No Reference Quality Assessment for Screen Content Images Using Stacked Autoencoders in Pictorial and Textual Regions

Jiachen Yang*, Member, IEEE, Yang Zhao*, Jiacheng Liu, Bin Jiang*, Qinggang Meng†, Senior Member, IEEE, Wen Lu*, Member, IEEE, and Xinbo Gao*, Senior Member, IEEE.

Abstract—Recently, the visual quality evaluation of screen content images (SCIs) has become an important and timely emerging research theme. This article presents an effective and novel blind quality assessment metric for SCIs by using stacked autoencoders (SAE) based on pictorial and textual regions. Since the SCI consists of not only the pictorial area but also the textual area, the human visual system (HVS) is not equally sensitive to their different distortion types. First, the textual and pictorial regions can be obtained by dividing an input SCI via an SCI segmentation metric. Next, we extract quality-aware features from the textual region and pictorial region, respectively. Then, two different SAEs are trained via an unsupervised approach for quality-aware features that are extracted from these two regions. After the training procedure of the SAEs, the quality-aware features can evolve into more discriminative and meaningful features. Subsequently, the evolved features and their corresponding subjective scores are input into two regressors for training. Each regressor can obtain one output predictive score. Finally, the final perceptual quality score of a test SCI is computed by these two predicted scores via a weighted model. Experimental results on two public SCI-oriented databases have revealed that the proposed scheme can compare favorably with the existing blind image quality assessment metrics.

Index Terms—Human visual system (HVS), quality-aware features, screen content image (SCI), stacked autoencoders (SAE), unsupervised approach.

I. INTRODUCTION

W ith the flying start and rapid popularization of online gaming, remote desktop control, mobile Web browsing, virtual screen sharing, and other application scenarios, the processing and transmission of screen content images (SCIs) have become more and more extensive [1]. The SCI is a composite image that contains not only computer-generated graphics and text but also natural scene images (NSIs) taken by digital cameras. In order to understand SCIs more intuitively, several SCI examples [2] are shown in Fig. 1. From the figure, we can see that SCIs have high contrast and sharp edges. For decades, image quality assessment (IQA) has received widespread attention from researchers, and we have witnessed several milestones research during this time. However, because the SCI has several easily distinguishable features, it should be treated differently. In the transmission process of SCIs, various distortion types can occur, such as the Gaussian noise (GN), Gaussian blur (GB), compression artifacts, and so on. Hence, it is urgent to propose a practical IQA model to evaluate the visual quality of distorted SCIs. Furthermore, it is essential for the service provider to know the qualities of SCIs to maintain the user’s quality of experience (QoE) [3].

In order to address the above-mentioned issues, it is necessary to use deep networks and extract effective features that are consistent with the human visual system (HVS). Meanwhile, the existing theory indicates that none of the existing IQA algorithms can achieve the best results in all cases. Therefore, the new score is set as a nonlinear combination of quality scores via multiple metrics and appropriate weights acquired through the training process [4]–[6]. For instance, different quality evaluation approaches can deal with different types of image distortions well. Hence, the proposed model applying multiple regressors is inspired by this concept. In this article, we introduce an image quality evaluator for SCIs via a stacked autoencoder (SAE) based on different regions. In order to model the framework, the textual and pictorial regions are obtained and two parallel SAEs are trained in an unsupervised manner, respectively. Compared with previous methods, the metric we put forward has the following contributions.

1) The proposed metric is based on the innovative utilization of SAEs in screen content IQA (SIQA). The deep belief network (DBN) and convolutional neural network (CNN) were exploited in the field of IQA, showing better performance than IQA metrics based on
the shallow architecture [7]–[9]. In CNN-based metrics, CNNs automatically extract features about the quality, so the resulting features are not as useful as elaborately extracted features by researchers for IQA [10]. In DBNs-based methods, both unsupervised pretraining and supervised fine tuning are implemented to train DBNs. However, we train SAEs in a purely unsupervised fashion to transform quality-aware features into more valuable features without quality labels.

2) Existing papers have revealed that a human perceives the textual region and pictorial region differently. Therefore, different quality-aware features need to be extracted in different regions. In the textual region, the edge information is extracted to assess the visual quality of the textual region, which can be exploited to represent the corresponding structure feature. In the pictorial region, image statistical features are extracted to predict the perceptual quality of the pictorial region, which can reflect its distortion change.

3) The proposed model has acquired higher satisfactory performance in contrast with several no reference (NR) models for SCIs. It demonstrates that the presented algorithm is more consistent with the HVS. Furthermore, there is a stronger convergence of the scatter in the proposed metric compared with the other existing approaches, which further reveals the fact that the perceptual quality scores predicted by our method are more consistent with subjective opinion scores.

The remainder of this article can be organized as follows. The related work is briefly introduced in Section II. Section III demonstrates the proposed method in detail. Section IV presents and discusses the performance results of contrast experiments conducted in the SIQAD and SCI database (SCID) databases. Finally, the conclusion is drawn and future plans are envisaged in Section V.

II. RELATED WORKS

This section outlines the existing IQA metrics and types which they belong to. We classify IQA metrics from different sources of information: natural scene IQA (NIQA), document IQA (DIQA), and SIQA. Several representative metrics for each category are described below.

