Subjective and Objective Quality Assessment of Compressed Screen Content Videos

Teng Li, *Member, IEEE*, Xiongkuo Min, *Member, IEEE*, Heng Zhao, Guangtao Zhai, *Senior Member, IEEE*, Yiling Xu, *Member, IEEE*, and Wenjun Zhang, *Fellow, IEEE*

***Abstract*—With the widespread of application scenarios such as remote office and cloud collaboration, Screen Content Video (SCV) and its processing which show different characteristics from Natural Scene Video (NSV) and its processing, are increasingly attracting researcher's attention. Among these processing techniques, quality evaluation plays an important role in various media processing systems. Despite extensive research on general Image Quality Assessment (IQA) and Video Quality Assessment (VQA), quality assessment of SCVs remains undeveloped. In particular, SCVs always suffer from compression degradations in all kinds of application scenarios. In this article, we first study subjective SCV quality assessment. Specifically, we first construct a Compressed Screen Content Video Quality (CSCVQ) database with 165 distorted SCVs compressed from 11 most common screen application scenarios using the H.264, HEVC and HEVC-SCC formats. Twenty subjects were recruited to participate in the subjective test on the CSCVQ database. Then we study objective SCV quality assessment and propose a SCV quality measure. We observe that localized protruding information such as curves and dots can be well captured by the local relative standard deviation which then can be used to measure the intra-frame quality. Base on this observation, we develop a MutiScale Relative Standard Deviation Similarity (MS-RSDS) model for SCV quality evaluation. In our model, the relative standard deviation similarity between the reference and distorted SCVs is measured from frame differences between two adjacent frames, which can capture the spatiotemporal distortions accurately. A multiscale strategy is also applied to strengthen the original single-scale model. Extensive experiments are performed to compare the proposed model with the most popular and state-of-the-art quality assessment models on the CSCVQ database. Experimental results show that our proposed MS-RSDS model which has relatively low computation complexity, outperforms other IQA/VQA models.**

***Index Terms*—Video quality assessment, screen content video, SCV quality database, relative standard deviation similarity, multiscale.**

Manuscript received July 15, 2020; revised July 28, 2020; accepted August 17, 2020. Date of publication October 21, 2020; date of current version June 5, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61901260, Grant 61831015, Grant 61521062, Grant 61527804, and Grant 61971282; in part by the National Key Research and Development Project of China Science and Technology Exchange Center under Grant 2018YFE0206700; and in part by the Scientific Research Plan of the Science and Technology Commission of Shanghai Municipality under Grant 18511105402. *(Corresponding authors: Guangtao Zhai; Yiling Xu.)*

The authors are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: liteng0521@sjtu.edu.cn; minxiongkuo@sjtu.edu.cn; z-heng@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn; yiling.xu@gmail.com; zhangwenjun@sjtu.edu.cn).

Digital Object Identifier 10.1109/TBC.2020.3028335

I. Introduction

With the tremendous increase in network transmission capacity and stability, extensive new functions are being implemented as never before through so many easily accessible multimedia services. In many scenarios, e.g., cloud-mobile applications, remote computing platforms, and cloud gaming, remote computing is facilitated based on the users interaction with the local display interface, which typically includes computer-generated screen content videos [1]. As the user requirements and expectations for high-quality screen content videos are increasingly rising, a reliable system to evaluate, control and improve the users' quality of experience (QoE) is urgently required [2]–[6]. Video Quality Assessment (VQA) technique is widely used by various multimedia services in all kinds of devices. Designing an efficient and accurate objective assessment model which can simulate Human Visual System (HVS) is of great significance for numerous image and video processing tasks [7]–[13].

In general, the assessed media can be classified into two categories according to their sources: Natural Scenes Video (NSV or Natural Scene Image, NSI) and Screen Content Video (SCV or Screen Content Image, SCI). SCVs usually contain combinations of texts, graphics and camera-captured images [14]. SCV is often identified as an important new category of video, which possesses distinct characteristics compared with natural videos captured directly by cameras [15]. SCVs tend to have radical changes resulting from sharp region transitions and texts, which will lead to the large size of edges [16]. These distinct characteristics call for new theories and techniques beyond the traditional techniques for the objective of better screen content processing. For example, screen content quality assessment and screen content coding are widely researched in the literature. With the rapid development of computer technologies, screen content becomes more and more indispensable for applications such as wireless display, video conference and cloud computing, where high quality screen content videos need to be transmitted to the client-side [14].

Quality assessment can be classified into subjective and objective quality assessment [2]. Subjective quality assessment, which requires human observers to assess each media content in a specific protocol, is a straightforward and reliable approach to evaluate media quality. In spite of costliness, inconvenience and being time-consuming, a comprehensive subjective quality study is required in many cases. It collects data from extensive user opinions through evaluating the perceived quality of SCVs and it can be utilized as a

benchmark to validate and compare the performance of objective VQA models. In order to verify the effectiveness of all kinds of VQA or IQA algorithms, various organizations and researchers have also developed databases as benchmarks. These databases are basically established following the instructions of the ITU standard [17],for example the Screen Image Quality Assessment Database (SIQAD) [18] and the Cross-Content-Type (CCT) database [19]. These databases, generally include several reference media contents and their associate distorted versions generated by several distortion methods such as compression, display distortion or noise superposition. Although there are a bunch of image quality databases targeting NSI and SCI, as well as video database targeting NSV, there is currently no database for SCV quality assessment.

From the objective modeling percept of view, researchers have developed numerous quality assessment models to predict the subjective scores collected from the subjective assessment in the aforementioned databases. The most well-known full reference model is Mean Square Error (MSE) or Peak Signal to Noise Ratio (PSNR), which is simple but criticized for not effective in perceived quality assessment. Another groundbreaking metric Structural Similarity Index Measure (SSIM) [20] was proposed by Wang *et al.*, which extracts structural features from images. Based on SSIM, a bunch of modification models such as Multi-Scale SSIM (MS-SSIM) [21] and Information Weighted SSIM (IW-SSIM) [22], assume that HVS tends to perceive the local structures in a multi-scale manner. In [23], the Feature SIMilarity (FSIM) applies the phase congruency and the gradient magnitude to further extract protruding edges and texture information. In [22], the author adopt the Gaussian Scale Mixture (GSM) model for natural images by using the statistical modeling of groups of neighboring pixels. The visual saliency is applied for IQA in the Visual Saliency-based Index (VSI) model [24]. In [25], A Perceptual SIMilarity (PSIM) measure is proposed by fusing the micro- and macro-structures. The Visual Information Fidelity (VIF) [26] and Information Fidelity Criteria (IFC) [27] take HVS as a communication channel and they predict the subjective image quality by computing how much the information within the perceived reference image is preserved in the perceived distorted one. The Gradient Magnitude Similarity Mean (GMSM) and the Gradient Magnitude Similarity Deviation (GMSD) [28] make full use of the pixel-wise similarity between the gradient magnitude maps of reference and distorted images.

