A Study of Subjective and Objective Quality Assessment of HDR Videos

Zaixi Shang, Joshua P. Ebenezer, Abhinau K. Venkataramanan, Yongjun Wu, Hai Wei, Sriram Sethuraman, Alan C. Bovik *Fellow, IEEE*

***Abstract*—As compared to standard dynamic range (SDR) videos, high dynamic range (HDR) content is able to represent and display much wider and more accurate ranges of brightness and color, leading to more engaging and enjoyable visual experiences. HDR also implies increases in data volume, further challenging existing limits on bandwidth consumption and on the quality of delivered content. Perceptual quality models are used to monitor and control the compression of streamed SDR content. A similar strategy should be useful for HDR content, yet there has been limited work on building HDR video quality assessment (VQA) algorithms. One reason for this is a scarcity of high-quality HDR VQA databases representative of contemporary HDR standards. Towards filling this gap, we created the first publicly available HDR VQA database dedicated to HDR10 videos, called the Laboratory for Image and Video Engineering (LIVE) HDR Database. It comprises 310 videos from 31 distinct source sequences processed by ten different compression and resolution combinations, simulating bitrate ladders used by the streaming industry. We used this data to conduct a subjective quality study, gathering more than 20,000 human quality judgments under two different illumination conditions. To demonstrate the usefulness of this new psychometric data resource, we also designed a new framework for creating HDR quality sensitive features, using a nonlinear transform to emphasize distortions occurring in spatial portions of videos that are enhanced by HDR, e.g., having darker blacks and brighter whites. We apply this new method, which we call HDRMAX, to modify the widely-deployed Video Multimethod Assessment Fusion (VMAF) model. We show that VMAF+HDRMAX provides significantly elevated performance on both HDR and SDR videos, exceeding prior state-of-the-art model performance.**

***Index Terms*—High dynamic range (HDR), video quality assessment (VQA), HDR VQA database, HDRMAX, full reference (FR) models**

I. Introduction

THE human visual system (HVS) is able to perceive luminance levels between 10^{-6} cd/m^2 and 10^{8} cd/m^2 using various mechanical, photochemical, and neuronal adaptive processes [1]. Traditional imaging and display systems produce content having much narrower ranges of luminance values than the vision system is able to perceive, due to limitations on sensor technology, processing, transmission, bandwidths, and display depths. These older content formats are commonly referred to as Standard Dynamic Range (SDR), and have specifications on brightness, contrast, and color that were originally designed for display on cathode ray tube (CRT) devices [2]. Although CRTs are obsolete, a considerable fraction of content continues to be produced according to SDR specifications. A device that displays SDR content, which has a bit depth of 8 bits/channel, can represent a maximum luminance of 100 nits (1 nit = 1 candela/meter^2) and a minimum luminance of 0.1 nits, using the Rec. 709/sRGB color gamut [3], which covers 35.6% of the CIE 1931 color space.

High Dynamic Range (HDR) is a set of techniques that extend the ranges of luminances and color that can be represented and displayed. "HDR" pictures are sometimes synthesized by combining photographs taken at multiple exposures into a single picture, then tone-mapping it to the 8 bit range that is compatible with SDR displays. What we will refer to as "true HDR" video content is captured using single exposures with advanced sensors, and compatible with HDR displays having wider dynamic ranges and higher average and peak brightness levels. True HDR content has a bit depth of at least 10 bits/channel. HDR10 is an open HDR standard announced by the Consumer Technology Association in 2015 [4] and remains the most widely used HDR format. HDR10 content must have a bit-depth of 10 bits, use the Rec. 2020 [5] color primaries (which cover 75.8% of the CIE 1931 color space), and must apply the SMPTE ST 2084 [6] opto-electronic Transfer Function (OETF) to the linear RGB signals, also known as the Perceptual Quantizer (PQ).

HDR10 has seen increasing adoption over the past few years. Streaming and video hosting services such as Amazon Prime, Netflix, and YouTube now offer content in HDR10. HDR10 is also used as the default standard for UHD Blu-Rays. Major TV manufacturers such as LG, Samsung, and Panasonic support HDR10 content, and manufacturers such as Lenovo and Apple have also recently released laptops that can display HDR10 content. HDR10 is now part of live broadcast and film production workflows and is progressing rapidly into an industry standard.

The adoption of HDR10 has created challenges related to the quality of user experience and the performance of compression algorithms. The increases in bit depth and the use of nonlinear transfer functions in HDR can change the visibility and severity of compression distortions. Being able to measure and control perceptual quality is a critical element of video compression and communication workflows. However, there are few video quality assessment (VQA) models that address the compression of HDR videos. Most existing VQA models can only operate on 8 bit luminance and color data, let alone account for HDR transfer functions and expanded color gamuts. For example, one of the most successful VQA models, the Video Multimethod Assessment Fusion (VMAF) algorithm [7] can be applied to 10 bit data, but it does not take

Zaixi Shang and Joshua Ebenezer contributed equally to this work.

into account the extended luminance range or transfer function of HDR10.

An important consideration is the nonlinear visual response to brightness. Because the vision system is more sensitive to luminance ratios than to absolute brightness values, the perception of differences between luminances is governed by the Weber-Fechner law [8]. The exponential function or "gamma," as specified in the industry standard BT. 709, has been traditionally applied to nonlinearly encode SDR images, but it fails to work with HDR imaging, due to the mismatch of quantization and human perception. Therefore, SDR VQA models, which operate under the assumption of gamma, are less effective on HDR content. This does not imply that SDR VQA models, developed under the assumption of gamma, are always ineffective for HDR content. Several studies, such as [9], [10], have demonstrated that these models can perform competitively even when applied to HDR content, depending on other aspects of the content, or the device it is displayed on, suggesting a nuanced landscape [11], [12]. Furthermore, the perception of brightness distortions is influenced by the viewing conditions, including the image background, the environmental light, the peak luminance, and the dynamic range of the display.

Towards advancing progress in this direction, we created a new HDR video quality database, on which we conducted a subjective quality study of how compression and scaling distortion combinations affect the perceived quality of HDR videos. Currently, there is no publicly accessible VQA database of HDR10 material. There are other HDR content databases, but these are either not publicly accessible, or are based on outdated standards. To address this, we built the first VQA database designed for contemporary HDR. The new "LIVE HDR" database is also the largest HDR VQA database to date. It consists of 310 HDR10 videos created from 31 reference contents that have been distorted by compression and scaling. The videos were presented to 66 human subjects under two ambient conditions using high-speed display hardware in a controlled environment. We conducted an extensive analysis of the collected subjective opinion scores and studied the possible differences in scores gathered under two ambient conditions.

To demonstrate the usefulness of our new psychometric HDR data resource, we used it to compare the HDR quality prediction performances of a number of leading VQA models. In addition, we designed a new framework for conceptualizing and developing new HDR-specific video features, which we call HDRMAX. We used HDRMAX to modify the widely-used VMAF model by supplementing it with HDRMAX features that sensitize it to the expanded luminance ranges, transfer functions, and large color gamuts of HDR video formats. Because of its excellent commercial success and extensive real-world validation, VMAF is an excellent platform to show how HDR-specific HDRMAX features can improve the performance of existing SDR VQA models.

The remainder of the paper is organized as follows. Section II reviews relevant literature on subjective and objective HDR quality studies. Section III details the design of the new LIVE HDR database's source contents, as well as the protocol and implementation of the psychometric study we conducted on the new database. Section IV analyzes the human subjective data gathered from the study, while Section V analyzes the effects of the different ambient environments. Section VI explains the design of new HDRMAX video features, while section VII applies HDRMAX to VMAF, examines its quality prediction performance and compares it against other objective VQA models on the new database. Finally, Section VIII concludes with a forward-looking discussion.

This paper is developed from a conference paper [13]. This paper includes additional details on the subjective study, analysis of the human ratings, the design and evaluation of the new HDRMAX model and the evaluation of other existing VQA models.

II. Related Work

A. Subjective HDR Video Quality Databases

Over the past few years, a number of efforts have been made to create video quality datasets for HDR, but all of these have limited usefulness, either because they have been rendered obsolete by the rapid pace of HDR standard development, or by the inability of authors to publicly release their data owing to copyright issues. Azimi *et al*. [14] conducted a study using 18 human subjects who viewed 5 different 12-bit YUV contents captured by a RED Scarlet-X Camera and afflicted by compression and four other types of distortion, yielding 30 videos. The videos were displayed on a non-standard HDR device the authors designed themselves, supporting the older, more limited BT. 709 gamut, rather than the HDR10 compliant BT. 2020 gamut, and the PQ OETF was not applied prior to compression. Moreover, the videos were of maximum resolution 1920×1080 (1080p), while most current HDR content is 4K. Pan *et al*. [15] conducted a study of the effects of compression on HDR quality using 6 source videos encoded using PQ and HLG and the BT. 2020 color space, but the codec used for compression was AVS2, which has seen little industry adoption. The study included 144 videos that were rated by 22 subjects, but unfortunately none of the video or subjective data has been made publicly available. Baroncini *et al*. [16] conducted a study of 12 compressed HDR videos evaluated by 40 human subjects. The source contents did not follow ITU Rec. BT 2020, the PQ OETF was not applied on the video data, and again, none of the data was made publicly available. Moreover, the resolution of all the videos was 1080p. Rerabek *et al*. [17] conducted a study of 5 HDR videos, each distorted by 4 compression levels, with the aim of comparing objective HDR VQA algorithms, but the data was not made publicly available. The videos were all only of resolution 944×1080, and the data was tone-mapped to 8-bit format before being displayed to the subjects. Athat *et al*. [18] conducted a subjective study of HDR10 content, but none of the data was publicly released because of copyright issues. The authors compressed 14 HDR10 source contents using H.264 and HEVC to generate 140 distorted videos, which were viewed and rated by 51 subjects.

The study that we report here advances the field in several ways: first, all of the source videos are compliant with the most widely used modern HDR standard (HDR10) and include wide

color gamut (WCG) and high frame rate (HFR) videos. Second, the new dataset contains almost twice as many videos as any prior HDR VQA dataset, and more than double the number of collected subjective opinion scores. Third, we conducted the largest and most contemporaneous HDR VQA study on it to date. Fourth, we compared the performances of leading HDR VQA models on it to validate the usefulness of the collected data. Lastly, unlike nearly all of the prior datasets, we are making the LIVE HDR dataset publicly available at http://live.ece.utexas.edu/research/LIVEHDR/LIVEHDR_index.html.

