IEEE Open Journal of Signal Processing

Received 24 January 2021; revised 4 May 2021; accepted 9 June 2021. Date of publication 23 June 2021;
date of current version 22 September 2021. The review of this article was arranged by Associate Editor Dr. M. R Hashemi.

Digital Object Identifier 10.1109/OJSP.2021.3090333

RAPIQUE: Rapid and Accurate Video Quality Prediction of User Generated Content

ZHENGZHONG TU [1] (Graduate Student Member, IEEE), XIANGXU YU [1], YILIN WANG[2], NEIL BIRKBECK[2], BALU ADSUMILLI[2], AND ALAN C. BOVIK[1] (Fellow, IEEE)

[1] Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA
[2] YouTube Media Algorithms Team, Google LLC, Mountain View, CA 94043 USA

CORRESPONDING AUTHOR: ZHENGZHONG TU (e-mail: zhengzhong.tu@utexas.edu)

This work was supported by Google.

ABSTRACT Blind or no-reference video quality assessment of user-generated content (UGC) has become a trending, challenging, heretofore unsolved problem. Accurate and efficient video quality predictors suitable for this content are thus in great demand to achieve more intelligent analysis and processing of UGC videos. Previous studies have shown that natural scene statistics and deep learning features are both sufficient to capture spatial distortions, which contribute to a significant aspect of UGC video quality issues. However, these models are either incapable or inefficient for predicting the quality of complex and diverse UGC videos in practical applications. Here we introduce an effective and efficient video quality model for UGC content, which we dub the Rapid and Accurate Video Quality Evaluator (RAPIQUE), which we show performs comparably to state-of-the-art (SOTA) models but with orders-of-magnitude faster runtime. RAPIQUE combines and leverages the advantages of both quality-aware scene statistics features and semantics-aware deep convolutional features, allowing us to design the first general and efficient spatial and temporal (space-time) bandpass statistics model for video quality modeling. Our experimental results on recent large-scale UGC video quality databases show that RAPIQUE delivers top performances on all the datasets at a considerably lower computational expense. We hope this work promotes and inspires further efforts towards practical modeling of video quality problems for potential real-time and low-latency applications.

INDEX TERMS Video quality assessment, natural scene statistics, temporal, video compression, perceptual quality, user-generated content, image quality assessment, deep learning.

I. INTRODUCTION

Recent years have witnessed an explosion of user-generated content (UGC) captured and streamed over social media platforms such as YouTube, Facebook, TikTok, and Twitter. Thus, there is a great need to understand and analyze billions of these shared contents to optimize video pipelines of efficient UGC data storage, processing, and streaming. UGC videos, which are typically created by amateur videographers, often suffer from unsatisfactory perceptual quality, arising from imperfect capture devices, uncertain shooting skills, and a variety of possible content processes, as well as compression and streaming distortions. In this regard, predicting UGC video quality is much more challenging than assessing the quality of synthetically distorted videos in traditional video databases. UGC distortions are more diverse, complicated, commingled, and no "pristine" reference is available.

Traditional video quality assessment (VQA) models have been widely studied [1] as an increasingly important toolset used by the streaming and social media industries. While full-reference (FR) VQA research is gradually maturing and several algorithms [2], [3] are quite widely deployed, recent attention has shifted more towards creating better no-reference (NR) VQA models that can be used to predict and monitor the quality of authentically distorted UGC videos. One intriguing property of UGC videos, from the data compression aspect, is that the original videos to be compressed often already suffer from artifacts or distortions, making it difficult to decide the compression settings [4]. Similarly, it is of great interest

to be able to deploy flexible video transcoding profiles in industry-level applications based on measurements of input video quality to achieve even better rate-quality tradeoffs relative to traditional encoding paradigms [5]. The decision tuning strategy of such an adaptive encoding scheme, however, would require the guidance of an accurate and efficient NR or blind video quality (BVQA) model suitable for UGC [6].

Many blind video quality models have been proposed to solve the UGC-VQA problem [4], [6]–[19]. Among these, BRISQUE [8], GM-LOG [11], FRIQUEE [12], V-BLIINDS [9], and VIDEVAL [6] have leveraged different sets of natural scene statistics (NSS)-based quality-aware features, using them to train shallow regressors to predict subjective quality scores. Another well-founded approach is to design a large number of distortion-specific features, whether individually [20]–[22], or combined, as is done in TLVQM [13] to achieve a final quality prediction score. Recently, convolutional neural networks (CNN) have been shown to deliver remarkable performance on a wide range of computer vision tasks [23]–[25]. Several deep CNN-based BVQA models have also been proposed [18], [19], [26], [27] by training them on recently created large-scale psychometric databases [28], [29]. These methods have yielded promising results on synthetic distortion datasets [1], but still struggled on UGC quality assessment databases [30]–[32].

Prior work has mainly focused on spatial distortions, which have been shown to indeed play a critical role in UGC video quality prediction [6]. The exploration of the temporal statistics of natural videos, however, has been relatively limited. The authors of [6] have shown that temporal- or motion-related features are essential components when analyzing the quality of mobile captured videos, as exemplified by those in the LIVE-VQC database [30]. Yet, previous BVQA models that account for temporal distortions, such as V-BLIINDS and TLVQM, generally involve expensive motion estimation models, which are not practical in many scenarios. Furthermore, while compute-efficient VQA models exist, simple BVQA models like BRISQUE [8], NIQE [33], GM-LOG [11] are incapable of capturing complex distortions that arise in UGC videos. Complex models like V-BLIINDS [9], TLVQM [13], and VIDEVAL [6], on the contrary, perform well on existing UGC video databases, but are much less efficient, since they either involve intensive motion-estimation algorithms or complicated scene statistics features. A recent deep learning model, VSFA [19], which extracts frame-level ResNet-50 [34] features followed by training a GRU layer, is also less practical due to the use of full-size, frame-wise image inputs and the recurrent layers.

We have made recent progress towards efficient modeling of temporal statistics relevant to the video quality problem, by exploiting and combining spatial and temporal scene statistics, as well as deep spatial features of natural videos. We summarize our contributions as follows:

- We created a rapid and accurate video quality predictor, called RAPIQUE, in an efficient manner, achieving superior performance that is comparable or better than state-of-the-art (SOTA) models, but with a relative 20x speedup on 1080p videos. The runtime of RAPIQUE also scales well as a function of video resolution, and is 60x faster than the SOTA model VIDEVAL on 4 k videos.
- We built a first-of-its-kind BVQA model that combines novel, effective, and easily computed low-level scene statistics features with high-level deep learning features. Aggressive spatial and temporal sampling strategies are used, exploiting content and distortion redundancies, to increase efficiency without sacrificing performance.
- We created a new spatial NSS feature extraction module within RAPIQUE, which is a highly efficient and effective alternative to the popular but expensive feature-based BIQA model, FRIQUEE. The spatial NSS features used in RAPIQUE are suitable for inclusion as basic elements of a variety of perceptual transforms, leading to significant efficiencies which might also be useful when developing future BVQA models.
- We designed *the first* general, effective and efficient temporal statistics model (beyond frame-differences) that is based on bandpass regularities of natural videos, and which can also be used as a standalone module to boost existing BVQA methods on temporally-distorted or motion-intensive videos.

The rest of this paper is organized as follows. Section II reviews previous literature relating to video quality assessment models, while Section III unfolds the details of the RAPIQUE model. Experimental results and concluding remarks are given in Section IV and Section V, respectively.

II. RELATED WORK

A. TRADITIONAL BVQA MODELS

Many early BVQA/BIQA models have been 'distortion specific' in that they were designed to quantify a specific type of distortion such as blockiness [35], blur [36], ringing [20], banding [21], [37], [38], or noise [22], [39] in compressed images and videos. Recent high-performing BIQA/BVQA models are almost exclusively learning-based, operating by training sets of generic quality-aware features, which are combined to conduct quality predictions [7]–[10], [12]–[15]. Learning-based BVQA models are more versatile and generalizable than 'distortion specific' models, in that the selected features are broadly perceptually relevant, while powerful regression models can adaptively map the features onto quality scores learned from the data in the context of a specific application.

The most popular BVQA algorithms deploy perceptually relevant, low-level features based on simple, yet highly regular parametric bandpass models of scene statistics [40]. These natural scene statistics (NSS) models often are predictably altered by the presence of distortions [41], although they have more limited power to characterize complex, commingled distortions. Successful picture quality models of this type have been explored in the wavelet (BIQI [42], DIIVINE [7],

C-DIIVINE [43]), DCT (BLIINDS [44], BLIINDS-II [45]) and spatial domains (NIQE [33], BRISQUE [8]), and have been further extended to encompass natural bandpass space-time video statistics models [9], [46]–[48], among which the most well-known model is Video-BLIINDS [9]. Other extensions of empirical NSS include models of the joint statistics of the gradient magnitude and Laplacian of Gaussian (GM-LOG [11]), in log-derivative and log-Gabor spaces (DESIQUE [49]), as well as in the gradient domain of LAB color transforms (HIGRADE [10]). The FRIQUEE model [12] achieves excellent performance on UGC/consumer video/picture databases [29], [31], [32], [50], [51] by leveraging a bag of NSS features drawn from diverse color spaces and perceptually motivated transform domains.

Time-domain behavior is the key attribute that differentiates videos from still pictures. The perception of video correlates highly with motion and temporal change [52]. Amongst BVQA models, Video-BLIINDS [9] was the first to explore the use of (spatio-) temporal scene statistics of video using DCT coefficient statistics in the time-differenced domain. V-BLIINDS also involves calculating motion coherence and global motion features, which requires expensive motion estimation, to account for temporal masking effects.

Instead of using DCT transforms, Mittal *et al.* proposed a completely blind model called VIIDEO [47], which inspects the divisively normalized spatial statistics of frame differences. Bandpass filtering followed by divisive normalization was applied to frame differences, after which the inter-subband correlations are modeled over the temporal variation of the extracted generalized Gaussian parameters. As a highly experimental temporal-only model, VIIDEO includes no spatial features, hence does not perform well on natural UGC video datasets [30], [31].

Regarding the joint modeling of spatiotemporal statistics, Li *et al.* proposed to adopt 3D-DCT transforms of local space-time regions from videos to extract quality-aware features [46]. More recently, the authors of [53] leveraged 3D divisive normalization transformed (DNT) and spatiotemporal Gabor-filtered responses of 3D-DNT coefficients of natural videos. The 3D transforms adopted therein, however, are too expensive for practical use; neither have these models been observed to perform well on UGC datasets [30], [32].

Another intriguing and more practical approach to integrating temporal features into BVQA models is to design separable spatial-temporal statistics [4], [48], [54], [55]. Spatial features can be modified to capture temporal effects within BIQA models like BRISQUE, whereby simple frame-differences or spatially displaced frame-differences are deployed [4], [48], [56], [57].

