Full Reference Image Quality Assessment Based on Visual Salience With Color Appearance and Gradient Similarity

CHENYANG SHI1 AND YANDAN LIN1,2
1Institute for Electric Light Sources, Fudan University, Shanghai 200433, China
2Institute of Future Lighting, Academy for Engineering and Technology, Fudan University, Shanghai 200433, China

Corresponding author: Yandan Lin (ydlin@fudan.edu.cn)

This work was supported by the National Key Research and Development Program of China under Project 2017YFB0403700.

ABSTRACT  Image quality assessment (IQA) model is designed to measure the image quality in consistent with subjective ratings by computational models. In this research, a reliable full reference color IQA model is proposed by combining the Visual saliency with Color appearance (VC) similarity, gradient similarity and chrominance similarity. Two new color appearance indices, vividness and depth, are selected to build the visual saliency similarity map. The structure and chrominance features are characterized by different channels of chosen color space. VC map plays two roles in the proposed model. One is utilized as feature to compute the local quality of distorted image, and the other is as a weight part to reflect the importance of local domain. The novel model is called visual saliency with color appearance and gradient similarity (VCGS). To quantify the specific parameters of VCGS, some experiments are conducted based on the statistical correlation indices. Massive experiments are performed on the publicly available benchmark single and multiple distortion databases, and the commonly evaluation criteria results prove that VCGS works with higher consistency with the subjective evaluations than the other state-of-the-art IQA models for the prediction accuracy. Besides, VCGS maintain a moderate computational complexity. The MATLAB source code of VCGS is publicly available online at https://github.com/AlAlien/VCGS.

INDEX TERMS  Full reference, image quality assessment, visual saliency with color appearance, gradient.

I. INTRODUCTION

With the rapid development of color image contents and imaging devices in various kinds of multimedia communication systems, conventional grayscale counterparts are replaced by chromatic ones. Under such a transition, perceptual image quality assessment (IQA) has played a significant role in numerous visual data processing application [1]. IQA methods can generally be divided into two categories: human-based subjective assessment and algorithm-based objective assessment [2]. In the past few decades, multiple objective IQA models have been developed to evaluate image quality [3]. There are three main well-established frameworks according to the availability of a reference image: full reference (FR) [4]–[13], reduced reference (RR), and no reference (NR) [14] or blind IQA. FR-IQAs quantify the visual quality of a distorted image with respect to its reference image, which is the main scope of this article.

Human visual system (HVS) is an ideal receiver of visual information and it is also the most reliable way to evaluate image quality with subjective judgement [15]. Whereas, subjective evaluation is infeasible under many conditions, i.e., psycho-visual experiments under standard protocols is laborious; subjective tests cannot be conducted for an automated system. Thus, the subjective opinion of human observers, for instance Mean Opinion Score (MOS) and Difference Mean Opinion Score (DMOS), can be predicted by an objective model [16].

For IQAs, mean squared error (MSE) and its variations, such as peak-signal-to-noise ratio (PSNR), are the conventional models and widely used due to their simplicity [17]. Their accuracy, however, is not as good as their efficiency. Hence, numerous IQA models have been proposed with better performance based on HVS [3]. Generally, these better models characterize the structural information, luminance
information, contrast information and color information in the spatial and frequency domains. The representative model for evaluating structural information is Structural SIMilarity (SSIM) index, which yields accurate performance on publicly available databases [4]–[6]. Geometric features are suitable for grayscale domain with relative insignificant color information and have high computational efficiency. But they cannot deal with color image with chromatic deviations, because they overlook the impact of chromatic information in visual quality.

Recently, some learning based models were also proposed. Chang et al. [18] introduced an Independent Feature Similarity (IFS) index by combining feature and luminance components. The feature detector was training by FastICA (Fast Independent Component Analysis) algorithm [19]. Besides, a Local Linear Model (LLM) was introduced to extract local information of images and a Convolutional Neural Network (CNN) was used for automatic distortion-specific compensation [20]. Learning based models can obtain higher prediction accuracy and it has been another direction for FR-IQA model designing.

In this research, a novel similarity-based FR-IQA model is introduced without learning procedure, which combines three feature information processing parts, i.e., visual saliency, structure and chrominance features. Two new color appearance indices, i.e., vividness and depth, are operated by Log-Gabor filter for dealing with visual saliency feature, and gradient similarity map characterize structure feature. The chrominance feature is extracted by the chrominance channels. The visual saliency part is also as a weight part to reflect the importance of local domain. This model has a moderate complexity and offers better quality predictions, in comparison to the other state-of-the-art models.

II. RELATED WORK

The most famous IQA models have the same design logic, named top-down strategy [15]. Firstly, similarity maps are calculated based on evaluation indices; then a pooling strategy is chosen; lastly, the values of those similarity maps are converted to a single quality score. Wang et al. [5] proposed an SSIM index, which can be thought as the milestone of IQA research. Some extensions, such as the Multi-Scale SSIM (MS-SSIM) [4] and information content weighted SSIM (IW-SSIM) [6], have been introduced to improve the accuracy.