B. Objective Video Quality Assessment Algorithms

Objective VQA algorithms aim to automatically predict the perceptual quality of videos. There are three categories of objective VQA models: full-reference (FR), reduced reference (RR), and no-reference (NR). FR VQA models operate by comparing pristine reference videos against distorted versions of them using perceptually motivated features and/or training data [19], [20]. Reduced reference VQA models use only partial reference information to achieve efficiencies [21], [22], [23], [24]. NR VQA models require no information regarding any reference videos, and instead predict perceptual video quality based only on information extracted form distorted videos [25], [26], [27], [28]. We use the new psychometric HDR VQA database to compare leading HDR VQA models that fall into the FR VQA category. The MSE (or equivalently, the PSNR) has long been used as a basic index of video quality. More recent popular VQA models include Structural Similarity (SSIM) [19], Multiscale SSIM (MS-SSIM) [29], Gradient Magnitude Similarity Deviation (GMSD) [30], most apparent distortion (MAD) [31], visual information fidelity (VIF) [20], and FSIM [32], among others [33], [34], [35], [36]. More recently, machine learning-based FR-VQA frameworks have become quite popular. For example, VMAF [7] combines features from two VQA models, using a Support Vector Regressor (SVR) to map their feature sets to video quality predictions. FR VQA models that rely on deep learning have recently achieved competitive performance, such as DeepVQA [37], and some even use unsupervised deep learning (UDL) [38].

HDR quality prediction research is still a nascent field, and there is only a small literature on the subject. [39] discusses HDR visual quality impairments and efforts at developing dedicated objective HDR video quality metrics. An early algorithm was HDR-VDP [40], which considers the nonlinear response to light of high contrast content and the full range of luminances. An improved version called HDR-VDP-2 [41] uses a model of all luminance conditions derived from contrast sensitivity measurements. Further improvements of HDR-VDP-2 include HDR-VDP2.2 [42], [43]) and HDR-VDP3 [44]. The author of [45] proposed PU, a nonlinear transform to extend normal SDR quality metrics to HDR. Recent developments such as the PU21 encoding function have further refined the field, providing an enhanced methodology for designing quality metrics specific to HDR images [11]. Other authors have focused on the chromatic aspects of HDR video quality by focusing on color fidelity [46], using HDR Uniform Color Spaces [47], and using color difference models [48]. Another method called HDR-VQM utilizes spatio-temporal analysis that simulates human perception [49].

Each of these prior methods has shortcomings. Most of them rely on simple transforms that map video features to quality predictions, such as, the root mean square error (RMSE) used in color difference models, spatial pooling in HDR-VDP-2, or the PU-SSIM and PU-PSNR models proposed in [45]. While these methods are effective on their intended applications, they were primarily designed for legacy HDR videos or HDR images. The modern HDR10 standard, however, introduces several significant changes, including the use of the Perceptual Quantizer (PQ) curve for encoding luminance information, the adoption of the BT.2020 color space, and the inclusion of metadata for accurate display of HDR content. Furthermore, our focus on a video database inherently includes temporal distortions, a factor not present in image databases. Given these changes, it is likely that the reliability of legacy-based quality metrics is reduced when applied to HDR10 content. Therefore, it is necessary to evaluate these existing methods within the context of HDR10 content to ensure their continued relevance and accuracy. Additionally, our study also emphasizes the use of the HEVC codec, which aligns with modern practice. This new codec may introduce different types of distortions, and the visibility of these distortions may also be different, further underscoring the need for evaluation.

III. Subjective Experiment Design

A. HDR Video Contents

We gathered a collection of high-quality, distortion-free HDR10 sequences from [50], [51] and nearly distortion-free content from [52]. These videos were captured by professionals using high-end cinematic HDR video cameras. These sequences were all progressively captured at resolution 3840×2160 with the audio signal removed. The sequences from [50], [51] were captured using Sony F55 or Sony F65 cameras with the dynamic range fixed to the S-log3 profile and are then transformed to PQ EOTF in the post-production process. The videos from [52] were provided in HDR10 format. The videos from [50], [51] have frame rates of 60 frames per second (fps) and those from [52] include both 50 fps and 60 fps. All of the source sequences are HDR-WCG-HFR videos. Following recent studies [53], [54], [55], [56], we segmented all of the video sequences into one or more clips of 7-10 seconds duration. This range was chosen to balance data collection efficiency and maintaining the integrity of the depicted scenes. The 31 source clips were generated from 19 different sources. When clipping the videos, care was taken to avoid awkward interruptions of content and to prevent similar clips from being taken from the same segments, ensuring a more coherent, diverse, and representative set of visual experiences for studying quality assessment.

Fig. 1 shows several sample frames from the source sequences we acquired. The videos span a wide range of contents. We directly applied the spatial information (SI), or integrated Sobel magnitude, and the temporal information (TI), or absolute average frame difference, both defined in [57],

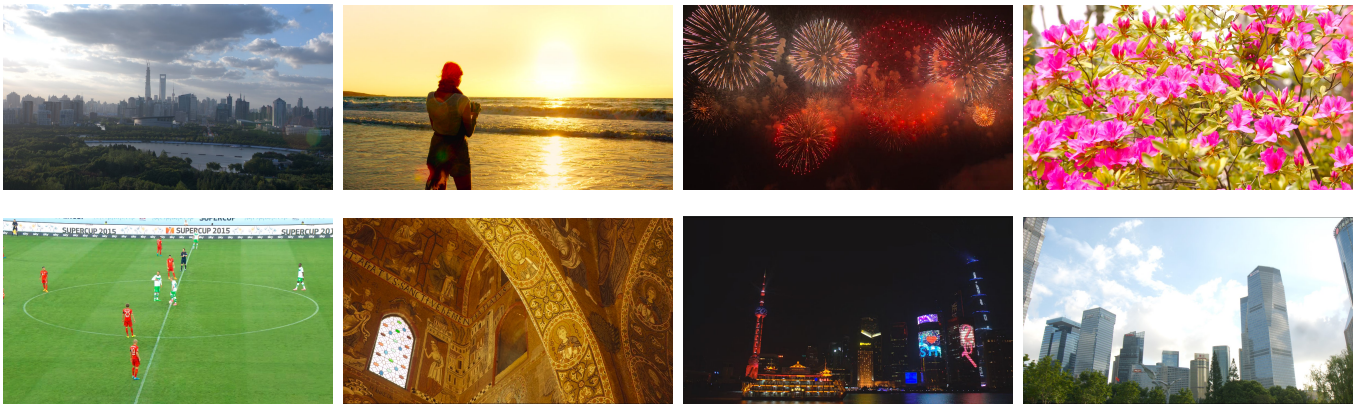


Fig. 1. Exemplar screenshots of frames from source sequences.

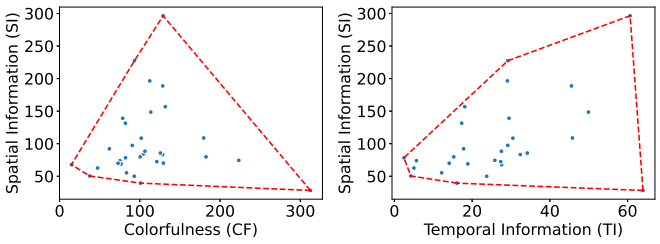


Fig. 2. Spatial Information (SI) versus (a) colorfulness (CF) and (b) Temporal Information (TI), measured on all of the source sequences in the new LIVE-HDR Database. The corresponding convex hulls are plotted by red lines.

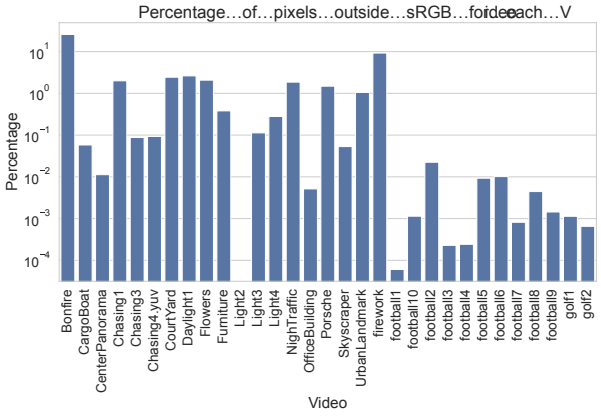


Fig. 3. Proportion of pixels outside of the sRGB color gamut, measured on all of the source sequences in the new LIVE-HDR Database.

to the 10-bit HDR data. Similarly, the colorfulness measure denoted as CF was computed as in [58]. Fig 2 plots the SI, TI, and CF of all of the source sequences in the LIVE-HDR database, indicating wide coverage of low-level content and activity in space and time.

Moreover, we included additional characteristics of the HDR content: min, max, mean, and median luminance, and the portion of pixels outside of the sRGB color gamut. These new metrics, visualized in Figs. 3 and 4, provide further insights into the diversity and coverage of the color and luminance in the HDR videos of our database.

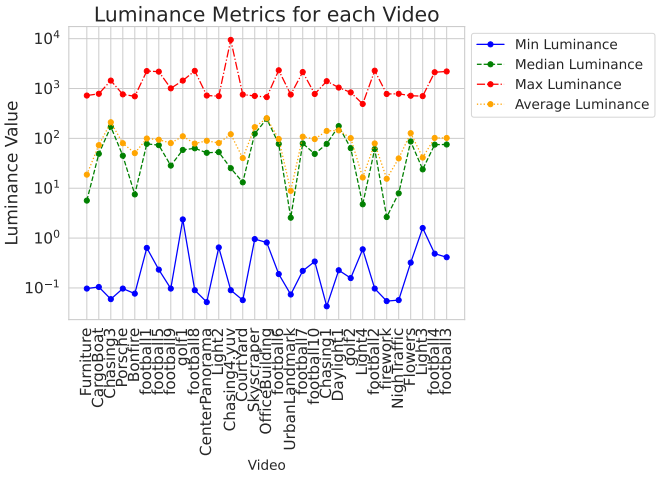


Fig. 4. Min, max, mean, and median luminance metrics measured on all of the source sequences in the new LIVE-HDR Database.

B. Test Sequences

We collected 9 distorted video sequences from each source sequence using the High Efficiency Video Coding (HEVC) Codec. The selection process was subjective but systematic, aiming to ensure that the videos are perceptually distinguishable while spanning a broad range of perceptual qualities. We initially generated a substantial set of videos using a range of bitrates and spatial resolutions, including but extending beyond common settings in the streaming industry. We manually reviewed all the videos and progressively reduced their number to make the total playback duration suitable for our human subjective study. The final bitrate and resolution settings that we used are listed in Table I.