Researchers have studied and proposed some possible solutions to conduct VQA. One simple but reasonably effective strategy is to compute the frame-level quality scores generated by applying IQA models, and then to apply temporal pooling on the frame-level quality scores. Such temporal pooling strategy is widely used in many VQA models [29]–[31]. Besides average pooling, other pooling strategies may include harmonic mean [32], Minkowski mean [33] and percentile pooling [34]. Obviously, the performance of these combination approaches is highly related to the efficiency of the IQA models and pooling strategies they used.

A considerable amount of studies have also designed models focusing on the quality of the complete video rather than pooling the quality of each frame, since temporal information is also considered as a significant factor for video quality assessment. Wang *et al.* [35] proposed a novel VQA model by considering both the motion information and the perceptual uncertainty in video signals. ST-MAD [36] employs spatial-temporal slices created by taking time-based slices of the original and distorted videos to achieve effective quality assessment for natural video. In STRRED [37], a Gaussian scale mixture model for the wavelet coefficients of frames and frame differences is used to measure the amount of spatial and temporal information differences between the reference and distorted videos, respectively. MOtion-based Video Integrity Evaluation (MOVIE) index [29] seeks a novel model by developing a general, spatio-spectrally localized multiscale framework for evaluating dynamic video fidelity, and it considers both spatial and temporal (and spatial-temporal) aspects of the video distortions.

However, the aforementioned models cannot fulfill the need of quality assessment for screen content due to the content differences as described above, while these content differences can lead to totally different visual perception. In [38], the authors proposed a Screen Content Perceptual Quality Assessment (SPQA) model by first roughly dividing the input images into pictorial and textual regions, followed by comparing and combining the perceptual differences of the aforementioned two types of regions between the corrupted and uncorrupted screen content images, finally yielding single-score quality estimation. In [39], the authors adopt a simple idea by deploying adaptive window sizes of local filters to modify the classical SSIM metric. In [40], the authors designed a gradient direction-based screen content image quality metric. This metric derives the final quality estimation by performing a deviation-inspired model for pooling after extracting the local gradient direction information. The SQMS [41] model extracts and integrates four types of features including picture complexity, screen content statistics, global brightness quality, and the sharpness of image details. The SVQI [42] model combines the measurements of variations in the global and local structures to yield the final quality estimation of screen content images. Besides the above quality measures specifically designed for SCIs, there are also some unified quality measures which are effective for both natural scene and screen content. For example, UCA [19] and SIRR [43] are specifically designed unified quality measures, while some general-purpose quality measures like BPRI [44] and BMPRI [5] are verified on both natural scene and screen content. These IQA models for SCIs, however, are not applicable for screen video quality evaluation. For the reason that there is currently no objective quality assessment model for SCVs in the literature, the development of a relevant model is needed.

In this article, we firstly establish a brand new Compressed Screen Content Video Quality Database, i.e., CSCVQ database, which will be introduced in details in Section II. Then we proposed a new screen content video quality model by measuring the multiscale localized relative standard deviation similarity, whose details will be described in Section III. In Section IV, we test our proposed model and compare it

Fig. 1. Sample frames for all 11 reference videos included in the proposed CSCVQ database. a. Chrome Browsing. b. Desktop. c. Multitask. d. Video Conferencing. e. Flying Graphics. f. Map. g. Programming. h. Robot. i. Slide Editing. j. Word Editing. k. Slide Showing.

with other popular IQA/VQA models on the CSCVQ database. Conclusion is drawn in Section V.

II. Subjective Compressed Screen Content Video Quality Assessment

For the purpose of developing and verifying quality assessment algorithms, a collection of screen videos with different quality grades are demanded. Thus we construct a database called Compressed Screen Content Video Quality (CSCVQ) database, which includes 11 screen content videos and 165 distorted videos generated by three commonly used compression methods. Each of them is scored by 20 human observers in a single stimulus setting with a continuous quality rating scale. The construction of the database and the details of the subjective study will be introduced in the following two subsections.

A. Reference and Distorted SCVs

Our source sequences contain 11 distortion free screen videos recorded directly by screen recording software (OBS studio). To fully demonstrate the characteristics of screen applications, the criteria of content selection is mainly based on three aspects. First, the sequences tend to have different types of motion characteristics. Motionless frames, partial movements and sudden changes are commonly existed in the screen content, and different motion characteristics can affect human observers in significantly different ways. Second, an abundant variety of graphics and colorful contents are suggested to be included in the sequence. Computer generated videos consist of texts, windows, symbols, patterns, etc., which make screen content videos be rich in sharp edges and high chroma colors. Third, various scenarios should be covered. Different types of application scenarios can result in different mental and physical states in the human subjects. Typical screen application scenarios include Web surfing, multitasking, video conferencing, etc.

We carefully tweak a bunch array of configuration options to capture these videos. All the videos last for 10 seconds with a frame rate of 30 frames per second. Though the screen we use to generate these videos has a resolution of 1920×1080, we truncate a part of the whole screen with a 1280×720 region for the convenience of the following subjective experiments. We capture videos in this way rather than down-sampling the high-resolutions videos because the down-sampling can introduce some additional distortions. The videos are saved with a YUV 4:2:0 format. Though it is widely proved that the audio and video will together shape the user-perceived QoE [45]–[47], we exclude the audio components and focus on the video in this article. Fig. 1 shows the sample frames for all source videos in our database and a short description of each video's content is provided below.

- *Chrome Browsing:* Web surfing with page rolling and tab switching. The main color changes extremely before and after tab switching.
- *Desktop:* Text showing and highlighting, then switching window to command lines and coding lines.
- *Multitask:* Four small windows (Web browsing, command window, compiler window and animation) are displayed simultaneously and then windows are flipped with a 3D stack look.
- *Video Conferencing:* Online video meeting and cloud collaboration scene. Cloud camera with conventioneer's view is located on the left. Text and sheet editing collaboration are located on the right.
- *Flying Graphics:* A high complexity scene which contains mix of rotating texts, symbols, command

lines and charts with changing background color and texture.
- *Map*: Map scaling and text size zooming step by step.
- *Programming:* A program running scene which produces 3D graphics change in the left bottom window.
- *Robot*: A computer game scene with view angle translation, shadows movement, snow falling and robot action.
- *Slide Editing:* Two parallel windows showing slide and text editing.
- *Word Editing:* Text editing with line selecting, highlighting and zooming. Windows' thumbnail and size zooming are also included.
- *Slide Showing:* Presenting a common slide which has a bunch of texts, graph and icons, as well as various kinds of animations. This video contains a certain amount of blank (all white) frames during the slides switching.

Screen video is usually generated directly by computer and transmitted via various relatively reliable networks, thus the screen content may rarely be affected by tradition distortions like Gaussian blur and Gaussian noise, or the display distortions such as contrast change and color change. Compared with traditional visual quality assessment for natural scene videos, quality assessment for screen content videos should focus more on compression distortions, since the main distortion source for screen content video is compression. Considering these, our CSCVQ database focuses on the compression loss. We create 15 distorted videos from each of the reference videos using three compression methods, including H.264 compression, High Efficiency Video Coding (HEVC) compression and Screen Content Compression (SCC). All of them are widely and commonly used in various kinds of video compression applications.

