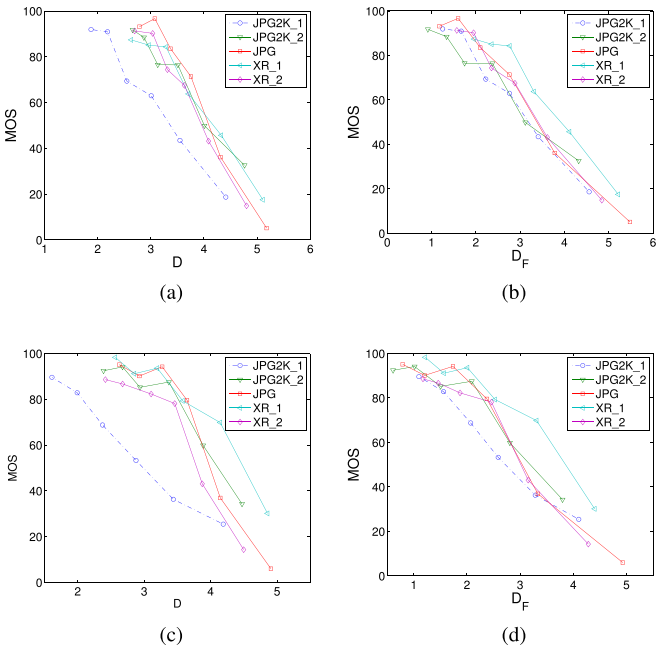


Fig. 6. The relation of MOS and distortion measurement for different coding methods. The images are coded with different coding methods: including encoding with JPEG2000 using two different setting, denoted as "JPG2K_1" and "JPG2K_2"; with JPEG XR using two different setting denoted as "XR_1" and "XR_2"; and JPEG coding denoted as "JPG". Details are included in [41]. (a) and (c) Without LPF for the image "bike" and "woman", respectively. (b) and (d) With LPF for the image "bike" and "woman", respectively.

that $\Delta\mathbf{I} = \mathbf{I} - \mathbf{I}_C$; $\mathbf{g}$ is the spatial low-pass filter of the CSF in Eq. (10). $\Delta\mathbf{I}_F$ reflects the observed distortion after initial visual processing. In this way, we could remove the high frequency noise that could not be perceived by humans.

Different encoding methods could yield distinct properties that MSE may not be able to capture. To investigate the effect of low-pass filtering, the distortion measurement before and after low-pass filtering are defined as

$$D = ln\,(\mathrm{MSE}), \quad D_F = ln\,(\mathrm{MSE_F}) \tag{12}$$

where MSE and MSE_F are the mean squared error without and with low-pass filtering, *i.e.*, mean squared value of $\Delta\mathbf{I}$ and $\Delta\mathbf{I}_F$. Fig. 6 (a) and 6 (c) show the plots of MOS vs. D,

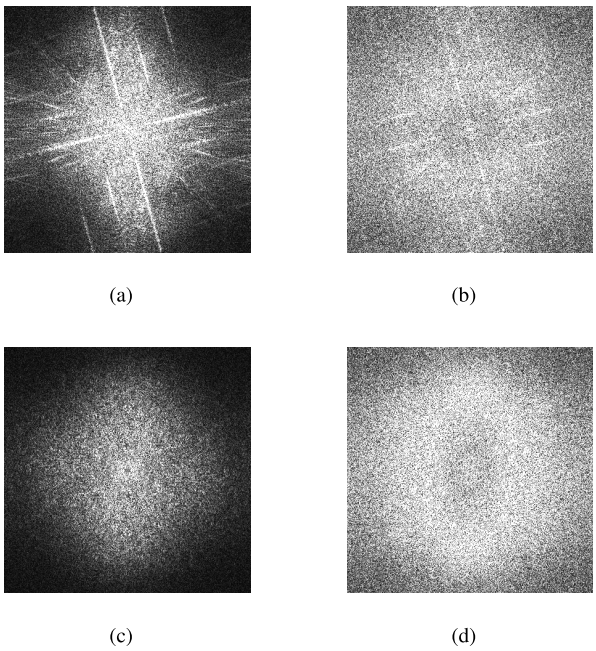


Fig. 7. Frequency magnitude of the distortion ΔI. DC component locates at the center. (a) and (b) Show the image "bike" coded with "JPG2K_1" and "JPG2K_2", respectively. (c) and (d) Show the image "woman" coded with "JPG2K_1" and "JPG2K_2", respectively.

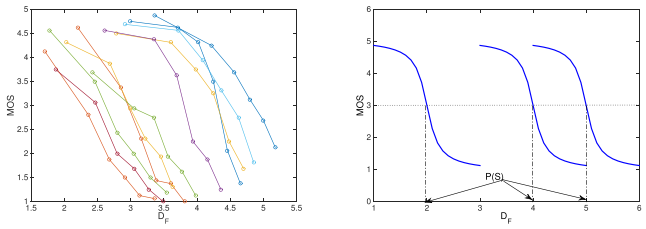


Fig. 8. Plot of MOS vs. D_F. Each line corresponds to one original image. (a) Actual plot of MOS vs. D_F from database Toyama. (b) Idealized plot of MOS vs. D_F.

where the images are coded with different coding methods at different quality levels. We could find that given the same D, the images coded with "JPG2K_1" has smaller MOS than with other coding methods, which means the distortion from "JPG2K_1" is more obvious. This is because as shown in Fig. 7, for "JPG2K_1", the most distortion energy locates on low frequencies while for "JPG2K_2" the distortion energy spreads out to higher frequencies at which humans are less sensitive. After low-pass filtering, the most parts of imperceivable distortion are removed, and hence D_F becomes more consistent among different coding methods as shown in Fig. 6 (b) and Fig. 6 (d).

B. Imagewise Masking Modulation

To investigate how the masking effect reduces the visibility of distortion at image level, The relationship between D_F and MOS is shown in Fig. 8 (a) for various images compressed at different quality levels. Each circle represents a coded image and the circles connected by the same lines share the same original images. In other words, the connected circles in Fig. 8 (a) are the images compressed from the same original images but with different compression levels, hence they are affected by the same masking effect.

TABLE I
AVERAGE MOS AND AVERAGE D_F OF EACH IMAGE

Image	MOS	D_F	Image	MOS	D_F
Kp01	3.0	2.44	Kp13	2.8	2.91
Kp03	2.6	0.85	Kp16	2.7	1.32
Kp05	3.0	2.55	Kp20	3.2	0.95
Kp06	3.0	1.93	Kp21	2.3	1.74
Kp07	3.0	1.13	Kp22	2.6	1.72
Kp08	2.8	2.62	Kp23	3.0	0.58
Kp12	2.6	0.99	Kp24	2.8	2.22

As we could see in Fig. 8 (a), for the image set sharing a particular original image, their MOS values monotonically decrease with D_F and each image set has similar MOS-D_F relation but with different horizontal displacement. The mean MOS and mean D_F of each set is calculated and summarized in Table I, where we could see the average perceptual quality of coded image is around at 3.0 in MOS, however the mean D_F is quite different from each other.

Such difference in horizontal displacement comes from the different masking effect of different images. Given the same MOS, the lines of the images on the right side have more distortion than the lines on left as shown in Fig. 8 (a), which means the image on the right side has more masking which makes it appear the same quality as the images on the left side. Therefore, the image sets with strong masking effect are more likely to have curves on the right side, and the relative displacement of these curves to the left reflects the significance of masking effect.