As for the encoding parameters, we used the libx265 encoder in constant bitrate mode with single-pass encoding, which is most commonly used in industrial streaming applications, owing to its simplicity and efficiency. While certain bitrates and resolutions may be less prevalent in practical applications, their inclusion remains advantageous. For instance, a 540p video with a 2.2 Mbps bitrate may exceed those encountered in real-world situations, yet it exemplifies a scenario

TABLE I
BITRATE AND RESOLUTION SETTINGS USED TO CREATE THE DISTORTED VIDEOS.

Number	resolution	bitrate (Mbps)
1	3840×2160	15
2	3840×2160	6
3	3840×2160	3
4	1920×1080	9
5	1920×1080	6
6	1920×1080	1
7	1280×720	4.6
8	1280×720	2.6
9	960×540	2.2

with pronounced scaling artifacts and reduced compression artifacts. Conversely, the 2160p video at 3 Mbps exhibits significant compression artifacts, devoid of any scaling issues. Lastly, the 720p video at 2.6 Mbps represents a confluence of both compression and scaling artifacts. In numerous past studies [59], [60] we have found this approach to be an effective way to cover the distortion space, helping to ensure subsequent model learning. The source videos were included in the database and subsequent psychometric study, to serve as labeled reference videos against which difference mean opinion scores (DMOS) can be calculated. The videos include four practical spatial resolutions. The higher-resolution 4K and 1080p videos were compressed using four and three bitrate targets, respectively, mimicking the bitrate ladders used in HDR video streaming. The videos compressed at the highest bitrate may be observed to present only slightly visible compression artifacts, while the videos compressed to the lowest bitrates exhibit obvious blocking, banding, temporal and scaling artifacts. The 1080p, 720p and 540p videos were all upscaled to 4K resolution when displayed to the human subjects, using bicubic interpolation. This method was selected for its balance between computational efficiency and performance, which minimizes distortion and delay during video playback, thereby maintaining the integrity of the HDR content. The overall video database contains 279 distorted videos and 31 reference videos, yielding a total of 310 videos that were presented to the human subjects.

C. Subjective Testing Design

The human study was conducted in the Laboratory for Image and Video Engineering (LIVE) subjective study room at The University of Texas at Austin. A 65 inch Samsung Class Q90T QLED 4K UHD HDR Smart TV [61] was used to display the HDR content to the participating subjects. The TV was calibrated for HDR by an Imaging Science Foundation (ISF) certified professional using a Calman Calibration kit.

After calibration, the TV had a peak luminance of approximately 1033 cd/m^2, and a minimum luminance below the measurement threshold of 0.7 cd/m^2. Color gamut coverages were 99.88% for BT.709, 88.86% for P3, and 66.33% for BT.2020. Our choice of display sought to mirror what typical consumers currently use in their homes. However, the limited coverage can potentially introduce clipping on HDR10 videos. We plotted the proportion of pixels of each video that fall outside the TV's gamut in Fig. 14 of the supplemental material. It may be observed that this percentage was very small on all but one video ('bonfire'), but even on it most consumer devices would display it similar to the display used in the study. All the measurement was made with a SpectraScan® Spectroradiometer PR-655. It was crucial to ascertain that the TV detected and displayed HDR input correctly, thereby avoiding any unintended tone mapping processes that might introduce distortions. To accomplish this, we made specific configurations and settings adjustments.

First, we enabled the "input signal plus function" in the TV settings, allowing the Samsung TV to receive an extended input signal range and enable HDR input. Subsequently, in the Windows 10 operating system, we activated HDR functionality in the Display settings. Additionally, in the Nvidia Control Panel, we modified the output format to yuv420p and 10-bit depth, while setting the refresh rate at 60Hz. These settings were meticulously reviewed and ensured to remain consistent throughout the entire study. The TV was connected to a workstation having a 12 GB Titan X Graphics Processing Unit (GPU), via an HDMI 2.0b cable allowing for smooth playback of the videos.The Potplayer Video Player with the MadVR renderer was used for playback. In the MadVR settings, we took additional measures to guarantee an authentic HDR viewing experience for the subjects. Specifically, we configured MadVR to pass through HDR content directly to the display. Moreover, we ensured that the "Send HDR metadata to the display" option was enabled. We also used the test pattern in [62] to verify the display. All advanced temporal processing options on the TV were disabled to avoid the introduction of any processing artifacts.

For all the subjects the viewing distance was about $1.5H$, where H is the height of the display. During a session, the subject would watch each video, then see a screen where they were asked to record a quality judgment on the video that they had just seen, using a visible slider on the screen they controlled with their mouse. While the rating scale was continuous, the user was guided by five Likert-like markers placed at uniform intervals labeled as "Bad," "Poor," "Fair," "Good," and "Excellent." The scores given by the subjects were sampled as integers on the interval [0, 100], although numerical values were not made visible to the subjects. In order to prevent bias due to initial positioning of the rating indicator, it would not appear on the sliding scale until the subject placed the cursor on the slider and clicked on it.

The first session shown to each subject was preceded by a briefer training session that presented six exemplar videos of two contents (different from those that followed) that generally spanned the range of distortions that would be seen. For each of the two contents, one reference video and two compressed versions were displayed. All of the training videos were played in a randomized order, each followed by the interactive rating screen, to allow the subjects to become familiar with the overall rating protocol. We utilized the Absolute Category Rating with Hidden Reference (ACR-HR) protocol [57] when displaying the training and test videos, hence the videos shown

in each session were displayed in randomized order. Each subject viewed the videos in a different random order.

D. Ambient Conditions

Two different lighting conditions were used to test the effects of ambient illumination on the perceived quality of HDR content. The first was a dark viewing condition, where the incident illumination on the television was measured to be 5 lux, following the recommendation in [63] for critical viewing of HDR content, and the recommendation in [64] describing general viewing conditions for a subjective study conducted in a laboratory environment. An incandescent table lamp and floor lamp were used to create the light necessary for this environment.

The second ambient condition was illuminated by a pair of yellow-filtered Neewer LED lights to produce an incident illumination on the TV of 200 lux, following the recommendation in [64] for general viewing conditions in a home environment. In this environment, a set of studio LED lights and a 95 W studio compact fluorescent light were placed behind and below the television in order to create a uniform, diffuse ambient illumination. In both environments, the lights were positioned so that their reflections off the television would not be visible to the viewers. The incident luminance on the TV was measured by a Dr. Meter LX1330B luxmeter.

E. Subjects

A total of 66 human subjects were recruited from the student population at The University of Texas at Austin. Each subject participated in two sessions separated by at least 24 hours. The subjects were divided into two groups, one for each ambient condition. Hence 33 subjects watched the videos in the darker environment and 33 watched the videos in the brighter environment. No subject was given any information about the ambient conditions. We applied the Snellen and Ishihara tests of test each subject's visual acuity and color perception, respectively. One subject was found to have a color deficiency, but no subjects had less than 20/30 visual acuity on the Snellen test, when wearing their corrective lenses (if needed). The color deficient subject was not rejected from the study following our common practice of promoting a more realistic subject pool, as explained on our website [65].

IV. Processing of Subjective Scores

There are a number of ways in which subjective scores can be converted into Mean Opinion Scores (MOS). We computed MOS as the average of subjective scores given by subjects (MOS), the average of z scores (ZMOS), and we also computed MOS using the statistical method proposed in [66].

A. MOS

Let i_d index those subjects that viewed videos in the dark environment, and i_b index the subjects who viewed the videos in the bright environment. MOS is calculated as the average of the scores given by a set of subjects, in [67]. We will also define separate MOS values for the dark and light environments. Let the scores given by a subject i_k on video j be $s_{i_k j}$. We will refer to the MOS of a video j whose scores were collected under the darker (brighter) ambient conditions as the respective average scores given under each condition: MOS_{dj} and MOS_{bj}, where

$$MOS_{kj} = \sum_{i_k=1}^{S_k} s_{i_k j}, \tag{1}$$

for $k = d, b$ (dark, bright), and $j = 1, 2 \dots N$.

B. ZMOS

We also define MOS calculated as the average of the z scores [68], given by

$$z_{i_k j} = \frac{s_{i_k j} - \mu_{i_k}}{\sigma_{i_k}} \tag{2}$$

for $k = b, d$, where the subjects under dark (bright) conditions are indexed $i_d = 1, 2 \dots S_d$ ($i_b = 1, 2 \dots S_b$) when rating videos indexed $j = 1, 2 \dots N$. In our database, $S_d = 33$, $S_b = 33$ and $N = 310$. In (2), μ_{i_k} and σ_{i_k} are the mean and standard deviation of the scores given by subject i_k across all videos:

$$\mu_{i_k} = \frac{\sum_{j=1}^{N} s_{i_k j}}{N} \tag{3}$$

and

$$\sigma_{i_k} = \sqrt{\frac{\sum_{j=1}^{N} (s_{i_k j} - \mu_{i_k})^2}{N}}. \tag{4}$$

Since there are two ambient conditions, for each video $j = 1, \dots, N$ we will refer to the MOS calculated from scores that were collected under darker (brighter) ambient conditions as $ZMOS_{dj}$ and $ZMOS_{bj}$, respectively, where

$$ZMOS_{kj} = \sum_{i_k=1}^{S_k} z_{i_k j} \tag{5}$$

for $k = d, b$ (dark, bright).

C. Consistency Analysis

We studied the internal consistency of the scores as follows. We randomly partitioned the subjects who participated under each ambient condition into two approximately equal sized groups and computed the correlations between the mean MOS computed separately from the two groups over 100 random divisions. We then computed the correlation across the 100 splits. As expected, the internal consistency of the $ZMOS$ was better than that of MOS. We applied the outlier rejection method suggested by ITU Rec. BT 500.11 on both the MOS and $ZMOS$, separately for each ambient condition. However, we found that the internal correlations did not improve when the outliers were removed, as shown in Table II. We also examined the scores of the color-deficient subject, and found that his scores correlated more highly against the other subjects who participated under the same ambient condition (0.88) than the average correlation between individual scores and group scores (0.82). In our analysis, we therefore chose not to remove the outliers when conducting the subsequent statistical analysis.

TABLE II
CONSISTENCY ANALYSIS OF THE SUBJECTIVE DATA.

