UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content

Zhengzhong Tu, *Graduate Student Member, IEEE*, Yilin Wang, *Member, IEEE*, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*— Recent years have witnessed an explosion of user-generated content (UGC) videos shared and streamed over the Internet, thanks to the evolution of affordable and reliable consumer capture devices, and the tremendous popularity of social media platforms. Accordingly, there is a great need for accurate video quality assessment (VQA) models for UGC/consumer videos to monitor, control, and optimize this vast content. Blind quality prediction of in-the-wild videos is quite challenging, since the quality degradations of UGC videos are unpredictable, complicated, and often commingled. Here we contribute to advancing the UGC-VQA problem by conducting a comprehensive evaluation of leading no-reference/blind VQA (BVQA) features and models on a fixed evaluation architecture, yielding new empirical insights on both subjective video quality studies and objective VQA model design. By employing a feature selection strategy on top of efficient BVQA models, we are able to extract 60 out of 763 statistical features used in existing methods to create a new fusion-based model, which we dub the VIDeo quality EVALuator (VIDEVAL), that effectively balances the trade-off between VQA performance and efficiency. Our experimental results show that VIDEVAL achieves state-of-the-art performance at considerably lower computational cost than other leading models. Our study protocol also defines a reliable benchmark for the UGC-VQA problem, which we believe will facilitate further research on deep learning-based VQA modeling, as well as perceptually-optimized efficient UGC video processing, transcoding, and streaming. To promote reproducible research and public evaluation, an implementation of VIDEVAL has been made available online: https://github.com/vztu/VIDEVAL.**

***Index Terms*— Video quality assessment, image quality assessment, no-reference/blind, user-generated content.**

I. Introduction

VIDEO dominates the Internet. In North America, Netflix and YouTube alone account for more than fifty percent of downstream traffic, and there are many other significant video service providers. Improving the efficiency of video encoding, storage, and streaming over communication networks is a principle goal of video sharing and streaming platforms. One relevant and essential research direction is the perceptual optimization of rate-distortion tradeoffs in video encoding and streaming, where distortion (or quality) is usually modeled using video quality assessment (VQA) algorithms that can predict human judgements of video quality. This has motivated years of research on the topics of perceptual video and image quality assessment (VQA/IQA).

VQA research can be divided into two closely related categories: subjective video quality studies and objective video quality modeling. Subjective video quality research usually requires substantial resources devoted to time- and labor-consuming human studies to obtain valuable and reliable subjective data. The datasets obtained from subjective studies are invaluable for the development, calibration, and benchmarking of objective video quality models that are consistent with subjective mean opinion scores (MOS).

Hence, researchers have devoted considerable efforts on the development of high-quality VQA datasets that benefit the video quality community. Table I summarizes the ten-year evolution of popular public VQA databases. The first successful VQA database was the LIVE Video Quality Database [1], which was first made publicly available in 2008. It contains 10 pristine high-quality videos subjected to compression and transmission distortions. Other similar databases targeting simulated compression and transmission distortions have been released subsequently, including EPFL-PoliMI [2], VQEG-HDTV [3], IVP [4], TUM 1080p50 [5], CSIQ [6], MCL-V [7], and MCL-JCV [8]. All of the above mentioned datasets are based on a small set of high-quality videos, dubbed "pristine" or "reference," then synthetically distorting them in a controlled manner. We will refer to these kinds of synthetically-distorted video sets as *legacy* VQA databases. Legacy databases are generally characterized by only a small number of unique contents, each simultaneously degraded by only one or at most two synthetic distortions. For most practical scenarios, these are too simple to represent the great variety of real-world videos, and hence, VQA models derived on these databases may be insufficiently generalizable to large-scale realistic commercial VQA applications.

Recently, there has been tremendous growth in social media, where huge volumes of user-generated content (UGC) is shared over the media platforms such as YouTube, Facebook, and TikTok. Advances in powerful and affordable mobile devices and cloud computing techniques, combined with significant advances in video streaming have made it easy for most consumers to create, share, and view UGC pictures/videos instantaneously across the globe. Indeed, the prevalence of UGC has started to shift the focus of video quality research from *legacy* synthetically-distorted databases to newer, larger-scale authentic UGC datasets, which are being used to create solutions to what we call the

Manuscript received May 18, 2020; revised December 20, 2020; accepted April 1, 2021. Date of publication April 15, 2021; date of current version April 22, 2021. This work was supported by Google. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Baoxin Li. *(Corresponding author: Zhengzhong Tu.)*

Zhengzhong Tu and Alan C. Bovik are with the Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: zhengzhong.tu@utexas.edu; bovik@utexas.edu).

Yilin Wang, Neil Birkbeck, and Balu Adsumilli are with the YouTube Media Algorithms Team, Google LLC, Mountain View, CA 94043 USA (e-mail: yilin@google.com; birkbeck@google.com; badsumilli@google.com).

Digital Object Identifier 10.1109/TIP.2021.3072221

1941-0042 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

TABLE I

EVOLUTION OF POPULAR PUBLIC VIDEO QUALITY ASSESSMENT DATABASES: FROM LEGACY LAB STUDIES OF SYNTHETICALLY DISTORTED VIDEO SETS TO LARGE-SCALE CROWDSOURCED USER-GENERATED CONTENT (UGC) VIDEO DATASETS WITH AUTHENTIC DISTORTIONS

DATABASE	YEAR	#CONT	#TOTAL	RESOLUTION	FR	LEN	FORMAT	DISTORTION TYPE	#SUBJ	#RATES	DATA	ENV
LIVE-VQA	2008	10	160	768×432	25/50	10	YUV+264	Compression, transmission	38	29	DMOS+σ	In-lab
EPFL-PoliMI	2009	12	156	CIF/4CIF	25/30	10	YUV+264	Compression, transmission	40	34	MOS	In-lab
VQEG-HDTV	2010	49	740	1080i/p	25/30	10	AVI	Compression, transmission	120	24	RAW	In-lab
IVP	2011	10	138	1080p	25	10	YUV	Compression, transmission	42	35	DMOS+σ	In-lab
TUM 1080p50	2012	5	25	1080p	50	10	YUV	Compression	21	21	MOS	In-lab
CSIQ	2014	12	228	832×480	24-60	10	YUV	Compression, transmission	35	N/A	DMOS+σ	In-lab
CVD2014	2014	5	234	720p, 480p	9-30	10-25	AVI	Camera capture (authentic)	210	30	MOS	In-lab
MCL-V	2015	12	108	1080p	24-30	6	YUV	Compression, scaling	45	32	MOS	In-lab
MCL-JCV	2016	30	1560	1080p	24-30	5	MP4	Compression	150	50	RAW-JND	In-lab
KoNViD-1k	2017	1200	1200	540p	24-30	8	MP4	Diverse distortions (authentic)	642	114	MOS+σ	Crowd
LIVE-Qualcomm	2018	54	208	1080p	30	15	YUV	Camera capture (authentic)	39	39	MOS	In-lab
LIVE-VQC	2018	585	585	1080p-240p	19-30	10	MP4	Diverse distortions (authentic)	4776	240	MOS	Crowd
YouTube-UGC	2019	1380	1380	4k-360p	15-60	20	MKV	Diverse distortions (authentic)	>8k	123	MOS+σ	Crowd

#CONT: Total number of unique contents. #TOTAL: Total number of test sequences, including reference and distorted videos.
RESOLUTION: Video resolution (p: progressive). FR: Framerate. LEN: Video duration/length (in seconds).
FORMAT: Video container. #SUBJ: Total number of subjects in the study. #RATES: Average number of subjective ratings per video.
ENV: Subjective testing environment. In-lab: study was conducted in a laboratory. Crowd: study was conducted by crowdsourcing.

UGC-VQA problem. UGC-VQA studies typically follow a new design paradigm whereby: 1) All the source content is consumer-generated instead of professional-grade, thus suffers from unknown and highly diverse impairments; 2) they are only suitable for testing and comparing no-reference models, since reference videos are unavailable; 3) the types of distortions are authentic and commonly intermixed, and include but are not limited to capture impairments, editing and processing artifacts, compression, transcoding, and transmission distortions. Moreover, compression artifacts are not necessarily the dominant factors affecting video quality, unlike legacy VQA datasets and algorithms. These unpredictable perceptual degradations make perceptual quality prediction of UGC consumer videos very challenging.

Here we seek to address and gain insights into this new challenge (UGC-VQA) by first, conducting a comprehensive benchmarking study of leading video quality models on several recently released large-scale UGC-VQA databases. We also propose a new fusion-based blind VQA (BVQA) algorithm, which we call the VIDeo quality EVALuator (VIDEVAL), which is created by the processes of feature selection from existing top-performing VQA models. The empirical results show that a simple aggregation of these known models can achieve state-of-the-art (SOTA) performance. We believe that our expansive study will inspire and drive future research on BVQA modeling for the challenging UGC-VQA problem, and also pave the way towards deep learning-based solutions.

The outline of this paper is as follows: Section II reviews and analyzes the three most recent large-scale UGC-VQA databases, while Section III briefly surveys the development of BVQA models. We introduce the proposed VIDEVAL model in Section IV, and provide experimental results in Section V. Finally, concluding remarks are given in Section VI.

II. UGC-VQA DATABASES

The first UGC-relevant VQA dataset containing authentic distortions was introduced as the Camera Video Database (CVD2014) [12], which consists of videos with in-the-wild distortions from 78 different video capture devices, followed by the similar LIVE-Qualcomm Mobile In-Capture Database [13]. These two databases, however, only modeled (camera) capture distortions on small numbers of not very diverse unique contents. Inspired by the first successful massive online crowdsourcing study of UGC picture quality [14], the authors of [10] created the KoNViD-1k video quality database, the first such resource for UGC videos. It consists of 1,200 public-domain videos sampled from the YFCC100M dataset [15], and was annotated by 642 crowd-workers. LIVE-VQC [9] was another large-scale UGC-VQA database with 585 videos, crowdsourced on Amazon Mechanical Turk to collect human opinions from 4,776 unique participants. The most recently published UGC-VQA database is the YouTube-UGC Dataset [11] comprising 1,380 20-second video clips sampled from millions of YouTube videos, which were rated by more than 8,000 human subjects. Table II summarizes the main characteristics of the three large-scale UGC-VQA datasets studied, while Figure 1 shows some representative snapshots of the source sequences for each database, respectively.

A. Content Diversity and MOS Distribution

As a way of characterizing the content diversity of the videos in each database, Winkler [16] suggested three quantitative attributes related to spatial activity, temporal activity, and colorfulness. Here we expand the set of attributes to include six low-level features including brightness, contrast, colorfulness [17], sharpness, spatial information (SI), and temporal information (TI), thereby providing a larger visual space in which to plot and analyze content diversities of the three UGC-VQA databases. To reasonably limit the computational cost, each of these features was calculated on every 10th frame, then was averaged over frames to obtain an overall feature representation of each content. For simplicity, we denote the features as $\{C_i\}, i = 1, 2, \ldots, 6$. Figure 2 shows the fitted kernel distribution of each selected feature. We also plotted the

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

(a) LIVE-VQC (b) KoNViD-1k (c) YouTube-UGC

Fig. 1. Sample frames of the video contents contained in the three large scale UGC-VQA databases: (a) LIVE-VQC [9], (b) KoNViD-1k [10], and (c) YouTube-UGC [11]. LIVE-VQC includes only natural contents captured by mobile devices, while KoNViD-1k and YouTube-UGC comprise of both natural videos, animations, and gaming sources. Note that YouTube-UGC video set is categorized whereas the others are not.