To investigate these horizontal displacement of these curves, the small difference in the shapes of curves is neglected by idealizing the curves as in Fig. 8 (b). Consequently the MOS-D_F relation could be expressed as

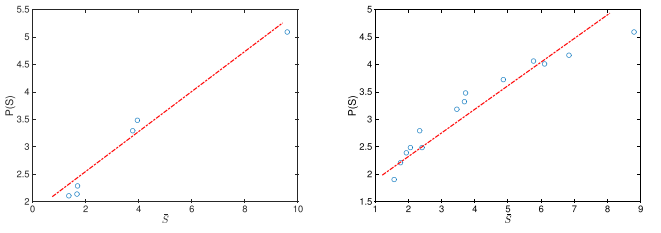
$$\widehat{\mathrm{MOS}} = F(D_F - P(S)), \tag{13}$$

where $\widehat{\mathrm{MOS}}$ is the predicted MOS; $F(\cdot)$ is a nonlinear monotonic decreasing function representing the shape of these curves and $P(S)$ is the displacement of the curves, which is a function of randomness S of the corresponding images, since S reflects the significance of masking effect.

The actual horizontal displacement of the curves could be measured by the intersection of the curves and any horizontal lines such as MOS $= 3.0$ as shown in Fig. 8 (b). Using other lines will result in a constant adding to $P(S)$, but it will not affect the following equations. To investigate the relation between $P(S)$ and randomness S, the image level randomness is calculated by averaging pixel level randomness as

$$\bar{S} = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} S(i,j) \tag{14}$$

and the plot of $P(S)$ vs. $\bar{S}$ is shown in Fig. 9. In Fig. 9 (a), for the Toyama database we could observe that $P(S)$ increase linearly with $\bar{S}$. The same observation could be obtained in



(a) (b)

Fig. 9. The linear relationship between mean randomness $\bar{S}$ and horizontal displacement $P(S)$. (a) On Toyama. (b) On MMSPG.

Fig. 9 (b) for the MMSPG database. Therefore their relationship could be expressed as

$$P(S) = \lambda \bar{S} + b, \tag{15}$$

where λ and b are model parameters. Then by substituting Eq. (12) and Eq. (15) into Eq. (13), we could have

$$\begin{aligned} \widehat{\text{MOS}} &= F\left(ln(\text{MSE}_\text{F} \cdot e^{-\lambda \bar{S}}) - b\right) \\ &= G\left(\text{MSE}_\text{F} \cdot e^{-\lambda \bar{S}}\right) \end{aligned} \tag{16}$$

where $G(\cdot) \equiv F(ln(\cdot) - b)$ is a nonlinear mapping. It is acceptable for a IQM to predict MOS through a nonlinear mapping, because the mapping is easy to be found and it depends on various environmental factors like the range of MOS and evaluation methodology. Therefore, in [38] and [39], a nonlinear mapping is not considered as part of IQM, rather it is left to the final stage of performance evaluation. $G(\cdot)$ could be obtained by fitting the objective prediction scores to the subjective quality scores as described in [38] and [39].

From Eq. (16), we can conclude that Image-wise Perceptually Weighted MSE (IPW-MSE) is a good indicator of MOS, which is calculated as

$$\text{IPW-MSE} = \text{MSE}_\text{F} \cdot e^{-\lambda \bar{S}} \tag{17}$$

Without considering the masking effect, MSE_F is not accurate enough to indicate the perceptual quality as we have observed in Fig. 8. Eq. (17) gives the exact relation how MSE_F should be modified with randomness S. It is also consistent with our intuition that the increase of image level randomness $\bar{S}$ will reduce the visibility of distortion MSE_F.

C. Pixelwise Masking Modulation

In the above section, we discuss the same distortion (*i.e.*, MSE_F) does not mean equal perceptual quality in different images due to the masking effect. Rather it should be modulated with randomness as in Eq. (17). Even within the image, the distortion is not equally perceived because of the various masking effect in different image regions. To obtain the precise IQM, we consider the masking effect at a finer level, *i.e.*, pixel level. Since the subjective test can be hardly conducted at pixel level, we assume that the obtained modulation relationship at image level in Eq. (17) is also applied to pixels. It is validated by the performance improvement in the experiments of Section IV. In Eq. (17), by replacing MSE_F and mean randomness ($\bar{S}$) with filtered squared error $\Delta I_F(i, j)^2$ and randomness $S(i, j)$ of each pixel measured in Eq. (6), we have modulated the squared error at each pixel as

$$SE_M(i, j) = \Delta I_F(i, j)^2 \cdot e^{-\lambda_2 \cdot k \cdot |y(i,j) - R_{YX} \tilde{R}_X^{-1} \mathbf{x}(i,j)|} \tag{18}$$

where λ_2 is a constant model parameter and k is related to image resolution, i.e. $k = 1$ if $W \times H > 768 \times 511$ and $k = 0.083$ elsewise. In this way, the normalized distortion at each pixel has equal perceptual effect.

Fig. 10 (a) and (b) show a original image and the compressed image. Fig. 10 (d) shows the filtered distortion, where we can see that even though the actual distortion in the sky area is much small compared to that in other parts, the perceived distortion is still comparable to other parts. This is because the sky area is smoother than other areas, and thus the masking effect is much weaker than other parts. That could be reflected by the corresponding randomness map as shown in Fig.10 (c). After modulating the actual distortion with the randomness map, we can see the distortion in the sky area is enhanced relatively. This is consistent with perceptual observation.

Since the modulated distortion is perceptually normalized, the perceptually weighted MSE (PW-MSE) is calculated by even pooling as

$$\text{PW-MSE} = ln\left(\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} SE_M(i, j)\right). \tag{19}$$

Similarly MOS could be predicted with PW-MSE through a proper nonlinear mapping.

IV. Experimental Results

To evaluate the performance of PW-MSE, six databases with various types of compression distortion are used, including Toyama [40], MMSPG [41], TID2008 [17], TID2013 [51] and CSIQ [52]. In the Toyama database, there are 14 original images with solution of 768×512. Each original image is encoded with JPEG [42] and JPEG2000 [43] at six different quality levels, generating 168 distorted images. In the MMSPG database, there are 6 original images with the solution of 1280×1600. Three different codecs JPEG, JPEG 2000 and JPEG XR are used in the database. For JPEG 2000 and JPEG XR two different coding strategies are adopted, which are denoted as "JPG2K_1" and "JPG2K_2", "XR_1" and "XR_2", respectively. For each coding method, original images are coded at 6 different quality levels. Therefore, there are totally 160 distorted images. There are a broad spectrum of distortion types in the TID2008, TID2013 and CSIQ databases. Since we are only intereted in compression distortion, only JPEG and JPEG 2000 distortion are investigated on these databases.

