Satisfied-User-Ratio Modeling for Compressed Video

Xinfeng Zhang, *Member, IEEE*, Chao Yang, Haiqiang Wang, Wei Xu, and C.-C. Jay Kuo, *Fellow, IEEE*

***Abstract*— With explosive increase of internet video services, perceptual modeling for video quality has attracted more attentions to provide high quality-of-experience (QoE) for end-users subject to bandwidth constraints, especially for compressed video quality. In this paper, a novel perceptual model for satisfied-user-ratio (SUR) on compressed video quality is proposed by exploiting compressed video bitrate changes and spatial-temporal statistical characteristics extracted from both uncompressed original video and reference video. In the proposed method, an efficient video feature set is explored and established to model SUR curves against bitrate variations by leveraging the Gaussian Processes Regression (GPR) framework. In particular, the proposed model is based on the recently released large-scale video quality dataset, VideoSet, and takes both spatial and temporal masking effects into consideration. To make it more practical, we further optimize the proposed method from three aspects including feature source simplification, computation complexity reduction and video codec adaption. Based on experimental results on VideoSet, the proposed method can accurately model SUR curves for various video contents and predict their required bitrates at given SUR values. Subjective experiments are conducted to further verify the generalization ability of the proposed SUR model.**

***Index Terms*— Satisfied-user-ratio, just-noticeable difference, video quality, rate estimation, visual masking.**

I. Introduction

PERCEPTUAL video quality modeling has been becoming more and more important especially for compressed videos due to explosive increase of video services over Internet, which is restricted by limited and unstable bandwidth. Video quality models (VQM) can be utilized as guidance in video compression optimization to achieve the same perceptual quality with fewer bitrates [1]. They are also useful for video streaming to provide better quality-of-experience (QoE) adapting to bandwidth fluctuation [2], [3]. Since human eyes are the ultimate receivers of images and videos, subjective testing method is the most reliable for image and video quality evaluation [4], [5]. However, subjective testing method needs many subjects to provide their perceptual opinions on image or video quality. Thus, it is not applicable for practical video service since it is time-consuming and expensive. The Peak signal-to-noise ratio (PSNR) is still widely utilized in many video coding algorithms [6], [7].

In recent years, lots of objective image and video quality models [8], [9] are proposed to predict the subjective results from existing calibrated datasets. In most of compressed image and video quality datasets, each content are compressed into about five distortion levels, *i.e.,* worst, bad, average, good and excellent. A quality score determined by each subject is assigned to each distorted image or video. Therefore, the Mean Opinion Score (MOS) is utilized as the final subjective results. Different from pixel-level quality metric PSNR, Wang *et al.* proposed a Structural SIMilarity (SSIM) metric to approach human vision perceived quality of images by measuring the correlation of two patches from reference and distorted images, which achieved more correlated results with MOS compared to PSNR. Afterwards, various features are proposed to measure image and video quality in structural distortion instead of pixel-level difference [10], [11].

Besides structure-based metrics, characteristics of human visual system (HVS), *e.g.,* uneven distribution of perceived distortions, are also widely utilized in different IQA methods. For example, visual masking effects and saliency distribution are utilized in VSNR [12] and VSI [13] to approach subjective quality. For each image or video, these metrics can only provide a score which is correlated with the corresponding MOS result, while it is still difficult to analyze viewers' average satisfaction from a scalar score for test image or video against the corresponding reference one. For example, in video streaming application, we would like to switch into video presentations with lower bitrates when bandwidth decreases but we hope more viewers cannot perceive the quality degradation. However, from a quality score, we cannot find the ratio of satisfied users for the switched videos directly.

Just-noticeable difference (JND) is another widely accepted concept in HVS, which is the minimum detectable difference threshold for 50% viewers. In other words, there are 50% viewers are satisfied with test video quality compared with the reference one due to they cannot perceive quality difference between them. There are several JND based compressed image and video quality assessment datasets released in recent years, including MCL-JCI [14], MCL-JCV [15]

Manuscript received July 18, 2019; revised November 22, 2019; accepted January 6, 2020. Date of publication January 17, 2020; date of current version January 30, 2020. This work was supported by the joint research project between the University of Southern California and Huawei Technologies Company Ltd. under Grant YBN2017060057. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel L. Lau. *(Xinfeng Zhang and Chao Yang are co-first authors.) (Corresponding author: Xinfeng Zhang.)*

Xinfeng Zhang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: xfzhang@ucas.ac.cn).

Chao Yang is with the Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: yangchaoie@shu.edu.cn).

Haiqiang Wang and C.-C. Jay Kuo are with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: haiqianw@usc.edu; cckuo@sipi.usc.edu).

Wei Xu is with Huawei Corporation, Nanjing 210012, China (e-mail: xuwei35@huawei.com).

Digital Object Identifier 10.1109/TIP.2020.2965994

and VideoSet [5]. Different from MOS oriented datasets, in JND quality datasets, images or videos are compressed by continuous quality control parameters, *e.g.,* quality factors (QF) for JPEG and quantization parameters (QP) for H.264/AVC or HEVC. A subject should compare a pair of compressed images/videos with the same content, compressed by different parameters, and determine whether there is perceived quality difference. To speed up this procedure, a bisection search is often adopted to reduce the number of comparisons. MCL-JCI and MCL-JCV are two small-scale JND-based image and video quality datasets built by the Media Communications Lab at the University of Southern California, which aim to measure JND locations for JPEG compressed images and H.264/AVC compressed videos respectively. Based on the experience on JND subjective experiments, a large-scale JND-based video quality dataset, VideoSet, was built and analyzed in [5].

Different from MOS, the cumulative distribution of JND locations directly reflects the satisfied-user-ratio (SUR) by regarding that users are satisfied with test video quality against the corresponding reference one when their quality difference is not beyond JND thresholds. The SUR modeling is more useful in both video streaming and video coding applications by providing users satisfaction, which is more direct and instructive compared with MOS-based quality score. Moreover, it provides a relative quality indicator against the reference video instead of the original one. This is more useful in optimization for transcoding application where the original video is absent. In this paper, we focus on the prediction of SUR curves against video bitrates, and propose an efficient SUR model for flexible reference videos to well predict SUR curves of different video contents. The main contributions of the paper are summarized as follows:

- An efficient SUR modeling method is proposed for flexible reference videos under machine learning framework. Herein, spatial-temporal masking effects and saliency values are utilized to reflect HVS characteristics, and variations of objective video quality and bitrate are utilized to reflect rate-distortion characteristics of videos. These features are fed into machine learning framework to learn SUR curves based on VideSet dataset.
- We further optimize the proposed model by limiting feature extraction only from reference video and reducing the computational complexity to make it more suitable for practical applications. Furthermore, we analyze the rate-distortion relationship between video coding standards H.264/AVC and HEVC, and design a simple model translation strategy by adjusting SUR curves with a scale factor to adapt our model to HEVC videos.
- Extensive experiments and analysis on the large-scale JND video quality dataset, VideoSet, are provided to show the superiority of the proposed SUR model. Moreover, subjective experiments on other kind of video streams compressed by different video codecs and configurations are conducted to further show efficiency and good generalization ability of the proposed method.

The remainder of this paper is organized as follows. Section II introduces the related work about the JND based dataset VideoSet and video quality models. Section III first describes the overall framework of the proposed SUR model, and then introduces the detailed features utilized in our model. In Section IV, optimization scheme in speed and extension to different video codecs is introduced. Extensive experimental results and discussions are reported in Section V, and finally we conclude this paper in Section VI.

II. Related Work

A. VideoSet and Satisfied-User-Ratio

VideoSet is a large-scale JND video quality dataset jointly constructed by both academic and industrial organizations in 2017. In VideoSet, there are four resolution video sequences with a duration of 5s, *i.e.,* 1920×1080, 1280×720, 960×540 and 640×360. For each resolution, there are 220 video sequences, which are compressed by x264 codec [16] with QPs ranging from 1 to 51. In total, there are 880 original video sequences and the corresponding 44,880 compressed video sequences (compressed with $QP = 1, \ldots, 51$). For each video sequence, the first three JND points are searched in subjective experiments with 30+ subjects.

In the subjective experiment, a reference video is first fixed and then each subject should compare the quality of two video sequences, *i.e.,* reference and test videos, where the reference and test videos are from the same video content. The subject should determine whether they can perceive the quality difference between reference and test videos. If subject can perceive quality difference for current comparison, another test video with better quality will replace the current one. To speed up the JND search procedure, a binary search method is utilized. For example, for the first JND point, the original video is utilized as reference, and the first test video is its corresponding compressed version with $QP = 51$. If the subject can perceive quality difference between them, JND point should be in the QP range of [1,26] and the test video is replace by the video compressed with $\text{QP} = 26$. This procedure is performed repeatedly until arriving JND point. The reference for the second JND point is the corresponding video at the first JND point, and the video at the second JND point is utilized as the reference of the third JND point.

