Study of the Subjective and Objective Quality of High Motion Live Streaming Videos

Zaixi Shang, *Student Member, IEEE*, Joshua Peter Ebenezer, Yongjun Wu, *Senior Member, IEEE*, Hai Wei, *Member, IEEE*, Sriram Sethuraman, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*— Video livestreaming is gaining prevalence among video streaming services, especially for the delivery of live, high motion content such as sporting events. The quality of these livestreaming videos can be adversely affected by any of a wide variety of events, including capture artifacts, and distortions incurred during coding and transmission. High motion content can cause or exacerbate many kinds of distortion, such as motion blur and stutter. Because of this, the development of objective Video Quality Assessment (VQA) algorithms that can predict the perceptual quality of high motion, live streamed videos is greatly desired. Important resources for developing these algorithms are appropriate databases that exemplify the kinds of live streaming video distortions encountered in practice. Towards making progress in this direction, we built a video quality database specifically designed for live streaming VQA research. The new video database is called the Laboratory for Image and Video Engineering (LIVE) Livestream Database. The LIVE Livestream Database includes 315 videos of 45 source sequences from 33 original contents impaired by 6 types of distortions. We also performed a subjective quality study using the new database, whereby more than 12,000 human opinions were gathered from 40 subjects. We demonstrate the usefulness of the new resource by performing a holistic evaluation of the performance of current state-of-the-art (SOTA) VQA models. We envision that researchers will find the dataset to be useful for the development, testing, and comparison of future VQA models. The LIVE Livestream database is being made publicly available for these purposes at https://live.ece.utexas.edu/research/LIVE_APV_Study/apv_index.html**

***Index Terms*— Live Streaming, video quality assessment, video quality database, objective VQA algorithm evaluation.**

I. Introduction

VIDEO traffic now occupies more than 70% of all total downstream Internet traffic and is still expected to grow [1], [2]. Major content providers such as Amazon Prime Video, YouTube, Netflix, and Hulu are providing increasing amounts of video on demand (VoD) content, as well as live streaming videos, to an expanding audience. live streaming, which is real-time audio and video transmission of live events, is gaining popularity very rapidly, especially for sporting events like the Super Bowl [3].

Although significant efforts have been made to enable the delivery of high-quality, high-resolution VoD, little effort has focused on live, high motion video streaming. In live streaming, there are still a variety of factors that can adversely affect the quality of live streaming videos. For example, bandwidth and stability may affect the received video source quality, causing distortion like blocking, banding, deinterlacing motion mismatches, aliasing and interpolation artifacts [4].

A discussion of prevalent transmission protocols is pertinent here. Both Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are extensively used for live video streaming. However, many live streaming companies are pursuing UDP for high-motion applications. End-to-end delays when using UDP is substantially less than TCP, which is critical for live sports events. UDP, on the other hand, depends on basic mechanisms and checksums to ensure data integrity with no handshakes, no delivery guarantees, and no duplicate protections. Thus, some video distortions, such as frame drops and flicker, are more likely to occur when using UDP. This helped motivate the design of the current database. The videos may also be distorted by stutter or motion blur, especially when there is rapid motion. By contrast with VoD streaming, a large portion of live streamed content is still interlaced and then deinterlaced, causing combing effects, flicker or noticeable line movements.

Video impairments like these can severely impair the delivered video quality and users' holistic levels of visual satisfaction. This is a pressing problem for high motion, action content such as sports videos. high motion videos generally contain richer temporal information and are harder to compress, hence compression artifacts are often more severe in sports videos. Other distortions can also be exacerbated by high motion. For example, at lower frame rates, high motion sports may appear discontinuous over time, and may exhibit obvious judder. Likewise, high motion can worsen the visual appearance of interlacing, by causing jagged moving edges.

While it is highly desirable to create algorithms that can successfully predict the visual impacts of these distortions, subjective data is necessary to understand these perceptual phenomenon and to design and test the underlying objective models. Human subjective studies make it possible to better

Manuscript received November 3, 2020; revised July 13, 2021 and November 9, 2021; accepted November 26, 2021. Date of publication December 24, 2021; date of current version January 12, 2022. This work was supported in part by Amazon Prime Video and in part by the National Science Foundation AI Institute for Foundations of Machine Learning (IFML) under Grant 2019844. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Giuseppe Valenzise. *(Zaixi Shang and Joshua Peter Ebenezer contributed equally to this work.) (Corresponding author: Alan C. Bovik.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB), University of Texas, Austin, under FWA No. 00002030 and Protocol No. 2007-11-0066.

Zaixi Shang, Joshua Peter Ebenezer, and Alan C. Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: bovik@ece.utexas.edu).

Yongjun Wu, Hai Wei, and Sriram Sethuraman are with Amazon Prime Video, Amazon.com Inc., Seattle, WA 98108 USA.

Digital Object Identifier 10.1109/TIP.2021.3136723

understand and model the specific factors that contribute to the perceived quality of streaming videos. This data can be used to design or learn objective Video Quality Assessment (VQA) models that are consistent with subjective human evaluations of quality. The development of subjective video quality assessment datasets has been an ongoing effort for two decades [5]–[14], yet none of these are specific to live streaming distortions. Among existing datasets, most include fewer than 20 pristine source video contents of Standard Definition (SD) or High Definition (HD) resolutions, along with various distorted versions of them. The distortions in these resources are largely limited to compression and aliasing, and the datasets lack other live streaming distortions. What is needed is a database of higher resolution (UHD), high-quality source videos that have been processed to include distortions characteristic of those encountered in live streaming scenarios.

Towards filling this gap, we have created a new resource that we call the LIVE Livestream Database, which includes a large number of high motion sports videos, impaired by the most common distortions that impact the perceptual quality of live streamed videos. The new database contains 315 videos, built from 45 source sequences from 33 original contents impaired by six types of common processing distortions. Unlike prior, legacy VQA databases, the LIVE Livestream database consists of Full High Definition (FHD) and Ultra High Definition (UHD) videos of high motion sports content captured by professional videographers. Using these videos, we conducted a large human subjective study, whereby we presented the videos to a large pool of volunteers to obtain Mean Opinion Scores (MOS). To demonstrate the usefulness of the new dataset, we used it to perform a holistic evaluation of current state-of-the-art VQA models, to compare their performance and to gain insights into potential future live streaming VQA problems.

The rest of the paper is organized as follows: In Section II we introduce prior work related to our study and in Section III we discuss the relevance and novelty of the work. In Section IV, we explain the details of the construction of new database and the protocol of the human study. Section V elaborates on the processing and analysis of the obtained subjective scores. Section VI compares the performances of various state of the art (SOTA) VQA models on the new database. Finally, Section VIII concludes the paper with thoughts regarding future efforts.

II. Related Work

Over the past decade, there have been many efforts to build subjective video quality databases. Among those, the LIVE VQA Database [5] includes 10 pristine videos processed with compression and packet loss distortions. Similarly, the later database in [15] contains 156 videos modified by H.264 compression artifacts and wireless packet losses. The LIVE QoE Database for HTTP-based Video Streaming [12] studies the quality of experience (QoE) of users who viewed compressed videos with simulated video stalls, which can arise when there is low channel throughput. This database models the perception of video quality on mobile devices, and the human study was performed on mobile phones and tablets. Another QoE database proposed in [20] aims to motivate QoE prediction in video streaming, with different bitrate levels and stalling events. Among 20 1080p source sequences, 5 videos contain high motion content. Another database [25] studied H.264 compressed videos transferred through an error-prone network, including 156 sequences at CIF and 4CIF spatial resolutions. The LIVE Mobile Video Quality Database [17] consists of 200 distorted videos created from 10 RAW HD reference videos, including compression and wireless packet-losses, with dynamically varying distortions. The MCL-V database [26] was designed for streaming video quality assessment, and contains 12 source video clips and 96 distorted video clips impaired by H.264 compression, as well as compression followed by spatial scaling. The TUM databases [27], [28], contain several synthesized videos with H.264 compression. Other exemplars include the MCL video quality database [29], ECVQ and EVVQ [30], and the Poly@NYU Video Quality Databases [31], [32].

More recently, novel databases have been introduced that contain user-generated-content (UGC) videos with authentic distortions. The LIVE-VQC database [7] contains 585 videos, all of unique contents captured by a large group of users deploying various camera devices, including smartphones of all brands. The LIVE-VQC videos cover a wide range of qualities, and include complex, often commingled authentic distortions. The large KoNViD-1k [33] video quality database contains 1,200 video sequences, covering a wide variety of contents and authentic distortions. The YouTube UGC Dataset [34] contains 1500 20-second video clips covering popular UGC video categories, including gaming and sports.

A number of deficiencies limit the usefulness of all of these databases for the study of the quality of live video streams. Older, legacy databases contain only limited numbers of SD source contents, which are not representative of current high-resolution live streaming. Although most databases consider compression distortions and packet loss, other prevalent distortions common to live streaming videos are rarely found in them. Given exploding interest in live streaming video, a comprehensive database that includes both ample video content and representative live streaming distortions is needed.