A. Types of IQA Methods

According to the availability of pristine images, IQA schemes can be divided into three types: 1) full reference (FR); 2) reduced reference (RR); and 3) NR/blind. When pristine images are available, FR-IQA approaches can be applied to directly evaluate the difference between the distorted image and its pristine image. As for RR-IQA methods, only partial information of the pristine image needs to be calculated for the visual quality score. In practice, pristine images are commonly unavailable, hence NR-IQA metrics should be used instead.

Blind IQA (BIQA) approaches do not always implement as well as FR metrics since the visual quality score is calculated solely via distorted images without reference images. Nevertheless, the BIQA method can be applied to a wider range of applications (e.g., image/video retargeting, streaming media, and computer graphics). Since no references need to be processed, computation requirements are usually low. Therefore, more and more researchers begin to develop BIQA methods.

B. Overview of NIQA Methods

Extensive research on NIQA has been conducted over the past few decades. Classical FR-IQA metrics include the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [11], multiscale SSIM (MS-SSIM) [12], feature similarity (FSIM) [13], visual information fidelity (VIF) [14], internal generative mechanism (IGM) [15], visual saliency-induced index (VSI) [16], etc. RR-IQA metrics include RR entropic differenting (RRED) [17], RR-SSIM [18], RR metric based on VIF (RR-VIF) [19], and so on.

In the past few years, many excellent BIQA approaches have been put forward, such as image verity and integrity evaluation (DIIVINE) [20], blind/referencelss image spatial quality evaluator (BRISQUE) [21], natural image quality evaluator (NIQE) [22], the integrated-local NIQE (IL-NIQE) [23], blind image integrity notator using DCT statistics (BLIINDS-II) [24], and the quality-aware clustering (QAC) [25]. Moreover, Gu et al. [26] proposed an NR model called NFERM by combining spatial natural statistical scene (NSS) features with free energy principle-based features. Ghadiyaram and Bovik [27] put forward an effective BIQA method by using a bag of features on authentically distorted images. Ye and Doermann [28] designed an approach to encode image by exploiting a visual codebook. Li et al. [29] presented an NR NIQA method by using structural and luminance features.

C. Overview of DIQA Methods

Due to the potential applications of DIQA and the growing demand for the document analysis and identification,
much attention has been drawn recently. Several document image databases \cite{30}, \cite{31} have been published. The document images in these databases mainly consist of the gray or binary text without any graphical information. Therefore, the properties of the document images are very different from those of the NSIs. The significant difference from NSIs is that the degradation of the document image is mainly provided by different quality factors of the creation, external degradation, and digitization stages, leading to completely different types of degradation (apart from the typical compression and blurring distortion in NSIs) \cite{32}. Unlike NIQA, DIQA focuses on predicting the accuracy of optical character recognition (OCR) or human perception without any reference. Kang et al. \cite{33} extracted more effective quality-aware features via CNN. Alaei et al. \cite{34} proposed an FR DIQA method according to the hypothesis that the human perception of the document can be influenced by the foreground information. Subsequently, Alaei et al. \cite{35} presented an NR DIQA metric via modification of the QAC method that integrates a patch selection strategy. Ye and Doermann \cite{36} designed an unsupervised feature learning method by learning features for predicting the accuracy of OCR. Lu et al. \cite{37} put forward a binary document image objective distortion measure by applying a reciprocal of distance. Peng et al. \cite{38} proposed a method based on sparse representation to evaluate the document image quality in terms of the OCR functionality. Generally, the effectiveness of DIQA metrics is certified by the accuracy of OCR rather than the manual selection judgment.

D. Overview of SIQA Methods

The SCI can be regarded as the combination of the natural content image and the document content image. However, NIQA or DIQA metrics cannot be directly used to assess the visual quality of SCIs. The reason is that the properties of SCIs are different from natural scene images (NSIs) with textual regions or document images with pictorial regions.

Although challenging, there have been several previous works on the technical standards system of SIQA. Ni et al. \cite{39} proposed an FR method by using edge similarity (ESIM) to predict the visual quality of SCIs and the basic idea is to use three main edge features, that is, edge contrast, edge width, and edge direction. Gu et al. \cite{40} developed an NR model by extracting four types of quality-aware features and using big data learning. Fang et al. \cite{41} represented global statistical luminance and local texture features (NRLT) with multiple histograms and proposed a no-reference algorithm, called NRLT. Lu and Li \cite{42} devised an NR metric by conducting the orientation selectivity mechanism. Wu et al. \cite{43} designed an NR metric via the most preferred structure feature inspired by the free-energy theory in the human brain, which represents the meaningful part of an image patch. Wu et al. \cite{44} put forward a BIQA method for SCIs based on global and local characteristics of the HVS. Zheng et al. \cite{45} proposed a BIQA approach by considering the hybrid region features fusion. Zhou et al. \cite{46} proposed an NR metric for SCIs by using the spatial continuity to describe statistical features in the form of histograms. Nevertheless, the shallow structure of these approaches cannot fit well with the mathematical functions of the HVS. Therefore, this article creatively utilizes the SAEs that can be regarded as the deep architecture.

In a few years, several BIQA approaches based on deep networks were presented. Yue et al. \cite{48} put forward an NR-IQA scheme for SCIs based on a CNN, which can
learn features automatically. Jiang et al. [49] proposed a quadratic-optimized metric via the deep CNN (QODCNN) for predicting the visual score of SCIs. Cheng et al. [50] devised a SIQA model via a fast CNN, which used an aggregation algorithm. Chen et al. [51] presented a deep-learning framework with a naturalization module to achieve an end-to-end solution for BIQA for SCIs.