1) H.264: H.264 is a block-oriented, compensation-based none lossless video compression standard that defines multiple profiles (tools) and levels (max bitrates and resolutions). We use the well known H.264 codec implemented by FFmpeg to generate 5 levels of distortions for each reference video. The level of the distortion is controlled by the quantization parameter (QP) which is set as 24, 30, 36, 42 and 48, respectively. Note that an increase of 1 in QP means an increase of the quantization step size by approximately 12%. An increase of 6 leads to an increase in the quantization step size by a factor of 2, which indicates a different level of distortion. For the universality, we set the GOP size at 8. Take the *Slide Showing* video as an example, the total file size is 2585kb (QP = 24), 2187kb (QP = 30), 1384kb (QP = 36), 838kb (QP = 42) and 474kb (QP = 48), respectively.

2) HEVC: HEVC, also known as H.265, is drawing more and more attention due to its superior compression efficiency compared to H.264. HEVC describes a large range of block size up to 64×64 pixels to employ adaptive quad-tree coding using Coding Tree Unit (CTU). HEVC can easily gain twice the efficiency as H.264 on the same source video. We used the reference software HM (HEVC Test Model Version 16.18) with the low delay REXt configuration made available by the Joint Collaborative Team on Video Coding (JCT-VC). We also set the GOP size as 8 and the QP to five levels from 24 to 48 to control the compression level.

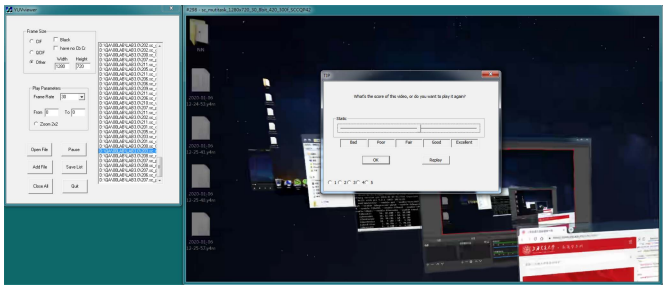


Fig. 2. User interface for the subjective tests.

3) HEVC-SCC [48]: To efficiently compress the screen content video, the Screen Content Coding (SCC) is developed based on the HEVC, which is regarded as the extension of the HEVC [49]. The HEVC-SCC improves compression capability for video containing a significant portion of screen content such as rendered (moving or static) graphics, text, or animation rather than camera-captured video scenes. We include this extension as one distortion type in our database to highlight the prominent features of computer generated content. Again, we set the GOP size as 8 and the QP as 5 levels from 24 to 48.

B. Subjective SCV Quality Evaluation Study

Our subjective test is designed under the guide of ITU-R BT.500-13, which is a widely accepted recommendation of subjective quality accessing process [17]. There are two basic categories of methodologies, single stimulus and double stimulus. The double stimulus method of laboratory testing does not reflect the every-day viewing conditions since it requires the observer to examine both the reference and the distorted content simultaneously. Compare to the double stimulus method, the single stimulus method is considered more useful and efficient. In a single stimulus test session, the subjects are asked to view the distorted videos without the references and then measure the quality of the videos continuously. Thus Single Stimulus Continuous Quality Evaluation (SSCQE) is adopted in our test.

We developed an user interface as shown in Fig. 2. The observers are asked to examine all the 165 test sequences one by one in a random order and assess the overall video quality of each sequence by adjusting a horizontal bar. Through this sliding bar, we can collect continuous rating scores. The scales are divided into five equal lengths which correspond to the ITU-R five-point quality scale (Excellent, Good, Fair, Poor and Bad). Any possibility of confusion between the scale divisions and the test results is avoided. The subjects are allowed to take as long as they want to evaluate the quality of every video and they can replay as many times as they want. But once they have already given a score, they cannot go back to the previous one and rescore it.

Before the test, all the subjects are instructed about the process of the assessment, the types of impairment or quality factors likely to occur, the grading scale, the lengths of the sequence and the whole test, etc. We also include some training sequences, which have the same length and resolution as those used in the formal test, and do not have any content

overlap with them, to demonstrate the range and the type of the impairments to be assessed. The whole accessing test lasts for 45 minutes or so and the subjects can take a rest during the testing to avoid fatigue.

We recruit 20 subjects from the campus, including 15 males and 5 females. All of them have normal or corrected-to-normal vision. They all have less expertise in the video processing field. The specific and detailed information of the experiments remains confidential to the subjects. To ensure perfect playback, all distorted sequences were processed and stored as raw YUV 4:2:0 files. The subjective tests are performed on a typical desktop PC with 20 GB RAM and 64-bit Windows operating system. The videos are displayed on a 23-inch 1920×1080 resolution 60 Hz monitor which is placed in a laboratory with normal indoor light. The test sequences were played without scaling on a low brightness and low chroma desktop background to ensure the test accuracy.

According to [17], a subject rejection procedure to discard scores from unreliable subjects is necessary before subjective score processing. We perform this process to discard observers who have produced scores significantly distant from the average scores. Scores assigned by every subject is examined by its kurtosis β_i, which is given by:

$$\beta = \frac{m_4}{m_2^2} \quad with \quad m_x = \frac{\sum_N^{j=1}(s_{ij} - \mu_i)^x}{N}, \tag{1}$$

where N denotes the number of subjects, s_{ij} denotes the score of jth video given by subject i, μ_i denotes the mean score of all videos assigned by subject i. For each observer i, the rejection algorithm count the number of scores which fall out of the confidence interval, i.e., P_i and Q_i. When $2 \leq \beta \leq 4$, then:

$$if \quad s_{ij} \geq \mu_i + 2\sigma_i, \quad then \quad P_i = P_i + 1; \tag{2}$$

$$if \quad s_{ij} \leq \mu_i - 2\sigma_i, \quad then \quad Q_i = Q_i + 1; \tag{3}$$

where σ_i denotes the score standard deviation of all videos assigned by subject i. When β falls out of the interval of $[2, 4]$, then:

$$if \quad s_{ij} \geq \mu_i + \sqrt{20}\sigma_i, \quad then \quad P_i = P_i + 1; \tag{4}$$

$$if \quad s_{ij} \leq \mu_i - \sqrt{20}\sigma_i, \quad then \quad Q_i = Q_i + 1. \tag{5}$$

Finally, the subject will be rejected if $(P_i + Q_i)/N > 0.05$ and $|(P_i - Q_i)/(P_i + Q_i)| < 0.3$. By this rejection procedure, one of the subjects in our test is rejected.

