Perceptual Quality Assessment of HEVC and VVC Standards for 8K Video

Charles Bonnineau, Wassim Hamidouche, Jerome Fournier, Naty Sidaty, Jean-Francois Travers, Olivier Deforges

Perceptual Quality Assessment of HEVC and VVC Standards for 8K Video

Charles Bonnineau, Wassim Hamidouche, *Member, IEEE*, Jérôme Fournier, Naty Sidaty, *Member, IEEE*, Jean-François Travers, and Olivier Déforges

***Abstract*—With the growing data consumption of emerging video applications and users' requirement for higher resolutions, up to 8K, a huge effort has been made in video compression technologies. Recently, versatile video coding (VVC) has been standardized by the moving picture expert group (MPEG), providing a significant improvement in compression performance over its predecessor high efficiency video coding (HEVC). In this paper, we provide a comparative subjective quality evaluation between VVC and HEVC standards for 8K resolution videos. In addition, we evaluate the perceived quality improvement offered by 8K over UHD 4K resolution. The compression performance of both VVC and HEVC standards has been conducted in random access (RA) coding configuration, using their respective reference software, VVC test model (VTM-11) and HEVC test model (HM-16.20). Objective measurements, using PSNR, MS-SSIM and VMAF metrics have shown that the bitrate gains offered by VVC over HEVC for 8K video content are around 31%, 26% and 35%, respectively. Subjectively, VVC offers an average of around 41% of bitrate reduction over HEVC for the same visual quality. A compression gain of 50% has been reached for some tested video sequences regarding a Student's t-test analysis. In addition, for most tested scenes, a significant visual difference between uncompressed 4K and 8K has been noticed.**

***Index Terms*—Subjective quality assessment, compression efficiency, VVC, HEVC, 8K, UHD (4K).**

I. Introduction

With the latest ultra-high definition television (UHDTV) system [1] deployment, the quality of experience (QoE) of users is expected to improve by introducing new features to the existing high definition television (HDTV) system [2], including high dynamic range (HDR), wider color gamut, high frame-rate (HFR), and higher spatial resolutions, with 4K (3840×2160) and 8K (7680×4320) [3], [4]. The delivery of these video formats on current broadcast infrastructures is a real challenge and requires efficient compression methods to reach the available bandwidth while ensuring a higher video quality.

Manuscript received September 17, 2021; revised December 15, 2021; accepted December 17, 2021. This work was supported by the French Government through the National Research Agency (ANR) Investment under Grant ANR-A0-AIRT-07. *(Corresponding author: Charles Bonnineau.)*

Charles Bonnineau is with the Institute of Research and Technology, b<>com, 35510 Cesson-Sévigné, France, also with Univ. Rennes, INSA Rennes, CNRS, IETR-UMR 6164, 35708 Rennes, France, and also with Direction Technique, TDF, 35510 Cesson-Sévigné, France (e-mail: charles.bonnineau@b-com.com).

Wassim Hamidouche is with the Institute of Research and Technology, b<>com, 35510 Cesson-Sévigné, France, and also with Univ. Rennes, INSA Rennes, CNRS, IETR-UMR 6164, 35708 Rennes, France (e-mail: whamidou@insa-rennes.fr).

Jérôme Fournier is with the Institute of Research and Technology, b<>com, 35510 Cesson-Sévigné, France, and also with Orange Labs, Orange, 35510 Cesson-Sévigné, France.

Naty Sidaty and Jean-François Travers are with TDF, 35510 Cesson-Sévigné, France.

Olivier Déforges is with Univ. Rennes, INSA Rennes, CNRS, IETR-UMR 6164, 35708 Rennes, France.

Digital Object Identifier 10.1109/TBC.2022.3140710

Contributions to video coding standards like HEVC [5] or its successor VVC, finalized in July 2020 as ITU-T H.266 — MPEG-I - Part 3 (ISO/IEC 23090-3) standard [6], [7], enable video signal compression to be continuously improved through the standardization bodies. Although HEVC has brought a significant bitrate reduction for 4K delivery, its efficiency is not enough for 8K applications. Several studies have shown that the bitrate required by HEVC for 8K applications in 60Hz and 120Hz (temporally scalable) is around 80Mbps [8]–[10]. In practice, an 8K 120Hz HEVC codec [11], [12] has been used for Japan's satellite broadcasting by using DVBS2X [13]. In that case, the use of a complete transponder or multiple bonded transponders can reach bandwidth in the range 70-80Mbps. For terrestrial transmission, such bandwidth requirements prevent the deployment of more than one 8K HEVC program per ultra high frequency (UHF) channel, as practical DVB-T2 [14] channels offer bandwidth in the range of 30-40Mbps over an 8MHz channel. Thus, significant compression gains need to be achieved to ensure the successful deployment of 8K video services.

This paper provides both subjective and objective quality assessments of the two latest MPEG video coding standards for 8K video coding. We selected 8K sequences with various spatial and temporal characteristics to provide a fair evaluation. The compression points have been generated using the random access (RA) mode of the VVC and HEVC reference software models, called VTM-11 and HM-16.20, respectively. For subjective quality assessment, we used the double stimulus continuous quality scale (DSCQS) method described in Recommendation BT.500-14 [15] standardized at ITU-R. This study includes rate-distortion (RD) curves, Bjontegaard-Delta (BD) bitrate evaluation, and a Student's t-test, offering a robust statistical analysis.