As for metrics of performance evaluation, the Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC) and root mean squared error (RMSE) are employed as described in [38] and [39]. PLCC generally indicates the goodness of linear relation. SROCC is computed on ranks and thus depicts the monotonic relationships. RMSE computes the prediction errors and thus depicts the prediction accuracy. To put the MOS and its prediction on the

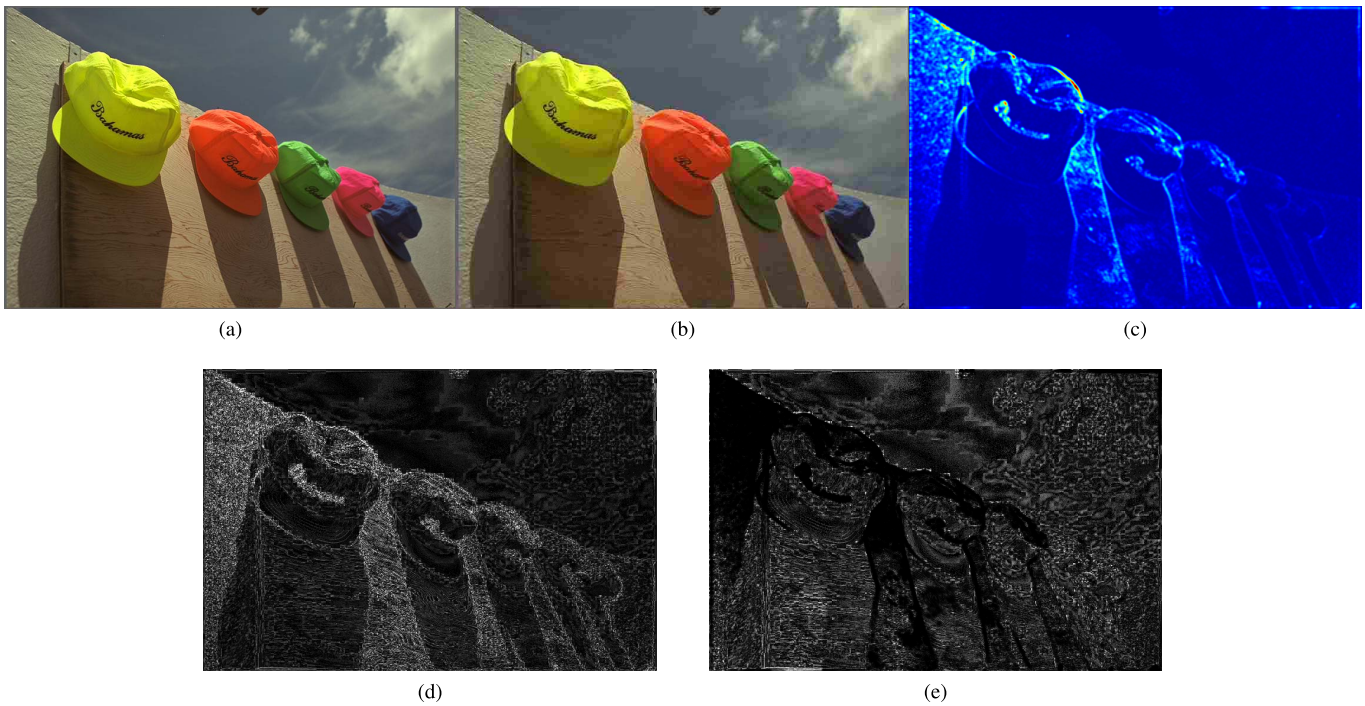


Fig. 10. Distortion modulated at pixel level. (a) Original image. (b) Distorted image. (c) Heat map of randomness. (d) Distortion before modulation. (e) distortion after modulation (properly scaled for better illustration).

TABLE II
PERFORMANCE EVALUATION AT EACH STEP

	PLCC				SROCC				RMSE			
	D	D_F	IPW-MSE	PW-MSE	D	D_F	IPW-MSE	PW-MSE	D	D_F	IPW-MSE	PW-MSE
Toyama	0.626	0.822	0.872	**0.926**	0.613	0.816	0.873	**0.922**	0.976	0.712	0.612	**0.470**
MMSPG	0.775	0.890	0.921	**0.954**	0.797	0.891	0.866	**0.927**	16.769	12.139	10.358	**7.965**
TID2008	0.870	0.952	0.961	**0.983**	0.866	0.949	0.963	**0.977**	0.933	0.577	0.478	**0.343**
TID2013	0.899	0.967	0.972	**0.983**	0.917	0.916	0.956	**0.970**	2.199	0.414	0.389	**0.300**
CSIQ	0.861	0.954	0.970	**0.973**	0.916	0.948	0.956	**0.963**	0.158	0.094	0.079	**0.072**

same scale for various algorithms, a monotonic logistic function is used to find nonlinear mapping between the prediction and subjective quality scores as [39]:

$$q(x) = \alpha_1 \left(0.5 - \frac{1}{1+exp(\alpha_2(x-\alpha_3))}\right) + \alpha_4 x + \alpha_5, \quad (20)$$

where α_1 to α_5 are the parameters obtained by regression between the input and output data.

A. Validation at Each Stage

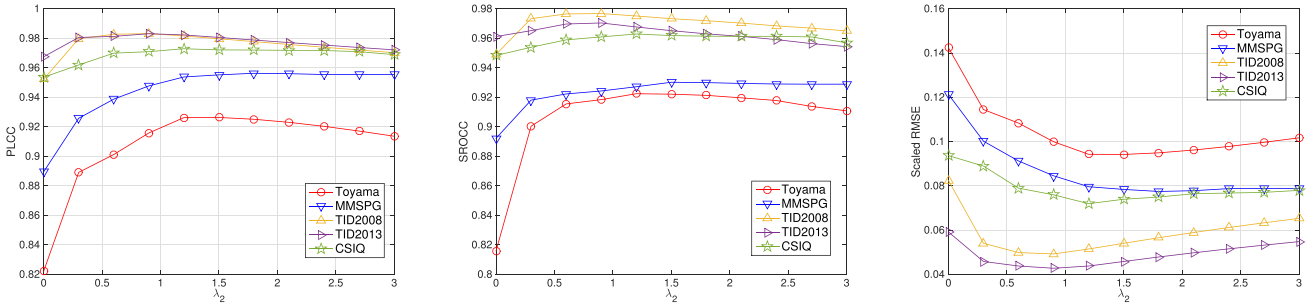
The proposed algorithm consists of several steps to simulate the different stages of HVS. To evaluate the effectiveness of the proposed IQM at each step, intermediate results are summarized in Table II for all six databases.

We evaluate the performance of D_F in Eq. (12) after applying low-pass filter. Then frame level masking effect is considered and the performance of IPW-MSE is measured, and finally the performance of PW-MSE is measured. As shown in Table II, the performance on compression distortion of all databases are presented, where we can see, as the starting point, MSE has the worst performance comparing to other steps of the proposed algorithm. This is expected because MSE does not incorporate any characteristics of HVS. Then from D_F to PW-MSE, the performance on the overall database is improved from 0.822 to 0.926 in PLCC for the Toyama database and from 0.890 to 0.954 in PLCC for the MMPSG database. Similarly, we can observe the similar trend on other databases and in other performance metrics, *i.e.*, SROCC and RMSE.

The performance of D_F is significant improved from MSE. This is because with the low-pass filtering, D_F removes the most parts of imperceivable distortion, making it more consistent with the human perception. IPW-MSE and PW-MSE improves the performance further, because in addition to low-pass filtering, the masking effect is considered. Moreover, we could find that the performance of PW-MSE is generally better than IPW-MSE either under each type of distortions or under the overall database. This is because in PW-MSE, the masking effect is considered at a finer scale than in IPW-MSE, as a consequence, the predication is more accurate.