UGC video quality databases usually include a large number of contents, but there is a lack of professionally captured content, and the distortions encountered in live streaming often significantly differ from those caused by typical casual social media users. The only existing publicly available VQA database designed for live streaming is the LIMP Video Quality Database [35]. The LIMP database consists of nine high-quality videos taken from the LIVE Video Quality Video Database [5], with simulated compression modeling transmitted in a controlled network. However, it suffers from the same problems mentioned above. Motivated by an apparent dearth of live streaming databases containing enough high-resolution video contents and sufficiently representative live streaming distortions, we have created a large new resource intended to address modern aspects of the live sports streaming video quality problem.

TABLE I
A SUMMARY OF PUBLISHED VQA DATABASES AND THE NEW DATABASE

Database	Year	# SRC	# DST	# total	resolution	distortions
EPFL-PoliMI [15]	2009	12	144	156	704×576	compression
LIVE VQA [5]	2010	10	150	160	768×432	compression, transmission error
IVP [16]	2011	10	128	138	1920×1088	compression, transmission error
LIVE-Mobile [17]	2012	10	200	210	1280×720	compression, transmission error, rate adaptation, frame-freeze
LIVE-Flicker [18]	2015	6	66	72	1280×720	flicker
CVD2014 [19]	2014	-	234	234	up to 1920×1080	authentic
LIVE QoE [12]	2014	8	18	18	1280×720	rate-adaptation
LIVE-Netflix [9]	2017	14	112	112	1920×1080	network fluctuations
Streaming QoE [20]	2017	20	200	220	1920×1080	compression, stalling
SMD Panoramic [21]	2018	10	50	60	-	compression
BVI-HD [22]	2018	32	384	416	1920×1080	compression
CSCVQ [23]	2020	11	165	176	1280×720	compression
LIVE-SJTU [24]	2020	14	336	336	1920×1080	video compression, scaling, audio compression
LIVE Livestreaming	2021	45	270	315	1920×1080, 3840×2160	compression, aliasing, judder, flicker, frame drop, interlacing

III. RELEVANCE AND NOVELTY

In recent years, the streaming of live high motion video content such as sports has exploded [1]. Live streaming high motion videos often suffer from severe distortions less often encountered in the streaming of generic content. In live streaming, considerations of network instability and bandwidth limitations imply greater challenges when attempting to control video quality. Moreover, the real-time requirement greatly limits the time available for post-processing to compensate for defects. The unique nature of live streaming introduces many obstacles that differ from those encountered in generic on-demand video streaming. For example, sports videos usually include content containing complex, large motions. Rapid and irregular camera motions occurs frequently, when tracking moving objects, such as balls or players. Temporal distortions often arise that are annoying and that adversely affect the viewer experiences.

The new psychometric database that we describe here has a number of unique attributes. It contains a larger number of unique source contents and distortion types. We summarize the attributes of public video quality databases in Table I. The new database includes 45 source sequences token from 33 unique contents. All of the videos contain complex, fast motions, which are rarely included in existing databases. The new resource contains a wide variety of distortion classes common to sports live streaming content that is not found in existing VQA databases. Although LIVE-Flicker and Live-Mobile include specific temporal distortions such as flicker or frame-freeze, neither contains a holistic collection of high motion, live streaming distortion types.

IV. DETAILS OF SUBJECTIVE STUDY

We constructed a new video quality database that consistent of 315 video sequences including 45 reference videos and 6 copies of synthetically distorted version of each reference video. Those videos are used as stimuli in the subjective study.

A. Source Sequences

We collected 33 uncompressed, high-quality source videos with sports content. These videos are freely available online from multiple sources, including from Tampere University [36], the MCML Group [37], the Netflix Public Dataset [38], the VQEG HD3 Dataset [39], the Consumer Digital Video Library (CDVL) [40], and the SJTU Media Lab [41]. All of the selected videos were captured with professional, high-end camera equipment and are distortion-free. The original pristine videos all have resolutions of 1920×1080 or 3840×2160 pixels, and were progressively scanned in YUV 4:2:0 format with audio components removed. The videos have frame rates at 30 fps. The video contents include 10 different types of sports, including running, football, and soccer, and one video of the audience in a stadium, as exemplified in Fig. 1.

The original 33 videos that we collected are of durations ranging from 5s to 26s. However, since viewing videos of such differences of durations could cause biases in subjective and objective judgments, longer videos may exhibit visible changes of distortion over time. While the effects of video duration is interesting and worthy of study, this also would increase the dimensionality of the study. Thus, we manually cropped the longer videos along the temporal dimension into one or two shorter clips of about 7 seconds with no overlap or close proximity between the clips. Based on internal studies at UT-LIVE, it has been observed that very short videos of sports videos may cause annoying content disruptions, such as incomplete "play," but these events are usually shorter than 8s. To avoid unpleasant cuts during action scenes, we allowed some flexibility of the video durations, hence the final set

Fig. 1. Exemplar screenshots of frames from source videos in the LIVE Livestream Database.

of original videos had lengths in the range 5s-8s, averaging 7.88s with a standard deviation of 1.36s. In this way, 45 video clips were created from the 33 originals, of which 22 clips are of resolution 1920×1080 and 23 clips are of resolution 3840×2160.

B. Synthetic Distortions

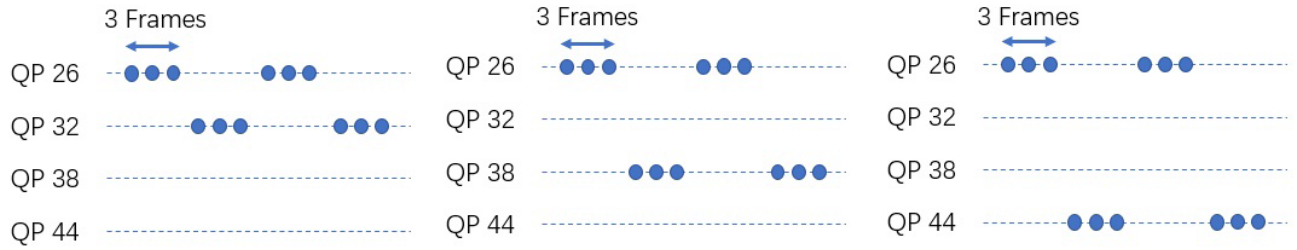
We created 6 distorted video sequences from each of the pristine sequences, using six different distortion processes. These included H.264 compression, aliasing, judder, flicker, frame drops, and interlacing. Since our primary goal is to model the visual quality high motion live sports videos, the distortions chosen were judged to be the most common and salient ones that are encountered during live sports events. During live streaming of high motion contents, certain distortions may produce more severe effects than on more generic video content. For example, a moving object may cause large pixel offsets between neighboring frames or fields. If the video is interlaced, then severe edge combing and blur may occur. If the frame rate is too slow, then judder from 3:2 conversion [42], [43] may be visible in high motion regions, which can seriously and adversely impact the appearances of sports videos. Purely temporal distortions, such as frame drops, which cause discontinuities and motion stalls, are difficult to detect.

When applying different levels of each distortion type, we sought to ensure that the distorted videos would be both perceptually separable and also cover a wide range of perceptual qualities, following successful practice in numerous previous studies [5]–[7]. However, given the large number of source sequences, it is not practical to include multiple copies of the same content, which can greatly increase the duration of the human study. Moreover, having larger number of unique contents can contribute to improved model building. Hence, given the fairly large number of source videos, we dictated that each would only have a single level of severity of each distortion type applied to it. For example, four levels of H.264 compression, corresponding to different constant rate factors (CRF) were defined. This was accomplished in a "round robin" sequential manner: the first reference video could only be compressed using the first CRF level, the second reference was only compressed using the second CRF level, and so on. The fifth source video then had the first level of distortion applied. However, to ensure that there would be no content-related quality bias, the first video in the quality level cycle was also sequenced as subsequent distortions were applied. In this way, each of the 45 clips taken from the original 33 pristine source videos has 6 associated distorted versions of it, yielding 315 videos including the 45 reference videos.

1) H.264 Compression: H.264 remains the most widely-accepted and used video compression standard. A 2020 streaming industry survey [44] found that 91% of streaming services use H.264. Although newer codecs exist, such as HEVC, VP9 and AV1, they are not yet as widely adopted. Browsers and devices also don't have full support for all codecs. The Apple Safari browser supports HEVC, but not VP9, while Chrome and Firefox support VP9 and AV1, but not HEVC. All browsers support H.264. Hence, when designing this VQA database, we deemed H.264 to be most representative of current practice. Moreover, even emerging standards still follow the basic hybrid codec method of distortion, viz., quantization of DCT blocks, while several distortions are not compression-related. Hence, we believe that the new database will retain usefulness as the compression standards evolve. We fixed four levels of H.264 compression using the criteria described earlier, by varying the CRF values. Similar to other successful VQA databases [5], [15], [26], we included a wide range of compression CRFs to ensure that the distorted videos cover a wide range of perceptual qualities, while also ensuring perceptual difference between the applied compression levels, to allow for improved model-building. Since in practice, the compression parameters differ on videos of different resolution, we selected different sets of CRFs for the 4K

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

Fig. 2. Simulation of motion judder from 3:2 pulldown. (a) Original frames at 23.94 fps. (b) Odd video fields. (c) Even video fields (d) Resulting frames at 29.97 fps.



