Quantitative Measurement of Perceptual Attributes and Artifacts for Tone-Mapped HDR Display

Mingxing Jiang, Liquan Shen, *Member, IEEE*, Min Hu, Ping An, *Member, IEEE*, Yu Gu, *Senior Member, IEEE*, and Fuji Ren, *Senior Member, IEEE*

***Abstract*— Measuring electronic display quality, as perceived by human observers, has attracted high attention in current consumer displays. With limited dynamic range of consumer-level standard dynamic range (SDR) displays, high dynamic range (HDR) scenes are often rendered by different tone mapping operators (TMOs). Due to the lack of quantitative measurement and analysis of tone mapping in the existing image quality measurement (IQM) methods, it is of considerable significance to establish new IQM protocols that can differentiate the display quality of SDR electronic devices. We first propose an IQM model to exhibit the essential perceptual attributes and artifacts that are peculiar to tone mapping. Furthermore, we characterize the overall image quality (OiQ) resulting from linear regression and various machine learning techniques. Finally, the execution of without HDR reference ablation experiments demonstrates the relative contribution of these attribute measurements to OiQ. The use of IQM protocols helps with well-founded quality measurement between TMOs during tone mapping processing. Our effort is not only useful to get into the tone mapping field or when implementing a TMO but also sets the stage for quantitative measurement of TMOs. By monitoring these attributes and artifacts after different tone mapping process, user-driven or optimal display is made possible.**

***Index Terms*— High dynamic range (HDR), image attributes, image quality measurement (IQM), machine learning, tone mapping operators (TMOs).**

I. Introduction

AS THE video industry begins deployment of ultrahigh-definition display in both professional and consumer markets, including support for high dynamic range (HDR) services has become mainstream focus. The HDR imaging technology provides users a brand new visual experience, but most consumer-level display devices have traditional standard dynamic range (SDR) and cannot directly render HDR images. Thus, the mainstream method is to use tone mapping to compress the dynamic range to fit SDR display devices [1]–[4]. Different tone mapping operators (TMOs) meet the requirements of diverse aspects during the conversion process: brightness, contrast, color, and details, as shown in Fig. 1. This process is qualitative in scene as it aims at tricking the observer into thinking that the image shown on an SDR medium has actually different scene attributes from the original HDR image [5]. Through observation, it is not one attribute, but multiple attributes that affect the overall image quality (OiQ). Since TMOs compress the dynamic range, information loss is inevitable, and therefore measurement of the quality of tone-mapped result is important. Nevertheless, as we will see below, the quantitative measurement of these attributes and related artifacts is possible and it fits with our perception of the scene.

The current image quality measurement (IQM) methods are challenged to quantify the performance of these electronic displays [6]–[10] without a comprehensive viewpoint for addressing the perceptual quality measurement problem. The existing research efforts about the quality measurement of tone-mapped HDR mainly focus on the following three aspects: 1) mainly exploiting novel natural scene statistics (NSSs) features to evaluate naturalness [11]–[14]; 2) mostly considering image structure [15], texture [11], and entropy [16] to measure the structural fidelity; and 3) rarely considering the exposure distortion information caused by tone mapping [17]–[20]. However, perceptual quality of display generally favors signal-independent metrics, including the evaluation of brightness, contrast, color, and details. Quantitative measurement and analysis of these attributes related to tone mapping are missing. Therefore, a systematic comparison of representative TMOs is highly desirable, and essential perceptual attributes and specific artifacts can be measured quantitatively.

The novelty of this article is to present new measurement protocols to assess the quality of SDR displays using tone mapping. The IQM protocols can study quantitatively these attributes (i.e., brightness, color, contrast, detail) and halo artifacts related to tone mapping. Meanwhile, these essential attributes are not independent and simultaneously influence the OIQ of tone-mapped images. Fig. 2 presents a description

Manuscript received 3 March 2022; revised 12 May 2022; accepted 31 May 2022. Date of publication 22 June 2022; date of current version 1 July 2022. This work was supported by the National Natural Science Foundation of China under Grant 61931022, Grant 61671282, and Grant 62176084. The Associate Editor coordinating the review process was Dr. Jing Lei. *(Corresponding authors: Liquan Shen; Min Hu.)*

Mingxing Jiang is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology, Hefei 230602, China (e-mail: mx0551@163.com).

Liquan Shen and Ping An are with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: jsslq@163.com; anping@shu.edu.cn).

Min Hu and Yu Gu are with the School of Computer and Information, Hefei University of Technology, Hefei 230602, China (e-mail: jsjxhumin@hfut.edu.cn; yugu.bruce@ieee.org).

Fuji Ren is with the School of Computer Science and Engineering, University of Electronic Science and Technology, Chengdu 610054, China, and also with the Faculty of Engineering, Tokushima University, Tokushima 770-8500, Japan (e-mail: ren@is.tokushima-u.ac.jp).

Digital Object Identifier 10.1109/TIM.2022.3185322

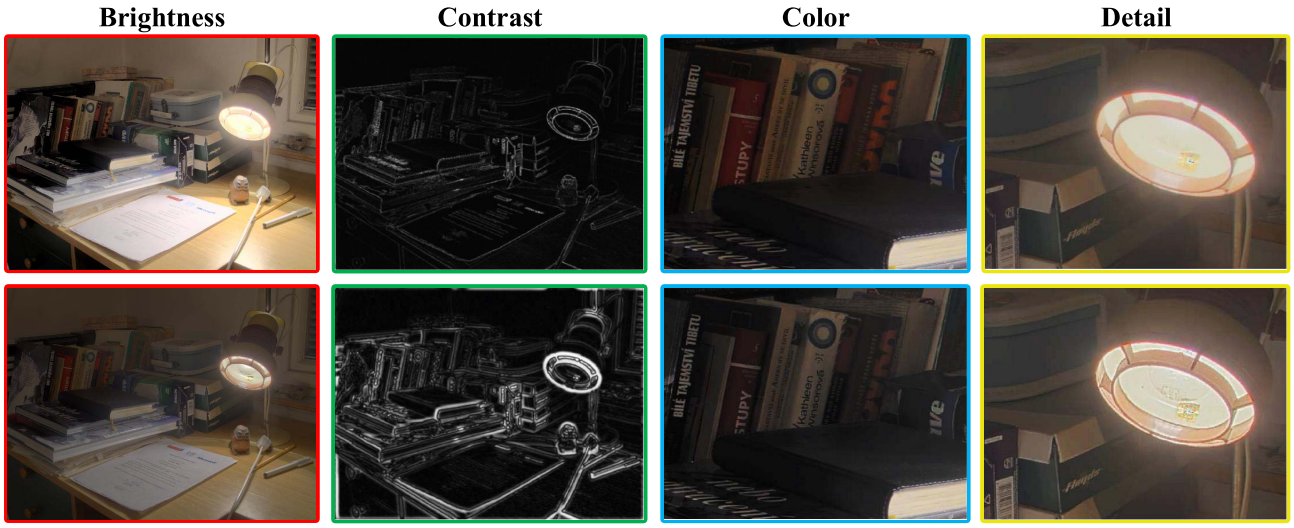


Fig. 1. Illustration of different attributes at each column after tone-mapped HDR imaging for electronic display: brightness, contrast, color, and details.

of the relationships between these attributes [21]. For each attribute, the OIQ is associated with it. First, human vision system (HVS) depends strongly on the change in brightness distribution because dynamic range compression during tone mapping processing is prone to under- or overexposed scenes. Moreover, simulating the appearance of color at the dark spot illumination level can change the perceived brightness to a great extent [22]. Similarly, the reproduction of contrast and details is certainly essential to enrich tone-mapped image naturalness. The simulation of visual acuity loss can effectively perceive the quality of dim night scenes, while the simulation of glare can only improve the perception of dark scenes with bright light sources. Finally, the emergence of disturbing artifacts also causes quality degradation.