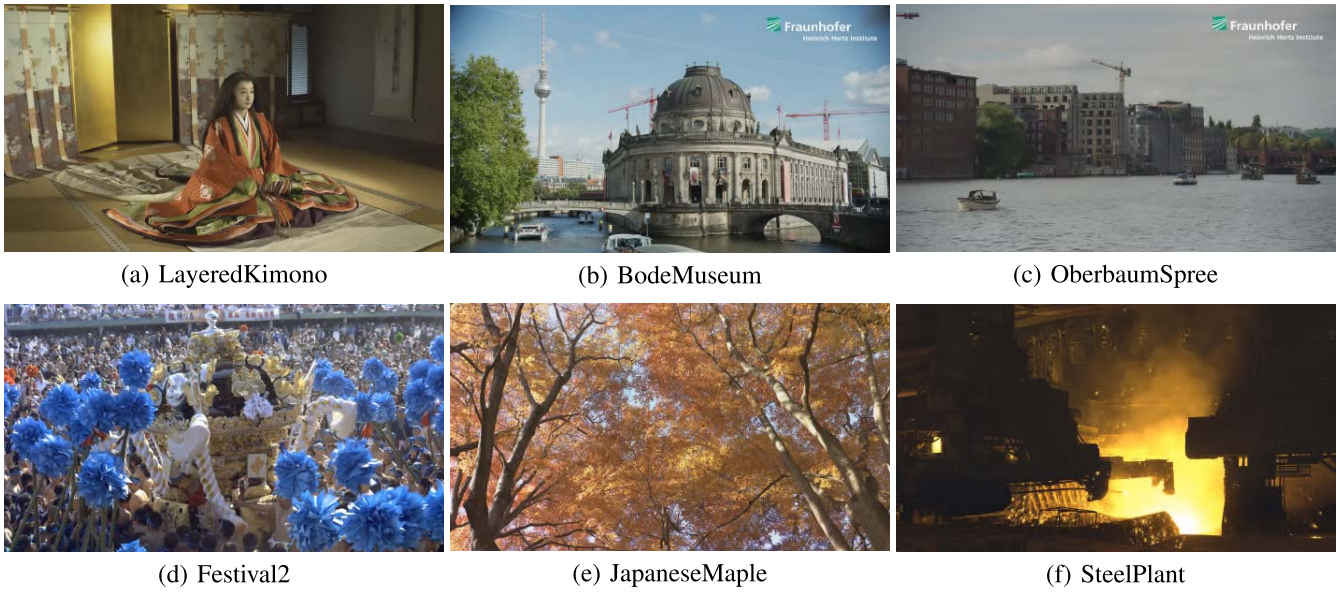
The contributions of this work are the following:

- Assess the compression gain offered by VVC over HEVC for 8K video contents. This gain represents approximately 41% of bitrate saving for the same visual quality,
- Determine the required bitrate for transparency, i.e., no visual difference is perceived between the source and decoded video,
- Confirm that non-expert viewers can see the difference between 4K and 8K resolutions and measure that difference,
- Evaluate several objective quality metrics based on the subjective test statistics collected on the 8K video dataset.

The rest of this paper is organized as follows. Section II provides an overview of existing studies for 4K and 8K video quality assessment. Section III describes the subjective test materials, including the test sequences, the codecs configuration, and the subjective test methodology. The results of both the objective and the subjective experiments are given in Section IV. Finally, Section V concludes the paper.

II. Related Works

Recently, a study was conducted to evaluate different scenarios for 8K video delivery with 4K backward compatibility relaying on objective quality metrics [16]. It was shown that



(a) LayeredKimono (b) BodeMuseum (c) OberbaumSpree

(d) Festival2 (e) JapaneseMaple (f) SteelPlant

Fig. 1. Snapshots of the six selected 8K test video sequences.

VVC offers around 40% of bitrate reduction over HEVC for the same peak signal to noise ratio (PSNR) quality on 8K video resolution [17]. Although recently developed objective quality metrics, like video multimethod assessment fusion (VMAF) [18], are more correlated to subjective test scores, it is acknowledged that these quality metrics still lack fidelity regarding the viewing conditions and the human visual system. Thus, rigorous perceptual quality assessment methodologies have been developed to fairly evaluate compression algorithms and ensure experiment reproducibility [19], [20].

For instance, Tan *et al.* [21] have demonstrated that a difference of 15% of compression gain is noticed depending on whether the objective or subjective quality is considered when evaluating HEVC over advanced video coding (AVC). This evaluation has been conducted using the respective reference implementations of both standards for resolutions ranging from 480p to 2160p. Another perceptual study has confirmed that a bitrate saving in the range 55-87% for the same perceived quality is enabled by HEVC over AVC on a bench of sequences, including 4K contents [22]. Regarding VVC and HEVC comparison, a recent subjective test has validated that VVC offers around 40% or bitrate reduction for the same perceived quality targeting 4K and HD contents [23]. In addition to HEVC and VVC, subjective quality assessment of AOMedia Video 1 (AV1) has been included in the work of Zhang *et al.* [24] for 4K video resolution. The results have shown that, at the same video bitrate level, AV1 and HM-16.20 are not significantly different in terms of perceived quality.

For 4K video resolution broadcasting with HEVC, a study has been conducted regarding target bitrates in the range 18-36Mbps [25]. This experiment has demonstrated that 4K resolution can reach a good perceptual quality at a bitrate of 18Mbps using HEVC.

Concerning 8K resolution videos, several studies have shown that the bitrate required for 8K applications is approximately 80Mbps using HEVC [8]–[10]. The QoE of 8K contents has also been assessed regarding different use-cases by using specific contents [26], e.g., food, people.

In this paper, we provide a subjective evaluation between HEVC and VVC for 8K resolution video. To the best of our knowledge, this is the first quality assessment study based on those two MPEG standards for 8K. Also, we provide an analysis on the gain in terms of quality enhancement offered by 8K over 4K for the uncompressed selected contents.

TABLE I
PARAMETERS OF THE 8K TEST VIDEO SEQUENCES. ALL SEQUENCES ARE IN 4:2:0 COLOR SUB-SAMPLING FORMAT

Sequence	Resolution (W × H)	Frame-rate	Frames	Color space	Bitdepth	Src
BodeMuseum	7680×4320	60fps	600	BT.709	10	HHI
OberbaumSpree	7680×4320	60fps	600	BT.709	10	HHI
LayeredKimono	7680×4320	60fps	300	BT.2020	10	ITE
Festival2	7680×4320	60fps	300	BT.2020	10	ITE
JapaneseMaple	7680×4320	60fps	300	BT.2020	10	ITE
SteelPlant	7680×4320	60fps	600	BT.2020	10	ITE

III. SUBJECTIVE QUALITY ASSESSMENT OF 8K RESOLUTION

This section provides details regarding the test sequences, the subjective test settings, and the experimental environment.

A. Test Video Sequences

In this study, we selected six test video sequences over multiple videos collected from the Institute of Image Information and Television Engineers (ITE)[1] and the Fraunhofer Heinrich-Hertz-Institut (HHI) [27] 8K video databases. The scenes were chosen based on video features like color, movement, texture, and homogeneous content, leading to different behaviors of the compression algorithms. We also considered the relevance of the 8K resolution in the scene selection. The details of the 8K test sequences are reported in Table I. Screenshots of the selected scenes are given in Fig. 1. To ensure homogeneity over video sequences and keep the same display parameters for the whole experiment, we performed a color space conversion from BT.709 [28] to BT.2020 [29] for *BodeMuseum* and *OberbaumSpree* scenes. Also, as the sequences *LayeredKimono*, *Festival2*, and *JapaneseMaple* contain fewer frames than the others, we played them back in mirror mode after 5 seconds to get 10 seconds videos while preserving the motion continuity of the scene. For those sequences, the motion direction change was coherent with the initial content.