Zhang et al. [11] combined the Phase Congruency (PC) and Gradient Magnitude (GM) to calculate the similarity maps, and proposed a Feature SIMilarity index (FSIM). PC was used again as a weighting map to evaluate the perceptually important of local domain in HVS, and the pooling quality score was derived. Additionally, gradient was chosen as the main similarity operator by Liu et al. [12], and a Gradient Similarity Metric (GSM) was proposed. In [21], gradient magnitude similarity was also an component in RVSIM (Riesz transform and Visual contrast sensitivity-based feature SIMilarity index) IQA model, combining with the monogenic components by Log-Gabor filter, Riesz transform and contrast sensitivity processing.

Visual saliency of an image can reflect how “salient” a local region is to HVS. Because the visual saliency and IQA are all related to how HVS perceives an image and visual attention attractor of suprathreshold distortions. The relationship between visual saliency and image quality has been integrated in IQA model, and better prediction results could be achieved. In [22], a Visual Saliency-based Index (VSI) was introduced and the visual saliency was computed by SDSP (Saliency Detection by combing Simple Priors). In addition, to compensate for the lack of contrast sensitivity, gradient modulus was chosen as an additional feature. In SDSP [23], frequency prior, color prior and location prior were integrated as a saliency algorithm. Recently, a perceptual IQA model was developed utilizing Contrast and Visual Saliency Similarity (CVSS) [13] maps. In CVSS, the spectral residual method [24] was operated as saliency map generator. This model had a better performance on grayscale images than the previous model did. But for the real HVS, visual detector receives colorful information. Therefore, objective IQAs should take chromatic information into consideration and put higher weights to chromatic information. In SDSP, the frequency prior was extracted by Log-Gabor filter on CIELAB color space three independent channels. Similar to SDSP, Achanta et al. adopted DoG (Difference of Gaussian) filter to mimic the salient region detection mechanism [25]. Therefore, it can be concluded that band-pass filter can be an effective way to extract visual saliency feature.

Normally, color image quality can be evaluated by using a grayscale model on individual RGB channels and connecting channel scores afterwards. Since such models are suboptimal, more elaborate models exploiting properties of color perception need to be proposed. In addition, color image quality can be evaluated using a similar approach as colorimetry. For example, uniform color spaces (e.g., CIELAB and CIELUV) have been developed to characterize the perceived color differences between pairs of colors [26], [27]. CIELAB was more uniform than CIELUV globally [27]. The cylindrical polar coordinates in CIELAB correlates to the color attributes of a stimulus. Therefore, the perceived distortion of color images can be better characterized on pixel levels using CIE formulations. Although these formulations are useful to compare uniformly color patches, they need to be modified to be consistency with complex real-world image data. With the better and better understanding on human perception to color appearance of stimuli under different viewing conditions, some novel indices were proposed and more accurate than the former ones by utilizing similarity-based formulation. Lee and Plataniotis [28], [29] proposed FR-IQA models based on normal three channels of color appearance, i.e., lightness, hue and chroma, which yields better performance for evaluation color images, namely Directional Statistics based Color Similarity Index (DSCSI). To obtain a higher consistency with the subjective evaluations, color appearance indices can be taken into account to assess chromatic information.
Because the color appearance indices can represent chromatic deviation for color image in HVS perception way.

### III. PROPOSED COLOR IMAGE QUALITY MODEL

In this section, a full reference model—the Visual saliency with Color appearance and Gradient Similarity index (VCGS)—to quantify the perceptual visual quality of color images is introduced. The proposed model is general purpose, as a matter of fact that it consistently performs well over commonly encountered chromatic and achromatic distortions. The inputs to the model are two RGB images with an identical spatial resolution, X and Y, called the original and the distorted images, respectively. They are hypothesized to be matching in bit depth and properly aligned. The output quality score, denoted as S(X, Y), has a range between 0 and 1, with 1 for the best quality when the two images are completely identical.

Three feature similarity maps are included in the proposed model. Visual saliency feature is extracted by the color appearance in CIELAB color space, which is more compatible with human intuition [30]. The chrominance similarity maps can express the color difference between two images in pixel level by the chrominance channels. For structural feature, gradient is used to derive another similarity map due to its superiority [11], [12]. Lastly, these three similarity maps are connected and pooled based on [2], [11].