According to the study in [5], [17], JND locations can be approached by a Gaussian distribution in the form of,

$$X \sim (\mu_x, \sigma_x^2), \tag{1}$$

where μ_x and σ_x are sample mean and sample standard deviation, respectively. Since a subject is satisfied with the quality of test videos when their quality is perceptually the same as the reference ones, the satisfied-user-ratio for the i^{th} test video v_i can be formulated as,

$$S_i = 1 - \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_n(v_i), \tag{2}$$

where N is the amount of subjects. $\mathbb{1}_n(v_i)$ is the indicator function for the n^{th} subject, where 1 means the n^{th} subject can perceive quality difference and 0 otherwise. In fact, SUR in Eq. (2) is the empirical cumulative distribution function (CDF) of random variable X as given in Eq. (1).

From the above introduction, we can see that SUR is a more straightforward measure for video quality compared with MOS, especially for QoE-oriented video streaming application, which aims to maximize the perceived quality for viewers at video presentation switchover [18]. Although MOS also follows a Gaussian-like distribution, it only provides a subjective score for a given video. From MOS and its distribution, we still cannot determine which score corresponds to a JND location against it reference video. In other words, we still cannot tell the percentage of satisfied users from MOS and its distribution. The SUR directly reflects the ratio of satisfied users, and provides a new quality evaluation methodology as compared with MOS. In terms of applications, a video streaming service provider can generate video presentations according to SUR to make a certain ratio of users not perceiving any quality change. Therefore, we think SUR is a new quality indicator to directly provide the ratio of satisfied users. In this paper, we explore the SUR modeling and hope to provide a better QoE prediction strategy from this perspective.

B. Video Quality Models

In recent years, many video quality models are proposed to predict video quality score based on MOS-oriented VQA datasets. In [19], Wang *et al.* proposed a structural distortion based VQA method instead of previous error sensitivity based methods [20]. Herein, structure distortions are measured by weighted SSIM values of extracted regions from each frame, and the video quality is derived by pooling these frame-level scores according to motion vector lengths. In [21], a standardized video quality model (VQM) is designed for estimating video quality and its associated calibration techniques, which is released by Video Quality Experts Group (VQEG). In this VQM, seven independent quality indicators are calculated based on features extracted from spatial and temporal domains of both reference and test videos. These features include spatial gradients, contrast and motion of luminance components, and color feature vectors constructed based mean and variance of chroma components. The parameters of these quality indicators are learned from training data. In [22], MOtion-based Video Integrity Evaluation (MOVIE) index is proposed to measure both spatial contrast masking and motion quality along motion trajectories from the Gabor filtered reference and test images instead of original ones. These Gabor filters make video features extracted for VQA be optimally localized.

Instead of utilizing spatial and temporal features, in [23], Lin *et al.* proposed a fusion-based video quality assessment (FVQA) framework by integrating multiple VQA methods into learning framework, support vector machine (SVM). Four FVQA models are learnt from the JND-based VQA dataset MCL-V by dividing these videos into four categories according to their spatial information, temporal information and distortion types. Each FVQA model took advantage of five existing VQA indices as input of SVM, including ADM [24], VIF [25], FSIM [10], PSNR and SNR. This idea is further improved by exploring more efficient and complementary VQA indices, which evolves into the widely accepted video quality metric, Video Multimethod Assessment Fusion (VMAF) [26] developed by Netflix in cooperation with the Multimedia Communications Lab (MCL), University of Southern California. After a series of optimization, VMAF has achieved more correlated results with HVS compared with other VQA methods, and has been widely utilized in various video service applications, especially in compressed video quality assessment task.

The above VQA models mainly focused on MOS based video quality evaluation by providing objective scores for videos, but they cannot directly determine the degree of user satisfaction for test videos against given reference ones, especially when the reference videos are also distorted *e.g.,* compressed videos. JND can be regarded as a user satisfaction indicator at specific point. In [27], Liu *et al.* took the picture wise JND prediction as a multi-class classification problem and utilized deep learning technique to predict JND position for compressed images. Zhang *et al.* proposed DEEPQoE model using a 3D convolutional neural network to predict video quality of experience (QoE) based on VideoSet where JND is utilized as a QoE metric [28]. In [29], Huang *et al.* utilized spatial contrast information, temporal sensitive information and saliency information to construct a set of features to predict the first JND location taking original video as reference. In [30], Wang *et al.* proposed a SUR based video quality assessment method by predicting SUR curves for test videos against the corresponding original videos as reference. The proposed model was further extended to predict SUR curves against videos at the second and third JND locations as reference in [31]. In this method, quantization parameters and histograms of spatial-temporal masking effects extracted from specific spatial-temporal regions are utilized to construct a SUR model via SVM learning. However, there are two weaknesses for this method, 1) the prediction for QPs of JND locations is rarely utilized in practice, where bitrate is a more widely utilized indicator for video streams compared with QPs; 2) the prediction accuracy is not satisfied when applying its model to predict SUR and bitrates. To improve the prediction accuracy, Fan *et al.* utilized the convolutional neural network to predict SUR curves against the corresponding original image as reference in [32]. However, this work aimed to predict SUR of compressed images based on MCL-JCI dataset without considering temporal information of videos.

III. SUR Oriented Video Quality Modeling

A. Motivation and Framework

The existing research works mainly focused on MOS oriented video quality assessment to predict quality scores for given videos with or without reference. The predicted quality scores can be utilized to compare quality of videos, but it cannot tell how much quality difference is perceived by users. According to the JND concept, quality difference below certain threshold cannot be perceived by most of viewers. For example, if the quality score of one video is larger than the other one by 0.1, this quality difference may be meaningless because most of viewers cannot perceive the quality difference. Therefore, a new SUR oriented video quality model is more instructive and meaningful for video service applications.

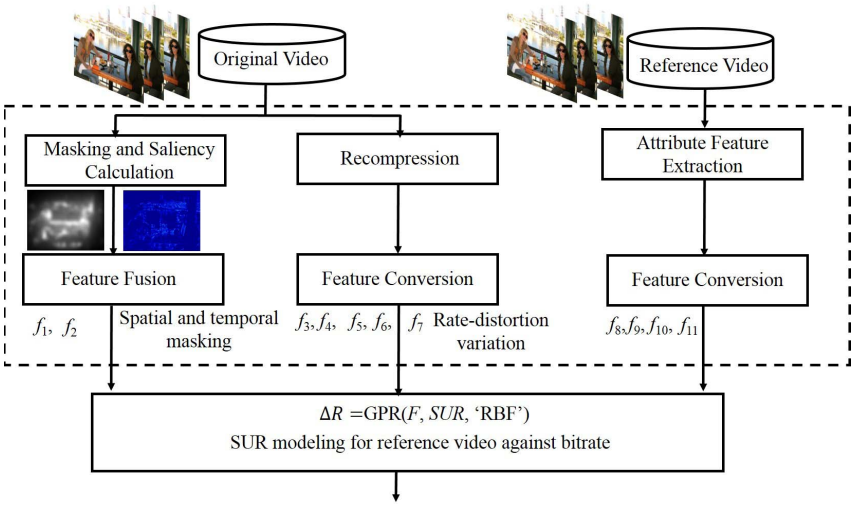


Fig. 1. The framework of the proposed model for SUR and bitrates of compressed videos.

In practice, SUR oriented video quality model is more desired in some video applications than MOS oriented VQM, *e.g.,* perceptual adaptive video streaming [33], [34] and transcoding [35]. In video streaming, to adapt bandwidth variation, each video is divided into small segments with a duration of 2∼10 seconds, which are compressed into multiple versions with various bitrates and stored in servers. Each compressed version is a presentation of the original video segment. Due to lack of SUR oriented VQMs, SUR is unknown for video presentations compared with a given reference. To reduce perceived quality fluctuation as possible, redundant video presentations may be generated in general for practical application which result in storage waste. On the other hand, QoE oriented video transcoding and transmission applications [36]–[38] also require efficient VQMs to optimize bitrate allocation for different video content, where the input videos may be compressed version instead of original ones.

