Image Quality Evaluation in Professional HDR/WCG Production Questions the Need for HDR Metrics

Yasuko Sugito, *Member, IEEE*, Javier Vazquez-Corral, Trevor Canham, and Marcelo Bertalmío

***Abstract*— In the quality evaluation of high dynamic range and wide color gamut (HDR/WCG) images, a number of works have concluded that native HDR metrics, such as HDR visual difference predictor (HDR-VDP), HDR video quality metric (HDR-VQM), or convolutional neural network (CNN)-based visibility metrics for HDR content, provide the best results. These metrics consider only the luminance component, but several color difference metrics have been specifically developed for, and validated with, HDR/WCG images. In this paper, we perform subjective evaluation experiments in a professional HDR/WCG production setting, under a real use case scenario. The results are quite relevant in that they show, firstly, that the performance of HDR metrics is worse than that of a classic, simple standard dynamic range (SDR) metric applied directly to the HDR content; and secondly, that the chrominance metrics specifically developed for HDR/WCG imaging have poor correlation with observer scores and are also outperformed by an SDR metric. Based on these findings, we show how a very simple framework for creating color HDR metrics, that uses only luminance SDR metrics, transfer functions, and classic color spaces, is able to consistently outperform, by a considerable margin, state-of-the-art HDR metrics on a varied set of HDR content, for both perceptual quantization (PQ) and Hybrid Log-Gamma (HLG) encoding, luminance and chroma distortions, and on different color spaces of common use.**

***Index Terms*— High dynamic range (HDR), wide color gamut (WCG), objective quality metric, image coding, visual perception.**

Manuscript received 14 June 2021; revised 7 December 2021 and 23 May 2022; accepted 27 June 2022. Date of publication 19 July 2022; date of current version 4 August 2022. This work was supported in part by the European Union's Horizon 2020 research and innovation programme, AdMiRe Project, under Agreement 952027. The work of Javier Vazquez-Corral was supported by Grant PID2021-128178OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF "A way of making Europe." The associate editor coordinating the review of this manuscript was Dr. Damon Chandler and approving it for publication was Dr. Alessandro Foi. *(Corresponding author: Yasuko Sugito.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee on Human Research in NHK (Japan Broadcasting Corporation) under Application Nos. 2018-17 and 2019-22, and performed in line with the Declaration of Helsinki.

Yasuko Sugito is with the Science and Technology Research Laboratories, NHK, Setagaya-ku, Tokyo 157-8510, Japan (e-mail: sugitou.y-gy@nhk.or.jp).

Javier Vazquez-Corral is with the Computer Vision Center, Computer Science Department, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain (e-mail: javier.vazquez@cvc.uab.cat).

Trevor Canham and Marcelo Bertalmío are with the Instituto de Óptica, CSIC, 28006 Madrid, Spain (e-mail: tcanham82@gmail.com; marcelo.bertalmio@csic.es).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2022.3190706, provided by the authors.

Digital Object Identifier 10.1109/TIP.2022.3190706

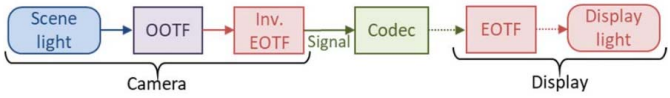

Fig. 1. HDR image coding diagram using the PQ method.

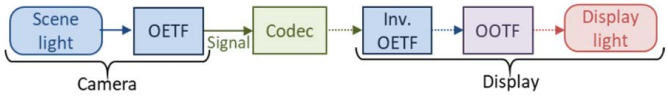

Fig. 2. HDR image coding diagram using the HLG method.

I. Introduction

HIGH dynamic range (HDR) imaging helps to express better detail in dark areas as well as much brighter highlights and is becoming an essential technology for video production. For HDR Television (HDR-TV), two different types of HDR methods are standardized, namely, perceptual quantization (PQ) and Hybrid Log-Gamma (HLG), and they define a non-linear transfer function (TF) between luminance and signal for capturing, displaying, recording, compressing, and transmitting purposes [1].

The PQ electro-optical TF (EOTF) was designed according to Barten's contrast sensitivity function (CSF) [2] and translates a non-linear PQ encoded signal value into an absolute display linear light that comes out of a monitor, whereas the HLG opto-electronic TF (OETF) was designed for backward compatibility with standard dynamic range (SDR) displays and translates relative scene linear light captured by a camera into a non-linear HLG signal value. See Figs. 1 and 2 for coding diagrams illustrating these concepts, and [3] for details on the concepts of EOTF, OETF, the opto-optical TF (OOTF), and HDR coding.

Regardless of the TF used, the codec is composed of a lossy encoding process followed by decoding, and therefore image deterioration is unavoidable. With a good compression quality the degree of deterioration might be unnoticeable, and conversely, low compression quality might produce annoying artifacts. A key element, then, for the evaluation of image coding techniques is the use of full-reference objective quality metrics, which measure the quality of a distorted image relative to an original reference image. Appropriate objective quality metrics should accurately emulate human perception, giving results similar to those of a subjective evaluation. The importance of having effective and precise quality metrics for video coding cannot be overstated: video streaming has

a very substantial environmental impact [4], and accurate image quality metrics do help to reduce the carbon footprint by allowing the content provider to minimize bandwidth use without compromising the user experience.

It is possible to extend SDR metrics so that they can be used in an HDR scenario as proposed in [5], simply by applying to the HDR content a particular form of transfer function termed Perceptually Uniform (PU) encoding. In this way we can resort to "classic" SDR metrics like peak signal-to-noise ratio (PSNR) or structure similarity index (SSIM) [6] and create variants for HDR: PU-PSNR, PU-SSIM, etc. We must point out that PU, like PQ, is based on experimental data on contrast sensitivity thresholds, not on (suprathreshold) brightness perception, making an assumption (peak sensitivity at each luminance level) that is incompatible with biological vision [7] and for this reason it is unable to reproduce some basic brightness perception phenomena like the crispening effect. There are also a number of objective metrics specifically dedicated to HDR image coding, the most popular ones being the HDR visual difference predictor (HDR-VDP-2) [8] and the HDR video quality metric (HDR-VQM) [9]; for other applications, like augmented reality and virtual reality (AR/VR) or rendering, HDR quality metrics have been proposed as well [10], [11].

Among all metrics that can be used in HDR, the ones that have the best performance in terms of their consistency with subjective evaluation results are the "native" HDR metrics HDR-VDP-2 and HDR-VQM [12]–[15], as they appear to be considerably better than the PU extensions of SDR metrics [16]. As a downside, these HDR-specific metrics can not be used in all situations because they are too complex, computationally intensive, unsuitable as loss functions in optimization problems because they are not differentiable [17], and they have a limited correlation with observer scores in applications like the evaluation of tone mapping results [18], [19]. Also, in [20], [21] we show that their ranking for HDR coding changes drastically depending on the experimental setting and on the TF (PQ or HLG) used for compression. In any case, recent works [22], [23] contend that these native HDR metrics are only surpassed in accuracy by convolutional neural network (CNN)-based visibility metrics for HDR.

Wide color gamut (WCG) technology allows for the reproduction of very vivid colors that fall outside the standard color gamut of traditional television, Rec. BT.709 [24], and are contained in the wider gamut prescribed in Rec. BT.2020 [25]. It is key to the advancement of realistic image presentation and is commonly associated with HDR given that brighter displays can produce more saturated colors [7]. HDR-specific metrics like the ones mentioned above consider only the luminance component, ignoring color. Some studies in the literature have investigated objective quality metrics for HDR/WCG images taking into account the chroma channels [26]–[28], and some color difference metrics have been introduced specifically for HDR/WCG images [29]. The computation of PU-encoded values for red, green, and blue color channels, followed by an SDR metric on each channel and a final aggregation of the three values, has been shown to perform substantially worse than the PU-SDR metric on luminance alone [16].

The major contribution and novelty of the present work is twofold. Firstly, we provide experimental data, thoroughly obtained in a practical use scenario of professional HDR/WCG production, that challenge several of the conclusions mentioned above and in particular question the need for HDR metrics in this context. Specifically, we show that:

- An SDR metric applied directly on PQ or HLG encoded content performs better than the HDR-specific metrics HDR-VDP-2 and HDR-VQM.
- In terms of color distortions, an SDR metric applied directly on the *luminance* channel of PQ or HLG encoded content performs better than any HDR/WCG *color* metric.

Secondly, we present a very simple framework for creating color HDR metrics, that uses only luminance SDR metrics, transfer functions (including a novel TF based on brightness perception) and weighted averages on typical color spaces. We show that metrics produced in this framework are able to consistently outperform, by a considerable margin, state-of-the-art HDR metrics on a varied set of HDR content, for both PQ and HLG encoding, for luminance and chroma distortions, and on different color spaces of common use. This approach is therefore quite more practical and effective than using HDR-specific metrics, offering higher accuracy with lower computational complexity, the ability to detect color distortions, and the possibility to be applied directly on PQ or HLG encoded content (which is a plus for HDR/WCG professional production). Furthermore, if the underlying SDR metric is differentiable, the metric created with this framework can be used as a loss function in optimization problems.