Encouraged by these findings, a qualitative measurement of perceptual attributes and artifacts is proposed to simulate the user to characterize the quantity of tone-mapped image attributes. To be specific, we aim at measuring quantitatively the perceptual attributes of tone-mapped HDR display with different machine learning techniques. This process can verify how well TMOs reproduce these attributes: brightness, contrast, color, detail, and artifacts. These selected attributes are related to HVS, but they also include cognitive and esthetic aspects. Beyond these attributes, which are related to color and spatial vision, there are other important artifacts and some special effects which can improve or modify the final appearance. Following are the contributions of the article.

1) We propose a set of quality measurement protocols without specific test patterns, which are used for automatic qualitative measurement of essential perceptual attributes and halo artifacts for tone-mapped HDR display. In other words, these protocols form a consolidated model to quantitatively compute display quality of tone-mapped image.
2) These measured attributes (i.e., brightness, color, contrast, and detail) are precisely defined; they reflect proven aspects of human visual system (HVS) such as exposure degree and color difference. Specifically, we calculate energy deviation and color distance after channel multiplication to quantitatively describe exposure and color. Instead of using space-domain statistics, these are the first attempt to use exposure-related naturalness for quality measurement of these attributes.
3) The halo artifacts caused by tone mapping are measured by calculating statistics of a defined edge map, described as mean, standard deviation, kurtosis, and skewness. The experimental results demonstrate that these described indexes can accurately monitor edge dispersion.

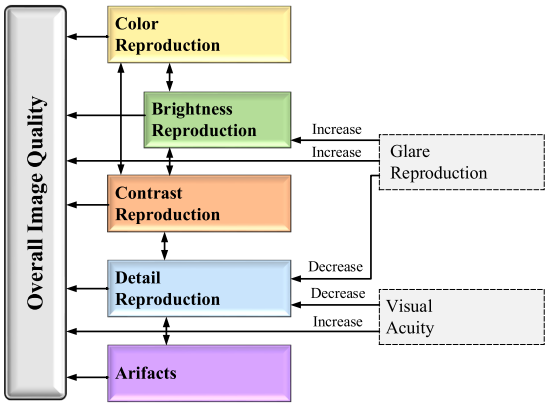


Fig. 2. Description of the relationship between image attributes [21].

The remainder of this article is organized as follows. In Section II, we describe the proposed quantitative measurement method. Section III presents the experimental results with the latest methods by applying them on public benchmark databases and visualizes the results of individual distortion types. In Section IV, we conclude the article with an expectation of future work.

II. Proposed Quantitative Measurement Method

Fig. 3 generally describes a framework of the proposed quantitative measurement metric for optimal tone-mapped

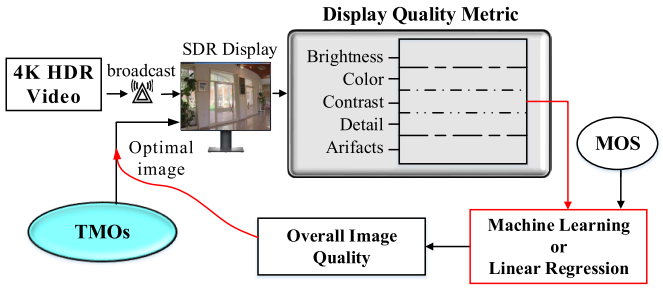


Fig. 3. Framework of the design of the proposed QmTm model.

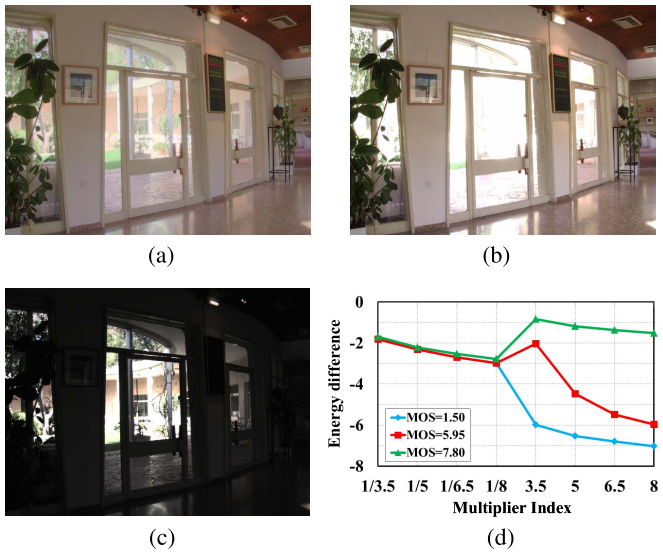


Fig. 4. Illustration of how energy difference $\triangle E$ varies with the change in multiplier M_i. Note that a smaller MOS value indicates a better quality here. (a) MOS = 1.50. (b) MOS = 5.95. (c) MOS = 7.80. (d) Multiplier index.

HDR display, abbreviated as QmTm. This display metric consists of a set of IQM protocols to quantify essential perceptual attributes and artifacts. TMOs render the dynamic effect HDR images in SDR display at the expense of some perceptual information. The QmTm presents clear protocols to quantitatively measure these attributes mentioned above. The measurement criteria include absolute energy deviation (AED), color difference, entropy, gradient similarity, and statistics of edge map so as to measure brightness, color, contrast, detail, and halo artifacts. Quality regression module based on machine learning techniques or linear regression (marked with red box) is adopted to establish the connection between these criterion features and subjective scores [i.e., mean opinion score (MOS)]. The optimal image can be automatically selected from a group of candidates by the proposed QmTm.

A. Brightness

Abnormal exposure is easy to occur at regions with inconsistent luminance distribution in a tone-mapped image, as shown in Fig. 4. Fig. 4(a) with normal exposure seems natural, such as the view outside the door and the paintings on the wall. Instead, the scenes rendered in Fig. 4(b) and (c) show overexposure and underexposure appearance, respectively. Therefore, we can perceive roughness or glossiness of object surface. Different brightness levels of electronic display could impact on the viewing experience [23]. Referring to the benchmark MOS in available tone mapping database, higher overexposure or underexposure tone-mapped image has worse perceptual quality. Meanwhile, it can be observed that luminance distribution shifts to the maximum/minimum value under abnormal exposure. In other words, the greater the intensity deviation, the more obvious the degree of abnormal exposure.