	Correlations before ITU BT 500.11 outlier removal.	Number of outliers according to ITU BT 500.11.	Correlations after ITU BT 500.11 outlier removal.
MOS_d	0.9481	0	0.9481
MOS_b	0.9528	2	0.9492
$ZMOS_d$	0.9636	7	0.9581
$ZMOS_b$	0.9669	6	0.9665

D. SUREAL Scores

A number of deficiencies in the ITU BT 500.11 outlier removal method have been observed in [66], along with an improved method called SUREAL that finds a Maximum Likelihood (ML) estimate of the scores. Using this method, represent the opinion scores $s_{i_k j}$ as random variables $S_{i_k r}$

$$S_{i_k j} = \psi_{kj} + \Delta_{i_k} + \nu_{i_k} X, \tag{6}$$

where ψ_{kj} is the true quality of video j under ambient condition k, Δ_{i_k} represents the bias of subject i_k, the non-negative term ν_{i_k} represents the inconsistency of subject i_k, and $X \sim N(0,1)$ are i.i.d. Gaussian random variables. The quantities $\psi_{kj}, \Delta_{i_k}, \nu_{i_k}$ are estimated by computing the log-likelihood of the observed scores, using the Newton-Raphson method to solve for the values of $\psi_{kj}, \Delta_{i_k}, \nu_{i_k}$ that maximize the log-likelihood. We plotted the estimated subject biases in Fig. 15 and their inconsistencies in Fig. 16 in the supplementary material. It may be observed that both the subject biases and inconsistencies are quite dispersed. In this way, subject biases are accounted for when estimating the true qualities ψ_{kj}, and the method is robust against subject inconsistencies.

V. EFFECT OF AMBIENT ILLUMINATION

We used all three types of summary subjective opinion scores to analyze the effects of ambient illumination on impressions of quality. It is worth noting that the MOS and SUREAL scores preserve the differences between the absolute values of the scores under the two ambient conditions, while the $ZMOS$ scores do not, since they are normalized. The distributions of MOS, $ZMOS$, and SUREAL are shown in Fig. 5. The MOS and SUREAL values under each of the ambient conditions cover a wide range, and it may be observed that the overall distributions of scores under the two ambient conditions are similar. Since SUREAL and MOS are absolute scores, one may deduce from Fig. 5 that the videos watched under darker ambient conditions were rated as being of slightly higher qualities than those watched under bright ambient conditions. The same conclusions cannot be drawn regarding $ZMOS$, which is a normalized score, suggesting that these results reflect a slight preference for viewing under the darker conditions, but the relative ratings remain largely unaffected. Fig. 6 plots MOS against spatial resolution and bitrate. It may be observed that the MOS recorded under both ambient conditions fell in similar ranges for each spatial resolution and bitrate combination, but the MOS recorded under brighter conditions were slightly lower than under darker conditions at most resolution and bitrate settings. These differences, however, were more pronounced at lower bitrates and resolutions.

To assess the possible significance of the differences that we observed in Fig. 6, we conducted Welch's two-sided t-test on the MOS under both ambient illumination settings. We compared the MOS at each resolution and bitrate setting, obtaining the p-values shown in Table III. As may be seen, none of the resolution and bitrate combinations yielded a p-value less than 0.05, indicating that, while differences may be discerned between the MOS obtained under the two different illumination settings, these differences were not statistically significant. Separately, we also tested the raw (non-averaged) scores that were recorded by the individual subjects under the two ambient conditions. From among 310 labeled videos, only 17 were associated with differences in quality judgments that were statistically significant.

We further investigated the influence of ambient illumination on perceived video quality through a permutation test as outlined in [69]. Despite 17 videos showing statistically significant differences in mean scores under different viewing conditions in our initial t-test analysis ($D = 17$), we sought to examine whether this could occur by chance. In the permutation test, subjects were randomly divided into two groups and mean scores for each video were recalculated. A paired t-test was then executed for each video. This process was replicated 10,000 times to construct a distribution of counts of significant differences, D', under random group assignment.

For ambient illumination to be considered significant, it must satisfy $\Pr(D' < D) \geq 0.95$. Our analysis revealed that the 95th percentile of the D' distribution was 41, greater than observed $D = 17$, leading to the conclusion that differences between bright and dark conditions were not statistically significant. The D' distribution, observed D, and the 95th percentile are shown in Fig. 7, illustrating the lack of significant impact of the ambient illumination on video quality ratings.

Further, we calculated the average luminances of each video which does not depend on the illumination. Fig. 8 shows a scatter plot of the p-values of videos in the raw score comparisons against the computed average luminances. There was no clear tendency of p-values against the average luminance. Indeed, the Pearson's correlation coefficient between the p-values and the average luminances were essentially nil (0.03).

We also used the confidence intervals of the SUREAL scores to study the effects of ambient illumination. The SUREAL method provides 95% confidence intervals on the subjective scores using the Cramer-Rao bound. The values of

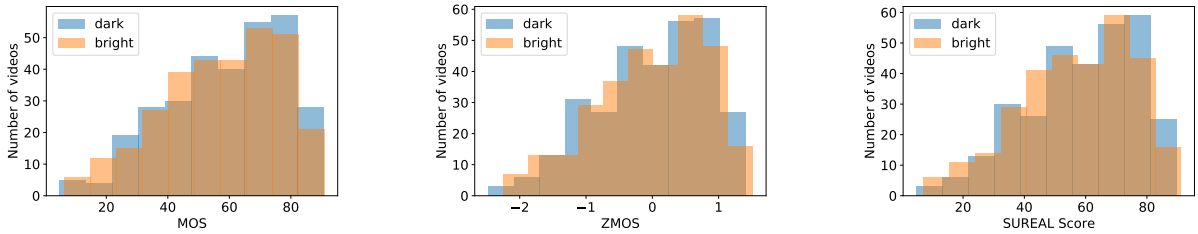


Fig. 5. Histograms showing distributions of MOS, $ZMOS$, and SUREAL scores.

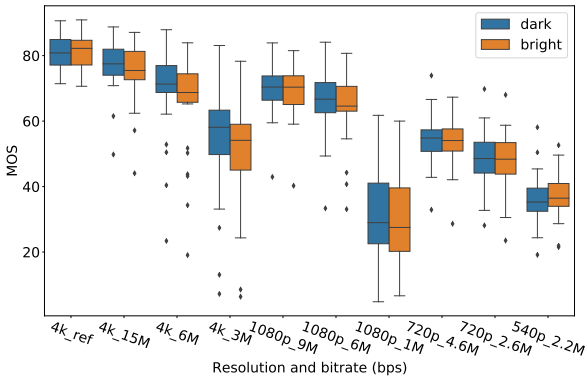


Fig. 6. A box plot showing the distribution of MOS under two ambient illumination settings for each distortion combination.

TABLE III
THE P-VALUE OF EACH BITRATE AND RESOLUTION SETTINGS FOR THE DISTORTED VIDEOS.

Number	resolution	bitrate (Mbps)	p-value
1	3840×2160	ref	0.5987
2	3840×2160	15	0.1539
3	3840×2160	6	0.1750
4	3840×2160	3	0.1538
5	1920×1080	9	0.3422
6	1920×1080	6	0.2856
7	1920×1080	1	0.3105
8	1280×720	4.6	0.4361
9	1280×720	2.6	0.3645
10	960×540	2.2	0.7095

ψ_{dj} and ψ_{bj} are plotted in Fig. 17 in the supplementary material. We found that for 10 of the 310 videos, the confidence intervals did not overlap, indicating statistically significant differences. We also computed the 95% confidence intervals of the MOS (assuming normality) and plotted the scores and their confidence intervals in Fig. 18 in the supplementary material.

VI. OBJECTIVE VIDEO QUALITY MODEL DESIGN

The goal of our model design is to find features that are expressive of distortions that are more noticeable in HDR videos. As compared to SDR videos, HDR videos contain lower black levels, higher peak luminances, and more brilliant

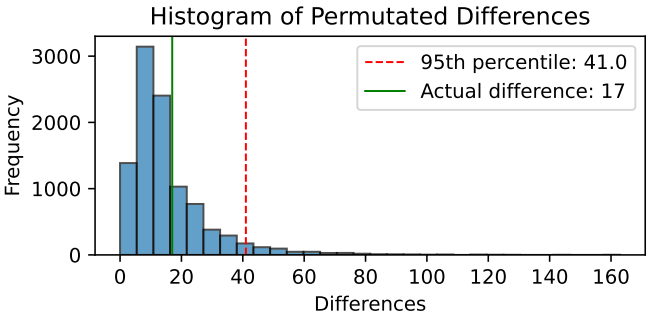


Fig. 7. Distribution of significant differences (D') under random group assignment for the permutation test. The observed value $D = 17$ and the 95th percentile of the D' distribution are also shown, indicating that the observed differences in scores under bright and dark ambient conditions are not statistically significant.

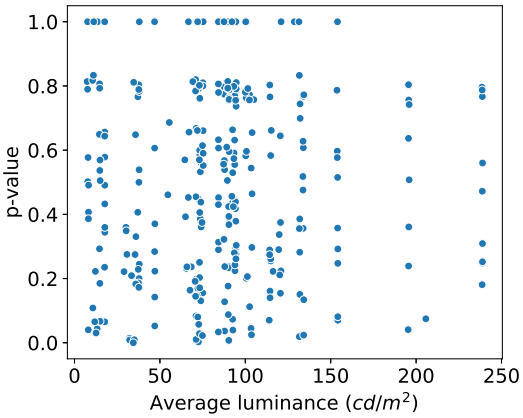


Fig. 8. Scatter plot of the p-values of the raw score comparison against the average luminances of each video .

colors. Rich visual information, and visible distortions, can be observed in the dark and bright zones, both affecting subjective quality; however, conventional SDR VQA models have difficulty capturing this information.

The reason for this is that the responses of conventional VQA feature sets are dominated by, or at least strongly affected by distortions on regions that are "SDR-like," *i.e.*, occupying the mid-range of brightnesses. The feature responses to very dark and bright regions become dilute, greatly reducing the sensitivity of standard VQA models to highly conspicuous "HDR" distortions. Moreover, the visual response

to luminance is highly nonlinear. The visual system is able to map large ranges of luminances onto much smaller ranges of perceived lightness, thereby achieving a high degree of compression [70]. For a distortion of fixed magnitude, the Weber ratio of luminance is higher on dark and is reduced as the luminance increases. Thus, small changes in luminance in dark regions will be more noticeable than in bright regions.

Because of these reasons, distortions on the darkest and brightest areas have distinct perceptual responses and contributions to perceptual quality. The perceptual distortion information in these areas is not effectively captured by conventional VQA feature sets. Thus, we introduce additional feature computation pathways to capture "HDR-specific" features in parallel with the traditional "SDR" features, to better account for perceived distortions in these areas.