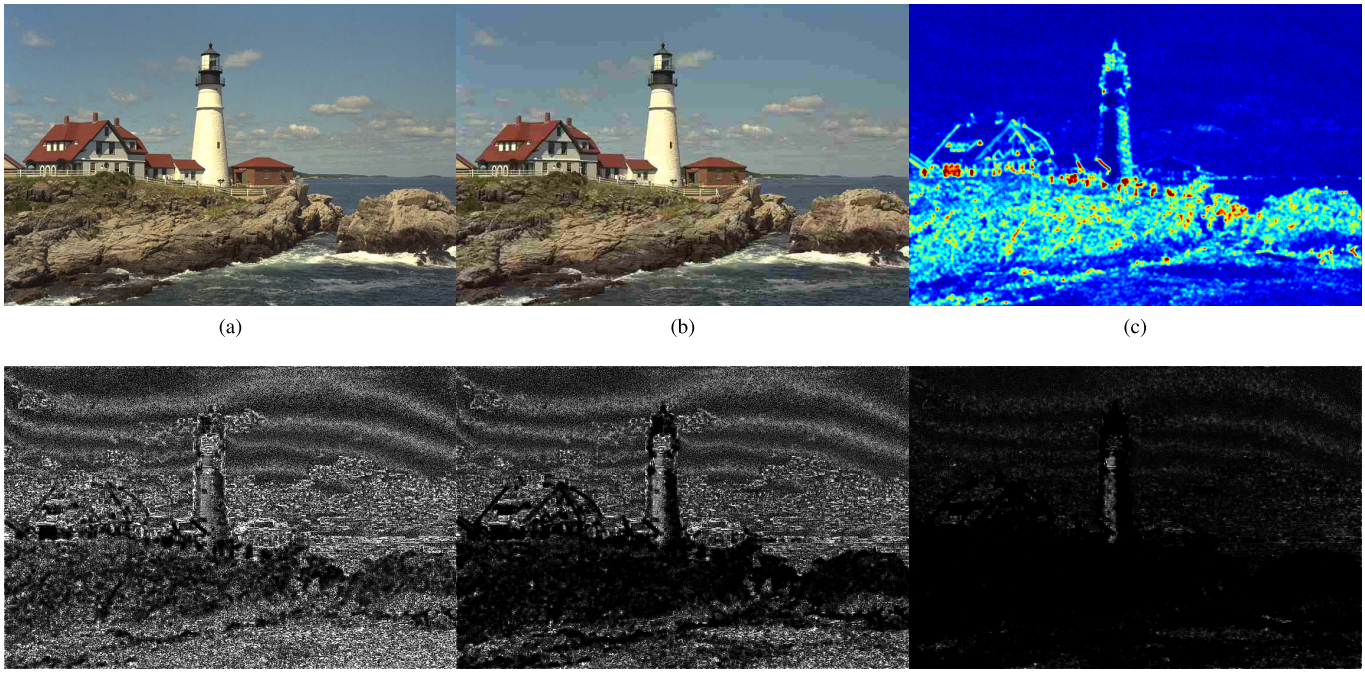
B. Parameter Investigation

Parameters are critical to the performance of the proposed algorithm. λ_2 in Eq. (18) is an important parameter that would affect the overall performance. To investigate its influence on



(a) (b) (c)

Fig. 11. The effect of model parameter λ_2 on various performances. (a) PLCC (b) SROCC (c) RMSE.



(a) (b) (c)

(d) (e) (f)

Fig. 12. Visual illustration of distortion modulation at pixel. (a) Original image. (b) Distorted image. (c) Randomness map. (d) Distortion modulated with $\lambda_2 = 0.2$. (e) Distortion modulation with $\lambda_2 = 1.2$. (f) Distortion modulation with $\lambda_2 = 2.2$.

the final performance, experiments are carried out by varying it in the range of [0, 3].

The curves of the overall performance on the six databases are shown in Fig. 11 for PLCC, SROCC and RMSE, respectively. When $\lambda_2 = 0$, the masking modulation with randomness is actually eliminated, resulting in the same performance as D_F. As shown in Fig. 11 (a), when λ_2 increases slightly, the performance increases significantly on all the databases. During this stage, the masking modulation starts affecting and the parts masked by strong maskers reduce its impacts on the overall quality index. When λ_2 becomes larger, after peaking at a certain value, the performance starts decreasing. This is because some distortion is over-masked and thus it is not consistent with the HVS. The same observation could be obtained in SROCC and RMSE in Fig. 11 (b) and (c). As for the best λ_2, it is almost constant on each database that it generally falls in the range [1, 2]. In the proposed algorithm, it is fixed at 1.2.

Fig. 12 visually illustrates the masked distortion with different parameters. We can see that in the distorted images in Fig. 12 (b), the distortion is more obvious in the sky region where the content is simple, while less obvious in the rock region. If the parameter λ_2 is too small as in Fig. 12 (d), the distortion in the complex region is not masked enough. Thus the measured quality index is not accurate enough. When λ_2 is too large as in Fig. 12 (f), the distortion in the complex region is over masked that it totally disappears, which is also inaccurate.

C. Validation of Effectiveness of Randomnness Map

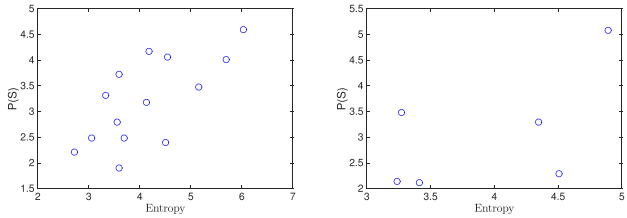
To further verify the effectiveness of proposed randomness, a entropy map and a masking map generated from division gain normalization [59] are used to replace randomness map in the proposed metric and their performance are compared. The entropy map is calculated based on 9×9 blocks, pixels within each non-overlap 9×9 block share the same entropy value.

TABLE III
OVERALL PERFORMANCE ON DIFFERENT DATABASES

	Database	PSNR	SSIM	MS-SSIM	VIFp	GSMD	FSIM	VSI	Entropy	DGN	Our
PLCC	Toyama	0.588	0.849	0.852	0.779	0.825	0.863	0.858	0.822	0.832	**0.926**
	MMSPG	0.790	0.927	0.936	0.853	0.898	0.899	0.926	0.849	0.886	**0.954**
	TID2008	0.869	0.963	0.974	0.953	0.982	0.975	0.981	0.935	0.951	**0.983**
	TID2013	0.916	0.962	0.970	0.952	0.975	0.971	0.981	0.938	0.967	**0.983**
	CSIQ	0.918	0.967	**0.981**	0.978	0.977	0.979	0.976	0.954	0.955	0.973
SROCC	Toyama	0.578	0.841	0.848	0.778	0.850	0.856	0.855	0.816	0.817	**0.922**
	MMSPG	0.797	0.904	0.897	0.820	0.914	0.892	0.900	0.760	0.892	**0.927**
	TID2008	0.866	0.961	0.969	0.949	**0.979**	0.969	0.978	0.929	0.949	0.977
	TID2013	0.917	0.948	0.955	0.938	0.968	0.958	0.968	0.931	0.961	**0.970**
	CSIQ	0.916	0.951	**0.968**	0.967	0.963	0.964	0.967	0.948	0.949	0.963
RMSE	Toyama	1.012	0.661	0.656	0.785	0.708	0.633	0.642	0.712	0.710	**0.472**
	MMSPG	16.277	9.925	9.345	13.851	11.656	11.611	10.014	15.563	12.288	**7.965**
	TID2008	0.937	0.510	0.431	0.571	0.354	0.424	0.372	0.777	0.583	**0.343**
	TID2013	0.658	0.445	0.397	0.502	0.366	0.393	0.318	0.731	0.418	**0.300**
	CSIQ	0.123	0.080	**0.060**	0.065	0.066	0.063	0.068	0.094	0.096	0.072

TABLE IV
RESULTS OF STATISTICAL SIGNIFICANCE TEST

	PSNR	SSIM	MS-SSIM	VIFp	GSMD	FSIM	VSI	PW-MSE
PSNR	–	11111	11111	11111	11111	11111	11111	11111
SSIM	00000	–	00111	00001	00111	11111	00111	11111
MS-SSIM	00000	00000	–	00000	00100	11000	00110	11110
VIFp	00000	11010	11110	–	01110	11110	11110	11110
GSMD	00000	11000	11000	10000	–	11000	11010	11010
FSIM	00000	00000	01000	00000	00100	–	00110	11110
VSI	00000	00000	00001	00000	00000	10000	–	11000
PW-MSE	00000	00000	00001	00000	00000	00000	00000	–



(a) (b)

Fig. 13. Relation between displacement of metric curves and entropy map. (a) On the Toyama database. (b) On the MMSPG database.