TABLE II

PUBLIC LARGE-SCALE USER-GENERATED CONTENT VIDEO QUALITY ASSESSMENT (UGC-VQA) DATABASES COMPARED: KONVID-1K [10], LIVE-VQC [9], AND YOUTUBE-UGC [11]

DATABASE ATTRIBUTE	KoNViD-1k	LIVE-VQC	YouTube-UGC
Number of contents	1200	585	1380
Video sources	YFCC100m (Flickr)	Captured (mobile devices)	YouTube
Video resolutions	540p	1080p,720p,480p,etc.	4k,1080p,720p,480p,360p
Video layouts	Landscape	Landscape,portrait	Landscape,portrait
Video framerates	24,25,30 fr/sec	20,24,25,30 fr/sec	15,20,24,25,30,50,60 fr/sec
Video lengths	8 seconds	10 seconds	20 seconds
Audio track included	Yes (97%)	Yes	No
Testing methodology	Crowdsourcing (CrowdFlower)	Crowdsourcing (AMT)	Crowdsourcing (AMT)
Number of subjects	642	4,776	>8,000
Number of ratings	136,800 (114 votes/video)	205,000 (240 votes/video)	170,159 (123 votes/video)
Rating scale	Absolute Category Rating 1-5	Continuous Rating 0-100	Continuous Rating 1-5
Content remarks	Videos sampled from YFCC100m via a feature space of blur, colorfulness, contrast, SI, TI, and NIQE; Some contents irrelevant to quality research; Content was clipped from the original and resized to 540p.	Videos manually captured by certain people; Content including many camera motions; Content including some night scenes that are prone to be outliers; Resolutions not uniformly distributed.	Videos sampled from YouTube via a feature space of spatial, color, temporal, and chunk variation; Contents categorized into 15 classes, including HDR, screen content, animations, and gaming videos.
Study remarks	Study did not account for or remove videos on which stalling events occurred when viewed; test methodology prone to unreliable individual scores.	Distribution of MOS values slightly skewed towards higher scores; standard deviation statistics of MOS were not provided.	Distribution of MOS values slightly skewed towards higher values; three additional chunk MOS scores with standard deviation were provided.

convex hulls of paired features, to show the feature coverage of each database, in Figure 3. To quantify the coverage and uniformity of these databases over each defined feature space, we computed the relative range and uniformity of coverage [16], where the relative range is given by:

$$\mathrm{R}_i^k = \frac{\max(\mathrm{C}_i^k) - \min(\mathrm{C}_i^k)}{\max_k(\mathrm{C}_i^k)}, \tag{1}$$

where C_i^k denotes the feature distribution of database k for a given feature dimension i, and $\max_k(\mathrm{C}_i^k)$ specifies the maximum value for that given dimension across all databases.

Uniformity of coverage measures how uniformly distributed the videos are in each feature dimension. We computed this as the entropy of the B-bin histogram of C_i^k over all sources for each database indexed k:

$$\mathrm{U}_i^k = -\sum_{b=1}^{\mathrm{B}} p_b \log_{\mathrm{B}} p_b, \tag{2}$$

where p_b is the normalized number of sources in bin b at feature i for database k. The higher the uniformity the more uniform the database is. Relative range and uniformity of coverage are plotted in Figure 4 and Figure 5, respectively, quantifying the intra- and inter-database differences in source content characteristics.

We also extracted 4,096-dimensional VGG-19 [18] deep features and embedded these features into 2D subspace using t-SNE [19] to further compare content diversity, as shown in Figure 7. Apart from content diversity expressed in terms of visual features, the statistics of the subjective ratings are another important attribute of each video quality database. The main aspect considered in the analysis here is the distributions of mean opinion scores (MOS), as these are indicative of the quality range of the subjective judgements. The analysis of standard deviation of MOS is not presented here since it is not provided in LIVE-VQC. Figure 6 displays the histogram of MOS distributions for the three UGC-VQA databases.

B. Observations

We make some observations from the above plots. As may be seen in Figures 2a and 2b, and the corresponding convex hulls in Figure 3, KoNViD-1k and YouTube-UGC exhibit

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

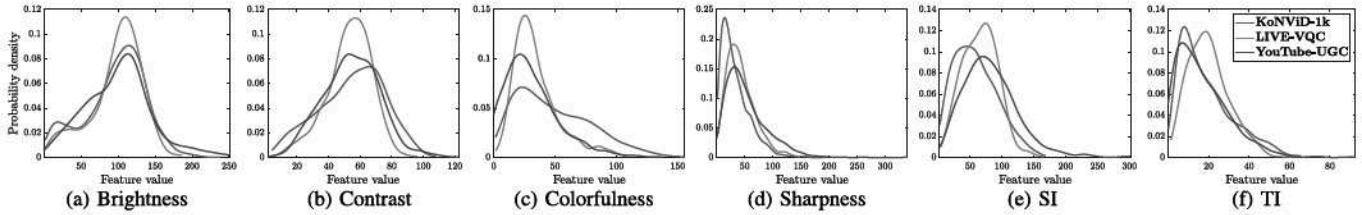


Fig. 2. Feature distribution comparisons among the three considered UGC-VQA databases: KoNViD-1k, LIVE-VQC, and YouTube-UGC.

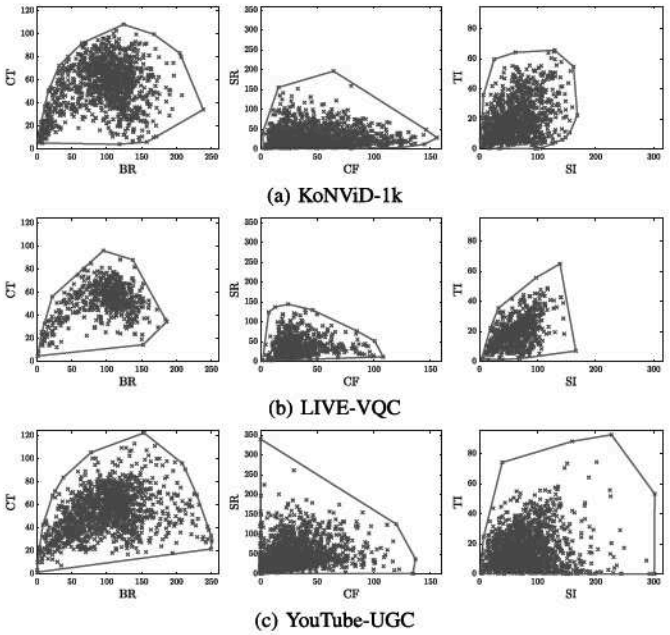


Fig. 3. Source content (blue 'x') distribution in paired feature space with corresponding convex hulls (orange boundaries). Left column: BR×CT, middle column: CF×SR, right column: SI×TI.

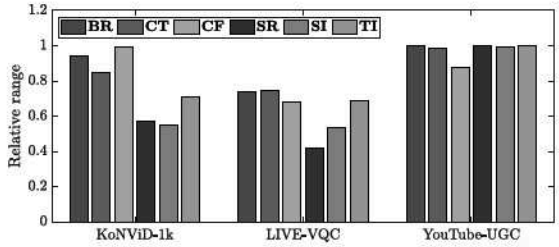


Fig. 4. Relative range R_i^k comparisons of the selected six features calculated on the three UGC-VQA databases: KoNViD-1k, LIVE-VQC, and YouTube-UGC.

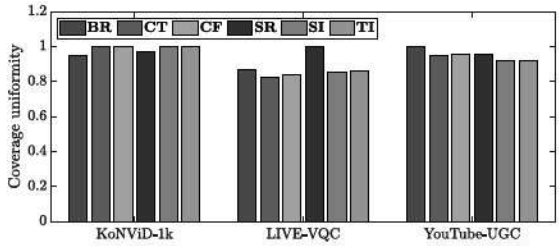


Fig. 5. Comparison of coverage uniformity U_i^k of the selected six features computed on the three UGC-VQA databases: KoNViD-1k, LIVE-VQC, and YouTube-UGC.

similar coverage in terms of brightness and contrast, while LIVE-VQC adheres closer to middle values. Regarding colorfulness, KoNViD-1k shows a skew towards higher scores than the other two datasets, which is consistent with the observations that Flickr users self-characterize as either professional video/photographers or as dedicated amateurs. On the sharpness and SI histograms, YouTube-UGC is spread most widely, while KoNViD-1k is concentrated on lower values. Another interesting finding from the TI statistics: LIVE-VQC is distributed more towards higher values than YouTube-UGC and KoNViD-1k, consistent with our observation that videos in LIVE-VQC were captured in the presence of larger and more frequent camera motions. We will revisit this interesting aspect of TI when evaluating the BVQA models in Section V. The visual comparison in Figure 7 shows that YouTube-UGC and KoNViD-1k span a wider range of VGG-19 feature space than does LIVE-VQC, indicating significant content diversity differences. Figure 6 shows the MOS distributions: all three databases have right-skewed MOS distributions, with KoNViD-1k less so, and LIVE-VQC and YouTube-UGC more so. The overall ranges and uniformity comparisons in Figures 4, 5, and 7 suggest that constructing a database by crawling and sampling from a large content repository is likely to yield a more content-diverse, uniformly-distributed dataset than one created from pictures or videos captured directly from a set of user cameras. Both cases may be argued to be realistic in some scenario.

III. UGC-VQA Models

The goal of subjective video quality studies is to motivate the development of automatic objective video quality models. Conventionally, objective video quality assessment can be classified into three main categories: full-reference (FR), reduced-reference (RR), and no-reference (NR) models. FR-VQA models require the availability of an entire pristine source video to measure visual differences between a target signal and a corresponding reference [20]–[23], while RR-VQA models only make use of a limited amount of reference information [24], [25]. Some popular FR-VQA models, including PSNR, SSIM [26], and VMAF [21] have already been successfully and massively deployed to optimize streaming and shared/uploaded video encoding protocols by leading video service providers. NR-VQA or BVQA models, however, rely solely on analyzing the test stimuli without the benefit of any corresponding "ground truth" pristine signal. It is obvious that only BVQA models are appropriate for the UGC-VQA problem. Here we briefly review the evolution of BVQA models, from conventional handcrafted feature-based approaches, on to convolutional neural network-based models.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

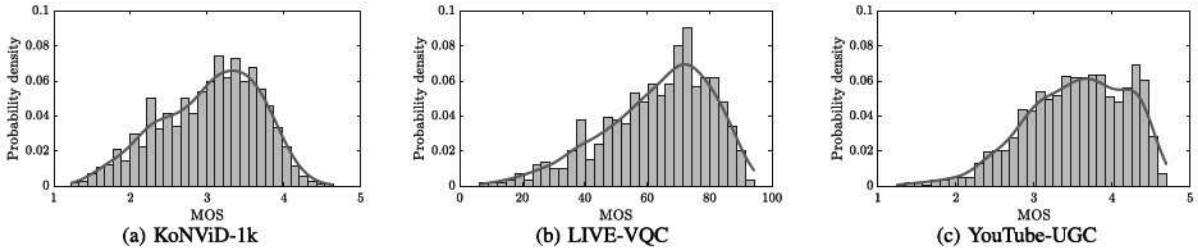


Fig. 6. MOS histograms and the fitted kernel distributions of the three UGC-VQA databases: KoNViD-1k, LIVE-VQC, and YouTube-UGC.