Specifically, we introduce HDRMAX, a simple but effective way to process bright and dark regions separately, and computing HDR-aware quality features on them, while avoiding complicated computations such as image segmentation. Instead, we define a pair of nonlinear transforms that expand the luminance ranges of very dark and bright regions, at the expense of the mid-range, which effectively amplifies the impact of "HDR" distortions on VQA feature responses. Following the transforms, we define separate and parallel feature extraction paths, to drive the quality-aware features specific to each of the areas, so that features computed on the nonlinearly altered frames can be used to augment conventional SDR VQA features.

A. Double Exponential Nonlinearity

The main characteristic of the nonlinear transform is to stretch the brightness values near the minimal (darker) and maximum (lighter) values, thereby enhancing the contrast there.

Neural responses are adequately modelled as sigmoidal functions [71]:

$$R = R_{max}\frac{I^n}{I^n + I_s^n}, \tag{7}$$

where R is the response to an input signal I, R_{max} is a maximum response, I_s is a semisaturation constant, and n depends on the type of neuron, but usually falls in the range $[1, 2]$ [72]. The sigmoidal function has the greatest slope for the smallest input magnitudes, gradually decreasing as the input increases.

We selected an exponential functions as a simple and effective way to amplify the brightness values at the extreme ends of the dynamic range in a nonlinear fashion, while gradually compressing the mid-range brightness values. This choice was guided by the simplicity of an exponential function's form and the control it provides over the degree of expansion through its parameters. The numerical stability it offers also contributed to its selection. While we do not claim that it accurately models the perceptual response, its use is quite perceptually relevant to VQA model design. The reason is that it is making perceptually relevant distortion information more available to VQA algorithms. It does this in a way that is copacetic with theories of distortion-sensitive natural video statistics. In this sense it may be viewed as a pooling preprocessing step that can remedy the defects of current learning-based VQA models. Since it is not meant to model a biological perceptual process, there may be other functional forms that are as effective, or more so, but our choice is a simple one. Moreover, HDRMAX incorporates a local adaptation operation, a process fundamental to vision, facilitating adjustment to a wide range of brightness values. Local adaptation adjusts the sensitivity of the visual system based on the local luminance level, acting specifically on each region of the retina [73]. A refined model of this process, building upon the Naka-Rushton equation, has been proposed to simulate the physiological adaptations of the retina. Particularly, it modifies the half-saturation parameter, depending on the local luminance level. Inspired by this model of local brightness adaptation, we integrated a mean debiasing operation into HDRMAX. This operation precedes the exponential transform, its purpose being to adjust the nonlinearity based on the local mean luminance, thereby preserving sensitivity across different local luminance levels within each frame.

In the context of the HDRMAX augmentation, the mean debiasing operation is positioned before the input into conventional SDR VQA models. This reflects the local adaptation model that simulates the initial stages of visual processing in the retina. Implementing this operation before later stages of the visual pathway modeled by existing SDR VQA models aligns with the natural flows of visual processing. As a result, HDRMAX ensures that the local nonlinear operation maintains sensitivity and responsiveness across varying local luminance levels.

The basic goal of HDRMAX is to address the inability of conventional SDR VQA models to capture some HDR distortion characteristics. Our method makes better available distortion information in the extreme range of luminance and color that are highly visible but not well accessed by current VQA models such as VMAF. We do this by introducing a separate processing pathway that expands the extreme ends of the dynamic range. This is accomplished by introducing an expansive nonlinearity whose outputs are nicely analyzable using natural video statistics model.

The nonlinearities are applied on the perceptually uniform PQ-encoded luma. The nonlinearities, while designed to capture aspects of relative luminance perception, are only a simple model of it. The interplay between the PQ transfer function and the exponential functions is intricate and not easily explained from a perceptual point of view. An advantage of applying the nonlinearities on perceptually uniform luma is that it allows for predictable modifications to video content. This predictability enables a clear understanding of how the nonlinearities stretch or compress bright/dark regions, providing a greater level of control over the quality assessment process.

Assume that the brightness values $I(x, y, t)$ fall within the range $[0, 1]$. If they don't, linearly scale the brightness range $[A, B] \mapsto [0, 1]$, where A and B represent the minimum and maximum brightness values within each frame, respectively. This scaling operation aligns the dynamic range of each frame's brightness values with the $[0, 1]$ interval, while controlling the strength of the applied exponential function,

maintaining uniformity across each frame and avoiding extreme values. While this approach to normalization diminishes the link to absolute luminance, it isolates the local contrast, which is highly relevant to visual distortion perception. We then apply point operations on the scaled brightness values, with the goal of nonlinearly expanding the dynamic ranges of the extreme high (bright) and dark ends. Once these operations are applied, feature extraction is conducted in the same way on two nonlinearly transformed frames, and the original frame. The nonlinear transformations are adaptive, since it includes local mean debiasing. The two nonlinearly transformed videos are given by:

$$\widetilde{I}_1^l(x,y,t) = exp[\delta_1(I(x,y,t) - \bar{I}^l(x,y,t))], \tag{8}$$

and

$$\widetilde{I}_2^l(x,y,t) = exp[-\delta_2(I(x,y,t) - \bar{I}^l(x,y,t))]. \tag{9}$$

The parameters $\delta_1, \delta_2 \in 0.5, 1, 2, 5$ in equations 8 and 9 control the expansion strength in the bright and dark areas, respectively. This choice, akin to a log grid search, offers a balance between model complexity and computational feasibility, and appropriately captures the inherent data patterns.

These parameters help modulate the representation of HDR details in dark and bright regions. Extreme values could lead to under-detailed or unnaturally contrasted images, emphasizing the need for careful selection of these parameters. In the experiments, we fixed $\delta_1 = 0.5$ and $\delta_2 = 5$ but we discuss these choices and how performance varies with them in the performance evaluation section. $\bar{I}^l(x,y,t)$ is the local mean brightness estimate:

$$\bar{I}^l(x,y,t) = \sum_{k=-K}^{K} \sum_{j=-L}^{L} w_{k,l} I_{k,l}(i,j), \tag{10}$$

where $w = \{w_{k,l} | k = -K, ..., K, l = -L, ..., L\}$ is a 2D circularly-symmetric unit-volume Gaussian weighting function sampled up to 3 standard deviations away from the mean. We used $K = L = 31$ in our experiments and we discuss the choice of the parameter later in the performance evaluation section.

We show plots of the exponential transforms in Fig. 9, illustrating the expansion of the extreme dark and bright ranges. We use separate transformations, because it allows flexibility when accessing information at the bright and dark ends. For example, we assume throughout that the luma values are expressed as luma, rather than luminance. In most HDR streaming video workflows, the PQ OETF is applied to the linear luminance signals received by RGB sensors to convert them to nonlinear color R'G'B', which are then weighted and summed to compute luma and color-difference channels ($Y'C'_BC'_R$, sometimes referred to as YUV.) The nonlinearities (8)-(10) are flexible enough to be used either on luma or on luminance, the latter of which has already been transformed by an asymmetric nonlinearity.

Two sample reference frames taken from the 'flower' and the 'firework' videos, as well as the result of applying the nonlinear transformations to the 'flower' and 'firework' frames, are shown in Fig. 10. The 'flower' video frame contains

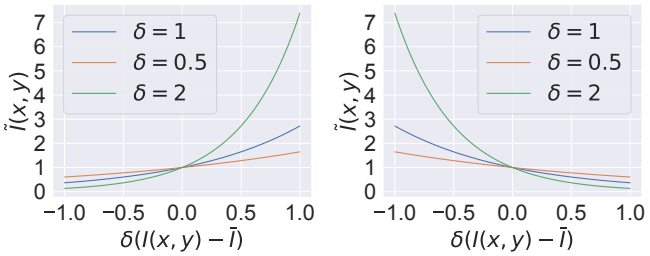


Fig. 9. The two exponential transforms in (8) (left) and (9) (right) plotted for several values of the expansion parameters δ_1 and δ_2.

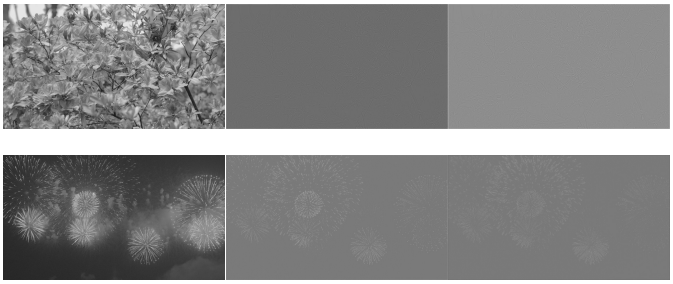
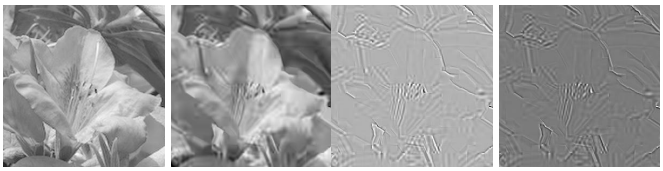


Fig. 10. The reference frames 'flower' and 'firework' (left), the transformed reference frames after processing with (8) (middle) and (9) (right).



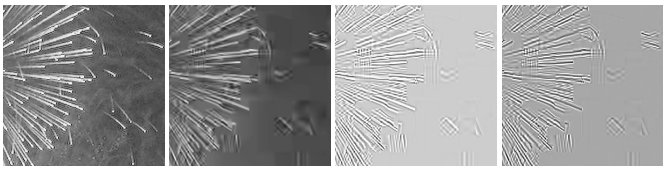
(a) (b) (c) (d)

Fig. 11. A patch from 'flower'. (a) from the reference frame; (b)-(d) from the compressed frame. (b) before nonlinear transformation; (c) after nonlinear transformation (8); (d) after nonlinear transformation (9).

areas containing mostly mid-range brightness values, while the 'firework' video frame contains very bright areas on a very dark background. As such, the nonlinearly processed 'firework' video will contain more heavily enhanced areas. Of course, these printed representations are not HDR and are being shown to give an idea of the applied effects. To illustrate the effects on distortion visibility, we also show magnified areas of 'flower' and 'firework' before and after compression and with nonlinearities applied in Fig. 11 and Fig. 12. To demonstrate the amplification of distortions on the bright areas, we also show the result of applying transformation (8) and (9). As may be observed, application of the nonlinear transformation greatly enhances the distortions in the bright regions of 'firework,' and less so on the mid-range distortions in 'flower.'

