Blind Night-Time Image Quality Assessment: Subjective and Objective Approaches

Tao Xiang, *Member, IEEE*, Ying Yang, and Shangwei Guo

***Abstract*—Blind image quality assessment (BIQA) aims to develop quantitative measures to automatically and accurately estimate the visual quality of an image without any prior information about its reference image. This issue has been attracting a great deal of attention for a long time; however, little work has been done on night-time images, which are crucially important for consumer photography and practical applications such as automated driving systems. In this paper, to the best of our knowledge, we conduct the first exploration on subjective and objective quality assessment of night-time images. First, we build a large-scale natural night-time image database (NNID) containing 2240 images with 448 different image contents captured by different photographic equipment in real-world scenarios. Subsequently, we carry out a subjective experiment to evaluate the perceptual quality of all the images in the NNID database. Thereafter, we perform objective assessment of night-time images by proposing a blind night-time image quality assessment metric using brightness and texture features (BNBT). Finally, extensive experiments are conducted to evaluate the performance and efficiency of the proposed BNBT metric on the NNID database. The experimental results demonstrate that this metric outperforms existing state-of-the-art BIQA methods in terms of all evaluation criteria and has an acceptable computational cost at the same time. We have made the NNID database publicly available for downloading at https://sites.google.com/site/xiangtaooo/.**

***Index Terms*—Blind image quality assessment, natural night-time images, superpixel, ranking-based weighting, gray-level co-occurrence matrix.**

I. Introduction

NIGHT-TIME images captured under the weakly illuminated night-time environment are characterized by low contrast, blurred details, and reduced visibility, thereby usually exhibiting impaired visual quality. Therefore, determining a well-designed image quality assessment (IQA) metric for faithfully predicting the quality of night-time images becomes a highly urgent and beneficial endeavor for consumer photography and practical image processing systems. With the increasing boom in consumer photography, consumers place high demands on the capability of capturing images in the night environment, and image quality has an immediate impact on their quality of experience (QoE). In this context, a proper IQA metric for night-time images can be used to quantify the QoE of consumers. Furthermore, the performance of many real-time image processing algorithms and image-driven applications, such as visual surveillance and autonomous driving, is strongly affected by the quality of the input images. IQA metrics can be adopted to effectively benchmark and optimize these algorithms and systems [1], [2].

The last few years have witnessed an explosive growth of research on IQA. However, to the best of our knowledge, there is little study of visual quality assessment of night-time images, which is a challenging task for two reasons. First, the absence of a public night-time image quality assessment database is one of obstacles to conduct this research because an appropriate image database is the cornerstone and basic requirement for image quality evaluation. Second, there is currently no specific quality assessment metric for night-time images. Images with different distortions may require different quality evaluation metrics, which are usually designed by considering the characteristics of distortions. For example, the distortion-specific blind image quality assessment (BIQA) metrics are designed based on the characteristics of each type of image distortion.

The existing IQA solutions suffer from many limitations when used in night-time image quality assessment. On the one hand, traditional BIQA methods focus on the simulated distortion rather than on the natural distortion of night-time images. Specifically, the distortions in night-time images do not originate from the introduction of graded simulated distortions into high-quality photographs but are derived from the image sampling process. On the other hand, although existing general-purpose BIQA [3]–[11] methods can be used to evaluate the visual quality of night-time images, their results are unsatisfactory and inconsistent with subjective assessment results. Specifically, a night-time image with a high mean opinion score (MOS) may not achieve a high predicted objective score. An example is given in Fig. 1, where we can see that the perceptual subjective qualities of Fig. 1(a) and Fig. 1(c) are obviously better than those of Fig. 1(b) and Fig. 1(d), respectively, but the latter two images have better objective scores predicted by existing BIQA methods. The reason for this inconsistency is that most existing general-purpose BIQA metrics are designed for images with simulated distortions. Moreover, they target and work well

Manuscript received January 27, 2019; revised May 28, 2019 and July 29, 2019; accepted August 19, 2019. Date of publication August 30, 2019; date of current version April 23, 2020. This work was supported by the National Natural Science Foundation of China under Grants 61672118 and 61932006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hantao Liu. *(Corresponding author: Tao Xiang.)*

T. Xiang and Y. Yang are with the College of Computer Science, Chongqing University, Chongqing 400044, China (e-mail: txiang@cqu.edu.cn; yingyang@cqu.edu.cn).

S. Guo is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: shangwei.guo@ntu.edu.sg).

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

Digital Object Identifier 10.1109/TMM.2019.2938612

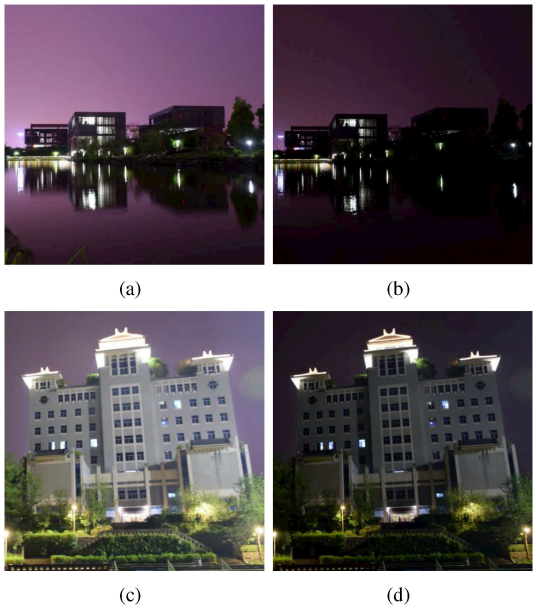


Fig. 1. The performance of BLIINDS-II [4], GM-LOG [5], MSGF [6] and NFERM [7] on images from the NNID database. (a) BLIINDS-II = 45.5, GM-LOG = 111.7. (b) BLIINDS-II = 36.5, GM-LOG = 94.1. (c) NFREM = 29.9, MSGF = 34.7. (d) NFERM = 26.0, MSGF = 32.8.

for relatively high-quality images but exhibit poor performance on night-time images, which are usually of low visual quality.

In this paper, we investigate the problem of visual quality assessment of night-time images from both subjective and objective aspects. Considering that no camera setting can result in a distortion-free and perfect-quality image to serve as a reference image in a relatively dark scenario, we thus focus on the BIQA approach in this paper. First, we build a dedicated large-scale natural night-time image database (NNID) containing 2240 night-time images with 448 different image contents captured by different photographic equipment with different exposure times and other settings, i.e., a digital camera (Nikon D5300), a mobile phone (iPhone 8plus) and a tablet (iPad mini2). Then, we carry out subjective assessment of the NNID database by adopting the single stimulus (SS) method recommended by the International Telecommunications Union (ITU). In a carefully prepared test setting, we invite 74 inexperienced observers to evaluate the images of the NNID database based on their own visual perception and then obtain the final MOS by averaging all the valid scores after necessary postprocessing procedures.

We further carry out objective assessment of NNID by proposing a blind night-time image quality assessment metric based on brightness and texture features (BNBT). The brightness features are extracted from superpixel segments of a night-time image, and the texture features are extracted from a gray-level co-occurrence matrix (GLCM). Both the brightness and texture features are combined as the input of an SVR to obtain the final quality score. Compared with the state-of-the-art BIQA methods, our proposed BNBT metric achieves much higher consistency with human visual perception when judging the quality of night-time images.

Our contributions can be summarized as follows:

- To the best of our knowledge, we build the first dedicated large-scale natural night-time image database (NNID), which contains 2240 images with 448 different image contents captured by different photographic equipment in real-world scenarios. For each image content, we use one device with five different settings to capture five images of different visual qualities. The five settings are different for different image contents. We also make this database publicly available for downloading at https://sites.google.com/site/xiangtaooo/.
- We conduct comprehensive subjective assessment of the images in the NNID database by adopting the single stimulus (SS) method recommended by the International Telecommunications Union (ITU). A ground truth, i.e., mean opinion score (MOS), is obtained for each image via carefully designed experiments and postprocessing procedures.
- We conduct objective assessment of night-time images by proposing a blind night-time image quality assessment metric using brightness and texture features (BNBT). Unlike traditional methods that usually extract brightness features directly from nonoverlapped image blocks, we extract brightness features from superpixel segments of the image for better consistency with visual perception. We also extract brightness features from both luminance and chrominance channels to make the proposed BNBT more sensitive to the distortions in night-time images.
- We conduct extensive experiments to evaluate the performance of our proposed BNBT metric on the NNID database, and compare this metric with many state-of-the-art general-purpose and contrast-distorted BIQA metrics. The results show that the BNBT metric outperforms the existing metrics, and it has acceptable computational cost at the same time.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III introduces the construction of the NNID database and its subjective testing methodology. Section IV describes the details of our proposed BNBT metric. Section V provides experimental results and their analysis. Finally, Section VI concludes the paper.

II. Related Work

In this section, we review existing IQA databases and some state-of-the-art BIQA metrics.