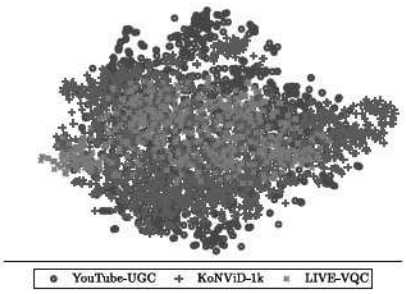


Fig. 7. VGG-19 deep feature embedding via t-SNE [19] on KoNViD-1k, LIVE-VQC, and YouTube-UGC, respectively.

A. Conventional Feature-Based BVQA Models

Almost all of the earliest BVQA models have been 'distortion specific,' meaning they were designed to quantify a specific type of distortion such as blockiness [27], blur [28], ringing [29], banding [30]–[32], or noise [33], [34] in distorted videos, or to assess multiple specific coincident distortion types caused by compression or transmission impairments [35], [36]. More recent top-performing BVQA models are almost exclusively learning-based, leveraging a set of generic quality-aware features, combined to conduct quality prediction by machine learning regression [37]–[45]. 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 good-quality scene statistics [46]. These natural scene statistics (NSS) models predictably deviate in the presence of distortions, thereby characterizing perceived quality degradations [47]. Successful blind picture quality assessment (BIQA) models of this type have been developed in the wavelet (BIQI [48], DIIVINE [37], C-DIIVINE [49]), discrete cosine transform (BLIINDS [50], BLIINDS-II [51]), curvelet [52], and spatial intensity domains (NIQE [53], BRISQUE [38]), and have further been extended to video signals using natural bandpass space-time video statistics models [39], [54]–[56], among which the most well-known model is the Video-BLIINDS [39]. Other extensions to empirical NSS include the joint statistics of the gradient magnitude and Laplacian of Gaussian responses in the spatial domain (GM-LOG [57]), in log-derivative and log-Gabor spaces (DESIQUE [58]), as well as in the gradient domain of LAB color transforms (HIGRADE [40]). The FRIQUEE model [41] has been observed to achieve SOTA performance both on UGC/consumer video/picture databases like LIVE-Challenge [14], CVD2014 [12], and KoNViD-1k [10] by leveraging a bag of NSS features drawn from diverse color spaces and perceptually motivated transform domains.

Instead of using NSS-inspired feature descriptors, methods like CORNIA [43] employ unsupervised learning techniques to learn a dictionary (or codebook) of distortions from raw image patches, and was further extended to Video CORNIA [59] by applying an additional temporal hysteresis pooling [60] of learned frame-level quality scores. Similar to CORNIA, the authors of [61] proposed another codebook-based general-purpose BVQA method based on High Order Statistics Aggregation (HOSA), requiring only a small codebook, yet yielding promising performance.

A very recent handcrafted feature-based BVQA model is the "two level" video quality model (TLVQM) [42], wherein a two-level feature extraction mechanism is adopted to achieve efficient computation of a set of carefully-defined impairment/distortion-relevant features. Unlike NSS features, TLVQM selects a comprehensive feature set comprising of empirical motion statistics, specific artifacts, and aesthetics. TLVQM does require that a large set of parameters (around 30) be specified, which may affect performance on datasets or application scenarios it has not been exposed to. The model currently achieves SOTA performance on three UGC video quality databases, CVD2014 [12], KoNViD-1k [10], and LIVE-Qualcomm [13], at a reasonably low complexity, as reported by the authors.

B. Deep Convolutional Neural Network-Based BVQA Models

Deep convolutional neural networks (CNNs or ConvNets) have been shown to deliver standout performance on a wide variety of low-level computer vision applications. Recently, the release of several "large-scale" (in the context of IQA/VQA research) subjective quality databases [10], [14] have sped the application of deep CNNs to perceptual quality modeling. For example, several deep learning picture-quality prediction methods were proposed in [62]–[65]. To conquer the limits of data scale, they either propose to conduct patch-wise training [62], [63], [66] using global scores, or by pretraining deep nets on ImageNet [67], then fine tuning. Several authors report SOTA performance on legacy synthetic distortion databases [68], [69] or on naturally distorted databases [14], [70].

Among the applications of deep CNNs to blind video quality prediction, Kim [71] proposed a deep video quality assessor (DeepVQA) to learn the spatio-temporal visual sensitivity

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

TABLE III
SUMMARY OF THE INITIAL FEATURE SET AND THE FINALIZED VIDEVAL SUBSET AFTER FEATURE SELECTION

FEATURE NAME	FEATURE INDEX	#($\mathcal{F}_{\text{INIT}}$)	#(VIDEVAL)
$\text{BRISQUE}_{\text{avg}}$	$f_1 - f_{36}$	36	3
$\text{BRISQUE}_{\text{std}}$	$f_{37} - f_{72}$	36	1
$\text{GM-LOG}_{\text{avg}}$	$f_{73} - f_{112}$	40	4
$\text{GM-LOG}_{\text{std}}$	$f_{113} - f_{152}$	40	5
$\text{HIGRADE-GRAD}_{\text{avg}}$	$f_{153} - f_{188}$	36	8
$\text{HIGRADE-GRAD}_{\text{std}}$	$f_{189} - f_{224}$	36	1
$\text{FRIQUEE-LUMA}_{\text{avg}}$	$f_{225} - f_{298}$	74	4
$\text{FRIQUEE-LUMA}_{\text{std}}$	$f_{299} - f_{372}$	74	8
$\text{FRIQUEE-CHROMA}_{\text{avg}}$	$f_{373} - f_{452}$	80	10
$\text{FRIQUEE-CHROMA}_{\text{std}}$	$f_{453} - f_{532}$	80	1
$\text{FRIQUEE-LMS}_{\text{avg}}$	$f_{533} - f_{606}$	74	1
$\text{FRIQUEE-LMS}_{\text{std}}$	$f_{607} - f_{680}$	74	0
$\text{FRIQUEE-HS}_{\text{avg}}$	$f_{681} - f_{684}$	4	0
$\text{FRIQUEE-HS}_{\text{std}}$	$f_{685} - f_{688}$	4	0
$\text{TLVQM-LCF}_{\text{avg}}$	$f_{689} - f_{710}$	22	5
$\text{TLVQM-LCF}_{\text{std}}$	$f_{711} - f_{733}$	23	3
TLVQM-HCF	$f_{734} - f_{763}$	30	6
$\mathcal{F}_{\text{ALL}}$	$f_1 - f_{763}$	763	60

*All the spatial features are calculated every two frames and aggregated into a single feature vector within 1-sec chunks. The overall feature vector for the whole video is then obtained by averaging all the chunk-wise feature vectors. Subscript *avg* means within-chunk average pooling, whereas subscript *std* means within-chunk standard deviation pooling.

maps via a deep ConvNet and a convolutional aggregation network. The V-MEON model [72] used a multi-task CNN framework which jointly optimizes a 3D-CNN for feature extraction and a codec classifier using fully-connected layers to predict video quality. Zhang [73] leveraged transfer learning to develop a general-purpose BVQA framework based on weakly supervised learning and a resampling strategy. In the VSFA model [74], the authors applied a pre-trained image classification CNN as a deep feature extractor and integrated the frame-wise deep features using a gated recurrent unit and a subjectively-inspired temporal pooling layer, and reported leading performance on several natural video databases [10], [12], [13]. These SOTA deep CNN-based BVQA models [71]–[74] produce accurate quality predictions on legacy (single synthetic distortion) video datasets [1], [6], but struggle on recent in-the-wild UGC databases [10], [12], [13].

IV. FEATURE FUSED VIDEO QUALITY EVALUATOR (VIDEVAL)

We have just presented a diverse set of BVQA models designed from a variety of perspectives, each either based on scene statistics, or motivated by visual impairment heuristics. As might be expected, and as we shall show later, the performances of these models differ, and also vary on different datasets. We assume that the features extracted from different models may represent statistics of the signal in different perceptual domains, and henceforce, a selected fusion of BVQA models may be expected to deliver better consistency against subjective assessment, and also to achieve more reliable performance across different databases and use cases. This inspired our new feature fused VIDeo quality EVALuator (VIDEVAL), as described next.

We begin by constructing an initial feature set on top of existing high-performing, compute-efficient BVQA models and features, distilled through a feature selection program. The goal of feature selection is to choose an optimal or sub-optimal feature subset $\mathcal{F}_k \in \mathbb{R}^k$ from the initial feature set $\mathcal{F}_{\text{INIT}} \in \mathbb{R}^N$ (where $k < N$) that achieves nearly top performance but with many fewer features.

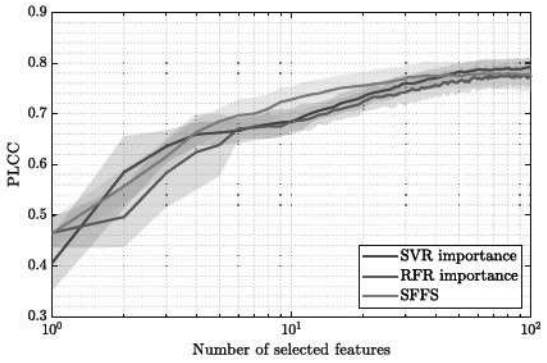


Fig. 8. Feature selection performance (PLCC) of three selected algorithms as a function of k on the All-Combined$_c$ dataset. The shaded error bar denotes the standard deviation of PLCC over 10 iterations.

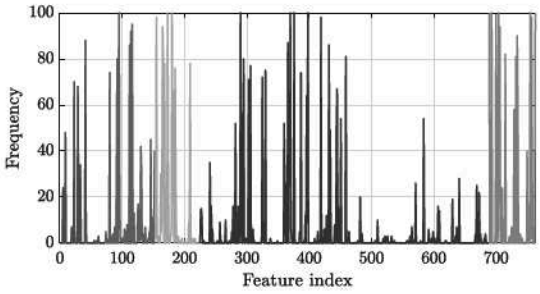


Fig. 9. Visualization of the second step in feature selection: frequency of each feature being selected over 100 iterations of train-test splits using SVR importance selection method with fixed $k = 60$.

A. Feature Extraction

We construct an initial feature set by selecting features from existing top-performing BVQA models. For practical reasons, we ignore features with high computational cost, e.g., certain features from DIIVINE, BLIINDS, C-DIIVINE, and V-BLIINDS. We also avoid using duplicate features in different models, such as the BRISQUE-like features in HIGRADE, and the C-DIIVINE features in V-BLIINDS. This filtering process yields the initial feature candidates, which we denote as BRISQUE, GM-LOG, HIGRADE-GRAD, FRIQUEE-LUMA, FRIQUEE-CHROMA, FRIQUEE-LMS, FRIQUEE-HS, TLVQM-LCF, and TLVQM-HCF.