[1] https://www.ite.or.jp/content/test-materials/

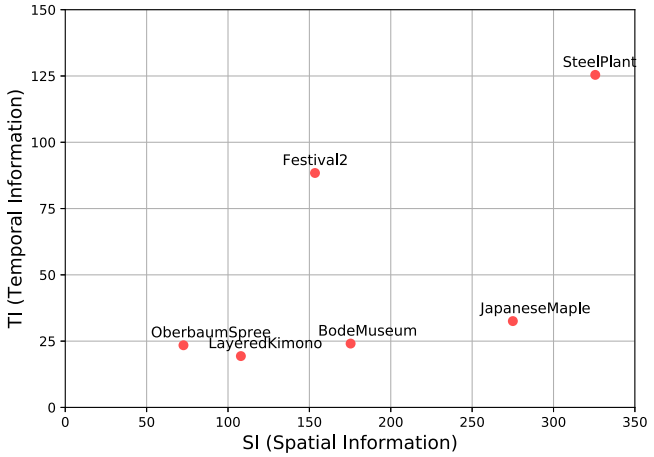


Fig. 2. SI-TI graph of the tested 8K video sequences.

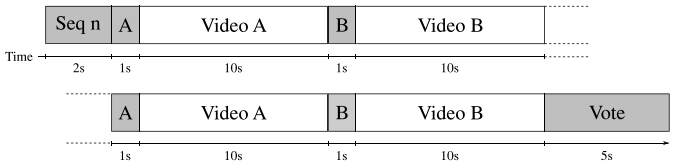


Fig. 3. Subjective basic test cell (BTC) structure according to the DSCQS evaluation methodology.

The spatial and temporal information (SI-TI) [15] of the selected sequences is plotted in Fig. 2. This 2D plan shows that the contents selected for the study are diverse regarding spatio-temporal features.

Based on these six uncompressed (raw) selected 8K video sequences (scenes), ten processed video sequences (PVSs) are generated per scene:

- one 8K (7680×4320) hidden reference uncompressed video.
- one 4K (4320×2160) uncompressed video. In that case, the source signal is first downscaled to 4K and then rescaled to 8K by using the *Lanczos3* [30] filter provided by *ffmpeg*[2] for both operations.
- 8K video encoded at four bitrates with HEVC.
- 8K video encoded at four bitrates with VVC.

In total, 60 video sequences are evaluated in this study.

The Common Test Conditions for VTM-11 [31] and HM-16.20 [32] in RA coding mode for main10 profile were used to perform a fair rate/distortion evaluation. These software models provide a reference implementation of the compression standards, representing their upper-bound coding performance with a moderate optimization level. For both codec, a GOP size of 16 and an Intra Period of 64 frames were used. For each scene, the test points are obtained using different fixed quantization parameter (QP) values. To cover a wide range of visual quality, we determined the highest bitrate value considering the transparency, i.e., the bitrate for which degradation starts to appear, as the highest bitrate point for each sequence. Also, the bitrates were carefully selected so that each bitrate R_i is approximately half of the next bitrate R_{i+1} and each VVC bitrate R_i^{VVC} is equal to the corresponding HEVC bitrate R_i^{HEVC} for $i \in \{1, 2, 3, 4\}$. The used QPs and bitrates for each sequence are given in Table II. We can note that the bitrate selected for transparency varies from 11Mbps to 180Mbps, depending on the test sequence.

[2]https://www.ffmpeg.org/

TABLE II
SELECTED QP AND CORRESPONDING BITRATES (MBPS), FOR BOTH VTM-11 AND HM-16.20 CODECS, ACCORDING TO THE TEST SEQUENCE

Sequence	Codec	R_1 (QP/Mbps)	R_2 (QP/Mbps)	R_3 (QP/Mbps)	R_4 (QP/Mbps)
LayeredKimono	HEVC	38/1.9	34/3.2	29/6.3	26/11.4
	VVC	37/1.8	32/3.4	27/6.5	24/10.8
BodeMuseum	HEVC	38/4.7	33/9.8	28/22.5	25/45.4
	VVC	37/4.8	32/10.1	27/22.6	24/42.9
OberbaumSpree	HEVC	38/3.3	33/7.4	28/17.5	24/40.5
	VVC	37/3.6	32/8.1	27/18.6	23/43.9
Festival2	HEVC	39/17.5	34/32.1	29/59.5	24/130.4
	VVC	37/17.4	32/32.2	27/61.1	22/135.5
JapaneseMaple	HEVC	43/15.2	38/34.9	33/76.1	28/168
	VVC	42/15.9	37/35.7	32/79.8	27/174.9
SteelPlant	HEVC	42/19.6	38/40.5	33/86.9	28/175.5
	VVC	42/18.0	37/42.9	32/91.1	27/180.5

B. Subjective Testing Procedure

In this study, we used the method described in the ITU-R Recommendation BT.500-14 [15], called double stimulus continuous quality scale (DSCQS), to collect the video quality scores from participants. This testing method requires a prior pseudo-random sequencing of the testing videos, as the observer has no interactivity with the player. Thus, each test session of the DSCQS method consists of different random series of basic test cells (BTCs) presentations. This method presents the test videos by pairs ("video A" and "video B") separated with annotated mid-greys. For each BTC, both "video A" and "video B" are repeated twice. An example of BTC used for evaluation is illustrated in Fig. 3. Each presented pair contains the implicit 8K uncompressed reference and one random PVS over all the ten configurations, i.e., the same scene encoded with HEVC or VVC at four bitrates or the uncompressed sequence in 4K or 8K resolution. Also, to prevent visual fatigue, the test is divided into three sessions of 20 minutes each. Before each experiment, participants receive clear explanations about the evaluation procedures.

After the first "video A/video B" pair presentation, the participant could report his opinion about the perceived video quality on two vertical lines with the corresponding sequence index for both "video A" and "video B". For this testing method, the vertical rating lines are divided into five segments of the same height and scaled from the lower to the higher quality with the labels *Bad*, *Poor*, *Fair*, *Good*, and *Excellent*. After each video pair visualization, participants can vote by annotating both videos along the continuous quality scale. The scores are then collected by converting the annotations into a value between 0 and 100.

C. Experimental Environment

This subjective study has been conducted in a controlled laboratory environment that follows the ITU-R Rec. BT.500-14 [15]. The objective is to offer visualization comfort to participants and ensure the reproducibility of the test. All the experimental setup details are reported in Table III. A picture illustrating the test conditions is given in Fig. 4. A total of 22 non-expert observers aged from 22 to 53 years have taken part in this experiment. All participants have been screened for normal visual acuity and color blindness using the Ishihara and Snellen vision tests, as described in the ITU-R

Fig. 4. Illustration of the laboratory environment, compliant with the ITU-R BT500-13 Recommendation [19].