The organization of the paper is as follows. In Section II we detail the procedures for the dataset generation and for performing the subjective evaluation experiments. In Section III we prove that HDR-specific metrics are outperformed by SDR metrics applied directly to HDR content. In Section IV we prove that HDR/WCG color difference metrics are outperformed by an achromatic SDR metric. And finally in Section V we present a simple framework to extend luminance SDR metrics into color HDR metrics, introducing a novel TF based on brightness perception and showing that the resulting metrics can be considerably more effective than HDR-specific metrics. We point out that our previous works [21] and [30] had preliminary results on what is discussed here in Sections III and IV, but the experiments in those works involved considerably less test data and less metrics than what we now present in this paper; on the other hand, the content of Section V has not been reported elsewhere.

II. Dataset and Subjective Evaluation Experiments

A. Dataset: HDR/WCG Test Images and Generation of Distorted Images

We generated a dataset consisting of PQ and HLG images, which is an extended version of our previous works, [21] and [30].

Figure 3 represents thumbnails of 20 HDR/WCG test still images, and Table I describes their specifications.

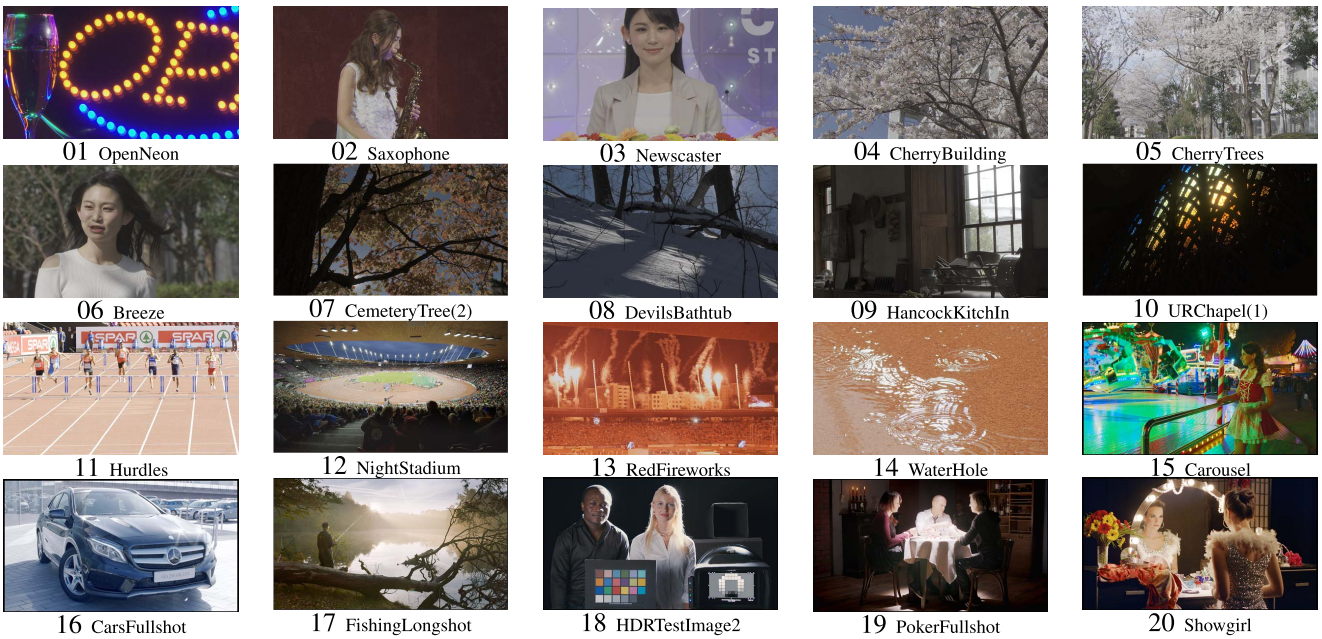

Fig. 3. 20 HDR/WCG test images.

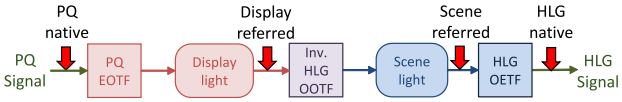

Fig. 4. Conversion process from PQ to HLG signal.

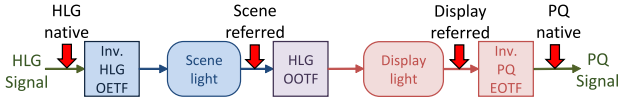

Fig. 5. Conversion process from HLG to PQ signal.

The spatial resolutions of the considered images are set to $1{,}920 \times 1.080$ pixels by cropping as needed. The images were in four HDR types and converted to both nonlinear PQ and HLG R'G'B' images with the peak luminance of 1,000 or 4,000 cd/m^2 as shown in Table I. Figures 4 and 5 illustrate the conversion process from PQ/HLG to HLG/PQ signals and the corresponding HDR types. The process between PQ and HLG was lossless by setting the nominal peak luminance (L_W) of the forward and inverse HLG OOTFs as the corresponding peak luminance. The details of the process were explained in [21].

Figure 6 illustrates the relationship between dynamic range (DR) and spatial information (SI) corresponding to the twenty test images. The vertical axis indicates $DR = \log_{10}(L_{max}/L_{min})$, where L_{max} and L_{min} are the maximum and minimum absolute display luminance in cd/m^2 after excluding 1% of the brightest and darkest pixels, respectively. The horizontal axis represents SI [34], which corresponds to the spatial complexity. SI is defined as the standard deviation of the pixel values of $\sqrt{(|\mathrm{Sobel_H}(Y_{10b})|^2 + |\mathrm{Sobel_V}(Y_{10b})|^2)}$, where

$$Y = 0.2627 \times R + 0.6780 \times G + 0.0593 \times B \quad (1)$$

$$Y_{10b} = round(1023 \times Y). \quad (2)$$

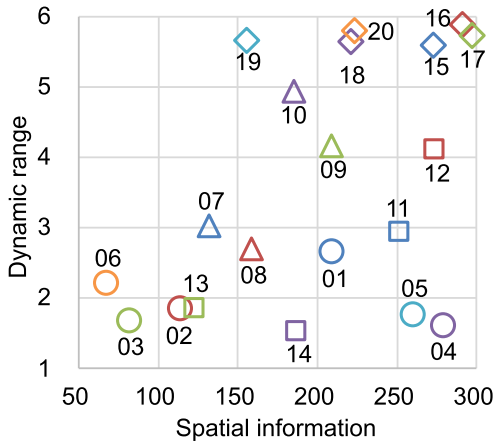

Fig. 6. Dynamic range and spatial information of 20 test images.

Here, Y_{10b} is the luminance component of an HLG image in 10-bit precision, and $\mathrm{Sobel_H}$ and $\mathrm{Sobel_V}$ are Sobel operators (3×3 convolution filters) in horizontal and vertical directions, respectively. A bar chart in Fig. 7 indicates colorfulness (CF) of the twenty images: $CF = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}$ where $rg = R - G$, $yb = 1/2(R + G) - B$, and σ^2 and μ are the variance and average of the pixel values, respectively [35]. Although it was originally defined on the BT.709 RGB color space [24], we directly applied the formula to the HLG images in 10-bit precision that used BT.2020 color space [25]. Furthermore, the chromaticity diagrams of these images are in [30], and 9 out of 20 images (01, 07, 10, and 15 – 20) have a color gamut exceeding BT.709. Overall, the resulting graphs indicate that the test images have a wide coding complexity.

To prepare various distorted images, we compressed each twenty PQ and HLG original images using two video coding

TABLE I
SPECIFICATIONS OF THE HDR/WCG TEST IMAGES

Test images	HDR type	Peak luminance	Source
01–06	HLG native	1,000 $\mathrm{cd/m^2}$	Our experimental content (a diffuse white level is in accordance with the HDR-TV production guideline [31])
07–10	Scene referred [3]	1,000 $\mathrm{cd/m^2}$	Fairchild's HDR photos [32]
11–14	Display referred [3]	4,000 $\mathrm{cd/m^2}$	Zurich Athletics 2014 test sequence [1]
15–20	PQ native	4,000 $\mathrm{cd/m^2}$	HdM-HDR-2014 content [33]

TABLE II
ENCODING CONDITIONS FOR HEVC AND VVC

	HEVC	VVC
Number of original images	40 (01–20 for PQ and HLG)	20 (01, 03, 05, 08, 10, 12, 14, 15, 17, and 19 for PQ and HLG)
Encoder	HEVC Test Model (HM) ver. 16.19 [39]	VVC Test Model (VTM) ver. 3.0 [40]
Configurations	Intra only, 4:2:0, 10-bit precision	
Target bitrates	100, 200, 300, and 400 kbits with the fixed quantization parameter (QP) setting	

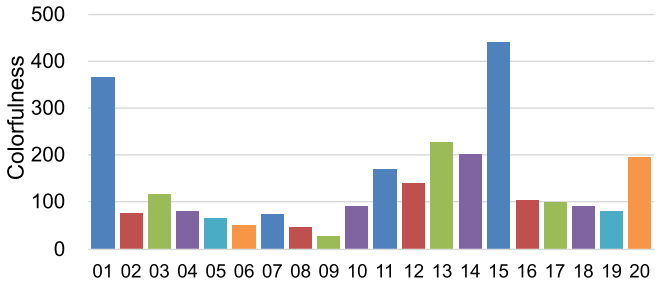

Fig. 7. Colorfulness of 20 test images.

schemes, high efficiency video coding (HEVC)/H.265 [36] and versatile video coding (VVC)/H.266 [37], and generated synthesized images based on the HEVC and VVC encoded images. The image coding procedures conform with the HEVC common test conditions concerning HDR/WCG images [38].