III. PROPOSED MODEL

According to the characteristics of the HVS and the mixed content type in SCIs, the article puts forward an effective NR metric for SCIs using SAEs based on different regions. The schematic diagram of the presented approach is shown in Fig. 2. Given a distorted SCI, the pictorial and textual regions are divided by an SCI segmentation method first. Second, the corresponding quality-aware features are extracted from these two regions. Third, we train two parallel SAE networks in an unsupervised manner, respectively, and next, input these extracted features into the well-trained SAEs to calculate the corresponding deep features. Then, two support vector regressions (SVRs) are trained by exploiting deep features of these two regions and the corresponding differential mean opinion scores (DMOS) value of the distorted SCI. Finally, the textual score and pictorial score obtained by two SVRs are pooled into an entire visual quality score by means of the weighted strategy. The following sections provide detailed information on each phase.

A. SCI Segmentation

Based on the subjective experiment in building the SIQAD database for assessing the perceptual quality of the SCI, Yang et al. [2] concluded that the textual area and pictorial area of an SCI bring different visual perception characteristics, especially when the SCI suffers from various distorted versions. Fig. 3 shows two distorted images corresponding to the same original SCI in the SIQAD database. The distortion type of one SCI is GB, while the distortion type of the other SCI is contrast change (CC). The distortion level of these two distorted SCIs is at the same analog processing level.

As can be seen from Fig. 3, when the SCI is affected by the distortion of GB, the distortion will make it difficult for the observer to obtain the visual information of the textual area of the SCI. But the observer can still tell what the pictorial area is roughly. So the effect of the GB on the textual area is greater than that on the pictorial area. When the SCI is affected by the distortion of the CC, the improvement of contrast makes the text and background better distinguish, which increases people’s comfort. Hence, the effect of CC in the pictorial area is greater than that in the textual area. The same situation is true for other distorted SCIs. Therefore, this article puts forward an NR evaluation approach for SCIs based on the different visual characteristics of the two regions in the SCI.

In order to divide the SCI into the textual and pictorial regions, this article uses the fast CNN-based document layout analysis algorithm in the paper [52]. The method first divides the content in the image into blocks and then calculates the horizontal and vertical projections of these image tiles as inputs to the 1-D CNN model, which classifies them into text, tables, and images. Then, it combines a simple voting pattern to output the classification results. An example of SCI segmentation is demonstrated in Fig. 4.

B. Feature Extraction of the Textual Region

As we all know, there are various characters in the content of the textual region, which are consisted of plenty of sharp edges. For the purpose of capturing the distorted change of edge information, the histogram of a local binary pattern computed on the gradient map (GMLBP) is exploited to represent the edge characteristics of the textual region, which reflects the structure of the textual region. For a textual map $T$, the gradient magnitude (GM) can be defined as

$$G(i, j) = \sqrt{G^2_H(i, j) + G^2_V(i, j)}$$  (1)

where

$$G_H = h_H \otimes T = \frac{1}{16} \begin{bmatrix} +3 & 0 & -3 \\ +10 & 0 & -10 \\ +3 & 0 & -3 \end{bmatrix} \otimes T$$  (2)

$$G_V = h_V \otimes T = \frac{1}{16} \begin{bmatrix} +3 & +10 & +3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix} \otimes T.$$  (3)

In the above formulas, $(i, j)$ denotes the pixel coordinate, $h$ denotes the gradient operator, $G_H$ and $G_V$ represent the gradient information in the horizontal and vertical directions, respectively, and $\otimes$ is the symbol of the convolution operation. Then, we compute the rotation invariant and uniform local binary pattern (LBP) [53] for each pixel in the GM. Through applying the LBP operation in the GM, the GMLBP code at one location is deduced as

$$G^{MLBP}_{U,S} = \sum_{k=0}^{U-1} \rho(G_k - G_c)2^k$$  (4)

where $U$ represents the number of neighbors, $S$ indicates the radius value of the neighborhood, $G_k$ and $G_c$ denote the GM values in the center coordination and its neighborhood, and $\rho(\cdot)$ is defined as the thresholding function. According to the study [38], the locally rotation invariant uniform GMLBP operator is expressed by

$$G^{MLBP}_{U,S} = \frac{1}{U+1} \sum_{k=0}^{U-1} \rho(G_k - G_c), \quad \Delta(G^{MLBP}_{U,S}) \leq 2$$ otherwise  (5)
where $\Delta$ denotes the uniform metric, and the superscript $\text{rui2}$ represents the pattern of rotation invariant uniform if $\Delta$ value less than 2. The uniform metric $\Delta$ is expressed as the quantity of bitwise transitions

$$
\Delta(\text{GMLBP}_{U,S}) = \|\rho(G_{U-1} - G_c) - \rho(G_0 - G_c)\|
$$

$$+ \sum_{k=0}^{U-1} \|\rho(G_k - G_c) - \rho(G_{k-1} - G_c)\|. \quad (7)
$$

It is observed that the rotation-invariant uniform GMLBP can contain $U + 2$ distinct patterns, which can be combined into one bin of the histogram. We set $U$ as 8 and therefore each GMLBP histogram would contain ten bins in our algorithm. The global priority principle reveals that the HVS uses the strategy from coarse to fine to perceive image edges [54]. Thus, our presented quality-aware features are extracted in a multiscale strategy. In addition to the pristine image scale, the coarser scale is formed by low-pass filtering by sampling a coefficient of 2 in each dimension. So we extract edge features at five scales. Therefore, extracted features have a total of 50 dimensions in the textual region of the SCI. The histograms of the structure feature are figured in Fig. 5. From this figure, it can be observed explicitly that the histogram of the GMLBP operator has the ability to distinguish various distortions. For instance, the GN decreases the central cusp of each statistical feature. By contrast, GB shapes the central apex more.