A very recent feature-based BVQA model called TLVQM [13] uses a two-level feature extraction mechanism to achieve efficient computation of a set of impairment/distortion-relevant features. Unlike NSS-based models, TLVQM relies on a comprehensive set of highly crafted features that measure motion, specific distortion artifacts, and aesthetics. TLVQM requires that a large number of parameters be specified by the user, which may affect its general performance on datasets or scenarios it has not been exposed to. The model currently achieves very good performance on natural video quality databases at a reasonable complexity.

VIDEVAL [6] is currently the SOTA feature-based BVQA model on recent large-scale video dataset like KoNViD-1k [31] and YouTube-UGC [32]. It employs feature selection and fusion on top of efficacious NSS-based models as well as distortion-based features. It is also a very compact model as it only utilizes 60 features. However, it has not been observed to efficiently scale to high-resolution and high-framerate videos.

B. DEEP LEARNING-BASED BVQA MODELS

Deep convolutional neural networks (CNNs) have been shown to deliver standout performance in a wide variety of low-level computer vision applications [17], [23], [25], [58]. Recently, the release of several large-scale psychometric visual quality databases [29]–[32], [51] have sped the application of deep CNNs to perceptual video and image quality modeling. To conquer the limits of small data size, researchers have either proposed to conduct patch-wise data-augmentation during training [59]–[61], or to pretrain deep nets on larger visual sets like ImageNet [62], then fine tune on target quality databases. Several authors report remarkable performance on synthetic distortion databases [63], [64] or on naturally distorted databases [29], [51].

Deep CNN models have also been employed for natural video quality prediction. Kim *et al.* [26] proposed a deep video quality assessor (DeepVQA) to learn spatio-temporal visual sensitivity maps via a deep CNN and a convolutional aggregation network. The V-MEON model [65] leveraged a multi-task CNN framework which jointly optimizes a 3D-CNN for feature extraction and a codec classifier, and using fully-connected layers to predict video quality. Zhang *et al.* [27] leveraged transfer learning to develop a general-purpose BVQA framework based on weakly supervised learning and a resampling strategy. In the VSFA model [19], the authors applied a pre-trained image classification CNN as a deep feature extractor, then integrated the frame-wise deep features using a gated recurrent unit and a subjectively-inspired temporal pooling layer, reporting leading performance on several natural video databases [31], [50], [66]. The authors then built an enhanced version of VSFA, dubbed MDVSFA [67], by employing a mixed datasets training strategy on top, training a single VQA model on multiple datasets, and reporting superior performance on publicly available datasets. Several other popular CNN-based BVQA models [19], [26], [27], [65], [67] produce accurate quality predictions on legacy (single synthetic distortion) video datasets [1], [68], but struggle on recent in-the-wild UGC databases [31], [50], [66].

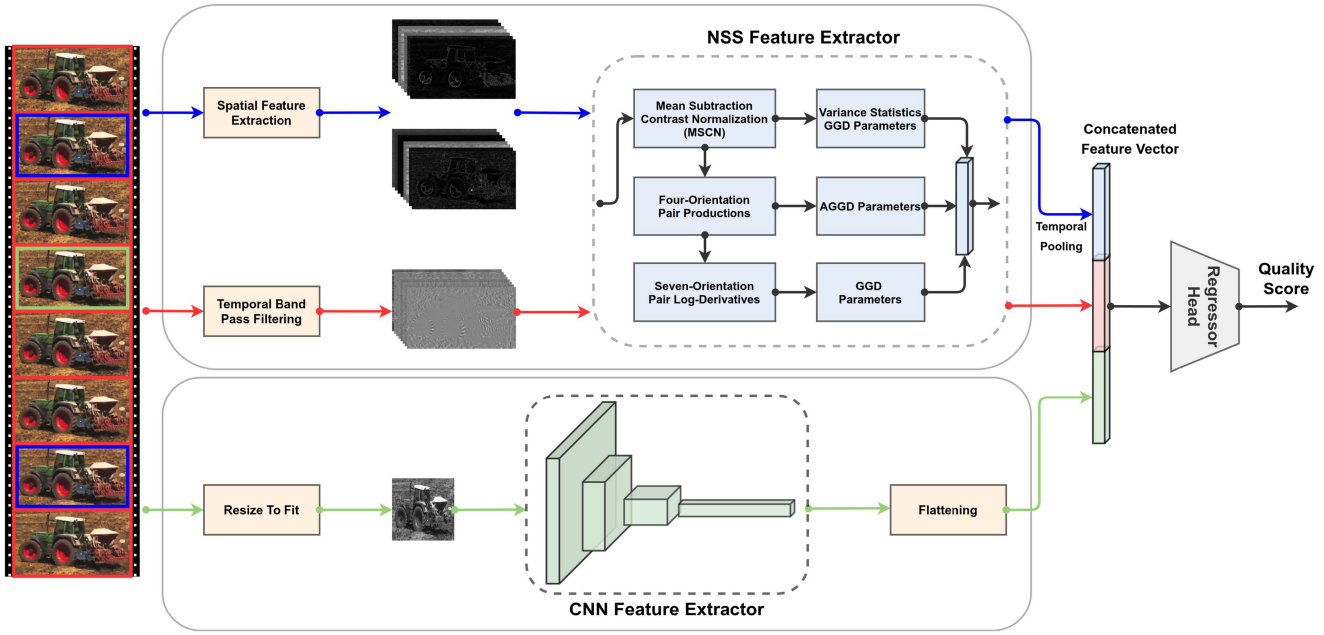


FIGURE 1. **Schematic overview of the proposed RAPIQUE model. Top block shows the spatial and temporal NSS feature extraction branch, while bottom block depicts the CNN feature extraction flow. The final feature vector is simply concatenated from the extracted spatial and temporal NSS and the CNN features, which is further used to train a regressor head.**

III. RAPID AND ACCURATE VIDEO QUALITY EVALUATOR (RAPIQUE)

Prior statistics-based video quality models have been shown to be capable of capturing complex UGC distortions, such as FRIQUEE [12], TLVQM [13], and VIDEVAL [6]. However, these models are subject to time-consuming executions since either expensive motion estimation or high-order statistical features are required. CNN models are able to efficiently capture high-level features, which have also been observed to be useful quality indicators [19], albeit directly applying a CNN on high-resolution video frames is expensive. Here we propose an efficient two-branch framework, as depicted in Fig. 1, which combines quality-aware, low-level NSS features with high-level, semantics-aware CNN features. The NSS features operate on higher-resolution spatial and temporal bandpass feature maps, while the CNN feature extractor is applied on a resized low-resolution frames for practical considerations. We also adopt a sparse frame sampling strategy when extracting features, which further accelerates the runtime. We present the details of RAPIQUE in the following.

A. NATURAL SCENE STATISTICS

It has been observed that the spatial wavelet/subband coefficients of natural images exhibit strong regularities (Gaussianity) following a divisive normalization transform (DNT) [7]. A simple but effective form of divisive normalization, called mean subtraction and contrast normalization (MSCN), has been observed to accurately characterize image naturalness in multiple feature transforms [8], [10], [12], [49]. We develop NSS-based features following the methodology of FRIQUEE [12], which leverages multiple perceptually-relevant feature transforms to extract a large number of statistical features. RAPIQUE uses simple yet effective low-order bandpass statistics, achieving comparable performance as the complex and time-consuming high-order features used in FRIQUEE. We were inspired by the successful and efficient basic features developed as products of spatially-adjacent MSCN responses, and log-derivative statistics in BRISQUE [8] and DESIQUE [49], respectively. Specifically, let $Y(i, j)$ be a given intensity image or a transformed feature map. The MSCN operator is applied on $Y(i, j)$ to further decorrelate and Gaussianize the local pixels:

$$\hat{Y}(i, j) = \frac{Y(i, j) - \mu(i, j)}{\sigma(i, j) + C}, \tag{1}$$

where, (i, j) are spatial indices and $C = 1$ is a constant that prevents instabilities caused by having a small variance in the denominators. The factors $\mu(i, j)$ and $\sigma(i, j)$ are the weighted local mean and standard deviation within a spatial window centered at location (i, j) calculated by:

$$\mu(i, j) = \sum_{k=-K}^{K} \sum_{\ell=-L}^{L} w_{k,l} Y(i - k, j - \ell) \tag{2}$$

$$\sigma(i, j) = \sqrt{\sum_{k=-K}^{K} \sum_{\ell=-L}^{L} w_{k,\ell} [Y(i - k, j - \ell) - \mu(i, j)]^2}, \tag{3}$$

where $\boldsymbol{w} = \{w_{k,\ell} | k = -K, \ldots, K, \ell = -L, \ldots, L\}$ is a 2D isotropic, truncated, unit-volume Gaussian weighting function. We used $K = L = 3$ in our implementations.

It has been empirically observed that the MSCN coefficients of images or video frames have characteristic statistical properties that are altered by the presence of distortion, and therefore, quantifying these deviations can help enable the prediction of perceived quality [8], [40]. A well-known model is the generalized Gaussian distribution (GGD) with zero mean [8]:

$$f(x;\alpha,\sigma^2)=\frac{\alpha}{2\beta\Gamma(1/\alpha)}\exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right), \tag{4}$$

where $\beta=\sigma\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$ and $\Gamma(\cdot)$ is the gamma function: $\Gamma(a)=\int_0^\infty t^{a-1}e^{-t}dt$. The two parameters are the shape α and the spread σ, of the zero-mean symmetric GGD, which are estimated using a popular moment-matching based method [69]. These are used as features to predict perceptual quality.

Another statistical observation is that the sample distributions of products of pairs of neighboring pixels in the MSCN coefficient map along four directions - horizontal (H) ($\hat{Y}(i,j)\hat{Y}(i,j+1)$), vertical (V) ($\hat{Y}(i,j)\hat{Y}(i+1,j)$), main-diagonal (D1) ($\hat{Y}(i,j)\hat{Y}(i+1,j+1)$), and secondary-diagonal (D2) ($\hat{Y}(i,j)\hat{Y}(i+1,j-1)$) also exhibit a regular statistical structure, which are well-modeled as following a zero mode asymmetric generalized Gaussian distribution (AGGD) [8], [47]:

$$f(x;\nu,\sigma_l^2,\sigma_r^2)=\begin{cases}\frac{\nu}{(\beta_l+\beta_r)\Gamma(\frac{1}{\nu})}\exp\left(-\left(\frac{-x}{\beta_l}\right)^{\nu}\right) & x<0\\ \frac{\nu}{(\beta_l+\beta_r)\Gamma(\frac{1}{\nu})}\exp\left(-\left(\frac{x}{\beta_r}\right)^{\nu}\right) & x>0,\end{cases} \tag{5}$$

where

$$\beta_l=\sigma_l\sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}}\quad\text{and}\quad\beta_r=\sigma_r\sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}}. \tag{6}$$

An AGGD model has four parameters: ν controls the shape of the distribution, while (σ_l,σ_r) are scale parameters that control the spread along each side of the mode; and η is the mean of the distribution, given by $\eta=(\beta_r-\beta_l)\frac{\Gamma(2/\eta)}{\Gamma(1/\eta)}$.