#### A. VISUAL SALIENCY WITH COLOR APPEARANCE SIMILARITY MAP

Inspired by [23] and [25], in this research, band-pass filter is used for saliency detection. Because of the better performance of Log-Gabor filter, it is also chosen as an operator to obtain visual saliency. In [28], color appearance indices had been directly utilized in IQA model. To better characterize the color appearance of the images perceived by the human visual system, the original RGB images are transformed to a color space that is more compatible with human’s intuition. Among various perceptual color models, CIELAB is utilized in this research. The strong link between CIELAB and the Munsell system and the separation of CIELAB into lightness, L\(^*\), and chromaticness, a\(^*\), b\(^*\) or \(C_{ab}^\ast\), \(H_{ab}^\ast\), has limited its utility for applications where describing colors that covary in lightness and chromaticness is advantageous, for example, art, design, comparing colorant strength, and color image evaluation. Because of this, two new CIELAB variables, vividness and depth [30], were introduced to extend the utility of CIELAB as color appearance variable, shown in Fig. 1. After RGB to CIELAB transformation, each pixel of X contains three color components: lightness \(L^*\), red-green \(a^*\) and blue-yellow \(b^*\). And vividness \(V_{ab}^*\) and depth \(D_{ab}^*\) are calculated via Eq. (1) and (2) [30].

\[
\begin{align*}
V_{ab}^* &= \sqrt{(L^*)^2 + (a^*)^2 + (b^*)^2} = \sqrt{(L^*)^2 + (C_{ab}^*)^2} \quad (1) \\
D_{ab}^* &= \sqrt{(100 - L^*)^2 + (a^*)^2 + (b^*)^2} \\
&= \sqrt{(100 - L^*)^2 + (C_{ab}^*)^2} \quad (2)
\end{align*}
\]

Depend on the Eq. (1) and (2), the local color appearance for the pristine image X and distorted one Y are represented by \(V_1\) and \(D_1\), \(V_2\) and \(D_2\), respectively. To quantify the visual saliency, Log-Gabor filter is used to operate on vividness and depth. The transfer function of a Log-Gabor filter \(g(x)\) \((x = (x, y) \in \mathbb{R}^2)\) in the frequency domain can be expressed as:

\[
G(u) = \exp\left(-\left(\log\left(\frac{\|u\|_2}{\omega_0}\right)^2 / 2\sigma_F^2\right)\right) \quad (3)
\]

where \(u = (u, v) \in \mathbb{R}^2\) is the coordinate in the frequency domain, \(\omega_0\) is the center frequency of filter, and \(\sigma_F\) is the bandwidth parameter of filter. \(g(x)\) cannot be analytically expressed due to the singularity in the log function at the origin. Instead, \(g(x)\) can only be approximately obtained by performing a numerical inverse Fourier transform to \(G(u)\). The Visual saliency with Color appearance (VC) of pristine image X and distorted one Y can be calculated by:

\[
VC = \left(\left(V \ast g\right)^2 + \left(D \ast g\right)^2\right)^{1/2} \quad (4)
\]

where \(\ast\) donates the convolution operation. Then the visual salience with color appearance similarity for the two images are compared as:

\[
S_{VC} = \frac{2V_1V_2 + K_{VC}}{V_1^2 + V_2^2 + K_{VC}} \quad (5)
\]

where parameter \(K_{VC}\) is a constant to control numerical stability [5]. In the following section, we set \(\omega_0 = 0.021, \sigma_F = 1.34\) based on [2] and [23].

To illustrate the validity of VC maps, an example is utilized in Fig. 2 from TID2008 database [31]. In Fig. 2, (a1-e1) in the first row represent the reference image R and Gaussian blur image with 4 distortion levels, (a2-e2) in the second row represent the VC maps of the first row images after normalization. It can be clearly seen that the higher distorted level is, the lower quality of VC map is. This is mainly due to the reason that the high-frequency signal of VC map is dependent on the quality of input image. Based on this reason, VC map can be utilized to design the proposed model.
B. GRADIENT AND CHROMINANCE SIMILARITY MAP

There are several different operators to compute the image gradient, such as the Prewitt operator [32], the Sobel operator [32], the Roberts operator [33] and the Scharr operator [33]. It is very common that gradient magnitude in the discrete domain is calculated on the basis of some operators which approximate derivatives of the image function using differences. These operators approximate vertical \( G_y \) and horizontal \( G_x \) gradients of an image \( X \) using convolution:

\[
G_x = h_x * X \quad \text{and} \quad G_y = h_y * X \quad \text{(see Eq. (5) and (7))},
\]

where \( h_x \) and \( h_y \) are horizontal and vertical gradient operators and \(*\) denotes the convolution. The first derivative magnitude is defined as:

\[
G(x) = \sqrt{G_x^2 + G_y^2}.
\]

Within the proposed IQA model, these operators perform almost the same.

\[
G_x = \frac{1}{16} \begin{bmatrix} 3 & 0 & -3 \\ 10 & 0 & -10 \\ 3 & 0 & -3 \end{bmatrix} * X \quad \text{(6)}
\]

\[
G_y = \frac{1}{16} \begin{bmatrix} 3 & 10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & -3 \end{bmatrix} * X \quad \text{(7)}
\]

In this paper, Scharr operator is used to compute gradient magnitudes of lightness \( L^* \) channels in CIELAB of reference and distorted images, \( G_1 \) and \( G_2 \). From which, gradient similarity \( (S_G) \) is computed by the following SSIM induced equation:

\[
S_G = \frac{2G_1 \cdot G_2 + K_G}{G_1^2 + G_2^2 + K_G}
\]

where parameter \( K_G \) is a constant to control numerical stability [5]. The gradient similarity \( (S_G) \) is widely used in the literature [2], [11], [12], [34]–[36] and its usefulness to measure image distortions was extensively investigated in [34]. As shown in Fig. 2, the gradient map of the first row is in the third row. We have the observation that, when the image becomes more distorted, the gradient map is of lower structural feature. Therefore, the gradient map is a useful judgment for the structural distortions of human visual system (HVS).