C. Feature Extraction of the Pictorial Region

As indicated in the above sections, image statistical features are extracted to predict the perceptual quality of the pictorial region. Different from NSIs, the structure of SCIs does not follow certain statistical characteristics that may be destructed by the introduction of distortions and most BIQA methods cannot be directly exploited for the perceptual quality evaluation for SCIs, such as GGD [20], AGGD [55], and MSCN [21]. In order to effectively reflect the introduction of distortions (e.g., the CC and GB) from the pictorial region of the SCI, we extract the exponential attenuation characteristics [56] in different wavelet subbands of SCIs, which include three types of image statistical features. They are magnitude (generalized spectral behavior), variance (the fluctuations of energy), and entropy (generalized information).

The following procedures introduce how to extract image statistical features in the pictorial region. First, the input pictorial region $P$ is decomposed into subband wavelet coefficients $y$ by the wavelet transform. As the frequency of the wavelet subband increases, multiple subbands are defined. Since low–high (LH) and high–low (HL) subbands have similar statistics at the identical scale, we will compute the characteristics (magnitude, variance, and entropy) of the two subbands, that is, we do not distinguish LH and HL subbands at the identical scale. The image is decomposed into four scales and eight wavelet subbands are obtained. For each subband, we compute the magnitude $m_y$, the variance $v_y$ and the entropy $e_y$, which are defined as follows:

$$m_y = \frac{1}{N_y \times M_y} \sum_{i=1}^{M_y} \sum_{j=1}^{N_y} \log_2|\Gamma_y(i,j)| \quad (8)$$

$$v_y = \frac{1}{N_y \times M_y} \sum_{i=1}^{M_y} \sum_{j=1}^{N_y} \log_2|\Gamma_y(i,j) - m_y| \quad (9)$$
Fig. 5. Example of our proposed structure feature. (a) Reference SCI. (b)–(d) Three types of distorted SCIs corrupted by the GN, GB, and JPEG compression, respectively. They are all processed in the same distorted level. (e)–(h) Textual map corresponding to (a)–(d). (i)–(l) Structure feature corresponding to (a)–(d), respectively.

\[ e_y = \sum_{i=1}^{M_y} \sum_{j=1}^{N_y} \Gamma_y(i,j) \ln p\left[\Gamma_y(i,j)\right] \]  

(10)

where \( \Gamma_y(i,j) \) represents the \((i,j)\) coefficient of the \(y\)th subband, \(M_y\) and \(N_y\) denote the length and width of the \(y\)th subband, respectively, and \(p\left[\cdot\right]\) is defined as the probability density function of the subband. The vertical subband and horizontal subband with the same mark at the identical scale are combined by the average of the above treatment. Finally, there are 24 features for the pictorial region.

D. Feature Evolution

Quality-aware features extracted from SCIs can be used as simple features to distinguish the original SCIs and distorted SCIs. Nevertheless, an intuitive idea is whether we can design a system that amplifies the differences between the original SCI features and the distorted SCI features before inputting this primary feature into the regressor. Recent research on the deep neural network inspires us to address this issue. In our algorithm, SAEs are applied as an amplifier to widen the gap between pristine and distorted SCI features and make them more recognizable and meaningful.

As a branch of the deep neural network structure, the SAE plays an important role in lots of fields, such as dimensionality reduction and feature learning [57], [58]. The basic theory of SAEs is to extract features from input samples and express an object through more basic features. As shown in Fig. 6, \(h\) autoencoders are trained from bottom to top in the hierarchical order. The input vector is fed into the bottom autoencoder, which is shown as the green color in the figure. After the training of the bottom autoencoder, the hidden representation of the output is propagated to a higher layer, which is shown as the red color in the figure. Sigmoid or tanh functions are commonly used for the activate function. Repeat the same process until all the autoencoders have been trained. After such a pretraining phase, the entire neural network is fine tuned to predefined standards. The hidden layer of the top-level autoencoder is the final output of the SAE, which can be further fed back to the support vector machine (SVM) or other applications for classification. Compared with the traditional random initialization method, unsupervised pretraining can automatically use a large number of unlabeled data to obtain better weight initialization.

To obtain the numbers of hidden layers and units (UN) of neurons fitted our approach, we change their quantities as
different models. Then, these models are trained on a training set and we assess their performance results on a test set. Finally, we adopt a configuration of 40-30-20 units each layer for the network of the textual region (SAE-T) and 20-15-10 units for the pictorial region (SAE-P), which can obtain the most superior performance. Therefore, this network structure is chosen as the validate model. In our proposed algorithm, the learning rate (LR) is selected to be 0.6 and the quantity of epoch (EP) is chosen to be 1000 for the entire network. In order to reduce the time of the learning process, the training process is stopped after that full-batch train error (FBTE) is less than 0.005 for the first hidden layer and 0.001 for the second and third hidden layers. The key parameters of SAEs are listed in Table I. The primary features are evolved into deep features after the training of SAEs and the final evolved features are input into the corresponding regressor.