To measure brightness attribute in the QmTm metric, the energy difference of a tone-mapped image after its multiplication is proposed to evaluate brightness distortion. Inspired by [11] and [15], the basic consideration of our metric is to straightforwardly measure the exposure degree in tone-mapped image and its generated new images. To be specific, we first darken and brighten the original tone-mapped image to produce a group of derived images. The derived image I_t can be generated by the original image I with multiplier M_i

$$I_t = I * M_i \tag{1}$$

where M_i indicates the ith multiplier. Second, AED can be estimated as the energy difference E_{d_i} between I and its derived images. Fig. 4(d) shows that E_{d_i} varies with the change in M_i, including the sharp difference in brightening. Since the information entropy reflects the average information preservation in images, we use information entropy of these derived images as its amount of energy. Specifically, the entropy value E of a tone-mapped image can be produced by

$$E = -\sum_{i=1}^{n} P(a_i)\log(P(a_i)) \tag{2}$$

where $P(a_i)$ represents the probability of occurrence of each component. Then, the overall exposure degree can be estimated as the energy difference $\triangle E$ between I and its derived images I_i. Formally, $\triangle E_i$ can be calculated as

$$\triangle E_i = E_{I_i} - E_I \tag{3}$$

where E_I and E_{I_i} indicate the entropy values of I and its derived images, respectively. AED is used as a criterion feature to evaluate brightness attribute, and lower values indicate moderate exposure. To balance efficacy and efficiency, we take advantage of eight AED values that are measured using $M = \{(1/n), n \mid n = 3.5, 5, 6.5, 8\}$ as features (marked as $\mathbf{F}_{\mathrm{Br}}$).

B. Reproduction of Color

Color is a quantity that measures the subjective consciousness of chromatic property under different luminance environments. HVS adapts differently in front of a real HDR scene and in front of a typical SDR display. Even with an accurate HDR image of a scene, the tone-mapped image may appear different from the real scene due to different adaptation of HVS [9], [24]. To describe colorfulness, we first convert the RGB space into the CIE LAB color space. By observing 2-D histograms of $a*$ and $b*$ channels in Fig. 5(a)–(c), color difference is reduced in underexposure/overexposure environments, as the light-sensitive rods take over for the color cast

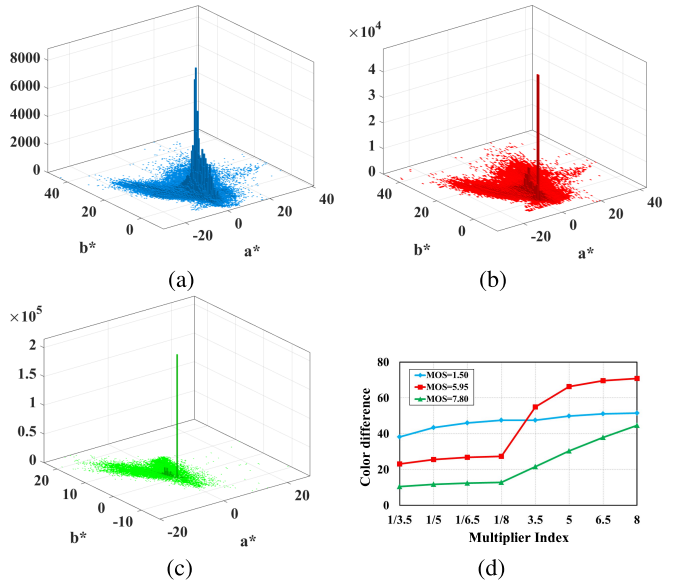


Fig. 5. (a)–(c) 2-D histograms of $a*$ and $b*$ channels corresponding to Fig. 4(a)–(c). (d) Corresponding color difference of (a)–(c) under different multipliers M_i.

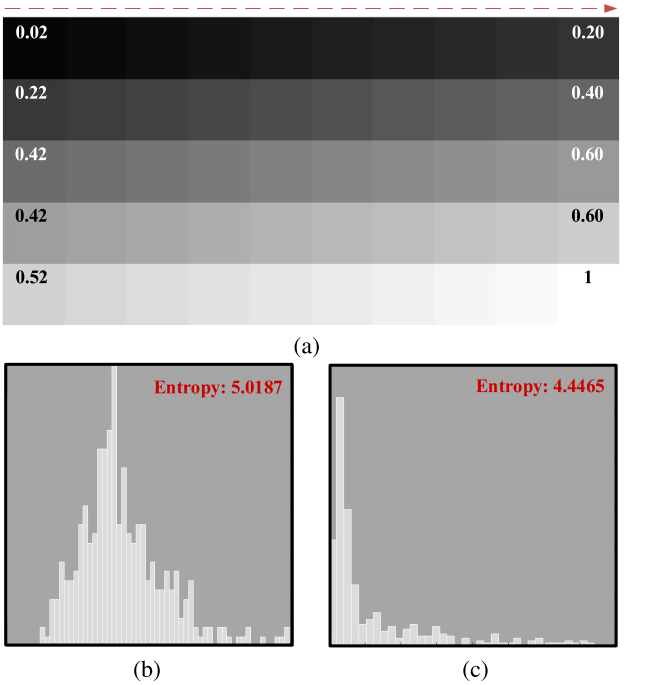


Fig. 6. Local contrast analysis using entropy. (a) Benchmark grayscale of the charts and (b) and (c) corresponding normalized histograms and entropy.

cone system. With normal exposure in Fig. 4(a), Fig. 5(a) exhibits multiple scattered peaks. Instead, Fig. 5(b) and (c) possesses a single peak or a few peaks that are concentrated and far from the neutral axis. These indicate that the color deviation is serious. As the above multiplier M_i is raised, the cone system becomes active and colors begin to be not seen [see Fig. 5(d)]. To simulate the appearance of color at different illumination levels, we compute the distance ΔLab to estimate image colorfulness by the following formula:

$$\Delta\text{Lab} = \sqrt{(L_{I_i} - L_I)^2 + (a_{I_i} - a_I)^2 + (b_{I_i} - b_I)^2} \quad (4)$$

where L_I, a_I, b_I and L_{I_i}, a_{I_i}, b_{I_i} are the channel values of I and its derived images using above multiplier M_i, respectively. From Fig. 5(d), the overall color difference of an overexposure/underexposure image increases with the increase in multiplier index. Therefore, the color difference can effectively quantify color information primarily, denoted by $\mathbf{F_{Cl}}$.

C. Local Contrast Preservation

Very present defects are halos and contrast loss resulting from tone mapping of HDR images, as illustrated in Fig. 1. We observe that local contrast perception along with a spatial/intensity coherence in the image is an important factor to account for a good HDR image. Since humans do not perceive absolute intensities but rather local contrast changes [7], [25], our measure uses a benchmark grayscale of the charts in Fig. 6(a), which is composed of 50 uniform patches between 0 and 1 of transmittance. This grayscale allows to measure how a tone-mapped HDR image preserves local dynamic range of each part of the image. To measure local contrast preservation, the entropy of the normalized histogram hist_{gs} of the grayscale chart is computed by

$$\text{Entropy}_{\text{gs}} = \sum_i \text{hist}_{\text{gs}}(i) \log \frac{1}{\text{hist}_{\text{gs}}(i)}. \quad (5)$$

The entropy can be used to quantify the information contained in the grayscale chart (feature marked as: $\mathbf{F}_{\text{Ct}}$). For a grayscale with many saturation values in the dark or bright part, its entropy will be lower than that of the uniformly distributed grayscale [as illustrated in Fig. 6(b) and (c)]. The entropy of a grayscale with uniformly distributed values is equal to the dynamic entropy of the grayscale. Meanwhile, the entropy has certain limitations which cannot express spatial information. For example, a grayscale with clear halos can have high entropy value but bad perceptual quality.