(a) Flicker level 1 (b) Flicker level 2 (c) Flicker level 3

Fig. 3. Three levels of flicker synthesis.

videos and the 1080p videos. The CRF values selected for the 4K videos were 9, 27, 39, and 43, while those for 1080p videos were 9, 25, 35, and 39. All of the compressed videos were generated using FFmpeg.

2) Aliasing: Aliasing was simulated by first downscaling each video, then upscaling it back to its original dimensions. The downscaling was performed by spatially downsampling the video to half the original size without the use of an anti-aliasing filter, while the upscaling was performed using a Lanczos filter.

3) Judder: Motion judder is an artifact that is introduced when scenes shot at 23.94 fps are converted to 29.97 fps by a process called 2:3 pulldown. The ratio of these frame rates is 4:5: for every 4 input frames, 5 output frames were created by temporally downsampling the video to 23.94 fps, then converting the frame rate to 29.97 by 2:3 pulldown. The odd video field of every 2nd frame, and the even video field of every 3rd frame of each group of 4 frames were combined to form an additional frame, for each group of 4 frames. This process is shown in Fig. 2. Classic 2:3 pulldown followed a slightly different pattern where the 2nd and 3rd frames of the original video would be interlaced to form the 3rd frame of the juddered video, and the 4th and 5th frames of the original video would be interlaced to form the 4th frame of the juddered video. This had the disadvantage of producing two "dirty" frames, which were the 3rd and 4th frames in each group, but was used in legacy systems where the buffer could not hold fields from more than one frame at a time. The version we use here is a more advanced pulldown, supported by cameras released after 2000 such as the Panasonic DVX100 [45] or the Canon XL2 [46]. The more advanced version of pulldown generates only one "dirty" frame and also allows for better compression and easier conversion back to 23.94 fps.

4) Flicker: We simulated flicker distortion from compression by alternating the H.264 quantization parameter (QP) on the video. The QP is fixed at a constant value by passing this parameter to libx264. These QP values were applied to each frame, regardless of the frame type, content and motion. Three pairs of QPs were chosen to form three flicker distortion levels: QP26 and QP32, QP26 and QP 38, and QP26 and QP44. The flicker rate, which is the number of QP alternations per second, was kept a constant roughly 5 Hz i.e. by alternating the QP every 3 frames. This process is depicted in Fig. 3.

5) Frame Drops: We simulated video frame losses that occur when a source video is transmitted over a channel, such as a wireless network. We simulated frame drop clusters of adjacent frames to account for 10%-30% of a group of pictures (GOP). When a cluster of frames was removed from a video, the previous frame was repeated as many times as needed so that the total video duration remained unchanged. Three levels of frame drop densities were chosen: 3, 6 and 9 frames per cluster, yielding a slight to severe impact on the perceptual qualities of the videos.

6) Interlacing: On each frame of the video, the even and odd lines were separated to form two fields, field A and field B. Field B from each current frame and field A from each next frame were then combined to create interlaced frames. In the presence of motion, combing effects become evident. Since interlaced video fields are captured at different moments in time, interlaced frames often exhibit motion combing artifacts, when objects move quickly enough to be at different positions in each field.

C. Subjective Testing Environment and Display

The human study was carried out in the LIVE Subjective study room at The University of Texas at Austin. The Lab was arranged to simulate a living room environment. The windows were covered, and background distractions were removed. A Samsung UN65RU7100FXZA Flat 65-Inch 4K UHD TV was used to display all of the videos. All advanced motion optimization options on the TV, including the anti-judder and anti-flicker functions, were disabled. The viewing distance was about 2H, where H is the height of the TV so that the subjects could comfortably view the videos and assess the video distortions. The level of illumination was set to be similar to a living room, using one stand-up incandescent lamp and two indirect white LED studio lights behind the viewer. The lights were positioned to eliminate reflections from the lights on the screen.

Since the TV is able to upscale 1080p content using an unknown algorithm, all of the 1080p videos were instead upscaled using the Lanczos resizing function in OpenCV [47], to avoid any unpredictable effects. The 1080p videos were upscaled to 4K, after the distortions were applied. To ensure perfect playback, all of the videos were stored as raw YUV 4:2:0 files. The powerful Venueplayer application developed by VideoClarity was used to guarantee smooth playback of the 4K videos, without introducing any additional artifacts that could impact the perception of video quality.

After displaying each of the test videos, a continuous rating bar was displayed on the screen with a randomly placed cursor. The quality bar was marked with labels "Bad," "Poor," "Fair," "Good," and "Excellent" quality to facilitate the subjects in making decisions. The scores given by the subjects were sampled as integers from [0, 100] although numerical values were not made visible to the subjects. A Palette gear console was provided to enable the subjects to move the cursor without distraction. After moving the cursor to each desired scoring position, the subject depressed the button next to the sliding bar to confirm the score, which was then recorded without any further change. After each score was stored, the system immediately began to play the next video on the playlist.

D. Subjective Testing Design

In the human study, a single-stimulus (SS) method was employed, as described in the ITU-R BT 500.13 recommendation [48]. The reference videos are included as "hidden reference", not explicitly marked as "distorted" or "reference." The subjects used a rating bar to record their subjective opinion scores. Video rating scores were given after watching each video on an (invisible) scale ranging from 0 to 100, where 0 indicates the worst quality and 100 indicates the best quality. Due to the large number of video sequences, each subject participated in two sessions. The 45 contents associated with the pristine videos were divided into two sessions, where the reference videos and their corresponding distorted versions were grouped into the same session. The playlists within each of the two sessions were placed in randomized order for each subject, where videos of the same content, were separated by at least one video. This was done to counter any visual memory effects that might affect the subjective quality judgments, or any bias caused by playing the videos in a particular order. Each session required about 40 minutes.

E. Subjects and Training

A total of 40 human subjects were recruited from the student population at The University of Texas at Austin. The male/female gender ratio of the subject pool was 4.0. The mean and standard deviation of the ages of the participants was 23.47 and 1.78. Each subject participated in two sessions separated by at least 24 hours. Two of the subjects finished only one of the two sessions, while the rest of the 38 human subjects finished both sessions. 180 of the videos were rated by 40 subjects, while 187 videos were rated by 38 subjects. The subject pool was inexperienced with the topic of video quality assessment and video distortions.

The Snellen test and the Ishihara test were performed to validate each subject's vision. Two subjects were found to have 20/30 visual acuity, while one subject was found to have a color deficiency. However, these subjects were allowed to participate since the overall subject pool was deemed to be a good representation of the general population, following our common practice [49]. We conducted the tests as a screen against an unusual percentage of deficient subjects. Before the study, each subject was presented with a brief introduction to the study. The introduction described the study's goals, and gave detailed instructions on how to operate the system and assign scores. Each subject was asked to rate each video by quality only, without regard to the appeal of the content. Before the actual study commenced, each subject participated in a training session on two videos, to familiarize themselves with the system. The training videos and their scores were not included in the final database.

V. Processing of Subjective Scores

Subjective Mean Opinion Scores (MOS) were computed using the formulas below: Let s_{ij} denote the score by subject i for the video j. The subject scores were then converted into Z-scores z_{ij} for each subject. Subject rejection was performed based on the ITU-R BR 500.11 recommendation [48]. The scores z_{ij} for each video were tested against the normal distribution using the β_2 test:

$$\beta_{2j} = \frac{m_4}{(m_2)^2}, \tag{1}$$

where

$$m_x = \frac{\sum_{i=1}^{N} (z_{ij} - \overline{z}_{ij})^x}{N_j} \tag{2}$$

for subject i and video j, where N_j is the number of subjects that viewed video j. A score was regarded as normally distributed if β_{2j} fell between 2 and 4. We calculated the quantities P_i and Q_i for each subject i, by comparing z_{ij} with the mean $\bar{z_j}$ standard deviation σ_j of video j: If the score for video j was found to be normally distributed then:

if $z_{ij} \geq \bar{z_j} + 2\sigma_j$, then $P_i = P_i + 1$

if $z_{ij} \leq \bar{z_j} - 2\sigma_j$, then $Q_i = Q_i + 1$

If the score for video j was found to not be normally distributed, then:

if $z_{ij} \geq \bar{z_j} + \sqrt{20}\sigma_j$, then $P_i = P_i + 1$

if $z_{ij} \leq \bar{z_j} - \sqrt{20}\sigma_j$, then $Q_i = Q_i + 1$.

A subject i was rejected if the following two conditions held:

$$\frac{P_i + Q_i}{N} > 0.05, \tag{3}$$

and

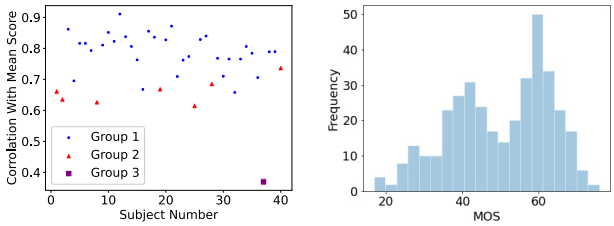
$$\left| \frac{P_i - Q_i}{P_i + Q_i} \right| < 0.3. \tag{4}$$

In our study, 8 of the 40 subjects satisfied these two conditions. However, since most of the rejected subjects fell close to the decision boundaries, we decided to revisit how the rejection criteria should be used. Given that the intent of subject rejection is to eliminate the outcomes of less engaged, distracted, or otherwise deficient subjects, we believed it worth considering whether any of the higher-deviation subjects were actually representative, as we have done in other recent studies [50]. We therefore computed the correlations between each subject's score and the MOS calculated using three different variations of the rejection criterion: 8 rejected, none rejected, and 1 (most anomalous) subject rejected, as shown in Fig. 4a. Specifically, the subjects were divided into three groups: Group 1 included all subjects not excluded by the ITU method. Group 2 and Group 3 included only the 8 subjects that were rejected, while Group 3 considered only of the single subject having the worst correlation against MOS. In the end, we chose to report all of the foregoing results by only excluding the single subject in Group 3. The objective NR VQA models were assessed using the three options mentioned, and the results are displayed in Table III. Section VI-B will go through the details of the NR VQA model evaluation. The objective VQA models lead to the same observation: nearly all VQA models have the highest correlation with the MOS computed when just the most anomalous subject is rejected. Including all of the original 8 rejected subjects except the one extreme outlier did not significantly affect either the inter-subject correlations that follow next, nor the algorithm results presented later. However, including the single extreme outlier affected both. Clearly, we are making a judgment that all but one of the 8 'rejected' subjects were serious and committed, and the extreme anomalous subject was not, i.e., was possibly distracted or simply inattentive. There is no way of knowing this for certain, and there is no formal way of defining such a subject. Instead, we are making a simple application of Occam's Razor that allows us the largest amount of reliable subjective data.

Table II shows our analysis of the data's internal consistency. Our modification of the typical outlier rejection criterion finds support in the analysis, and allows for a larger amount of likely representative data for model-building. We randomly divided the subjects into two equally sized groups and computed the Pearson correlation coefficient (PLCC) between the two groups' scores. We repeated this calculation over 1000 results, and report the mean and median correlations in Table. II. As may be seen, the best results were attained by removing the single very anomalous subject. We also observed negligible effect of the choice of rejection criteria on the objective algorithm performances reported later.

TABLE II
INTERNAL CONSISTENCY

	PLCC mean	PLCC median
All subjects	0.9616	0.9632
Reject Group 2 & 3	0.9603	0.9618
Reject Group 3	0.9635	0.9647



(a) (b)

Fig. 4. (a) The correlations against MOS of all subjects. Group 1 are subjects not rejected by the ITU method. Group 2 are all subjects that were rejected by the ITU method. Group 3 consists of the single lowest-correlating subject. (b) Distribution of MOS over all videos in the LIVE Livestream Database.

The Z-scores were then linearly rescaled from [-3,3] to [0,100]:

$$z'_{ij} = \frac{100(z_{ij} + 3)}{6}. \tag{5}$$

Finally the Mean Opinion Score (MOS) of each video was calculated:

$$MOS_j = \frac{1}{N_j} \sum_{i=1}^{N_j} z'_{ij}. \tag{6}$$

The converted MOS score is shown in Fig. 4b. Fig. 5 shows the distributions of scores for each individual video distortion class. The shapes of the MOS distributions of the reference videos are more Gaussian-like. The distorted video classes exhibit different distribution shapes, since they reflect different types and levels of distortion. Further, the MOS of the different levels of compression, flicker, and frame drop distortions are shown in Fig. 6. Generally, the MOS ranges of different distortion levels are mostly well-separated, but there are overlaps between distortion levels, largely because of the different interactions that occur between content and distortion. The perceptual quality of distorted (compressed videos) is affected by content masking, e.g. in regions containing significant high frequency spatial energy or high motion. While spatial masking is well-understood, temporal masking is less so, although it is known that motion has a silencing effect on flicker [51].

Fig. 7 plots the MOS against spatial resolution for each distortions class. The purely temporal distortions: judder and frame drops yielded similar ranges of MOS for 1080p and 4K videos. However, aliasing resulted in very different MOS ranges, likely because of the additional upscaling of 1080p videos when displayed on the 4K TV.

TABLE III
INTERNAL CONSISTENCY

	BRISQUE	CORNIA	HIGRADE	V-BLIINDS	TLVQM	ChipQA
All subjects	0.6229	0.6712	0.7031	0.7246	0.7422	0.7931
Reject Group 2 & 3	0.6239	0.6744	0.6818	0.7262	0.7496	0.7929
Reject Group 3	0.6381	0.6778	0.6916	0.7330	0.7503	0.7994

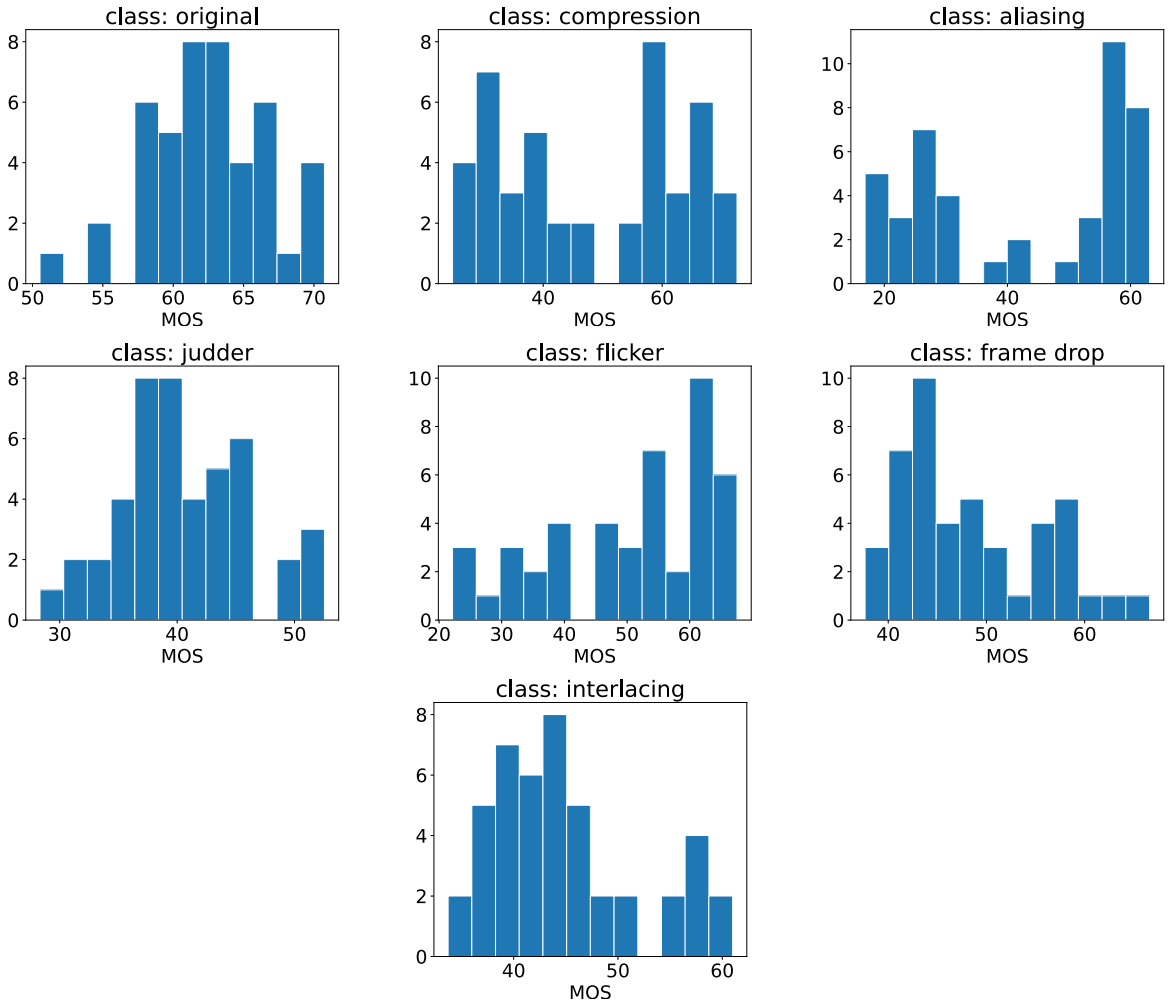


Fig. 5. Distribution of MOS of original, and synthetically distorted videos.

Tables IV and V show measurements of the consistency of human scores for each of the different distortion types. The Tables list the Spearman's Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC) computed on the entire database and for each distortion type, again by randomly dividing the subjects into two groups. It may be observed that the SROCC was slightly lower than the PLCC, which might be explained by subjects having difficulty supplying correctly ordered ratings of videos of very similar quality. but still generally able to make predictions in a linear manner. Overall, the results of the results indicate a very high degree of internal consistency and agreement amount the human subjects on all of the distorted video types.

Although MOS is a good representation of the subjective quality of videos and is necessary for the development and evaluation of NR VQA algorithms, the Difference MOS (DMOS) is more commonly used in the development and evaluation of FR VQA models, since it allows a way to reduce content dependencies of quality labels. Since we are supplying this resource for the study of both NR and FR models, we also calculated the DMOS of the videos with references. We calculated the DMOS according to:

$$DMOS_j = MOS_j^{ref} - MOS_j, \tag{7}$$

where MOS_j is the MOS of video j, and MOS_j^{ref} is the MOS of the reference video j, which is regarded as a "hidden reference," since it is not identified as such to the subjects.

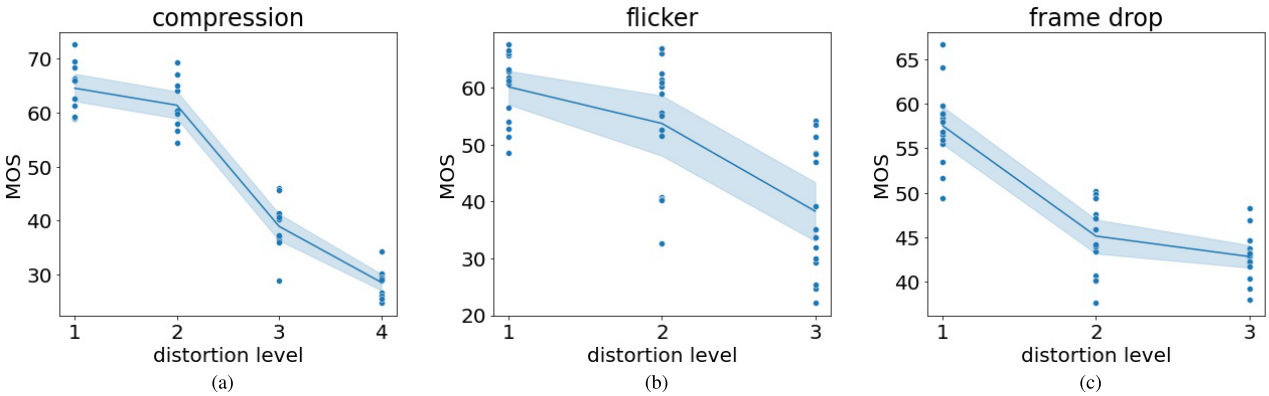


Fig. 6. MOS of videos affected by compression, flicker, and frame drops. Each dot in the plot is a video and the line and the shadowed regions indicate the average MOS and the 95% confident interval. (a)Compression, (b) Flicker, and (c) Frame drops.

TABLE IV
MIN, MEDIAN, AND MAX SROCC OF HUMAN SCORES DIVIDED INTO TWO GROUPS

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
MIN	0.9326	0.8285	0.8097	0.9092	0.8071	0.8543	0.8908
MEDIAN	0.9552	0.8967	0.8714	0.9425	0.8860	0.9151	0.9283
MAX	0.9685	0.9470	0.9250	0.9701	0.9373	0.9565	0.9651

TABLE V
MIN, MEDIAN, AND MAX PLCC OF HUMAN SCORES DIVIDED INTO TWO GROUPS

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
MIN	0.9292	0.9547	0.9688	0.9334	0.9287	0.8891	0.9253
MEDIAN	0.9648	0.9728	0.9792	0.9633	0.9607	0.9383	0.9524
MAX	0.9741	0.9870	0.9875	0.9795	0.9763	0.9644	0.9725

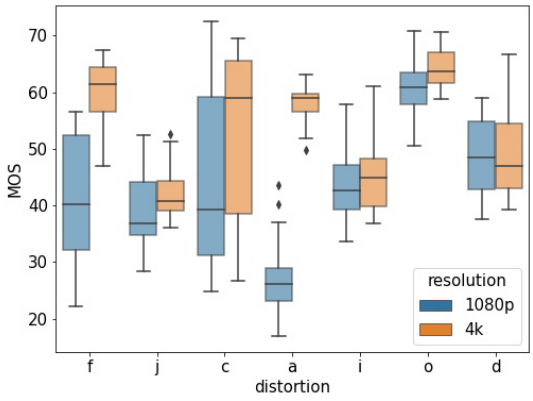


Fig. 7. Box plot comparing MOS against distortion type for both considered video resolutions. The labels on the horizontal axis represent: f: flicker; j: judder; c: compression; a: aliasing; i: interlacing; d: frame drop and o: original (reference videos).

VI. OBJECTIVE VQA MODEL COMPARISON

We evaluated several publicly available objective VQA algorithms on the LIVE Livestream Database to demonstrate the usefulness of the new resource. Given MOS and DMOS, we are able to test and compare both FR and NR VQA models. The performances of the objective VQA algorithms were evaluated using three standard metrics: the Spearman's Rank Order Correlation Coefficient (SROCC), the Pearson Linear Correlation Coefficient (PLCC), and the Root Mean Square Error (RMSE).

A. Performances of FR VQA Models

Here we present the results for the following seven popular FR VQA models: PSNR, SSIM, MS-SSIM, SpEEDQA, ST-RRED, FAST, and VMAF. The distorted versions of the 45 reference contents (270 videos in total) were processed to produce predictions that were cast against the DMOS. Note that most FR VQA models require that there be an equal number of frames between each reference video and its corresponding compared distorted video. However, the videos subjected to interlacing distortions have one less frame than the originals they derive from. Hence, the final frame of the interlaced video is duplicated to match the reference. The predicted scores s were passed though a five-parameter nonlinear logistic regression function before the PLCC and MSE were computed:

$$f(s) = \beta_1\left(\frac{1}{2} - \frac{1}{(1 + exp(\beta_2(s - \beta_3))}\right) + \beta_4 s + \beta_5, \quad (8)$$

TABLE VI
SROCC OF THE COMPARED FR VQA MODELS. THE SCORES OF THE TOP PERFORMING ALGORITHM ARE BOLDFACED

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
PSNR	0.3760	0.8750	0.4012	0.2117	0.5264	0.3024	**0.7507**
SSIM	**0.6976**	0.9171	0.7341	**0.3933**	0.8758	0.5291	0.6623
MS-SSIM	0.6757	0.9154	0.7335	0.3622	0.8652	0.5997	0.6179
SpEEDQA	0.6894	0.8979	0.8124	0.3165	0.8780	0.5993	0.6130
ST-RRED	0.6564	0.8943	0.8269	0.2968	0.8653	0.5635	0.7121
FAST	0.6192	**0.9283**	0.7269	0.2769	**0.9391**	0.7733	0.5960
VMAF	0.6434	0.9135	**0.9153**	0.3039	0.9243	**0.7843**	0.5346

TABLE VII
PLCC OF THE COMPARED FR VQA MODELS. THE SCORES OF THE TOP PERFORMING ALGORITHM ARE BOLDFACED

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
PSNR	0.4192	0.9586	0.4885	0.3452	0.5723	0.5886	**0.7840**
SSIM	0.7107	0.9659	0.7483	**0.5679**	0.8308	0.5770	0.7460
MS-SSIM	0.6907	**0.9690**	0.7696	0.5259	0.8589	0.6421	0.7105
SpEEDQA	**0.7235**	0.9526	0.9234	0.5037	0.8432	0.6183	0.7806
ST-RRED	0.6694	0.9483	**0.9425**	0.3952	0.8358	0.5915	0.7465
FAST	0.6520	0.9587	0.8329	0.4142	**0.9391**	0.8298	0.6978
VMAF	0.6355	0.9675	0.9296	0.3043	0.9242	**0.8654**	0.6242

TABLE VIII
RMSE OF THE COMPARED FR VQA MODELS. THE SCORES OF THE TOP PERFORMING ALGORITHM ARE BOLDFACED

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
PSNR	10.3355	4.2304	13.6253	4.3601	9.5390	5.2311	**3.9024**
SSIM	8.0082	3.8493	10.3588	**3.8237**	6.4740	5.2852	4.1864
MS-SSIM	8.2324	**3.6708**	9.9705	3.9510	5.9578	4.9605	4.4235
SpEEDQA	**7.8589**	4.5223	5.9924	4.0129	6.2531	5.0856	3.9294
ST-RRED	8.4573	4.7155	**5.2190**	4.2673	6.3858	5.2174	4.1832
FAST	8.6315	4.2267	8.6452	4.2282	**3.7669**	3.6114	4.5028
VMAF	8.7894	3.7600	5.7557	4.4251	4.4430	**3.2435**	4.9154

TABLE IX
SROCC OF THE COMPARED NR VQA MODELS. THE SCORES OF THE TOP PERFORMING ALGORITHM ARE BOLDFACED

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
NIQE	0.3232	0.2775	0.2860	0.2863	0.2832	0.2842	0.2780
BRISQUE	0.6381	0.6409	0.7482	0.8039	0.6440	0.4180	0.8720
CORNIA	0.6778	0.7399	**0.8142**	0.7049	0.7193	0.0000	**0.8782**
HIGRADE	0.6916	0.7234	0.7337	0.5784	0.6429	0.5748	0.8060
V-BLIINDS	0.7330	0.7131	0.7482	**0.8679**	0.5769	0.7513	0.7936
TLVQM	0.7503	0.6574	0.7915	0.8246	0.6966	**0.8927**	0.8369
ChipQA	**0.7994**	**0.7482**	0.7998	0.8514	**0.7668**	0.7874	0.8111

where s are the predicted scores produced by the tested algorithm and $f(s)$ is the mapped score. By fitting parameters β_i $(i = 1, 2, 3, 4, 5)$, the MSE between the mapped and subjective scores is minimized. The SROCC, PLCC, and RMSE for each category of distortions are calculated by comparing the predictions made by the FR models and the ground truth

TABLE X
PLCC OF THE COMPARED NR VQA MODELS. THE SCORES OF THE TOP PERFORMING ALGORITHM ARE BOLDFACED

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
NIQE	0.4962	0.2805	0.2865	0.2860	0.2848	0.2849	0.2850
BRISQUE	0.6698	0.7616	0.9415	0.8362	0.7166	0.4265	0.9185
CORNIA	0.7257	0.8197	0.9595	0.7409	0.7841	0.0000	**0.9234**
HIGRADE	0.6990	0.8395	0.9426	0.6310	0.6938	0.5806	0.8528
V-BLIINDS	0.7477	0.8313	0.9277	**0.9239**	0.6238	0.7850	0.8826
TLVQM	0.7513	0.6991	0.9550	0.8850	0.8037	**0.9153**	0.8648
ChipQA	**0.8156**	**0.8408**	**0.9613**	0.9040	**0.8608**	0.8470	0.8587

TABLE XI
RMSE OF THE COMPARED NR VQA MODELS. THE SCORES OF THE TOP PERFORMING ALGORITHM ARE BOLDFACED

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
NIQE	50.4055	45.7805	45.8797	50.8770	49.9032	50.8316	50.9243
BRISQUE	9.6376	9.1434	5.5712	7.1173	8.4869	9.1254	4.3474
CORNIA	9.6960	8.0173	4.6140	8.5343	7.3121	9.7778	**4.3074**
HIGRADE	9.6469	**7.7381**	5.2704	9.9567	8.6036	8.0093	5.8163
V-BLIINDS	8.4058	7.7836	6.1912	**4.7971**	9.3751	5.8211	5.1704
TLVQM	8.7217	10.0801	4.8209	6.1302	7.1367	**4.0113**	5.5803
ChipQA	**7.2874**	7.7510	**4.3791**	5.3599	**5.6626**	5.0459	5.6679

for each of those distortions separately. Table VI, VII, and VIII show the performance metrics of the compared algorithms, which will be discussed shortly.

B. Performance of NR VQA Models

We compared the quality predictions made by a variety of NR models against the MOS. The NR VQA algorithms that were tested include NIQE [52], BRISQUE [53], HIGRADE [54], CORNIA [55], TLVQM [56], V-BLIINDS [57], and ChipQA [58], [59]. BRISQUE, HIGRADE, CORNIA, TLVQM, V-BLIINDS, and ChipQA are supervised learning algorithms that use a support vector regressor (SVR) to learn mappings from 'quality-aware' features to mean opinion scores. These algorithms were tested on 1000 random train-test splits. On each split, 80% of the data was used for training, and 20% for testing. Following common practice, 5-fold cross-validation was applied within each training set to find the best parameters for the SVR. Care was taken to ensure that no content could appear in both the training and testing set, the same is also true for the 5-fold cross-validation.

NIQE, BRISQUE and HIGRADE are image quality assessment (IAQ) algorithms, so they were used to extract features frame by frame, followed by temporal average pooling.

For the unsupervised methods (NIQE), the scores s were passed through the same nonlinear logistic regression process before the PLCC and MSE were computed, as described earlier. The performances of the compared VQA models on the entire database, as well as for each synthetic distortion, are shown in Tables IX, X, and XI, where the best performing model on each distortion category is boldfaced. The results for each specific distortion were acquired by training the SVR on the reference sequences and the specific distorted sequences. Scatter plots of some selected objective VQA models against MOS are shown in Fig. 8.

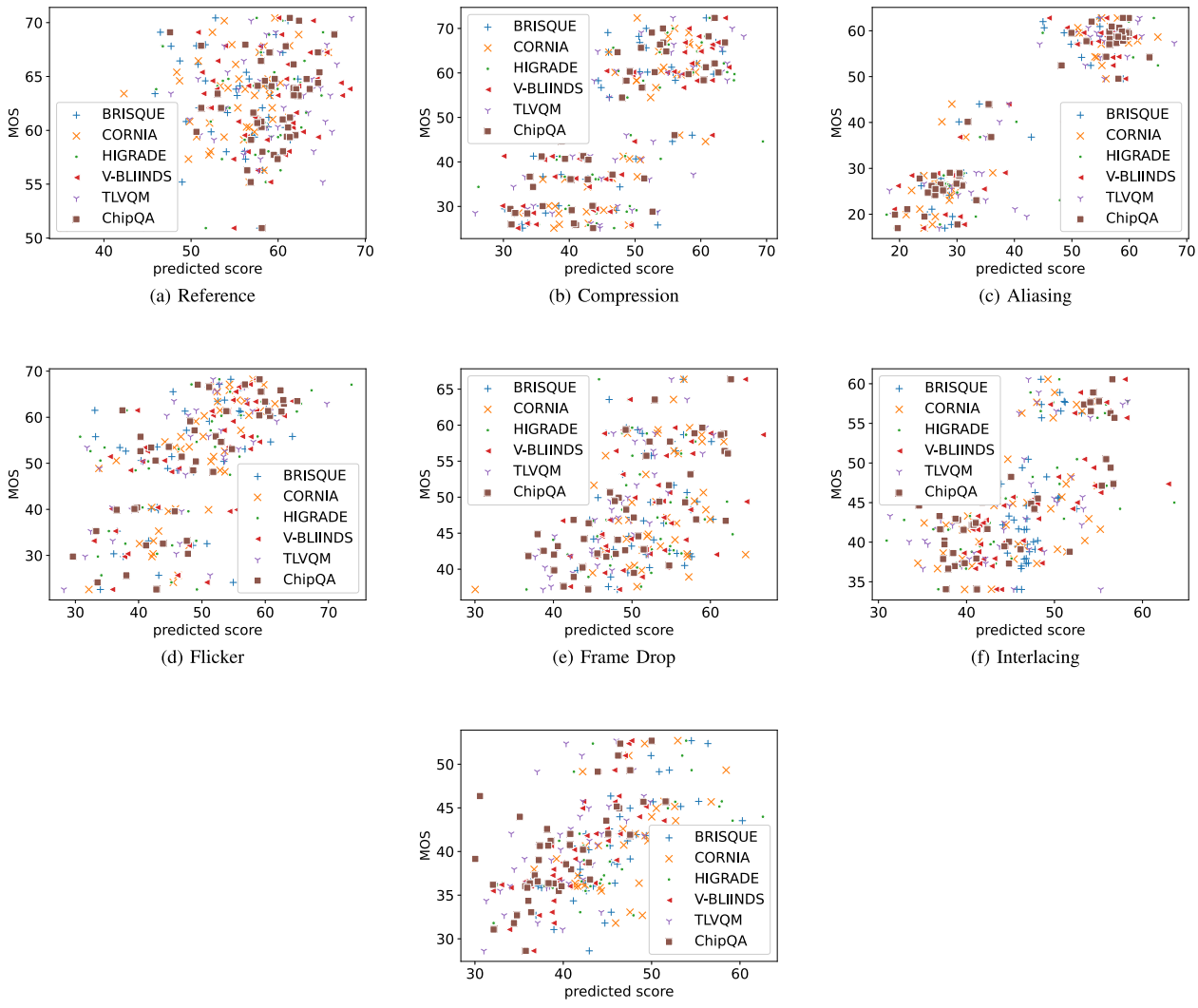
TABLE XII
COMPUTATION TIME ON A SINGLE 3840 × 2160 VIDEO WITH 210 FRAMES FROM THE LIVE LIVESTREAM VQA DATABASE

Algorithm	Time (s)	GFLOPS	Complexity
NIQE	1008	3094	$\mathcal{O}(k^2NT)$
BRISQUE	301	352	$\mathcal{O}(k^2NT)$
TLVQM	1002	477	$\mathcal{O}(k_1^2NT+(\log(N)+k_2^2)NT_2))$
CORNIA	2056	4480	$\mathcal{O}(k^2MNT)$
VBLIINDS	3086	465	$\mathcal{O}((k^2N+\log(w)N+w^2d^3)T)$
HIGRADE	16240	9604	$\mathcal{O}(3(2k^2+k_2)NT)$
ChipQA	814	700	$\mathcal{O}(k^2+Q/R+(Q/R^3)\log Q)\frac{NT}{D^2}$

k: window size; N pixel number per frame; T: number of frames;
TLVQM: k_1,k_2: filter size, T_2: number of representative frames;
CORNIA, M: codebook size;
V-BLIINDS, w: window size, d: motion vector tensor size;
HIGRADE, k_2: gradient kernel size;
ChipQA, D: downsampling factor. Q: chip search quantization factor. R: chip dimension.