B. Modifying VMAF Using HDRMAX Features

VMAF is a data driven video quality framework that extracts several highly successful VQA features, then uses a trained SVR to map the features to human judgments. The features



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

Fig. 12. A patch from 'firework'. (a) from the reference frame; (b)-(d) from the compressed frame. (b) before nonlinear transformation; (c) after nonlinear transformation (8); (d) after nonlinear transformation (9).

TABLE IV
DESCRIPTIONS OF FEATURES

Feature index	Description
$f_1 - f_5$	VIF and DLM features from the original frame.
f_6	Motion feature
$f_7 - f_{16}$	VIF and DLM features from the frames following the nonlinear transformation.

used in VMAF 2.3.0 include the Detail Loss Metric (DLM), four Visual Information Fidelity (VIF) features computed on different oriented frequency bands, and a simple frame difference feature, all of which are applied on the PQ luma component only. Modifying VMAF to include HDRMAX features is quite simple. On the brightness component of each video frame, also compute the nonlinearly transformed frames $\widetilde{I}_1^l$ and $\widetilde{I}_2^l$, along with the usual VMAF features computed on I. Table IV summarizes the features used.

VII. Objective Video Quality Assessment Experiments

As a way of demonstrating the usefulness of the new LIVE HDR Database, we used it to study the performance of several existing HDR VQA models, as well as state-of-the-art (SOTA) SDR VQA models. We also studied the performance of VMAF augmented by HDRMAX features as its parameters were varied.

A. Evaluation Criteria

We used the SUREAL scores owing to their statistical reliability. Since they are absolute quality scores, we obtained quality differences referred to as difference MOS (DMOS). Given a video indexed j with SUREAL score ψ_{dj}, compute the difference score

$$D\psi_{dj} = \psi_{dj}^{ref} - \psi_{dj}. \tag{11}$$

The performances of the compared algorithms, including VMAF+HDRMAX, were evaluated using three standard metrics: the Spearman's Rank Order Correlation Coefficient (SROCC), the Pearson Linear Correlation Coefficient (PLCC), and the Root Mean Square Error (RMSE). Following common practice [74], we fit the predicted scores to the real scores using a logistic function

$$f(s) = \beta_1(\frac{1}{2} - \frac{1}{(1+\exp(\beta_2(s-\beta_3))}) + \beta_4 s + \beta_5 \tag{12}$$

before computing the PLCC and the RMSE.

B. Evaluation Protocol

We used an SVR to learn the mappings from features to DMOS. The SVR was implemented using the linear kernel. All of the compared algorithms were evaluated using 1000 random train-test splits. On each split, 80% of the data was used for training, and the other 20% for testing, while not allowing any sharing of content between training and testing subsets. Notably, the new dataset includes several videos derived from the same longer clips, specifically, the football videos (football 1-8) and golf videos (golf 1-2). We diligently ensured that these videos were not split between the training and testing sets, to avoid any potential leakage of similar content between the sets. We applied 5-fold cross-validation to find the optimal SVR parameters for each training set.

C. Performance Evaluation of VMAF+HDRMAX

We tested the performance of VMAF+HDRMAX against different choices of the expansion parameters δ_1 and δ_2. For each parameter combination, we computed the 16 features in Table IV, on the LIVE HDR Database and conducted 1000 train-test splits. The median values of the obtained performance metrics SROCC, PLCC and RMSE are given in Table V. For better visualization, a heatmap of the SROCC as the parameters δ_1, δ_2 were varied is shown in Fig. 13. As may be observed, smaller values of δ_1 and larger value of δ_2 generally resulted in higher SROCC, while $(\delta_1, \delta_2) = (0.5, 5.0)$ yielded the best SROCC. One possible explanation for this is that HDR10 videos extend the original SDR luminance range from 0.01-100 nits to 0.0001-10000 nits. The difference between the darkest blacks of SDR and HDR is much less than between the brightest SDR and HDR values, suggesting that greater expansion is required on the darker end. However, although the choice of the parameter selection does influence the measured model efficacy, the differences are not large, and every choice and combination resulted in excellent performance relative to other, prior models. This demonstrates the efficacy of the nonlinear transformation and HDR features.

We also conducted experiments on the patch size W used in transformation (8) and (9). The results for $W = 9, 17, 31$ and 63 are reported in Table VI using $\delta_1 = 0.5$ and $\delta_2 = 5$. We avoided W values that are multiples of 4 to avoid alignments of the transformation window edges with compression block boundaries. The choice of window size had a minor effect on performance, but we chose the one giving the highest degree of correlation between predicted quality against human judgments.

We also studied other design choices. First, we extended the nonlinear transformation to the components of three color spaces: the BT.2020 RGB color space, the YC_BC_R [5] color space, and the $HDR - Lab$ [75] color space. The RGB

TABLE V
PERFORMANCE OF LUMA VMAF+HDRMAX AS THE EXPANSION PARAMETERS δ_1 AND δ_2 VARIED, FOR USING THE NONLINEAR TRANSFORM (8)-(10). THE TOP PERFORMING COMBINATION IS BOLDFACED.

δ_1	δ_2	SROCC	PLCC	RMSE
0.5	0.5	0.8470	0.8056	11.9296
0.5	1	0.8238	0.7918	11.4521
0.5	2	0.8610	0.8167	10.8815
0.5	5	**0.8755**	**0.8397**	**10.1410**
1	0.5	0.8516	0.8099	11.6104
1	1	0.8500	0.8125	11.5388
1	2	0.8628	0.8303	10.9217
1	5	0.8584	0.8213	11.2416
2	0.5	0.8335	0.7861	11.5870
2	1	0.8282	0.7953	11.5907
2	2	0.8433	0.8200	10.1993
2	5	0.8540	0.8268	10.1404
5	0.5	0.8378	0.8003	11.8099
5	1	0.8422	0.8086	11.3328
5	2	0.8203	0.8081	11.6327
5	5	0.8216	0.7958	11.6869

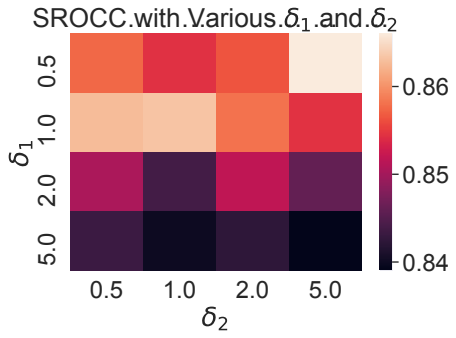


Fig. 13. A heatmap visualizing median SROCC as (δ_1, δ_2) are varied for the nonlinear transformation (8)-(10).

space is associated with acquisition and display. YC_BC_R is a common format for HDR videos. In $HDR-Lab$, the $L*$ component captures the perceived lightness of a color as compared to a white reference. The $a*$ and $b*$ components represent the position of the color between red/magenta and green, yellow and blue respectively. For each variant model, we extracted the original six VMAF features on each channel, and also extracted the four VIF features and the DLM feature on the nonlinearly transformed frames of each component. Thus, each color variant of VMAF+HDRMAX utilizes 46 features. As a final comparison model, we applied the nonlinearity (8)-(10) on the linear luminances instead of the PQ luma values, but without any color components. The performance results for these four variants of HDRMAX are shown in Table VII. The results for all models were quite good, but not as high

TABLE VI
PERFORMANCE OF THE NONLINEAR TRANSFORM FOR VARIOUS WINDOW SIZES. TOP PERFORMANCE IS BOLDFACED.

W	SROCC	PLCC	RMSE
9	0.8601	0.8265	11.1056
17	0.8552	0.8354	11.0654
31	**0.8755**	**0.8397**	**10.1410**
63	0.8675	0.8205	11.1852

TABLE VII
PERFORMANCE OF COLOR VARIANTS OF VMAF+HDRMAX. THE "SETTING" COLUMN INDICATES THE COLOR SPACE. "LINEAR" INDICATES THE TWO-EXPONENTIAL TRANSFORM AND FEATURES ARE PERFORMED ON THE LINEAR LUMINANCE VALUES. THE TOP PERFORMANCE IN EACH DOMAIN IS BOLDFACED.

Setting	SROCC	PLCC	RMSE
$HDR-Lab$	0.7850	0.7348	14.3641
RGB	0.7986	0.7477	13.3448
YC_BC_R	0.8025	0.7502	13.7340
linear	**0.8355**	**0.8068**	**11.3307**

as for the luma-only VMAF+HDRMAX results. Since the database contains videos that have excellent color diversity and coverage, this suggests that most of the distortion artifacts can be captured and analyzed within the luma channel, while increasing the dimension of the feature space slightly reduces the model performance.

D. Comparison Against Other VQA Models

We also evaluated several other FR HDR and FR SDR VQA models on the new database and compared them against the VMAF+HDRMAX. The existing HDR algorithms we studied are the latest PU21 enhanced models, including PSNR, SSIM, MS-SSIM, FSIM [32] and VSI [76], HDR-VDP2.2, HDR-VDP3, and HDR-VQM, while the compared SDR methods are PSNR, SSIM, MS-SSIM, STRRED, SpEED-QA, and VMAF. Most of these models are not trained. We listed both the pre-trained and retrained VMAF for comparison. The results of the comparison are shown in Table VIII and Table IX against the DMOS obtained from the dark environment and bright environment respectively. It may be seen that VMAF modified using HDRMAX was able to significantly outperform the other models, including retrained VMAF. The fact that VMAF+HDRMAX outperforms VMAF by a large margin implies that the unmodified VMAF largely captures distortions from the usually dominant mid-range of brightness.

To further substantiate our claim, we performed a one-sided t-test for statistical analysis. For each model, we used 1000 SROCC values, obtained from individual train-test splits. In the case of models that do not require training, we randomly selected a 20% video sample to calculate a comparable SROCC sample. The single-sided t-test was then performed on the SROCCs between our proposed VMAF+HDRMAX method and the rest of the models, under both bright and

TABLE VIII
PERFORMANCE OF THE COMPARED HDR AND SDR QUALITY MODELS EVALUATED USING THE SCORES FROM THE DARK ENVIRONMENT. THE TOP PERFORMANCE IS BOLDFACED.