D. Reproduction of Detail

Abnormal exposure also induces detail distortions, which makes local structure distortion conspicuous. Because the dynamic range of the SDR device is limited, the overflow values in these areas are most likely to be clipped. Common tone methods, e.g., linear scaling or clamping, are applied uniformly over the texture [see Fig. 4(b) and (c)]. Inspired by the above analysis, we capture some parameters of gradient distributions to quantify these detail distortions. More concretely, the gradient magnitude G can be expressed as

$$G = \sqrt{(I \otimes p_x)^2 + (I \otimes p_y)^2} \quad (6)$$

where $\otimes$ is the convolution filtering, and p_x and p_y are the horizontal and vertical filter kernels, respectively. Here, we specially design a modified Prewitt filter, which can be defined as

$$p_x = \begin{bmatrix} 1/5 & 0 & -1/5 \\ 3/5 & 0 & -3/5 \\ 1/5 & 0 & -1/5 \end{bmatrix}, \quad p_y = p_x^T \quad (7)$$

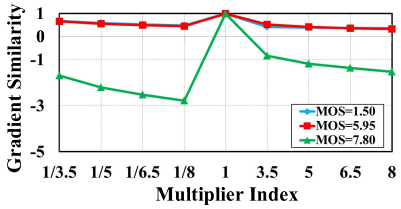


Fig. 7. Illustration of gradient similarity varies with the change in the multiplier M.

where ^{T} denotes the transpose operation. Then, as abnormal exposure causes structure distortion, we compare I and its derived image I^+ (using above darken/brighten multiplier) via the following gradient similarity:

$$S(i,j) = \frac{2G_I(x,y)\cdot G_{I^+}(x,y) + C_1}{G_I^2(x,y)\cdot G_{I^+}^2(x,y) + C_1} \tag{8}$$

where C_1 is a small constant to keep $S(i,j)$ stable; G_I and G_{I^+} express the gradient magnitudes of I and its derived images, respectively. Fig. 7 conveys the gradient similarity variation with different multiplier indexes. It is clear that gradient similarity of Fig. 4(c) falls faster than those of Fig. 4(a) and (b). This explains that underexposed image loses more detail information than overexposed ones. Furthermore, we use CS-LBP descriptor [26] (with radius $R = 1$ and sampling points $P = 8$) to measure the local structure detail of tone-mapped image. As a result, 16 features can be obtained from each CS-LBP map. Overall, the detail features can be measured by the gradient similarity and local structure information, denoted by $\mathbf{F}_{\mathrm{De}}$.

E. Halo Artifacts

In addition, we further explore halo artifacts related to local tone mapping algorithms present in tone mapping technology, as shown in Fig. 8. These artifacts are caused by pursuing high contrast edge, typically happens for small bright features or sharp high contrast edges, and exhibit dark bands or halos [27]. Compared with Fig. 8(b), Fig. 8(a) shows stratification and demarcation, especially in the edge region. This halo effect makes the edge blur and spread the blur distortions to the surrounding area. As mentioned above, due to the limitation of entropy to express clear halos, we compute the statistics of a defined edge map to quantify the halo artifacts. First, the edge map of a tone-mapped is calculated by the Canny descriptor (with default settings in MATLAB software). The generated edge maps are shown in Fig. 8(c) and (d). Fig. 8(d) exhibits a clear structure and edge than Fig. 8(c). Then, we consider using statistical indicators over all the patches in the edge map to quantify the halo artifacts. Here, we adopt 8×8 image patch at each edge point. Finally, the mean, standard deviation, kurtosis, and skewness of these patches are calculated to measure the edge dispersion (feature marked as: $\mathbf{F}_{\mathrm{Ar}}$). In particular, the mean value represents the holistic tendency, and the variance value indicates the fluctuation of local contrast. As shown in Fig. 8(e) and (f), Fig. 8(a)

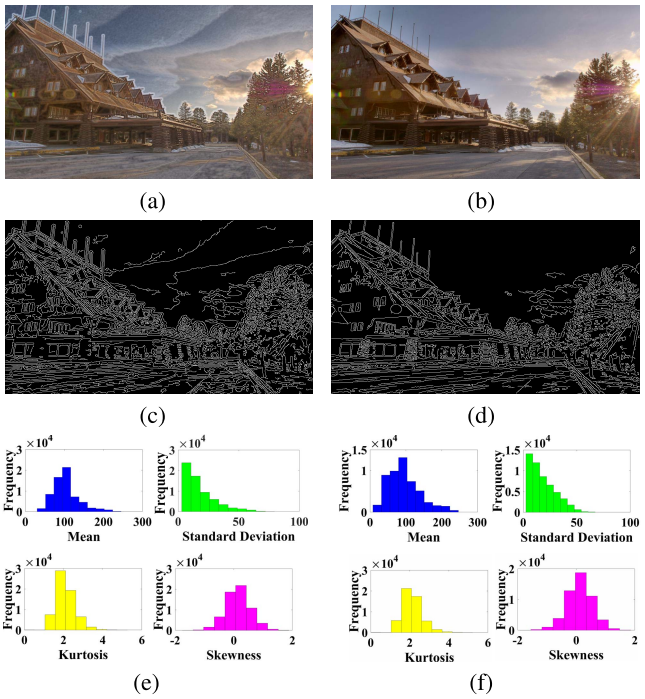


Fig. 8. Edge dispersion measurement. (a)–(b) Different tone-mapped images with the same scene, (c)–(d) corresponding edge maps, and (e)–(f) statistical indicators based on the edge map.

with halo artifacts has higher mean value and lower higher standard deviation value. In general, the mean only reflects the overall strength of edge, and the standard deviation effectively captures local detailed strength. Compared with Fig. 8(b), the kurtosis distribution of Fig. 8(a) is more abrupt, and the skewness distribution of Fig. 8(a) deviates from the central axis (i.e., 0).

F. Overall Image Quality Estimation

With the above description, we have captured 29 criterion features via quantitative measurement of essential perceptual attributes and artifacts. Since these attributes and artifacts are not mutually independent, we first combine them into a feature vector, denoted as $F = \{\mathbf{F}_{\mathrm{Br}}, \mathbf{F}_{\mathrm{Cl}}, \mathbf{F}_{\mathrm{Ct}}, \mathbf{F}_{\mathrm{De}}, \text{and } \mathbf{F}_{\mathrm{Ar}}\}$. Next, we construct an OiQ prediction model by establishing the connection between these criterion features and subjective scores [i.e., mean opinion score (MOS)]. For prediction, we compare linear regression and various machine learning techniques to model this connection.

III. Experimental Result and Analysis

A. Experimental Protocol

1) Database: The ESPL-LIVE HDR database [28] is currently the abundant and most challenging tone mapping image database. It has three distortion categories, i.e., TMO, multi-exposure fusion (MEF), and postprocessing (PP), as described in Table I. This database includes 1811 images generated from hundreds of HDR scene images.