E. Visual Quality Pooling for SCI

Without the procedure of supervised fine tuning in the deep networks, we exploited two SVRs (SVR-T and SVR-P) to obtain the textual and pictorial quality scores of SCIs, respectively. After the pretraining, the SAE network can be utilized to mine potential data information, which is closely related to the perceptual quality and can obtain deep feature representations. In the final stage, those deep features can be exploited to train the SVM regressor to predict the corresponding visual quality score. Note that the SVM regressor is regarded as a shallow learning method. However, the SVM regressor would have a greater performance when small sample data need to be dealt with, particularly in the field of predictive accuracy and generalization ability.

The perceptual quality score of the SCI is acquired by the combination of the textual quality score $Q_T$ and pictorial quality score $Q_P$. We utilized a linear equation to calculate the final quality score of the SCI, which is expressed by

$$Q = \omega_T \cdot Q_T + \omega_P \cdot Q_P$$

$$\omega_T = \frac{s_T}{s_T + s_P}$$

$$\omega_P = \frac{s_P}{s_T + s_P}$$

where $s_P$ and $s_T$ denote the information entropy of the textual region and pictorial region of the SCI, respectively. The information entropy can be defined as

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

where $p_i$ is the probability of a certain grayscale appearing in the image, which can be obtained by the gray histogram.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, several contrast experiments are conducted to prove the superiority of the presented metric and we analyze the impacts of the weighted model and the deep architecture in the designed model.

A. Testing Databases

The SIQAD database [2] is the first database to be established that is comprised of 20 original SCIs and 980 distorted SCIs corrupted by seven types of distorted versions. As for each distorted type, there are seven distorted levels. In contrast, the SCID database [39] includes 1840 SCIs. They are 40 original SCIs and 1800 distorted SCIs corrupted by nine types of distorted versions, but the distorted level is 5 for the individual distortion type. These original images are usually from electronic posters, computer screenshots, PPT files, and electronic magazines. These two databases both involve the GN, GB, motion blur (MB), CC, JPEG compression (JPEG), JPEG2000 compression (JP2K), and layer segmentation-based coding (LSC). Moreover, the SCID database has two more distortion types, which are color quantization with dithering (CQD) and high-efficiency video coding (HEVC).

B. Assessment Criteria

In this article, three common-used criteria, namely, the Pearson linear correlation coefficient (PLCC), Spearman order correlation coefficient (SRCC), and root mean square error (RMSE), are used to assess the performance of this metric. PLCC, SRCC, and RMSE estimate the accuracy, monotonicity, and consistency of the prediction results, respectively. Generally speaking, the better IQA approach must have higher

| Network  | The pre-training phase | The first layer | The second layer | The third layer |
|----------|------------------------|----------------|-----------------|----------------|
| SAE-T    |                        |                |                 |                |
|          |                        | 40             | 30              | 20             |
|          |                        | 0.6            | 0.6             | 0.6            |
|          |                        | 1000           | 1000            | 1000           |
|          |                        | FBTE ≤ 0.005   | FBTE ≤ 0.001    | FBTE ≤ 0.005   |
| SAE-P    |                        |                |                 |                |
|          |                        | 20             | 15              | 10             |
|          |                        | 0.6            | 0.6             | 0.6            |
|          |                        | 1000           | 1000            | 1000           |
|          |                        | FBTE ≤ 0.005   | FBTE ≤ 0.001    | FBTE ≤ 0.001   |

| Network  | The pre-training phase | The first layer | The second layer | The third layer |
|----------|------------------------|----------------|-----------------|----------------|
| SAE-T    |                        |                |                 |                |
|          |                        | 40             | 30              | 20             |
|          |                        | 0.6            | 0.6             | 0.6            |
|          |                        | 1000           | 1000            | 1000           |
|          |                        | FBTE ≤ 0.005   | FBTE ≤ 0.001    | FBTE ≤ 0.005   |
| SAE-P    |                        |                |                 |                |
|          |                        | 20             | 15              | 10             |
|          |                        | 0.6            | 0.6             | 0.6            |
|          |                        | 1000           | 1000            | 1000           |
|          |                        | FBTE ≤ 0.005   | FBTE ≤ 0.001    | FBTE ≤ 0.001   |

| Network  | The pre-training phase | The first layer | The second layer | The third layer |
|----------|------------------------|----------------|-----------------|----------------|
| SAE-T    |                        |                |                 |                |
|          |                        | 40             | 30              | 20             |
|          |                        | 0.6            | 0.6             | 0.6            |
|          |                        | 1000           | 1000            | 1000           |
|          |                        | FBTE ≤ 0.005   | FBTE ≤ 0.001    | FBTE ≤ 0.005   |
| SAE-P    |                        |                |                 |                |
|          |                        | 20             | 15              | 10             |
|          |                        | 0.6            | 0.6             | 0.6            |
|          |                        | 1000           | 1000            | 1000           |
|          |                        | FBTE ≤ 0.005   | FBTE ≤ 0.001    | FBTE ≤ 0.001   |