Apart from the second-order pair-product statistics, we also extract another supplementary set of features by modeling the log-derivative statistics of neighboring MSCN coefficient pairs as introduced in [49]. Specifically, the absolute pixel values of $\hat{Y}(i,j)$ are first logarithmically transformed:

$$Z(i,j)=\log[|\hat{Y}(i,j)|+0.1], \tag{7}$$

then seven types of log-derivative statistics (Eqs. (8)) along six paired orientations - horizontal (H: $\nabla_x Z(i,j)$), vertical (V: $\nabla_y Z(i,j)$), main-diagonal (MD: $\nabla_{xy}Z(i,j)$), secondary-diagonal (SD: $\nabla_{yx}Z(i,j)$), horizontal-vertical (HV: $\nabla_x\nabla_y Z(i,j)$), and two combined-diagonals (CDs: $\nabla_{cx}\nabla_{cy}Z(i,j)_1$, $\nabla_{cx}\nabla_{cy}Z(i,j)_2$), are modeled as GGD, respectively, after which the estimated GGD parameters are gathered as additional statistical features for learning the eventual quality predictor.

$$\begin{aligned}
D_1&:\nabla_x Z(i,j)=Z(i,j+1)-Z(i,j)\\
D_2&:\nabla_y Z(i,j)=Z(i+1,j)-Z(i,j)\\
D_3&:\nabla_{xy}Z(i,j)=Z(i+1,j+1)-Z(i,j)\\
D_4&:\nabla_{yx}Z(i,j)=Z(i+1,j-1)-Z(i,j)\\
D_5&:\nabla_x\nabla_y Z(i,j)=Z(i-1,j)+Z(i+1,j)\\
&\quad -Z(i,j-1)-Z(i,j+1)\\
D_6&:\nabla_{cx}\nabla_{cy}Z(i,j)_1=Z(i,j)+Z(i+1,j+1)\\
&\quad -Z(i,j+1)-Z(i+1,j)\\
D_7&:\nabla_{cx}\nabla_{cy}Z(i,j)_2=Z(i-1,j-1)+Z(i+1,j+1)\\
&\quad -Z(i-1,j+1)-Z(i+1,j-1)
\end{aligned} \tag{8}$$

For each pair log-derivative feature map, a single scale NSS model is used to derive two parameters (α,σ) by fitting a GGD distribution using the same moment-matching procedure, yielding a total of 14 additional features.

The variance field (or 'sigma' field) in Eq. (3) has been previously shown to provide effective quality-aware features deriving from the same NSS/retinal model [10], [12]. We extract two additional quantities from the variance field (Eq. (3)): the mean ϕ_σ and square of the reciprocal of the coefficient of variation (CoV):

$$\phi_\sigma=\frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\sigma(i,j) \tag{9}$$

where the CoV is $\rho=(\phi_\sigma/\omega_\sigma)^2$ and where

$$\omega_\sigma=\sqrt{\frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[\sigma(i,j)-\phi_\sigma]^2}. \tag{10}$$

In order to visualize how these NSS regularities are perturbed by *UGC* distortions, we selected four pictures ranging from high quality to low quality - `10004473376.jpg` (MOS=3.82), `6462096609.jpg` (MOS=2.93) from KoNIQ-10k [51] and `t4.bmp` (MOS=35.8), `12.bmp` (MOS=15.9) from LIVE-IQC [29], as shown in Fig. 2. Fig. 3 plots the histograms of the MSCN and variance field coefficients on these images of diverse perceptual quality. Note that it is extremely difficult to isolate specific distortion types on authentically distorted UGC pictures, since several complex, commingled distortions usually co-exist, hence it is difficult to predict how a given GGD histogram will vary in the presence of quality degradations. However, we may still observe in Fig. 3 that MSCN and sigma coefficients can differentiate images having different perceptual qualities. In this regard, the estimated parameters from these distributions are good indicators of perceptual quality for UGC pictures or videos.

We also plotted the histograms of adjacent pair products and log-derivative statistics in Fig. 4 and 5, respectively. It may be observed that the product statistics exhibit similar

(a) img1 (MOS=3.82, best quality) (b) img2 (MOS=2.93, good quality)

(c) img3 (MOS=35.8, bad quality) (d) img4 (MOS=15.9, worst quality)

FIGURE 2. Exemplar test images exhibiting four categories of quality: (a) img1 (best) and (b) img2 (good) are two good-quality images from KonIQ-10 k (MOS range: [1,5]) [51], while (c) img3 (bad) and (d) img4 (worst) are two bad-quality pictures from CLIVE (MOS range: [0100]) [29].

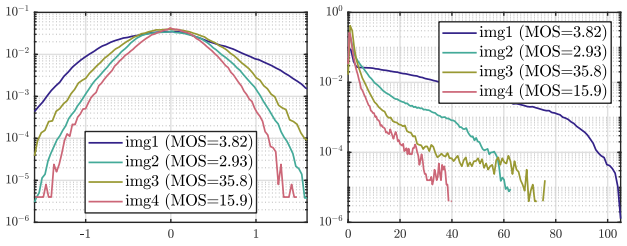


FIGURE 3. Histograms of MSCN (left) and variance map (right) of the four images shown in Fig. 2.

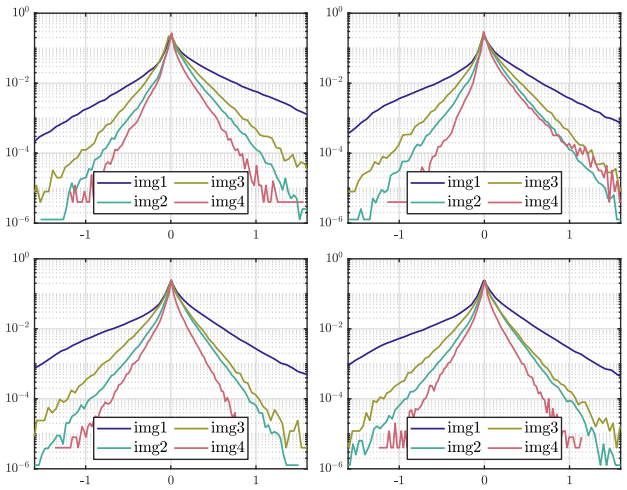


FIGURE 4. Histograms of four-orientation (H, V, D1, D2) MSCN pair-production of the four images shown in Fig. 2.

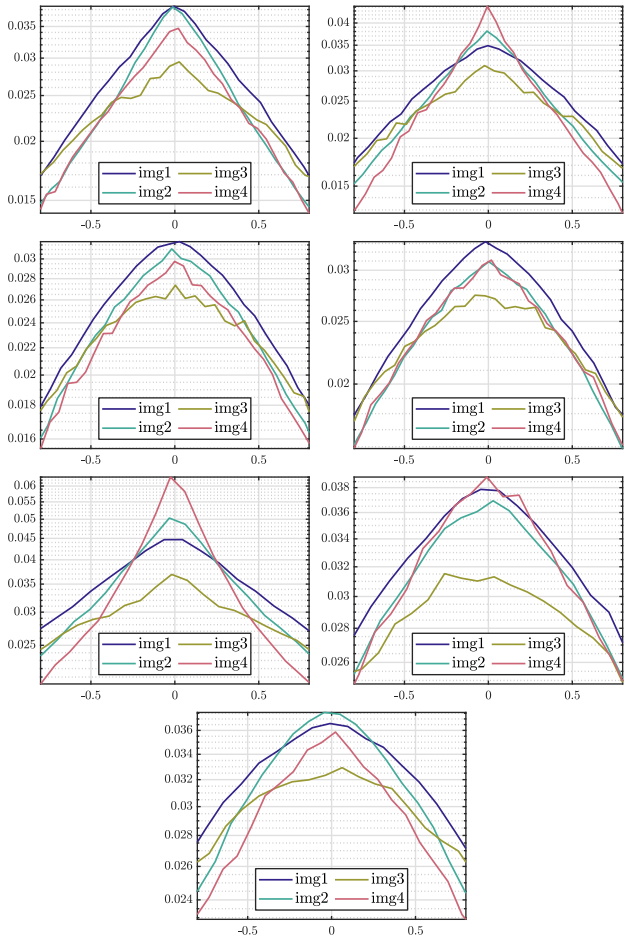


FIGURE 5. Histograms of seven types of MSCN paired log-derivative (Eqs. (8)) of the four images shown in Fig. 2.

TABLE 1. Summary of the Proposed NSS-34 Feature Extraction Module

INDEX	DESCRIPTION	COMPUTATION PROCEDURE
$f_1 - f_2$	(α, σ)	Fit GGD to MSCN coefficients
$f_3 - f_4$	$(\phi_\sigma, \omega_\sigma)$	Compute statistics on 'sigma' map
$f_5 - f_8$	$(\nu, \eta, \sigma_l, \sigma_r)$	Fit AGGD to H pairwise products
$f_9 - f_{12}$	$(\nu, \eta, \sigma_l, \sigma_r)$	Fit AGGD to V pairwise products
$f_{13} - f_{16}$	$(\nu, \eta, \sigma_l, \sigma_r)$	Fit AGGD to D1 pairwise products
$f_{17} - f_{20}$	$(\nu, \eta, \sigma_l, \sigma_r)$	Fit AGGD to D2 pairwise products
$f_{21} - f_{22}$	(α, σ)	Fit GGD to D_1 pairwise log-derivative
$f_{23} - f_{24}$	(α, σ)	Fit GGD to D_2 pairwise log-derivative
$f_{25} - f_{26}$	(α, σ)	Fit GGD to D_3 pairwise log-derivative
$f_{27} - f_{28}$	(α, σ)	Fit GGD to D_4 pairwise log-derivative
$f_{29} - f_{30}$	(α, σ)	Fit GGD to D_5 pairwise log-derivative
$f_{31} - f_{32}$	(α, σ)	Fit GGD to D_6 pairwise log-derivative
$f_{33} - f_{34}$	(α, σ)	Fit GGD to D_7 pairwise log-derivative

behavior as the MSCN coefficients, in that their shapes are significantly altered as a function of quality. These statistics better characterize correlations introduced or lost by distortion as compared to first-order MSCN and variance features. The seven types of log-derivative statistics shown in Fig. 5 exhibit distinct deviations against distortion, and all are effective at capturing quality variations on UGC pictures. Therefore, it is informative to include all these statistical features in the final prediction model.