TABLE I
ABBREVIATIONS OF EVALUATED TMOS ON THE ESPL-LIVE HDR DATABASE

Abbreviation	Method description	Number of images	Type
Durand	Fast bilateral filtering for the display of high dynamic-range images	187	TMO
Fattal	Gradient domain high dynamic range compression	187	TMO
Reinhard	Photographic tone reproduction for digital images	192	TMO
Ward	A visibility matching tone reproduction operator for high dynamic range scenes	181	TMO
Global energy weighting	-	89	MEF
Local energy weighting	-	89	MEF
Paul	Multi-exposure and multifocus image fusion in gradient domain	179	MEF
Pece Kautz	Bitmap movement detection: HDR for dynamic scenes	176	MEF
Raman	Bilateral filter based compositing for variable exposure photography	177	MEF
Grunge	-	178	PP
Surreal	-	176	PP

2) Performance Criteria: Three common indicators are applied to evaluate the prediction accuracy and monotonicity of IQM methods, including Pearson linear correlation coefficient (PLCC), spearman rank correlation coefficient (SRCC), and root-mean-squared error (RMSE). Specifically, PLCC and RMSE measure the prediction linearity and accuracy, while SRCC is a nonparametric rank-order-based correlation metric to measure the prediction monotonicity. Higher PLCC and SRCC values indicate a better performance, and the trend is reversed for RMSE value. To implement nonlinear mapping, a following five-parameter logistic regression function is required to be performed as follows:

$$Q(p) = \zeta_1 \cdot \left[\frac{1}{2} - \frac{1}{1+\exp(\zeta_2 \cdot (p-\zeta_3))}\right] + \zeta_4 \cdot p + \zeta_5 \quad (9)$$

where p and $Q(p)$ denote the scores estimated by an IQM model and the corresponding mapped scores, respectively; ($\zeta_1, \zeta_2, \zeta_3, \zeta_4$, and ζ_5) are the parameters of nonlinear function which are guided by Video Quality Experts Group [29] for initialization settings.

3) Parameter Selection: Since our QmTm metric included each attributes and artifacts' measure, several parameters may affect the final performance of quantitative measurement. From the description in Section II, multiplier index M_i used in image channel multiplication is a key parameter, where i denotes the ith multiplier. Here, we test QmTm with different M_i quantities, including $M_2 \in \{(1/3.5), 3.5\}$, $M_4 \in \{(1/3.5), (1/5), 3.5, 5\}$, $M_6 \in \{(1/3.5), (1/5), (1/6.5), 3.5, 5, 6.5\}$, $M_8 \in \{(1/3.5), (1/5), (1/6.5), (1/8), 3.5, 5, 6.5, 8\}$, and $M_{10} \in \{(1/3.5), (1/5), (1/6.5), (1/8), (1/9.5), 3.5, 5, 6.5, 8, 9.5\}$. Fig. 9(a) shows the experimental results where characters on the x-axis represent the number of M_i corresponding to the above five conditions. With the increase in M_i quantity, it can be observed that the SRCC values increase first and then decrease, including the maximum when M_8. We can conclude that the choice of M_i significantly affects the final performance, and more quantity does not necessarily lead to consistently better results. Besides, block size used in artifact measurement is also important. To find optimal block size, we select four different block sizes to measure the performance, including 8×8, 16×16, 32×32, and 64×64. The results are shown in Fig. 9(b). One can see that QmTm has little sensitivity to block size on the final performance. Here, we fix the block size as 8×8.

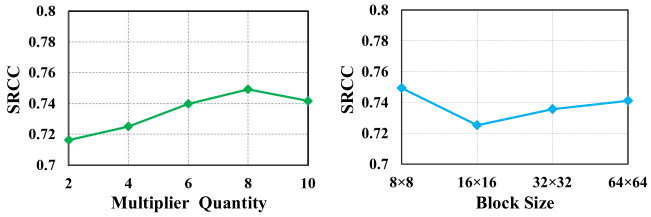


Fig. 9. SRCC values of the proposed QmTm model under different parameter settings: (a) multiplier quantity and (b) block size.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT QUALITY REGRESSION MODELS

Quality Regression Model	PLCC	SRCC	RMSE
Linear Regression	0.6456	0.6357	7.6843
Random Forest	0.7331	0.7230	6.8422
SVR	**0.7538**	**0.7492**	**6.6220**
Extreme Learning Machine	0.7389	0.7255	6.8094

B. Impacts of Machine Learning Methods

In general, it is difficult to reach a consensus on the objective measurement of OIQ. However, several HVS-related image attributes can prove their effectiveness in experiments. Consequently, we use linear regression and various machine learning techniques to predict OIQ score, including linear regression, random forest, support vector regression (SVR), and extreme learning machine. From Table II, we can find that machine learning techniques are generally better at prediction than the simple linear regression method. Also, among these machine learning techniques, SVR shows better performance than random forest. Although extreme learning machine has faster speed than SVR, its performance is worse than SVR. By testing with these machine learning techniques, higher performance was achieved by SVR, and thus SVR was chosen as the default regression model in QmTm.

TABLE III
PERFORMANCE COMPARISON ON THE ESPL-LIVE HDR DATABASE

Method	GM-LOG [30]	BLIINDS-II [31]	BRISQUE [32]	NIQE [33]	C-DIIVINE [34]	DIIVINE [35]
PLCC	0.5075	0.4418	0.3901	0.0768	0.8584	0.4840
SRCC	0.4493	0.4291	0.3493	0.0904	0.8416	0.4820
RMSE	9.6135	9.0157	9.2159	10.001	7.3421	8.8210
Method	BTMQI [11]	HIGRADE [12]	BLIQUE-TMI [13]	Yue [20]	Wang [19]	QmTm
PLCC	0.6175	0.7304	0.7120	0.7384	0.7430	**0.7538**
SRCC	0.6152	0.7295	0.7040	0.7328	0.7500	**0.7492**
RMSE	7.8674	9.6135	-	6.8354	-	**6.6220**

C. Performance Evaluation

1) Overall Performance Comparison for OiQ Prediction: In this section, we evaluate the performance using the underlying data gathered from the ESPL-LIVE HDR Database. Using common conventions of the learning-based methods, we randomly divide the dataset into 80% training set and 20% testing set. To ensure the credibility of the reported results, such training–testing separation is repeated 1000 times without overlap. In QmTm, the SVR is adopted to learn an OiQ predictor from the feature space to quality space. To measure the performance of the proposed QmTm, we compare the proposed QmTm with two categories of the state-of-the-art methods. The first category includes those method proposed for evaluating ordinary SDR image quality, such as GM-LOG [30], BLIINDS-II [31], BRISQUE [32], NIQE [33], C-DIIVINE [34], and DIIVINE [35]. These metrics are designed for measuring common distortions, such as blur, noise, and JPEG compression, which have intrinsically different characteristics from tone mapping distortions. The second category contains those proposed for tone-mapped images, including BTMQI [11], HIGRADE [12], BLIQUE-TMI [13], Yue's method [20], and Wang's method [19]. Specifically, BTMQI [11] focused more on measuring detail preservation with information entropy, and only used the edge magnitude to quantify the structure information. HIGRADE [12] emphasized naturalness more and proposed a leading no-reference method using space-domain NSS features and HDR-specific gradient-based features. BLIQUE-TMI [13] quantified image naturalness via moment statistics and NSS statistics in color channels. Yue *et al.* [20] used multiple quality-sensitive features from colorfulness, exposure, and naturalness aspects. Wang *et al.* [19] exploited both local degradation characteristics and global statistical properties for feature extraction, such as texture, structure, colorfulness, and naturalness.