TABLE III
TEST LOGISTICS

Monitor	SONY 85" KD-85ZG
Player	Zaxel's Zaxtar 5 8K
Peak luminance	120 cd/m^2
Video Format	7680x4320/60p/YUV4:2:0/10bits
Viewing distance	0.8H (approximtely 0.8m)
Background color	D65 mid-grey
Background luminance	15% of the screen maximum luminance

Recommendation BT.500-14 [15]. To detect outliers, the rejection method based on the Kurtosis coefficient from this same recommendation has been applied and has validated the overall participant's reported votes.

D. Subjective Quality Assessment

At the end of the subjective test sessions, the results for each scene are assessed by the differential mean opinion score (DMOS), corresponding to the average of the difference between the hidden reference and the corresponding PVS scores computed by:

$$\bar{x}_a = \frac{1}{n}\sum_{i=1}^{n} x_{i,a}, \tag{1}$$

where n is the total number of valid participants, $\bar{x}_a$ is the DMOS value of the tested configuration a, $a \in \{R_j^m$, 4K, 8K (ref)$\}$ for $j \in \{1, 2, 3, 4\}$ and $m \in \{VVC, HEVC\}$ and $x_{i,a}$ is the differential score computed as:

$$x_{i,a} = 100 - \left(y_{i,ref} - y_{i,a}\right), \tag{2}$$

with the pair $(y_{i,ref}, y_{i,a})$ representing the scores attributed by the participant i, $i \in \{1, \ldots, n\}$, to respectively the hidden reference (8K) and the tested configuration a, i.e., both videos of a given BTC.

To ensure that the vote distributions are normal, the bias reduction technique described in the ITU-T P.913 Recommendation [33] has been applied. Thus, from each resulting DMOS $\bar{x}_a$, the associated confidence intervals at 95% $(\bar{x}_a - c_a, \bar{x}_a + c_a)$ can be computed as follows:

$$c_a = 1.96\,\frac{s_a}{\sqrt{n}}, \tag{3}$$

where s_a is the standard deviation of the tested configuration a computed as:

$$s_a = \sqrt{\sum_{i=1}^{n} \frac{\left(x_{i,a} - \bar{x}_a\right)^2}{(n-1)}}, \tag{4}$$

with $x_{i,a}$ and $\bar{x}_a$ corresponding to the differential score of the observer i, $i \in \{1, \ldots, n\}$, and the DMOS score of the tested configuration a, respectively.

In addition, a Student's t-test with a two-tailed distribution is performed to provide a more rigorous analysis. More details are given in Section IV-B

IV. EXPERIMENTAL RESULTS

This section presents and discusses the results of both objective and subjective evaluation scores. An assessment of the objective metrics performance compared to the subjective scores for 8K video contents is also investigated.

A. Objective Results

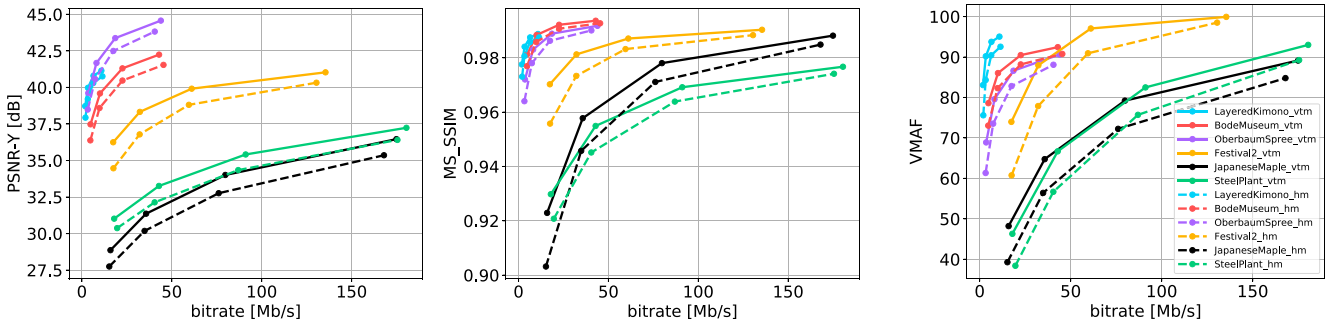
In this experiment, objective quality metrics, including PSNR, multi-scale structural similarity (MS-SSIM) [34], and VMAF [18], are used to measure the distortion between the 8K reconstructed signal and the source video. VMAF is an objective metric with reference, based on machine learning (ML) which evaluates the quality between the source and the tested content by giving a score between 0 and 100. This metric is trained to produce a score computed from different features (motion, spatial, texture) that maximize the correlation with mean opinion score (MOS) scores. Although VMAF was initially optimized for visual quality estimation of 4K contents, we have integrated it into the study as it achieves a high correlation with subjective scores. In this experiment, the VMAF scores are computed with the provided set of parameters *vmaf_v0.6.1.pkl*.[3] The PSNR is assessed on the luma component only. The RD curves are depicted in Fig. 5. It can be noted that the bitrates selected for transparency lead to quite different PSNR values depending on the sequence. In contrast, for more perceptually correlated objective metrics like MS-SSIM or VMAF, the predicted quality converges to the maximum value for all 8K sequences. Also, those curves confirm the observation made on the scene complexity with the SI-TI graph in Fig. 2. Three categories of sequences can be distinguished by scene complexity: Group 1 includes *LayeredKimono, OberbaumSpree, BodeMuseum* sequences, Group 2: *Festival2*, and Group 3: *JapaneseMaple, SteelPlants*.

We use the Bjontegaard-Delta (BD) computation method described in [35] to quantify the average gain in bitrate and visual quality offered by the VTM-11 over the HM-16.20 codec. The results are summarized in Table IV. In average, the VTM-11 codec enables around 31%, 26% and 35% of bitrate saving over the HM-16.20 codec, regarding PSNR, MS-SSIM and VMAF, respectively. However, the area between the interpolated curves covered using the BD-BR approach is limited as the selected bitrates are the same for both VVC and HEVC. Thus, to bring more details on the performance and consider a wider area between the curves, we compute the gain in quality of the VTM-11 over the HM-16.20 for the same bitrate using the BD method. By considering this approach, 0.91dB, 0.005 and 5.48 of quality improvement is offered by the VTM-11 over the HM-16.20 codec for the same bitrate, regarding PSNR, MS-SSIM and VMAF quality metrics, respectively.