Inspired by the efficacy of standard deviation pooling as first introduced in GMSD [75] and later also used in TLVQM [42], we calculate these spatial features every second frame within each sequentially cut non-overlapping one-second chunk, then we enrich the feature set by applying average and standard deviation pooling of frame-level features within each chunk, based on the hypothesis that the variation of spatial NSS features also correlates with the temporal properties of the video. Finally, all the chunk-wise feature vectors are average pooled [76] across all the chunks to derive the final set of features for the entire video. Table III indexes and summarizes

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

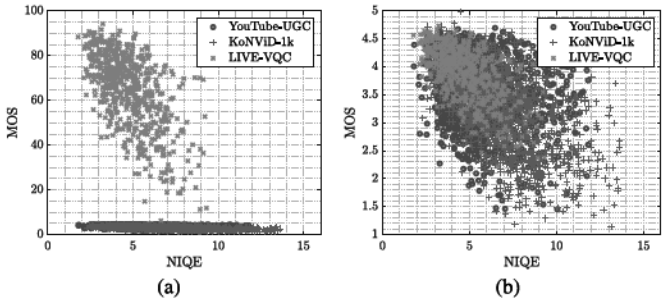


Fig. 10. Scatter plots of MOS versus NIQE scores (a) before, and (b) after INLSA calibration [77] using YouTube-UGC as the reference set.

the selected features in the initial feature set, yielding an overall 763-dimensional feature vector, $\mathcal{F}_{\text{INIT}} \in \mathbb{R}^{763}$.

B. Feature Selection

We deploy two types of feature selection algorithms to distill the initial feature set. The first method is a model-based feature selector that utilizes a machine learning model to suggest features that are important. We employed the popular random forest (RF) to fit a regression model and eliminate the least significant features sorted by permutation importance. We also trained a support vector machine (SVM) with the linear kernel to rank the features, as a second model selector. Another sub-optimal solution is to apply a greedy search approach to find a good feature subset. Here we employed Sequential Forward Floating Selection (SFFS), and used SVM as the target regressor with its corresponding mean squared error between the predictions and MOS as the cost function. The mean squared error is calculated by cross-validation measures of predictive accuracy to avoid overfitting.

One problem with feature selection is that we do not know *a priori* what k to select, i.e., how many features are needed. Therefore, we conducted a two-step feature selection procedure. First, we evaluated the feature selection methods as a function of k via 10 train-test iterations, to select the best algorithm with corresponding optimal k. Figure 8 shows the median PLCC (defined in Section V-A) performance with respect to k for different feature selection models, based on which we finally chose the SVM importance method with $k = 60$ in our next experiments. In the second step, we applied the best feature selection algorithm with the fixed best k over 100 random train-test splits. On each iteration, a subset is selected from the feature selector, based on which the frequency of each feature over the iterations is counted, and the j most frequently occurring features are included into the final feature set. Figure 9 shows the frequency of each feature being selected over 100 random splits in the second step. This selection process is implemented on a combined dataset constructed from three independent databases, as described in Section V-A. Table III summarizes the results of the feature selection procedure (SVR importance with $k = 60$), yielding the final proposed VIDEVAL model.

V. Experimental Results

A. Evaluation Protocol

1) UGC Dataset Benchmarks: To conduct BVQA performance evaluation, we used the three UGC-VQA databases:

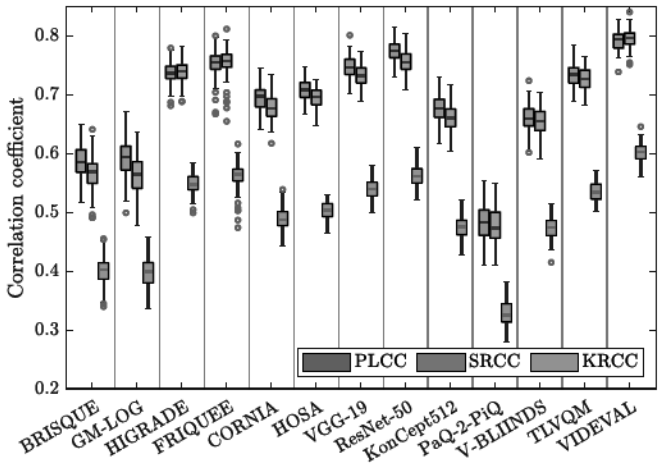


Fig. 11. Box plots of PLCC, SRCC, and KRCC of evaluated learning-based BVQA algorithms on the All-Combined$_c$ dataset over 100 random splits. For each box, median is the central box, and the edges of the box represent 25th and 75th percentiles, while red circles denote outliers.

TABLE IV
Performance Comparison of Evaluated Opinion-Unaware "Completely Blind" BVQA Models

Dataset	Model \ Metric	SRCC↑	KRCC↑	PLCC↑	RMSE↓
KoNViD	NIQE (1 fr/sec)	0.5417	0.3790	0.5530	0.5336
	ILNIQE (1 fr/sec)	0.5264	0.3692	0.5400	0.5406
	VIIDEO	0.2988	0.2036	0.3002	0.6101
LIVE-C	NIQE (1 fr/sec)	0.5957	0.4252	0.6286	13.110
	ILNIQE (1 fr/sec)	0.5037	0.3555	0.5437	14.148
	VIIDEO	0.0332	0.0231	0.2146	16.654
YT-UGC	NIQE (1 fr/sec)	0.2379	0.1600	0.2776	0.6174
	ILNIQE (1 fr/sec)	0.2918	0.1980	0.3302	0.6052
	VIIDEO	0.0580	0.0389	0.1534	0.6339
All-Comb	NIQE (1 fr/sec)	0.4622	0.3222	0.4773	0.6112
	ILNIQE (1 fr/sec)	0.4592	0.3213	0.4741	0.6119
	VIIDEO	0.1039	0.0688	0.1621	0.6804

KoNViD-1K [10], LIVE-VQC [9], and YouTube-UGC [11]. We found that the YouTube-UGC dataset contains 57 grayscale videos, which yield numerical errors when computing the color model FRIQUEE. Therefore, we extracted a subset of 1,323 color videos from YouTube-UGC, which we denote here as the YouTube-UGC$_c$ set, for the evaluation of color models. In order to study overall model performances on all the databases, we created a large composite benchmark, which is referred to here as All-Combined$_c$, using the iterative nested least squares algorithm (INLSA) suggested in [77], wherein YouTube-UGC is selected as the anchor set, and the objective MOS from the other two sets, KoNViD-1k and LIVE-VQC, are linearly mapped onto a common scale ($[1, 5]$). Figure 10 shows scatter plots of MOS versus NIQE scores before (Figure 10a) and after (Figure 10b) INLSA linear mapping, calibrated by NIQE [53] scores. The All-Combined$_c$ (3,108) dataset is simply the union of KoNViD-1k (1,200), LIVE-VQC (575), and YouTube-UGC$_c$ (1,323) after MOS calibration:

$$y_{\text{adj}} = 5 - 4 \times \left[(5 - y_{\text{org}})/4 \times 1.1241 - 0.0993\right] \quad (3)$$

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

where (3) and (4) are for calibrating KoNViD-1k and LIVE-VQC, respectively. y_{adj} denotes the adjusted scores, while y_{org} is the original MOS.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

TABLE V

PERFORMANCE COMPARISON OF EVALUATED BVQA MODELS ON THE FOUR BENCHMARK DATASETS. THE UNDERLINED AND BOLDFACED ENTRIES INDICATE THE BEST AND TOP THREE PERFORMERS ON EACH DATABASE FOR EACH PERFORMANCE METRIC, RESPECTIVELY

DATASET	KoNViD-1k				LIVE-VQC			
MODEL \ METRIC	SRCC↑ (STD)	KRCC↑ (STD)	PLCC↑ (STD)	RMSE↓ (STD)	SRCC↑ (STD)	KRCC↑ (STD)	PLCC↑ (STD)	RMSE↓ (STD)
BRISQUE (1 fr/sec)	0.6567 (.035)	0.4761 (.029)	0.6576 (.034)	0.4813 (.022)	0.5925 (.068)	0.4162 (.052)	0.6380 (.063)	13.100 (.796)
GM-LOG (1 fr/sec)	0.6578 (.032)	0.4770 (.026)	0.6636 (.031)	0.4818 (.022)	0.5881 (.068)	0.4180 (.052)	0.6212 (.063)	13.223 (.822)
HIGRADE (1 fr/sec)	0.7206 (.030)	0.5319 (.026)	0.7269 (.028)	0.4391 (.018)	0.6103 (.068)	0.4391 (.054)	0.6332 (.065)	13.027 (.904)
FRIQUEE (1 fr/sec)	0.7472 (.026)	0.5509 (.024)	0.7482 (.025)	0.4252 (.017)	0.6579 (.053)	0.4770 (.043)	0.7000 (.058)	12.198 (.914)
CORNIA (1 fr/sec)	0.7169 (.024)	0.5231 (.021)	0.7135 (.023)	0.4486 (.018)	0.6719 (.047)	0.4849 (.039)	0.7183 (.042)	11.832 (.700)
HOSA (1 fr/sec)	0.7654 (.022)	0.5690 (.021)	0.7664 (.020)	0.4142 (.016)	0.6873 (.046)	0.5033 (.039)	**0.7414 (.041)**	**11.353 (.747)**
VGG-19 (1 fr/sec)	**0.7741 (.028)**	**0.5841 (.027)**	**0.7845 (.024)**	**0.3958 (.017)**	0.6568 (.053)	0.4722 (.044)	0.7160 (.048)	11.783 (.696)
ResNet-50 (1 fr/sec)	**0.8018 (.025)**	**0.6100 (.024)**	**0.8104 (.022)**	**0.3749 (.017)**	0.6636 (.051)	0.4786 (.042)	0.7205 (.043)	11.591 (.733)
KonCept512 (1 fr/sec)	0.7349 (.025)	0.5425 (.023)	0.7489 (.024)	0.4260 (.016)	0.6645 (.052)	0.4793 (.045)	0.7278 (.046)	11.626 (.767)
PaQ-2-PiQ (1 fr/sec)	0.6130 (.032)	0.4334 (.026)	0.6014 (.033)	0.5148 (.019)	0.6436 (.045)	0.4568 (.035)	0.6683 (.044)	12.619 (.848)
V-BLIINDS	0.7101 (.031)	0.5188 (.026)	0.7037 (.030)	0.4595 (.023)	**0.6939 (.050)**	**0.5078 (.042)**	0.7178 (.050)	11.765 (.828)
TLVQM	0.7729 (.024)	0.5770 (.022)	0.7688 (.023)	0.4102 (.017)	**0.7988 (.036)**	**0.6080 (.037)**	**0.8025 (.036)**	**10.145 (.818)**
VIDEVAL	**0.7832 (.021)**	**0.5845 (.021)**	**0.7803 (.022)**	**0.4026 (.017)**	**0.7522 (.039)**	**0.5639 (.036)**	**0.7514 (.042)**	**11.100 (.810)**