The image format of the encoder input and decoder output is Y'CbCr 4:2:0 10-bit. During the preprocessing step of the encoder, a PQ or HLG encoded signal in BT.2100 R'G'B' 4:4:4 is transferred to Y'CbCr 4:2:0 10 bits, which has one luma and two subsampled chroma components. Due to subsampling, the image sizes of the Cb and Cr components are decreased by half with respect to the original image both horizontally and vertically. This subsampling process is lossy, yet difficult to detect the difference perceptually. We denote this subsampled original Y'CbCr image as $\mathrm{Y_{org}C_{org}}$.

Table II describes the encoding conditions for HEVC and VVC. After encoding, image deterioration can be observed in both luma and chroma components. Therefore, we denote the compressed Y'CbCr images as $\mathrm{Y_{dis}C_{dis}}$. A total of 240 $\mathrm{Y_{dis}C_{dis}}$ images (20 images $\times$ 4 bitrates HEVC and 10 images $\times$ 4 bitrates VVC encoded images for PQ and HLG) were generated. Also, the synthesized images from $\mathrm{Y_{org}C_{org}}$ and $\mathrm{Y_{dis}C_{dis}}$ were generated, including $\mathrm{Y_{dis}C_{org}}$, composed of the compressed Y' and uncompressed Cb and Cr components, and $\mathrm{Y_{org}C_{dis}}$, comprised of the opposite components. For the four types of distorted Y'CbCr 4:2:0 10-bit images (i.e., $\mathrm{Y_{org}C_{org}}$, $\mathrm{Y_{dis}C_{dis}}$, $\mathrm{Y_{dis}C_{org}}$, and $\mathrm{Y_{org}C_{dis}}$), the postprocessing phase (the inverse of the preprocessing stage) was applied, and distorted images in nonlinear PQ and HLG R'G'B' 4:4:4 were prepared. An example of original and distorted HLG images is shown in Fig. 8.

TABLE III
EXPERIMENTAL CONDITIONS FOR SUBJECTIVE ASSESSMENTS

	1st batch	2nd batch
Monitor	31.1-in. HDR/WCG LCD monitor (approximately 0.70 m wide $\times$ 0.37 m high) 4,096 $\times$ 2,160/10-bit/1,000 $\mathrm{cd/m^2}$	
Viewing distance	1.5 picture height (approximately 0.55 m)	
Surround luminance	5 $\mathrm{cd/m^2}$	
Presentation method	SDSCE method	
Grading method	DSIS method	
Date of experiment	December 2018	March 2020
Observers	16 video experts	15 video experts
Test images	encoded images	other distorted images

B. Subjective Evaluation Experiments

A subjective evaluation experiment was performed following Rec. BT.500 [34] and BT.2100 [1]. Table III represents the experimental conditions used in the performed subjective assessments.

We used a 31.1-inch 4K HDR/WCG liquid crystal display (LCD) monitor (i.e., EIZO CG-3145[2]), supporting both PQ and HLG methods and functions as "Display" in the diagrams of Figs. 1 and 2. The monitor can display an all-white background at 1,000 $\mathrm{cd/m^2}$ and controls the brightness by pixel. The color gamut covers 99% of DCI-P3 [41] and more than 80% of BT.2020 [25].

The viewing environment was set following Table 3 of BT.2100, which establishes a reference viewing environment for the critical viewing of HDR program material or completed

[1] https://tech.ebu.ch/testsequences/zurich_athletics

[2] https://www.eizo.com/products/coloredge/cg3145/

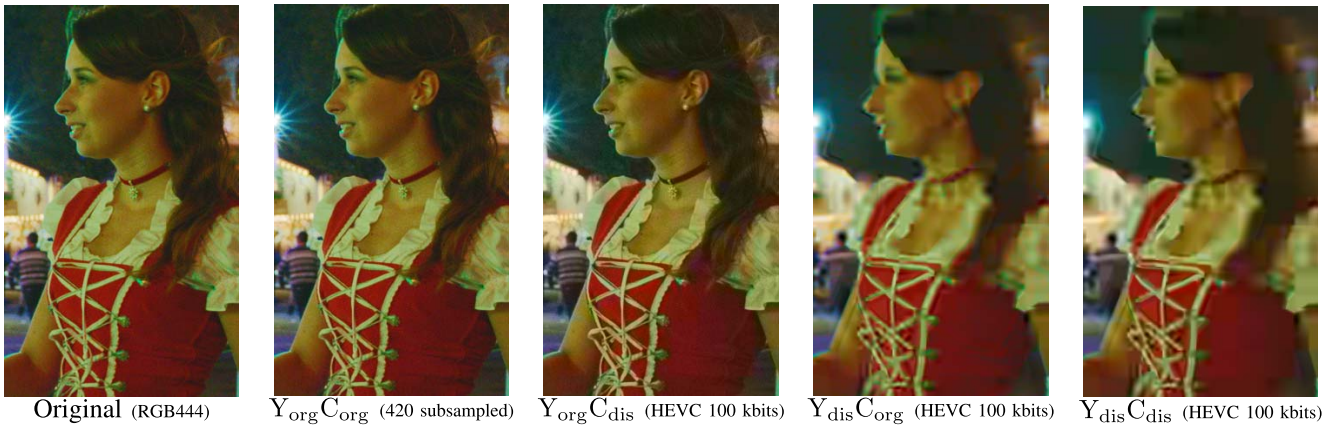

Fig. 8. Example of original and distorted HLG images.

programs to provide repeatable results from one facility to another when viewing the same material; the viewing distance and the surround luminance were set to 1.5 times the picture height and 5 cd/m^2, respectively.

The presentation method follows the simultaneous double stimulus for the continuous evaluation (SDSCE) method [34]. A pristine RGB image (without deterioration caused by the subsampling) and the corresponding distorted one to be evaluated were displayed at the original scale side by side with the 80-px padding between the two images on a mid-gray background (approximately 50 cd/m^2) during 10 s. Considering the order effect, the position of the original reference images differs for different subjects of which half receive the original on the left side and the other half on the right side. Each observer evaluates the deterioration level of a test image relative to the reference image using the five-grade impairment scale corresponding to the double stimulus impairment scale (DSIS) method [34]: 5 – imperceptible; 4 – perceptible, but not annoying; 3 – slightly annoying; 2 – annoying; and 1 – very annoying. Psychtoolbox-3 [42] with a 10-bit frame buffer mode was used to present the images and input, and record the scores. In this experiment, we focused on luminance reproducibility: e.g., a signal value corresponding to 800 cd/m^2 should be displayed at 800 cd/m^2. The monitor displayed several PQ images (11–20) after clipping at the peak luminance, 1,000 cd/m^2 (in our experiments this clipping procedure does not appear to have any effect on the perceived deterioration).

The experiment was conducted in two batches by research experts in HDR videos, with 16 experts taking part in the first batch and 11 of them, plus 4 other experts, participating in the second. Each batch was comprised of two 30-min sessions. Between each session, they took a break of at least 30 min. The evaluation was conducted one person at a time. Before the experiment started, verbal instructions of the evaluation method were provided to the subjects, and then they tested the training samples using different images from the test set to become familiar with the operation. The first and second sessions were for either HLG or PQ images, respectively (this information was not provided to the subjects before the experiment). A total 240 encoded images (120 $Y_{dis}C_{dis}$ generated from 20 original images (01–20) for PQ and HLG)

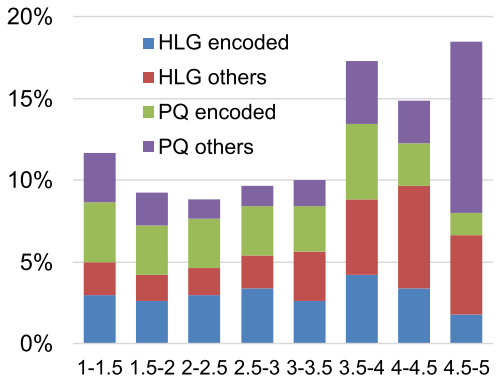

Fig. 9. MOS distribution of the distorted images.

and 258 other distorted images (9 $Y_{org}C_{org}$, 60 $Y_{dis}C_{org}$, and 60 $Y_{org}C_{dis}$ generated from 9 original images (01, 05, 10–11, 14–16, 19–20) for PQ and HLG) were assessed in the first and second batches, respectively. These other distorted images were assessed to investigate the perceptual influence on the deterioration of color components in Section IV and the performance of HDR/WCG metrics in Section IV-A. Additionally, 10 and 8 original images were, respectively, assessed for each batch for screening purposes. These items were randomly displayed.