B. Subjective Results

For the subjective quality evaluation, the rectified DMOS scores and their associated 95% confidence interval are collected following the method described in Section III-D. The resulting RD curves are depicted in Fig. 6 for all 8K sequences. These curves also display the scores obtained for the 8K hidden reference videos and the 4K sequences, with their associated 95% confidence interval represented by transparent areas.

[3]https://github.com/Netflix/vmaf



(a) PSNR (b) MS-SSIM (c) VMAF

Fig. 5. Objective quality comparison, using PSNR, MS-SSIM, and VMAF quality metrics for the 8K test video sequences.

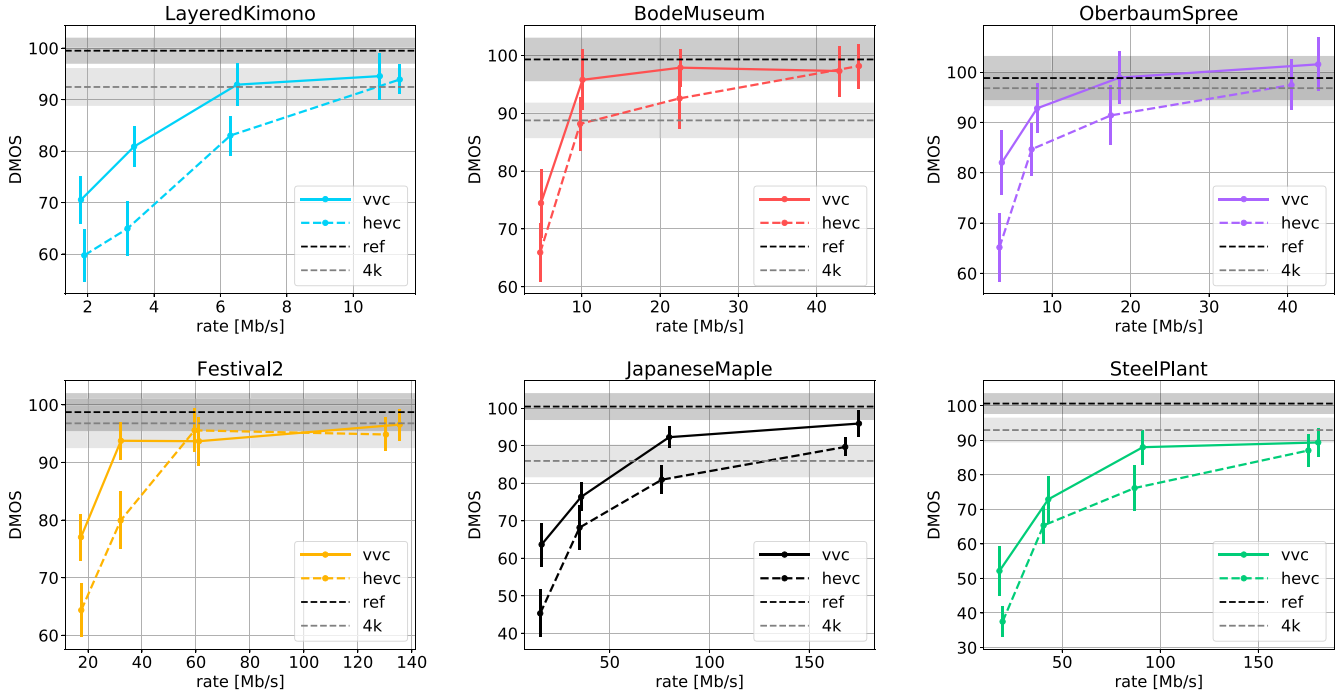


Fig. 6. DMOS-based comparison, with associated 95% confidence interval, for the six selected 8K video sequences.

TABLE IV

BD-BR SCORES OF THE VTM-11 CODEC COMPARED TO THE ANCHOR HM-16.20. THE LEFT PART OF THE TABLE REPRESENTS THE BITRATE SAVINGS (%) FOR THE SAME QUALITY COMPUTED BY OBJECTIVE METRICS AND DMOS. NEGATIVE VALUES REPRESENT COMPRESSION GAIN OFFERED BY VVC OVER HEVC. THE RIGHT PART OF THE TABLE ILLUSTRATES THE GAIN IN QUALITY REGARDING EACH METRIC FOR THE SAME BITRATE. POSITIVE VALUES REPRESENT A GAIN IN QUALITY (REPRESENTED IN THE SCALE OF THE CONSIDERED METRIC) ENABLED BY VVC OVER HEVC

Sequence	BD-BR (PSNR)	BD-BR (MS-SSIM)	BD-BR (VMAF)	BD-BR (DMOS *upper* and *lower* limits)	BD-PSNR	BD-MS-SSIM	BD-VMAF	BD-DMOS (*upper* and *lower* limits)
LayeredKimono	-29.77%	-21.05%	-33.30%	-44.99% [-60.92%, -20.04%]	+0.61dB	+0.003	+4.63	+10.76 [+19.3, +2.22]
BodeMuseum	-32.75%	-25.05%	-34.70%	-36.43% [-74.71%, +21.12%]	+0.88dB	+0.002	+3.06	+5.79 [+15.21, -3.63]
OberbaumSpree	-32.07%	-27.00%	-33.41%	-55.59% [-87.15%, +28.59%]	+0.81dB	+0.003	+7.55	+7.87 [+18.44, -3.35]
Festival2	-36.40%	-33.36%	-28.24%	-28.89% [-59.43%, +37.28%]	+1.22dB	+0.006	+7.37	+5.13 [+12.98, -2.72]
JapaneseMaple	-28.33%	-23.37%	-30.86%	-43.36% [-64.42%, -6.69%]	+1.04dB	+0.009	+6.63	+9.79 [+18.27, +1.31]
SteelPlant	-28.30%	-24.40%	-27.57%	-37.41% [-67.61%, +13.31%]	+0.91dB	+0.007	+7.10	+8.83 [+20.40, -2.74]
Average	-31.27%	-25.7%	-35.30%	-41.11% [-69.04%, +12.26%]	+0.91dB	+0.005	+5.48	+8.03 [+17.43, -1.49]

In order to confidently evaluate the statistical significance of the similarity (or not) between different tested sequences, we also performed a two-sample unequal variance Student's t-test with a two-tailed distribution. This study allows us to determine, for each scene, if the perceived quality between each pair of tested configurations is significantly different or not.