DATASET	YouTube-UGC				All-Combined$_c$†			
MODEL \ METRIC	SRCC↑ (STD)	KRCC↑ (STD)	PLCC↑ (STD)	RMSE↓ (STD)	SRCC↑ (STD)	KRCC↑ (STD)	PLCC↑ (STD)	RMSE↓ (STD)
BRISQUE (1 fr/sec)	0.3820 (.051)	0.2635 (.036)	0.3952 (.048)	0.5919 (.021)	0.5695 (.028)	0.4030 (.022)	0.5861 (.027)	0.5617 (.016)
GM-LOG (1 fr/sec)	0.3678 (.058)	0.2517 (.041)	0.3920 (.054)	0.5896 (.022)	0.5650 (.029)	0.3995 (.022)	0.5942 (.030)	0.5588 (.014)
HIGRADE (1 fr/sec)	**0.7376 (.033)**	**0.5478 (.028)**	**0.7216 (.033)**	**0.4471 (.024)**	0.7398 (.018)	0.5471 (.016)	0.7368 (.019)	0.4674 (.015)
FRIQUEE* (1 fr/sec)	**0.7652 (.030)**	**0.5688 (.026)**	**0.7571 (.032)**	**0.4169 (.023)**	**0.7568 (.023)**	**0.5651 (.021)**	**0.7550 (.022)**	**0.4549 (.018)**
CORNIA (1 fr/sec)	0.5972 (.041)	0.4211 (.032)	0.6057 (.039)	0.5136 (.024)	0.6764 (.021)	0.4846 (.017)	0.6974 (.020)	0.4946 (.013)
HOSA (1 fr/sec)	0.6025 (.034)	0.4257 (.026)	0.6047 (.034)	0.5132 (.021)	0.6957 (.018)	0.5038 (.015)	0.7082 (.016)	0.4893 (.013)
VGG-19 (1 fr/sec)	0.7025 (.028)	0.5091 (.023)	0.6997 (.028)	0.4562 (.020)	0.7321 (.018)	0.5399 (.016)	0.7482 (.017)	0.4610 (.013)
ResNet-50 (1 fr/sec)	0.7183 (.028)	0.5229 (.024)	0.7097 (.027)	0.4538 (.021)	**0.7557 (.017)**	**0.5613 (.016)**	**0.7747 (.016)**	**0.4385 (.013)**
KonCept512 (1 fr/sec)	0.5872 (.039)	0.4101 (.030)	0.5940 (.041)	0.5135 (.022)	0.6608 (.022)	0.4759 (.018)	0.6763 (.022)	0.5091 (.014)
PaQ-2-PiQ (1 fr/sec)	0.2658 (.047)	0.1778 (.032)	0.2935 (.049)	0.6153 (.019)	0.4727 (.029)	0.3242 (.021)	0.4828 (.029)	0.6081 (.015)
V-BLIINDS	0.5590 (.049)	0.3899 (.036)	0.5551 (.046)	0.5356 (.022)	0.6545 (.023)	0.4739 (.019)	0.6599 (.023)	0.5200 (.016)
TLVQM	0.6693 (.030)	0.4816 (.025)	0.6590 (.030)	0.4849 (.022)	0.7271 (.018)	0.5347 (.016)	0.7342 (.018)	0.4705 (.013)
VIDEVAL*	**0.7787 (.025)**	**0.5830 (.023)**	**0.7733 (.025)**	**0.4049 (.021)**	**0.7960 (.015)**	**0.6032 (.014)**	**0.7939 (.015)**	**0.4268 (.015)**

*FRIQUEE and VIDEVAL were evaluated on a subset of 1,323 color videos in YouTube-UGC, denoted YouTube-UGC$_c$, since it yields numerical errors when calculating on the remaining 57 grayscale videos. For the other BVQA models evaluated, no significant difference was observed when evaluated on YouTube-UGC$_c$ versus YouTube-UGC, and hence we still report the results on YouTube-UGC.

†For a fair comparison, we only combined and calibrated (via INLSA [77]) all the color videos from these three databases to obtain the combined dataset, i.e., All-Combined$_c$ (3,108) = KoNViD-1k (1,200) + LIVE-VQC (585) + YouTube-UGC$_c$ (1,323).

2) BVQA Model Benchmarks: We include a number of representative BVQA/BIQA algorithms in our benchmarking evaluation as references to be compared against. These baseline models include NIQE [53], ILNIQE [78], VIIDEO [55], BRISQUE [38], GM-LOG [57], HIGRADE [40], FRIQUEE [41], CORNIA [43], HOSA [61], KonCept512 [70], PaQ-2-PiQ [64], V-BLIINDS [39], and TLVQM [42]. Among these, NIQE, ILNIQE, and VIIDEO are "completely blind" (opinion-unaware (OU)), since no training is required to build them. The rest of the models are all training-based (opinion-aware (OA)) and we re-train the models/features when evaluating on a given dataset. We also utilized the well-known deep CNN models VGG-19 [18] and ResNet-50 [79] as additional CNN-based baseline models, where each was pretrained on the ImageNet classification task. The fully-connected layer (4,096-dim) from VGG-19 and average-pooled layer (2,048-dim) from ResNet-50 served as deep feature descriptors, by operating on 25 227×227 random crops of each input frame, then average-pooled into a single feature vector representing the entire frame [63]. Two SOTA deep BIQA models, KonCept512 [70] and PaQ-2-PiQ [64], were also included in our evaluations. We implemented the feature extraction process for each evaluated BVQA model using its initial released implementation in MATLAB R2018b, except that VGG-19 and ResNet-50 were implemented in TensorFlow, while KonCept512[1] and PaQ-2-PiQ[2] were implemented in PyTorch. All the feature-based BIQA models extract features at a uniform sampling rate of one frame per second, then temporally average-pooled to obtain the overall video-level feature.

3) Regression Models: We used a support vector regressor (SVR) as the back-end regression model to learn the feature-to-score mappings, since it achieves excellent performance in most cases [38], [39], [41], [42], [59], [63]. The effectiveness of SVR, however, largely depends on the selection of its hyperparameters. As recommended in [80], we optimized the SVR parameter values (C, γ) by a grid-search of 10×10 exponentially growing sequences (in our experiments, we used a grid of $C = 2^1, 2^2, \ldots, 2^{10}, \gamma = 2^{-8}, 2^{-7}, \ldots, 2^1$) using cross-validation on the training set. The pair (C, γ) yielding the best cross-validation performance, as measured by the root mean squared error (RMSE) between the predicted scores and the MOS, is picked. Afterward, the selected model parameters are applied to re-train the model on the entire training set, and we report the evaluation results on the test set. This

[1]https://github.com/ZhengyuZhao/koniq-PyTorch

[2]https://github.com/baidut/paq2piq

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

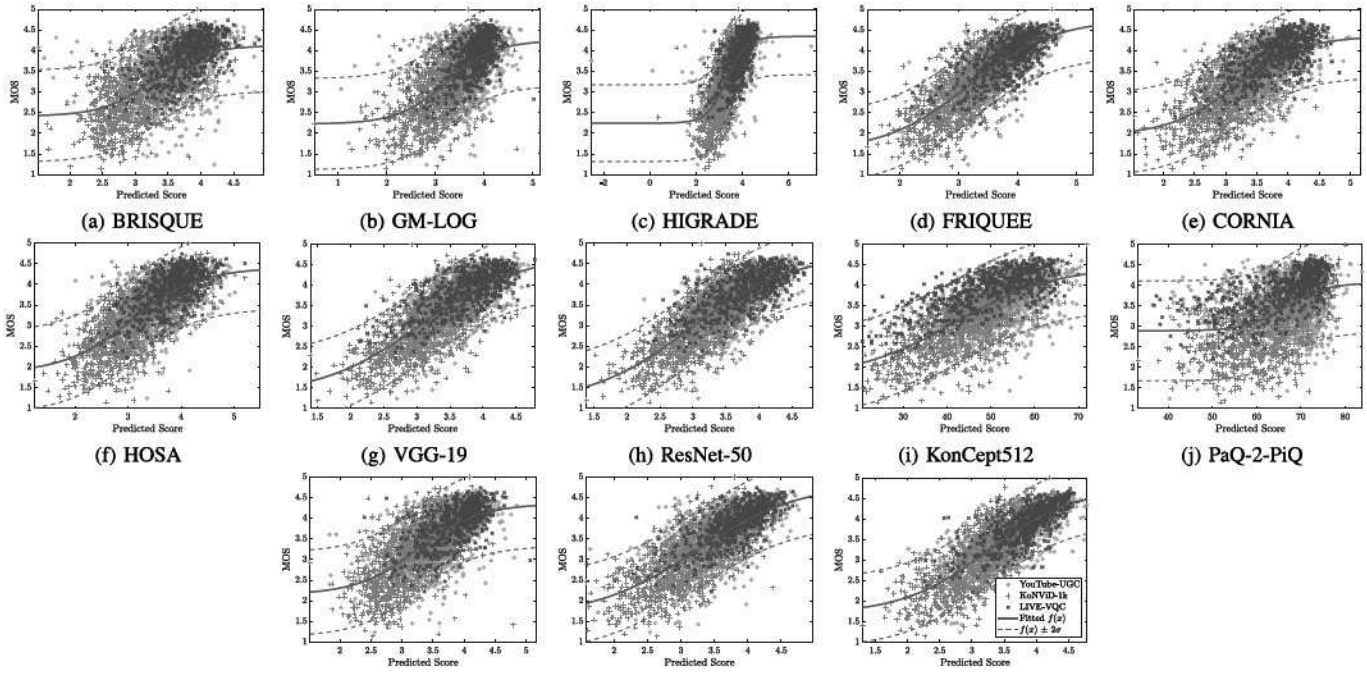


Fig. 12. Scatter plots and nonlinear logistic fitted curves of VQA models versus MOS trained with a grid-search SVR using k-fold cross-validation on the All-Combined$_c$ set. (a) BRISQUE (1 fr/sec), (b) GM-LOG (1 fr/sec), (c) HIGRADE (1 fr/sec), (d) FRIQUEE (1 fr/sec), (e) CORNIA (1 fr/sec), (f) HOSA (1 fr/sec), (g) VGG-19 (1 fr/sec), (h) ResNet-50 (1 fr/sec), (i) KonCept512 (1 fr/sec), (j) PaQ-2-PiQ (1 fr/sec), (k) V-BLIINDS, (l) TLVQM, and (m) VIDEVAL.

kind of cross-validation procedure can prevent over-fitting, thus providing fair evaluation of the compared BVQA models. We chose the linear kernel for CORNIA, HOSA, VGG-19, and ResNet-50, considering their large feature dimension, and the radial basis function (RBF) kernel for all the other algorithms. We used Python 3.6.7 with the scikit-learn toolbox to train and test all the evaluated learning-based BVQA models.

4) Performance Metrics: Following convention, we randomly split the dataset into non-overlapping training and test sets (80%/20%), where the regression model was trained on the training set, and the performance was reported on the test set. This process of random split was iterated 100 times and the overall median performance was recorded. For each iteration, we adopted four commonly used performance criteria to evaluate the models: 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) are 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 of MOS [1].

B. Performance on Individual and Combined Datasets

Table IV shows the performance evaluation of the three "completely blind" BVQA models, NIQE, ILNIQE, and VIIDEO on the four UGC-VQA benchmarks. None of these methods performed very well, meaning that we still have much room for developing OU "completely blind" UGC video quality models.

Table V shows the performance evaluation of all the learning-based BVQA models trained with SVR on the four datasets in our evaluation framework. For better visualization, we also show box plots of performances as well as scatter plots of predictions versus MOS on the All-Combined$_c$ set, in Figures 11 and 12, respectively. Overall, VIDEVAL achieves SOTA or near-SOTA performance on all the test sets. On LIVE-VQC, however, TLVQM outperformed other BVQA models by a notable margin, while it significantly underperformed on the more recent YouTube-UGC database. We observed in Section II-B that LIVE-VQC videos generally contain more (camera) motions than KoNViD-1k and YouTube-UGC, and TLVQM computes multiple motion relevant features. Moreover, the only three BVQA models containing temporal features (V-BLIINDS, TLVQM, and VIDEVAL) excelled on LIVE-VQC, which suggests that it is potentially valuable to integrate at least a few, if not many, motion-related features into quality prediction models, when assessing on videos with large (camera) motions.