A. IQA Databases

One important task of IQA studies is constructing a publicly available IQA database. There are two kinds of IQA databases. One is simulated-distorted image databases, such as LIVE [12], CSIQ [13], TID2013 [14] and MLIVE [15], which are generated by introducing graded simulated distortions into high-quality photographs. The distortion process is controlled by image processing techniques. The other is naturally-distorted image databases, such as CID2013 [16], CCRIQ [17], [18], CLIVE [19] and BID [20], which contain realistic distortions generated

during the process of image acquisition, processing and storage. The latter database is more difficult to construct than the former because various naturally distorted images must be manually captured.

Although these databases are often used in IQA and can achieve satisfactory results, they are not suitable for night-time images. Specifically, the images in simulated-distorted image databases are generated by image processing techniques instead of directly coming from cameras. Moreover, the distortions of images in existing naturally distorted image databases are different from the distortions of night-time images. Therefore, constructing a specific night-time image database is a must for night-time image quality assessment research.

B. BIQA Metrics

Objective IQA is classified into full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and blind IQA (BIQA). BIQA is most convenient in practice because it requires no knowledge about reference images. The existing BIQA metrics are usually designed based on two different considerations: different image distortions or different image contents.

1) BIQA for Image Distortions: The existing BIQA metrics designed for image distortions can be divided into two categories: distortion-specific BIQA and general-purpose BIQA.

The distortion-specific BIQA methods have prior knowledge of specific image distortion. Common types of distortions include compression [21], noise distortion [22], blur distortion [23], contrast distortion [24]–[26], etc. In [21], the quality of JPEG2000 (JP2K) images is evaluated by using pixel distortions and edge information in the discrete cosine transform (DCT) domain of general image blocks. Yang *et al.* [22] proposed a BIQA metric for noise-distorted images based on frequency mapping, where the metric measures the similarity between distorted images and denoised images. The gradient is widely used in image sharpness assessment, and Zhang *et al.* [23] employed the maximum gradient and the variability of gradients to predict the quality of blurry images. Fang *et al.* [24] presented a BIQA method for evaluating the quality of contrast-distorted images based on moment and entropy features. The moment and entropy features are extracted from a large-scale image database to build a natural scene statistics (NSS) model, and then, the unnaturalness of a test image compared with the NSS model is calculated as the quality index. There are also some application-specific distortions considered in BIQA, such as tone mapping [27], [28], compressive sensing (CS) [29] and enhancement [30].

In contrast, general-purpose BIQA approaches do not require any prior knowledge of distortion types. Most of the existing general-purpose approaches [3]–[9] require a training process. BRISQUE [3] measures the unnaturalness of an image based on the NSS of locally normalized luminance coefficients in the spatial domain. BLIINDS-II [4] predicts the quality via training a probabilistic model using the contrast and structure features extracted in the DCT domain. GM-LOG [5] extracts histogram-based statistical properties from the joint statistics of the image gradient and Laplacian of Gaussian response. NFERM [7] uses important human visual system (HVS)-inspired features, such as structural information, gradient magnitude and possible losses of naturalness, to predict the quality of a distorted image. MSGF [6] addresses the quality assessment problem by introducing multidomain structural information and piecewise regression. NRSL [8] employs the statistical structural feature and luminance feature to construct opinion-aware features. Structural information is characterized by the spatial distribution of a local binary pattern (LBP) based on a normalized luminance map, while the distribution of normalized luminance magnitudes is extracted as a luminance feature. Liu *et al.* [9] proposed a method to deal with more distortion types and image contents by extracting a rich diversity of features from wide perceptual domains to train a prediction model (scorer). There are also two general-purpose BIQA methods that do not require training samples and subjective scores to learn a predictive model [10], [11].

Using popular image databases, e.g., LIVE, CSIQ and TID2013, the majority of general-purpose BIQA models described above have been proven to be of fairly high performance in accordance with subjective assessment. However, many excellent general-purpose BIQA metrics exhibit poor performance when used for night-time image visual quality estimation. For example, Figs. 1(a)-1(d) show the performance of the BLIINDS-II, GM-LOG, NFERM and MSGF methods on several images from the NNID database. Compared to Fig. 1(a), it is clear that less intelligible information can be obtained from Fig. 1(b), but this image is found to have better visual quality as assessed by BLIINDS-II and GM-LOG. Similarly, Fig. 1(d) has poor visual quality compared to Fig. 1(c), but has better NFERM and MSGF values.

2) BIQA for Image Contents: Images with different contents are portrayed by different characteristics, and they require different IQA metrics because traditional IQA metrics may not be well applicable in this situation. Several BIQA metrics have been proposed for specific image contents, such as screen content images [31]–[33], medical images [34], [35] and camera images [36].

Unlike natural scene images captured by modern high-fidelity cameras, screen content images are typically composed of fewer colors, simpler shapes, and a larger frequency of thin lines. Gu *et al.* [32] developed a BIQA model for screen content images by extracting four types of features to describe image complexity, screen content statistics, global brightness quality, and the sharpness of details. Shao *et al.* [37] proposed a blind quality predictor for SCI (BLIQUP-SCI) from the perspective of sparse representation. Inspired by the perceptual property of the HVS, which is sensitive to luminance change and texture information for image perception, Fang *et al.* [33] proposed a blind quality assessment method by incorporating statistical luminance and texture features for screen content images with both local and global feature representation. Medical images are usually generated by special imaging techniques such as magnetic resonance imaging (MRI), computed tomography (CT), and single-photon emission computerized tomography (SPECT). Each of them has its own specific features and needs a specific IQA metric to evaluate their quality. Nakhaie *et al.* [38] presented a watermarking method based on region of interest (ROI) as a BIQA metric for medical images, where the peak signal-to-noise ratio (PSNR)

TABLE I
IMAGE INFORMATION OF NNID

Size	Device-I (Nikon D5300)	Device-II (iPhone 8plus)	Device-III (iPad mini2)
512×512	800	340	200
1024×1024	300	150	-
2048×2048	300	150	-

and mean square error (MSE) of the extracted watermark were used to evaluate the quality of original medical images. Kalayeh *et al.* [39] proposed two new numerical observer models based on Bayesian learning regression, the channelized relevance vector machine (CRVM) and the multi-kernel channelized relevance vector machine (MKCRVM), to predict the quality of SPECT images. Saad *et al.* [36] proposed VIQET to evaluate the quality of consumer photos with realistic distortions and realistic quality ranges. VIQET utilizes NSS modeling and the consumer-centric, quality aware interpretable features for image quality prediction.

III. NNID DATABASE

In this section, we describe the construction of our night-time image database and the subjective evaluation of images to obtain their subjective quality scores.

A. Construction

To investigate the visual quality of night-time images, we construct a dedicated large-scale natural night-time image database (NNID). The images in the NNID database cover many night-time scenes that often appear in daily life, including humans, buildings, natural scenery, traffic signs, etc. We employ a digital camera (Nikon D5300), a mobile phone (iPhone 8plus) and a tablet (iPad mini2) to capture images in the night environment. In subsequent statements, we use Device-I, Device-II and Device-III to represent these three devices.

NNID contains 2240 night-time images with 448 different image contents. For each image content, we use one device with five different settings to capture five images of different visual qualities. Specifically, we fix the shooting position, shooting angle, depth of field, and focal length while randomly changing the exposure, aperture, shutter, ISO, etc. The five settings are different for different image contents. In NNID, 1400 images with 280 different image contents are captured by Device-I, 640 images with 128 different image contents are captured by Device-II, and 200 images with 40 different image contents are captured by Device-III. The resolutions of the images in NNID include 512 × 512, 1024 × 1024, and 2048 × 2048. We obtain the desired size by image cropping to keep the properties of the original images. To include more real-world night scenarios, the image contents of different resolutions do not overlap. Table I shows the statistical information about the images in NNID. Fig. 2 demonstrates 18 night-time images from NNID. The images in the first, second and third rows are captured from Device-I, Device-II and Device-III, respectively.

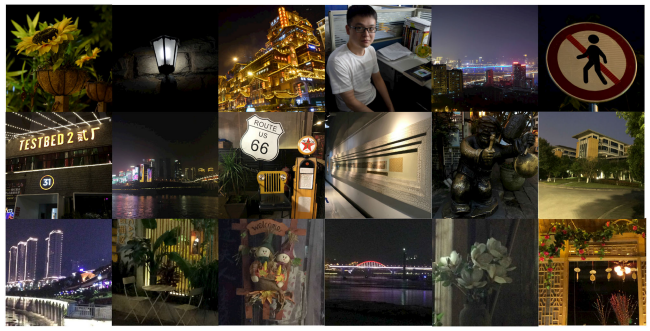


Fig. 2. Sample images in NNID. The images in the first, second and third rows are captured from Device-I, Device-II and Device-III, respectively.

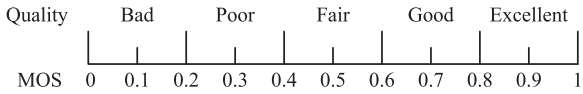


Fig. 3. Rating scales and quality levels.