Method		SROCC	PLCC	RMSE
SDR Quality Models	PSNR	0.5798	0.6229	13.6735
	SSIM	0.4982	0.4925	15.2124
	MS-SSIM	0.5139	0.5252	14.8741
	STRRED	0.5670	0.5506	14.5913
	SpEED-QA	0.5716	0.5685	14.6258
	VMAF (original)	0.7628	0.7492	12.2953
	VMAF (retrained)	0.7940	0.7679	11.4522
HDR Quality Models	HDR-VDP2.2	0.5868	0.5128	15.0052
	HDR-VDP3.0.7	0.7363	0.7307	11.9332
	HDR-VQM	0.5543	0.5450	14.3890
	PU21-PSNR	0.5841	0.5767	14.2798
	PU21-SSIM	0.6019	0.6065	13.8971
	PU21-MSSSIM	0.6593	0.6564	13.1868
	PU21-FSIM	0.6470	0.6372	13.4705
	PU21-VSI	0.6795	0.6667	13.0284
	VMAF+ HDRMAX	**0.8755**	**0.8397**	**10.1410**

TABLE IX
PERFORMANCE OF THE COMPARED HDR AND SDR QUALITY MODELS EVALUATED USING THE SCORES FROM THE BRIGHT ENVIRONMENT. THE TOP PERFORMANCE IS BOLDFACED.

Method		SROCC	PLCC	RMSE
SDR Quality Models	PSNR	0.6268	0.6621	13.0476
	SSIM	0.5493	0.5406	14.6461
	MS-SSIM	0.5740	0.5831	14.1442
	STRRED	0.6373	0.6167	13.7048
	SpEED-QA	0.6435	0.6254	13.6944
	VMAF (original)	0.8184	0.7947	11.0224
	VMAF (retrained)	0.8133	0.7890	11.0915
HDR Quality Models	HDR-VDP2.2	0.6472	0.6254	13.9861
	HDR-VDP3.0.7	0.8080	0.8098	10.2139
	HDR-VQM	0.6315	0.6144	13.5114
	PU21-PSNR	0.6117	0.5963	13.9762
	PU21-SSIM	0.6403	0.6301	13.5188
	PU21-MSSSIM	0.7120	0.6969	12.4859
	PU21-FSIM	0.7116	0.6904	12.5951
	PU21-VSI	0.7290	0.7058	12.3334
	VMAF+ HDRMAX	**0.8693**	**0.8256**	**10.6864**

dark conditions. The details of these t-test analyses can be found in Table X. It may be observed that the SROCC values for pretrained and retrained VMAF appear to be similar in Tables VIII and IX, but show some difference in Table X. This minor difference arises from the fact that we sample 20% of the videos for the pretrained VMAF in the process of t-test, leading to slightly varied SROCC values obtained from these samples. This provided statistical evidence of our method's superior performance, with all p-values below the threshold of 0.05, denoting statistical significance.

E. Evaluation on SDR Database

We also trained and evaluated VMAF+HDRMAX on the SDR-only LIVE Livestream Database [77] to study the efficacy of the nonlinear transformation prior to conducting SDR VQA. We also re-trained the original (SDR) VMAF in a similar manner for a fair comparison. The LIVE Livestream Database was selected because it is both modern and very diverse. It contains 315 videos of varying resolutions (1080p and 4K) multiple types of distortions and significant high-motion temporal content. It offers professional-quality videos captured under controlled lab conditions, similar to the anticipated application scenarios of the HDRMAX model. Moreover, there is no content overlap with the LIVE-HDR database, ensuring independent evaluation.

Our findings, displayed in Table XI, indicate that HDRMAX notably enhances performance on SDR content as well, underscoring the value of focusing on dark and bright regions during VQA. This improvement does not merely result from an increase in the size of the feature space. In the context of machine learning, it is widely recognized that adding more features does not inherently enhance model performance. Instead, the efficacy of a feature lies in its discriminative power and its relevance to the task at hand. The features added by HDRMAX are both discriminitive and highly sensitive to video quality characteristics, thus contributing to improved performance. Recognizing potential interest in the contribution of HDRMAX features, we also include the standalone performance of these features.

F. Evaluation on HDR Inage Database

To better illustrate the generalizability of our method, we conducted additional testing on the Unified Photometric Image Quality dataset (UPIQ) [12]. UPIQ is an expansive collection of over 4000 HDR and SDR images, and has proven to be a valuable resource for developing and validating HDR metrics.

TABLE X
STATISTICAL ANALYSIS OF MODEL COMPARISONS

	Test Condition	Dark		Bright	
	Model	t-statistic	p-Value	t-statistic	p-Value
SDR Quality Models	PSNR	7.32	1.78E-13	3.72	1.01E-04
	SSIM	23.67	1.07E-109	21.42	4.43E-92
	MS-SSIM	31.76	6.76E-180	25.71	1.72E-126
	ST-RRED	14.20	5.87E-44	12.60	2.18E-35
	SpEED-QA	19.32	1.02E-76	14.31	1.30E-44
	VMAF (original)	29.59	3.34E-160	17.93	4.18E-67
	VMAF (retrained)	2.63	4.30E-03	3.52	2.23E-04
HDR Quality Models	HDR-VDP 2.2	13.46	6.78E-40	12.46	1.19E-34
	HDR-VDP3.0.7	40.53	5.86E-263	24.05	9.69E-113
	HDR-VQM	405.60	0	444.17	0
	PU21-PSNR	74.90	0	74.11	0
	PU21-FSIM	61.62	0	48.35	0
	PU21-MSSSIM	57.80	0	74.55	0
	PU21-SSIM	73.21	0	69.63	0
	PU21-VSI	53.53	0	45.73	2.44E-313

TABLE XI
PERFORMANCE OF THE EVALUATED ALGORITHMS ON LIVE LIVESTREAM DATABASE. THE TOP PERFORMANCE IS BOLDFACED.

Algorithms	SROCC	PLCC	RMSE
PSNR	0.3760	0.4192	10.3355
SSIM	0.6976	0.7107	8.0082
MS-SSIM	0.6757	0.6907	8.2324
STRRED	0.6564	0.6694	8.4573
SpEED-QA	0.6894	0.7235	7.8589
VMAF (original)	0.6434	0.6355	8.7894
VMAF (retained)	0.6836	0.6912	8.2712
HDRMAX	0.6613	0.6755	8.9744
VMAF+HDRMAX	**0.7632**	**0.7743**	**7.2468**

TABLE XII
PERFORMANCE OF THE EVALUATED ALGORITHMS ON UPIQ DATABASE. THE TOP PERFORMANCE IS BOLDFACED.

	SROCC	PLCC	RMSE
HDR-VDP 3.0.7	0.8448	0.8426	0.3528
HDR-VQM	**0.8893**	**0.8824**	**0.3082**
PU21-FSIM	0.7358	0.71944	0.4551
PU21-MSSSIM	0.8192	0.8193	0.3757
PU21-PSNR	0.4903	0.4192	0.5950
PU21-SSIM	0.7215	0.7270	0.4499
PU21-VSI	0.6792	0.6713	0.4857
VMAF+HDRMAX	0.8485	0.8417	0.3680

However, given the scope of our study, we focused exclusively on the 380 HDR images in UPIQ.

It is noteworthy that the images in UPIQ are represented in absolute photometric and colorimetric units, reflecting light emitted from a display. To make these images compatible with our method, we transformed the pixel values into PQ before applying our models. We show the results in Table XII. Although our model didn't outperform all of the existing HDR metrics on this dataset, it still demonstrated commendable performance. This extra evaluation indicates the potential of our approach on diverse HDR contents and highlights its applicability to real-world scenarios.

VIII. DISCUSSION

We have created and made available the first public domain HDR10 VQA database. The new database contains 310 videos and subjective evaluations of these videos under two illumination conditions. The database can be effectively used for HDR algorithm design, evaluation, and comparison. However, it is important to note that our current focus is exclusively on the HDR10 format. Further research could extend this work to encompass other HDR formats such as HDR10+, Dolby Vision, and HLG. It is important to mention that 8 of the 31 source videos focus on football. While our goal was to highlight popular consumer sports content, this overweight could conceivably introduce biases. We also developed a framework for defining HDR quality aware features, which

when used to modify the widely-used VQA model, VMAF, achieves improved quality predictions against human subjective judgments on both HDR and SDR content. The enhanced HDRMAX performance across different HDR parameters, and on SDR data, suggests that this parallel pathway with nonlinear transforms adeptly captures perceived distortions on both HDR and SDR content. The extra feature extraction steps have only a minor impact on the computational complexity relative to the original VMAF. The nonlinear transform requires a single convolution to obtain a local average and is $\mathcal{O}(k^2NT)$, where k is the local window size, N is the number of pixels in each frame, and T is the number of frames. Because the feature extraction stage can occur in parallel on the transformed and original frames, parallel computation may substantially speed up feature extraction.

ACKNOWLEDGMENT

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

REFERENCES

[1] T. Kunkel, S. Daly, S. Miller, and J. Froehlich, "Perceptual design for high dynamic range systems," in *High Dynamic Range Video*. Elsevier, 2016, pp. 391–430.

[2] ITU, "BT.1886 : Reference electro-optical transfer function for flat panel displays used in HDTV studio production," Intl. Telecomm. Union, Tech. Rep., 2011.

[3] ITU, "BT.709 : Parameter values for the hdtv standards for production and international programme exchange," Intl. Telecomm. Union, Tech. Rep., 2011.

[4] CTA, "Television technology consumer definitions," https://cdn.cta.tech/cta/media/media/membership/pdfs/video-technology-consumer-definitions.pdf.

[5] (2015) BT.2020 : Parameter values for ultra-high definition television systems for production and international programme exchange. [Online]. Available: https://www.itu.int/rec/R-REC-BT.2020

[6] S. Standard, "High dynamic range electro-optical transfer function of mastering reference displays," *SMPTE ST*, vol. 2084, no. 2014, p. 11, 2014.

[7] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Apr 2017. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[8] T. N. Cornsweet and H. Pinsker, "Luminance discrimination of brief flashes under various conditions of adaptation," *The Journal of Physiology*, vol. 176, no. 2, p. 294, 1965.

[9] Y. Sugito, J. Vazquez-Corral, T. Canham, and M. Bertalmío, "Image quality evaluation in professional hdr/wcg production questions the need for hdr metrics," *IEEE Transactions on Image Processing*, vol. 31, pp. 5163–5177, 2022.

[10] L. Krasula, A. Choudhury, S. Daly, Z. Li, R. Atkins, L. Malfait, and A. Mavlankar, "Subjective video quality for 4k hdr-wcg content using a browser-based approach for" at-home" testing," *Electronic Imaging*, vol. 35, pp. 263–1, 2023.

[11] R. K. Mantiuk and M. Azimi, "Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[12] A. Mikhailiuk, M. Pérez-Ortiz, D. Yue, W. Suen, and R. K. Mantiuk, "Consolidated dataset and metrics for high-dynamic-range image quality," *IEEE Transactions on Multimedia*, vol. 24, pp. 2125–2138, 2022.