The similarity of the chrominance components in CIELAB color space, i.e., red-green \( a^* \) and blue-yellow \( b^* \) channels, can be simply defined as:

\[
S_C = \frac{2a_1 \cdot a_2 + K_C}{a_1^2 + a_2^2 + K_C} \cdot \frac{2b_1 \cdot b_2 + K_C}{b_1^2 + b_2^2 + K_C}
\]

where parameter \( K_C \) is a constant to control numerical stability [5].

C. VCGS INDEX

With the extracted visual salience with color appearance similarity, gradient similarity and chrominance similarity maps, a novel model is defined in IQA task, named Visual saliency with Color appearance and Gradient Similarity index (VCGS), shown as follow:

\[
\text{VCGS} = \frac{\sum \Omega S_V \cdot (S_G)^\alpha \cdot (S_C)^\lambda \cdot VC_m}{\sum \Omega VC_m}
\]

where \( \Omega \) means the spatial domain, and \( VC_m \) \( (VC_m = \max (VC_1, VC_2)) \) is used to weight the importance of the two maps in the overall similarity. \( \alpha \) and \( \lambda \) represent the relative importance among visual saliency, structure and chrominance features.

It has been widely accepted that different locations can have different contributions to the human vision perception.
of the image quality, so it would be better if the fixation point of the visual system is considered in deriving the score. Since human visual cortex is sensitive to lightness and chromatic, the visual saliency with color appearance at a location can reflect how likely it is a perceptibly significant point. Consequently, in our framework, it is natural to choose the visual saliency with color appearance map (VC\(m\)) to characterize the visual importance of a local region. The procedures to compute VCGS are illustrated by an example in Fig. 3. In Fig. 2, the quality scores of VCGS for different distortion level images are given. It can be concluded that the quality scores are consistent with the distorted image.

In this paper, \(K_{VC}\), \(K_G\) and \(K_C\) are all fixed so that the proposed model can be conveniently applied in all databases. Besides, \(\alpha\) and \(\lambda\) need also to be determined for all databases. In the previous researches, trial-and-error method is a popular way to deal with this kind of problem. In the next section, these parameters will be determined by trial-and-error method.

### IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

#### A. ASSESSMENT CRITERIA AND DATABASES

In our research, four publicly available databases are used for model validation and comparison, i.e., TID2013 [37], TID2008 [31], CSIQ [38], and LIVE [39]. The information of each databases is introduced in Table 1. These four databases are the most commonly utilized collections in IQA research covering a wide range of ordinarily encountered distortions in real world application. They are annotated with subjective ratings, i.e., MOS or DMOS, facilitating reasonable benchmarking of the proposed model with others. Distorted images in the aforementioned databases are processed from a set of source images that reflects adequate multiplicity in color complexity and edge/texture details, including pictures of humans, natural scenes and man-made objects.

In order to evaluate whether a model is able to predict the perception of human observers, the comparisons are made between the calculated scores using the proposed model and the values rated by the observers. Four commonly evaluation criteria for IQA model are employed: Spearman rank-order correlation coefficient (SROCC), Pearson linear correlation coefficient (PLCC), Kendall rank-order correlation coefficient (KROCC) and root mean squared error (RMSE) [3], [40].

The SROCC measures the prediction monotonicity, i.e., the degree to which model agrees with the rank of the subjective ratings, and is defined as follow:

\[
SROCC = 1 - \frac{6 \sum_{i=1}^{n} [R(p_i) - R(s_i)]^2}{n(n^2 - 1)}
\]

where \(R(p_i)\) and \(R(s_i)\) represent ranks of the prediction score \(p_i\) and the subjective score \(s_i\), respectively, and \(n\) is the number of score sets. The KROCC is also a correlation coefficient to measure the prediction monotonicity, and

| Database | Source Images | Distorted Images | Distortion Types | Observers |
|----------|---------------|-----------------|-----------------|-----------|
| TID2013 | 25            | 3000            | 24              | 971       |
| TID2008 | 25            | 1700            | 17              | 838       |
| CSIQ    | 30            | 866             | 6               | 35        |
| LIVE    | 29            | 779             | 5               | 161       |
given by:

$$KROCC = \frac{n_c - n_d}{n(n-1)/2}$$  \(12\)

where \(n_c\) and \(n_d\) express the numbers of concordant and discordant pairs in scores, respectively. Both SROCC and KROCC take into account the rank of the score and neglect the relative distance between scores, which are different from following PLCC. These all three coefficients range from −1 to 1, and the absolute value of each is close to unity 1 means that the fidelity of an objective model is considered high.