The overall comparison results are listed in Table III where the best performance is in bold. First, the first six metrics designed for evaluating ordinary SDR images cannot keep good performance for measuring tone-mapped HDR quality. This can be summarized that these methods cannot effectively evaluate tone-mapped HDR perceptual quality. Meanwhile, the last five metrics designed for evaluation of tone-mapped images, i.e., BTMQI [11], HIGRADE [12], BLIQUE-TMI [13], Yue's method [20], and Wang's method [19], achieved better performance than the first six methods. Among them, BTMQI pays emphasis on evaluating the distortions of naturalness but ignores the measurement of color information. Similarly, HIGRADE and BLIQUE-TMI mainly measure the naturalness and structure but pay less on measuring color information. In contrast, Yue's method, Wang's method, and QmTm achieve better prediction effect than BTMQI which validates the advantage of incorporating the measurement of color distortion and halo artifacts. Fully analyzing the dependencies of OiQ on the quality of reproduction of the four perceptual attributes and halo artifacts, QmTm exhibits higher performance than others. Hence, QmTm is competent to address the quality measurement problem of tone-mapped HDR images.

TABLE IV
PERFORMANCE RESULTS WITH DIFFERENT ATTRIBUTE COMBINATIONS

$\mathbf{F}_{Br}$	$\mathbf{F}_{Cl}$	$\mathbf{F}_{Ct}$	$\mathbf{F}_{De}$	$\mathbf{F}_{Ar}$	PLCC	SRCC
$\surd$	$\times$	$\times$	$\times$	$\times$	0.5560	0.5284
$\times$	$\surd$	$\times$	$\times$	$\times$	0.5794	0.5548
$\times$	$\times$	$\surd$	$\times$	$\times$	0.5537	0.5228
$\times$	$\times$	$\times$	$\surd$	$\times$	0.5897	0.5778
$\times$	$\times$	$\times$	$\times$	$\surd$	0.6108	0.5970
$\surd$	$\surd$	$\times$	$\times$	$\times$	0.6071	0.5846
$\times$	$\times$	$\surd$	$\surd$	$\times$	0.7125	0.7036
$\surd$	$\surd$	$\surd$	$\times$	$\times$	0.6396	0.6244
$\surd$	$\surd$	$\times$	$\surd$	$\times$	0.7311	0.7201
$\surd$	$\surd$	$\times$	$\surd$	$\surd$	0.7458	0.7387
$\surd$	$\surd$	$\surd$	$\surd$	$\times$	0.7407	0.7335
$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	**0.7538**	**0.7492**

Although we have conducted OiQ measure using these criterion features, it is still unclear about the role of each attributes and artifacts. Therefore, it is necessary to conduct a feature combination analysis to investigate the relative contributions of attribute measurement to the overall performance. Table IV shows the performance results of each attribute criteria component via learning SVR model. Here, the symbol "$\surd$" ("$\times$") represents that this criteria feature is added (not added). In Table IV, $\mathbf{F}_{\mathrm{Br}}$ represents brightness criteria features, and $\mathbf{F}_{\mathrm{Cl}}$ represents reproduction of color. $\mathbf{F}_{\mathrm{Ct}}$ represents local contrast preservation, and $\mathbf{F}_{\mathrm{De}}$ and $\mathbf{F}_{\mathrm{Ar}}$ represent reproduction of detail and halo artifacts, respectively. First, $\mathbf{F}_{\mathrm{Cl}}$ and $\mathbf{F}_{\mathrm{Ar}}$

TABLE V
PERFORMANCE MEASUREMENTS ON DIFFERENT COLOR SPACES

Distortion Category	LAB Space			YUV Space			HSV Space		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
Tone Mapping	0.8050	0.7711	6.1217	0.7891	0.7538	6.3221	0.7855	0.7485	6.3642
Multi-Exposure Fusion	0.6947	0.6910	6.8434	0.6775	0.6686	6.9821	0.6651	0.6510	7.0858
Post Processing	0.6488	0.6033	6.9062	0.6386	0.6001	6.9747	0.6299	0.5943	7.0171
Overall	**0.7538**	**0.7492**	**6.6220**	0.7471	0.7464	6.7032	0.7396	0.7345	6.7183

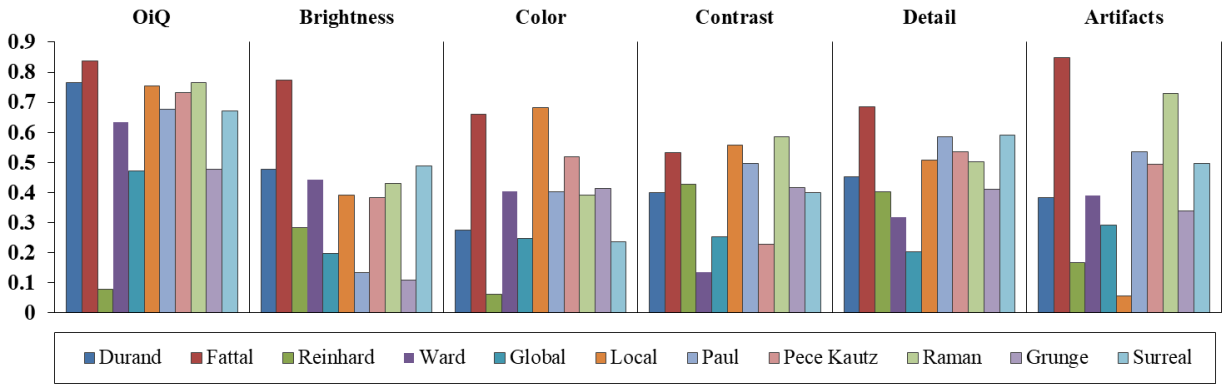


Fig. 10. Predicted SRCC values by 11 different distortion types. Left to right: OIQ, brightness, color, contrast, detail, and artifacts. In each chart, higher value represents higher prediction quality.

produce better performances than $\mathbf{F}_{\mathrm{Br}}$ and $\mathbf{F}_{\mathrm{De}}$. Second, $\mathbf{F}_{\mathrm{De}}$ which fully describes the detail distortion can obtain relatively good performance. Third, first four protocols perform more complementary to naturalness features. Best performance can be obtained when essential perceptual attributes and artifacts form a whole for regression.

2) Performance Measurement on Different Color Spaces: To measure the impact of choosing different color spaces on the performance of the proposed QmTm model, the above criterion features are also extracted from same tone-mapped image represented in YUV and HSV color spaces. The experimental results of QmTm using different color spaces are summarized in Table V. As can be seen, using LAB color space achieves a superior performance than other spaces. Interestingly, when testing on images represented in YUV and HSV color spaces, the experimental results are slightly lower than LAB space. Based on LAB color space, using only the L-component resulted in a higher degree of correlation with human subjective scores when compared with o only extracting features from the A or B chromatic channels. Overall, LAB color is most suitable for measuring the quality of tone-mapped images.

3) Performance Measurement on Individual Distortions: As stated above, the ESPL-LIVE database consists of three categories (i.e., tone-mapping, MEF, and PP) of SDR images, including 11 different distortion types. Next, we also explored the overall performance for each distortion types using the proposed criterion features. To test such tone-mapped HDR quality, we train our QmTm method on these different distortions. In Fig. 10, we show OIQ result and individual distortions results. These results testify how well QmTm performed in rating (without a reference HDR) experiments. Similar to OIQ result, TMO Fattal and TMO Raman exhibit better scores than other TMOs in all the attributes, while TMO Reinhard performs relatively poor. In the rating experiment, local energy weighting (abbreviated as Local) shows moderate performance but inferior in artifacts.