| Network  | The pre-training phase | The first layer | The second layer | The third layer |
|----------|------------------------|----------------|-----------------|----------------|
| SAE-T    |                        |                |                 |                |
|          |                        | 40             | 30              | 20             |
|          |                        | 0.6            | 0.6             | 0.6            |
|          |                        | 1000           | 1000            | 1000           |
|          |                        | FBTE ≤ 0.005   | FBTE ≤ 0.001    | FBTE ≤ 0.005   |
| SAE-P    |                        |                |                 |                |
|          |                        | 20             | 15              | 10             |
|          |                        | 0.6            | 0.6             | 0.6            |
|          |                        | 1000           | 1000            | 1000           |
|          |                        | FBTE ≤ 0.005   | FBTE ≤ 0.001    | FBTE ≤ 0.001   |
Correlations about PLCC, SRCC, and RMSE values between the subjective and predictive objective scores are calculated based on the comparison metrics used. Table II lists the experimental performance, where the highest two results in each type are emphasized in boldface. We download all program source codes of comparison models from the original address provided by the author. Since codes of the NRLT and HRFF methods are not available, the relevant results are reproduced directly from the original papers.

As shown in Table II, in contrast with all the BIQA methods, the proposed method can acquire superior performance results in predicting the visual quality of SCIs. Reasonable explanations about the above phenomena can be summarized in the following aspects. For one thing, the SCI is more complex compared with the NSI. It is acceptable that those approaches proposed for NSIs cannot assess the quality of SCIs solely by means of characteristics of NSIs. For example, NIQE is based on the assumption that high-quality NSIs satisfy a regularity that is undermined by distortions. The degree of distortion can be quantified by the distance between the extracted distorted image features and the established model from high-quality images. Unfortunately, such regularity cannot be applied to SCIs [41]. As a result, NIQE fails to achieve satisfactory performance. As for BQMS, SIQE, and ASIQE, they extract a large quantity of quality-aware features according to the property of SCIs and apply these features and quality labels to establish the prediction model. Nevertheless, the quality label is estimated via several FR metrics. There is no doubt that this approach is limited since it can be influenced by the performance of the chosen FR approach [32]. On the contrary,
the designed model aims to solve the problem of quality evaluation by analyzing the main characteristics of SCIs.

As for NRLT and HRFF, the computation metrics for the textual and pictorial regions are identical without considering their different characteristics. On the contrary, the proposed method uses different measures to predict the quality scores of the textual and pictorial regions, respectively, according to the property of the HVS. The edge information in the textual region is exploited as the structure feature because the HVS is highly sensitive to it. Image statistical features can be applied to the perceptual quality estimation of the pictorial regions, which can reflect the distortion change of the pictorial region. In addition, we determine the weights of the textual region and pictorial region based on their information entropy rather than the average weight.

Furthermore, to validate the stability and superiority of the proposed metric, an error rate statistical experiment is conducted between the proposed metric and other approaches for different distorted SCIs. Fig. 7 lists the experimental results. It is obviously observed that the proposed metric computes the lowest error rate except for the case of DMOS 38.9563. SIQE predicts the lowest error rate than the remainder on the DMOS 38.956, but it has poor performance about other DMOS values. Overall, our proposed method has relatively stable predictive performance than other approaches with unstable error rates.

In addition, we also study the scatter plots between predictive quality scores against subjective opinion scores, as illustrated in Fig. 8. The contrast algorithms studied include five BIQA models: 1) NIQE; 2) DIVINE; 3) BQMS; 4) SIQE; and 5) ASIQE. As we can see, there is a stronger convergence of the scatter in the proposed metric compared with other existing approaches, which further reveals the fact that the perceptual quality scores predicted by our method are more consistent with subjective opinion scores.

To confirm the effectiveness of our presented scheme, an extra contrast experiment is added for NR metrics in the SCID database. In this contrast experiment, we add BRISQUE [21] and NFERM [26] for the result comparison. The experimental results are listed in Table III. Meanwhile, the top result in the individual case is emphasized with boldface. It can be concluded that the proposed model can acquire the highest accuracy in predicting the perceptual quality of SCIs.

D. Comparison Results on Individual Distortion Version

For comprehensively certifying the performance of the proposed metric on the individual distortion version, we conduct the contrast experiment in the SIQAD database. Specifically, the performance results of PLCC, SRCC, and RMSE are reported in Table II. The top two results are emphasized in boldface. It can be observed that the performance of the proposed method on most distortion categories can achieve the best results among the compared approaches. The reasonable explanation is that the extracted quality-sensitive features can effectively detect different distortion categories in SCIs. The proposed model might obtain worse performance than certain contrast approaches, such as LSC, since it may not be suitable for all distortion categories due to the property of different perceptual distortion categories.

E. Impacts of Different Features in Textual Regions

Many structure-based or texture-based features have been proposed for a variety of images, such as NSIs, sonar images, etc. In this section, we adopt BRISQUE features and GLCM features in the textual region of the SCI for comparison. The BRISQUE feature comes from the structure distribution of normalized brightness and the product of adjacent normalized brightness in the spatial domain. Meanwhile,
Haralick et al. [60] exploited GLCM to obtain the spatial division of image textures and acquired statistical features based on this matrix through texture analysis. Experimental results are reported in Table IV. It can be observed that the proposed GMLBP features can achieve the best performance. The main reason is that the difference between NSIs and SCIs leads to the failure of the NSS model of BRISQUE to address the issue of SIQA. Although GLCM features provide information about the gray-level direction, interval, and change amplitude of the image, it cannot directly provide the characteristics that distinguish the texture of the textual region.