To solve the above-mentioned problems, we propose a more efficient SUR oriented VQM, which not only can predict SUR curves by taking original video as reference but is also applicable for compressed reference videos. In the proposed method, three kinds of features are extracted from both original and reference videos as illustrated in Fig.1, including visual masking features (MF), recompression features (RF) and basic attribute features (BF). Herein, visual masking features mainly reflect perceived quality variation on different video content, recompression features aim to capture quality variation against bitrate changes, and basic attribute features are the intrinsic properties of reference videos. In the following, we will introduce the details on the three kinds of features and the proposed SUR model using Gaussian Process Regression (GPR).

B. Visual Masking Features

Based on previous research [9], [13], [39]–[41], we find that visual masking effect and saliency distribution have significant influence on perceptual image and video quality assessment. Herein, visual masking effect refers to that the visibility of distortions can be significantly reduced or completely removed for human visual system due to a spatially or temporally complex background. Although the visual masking mechanism remains not well understood, various visual masking models have been proposed, *e.g.*, spatial visual masking in [39], [42] and temporal visual masking in [40], [41]. They have been applied to image/video quality assessment with great success. For example, in [39], [40], spatial and temporal masking effects are utilized to measure compressed image and video quality. On the other hand, visual saliency refers to that humans have a remarkable ability to automatically pay more attention to salient regions of visual scenes than other regions. It implies that more distortions can be tolerated in non-salient regions while subtle distortions may be perceived in salient regions. In recent years, different saliency models have been proposed and applied in image/video quality assessment by adjusting pooling weights of distortions [43], [44]. However, in most cases, the two features are separately utilized in different models.

In fact, visual masking effect and saliency are two complementary features for human perceived quality since the same visual masking effect may show different perceptual quality influence on regions with various saliency values. In this work, we propose a weighted spatial visual masking effect as the first feature in our proposed SUR model, where the weights are derived according to visual saliency map. Firstly, the spatial visual masking effects are modeled by prediction errors from neighborhood using auto-regressive (AR) model [39], which are calculated via the following equation,

$$S(i, j) = \|y(i, j) - \mathbf{R}_{YX}\tilde{\mathbf{R}}_X^{+}\mathbf{x}(i, j)\|, \tag{3}$$

where $S(i, j)$ represents the value of spatial visual masking effect at location (i, j) for a given image, *e.g.,* as shown in Fig.2. The larger value of $S(i, j)$ means that the region contains more randomness and human vision is difficult to perceive its quality degeneration. $\mathbf{x}(i, j)$ is a vector composed of neighboring pixels (solid circles with black color) around $y(i, j)$ as shown in Fig. 2(b). $\mathbf{R}_{YX}$ and $\tilde{\mathbf{R}}_X$ are the covariance

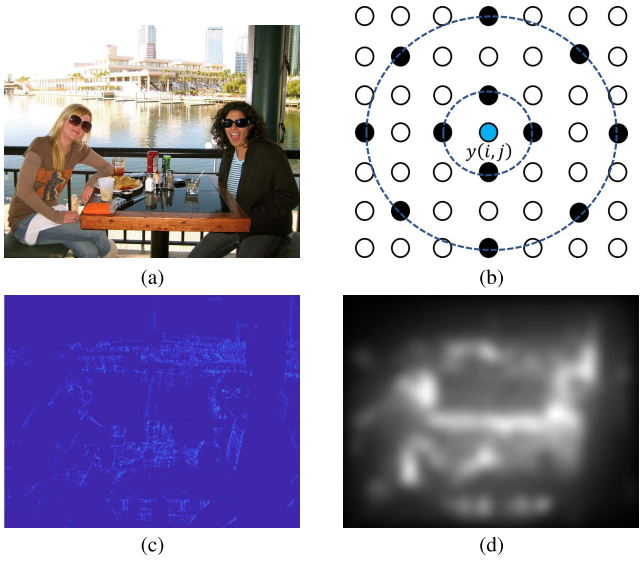


Fig. 2. Illustration of spatial visual masking effects and saliency map, (a) an image from MIT300 [45], (b) sample configuration for spatial visual masking calculation, where the blue circle represents predicting pixel $y(i, j)$ and the solid circles with black color compose the corresponding vector $\mathbf{x}(i, j)$, (c) spatial visual masking map, (d) saliency map calculated from GBVS [46].

matrices estimated from pixels in local region as,

$$\tilde{\mathbf{R}}_X = \frac{1}{N-1} X_S X_S^T, \quad \mathbf{R}_{YX} = \frac{1}{N-1} Y_S X_S^T. \tag{4}$$

$\tilde{\mathbf{R}}_X^+$ is the pesudo-inverse of $\tilde{\mathbf{R}}_X$, X_S and Y_S are the corresponding matrix and vector composed of local pixels. Fig. 2(c) shows an example of spatial visual masking map calculated from the image in Fig. 2(a). We can see that the areas with fine-grained textures show larger visual masking values, while areas with smooth and regular structures show smaller visual masking values.

However, from the formulation of spatial visual masking calculation in Eq. (3), the calculation of spatial visual masking only considers image local structures without considering human visual attention distribution from a global perspective. Although some areas with complex structures show strong visual masking effects, the distortions in these areas may be still perceived more obviously than those in areas with the same visual masking effects but smaller salient values. This is because viewers pay fewer attentions on these non-salient areas. As shown in Fig. 2(d), on one hand, although the objects on the desk are with strong visual masking effects, viewers pay more attentions to these areas which decreases the perceivable distortion thresholds compared with areas with the same visual masking effect but fewer attentions. On the other hand, the sky and lake areas are with weak visual masking effects, but fewer attention allocation will increase the perceivable distortion thresholds compared with areas with the same visual masking effects but more attentions. To deal with this problem, *i.e.* the coupling effect between saliency and spatial visual masking effect distribution, we pool the spatial visual masking values according to saliency map as the final spatial visual masking

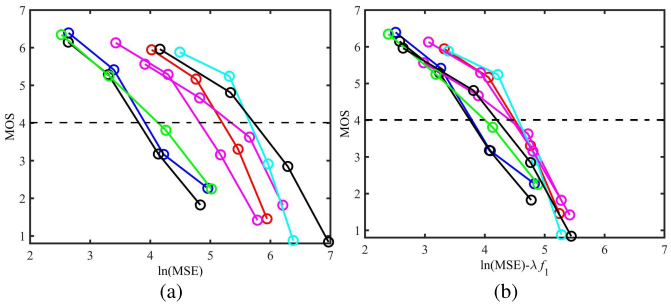


Fig. 3. Illustration of the relationship between MOS and MSE on IQA dataset, TID2008 [47], (a) MOS vs. *ln(MSE)*, (b) MOS vs. *ln(MSE)*-λf_1.

feature, which is calculated as,

$$f_1 = \frac{1}{N} \sum_{n=0}^{N-1} \Big(\sum_{i=0}^{H-1} \sum_{j=0}^{W-1} S(i, j) * (1 - Sal(i, j)) \Big), \tag{5}$$

where $Sal(i, j)$ represents the saliency value corresponding to sample $y(i, j)$, W, H and N are the video frame width, height and number of frames in a GOP (Group of Pictures). The Graph-based visual saliency (GBVS) model [46] is utilized in this paper.

To analyze efficiency of the weighted spatial visual masking feature, we utilize it as a factor to adjust sample distortion, mean square error (MSE). As illustrated in Fig.3(a), we calculate the *MSE* for JPEG images in IQA dataset TID2008, and show the relationship between MOS and *ln(MSE)*. We can see that the same MOS corresponds to various *MSE* for different images, which illustrates that *MSE* is not well correlated with perceived quality. The horizontal offsets among different MOS-*ln(MSE)* curves may be caused by the variations of spatial visual masking effects and saliency distribution for different image content. We further utilize the proposed saliency based weighted spatial visual masking feature f_1 to adjust the *MSE* as follows,

$$D_{ws} = ln(MSE \times e^{-\lambda f_1}) = ln(MSE) - \lambda f_1, \tag{6}$$

where f_1 is calculated from Eq.(5) with $\lambda = 2$. From Fig.3(b), we can see that the proposed feature can shrink the horizontal offsets of MOS-*ln(MSE)* curves for different images, which verifies that the weighted spatial visual masking feature is effective in describing HVS perceived quality. Furthermore, we also compare the proposed weighted spatial visual masking feature with the traditional spatial visual masking utilization strategy in [39] as,

$$D_s = ln(MSE \times e^{-\lambda \bar{S}}) = ln(MSE) - \lambda \bar{S}, \tag{7}$$

where $\bar{S}$ is the average of spatial visual masking value $S(i, j)$. In Table I, we show the three widely utilized correlation coefficients in IQA field, including PLCC, SRCC and KRCC, to compare correlations between MOS and various distortion metrics. We can see that the proposed feature achieves improvement compared with others on the three correlation coefficients, which further proves that the proposed feature is more efficient than individual spatial visual masking effect.