[13] Z. Shang, J. P. Ebenezer, A. C. Bovik, Y. Wu, H. Wei, and S. Sethuraman, "Subjective assessment of high dynamic range videos under different ambient conditions," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2021.

[14] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content," *arXiv preprint arXiv:1803.04815*, 2018.

[15] X. Pan, J. Zhang, S. Wang, S. Wang, Y. Zhou, W. Ding, and Y. Yang, "HDR video quality assessment: Perceptual evaluation of compressed HDR video," *J. Vis. Comm. Image Rep.*, vol. 57, pp. 76–83, 2018.

[16] V. Baroncini, K. Andersson, A. Ramasubramonian, and G. Sullivan, "Verification test report for HDR/WCG video coding using HEVC main 10 profile," in *Proc. JCTVC-X1018 24th JCT-VC Meeting*, 2016.

[17] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective and objective evaluation of HDR video compression," in *9th Intl. Workshop Video Process. Qual. Metrics Consum. Electron. (VPQM)*, 2015.

[18] S. Athar, T. Costa, K. Zeng, and Z. Wang, "Perceptual quality assessment of UHD-HDR-WCG videos," in *IEEE Int. Conf. Image Process.*, 2019, pp. 1740–1744.

[19] 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, 2004.

[20] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Int. Workshop Video Process. Quality Metrics for Consumer Electron.*, vol. 7. sn, 2005, p. 2.

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

[22] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. image Process.*, vol. 15, no. 6, pp. 1680–1689, 2006.

[23] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, 2011.

[24] J. Wu, W. Lin, G. Shi, Y. Zhang, W. Dong, and Z. Chen, "Visual orientation selectivity based structure description," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4602–4613, 2015.

[25] 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, 2012.

[26] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, 2015.

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

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

[29] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The 37th ACSSC*, vol. 2, 2003, pp. 1398–1402.

[30] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, 2013.

[31] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, p. 011006, 2010.

[32] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[33] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *IEEE ICIP*, 2011, pp. 2505–2508.

[34] P. V. Vu and D. M. Chandler, "Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imaging*, vol. 23, no. 1, p. 013016, 2014.

[35] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "Speed-qa: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1333–1337, 2017.

[36] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. image Process.*, vol. 19, no. 2, pp. 335–350, 2009.

[37] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *the 15th ECCV*, 2018, pp. 219–234.

[38] M. T. Vega, D. C. Mocanu, J. Famaey, S. Stavrou, and A. Liotta, "Deep learning for quality assessment in live video streaming," *IEEE Signal Process. Lett*, vol. 24, no. 6, pp. 736–740, 2017.

[39] Z. Wang, H. Yeganeh, K. Zeng, and J. Wang, "Diagnosing visual quality impairments in high dynamic-range/wide-color-gamut videos," *Journal of Digital Video*, vol. 5, pp. 74–83, 2020.

[40] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel, "Predicting visible differences in high dynamic range images: model and its calibration," in *Human Vision and Electronic Imaging X*, B. E. Rogowitz, T. N. Pappas, and S. J. Daly, Eds., vol. 5666, International Society for Optics and Photonics. SPIE, 2005, pp. 204 – 214. [Online]. Available: https://doi.org/10.1117/12.586757

[41] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, jul 2011. [Online]. Available: https://doi.org/10.1145/2010324.1964935

[42] M. Narwaria, R. Mantiuk, M. P. D. Silva, and P. L. Callet, "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging*, vol. 24, no. 1, pp. 1 – 3, 2015. [Online]. Available: https://doi.org/10.1117/1.JEI.24.1.010501

[43] M. Narwaria, M. P. D. Silva, P. L. Callet, and R. Pepion, "On improving the pooling in HDR-VDP-2 towards better HDR perceptual quality assessment," in *Human Vision and Electronic Imaging XIX*, B. E. Rogowitz, T. N. Pappas, and H. de Ridder, Eds., vol. 9014, International Society for Optics and Photonics. SPIE, 2014, pp. 143 – 151. [Online]. Available: https://doi.org/10.1117/12.2045436

[44] R. K. Mantiuk, D. Hammou, and P. Hanji, "Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content," 2023.

[45] T. O. Aydın, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," in *Human Vision and Electronic Imaging XIII*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 6806, International Society for Optics and Photonics. SPIE, 2008, pp. 109 – 118. [Online]. Available: https://doi.org/10.1117/12.765095

[46] M. A. Abebe, T. Pouli, and J. Kervec, "Evaluating the color fidelity of itmos and hdr color appearance models," *ACM Trans. Appl. Percept.*, vol. 12, no. 4, sep 2015. [Online]. Available: https://doi.org/10.1145/2808232

[47] M. Rousselot, O. Le Meur, R. Cozot, and X. Ducloux, "Quality assessment of hdr/wcg images using hdr uniform color spaces," *Journal of Imaging*, vol. 5, no. 1, 2019. [Online]. Available: https://www.mdpi.com/2313-433X/5/1/18

[48] A. Choudhury, R. Wanat, J. Pytlarz, and S. Daly, "Image quality evaluation for high dynamic range and wide color gamut applications using visual spatial processing of color differences," *Color Research & Application*, vol. 46, no. 1, pp. 46–64, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/col.22588

[49] M. Narwaria, M. Perreira Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46–60, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596515000703

[50] "The Consumer Digital Video Library," https://www.cdvl.org/. [Online]. Available: https://www.cdvl.org/

[51] L. Song, Y. Liu, X. Yang, G. Zhai, R. Xie, and W. Zhang, "The SJTU HDR video sequence dataset," in *Proc. Int. Conf. Qual. Multim. Exp. (QoMEX)*, 2016, p. 100.

[52] "Free Ultra-HD / HDR / HLG / Dolby Vision 4K video demos," https://4kmedia.org/, 2020. [Online]. Available: https://4kmedia.org/

[53] F. Mercer Moss, K. Wang, F. Zhang, R. Baddeley, and D. R. Bull, "On the optimal presentation duration for subjective video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 1977–1987, 2016.

[54] F. Mercer Moss, C.-T. Yeh, F. Zhang, R. Baddeley, and D. R. Bull, "Support for reduced presentation durations in subjective video quality assessment," *Signal Processing: Image Communication*, vol. 48, pp. 38–49, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596516301126

[55] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "Bvi-hd: A video quality database for hevc compressed and texture synthesized content," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2620–2630, 2018.

[56] P. Paudyal, F. Battisti, and M. Carli, "Reduced reference quality assessment of light field images," *IEEE Transactions on Broadcasting*, vol. 65, no. 1, pp. 152–165, 2019.

[57] ITU, "ITU. 910 : Subjective video quality assessment methods for multimedia applications," Intl. Telecomm. Union, Tech. Rep., 2008.

[58] D. Hasler and S. Suesstrunk, "Measuring colorfulness in natural images," in *Human Vis. Electr. Imaging VIII*, vol. 5007. Intl. Soc. Opt. Photon., 2003, pp. 87–95.

[59] Z. Shang, J. P. Ebenezer, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "Study of the subjective and objective quality of high motion live streaming videos," *IEEE Transactions on Image Processing*, vol. 31, pp. 1027–1041, 2021.

[60] Z. Shang, J. P. Ebenezer, A. C. Bovik, Y. Wu, H. Wei, and S. Sethuraman, "Assessment of subjective and objective quality of live streaming sports videos," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[61] "65" class Q90T QLED 4K UHD HDR smart TV 2020 TVs - QN65Q90TAFXZA: Samsung US." [Online]. Available: https://www.samsung.com/us/televisions-home-theater/tvs/qled-4k-tvs/65-class-q90t-qled-4k-uhd-hdr-smart-tv-2020-qn65q90tafxza

[62] E. TECHNICAL, "Eotf chart for calibration and monitoring," European Broadcasting Union, Tech. Rep. Tech. 3374, Dec 2020. [Online]. Available: https://tech.ebu.ch/publications/tech3374

[63] ITU, "BT.2100 : Image parameter values for high dynamic range television for use in production and international programme exchange," Intl. Telecomm. Union, Tech. Rep., 2018.

[64] ITU, "BT.500 : Methodologies for the subjective assessment of the quality of television images," Tech. Rep.

[65] "Laboratory for image amp; video engineering." [Online]. Available: https://live.ece.utexas.edu/research/Quality/visualScreening.htm

[66] Z. Li, C. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," *Electron. Imag.*, vol. 2020, no. 11, pp. 131–1, 2020.

[67] ITU, "Methodology for the subjective assessment of the quality of television pictures ITU-R recommendation BT. 500-13," Tech. Rep., 2012.

[68] Z. Shang, J. P. Ebenezer, A. C. Bovik, Y. Wu, H. Wei, and S. Sethuraman, "Assessment of subjective and objective quality of live streaming sports videos," in *Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[69] J. Li, L. Krasula, Y. Baveye, Z. Li, and P. Le Callet, "Accann: A new subjective assessment methodology for measuring acceptability and annoyance of quality of experience," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2589–2602, 2019.

[70] A. Radonjić, S. R. Allred, A. L. Gilchrist, and D. H. Brainard, "The dynamic range of human lightness perception," *Current Biology*, vol. 21, no. 22, pp. 1931–1936, 2011.

[71] V. A. Billock and B. H. Tsou, "To honor fechner and obey stevens: relationships between psychophysical and neural nonlinearities." *Psychological bulletin*, vol. 137, no. 1, p. 1, 2011.

[72] M. Bertalmío, "Chapter 5 - brightness perception and encoding curves," in *Vision Models for High Dynamic Range and Wide Colour Gamut Imaging*, ser. Computer Vision and Pattern Recognition, M. Bertalmío, Ed. Academic Press, 2020, pp. 95–129. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128138946000107

[73] P. Ledda, L. P. Santos, and A. Chalmers, "A local model of eye adaptation for high dynamic range images," in *Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, 2004, pp. 151–160.

[74] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

[75] M. D. Fairchild and P.-H. Chen, "Brightness, lightness, and specifying color in high-dynamic-range scenes and images," in *Image Quality and System Performance VIII*, S. P. Farnand and F. Gaykema, Eds., vol. 7867, International Society for Optics and Photonics. SPIE, 2011, pp. 233 – 246. [Online]. Available: https://doi.org/10.1117/12.872075

[76] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[77] Z. Shang, J. P. Ebenezer, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "Study of the subjective and objective quality of high motion live streaming videos," *IEEE Transactions on Image Processing*, vol. 31, pp. 1027–1041, 2022.