To compute the PLCC and RMSE indices, a logistic regression is adopted to get the same scale values with subjective judgments by using:

$$p(x) = \beta_0 \left[ \frac{1}{2} - \frac{1}{1 + \exp(\beta_2 (x - \beta_3))} \right] + \beta_4 x + \beta_5$$  \(13\)

where \(\beta_1, \ldots, \beta_5\) are the parameter to be fitted, \(x\) represents the original IQA scores, and \(p(x)\) is the IQA score after the regression [39]. The PLCC is a correlation coefficient used to measure the prediction accuracy of a model, i.e., the ability of predicting the subjective ratings with low error. For \(n\) pairs of model and subjective scores \((p_i, s_i)\), PLCC is calculated by:

$$PLCC = \frac{\sum_{i=1}^{n} (p_i - \bar{p})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n} (p_i - \bar{p})^2 \sum_{i=1}^{n} (s_i - \bar{s})^2}}$$  \(14\)

where \(\bar{p}\) and \(\bar{s}\) denote the mean value of model prediction scores and subjective scores, respectively. For the RMSE (see Eq. (15)), a smaller value indicates better performance.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (s_i - p_i)^2}$$  \(15\)

### B. PARAMETERS OPTIMIZATION FOR VCGS

In this work, five parameters are involved, including \(K_{VC}\), \(K_G\), \(K_C\), \(\alpha\) and \(\lambda\). Among them, the evaluation criteria for IQA model is not sensitive to \(K_{VC}\) changes and it can be set as \(K_{VC} = 1.25\) based on [2]. The other parameters are investigated in the following part. It should be note that when the current parameter is under studied, the others remain unchanged. Because SROCC, KROCC, PLCC and RMSE have the similar performance with parameters changing, SROCC is selected as the criteria to optimizing the parameters.

Parameters \(K_G\) and \(K_C\) serve as numerical stability controller for \(S_G\) and \(S_C\). In Fig. 4 (a), SROCC curves against \(K_G\) on the four databases are presented. It can be found that for all databases, the performance is stable high, when \(K_G\) is in the interval [50, 70]. In our experiments, we set \(K_G = 60\). In Fig. 4 (b), SROCC curves against \(K_C\) on the four databases are presented. It can be observed that for all databases, when \(K_C\) is in the interval [150, 200], the performance can be stable high. In our experiments, we set \(K_C = 200\).

The last two parameters, namely, \(\alpha\) and \(\lambda\), against the weights of gradient and chrominance similarity measurement. Similar to the parameters mentioned above, their influences on the performance of VCGS are further discussed. The results are expressed in Fig. 5. To make the changes visible and distinguishable in the figures, SROCC is transformed to NSROCC \((NROCC = 10 \times (SROCC - \text{min}(SROCC)))\) [41]. It can be observed that the optimal \(\alpha\) and \(\lambda\) of TID2013, TID2008, CSIQ and LIVE are in intervals [0.3, 0.4] \(\times\) [0.015, 0.03], [0.35, 0.5] \(\times\) [0.01, 0.025], [0.3, 0.4] \(\times\) [0.01, 0.025], [0.35, 0.5] \(\times\) [0.01, 0.025]. In all databases, the optimal \(\alpha\) and \(\lambda\) are in similar intervals, which is consistent with that gradient and chrominance changes are visually perception with certain weights in image quality evaluation. In this research, \(\alpha\) and \(\lambda\) are fixed as 0.4 and 0.02, respectively. In the future work, the weights of gradient and chrominance for psychovisual experiments need to be discussed. In the next section, the best setting of \(\alpha\) and \(\lambda\) for each database will be displayed.

### C. PERFORMANCE COMPARISON AMONG DIFFERENT MODELS

An ideal IQA model should yield good performance and predict consistently well under different types of distortion. In this section, the proposed model was compared with other state-of-the-art models, including SSIM [5], IW-SSIM [6], FSIM [11], GSM [12] and VSI [2], and the latest DCS [28], IFS [18], LLM [20], RVSIM [21] and CVSS [13], published in 2015, 2015, 2016, 2018 and 2018, respectively. The best three results are highlighted in bold using the four indices in Table 2. Besides, according to Wang and Li [6], the weighted and direct average values of SROCC, PLCC and KROCC results of the four databases were also presented in Table 2 to evaluate the overall performance. The weight of each database is linearly based on the number of the distortion images the database contained.