### F. Impacts of the Weighted Model

To illustrate the importance of the weighted model presented in this article, we adopt the average score of the textual and pictorial regions as the perceptual quality score of SCI and conduct comparative experiments. The performance results are listed in Table V. The existing theory has revealed that there are three factors jointly affecting the overall quality of SCIs, including the perceived quality of the textual and pictorial regions, the importance of these two regions, and the plain text [61]. So in the proposed metric, we compute the corresponding region complexity via the information entropy, which can be regarded as the weight value, respectively. It can be observed that the comparison results on the two databases confirm the feasibility of the proposed weighted model. In brief, compared with the measurement using the average weight, the weighted model with the information entropy has better performance.

### G. Impacts of the Deep Architecture

To verify the effectiveness of the deep architecture in the designed model, we conduct a comparison experiment by using the shallow architecture version that only uses SVRs without SAEs (Proposed-SVR). Proposed-SVR directly inputs the prior features into the SVRs. Table VI reports the performance results. It is observed that on both SIQAD and SCID databases, the deep version of the designed algorithm achieves the superior predictive accuracy compared with the Proposed-SVR. Therefore, we can confirm the fact that the function of SAEs is to transform the original features into deeper features. The main reason why the performance of the proposed model in the individual database has not been improved significantly is that there are not enough samples to train SAEs. Moreover, since the SIQAD and SCID databases contain 980 and 1800 SCIs, respectively, the utilization of SAEs brings a larger improvement in the predictive performance of the SCID database. It is well-known that the quantity of data samples is the key factor to process...
deep learning successfully. Therefore, the issue would be addressed easily if the much larger SCI-oriented database can be established in the future.

V. CONCLUSION

In this work, we presented an effective BIQA algorithm for SCIs by using SAEs based on textural and pictorial regions. The novelty of the proposed model is that we trained SAEs in a purely unsupervised fashion to transform quality-aware features into more valuable features. As for the textural region and the pictorial region, we extracted different features for their characteristics. The experimental results demonstrated that the designed model can acquire superior results in contrast with state-of-the-art models. Furthermore, we also testified that the utilization of the deep network is more superior than that of the shallow architecture in the SIQA.

The future work will be carried out in two aspects: 1) developing a more accurate and robust SCI segmentation approach, which can faithfully improve the SIQA and 2) expecting to apply a more effective strategy of the dynamic weighting sum into the SIQA.

REFERENCES

[1] Y. Bai, M. Yu, Q. Jiang, G. Jiang, and Z. Zhu, “Learning content-specific codebooks for blind quality assessment of screen content images,” Signal Process., vol. 161, pp. 248–258, Aug. 2019.

[2] H. Yang, Y. Fang, and W. Lin, “Perceptual quality assessment of screen content images,” IEEE Trans. Image Process., vol. 24, no. 11, pp. 4408–4421, Nov. 2015.

[3] M. A. Layek, N. T. Thu, S. Y. Tu, C. Hung, and E-N. Huh, “Center-emphasized gradient-feature based quality assessment method for screen content images,” in Proc. IEEE 11th Int. Conf. Ubiquitous Future Netw. (ICUFN), Zagreb, Croatia, 2019, pp. 257–262.

[4] T.-J. Liu, W. Lin, and C.-C. J. Kuo, “Image quality assessment using multi-method fusion,” IEEE Trans. Image Process., vol. 22, no. 5, pp. 1793–1807, May 2013.

[5] T.-J. Liu, K.-H. Liu, J. Y. Lin, W. Lin, and C.-C. J. Kuo, “A paraBoost method to image quality assessment,” IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 1, pp. 107–121, Jan. 2017.

[6] H. Ko, R. Song, and C.-C. J. Kuo, “A paraBoost stereoscopic image quality assessment (PBSIIQA) system,” 2016. [Online]. Available: arXiv:1609.09469.

[7] W. Hou, X. Gao, D. Tao, and X. Li, “Blind image quality assessment via deep learning,” IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 6, pp. 1275–1286, Jun. 2015.

[8] B. Jiang, Y. Jiachen, L. Zhihan, and S. Houbing, “Wearable vision assistance system based on binocular sensors for visually impaired users,” IEEE Internet Things J., vol. 6, no. 2, pp. 1375–1383, Apr. 2019.

[9] D. Ghadiyaram and A. C. Bovik, “Blind image quality assessment on real distorted images using deep belief nets,” in Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP), Atlanta, GA, USA, 2014, pp. 946–950.

[10] J. Yang, K. Sim, W. Lu, and B. Jiang, “Predicting stereoscopic image quality via stacked auto-encoders based on stereopsis formation,” IEEE Trans. Multimedia, vol. 21, no. 7, pp. 1750–1761, Jul. 2019.

[11] 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.

[12] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in Proc. IEEE 37th Asilomar Conf. Signals Syst. Comput., vol. 2. Pacific Grove, CA, USA, 2003, pp. 1398–1402.

[13] 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.

[14] 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.
[39] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K. K. Ma, “ESIM: Edge similarity for screen content image quality assessment,” IEEE Trans. Image Process., vol. 26, no. 10, pp. 4818–4831, Oct. 2017.