It is also worth mentioning that the deep CNN baseline methods (VGG-19 and ResNet-50), despite being trained as picture-only models, performed quite well on KoNViD-1k and All-Combined$_c$. This suggests that transfer learning is a promising technique for the blind UGC-VQA problem, consistent with conclusions drawn for picture-quality prediction [63]. Deep models will perform even better, no doubt, if trained on temporal content and distortions.

The two most recent deep learning picture quality models, PaQ-2-PiQ, and KonCept512, however, did not perform very well on the three evaluated video datasets. The most probable reason would be that these models were trained on picture quality datasets [64], [70], which contain different types of

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

TABLE VI

PERFORMANCES ON DIFFERENT RESOLUTION SUBSETS: 1080P (427), 720P (566), AND ≤480P (448)

SUBSET	1080p		720p		≤480p	
MODEL	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.4597	0.4637	0.5407	0.5585	0.3812	0.4065
GM-LOG	0.4796	0.4970	0.5098	0.5172	0.3685	0.4200
HIGRADE	0.5142	0.5543	0.5095	0.5324	0.4650	0.4642
FRIQUEE	0.5787	0.5797	0.5369	0.5652	0.5042	0.5363
CORNIA	0.5951	**0.6358**	0.6212	0.6551	0.5631	0.6118
HOSA	0.5924	0.6093	**0.6651**	**0.6739**	**0.6514**	**0.6652**
VGG-19	**0.6440**	0.6090	0.6158	**0.6568**	**0.5845**	**0.6267**
ResNet-50	**0.6615**	**0.6644**	**0.6645**	**0.7076**	**0.6570**	**0.6997**
KonCept512	**0.6332**	**0.6336**	0.6055	0.6514	0.4271	0.4612
PaQ-2-PiQ	0.5304	0.5176	0.5768	0.5802	0.3646	0.4748
V-BLIINDS	0.4449	0.4491	0.5546	0.5719	0.4484	0.4752
TLVQM	0.5638	0.6031	**0.6300**	0.6526	0.4318	0.4784
VIDEVAL	0.5805	0.6111	0.6296	0.6393	0.5014	0.5508

TABLE VII

PERFORMANCES ON DIFFERENT CONTENT SUBSETS: SCREEN CONTENT (163), ANIMATION (81), AND GAMING (209)

SUBSET	Screen Content		Animation		Gaming	
MODEL	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.2573	0.3954	0.0747	0.3857	0.2717	0.3307
GM-LOG	0.3004	0.4244	0.2009	0.4129	0.3371	0.4185
HIGRADE	0.4971	0.5652	0.1985	0.4140	0.6228	0.6832
FRIQUEE	**0.5522**	**0.6160**	0.2377	0.4574	**0.6919**	**0.7193**
CORNIA	0.5105	0.5667	0.1936	0.4627	0.5741	0.6502
HOSA	0.4667	0.5255	0.1048	0.4489	0.6019	**0.6998**
VGG-19	0.5472	0.6229	0.1973	0.4700	0.5765	0.6370
ResNet-50	**0.6199**	**0.6676**	**0.2781**	**0.4871**	**0.6378**	0.6779
KonCept512	0.4714	0.5119	**0.2757**	**0.5229**	0.4780	0.6240
PaQ-2-PiQ	0.3231	0.4312	0.0208	0.4630	0.2169	0.3874
V-BLIINDS	0.3064	0.4155	0.0379	0.3917	0.5473	0.6101
TLVQM	0.3843	0.4524	0.2708	0.4598	0.5749	0.6195
VIDEVAL	**0.6033**	**0.6610**	**0.3492**	**0.5274**	**0.6954**	**0.7323**

(strictly spatial) distortions than UGC-VQA databases. Models trained on picture quality sets do not necessarily transfer very well to UGC video quality problems. In other words, whatever model should be either trained or fine-tuned on UGC-VQA datasets in order to obtain reasonable performance. Indeed, if temporal distortions (like judder) are present, they may severely underperform if the frame quality is high [81].

C. Performance Evaluation on Categorical Subsets

We propose three new categorical evaluation methodologies - resolution, quality, and content-based category breakdown. These will allow us to study the compared BVQA models from additional and practical aspects in the context of real-world UGC scenarios, which have not been, nor can it be accounted in previous legacy VQA databases or studies.

For resolution-dependent evaluation, we divided the All-Combined$_c$ set into three subsets, based on video resolution: (1) 427 1080p-videos (110 from LIVE-VQC, 317 from YouTube-UGC), (2) 566 720p-videos (316 from LIVE-VQC, 250 from YouTube-UGC), and (3) 448 videos with resolution ≤480p (29 from LIVE-VQC, 419 from YouTube-UGC), since

TABLE VIII

PERFORMANCES ON DIFFERENT QUALITY SUBSETS: LOW QUALITY (1558) AND HIGH QUALITY (1550)

SUBSET	Low Quality		High Quality	
MODEL	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.4312	0.4593	0.2813	0.2979
GM-LOG	0.4221	0.4715	0.2367	0.2621
HIGRADE	0.5057	0.5466	0.4714	0.4799
FRIQUEE	**0.5460**	**0.5886**	**0.5061**	**0.5152**
CORNIA	0.4931	0.5435	0.3610	0.3748
HOSA	**0.5348**	**0.5789**	0.4208	0.4323
VGG-19	0.3710	0.4181	0.3522	0.3614
ResNet-50	0.3881	0.4250	0.2791	0.3030
KonCept512	0.3428	0.4497	0.2245	0.2597
PaQ-2-PiQ	0.2438	0.2713	0.2013	0.2252
V-BLIINDS	0.4703	0.5060	0.3207	0.3444
TLVQM	0.4845	0.5386	**0.4783**	**0.4860**
VIDEVAL	**0.5680**	**0.6056**	**0.5546**	**0.5657**

TABLE IX

BEST MODEL IN TERMS OF SRCC FOR CROSS DATASET GENERALIZATION EVALUATION

TRAIN\TEST	LIVE-VQC	KoNViD-1k	YouTube-UGC$_c$
LIVE-VQC	-	ResNet-50 (0.69)	ResNet-50 (0.33)
KoNViD-1k	ResNet-50 (0.70)	-	VIDEVAL (0.37)
YouTube-UGC$_c$	HOSA (0.49)	VIDEVAL (0.61)	-

TABLE X

BEST MODEL IN TERMS OF PLCC FOR CROSS DATASET GENERALIZATION EVALUATION

TRAIN\TEST	LIVE-VQC	KoNViD-1k	YouTube-UGC$_c$
LIVE-VQC	-	ResNet-50 (0.70)	VIDEVAL (0.35)
KoNViD-1k	ResNet-50 (0.75)	-	VIDEVAL (0.39)
YouTube-UGC$_c$	HOSA (0.50)	VIDEVAL (0.62)	-

we are also interested in performance on videos of different resolutions. We did not include 540p-videos, since those videos are almost exclusively from KoNViD-1k. Table VI shows the resolution-breakdown evaluation results. Generally speaking, learned features (CORNIA, HOSA, VGG-19, KonCept512, and ResNet-50) outperformed hand-designed features, among which ResNet-50 ranked first.

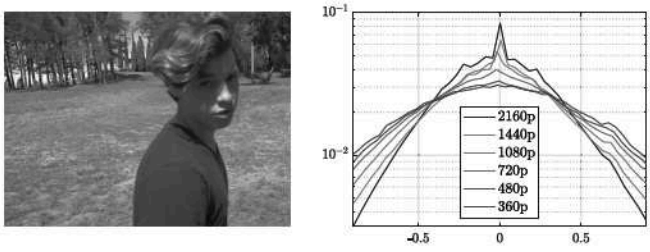
Here we make two arguments to try to explain the observations above: (1) video quality is intrinsically correlated with resolution; (2) NSS features are implicitly *resolution-aware*, while CNN features are not. The first point is almost self-explanatory, no matter to what degree one agrees. To further justify this, we trained an SVR only using resolution (height, width) as features to predict MOS on YouTube-UGC, which contains balanced samples across five different resolutions. This yielded surprisingly high values 0.576/0.571 for SRCC/PLCC, indicating the inherent correlation between video quality and resolution. Secondly, we selected one 2160p video from YouTube-UGC, namely 'Vlog2160P-408f.mkv,' and plotted, in Figure 13, the mean-subtracted contrast-normalized (MSCN) distributions of its downscaled versions: 2160p, 1440p, 1080p, 720p, 480p, and 360p. It may be observed that resolution can be well separated by MSCN statistics, based on which most feature-based

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

TABLE XI

PERFORMANCE COMPARISON OF A TOTAL OF ELEVEN TEMPORAL POOLING METHODS USING TLVQM AND VIDEVAL AS TESTBEDS ON KONVID-1K, LIVE-VQC, AND YOUTUBE-UGC. THE THREE BEST RESULTS ALONG EACH COLUMN ARE **BOLDFACED**

DATABASE	KoNViD-1k				LIVE-VQC				YouTube-UGC			
MODEL	TLVQM		VIDEVAL		TLVQM		VIDEVAL		TLVQM		VIDEVAL	
POOLING	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Mean	**0.7511**	**0.7475**	0.7749	**0.7727**	**0.7917**	**0.7984**	**0.7396**	0.7432	**0.6369**	**0.6310**	0.7447	0.7332
Median	0.7483	0.7437	0.7650	0.7698	0.7708	0.7887	0.7236	0.7308	0.6127	0.6090	0.7452	**0.7448**
Harmonic	0.7458	0.7392	**0.7772**	0.7681	0.7845	0.7890	0.7312	0.7250	0.6119	0.6038	0.7449	0.7318
Geometric	0.7449	0.7461	0.7566	0.7592	**0.7878**	**0.7964**	**0.7412**	0.7487	0.6347	0.6236	**0.7508**	**0.7437**
Minkowski	0.7498	**0.7481**	**0.7775**	**0.7727**	0.7863	0.7908	0.7371	**0.7558**	**0.6368**	**0.6311**	**0.7542**	**0.7508**
Percentile	0.7078	0.7000	0.7161	0.7049	0.7378	0.7313	0.6596	0.6576	0.4871	0.4996	0.6443	0.6465
VQPooling	0.7240	0.7196	0.7366	0.7296	0.7696	0.7895	0.7240	0.7311	0.5654	0.5618	0.6942	0.6862
Primacy	0.7456	0.7451	0.7711	0.7700	0.7751	0.7851	0.7349	**0.7523**	0.5734	0.5692	0.7221	0.7156
Recency	**0.7528**	0.7470	0.7683	0.7677	0.7715	0.7857	**0.7405**	**0.7584**	0.5821	0.5695	0.7176	0.7116
Hysteresis	0.7434	0.7430	0.7612	0.7554	0.7856	0.7901	0.7226	0.7433	0.6092	0.6109	0.7370	0.7306
EPooling	**0.7641**	**0.7573**	**0.7831**	**0.7867**	**0.7925**	**0.7917**	0.7371	0.7372	**0.6452**	**0.6592**	**0.7517**	0.7379



(a) Vlog_2160P-408f.mkv (b) MSCN distributions

Fig. 13. (a) An examplary 2160p video from YouTube-UGC and (b) the mean-subtracted contrast-normalized (MSCN) distributions of its downscaled versions: 2160p, 1440p, 1080p, 720p, 480p, and 360p.

methods are built. We may infer, from these two standpoints, that including various resolutions of videos is favorable to the training of NSS-based models, since NSS features are resolution-aware, and resolution is further well correlated with quality. In other words, the resolution-breakdown evaluation shown in Table VI, which removes this important implicit feature (resolution), would possibly reduce the performance of NSS-based models, such as FRIQUEE and VIDEVAL.