To facilitate the explanation, we use visualization results to suggest the effect of input scenes using different TMOs (e.g., Raman, Reinhard, and Ward). The linear correlations between subjective scores (i.e., MOS) and objective scores via QmTm are presented in in Fig. 11. To better visualize the results, we use different color points to indicate 11 different TMOs (three categories as mentioned above). It is clear that the prediction scores correlate well with MOS, especially tone mapping category. However, PP category shows poor performance. The scatter plots of correlation demonstrate the great monotonicity and accuracy of QmTm.

4) Computational Complexity Comparison: In the section, we discuss and relate computational complexity to other tone-mapped IQM methods. Our experiment platform is MATLAB, and the processor is a 2.9-GHz CPU with 4-GB memory. For each method, we record the total execution time consumed on a heap of tone-mapped images using tic and toc functions. After computing the average time (seconds per image), the run time of different methods is shown in Fig. 12. We find that the complexity of those methods varies greatly, and QmTm consumes a moderate running time among all the competing

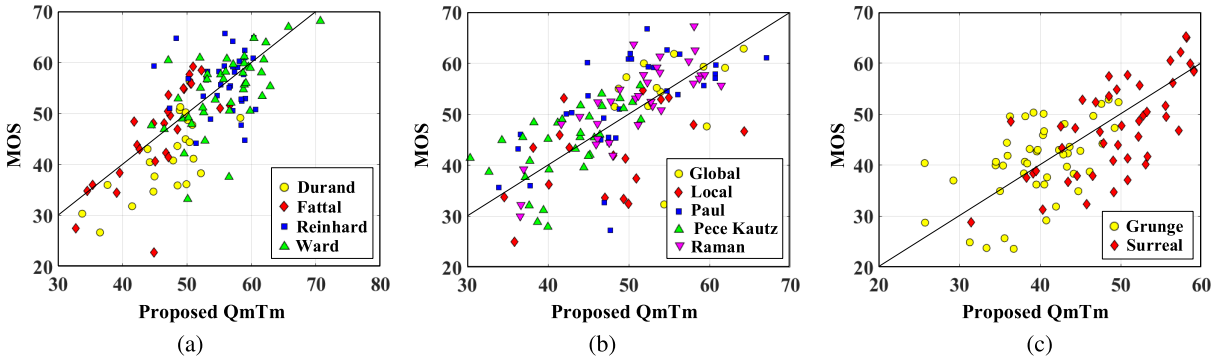


Fig. 11. Scatter plots of correlation between predict score by QmTm and MOS on different categories: (a) tone mapping, (b) MEF, and (c) PP.

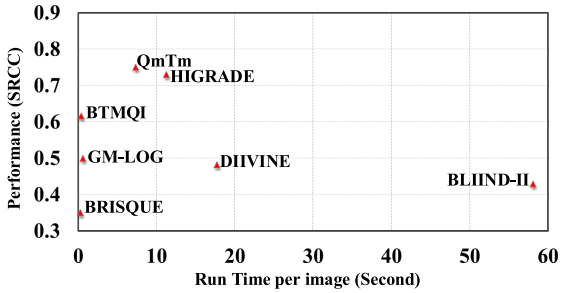


Fig. 12. Comparison of prediction performance under different test times.

methods. Considering the perfect prediction performance of QmTm in the tone-mapped IQM methods, QmTm has the applicability in real-time quality monitoring.

5) Performance Measurement in the Real SDR Display System: A strong demand for high-quality display is on the rise in various domains such as security and surveillance, automotive, medical, AR-VR, drones, and robotics imaging systems [24], [36]–[38]. For displaying a set of HDR images in a common SDR monitoring control system, ten representative TMOs are assembled into this display measurement program without specific hardware, including Mantiuk06, Mantiuk08, Fattal, Drago, Durand, and Reinhard. Next, we try to measure the performance of QmTm for optimal display output. Using these TMOs, 20 HDR images can generate 20 batches, 200 scene images in total [17]. Then, QmTm is tested on each batch whether automatically selecting optimal tone-mapped image with the best quality. Through overall measurement, the accuracy of image selection with optimal quality exceeds 80%. Meanwhile, the overall performance (PLCC = 0.7582, SRCC = 0.6889) shows that the QmTm model effectively filters high-quality tone-mapped images across different image contents and different TMOs. These results validate the capability of QmTm in meeting the requirement of real monitoring control system.

IV. Conclusion

This article aims toward a quantitative evaluation to measure the tone-mapped HDR display quality and transform that into a single number pertaining to OiQ. We characterize a perceptual model based on user-driven attribute computing and evaluate the prediction capabilities of various machine learning methods (i.e., random forest, SVR, and extreme learning machine). The experimental results show that the OiQ is determined by a composition of these attributes, but rather a single one. The precisely defined attributes are helpful for getting into the tone mapping field or when implementing a tone mapping method. Furthermore, using machine learning techniques for consolidation results in better prediction keeping fit with the subjective score when compared with linear regression. Future work includes conducting more experiments to determine the significance of these results, including more test images, subjects, and displays. We also explore other perceptual models to better the individual dimensions.

References

[1] F. Drago, K. Myszkowski, T. Annen, N. Chiba, P. Brunet, and D. Fellner, "Adaptive logarithmic mapping for displaying high contrast scenes," in *Computer Graphics Forum*. Hoboken, NJ, USA: Blackwell Publishing, Sep. 2003, pp. 419–426.

[2] M. H. Conde, K. Hartmann, and O. Loffeld, "Adaptive high dynamic range for time-of-flight cameras," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 7, pp. 1885–1906, Jul. 2015.

[3] Y. Que, Y. Yang, and H. J. Lee, "Exposure measurement and fusion via adaptive multiscale edge-preserving smoothing," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 12, pp. 4663–4674, Dec. 2019.

[4] A. R. Várkonyi-Kóczy and A. Rövid, "High-dynamic-range image reproduction methods," *IEEE Trans. Instrum. Meas.*, vol. 56, no. 4, pp. 1465–1472, Aug. 2007.

[5] D. M. Eagleman, "Visual illusions and neurobiology," *Nature Rev. Neurosci.*, vol. 2, no. 12, pp. 920–926, Dec. 2001.

[6] Q. Jiang *et al.*, "Blind image quality measurement by exploiting high-order statistics with deep dictionary encoding network," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7398–7410, Oct. 2020.

[7] G. Yue, C. Hou, T. Zhou, and X. Zhang, "Effective and efficient blind quality evaluator for contrast distorted images," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 8, pp. 2733–2741, Aug. 2019.

[8] L. Angrisani, D. Capriglione, L. Ferrigno, and G. Miele, "An internet protocol packet delay variation estimator for reliable quality assessment of video-streaming services," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 5, pp. 914–923, May 2013.

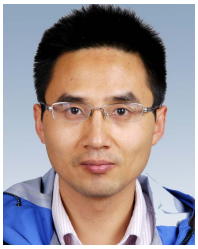
[9] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Burlington, MA, USA: Morgan Kaufmann, 2010.