Mean Opinion Score (MOS) is calculated to describe the quality of each video. Let s_{ij} denotes score of subject i to video j, and we convert it to Z-scores [50]:

$$\mu_j = \frac{1}{N_j}\sum_{j=1}^{N_j} s_{ij}, \tag{6}$$

$$\sigma_j = \sqrt{\frac{1}{N_j}\sum_{j=1}^{N_j}(s_{ij} - \mu_i)^2}, \tag{7}$$

$$Z_{ij} = \frac{s_{ij} - \mu_j}{\sigma_j}, \tag{8}$$

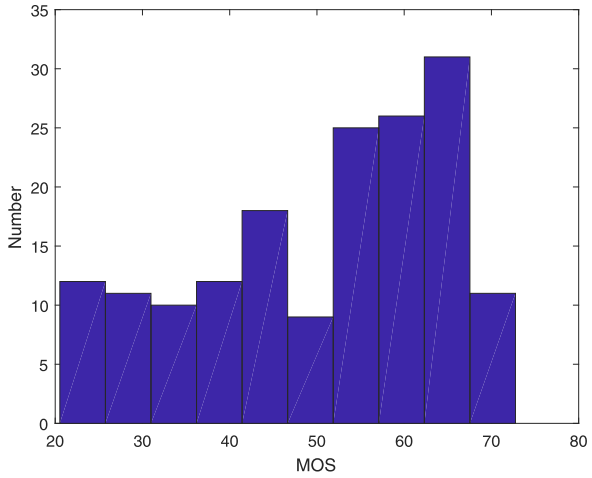


Fig. 3. Histogram of MOSs in the CSCVQ database.

where N_j is the number of the distorted videos scored by the subject i (without the subject who has been rejected). The Z-scores can reduce the influence of the rating range of each individual. We finally compute the mean of Z-scores as the MOS:

$$MOS_j = \frac{1}{N_i}\sum_{i=1}^{N_i} s_{ij}. \tag{9}$$

Fig. 3 shows the histogram of MOSs of all videos from our created database. It is observed that the MOSs of the CSCVQ database spread in a wide range uniformly.

III. Objective Screen Content Video Quality Assessment

VQA algorithms in most cases take both intra-frame spatial information and temporal information among frames into account. A series of IQA algorithms can be generalized to VQA algorithms with appropriate pooling strategy, e.g., mean pooling. Following this path, a VQA algorithm can be developed by first focusing on intra-frame feature extraction, and then considering inter-frame features. Additionally, computational complexity should also be considered as an essential factor due to the fact that VQA always deals with huge-scale high-definition media content. We propose a novel algorithm which consists of intra-frame local feature capturing and scoring, multi-scale processing, frame difference calculation, and inter-frame pooling. The proposed MS-RSDS model is illustrated in Fig. 4.

Screen content videos are full of thin lines, limited colors and regular shapes with more noise-free smooth areas or extremely sharp texts [19], [41]. These features can significantly influence subjects when accessing information in videos. We extract these features by adopting the concept of relative standard deviation, defined as the ratio of standard deviation σ to the mean μ:

$$C_v = \frac{\sigma}{\mu}. \tag{10}$$

Relative standard deviation is a dimensionless measurement, which indicates the dispersion degree of the data. In

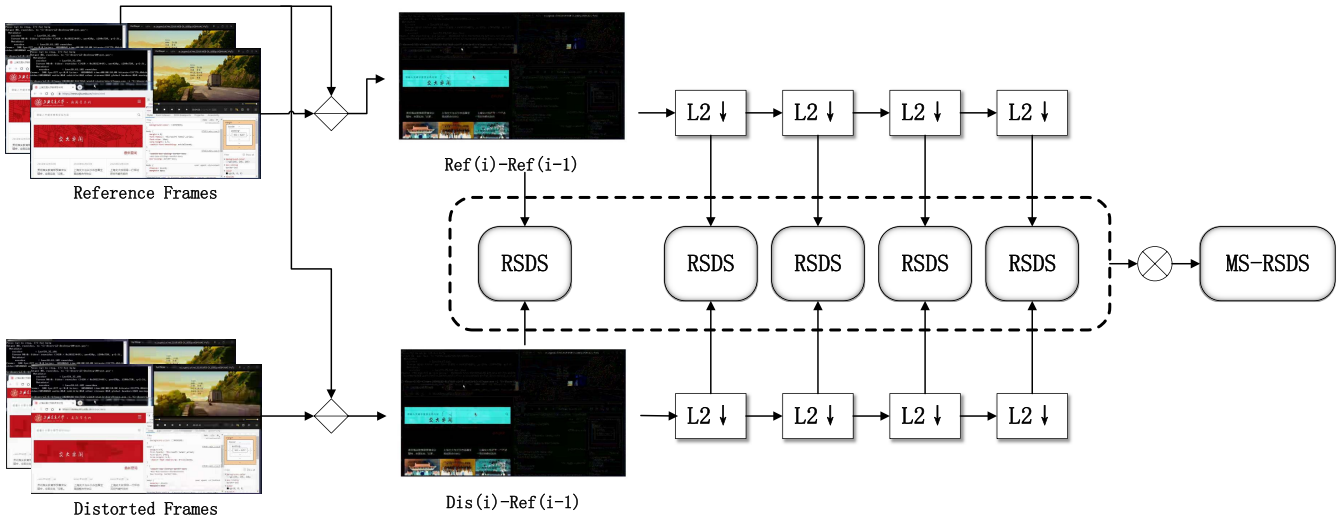


Fig. 4. Framework of the proposed MS-RSDS model. *L2* ↓: downsampling by factor 2 with low-pass filter.

the field of image processing, relative standard deviation can be regarded as a measurement of local contrast or saliency. Thus, we apply this concept to local image feature extractions by calculating the luminance relative standard deviation in a small window. Note that this measurement is more sensitive to small changes in the areas with small local mean luminance. This characteristic meets the need of extracting texture features which are pretty effective for the task of visual quality assessment [51], [52].

Assume W as a local window in a frame, the local relative standard deviation of the center pixel $RSD(i,j)$ in W is defined as:

$$RSD(i,j) = \frac{(Y(i,j) - \mu_W)^2 + c}{\mu_W + c}, \tag{11}$$

where μ_W is the mean luminance of all pixels within the local window, $Y(i,j)$ is the luminance at pixel (i,j), c is a constant to maintain stabilization. In our implementation, we adopt a Gaussion filtered window to gain the average instead of arithmetic mean to amplify the weights of the central pixels. Assuming $\mathbf{Y}$ is the frame to be processed, then the Guassion filtered average map $\mathbf{Y}_g$ and RSD Map $\mathbf{RSD}_m$ can be calculated by:

$$\mathbf{Y}_g = \mathbf{Y} \otimes \mathbf{G}_k, \tag{12}$$

$$\mathbf{RSD}_m = \frac{(\mathbf{Y} - \mathbf{Y}_g)^2 + c}{\mathbf{Y}_g + c}, \tag{13}$$

where $\mathbf{G}_k$ is a 2D rotationally symmetric normalized Gaussian window and $\otimes$ means convolution.

This RSD map can share some similarity with the gradient map. Fig. 5 shows the RSD map and gradient map extracting from the two different levels of distortion frames of the *Video Conferencing* video. The RSD map reveals the saliency region such as image textures, texts and icons, which are easy to attract human observers [53], [54]. Distortions in these regions are more likely to influence the final score since human subjects evaluate the quality of images based on the reached information, rather than other non-reached information such as those in the smooth background regions. From Fig. 5, we can observe that in gradient map lots of unexpected edges are located out of the salient zone, whereas the RSD map has captured the most salient areas. We can expect that RSD map is more appropriate for VQA and IQA when dealing with compressed screen content.