B. Subjective Assessment Methodology

Subjective assessment methodologies of image quality evaluation have been recommended by the International Telecommunications Union (ITU) [40], including single stimulus (SS), double stimulus and paired comparison. Because the SS methodology is close to the viewing experience in a dark scenario where there is no access to a distortion-free and perfect-quality reference image, in our study, experiments are conducted using an SS method with 11 discrete rating scales from 0 to 1, which correspond to five quality levels: bad, poor, fair, good and excellent. The relation between rating scales and quality levels is shown in Fig. 3. Given an image displayed on the screen, an observer is asked to give a rating score on the image quality based on her/his visual perception. During the rating process, each image is presented in a random order to reduce the memory effect on opinion scores.

In the subjective test, the images in NNID are randomly divided into two subsets of equal size. We invite 74 observers, including 33 females and 41 males, for our test. Half of the observers participate in the test of one subset, and the other half participate in the test of the other subset. All of these observers are college students majoring in art, science, engineering, and business; none of them has background knowledge of image processing. They all have normal or corrected vision, aged from 20 to 30 years old.

We design an interactive system to automatically display test images and collect subjective opinion scores using a graphical user interface (GUI). The interface for rating image quality is shown in Fig. 4. The viewing conditions in our test are in accordance with the ITU Recommendation [40]. The subjective experience is conducted using two monitors in the laboratory with a normal indoor light condition. One monitor is configured for HD resolution (1920 × 1080), and the other is configured for 4 K resolution (3840 × 2160). All the images are displayed in their original resolutions so that no extra distortions are introduced. The incident light falling on the screen is 80 lux, and the

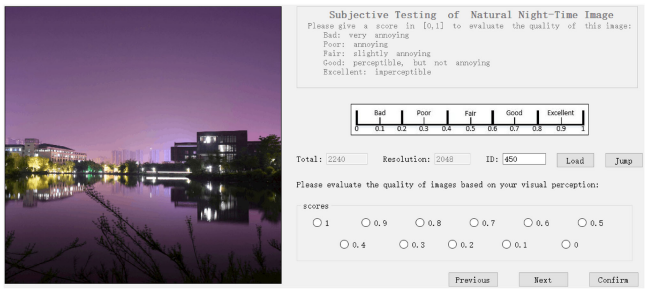


Fig. 4. Graphical user interface for subjective test.

TABLE II
SUBJECTIVE TEST SETUP

Method	Single stimulus (SS)
Evaluation Scales	11 discrete scale from 0 to 1
Subjects	74 inexperienced observers
Age Range	20-30 years old
Image Resolution	512×512, 1024×1024, 2048×2048
LCD Monitor	1920×1280, 3840×2160
Light Condition	normal indoor lighting

environmental illumination from behind the monitor is 240 lux. All observers are required to sit at a viewing distance of four times the image height. Table II summarizes major information about the test conditions and parameter configuration.

Before starting the testing, a training session is presented to all the observers. In the training session, some examples with representative quality levels are presented. These examples are not included in the testing stage. The observers are told to assess their quality by mainly taking three aspects into consideration: content recognizability, clarity, and viewing comfort [31]. Content recognizability is used to check whether the content of a natural night-time image can be recognized. Content clarity is used to judge the impairment appearance of an image. Viewing comfort reflects the viewing experience of an observer.

In the testing stage, given an image displayed on the screen, the observers are asked to rate its quality and provide a score. After an observer clicks the confirm button, as shown in Fig 4, the subjective score is automatically saved. Only after the observer rates all the images in the NNID database will the subjective test be finished, and a record file for each observer will be generated and saved automatically. On average, it takes approximately two hours for each observer to finish all the ratings in our experiments. During the test, the observers are required to take a break every 30 minutes to avoid fatigue. The observers are also able to take a break if they feel they need one within 30 minutes.

C. Processing of Raw Data

After collecting the raw scores rated by all the observers, we further process them to obtain the final mean opinion score (MOS) for each night-time image. The processing includes outlier detection and observer rejection, and a simple yet efficient method is adopted here, which is also used in [31], [41], [42].

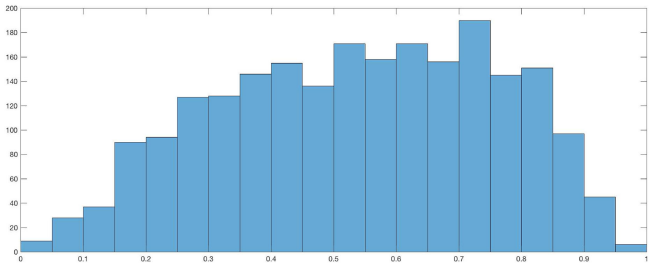


Fig. 5. Histogram of MOS values in NNID.

In outlier detection, a raw rating score is considered to be an outlier if it is outside its confidence interval. Specifically, for each image, the average value of subjective rating scores is first computed by

$$\overline{\mu_i} = \frac{1}{N}\sum_{j=1}^{N}\mu_{i,j} \tag{1}$$

where $\mu_{i,j}$ is the subjective rating score of image i from observer j, and N is the total number of observers. Then, the standard deviation is computed by

$$\sigma_i = \sqrt{\sum_{j=1}^{N}\frac{(\mu_{i,j}-\overline{\mu_i})^2}{(N-1)}} \tag{2}$$

The confidence interval of image i is defined as $[\overline{\mu_i}-\delta_i, \overline{\mu_i}+\delta_i]$ where $\delta_i = 1.96 * \frac{\sigma_i}{\sqrt{N}}$ when the confidence interval is set to 95% [40].

In observer rejection, all quality ratings of an observer are rejected if the number of his/her discarded rating scores exceeds a threshold R in outlier detection. R is determined by

$$R = OC \times N_I \tag{3}$$

where N_I denotes the number of images rated by each observer, and OC is the outlier coefficient defined below, which quantifies the subjective agreement of the database:

$$OC = \frac{N_O}{N_T} \tag{4}$$

where N_O denotes the number of outliers of all the observers, and N_T denotes the total number of rating scores of all the observers. By this calculation, $R = 16$ in our experiments, and four observers are rejected.

D. Analysis of MOS

Generally, the MOS values of images in the NNID database should exhibit good separation of perceptual quality and cover the entire range of visual quality (from bad to excellent) [31] so that they can be regarded as the ground truth for performance evaluation of objective quality metrics.

In Fig. 5, we demonstrate the histogram distribution of all the MOS values in the NNID database. We can observe that the MOS values span the entire visual quality scale [0, 1] and have a good spread at different visual levels. In addition, the distribution of MOS values indicates the diversity of images in the constructed NNID database. Moreover, we evaluate the

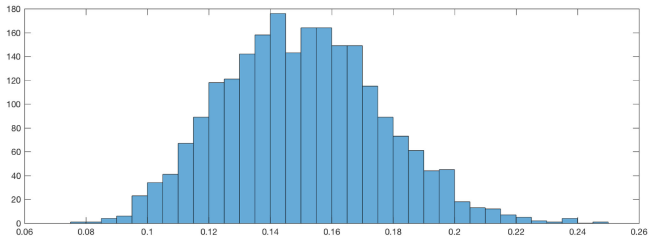


Fig. 6. Histogram of confidence intervals in NNID.

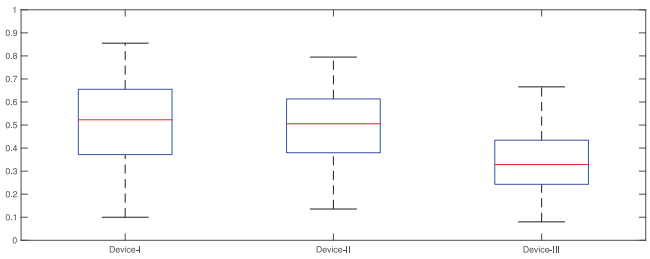


Fig. 7. Box plot of MOS values of images captured by different devices.

consistency among all the observers, which is measured by the 95% confidence interval derived from the mean and standard deviation of rating scores for each image [40]. The distribution of the confidence intervals is shown in Fig. 6. It is clear that the confidence intervals mainly distribute between 0.12 and 0.18. This is useful evidence indicating high agreement among all the observers in evaluating the visual quality of night-time images in the NNID database.

Moreover, Fig. 7 shows the box plot of MOS values of images captured by different devices. The MOS values of images captured by Device-I and Device-II span most of the range of [0,1], and those captured by Device-III range in [0,0.6]. This result occurs because Device-I and Device-II are high-end devices, but Device-III is a low-end camera, and its captured images have relatively poor visual quality with excessive noise. This observation validates the results of our subjective test and their post-analysis.

IV. The Proposed BNBT Metric

In this section, we propose our blind night-time image quality assessment metric based on brightness and texture features (BNBT). First, we introduce the basic idea and framework of the proposed scheme. Then, we describe the three main parts of our proposed BNBT in detail, i.e., brightness feature extraction, texture feature extraction, and prediction model regression.