The linear relation in Eq. (15) is critical to the accuracy of proposed quality metric. We can see that the randomness could generally achieve a good linear relation as shown in Fig. 9. The relation between entropy map and displacement of metric curves are visually shown in Fig. 13, where we can see that there is neither strong linear relation nor other proper relation.

The performance of the proposed metric with different masking maps is evaluated on various databases. The results are shown in Table III and it is obvious that the proposed metric with randomness map has better performance. This is because randomness has better prediction for the displacement of metric curves and the performance of the proposed metric significantly relies on such relation, otherwise the metric could not effectively estimate the masking effect.

D. Comparison With Benchmark Algorithms

In this section, the performance of PW-MSE is compared with that of the seven benchmarks including: PSNR, SSIM [9], MS-SSIM [11], VIFp [13], GSMD [45], FSIM [14] and VSI [46]. Default setting is used for all the benchmark IQMs. FSIM and VSI are computed in color space and the rest IQMs are computed in gray images, where color images in RGB space are converted into YCbCr color space and only the luminance component Y is used. In TID2008, TID2013, and CSIQ databases, only the images with compression distortion, *i.e.,* JPEG and JPEG 2000 distortion are used for evaluation.

TABLE V
COMPRESSION DISTORTION AND ITS VISUAL ARTIFACTS

Compression distortion	Visual distortion
JPEG	Blocking, Ringing
JPEG 2000	Blurriness, Ringing
JPEG XR	Blocking, Blurriness, Ringing

Generally PLCC, SROCC and RMSE are consistent in performance evaluation, but not always. For example, in Table III, PW-MSE achieve the best performance on TID2008 in terms of PLCC, but not the best in terms of SROCC. That is because these evaluation methods measure different aspects of performance, and they are not exactly the same.

For the overall performance, from Table III, we can see that PSNR has the worst performance in PLCC among all IQMs. This is reasonable, because all the other IQMs incorporates with the characteristics of HVS while PSNR merely computes the pixel errors. We can have the similar observation in other performance metrics, *i.e.,* SROCC and RMSE.

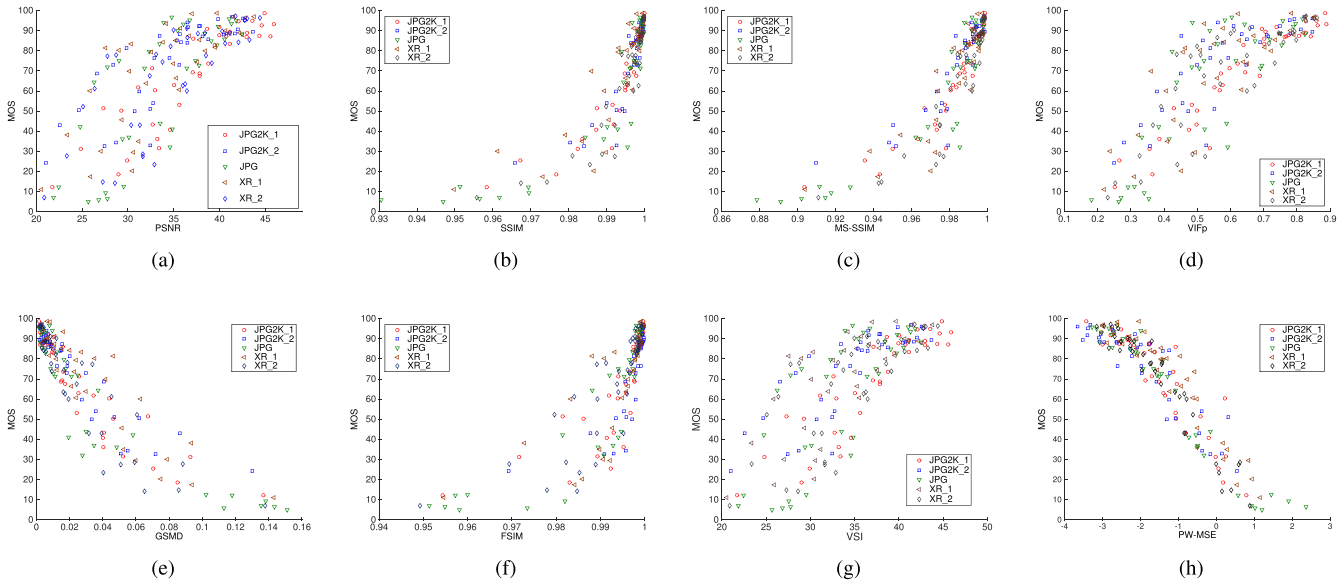


Fig. 14. Scatter plot of MOS vs. IQMs. (a) PSNR (b) SSIM (c) MS-SSIM (d) VIFp (e) GSMD (f) FSIM (g) VSI (h) PW-MSE.

TABLE VI
PERFORMANCE ON JPEG2000 DISTORTION

		PSNR	SSIM	MS-SSIM	VIFp	GSMD	FSIM	VSI	Proposed
PLCC	Toyama	0.856	0.853	0.858	0.833	0.865	0.839	0.912	**0.926**
	MMSPG	0.834	0.938	0.883	0.876	0.905	**0.975**	0.948	0.902
	TID2008	0.867	0.968	0.976	0.965	0.986	0.98	0.986	**0.987**
	TID2013	0.917	0.967	0.971	0.961	0.979	0.973	**0.983**	0.974
	CSIQ	0.947	0.963	**0.982**	0.978	0.980	0.981	0.975	0.978
SROCC	Toyama	0.865	0.845	0.848	0.83	0.892	0.826	0.908	**0.939**
	MMSPG	0.826	0.937	0.926	0.864	0.940	**0.956**	0.933	0.915
	TID2008	0.813	0.964	0.970	0.958	0.981	0.977	**0.985**	0.980
	TID2013	0.884	0.949	0.954	0.941	0.967	0.958	**0.971**	**0.971**
	CSIQ	0.936	0.956	**0.973**	0.97	0.972	0.969	0.969	0.970
RMSE	Toyama	0.652	0.66	0.648	0.699	0.633	0.687	0.517	**0.463**
	MMSPG	12.768	7.999	10.87	11.177	9.829	**5.093**	7.353	9.995
	TID2008	0.972	0.492	0.428	0.514	0.327	0.387	0.320	**0.312**
	TID2013	0.679	0.435	0.407	0.473	0.351	0.392	**0.312**	0.385
	CSIQ	0.102	0.085	**0.060**	0.066	0.063	0.062	0.071	0.066

SSIM and MS-SSIM have similar performances on both databases, this is because both of them measure the structure distortion. In PLCC, PW-MSE outperforms other seven benchmarks, except on the CSIQ database, where it also achieves close performance to the best performer MS-SSIM. In general, PW-MSE has excellent performance comparing with other benchmarks under various evaluation methods.