Besides spatial visual masking, temporal visual masking effect is also an important feature influencing both video

TABLE I

THE CORRELATION COEFFICIENTS BETWEEN MOS AND MSE AND ITS VARIATIONS WITH SPATIAL VISUAL MASKING AND SALIENCY ON IQA DATASET TID 2008

	PLCC	SRCC	KRCC
ln(MSE)	0.8703	0.8717	0.6847
$ln(MSE)+\lambda\bar{S}$	0.9539	0.9390	0.7781
$ln(MSE)+\lambda f_1$	0.9591	0.9494	0.7991

compression performance and QoE sensitivity. For video clips with fast and complex object motions, human visual system is difficult to perceive their quality changes, which is regarded as temporal visual masking effect in videos. In this work, to estimate temporal visual masking effect, we take video sequences as a linear dynamical system (LDS) and each frame is regarded as a state in temporal domain, which is similar with that in [40]. Then the state changes along with temporal axis can be described by the following equations:

$$\begin{cases} Y_{1:L} = AX_{1:L} + W_{1:L}, \\ X_{1:L} = CX_{0:L-1} + V_{0:L-1}, \end{cases} \tag{8}$$

where $X_{1:L} = [x_1, \cdots, x_l, \cdots, x_L]$ and $X_{0:L-1} = [x_0, \cdots, x_l, \cdots, x_{L-1}]$ are the state sequences of $Y_{1:L}$ and $Y_{0:L-1}$ respectively. Herein, y_l in $Y_{1:L}$ represents the vectorization of the l^{th} frame with M pixels, while x_l can be regarded as a compact presentation of y_l in spatial domain with K elements, $K < M$. The above system in Eq.(8) describes a temporal prediction model, where the matrix C reflects a spatial redundancy removal process and A reflects temporal motion information. W and V are the prediction errors for model A and C respectively.

The most straightforward technique calculating LDS parameters A and C is to utilize singular value decomposition (SVD) method to decompose a matrix of observations to yield an estimate of the underlying state sequence [48]. By applying SVD to the observations $X_{0:L-1}$, we can get $\mathcal{D} \approx U\Sigma V^T$, and obtain the estimates of C and X:

$$\hat{C} = U, \quad \hat{X} = \Sigma V^T. \tag{9}$$

Furthermore, the parameter A can be derived by solving the following optimization problem,

$$\hat{A} = \arg\min_A ||AX_{0:L-1} - X_{1:L}||_F^2 = X_{1:L}X_{0:L-1}^{\dagger}, \tag{10}$$

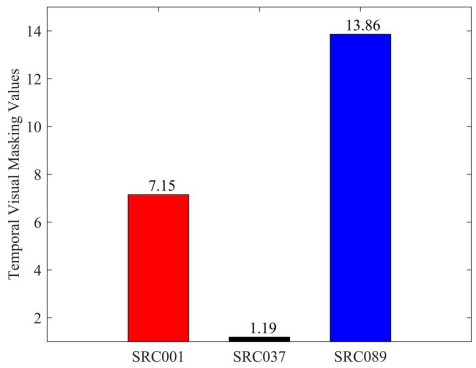
where $||\cdot||_F$ is the Frobenius norm and $\dagger$ is the Moore-Penrose inverse. Therefore, the temporal visual masking effect can be represented by the average of temporal prediction errors for a GOP, which is utilized as another feature in our work,

$$f_2 = \frac{1}{L}\sum_{t=0}^{L-1} mean(\|y_{l+1} - C_l A_l x_l\|), \tag{11}$$

where $mean(\cdot)$ denotes average operation.

Fig.4 illustrates temporal visual masking effect on different videos. As shown in Fig.4(a), the three video clips in VideoSet have distinct motion characteristics, regular and not fast motions in SRC001, very slow motions in SRC037 and irregular and fast motions in SRC089. We can see that the temporal visual masking feature in Eqn.(11) is effective in reflecting the video characteristics to HVS as shown in Fig.4(b).

(a) Video frames



(b) Temporal visual masking values

Fig. 4. (a) The video frames from top to down are extracted in every 6 frames from SRC001, SRC037, SRC089 respectively, (b) the average of temporal masking values for different video contents.

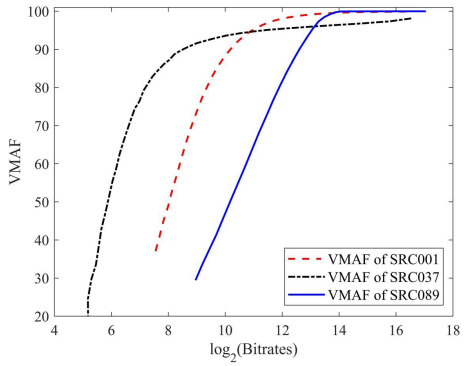


Fig. 5. The video quality variations along with bitrate for different video contents.

Recompression Features: The relationship between perceptual video quality and bitrates is highly correlated with video content. Although existing objective video quality metrics are not perfectly consistent with perceptual video quality, they have been able to predict the behavior of perceptual quality to some extent after years of efforts. For example, VMAF [49] has shown good correlation with human perceptual quality, especially for compressed videos. In Fig.5, we illustrate the relationship between VMAF scores and compressed bitrates for video clips, SRC001, SRC037 and SRC089 respectively. We can see that the VMAF-bitrate curves are benefit to predict SUR curves due to the correlation between VMAF and perceptual quality.

However, it is impractical to compress original videos with all the possible QPs due to computational costs. To reflect the above relationship, we propose to utilize the bitrates and VMAF scores of re-compressed video streams at four constant QPs *i.e.,* 22, 27, 32 and 37, as features in our prediction model. The four QPs can cover commonly utilized bitrate range, which are adopted in common test conditions of video coding standards. Thus, the four corresponding bitrates are further transformed into logarithmic domain as the final four features in our model, *i.e.,*

$$[f_3, f_4, f_5, f_6] = [log_2(R_{QP22}), log_2(R_{QP27}), log_2(R_{QP32}), log_2(R_{QP37})], \quad (12)$$

where R_{QP22} is the bitrate of compressed video stream at QP $=$ 22. Besides the bitrates, VMAF score variation is also utilized as the 7^{th} feature in the proposed prediction model, which is defined as,

$$f_7 = VMAF_{QP22} - VMAF_{QP37}. \quad (13)$$

Basic Attribute Features: Besides the above features reflecting characteristics of video content and HVS, the basic attributes of reference video are also adopted in our work, including the objective quality of reference video measured by VMAF (denoted as f_8), the frame rate and resolution (denoted as f_9 and f_{10}) and the bitrate of reference video in logarithmic domain (denoted as f_{11}).

C. SUR Modeling via Regression

In this work, the Gaussian Process Regression (GPR) [50], [51] is utilized to model the relationship between video bitrate and SUR. GPR is a nonparametric kernel-based probabilistic model to predict the value of a response variable y given input vector $\boldsymbol{x}$. A GPR model explains the response variable by introducing latent variables, $f(\boldsymbol{x}_i)$, $i = 1, 2, \cdots, n$, from a Gaussian process (GP) as follows,

$$y = h(\boldsymbol{x})^T \beta + f(\boldsymbol{x}), \quad (14)$$

where $f(\boldsymbol{x})$ is from a zero mean GP with covariance function, $k(\boldsymbol{x}, \boldsymbol{x}')$, *i.e.,* $f(\boldsymbol{x}) \sim GP(0, k(\boldsymbol{x}, \boldsymbol{x}'))$. $h(\boldsymbol{x})$ represents a set of basis functions that transform original feature vector $\boldsymbol{x}$ into a new feature space, and β is a vector of basis function coefficients. As such, the response y follows the Gaussian distribution,

$$P(y_i | f(\boldsymbol{x}_i), \boldsymbol{x}_i) \sim N(y_i | h(\boldsymbol{x}_i)^T \beta + f(\boldsymbol{x}_i), k(\boldsymbol{x}_i, \boldsymbol{x}_i')), \quad (15)$$

where the covariance function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is usually parameterized by a set of hyperparameters, Θ, *i.e.,* $k(\boldsymbol{x}_i, \boldsymbol{x}_j | \Theta)$.