In this experiment, regarding two different tested configurations a_1 and a_2 for a given scene, the null hypothesis, H_0, corresponds

TABLE V

p-VALUE PROBABILITIES RESULTING FROM TWO-SAMPLE UNEQUAL VARIANCE BILATERAL STUDENT'S T-TEST ON DMOS VALUES FOR EACH PAIR OF TESTED CONFIGURATIONS AND EACH TEST SEQUENCE. $p \geq 0.05$ (GREEN) MEANS THERE IS NO SIGNIFICANT DIFFERENCE BETWEEN THE DMOS VALUE OF THE ROW AND COLUMN LABELS. IN CONTRAST, $p < 0.05$ (RED) INDICATES THAT THE DMOS VALUE OF THE ROW LABEL IS SIGNIFICANTLY DIFFERENT THAN THE COLUMN LABEL. THE VALUES REFERRED IN SECTION IV-B ARE REPRESENTED IN BOLD

(a) LayeredKimono

HEVC \ VVC	R1	R2	R3	R4	4K	REF
R1	**0.01**	0.00	0.00	0.00	0.00	**0.00**
R2	**0.15**	**0.00**	0.00	0.00	0.00	**0.00**
R3	0.00	**0.44**	**0.00**	0.00	0.00	**0.00**
R4	0.00	0.00	**0.79**	**0.70**	0.65	**0.01**
4K	0.00	0.00	0.88	0.47	**1.00**	**0.01**
REF	**0.00**	**0.00**	**0.02**	**0.10**	**0.01**	1.00

(b) BodeMuseum

HEVC \ VVC	R1	R2	R3	R4	4K	REF
R1	**0.04**	0.00	0.00	0.00	0.00	**0.00**
R2	**0.00**	**0.06**	0.00	0.01	0.90	**0.00**
R3	0.00	**0.44**	**0.13**	0.21	0.28	**0.07**
R4	0.00	0.56	**0.98**	**0.86**	0.00	**0.62**
4K	0.00	0.05	0.00	0.01	**1.00**	**0.00**
REF	**0.00**	**0.32**	**0.58**	**0.53**	**0.00**	1.00

(c) OberbaumSpree

HEVC \ VVC	R1	R2	R3	R4	4K	REF
R1	**0.00**	0.00	0.00	0.00	0.00	**0.00**
R2	**0.61**	**0.04**	0.00	0.00	0.04	**0.01**
R3	0.06	**0.74**	**0.09**	0.02	0.16	**0.07**
R4	0.00	0.23	**0.71**	**0.31**	0.85	**0.71**
4K	0.00	0.23	0.55	0.18	**1.00**	**0.52**
REF	**0.00**	**0.10**	**0.98**	**0.47**	**0.52**	1.00

(d) Festival2

HEVC \ VVC	R1	R2	R3	R4	4K	REF
R1	**0.00**	0.00	0.00	0.00	0.00	**0.00**
R2	**0.41**	**0.00**	0.00	0.00	0.00	**0.00**
R3	0.00	**0.34**	**0.55**	0.70	0.73	**0.26**
R4	0.00	0.42	**0.68**	**0.44**	0.53	**0.11**
4K	0.00	0.21	0.37	0.98	**1.00**	**0.48**
REF	**0.00**	**0.02**	**0.09**	**0.36**	**0.48**	1.00

(e) JapaneseMaple

HEVC \ VVC	R1	R2	R3	R4	4K	REF
R1	**0.00**	0.00	0.00	0.00	0.00	**0.00**
R2	**0.33**	**0.04**	0.00	0.00	0.00	**0.00**
R3	0.00	**0.13**	**0.00**	0.00	0.12	**0.00**
R4	0.00	0.00	**0.24**	**0.00**	0.18	**0.00**
4K	0.00	0.00	0.04	0.00	**1.00**	**0.00**
REF	**0.00**	**0.00**	**0.00**	**0.14**	**0.00**	1.00

(f) SteelPlant

HEVC \ VVC	R1	R2	R3	R4	4K	REF
R1	**0.00**	0.00	0.00	0.00	0.00	**0.00**
R2	**0.00**	**0.11**	0.00	0.00	0.00	**0.00**
R3	0.00	**0.55**	**0.01**	0.00	0.00	**0.0**
R4	0.00	0.00	**0.91**	**0.50**	0.07	**0.00**
4K	0.00	0.00	0.11	0.24	**1.00**	**0.00**
REF	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	1.00

to the case that a_1 and a_2 have the same perceived quality. On the contrary, the alternate hypothesis, H_a, would be that a difference between the tested configurations a_1 and a_2 is noted. The t-statistic can be estimated to quantify the degree of significance of the alternate hypothesis H_a. By considering the sample populations x_{a_1} and x_{a_2} from attributed scores for the tested configuration a_1 and a_2, respectively, the t-statistic can be computed as follows:

$$t_{a_1,a_2} = \frac{\bar{x}_{a_1} - \bar{x}_{a_2}}{\sqrt{\frac{s^2_{a_1}}{n_{a_1}} + \frac{s^2_{a_2}}{n_{a_2}}}}, \tag{5}$$

with $\bar{x}_{a_j}$, $s^2_{a_j}$ and n_{a_j} denoting the mean, the variance and the size of the sample population x_{a_j}, with $j \in \{1, 2\}$.

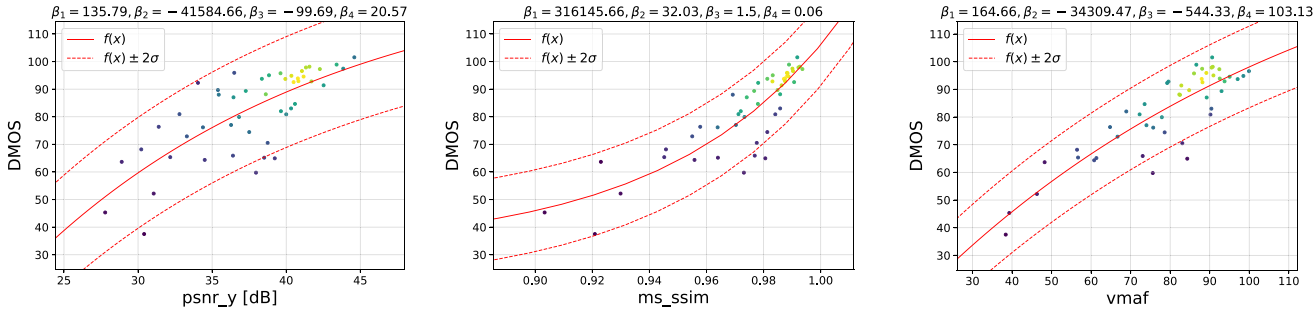
Then, by approximating the t-statistic with a Student's t-distribution, a value p, which indicates the degree of correlation between the means of the two sample populations, can be computed from the t-statistic. The higher the p-value is, the more significant the similarity between the distributions of the two populations is. A p-value lower than 0.05 indicates that there is a statistical significance that the two sample populations x_{a_1} and x_{a_2} have a different perceived quality. Indeed, there is a low probability of committing a type-I error, i.e., rejecting the null hypothesis when it is true, meaning that the null hypothesis can be confidently rejected. On the contrary, if the p-value is greater than or equal to 0.05, the null hypothesis cannot be safely rejected and both sample populations x_{a_1} and x_{a_2} can be considered to have the same perceived quality. The results for all scenes are given in Table V.