A. Basic Idea

The basic concept of our method is to measure the quality degree of night-time images by extracting visual features that are critical for visualization and may still be preserved in a weakly illuminated environment. Because night-time images usually exhibit low contrast, blurred details and reduced visibility, their visual quality evaluation metric should have the ability to measure the changes in visual features between night-time images of different qualities. Based on the psychovisual properties of the HVS, brightness and texture are the most two important visual features for night-time images. First, brightness information is of primary importance in visual perception, especially for weakly illuminated night-time images. For example, in night-time image acquisition, a too short or too long exposure time may lead to darker or brighter images; the brightness determines the basic quality level of the image. Second, texture information is also significant for the visual quality of night-time images because it contains the details of visual contents. Given the same brightness information, the texture information serves as an additional indicator of distortion strength. Therefore, by combing brightness and texture information, we can assess the visual quality of night-time images.

B. Framework

After introducing our basic idea, we now propose a blind night-time image quality assessment metric based on brightness and texture features (BNBT). The framework of our proposed BNBT is illustrated in Fig. 8. In brief, our scheme works as follows. First, we extract the brightness information of a night-time image by computing the mean value over superpixels from a perceptual color space and obtain carefully designed weighted brightness features. Then, we extract the texture features of the night-time image using the gray-level co-occurrence matrix (GLCM). Both brightness and texture features are calculated at two scales to reduce the variation in viewing distance and image resolution. Finally, we combine the brightness and texture features into support vector regression (SVR) to simulate the nonlinear mapping from feature space to an objective quality score. In the subsequent statements of this section, we will describe the detail of these three procedures.

C. Brightness Feature Extraction

We first extract the brightness feature of a night-time image because it is of primary importance in visual perception for weakly illuminated night-time images.

Traditional IQA methods usually divide images into nonoverlapped blocks and extract corresponding visual features from these blocks [25], [27]. This approach is convenient for feature computation, but it is not consistent with visual perception. Image contents are usually contained in different regions and have different impacts on visual quality, but the shapes of these regions are usually not rectangles and cannot be exactly covered by blocks; therefore, image blocks and their extracted features are usually meaningless for visual perception.

To solve this problem, we extract the brightness features on superpixels [43] that group perceptually similar pixels to create visually meaningful regions. Specifically, a superpixel is composed of spatially neighboring pixels sharing similar low-level features such as colors, intensities or structures. Regarding this property, the superpixel has received much attention and has been widely used in many vision tasks such as image decomposition [44], multisensory video fusion [45], and image synthesis [46].

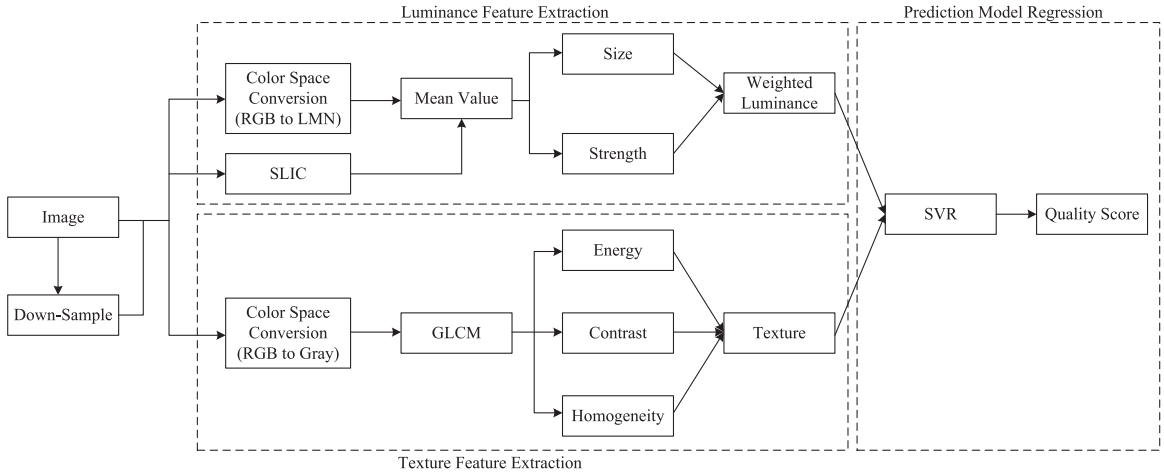


Fig. 8. Framework of the proposed BNBT method.

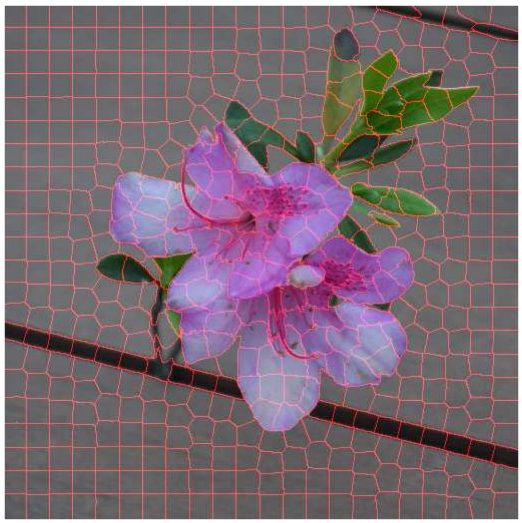


Fig. 9. Illustration of SLIC segmentation.

A computationally efficient superpixel algorithm, simple linear iterative clustering (SLIC) [47], is used to segment superpixels in our study. SLIC is an adaptation of k-means clustering and can show image boundary adherence. SLIC first initializes a number of cluster centers (N_c), and each pixel is assigned to its nearest center according to a predefined distance measure. Then, each cluster center is updated by the mean attribute of its corresponding elements. Finally, superpixels are generated by repeating the above steps. We present an example of superpixels generated by the SLIC algorithm in Fig. 9 to show the peculiarities of image regions with superpixels, where the number of cluster centers $N_c = 400$.

Because the brightness feature extracted from the luminance channel of images cannot well characterize color distortions, we consider both luminance and chrominance channels. After we acquire the superpixel segments of a night-time image by the SLIC algorithm, we then convert the RGB color space of the image into LMN color space [48] to compute its brightness features. The LMN color space is well-defined on a physical basis and optimized on the HVS [49], and the conversion can remove the correlation between the luminance component (L) and the chrominance components (M, N). The conversion can be formulated as

$$\begin{bmatrix} L \\ M \\ N \end{bmatrix} = \begin{bmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5}$$

After color space conversion, we compute the brightness feature on each channel L, M, and N. Without loss of generality, we take the L channel as an example. Suppose that the SLIC method divides a night-time image into η superpixels; we first calculate the mean intensity of each superpixel by

$$m_{L,i} = \frac{1}{|S_i|} \sum_{(x,y)\in S_i} I_L(x,y), i = 1, 2, \ldots, \eta \tag{6}$$

where $I_L(x, y)$ denotes the pixel value at coordinate (x, y) on channel L, S_i denotes the set of coordinates of the i-th superpixel, and $|S_i|$ is the number of elements in S_i.

We then propose to use a ranking-based weighting function to weight the brightness strength. The mean value of each superpixel is used to generate the weight. Specifically, the mean value $m_{L,i}$ of superpixels is sorted in ascending order. Suppose the order index after sorting is $Rank_i$; then, the weight of a superpixel is defined as

$$w_{L,i} = log_2 \left(1 + \frac{Rank_i}{\eta}\right), i = 1, 2, \ldots, \eta \tag{7}$$

Since $Rank_i \in 1, 2, \ldots, \eta$, it is easy to know that $w_i \in [0, 1]$, and a larger mean value of a superpixel corresponds to a larger weight.

Finally, the brightness feature on channel L is obtained by

$$f_L = \frac{\sum_{i=1}^{\eta} w_{L,i} m_{L,i}}{\sum_{i=1}^{\eta} w_{L,i}} \tag{8}$$

Similarly, we can obtain the other two brightness features f_M and f_N on channels M and N.

D. Texture Feature Extraction

In addition to the brightness feature, we also consider the texture feature because it is an indispensable element for human visual perception and is necessary for predicting the quality of night-time images. Night-time images are under the constraint of illumination, and they are usually characterized by low contrast. In this case, the visual perception of image details, which can be described by textures, has a close relation to the image quality. Therefore, it is reasonable to consider texture features when assessing the quality of night-time images.

The texture features are extracted by utilizing the gray-level co-occurrence matrix (GLCM) [50]. GLCM is a well-known method for describing textures by measuring the specific spatial correlation properties between pixels within a certain distance. Given an image I with size $m \times n$, the GLCM can be defined as follows:

$$GLCM(i,j) = \frac{1}{mn} \sum_x \sum_y \begin{cases} 1, & \text{if } I(x,y) = i \text{ and} \\ & I(x+\Delta x, y+\Delta y) = j \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where (x, y) is the spatial coordinates in I and $(\Delta x, \Delta y)$ is the offset.