The individual mean opinion score (MOS) of the original images and the Pearson linear correlation coefficient (PLCC) between the MOS and individual score for all evaluation items in each session were confirmed for screening the subjects. The individual MOS was between 4.5 and 5.0, and the PLCC was between 0.84 and 0.95. Thus, no outlier was present. Figure 9 shows the MOS distribution of the 498 distorted images. The distribution of MOS values is spread evenly, and the median value is 3.56. Since deterioration in $Y_{org}C_{org}$ and $Y_{org}C_{dis}$ images is hardly detectable (see Fig. 8), the ratio of other distorted images in MOS of 4 or larger becomes higher than that of encoded images.

III. HDR-Specific Metrics Are Outperformed by SDR Metrics Applied Directly on HDR Content

The performance of objective quality metrics on compressed HDR images using the first batch of our dataset was

TABLE IV
HDR METRIC RESULTS FOR PQ, HLG, AND ALL ENCODED IMAGES

	PQ			HLG			All (PQ and HLG)		
Metric	**PLCC**	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**
VIF	**0.9405**	**0.9377**	**0.3689**	**0.9336**	**0.9341**	**0.3934**	**0.9347**	**0.9323**	**0.3886**
VMAF	0.9365	0.9351	0.3808	0.9191	0.9173	0.4326	0.9272	0.9257	0.4097
MS-SSIM	0.9305	0.9276	0.3978	0.9180	0.9191	0.4354	0.9224	0.9206	0.4222
HDR-VDP-2	*0.9204*	*0.9198*	*0.4245*	*0.9137*	*0.9100*	*0.4461*	*0.9154*	*0.9150*	*0.4400*
HDR-VDP-3	*0.9084*	*0.9037*	*0.4540*	*0.9021*	*0.8980*	*0.4738*	*0.9040*	*0.9000*	*0.4674*
FSIM	0.8725	0.8716	0.5304	0.8775	0.8651	0.5265	0.8721	0.8666	0.5349
HDR-VQM	*0.8176*	*0.8073*	*0.6251*	*0.7706*	*0.7650*	*0.6996*	*0.7929*	*0.7866*	*0.6659*

studied in [21]. Results showed that directly applying SDR metrics, visual information fidelity (VIF) [43], and multi-scale SSIM (MS-SSIM) [44], to the luminance component of PQ/HLG images in 10-bit precision (calculated from (1) and (2)) showed better results than those of HDR dedicated metrics, HDR-VDP-2 [8] and HDR-VQM [9]. Note that HDR-VQM was selected because it showed an excellent result for HDR still images in a past study [13] though it was developed for videos, i.e., it considers deterioration in both spatial and temporal dimensions. Also, we considered two additional SDR metrics: feature similarity index (FSIM) [45], recommended in [27], and video multimethod assessment fusion (VMAF) [46], that is the state-of-the-art support vector regression (SVR)-based metric. For the VMAF calculation, we used VMAF v.2.1.1 released in January 2021 and the v.0.6.1 model. In addition, we tested another HDR dedicated metric, HDR-VDP-3,[3] which is the latest version of HDR-VDP.

We evaluated the performance of the considered objective quality metrics in the same way as previous related works [20], [21], [26]–[28], [30]. The consistency between the objective quality metric and the results of a subjective evaluation was investigated by the curve fitting of the logistic function based on the least square method:

$$\hat{y} = a + \frac{b}{1 + \exp(-c(x - d))} \tag{3}$$

where x and $\hat{y}$ denote the result of the objective metric and the predicted MOS, respectively. The true MOS y corresponding to x can be obtained from a subjective evaluation. The variables a, b, c, and d are selected to minimize $\sum_i (y_i - \hat{y}_i)^2$ for all items i. The number of items of the first batch was 120 each for PQ and HLG images, totaling 240. We assessed the performance in terms of PLCC, the Spearman rank-order correlation coefficient (SRCC), and the root-mean-square error (RMSE), concerning the corresponding relationship between y_i and $\hat{y}_i$. PLCC, SRCC, and RMSE measure the linearity, monotonicity, and accuracy, respectively. Ideally, the correlation coefficients should be 1, whereas the RMSE should be 0.

The performance results of the HDR metrics for PQ, HLG, and all images are shown in Table IV. The metrics are sorted in descending order of PLCC, while the orders of SRCC and RMSE are almost the same. The figures in bold indicate the best results, and the HDR dedicated metrics are in italics for reference. Although VMAF showed better results than those of HDR-VDP-2, HDR-VDP-3, and HDR-VQM, VIF remained the best among all the metrics (as in [12], and for the case of HDR video, VIF with PU encoding was also shown to outperform HDR-VDP-2).

[3] https://sourceforge.net/projects/hdrvdp/files/hdrvdp/3.0.6/

TABLE V
NUMBER OF CONDITIONS THAT EXHIBIT SIGNIFICANT DIFFERENCE FOR PQ IMAGES

Test image	$\mathbf{Y_{org}C_{org}} > \mathbf{Y_{org}C_{dis}}$	$\mathbf{Y_{dis}C_{org}} < \mathbf{Y_{dis}C_{dis}}$
01	1/8	
05	8/8	2/8
10		
11	1/4	
14		5/8
15		
16		2/4
19		
20		

IV. HDR/WCG COLOR DIFFERENCE METRICS ARE OUTPERFORMED BY AN ACHROMATIC SDR METRIC

Previously, we introduced the performance of HDR metrics on PQ and HLG images. However, all these HDR metrics rely on an achromatic component and ignore additional color information. To prove the necessity of incorporating chroma components in objective quality metrics for HDR/WCG images, we verified whether there is a significant difference between subjective evaluation results of Y_xC_{dis} and Y_xC_{org} with the same luma component $Y_x \in (Y_{org} \cup Y_{dis})$, even though VIF shows almost the same scores in such cases (i.e., VIF can not detect deterioration on chroma components).

Comparing MOS values of Y_xC_{dis} and Y_xC_{org} ($Y_x \in (Y_{org} \cup Y_{dis})$), we conducted Welch's t-test for the hypothesis that the two groups have equal mean, at the 5% significance level from the individual subjective test scores. Tables V and VI represent the number of conditions demonstrating a significant difference between MOS values in PQ and HLG images, respectively. For instance, 2/8 indicates that 2 of 8 conditions exhibit a significant difference between the MOS values corresponding to Y_xC_{org} and Y_xC_{dis}. The differences are always as follows: $Y_{org}C_{org} > Y_{org}C_{dis}$ for the Y_{org} images, and $Y_{dis}C_{org} < Y_{dis}C_{dis}$ for the Y_{dis} images.

TABLE VI
NUMBER OF CONDITIONS THAT EXHIBIT SIGNIFICANT DIFFERENCE FOR HLG IMAGES

Test image	$\mathbf{Y_{org}C_{org} > Y_{org}C_{dis}}$	$\mathbf{Y_{dis}C_{org} < Y_{dis}C_{dis}}$
01	1/8	
05	8/8	
10		
11	4/4	
14	5/8	2/8
15	4/8	
16	3/4	3/4
19	3/8	
20	4/4	

The subjective assessment results indicated that distortion in chroma components can be distinguished perceptually. Overall, PQ images showed less number of conditions that presented a significant difference. This can be due to a special encoder setting for PQ images that balances bit amounts on the luma and chroma components [38]. Considering that the difference was mainly in Y_{org} for HLG images, the degradation level on color components could be easily detected if a luma component was not significantly distorted. Therefore, we demonstrated the necessity of incorporating chroma components into the objective quality metrics.

A. Performance Evaluation of Existing HDR/WCG Metrics

We selected nine objective quality metrics that used both achromatic and chromatic components, including five color difference metrics, an SVR-based HDR/WCG metric, and two HDR/WCG metrics used for VVC standardization. We also included an achromatic metric, PSNR-L100, which was also adopted for VVC standardization. Additionally, VIF in PQ/HLG signal was calculated for reference. In what follows, we provide a short description for each metric.

1) ΔE_{00}: Commission International de l'Éclairage (CIE) DE2000 (ΔE_{00}) is a color difference metric calculated according to (4), as shown at the bottom of the next page [47].

This metric uses the CIE L*a*b* color space that is designed such that the same amount of numerical change in value corresponds to roughly the same amount of perceptual change. Although it is not intended for HDR/WCG images, it achieves good performance for an HDR/WCG image database [26]. We calculated ΔE_{00} for each pixel between an original and distorted image and then took the average value across all pixels in an image. Other color difference metrics are calculated similarly.