In this paper, we take GPR to model the relationship between video bitrate and SUR, which is an efficient machine learning technique with small-scale samples. Based on our experiments, the radial-basis function (RBF) kernel also known as the "squared exponential" kernel can well model their relationship, which is defined as,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp\left(-\frac{1}{2} d(\boldsymbol{x}_i / m, \boldsymbol{x}_j / m)^2 \right), \quad (16)$$

where $m > 0$ is a length-scale parameter. For the parameters, β and Θ, we partition the VideSet into training subset and test subset with 80% and 20% samples respectively, and jointly utilize the original videos, the 1^{st} JND videos and the 2^{nd} JND videos as references and train the model parameters. Herein, the data partition is based on video content instead of all the compressed videos. That is to say, we choose 80% video content (*i.e.,* 704 video clips with various contents in four resolutions) and their related compressed version in different resolutions as training set. The remaining 20% videos (*i.e.,* 176 video clips with various contents in four resolutions) are utilized as test set. 5-fold validation is applied by selecting training and test sets randomly. The bitrate distribution of reference videos except for original ones in VideoSet are illustrated in Fig. 6, which covers most commonly utilized bitrates in practical applications.

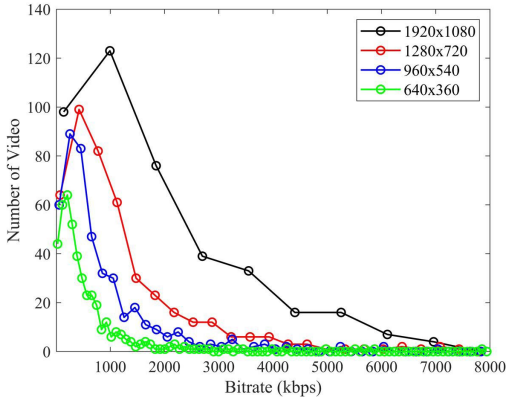


Fig. 6. The reference bitrate distribution in VideoSet (except for original videos), where each point represents the video number in the corresponding bitrate range (or a bitrate bin).

IV. Optimization for Practical Applications

For some practical applications, there are three limitations of the proposed SUR model, *i.e.,* the absence of original videos, the computational costs and diverse video codecs. We further optimize the proposed SUR model to make it applicable to practical scenarios by removing original videos and changing feature construction. First, due to the absence of original videos in some practical applications, VMAF score of reference videos is removed from the feature set. Second, for the re-compression features, reference videos are utilized to replace the original ones, and its recompression PSNR values at QP $=$ 22 and 37 are utilized instead of the corresponding VMAF to reduce computation costs.

To further speed up the computation, the calculations for spatial visual masking, temporal visual masking and saliency map are performed on downsampled reference videos adaptively according to their resolutions. In our implementation, the spatial and temporal visual masking calculations are performed on 8 times downsampled reference videos in both spatial and temporal dimensions for the height of video frame equal or larger than 720, and are performed on 4 times downsampled reference videos for the height of video frame smaller than 720. The downsampling ratios are determined based on the trade-off between performance and efficiency, as illustrated in Table.II. We can see that the prediction errors increase along with downsampling ratio continuously but the

TABLE II

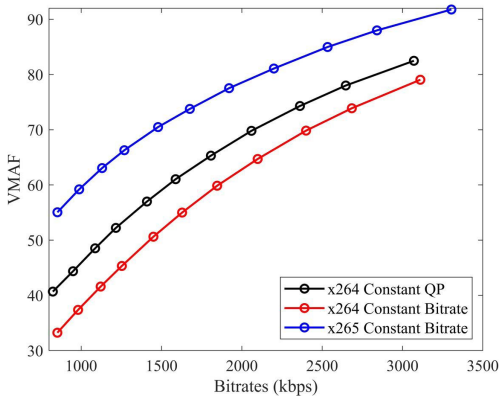
THE RUNNING TIME AND BITRATE PREDICTION ERRORS OF THE PROPOSED MODEL AT DIFFERENT DOWNSAMPLING RATIOS FOR SPATIAL AND TEMPORAL MASKING FEATURE EXTRACTION

Downsampling Ratio	1920×1080		960×540	
	Run time (s)	ARE	Run time (s)	ARE
1	460.3	0.05	66.5	0.047
2	188.2	0.063	43.0	0.060
4	30.4	0.077	9.8	0.071
8	19.8	0.087	6.1	0.085
16	19.4	0.110	5.1	0.108
32	18.6	0.135	4.8	0.131

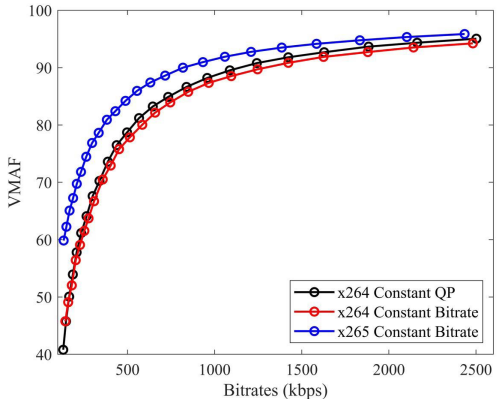
running time decrease becomes slow when the downsampling ratio beyond a specific value. Since saliency map calculation can afford more resolution reduction [46], [52], we downsample the height of video frames into 64 pixels by keeping the same aspect ratio as the original ones. In this paper, to distinguish the optimized version from the original one, we denote the optimized SUR model as *fast version*, and the original one as *high efficiency version*.

Furthermore, VideoSet is constructed based on videos compressed by x264 codec [53] under constant QP configuration, while practical video streams are compressed using rate control. The prediction performance may degenerate if we directly apply the proposed method to practical video streams. In addition, although H.264/AVC is a popular video coding standard at present, HEVC is the latest video coding standard with 50% bitrate saving compared with H.264/AVC and it is possible to be widely utilized in practical video services in the nearly future [54]. The trained model from x264 video streams may be inefficient to deal with HEVC video streams.

To solve the above mentioned problems deploying the proposed method in practical scenarios, we further analyze the quality variations over video bitrates for the following three cases, including x264 compression with constant QP, x264 constant bitrate compression with rate control and x265 constant bitrate [55] compression with rate control. Fig.7 shows the quality-bitrate relationships for video sequences SRC090 and SRC189 of resolution 1280×720 in VideoSet. We can see that although these coding configurations show different quality-bitrate performances, the quality change rates against bitrate are similar because x264 and x265 codecs follow the same video coding framework, *i.e.,* hybrid video coding framework. In this work, we design a simple but effective method to extend the proposed SUR model to other coding configurations by adjusting SUR curves via a scale factor. When applying the proposed model to x264 video streams with rate control, a scale factor, 1.1, is multiplied to the estimated bitrate. A scale factor, 1.3, is multiplied for x265 video streams with rate control. The two scale factors are designed according to their quality decreasing rate compared with compressed videos in VideoSet. As shown in Fig.7, we utilized the VMAF as approximate subjective quality and the average rate of quality change can be calculated in a common bitrate range, which is utilized to calculate the scale factors by regarding the scale factor of x264 codec with constant QP as 1.



(a) VideoSRC090



(b) VideoSRC189

Fig. 7. The quality variations over video bitrates for x264 codec with constant QP, x264 codec with rate control and x265 codec with rate control, respectively. The video resolution is 1280×720.

V. EXPERIMENTS AND ANALYSES

The proposed model formulates the relationship between SUR and video bitrate using GPR, and can predict required bitrates of input video at given SUR values, which is useful for video presentation generation in video streaming and trascoding applications. At the same time, it also can predict the SUR for a given compressed video. To verify the efficiency of the proposed model, both objective and subjective evaluation experiments are carried out and analyzed from the two aspects in this section.