B. SPATIAL FEATURES

We built a basic statistical feature extraction module using the NSS features mentioned in the previous section, as summarized in Table 1. Given an input image or feature map,

it extracts two features (α, σ) from the MSCN transforms, two features $(\phi_\sigma, \omega_\sigma)$ from the variance field, 16 features $4 \times (\nu, \eta, \sigma_l, \sigma_r)$ from the AGGD fit of MSCN adjacent pair products along 4 directions, and 14 features $7 \times (\alpha, \sigma)$ from GGD fits of paired log-derivatives along 7 directions, yielding a total of 34 features, which we dub NSS-34 for simplicity.

We also hypothesized that the NSS-34 operator is able to extract valuable color quality information if applied in a chromatic space, which we deem to be a more unified and efficient approach than crafting specific, complex features in different feature spaces, as is done in FRIQUEE [12]. The proposed NSS-34 feature set is based on fast, low-level statistics derived from GGD and AGGD models only, which we will show to deliver superior efficiency in the experimental section.

The gradient magnitude (GM) of a video frame is defined as the root mean square of directional gradients along two orthogonal spatial directions. GM is computed by convolving with a linear filter such as the Roberts, Sobel, or Prewitt. We utilized the Sobel kernels:

$$h_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} \text{ and } h_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{11}$$

whereby the GM of an image or frame $I(i, j)$ is calculated by:

$$GM = \sqrt{(I * h_x)^2 + (I * h_y)^2}, \tag{12}$$

where $*$ denotes the convolution operator.

It has been observed that two dimensional difference-of-Gaussian (DoG) or Laplacian-of-Gaussian (LoG) operators well-characterize the multiscale receptive fields of retinal ganglion cells [70]. We also extract two bandpass maps, using LoG and DoG, and extract their corresponding NSS-34 features, respectively. The LoG of image I is:

$$LoG = I * h_{LoG}, \tag{13}$$

where the LoG kernel is defined as:

$$\begin{aligned} h_{LoG} &= \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) g_\sigma(x, y) \\ &= \frac{x^2 + y^2 - 2\sigma^2}{2\pi\sigma^6} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \end{aligned} \tag{14}$$

where $g_\sigma(x, y)$ is an isotropic Gaussian function with scale parameter σ. We used a window size of 9×9 for LoG filtering.

While the GM and LoG are used by RAPIQUE to amplify high-frequency responses relating to local frame structures, the DoG is configured to capture mid-frequencies, expressive of structure at larger bandpass scales. The DoG response is defined as the difference of the responses of two Gaussian filters with different standard deviations

$$DoG = I * g_{\sigma_1} - I * g_{\sigma_2} = I * (g_{\sigma_1} - g_{\sigma_2}). \tag{15}$$

To avoid redundant information between the LoG and DoG, only the first level of an N-level DoG decomposition with

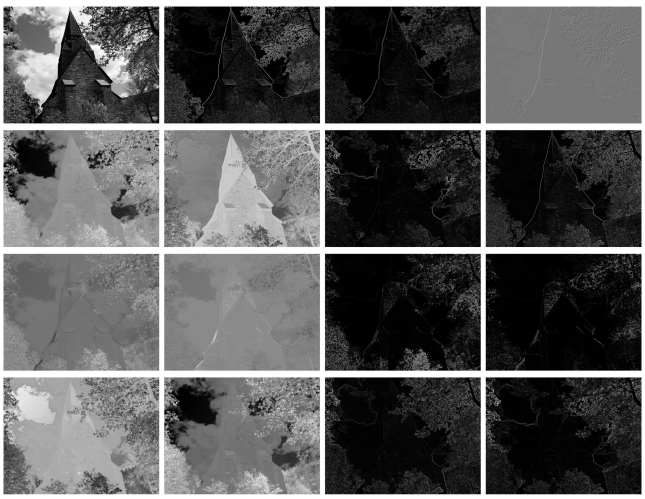


FIGURE 6. Extracted feature maps defined in Section III-B. 1st row: *Y*, *GM*, *LoG*, *DoG*; 2nd row: O_2, O_3, GMO_2, GMO_3; 3 rd row: *BY*, *RG*, *GMBY*, *GMRG*; 4th row: *A*, *B*, *GMA*, *GMB*.

$k = 1.6$, $\sigma_i = k^{i-1}, i = 1, \ldots, N-1$ is utilized. Fig. 6 shows the differences between the GM, LoG, and DoG responses on a sample video frame. Overall, the four luma channel feature maps (Y, GM, LoG, DoG) (where $Y = 0.299R + 0.587G + 0.114B$) are fed into the NSS-34 module to obtain useful statistical features.

Most previous BVQA models have overlooked the importance of chromatic features, whereas recent work [6], [10], [55], [66] has shown the efficacy of color components for UGC video quality prediction. Previous efforts on the chromatic statistics of quality models involve opponent color spaces such as YIQ/YUV [15], [71], $O_1O_2O_3$ [72], LMS [12], [72], perceptual color spaces like CIELAB [10], [12], [73], HSI [12], [74], Yellow color [12], and "colorfulness" features [6], [13], [75]. Here we deploy perceptually relevant color transforms from RGB frames (where $R(i, j), G(i, j), B(i, j)$ are red, green, and blue channels) to $O_1O_2O_3$, red-green (RG), and blue-yellow (BY) as follows:

$$\begin{bmatrix} O_1 \\ O_2 \\ O_3 \end{bmatrix} = \begin{bmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.60 & 0.17 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{16}$$

and

$$\begin{aligned} \mathcal{R}(i, j) &= \log[R(i, j) + 0.1] - \mu_R \\ \mathcal{G}(i, j) &= \log[G(i, j) + 0.1] - \mu_G, \\ \mathcal{B}(i, j) &= \log[B(i, j) + 0.1] - \mu_B \end{aligned} \tag{17}$$

where μ_R, μ_G, and μ_B are the average values of $\log[R(i, j) + 0.1]$, $\log[G(i, j) + 0.1]$, and $\log[B(i, j) + 0.1]$, respectively, over each entire frame. Then the RG and BY opponent color

space values are

$$\begin{aligned} \hat{L} &= (\mathcal{R} + \mathcal{G} + \mathcal{B})/\sqrt{3} \\ BY &= (\mathcal{R} + \mathcal{G} - 2\mathcal{B})/\sqrt{6}. \\ RG &= (\mathcal{R} - \mathcal{G})/\sqrt{2} \end{aligned} \tag{18}$$

We also included chroma maps A, B from the most widely used CIELAB perceptual color space [10], [12], which can be converted from RGB via CIEXYZ [76]. Note that we extract both chroma maps as well as their corresponding gradient maps (via Eq. (12)) following the suggestions in [10]. The above defined luma and chroma feature transforms are visualized in Fig. 6.

Images are naturally multiscale, and distortions affect image structures across scales. Incorporating multiscale information in quality models provides significant performance improvements [8], [77]. Hence, we extract NSS-34 features from the four luma feature maps (Y, GM, LoG, DoG) at two scales: the original image scale and a reduced (by a factor two) resolution. However, we only extract NSS-34 features at the half scale on the twelve chroma feature maps, since frames are often compressed in YUV420 format, which already contain chroma information in reduced scale; additionally, it has also been observed that humans are more sensitive to luma distortions than chroma distortions [55]. To sum up, the entire collection of spatial features is collected by applying two-scales NSS-34 to the luma feature maps and single-scale NSS-34 to the chromatic maps, yielding a total of $34 \times 2 \times 4 + 34 \times 1 \times 12 = 680$ features. It is worth noting that this 680-dim spatial model is an improved alternative to the SOTA 560-dim FRIQUEE model [12] since it achieves comparable performance as FRIQUEE, but is **20x** faster.

C. TEMPORAL FEATURES

Prior BVQA methods accounting for temporal distortions, however, either rely on expensive motion estimation [9], [13], or underperform on UGC videos by only accounting for simple frame-difference statistics [4], [47], [48], even including complex CNN models [19], [26]. Here we attempt to exploit more general temporal scene statistics of natural videos to develop and improve BVQA models. To the best of our knowledge, we propose the *first general, effective and efficient* temporal statistics model based on bandpass regularities of natural videos along the time dimension, going beyond simpler frame-difference models [78].

Inspired by the efficacy of temporal bandpass statistics in the prediction of frame rate-dependent video quality [54], our proposed temporal model utilizes 1D temporal bandpass representations. Specifically, consider a bank of K temporal bandpass filters denoted $h_k,\ k \in \{0, \ldots, K-1\}$, where k denotes the subband index. The temporal bandpass responses of a video $F(\mathbf{x}, t)$ (where $\mathbf{x} = (x, y)$ and t represents spatial and temporal co-ordinates, respectively) is

$$Y_k(\mathbf{x}, t) = F(\mathbf{x}, t) * h_k(t) \ \ k = 0, \ldots, K-1, \tag{19}$$

where $*$ and Y_k are 1D temporal convolution operations and the bandpass response of the k^{th} filter, respectively. Note that frame differences are a special case of Eq. (19) (the high-pass component of a 2-tap Haar wavelets). Fig. 7 visually illustrates the bandpass responses of a natural video from the LIVE-VQA dataset [1] using 3-level Haar wavelet filters.

Attempting to generalize the spatial NSS as mentioned in Section III-A to the temporal domain, we instead analyze the statistics of the temporal bandpass coefficients $Y_k(\mathbf{x}, t)$, $k = 1, \ldots, 7$ (ignoring the lowest band $k = 0$) by again applying MSCN transforms, as in Eq. (1), to further decorrelate the subband representations, over a set of frame time samples $t \in \{t_0, t_1, \ldots, t_N\}$ (note that t does not need to be densely sampled). Note that in Eq. (1), $\mu_k(\mathbf{x}, t)$ and $\sigma_k(\mathbf{x}, t)$ are replaced by the local mean and standard deviation within a spatial window centered at location $(\mathbf{x}, t)$, for each subband k.

We have found that the MSCN coefficients of the temporal bandpass coefficients of natural videos also exhibit a Gaussian-like appearance, as shown in Fig. 8, while the regularities are modified by the presence of distortion, strongly suggesting the possibility of quantifying deviations to predict perceived video quality. We model the distributions of subband MSCN coefficients again using the pre-defined GGD and AGGD distributions, by merely passing them into the NSS-34 feature extractor (Section III-B). Similar to the spatial feature processing, we also extract the temporal statistical features over two scales (original and half scale), yielding a 476-dim feature vector ((34 features/band)$\times$(7 subbands)$\times$(2 scales) $= 476$).

Inspired by the efficacy of standard deviation pooling as first introduced in GMSD [79] and later also shown effective when utilized for temporal pooling in [6], [13], we calculate the 680 spatial features at two frames per second within each non-overlapping one-second chunk, then enrich the feature set by applying average and absolute difference pooling [80] of the frame features within each chunk, based on the hypothesis that the variation of spatial features also correlates with the temporal properties of the video. Finally, all of the chunk-wise feature vectors are average pooled across all chunks to derive the final set of features over the entire video.