2) ΔE_S: S-CIELAB (ΔE_S) simulates spatial blurring by the human visual system (HVS) [48]. To realize this, a spatial Gaussian filter is applied to input images before calculating the color difference ΔE in the CIE L*a*b* color space. Similar to ΔE_{00}, an existing study indicated that ΔE_S achieves a good result for HDR/WCG images [27].

3) ΔE_{ITP}: ΔE_{ITP}, described in BT.2124 [29], has been introduced specifically for HDR/WCG images, and is the standardized version of the $\Delta E_{IC_tC_p}$ metric introduced in [49]. This version differs only in that the C_t channel is multiplied by a scalar value of 0.5 to convert to an ITP representation. It relies on the display referenced PQ IC_TC_P color space defined in Table 7 of BT.2100 [1], as shown in this equation:

$$\Delta E_{ITP} = 720 \times \sqrt{(I_1 - I_2)^2 + (T_1 - T_2)^2 + (P_1 - P_2)^2} \tag{5}$$

where $I = I$, $T = 0.5 \times C_T$, and $P = C_P$ in IC_TC_P.

For computing this metric, an estimate of the absolute display light in cd/m^2, as seen by human eyes, is required. To this end, for HLG images, we apply the HLG OOTF for mapping the relative scene light to the display light and set the peak luminance L_W to 1,000 cd/m^2, adapting to the monitor used in the subjective assessment.

4) ΔITP_R: ΔITP_R described in BT.2124 [29] is an extension of ΔE_{ITP} and can be directly applied to the scene-referred relative signals, such as those considered in the HLG method. For PQ images, we converted them to HLG signals before the calculation using the process of Fig. 4. Converting from HLG IC_TC_P to ITP, specific parameters are defined as follows: $I = I$; $T = 0.5 \times 1.823698 \times C_T$; $P = 1.887755 \times C_P$. Subsequently, those values are assigned to (5).

5) ΔE_z: ΔE_z is a color difference metric for HDR/WCG images [50]. The metric is calculated from the $J_za_zb_z$ perceptually uniform color space, which was based on PQ IC_TC_P color space and has more uniformity than that of IC_TC_P. As in the case of IC_TC_P, the PQ inverse EOTF is included in the conversion from display light to $J_za_zb_z$.

6) $FSIM_C$: FSIM has been developed from the observation that the HVS considers an image mainly according to its low-level features: specifically, a phase congruency (PC) and an image gradient magnitude [45]. $FSIM_C$ incorporates the chromatic information into the calculation procedure and is defined using this equation:

$$FSIM_C = \frac{\sum_{x \in \Omega} S_L(x) \cdot [S_C(x)]^{\lambda} \cdot PC_m(x)}{\sum_{x \in \Omega} PC_m(x)}. \tag{6}$$

Here, S_L and S_C denote the luminance and chrominance similarity measures, respectively. The paper experimentally determined the weight of the chrominance components λ as 0.03. We input PQ/HLG R'G'B' images in a 10-bit precision.

7) SVR HDR/WCG Metric [28]: This metric was developed by aggregating existing metrics, SI [34], HDR-VDP-2 [8], FSIM [45], and SSIM [6], using SVR. The SVR model has five explanatory variables as the inputs: (i) SI in J_z of $J_za_zb_z$, which indicates a spatial complexity of an original image; (ii) FSIM in J_z, which is calculated by omitting the term $[S_C(x)]^{\lambda}$ from (6); (iii) HDR-VDP-2; (iv) and (v) MCS_5 in a_z and b_z of $J_za_zb_z$, which signify the mean of contrast (c) $\times$ structure (s) defined in SSIM after down-sampling 5 times with a ratio of 2 on the a_z and b_z components, respectively. When inputting a component of the $J_za_zb_z$ color space to the four SDR metrics, (i), (ii), (iv), and (v), a scaling process in [27] was applied to adapt 8-bit SDR values. The SVR model was trained using four HDR still image databases to output an optimized metric score from the five input variables. We implemented this metric using MATLAB and confirmed that the median SRCC after conducting 1000 trials of the cross-validation using the four databases constantly marks approximately 0.94, whereas the median in the paper is 0.9421.

TABLE VII
HDR/WCG METRIC RESULTS FOR PQ IMAGES

Metric	PLCC	SRCC	RMSE
VIF	**0.9435**	0.8869	**0.4165**
SVR	0.9243	**0.9134**	0.4795
wPSNR	0.9159	0.8872	0.5044
FSIM_C	0.8952	0.8685	0.5598
PSNRL100	0.7548	0.7263	0.8241
$\Delta\text{E}_\text{ITP}$	0.6657	0.6626	0.9374
ΔE_Z	0.6191	0.6140	0.9866
ΔE_S	0.5809	0.5381	1.0225
$\Delta\text{ITP}_\text{R}$	0.5538	0.5220	1.0460
ΔE_{100}	0.5190	0.5057	1.0739
ΔE_{00}	0.3854	0.3899	1.1592

TABLE VIII
HDR/WCG METRIC RESULTS FOR HLG IMAGES

Metric	PLCC	SRCC	RMSE
VIF	**0.9202**	0.8424	**0.4414**
SVR	0.8832	**0.8563**	0.5289
FSIM_C	0.8618	0.8047	0.5718
wPSNR	0.8574	0.8163	0.5803
PSNRL100	0.6703	0.6024	0.8366
$\Delta\text{ITP}_\text{R}$	0.6615	0.6620	0.8455
ΔE_S	0.6443	0.6549	0.8623
$\Delta\text{E}_\text{ITP}$	0.6406	0.6282	0.8657
ΔE_Z	0.6327	0.6335	0.8731
ΔE_{00}	0.6058	0.6020	0.8970
ΔE_{100}	0.5507	0.5598	0.9411

8) $wPSNR$: wPSNR is a block-based calculation of PSNR, where the error values are weighted by a contrast sensitivity function, given the luminance value of the corresponding block. It is used by the Joint Video Experts Team (JVET) for evaluating the VVC encoding efficiency of HDR/WCG content, according to a very recent report [51]. To account for color, the metric is applied individually to YU'V' color channels and the channels are given weights [6,1,1]/8.

9) ΔE_{100}: The ΔE_{100} based metric is calculated by taking the color difference between Luma, Chroma, and Hue values from a CIE L*a*b* encoded signal and incorporating the resulting ΔE value in a PSNR calculation, where the peak value is set to 10,000 for PQ and 1,000 for HLG signals [51].

10) $PSNR-L100$: In this metric, also used by JVET for evaluating coding efficiency in HDR/WCG, the distortion is calculated as the ratio between the mean absolute error in the luminance channel and the peak signal value, which is set to 10,000 for PQ and 1,000 for HLG signals, respectively [51].

11) VIF: In Section III, VIF, derived from a statistical model for natural scenes, a model for image distortion, and an HVS model in an information-theoretical setting [43], demonstrated excellent performance compared with the other considered HDR metrics using only an achromatic component. We have inputted $\text{Y}_{10\text{b}}$, a 10-bit luminance signal derived from PQ/HLG R'G'B' images, calculated from (1) and (2).

Tables VII and VIII represent PLCC, SRCC, and RMSE of each metric for PQ and HLG images, respectively. The number of items was 249 for each PQ and HLG images (120 encoded and 129 other distorted images).

Tables VII and VIII show that, among the metrics considering color, the proposed SVR aggregation of HDR and WCG metrics is the most effective. The SVR metric outperforms the HDR-specific metrics HDR-VDP-2, HDR-VDP-3, and HDR-VQM, as can be observed by comparing with the data in Table IV. However, Tables IV, VII, and VIII also show that *VIF is still better than all other metrics*, including SVR, despite the fact that VIF is an SDR metric that ignores color because it is only applied to the luminance channel.

V. A SIMPLE FRAMEWORK TO DEVELOP HDR/WCG METRICS

A. Key Ideas

From the experimental results reported in the previous sections, we can see that there is still a need for a full-reference image quality metric adapted to HDR and WCG content that is sensitive to chromatic distortions.

We now make a series of observations that will be the basis of our proposed framework to develop HDR/WCG metrics, and which hopefully provide some insights on the reasoning to justify the novelty of our approach:

- SDR metrics simply are better metrics than HDR-native metrics: their correlation with observers' scores is higher.[4] SDR metrics have a much longer and richer history, and they have been much more extensively validated and optimized; it is easy to see that in the SDR case the validation experiments are simpler to perform, as they do not require the "specialized hardware" of HDR displays.
- The purpose of the TF nonlinearity is to allow for perceptual quantization, and for this reason, it has to emulate brightness perception [7]. This suggests that a modular approach with a TF that better emulates the perception of HDR images, followed by the application of an SDR metric, could yield a better HDR metric.
- In human vision, the contrast sensitivity is different for each color channel, so it makes sense when creating a metric that is intended to measure color differences to do this measurement on each channel, and then to combine these values through a weighted average.