We also divided the All-Combined$_c$ into subsets based on content category: Screen Content (163), Animation (81), Gaming (209), and Natural (2,667) videos. We only reported the evaluation results on the first three subsets in Table VII, since we observed similar results on the Natural subset with the entire combined set. The proposed VIDEVAL model outperformed over all categories, followed by ResNet-50 and FRIQUEE, suggesting that VIDEVAL features are robust quality indicatives across different content categories.

The third categorical division is based on quality scores: we partitioned the combined set into Low Quality (1,558) and High Quality (1,550) halves, using the median quality value 3.5536 as the threshold, to see the model performance only on high/low quality videos. Performance results are shown in Table VIII, wherein VIDEVAL still outperformed the other BVQA models on both low and high quality partitions.

D. Cross Dataset Generalizability

We also performed a cross dataset evaluation to verify the generalizability of BVQA models, wherein LIVE-VQC, KoNViD-1k, and YouTube-UGC$_c$ were included. That is, we trained the regression model on one full database and report the performance on another. To retain label consistency, we linearly scaled the MOS values in LIVE-VQC from raw $[0, 100]$ to $[1, 5]$, which is the scale for the other two datasets. We used SVR for regression and adopted k-fold cross validation using the same grid-search as in Section V-A for hyperparameter selection. The selected parameter pair were then applied to re-train the SVR model on the full training set, and the performance results on the test set were recorded. Table IX and X show the best performing methods with cross domain performances in terms of SRCC and PLCC, respectively.

We may see that the cross domain BVQA algorithm generalization between LIVE-VQC and KoNViD-1k was surprisingly good, and was well characterized by pre-trained ResNet-50 features. We also observed better algorithm generalization between KoNViD-1k and YouTube-UGC than LIVE-VQC, as indicated by the performances of the best model, VIDEVAL. This might be expected, since as Figure 7 shows, YouTube-UGC and KoNViD-1k share overlapped coverage of content space, much larger than that of LIVE-VQC. Therefore, we may conclude that VIDEVAL and ResNet-50 were the most robust BVQA models among those compared in terms of cross domain generalization capacities.

E. Effects of Temporal Pooling

Temporal pooling is one of the most important, unresolved problems for video quality prediction [42], [60], [76], [82], [83]. In our previous work [76], we have studied the efficacy of various pooling methods using scores predicted by BIQA models. Here we extend this to evaluate on SOTA BVQA models. For practical considerations, the high-performing TLVQM and VIDEVAL were selected as exemplar models. Since these two models independently extract features on each one-second block, we applied temporal pooling of chunk-wise quality predictions. A total of eleven pooling methods were tested: three Pythagorean means (arithmetic, geometric, and harmonic mean), median, Minkowski ($p = 2$) mean, percentile pooling (20%) [84], VQPooling [82], primacy and recency pooling [85], hysteresis pooling [60], and our previously proposed

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

TABLE XII

FEATURE DESCRIPTION, DIMENSIONALITY, COMPUTATIONAL COMPLEXITY, AND AVERAGE RUNTIME COMPARISON (IN SECONDS EVALUATED ON TWENTY 1080p VIDEOS FROM LIVE-VQC) AMONG MATLAB-IMPLEMENTED BVQA MODELS

CLASS	MODEL	FEATURE DESCRIPTION	DIM	COMPUTATIONAL COMPLEXITY	TIME (SEC)
IQA	NIQE (1 fr/sec)	Spatial NSS	1	$\mathcal{O}(d^2NT)$ d: window size	6.3
	ILNIQE (1 fr/sec)	Spatial NSS, gradient, log-Gabor, and color statistics	1	$\mathcal{O}((d^2+h+gh)NT)$ d: window size; h: filter size; g: log-Gabor filter size	23.3
	BRISQUE (1 fr/sec)	Spatial NSS	36	$\mathcal{O}(d^2NT)$ d: window size	1.7
	GM-LOG (1 fr/sec)	Joint statistics of gradient magnitude and laplacian of gaussian coefficients	40	$\mathcal{O}(((h+k)NT)$ d: window size; k: probability matrix size	2.1
	HIGRADE (1 fr/sec)	Spatial NSS, and gradient magnitude statistics in LAB color space	216	$\mathcal{O}(3(2d^2+k)NT)$ d: window size; k: gradient kernel size	11.6
	FRIQUEE (1 fr/sec)	Complex streerable pyramid wavelet, luminance, chroma, LMS, HSI, yellow channel, and their transformed domain statistics	560	$\mathcal{O}((fd^2N+4N(\log(N)+m^2))T)$ d: window size; f: number of color spaces; m: neighborhood size in DNT	701.2
	CORNIA (1 fr/sec)	Spatially normalized image patches and max min pooling	10k	$\mathcal{O}(d^2KNT)$ d: window size K: codebook size	14.3
	HOSA (1 fr/sec)	Local normalized image patches based on high order statistics aggregation	14.7k	$\mathcal{O}(d^2KNT)$ d: window size K: codebook size	1.2
VQA	VIIDEO	Frame difference spatial statistics, inter sub-band statistics	1	$\mathcal{O}(N\log(N)T)$	674.8
	V-BLIINDS	Spatial NSS, frame difference DCT coefficient statistics, motion coherency, and egomotion	47	$\mathcal{O}((d^2N+\log(k)N+k^2w^3)T)$ d: window size; k: block size; w: motion vector tensor size	1989.9
	TLVQM	Captures impairments computed at two computation levels: low complexity and high complexity features	75	$\mathcal{O}((h_1^2N+k^2K)T_1+(\log(N)+h_2^2)NT_2))$ h_1,h_2: filter size; k: motion estimation block size; K: number of key points	183.8
	VIDEVAL	Selected combination of NSS features in multiple perceptual spaces and using visual impairment features from TLVQM	60	$\mathcal{O}((fh_1^2N+k^2K)T_1+h_2^2NT_2)$ h_1,h_2: filter size; f: number of color spaces; k: motion estimation block size; K: number of key points	305.8

N: number of pixels per frame; T: number of frames computed for feature extraction. Note that for VIIDEO and V-BLIINDS, T is the total number of frames, whereas for IQA models, T equals the total number of frames sampled at 1 fr/sec. For TLVQM and VIDEVAL, T_1 is total number of frames divided by 2, while T_2 is the number of frames sampled at 1 fr/sec.

ensemble method, EPooling [76], which aggregates multiply pooled scores by training a second regressor on top of mean, Minkowski, percentile, VQPooling, variation, and hysteresis pooling. We refer the reader to [76] for detailed algorithmic formulations and parameter settings thereof.

It is worth noting that the results in Table XI are only *self-consistent*, meaning that they are not comparable to any prior experiments - since we employed chunk-wise instead of previously adopted video-wise quality prediction to be able to apply temporal quality pooling, which may affect the base performance. Here we observed yet slightly different results using BVQA testbeds as compared to what we observed on BIQA [76]. Generally, we found the mean families and ensemble pooling to be the most reliable pooling methods. Traditional sample mean prediction may be adequate in many cases, due to its simplicity. Pooling strategies that more heavily weight low-quality parts, however, were not observed to perform very well on the tested BVQA, which might be attributed to the fact that not enough samples ($8 \sim 20$) can be extracted from each video to attain statistically meaningful results.

F. Complexity Analysis and Runtime Comparison

The efficiency of a video quality model is of vital importance in practical commercial deployments. Therefore, we also tabulated the computational complexity and runtime cost of the compared BVQA models, as shown in Tables XII, XIII. The experiments were performed in MATLAB R2018b and Python 3.6.7 under Ubuntu 18.04.3 LTS system on a Dell OptiPlex 7080 Desktop with Intel Core i7-8700 CPU@3.2GHz, 32G RAM, and GeForce GTX 1050 Graphics Cards. The average feature computation time of MATLAB-implemented BVQA models on 1080p videos are reported in Table XII. The proposed VIDEVAL method achieves a reasonable complexity among the top-performing algorithms, TLVQM, and FRIQUEE. We also present theoretical time complexity in Table XII for potential analytical purposes.

We also provide in Table XIII an additional runtime comparison between MATLAB models on CPU and deep learning models on CPU and GPU, respectively. It may be observed that top-performing BVQA models such as TLVQM and VIDEVAL are essentially slower than deep CNN models, but we expect orders-of-magnitude speedup if re-implemented in pure C/C++. Simpler NSS-based models such as BRISQUE and HIGRADE (which only involve several convolution operations) still show competitive efficiency relative to CNN models even when implemented in MATLAB. We have also seen a $5 \sim 10$ times speedup switching from CPU to GPU for the CNN models, among which KonCept512 with PyTorch-GPU was the fastest since it requires just a single pass to the CNN backbone, while the other three entail multiple passes for each input frame.

Note that the training/test time of the machine learning regressor is approximately proportional to the number of features. Thus, it is not negligible compared to feature computation given a large number of features, regardless of the regression model employed. The feature dimension of each model is listed in Table XII. As may be seen, codebook-based algorithms (CORNIA ($10k$) and HOSA ($14.7k$)) require

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

TABLE XIII

RUN TIME COMPARISON OF FEATURE-BASED AND DEEP LEARNING BVQA MODELS (IN SECONDS EVALUATED ON TWENTY 1080*p* VIDEOS FROM LIVE-VQC). MODEL LOADING TIME FOR DEEP MODELS ARE EXCLUDED

MODEL		TIME (SEC)
BRISQUE (1 fr/sec)	MATLAB-CPU	1.7
HOSA (1 fr/sec)	MATLAB-CPU	1.2
TLVQM	MATLAB-CPU	183.8
VIDEVAL	MATLAB-CPU	305.8
VGG-19 (1 fr/sec)	TensorFlow-CPU	27.8
	TensorFlow-*GPU*	5.7
ResNet-50 (1 fr/sec)	TensorFlow-CPU	9.6
	TensorFlow-*GPU*	1.9
KonCept512 (1 fr/sec)	PyTorch-CPU	2.8
	PyTorch-*GPU*	0.3
PaQ-2-PiQ (1 fr/sec)	PyTorch-CPU	6.9
	PyTorch-*GPU*	0.8

significantly larger numbers of features than other hand-crafted feature based models. Deep ConvNet features ranked second in dimension (VGG-19 (4,080) and ResNet-50 (2,048)). Our proposed VIDEVAL only uses 60 features, which is fairly compact, as compared to other top-performing BVQA models like FRIQUEE (560) and TLVQM (75).

G. Ensembling VIDEVAL With Deep Features

We also attempted a more sophisticated ensemble fusion of VIDEVAL and deep learning features to determine whether this could further boost its performance, which could give insights on the future direction of this field. Since PaQ-2-PiQ aimed for local quality prediction, we included the predicted 3×5 local quality scores as well as a single global score, as additional features. For KonCept512, the feature vector (256-dim) immediately before the last linear layer in the fully-connected head was appended. Our own baseline CNN models, VGG-19 and ResNet-50, were also considered, because these are commonly used standards for downstream vision tasks.