D. DEEP LEARNING FEATURES

CNN-based solutions have been observed to generally perform well on UGC picture quality problems [17], [27], [51] thanks to several recently released large-scale picture quality datasets [17], [51], [82]. Still, none of them have proven effective on UGC video quality databases [30]–[32]. However, the authors of [6] have shown that the simple feature vector from an FC-layer, without fine-tuning, to be a useful quality indicator if training a shallow regressor on top. Therefore, we, *for the first time*, propose to leverage the best of both worlds, by combining powerful quality-aware NSS features as described in Section III-A, III-B, III-C, with pre-trained deep learning features, by jointly training a regressor on them to predict the final quality score.

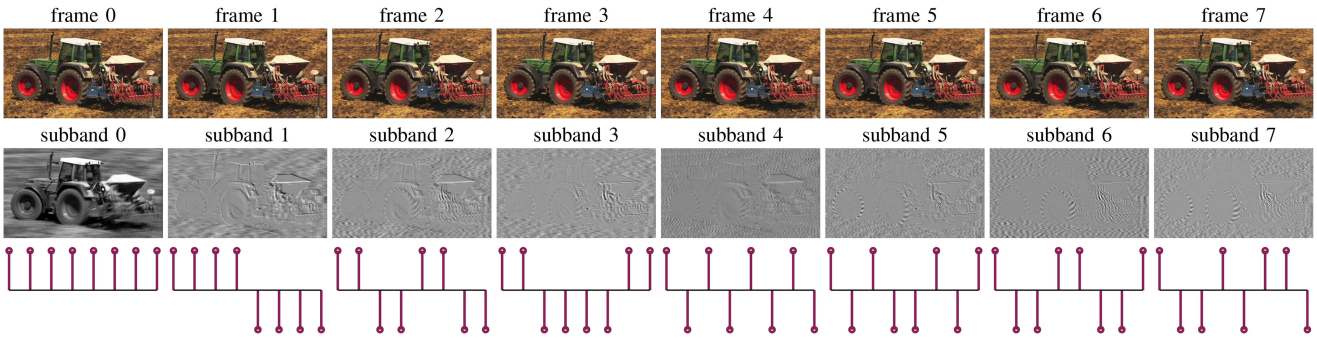


FIGURE 7. Top row: eight exemplar consecutive frames sampled from sequence `Tractor` in LIVE-VQA [1]. Middle row: temporal bandpass-filtered responses by convolving with the filters in an 8-subband Haar-wavelet filter bank, which are shown in the bottom row. The subband frequency increases from left to right: $k = 0, \ldots, 7$, for both the responses and wavelet functions.

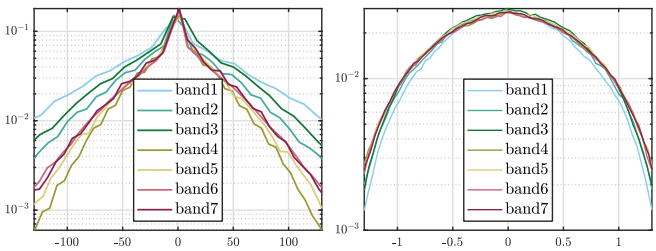


FIGURE 8. Histograms of raw subband (left) and the corresponding MSCN normalized coefficients (right) of a natural video `Tractor`, where the normalized coefficients exhibit homogeneous regularities across bands.

One issue encountered when dealing with quality prediction problems is the mismatch of picture sizes between the standard inputs of CNN models such as VGG-16 [83], ResNet-50 [34], and IQA-valid high-resolution images. Two possible solutions have been attempted to solve this. The authors of [17], [61] suggested applying a CNN on spatially sampled small patches, then aggregating the locally predicted scores to obtain global quality scores. The authors of [51] presented a CNN operating on full-sized images, but with either global average pooling (GAP) or spatial pyramid pooling (SPP) [84], feeding FC layers. These two schemes, however, increase the computation load of the CNN models. Since our proposed model is already armed with powerful spatial and temporal quality-aware features, we added CNN features only to exploit its ability to capture high-level semantic information, supplementing the low-level NSS features. In this regard, we aggressively downscaled the frames to fit the CNN model inputs when extracting these semantic-aware features, yielding greater efficiency than previous CNN VQA models. Another reason to use a pre-trained CNN without fine-tuning is to prevent overfitting, since existing video quality datasets are of limited sizes. In our implementation, we used a ResNet-50 (2048-dim) as a semantic feature extractor.

E. LEARNING A VIDEO QUALITY PREDICTOR

We summarize the feature extraction process as follows. Since our goal is to build an efficient BVQA model, we devised spatial and temporal sampling strategies to further improve its speed. Specifically, given an input video $F(\mathbf{x}, t)$, RAPIQUE uniformly samples 2 frames per second, based on which the 680-dim spatial NSS features (in Section III-B) are extracted, then average and absolute-difference pools these to obtain 680 spatial and 680 temporal variation features, respectively. RAPIQUE also uniformly samples 8 consecutive frames each second, then applies temporal Haar filter (Eq. (7)) to extract 7 bandpass responses, from which 476 features are calculated at each time sample. The above features are calculated at a higher resized resolution while maintaining the aspect ratio (we used 512p in our experiments). However, the CNN backbone (ResNet-50) operates on resized frames at a sparse temporal sampling of 1 frame/sec to attain an additional 2048 features.

After obtaining all the spatial, temporal, and CNN features within each one-second chunk, we adopt a simple approach to concatenate them all into a totally 3884-dimensional feature vector for each video chunk, then average-pool each to obtain a single 3884 feature vector over the entire video. A shallow or deep regressor head can then be trained on the aggregated feature vector to predict the final video quality scores.

TABLE 2. Summary of the Tested BVQA Datasets

Database	# Vid	Reso	Sec	Label	Range
KoNViD-1k'17 [31]	1,200	540p	8	MOS+σ	[1,5]
LIVE-VQC'18 [30]	585	1080p-240p	10	MOS	[0,100]
YouTube-UGC'20 [32]	1,380	4k-360p	20	MOS+σ	[1,5]

IV. EXPERIMENTS

A. EXPERIMENT SETTINGS

Datasets and Baselines: We used three recent BVQA datasets as testbeds for the performance evaluations: KoNViD-1k [31], LIVE-VQC [30], and YouTube-UGC [32], as summarized in Table 2. We also used the combined set (denoted All-Combined) as introduced in [6] as an additional composite benchmark. The All-Combined dataset is simply the union of KoNViD-1 k (1200), LIVE-VQC (575), and

TABLE 3. Performance Comparison of the Evaluated BVQA Models on the Four BVQA Datasets

Dataset	KoNViD-1k [31]			LIVE-VQC [30]			YouTube-UGC [32]			All-Combined [6]		
Model	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓
BRISQUE [8]	0.6567	0.6576	0.4813	0.5925	0.6380	13.100	0.3820	0.3952	0.5919	0.5695	0.5861	0.5617
GM-LOG [11]	0.6578	0.6636	0.4818	0.5881	0.6212	13.223	0.3678	0.3920	0.5896	0.5650	0.5942	0.5588
HIGRADE [10]	0.7206	0.7269	0.4391	0.6103	0.6332	13.027	0.7376	0.7216	0.4471	0.7398	0.7368	0.4674
FRIQUEE [12]	0.7472	0.7482	0.4252	0.6579	0.7000	12.198	**0.7652**	**0.7571**	**0.4169**	**0.7568**	**0.7550**	**0.4549**
CORNIA [14]	0.7169	0.7135	0.4486	0.6719	0.7183	11.832	0.5972	0.6057	0.5136	0.6764	0.6974	0.4946
HOSA [81]	0.7654	0.7664	0.4142	0.6873	0.7414	11.353	0.6025	0.6047	0.5132	0.6957	0.7082	0.4893
KonCept512 [51]	0.7349	0.7489	0.4260	0.6645	0.7278	11.626	0.5872	0.5940	0.5135	0.6608	0.6763	0.5091
PaQ-2-PiQ [17]	0.6130	0.6014	0.5148	0.6436	0.6683	12.619	0.2658	0.2935	0.6153	0.4727	0.4828	0.6081
V-BLIINDS [9]	0.7101	0.7037	0.4595	0.6939	0.7178	11.765	0.5590	0.5551	0.5356	0.6545	0.6599	0.5200
TLVQM [13]	0.7729	0.7688	0.4102	**0.7988**	**0.8025**	**10.145**	0.6693	0.6590	0.4849	0.7271	0.7342	0.4705
VMEON [65]	0.1118	0.1958	0.6322	0.4024	0.4088	15.524	0.0634	0.1100	0.6304	0.2578	0.2594	0.6657
VSFA [19]	0.7728*	0.7754*	0.4205*	0.6978*	0.7426*	11.649*	-	-	-	-	-	-
MDVSFA [67]	**0.7812***	**0.7856***	-	0.7382*	**0.7728***	-	-	-	-	-	-	-
VIDEVAL [6]	**0.7832**	**0.7803**	**0.4026**	**0.7522**	0.7514	**11.100**	**0.7787**	**0.7733**	**0.4049**	**0.7960**	**0.7939**	**0.4268**
RAPIQUE	**0.8031**	**0.8175**	**0.3623**	**0.7548**	**0.7863**	**10.518**	**0.7591**	**0.7684**	**0.4060**	**0.8070**	**0.8229**	**0.3968**

The **Underlined** and **Boldfaced** Entries Indicate the Best and Top Three Performers on Each Database for Each Performance Metric, Respectively.
*The results are cited from experiments reported in their original papers

YouTube-UGC (1380) after MOS calibration:

$$y_{\text{adj}} = 5 - 4 \times \left[(5 - y_{\text{org}})/4 \times 1.1241 - 0.0993\right] \quad (20)$$

$$y_{\text{adj}} = 5 - 4 \times \left[(100 - y_{\text{org}})/100 \times 0.7132 + 0.0253\right] \quad (21)$$

where equations (20) and (21) are used to calibrate KoNViD-1 k and LIVE-VQC, respectively (YouTube-UGC does not need to be changed). Here y_{adj} denotes the adjusted scores, while y_{org} is the original MOS. We refer the reader to [6] for details regarding assumptions and derivations of the calibration process.

The baseline models used for comparison are BRISQUE [8], GM-LOG [11], HIGRADE [10], FRIQUEE [12], the codebook-based models CORNIA [14] and HOSA [81], and the deep learning models, KonCept512 [51], PaQ-2-PiQ [17] which are all spatial-only models. All the spatial models extract features at 1 fps, which were average-pooled to obtain the final video-level feature vector used for training. We also compared against three feature-based BVQA models, V-BLIINDS [9], TLVQM [13], and VIDEVAL [6], and the deep learning-based models, V-MEON [65] and VSFA [19] as well as its enhanced version, MDVSFA [67]. Since 'completely blind' models such as NIQE [33] and VIIDEO [47] were not observed to perform reasonably well on these natural video datasets [6], we did not include them.