[4]We are talking of course of SDR metrics applied to SDR images, and HDR metrics applied to HDR images.

$$\Delta E_{00}(L_1^*, a_1^*, b_1^*, L_2^*, a_2^*, b_2^*) = \sqrt{\left(\frac{\Delta L'}{K_L S_L}\right)^2 + \left(\frac{\Delta C'}{K_C S_C}\right)^2 + \left(\frac{\Delta H'}{K_H S_H}\right)^2 + R_T \left(\frac{\Delta C'}{K_C S_C}\right)\left(\frac{\Delta H'}{K_H S_H}\right)} \tag{4}$$

- But in biological vision and for natural images there are so many factors affecting contrast sensitivity [7] that there is no effective model from color science that could tell us how to weight color channels: therefore, it makes sense to optimize these weights so that the resulting metric has a maximum correlation with the opinions of observers.

B. Proposed Framework

Given a linear-light decoded HDR image H_d (with distortions due to lossy compression) and an uncompressed original HDR image H_o, also in linear-light, our proposed framework to develop HDR/WCG metrics that compare H_d with H_o consists of three stages:

1) For each channel c in color images H_o and H_d apply a nonlinear TF that emulates brightness perception, yielding single-channel images $TF(H_d^c)$ and $TF(H_o^c)$.
2) For each channel c apply an SDR image quality metric M comparing images $TF(H_d^c)$ and $TF(H_o^c)$, yielding a single-channel image quality score V^c:

$$V^c = M(TF(H_d^c), TF(H_o^c)) \tag{7}$$

3) The final score V of the HDR/WCG metric is computed as a weighted average of the image quality scores obtained for the three channels:

$$V = \frac{\sum_c \alpha_c V^c}{\sum_c \alpha_c} \tag{8}$$

In summary, what we propose is to develop HDR/WCG metrics $V(H_d, H_o)$ whose output is given by (8). Any particular instance of a metric in our framework is determined by the choice of color space (i.e., what the channels c represent), nonlinearity TF, SDR metric M, and weights α_c. A schematic of our approach is presented in Fig. 10.

Although there are some relevant differences with previous approaches (e.g., our TF is not limited to PU and we work in color, unlike [5]) we must remark that our major contribution/novelty is not so much in the framework itself, whose core ideas are extremely simple, but in showing that it works and consistently outperforms HDR-specific metrics (as we will see in the following section), which is rather counter-intuitive as it challenges commonly held assumptions in the field: the assumptions that HDR-specific metrics are always better than the extensions of SDR metrics, and that working in color gives lower performance than working in luminance.

C. Experimental Validation of New Metrics From Our Proposed Framework

To illustrate the effectiveness and potential of our proposed framework, we evaluate the performance of the 45 metrics obtained from combinations of the following choices:

1) The nonlinearity TF can be HLG, PQ, PU, PU21 [17] (a very recent update on PU), or a novel image-adaptive function termed TMG_2 that we have developed based on an algorithm inspired by vision models for (suprathreshold) brightness perception [19]. The details of the TMG_2 derivation are included in the Appendix.
2) The SDR metric M can be VIF, VMAF, or MS-SSIM, chosen because they are the three best performing metrics for HDR in Table IV.
3) The color space can be RGB, ITP, or YC_bC_r.

In all cases, the weights α_c have been chosen to maximize the PLCC correlation of V with the MOS scores from the experiments described above - a variety of compressed images, with either HLG or PQ encoding at the time of compression, and two different compression modalities - HEVC and VVC.

1) Comparison With the State of the Art in HDR Coding: We split the source dataset into two 120-image sets, where the channel weights α_c are trained on one and the metrics tested on the other. The 20 original reference images were separated into two groups such that each set of corresponding distorted images reflected the variation in dynamic range and spatial information represented in the full dataset (shown in Fig. 6), while still preserving variance in image sourcing and content between both sets. Each 120 image set is subsequently split into two 60-image subsets, one composed of PQ-encoded images and the other where the images are encoded with HLG.

The results from the state-of-the-art HDR metrics from the literature, including the very recent deep learning metric, deep photometric visibility metric (DPVM) [23], are shown in Table IX. The results for all 45 metrics from our framework on the whole 120-image testing set are presented in Table X for RGB, Table XI for ITP, and Table XII for YC_bC_r.

TABLE IX
HDR METRIC RESULTS ON THE 120-IMAGE TEST SET

Metric	PLCC	SRCC	RMSE
HDR-VDP-2	0.9248	0.9240	0.4601
HDR-VDP-3	0.9214	0.9118	0.4702
HDR-VQM	0.8909	0.9029	0.5492
DPVM [23]	0.8662	0.8889	0.6047

For completeness, the separate results for PQ and HLG encodings are presented as Supplementary Material, but there are no important differences in the trends and rankings with respect to those that can be observed on the 120-image set. In the Appendix, Tables XV, XVI, and XVII show the optimized channel weights α_c for each metric, and in Fig. 12 we can see a detailed schematic of the color transformation pipelines used in our tests.

From the comparison of these tables we can make a number of very interesting observations:

1) There are many instances of our framework, 15 in total, where our metrics surpass all HDR-specific metrics.
2) For each color space, the best performing metric is obtained with a different TF; these three metrics all have comparable scores (in PLCC, SRCC, and RMSE) that clearly surpass those of HDR-VDP-2, which is the best HDR-specific metric.
3) The deep-learning HDR metric DPVM does not improve on HDR-VDP-2, HDR-VDP-3, nor HDR-VQM.
4) For the RGB color space our custom, vision-based transfer function TMG_2 clearly outperforms the other TFs tested. These results show that an image-adaptive

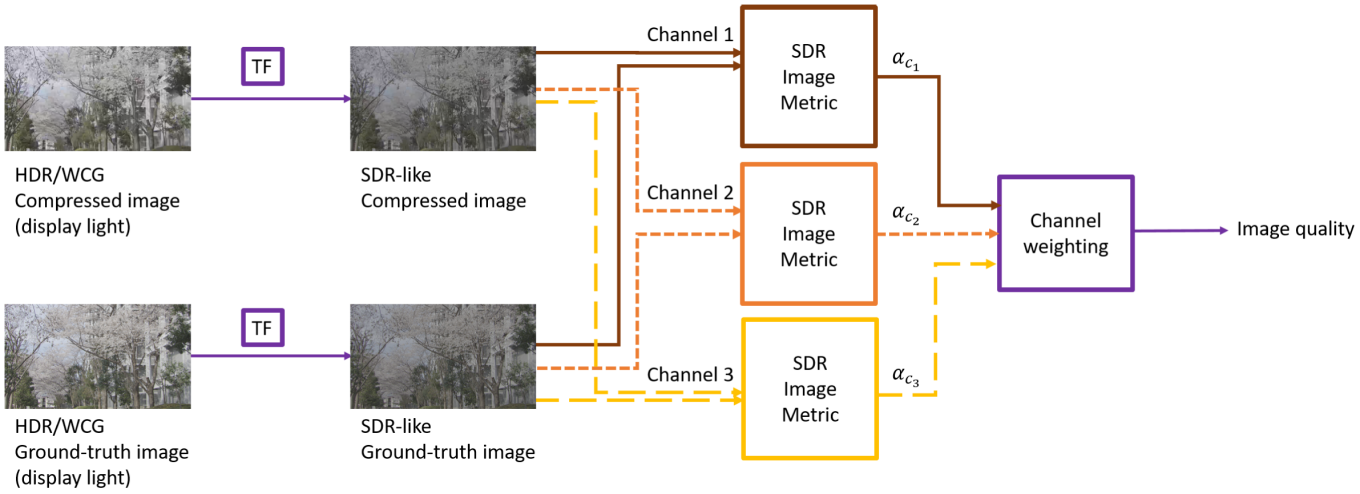

Fig. 10. Schematic of our proposed approach.