Objective Evaluation: First, the prediction accuracy of high efficiency version and fast version are evaluated respectively, and the average relative errors (ARE) are utilized to evaluate their prediction accuracy. The ARE and its corresponding standard deviation (STD) are calculated using the following equations:

$$ARE = \frac{1}{N \cdot |\Omega|} \sum_{n=1}^{N} \sum_{m \in \Omega} \frac{|R'_{n,m} - R_{n,m}|}{R_{n,m}} \tag{17}$$

$$STD = \sqrt{\frac{1}{N \cdot |\Omega|} \sum_{n=1}^{N} \sum_{m \in \Omega} (\frac{|R'_{n,m} - R_{n,m}|}{R_{n,m}} - ARE)^2} \tag{18}$$

where $R'_{n,m}$ is the predicted bitrate from the n^{th} reference video at the SUR value m, and Ω is the SUR value set in

TABLE III

THE AVERAGE RELATIVE ERRORS OF BITRATE PREDICTION USING THE PROPOSED HIGH EFFICIENCY VERSION AND THE METHOD IN [30] IN 5-FOLD CROSS-VALIDATION ON DIFFERENT REFERENCE VIDEOS IN VIDEOSET

Resolution &Methods &Reference	The Proposed Method						The Method in [30]					
	Original Video		1^{st} JND Video		2^{nd} JND Video		Original Video		1^{st} JND Video		2^{nd} JND Video	
	ARE	STD	ARE	STD	ARE	STD	ARE	STD	ARE	STD	ARE	STD
640×360	0.032	0.049	0.05	0.058	0.074	0.106	0.117	0.103	0.121	0.107	0.128	0.126
960×540	0.03	0.044	0.049	0.054	0.042	0.062	0.109	0.111	0.116	0.117	0.127	0.143
1280×720	0.042	0.06	0.056	0.06	0.061	0.083	0.106	0.109	0.112	0.117	0.130	0.205
1920×1080	0.041	0.093	0.06	0.093	0.05	0.064	0.107	0.109	0.119	0.129	0.150	0.195
Average	0.036	0.062	0.054	0.066	0.057	0.079	0.110	0.108	0.117	0.118	0.134	0.167
Overall Results	ARE: 0.049 STD: 0.069						ARE: 0.120 STD: 0.131					

TABLE IV

THE MEAN ABSOLUTE ERROR OF SUR PREDICTION USING THE PROPOSED HIGH EFFICIENCY VERSION AND THE METHOD IN [30] IN 5-FOLD CROSS-VALIDATION ON DIFFERENT REFERENCE VIDEOS IN VIDEOSET

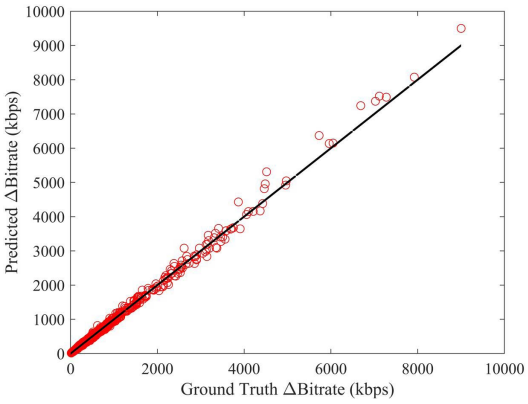
Resolution &Methods &Reference	The Proposed Method						The Method in [30]					
	Original Video		1^{st} JND Video		2^{nd} JND Video		Original Video		1^{st} JND Video		2^{nd} JND Video	
	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD
640×360	0.012	0.005	0.013	0.005	0.016	0.009	0.041	0.13	0.042	0.017	0.045	0.016
960×540	0.011	0.004	0.012	0.005	0.013	0.005	0.035	0.011	0.036	0.016	0.038	0.01
1280×720	0.012	0.005	0.015	0.005	0.015	0.007	0.036	0.009	0.038	0.017	0.04	0.015
1920×1080	0.013	0.008	0.015	0.008	0.014	0.006	0.036	0.01	0.039	0.019	0.041	0.013
Average	0.0120	0.0055	0.0138	0.0058	0.0145	0.0068	0.037	0.04	0.0386	0.0173	0.0236	0.0135
Overall Results	MAE: 0.0134 STD: 0.006						MAE: 0.0389 STD:0.0236					

VideoSet. $R_{n,m}$ is the ground-truth bitrate at SUR value m. The SUR prediction accuracy is directly measured using mean absolute error (MAE) and the corresponding standard deviation.

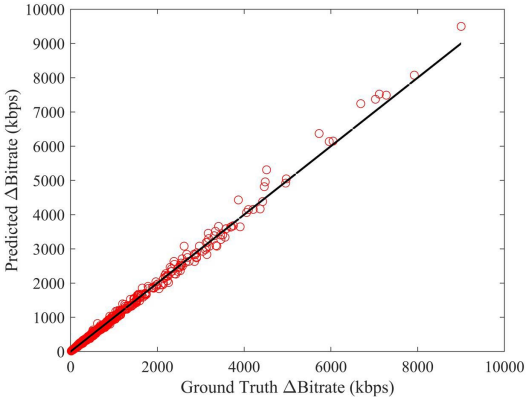
To fully evaluate the performance of the proposed method, we conduct the cross-validation using 80% training and 20% testing randomly and repeat the evaluation procedure 5 times to collect the average results. To avoid content biases, we randomly select the videos from the original contents instead of the huge of compressed videos, which is content-based selection. The original videos, 1^{st} JND videos and 2^{nd} JND videos are utilized as references to calculate the relative errors. To make the comparison comprehensive, we take 5-fold cross-validation and the average results are illustrated in Table III and Table IV. Table III shows the ARE and its standard deviation of predicted bitrates using the proposed method and the method in [30]. From the results, we can see that the proposed method achieves very high accuracy, and the AREs of predicted bitrates are only 3%$\sim$ 7.4% for different resolution videos. In particular, for low resolution videos, *i.e.,* $960\times$ 540 and 640×360, the proposed method achieves better performance than that on high resolution videos. We think there are two reasons. First, the existing masking and saliency features are both designed and evaluated on low-resolution videos, which may make the feature representation in low-resolution video more efficient. Second, the bitrate dynamic range of high-resolution videos is much larger than that of low-resolution video, which makes the relationship between bitrate and SUR on high resolution video more challenge.

For reference video with high bitrates, the proposed method performs much better than that of reference videos with low bitrates, *e.g.,* reference videos at the 2^{nd} JND position. This is because the dramatic degeneration of video quality make it more difficult to model than that at high bitrate scenarios. Moreover, the overall ARE of predicted bitrates for all the resolutions is only 4.9%, which verifies that the proposed method is robust to both video resolution and reference bitrate variation. We also show the prediction performance in [30], which is designed based on VideoSet for SUR modeling using QPs. We can see that overall ARE of the method in [30] is up to 12%, which is obviously inferior to our method due to the efficiency of our proposed features. Table IV shows the MAE and its standard deviation for the predicted SUR values using the proposed method and the method in [30]. The similar conclusions can be derived based on the superior performance of our proposed method over the method in [30].

Fig. 8 visualizes the relationship between the prediction bitrate and the ground-truth bitrate for all the test videos. We can see that most of the points are concentred around the diagonal line, which further shows the accuracy and robustness of the proposed method without any obvious outliers that are far from the ground-truth. The small standard deviation of the relative errors also verifies this conclusion from numerical perspective. To analyze the efficiency of the proposed three kinds of features, we conduct ablation study by adding features in sequence. Fig.9 illustrates the results of ablation study, where BF represents the basic attribute features, MF represents the visual masking features and RF refers to the re-compression features. Herein, BF(R) means that the bitrates are utilized as features directly, and BF($log_2(R)$) means that the bitrates in log domain are utilized as features. From the results,



(a) High efficiency version



(b) Fast version

Fig. 8. Scatter plot of predicted bitrate and the ground-truth bitrate.

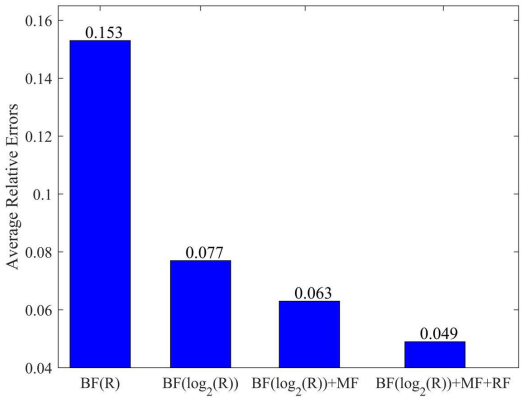


Fig. 9. Illustration for the proposed feature efficiency by ablation study.

we can see that the prediction errors are obviously reduced by combining the proposed features, MF and RF, with BF.