To obtain statistical conclusions on the performance of PW-MSE, we followed similar approaches of hypothesis testing in [45] and [47]. The hypothesis tests are carried out on the MOS prediction residual of two quality metrics, which is assumed to follow Gaussian distribution. The left-tailed F-test to the residuals of every two metrics on different databases and the results are shown in Table IV. A test result of $H = 1$ for the left-tailed F-test at a significance level of 0.05 means that the metric in the column has better performance than the model in rows with a confidence greater than 95%. A value of $H = 0$ means the metric in the column has indistinguishable or significant worse performance than the metrics in rows. Each cell of Table IV contains 5 flags, which from left to right stand for the test results on the Toyama, the MMSPG, the TID2008, the TID2013, and the CSIQ databases, respectively. We can see that PW-MSE has the most positive flags, *i.e.,* 1, indicating it has significant better performance than other metrics on most databases.

To provide a visual comparison among the benchmark IQMs and the proposed algorithm, the scatter plots of the quality index versus the MOS are shown in Fig. 14, where each point corresponds to a distorted image. We could see that for SSIM, MS-SSIM, GSMD and FSIM, the quality scores of the good quality images are very close to each other. For example, in SSIM, for the images with quality higher than 50 in MOS, its SSIM scores are in the range of 0.99 to 1.00. For PW-MSE, quality scores are evenly distributed.

E. Performance on Individual Distortion Types

The compression distortion consists of various visual distortion types, *e.g.,* blurriness, blocking and ringing artifacts.

TABLE VII
PERFORMANCE ON JPEG DISTORTION

		PSNR	SSIM	MS-SSIM	VIFp	GSMD	FSIM	VSI	Proposed
PLCC	Toyama	0.391	0.849	0.849	0.736	0.786	0.892	0.809	**0.954**
	TID2008	0.868	0.957	0.97	0.939	0.977	0.974	**0.986**	0.969
	TID2013	0.914	0.957	0.968	0.941	0.97	0.971	**0.985**	0.981
	CSIQ	0.847	0.976	**0.984**	0.982	0.984	0.984	0.981	0.971
SROCC	Toyama	0.332	0.844	0.853	0.730	0.814	0.899	0.809	**0.951**
	MMSPG	0.764	0.882	0.870	0.769	0.916	0.905	0.914	**0.944**
	TID2008	0.876	0.930	0.941	0.916	0.953	0.937	**0.962**	0.956
	TID2013	0.919	0.922	0.933	0.916	0.951	0.938	0.954	**0.959**
	CSIQ	0.888	0.953	0.966	**0.967**	0.965	0.965	0.962	0.955
RMSE	Toyama	1.138	0.654	0.653	0.838	0.764	0.558	0.728	**0.370**
	MMSPG	19.991	10.364	10.817	14.766	13.012	8.543	9.092	**4.678**
	TID2008	0.847	0.495	0.416	0.587	0.361	0.384	**0.284**	0.420
	TID2013	0.611	0.437	0.375	0.508	0.366	0.358	**0.256**	0.295
	CSIQ	0.163	0.066	**0.055**	0.057	0.055	0.055	0.060	0.073

TABLE VIII
PERFORMANCE ON JPEG XR DISTORTION

		PSNR	SSIM	MS-SSIM	VIFp	GSMD	FSIM	VSI	Proposed
PLCC	MMSPG	0.783	0.915	0.927	0.829	0.883	0.933	0.901	**0.956**
SROCC	MMSPG	0.775	0.878	0.883	0.806	0.885	0.908	0.88	**0.928**
RMSE	MMSPG	16.212	10.503	9.769	14.552	12.227	9.394	11.314	**7.618**

As pointed out in [48]–[50], different compression distortion types may be dominated by very different visual distortion types. For example, JPEG distortion mainly include blocking and ringing artifacts, while JPEG 2000 distortion include blurriness and ringing artifacts. Table V summarizes the compression distortion and their main visual distortion types.

To have a comprehensive understanding of the performance of the proposed metric on individual type of distortion, especially on the distortion types that are visually different, we compare the performance with benchmark metrics on JPEG 2000, JPEG and JPEG XR, respectively and the results are listed in Table VI, VII, and VIII, respectively. We can see that for JPEG 2000, PW-MSE hits the top 8 times, which is better than other quality metrics. Similarly for JPEG and JPEG XR, PW-MSE also has the best performance in terms of being the best metric on a specific database.

Besides, we also compare the performance on other non-compression distortions such as Gaussian blur and white additive noise. The results are shown in Table IX and X, respectively and the top 3 performers are highlighted in bold font. As we can see, the proposed metric still has the comparable performance with other benchmark metrics.

F. Computational Complexity

The computational complexity of the proposed PW-MSE is also analyzed in this section. Since PW-MSE consists of three stages: namely they are low-pass filtering, randomness calculation and modulation, their time consumption is investigated respectively. The average processing time over all images of each database was measured for each stage. The results are illustrated in Fig. 15, where we can see that, because of the larger image resolution, the time consumption on the MMSPG database is higher than on the Toyama database. Moreover, on both databases, we can find that the randomness calculation takes a large portion of computation in the proposed algorithm.

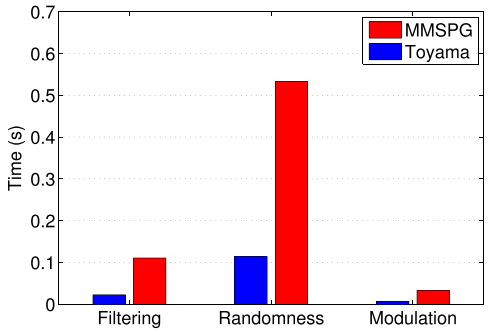
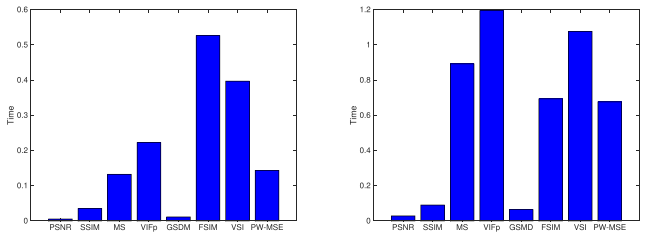


Fig. 15. Average consumed time in each stage of the PW-MSE.



(a) (b)

Fig. 16. Average total consumed time of the benchmark algorithms and the PW-MSE. (a) On the Toyama database (b) On the MMSPG database.