Evaluation Method: We used a support vector regressor (SVR) as the back-end regression model to learn the feature-to-score mappings [8], [9], [12], [13], [61]. We optimized the SVR parameters (C, γ) via a randomized grid-search on the training set. Following convention, we randomly split the dataset into training and test sets (80%/20%) over 20 iterations, and the overall median test performance was reported. All of the evaluated methods were implemented using the original release by the respective authors. Four performance metrics were used: the Spearman Rank-Order Correlation Coefficient (SRCC) and the Kendall Rank-Order Correlation Coefficient (KRCC) are non-parametric measures of prediction monotonicity, while the Pearson Linear Correlation Coefficient (PLCC) with corresponding Root Mean Square Error (RMSE) were computed to assess prediction accuracy. Note that PLCC and RMSE are computed after performing a nonlinear four-parametric logistic regression to linearize the objective predictions to be on the same scale as MOS [1]:

$$f(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp\left(-x + \beta_3/|\beta_4|\right)}. \quad (22)$$

B. MAIN EVALUATION RESULTS

Table 3 shows the main comparison results on the four evaluated datasets. It may be observed that RAPIQUE achieved the best performance on KoNViD-1 k, even outperforming the most recent, dense deep learning models such as VSFA and MDVSFA. On LIVE-VQC, which contains many mobile videos exhibiting large camera motions [6], TLVQM, which contains numerous heavily crafted motion-relevant features, was the best performer. However, RAPIQUE ranked a clear second, indicating that the temporal NSS features in RAPIQUE are powerful indications of temporal and motion-related distortions.

The most recent deep still picture quality models, KonCept512 and PaQ-2-PiQ, have been observed to perform poorly on UGC-VQA datasets. One reason for this is that these models were trained on picture quality datasets [17], [51], containing strictly spatial content and distortions. A leading blind deep video quality model, V-MEON, also does not perform well, likely because it was trained on compression artifacts rather than on complex combinations of UGC distortions.

On the larger datasets, RAPIQUE delivered the second-best correlation against the subjective data on YouTube-UGC, only slightly worse than the current SOTA model VIDEVAL, while RAPIQUE ranked the best on the 3165-video composite set, All-Combined. Since VIDEVAL was created by a supervised feature selection process (using subjective labels) on the composite combined set, wherein YouTube-UGC accounts for a

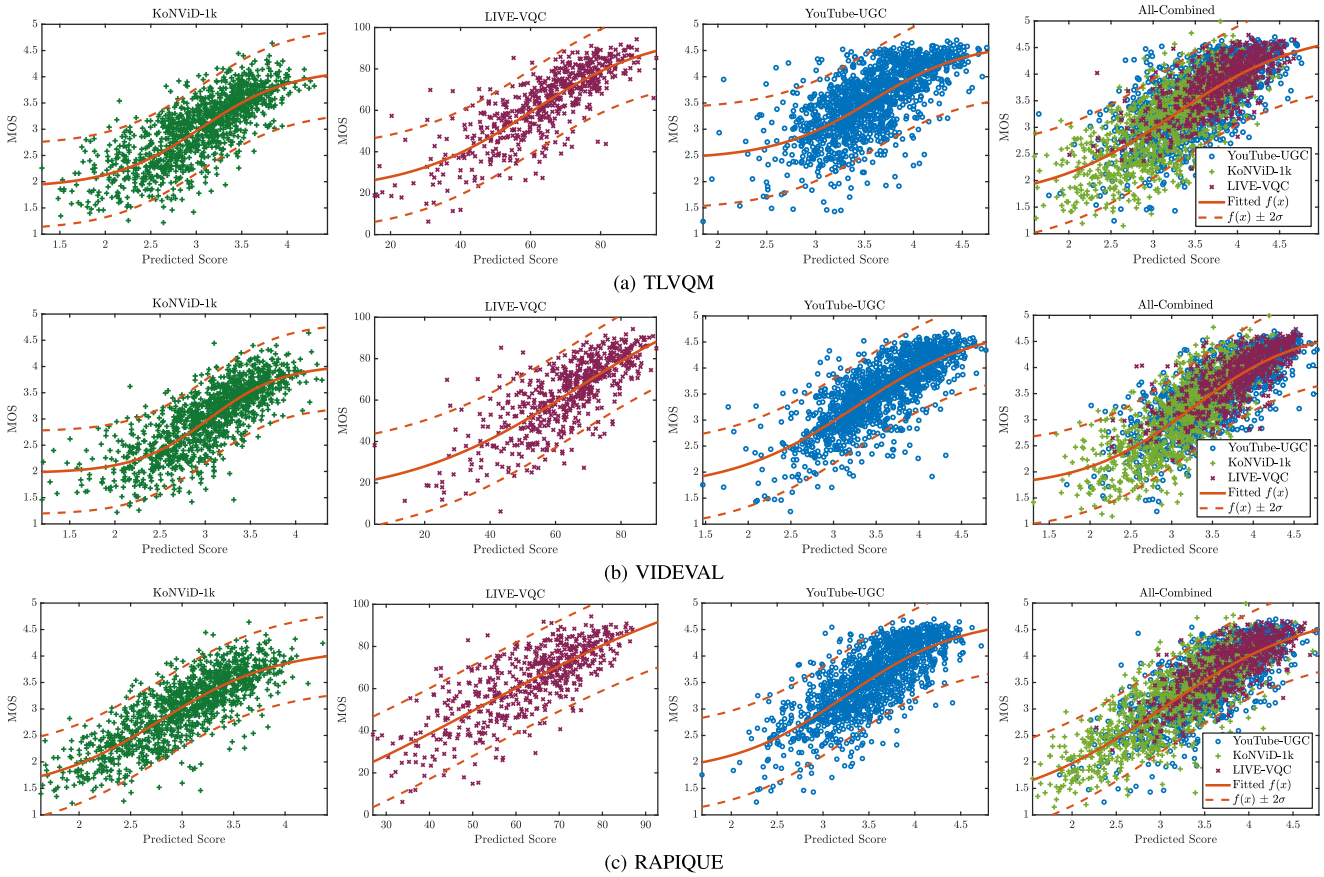


FIGURE 9. Scatter plots and nonlinear logistic fitted curves of (c) RAPIQUE versus MOS, compared against (a) TLVQM [13] and (b) VIDEVAL [6], using a grid-search SVR using *k*-fold cross-validation on KoNViD-1k [31], LIVE-VQC [30], YouTube-UGC [32], and the All-Combined set (Section IV-A), respectively.

large portion, it would be expected to outperform on these two sets. The RAPIQUE model, on the contrary, is database-agnostic, and also exhibited uniformly well performance on all four test sets. In this regard, RAPIQUE has the potential to perform better on future larger-scale datasets and in real-world application scenarios it has not been exposed to. The scatter plots and fitted curves of RAPIQUE predictions versus MOS in Fig. 9 visually demonstrate that the performance of RAPIQUE remains stable on video sequences from different databases, achieving smaller RMSE on larger databases.

C. EFFECTS OF TRAINING DATA SIZE

To study the degree of performance variation by the compared algorithms, we vary the training-test splits from 10% to 90% of the content used for training, using the rest for testing on the composite combined set. As might be seen in Fig. 10, the RAPIQUE model was able to already achieve better than 0.8 in PLCC provided only 50% of the data for training. When compared to SOTA methods, although RAPIQUE was not observed to outperform VIDEVAL when the fraction of training data was less than 40%, it delivered improved performances relative to VIDEVAL and TLVQM as the proportion of training data was increased, as shown in Fig. 10. This suggests that RAPIQUE is very data-efficient, with the potential to achieve even better results when larger-scale datasets become available.

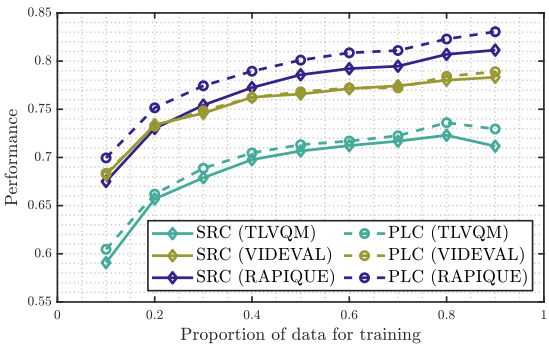


FIGURE 10. Performance comparison of SRCC / PLCC as a function of the percentage of the content used to train the compared blind VQA models on the composite All-Combined set. Note that this result is self-explanatory as we used a slightly different evaluation method (20 iterations, SVR with randomized search cross-validation) compared to previous experiments.

D. ABLATION STUDY

To analyze the importance of each module in RAPIQUE, we conducted an ablation study. Fig. 11 shows the incremental performance attained when adding each module sequentially.

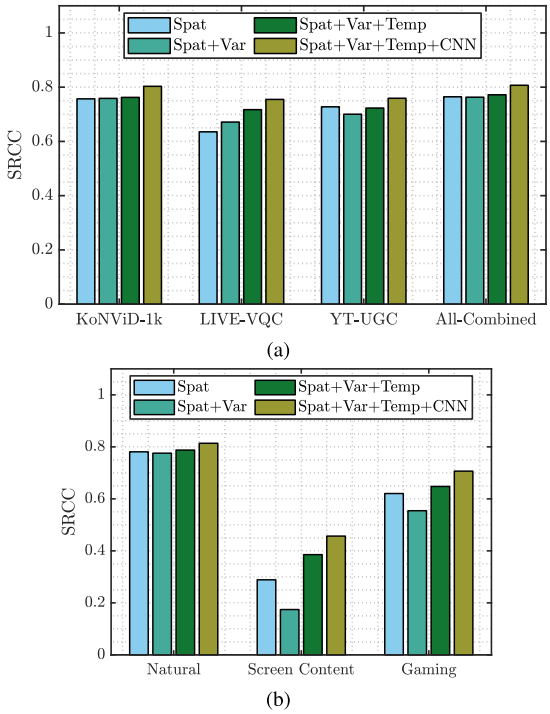


FIGURE 11. **Ablation study of RAPIQUE on (a) four benchmarks and (b) different content types - *Spat* denotes the spatial model (Section III-B), *Var* is the temporal difference-pooled spatial models (Section III-C), *CNN* represents deep features (Section III-D), and *Temp* presents the Haar bandpass-filtered features introduced in Section III-C.**

It is worth mentioning the dataset biases of the evaluated benchmarks. For example, the authors of [6] observed that the LIVE-VQC videos generally contain more (camera) motions and temporal distortions than other databases, while spatial distortions predominate on KoNViD-1 k and YouTube-UGC. It may be observed in Fig. 11(a) that the spatial NSS module (Section III-B) performs quite well on the UGC databases that mainly present spatial distortions, like KoNViD-1 k and YouTube-UGC, indicating its efficacy in capturing authentic spatial distortions. LIVE-VQC, which mainly contains videos with large motions, challenges the spatial NSS module, aligning with the empirical observations made above [6]. Adding spatial variation and temporal NSS features (Section III-C) improves the performance of RAPIQUE on LIVE-VQC, indicating that these two types of temporal features capture important attributes of motion-intensive videos. Interestingly, we also noticed that including the SpatialNSS-Var features degraded performance on YouTube-UGC. It is possible that the SpatialNSS-Var features are redundant with SpatialNSS features on YouTube-UGC, causing the training algorithm to underperform. We also observed that temporal statistics did not contribute much to the assessment of Internet UGC videos from YouTube and KoNViD-1 k (Flickr).