The results demonstrate that the perceived quality between uncompressed 8K and 4K formats depends on the scene content. For the sequences *JapaneseMaple*, *SteelPlant*, *BodeMuseum*, and *LayeredKimono*, the visual quality between both resolutions is significantly different as the p-value between the configurations 4K and REF is lower than 0.05. For those sequences, the global motion in the scene is low, which facilitate the sampling of 8K details by sensors. In contrast, for the sequences *Festival2* and *OberbaumSpree*, the motion in the scene can explain the 8K definition loss at 60fps. Indeed, the global motion in *Festival2* video sequence prevents from perceiving the details. For the *OberbaumSpree* motion blur appears on the scene due to a continuous horizontal camera traveling. It shows that higher framerates, e.g., 100/120fps, must be considered to fully benefit from the 8K resolution.

In complement to the objective study conducted in Section IV-A, we observe that the bitrate required to obtain transparency with the uncompressed 8K videos is highly content-dependent. Using VVC, the bitrates needed to reach the reference's quality are between 10Mbps to 180Mbps depending on the sequence. For the *SteelPlant* scene, the quality degradation with the source is always perceived on the selected bitrate range. Indeed, the p-values obtained between all R_i^{VVC} and REF configurations are lower than 0.05 for this sequence. It can be explained by the smoke in the scene, which is hard to compress and causes blocking artifacts. In comparison, the 8K source quality is obtained only for three scenes using HEVC: *BodeMuseum*, *Festival2*, *OberbaumSpree*. However, two of them are not critical (*Festival2*, *OberbaumSpree*), as no significant difference between REF and 4K is perceived ($p > 0.05$).

In addition, we can notice that, at the same bitrate, VVC offers perceived quality closer to the 8K reference video comparing to HEVC. For both *JapaneseMaple* and *LayeredKimono* scenes, a bitrate reduction of 50% is reached for the same level of visual quality. Indeed, we can observe in Table V that, for those two scenes, each VVC test point of bitrate R_i^{VVC} is statistically similar in terms of visual quality with respect to its corresponding HEVC test point at bitrate R_{i+1}^{HEVC} and significantly better at bitrate R_i^{HEVC}. Nevertheless, the results obtained with the rest of the 8K sequences with lower spatial textures do not follow this observation.

Finally, we applied the BD-BR method to the DMOS scores. Inspired by [21], we also compute the *upper* and *lower* limits for the BD-BR based on the confidence intervals. These scores are computed by comparing D_{max}^{VVC} with D_{min}^{HEVC} and D_{min}^{VVC} with D_{max}^{HEVC}, respectively, where $[D_{min}, D_{max}]$ represents the 95% confidence interval. All the results are reported in Table IV. These results demonstrate that VVC offers a compression gain over HEVC for the same perceived quality from 28.89% to 55.59% with an average of 41.11% over the whole 8K dataset.



(a) PSNR (b) MS-SSIM (c) VMAF

Fig. 7. Scatter plots and nonlinear logistic fitted curves of PSNR, MS-SSIM and VMAF quality metrics versus DMOS scores of the considered 8K video sequences. The logistic model coefficients are given for each tested objective metric.

TABLE VI
SROCC, PLCC, KROCC AND RMSE PERFORMANCE OF THE OBJECTIVE QUALITY METRICS MS-SSIM, SSIM, VMAF AND PSNR ON THE CONSIDERED 8K VIDEO SEQUENCES

Objective metric	SROCC	PLCC	KROCC	RMSE
MS-SSIM	**0.887**	0.871	**0.725**	7.409
SSIM	0.767	0.777	0.599	9.499
VMAF	0.806	**0.873**	0.603	**7.375**
PSNR	0.754	0.747	0.564	10.042

C. Correlation Consistency

In this section, the consistency of objective quality metrics with subjective scores is evaluated. Fig. 7 illustrates scatter plots with non-linear logistic fitted curves $f(x)$ and corresponding standard deviations intervals $f(x) \pm 2\sigma$ for PSNR, MS-SSIM, and VMAF quality metrics versus DMOS scores. The interpolated curves $f(x)$ are computed using the following logistic model:

$$f(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-\frac{x-\beta_3}{|\beta_4|}}}. \quad (6)$$

The more the standard deviation intervals are close to the logistic fitted curve, the more the metric is correlated to the DMOS score. In order to quantify the correlation of the objective metrics with the subjective scores, we use the Spearman's rank ordered correlation (SROCC), Pearson's linear correlation coefficient (PLCC), Kendall's rank-order correlation coefficient (KROCC), and root mean-squared error (RMSE). The results are reported in Table VI. As expected, it shows that MS-SSIM and VMAF are more correlated to subjective test ratings than PSNR, which gets the lowest performance regarding all indicators. In addition to the three considered objective quality metrics, we provide correlation scores with the SSIM metric. This latter shows slightly higher correlation with DMOS compared to PSNR, while it is outperformed by both MS-SSIM and VMAF. Finally, we can notice that VMAF is a relevant quality metric for 8K resolution evaluation although being optimized for 4K resolution.

V. CONCLUSION

In this paper, we evaluated the VVC compression performance over its predecessor HEVC for 8K video resolution. The subjective and objective quality assessments have been conducted on a selection of 8K video sequences in RA configuration. Objective results have demonstrated that the VTM-11 codec enables 31%, 26% and 35% of bitrate saving over the HM-16.20 codec, for PSNR, MS-SSIM and VMAF quality metrics, respectively. On the subjective side, VVC offers 41.11% of bitrate reduction over HEVC for the same visual quality, regarding the BD-BR method. Regarding the Student's t-test results, a bitrate reduction of about 50% is reached for two of the overall tested scenes. We have also demonstrated that the bitrate required to obtain transparency with the 8K source is highly content-dependent. Indeed, for VVC, a bitrate from 11Mbps to 180Mbps is needed, depending on the complexity of the scene. In addition, we demonstrated that the participants had noted a difference between uncompressed 4K and 8K for most of the tested sequences. However, sequences with high motion do not benefit from the 8K definition at 60fps. Finally, a higher correlation consistency between subjective and objective results can be noticed, particularly for the VMAF and MS-SSIM quality metrics.