Meanwhile, we also compared the total time consumption of PW-MSE with other benchmark algorithms. The mean of consumed time for each image was measured and the results on both databases are shown in Fig. 16. Among these IQMs, since PSNR is the simplest in computation complexity, it has

TABLE IX
PERFORMANCE ON GAUSSIAN BLUR

		PSNR	SSIM	MS-SSIM	VIFp	GSMD	FSIM	VSI	Proposed
PLCC	TID2008	**0.934**	0.818	0.821	0.781	0.885	0.783	**0.924**	**0.914**
	TID2013	**0.952**	0.88	0.882	0.859	0.911	0.904	**0.952**	**0.937**
	CSIQ	0.952	0.953	0.954	**0.957**	**0.968**	0.929	**0.964**	0.951
SROCC	TID2008	**0.908**	0.827	0.830	0.805	**0.923**	0.857	**0.924**	0.910
	TID2013	**0.929**	0.878	0.879	0.855	**0.949**	0.898	**0.946**	0.925
	CSIQ	0.936	0.953	0.954	0.957	**0.969**	0.926	**0.964**	0.947
RMSE	TID2008	**0.219**	0.351	0.349	0.381	0.285	0.540	**0.234**	**0.248**
	TID2013	**0.217**	0.337	0.334	0.363	0.293	0.304	**0.218**	**0.247**
	CSIQ	0.051	0.051	0.05	**0.048**	**0.042**	0.062	**0.045**	0.052

TABLE X
PERFORMANCE ON WHITE NOISE

		PSNR	SSIM	MS-SSIM	VIFp	GSMD	FSIM	VSI	Proposed
PLCC	TID2008	0.872	**0.947**	**0.951**	0.943	0.887	0.945	0.946	**0.947**
	TID2013	0.895	0.880	**0.964**	**0.962**	0.892	0.955	**0.956**	0.948
	CSIQ	0.908	0.939	0.866	**0.957**	0.969	**0.957**	0.876	**0.958**
SROCC	TID2008	0.879	0.879	**0.955**	0.943	0.901	0.901	**0.953**	**0.947**
	TID2013	0.915	0.915	**0.968**	**0.964**	0.915	0.915	**0.961**	0.953
	CSIQ	0.929	0.929	**0.975**	0.967	**0.971**	**0.971**	0.968	0.968
RMSE	TID2008	0.575	**0.378**	**0.361**	0.389	0.541	0.382	0.381	**0.372**
	TID2013	0.556	0.592	**0.33**	**0.342**	0.565	0.370	**0.365**	0.397
	CSIQ	0.120	0.098	0.143	**0.083**	**0.071**	0.083	0.304	**0.082**

the least computing time as expected. Because SSIM and GSMD calculate the similarity of pixel and edge information respectively, their time consumption is slightly larger than PSNR and less than other algorithms. For PW-MSE, since the randomness is computed for the entire image, it increases the computational complexity, but it still has less or comparable time consumption comparing with the rest IQMs.

V. CONCLUSION

In this paper, PW-MSE is proposed for compressed images. The masking effect as well as the low-passing filter characteristics of the initial process of HVS is explored. To mathematically model and simulate the initial process in HVS, the CSF is adopted as the transfer function in frequency domain. The error signal from the compression distortion is filtered with the proposed transfer function in spatial domain, which removed most errors in high frequency that can not be perceived by humans. Furthermore, after processing through the initial part of HVS, the error signal is highly affected by various masking effects from different image contents. To study the masking effect quantitatively, the randomness is proposed to measure it by considering the spatial correlations. Moreover, a modulation relation among the randomness and the distortion before masking and after masking is investigated. By observing the relation of MOS and the distortion before masking effect, a modulation model is proposed at image level. Later, it is extended into pixel level, providing finer scale masking analysis. PW-MSE is tested on the databases with various compression distortions. By validating at every step, we could found that each step of PW-MSE contributes to overall performance improvement. The performance comparison with other benchmark IQMs demonstrates the effectiveness of PW-MSE.

APPENDIX
OPTIMAL ESTIMATION IN SINGULAR CASE

Since there is redundency in neighborhood information and thus $\mathbf{X}(u)$ is linear dependent, we could reduce the redundent information by transform the $\mathbf{X}(u)$ into linear independent vector as

$$\tilde{\mathbf{X}}(u) = \mathbf{Q}\mathbf{X}(u), \tag{21}$$

where $\mathbf{Q} = \Lambda_m^{-\frac{1}{2}}\mathbf{U}_m^T$ and Λ_m and $\mathbf{U}_m$ are the same as in Eq. (5). Instead of estimating with $\mathbf{X}(u)$, we estimate $Y(U)$ with $\tilde{\mathbf{X}}(u)$. Since $\mathbf{X}(u)$ can be fully recovered from $\tilde{\mathbf{X}}(u)$, the optimal estimation with $\tilde{\mathbf{X}}(u)$ is also optimal with $\mathbf{X}(u)$. The correlation matrix of $\tilde{\mathbf{X}}(u)$ is

$$\begin{aligned}\mathbf{R}_{\tilde{X}} &= E[\mathbf{Q}\mathbf{X}\mathbf{X}^T\mathbf{Q}^T] \\ &= \mathbf{Q}\mathbf{R}_X\mathbf{Q}^T \\ &= \mathbf{I},\end{aligned} \tag{22}$$

and the cross-correlation matrix of $Y(u)$ and $\tilde{\mathbf{X}}(u)$ is

$$R_{Y\tilde{X}} = R_{YX}Q^T, \tag{23}$$

Therefore, using Eq. (4), we could have the optimal estimation with $\tilde{\mathbf{X}}(u)$ as

$$\begin{aligned}\tilde{Y}(u) &= R_{Y\tilde{X}}\mathbf{R}_{\tilde{X}}^{-1}\tilde{\mathbf{X}}(u) \\ &= R_{YX}Q^TQX(u) \\ &= R_{YX}U_m\Lambda_m^{-1}U_m^TX(u),\end{aligned} \tag{24}$$

where $U_m\Lambda_m^{-1}U_m^T$ is the psudo-inverse as expressed in Eq. (5). When there is reduedency in neighboring pixels, we could use Eq. (24) to estimate current pixel.

REFERENCES

[1] J. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 525–536, Jul. 1974.

[2] A. P. Bradley, "A wavelet visible difference predictor," *IEEE Trans. Image Process.*, vol. 8, no. 5, pp. 717–730, May 1999.

[3] Y.-K. Lai and C.-C. J. Kuo, "A Haar wavelet approach to compressed image quality measurement," *J. Vis. Commun. Image Represent.*, vol. 11, no. 1, pp. 17–40, Mar. 2000.

[4] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.

[5] J. Lubin, "A human vision system model for objective picture quality measurements," in *Proc. Int. Broadcast. Conf.*, Sep. 1997, pp. 498–503.

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

[7] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, Jul. 2011.

[8] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 200–213, Jan. 2014.

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

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

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

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

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

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

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

[16] J. Zhu and N. Wang, "Image quality assessment by visual gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 919–933, Mar. 2012.

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

[18] L, Jin, K. Egiazarian, and C.-C. J. Kuo, "JPEG-based perceptual image coding with block-based image quality metric," in *Proc. 19th IEEE ICIP*, Sep./Oct. 2012, pp. 1053–1056.

[19] Y.-H. Huang, T.-S. Ou, P.-Y. Su, and C. H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1614–1624, Nov. 2010.

[20] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-Motivated rate-distortion optimization for video coding," *Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 516–529, Apr. 2012.

[21] V. Laparra, J. Muñoz-Marí, and J. Malo, "Divisive normalization image quality metric revisited," *J. Opt. Soc. Amer. A*, vol. 27, no. 4, pp. 852–864, Apr. 2010.

[22] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer. A*, vol. 14, no. 9, pp. 2379–2391, Sep. 1997.

[23] D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *J. Opt. Soc. Amer. A*, vol. 20, no. 7, pp. 1164–1180, Jul. 2003.

[24] A. B. Watson, R. Borthwick, and M. Taylor, "Image quality and entropy masking," *Proc. SPIE*, vol. 3016, pp. 2–12, Jun. 1997.

[25] S. He, P. Cavanagh, and J. Intriligator, "Attentional resolution and the locus of visual awareness," *Nature*, vol. 383, no. 6598, pp. 334–337, Sep. 1997.