It is also important to note that including deep learning features (Section III-D) significantly boosts the performance over only using NSS features on all these UGC datasets, further validating our assumptions expressed in Section II-B, that high-level semantic features are also informative when conducting UGC video quality prediction. To better understand which types of videos are advantageously analyzed by the CNN features, we divided the combined set into three subsets of differing contents: 2667 natural videos, 163 screen contents, and 209 gaming videos, as shown in Fig. 11(b). Notably, we observed that the CNN features provided more benefits on screen content and gaming videos than on natural videos. The new temporal statistical features yielded noticeable improvements relative to using only spatial features. Lastly, our deployment of CNN modules is essentially different from other methods [17], [51] in that RAPIQUE only requires a single pass of the resized frames (224x224), making it highly advantageous in application scenarios having high-speed requirements.

TABLE 4. **Performance of RAPIQUE Combined With Different Deep Learning Features. RAPIQUE ((w/ ResNet-50) is the Default Version Proposed in This Paper**

DATASET	MODEL / METRIC	SRCC↑	PLCC↑	RMSE↓
KoNViD-1k	RAPIQUE (w/ ResNet-50)	**0.8031**	**0.8175**	**0.3623**
	RAPIQUE (w/ VGG-19)	0.7554	0.7389	0.4238
	RAPIQUE (w/ KonCept512)	0.7802	0.7793	0.3975
	RAPIQUE (w/ PaQ-2-PiQ)	0.7726	0.7672	0.4026
LIVE-VQC	RAPIQUE (w/ ResNet-50)	**0.7548**	**0.7863**	**10.518**
	RAPIQUE (w/ VGG-19)	0.6888	0.7048	12.228
	RAPIQUE (w/ KonCept512)	0.7497	0.7611	11.231
	RAPIQUE (w/ PaQ-2-PiQ)	0.7147	0.7308	11.599
YouTube-UGC	RAPIQUE (w/ ResNet-50)	0.7591	**0.7684**	**0.4060**
	RAPIQUE (w/ VGG-19)	0.7379	0.7398	0.4365
	RAPIQUE (w/ KonCept512)	**0.7668**	0.7678	0.4190
	RAPIQUE (w/ PaQ-2-PiQ)	0.7596	0.7606	0.4200
All-Combined	RAPIQUE (w/ ResNet-50)	**0.8070**	**0.8229**	**0.3968**
	RAPIQUE (w/ VGG-19)	0.6888	0.7048	12.228
	RAPIQUE (w/ KonCept512)	0.7924	0.7976	0.4169
	RAPIQUE (w/ PaQ-2-PiQ)	0.7742	0.7809	0.4312

E. PERFORMANCE ON DIFFERENT DEEP FEATURES

To determine which kinds of deep features most effectively complement the proposed NSS features, we conducted another ablation study. We compared the performance of RAPIQUE variants that use features from different backbones: VGG-19, ResNet-50, PaQ-2-PiQ (trained on LIVE-FB [17]), and KonCept512 (trained on KonIQ-10k [51]). Since PaQ-2-PiQ was designed for local quality prediction, we included the predicted 3×5 local quality scores along with the single global score. For KonCept512, the 256-dim feature vector immediately before the last linear layer in the fully connected head was included. We also included VGG-19 and ResNet-50, except for they were pre-trained on ImageNet classification.

The overall performance results are tabulated in Table 4. It may be observed that combining NSS features with ResNet-50 yielded the best or top performances on all benchmarks, slightly better than KonCept512, suggesting that features pre-trained on classification tasks provide valuable high-level semantic information to the quality assessment process.

TABLE 5. Feature Dimensionality and Average (CPU/GPU) Runtime Comparison (In Seconds) Evaluated on 1080*p* Videos

Model	Dim	Runtime	
		CPU	GPU
BRISQUE (1 fr/sec)	36	1.7	-
GM-LOG (1 fr/sec)	40	2.1	-
HIGRADE (1 fr/sec)	216	11.6	-
FRIQUEE (1 fr/sec)	560	701.2	-
CORNIA (1 fr/sec)	10k	14.3	-
HOSA (1 fr/sec)	14.7k	1.2	-
KonCept512 (1 fr/sec)	-	2.8	0.3
PaQ-2-PiQ (1 fr/sec)	-	6.9	0.8
V-BLIINDS	47	1989.9	-
V-MEON	-	16.4	2.6
TLVQM	75	183.8	-
VSFA	-	1288.7	157.9
MDVSFA	-	1319.4	162.5
VIDEVAL	60	305.8	-
RAPIQUE (proposed)	3.8k	17.3	-

Moreover, using features pre-trained on a specific IQA dataset may limit model generalizability to future, unseen distortions. Another important reason why we prefer ResNet-50 over KonCept512 is the gigantic model size of the InceptionResNetV2 [85], used as the backbone of KonCept512.

F. COMPLEXITY AND RUNTIME COMPARISON

Apart from performance analysis, computational efficiency is also of great importance for BVQA models. Thus, we also study the model (feature) dimension and runtime comparisons in Table 5. For a fair comparison, all the experiments were carried out in the same desktop computer, a Dell OptiPlex 7080 Desktop with Intel Core i7-8700 CPU@3.2 GHz, 32 G RAM, and GeForce GTX 1050 Graphics Cards. The models were implemented using their original releases on MATLAB R2018b and Python 3.6.7 under Ubuntu 18.04.3 LTS system. It should be noted that our comparison of computing complexity involves methods that use different sampling rate (or FPS), which is critical to model efficiency. However, we regard the design of FPS itself as an important aspect of BVQA algorithms, and thus our comparison still provides insights on developing more efficient BVQA models.

It may be seen that RAPIQUE is extremely efficient as compared to other complex top-performing BVQA models like TLVQM and VIDEVAL. Specifically, RAPIQUE is **10x** faster than TLVQM, which also aims to efficiency. Fig. 12 shows the scatter plots of SRCC versus runtime, which indicates that RAPIQUE achieves comparable prediction accuracy, but with **20x** less computational expense as compared to VIDEVAL, the current SOTA model on the UGC-VQA problem [6]. We observe however that, CNN models that benefit from optimized low-level implementations are generally faster than NSS models executed in MATLAB; we have observed a ∼10x speedup by switching from CPU to GPU on the CNN-based models, KonCept512, PaQ-2-PiQ, V-MEON, VSFA, and MDVSFA.

Predicting the quality of videos having multiple diverse resolutions is also a pressing problem, but has barely been discussed, since most video datasets only contain single-resolution contents. Thanks to the large-scale dataset, YouTube-UGC [32], which contains videos at five different resolutions, we were able to extend the complexity analysis to videos ranging from 540p to 4 k, to study computational scalability with respect to video size. Fig. 13 compares computation time as a function of video resolution. We may observe that RAPIQUE has superior computational scalability in terms of data sizes, making it attractive and preferable for potential real-time, low-latency, and light-weight applications requiring high-resolution video inputs. Particularly, as seen in the plot our model is 2-150x faster than baselines, depending on resolution, and the higher, the faster. In Table 6 we list the partial compute time of each sub-module in RAPIQUE on 1080p

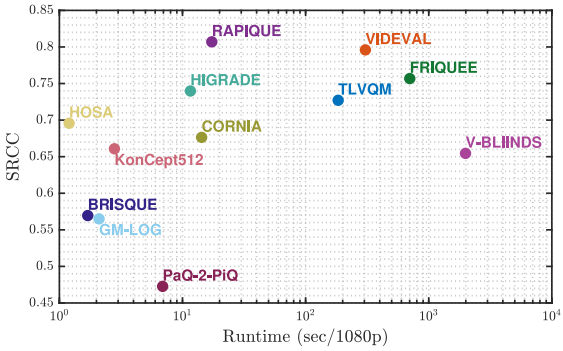


FIGURE 12. Scatter plots of SRCC (on All-Combined) of selected BVQA algorithms versus CPU runtime (per 1080p video on average). Purple indicates the proposed RAPIQUE model.

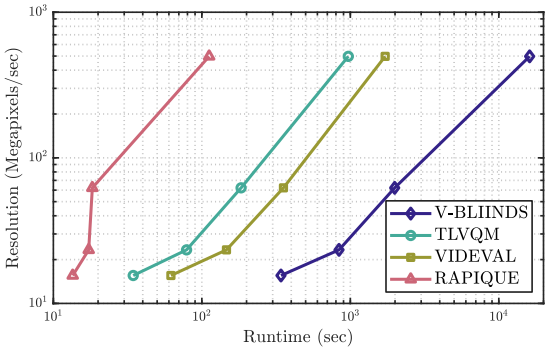


FIGURE 13. Our proposed RAPIQUE model enables high-resolution video quality prediction at significantly lower runtimes than existing BVQA methods. Particularly, as seen in the plot our model is 2-150x faster than baselines, depending on resolution, and the higher, the faster.

TABLE 6. Complexity Analysis of RAPIQUE. Tabulated Values Reflect the Partial Time Devoted to Each Sub-Component in RAPIQUE

Module	Dim	Runtime
SpatialNSS (Sec. III-B)	680	11.1
SpatialNSS-Var (Sec. III-C)	680	
TemporalNSS (Sec. III-C)	476	5.8
CNN (Sec. III-D)	2.0k	0.4
RAPIQUE (Full model)	3.8k	17.3

videos. Since all of the NSS-based features are implemented in MATLAB, a high-level prototyping tool, we would expect further accelerations to be possible (by orders-of-magnitude) if implemented in low-level languages like C/C++, or GPU-friendly frameworks such as Tensorflow or PyTorch.