We compute the energy f_e, the contrast f_c, and the homogeneity f_h of the GLCM as follows to describe the texture features of an image.

$$f_e = \sum_i \sum_j GLCM(i,j)^2 \tag{10}$$

$$f_c = \sum_i \sum_j (i-j)^2 GLCM(i,j) \tag{11}$$

$$f_h = \sum_i \sum_j \frac{1}{1+(i-j)^2} GLCM(i,j) \tag{12}$$

where $f_e, f_c, f_h \in [0, 1]$. The energy measures the amount of information in an image and indicates the complexity of the image, and $f_e = 1$ indicates a constant image. The contrast measures the intensity contrast between a pixel and its neighbors over the whole image. The homogeneity measures the spatial closeness to the diagonal of the distribution of the elements of the GLCM, and it reaches its maximum when the GLCM is a diagonal matrix.

Then, the texture features are obtained by computing the mean and standard deviation of the energy ($f_{\bar{e}}, f_{\delta_e}$), contrast ($f_{\bar{c}}, f_{\delta_c}$) and homogeneity ($f_{\bar{h}}, f_{\delta_h}$) features. In total, six texture features can be extracted from a scale of a night-time image.

E. Prediction Model Regression

After we extract brightness and texture features from two scales of an image, we adopt support vector regression (SVR) [51] to learn the mapping function from the feature space to the value of the quality measure for feature pooling. Please note that traditional deep learning methods are not appropriate for our work because they require considerable data samples for training, but we only extract 18 features here (9 features, $f_L, f_M, f_N, f_{\bar{e}}, f_{\delta_e}, f_{\bar{c}}, f_{\delta_c}, f_{\bar{h}}$, and f_{δ_h}, at each scale).

Consider a set of training data $D = \{(x_i, y_i) | i = 1, \ldots, r\}$, where r is the number of images, $x_i \in R^{18}$ is the extracted features, and y_i is the corresponding MOS. Suppose the parameters are $t > 0$ and $p > 0$; we can achieve the standard form of SVR as

$$\begin{aligned} \underset{\mathbf{w},\delta,\mathbf{b},\mathbf{b}'}{minimize} & \frac{1}{2} \|\mathbf{w}\|_2^2 + t \sum_{i=1}^{r} (b_i + b_i') \\ \text{s.t. } & \mathbf{w}^T \phi(x_i) + \delta - y_i \le p + b_i \\ & y_i - \mathbf{w}^T \phi(x_i) - \delta \le p + b_i' \\ & b_i, b_i' \ge 0, i = 1, \ldots, r \end{aligned} \tag{13}$$

where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel function. We use the radial basis function (RBF) kernel, which is defined as $K(x_i, x_j) = exp(-k \|x_i - x_j\|^2)$ in this work. Based on the training samples, our target is to determine the parameters t, p, and k and thus find the associated regression model.

V. Experimental Results

In this section, we conduct experiments to evaluate the performance of our proposed BNBT metric on the NNID database and compare it with several state-of-the-art methods. Considering that no camera exposure setting can result in a distortion-free and perfect quality image to serve as a reference image in a relatively dark scenario, only existing BIQA approaches are involved in our experiments.

A. Experiment Protocol

1) BIQA Methods: We investigate the performance of our proposed BNBT and other existing BIQA methods on our constructed NNID database. We compare our proposed BNBT metric with 14 existing BIQA methods, including 11 general-purpose and 3 contrast-distorted BIQA methods. The 11 general-purpose BIQA methods are BRISQUE [3], BLIINDS-II [4], NIQE [10], GM-LOG [5], ILNIQE [11], MSGF [6], GWH-GLBP [52], SSEQ [53], NFERM [7], NRSL [8], and VIQET [36]. The three contrast-distorted BIQA methods NR-CDIQA [24], NR-SPL [25] and NIQMC [26] are also considered because night-time images in the NNID database usually have low contrast. NR-CDIQA [24] sets up an NSS model based on five features extracted from the SUN2012 database [54] and subsequently uses SVR to predict the perceived quality of contrast-distorted images. NR-SPL [25] extracts human perception features, including the perceptual contrast of the image and the skewness and variance of the intensity distribution histogram. Then, a backpropagation (BP) network is used to find the mapping function between the feature set and visual quality scores. NIQMC [26] performs the IQA task for contrast-distorted images by using two types of features based on the concept of maximum information. The first type involves computing the entropy of the salient region, and the second type involves measuring the Kullback-Leibler divergence between

the image histogram and the uniformly distributed histogram of maximum information.

2) Database: All experiments are conducted on our proposed NNID database. To further check the performance of our proposed BNBT, we also conduct comparative experiments on the 44 night-time images in the CCRIQ [17], [18] database and the 79 low-contrast images in the CID2013 [16] database. Since all the involved methods except NIQE [10] and ILNIQE [11] require a training stage to calibrate regression models, we randomly divide the NNID database into two nonoverlapping subsets based on image content, a training set and a test set, such that the image contents of these two sets do not overlap. The training set consists of 80% of the images in the NNID database, and the test set consists of the remaining 20% images. Furthermore, to eliminate the performance bias due to a specific training-test split, we repeat the random splitting evaluation 1000 times on the NNID database, and the medians of the results across these 1000 iterations are reported.

3) Evaluation Criteria: As suggested by the Video Quality Experts Group (VQEG) Phase I FR-TV test [55], four commonly used performance evaluation criteria are employed to evaluate the performance of different BIQA metrics. They are the Pearson linear correlation coefficient (PLCC), the Spearman rank order correlation coefficient (SRCC), Kendall's rank-order correlation coefficient (KRCC), and the root mean-squared error (RMSE). SRCC and KRCC evaluate the prediction monotonicity, while PLCC and RMSE evaluate the prediction accuracy. To compute PLCC and RMSE, a five-parameter nonlinear fitting function is adopted to map predicted scores to subjective scores [55], whose coefficients are solved via the iterative least squares estimation [56]:

$$f(x) = \tau_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\tau_2 (x - \tau_3)}} \right) + \tau_4 x + \tau_5 \qquad (14)$$

where x denotes the predicted score, and $f(x)$ represents the corresponding subjective score. $\tau_i (i = 1, 2, \ldots, 5)$ are the parameters to be fitted. A better objective quality measure is expected to achieve a value closer to 1 for PLCC, SRCC and KRCC but closer to 0 for RMSE, indicating a superior correlation between subjective scores and predicted scores.

B. Evaluation of Features

Because both brightness features and texture features are used in our proposed BNBT metric, to validate the superiority of our proposal, we first conduct experiments to analyze the contribution of each feature component and their combination to the performance on the constructed NNID database. It is worth noting that the performance of each feature component is obtained after training the SVR model using the corresponding subset of features. For example, only the extracted brightness features are used to train the SVR model when evaluating the contribution of brightness features alone.

To identify how well the features correlate with the perceptual quality of human judgment, we list the results of PLCC, SRCC, KRCC and RMSE in Table III. The best results in terms of each evaluation criterion are marked in bold. It is observed that each feature component contributes to the overall performance, and the brightness/texture feature component itself has very encouraging results in terms of all performance criteria. More precisely, the weighted brightness features are more effective than the texture features, indicating that brightness and contrast are the most important distortions in natural night-time images. Furthermore, the combination of brightness and texture features produces much better performance than a single feature component for all evaluation criteria. This result validates the superiority of the feature combination in our proposed BNBT metric.

TABLE III
PERFORMANCE OF DIFFERENT FEATURES ON THE NNID DATABASE

Feature	SRCC	KRCC	PLCC	RMSE
Brightness	0.8536	0.6636	0.8598	0.1115
Texture	0.8343	0.6400	0.8364	0.1198
Brightness & Texture	**0.8769**	**0.6822**	**0.8784**	**0.1061**

C. Comparative Analysis

We now compare the performance of our proposed BNBT metric with existing general-purpose and contrast-distorted BIQA methods on the NNID database and subsets of the CCRIQ and CID2013 databases. To investigate the effect of different camera devices, we perform the training-testing process on the subsets of images captured by three devices (Device-I, Device-II and Device-III) and the entire database.

1) Comparison With General-Purpose BIQA Metrics: Table IV illustrates the performance results of our proposed BNBT metric and the other existing 11 general-purpose BIQA methods. We can make the following two observations. First, from Table IV, we can see that the results of all metrics on the images captured by Device-I, Device-II, Device-III and the entire NNID database are relatively stable, and no significant difference can be found. Therefore, the effect of different equipment used for acquiring night-time images on performance is negligible, which further proves the feasibility of using the NNID database for night-time image quality assessment. Second, Table IV shows that compared with the existing general-purpose BIQA methods, the proposed BNBT metric has the best performance on the NNID database because it has the highest PLCC, SRCC and KRCC values and the smallest RMSE value. In other words, the proposed BNBT method correlates highly with human visual perception of the quality of night-time images. We also show the scatter plots of the subjective scores (MOS) versus the predicted quality scores of different metrics in Fig. 10, where a point denotes a night-time image in the NNID database. A good metric is expected to produce scatter points that are close to the fitted curve. It can be easily found from Fig. 10 that the proposed metric produces the best fitting result on the NNID database.