TABLE X
RESULTS FOR DIFFERENT INSTANCES OF OUR FRAMEWORK ON 120 IMAGES IN RGB ENCODING

	VIF			VMAF			MS-SSIM		
	PLCC	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**
TMG_2	**0.9665**	**0.9666**	**0.3094**	0.9188	0.9183	0.4777	**0.9047**	**0.8989**	**0.5156**
HLG	0.9465	0.9457	0.3905	0.9125	0.9165	0.4944	0.8280	0.8343	0.6782
PQ	0.9558	0.9554	0.3538	**0.9401**	**0.9441**	**0.4132**	0.8756	0.8693	0.5839
PU	0.9215	0.9196	0.4593	0.9073	0.9092	0.5237	0.7060	0.7035	0.8436
PU21	0.9577	0.9559	0.3362	0.9036	0.9069	0.5380	0.8186	0.8898	0.6994

TABLE XI
RESULTS FOR DIFFERENT INSTANCES OF OUR FRAMEWORK ON 120 IMAGES IN ITP ENCODING

	VIF			VMAF			MS-SSIM		
	PLCC	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**
TMG_2	0.8714	0.8665	0.5934	0.7972	0.8491	1.0554	0.7876	0.8654	1.2100
HLG	0.9403	0.9390	0.4120	0.8672	0.8909	0.6971	0.5321	0.4853	1.0241
PQ	**0.9632**	**0.9614**	**0.3251**	0.8665	0.9083	0.8888	0.8291	0.9108	0.7317
PU	0.8689	0.8668	0.5425	**0.9151**	**0.9243**	**0.5330**	**0.9223**	**0.9150**	**0.4565**
PU21	0.9261	0.9127	0.4139	0.8529	0.8717	0.9577	0.8289	0.9075	0.7059

TABLE XII
RESULTS FOR DIFFERENT INSTANCES OF OUR FRAMEWORK ON 120 IMAGES IN YC_bC_r ENCODING

	VIF			VMAF			MS-SSIM		
	PLCC	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**
TMG_2	0.9220	0.9233	0.4684	0.8082	0.8558	0.8779	0.7323	0.7884	0.8463
HLG	0.9106	0.9054	0.4864	0.8755	0.9084	0.8231	0.8538	0.8920	1.0252
PQ	0.9633	0.9620	0.3252	**0.9112**	**0.9413**	**0.6843**	0.9126	0.9209	**0.5527**
PU	0.9449	0.9424	0.3909	0.8949	0.9184	0.7288	**0.9457**	**0.9447**	0.8093
PU21	**0.9745**	**0.9722**	**0.2711**	0.8797	0.9186	0.7986	0.8483	0.9038	0.6559

transfer function based on brightness perception can be a more robust option for the representation of HDR signals to be used as input for SDR metrics than standardized transfer functions based on contrast sensitivity thresholds.

In short, these results evince that several metrics developed under the proposed framework are shown to outperform by a considerable margin the current state-of-the-art HDR metrics, as demonstrated on a varied set of HDR content, for both PQ and HLG coding, for luminance and chroma distortions, and for common color spaces. Our conclusion then is to propose for image quality evaluation of HDR/WCG images the following metrics, which are the best-performing ones in our tests: VIF as SDR metric, with TMG_2 as TF when working in RGB, PQ for the ITP color space, and PU21 for the YC_bC_r color space.

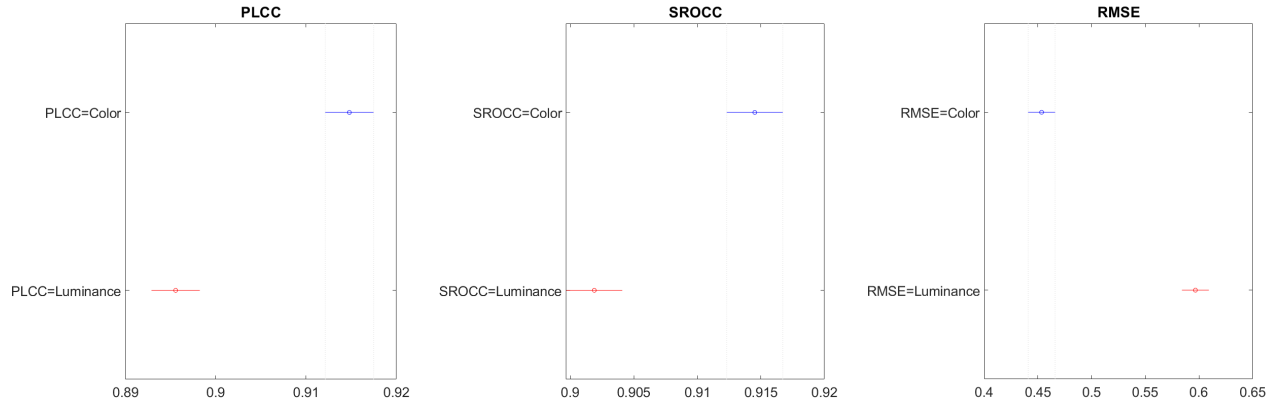

Fig. 11. Results for the ANOVA test comparing 9 metric instances when color is considered (the default form of our framework) versus the case when only luminance is considered (i.e. the achromatic case, $c \in \{Y\}$). The color processing presents a significant improvement for the three different measures.

TABLE XIII
RESULTS FOR DIFFERENT INSTANCES OF OUR FRAMEWORK ON 120 IMAGES WHEN WORKING ONLY ON LUMINANCE

	VIF			VMAF			MS-SSIM		
	PLCC	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**	**PLCC**	**SRCC**	**RMSE**
TMG_2	0.9490	0.9472	0.3814	0.9080	0.9072	0.5082	0.8511	0.9017	0.6433
HLG	0.9580	0.9586	0.3435	0.9074	0.9095	0.5073	**0.8895**	0.9134	**0.5296**
PQ	0.9621	**0.9629**	0.3303	**0.9292**	0.9376	0.4586	0.8888	**0.9284**	0.5718
PU	0.9499	0.9484	0.3749	0.9268	**0.9429**	**0.4506**	0.8098	0.8222	0.8995
PU21	**0.9631**	0.9606	**0.3259**	0.9110	0.9256	0.5485	0.8419	0.9184	0.6634

We must point out that while HDR-VDP and HDR-VQM are (by design) color blind, our metrics consistently pick up color distortions, as the metric value when comparing images $\mathrm{Y_xC_{dis}}$ and $\mathrm{Y_xC_{org}}$ that have the same luma component $\mathrm{Y_x} \in (\mathrm{Y_{org}} \cup \mathrm{Y_{dis}})$ is never the maximum score. In some cases, the metric value can correlate well with observer MOS data on these chroma distortions (e.g., $PLCC = 0.69$ for the metric given by the choices $TF = \mathrm{PQ}$ and $M = \mathrm{VMAF}$ in RGB), and this highlights another interesting aspect of our formulation, its modularity: the user can choose the color space, nonlinearity TF and SDR metric M that better suit her needs and better fit her data.

2) Comparison Against Using Luminance: We want to address the following point: the common assumption in the literature for HDR image quality assessment is that, when extending luminance SDR metrics to HDR, their performance on luminance is substantially better than if they are applied to the color channels and then the scores are averaged [16].

So to begin with, we conduct an ANOVA analysis to demonstrate the importance of considering color (the default form of our framework) versus the case when only luminance is considered (i.e., the achromatic case, $c \in \{Y\}$). For this analysis we take as an example 9 metrics from our framework, the ones produced by the combination of three TFs (TMG_2, PQ, and HLG) and the three SDR metrics under consideration, for the RGB color space, and full set of the images.

The left plot of Fig. 11 shows the general comparison of looking together at the PLCC for the 9 different metric instances, only splitting between color and luminance. As it can be seen, there is an advantage to the color scheme, which is statistically significant. Similarly, the center and right plots of the same figure show the results for the SRCC and the RMSE measures in which we can also see that the effect of optimizing for color channel weightings is clearly positive, and statistically significant when averaging overall results.

TABLE XIV
COMPARISON BETWEEN OPTIMIZED WEIGHTS AND CHANNEL AVERAGE FOR PU21 WITH VIF IN RGB

Metric	**PLCC**	**SRCC**	**RMSE**
Optimized weights	0.9577	0.9559	0.3362
Average	0.9449	0.9461	0.3961

To further highlight the importance and advantages of considering color, we can compare the results obtained for each color space by all combinations of TF and SDR metric (Table X for RGB, Table XI for ITP, Table XII for YC_bC_r) with the results obtained by those same metrics when working only on luminance (Table XIII). As we can see, for each color space the best result is produced with a different TF, and this color result is always better than the corresponding result for luminance. Specifically, for RGB the best results for VIF, VMAF, and MS-SSIM are produced by TMG_2 and they all surpass the results when using TMG_2 on luminance; for ITP, the best result is obtained with PQ, and for YC_bC_r the best performing TF is PU21, and in both cases their numbers are better than those for the luminance case.

Finally, we would like to point out that previous attempts at applying HDR-extended SDR metrics to the RGB color

space have used equal weights, whereas in our framework the weights are optimized so that the metric best fits the observers' scores, and this gives us a clear advantage in terms of performance, as exemplified in Table XIV for the case of PU21 with VIF in RGB.

VI. Conclusion

By performing thorough image quality experiments in a professional HDR/WCG production scenario, we have been able to demonstrate that the following commonly held assumptions are incorrect:

- *"HDR-native metrics provide the best results."* Our results in Section III show that the performance of HDR metrics is worse than that of a classic, simple SDR metric applied directly to the HDR content.
- *"HDR extensions of SDR metrics may be practical but they are not as good as HDR-native metrics."* Our results in Section V show that there are many instances in which HDR-extended SDR metrics outperform HDR-native metrics.
- *"The best results are obtained with a Deep Learning metric."* Our results in Section V show that the state-of-the-art Deep Learning HDR metric is outperformed by most of the metrics that we have considered.
- *"For color differences, HDR/WCG metrics provide the best results."* We show in Section IV how the chrominance metrics specifically developed for HDR/WCG imaging have poor correlation with observer scores and are also outperformed by an SDR metric.
- *"If working with an HDR-extended SDR metric, it is better to work only on the luminance channel than working on the color channels and performing a weighted average."* Our results in Section V show that performing a weighted average of the metric values computed on each color channel provides a better fit to the observers' scores than considering only the luminance channel.