[26] S. T. L. Chung, G. E. Legge, and B. S. Tjan, "Spatial-frequency characteristics of letter identification in central and peripheral vision," *Vis. Res.*, vol. 42, no. 18, pp. 2137–2152, Aug. 2002.

[27] P. G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. Bellingham, WA, USA: SPIE, 1999.

[28] A. B. Watson and A. J. Ahumada, Jr., "A standard model for foveal detection of spatial contrast," *J. Vis.*, vol. 5, no. 9, pp. 717–740, Oct. 2005.

[29] S. Park, E. Clarkson, M. A. Kupinski, and H. H. Barrett, "Efficiency of the human observer detecting random signals in random backgrounds," *J. Opt. Soc. Amer. A*, vol. 22, no. 1, pp. 3–16, Jan. 2005.

[30] D. H. Kelly, "Motion and vision. II. Stabilized spatio-temporal threshold surface," *J. Opt. Soc. Amer. A*, vol. 69, no. 10, pp. 1340–1349, Oct. 1979.

[31] S. J. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," *Proc. SPIE*, vol. 3299, pp. 180–191, Jul. 1998.

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

[33] J. M. Foley and G. M. Boynton, "New model of human luminance pattern vision mechanisms: Analysis of the effects of pattern orientation, spatial phase, and temporal frequency," *Proc. SPIE*, vol. 2054, pp. 32–42, Mar. 1994.

[34] K. N. Ngan, K. S. Leong, and H. Singh, "Adaptive cosine transform coding of images in perceptual domain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1743–1750, Nov. 1989.

[35] N. B. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.*, vol. 33, no. 6, pp. 551–557, Jul. 1985.

[36] C. J. van den Branden Lambrecht and M. Kunt, "Characterization of human visual sensitivity for video imaging applications," *Signal Process.*, vol. 67, no. 3, pp. 255–269, Jun. 1998.

[37] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 337–346, Mar. 2009.

[38] (Mar. 2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase I*. [Online]. Available:http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

[39] (Aug. 2003). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II*. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII

[40] *Toyoma Database*. [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html, accessed Oct. 3, 2015.

[41] F. De Simone, L. Goldmann, V. Baroncini, and T. Ebrahimi, "Subjective evaluation of JPEG XR image compression," *Proc. SPIE*, vol. 7443, p. 74430L, Aug. 2009.

[42] *Information Technology—Digital Compression and Coding of Continuous-Tone Still Images: Requirements and Guidelines*, document ITU T.81, 1993.

[43] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 36–58, Sep. 2001.

[44] S. Srinivasan, C. Tu, S. L. Regunathan, and G. J. Sullivan, "HD photo: A new image coding technology for digital photography," *Proc. SPIE*, vol. 6696, p. 66960A, Sep. 2007.

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

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

[47] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3441–3452, Nov. 2006.

[48] H. Tao, N. Klomp, and I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 529–539, Apr. 2010.

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

[50] F. Pan *et al.*, "A locally adaptive algorithm for measuring blocking artifacts in images and videos," *Signal Process., Image Commun.*, vol. 19, no. 6, pp. 499–506, Jul. 2004.

[51] N. Ponomarenko *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *Proc. 4th Eur. Workshop Vis. Inf. Process.*, Jun. 2013, pp. 106–111.

[52] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006-1–011006-21, Jan. 2010.
[53] S. Winkler, "Visual quality assessment using a contrast gain control model," in *Proc. IEEE 3rd Workshop Multimedia Signal Process.*, Sep. 1999, pp. 527–532.
[54] P. C. Teo and D. J. Heeger, "Perceptual image distortion," *Proc. SPIE*, vol. 2179, pp. 127–141, May 1994.
[55] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, "Nonlinear image representation for efficient perceptual coding," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 68–80, Jan. 2006.
[56] A. Watson and C. Ramirez, "A standard observer for spatial vision," *Invest. Ophthalmol. Vis. Sci.*, vol. 41, no. 4, p. 713, 2000.
[57] M. Ramasubramanian, S. N. Pattanaik, and D. P. Greenberg, "A perceptually based physical error metric for realistic image synthesis," in *Proc. 26th Annu. Conf. SIGGRAPH*, 1999, pp. 73–82.
[58] Z. Wang, A. C. Bovik, and L. Lu, "Wavelet-based foveated image quality measurement for region of interest image coding," in *Proc. ICIP*, 2001, pp. 89–92.
[59] Y. Liu and J. P. Allebach, "A computational texture masking model for natural images based on adjacent visual channel inhibition," *Proc. SPIE*, vol. 9016, p. 90160D, Jan. 2014.

Sudeng Hu received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2007, and the M.Phil. degree from the Department of Computer Science, City University of Hong Kong, Hong Kong, in 2010. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, University of Southern California, Los Angeles. From 2010 to 2011, he was a Research Associate with the Department of Computer Science, City University of Hong Kong. His research interests include image and video compression, scalable video coding, 3-D video coding, image and video quality assessment. He received the 2014 Chinese Government Award for Outstanding Self-Financed Students Abroad.

Lina Jin received the B.S. degree from Jilin University, Changchun, China, in 2005, and the M.Sc. and Ph.D. degrees from the Tampere University of Technology (TUT), Tampere, Finland, in 2010 and 2015, respectively. From 2009 to 2014, she was a Researcher with TUT. She joined the Multimedia Communication Laboratory, University of Southern California, as a Research Assistant in 2013. Her research interests include image and video quality metrics, quality of experience for multimedia, image and video compression, and image enhancement.

Hanli Wang (M'08–SM'12) received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2007. From 2007 to 2008, he was a Research Fellow with the Department of Computer Science, City University of Hong Kong. From 2007 to 2008, he was also a Visiting Scholar with Stanford University, Palo Alto, CA, invited by Prof. C. K. Chui. From 2008 to 2009, he was a Research Engineer with Precoad, Inc., Menlo Park, CA. From 2009 to 2010, he was an Alexander von Humboldt Research Fellow with the University of Hagen, Hagen, Germany. In 2010, he joined the Department of Computer Science and Technology, Tongji University, Shanghai, China, as a Professor. His current research interests include digital video coding, image processing, computer vision, and machine learning. He has authored over 80 papers in these fields.

Yun Zhang (M'12) received the B.S. and M.S. degrees in electrical engineering from Ningbo University, Ningbo, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, in 2010. From 2009 to 2014, he was a Post-Doctoral Researcher with the Department of Computer Science, City University of Hong Kong, Hong Kong. In 2010, he became an Assistant Professor with the Shenzhen Institutes of Advanced Technology, CAS, where he has served as an Associate Professor since 2012. His research interests are video compression, 3-D video processing, and visual perception.

Sam Kwong (F'13) received the B.S. degree in electrical engineering from the State University of New York at Buffalo, in 1983, the M.S. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from the University of Hagen, Germany, in 1996. From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada. He joined Bell Northern Research Canada as a member of the Scientific Staff. In 1990, he became a Lecturer with the Department of Electronic Engineering, City University of Hong Kong, where he is currently a Professor with the Department of Computer Science. His research interests are video and image coding and evolutionary algorithms.

C.-C. Jay Kuo (F'99) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively. He is currently the Director of the Multimedia Communications Laboratory and a Professor of Electrical Engineering, Computer Science and Mathematics with the Ming-Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles. He has co-authored about 200 journal papers, 850 conference papers, and ten books. His research interests include digital image/video analysis and modeling, multimedia data compression, communication and networking, and biological signal/image processing. He is a fellow of the American Association for the Advancement of Science and the International Society for Optical Engineers.