From the Table 2, it can be observed that the proposed model performs consistently well for all the databases. Particularly, the proposed model is always among the top three models for the TID2013 and LIVE databases. For TID2008 and CSIQ databases, the proposed model performs...
TABLE 2. Performance comparison of IQA models on four databases.

| Database Criteria | SSIM | IW-SSIM | FSIMc | VSI | GSM | DSCSI | IF | LLM | RVSIM | CVSS | Proposed |
|-------------------|------|---------|-------|-----|-----|-------|----|-----|-------|------|----------|
| TID2013           |      |         |       |     |     |       |    |     |       |      |          |
| SROCC             | 0.7417 | 0.7779 | 0.8510 | 0.8965 | 0.7946 | 0.8744 | 0.8697 | 0.9037 | 0.6757 | 0.8069 | 0.8926 |
| PLCC              | 0.7895 | 0.8319 | 0.8769 | 0.8999 | 0.8464 | 0.8782 | 0.8791 | 0.9068 | 0.7825 | 0.8406 | 0.9000 |
| RMSE              | 0.7608 | 0.6880 | 0.5959 | 0.5404 | 0.6603 | 0.5930 | 0.5909 | 0.5277 | 0.7719 | 0.6715 | 0.5404 |
| KROCC             | 0.5588 | 0.5977 | 0.6665 | 0.7183 | 0.6255 | 0.6875 | 0.5086 | 0.7209 | 0.5146 | 0.6331 | 0.7166 |
| TID2008           |      |         |       |     |     |       |    |     |       |      |          |
| SROCC             | 0.7749 | 0.8559 | 0.8840 | 0.8979 | 0.8504 | 0.8634 | 0.8903 | 0.9077 | 0.7375 | 0.9001 | 0.8975 |
| PLCC              | 0.7732 | 0.8579 | 0.8762 | 0.8762 | 0.8422 | 0.8445 | 0.8810 | 0.8971 | 0.7954 | 0.8961 | 0.8776 |
| RMSE              | 0.8511 | 0.6895 | 0.6468 | 0.6466 | 0.7235 | 0.7187 | 0.6349 | 0.5982 | 0.8133 | 0.5956 | 0.6433 |
| KROCC             | 0.5768 | 0.6636 | 0.6991 | 0.7123 | 0.6596 | 0.6651 | 0.7009 | 0.7368 | 0.5628 | 0.7215 | 0.7171 |
| SROCC             | 0.8756 | 0.9213 | 0.9310 | 0.9423 | 0.9108 | 0.9417 | 0.9581 | 0.9605 | 0.8979 | 0.9580 | 0.9443 |
| PLCC              | 0.8613 | 0.9144 | 0.9192 | 0.9279 | 0.8964 | 0.9313 | 0.9576 | 0.9000 | 0.9236 | 0.9589 | 0.9301 |
| RMSE              | 0.1334 | 0.1063 | 0.1034 | 0.0979 | 0.1164 | 0.0956 | 0.0757 | 0.1232 | 0.1007 | 0.0745 | 0.0964 |
| KROCC             | 0.6907 | 0.7529 | 0.7690 | 0.7857 | 0.7374 | 0.7787 | 0.8158 | 0.7238 | 0.7234 | 0.8171 | 0.7906 |
| LIVE              |      |         |       |     |     |       |    |     |       |      |          |
| SROCC             | 0.9479 | 0.9567 | 0.9599 | 0.9524 | 0.9561 | 0.9487 | 0.9599 | 0.9068 | 0.9600 | 0.9672 | 0.9768 |
| PLCC              | 0.9449 | 0.9522 | 0.9503 | 0.9482 | 0.9512 | 0.9434 | 0.9586 | 0.9578 | 0.9570 | 0.9691 | 0.9676 |
| RMSE              | 8.9455 | 8.3473 | 7.1997 | 8.6816 | 8.4327 | 9.0635 | 7.7764 | 7.7678 | 7.9274 | 7.1573 | 7.9035 |
| KROCC             | 0.7963 | 0.8175 | 0.8366 | 0.8058 | 0.8150 | 0.7982 | 0.8254 | 0.8230 | 0.8203 | 0.8406 | 0.8752 |
| SROCC             | 0.7990 | 0.8439 | 0.8865 | 0.9113 | 0.8487 | 0.8916 | 0.9003 | 0.9135 | 0.7638 | 0.8751 | 0.9133 |
| PLCC              | 0.8181 | 0.8676 | 0.8933 | 0.9047 | 0.8676 | 0.8863 | 0.9019 | 0.9110 | 0.8307 | 0.8893 | 0.9083 |
| KROCC             | 0.6165 | 0.6683 | 0.7140 | 0.7388 | 0.6776 | 0.7098 | 0.7245 | 0.7407 | 0.6096 | 0.7115 | 0.7503 |
| SROCC             | 0.8350 | 0.8870 | 0.9065 | 0.9223 | 0.8780 | 0.9071 | 0.9195 | 0.9193 | 0.8178 | 0.9080 | 0.9278 |
| PLCC              | 0.8422 | 0.8891 | 0.9056 | 0.9131 | 0.8841 | 0.8994 | 0.9191 | 0.9154 | 0.8646 | 0.9152 | 0.9188 |
| KROCC             | 0.6557 | 0.7079 | 0.7428 | 0.7555 | 0.7094 | 0.7321 | 0.7552 | 0.7511 | 0.6553 | 0.7531 | 0.7749 |