V. CONCLUSION

We have proposed an effective and efficient model for predicting the subjective quality of user-generated videos, which we call the Rapid and Accurate Video Quality Evaluator (RAPIQUE). The model, for the first time, leverages a composite of spatio-temporal scene statistics features and deep CNN-based high-level features in a two-branch framework, then jointly learns a regressor head for video quality prediction. Within the model, we developed new spatial scene statistics models in an efficient way and further extended the overall model to include normalized temporal bandpass responses, yielding the first general efficacious temporal NSS model for UGC video quality problems. Experiments on recent large-scale UGC video databases show the superior accuracy and efficiency of the proposed model in that it achieves competitive or substantially higher accuracy than both SOTA conventional as well as deep learning video quality models. RAPIQUE is computationally less expensive by orders-of-magnitude than the most accurate benchmark methods and scales remarkably well with video resolution. To support reproducible research, an implementation of RAPIQUE is available.[1]

REFERENCES

[1] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

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

[3] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *Netflix Tech Blog*, vol. 6, 2016. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[4] X. Yu, N. Birkbeck, Y. Wang, C. G. Bampis, B. Adsumilli, and A. C. Bovik, "Predicting the quality of compressed videos with pre-existing distortions," 2020, *arXiv:2004.02943*.

[5] Y. Wang *et al.*, "Video transcoding optimization based on input perceptual quality," in *Appl. Digit. Image Process. XLIII*, Aug. 2020, Art. no. 1151015.

[6] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, Apr. 2021.

[7] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

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

[9] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[10] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, Jun. 2017.

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

[12] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, pp. 32–32, 2017.

[13] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, Dec. 2019.

[14] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1098–1105.

[15] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual DOG model fused with random forest," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3282–3292, Nov. 2015.

[16] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, and A. C. Bovik, "No-reference video quality assessment using space-time chips,"in *IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, 2020, pp. 1–6.

[17] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3575–3585.

[18] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ:'Patching up'the video quality problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14019–14029.

[19] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. ACM Multimedia Conf.*, 2019, pp. 2351–2359.

[20] X. Feng and J. P. Allebach, "Measurement of ringing artifacts in JPEG images,"*Digit. Pub.*, vol. 6076, 2006, Art. no. 60760A.

[21] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik, "Bband index: A no-reference banding artifact predictor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 2712–2716.

[22] A. Norkin and N. Birkbeck, "Film grain synthesis for AV1 video codec," in *Data Compress. Conf.*, 2018, pp. 3–12.

[23] L.-H. Chen, C. G. Bampis, Z. Li, A. Norkin, and A. C. Bovik, "Prox-IQA: A proxy approach to perceptual optimization of learned image compression," *IEEE Trans. Image Process.*, vol. 30, pp. 360–373, Nov. 2020, doi: 10.1109/TIP.2020.3036752.

[24] H. Wang, T. Chen, Z. Wang, and K. Ma, "I am going mad: Maximum discrepancy competition for comparing classifiers adaptively," in *Proc. Int. Conf. Learn. Representations*, 2019.

[25] L.-H. Chen, C. Bampis, Z. Li, and A. C. Bovik, "Learning to distort images using generative adversarial networks," *IEEE Signal Process. Lett.*, vol. 27, pp. 2144–2148, Nov. 2020, doi: 10.1109/LSP.2020.3040656.

[26] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 219–234.

[27] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2244–2255, Aug. 2019.

[28] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10 K: Towards an ecologically valid and large-scale IQA database," 2018, *arXiv:1803.08489*.

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

[30] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, Feb. 2019.

[31] V. Hosu *et al.*, "The konstanz natural video database (KoNViD-1 k)," in *Proc. 9th Int. Conf. Qual. Multimedia Exper*, 2017, pp. 1–6.

[32] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, 2019, pp. 1–5.

[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[1] [Online]. Available: https://github.com/vztu/RAPIQUE

[35] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Process.*, 2000, pp. 981–984.

[36] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. IEEE Int. Conf. Image Process.*, 2002, vol. 3, pp. III-III.

[37] Y. Wang, S.-U. Kum, C. Chen, and A. Kokaram, "A perceptual visibility metric for banding artifacts," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 2067–2071.

[38] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik, "Adaptive debanding filter," *IEEE Signal Process. Lett.*, vol. 27, pp. 1715–1719, Sep. 2020, doi: 10.1109/LSP.2020.3024985.

[39] A. Amer and E. Dubois, "Fast and reliable structure-oriented video noise estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 113–118, Jan. 2005.

[40] D. L. Ruderman, "The statistics of natural images," *Netw.: Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.

[41] 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.

[42] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[43] Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Signal Process Image Commun.*, vol. 29, no. 7, pp. 725–747, 2014.

[44] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, Jun. 2010.

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

[46] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, Jul. 2016.

[47] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Jan. 2016.

[48] Z. Sinno and A. C. Bovik, "Spatio-temporal measures of naturalness," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1750–1754.

[49] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *J. Electron. Imag.*, vol. 22, no. 4, 2013, Art. no. 043025.

[50] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014-a database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073–3086, Jul. 2016.

[51] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10 k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, Jan. 2020, doi: 10.1109/TIP.2020.2967829.

[52] R. T. Born and D. C. Bradley, "Structure and function of visual area MT," *Annu. Rev. Neurosci.*, vol. 28, pp. 157–189, 2005.

[53] S. V. R. Dendi and S. S. Channappayya, "No-reference video quality assessment using natural spatiotemporal scene statistics," *IEEE Trans. Image Process.*, vol. 29, pp. 5612–5624, Apr. 2020, doi: 10.1109/TIP.2020.2984879.

[54] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction," 2020, *arXiv:2010.13715*.

[55] L.-H. Chen, C. G. Bampis, Z. Li, J. Sole, and A. C. Bovik, "Perceptual video quality prediction emphasizing chroma distortions," *IEEE Trans. Image Process.*, vol. 30, pp. 1408–1422, Dec. 2020.

[56] D. Y. Lee, H. Ko, J. Kim, and A. C. Bovik, "Video quality model for space-time resolution adaptation," in *Proc. IEEE Int. Conf. Image Process. Appl. Syst.*, 2020, pp. 34–39.

[57] D. Y. Lee, H. Ko, J. Kim, and A. C. Bovik, "On the space-time statistics of motion pictures," *J. Opt. Soc. Amer. A*, vol. 38, no. 7, pp. 908–923, 2021.

[58] Y. Jiang *et al.*, "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, Jan. 2021.

[59] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1733–1740.

[60] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3773–3777.

[61] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[63] 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.

[64] N. Ponomarenko *et al.* "Color image database TID2013: Peculiarities and preliminary results," in *Proc. 4th Eur. Workshop Vis. Inf. Process.*, 2013, pp. 106–111.

[65] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *Proc. ACM Multimedia Conf.*, 2018, pp. 546–554.

[66] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2061–2077, Sep. 2018.

[67] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1238–1257, 2021.

[68] 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.

[69] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, Feb. 1995.

[70] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," *J. Physiol.*, vol. 197, no. 3, pp. 551–566, 1968.

[71] 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.

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

[73] U. Rajashekar, Z. Wang, and E. P. Simoncelli, "Perceptual quality assessment of color images using adaptive signal representation," in *Hum. Vis. Electron. Imag. XV*, 2010, p. 75271L.

[74] D. Lee and K. N. Plataniotis, "Toward a no-reference image quality assessment using statistics of perceptual color descriptors," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3875–3889, Aug. 2016.

[75] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Proc. SPIE Hum. Vis. Electron. Imag.*, 2003, pp. 87–96.

[76] "Cielab Color Space - Wikipedia, the Free Encyclopedia," Accessed: Oct. 5, 2020. [Online]. Available: https://en.wikipedia.org/wiki/CIELAB_color_space

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

[78] 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, Apr. 2013.

[79] 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, Feb. 2014.

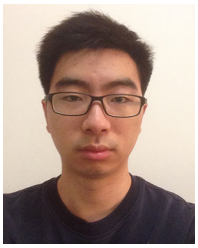
[80] Z. Tu, C. J. Chen, L. H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 141–145.

[81] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.

[82] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3677–3686.

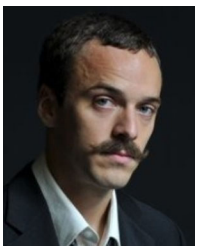
[83] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[84] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[85] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017.

ZHENGZHONG TU (Graduate Student Member, IEEE) received the B.S. and M.Eng. degrees in microelectronics from Fudan University, Shanghai, China, in 2016 and 2018, respectively. He is currently working toward the Ph.D. degree with the Laboratory for Image and Video Engineering, The University of Texas at Austin, Austin, TX, USA. His research interests include perceptual image and video quality assessment, computer vision, and machine learning.



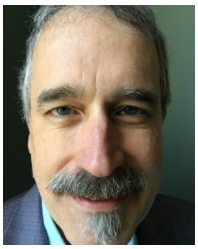
XIANGXU YU received the B.Eng. in electronic and information engineering from The Hong Kong Polytechnic University, Hongkong, Hong Kong, and the M.S. degree in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2015 and 2018, respectively. He is currently working toward the Ph.D. degree with the Laboratory for Image and Video Engineering, The University of Texas at Austin. His research interests include image and video processing and machine learning.

YILIN WANG received B.S. and M.S. degrees in computer science from Nanjing University, Nanjing, China, in 2005 and 2008 respectively, and the Ph.D. degree in computer science from the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, in 2014, working on topics in computer vision and image processing. After graduation, he joined the Media Algorithm Team, Youtube/Google. His research interests include video processing infrastructure, video quality assessment, and video compression.



NEIL BIRKBECK received the Ph.D. from the University of Alberta, Edmonton, AB, Canada, in 2011, working on topics in computer vision, graphics and robotics, with a specific focus on image-based modeling and rendering. He went on to become a Research Scientist with Siemens corporate research working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Software Engineer with Media Algorithms Team, YouTube/Google, with research interests in perceptual video processing, video coding, and video quality assessment.

BALU ADSUMILLI received the master's degree from the University of Wisconsin Madison, Madison, WI, USA, in 2002, and the Ph.D. degree from the University of California Santa Barbara, Santa Barbara, CA, USA, in 2005, on watermark-based error resilience in video communications. He manages and leads the Media Algorithms Group with YouTube/Google. From 2005 to 2011, he was a Sr. Research Scientist with Citrix Online, and from 2011 to 2016, he was the Sr. Manager Advanced Technology with GoPro, at both places developing algorithms for images or video quality enhancement, compression, capture, and streaming. He has coauthored more than 120 papers and patents. His research interests include image or video processing, machine vision, video compression, spherical capture, VR/AR, visual effects, and related areas. He is an active Member of IEEE (and MMSP TC), ACM, SPIE, and VES.



ALAN C. BOVIK (Fellow, IEEE) is currently the Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin, Austin, TX, USA. His research interests include image processing, digital photography, digital television, digital streaming video, and visual perception. For his work in these areas he was the recipient of the 2019 Progress Medal from The Royal Photographic Society, the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from The Optical Society, the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, the 2020 Technology and Engineering Emmy Award from the National Academy for Television Arts and Sciences, and the Norbert Wiener Society Award and the Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. He was also the recipient of about ten Best Journal Paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. His books include *The Essential Guides to Image and Video Processing*. He co-founded and was the longest-serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING, and also created/Chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, 1994.