Table V tabulates the results of our proposed BNBT metric and the other existing 11 general-purpose BIQA methods on the subsets of the CCRIQ and CID2013 databases. From Table V, we can find that the proposed BNBT has the best performance on the 44 night-time images of the CCRIQ database. In addition,

TABLE IV
PERFORMANCE COMPARISON OF BNBT AND EXISTING GENERAL-PURPOSE BIQA METRICS ON THE NNID DATABASE

Metric	Entire database (2240)				Device-I (1400)				Device-II (640)				Device-III (200)			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
BRISQUE [3]	0.7445	0.5473	0.7600	0.1416	0.7471	0.5531	0.7633	0.1467	0.7663	0.5678	0.7788	0.1402	0.7406	0.5632	0.7853	0.1387
BLIINDS-II [4]	0.7511	0.5500	0.7580	0.1415	0.7823	0.5820	0.7967	0.1376	0.7478	0.5545	0.7823	0.1531	0.7437	0.5597	0.7676	0.1451
NIQE [10]	0.5983	0.4220	0.5701	0.1803	0.6007	0.4240	0.5859	0.1847	0.5772	0.4017	0.5874	0.1811	0.6591	0.4694	0.6092	0.1842
ILNIQE [11]	0.7115	0.5183	0.6335	0.1691	0.6712	0.4831	0.6766	0.1679	0.6949	0.5018	0.6809	0.1639	0.7983	0.6086	0.6721	0.1720
SSEQ [53]	0.7675	0.5707	0.7718	0.1385	0.8126	0.6171	0.8213	0.1300	0.7534	0.5637	0.7731	0.1407	0.7057	0.5268	0.7469	0.1505
GM-LOG [5]	0.8180	0.6240	0.8264	0.1228	0.8359	0.6447	0.8496	0.1198	0.7968	0.6065	0.8154	0.1285	0.7912	0.6077	0.8198	0.1297
MSGF [6]	0.8571	0.6710	0.8603	0.1096	0.8768	0.7055	0.8924	0.1025	0.8353	0.6411	0.8407	0.1201	0.8310	0.6513	0.8628	0.1150
GWH-GLBP [52]	0.7721	0.5762	0.7862	0.1350	0.7795	0.5802	0.7948	0.1382	0.6517	0.4640	0.6790	0.1614	0.7150	0.5807	0.8019	0.1343
NFERM [7]	0.8596	0.6689	0.8637	0.1099	0.8692	0.6801	0.8732	0.1110	0.8096	0.6169	0.8248	0.1257	0.8115	0.6201	0.8371	0.1231
NRSL [8]	0.8412	0.6482	0.8470	0.1159	0.8575	0.6675	0.8659	0.1134	0.7904	0.5980	0.8091	0.1305	0.7619	0.5760	0.7981	0.1364
VIQET [36]	0.7486	0.5664	0.7203	0.1566	0.7476	0.5564	0.7359	0.1489	0.7463	0.5727	0.7457	0.1492	0.7515	0.5797	0.7472	0.1544
BNBT	**0.8769**	**0.6822**	**0.8784**	**0.1061**	**0.8866**	**0.7066**	**0.8939**	**0.1020**	**0.8632**	**0.6737**	**0.8698**	**0.1157**	**0.8517**	**0.6890**	**0.8576**	**0.1137**

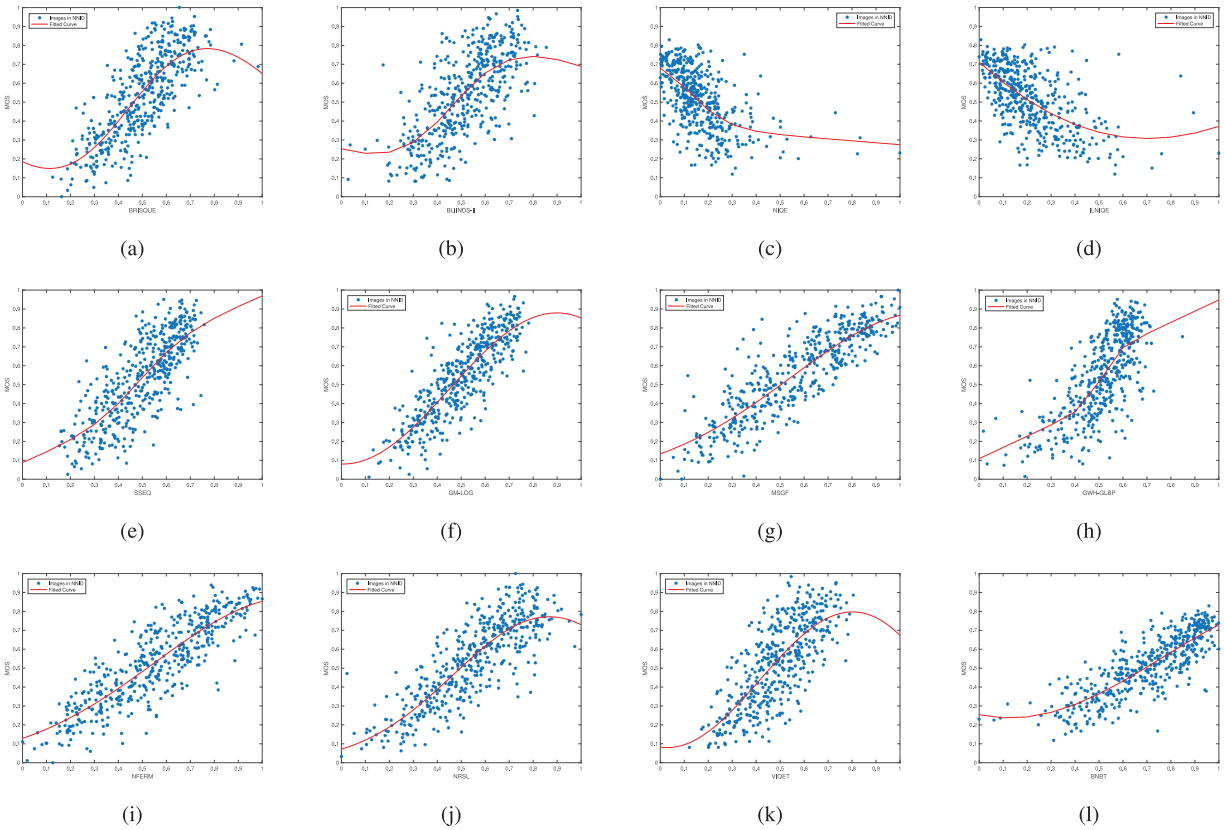


Fig. 10. Scatter plots of the performance of BNBT and general-purpose BIQA metrics on the NNID database. (a) BRISQUE [3], (b) BLIINDS-II [4], (c) NIQE [10], (d) ILNIQE [11], (e) SSEQ [53], (f) GM-LOG [5], (g) MSGF [6], (h) GWH-GLBP [52], (i) NFERM [7], (j) NRSL [8], (k) VIQET [36], and (l) BNBT.

BNBT has the second best (close to the best) performance on the 79 low-contrast images of the CID2013 database. These results demonstrate that our proposed scheme is effective for the quality assessment of night-time images, as well as for other images captured in dark lighting conditions.

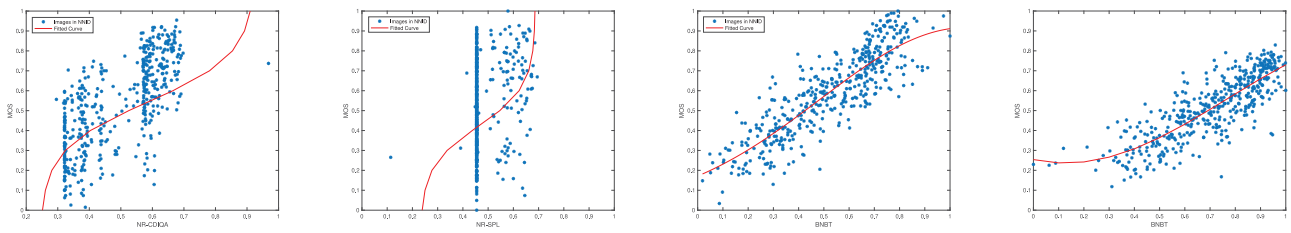
2) Comparison With Contrast-Distorted BIQA Metrics: Because night-time images in NNID are characterized by low contrast, contrast distortion is an important factor affecting the visual quality of night-time images. In this part, we compare the performance of the proposed BNBT metric with three prevalent contrast-distorted BIQA methods to verify the robustness of the proposed BNBT. Table VI lists the performance of the comparison methods, and Fig. 11 shows the scatter plots between the predicted scores and the corresponding MOS values, where a point denotes one night-time image in the NNID database. It can be clearly observed from both Table VI and Fig. 11 that the proposed BNBT metric significantly outperforms all the considered contrast-distorted BIQA methods in terms of every evaluation criterion. This result indicates that traditional contrast-distorted BIQA methods are inappropriate for night-time images. The reason for this result is that although low contrast is an important characteristic of night-time images, other characteristics, such