[10] J. Xu, W. Zhou, Z. Chen, S. Ling, and P. L. Callet, "Binocular rivalry oriented predictive autoencoding network for blind stereoscopic image quality measurement," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[11] K. Gu *et al.*, "Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 432–443, Mar. 2016.
[12] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, Jun. 2017.
[13] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "BLIQUE-TMI: Blind quality evaluator for tone-mapped images based on local and global feature analyses," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 323–335, Feb. 2019.
[14] L. Cao, G. Jiang, Z. Jiang, M. Yu, Y. Qi, and Y.-S. Ho, "Quality measurement for high dynamic range omnidirectional image systems," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.
[15] G. Yue, C. Hou, K. Gu, S. Mao, and W. Zhang, "Biologically inspired blind quality assessment of tone-mapped images," *IEEE Trans. Ind. Electron.*, vol. 65, no. 3, pp. 2525–2536, Mar. 2018.
[16] P. Chen, L. Li, X. Zhang, S. Wang, and A. Tan, "Blind quality index for tone-mapped images based on luminance partition," *Pattern Recognit.*, vol. 89, pp. 108–118, May 2019.
[17] M. Jiang, L. Shen, L. Zheng, M. Zhao, and X. Jiang, "Tone-mapped image quality assessment for electronics displays by combining luminance partition and colorfulness index," *IEEE Trans. Consum. Electron.*, vol. 66, no. 2, pp. 153–162, May 2020.
[18] M. Jiang, L. Shen, M. Hu, P. An, and F. Ren, "Blind quality evaluator of tone-mapped HDR and multi-exposure fused images for electronic display," *IEEE Trans. Consum. Electron.*, vol. 67, no. 4, pp. 350–362, Nov. 2021.
[19] X. Wang, Q. Jiang, F. Shao, K. Gu, G. Zhai, and X. Yang, "Exploiting local degradation characteristics and global statistical properties for blind quality assessment of tone-mapped HDR images," *IEEE Trans. Multimedia*, vol. 23, pp. 692–705, 2021.
[20] G. Yue, W. Yan, and T. Zhou, "Referenceless quality evaluation of tone-mapped HDR and multiexposure fused images," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1764–1775, Mar. 2020.
[21] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Comput. Graph.*, vol. 32, no. 3, pp. 330–349, Jun. 2008.
[22] M. D. Fairchild, *Color Appearance Models*, 2nd ed. Hoboken, NJ, USA: Wiley, 2005.
[23] Y. Gan and Q. Zhao, "An effective defect inspection method for LCD using active contour model," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 9, pp. 2438–2445, Sep. 2013.
[24] G. Nam, H. Lee, S. Oh, and M. H. Kim, "Measuring color defects in flat panel displays using HDR imaging and appearance modeling," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 2, pp. 297–304, Feb. 2016.
[25] K. Gu, Z. Xia, and J. Qiao, "Stacked selective ensemble for $PM_{2.5}$ forecast," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 660–671, Mar. 2020.
[26] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognit.*, vol. 42, no. 3, pp. 425–436, Mar. 2009.
[27] Y. Li, L. Sharan, and E. H. Adelson, "Compressing and companding high dynamic range images with subband architectures," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 836–844, Jul. 2005.
[28] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-scale crowdsourced study for tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4725–4740, Oct. 2017.
[29] Video Quality Experts Group (VQEG). (2003). *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*. [Online]. Available: http://www.vqeg.org/2003
[30] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.
[31] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.
[32] E. V. Tolstaya, M. N. Rychagov, S. H. Kim, and D. C. Choi, "Removal of blocking and ringing artifacts in JPEG-coded images," *Proc. SPIE*, vol. 7537, pp. 1–12, Jan. 2010.
[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blin' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[34] Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Signal Process., Image Commun.*, vol. 29, no. 7, pp. 725–747, Aug. 2014.
[35] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
[36] A. R. Várkonyi-Kóczy, A. Rövid, and T. Hashimoto, "Gradient-based synthesized multiple exposure time color HDR image," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 8, pp. 1779–1785, Aug. 2008.
[37] X. Dong, L. Shen, M. Yu, and H. Yang, "Fast intra mode decision algorithm for versatile video coding," *IEEE Trans. Multimedia*, vol. 24, pp. 400–414, 2021.
[38] H. Yang, L. Shen, X. Dong, Q. Ding, P. An, and G. Jiang, "Low-complexity CTU partition structure decision and fast intra mode decision for versatile video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1668–1682, Jun. 2020.

Mingxing Jiang received the M.S. degree from the Hefei University of Technology, Hefei, China, in 2013. He is currently pursuing the Ph.D. degree in communication and information systems with Shanghai University, Shanghai, China.

He is also a Visiting Student with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology. His research interests include multimedia quality assessment, affective computing, and machine learning.



Liquan Shen (Member, IEEE) received the B.S. degree in automation control from Henan Polytechnic University, Henan, China, in 2001, and the M.E. and Ph.D. degrees in information and communication engineering from Shanghai University, Shanghai, China, in 2005 and 2008, respectively.

He is currently a Professor with the School of Communication and Information Engineering, Shanghai University. He has authored or coauthored more than 100 refereed technical papers in international journals and conferences in the field of video coding and image processing. His research interests include high-efficiency video coding (HEVC), perceptual coding, video codec optimization, 3DTV, and video quality assessment.

Min Hu received the M.S. degree in industrial automation and the Ph.D. degree in computer science from the Hefei University of Technology, Hefei, China, in 1994 and 2004, respectively, where she is currently a Professor with the School of Computer and Information.

Her research interests include digital image processing, artificial intelligence, and data mining.

Ping An (Member, IEEE) received the B.E. and M.E. degrees from the Hefei University of Technology, Hefei, China, in 1990 and 1993, respectively, and the Ph.D. degree from Shanghai University, Shanghai, China, in 2002.

In 1993, she joined Shanghai University. From 2011 to 2012, she was a Visiting Professor with the Communication Systems Group, Technical University of Berlin, Berlin, Germany. She is currently a Professor with the Video Processing Group, School of Communication and Information Engineering, Shanghai University. She has finished more than 16 projects supported by the National Natural Science Foundation of China, the National Science and Technology Ministry, and the Science and Technology Commission of Shanghai Municipality. Her research interests include image and video processing, with a focus on immersive image processing and computer vision.

Yu Gu (Senior Member, IEEE) received the B.E. degree from the Special Classes for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2004, and the D.E. degree from the University of Science and Technology of China in 2010.

In 2006, he was an Intern with Microsoft Research Asia, Beijing, China, for seven months. From 2007 to 2008, he was a Visiting Scholar with the University of Tsukuba, Tsukuba, Japan. From 2010 to 2012, he was a JSPS Research Fellow with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor with the School of Computer and Information, Hefei University of Technology, Hefei. His current research interests include pervasive computing and affective computing.

Dr. Gu is also a member of ACM. He was a recipient of the IEEE Scalcom2009 Excellent Paper Award and the NLP-KE2017 Best Paper Award.

Fuji Ren (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Engineering, Hokkaido University, Sapporo, Japan, in 1991.

He is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology, Chengdu, China. He is also a Professor with the Department of Information Science and Intelligent Systems, University of Tokushima, Tokushima, Japan. His current research interests include natural language processing, machine translation, artificial intelligence, language understanding and communication, robust methods for dialog understanding, and affective information processing and knowledge engineering.