In the state-of-the-art full-reference IQA and VQA models, the similarity between the features extracted from the reference and distorted signals are calculated to measure the quality. Let $\mathbf{RSD}_m^r$ and $\mathbf{RSD}_m^d$ denote the RSD maps calculated from reference frame and its distorted frame. We measure the similarity between these two maps for a specific frame:

$$\mathbf{S} = \frac{2 * \mathbf{RSD}_m^r * \mathbf{RSD}_m^d + p}{\left(\mathbf{RSD}_m^r\right)^2 + \left(\mathbf{RSD}_m^d\right)^2 + p}, \tag{14}$$

$$RSDS = std[\mathbf{S}], \tag{15}$$

where p is also a constant to maintain stabilization. Note that the RSDS value reveals the degree of inconformity of the distorted frame from the reference one. That means high RSDS value indicates low quality of the distorted frame.

Multiscale processing is widely accepted as a useful strategy to enhance the performance of quality prediction models. The perceivability of image details depends on several viewing conditions, e.g., display resolution, viewing distance, sampling density etc. Considering this perceivability differences, mutiscale processing model simulates the mechanism of HVS by calibrating the parameters to weight the relative importance between different scales [21]. Thus, we also include this multiscale strategy in our model:

$$MS - RSDS = \prod_{k=0}^{M} [RSDS_k]^{\alpha_k}. \tag{16}$$

Fig. 4 shows the procedure of calculating MS-RSDS. The system iteratively applies a low-pass filter and downsamples the filtered frame by a factor of 2. We index the original frame as Scale 0, and the highest scale as Scale M, which is obtained after $M-1$ iterations. The exponents vector α_k is used to adjust

Fig. 5. Comparison of RSD map and gradient map. The first row shows two same frames from two different distortion levels of the *Video Conferencing* video. The left column of the other rows shows gradient maps of aforementioned frames and the gradient similarity map. The right column shows the associated RSD maps and RSD similarity map. a. Distorted Frame 1 (HEVC-SCC, $QP = 24$). b. Distorted Frame 2 (HEVC-SCC, $QP = 48$). c. Gradient map of Distorted Frame 1. d. RSD map of Distorted Frame 1. e. Gradient map of Distorted Frame 2. f. RSD map of Distorted Frame 2. g. Similarity map of subfigure c and reference frame's gradient map. h. Similarity map of subfigure d and reference frame's RSD map.

the relative importance of different scales. We will discuss the parameters c, p, M and α_k in detail in the following section.

Up to now, we have gained a useful frame evaluation model which can be regarded as a new IQA algorithm to evaluate the quality of SCIs. In the process of designing this algorithm, we obtain quality evaluation features based on the characteristics of screen content. At the same time, we also took into account the computational complexity and operability of the algorithm. Compared with a large number of IQA algorithms based on frequency domain characteristics, this algorithm achieves a good balance between computational complexity and effectiveness, and it gives us the possibility to further extend the algorithm for VQA.

To design a VQA algorithm, temporal information and spatial information ought to be integrated when evaluating the final video quality. Static content draws more attention than dynamic content. In screen content scenarios, static contents such as texts, charts, etc. play a decisive role in the quality evaluation process. Thus, a subtle weighting strategy should be adopted between dynamic and static parts of continuous time frames. The characteristics of the RSD measurement can well match for such weighting strategy appropriately, since RSD can be conducted more sensitively in areas with small local mean luminance. Additionally, frame difference information is widely used in various fields of video processing such as compression, transmission and quality evaluation. Frame difference information often reflects the spatio-temporal motion information of a video. In our model, the MS-RSDS is performed on frame difference image:

$$\mathbf{R}_d = \mathbf{Ref}_{k+1} - \mathbf{Ref}_k, \tag{17}$$

$$\mathbf{D}_d = \mathbf{Dis}_{k+1} - \mathbf{Ref}_k, \tag{18}$$

$$S_{MS-RSDS} = \frac{1}{N-1}\sum_{k=1}^{N-1} \mathbf{R}_d \bigodot \mathbf{D}_d, \tag{19}$$

where $\mathbf{Ref}_k$ and $\mathbf{Dis}_k$ are the kth frame of reference video and distorted video. Note that they are both obtained by the difference from the reference video's previous frame, so that this difference can extract the distortion information precisely. We then perform MS-RSDS on R_k and D_k (denoted as $\bigodot$), and get the final score of distorted video by average pooling.

IV. Experimental Validation

A. Experimental Settings

To evaluate the quality prediction models, we need to measure the consistency between the scores predicted by the objective quality assessment model and the subjective MOSs. A higher degree of consistency indicates better simulation of the HVS characteristics, which means a better objective model. We use three indicators widely used by the IQA and VQA community to quantify the consistency between various assessment models and MOS values, including Spearman Rank-Order Correlation Coefficient (SROCC) for rank correlation, Pearson Linear Correlation Coefficient (PLCC) for linear correlation and Root Mean Squared Error (RMSE) for consistency:

$$SROCC = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2-1\right)}, \tag{20}$$

$$PLCC = \frac{cov(P, MOS)}{\delta_P \delta_{MOS}}, \tag{21}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - MOS_i)}, \tag{22}$$

where N denotes the number of test sequences, d_i denotes the rank difference of video i, P and MOS denote the predicted scores and MOSs. Note that higher SROCC, PLCC and lower RMSE indicate better performance of IQA or VQA model.

Following the recommendation of Video Quality Experts Group (VQEG) [55], [56], a logistic function is performed before correlation computation, in order to map the predicted scores into a common scale:

$$P_i = \varepsilon_1\left(\frac{1}{2} - \frac{1}{1+e^{\varepsilon_2(S_i-\varepsilon_3)}}\right) + \varepsilon_4 S_i + \varepsilon_5, \tag{23}$$

where $\varepsilon_i (i = 1, 2, 3, 4, 5)$ are fitting parameters to minimize the square difference between subjective and predicted scores, S_i is the predicted scores and P_i is the mapped scores.

TABLE I
PERFORMANCE COMPARISON IN TERMS OF SROCC, PLCC AND RMSE ON THE CSCVQ DATABASE. THE TOP THREE MODELS FOR EACH CRITERION ARE HIGHLIGHTED IN RED, BLUE AND GREEN