TABLE V
PERFORMANCE COMPARISON OF BNBT AND GENERAL-PURPOSE BIQA METRICS ON SUBSETS OF THE CCRIQ AND CID2013 DATABASES

Metric	CCRIQ Subset (44)				CID2013 Subset (79)			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
BRISQUE [3]	0.6091	0.4547	0.7434	0.4343	0.3802	0.2674	0.5334	14.0366
BLIINDS-II [4]	0.4665	0.3487	0.6027	0.5182	0.5725	0.4386	0.6708	12.3067
NIQE [10]	0.4763	0.3275	0.4033	0.5469	0.6910	0.5058	0.7323	12.1892
ILNIQE [11]	0.5536	0.3847	0.4713	0.5927	0.5624	0.3946	0.6755	14.1954
SSEQ [53]	0.6237	0.4229	0.6507	0.4935	0.5941	0.4239	0.6082	11.6468
GM-LOG [5]	0.5198	0.3741	0.5245	0.5529	0.3086	0.2408	0.5528	13.8285
MSGF [6]	0.3724	0.2788	0.3081	0.5910	0.2717	0.1848	0.1496	15.5023
GWH-GLBP [52]	0.7887	0.5882	0.7761	0.4790	**0.7879**	**0.5971**	**0.8576**	**8.5354**
NFERM [7]	0.6831	0.5034	0.6820	0.5750	0.4491	0.3107	0.7232	16.5983
NRSL [8]	0.6282	0.4483	0.6241	0.6699	0.5074	0.3420	0.6628	12.4263
VIQET [36]	0.8059	0.6123	0.8023	0.4713	0.7027	0.5239	0.7180	10.1844
BNBT	**0.8300**	**0.6301**	**0.8367**	**0.4658**	0.7482	0.5671	0.8137	9.5919

TABLE VI
PERFORMANCE COMPARISON OF BNBT AND EXISTING CONTRAST-DISTORTED BIQA METRICS ON THE NNID DATABASE

Metric	Entire database (2240)				Device-I (1400)				Device-II (640)				Device-III (200)			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
NR-CDIQA [24]	0.7007	0.4977	0.7036	0.1553	0.7144	0.5370	0.7198	0.1501	0.6963	0.4793	0.6988	0.1603	0.6979	0.4200	0.6519	0.1716
NR-SPL [25]	0.4110	0.2782	0.4057	0.2031	0.4758	0.3234	0.4131	0.2080	0.5176	0.3512	0.3016	0.2137	0.4615	0.3143	0.5073	0.1929
NIQMC [26]	0.8164	0.6198	0.8187	0.1255	0.8446	0.6515	0.8492	0.1204	0.7918	0.5960	0.7965	0.1353	0.7873	0.5935	0.7918	0.1419
BNBT	**0.8769**	**0.6822**	**0.8784**	**0.1061**	**0.8866**	**0.7066**	**0.8939**	**0.1020**	**0.8632**	**0.6737**	**0.8698**	**0.1157**	**0.8517**	**0.6890**	**0.8576**	**0.1137**



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

Fig. 11. Scatter plots of the performance of BNBT and contrast-distorted BIQA metrics on the NNID database. (a) NR-CDIQA [24], (b) NR-SPL [25], (c) NIQMC [26], and (d) BNBT.

TABLE VII
PERFORMANCE COMPARISON OF BNBT AND CONTRAST-DISTORTED BIQA METRICS ON SUBSETS OF THE CCRIQ AND CID2013 DATABASES

Metric	CCRIQ Subset (44)				CID2013 Subset (79)			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
NR-CDIQA [24]	0.5675	0.4081	0.5629	0.5368	0.3609	0.2428	0.6225	12.9873
NR-SPL [25]	0.2756	0.1940	0.4271	0.5872	0.2016	0.1076	0.2430	16.0967
NIQMC [26]	0.6362	0.4653	0.5772	0.5154	0.5743	0.4113	0.7122	11.8608
BNBT	**0.8300**	**0.6301**	**0.8367**	**0.4658**	**0.7482**	**0.5671**	**0.8137**	**9.5919**

as blurred details and reduced visibility, should also be considered. However, the contrast-distorted BIQA methods usually only consider one aspect of distortions, so their performance on the NNID database is not very satisfactory. In conclusion, our proposed BNBT metric is much more appropriate for predicting the visual quality of night-time images than the existing contrast-distorted BIQA methods.

Table VII reports the experimental results on the subsets of the CCRIQ and CID2013 databases. From Table VII, we can see that our proposed metric outperforms all the contrast-distorted methods on the subsets of both the CCRIQ and CID2013 databases. This observation demonstrates that our proposed BNBT metric has satisfactory ability to evaluate the visual quality of night-time images and shows stable and superior performance on other similar image datasets.

3) Statistical Significance Comparison: Statistical significance testing is a method for making quantitative statements about the performance of different IQA metrics from the viewpoint of sample numbers. One IQA metric may be statistically inferior when tested on only a small amount of images, while

TABLE VIII
COMPUTATIONAL COST OF BNBT AND OTHER EXISTING BIQA METHODS

Metric	BRISQUE [3]	BLIINDS-II [4]	NIQE [10]	ILNIQE [11]	SSEQ [53]	GM-LOG [5]	MSGF [6]	GWH-GLBP [52]
Time(s)	0.3715	58.3649	0.2312	4.1027	1.754	**0.1021**	1.8314	1.3097
Metric	NFERM [7]	NRSL [8]	VIQET [36]	NR-CDIQA [24]	NR-SPL [25]	NIQMC [26]	BNBT	
Time(s)	57.5746	0.1303	10.5137	3.6105	0.6018	3.2756	1.3418	

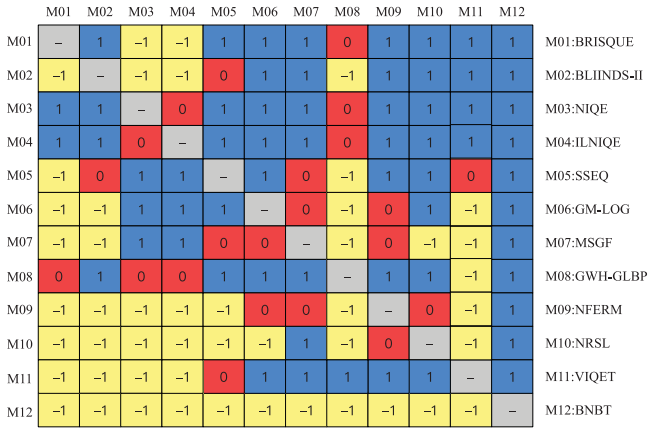


Fig. 12. Statistical significance comparison by the F-test between BNBT and general-purpose BIQA methods.

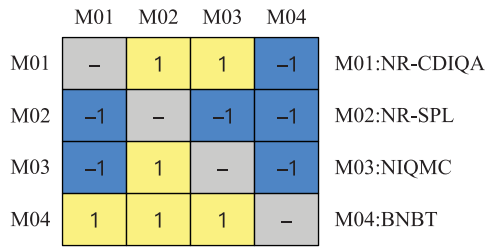


Fig. 13. Statistical significance comparison by the F-test between BNBT and contrast-distorted BIQA methods.

another IQA metric may be statistically inferior when tested on thousands of images. To make a statistical judgment regarding the superiority of one objective metric against another, the F-test [41] is employed to measure the statistical significance, which is based on variance-based hypothesis testing. The test measures the Gaussianity of the residual difference between the predicted objective scores (after nonlinear fitting) and the subjective scores (MOS), and the goal is to determine whether these two sets come from the same distribution or not. In our experiments, if the prediction residuals have a kurtosis between 2 and 4, they are considered to have Gaussianity.

Fig. 12 and Fig. 13 present the results of the hypothesis test for our proposed BNBT and the existing general-purpose and contrast-distorted BIQA methods on the NNID database, respectively. A symbol of 1 at (i, j) in these two figures indicates that the metric at row i performs statistically better than the metric at column j, a symbol of -1 denotes that the metric at row i is statistically worse than the metric at column j, and 0 means that these two metrics are statistically indistinguishable. We also fill the cells of different values with different colors for viewing convenience. It can be easily found from Fig. 12 and Fig. 13 that the performance of the proposed BNBT metric on the NNID database is statistically superior to those of all considered existing state-of-the-art BIQA methods, regardless of whether they are general-purpose or contrast-distorted.

D. Computational Complexity

It is necessary to analyze the computational complexity of an IQA algorithm because the running time is crucial in many real-time applications. In our experiments, the running time of predicting the quality of an image with a size of 512×512 is utilized for measuring the computational complexity of BNBT and comparative BIQA methods. Experiments are performed on a personal computer with an Intel Core i5 CPU @ 3.20 GHz, with 4 GB RAM. The software platform is MATLAB R2014b (8.4) under Windows 10 Home Premium. To eliminate the bias due to specific image selection, we randomly select 100 images from the NNID database, and the average running time of each method is used as its computational time cost.