We have proposed for image quality evaluation in HDR/WCG coding a very simple framework for creating color HDR metrics, that uses only luminance SDR metrics, transfer functions, and common color spaces. The advantages of our proposed framework over HDR-native metrics are the following:

- It provides a better match to observers' scores.
- It is able to detect both luminance and chroma distortions.
- It has much less computational complexity.
- It can be applied directly on PQ/HLG encoded content without having to do any transforms, which is an important plus for HDR/WCG professional production.
- Its modularity allows the user to choose the building blocks (color space, nonlinearity, and SDR metric) that are most suitable for each particular scenario.
- If the underlying SDR metric is differentiable, the metric created with this framework can be used as a loss function in optimization problems.

From our results we propose for image quality evaluation of HDR/WCG images the following metrics within our framework, which are the best-performing ones in our tests: VIF as SDR metric, with TMG_2 as TF when working in RGB, PQ for the ITP color space, and PU21 for the YC_bC_r color space.

Our conclusion then is to question the need for HDR metrics.

We are currently working on extending this study to videos.

Appendix I

In this Appendix, we explain the computation of our proposed TF non-linearity TMG_2, which is a slight adaptation of the vision-based transform TMG recently introduced in [19] for the purpose of tone mapping of graded content, i.e., for converting HDR images into SDR ones by emulating the brightness perception of HDR pictures. The transform TMG is a power law with a variable exponent that changes depending on the local pixel intensity, going from a value of γ_L for low intensities to a value of γ_H for high intensities. The values of the parameters that determine TMG are image-dependent, and for our proposed transform TMG_2 we have simply modified the way they are computed, while preserving the basic structure of the TMG model (see [19] for details). We will now describe step by step the transform TMG_2.

Let I be a linear HDR image, graded for some peak intensity level (e.g., 1,000 cd/m^2). For simplicity, let us assume I is a single-channel, luminance image, μ_1 is its associated median: $\mu_1 = median(I)$, and σ_1 is its associated standard deviation.

Let $\hat{I}$ be the HLG encoding of I:

$$\hat{I} = HLG(I), \tag{9}$$

and μ_2 its associated median: $\mu_2 = median(\hat{I})$.

By design, the HLG encoding is backwards-compatible with SDR displays, i.e., we can directly show an HLG image on an SDR display and it will look good. This tells us that the median luminance of $\hat{I}$ is the one that we should try to achieve with our TF, which gives us the following expression:

$$\mu_1^{\gamma} = \mu_2. \tag{10}$$

Solving for γ, we have this new equation:

$$\gamma = \frac{log(\mu_2)}{log(\mu_1)}. \tag{11}$$

For γ_L (corresponding to intensity levels below μ_1), and γ_H (corresponding to intensity levels above μ_1), we go back to the original formulation of [19]:

$$\gamma_L = (1+k)\gamma\,;\, \gamma_H = (1-k)\gamma\,, \tag{12}$$

but now we make the parameter k depend on the standard deviation:

$$k = 0.4 - 8.12 \cdot \sigma_1. \tag{13}$$

The value 0.4 in this equation comes from the crispening data in [52], and the value 8.12 comes from optimizing an instance metric from our framework (the one given by choosing $TF = TMG_2$ and $M = VIF$) so as to provide the best possible fit to the MOS data used in the experiments in the present paper.

For the interpolation/transition from γ_L to γ_H we use a classical sigmoid:

$$s(I) = \frac{1}{1 + e^{3.25(I-\mu_1)}}, \tag{14}$$

TABLE XV
OPTIMIZED CHANNEL WEIGHTING VALUES FOR FRAMEWORK INSTANCES IN RGB

	VIF			VMAF			MS-SSIM		
	R	G	B	R	G	B	R	G	B
TMG_2	0.82	1.00	-1.16	-0.34	1.00	-0.03	0.51	1.00	-0.51
HLG	0.97	1.00	-1.14	-0.11	1.00	-0.37	1.00	0.58	-0.62
PQ	1.00	0.51	-0.94	-0.32	1.00	-0.05	1.00	0.22	-0.46
PU	1.00	0.16	-0.54	-0.21	1.00	0.64	1.00	-0.20	-0.34
PU21	1.00	0.31	-0.59	1.00	0.42	-0.96	1.00	0.40	0.25

TABLE XVI
OPTIMIZED CHANNEL WEIGHTING VALUES FOR FRAMEWORK INSTANCES IN ITP

	VIF			VMAF			MS-SSIM		
	I	T	P	I	T	P	I	T	P
TMG_2	0.34	1.00	-0.97	1.00	-0.08	0.05	1.00	0.95	0.89
HLG	1.00	0.41	-0.44	1.00	-0.02	-0.03	1.00	-0.19	0.12
PQ	1.00	0.06	-0.25	1.00	-0.03	0.02	1.00	-0.27	0.06
PU	1.00	-0.19	0.41	1.00	0.01	-0.04	1.00	0.63	0.51
PU21	1.00	0.19	-0.27	1.00	-0.08	0.04	1.00	-0.13	0.07

TABLE XVII
OPTIMIZED CHANNEL WEIGHTING VALUES FOR FRAMEWORK INSTANCES IN YC_bC_r

	VIF			VMAF			MS-SSIM		
	Y	C_b	C_r	Y	C_b	C_r	Y	C_b	C_r
TMG_2	1.00	-0.23	0.50	1.00	-0.13	-0.50	1.00	0.63	-1.30
HLG	1.00	0.04	0.39	1.00	-0.06	-0.31	0.99	0.94	1.00
PQ	1.00	0.98	0.96	1.00	0.04	0.86	1.00	0.98	0.96
PU	1.00	-0.37	0.13	1.00	-0.20	0.81	1.00	-0.36	0.78
PU21	1.00	-0.46	0.12	1.00	-0.09	-0.32	1.00	0.96	0.94

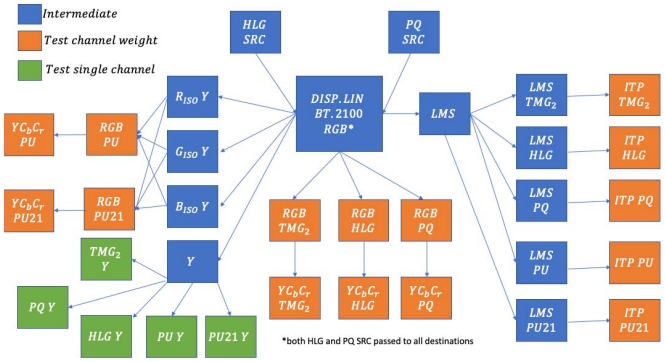

Fig. 12. Color transformation pipelines for RGB, Luminance, ITP, and YC_bC_r tests. $RGB_{ISO}Y$ signifies the display linear luminance of isolated RGB channels.

which goes from $s = 1$ when $I = 0$ to $s = 0.5$ when $I = \mu_1$ to $s = 0$ when $I \gg \mu_1$, and where 3.25 regulates the slope (this parameter was also optimized to fit the MOS data).

Finally, our transfer function $TMG_2(I)$ is computed as a power-law of variable exponent:

$$TMG_2(I) = I^{\gamma(I)}, \tag{15}$$

where the expression for the exponent is:

$$\gamma(I) = s(I)\gamma_L + (1 - s(I))\gamma_H. \tag{16}$$

APPENDIX II

Tables XV to XVII show the optimized channel weights α_c for each metric, and in Fig. 12 we have a detailed schematic of the color transformation pipelines used in our tests.

ACKNOWLEDGMENT

The authors would like to thank Dr. Maxime Rousselot and Dr. Olivier Le Meur for providing helpful information on their HDR/WCG metric. They also thank Prof. Rafał Mantiuk for his comments on the DPVM metric.

REFERENCES

[1] *Image Parameter Values for High Dynamic Range Television for Use in Production and International Programme Exchange*, Recommendation ITU-R BT.2100-2, Int. Telecommun. Union, Geneva, Switzerland, Jul. 2018.
[2] P. G. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. Bellingham, WA, USA: SPIE Press, 1999.
[3] T. Borer, A. Cotton, M. Pindoria, and S. Thompson, "Approaches to high dynamic range video," in *Proc. Digit. Media Ind. Academic Forum (DMIAF)*, Jul. 2016, pp. 71–76.
[4] D. Schien, P. Shabajee, J. Chandaria, D. Williams, and C. Preist, "Using behavioural data to assess the environmental impact of electricity consumption of alternate television service distribution platforms," *Environ. