A Subjective and Objective Study of Space-Time Subsampled Video Quality

Dae Yeol Lee, Somdyuti Paul, *Graduate Student Member, IEEE*, Christos G. Bampis, *Member, IEEE*, Hyunsuk Ko, *Member, IEEE*, Jongho Kim, Se Yoon Jeong, Blake Homan, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*— Video dimensions are continuously increasing to provide more realistic and immersive experiences to global streaming and social media viewers. However, increments in video parameters such as spatial resolution and frame rate are inevitably associated with larger data volumes. Transmitting increasingly voluminous videos through limited bandwidth networks in a perceptually optimal way is a current challenge affecting billions of viewers. One recent practice adopted by video service providers is space-time resolution adaptation in conjunction with video compression. Consequently, it is important to understand how different levels of space-time subsampling and compression affect the perceptual quality of videos. Towards making progress in this direction, we constructed a large new resource, called the ETRI-LIVE Space-Time Subsampled Video Quality (ETRI-LIVE STSVQ) database, containing 437 videos generated by applying various levels of combined space-time subsampling and video compression on 15 diverse video contents. We also conducted a large-scale human study on the new dataset, collecting about 15,000 subjective judgments of video quality. We provide a rate-distortion analysis of the collected subjective scores, enabling us to investigate the perceptual impact of space-time subsampling at different bit rates. We also evaluated and compare the performance of leading video quality models on the new database. The new ETRI-LIVE STSVQ database is being made freely available at (https://live.ece.utexas.edu/research/ETRI-LIVE_STSVQ/index.html).**

Manuscript received September 9, 2020; revised September 5, 2021; accepted December 6, 2021. Date of publication December 29, 2021; date of current version January 6, 2022. This work was supported by the Institute for Information & Communications Technology Promotion (IITP) Grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] (Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Teramedia) under Grant 2017-0-00072. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Giuseppe Valenzise. *(Corresponding author: Dae Yeol Lee.)*

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) of The University of Texas at Austin under Protocol No. 2007-11-0066 - PI: Alan C. Bovik - Title: Intelligent Autonomous Video Quality Agents (Subjective Quality Assessment of Digital Video) - Period: 05/11/2020 to 05/10/2023.

Dae Yeol Lee, Somdyuti Paul, 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: daelee711@utexas.edu; somdyuti@utexas.edu; bovik@ece.utexas.edu).

Christos G. Bampis is with Netflix Inc., Los Gatos, CA 95032 USA (e-mail: christosb@netflix.com).

Hyunsuk Ko is with the School of Electrical Engineering, Hanyang University ERICA, Ansan 15588, South Korea (e-mail: hyunsuk@hanyang.ac.kr).

Jongho Kim and Se Yoon Jeong are with the Media Coding Research Section, ETRI, Daejeon 34129, South Korea (e-mail: pooney@etri.re.kr; jsy@etri.re.kr).

Blake Homan is with Video Clarity, Inc., Campbell, CA 95008 USA (e-mail: blake@videoclarity.com).

Digital Object Identifier 10.1109/TIP.2021.3137658

***Index Terms*— Video quality database, space-time subsampled video coding, human study, perceptual quality, video quality assessment.**

I. Introduction

THE streaming and social media industry is continuously progressing towards providing more realistic and immersive experiences to video consumers. Display companies and content providers are enabling higher spatial resolutions, frame rates, and high dynamic range (HDR). Televisions and monitors are now available that support 8K HDR and/or true 120Hz 10-bit input and playout. Popular media streaming services, such as YouTube, Netflix, and Amazon, now provide contents at 4K/60fps/HDR, and it is expected that increases in these video parameters will be met by even larger, faster, and deeper displays and streamed video content. However, increases in video dimensions inevitably increase streamed data volume, hence service providers are increasingly challenged to deliver high-quality videos with limited bandwidths. while providing the highest possible quality.

Video compression is the principal technology that enables bandwidth-constrained video streaming, as exemplified by the global ITU standards H.264 [1], HEVC [2], and the emerging Versatile Video Coder (VVC), as well as the open source standards VP9 [3] and AV-1 [4].

Given increases in video dimensions, a recent approach taken by streaming video providers is to combine resolution adaptation with compression. For example, a spatially subsampled video may require less quantization (compression) to meet a given bit rate requirement, and possibly resulting in a perceptually less degraded video, depending on the content. Thus far, this practice has been largely limited to spatial subsampling, but temporal, and more generally, space-time subsampling, also offer the potential for increased efficiencies.

There have been some studies that have investigated the combined effects of spatial subsampling and compression on perceptual video quality [5], [6]. The authors of [7] investigated the perceptual quality of videos with varying spatial adaptation filters including the nearest-neighbor, bicubic, and Convolutional Neural Network (CNN) based super resolution filter.

Other authors have studied temporal subsampling and its effects on subjective video quality, but without considering coincident compression or spatial subsampling. They used these results to motivate resolution adaptation methods which

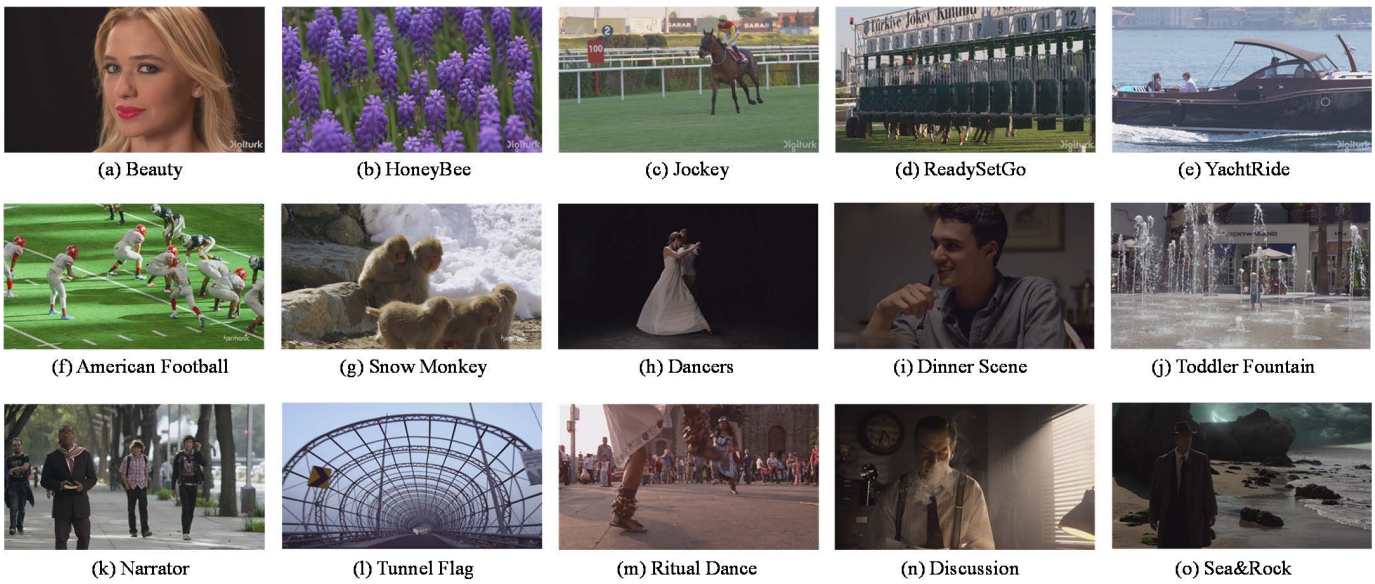

Fig. 1. Sample frames from source contents in the ETRI-LIVE space-time subsampled video quality database.

reduce video frame rate if the content does not perceptually benefit from a higher frame rate [8], [9]. In [10], a spatio-temporal resolution adaptation method for video compression was proposed, but quality prediction and consequent down-sampling decisions were conducted separately in space and time. The authors of [11] did study the joint application of space-time subsampling and compression, but the codec was confined to H.264, and the maximum considered frame rate was 60 Hz.

These prior efforts have helped us to understand how spatial and temporal video density affect perceptual quality. However, given that spatial and temporal (space-time) subsampling and compression are likely to be applied in concert, studies are needed to be able to understand and model how they affect perceived video quality when they are jointly applied. Towards advancing progress in this direction, we have constructed a large-scale video quality database entitled the "ETRI-LIVE Space-Time Subsampled Video Quality (ETRI-LIVE-STSVQ)" database, which contains a large number of videos operating at different space resolutions, temporal frame rates, and levels of compression, along with collected subjective human opinion scores on all of them. The contributions that we make include:

- The first database with subjective quality scores rendered on 4K 10-bit videos at frame rates up to 120Hz, subjected to simultaneous space-time subsampling and compression (HEVC) distortions applied at multiple levels. A total of 437 space-time subsampled and compressed videos were created.
- We conducted a large-scale laboratory human study on the videos, using a high-speed video playout system capable of displaying true 120 Hz 10-bit video signals in real-time.
- Since the new database can be uniquely used to design and compare video quality models that can predict the perceptual quality of space-time subsampled and compressed videos, we conducted a comparative study of relevant popular VQA models on the prediction problem.
- The new database is a unique psychometric resource for understanding the perceptual effects of space-time subsampling and compression, and for designing strategies for subsampling and compression parameter control to achieve perceptually optimized streaming.

The ETRI-LIVE STSVQ database is being made freely and publicly available at (https://live.ece.utexas.edu/research/ETRI-LIVE_STSVQ/index.html) to assist future research and development on space-time video quality modeling and perceptually optimized video coding. The rest of the paper is organized as follows: Section II provides a detailed description of the construction of the database. Section III describes the subjective experiment protocol. Section IV describes the data processing and analysis of the subjective opinion scores. Section V compares the performances of various relevant high-performance video quality models on the new database. Finally, conclusions are drawn in Section VI.

II. Construction of the Database

A. Source Contents

We collected 15 high quality 4K 10-bit source contents having wide variety of spatiotemporal properties. Of the 15 contents, five are from the Ultra Video Group (UVG) dataset [12], two are Harmonic 4K footages, and eight are from the Netflix public video library [13]. Fig. 1 shows sample frames of each of the source contents, while Table I the formats of the source videos. As shown in the table, all of the source contents are of high spatial resolutions and frame rates of at least 3840×2160 and 60 fps, respectively. We set the video format of the source contents as 3840×2160, YUV420p, and 10 bits. A few contents of spatial resolution 4096×2160 were slightly cropped, and those of format YUV422p were chroma

TABLE I
SUMMARY OF SOURCE CONTENT VIDEO FORMATS

Video	Bit Depth	Chroma Format		Spatial Resolution		Frame rate (fps)	No. of Frames	Duration (sec)
		Original	Processed	Original	Processed			
Beauty	10	YUV420p	YUV420p	3840×2160	3840×2160	120	600	5
HoneyBee	10	YUV420p	YUV420p	3840×2160	3840×2160	120	600	5
Jockey	10	YUV420p	YUV420p	3840×2160	3840×2160	120	600	5
ReadySetGo	10	YUV420p	YUV420p	3840×2160	3840×2160	120	600	5
YachtRide	10	YUV420p	YUV420p	3840×2160	3840×2160	120	600	5
Snow Monkey	10	YUV420p	YUV420p	3840×2160	3840×2160	60	360	6
American Football	10	YUV420p	YUV420p	3840×2160	3840×2160	60	420	7
Dancers	10	YUV420p	YUV420p	4096×2160	3840×2160	60	404	6.7
Dinner Scene	10	YUV420p	YUV420p	4096×2160	3840×2160	60	360	6
Toddler Fountain	10	YUV420p	YUV420p	4096×2160	3840×2160	60	420	7
Narrator	10	YUV420p	YUV420p	4096×2160	3840×2160	60	300	5
Tunnel Flag	10	YUV420p	YUV420p	4096×2160	3840×2160	60	360	6
Ritual Dance	10	YUV420p	YUV420p	4096×2160	3840×2160	60	280	4.7
Discussion	10	YUV422p	YUV420p	3840×2160	3840×2160	60	380	6.3
Sea&Rock	10	YUV422p	YUV420p	3840×2160	3840×2160	60	272	4.5

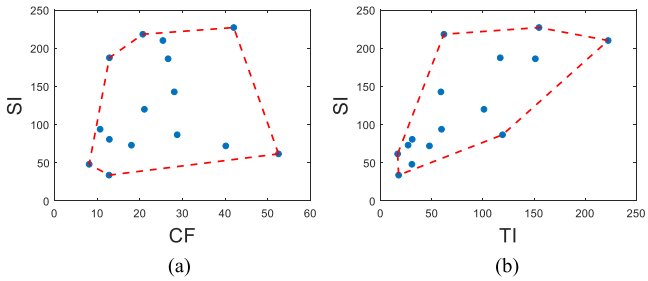

Fig. 2. Spatial information (SI) versus colorfulness (CF), and (b) temporal information (TI) versus colorfulness (CF) measured on the source contents of the ETRI-LIVE STSVQ database. The convex hull is indicated in red boundaries.

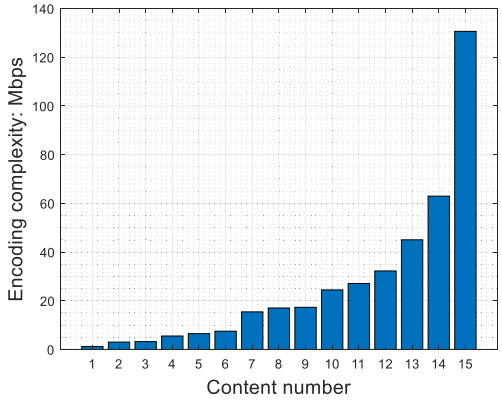

Fig. 3. Encoding complexity across contents, expressed in terms of Mbps when encoding using HEVC (libx265) at QP 29.

subsampled. Among the 15 source contents, five taken from the UVG dataset have frame rates of 120fps, while the other ten have frame rates of 60fps. Each video content was clipped to the range 5∼7 second duration, taking care to exclude scene changes or disruptions of content, such as a sport play or an actor speaking. The average duration of the video contents is 5.61 seconds.

The diversity of the source contents was confirmed by measuring the spans of (i) low-level space-time video features and (ii) encoding complexities. The low-level feature measurements included the spatial information (SI) and temporal information (TI) suggested in [14], representing the complexity of spatial details and temporal change of the videos, respectively. Another low-level feature that was used is the colorfulness (CF) measure proposed in [15]. Figs. 2(a) and (b) show plots of SI against CF and SI against TI with their corresponding convex hulls superimposed. The plots illustrate a diverse span of spatial and temporal characteristics covered by the source contents. We also computed the relative range and the uniformity of coverage [16] on each low-level feature, to quantify how well the feature space is covered by the selected source contents. The relative ranges of SI, TI, and CF were 0.85, 0.93, and 0.85, respectively, and the uniformity of coverage values for SI, TI, and CF were 0.84, 0.80, and 0.81, respectively. Again, the values illustrate the diverse space-time characteristics of the selected contents. We also considered content complexity as measured by encoded bitrate [17]. We encoded all of the source contents using HEVC (libx265) with a fixed quantization parameter (QP) of 29, then measured the bit rate of each content. As shown in Fig. 3, the source contents span a wide range of encoding complexities, ranging from less than 1Mbps to 130Mbps.

B. Distorted Video Generation

Each of the 15 source contents was subjected to various levels of distortion, in the form of space-time subsampling and compression. Since a main goal of our study is to understand the joint effects of space-time subsampling and compression, with an aim to improve perceptually optimal video coding strategies in practical settings, we constrained the videos used in the experiments to each approximate one of five "target" bit rates. These cover a range of perceived video qualities from

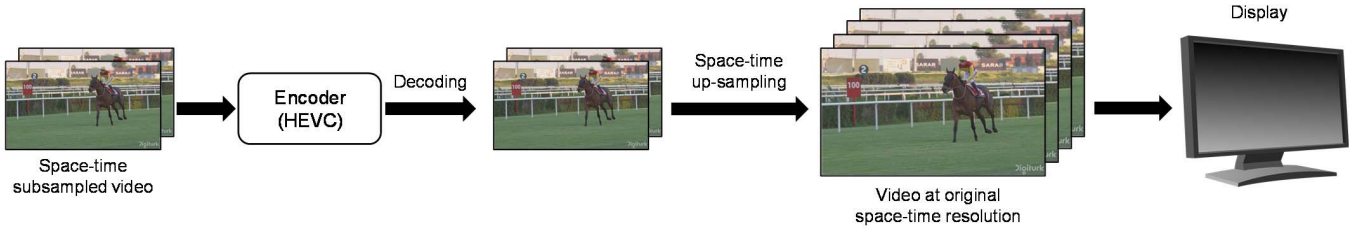

Fig. 4. Workflow for viewing space-time subsampled videos in the subjective experiment described in Section III.

very high to very low, while allowing for noticeable perceptual separations between bit rate levels. We then generated distorted videos having various combinations of space-time subsampling and degree of compression to approximately meet the determined bit rates. In this way, we generated 437 distorted videos affected by space-time subsampling and compression.

In the subjective study to be described shortly, all of the videos that were rated were viewed on a display supporting the video format of 3840 × 2160, 60/120 fps, YUV420p, and 10 bits. Hence, subsampled videos were up-sampled back to this format before being viewed. Fig. 4 shows the processing flow on space-time subsampled videos for our subjective experiment. As shown in the figure, space-time subsampled videos are restored to the original space-time resolutions before being viewed, thereby avoiding visual effects from the display's space-time up-sampling engines. Next, we explain how each distortion was applied to the source videos.

1) Spatial Subsampling: The database includes videos of four different spatial resolutions, including the source resolution (3840 × 2160) and three subsampled resolution (1920 × 1080, 1280 × 720, and 960 × 540). The videos were downsampled prior to encoding using the Lanczos kernel [18]. The spatially subsampled videos were then up-sampled back to the original spatial resolution (3840 × 2160), also using the Lanczos kernel, prior to displaying them.

2) Temporal Subsampling: The database contains original source videos that have frame rates of 120 or 60 fps, and temporally downsampled ("half frame rate") versions of them at 60 or 30 fps, respectively. The "full frame rate" videos were temporally downsampled to half frame rate, by simply dropping alternate frames, analogous to capturing the video at a lower shutter speed [19], without introducing motion blur.

However, when upsampling videos for viewing by the human subjects (Fig.4), we did not apply simple frame duplication, since this tends to produce visually unpleasant stuttering effects. We also did not rely on the frame rate interpolation engine of the display, since although it is designed to promote motion smoothness, it can produce severe and unexpected distortions. While Motion Compensated Interpolation (MCI) methods can deliver results having high visual quality, they can also introduce severe quality degradations when motion estimation failures occur [20], and they are major contributor to the "soap opera effect". Moreover, there are no agreed-upon best methods of MCI, which varies across display manufacturers. To avoid the severe distortions that can be produced by frame duplication or MCI, we instead utilized Linear Filter Interpolation [21], which linearly interpolates between adjacent frames. Generally, LFI yields stable and consistent results that may be regarded as a lower bound of the best results provided by modern displays, without producing the more severe artifacts that can occur.

3) Video Compression: The videos were compressed by the Main 10 profile of HEVC, using the FFmpeg libx265 encoder. We fixed the intra period to 1 second, to ensure that I-pictures would be regularly inserted as in the Random Access (RA) configuration of the reference software [22]. The compression levels were controlled by varying the QP parameters, where higher values of QP increased the degrees of compression.

The source videos exhibit different space-time characteristics, and consequently, the bitrates arrived at by processing each content with spatial and temporal downsampling and compression were varied by content to span a wide range of perceptual qualities. Instead of imposing the same bit rates on all contents, we determined a set of five bit rates on each source video so that a wide range of well-separated perceptual qualities were represented. Each source content was compressed at their full space-time resolutions using QP values ranging from 1 to 51. Among these, we visually selected five videos judged to fall into each of the following approximate quality levels resulting from compression:

- Level 1: Excellent quality video, perceptually difficult to distinguish from the uncompressed video.
- Level 2: Good quality video, having light compression-induced blur or flattening.
- Level 3: Fair quality video, having moderate blur and subtle but visible blocking artifacts.
- Level 4: Poor quality video, having noticeable levels of blur and blocking.
- Level 5: Bad quality video, with severe blur and blocking.

We did not use image/video quality models like PSNR, SSIM or VMAF to determine the compression qualities, since this would bias video selection relative to the applied quality model. Instead, our goal, as in many prior studies, was to produce a set of videos covering a wide range of well-separated perceptual quality levels. While the range of bitrates produced adequately encompasses most applications, we did not predetermine any target bitrate, since our goal was to model perceptual principles rather than any specific application. Thus, each source videos "target bitrates" were determined by this process.

Once these videos and their corresponding bit rates were determined for each source content, the space-time subsampled

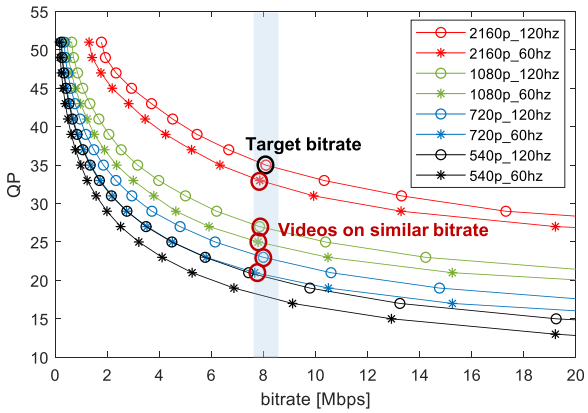

Fig. 5. Example of QP determination on space-time subsampled versions of the 'Jockey' video.

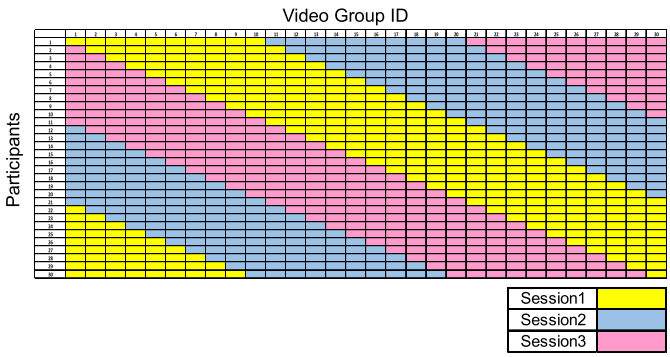

Fig. 6. Illustration of the round-robin method used to determine which video groups are presented to a given participant.

videos were then compressed (via QP selection) to have bit rates as similar as possible as these "target" bit rates. Fig. 5 shows an example of the QP determination process on space-time subsampled versions of the 'Jockey' sequence. The top curve indicates the source content having full space-time resolution, while the lower curves indicate various combinations of space-time subsampling. Once a bit rate is selected, the QP values of each space-time subsampled video was determined to be most similar to the "target" bitrate. Detailed listings of the bit rates associated with each source video, along with the corresponding QP and bit rates of the space-time subsampled videos, are also provided at the ETRI-LIVE STSVQ webpage (https://live.ece.utexas.edu/research/ETRI-LIVE_STSVQ/index.html).

III. Subjective Experiments

A. Experiment Design

In subjective experiments, we adopted a Single-Stimulus Continuous Quality Evaluation (SSCQE) procedure with hidden reference [23]. The participants delivered the subjective quality scores using a continuous scale score bar after viewing the video once. The original reference videos are presented but 'hidden', i.e., without being identified as such. The scores on reference videos are useful as high-quality anchors, and are used to calculate difference mean opinion scores (DMOS) as a way of removing content biases.

Given that the average duration of each presented video is 5.61 seconds, and the average time subjects expend scoring each video is about 6 seconds, a participant requires approximately 90 minutes to view and score all of the 437 videos in the database. To avoid viewer fatigue, we therefore divided the study into three 30-minute sessions, each comprising 145 or 146 distorted videos and 15 hidden references. The subjects each participated in three sessions separated by least 24 hours, hence every subject evaluated all of the videos in the database.

When constructing the playlist of videos for each session, we sought to eliminate any biases introduced if the videos were displayed in the same order to every subject. We also avoided the successive presentations of same contents. Therefore, to minimize contextual and memory effects, which can affect judgements of video quality [17], we randomized the playlist using the following procedures.

1) Initial Randomization: An initial randomization was used to reduce any clustering of the source contents. The list was shuffled repeatedly until there were less than 10 occurrences of any video content (distorted or not) that appeared consecutively. Of course, this does not adequately remove the effects of memory, so additional randomization was applied in the final step (to follow).

2) Video Groups: The randomized 437 videos from the previous steps were then divided into 30 'video groups', each containing 14 or 15 videos. Since each subject participated in three sessions, 10 video groups were viewed in each session.

3) Round-Robin Ordering: We also employed a round-robin presentation ordering, to minimize the possibility of subjective opinions being affected by contextual factors, such as combinations of videos being shown together to all of the participants. Fig. 6 illustrates how the round-robin method was applied to decide which video groups were presented to each participant within each session. As shown in the figure, the video groups comprising sessions 1, 2, and 3 will be different for all participants. This round-robin approach also guarantees each video group will appear an equal number of times within each session (across all subjects). For example, in a study with 30 subjects, the videos in group 1 will appear 10 times in each of sessions 1, 2 and 3.

4) Final Randomization: After the 10 video groups were selected for a given participant in a given session, the 15 undistorted hidden references were included, and the entire collection of references and 10 video groups were collectively randomly shuffled again into a single session playlist. However, during the shuffling, videos having the same content were constrained to have at least three different contents lie between them, i.e., to be separated by at least four display periods. Proceeding in exactly this way throughout, we generated distinct playlists for every session created throughout the study.

B. Experimental Set-up

As explained previously, all of the videos included in the study are stored in 10-bit raw YUV format. To be able to successfully play out the UHD high frame rate videos without

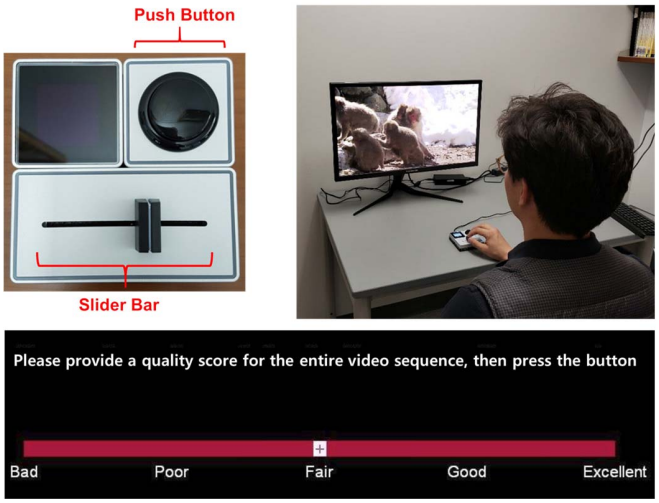

Fig. 7. Hardware slider interface (top) and video score voting screen (bottom).

hitches, we relied on a powerful ClearView system provided by Video Clarity [24], which enables real-time playout of raw 4K 10-bit YUV format videos at 120fps. The system is connected to a 27-inch Acer Predator X27 display which also supports true 120fps 10-bit video signal input and playout [25]. Before each subject entered the study room, the entire experimental dataset and script was preloaded to present the session using the playlist generated for that session. After viewing each test video, a voting screen appeared and the subject used a Pallete hardware slider [26] to control and select the scores from an onscreen slider bar, as shown in Fig. 7. The quality bar is marked with five Likert labels ranging from 'Bad' to 'Excellent.' However, the quality scale is continuous, and the subjects were instructed that they could move the slider bar to any position between the labels. Once the score was selected, it was converted to a numerical value ranging from 0 to 39, where 0 is 'Bad' and 39 is 'Excellent.'

C. Experimental Procedure

Each subject was presented with a brief oral summary of the overall experiment, and given written instructions explaining how to use the hardware slider to assign scores to each video, and that scores should reflect the degree of satisfaction they felt regarding the level of overall video quality, while discounting the aesthetic value or interestingness of the content. Each subject was seated in front of the display at a distance of about 1.5 times the height of the display, as recommended in [27] for 4K videos. The subjects then participated in a short training session using 10 videos different from those viewed in the actual study, but also covering a wide range of perceptual qualities and distortions representation of those seen during the actual experiment. The training session enabled the subjects to become familiar with the types and qualities of videos to be judged, to attain facility with the hardware interface used to score them. Following the training session, each subject proceeded immediately to the actual session where the subjective data was collected. The training session was only given prior to each subject's first session.

IV. Data Processing and Analysis

A. Processing of Subjective Scores

A total of 34 naïve subjects from The University of Texas at Austin took part in the study. Each subject participated in all three sessions, and thus, each video was rated by all 34 subjects. The collected scores were converted to DMOS according to [28]. Let s_{ijk} refer to the score given by subject i to video j during session $k \in \{1, 2, 3\}$. Then, the difference score d_{ijk} for video j at session k is computed by subtracting the score given to the video from that given to the corresponding (same content) reference video j_{ref} *in the same session* k

$$d_{ijk} = s_{ij_{ref}k} - s_{ijk}. \tag{1}$$

The difference scores for the reference videos are, of course, 0, and were then removed from all sessions. To normalize the scores collected on different sessions, we computed the Z-score per session [29] as:

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} d_{ijk}, \tag{2}$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik} - 1} \sum_{j=1}^{N_{ik}} \left(d_{ijk} - \mu_{ijk}\right)^2}, \tag{3}$$

and

$$z_{ijk} = \frac{d_{ijk} - \mu_{ijk}}{\sigma_{ik}}, \tag{4}$$

where N_{ik} is the number of videos the subject i viewed during session k. We collected the Z-scores from all sessions and formed a matrix z_{ij} with element z_{ij} corresponding to the Z-score assigned by subject i to video j, where $j \in \{1, 2, \ldots, 437\}$.

Subject rejection was performed according to the procedure from [23]. The normality of the Z-scores for each content was evaluated by computing the kurtosis β_2 via

$$\bar{z}_j = \frac{1}{N} \sum_{i=1}^{N} z_{ij}, \tag{5}$$

$$m_{xj} = \frac{\sum_{i=1}^{N} \left(z_{ij} - \bar{z}_j\right)^x}{N}, \tag{6}$$

and

$$\beta_{2j} = \frac{m_{4j}}{\left(m_{2j}\right)^2}, \tag{7}$$

where N refers to the number of subjects that evaluated video j, which in our case, is 34. If $2 \leq \beta_{2j} \leq 4$, we considered the scores for the video j to be normally distributed. We identified potential outlier subjects according to the predicate

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 $Q_i = Q_i + 1$,

where

$$\sigma_j = \sqrt{\sum_{i=1}^{N} \frac{\left(z_{ij} - \bar{z}_j\right)^2}{N - 1}}. \tag{8}$$

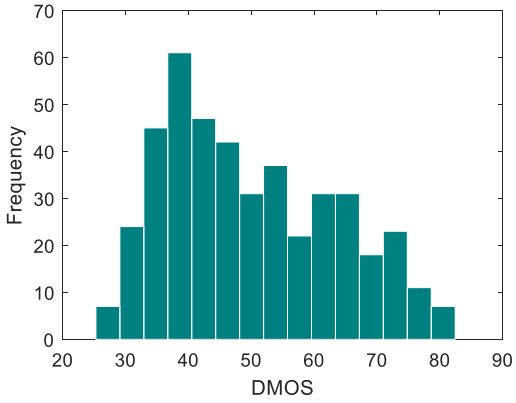

Fig. 8. Histogram of DMOS in 15 equally spaced bins.

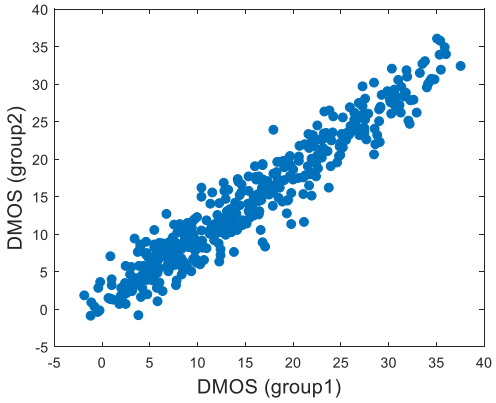

Fig. 9. Scatter plot of DMOS of a random division of the human subjects into two non-overlapping groups of equal size. The plot indicates a strong agreement between the subjects.

If β_{2j} did not fall between 2 and 4, we instead used

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 $Q_i = Q_i + 1$.

In either case, for each subject i, we determined if the following conditions:

$$\frac{P_i + Q_i}{N} > 0.05, \tag{9}$$

and

$$\left| \frac{P_i - Q_i}{P_i + Q_i} \right| < 0.3 \tag{10}$$

were met. For a certain subject i, if both (9) and (10) were true, then we rejected the subject. Overall, only four out of the 34 subjects were rejected. We linearly rescaled the Z-scores of the remaining 30 subjects to final DMOS values in the range [0, 100] using

$$z'_{ij} = \frac{100(z_{ij} + 3)}{6}. \tag{11}$$

Fig. 8 shows the histogram of the resulting DMOS values, showing a distribution of opinions. The extreme DMOS values were 25.30 and 82.45, while the mean and standard deviation of DMOS were found to be 50 and 13.62, respectively.

To check the consistency of the collected subjective data, we randomly split the subjects into two randomly selected, non-overlapping groups of equal size, and measured the Spearman's Rank Correlation Coefficient (SRCC) between their scores. Fig. 9 shows a scatter plot of DMOS between a pair of such randomly split groups, exhibiting an approximately linear unit slope. The SRCC between the subject halves for this split was 0.958, indicating a high consistency between the groups. This random procedure was repeated 1000 times, yielding SRCC values lying between 0.941 and 0.972 with a median value of 0.960, indicating a reliably high degree of inter-subject consistency.

B. Analysis of Opinion Scores

We observed how the average DMOS values varied with content across various levels of distortions. Fig. 10(a) plots the average DMOS values of each content across the bit rates. The labels denoted Lv. refer to distortion levels, which in Fig. 10(a) refers to H.265 compression. More specifically, Lv. 1 refers to videos that were compressed to the highest bit rate, while Lv. 5 indicates the highest degree of compression (lowest bit rate). The target bit rate decreases with increases of the compression level. Generally, DMOS increased with increased compression, although, as expected, the relationship is not monotonic because of content effects (e.g., masking).

When observing the perceptual effects of different levels of spatial and/or temporal subsampling, consideration must be given to an assumed available bit rate. This is because the spatial and/or temporal subsampling can drive the perceptual quality of a video in different directions depending on an imposed bit budget. For example, videos subjected to little or no space-time subsampling may yield very high levels of perceived quality given a large enough budget. However, the quality may be severely degraded if the budget is small (hence compression is heavy). Figs. 10(b)-(e) depict the effects of different levels of spatial and temporal subsampling of the videos, conditioned on low and high bit rates, where the low bit rates were taken to be levels 4 and 5, and levels 1-3 were regarded as high bit rates.

Figs. 10(b) and (c) plots the average DMOS across full and half frame rates. Fig. 10(c) shows that, at sufficiently high bit rates, the full frame rate videos, generally, yielded lower DMOS (higher perceptual quality) probably in large part because of smoother motion. However, as shown in Fig. 10(b), when the available bit budget was low, the tendency was reversed, and the half frame rate videos resulted in better reported quality than the full frame rate videos. This is because videos at full frame rate require heavy compression to meet a given low bit budget, which will introduce severe compression artifacts that significantly degrade perceptual quality. The videos at half frame rate, on the other hand, may involve loss of some temporal information, possibly resulting in less smooth motion or stutter artifacts, but they also require much less compression to meet the bit budget. Because of this, videos at half frame rate can present better overall perceptual experiences than full frame rate videos having severe compression artifacts.

Figs. 10(d) and (e) plot the DMOS across different spatial resolutions. A similar trend may be observed,

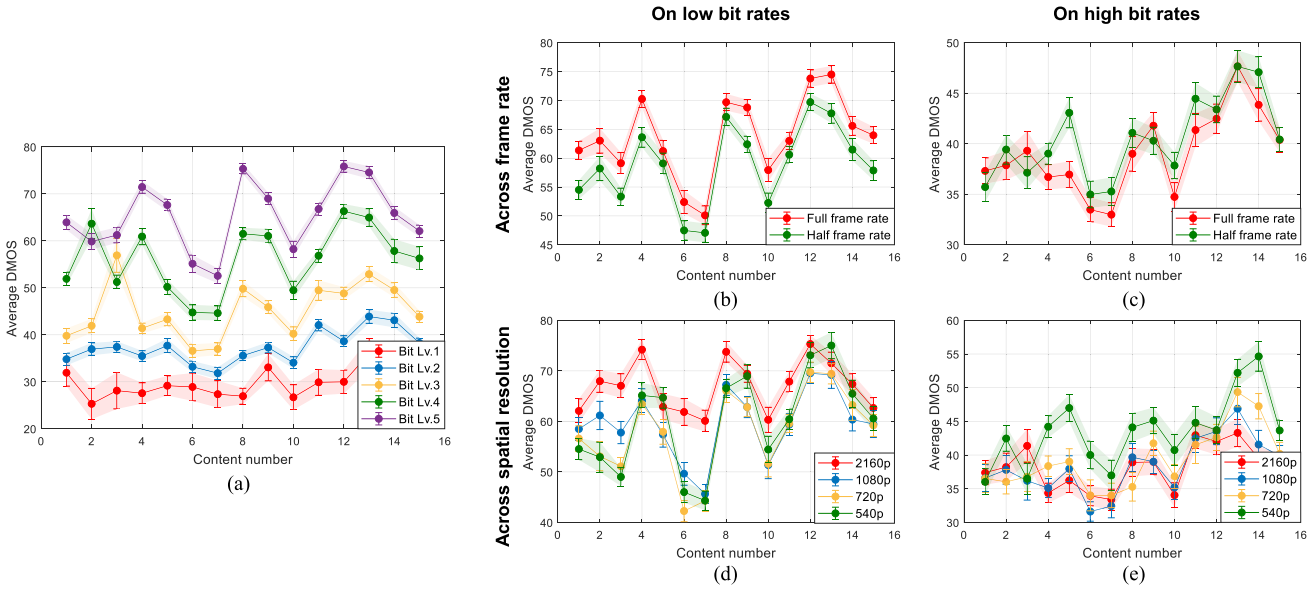

Fig. 10. Variation of average DMOS of each content for each (a) compression level, (b) temporal subsampling level at low bit rates, (c) temporal subsampling level at high bit rates, (d) spatial subsampling level at low bit rates, and (e) spatial subsampling level at high bit rates. The error bars represent 95% confidence intervals.

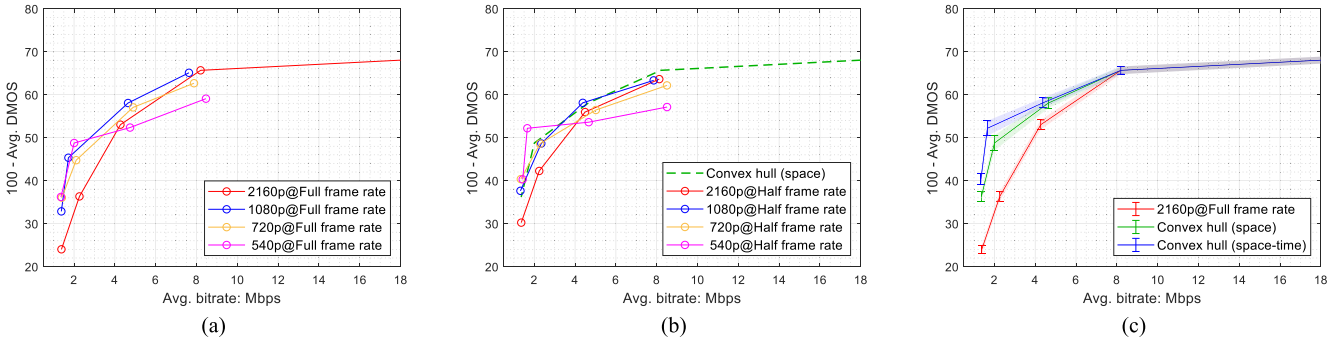

Fig. 11. Rate distortion curves plotted on the entire database to observe the effects of (a) spatial subsampling only, and (b) space-time subsampling. Plot (c) depicts the rate distortion curves of the video at the original space-time resolutions, using the convex hull constructed via spatial subsampling, and the convex hull constructed via space-time subsampling. The error bars represent 95% confidence intervals.

e.g., in Fig. 10(e), low perceptual quality (higher DMOS) was reported on heavily subsampled 540p videos. However, as shown in Fig. 10(d), at low bit rates, the tendency again reversed, and the 540p videos generally provided better perceptual qualities as compared to the full resolution (2160p) videos. These interesting results nicely exemplify the somewhat complex relationships between spatial and temporal resolution, compression, and perceived quality, and provide evidence that it should be possible to perceptually optimize video coding strategies by considering spatial and/or temporal subsampling combined with compression, especially when trying to attain lower bit rates.

Fig. 11 plots rate distortion (RD) curves on the entire ETRI-LIVE database, further revealing the effects of space-time subsampling. The bit rate and DMOS values of each point on a curve corresponds to the average DMOS of all videos in the database having the same space-time subsampling configuration, as specified in the legend. Note that the vertical axis is set $100 - Avg.DMOS$, hence higher values correspond to higher perceptual quality.

Fig. 11(a) focuses on the effects of spatial subsampling on the RD curve. As the legend of Fig. 11(a) indicates, the considered videos were at full frame rate while the spatial resolutions were varied. The plot reveals the interesting tendency that lower spatial resolution videos are favored over the higher resolution videos as the bit rate is reduced (increased compression). This is not unexpected, since retaining full video spatial dimension is not the best strategy at very low bit rates, as heavy compression artifacts are introduced. While spatial subsampling results in a loss of information and degradations of quality, much less compression is required to meet the bit budget, yielding less perceptually degraded videos. It is possible to construct a convex hull that can be used to help choose the best spatial subsampling strategy at each bit rate to maintain the best possible perceptual quality. We indicate such a spatial convex hull using a green dashed curve, in Fig. 11(b).

Fig. 11(b) also considers the effects of temporal subsampling. The videos were subsampled in both space and time. As shown in the figure, the perceptual quality at low bit rates can be further improved using simultaneous spatial and

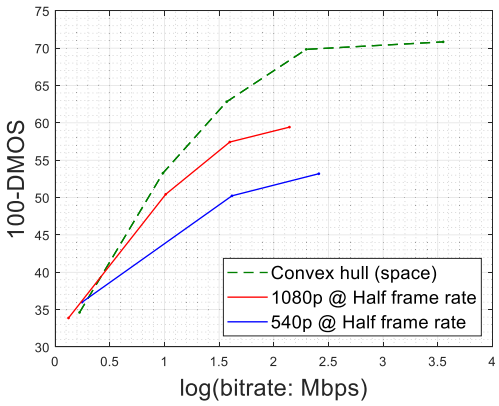

Fig. 12. Rate distortion curves of the video content 'American Football' at half frame rate, compared against the convex hull constructed using spatial subsampling.

temporal subsampling, as compared to the convex hull constructed from just spatial subsampling. While this additional temporal subsampling may introduce some level of visual degradation arising from loss of temporal information (such as stutter) it significantly reduces the amount of applied compression, thereby improving the overall perceptual quality. Fig. 11(c) depicts comprehensive RD curves obtained at original space-time resolutions, from the convex hull constructed using spatial subsampling only, and from the convex hull constructed using both space and time subsampling, along with 95% confidence intervals superimposed. It is easily observed that perceptual quality can be significantly improved at low bit rates by using space and/or time subsampling. Indeed, statistically superior quality improvements are obtained by considering both space and time subsampling, as opposed to considering only one kind of subsampling.

Of course, the aforementioned observations are comprehensive on the whole database, and the results may vary depending on the content characteristics. For example, Fig. 12 shows the RD curve of the 'American Football' sequence, which contains significant and rapid motions, including those arising from camera movements and from the action of football players. In this kind of video, the impact of temporal information loss can be much more significant than on more static contents. As shown in Fig. 12, the convex hull constructed from only spatial subsampling on this high-motion video produced better subjective quality results as compared to deploying any temporal subsampling. An optimal strategy for this content would likely employ only spatial subsampling, even at low bit rates. Understanding the effects of space-time subsampling as a function of content characteristics, and devising perceptually optimal strategies are among the topics that could be fruitfully investigated by analyzing the rich information available in the new database.

V. Evaluation of Objective Video Quality Models

As a way of both demonstrating and exploiting the usefulness of the new ETRI-LIVE STSVQ database we evaluated and compared the performances of a variety of relevant and widely-used image/video quality assessment models on it. We studied both Reference (including full and reduced reference) and No-reference models. The former assume that at least some information is derived from an undistorted reference image/video to compare against, while the latter predict image/video quality without using any reference information, which is often not available.

The reference models that we evaluated include the image quality assessment (IQA) models such as PSNR, SSIM [30], MSSSIM [31], and VIF [32], computed on each frame yielding predictions that were averaged (pooled) over all frames to obtain overall video quality scores. We also considered video quality assessment (VQA) models that utilize both spatial and temporal video features. Among these, ST-RRED [33] and SpEED [34] are natural scene statistics (NSS) based models that measure the statistical space-time statistical deviations between a distorted videos and their references. VMAF [35] is a learning-based VQA model that fuses a set of quality-aware video features using a Support Vector Regressor (SVR). We also included a popular deep learning-based quality model called LPIPS [36], which utilizes the intermediate coefficients of a deep learning-based classification network to capture information relevant to perceptual similarity. LPIPS calibrates the AlexNet based classification network by adding a linear layer on top, to measure the perceptual similarities between a reference and a distorted video. Finally, we also include a results of a recent prototype model we have developed [37], which is a space-time NSS based model that is based on statistical measurements of Video's Space-Time Regularity (VSTR). The prediction performance of the reference models is evaluated by comparing them to DMOS.

The no-reference models that we evaluated include BRISQUE [38] and NIQE [39], which are NSS-based, and TLVQM [40], which is a learning-based model that uses a large set of hand-designed features, including motion statistics and estimates of specific distortions. Since no-reference methods evaluate the intrinsic quality of videos, we evaluated no-reference model performance against MOS. MOS computation is similar to that for DMOS in Section IV-A, with (1) modified to read $d_{ijk} = s_{ijk}$.

The quality prediction performances of the compared models was evaluated using Spearman's rank order correlation coefficient (SRCC), the Kendall rank correlation coefficient (KRCC), the Pearson linear correlation coefficient (PLCC), and the root mean squared error (RMSE). SRCC and KRCC measure ordinal correlations, while PLCC measures linear correlation between variables. Higher values are favorable for SRCC, PLCC, and KRCC, and lower values are favorable for RMSE. Before computing PLCC and RMSE, we used logistic regression to linearize the model prediction following the procedure in [23].

A. Overall Performance Comparison

We evaluated the prediction performance of the compared models on the new ETRI-LIVE STSVQ database. For models that involve machine learning, we first tested the model as pre-trained on an other database, or used one of their representative features. In this way, we measured holistic model performance over all 437 videos of our database

TABLE II

PERFORMANCE COMPARISON OF VQA/IQA MODELS ON THE ETRI-LIVE STSVQ DATABASE ACROSS DIFFERENT TEMPORAL SUBSAMPLING LEVELS. THE TWO BEST MODELS IN EACH COLUMN ARE BOLDFACED

Model		Full frame rate				Half frame rate				Overall (full + half frame rate)			
		SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
Reference	PSNR	0.57	0.41	0.57	13.09	0.33	0.23	0.28	11.82	0.42	0.29	0.40	12.49
	SSIM [30]	0.76	0.55	0.68	14.16	0.30	0.21	0.17	12.07	0.50	0.34	0.24	13.19
	MSSSIM [31]	0.65	0.46	0.61	13.65	0.29	0.21	0.23	12.02	0.44	0.30	0.32	12.89
	VIF [32]	0.68	0.49	0.67	12.33	**0.39**	**0.28**	0.33	11.81	0.52	0.36	0.46	12.08
	ST-RRED [33]	0.76	0.55	0.64	14.55	0.32	0.23	0.18	11.88	0.49	0.34	0.20	13.34
	SpEED [34]	**0.82**	**0.62**	0.68	14.74	0.25	0.17	0.15	11.86	0.46	0.33	0.16	13.43
	VMAF [35]	0.74	0.56	**0.74**	**11.18**	**0.50**	**0.37**	**0.50**	**10.90**	**0.59**	**0.43**	**0.58**	**11.05**
	LPIPS [36]	0.38	0.27	0.35	14.07	0.31	0.22	0.27	**11.51**	0.34	0.24	0.32	12.90
	VSTR-ED [37]	**0.88**	**0.70**	**0.88**	**11.92**	0.36	0.25	**0.35**	12.14	**0.54**	**0.39**	**0.47**	**12.03**
No Reference	BRISQUE [38]	0.39	0.27	0.37	14.49	0.30	0.22	0.29	11.70	0.35	0.24	0.34	13.23
	NIQE [39]	0.29	0.20	0.26	15.09	0.20	0.14	0.13	12.19	0.25	0.17	0.21	13.77
	TLVQM-HCF [40]	0.30	0.27	0.27	15.03	0.25	0.17	0.23	11.92	0.27	0.18	0.25	13.63

without applying train-test split procedures. Of course, we also obtained cross-validation results of all learned models trained on our database, in the following section. For VMAF, we used the vmaf_4k_v.0.6.1 model available at [41]. For LPIPS implementation, we used version 0.1 with pre-trained AlexNet weights available at [42]. For VSTR, we used one of its representative features, an entropic difference (ED) feature computed on space-time displaced frame differences. For BRISQUE, we used the model trained on the LIVE image database provided by the authors. For TLVQM, we used the average of thirty high complexity features (HCF) to capture a wide variety of distortions.

Table II reports the performances of the compared models over different temporal subsampling levels and overall. As shown in the Table, the no-reference models performed much worse than the reference models. It is interesting that the references models attained very high performance on full frame rate videos, which they have been amply validated on in the past, but performed much worse on temporally subsampled videos. This suggests that there is ample room for improvement of solutions to the underdeveloped topic of assessing the quality of temporally subsampled and compressed videos. Overall, VIF, VMAF and VSTR-ED yielded the highest prediction performances. Fig. 13 shows scatter plots of the model predictions against the subjective opinion scores. The orange data points refer to videos that were temporally subsampled to half frame rate, while the blue data points refer to full frame rate videos. One notable tendency is that for no-reference (NR) models, the orange and blue points tend to significantly overlap, while the reference models result in less overlap and somewhat parallel distributions of blue and orange points. This is because half frame rate videos have in-between frames that are generated by LFI, and the reference models explicitly compare those frames with original frames, which can introduce discrepancies. This may lead the models to underestimate quality, sending the orange points to the left side. The amount of underestimation may vary depending on the content characteristics, whereby videos with high motion may suffer more than static videos. The separation of blue and orange points is less apparent for reference models that employ a feature fusion stage. For instance, VMAF and LPIPS may also extract features that underestimate the quality of half frame rate videos. However, since the models fuse multiple features using a pre-trained regressor or multi-layer perceptron, it penalizes the half frame rate videos less. On the other hand, reference models that rely on a single feature or that use simple averaging of features may result in a separation of orange and blue points, as shown for ST-RRED, SpEED, and VSTR-ED in Fig. 13. Of course, the underestimation effect on half frame rate videos can be alleviated by extracting richer sets of features and by fusing them into a video quality model. We explore this in more detail on Section V.C, where we also show how the prediction power of VSTR can be improved relative to the single feature VSTR-ED model through multiple feature fusion.

Table III tabulates the prediction performance against amount of spatial subsampling. The no-reference models again underperformed against the reference models. However, unlike the results in Table II, similar prediction performance was obtained across spatial resolutions. VMAF yielded good performance across all resolutions, VIF delivered good performance at lower spatial resolutions (540p and 720p), and VSTR-ED delivered good performance at higher spatial resolutions (1080p and 2160p).

B. Statistical Significance

We verified the statistical significance of the performance differences among the compared models in Tables II and III via an F-test. Table IV shows the F-test results performed on the residuals between the model predictions and the subjective opinion scores. The underlying assumption is that the distribution of residuals follows a zero mean Gaussian distribution. The F-test evaluates the ratio of variances of the residuals, and determines whether the variances are equal at the 95% confidence level. Table IV contains 7 entries, corresponding to half frame rate, full frame rate, 540p, 720p, 1080p, 2160p, and all of the videos, in that order. As shown in the Table, VIF, VMAF and VSTR-ED attained statistically superior prediction performances as compared to the other methods.

C. Cross-Validation Results on Learning-Based Models

We also evaluated the cross-validation performances of the learning-based models trained specifically on our database.

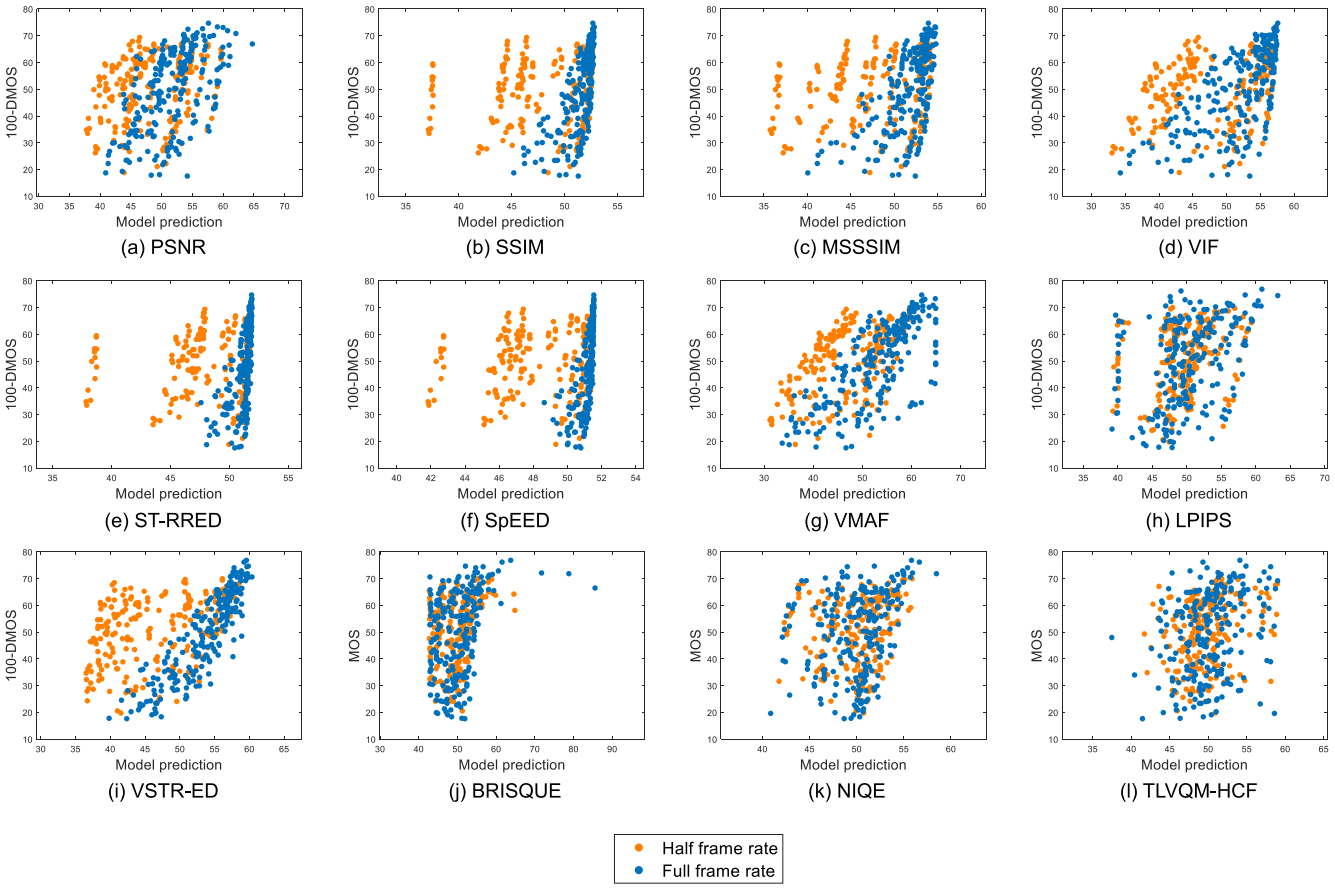

Fig. 13. Scatter plots of VQA/IQA model predictions plotted against subjective opinion scores of all of the distorted videos in the ETRI-LIVE STSVQ database. The orange data points refer to videos that were temporally subsampled to half frame rate, while the blue data points refer to full frame rate videos.

TABLE III

PERFORMANCE COMPARISON OF VQA/IQA MODELS ON THE ETRI-LIVE STSVQ DATABASE ACROSS DIFFERENT SPATIAL SUBSAMPLING LEVELS. THE TWO BEST MODELS IN EACH COLUMN ARE BOLDFACED

Model		540p		720p		1080p		2160p		Overall	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Reference	PSNR	0.45	0.43	0.39	0.37	0.36	0.32	0.49	0.46	0.42	0.40
	SSIM [30]	0.51	0.29	0.46	0.22	0.44	0.18	0.58	0.30	0.50	0.24
	MSSSIM [31]	0.46	0.38	0.40	0.31	0.38	0.25	0.53	0.36	0.44	0.32
	VIF [32]	**0.56**	**0.52**	**0.51**	**0.44**	0.45	0.38	0.59	0.51	0.52	0.46
	ST-RRED [33]	0.47	0.26	0.43	0.17	0.44	0.15	0.59	0.23	0.49	0.20
	SpEED [34]	0.43	0.21	0.41	0.13	0.41	0.11	0.57	0.20	0.46	0.16
	VMAF [35]	**0.52**	**0.52**	**0.56**	**0.55**	**0.54**	**0.52**	**0.69**	**0.67**	**0.59**	**0.58**
	LPIPS [36]	0.17	0.12	0.29	0.25	0.34	0.31	0.47	0.44	0.34	0.32
	VSTR-ED [37]	0.48	0.43	0.49	0.40	**0.49**	**0.40**	**0.65**	**0.58**	**0.54**	**0.47**
No Reference	BRISQUE [38]	0.24	0.20	0.27	0.24	0.31	0.30	0.49	0.46	0.35	0.34
	NIQE [39]	0.21	0.08	0.13	0.04	0.21	0.14	0.38	0.37	0.25	0.21
	TLVQM-HCF [40]	0.12	0.13	0.26	0.26	0.34	0.30	0.32	0.31	0.27	0.25

We split the database into non-overlapping train and test sets. The learning-based models learned to map features to video quality by training on the train set, and their prediction performances were evaluated on the test set. While the prediction performances here don't represent the holistic performance on all 437 videos in the database, cross-validation results are important for understanding how well learning-based models can generalize to unseen test data, which is especially important in practical application.

The trained models include VMAF, VSTR, BRISQUE, and TLVQM which each consists of 6, 16, 36 and 75 features, respectively. In addition, we also evaluated the VIDEVAL model [43], which uses a feature ensemble and selection procedure to extract 60 features from among 763 perceptual features. The VIDEVAL is a highly optimized feature-fusion based method that is known to deliver robust state-of-the-art performances on many video quality databases. The features from the models were pre-processed by min-max normalization and were then used to train an SVR with a radial basis function (RBF) kernel. The SVR-RBF parameters were determined using cross-validation within the training set, as in [44].

TABLE IV

RESULT OF F-TEST ON RESIDUALS OF MODEL PREDICTION AND OPINION SCORES AT 95% CONFIDENCE. EACH CELL CONTAINS 7 ENTRIES CORRESPONDING TO THE HALF FRAME RATE, FULL FRAME RATE, 540P, 720P, 1080P, 2160P AND ALL VIDEOS. A SYMBOL '-' INDICATES STATISTICAL EQUIVALENCE BETWEEN THE ROW AND THE COLUMN. A VALUE '1' INDICATES THE ROW HAVING LESS RESIDUAL VARIANCE (BETTER QUALITY PREDICTION) THAN THE COLUMN. A VALUE '0' INDICATES COLUMN HAVING LESS RESIDUAL VARIANCE THAN THE ROW

	PSNR	SSIM	MSSSIM	VIF	ST-RRED	SpEED	VMAF	LPIPS	VSTR-ED	BRISQUE	NIQE	TLVQM-HCF
PSNR	-------	-------	-------	-------	-------	-------	-0----0	-------	-0-----	-------	-1----1	-1-----
SSIM	-------	-------	-------	-0-----	-------	-------	00---00	-------	-0-----	-------	-------	-------
MSSSIM	-------	-------	-------	-------	-------	-------	-0---00	-------	-0-----	-------	-------	-------
VIF	-------	-1-----	-------	-------	-1----1	-1----1	-------	-1-----	-------	-1-----	-1----1	-1----1
ST-RRED	-------	-------	-------	-0----0	-------	-------	-0---00	-------	-0----0	-------	-------	-------
SpEED	-------	-------	-------	-0----0	-------	-------	-0---00	-------	-0----0	-------	-------	-------
VMAF	-1----1	11---11	-1---11	-------	-1---11	-1---11	-------	-1---11	-------	-1---11	11-1-11	11---11
LPIPS	-------	-------	-------	-0-----	-------	-------	-0---00	-------	-0-----	-------	-------	-------
VSTR-ED	-1-----	-1-----	-1-----	-------	-1----1	-1----1	-------	-1-----	-------	-1----1	-1---11	-1---11
BRISQUE	-------	-------	-------	-0-----	-------	-------	-0---00	-------	-0----0	-------	-------	-------
NIQE	-0----0	-------	-------	-0----0	-------	-------	00-0-00	-------	-0---00	-------	-------	-------
TLVQM-HCF	-0-----	-------	-------	-0----0	-------	-------	00---00	-------	-0---00	-------	-------	-------

TABLE V

CROSS-VALIDATION PERFORMANCE COMPARISON OF VQA/IQA MODELS ON THE ETRI-LIVE STSVQ DATABASE ACROSS DIFFERENT TEMPORAL SUBSAMPLING LEVELS. THE NUMBERS DENOTE MEDIAN VALUES OVER 1000 ITERATIONS OF RANDOMLY SPLIT TRAIN AND TEST SETS. THE VALUES INSIDE THE BRACKETS DENOTE STANDARD DEVIATIONS. THE TWO BEST MODELS IN EACH COLUMN ARE BOLDFACED

Model		Full frame rate		Half frame rate		Overall (full + half frame rate)	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Reference	PSNR	0.68 (0.15)	0.65 (0.15)	0.51 (0.18)	0.40 (0.22)	0.52 (0.13)	0.48 (0.15)
	SSIM [30]	0.81 (0.09)	0.76 (0.14)	0.47 (0.19)	0.27 (0.23)	0.53 (0.14)	0.33 (0.22)
	MSSSIM [31]	0.75 (0.14)	0.71 (0.14)	0.47 (0.18)	0.32 (0.21)	0.53 (0.13)	0.38 (0.19)
	VIF [32]	0.76 (0.12)	0.73 (0.14)	0.53 (0.18)	0.46 (0.24)	0.60 (0.13)	0.54 (0.17)
	ST-RRED [33]	0.83 (0.12)	0.74 (0.17)	0.47 (0.19)	0.27 (0.23)	0.52 (0.15)	0.27 (0.20)
	SpEED [34]	**0.87 (0.08)**	**0.77 (0.14)**	0.44 (0.19)	0.25 (0.20)	0.49 (0.12)	0.22 (0.17)
	VMAF [35]	0.78 (0.19)	0.75 (0.20)	**0.58 (0.21)**	**0.55 (0.24)**	**0.65 (0.18)**	**0.63 (0.19)**
	LPIPS [36]	0.54 (0.21)	0.56 (0.25)	0.46 (0.22)	0.51 (0.26)	0.49 (0.21)	0.53 (0.25)
	VSTR [37]	**0.88 (0.08)**	**0.88 (0.08)**	**0.65 (0.16)**	**0.64 (0.18)**	**0.75 (0.10)**	**0.73 (0.10)**
No Reference	BRISQUE [38]	0.37 (0.1981)	0.36 (0.19)	0.28 (0.21)	0.26 (0.20)	0.31 (0.20)	0.31 (0.19)
	NIQE [39]	0.40 (0.1957)	0.42 (0.22)	0.35 (0.20)	0.32 (0.22)	0.37 (0.20)	0.38 (0.22)
	TLVQM [40]	0.43 (0.1749)	0.40 (0.19)	0.25 (0.17)	0.22 (0.18)	0.32 (0.17)	0.30 (0.17)
	VIDEVAL [43]	0.35 (0.1889)	0.32 (0.19)	0.30 (0.20)	0.28 (0.20)	0.33 (0.19)	0.30 (0.19)

TABLE VI

CROSS-VALIDATION PERFORMANCE COMPARISON OF VQA/IQA MODELS ON THE ETRI-LIVE STSVQ DATABASE ACROSS DIFFERENT SPATIAL SUBSAMPLING LEVELS. THE NUMBERS DENOTE MEDIAN VALUES OVER 1000 ITERATIONS OF RANDOMLY SPLIT TRAIN AND TEST SETS. THE VALUES INSIDE THE BRACKETS DENOTE STANDARD DEVIATIONS. THE TWO BEST MODELS IN EACH COLUMN ARE BOLDFACED

Model		540p		720p		1080p		2160p		Overall	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Reference	PSNR	0.51 (0.17)	0.47 (0.19)	0.48 (0.18)	0.44 (0.20)	0.48 (0.17)	0.41 (0.18)	0.59 (0.13)	0.53 (0.14)	0.52 (0.13)	0.48 (0.15)
	SSIM [30]	0.54 (0.18)	0.37 (0.23)	0.50 (0.16)	0.29 (0.23)	0.50 (0.17)	0.25 (0.23)	0.64 (0.13)	0.40 (0.21)	0.53 (0.14)	0.33 (0.22)
	MSSSIM [31]	0.53 (0.17)	0.42 (0.21)	0.49 (0.17)	0.36 (0.22)	0.47 (0.16)	0.31 (0.20)	0.61 (0.12)	0.43 (0.18)	0.53 (0.13)	0.38 (0.19)
	VIF [32]	**0.63 (0.17)**	0.59 (0.20)	0.60 (0.17)	0.53 (0.22)	0.56 (0.17)	0.47 (0.21)	0.68 (0.12)	0.58 (0.16)	0.60 (0.13)	0.54 (0.17)
	ST-RRED [33]	0.52 (0.18)	0.31 (0.19)	0.47 (0.16)	0.25 (0.20)	0.49 (0.17)	0.23 (0.21)	0.63 (0.14)	0.36 (0.21)	0.52 (0.15)	0.27 (0.20)
	SpEED [34]	0.46 (0.17)	0.26 (0.17)	0.43 (0.15)	0.17 (0.17)	0.45 (0.16)	0.17 (0.17)	0.60 (0.10)	0.27 (0.17)	0.49 (0.12)	0.22 (0.17)
	VMAF [35]	0.60 (0.22)	**0.60 (0.23)**	**0.64 (0.22)**	**0.61 (0.22)**	**0.61 (0.19)**	**0.58 (0.21)**	**0.72 (0.16)**	**0.70 (0.17)**	**0.65 (0.18)**	**0.63 (0.19)**
	LPIPS [36]	0.33 (0.25)	0.39 (0.27)	0.46 (0.25)	0.52 (0.27)	0.50 (0.22)	0.53 (0.25)	0.61 (0.19)	0.61 (0.23)	0.49 (0.21)	0.53 (0.25)
	VSTR [37]	**0.73 (0.19)**	**0.72 (0.17)**	**0.71 (0.14)**	**0.71 (0.14)**	**0.73 (0.11)**	**0.71 (0.12)**	**0.82 (0.08)**	**0.79 (0.09)**	**0.75 (0.10)**	**0.73 (0.10)**
No Reference	BRISQUE [38]	0.29 (0.21)	0.24 (0.21)	0.28 (0.21)	0.25 (0.21)	0.35 (0.23)	0.32 (0.23)	0.45 (0.21)	0.44 (0.18)	0.31 (0.20)	0.31 (0.19)
	NIQE [39]	0.35 (0.25)	0.27 (0.25)	0.34 (0.22)	0.28 (0.24)	0.39 (0.24)	0.34 (0.25)	0.44 (0.21)	0.45 (0.22)	0.37 (0.20)	0.38 (0.22)
	TLVQM [40]	0.20 (0.15)	0.22 (0.15)	0.28 (0.20)	0.27 (0.20)	0.33 (0.19)	0.27 (0.19)	0.48 (0.15)	0.44 (0.16)	0.32 (0.17)	0.30 (0.17)
	VIDEVAL [43]	0.20 (0.17)	0.18 (0.17)	0.23 (0.19)	0.21 (0.20)	0.33 (0.22)	0.30 (0.22)	0.52 (0.21)	0.52 (0.20)	0.33 (0.19)	0.30 (0.19)

Since the ETRI-LIVE database consist of videos afflicted by various distortions applied on the same source contents, we took particular care to separate the train and test sets 'content-wise.' Hence, the train and test sets did not share any videos derived from the same source contents. For the performance evaluation, we used 5-fold cross validation. Since the database contains 15 unique source contents, the model was trained on all the distorted versions of 12 source contents (and their DMOS or MOS, as appropriate), and tested on the distorted videos derived from the other three source contents. We ran 1000 train/test iterations, in this manner, where the train/test sets were randomly divided at each iteration while following the content-wise separation. The results in Table V and VI show the medians and standard deviations of PLCC and SRCC of the compared learning-based models across the 1000 iterations. We also list the median

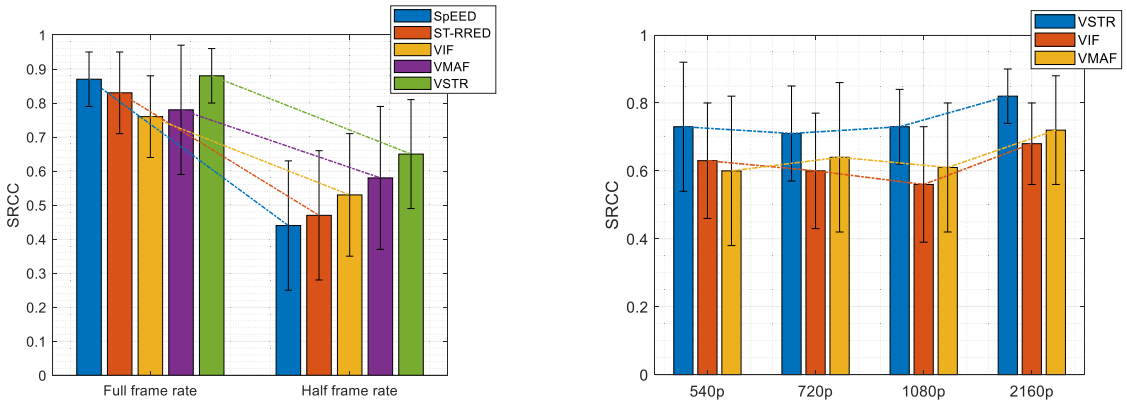

Fig. 14. Bar plots of SRCC performances of the compared quality models across (a) different frame rates, where the three best models in each SRCC column of Table V (VSTR, VMAF, and VIF for half frame rate and SpEED, ST-RRED, and VSTR for full frame rate) are compared, and across (b) different spatial resolution, where three best models in each SRCC column of Table VI (VSTR, VIF, and VMAF for all resolutions) are compared.

performances of the other non-learning-based models, but on the *same* randomized splits for comparison.

Table V shows the cross-validation performance over different temporal subsampling levels and overall. No-reference models underperformed against the reference models. Reference models again attained high performances on full frame rate videos, but performed comparatively worse on temporally subsampled videos. Fig. 14(a) shows a bar plot of SRCC performance across frame rates. The compared models were chosen to include the three best models in each column of Table V. For the full frame rate case, SpEED, ST-RRED, and VSTR attained the best SRCC performances (higher than 0.8), suggesting that they are reliable models when assessing the quality of videos afflicted only by compression and spatial subsampling. However, the performances of SpEED and ST-RRED fell steeply when temporal subsampling was introduced, where the performance decreased by 49% and 43%, respectively. For the half frame rate case, the learning-based reference models, such as VMAF and VSTR, were able to effectively predict the video quality. These models still performed comparatively worse than on the full frame rate case, but the performance decrease was less severe, at 26% for both models. These learning-based reference models are likely more reliable predictors when the application scenario includes frame rate adaptation.

Table VI presents the cross-validation performance over different spatial subsampling level and overall. Fig. 14(b) shows a bar plot of the SRCC performances of the three best models in each column of Table VI, across different spatial resolutions. As may be seen in Table III and Fig. 14(b), similar prediction performance was obtained across the different spatial resolutions, suggesting that the prediction power of the models was not very affected by the different levels of spatial subsampling.

For most quality models, the attained prediction performance mainly depends on whether the frame rate variations are present or not. This suggests the need for further investigations into the underdeveloped topic of temporally subsampled and compressed video quality. Overall, the learning-based reference models, VMAF and VSTR, outperformed the other models.

VI. Conclusion and Future Work

We conducted a large-scale human study to more generally understand combined space-time subsampling and compression affect the perceptual quality of videos. The new ETRI-LIVE STSVQ database contains 15 unique source contents and 437 distorted versions of them, on which almost 15,000 subjective opinion scores were collected. This study is the first to include the subjective scores on 4K 10bit videos with frame rates up to 120Hz, subjected to simultaneous space-time subsampling and compression.

Analysis of the subjective scores reveals that, while space-time subsampling inevitably results in a loss of information and subsequent degradations on quality, it may be a good tradeoff against increased compression, given a fixed bit rate budget. A rate distortion analysis of the subjective scores showed that space-time subsampling prior to video compression can significantly improve video perceptual quality at low bit rates. An interesting topic for further study, would be to understand content-wise trade-offs between space-time subsampling and compression. For instance, contents containing large motions may be affected more significantly by temporal information loss and artifacts, and thus, the optimal strategy for this kind of content may be to not include temporal subsampling, even at low bit rates. On the other hand, static contents may benefit from simultaneous space-time subsampling, where less compression is required to meet a limited bit budget. The new database may be further studied towards developing space-time resolution adaptation algorithms, whereby the space-time characteristics of videos and the available bit rate are jointly considered to determine optimal space-time resolution parameters prior to perceptual video coding. Successfully addressing this problem could provide significant optimization of the workflows of streaming media providers.

We also evaluated several high-performance objective image/video quality models on the new database. The results from this benchmark study indicate that, while the compared models can effectively predict the quality of videos subjected to spatial subsampling and compression, they are much less effective if temporal subsampling is included in the mix of distortions. This suggests that further study could lead to significant improvements of existing models, or new models altogether, more capable of capturing the deleterious perceptual effects of temporal subsampling. Such quality models could be utilized in media streaming applications where the space-time sensitive VQA models are used to monitor streaming video quality while delivering feedback to the server to affect perceptually optimal video encoding.

The ETRI-LIVE STSVQ database is being made publicly and freely available at (https://live.ece.utexas.edu/research/ETRI-LIVE_STSVQ/index.html) with the desire to improve future research and development on topics such as video quality modeling and perceptual video coding.

ACKNOWLEDGMENT

The research was IRB exempt 2007-11-0066.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[3] D. Mukherjee *et al.*, "A technical overview of VP9—The latest open-source video codec," *SMPTE Motion Imag. J.*, vol. 124, no. 1, pp. 44–54, Jan. 2015.

[4] Y. Chen *et al.*, "An overview of core coding tools in the AV1 video codec," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2018, pp. 41–45.

[5] 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.

[6] 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.

[7] A. Mackin, M. Afonso, F. Zhang, and D. Bull, "A study of subjective video quality at various spatial resolutions," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2830–2834.

[8] Q. Huang *et al.*, "Perceptual quality driven frame-rate selection (PQD-FRS) for high-frame-rate video," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 640–653, Sep. 2016.

[9] A. V. Katsenou, D. Ma, and D. R. Bull, "Perceptually-aligned frame rate selection using spatio-temporal features," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2018, pp. 288–292.

[10] M. Afonso, F. Zhang, and D. R. Bull, "Video compression based on spatio-temporal resolution adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 275–280, Jan. 2019.

[11] R. R. R. Rao, S. Goring, W. Robitza, B. Feiten, and A. Raake, "AVT-VQDB-UHD-1: A large scale video quality database for UHD-1," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 1–8.

[12] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120 fps 4K sequences for video codec analysis and development," in *Proc. 11th ACM Multimedia Syst. Conf.*, May 2020, pp. 297–302.

[13] C. Montgomery. (1994). *Xiph.org Video Test Media (Derf's Collection), the Xiph Open Source Community*. Accessed: May 2020. [Online]. Available: https://media.xiph.org/video/derf

[14] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T Rec. P.910, International Telecommunication Union, Geneva, Switzerland, 2009, pp. 1–42.

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

[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] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2019.

[18] J. Li, Y. Koudota, M. Barkowsky, H. Primon, and P. Le Callet, "Comparing upscaling algorithms from HD to ultra HD by evaluating preference of experience," in *Proc. 6th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 208–213.

[19] A. B. Watson, "High frame rates and human vision: A view through the window of visibility," *SMPTE Motion Imag. J.*, vol. 122, no. 2, pp. 18–32, Mar. 2013.

[20] B.-D. Choi, J.-W. Han, C.-S. Kim, and S.-J. Ko, "Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 407–416, Apr. 2007.

[21] Y. Matsuo and S. Sakaida, "Frame-rate conversion method by linear-filtering interpolation using spatio-temporal contrast compensation," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2017, pp. 243–244.

[22] *HEVC Reference Software*. Accessed: Nov. 2019. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

[23] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Rec. BT.500-11, ITU-R, 2012.

[24] *ClearView Player*. Accessed: May 2020. [Online]. Available: https://videoclarity.com/PDF/ClearView-Player-DataSheet.pdf

[25] *Acer Predator X27*. Accessed: May 2020. [Online]. Available: https://www.acer.com/ac/en/U.S./content/predator-series/predatorx27

[26] *Palette Gear Console*. Accessed: Jul. 2020. [Online]. Available: https://monogramcc.com

[27] *Parameter Values for an Expanded Hierarchy of LSDI Image Formats for Production and International Programme Exchange*, document Rec. BT.1769, ITU-R, 2008.

[28] 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.

[29] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," *Proc. SPIE, Adv. Image Video Commun. Storage Technol.*, vol. 2451, pp. 90–101, Feb. 1995.

[30] 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.

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

[32] 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.

[33] 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. 2012.

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

[35] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. *Toward a Practical Perceptual Video Quality Metric*. Accessed: May 2019. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[36] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[37] D. Y. Lee, H. Ko, J. Kim, and A. C. Bovik, "Space-time video regularity and visual fidelity: Compression, resolution and frame rate adaptation," 2021, *arXiv:2103.16771*.

[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] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2013.

[40] 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.

[41] *VMAF Pre-Trained Models*. Accessed: May 2020. [Online]. Available: https://github.com/Netflix/vmaf/tree/master/model
[42] *Perceptual Similarity Metric and Dataset Project Page*. Accessed: Aug. 2021. [Online]. Available: https://github.com/richzhang/PerceptualSimilarity
[43] 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.
[44] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci. Inf. Eng., Univ. Nat. Taiwan, Taipei, Taiwan, Tech. Rep., 2003, pp. 1–12.

Dae Yeol Lee received the B.S. and M.Eng. degrees in electrical and computer engineering from Cornell University in 2012 and 2013, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with The University of Texas at Austin. Since 2013, he has been a Researcher at the Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include image/video quality assessment and enhancement, immersive media, machine learning, and computer vision.

Somdyuti Paul (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the West Bengal University of Technology in 2012 and the M.Tech. degree in electrical engineering from the Indian Institute of Technology Kanpur in 2015. She is currently pursuing the Ph.D. degree in electrical and computer engineering with The University of Texas at Austin. She is also with the Laboratory for Image and Video Engineering, The University of Texas at Austin. Her research interests include image and video processing, computer vision, and machine learning.

Christos G. Bampis (Member, IEEE) received the M.Eng. (Diploma) degree in electrical engineering from the National Technical University of Athens in 2014 and the Ph.D. degree in electrical engineering from The University of Texas at Austin in 2018. He is currently a Senior Software Engineer with the Video Algorithms Team, Netflix, Inc. His work focuses on research and development of perceptual video quality prediction systems. His research interests include image and video quality assessment, computer vision, and machine learning.

Hyunsuk Ko (Member, IEEE) received the B.S. and M.S. degrees from the Department of Electrical Engineering, Yonsei University, Seoul, South Korea, in 2006 and 2009, respectively, and the Ph.D. degree from the Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA, in 2015. From 2015 to 2020, he had been a Senior Researcher with the Electronics and Telecommunications Research Institute. He is currently an Assistant Professor with the School of Electrical Engineering, Hanyang University ERICA, Ansan, South Korea. His research interests include video coding, perceptual visual quality assessment, and deep learning-based multimedia signal processing.

Jongho Kim received the B.S. degree from the Control and Computer Engineering Department, Korea Maritime University, in 2005, and the M.S. degree from the University of Science and Technology (UST) in 2007. He is currently pursuing the Ph.D. degree with the Korea Advanced Institute of Science and Technology (KAIST). He joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, in 2008, and has been a Senior Researcher since 2016. His research interests include video standardization activities of ITU-T VCEG and ISO/IEC MPEG, perceptual video coding, and machine learning-based compression.

Se Yoon Jeong received the B.S. and M.S. degrees from Inha University in 1995 and 1997, respectively, and the Ph.D. degree from the Korea Advanced Institute of Science and Technology in 2014. In 1996, he joined the Electronics and Telecommunications Research Institute, as a Principal Researcher. He has many contributions to development of international standards, such as scalable video coding and high efficient video coding. His current research interests include video coding for machine and AI-based video coding.

Blake Homan is an Entrepreneur and Industry Veteran, with over 30 of experience in a variety of senior engineering and marketing positions. Prior to founding Video Clarity, Inc., he was the Vice President of Technical Marketing at Optibase (OBAS), an IPTV streaming, encoding, and decoding solutions' company. He joined Optibase with the acquisition of Viewgraphics Inc., a digital media transmission company. During his tenure at Optibase, he was instrumental in setting up world-wide support and defining application engineering. He founded Video Clarity, Inc., in 2003, and continues to lead the company in efforts to assist media companies with video quality test and measurement.

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 was a recipient of the 2022 IEEE Edison Medal, 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 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 longest-serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING, and also created/chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, in 1994.