		H.264			HEVC			HEVC SCC			Overall		
		SROCC	PLCC	RMSE	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
IQA	PSNR	0.8304	0.8591	7.2164	0.7826	0.8266	8.0915	0.7681	0.8027	8.2075	0.7935	0.8254	7.9524
	SSIM	0.7531	0.7706	8.9848	0.7003	0.7533	9.4555	0.6985	0.7233	9.5035	0.7196	0.7478	9.3520
	FSIM	0.9271	0.9271	5.2848	0.8762	0.9120	5.8967	0.8517	0.8912	6.2415	0.8771	0.9091	5.8666
	MS-SSIM	0.7676	0.7924	8.5997	0.7343	0.7877	8.8566	0.7385	0.7747	8.7018	0.7508	0.7842	8.7399
	IW-SSIM	0.8362	0.8642	7.0945	0.8060	0.8565	7.4205	0.7864	0.8267	7.7433	0.8113	0.8487	7.4503
	VIFP	0.8247	0.8563	7.2814	0.7715	0.8343	7.9289	0.7346	0.7557	9.0123	0.7744	0.8072	8.3148
	GSM	0.8441	0.8678	7.0059	0.8231	0.8721	7.0353	0.8152	0.8560	7.1154	0.8304	0.8643	7.0851
	VSI	0.9015	0.9243	5.3801	0.8696	0.9205	5.6193	0.8603	0.9014	5.9593	0.8755	0.9151	5.6784
	SQMS	0.8879	0.9091	5.8739	0.8203	0.8790	6.8563	0.7900	0.8245	7.7881	0.8283	0.8647	7.0763
	SVQI	0.8408	0.8810	6.6751	0.8021	0.8540	7.4794	0.7695	0.8073	8.1222	0.7991	0.8370	7.7081
	MS-RSDS (intra)	0.8819	0.9122	5.7774	0.8993	0.9253	5.4526	0.9081	0.9328	4.9595	0.8990	0.9217	5.4649
VQA	STMAD	0.8128	0.7856	8.7230	0.8485	0.8561	7.4306	0.8751	0.8606	7.0080	0.8434	0.8408	7.6258
	STRRED	0.8535	0.8499	7.4288	0.9214	0.9205	5.6183	0.9328	0.9347	4.8908	0.8896	0.8729	6.8736
	MOVIE	0.8703	0.8654	7.0658	0.8593	0.8737	6.9945	0.8522	0.8726	6.7216	0.8551	0.8689	6.9713
	MS-RSDS	0.9093	0.9280	5.2544	0.9255	0.9498	4.4996	0.9271	0.9490	4.3387	0.9202	0.9398	4.8138

B. Comparison With State-of-the-Art Models

We test 10 most popular IQA models adopting average pooling among frames and 3 latest VQA models, and compare them with our proposed MS-RSDS, to prove the effectiveness of our model with both intra- and inter-frames setting. All these tests are performed on the proposed CSCVQ database. Table I summarizes the performance comparison results. The top three results for each measurement criterion (i.e., SROCC, PLCC and RMSE) are highlighted in red, blue and green. In the IQA model comparison, we perform the MS-RSDS model excluding frame difference calculation which is denoted as MS-RSDS (intra). In the VQA model comparison, we used our complete MS-RSDS model including multiscale and frame difference strategies with the same parameters as used in the IQA model comparison. It is observed that the SROCC values of all prediction models fall within the interval of [0.7, 0.95]. The results show high consistency in three different kinds of compression distortions. These results prove the rationality of the CSCVQ database.

Among all models we compared, SQMS and SVQI are designed specially for SCIs. These two models achieved relatively fair predictions, but they were not particularly prominent. Both of the models aim at predicting perceived quality of images which suffer from various types of distortions, instead of focusing on compression loss in image quality. These models are better at extracting features from a mixture of computer-generated contents and nature scene contents. However, this situation is not common in screen content scenarios.

FSIM and VSI which are good at extracting texture features have better predictions. This proves that prominent points and lines are crucial factors which affect the quality assessment in screen content scenario. Our proposed MS-RSDS(intra) also obtains a high SROCC since it is a local salient feature based algorithm. We can observe that MS-SSIM gains a certain degree of improvement over SSIM, indicating that the multi-scale strategy is very effective to improve the performance of quality assessing model.

TABLE II
RUNTIME FOR VQA MODELS (ONE VIDEO IN SECONDS)

Model	STMAD	STRRED	MOVIE	MS-RSDS
Runtime	1779	313	6240	50

MS-RSDS shows a leap after performing frame difference calculation (SROCC from 0.8990 to 0.9202), which meets our expectations. In VQA model comparison, our proposed MS-RSDS ranks at the first place in all 3 criteria, while STRRED also gains a relatively high performance. It is noteworthy that we do not optimize the parameters of the complete MS-RSDS model. Instead, we used the parameters from the aforementioned MS-RSDS (intra) model without frame difference calculation. Therefore, there are reasons to believe that the complete MS-RSDS might achieve higher performance if it is specially tuned.

Considering the computational complexity and the time consumed, MS-RSDS has shown obvious advantages, as shown in Table II in detail. It is a low computational complexity model without complex frequency domain calculation. We observe that running VQA model is extremely time consuming, which leads to a problem of trading off between computational complexity and prediction accuracy. According to the possible application scenarios of the VQA algorithm,

TABLE III
TEST RESULTS OF VERIFYING THE CONTRIBUTION OF EACH FEATURE IN THE MS-RSDS MODEL

	SROCC	PLCC	RMSE
a. The complete model	0.9202	0.9399	4.8138
b. Performing the MS-RSDS (intra) model on all frames	0.8990	0.9217	5.4649
c. Performing the MS-RSDS (intra) model with a frame skip of 8	0.8944	0.9185	5.5711
d. Performing the MS-RSDS (intra) model with a frame skip of 15	0.8931	0.9127	5.7562
e. Performing the MS-RSDS (intra) model with mean instead of std with a frame skip of 8	0.8381	0.8445	7.5440
f. The RSDS (intra) model excluding multiscale processing with a frame skip of 8	0.8326	0.8317	7.8201
g. The RSDS (intra) model excluding weighting factor between scales with a frame skip of 8	0.8748	0.3389	13.253

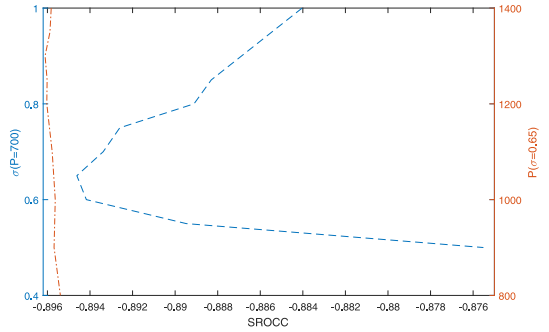


Fig. 6. Performance of the curve with various P and σ values. The blue curve shows the SROCC associated with different σ, under the fixed setting $P = 700$. The red curve shows the SROCC associated with different P, under setting $\sigma = 0.65$.

the time consumption in hours when evaluating a video with only several seconds is unacceptable.

C. Model Analysis

In order to verify the validity and contribution of each feature in our proposed MS-RSDS model, we separately tested the model in several parts. Table III shows some of the results of these tests. We first deliberately perform MS-RSDS (intra) on the original video frames rather than frame differences with average pooling which excludes inter-frame information. This evaluation method can reflect the image scale effect of the MS-RSDS algorithm and also provide a basis for later comparison with other IQA models.

For efficiency, instead of testing all the video frames, we perform MS-RSDS (intra) model with a frame skip of 8. To evaluate the effect of this, the sampling frequency was subsequently adjusted to every 15 frames. Obviously, the increase in sampling frequency has led to an improvement in the performance of the model. We believe that this should be the general case, since an increase in sampling frequency means that the model gets more valid information. However, such an increase may not be effective in all cases, because it also means a rise of computational complexity. Also, there is a possibility of performance degradation with an increase in sampling frequency when the number of video frames is limited in a small range. In practice, we can adjust the sampling frequency to obtain a balance between the model prediction accuracy and the computational complexity.