Future works will focus on evaluating the subjective quality offered by recent deep-learning-based tools for 8K video compression, such as super-resolution, quality enhancement, and learning-based compression methods.

ACKNOWLEDGMENT

This work has been achieved within the Institute of Research and Technology b<>com, dedicated to digital technologies.

REFERENCES

[1] "Parameter values for ultra-high definition television systems for production and international programme exchange," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation BT.2020-1, 2015.
[2] "Parameters values for the HDTV standards for production and international programme exchange," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation BT.709-5, 2002.
[3] M. Nilsson, "Ultra high definition video formats and standardisation," London, U.K., BT Media Broadcast, Research Paper, 2015.
[4] M. Sugawara and K. Masaoka, "UHDTV image format for better visual experience," *Proc. IEEE*, vol. 101, no. 1, pp. 8–17, Jan. 2013.
[5] 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.
[6] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC)," *Proc. IEEE*, vol. 109, no. 9, pp. 1463–1493, Sep. 2021.
[7] W. Hamidouche *et al.*, "Versatile video coding standard: A review from coding tools to consumers deployment," 2021, *arXiv:2106.14245*.
[8] Y. Sugito *et al.*, "Video bit-rate requirements for 8K 120-Hz HEVC/H.265 temporal scalable coding: Experimental study based on 8K subjective evaluations," *APSIPA Trans. Signal Inf. Process.*, vol. 9, no. 1, p. e5, 2020.
[9] A. Ichigaya and Y. Nishida, "Required bit rates analysis for a new broadcasting service using HEVC/H.265," *IEEE Trans. Broadcast.*, vol. 62, no. 2, pp. 417–425, Jun. 2016.
[10] S. Iwasaki *et al.*, "The required video bitrate for 8k120-hz real-time temporal scalable coding," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Las Vegas, NV, USA, 2020, pp. 1–5.

[11] Y. Sugito *et al.*, "HEVC/H.265 codec system and transmission experiments aimed at 8K broadcasting," in *Proc. Techn. Paper Int. Broadcast. Conv. (IBC)*, Amsterdam, The Netherlands, 2015, pp. 24–29,

[12] Y. Sugito *et al.*, "A study on the required video bit-rate for 8K 120-Hz HEVC/H.265 temporal scalable coding," in *Proc. Picture Coding Symp. (PCS)*, San Francisco, CA, USA, 2018, pp. 106–110.

[13] *Digital Video Broadcasting (DVB); Second Generation Framing Structure, Channel Coding and Modulation Systems for Broadcasting, Interactive Services, News Gathering and Other Broadband Satellite Applications; Part 2: DVB-S2 Extensions (DVB-S2X)*, ETSI Standard EN 302 307-2 V1.2.1, 2014.

[14] *Digital Video Broadcasting (DVB); Implementation Guidelines for a Second Generation Digital Terrestrial Television Broadcasting System (DVB-T2)*, ETSI Standard TS 102 831, 2012.

[15] "Methodologies for the subjective assessment of the quality of television images," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation BT.500-14, 2019.

[16] C. Bonnineau, W. Hamidouche, J.-F. Travers, and O. Déforges, "Versatile video coding and super-resolution for efficient delivery of 8k video with 4k backward-compatibility," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Barcelona, Spain, 2020, pp. 2048–2052.

[17] C. Bonnineau, J.-Y. Aubié, W. Hamidouche, O. Déforges, J. Travers, and N. Sidaty, "An objective evaluation of codecs and post-processing tools for 8K video compression," in *Proc. Techn. Paper Int. Broadcast. Conv. (IBC)*, Amsterdam, The Netherlands, 2020.

[18] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, *Toward a Practical Perceptual Video Quality Metric*, Netflix TechBlog, Los Gatos, CA, USA, Jun. 2016.

[19] "Methodologies for the subjective assessment of the quality of television images," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation BT.500-13, 2019.

[20] R. Sotelo, J. Joskowicz, M. Anedda, M. Murroni, and D. D. Giusto, "Subjective video quality assessments for 4k UHDTV," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Cagliari, Italy, 2017, pp. 1–6.

[21] T. K. Tan *et al.*, "Video quality evaluation methodology and verification testing of HEVC compression performance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 76–90, Jan. 2016.

[22] A. Tabatabai *et al.*, "Compression performance analysis in HEVC," in *High Efficiency Video Coding (HEVC)*. Cham, Switzerland: Springer, 2014, pp. 275–302.

[23] N. Sidaty, W. Hamidouche, O. Déforges, P. Philippe, and J. Fournier, "Compression performance of the versatile video coding: HD and UHD visual quality monitoring," in *Proc. Picture Coding Symp. (PCS)*, Ningbo, China, 2019, pp. 1–5.

[24] F. Zhang, A. V. Katsenou, M. Afonso, G. Dimitrov, and D. R. Bull, "Comparing VVC, HEVC and AV1 using objective and subjective assessments," 2020, *arXiv:2003.10282*.

[25] S.-H. Bae, J. Kim, M. Kim, S. Cho, and J. S. Choi, "Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services," *IEEE Trans. Broadcast.*, vol. 59, no. 2, pp. 209–222, Jun. 2013.

[26] Y. Shishikui, "Quality-of-experience evaluation of 8K ultra-high-definition television," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2021, pp. 1404–1408.

[27] B. Bross, H. Kirchhoffer, C. Bartnik, and M. Palkow, *Multiformat Berlin Test Sequences*, document JVET-Q0791, JVET, Brussels, Belgium, Jan. 2020.

[28] "Parameter values for the HDTV standards for production and international programme exchange," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation BT.709-6, 2015.

[29] "Parameter values for the ultra-high definition television systems for production and international programme exchange," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation BT.2020-2, 2015.

[30] C. E. Duchon, "Lanczos filtering in one and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, 1979.

[31] F. Bossen, J. Boyce, K. Suehring, X. Li, and V. Seregin, *JVET Common Test Conditions and Software Reference Configurations for SDR Video*, document JVET-K1010, JVET meeting, Geneva, Switzerland, Mar. 2019.

[32] F. Bossen, *Common Test Conditions and Software Reference Configurations*, document JCTVC-L1100, Joint Collaborative Team Video Coding (JCT-VC), Geneva, Switzerland, May 2012.

[33] "Methods for the subjective assessment of video for quality, audio and audiovisual quality of Internet Video and distribution quality television in any environment," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation BT.913, 2021.

[34] 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. Pacific Grove, CA, USA, 2003, pp. 1398–1402.

[35] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33 ITU-T Q6/16, Int. Telecommun. Union, Geneva, Switzerland, Apr. 2001.