In Table V and Fig. 8(b), we further show the corresponding AREs for predicted bitrates using the proposed fast version. Although the optimization makes the prediction accuracy of the fast version degenerate a little, we think it is still acceptable for practical applications. The overall ARE for all the resolutions is 7.9%, and the worse case is 10.3% for the 640 × 360 reference videos at 2^{nd} JND position. On the other hand, we also compare the running time for the two versions on the Dell Server with two CPUs, Intel(R) Xeon(R) and CPU E5-2680 v4, where the frequency of CPU is 2.40 GHz running on Ubuntu 16.04.3 LTS. Compared with

TABLE V

THE AVERAGE RELATIVE ERRORS OF THE PROPOSED FAST VERSION USING 5-FOLD CROSS-VALIDATION ON DIFFERENT REFERENCE VIDEOS IN VIDEOSET

Resolution	Original Video		1^{st} JND Video		2^{nd} JND Video	
&Reference	ARE	STD	ARE	STD	ARE	STD
640 × 360	0.071	0.057	0.070	0.063	0.103	0.110
960 × 540	0.068	0.056	0.074	0.068	0.071	0.067
1280 × 720	0.072	0.066	0.078	0.067	0.081	0.087
1920 × 1080	0.081	0.087	0.097	0.097	0.082	0.075
Average	0.073	0.067	0.080	0.074	0.084	0.085
Overall Results	ARE: 0.079 STD: 0.075					

TABLE VI

REFERENCE VIDEO BITRATES FOR DIFFERENT RESOLUTION VIDEOS COMPRESSED BY X264 AND X265 CODECS (MBPS)

Resolution&Codec	x264	x265
1920 × 1080	6	4.9
1280 × 720	2	1.4
854 × 480	0.6	0.5

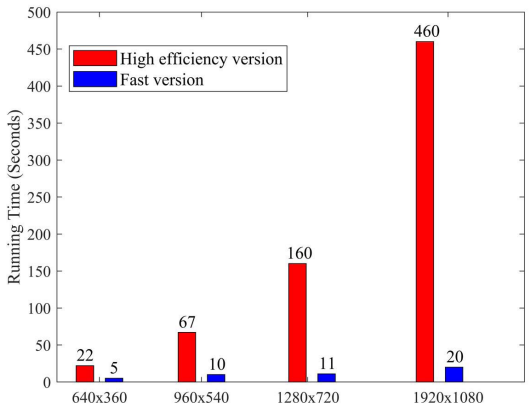


Fig. 10. The running time comparison between high efficiency version and fast version.

high efficiency version, the fast version not only get rid of the restriction of original videos, but also achieves significant running time reduction as shown in Fig. 10. In particular, for 1920 × 1080 video sequence, more than 95% running time reduction is achieved. Although the computation cannot satisfy real-time applications only based serial implementation on CPU yet, the proposed prediction model can be utilized to decrease the number of video presentations by generating them according to the estimated bitrates at given SUR values. This can be performed offline without real-time requirement. In addition, the proposed method can be further speeded up by utilizing popular parallel computation platforms, e.g., GPU.

Subjective Evaluation: To verify the practicability of the proposed method, we further carry out subjective experiments on commonly used bitrates and video codecs. The reference video bitrates for different resolutions and different video codecs are illustrated in Table VI, which are utilized in practical video services. To show the generalization ability of the proposed method, we not only utilize the x264 codec with rate control to compress both reference videos, but also utilize

TABLE VII
THE SATISFIED USER RATIO FOR DIFFERENT VIDEOS

Sequences	x264			x265		
	1920×1080	1280×720	854×480	1920×1080	1280×720	854×480
SRC008	0.82	0.76	0.79	0.73	0.67	0.42
SRC109	0.91	0.85	0.76	0.73	0.79	0.55
New093	0.82	0.73	0.61	0.67	0.64	0.36
New198	0.55	0.70	0.55	0.55	0.39	0.58
New271	0.70	0.97	0.73	0.61	0.67	0.55
New279	0.85	0.91	0.73	0.76	0.67	0.58
New401	0.85	0.79	0.79	0.91	0.70	0.64
New403	0.76	0.79	0.73	0.70	0.76	0.67
Average	0.78	0.81	0.71	0.71	0.66	0.54

TABLE VIII
THE BITRATE REDUCTION OF TARGET VIDEOS COMPARED WITH REFERENCE VIDEOS

Sequences	x264			x265		
	1920×1080	1280×720	854×480	1920×1080	1280×720	854×480
SRC008	51%	42.1%	29.4%	50.6%	43.4%	30.6%
SRC109	48.3%	45.3%	26.3%	38.8%	30%	18.7%
New093	48.1%	41.7%	22.2%	45.4%	38.4%	22.8%
New198	27.8%	25.6%	11.5%	18.5%	19.3%	9.4%
New271	45%	39.5%	25.2%	46.4%	45.2%	35.7%
New279	43.4%	40.1%	24%	46.9%	40%	33.3%
New401	42.8%	38.6%	21.1%	38.3%	38.7%	24.9%
New403	43.6%	44.4%	30.8%	38.6%	38.4%	31.3%
Average	43.8%	39.7%	23.8%	40.4%	36.7%	25.9%

x265 codec to compress them, which is an open source codec for HEVC. To approach practical applications, the two-pass rate control strategy is utilized to compress these videos at given bitrates. In addition, beside the video clips in VideoSet, SRC008 and SRC109, we also take another 6 videos in our subjective experiments, which are not included in VideoSet.

The original resolution of the test videos is 3840×2160, and the duration of each video is 5 seconds. These test videos are downsampled into 1920×1080, 1280×720 and 854×480, where 854×480 is a new resolution not included in VideoSet. These test videos are compressed into the specified bitrates according to Table VI as reference videos, V_r. Then, the fast version of the proposed method is applied to these reference videos to estimate the target bitrates at SUR value 0.75, and the reference videos are further re-compressed into the estimated target bitrates as test videos V_t. To find the satisfied user ratio, the two video clips in each video pair (V_r, V_t) are displayed sequentially but with random order. Moreover, video clip pairs with different resolutions are also displayed in random order. For each pair comparison, subjects should provide their determination from the following three options: the first one is better, the same quality and the second one is better. The display is DELL P2717H 27-inch Screen LED-Lit Monitor with 1080p resolution. 33 subjects attended the subjective experiments. They come from different countries including China, Indian, South Korea and USA, and their ages are from 22 to 40. There are 13 female and 20 male.

The satisfied user ratio for these video pairs are collected from the subjective experiments and shown in Table VII. Since test videos use fewer bitrates than the reference videos, their quality should not be better than the reference ones. Therefore, beside "the same quality", we regard the wrong quality selection, *i.e.*, test video is better than reference video, as "the same quality". The corresponding bitrate saving for each video sequence is shown in Table VIII. Herein, the column titles with x264 and x265 represent that the video streams are compressed by x264 and x265 codecs respectively.

From the results, we can see that the proposed SUR oriented model performs well on x264 video streams, especially on the videos with resolutions 1920×1080 and 1280×720, which are included in VideoSet. On average, test videos obtain about 43.8% and 39.7% bitrate saving compared with the reference ones and more than 75% users are satisfied with the quality of test videos. Although the SUR is a little lower than 75% for resolution 854×480 which is not included in VideoSet, for most of the viewers are still satisfied with these test videos, about 71%. For x265 video streams with resolution of 854×480 which is far from the training cases, the test videos also achieve more than 50% SUR on average, and there are only two videos with SUR value smaller than 50%. The subjective experiments have further verified the excellent generalization ability of the proposed model on various video contents, resolutions and different video codecs.

VI. CONCLUSION

In this paper, we have proposed a data-driven compressed video quality model directly oriented to formulate the relationship between satisfied user ratio and compressed video bitrates.

In the proposed method, a set of efficient features are designed to bridge the relationship between reference video and SUR variation. The Gaussian process regression method was utilized to model the relationship between video bitrate and SUR values. Moreover, although the proposed model was trained on videos compressed by x264 codec, it also can be extended to video streams compressed by the latest video standard HEVC. Both of the objective and subjective experimental results have verified the efficiency of the proposed method. In particular, the subjective experiments further illustrated the generalization ability of the proposed model on video streams compressed by various video codecs under different configurations from VideoSet. Compared with related work, the proposed method achieved obvious performance improvement on bitrate and SUR estimation. In the future work, we will further explore the perceptual quality adaptive video streaming algorithms based on the proposed SUR model to optimize the QoE for video streaming service.

REFERENCES

[1] X. Zhang, S. Wang, K. Gu, W. Lin, S. Ma, and W. Gao, "Just-noticeable difference-based perceptual optimization for JPEG compression," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 96–100, Jan. 2017.

[2] Z. Wang, K. Zeng, A. Rehman, H. Yeganeh, and S. Wang, "Objective video presentation QoE predictor for smart adaptive video streaming," *Proc. SPIE*, vol. 9599, Sep. 2015, Art. no. 95990Y.

[3] Z. Duanmu, K. Ma, and Z. Wang, "Quality-of-experience of adaptive video streaming: Exploring the space of adaptations," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1752–1760.