The overall results are summarized in Table XIV. We may observe that ensembling VIDEVAL with certain deep learning models improved the performance by up to $\sim 4\%$ compared to the vanilla VIDEVAL, which is very promising. Fusion with either ResNet-50 or KonCept512 yielded top performance. It should be noted that the number of fused features is also an essential aspect. For example, blending VIDEVAL (60-dim) with VGG-19 (4,096-dim) may not be recommended, since the enormous number of VGG-19 features could possibly dominate the VIDEVAL features, as suggested by some performance drops in Table XIV.

H. Summarization and Takeaways

Finally, we briefly summarize the experimental results and make additional observations:

1) Generally, spatial distortions dominated quality prediction on Internet UGC videos like those from YouTube and Flickr, as revealed by the remarkable performances of picture-only models (e.g., HIGRADE, FRIQUEE, HOSA, ResNet-50) on them. Some motion-related features (as in TLVQM) may not apply as well in this scenario.
2) On videos captured with mobile devices (e.g., those in LIVE-VQC), which often present larger and more frequent camera motions, including temporal- or motion-related features can be advantageous (e.g., V-BLIINDS, TLVQM, VIDEVAL).
3) Deep CNN feature descriptors (VGG-19, ResNet-50, etc.) pre-trained for other classical vision tasks (e.g. image classification) are transferable to UGC video quality predictions, achieving very good performance, suggesting that using transfer learning to address the general UGC-VQA problem is very promising.
4) It is still a very hard problem to predict UGC video quality on non-natural or computer-generated video contents: screen contents, animations, gaming, etc. Moreover, there are no sufficiently large UGC-VQA datasets designed for those kinds of contents.
5) A simple feature engineering and selection implementation built on top of current effective feature-based BVQA models is able to obtain excellent performance, as exemplified by the compact new model (VIDEVAL).
6) Simple temporal mean pooling of chunk-wise quality predictions by BVQA models yields decent and robust results. Furthermore, an ensemble pooling approach can noticeably improve the quality prediction performance, albeit with higher complexity.
7) Ensembling scene statistics-based BVQA models with additional deep learning features (e.g., VIDEVAL plus KonCept512) could further raise the performance upper bound, which may be a promising way of developing future BVQA models.

TABLE XIV

PERFORMANCE OF THE ENSEMBLE VIDEVAL MODELS FUSED WITH ADDITIONAL DEEP LEARNING FEATURES

DATASET	MODEL \ METRIC	SRCC	KRCC	PLCC	RMSE
KoNViD	VIDEVAL	0.7832	0.5845	0.7803	0.4024
	VIDEVAL+VGG-19	0.7827	0.5928	0.7913	0.3897
	VIDEVAL+ResNet-50	0.8129	0.6212	**0.8200**	**0.3659**
	VIDEVAL+KonCept512	**0.8149**	**0.6251**	0.8169	0.3670
	VIDEVAL+PaQ-2-PiQ	0.7844	0.5891	0.7793	0.4018
LIVE-VQC	VIDEVAL	0.7522	0.5639	0.7514	11.100
	VIDEVAL+VGG-19	0.7274	0.5375	0.7717	10.749
	VIDEVAL+ResNet-50	0.7456	0.5555	0.7810	10.385
	VIDEVAL+KonCept512	**0.7849**	**0.5953**	**0.8010**	**10.145**
	VIDEVAL+PaQ-2-PiQ	0.7677	0.5736	0.7686	10.787
YT-UGC	VIDEVAL	0.7787	0.5830	0.7733	0.4049
	VIDEVAL+VGG-19	0.7868	0.5930	0.7847	0.3993
	VIDEVAL+ResNet-50	**0.8085**	0.6128	**0.8033**	**0.3837**
	VIDEVAL+KonCept512	0.8083	**0.6139**	0.8028	0.3859
	VIDEVAL+PaQ-2-PiQ	0.7981	0.6015	0.7941	0.3959
All-Comb	VIDEVAL	0.7960	0.6032	0.7939	0.4268
	VIDEVAL+VGG-19	0.7859	0.5912	0.7962	0.4202
	VIDEVAL+ResNet-50	0.8115	**0.6207**	**0.8286**	**0.3871**
	VIDEVAL+KonCept512	**0.8123**	0.6193	0.8168	0.4017
	VIDEVAL+PaQ-2-PiQ	0.7962	0.5991	0.7934	0.4229

VI. CONCLUSION

We have presented a comprehensive analysis and empirical study of blind video quality assessment for user-generated

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

content (the **UGC-VQA problem**). We also proposed a new fusion-based BVQA model, called the VIDeo quality EVALuator (VIDEVAL), which uses a feature ensemble and selection procedure on top of existing efficient BVQA models. A systematic evaluation of prior leading video quality models was conducted within a unified and reproducible evaluation framework and accordingly, we concluded that a selected fusion of simple distortion-aware statistical video features, along with well-defined visual impairment features, is able to deliver state-of-the-art, robust performance at a very reasonable computational cost. The promising performances of baseline CNN models suggest the great potential of leveraging transfer learning techniques for the UGC-VQA problem. We believe that this benchmarking study will help facilitate UGC-VQA research by clarifying the current status of BVQA research and the relative efficacies of modern BVQA models. To promote reproducible research and public usage, an implementation of VIDEVAL has been made available online: **https://github.com/vztu/VIDEVAL**. In addition to the software, we are also maintaining an ongoing performance leaderboard on Github: **https://github.com/vztu/BVQA_Benchmark**.

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] F. D. Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 2430–2433.

[3] (2010). *VQEG HDTV Phase I Database*. Accessed: Nov. 9, 2019. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

[4] F. Zhang, S. Li, L. Ma, W. Y. Chung, and K. N. Ngan. (2011). *IVP Subjective Quality Video Database*. Accessed: Nov. 9, 2019. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/

[5] C. Keimel, A. Redl, and K. Diepold, "The TUM high definition video datasets," in *Proc. 4th Int. Workshop Qual. Multimedia Exper.*, Jul. 2012, pp. 97–102.

[6] 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, Feb. 2014, Art. no. 013016.

[7] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C.-J. Kuo, "MCL-V: A streaming video quality assessment database," *J. Vis. Commun. Image Represent.*, vol. 30, pp. 1–9, Jul. 2015.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] B. Thomee *et al.*, "YFCC100M: The new data in multimedia research," 2015, *arXiv:1503.01817*. [Online]. Available: http://arxiv.org/abs/1503.01817

[16] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[17] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Proc. SPIE*, vol. 5007, pp. 87–96, Jun. 2003.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[19] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

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

[21] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. (2016). *Toward a Practical Perceptual Video Quality Metric*. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652

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

[23] L.-H. Chen, C. G. Bampis, Z. Li, J. Sole, and A. C. Bovik, "Perceptual video quality prediction emphasizing chroma distortions," 2020, *arXiv:2009.11203*. [Online]. Available: http://arxiv.org/abs/2009.11203

[24] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Proc. SPIE*, vol. 5666, pp. 149–159, Mar. 2005.

[25] 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.

[26] 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.

[27] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proc. Int. Conf. Image Process.*, vol. 3, Sep. 2000, pp. 981–984.

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

[29] X. Feng and J. P. Allebach, "Measurement of ringing artifacts in JPEG images," *Proc. SPIE*, vol. 6076, Feb. 2006, Art. no. 60760A.

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

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

[32] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik, "Adaptive debanding filter," *IEEE Signal Process. Lett.*, vol. 27, pp. 1715–1719, 2020.

[33] 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.

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

[35] J. E. Caviedes and F. Oberti, "No-reference quality metric for degraded and enhanced video," in *Proc. Digit. Video Image Qual. Perceptual Coding*, 2017, pp. 305–324.

[36] C. Keimel, T. Oelbaum, and K. Diepold, "No-reference video quality evaluation for high-definition video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 1145–1148.

[37] 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.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.

[38] 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.

[39] 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.

[40] 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.

[41] 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, p. 32, Jan. 2017.

[42] 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.

[43] 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.*, Jun. 2012, pp. 1098–1105.

[44] 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.

[45] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," 2021, *arXiv:2101.10955*. [Online]. Available: http://arxiv.org/abs/2101.10955

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

[47] 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.

[48] 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.

[49] 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, Aug. 2014.

[50] 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.

[51] 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.

[52] L. Liu, H. Dong, H. Huang, and A. C. Bovik, "No-reference image quality assessment in curvelet domain," *Signal Process., Image Commun.*, vol. 29, no. 4, pp. 494–505, Apr. 2014.

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

[54] 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.

[55] 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.

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

[57] 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.

[58] 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, Dec. 2013, Art. no. 043025.

[59] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 491–495.

[60] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 1153–1156.

[61] 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.

[62] 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.*, Jun. 2014, pp. 1733–1740.

[63] 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.

[64] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3575–3585.

[65] L.-H. Chen, C. G. Bampis, Z. Li, A. Norkin, and A. C. Bovik, "ProxIQA: A proxy approach to perceptual optimization of learned image compression," *IEEE Trans. Image Process.*, vol. 30, pp. 360–373, 2021.

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

[67] 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.*, Jun. 2009, pp. 248–255.

[68] 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.

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

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

[71] 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. (ECCV)*, 2018, pp. 219–234.

[72] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 546–554.

[73] 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.

[74] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2351–2359.

[75] 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.

[76] 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. (ICIP)*, Oct. 2020, pp. 141–145.

[77] M. H. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," in *Proc. Vis. Commun. Image Process.*, vol. 5150, Jun. 2003, pp. 583–592.

[78] 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.

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

[80] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[81] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," 2020, *arXiv:2007.11634*. [Online]. Available: http://arxiv.org/abs/2007.11634

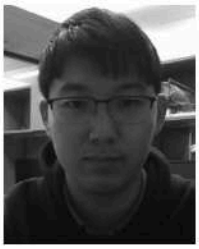
[82] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, Feb. 2013.

[83] C. G. Bampis, Z. Li, I. Katsavounidis, and A. C. Bovik, "Recurrent and dynamic models for predicting streaming video quality of experience," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3316–3331, Jul. 2018.

[84] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[85] B. B. Murdock, "The serial position effect of free recall," *J. Exp. Psychol.*, vol. 64, no. 5, p. 482, 1962.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.



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 pursuing the Ph.D. degree with the Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin. His research interests include perceptual image and video quality assessment, computer vision, and machine learning.

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

Neil Birkbeck received the Ph.D. degree from the University of Alberta 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 at Siemens Corporate Research, working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Software Engineer with the 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 in 2002, and the Ph.D. degree from the University of California at Santa Barbara in 2005, on watermark-based error resilience in video communications. From 2005 to 2011, he was a Senior Research Scientist at Citrix Online, and he was a Senior Manager Advanced Technology at GoPro from 2011 to 2016, developing algorithms for images/video quality enhancement, compression, capture, and streaming. He currently manages and leads the Media Algorithms group at YouTube/Google. He has coauthored more than 120 articles and patents. His fields of research interests include image/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. His research interests include image processing, digital photography, digital television, digital streaming video, and visual perception. His books include *The Essential Guides to Image and Video Processing*. For his work in these areas, he was a 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 has also received over ten best journal paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. He co-founded and was the longest-serving Editor-in-Chief for 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, in 1994.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 19,2021 at 00:51:58 UTC from IEEE Xplore. Restrictions apply.