In the MS-RSDS model, using the standard deviation of the similarity map as the final evaluation score is an effective strategy. According to [28], different local structures might suffer from different degradations in gradient magnitude in distorted frames. These differences are not reflected by the mean value of the similarity map. To verify this, we likewise used the mean value in Eq. (15) instead of the standard deviation as the quality score. The result shows that the use of standard deviation significantly improves the performance of the model compared with using the mean.

We also analyzed the influence of adopting the multiscale strategy on our proposed model. We first performed RSDS (intra) only on the original frames. The test result shows that the prediction accuracy dropped noticeably from 0.8944 to 0.8326 when adopting single scale strategy instead of multiscale strategy. We also conducted an experimental investigation on how much does each scale contribute to the final scores. Due to the space constraints, detailed results are not fully listed. Table III lists the test result without weighting of each scale. The exploration shows that different scales have all contributed to the accuracy of prediction, but the degree of the contribution varies. In particular, medium scales contribute more than small or large scales.

D. Parameters Sensitivity

There are several aforementioned parameters, i.e., c in Eq. (13), p in Eq. (14), M and α in Eq. (16) to be determined in the MS-RSDS model. Moreover, the window size w and the standard deviation σ of the Guassian window $\mathbf{G}_k$ we mentioned in Eq. (12) are also need to be determined.

The approximate ranges of all these parameters come from empirically settings. We set the value of c to 0.0001 without any tuning. M is the multi-scale coefficient, i.e., the number of downsampling iterations from the origin videos. It represents how many levels of visual information are included in the model. In practice, different M can be chosen depending on the environmental factors and the sensitivity to details of subjects. Considering the size of the original frame, an excessively large M will cause the down-sampled image to exceed the perceptibility of the HVS, which leads to unstable prediction. In our implementation, we set M as 5, which is the same as the setting of the MS-SSIM model. α reveals the relative importance of different scales. Generally, HVS is more sensitive to information in intermediate scales and less

sensitive to information in large or small scales, for which reason we set α as $[0.15, 0.05, 0.05, 0.2, 0.55]$. Similar to the setting in the MS-SSIM model, this setting gives larger weights to the middle scales and smaller weights to the small and large scales.

In order to obtain better prediction, we have adjusted p, w and σ. Fig. 6 shows the SROCC results when the p and σ are tuned, respectively. In the experiment, we test the parameter sensitivity of p and σ with one parameter fixed and the other one varied. For the efficiency, the model is performed with a frame skip of 8 without multiscale processing. Results indicate that a minor change occurred when these parameters are tuned and the curve takes an extreme value at $p = 1300$ or $\sigma = 0.65$. The analysis reveals that the precise values of p, σ and w will not influence the overall performance of the proposed model in quite wide ranges. In our implementation, we set $p = 1300$, $w = 9$ and $\sigma = 0.65$.

V. Conclusion

In this article, we constructed a subjective Compressed Screen Content Video Quality (CSCVQ) database, which covers most common screen content application scenarios. The CSCVQ database provides a basis for future research on quality assessment of SCVs. We further proposed a VQA model, the Multiscale Relative Standard Deviation Similarity (MS-RSDS) model, specifically designed for quality assessment of SCVs. This model we proposed focuses on local relative standard deviation features, which can extremely influence subjects in accessing information in screen content scenario. We adopt multiscale strategy and frame difference calculation in our model which improve the performance explicitly. Experiments show that our proposed MS-RSDS model performs splendidly compared to other IQA/VQA models with relatively low computation complexity. There are several promising future research directions in line with this research. First, the relative standard deviation might also be useful in the quality assessment of screen content images, or other quality assessment scenarios such as point cloud quality assessment and solid experiments in these scenarios should be conducted to verify the effectiveness of this method. Second, the proposed MS-RSDS model is still a full reference model, no reference SCV quality model could be developed and verified on the constructed CSCVQ database.

References

[1] S. Wang and S. Dey, "Adaptive mobile cloud computing to enable rich mobile multimedia applications," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 870–883, Jun. 2013.

[2] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, Apr. 2020, Art. no. 211301.

[3] G. Zhai, X. Min, and N. Liu, "Free-energy principle inspired visual quality assessment: An overview," *Digit. Signal Process.*, vol. 91, pp. 11–20, Aug. 2019.

[4] M. Liu, K. Gu, G. Zhai, P. Le Callet, and W. Zhang, "Perceptual reduced-reference visual quality assessment for contrast alteration," *IEEE Trans. Broadcast.*, vol. 63, no. 1, pp. 71–81, Sep. 2017.

[5] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Mar. 2018.

[6] Q. Yang, Z. Ma, Y. Xu, L. Yang, W. Zhang, and J. Sun, "Modeling the screen content image quality via multiscale edge attention similarity," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 310–321, Mar. 2020.

[7] M. Hu *et al.*, "A wavelet-predominant algorithm can evaluate quality of Thz security image and identify its usability," *IEEE Trans. Broadcast.*, vol. 66, no. 1, pp. 140–152, Mar. 2020.

[8] X. Min, J. Zhou, G. Zhai, P. Le Callet, X. Yang, and X. Guan, "A metric for light field reconstruction, compression, and display quality evaluation," *IEEE Trans. Image Process.*, vol. 29, pp. 3790–3804, Jan. 2020.

[9] X. Liu, D. Zhai, D. Zhao, G. Zhai, and W. Gao, "Progressive image denoising through hybrid graph Laplacian regularization: A unified framework," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1491–1503, Jan. 2014.

[10] J. Zhou, O. C. Au, G. Zhai, Y. Y. Tang, and X. Liu, "Scalable compression of stream cipher encrypted images through context-adaptive sampling," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 11, pp. 1857–1868, Aug. 2014.

[11] X.-P. Zhang and M. D. Desai, "Adaptive denoising based on sure risk," *IEEE Signal Process. Lett.*, vol. 5, no. 10, pp. 265–267, Oct. 1998.

[12] T. K. Tsui, X. Zhang, and D. Androutsos, "Color image watermarking using multidimensional Fourier transforms," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 16–28, Feb. 2008.

[13] X. Liu, D. Zhao, J. Zhou, W. Gao, and H. Sun, "Image interpolation via graph-based Bayesian label propagation," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1084–1096, Dec. 2014.

[14] W. Zhu *et al.*, "Inter-palette coding in screen content coding," *IEEE Trans. Broadcast.*, vol. 63, no. 4, pp. 673–679, Jun. 2017.

[15] K. Gu, 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, Jun. 2016.

[16] S. Wang, L. Ma, Y. Fang, W. Lin, S. Ma, and W. Gao, "Just noticeable difference estimation for screen content images," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3838–3851, May 2016.

[17] "Methodology for the subjective assessment of the quality of television pictures," IETF, Geneva, Switzerland, ITU Recommendation BT.500-13, 2012.

[18] H. Yang, Y. Fang, W. Lin, and Z. Wang, "Subjective quality assessment of screen content images," in *Proc. 6th Int. Workshop Qual. Multimedia Exp. (QoMEX)*, 2014, pp. 257–262.