The running times of generating quality scores consumed by each considered method are listed in Table VIII. From this table, we can find that GM-LOG [5] and NRSL [8] have relatively low computational cost because they extract features based on statistical histograms, which are easy to compute. NFERM [7] and BLIIND-II [4] are the two most computationally demanding methods, as both of them predict image quality by constructing a model needing complex calculations. The proposed BNBT shows a moderate running speed among all compared approaches. In contrast to other BIQA methods that usually include two procedures, feature extraction and quality prediction, BNBT has a superpixel segmentation procedure. The experimental results indicate that this procedure does not induce unacceptable heavy computational overhead for improving the prediction accuracy.

VI. CONCLUSION

We have conducted a comprehensive study on the subjective and objective quality assessment of night-time images. We construct a natural night-time image database (NNID), which contains 2240 images with 448 different image contents captured by different photographic equipment with different exposure times and other settings in different real-world night scenarios. The single stimulus (SS) method recommended by the International Telecommunications Union (ITU) is utilized to assess the subjective quality of the images in the NNID database. Furthermore, a blind night-time image quality assessment metric using brightness and texture features (BNBT) is proposed for objective assessment of night-time images. In the proposed

BNBT metric, a night-time image is first segmented into superpixels, and the mean value of each superpixel is calculated; then, brightness features are generated by weighting the mean value of each superpixel in the LMN color space. The energy, contrast and homogeneity of the gray-level co-occurrence matrix (GLCM) are extracted to form texture features. Both the brightness and texture features are combined for mapping to a quality score via support vector regression (SVR). Extensive experiments are conducted to evaluate the performance of our proposed BNBT metric on the NNID database. The results show that it outperforms the existing state-of-the-art general-purpose and contrast-distorted BIQA metrics at a moderate computational cost.

REFERENCES

[1] Z. Wang, "Applications of objective image quality assessment methods," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 137–142, Nov. 2011.

[2] X. Zhu and P. Milanfar, "Automatic parameter selection for denoising algorithms using a no-reference measure of image content," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3116–3132, Dec. 2010.

[3] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[4] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[5] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[6] Q. Wu, H. Li, F. Meng, K. N. Ngan, and S. Zhu, "No reference image quality assessment metric via multi-domain structural information and piecewise regression," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 205–216, 2015.

[7] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

[8] Q. Li, W. Lin, J. Xu, and Y. Fang, "Blind image quality assessment using statistical structural and luminance features," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2457–2469, Dec. 2016.

[9] T.-J. Liu and K.-H. Liu, "No-reference image quality assessment by wide-perceptual-domain scorer ensemble method," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1138–1151, Mar. 2018.

[10] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[11] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[12] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," 2005. [Online]. Available: http://live.ece.utexas.edu/research/quality

[13] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 011006-1–011006-21, Mar. 2010.

[14] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, 2015.

[15] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2012, pp. 1693–1697.

[16] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 390–402, Jan. 2015.

[17] M. A. Saad *et al.*, "Impact of camera pixel count and monitor resolution perceptual image quality," in *Proc. Colour Vis. Comput. Symp.*, 2015, pp. 1–6.

[18] M. A. Saad *et al.*, "Image quality of experience: A subjective test targeting the consumers experience," *Electron. Imag.*, vol. 2016, pp. 1–6, 2016.

[19] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[20] A. Ciancio *et al.*, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, Jan. 2011.

[21] J. Zhang, S. H. Ong, and T. M. Le, "Kurtosis-based no-reference quality assessment of JPEG2000 images," *Signal Process., Image Commun.*, vol. 26, no. 1, pp. 13–23, 2011.

[22] G. Yang, Y. Liao, Q. Zhang, D. Li, and W. Yang, "No-reference quality assessment of noise-distorted images based on frequency mapping," *IEEE Access*, vol. 5, pp. 23146–23156, 2017.

[23] Y. Zhan and R. Zhang, "No-reference image sharpness assessment based on maximum gradient and variability of gradients," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1796–1808, Jul. 2018.

[24] Y. Fang *et al.*, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 838–842, Jul. 2015.

[25] M. Xu and Z. Wang, "No-reference quality assessment of contrast-distorted images," in *Proc. IEEE Int. Conf. Signal Image Process.*, 2016, pp. 362–367.

[26] K. Gu *et al.*, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4559–4565, Dec. 2017.

[27] K. Gu *et al.*, "Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 432–443, Mar. 2016.

[28] G. Yue, C. Hou, and T. Zhou, "Blind quality assessment of tone-mapped images considering colorfulness, naturalness and structure," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3784–3793, May 2019.

[29] B. Hu *et al.*, "No-reference quality assessment of compressive sensing image recovery," *Signal Process., Image Commun.*, vol. 58, pp. 165–174, 2017.

[30] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1301–1313, Apr. 2018.

[31] 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.

[32] K. Gu *et al.*, "No-reference quality assessment of screen content pictures," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 4005–4018, Aug. 2017.

[33] Y. Fang, 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.

[34] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomed. Signal Process. Control*, vol. 27, pp. 145–154, 2016.

[35] J. Dutta, S. Ahn, and Q. Li, "Quantitative statistical methods for image quality assessment," *Theranostics*, vol. 3, no. 10, pp. 741–756, 2013.

[36] M. A. Saad, P. Corriveau, and R. Jaladi, "Objective consumer device photo quality evaluation," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1516–1520, Oct. 2015.

[37] F. Shao, Y. Gao, F. Li, and G. Jiang, "Toward a blind quality predictor for screen content images," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 9, pp. 1521–1530, Sep. 2018.

[38] A. A. Nakhaie and S. B. Shokouhi, "No reference medical image quality measurement based on spread spectrum and discrete wavelet transform using ROI processing," in *Proc. Can. Conf. Elect. Comput. Eng.*, 2011, pp. 121–125.

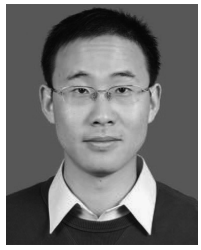
[39] M. M. Kalayeh, T. Marin, and J. G. Brankov, "Generalization evaluation of machine learning numerical observers for image quality assessment," *IEEE Trans. Nucl. Sci.*, vol. 60, no. 3, pp. 1609–1618, Jun. 2013.

[40] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Rec. ITU-R BT.500-11, International Telecommunications Union, Geneva, Switzerland, Jun. 2002.

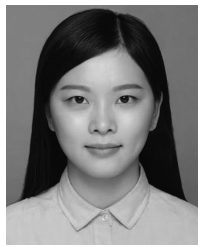
[41] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[42] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE Trans. Broadcast.*, vol. 61, no. 3, pp. 520–531, Sep. 2015.

[43] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 10–17.
[44] X. Jin and Y. Gu, "Superpixel-based intrinsic image decomposition of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4285–4295, Aug. 2017.
[45] V. N. Gangapure, S. Nanda, and A. S. Chowdhury, "Superpixel-based causal multisensor video fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1263–1272, Jun. 2018.
[46] C. Peng, X. Gao, N. Wang, and J. Li, "Superpixel-based face sketch–photo synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 2, pp. 288–299, Feb. 2017.
[47] R. Achanta *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[48] J.-M. Geusebroek, R. Van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1338–1350, Dec. 2001.
[49] J.-M. Geusebroek, R. Van Den Boomgaard, A. W. Smeulders, and A. Dev, "Color and scale: The spatial structure of color images," in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 331–341.
[50] R. M. Haralick *et al.*, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
[51] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.
[52] Q. Li, W. Lin, and Y. Fang, "No-reference quality assessment for multiply-distorted images in gradient domain," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 541–545, Apr. 2016.
[53] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process., Image Commun.*, vol. 29, no. 8, pp. 856–863, 2014.
[54] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3485–3492.
[55] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Mar. 2000. [Online]. Available: http://www.vqeg.org/
[56] K. Ma *et al.*, "Group mad competition? A new methodology to compare objective image quality models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1664–1673.



Tao Xiang (M'11) received the B.Eng., M.S., and Ph.D. degrees in computer science from Chongqing University, Chongqing, China, in 2003, 2005, and 2008, respectively. He is currently a Professor with the College of Computer Science, Chongqing University. He has also served as a referee for numerous international journals and conferences. He has authored or coauthored more than 90 papers in international journals and conferences. His research interests include multimedia security, cloud security, data privacy, and cryptography.



Ying Yang received the B.Sc. degree in computer science from Chongqing University, Chongqing, China, in 2016. She is currently working toward the Ph.D. degree at the College of Computer Science, Chongqing University. Her current research interests include multimedia security and image quality assessment.

Shangwei Guo received the Ph.D. degree in computer science from Chongqing University, Chongqing, China, in 2017. From 2018 to 2019, he was a Postdoctoral Research Fellow with Hong Kong Baptist University, Hong Kong. He is currently a Postdoctoral Research Fellow with Nanyang Technological University, Singapore. His research interests include multimedia security, cloud security, and data privacy.