C. Statistical Evaluation

A one-sided t-test was performed on the 1000 SROCC scores of the NR VQA models computed on the LIVE



(a) Reference (b) Compression (c) Aliasing

(d) Flicker (e) Frame Drop (f) Interlacing

(g) Judder

Fig. 8. Scatter plots of the predicted scores produced by several NR VQA models against MOS for each class of distorted videos.

Livestream Database, using the 95% confidence level to evaluate whether one VQA algorithm was statistically superior to another. The results are shown in Table XIII. Results on the entire database and on individual distortions are both included. Each entry in the table consists of 7 symbols corresponding to the entire database, and the 6 distortions, in the order of compression, aliasing, judder, flicker, frame drop, and interlacing. A symbol '1' indicates using the performance of the algorithm on the row was statistically superior to that of the column, while a symbol '0' indicates that the column algorithms was statistically better than the row algorithm. A symbol of '-' indicates that the performances of the row and the column algorithms were statistically equivalent.

D. Computational Cost

Since we are interested in live streaming use scenarios, we studied the computational costs, the number of giga floating point operations (GFLOPS), and complexity of the compared models, as shown in Table XII. The $\mathcal{O}(\cdot)$ figures make clear that all of the compared algorithms could be implemented as real-time hardware realizations. To measure computation time, we used a single 4K video having 210 frames. Of the compared algorithms, V-BLIINDS, and ChipQA were implemented in Python. All other algorithms were implemented in MATLAB®. All the algorithms were run on an Intel Xeon E5-2620 CPU with a maximum frequency of 3 GHz.

While none of the tested algorithms runs in real time in their current implementations, they may be optimized to do so. In most of the algorithms, the most expensive step is filtering. For example, in BRISQUE the largest computation is computing the mean subtracted contrast normalized (MSCN) coefficients. However, filtering scales up linearly and is highly parallelizable. Frame based algorithms can be applied at a lower frame rate with little loss of prediction efficacy [60]. While V-BLIINDS expends considerable computation on motion computation, motion vectors can be re-used from those produced by the involved codec. The complexity of CORNIA, which computes dot-products between local descriptors and visual codewords, is affected by the codebook size, which can be quite large.

TABLE XIII

RESULTS OF ONE-SIDED T-TEST PERFORMED BETWEEN SROCC VALUES OF VARIOUS ALGORITHMS ON THE LIVE LIVESTREAM DATABASE. EACH CELL CONTAINS 7 ENTRIES: THE ENTIRE DATABASE, 6 DISTORTIONS IN THE ORDER: COMPRESSION, ALIASING, JUDDER, FLICKER, FRAME DROP, AND INTERLACING. A VALUE OF '1' INDICATES THAT THE ROW IS STATISTICALLY SUPERIOR (BETTER VISUAL QUALITY) THAN THE COLUMN, WHILE A VALUE OF '0' INDICATES THAT THE COLUMN IS STATISTICALLY SUPERIOR THAN THE ROW. A VALUE OF '-' INDICATES STATISTICAL EQUIVALENCE BETWEEN ROW AND COLUMN

ALGORITHM	NIQE	BRISQUE	CORNIA	HIGRADE	V-BLIINDS	TLVQM	ChipQA
NIQE	-------	0000000	0000000	0000000	0000000	0000000	0000000
BRISQUE	1111111	-------	00-101-	0011-01	00-0101	01--101	0000001
CORNIA	1111111	11-010-	-------	011-001	01-0001	01-0101	0-10001
HIGRADE	1111111	1100-10	100-110	-------	0000101	010010-	0000000
V-BLIINDS	1111111	11-1010	10-1110	1111010	-------	11-1-0-	0001000
TLVQM	1111111	10--010	10-1010	101101-	00-0-1-	-------	0000011
ChipQA	1111111	111110	1-01110	1111111	1110111	1111100	-------

VII. DISCUSSION

The results presented in Tables VI, VII and VIII suggest that, other than PSNR, the compared FR VQA models generally delivered similar overall performances on the entire database, but some algorithms yielded better performances on certain distortions. For example, SSIM, which performed well overall, obtained the highest correlation against DMOS on the compressed videos, but low correlation on the judder videos. The main reason that SSIM delivers low performance on judder videos is that it is a frame-based model. Judder is a temporal distortion that arises when high motion is present in a video. The greater the magnitude of the motion, the more apparent the distortion is likely to be. While SSIM effectively captures spatial distortions (like compression), it is unable to capture the temporal effects of judder. ST-RRED does include limited temporal information expressed as NSS features from adjacent frame differences, which is inadequate to model complex or longer-duration temporal distortions, hence it does not outperform the other compared FR models. VMAF yielded the highest correlation on the aliased and flicker videos, but low correlations on the interlaced videos. The FR VQA models tended to deliver decent performances on common distortions found in other VQA databases, such as compression, and also flicker, which is compression based. These distortions are better studied and easier to catch with the presence of the reference. However, when tested on the purely temporal distortions, all of the compared FR VQA models delivered low correlations against DMOS. This suggests ample room for research on developing better models of temporal and motion-related distortions.

From Tables IX, X, and XI, it may be observed that ChipQA performed the best among the compared NR VQA algorithms, while TLVQM and V-BLIINDS also achieved relatively higher correlations against the human judgments. TLVQM achieved the top performance on flicker and frame drops, likely because of the large number of temporal features it uses. ChipQA builds a statistical representation of local spatiotemporal data that is attuned to local orientations of motion over large spatial fields, motivated by processes in areas V1 and MT of the brain. The explicit modeling of deviations from statistical regularity in the spatiotemporal domain allows it to perform well on both spatial and temporal distortions. NIQE and BRISQUE are similar methods, but BRISQUE is trained while NIQE is completely blind, hence BRISQUE usually can deliver predictions having higher correlations against human quality judgments. Similar statistical features are used in V-BLIINDS and HIGRADE. The frame-based models NIQE, BRISQUE, HIGRADE, and CORNIA do not access any motion information, which greatly limits their performance. CORNIA yielded top performances on compression, aliasing, and interlacing, all of which present strong spatial aspects of distortion. However, the overall performance of CORNIA was lower than that of V-BLIINDS, TLVQM, and ChipQA, due to the lack of temporal information.

VIII. CONCLUSION

We created a large scale video quality database targeting high motion, live streaming scenarios. The new resource includes 45 source sequences from 33 original contents and 6 different distortion types. The new database can be used to create, test, and compare both NR and FR VQA models. We are making the new LIVE Livestream database publicly available. Future steps include developing new NR VQA models using the proposed database.

ACKNOWLEDGMENT

The authors acknowledge the Texas Advanced Computing Center (TACC), The University of Texas at Austin for providing HPC, visualization, database, and grid resources that have contributed to the research results reported in this paper. URL: http://www.tacc.utexas.edu

REFERENCES

[1] *Cisco Global Cloud Index Forecast and Methodology 2015-2020*. Accessed: Dec. 29, 2021. [Online]. Available: https://www.cisco.com/c/dam/m/en_us/service-provider/ciscoknowledgenetwork/files/622_11_15-16-Cisco_GCI_CKN_2015-2020_AMER_EMEAR_NOV2016.pdf

[2] (2018). *2018 Global Internet Phenomena Report*. [Online]. Available: https://www.sandvine.com/phenomena

[3] C.-C. Chen and Y.-C. Lin, "What drives live-stream usage intention? The perspectives of flow, entertainment, social interaction, and endorsement," *Telematics Informat.*, vol. 35, no. 1, pp. 293–303, Apr. 2018.

[4] S. M. Keating, "Image signal processing with digital filtering to minimize aliasing caused by image manipulation," U.S. Patent 5 206 919, Apr. 27, 1993.

[5] 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, Jan. 2010.

[6] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[7] 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.

[8] S. Tomar, "Converting video formats with FFmpeg," *Linux J.*, vol. 2006, no. 146, p. 10, 2006.

[9] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5217–5231, Nov. 2017.

[10] 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.

[11] C. Chen, X. Zhu, G. de Veciana, A. C. Bovik, and R. W. Heath, Jr., "Adaptive video transmission with subjective quality constraints," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 2477–2481.

[12] C. Chen, L. Kwon Choi, G. de Veciana, C. Caramanis, R. W. Heath, Jr., and A. C. Bovik, "Modeling the time—Varying subjective quality of HTTP video streams with rate adaptations," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2206–2221, May 2014.

[13] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, Jr., and A. C. Bovik, "A dynamic system model of time-varying subjective quality of video streams over HTTP," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3602–3606.

[14] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Conf. Rec. Forty 6th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 1693–1697.

[15] F. De 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.*, Jan. 2010, pp. 2430–2433.

[16] *IVP Subjective Quality Video Database*. Accessed: Dec. 29, 2021. [Online]. Available: https://qualinet.github.io/databases/video/ivp_subjective_quality_video_database/

[17] A. K. Moorthy, L. K. Choi, G. de Veciana, and A. Bovik, "Mobile video quality assessment database," in *Proc. ICC Workshop Realizing Adv. Video Optimized Wireless Netw.*, 2012, pp. 7055–7059.

[18] L. K. Choi, L. K. Cormack, and A. C. Bovik, "On the visibility of flicker distortions in naturalistic videos," in *Proc. 5th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 164–169.

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

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

[21] Y. Zhang *et al.*, "Subjective panoramic video quality assessment database for coding applications," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 461–473, Jun. 2018.

[22] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2620–2630, Oct. 2018.

[23] T. Li, X. Min, H. Zhao, G. Zhai, Y. Xu, and W. Zhang, "Subjective and objective quality assessment of compressed screen content videos," *IEEE Trans. Broadcast.*, vol. 67, no. 2, pp. 438–449, Jun. 2021.

[24] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Trans. Image Process.*, vol. 29, pp. 6054–6068, 2020.

[25] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective quality assessment of H.264/AVC video streaming with packet losses," *EURASIP J. Image Video Process.*, vol. 2011, pp. 1–12, Jan. 2011.

[26] J. Y. Lin, R. Song, 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.

[27] C. Keimel, J. Habigt, T. Habigt, M. Rothbucher, and K. Diepold, "Visual quality of current coding technologies at high definition IPTV bitrates," in *Proc. IEEE MMSP*, Oct. 2010, pp. 390–393.

[28] C. Keimel, A. Redl, and K. Diepold, "The TUM high definition video datasets," in *Proc. QoMEX*, 2012, pp. 97–102.

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

[30] M. Vranješ, S. Rimac-Drlje, and K. Grgić, "Review of objective video quality metrics and performance comparison using different databases," *Signal Process., Image Commun.*, vol. 28, no. 1, pp. 1–19, 2013.

[31] Y.-F. Ou, Y. Zhou, and Y. Wang, "Perceptual quality of video with frame rate variation: A subjective study," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Jan. 2010, pp. 2446–2449.

[32] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2473–2486, Jun. 2014.

[33] 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.

[34] 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.

[35] M. Torres Vega, V. Sguazzo, D. C. Mocanu, and A. Liotta, "An experimental survey of no-reference video quality assessment methods," *Int. J. Pervasive Comput. Commun.*, vol. 12, no. 1, pp. 66–86, Apr. 2016.

[36] A. Mercat, M. Viitanen, and J. Vanne, "Uvg dataset: 50/120fps 4K sequences for video codec analysis and development," in *Proc. MMSys*, 2020, pp. 297–302.

[37] M. Cheon and J.-S. Lee, "Subjective and objective quality assessment of compressed 4k UHD videos for immersive experience," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 7, pp. 1467–1480, Jul. 2018.

[38] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *Netflix Tech. Blog*, vol. 6, p. 2, Jun. 2016.

[39] Video Quality Experts Group. (2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment*. Accessed: Oct. 31, 2020. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtvphaseI

[40] W. Yodel. (2011). *The Consumer Digital Video Library*. [Online]. Available: https://cdvl.org/

[41] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *Proc. QoMEX*, 2013, pp. 34–35.

[42] S. Daly, N. Xu, J. Crenshaw, and V. J. Zunjarrao, "A psychophysical study exploring judder using fundamental signals and complex imagery," *SMPTE Motion Imag. J.*, vol. 124, no. 7, pp. 62–70, Oct. 2015.

[43] S. R. Oh, D. Kim, P. G. Heo, and H. Park, "A new metric for judder in high frame-rate video," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3802–3806.

[44] (Sep. 2020). *2020 Bitmovin Video Developer Report*. [Online]. Available: https://f.hubspotusercontent30.net/hubfs/3411032/Bitmovin%20Developer

[45] *Panasonic DVX-100b Manual*. Accessed: Dec. 29, 2021. [Online]. Available: https://tfma.temple.edu/sites/tfma/files/site-pdfs/DVX100aManual.pdf

[46] *Canon XL2 Manual*. Accessed: Dec. 29, 2021. [Online]. Available: https://www.usa.canon.com/internet/portal/us/home/support/details/camcorders/support-professional-camcorders/xl2/xl2?tab=manuals

[47] G. Bradski, "The openCV library," *Dr. Dobb's J. Softw. Tools*, vol. 25, no. 11, pp. 120–123, 2000.

[48] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Recommendation BT. 500-13, 2012.

[49] *Laboratory for Image and Video Engineering*. Accessed: Dec. 2021. [Online]. Available: https://live.ece.utexas.edu/research/Quality/visualScreening.htm

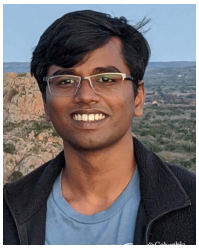
[50] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," 2020, *arXiv:2007.11634*.

[51] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen, "Flicker effects in adaptive video streaming to handheld devices," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 463–472.

[52] 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.

[53] 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.
[54] 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.
[55] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. CVPR*, 2012, pp. 1098–1105.
[56] 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.
[57] 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.
[58] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, and A. C. Bovik, "No-reference video quality assessment using space-time chips," 2020, *arXiv:2008.00031*.
[59] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "ChipQA: No-reference video quality prediction via space-time chips," *IEEE Trans. Image Process.*, vol. 30, pp. 8059–8074, 2021.
[60] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.

Zaixi Shang (Student Member, IEEE) received the B.E. degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China. He is currently pursuing the Ph.D. degree with The University of Texas at Austin. He joined the Laboratory for Image and Video Engineering, The University of Texas at Austin, in 2019. His research interests include image and video processing, visual perception, machine learning, and computer vision.

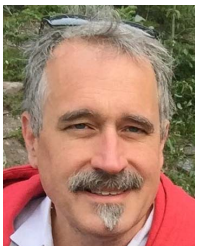


Joshua Peter Ebenezer received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology Kharagpur in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of Texas at Austin. He joined the Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin, in 2019, where he has been serving as an Assistant Director since 2020. He was a recipient of the Cockrell Graduate Engineering Fellowship from The University of Texas at Austin from 2019 to 2023, and the Nilanjan Ganguly Memorial Award and the Goralal Syngal Memorial Scholarship from the Indian Institute of Technology Kharagpur in 2018.

Yongjun Wu (Senior Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering and the areas of video streaming and compression from Rensselaer Polytechnic Institute, Troy, NY, in 2005. Since 2015, he has been working at Amazon Prime Video, Seattle, WA, as the Chief Architect of video playback end-to-end from camera to screen. From 2005 to 2015, he worked at Microsoft, Redmond, WA, USA, working on algorithms and projects related to Windows media. He has about 160 patent applications, most of which are approved, in the fields of audio and video streaming, compression, processing, quality evaluation, media platform optimizations and designs, and computer vision, from his Ph.D. work and industry projects. He published tens of conference and journal articles and MPEG contributions and two book chapters. He is a member of SMPTE. He received a few awards for his work at Microsoft.

Hai Wei (Member, IEEE) received the master's degree from Northwest University in 1997 and the Ph.D. degree from the Beijing University of Technology in 2001. From 2001 to 2005, he was a Post-Doctoral Research Fellow at the University of Hawaii at Manoa. Before joining Amazon, he served as the Director for the vision technology at Utopia Compression, leading research programs in advanced image compression, computer vision, and learning algorithms for unmanned systems. He is currently a Principal Research Scientist at Amazon Prime Video. He has coauthored more than 50 articles and patents. His research interests include image processing, video compression, video quality analysis, content-based visual learning, and machine vision. He is a member of SPIE.

Sriram Sethuraman received the Ph.D. degree from CMU. He is currently a Senior Principal Scientist at Prime Video Playback Organization, leading efforts related to encoding optimization, video quality measurement, ML-based restoration, and next-generation video compression. He joined PV in July 2019, prior to which he was the CTO and the Senior VP at Ittiam Systems, a Bengaluru-based multimedia technology venture. During his 17-year tenure at Ittiam, he was the Architect of its technologies and products in the fields of video compression, video communication, media broadcast, and computer vision/machine learning. He has been part of MPEG-4, MPEG-7, and VVC standardization efforts. Prior to joining Ittiam, he served as a Senior Member of Technical Staff at Sarnoff Corporation. He has 34 issued patents (and several pending patents) and the author of more than 35 publications.



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