[19] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5462–5474, Nov. 2017.

[20] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, 2004.

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

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

[23] 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, Jan. 2011.

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

[25] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro- and macro-structures," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 3903–3912, Jan. 2017.

[26] 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.

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

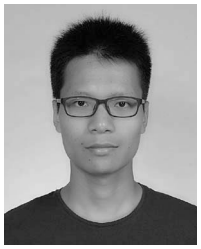
[28] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Dec. 2013.

[29] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Oct. 2010.

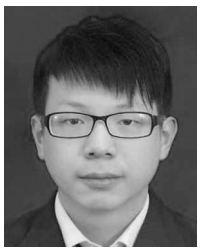
[30] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpeedQA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Jul. 2017.

[31] P. V. Vu and D. M. Chandler, "VIS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, 2014, Art. no. 013016.
[32] Z. Li *et al.*, *VMAF: The Journey Continues*, Netflix Technol., Los Gatos, CA, USA, 2018.
[33] S. Rimac-Drlje, M. Vranjes, and D. Zagar, "Influence of temporal pooling method on the objective video quality evaluation," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, 2009, pp. 1–5.
[34] C. Chen, M. Izadi, and A. Kokaram, "A perceptual quality metric for videos distorted by spatially correlated noise," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1277–1285.
[35] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. America A Opt. Image Sci. Vis.* vol. 24, no. 12, pp. B61–B69, Dec. 2007.
[36] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. 18th IEEE Int. Conf. Image Process.*, 2011, pp. 2505–2508.
[37] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Aug. 2013.
[38] 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, Aug. 2015.
[39] S. Wang, K. Gu, K. Zeng, Z. Wang, and W. Lin, "Objective quality assessment and perceptual compression of screen content images," *IEEE Comput. Graph. Appl.*, vol. 38, no. 1, pp. 47–58, May 2016.
[40] Z. Ni, L. Ma, H. Zeng, C. Cai, and K.-K. Ma, "Gradient direction for screen content image quality assessment," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1394–1398, Aug. 2016.
[41] K. Gu, J. Zhou, J. 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, Jun. 2017.
[42] K. Gu, J. Qiao, X. Min, G. Yue, W. Lin, and D. Thalmann, "Evaluating quality of screen content images via structural variation analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 10, pp. 2689–2701, Nov. 2017.
[43] X. Min, K. Gu, G. Zhai, M. Hu, and X. Yang, "Saliency-induced reduced-reference quality index for natural scene and screen content images," *Signal Process.*, vol. 145, pp. 127–136, Apr. 2018.
[44] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2049–2062, Jan. 2018.
[45] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Trans. Image Process.*, vol. 29, pp. 6054–6068, Apr. 2020.
[46] X. Min, G. Zhai, J. Zhou, X. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Trans. Image Process.*, vol. 29, pp. 3805–3819, Jan. 2020.
[47] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Trans. Multimedia Comput. Commun.*, vol. 13, no. 1, p. 6, 2017.
[48] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging HEVC screen content coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 50–62, Sep. 2016.
[49] J. Lei, D. Li, Z. Pan, Z. Sun, S. Kwong, and C. Hou, "Fast intra prediction based on content property analysis for low complexity hevc-based screen content coding," *IEEE Trans. Broadcast.*, vol. 63, no. 1, pp. 48–58, Nov. 2017.
[50] A. M. V. Dijk, J. B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," in *Proc. SPIE Int. Soc. Opt. Eng.*, vol. 2451, 1995, pp. 90–101.
[51] X. Min, G. Zhai, K. Gu, X. Yang, and X. Guan, "Objective quality evaluation of dehazed images," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2879–2892, Oct. 2019.
[52] X. Min *et al.*, "Quality evaluation of image dehazing methods using synthetic hazy images," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2319–2333, Feb. 2019.
[53] Y. Zhu, G. Zhai, X. Min, and J. Zhou, "The prediction of saliency map for head and eye movements in 360 degree images," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2331–2344, Sep. 2020.
[54] Y. Zhu, G. Zhai, and X. Min, "The prediction of head and eye movement for 360 degree images," *Signal Process. Image Commun.*, vol. 69, pp. 15–25, Nov. 2018.
[55] 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.
[56] J. Antkowiak *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment," in *Proc. Video Qual. Experts Group Meeting*, Ottawa, ON, Canada, Mar. 2000.

Teng Li (Member, IEEE) received the B.E. degree in information engineering from Shanghai Jiao Tong University, Shanghai, China, in 2015, where he is currently pursuing the Ph.D. degree with the Cooperative Medianet Innovation Center. His research focuses on visual quality assessment, perceptual signal processing, and machine learning.



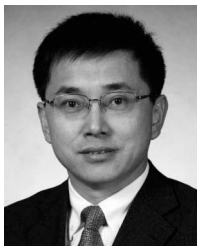
Xiongkuo Min (Member, IEEE) received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018. From January 2016 to January 2017, he was a visiting student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Postdoctoral Fellow with Shanghai Jiao Tong University. His research interests include visual quality assessment, visual attention modeling, and perceptual signal processing. He received the Best Student Paper Award at IEEE ICME 2016.



Heng Zhao received the B.E. degree in information engineering from Xidian University, Xi'an, China, in 2019. He is currently pursuing the master's degree with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, China. His research focuses on point cloud quality assessment for machine vision and perceptual signal processing.

Guangtao Zhai (Senior Member, IEEE) received the B.E. and M.E. degrees from Shandong University, Jinan, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. From 2008 to 2009, he was a visiting student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Postdoctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. His research interests include multimedia signal processing and perceptual signal processing. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012.

Yiling Xu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 1999, 2001, and 2004, respectively. She is a Full Researcher with the School of Electronic Information and Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. From 2004 to 2013, she was with Multimedia Communication Research Institute, Samsung Electronics Inc., Suwon, South Korea. Her main research interests include architecture design for next generation multimedia systems, dynamic data encapsulation, adaptive cross layer design, dynamic adaption for heterogeneous networks, and N-screen content presentation.



Wenjun Zhang (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1984, 1987, and 1989, respectively. From 1990 to 1993, he was a Postdoctoral Fellow with Philips Kommunikation Industrie AG, Nuremberg, Germany, where he was actively involved in developing the HD-MAC system. He joined the Faculty of Shanghai Jiao Tong University, in 1993, and became a Full Professor with the Department of Electronic Engineering in 1995. He holds more than 40 patents and authored or coauthored more than 90 papers in international journals and conferences. His main research interests include digital video coding and transmission, multimedia semantic processing, and intelligent video surveillance. He is a Chief Scientist with the Chinese National Engineering Research Centre of Digital Television (NERC-DTV), an industry/government consortium in DTV technology research and standardization, and the Chair of the Future of Broadcast Television Initiative Technical Committee. As the national HDTV TEEG project leader, he successfully developed the first Chinese HDTV prototype system in 1998. He was one of the main contributors to the Chinese Digital Television Terrestrial Broadcasting Standard issued in 2006. He has been leading team in designing the next generation of broadcast television system in China since 2011.