[4] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, and W. Gao, "Fine-grained quality assessment for compressed images," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1163–1175, Mar. 2019.

[5] H. Wang *et al.*, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 292–302, Jul. 2017.

[6] 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, Mar. 2017.

[7] J. Lei, X. He, H. Yuan, F. Wu, N. Ling, and C. Hou, "Region adaptive R-λ model-based rate control for depth maps coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1390–1405, Jun. 2018.

[8] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

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

[10] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

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

[12] D. Chandler and S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

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

[14] L. Jin *et al.*, "Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis," *Electron. Imag.*, vol. 2016, no. 13, pp. 1–9, Feb. 2016.

[15] H. Wang *et al.*, "MCL-JCV: A JND-based H. 264/AVC video quality assessment dataset," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1509–1513.

[16] L. Aimar *et al.* (2005). *X264-A Free H264/AVC Encoder*. Accessed: Apr. 1, 2007. [Online]. Available: http://www.videolan.org/developers/x264.html

[17] H. Wang, I. Katsavounidis, X. Zhang, C. Yang, and C.-C. J. Kuo, "A user model for JND-based video quality assessment: Theory and applications," *Proc. SPIE*, vol. 10752, Sep. 2018, Art. no. 107520M.

[18] L. Toni, R. Aparicio-Pardo, K. Pires, G. Simon, A. Blanc, and P. Frossard, "Optimal selection of adaptive streaming representations," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 2s, pp. 1–26, Feb. 2015.

[19] 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, Feb. 2004.

[20] Z. Wang *et al.*, "Why is image quality assessment so difficult?" in *Proc. ICASSP*, vol. 4, 2002, pp. 3313–3316.

[21] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

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

[23] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, "A fusion-based video quality assessment (FVQA) index," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA), Asia–Pacific*, 2014, pp. 1–5.

[24] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[25] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[26] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," in *The Netflix Tech Blog*, vol. 6. 2016.

[27] H. Liu *et al.*, "Deep learning-based picture-wise just noticeable distortion prediction model for image compression," *IEEE Trans. Image Process.*, vol. 29, pp. 641–656, 2020.

[28] H. Zhang, H. Hu, G. Gao, Y. Wen, and K. Guan, "DeepQoE: A unified framework for learning to predict video QoE," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[29] Q. Huang, H. Wang, S. C. Lim, H. Y. Kim, S. Y. Jeong, and C.-C. J. Kuo, "Measure and prediction of HEVC perceptually lossy/lossless boundary QP values," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017.

[30] H. Wang, I. Katsavounidis, Q. Huang, X. Zhou, and C.-C. J. Kuo, "Prediction of satisfied user ratio for compressed video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018.

[31] H. Wang, X. Zhang, C. Yang, and C.-C. J. Kuo, "Analysis and prediction of JND-based video quality model," in *Proc. Picture Coding Symp. (PCS)*, 2018.

[32] C. Fan *et al.*, "SUR-Net: Predicting the satisfied user ratio curve for image compression with deep learning," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.

[33] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 154–166, Feb. 2017.

[34] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards perceptually optimized end-to-end adaptive video streaming," 2018, *arXiv:1808.03898*. Accessed: Aug. 2018. [Online]. Available: https://arxiv.org/abs/1808.03898

[35] G. Gao, Y. Wen, and H. Hu, "QDLCoding: QoS-differentiated low-cost video encoding scheme for online video service," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.

[36] Y. Liu, S. Ci, H. Tang, Y. Ye, and J. Liu, "QoE-oriented 3D video transcoding for mobile streaming," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, no. 3s, pp. 1–20, Sep. 2012.

[37] A. El Essaili, Z. Wang, E. Steinbach, and L. Zhou, "QoE-based cross-layer optimization for uplink video transmission," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 1, pp. 1–22, Aug. 2015.

[38] N. Barman, S. Schmidt, S. Zadtootaghaj, M. G. Martini, and S. Möller, "An evaluation of video quality assessment metrics for passive gaming video streaming," in *Proc. 23rd Packet Video Workshop*, 2018, pp. 7–12.

[39] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, and C.-C.-J. Kuo, "Compressed image quality metric based on perceptually weighted distortion," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5594–5608, Dec. 2015.

[40] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, and C.-C.-J. Kuo, "Objective video quality assessment based on perceptually weighted mean squared error," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 9, pp. 1844–1855, Sep. 2017.

[41] L. K. Choi and A. C. Bovik, "Video quality assessment accounting for temporal visual masking of local flicker," *Signal Process., Image Commun.*, vol. 67, pp. 182–198, Sep. 2018.
[42] X. Yang, W. Lin, Z. Lu, E. P. Ong, and S. Yao, "Just-noticeable-distortion profile with nonlinear additivity model for perceptual masking in color images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP).*, vol. 3, Apr. 2003, p. III-609.
[43] Y. Zhang, Y. Fang, W. Lin, X. Zhang, and L. Li, "Backward registration-based aspect ratio similarity for image retargeting quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4286–4297, Sep. 2016.
[44] Y. Zhang, W. Lin, Q. Li, W. Cheng, and X. Zhang, "Multiple-level feature-based measure for retargeted image quality," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 451–463, Jan. 2018.
[45] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT, Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 2012.
[46] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 545–552.
[47] N. Ponomarenko *et al.*, "TID2008-A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.
[48] S. M. Siddiqi, B. Boots, and G. J. Gordon, "A constraint generation approach to learning stable linear dynamical systems," Dept. School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-ML-08-101, 2008.
[49] Netflix. (2018). *VMAF: Perceptual Video Quality Assessment Based on Multi-Method Fusion*. Accessed: Jun. 1, 2018. [Online]. Available: https://github.com/Netflix/vmaf
[50] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*. Berlin, Germany: Springer, 2004, pp. 63–71.
[51] M. Seeger, "Gaussian processes for machine learning," *Int. J. Neural Syst.*, vol. 14, no. 2, pp. 69–106, 2004.
[52] S. Yohanandan, A. Song, A. G. Dyer, and D. Tao, "Saliency preservation in low-resolution grayscale images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018.
[53] L. Aimar *et al. X264-A Free H264/AVC Encoder*. Accessed: Sep. 26, 2018. [Online]. Available: https://www.videolan.org/developers/x264.html
[54] G. J. Sullivan *et al.*, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[55] *x265: A Free HEVC Encoder*. Accessed: Sep. 26, 2018. [Online]. Available: https://www.videolan.org/developers/x265.html

Xinfeng Zhang (Member, IEEE) received the B.S. degree in computer science from the Hebei University of Technology, Tianjin, China, in 2007, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014. He is currently an Assistant Professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. He authored more than 100 refereed journal/conference papers. His research interests include video compression, image/video quality assessment, and image/video analysis. He received the Best Paper Award of the IEEE Multimedia 2018, the Best Paper Award at the 2017 Pacific-Rim Conference on Multimedia (PCM), and the Best Student Paper Award in the IEEE International Conference on Image Processing 2018.

Chao Yang received the B.E. and Ph.D. degrees from the School of Communication and Information Engineering, Shanghai University, Shanghai, China, in 2012 and 2017, respectively. From 2017 to 2018, he was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles. He joined the School of Communication and Information Engineering, Shanghai University, in 2019, where he is currently a Lecturer. His current research interests include video processing, video compression, and image quality assessment.

Haiqiang Wang received the B.S. and M.S. degrees in electrical engineering from Northwestern Polytechnical University, Xi'an, China, in 2010 and 2013, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California in 2018. He is currently a Senior Researcher with the Tencent Media Lab. His research interests include image and video processing, compression, and quality assessment. He was a recipient of the 2017 Capocelli Prize from the Data Compression Conference (DCC).

Wei Xu received the B.S. degree in computer science and technology and the Ph.D. degree in pattern recognition and artificial intelligence from the Nanjing University of Science and Technology, Nanjing, China, in 2009 and 2015, respectively. He is currently a Senior Engineer with Huawei Device Company, Ltd. His current research interests include image/video enhancement and image/video understanding.

C.-C. Jay Kuo (Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively. He is currently the Director of the Multimedia Communications Laboratory and a Distinguished Professor of electrical engineering and computer science with the University of Southern California, Los Angeles. He is the coauthor of about 290 journal articles, 940 conference papers, and 14 books. His research interests include digital image/video analysis and modeling, multimedia data compression, communication and networking, and machine learning. He is a fellow of the American Association for the Advancement of Science (AAAS) and The International Society for Optical Engineers (SPIE).