D. PERFORMANCE COMPARISONS AMONG DIFFERENT DISTORTION TYPES AND STATISTICAL SIGNIFICANCE COMPARISONS

A good IQA model need yield good performance to predict consistently well on each distortion type. In this section, the performance of the models for each type of distortion was examined and results are all summarized in Table 3, where the results on TID2008 are not presented since TID2013 database contains all distortion types of TID2008. SROCC was used as the performance measure, since the other measures, i.e., PLCC, RMSE and KROCC, have similar performance. Thus, 35 groups of distorted images in the main three databases were selected. For each database and each distortion type, the IQA indices producing the top three SROCC values are highlighted in bold. The best SROCC-performance winner is the proposed model (22 times), followed by VSI (16 times), LLM (15 times) and CVSS (14 times). Moreover, they perform much better the other IQA models. It can be concluded that the proposed IQA model performs quite well on the whole databases with different distortion types.

In order to better illustrate the better performance of the proposed IQA models on different types of distortions, scatter plots are shown in Figure 6 using TID2013 database. It can be observed that the objective scores of the proposed model is highly correlated to the subjective ratings, compared to the other models.

Furthermore, the results of statistical significance tests to evaluate the performances of the competing models are presented and illustrated in Fig.7, which is achieved.
by performing a series of hypothesis tests based on the prediction residuals of each model after nonlinear regression [3], [20], [34]. Specifically, the left-tailed \( F \)-test is employed to compare every two competing models. A value of \( H = 1 \) for the left-tailed \( F \)-test at a significance level of 0.05 represents that the first model (the model in row) is superior in IQA performance to the second model (model in column) with a confidence greater than 95%. A value of \( H = 0 \) shows that these two competing models have no significance difference in IQA performance. Based on the results in Fig.7, it can be observed that the statistical significance tests of IQA performance are consistent with the results shown in Table 2 and 3. Meanwhile, the proposed model always yields the top three performance in overall databases. Especially, note that no IQA model performs significantly better than proposed model on LIVE. Two learning-based models, i.e., IFS and LLM, perform better than proposed model in other databases (part as a training set). It can be easily concluded that the proposed model is superior to IFS in most cases. LLM produces unsatisfactory results on CSIQ (almost 4 percent lower than proposed model in SROCC). As for CVSS, it cannot perform satisfactory results on TID2013 (7-8 percent lower than proposed model in SROCC). In all, the proposed method is superior to others due to its excellent performance and universality.

From Fig. 5, the best setting of \( \alpha \) and \( \lambda \) for each database will not be the same, so we also provide the best settings of \( \alpha \) and \( \lambda \), and their corresponding results in Table 4.

### E. PERFORMANCE COMPARISONS AMONG MULTI-DISTORTION DATABASES

In this subsection, two multi-distortion databases, i.e., LIVE MD [42] and MDID [43], are selected to measure the performance of the IQA models. LIVE MD consists of 450 images
corrupted by two multiple distortion scenarios: (1) blur + JPEG compression; (2) blur + Gaussian noise. MDID contains 1600 distorted images of random distortion types and levels. Table 5 lists the SROCC and PLCC performances of proposed model and the other seven models on these two multi-distortion databases. The best performance is highlighted in boldface. From Table 5, IW-SSIM, RVSIM and IFS achieve the best results, but many IQA models yield similar performance, including the proposed model. Moreover, performances of these models are all lower than the results in single distortion databases. This indicates that IQA models, which can effectively measure the multi-distortion situations, are still lacking [3].

### F. COMPUTATIONAL COST

The efficiency of an IQA model is also an important consideration. We compared the running time of different models on a PC with a 2.5GHz Intel Core i5 CPU and an 8G RAM. The software platform was MATLAB R2013b. Table 6 lists the running time that each model used to compared a pair of color images on TID2013 database, with a resolution of $512 \times 384$. It can be concluded that the proposed model has a moderate computational complexity. In experiments, we find that the time cost of color space conversion is about 0.3148s and this time can be reduced by off-line conversion. And then, the efficiency of proposed model will improve obviously.

| Database Criteria | SSIM | IW-SSIM | FSIMc | VSI | GSM | DSCSI | IFS | RVSIM | CVSS | Proposed |
|-------------------|------|---------|-------|-----|-----|-------|-----|-------|------|----------|
| LIVE MD           | SROCC | 0.8366  | 0.8759 | 0.8635 | 0.8414 | 0.8454 | 0.8287 | 0.8510 | 0.8687 | 0.8666 | 0.8540 |
|                   | PLCC  | 0.7980  | 0.8448 | 0.8185 | 0.8789 | 0.8808 | 0.8719 | 0.8579 | 0.8990 | 0.8930 | 0.8846 |
| MDID              | SROCC | 0.8069  | 0.8911 | 0.8872 | 0.8569 | 0.8137 | 0.8649 | **0.9070** | 0.8835 | 0.7882 | 0.8683 |
|                   | PLCC  | 0.8139  | 0.8983 | 0.8969 | 0.8703 | 0.8352 | 0.8767 | **0.9132** | 0.8953 | 0.7766 | 0.8816 |

### FIGURE 6. Scatter plots of subjective MOS against scores calculated by models’ prediction for TID2013 databases.
FIGURE 7. Results of statistical significance tests of the competing IQA models on the databases of (a) LIVE, (b) CSIQ, (c) TID2008, and (d) TID2013. A value of ‘1’ (highlighted in green) indicates that the model in the row is significantly better than the model in the column, while a value of ‘0’ (highlighted in red) represents that the first model is not significantly better than the second model.