[40] K. Gu, J. Zhou, J.-F. Qiao, G. Zhai, W. Lin, and A. C. Bovik, “No-reference quality assessment of screen content pictures,” IEEE Trans. Image Process., vol. 26, no. 8, pp. 4005–4018, Aug. 2017.

[41] Y. J. Fan, J. Yan, L. Li, J. Wu, and W. Lin, “No reference quality assessment for screen content images with both local and global feature representation,” IEEE Trans. Image Process., vol. 27, no. 4, pp. 1600–1610, Apr. 2018.

[42] N. Lu and G. Li, “Blind quality assessment for screen content images by orientation selectivity mechanism,” Signal Process., vol. 145, pp. 225–232, Apr. 2018.

[43] J. Wu, H. Li, Z. Xia, and Z. Xia, “Screen content image quality assessment based on the most preferred structure feature,” J. Electron. Imag., vol. 27, no. 3, pp. 1–14, 2018.

[44] J. Wu, Z. Xia, H. Zhang, and H. Li, “Blind quality assessment for screen content images by combining local and global features,” Digit. Signal Process., vol. 91, pp. 31–40, Aug. 2019.

[45] L. Zheng, L. Shen, J. Chen, P. An, and J. Luo, “No-reference quality assessment for screen content images based on hybrid region feature fusions,” IEEE Trans. Multimedia, vol. 21, no. 8, pp. 2057–2070, Aug. 2019.

[46] W. Zhou, L. Yu, Y. Zhou, W. Qiu, J. Xiang, and Z. Zhai, “Blind screen content image quality measurement based on sparse feature learning,” Signal Image Video Process., vol. 13, no. 3, pp. 525–530, 2019.

[47] Y. Fang, R. Du, Y. Zuo, W. Wen, and L. Li, “Perceptual quality assessment for screen content images by spatial continuity,” IEEE Trans. Circuits Syst. Video Technol., early access, Nov. 5, 2019, doi: 10.1109/TCSVT.2019.2951747.

[48] G. Yue, C. Hou, W. Yan, L. K. Choi, T. Zhou, and Y. Hou, “Blind quality assessment for screen content images via convolutional neural network,” Digit. Signal Process., vol. 91, pp. 21–30, Aug. 2019.

[49] X. Jiang, L. Shen, G. Feng, L. Yu, and P. An, “Deep optimization model for screen content image quality assessment using neural networks,” 2019 [Online]. Available: arXiv:1903.00705.

[50] Z. Chen, M. Takeuchi, K. Kanai, and J. Katou, “A fast no-reference screen content image quality prediction using convolutional neural networks,” in Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW), San Diego, CA, USA, 2018, pp. 1–6.

[51] J. Chen, L. Shen, L. Zheng, and X. Jiang, “Naturalization module in neural networks for screen content image quality assessment,” IEEE Signal Process. Lett., vol. 25, no. 11, pp. 1685–1689, Nov. 2018.

[52] M. P. Viana and D. A. B. Oliveira, “Fast CNN-based document layout analysis,” in Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW), Venice, Italy, 2017, pp. 1173–1180.

[53] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.

[54] H. C. Hughes, G. Nozawa, and F. Kitterle, “Global precedence, spatial frequency channels, and the statistics of natural images,” J. Cogn. Neurosci., vol. 8, no. 3, pp. 197–230, 1996.

[55] C. Ming-Jun, L. K. Cormack, and A. C. Bovik, “No-reference quality assessment of natural stereopairs,” IEEE Trans. Image Process., vol. 22, no. 9, pp. 3379–3391, Sep. 2013.

[56] L. He, D. Tao, X. Li, and X. Gao, “Sparse representation for blind image quality assessment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Providence, RI, USA, 2012, pp. 1146–1153.

[57] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, “Contractive auto-encoders: Explicit invariance during feature extraction,” in Proc. 28th Int. Conf. Mach. Learn., 2011, pp. 833–840.

[58] B. Jiang, J. Yang, Q. Meng, B. Li, and W. Lu, “A deep evaluator for image retargeting quality by geometrical and contextual interaction,” IEEE Trans. Cybern., vol. 50, no. 1, pp. 87–99, Jan. 2020.

[59] B. Liu, G. Zhai, W. Lin, X. Yang, and W. Zhang, “Learning a blind quality evaluation engine of screen content images,” Neurocomputing, vol. 196, pp. 140–149, Jul. 2016.

[60] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” IEEE Trans. Syst., Man, Cybern., vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[61] Y. Zhang, D. M. Chandler, and X. Mou, “Quality assessment of screen content images via convolutional-neural-network-based synthetic/natural segmentation,” IEEE Trans. Image Process., vol. 27, no. 10, pp. 5113–5128, Oct. 2018.
Wen Lu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 2002, 2006, and 2009, respectively.

He is currently a Professor with Xidian University and a Postdoctoral Researcher with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA. His research interests include image quality metric, human vision system, and computational vision. He has published two books and around 30 technical articles in refereed journals and proceedings including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, Neurocomputing, and Signal Processing.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi’an, China, in 1994, 1997, and 1999, respectively.

He is currently a Cheung Kong Professor of Ministry of Education, China, a Professor of Pattern Recognition and Intelligent System, and the Dean of Graduate School, Xidian University. He has published six books and around 300 technical articles in refereed journals and proceedings. His current research interests include Image processing, computer vision, multimedia analysis, machine learning and pattern recognition.

Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or a PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology and the Chinese Institute of Electronics.