TABLE 6. Time cost of IQA models.

| IQA model | SSIM | IW-SSIM | FSIMc | VSI | GSM | DSCSI | IFS | RVSIM | CVSS | Proposed |
|-----------|------|---------|-------|-----|-----|-------|-----|-------|------|----------|
| Running time | 0.0601 | 0.4633 | 0.3110 | 0.2727 | 0.0715 | 1.0759 | 0.0771 | 1.0819 | 0.0396 | 0.6240 |

V. CONCLUSION

In this research, a novel and good performance FR-IQA model was proposed, namely visual saliency with color appearance and gradient similarity (VCGS). The model consists of a VC similarity map, a gradient similarity map and a chrominance similarity map, characterizing visual salience feature, structure feature and chrominance feature, respectively. Specifically, the VC map comprises of the high-frequency parts for vividness and depth, which are the two new indices for characterizing the visual attention features. VC map was chosen as a weighting function to define the importance of a local image part. After combining the three maps, the main parameters are determined by experiments. To illustrate the outstanding performance of the proposed model, other 10 state-of-the-art or widely cited IQA models were compared using four large scale IQA databases. The results suggested that VCGS yielded statistically better performance of prediction accuracy than the other models with a moderate computational complexity. In the multiple distortion IQA experiment, the proposed model has a similar performance with other models. In the future, to improve the performance of multi-distortion IQA problems is the development direction for all IQA models, including proposed model.

REFERENCES

[1] Y. Zhan, R. Zhang, and Q. Wu, “A structural variation classification model for image quality assessment,” IEEE Trans. Multimedia, vol. 19, no. 8, pp. 1837–1847, Aug. 2017.
[2] 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, Oct. 2014.
[3] S. Athar and Z. Wang, “A comprehensive performance evaluation of image quality assessment algorithms,” IEEE Access, vol. 7, pp. 140030–140070, 2019.
[4] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural sim-
ilarity index for image quality assessment,” in Proc. Conf. Rec. 37th Asilomar
Conf. Signals, Syst. Comput., vol. 1, and 2, pp. 1398–1402, 2003.

[5] 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.

[6] 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.

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

[8] 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.

[9] D. M. Chandler, “Most apparent distortion: Full-reference image quality
assessment and the role of strategy,” J. Electron. Imag., vol. 19, no. 1,
Jan./Mar. 2010, Art. no. 011006.

[10] L. Zhang, L. Zhang, and X. Mou, “RFSIM: A feature based image quality
assessment metric using Riesz transforms,” in Proc. IEEE Int. Conf. Image
Process., Sep. 2010, pp. 321–324.

[11] 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.

[12] 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.

[13] H. Jia, L. Zhang, and T. Wang, “Contrast and visual saliency similarity-
induced index for assessing image quality,” IEEE Access, vol. 6, pp.
65885–65893, 2018.

[14] B. Zhou, F. Shao, X. Meng, R. Fu, and Y.-S. Ho, “No-reference quality
assessment for pan sharpened images via opinion-unaware learning,” IEEE
Access, vol. 7, pp. 40388–40401, 2019.

[15] W. Lin and C.-C. Jay Kuo, “Perceptual visual quality metrics: A survey,”
J. Vis. Commun. Image Represent., vol. 22, no. 4, pp. 297–312, May 2011.

[16] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, “A Haar wavelet-
based perceptual similarity index for image quality assessment,” Signal
Process., Image Commun., vol. 61, pp. 33–43, Feb. 2018.

[17] R. Wang, H. Yang, Z. Pan, B. Huang, and G. Hou, “Screen content image
quality assessment with edge features in gradient domain,” IEEE Access,
vol. 7, pp. 5285–5295, 2019.

[18] H.-W. Chang, Q.-W. Zhang, Q.-G. Wu, and Y. Gan, “Perceptual image
quality assessment by independent feature detector,” Neurocomputing,
vol. 151, pp. 1142–1152, Mar. 2015.

[19] A. Hyvarinen, “Fast and robust fixed-point algorithms for indepen-
dent component analysis,” Proc. Inst. Elect. Eng., vol. 150, no. 4,
pp. 429–448, Jun. 2003.

[20] M. Sonka, V. Hlavac, and R. Boyle, Machine Vision. New York, NY
USA: McGraw-Hill, 1995.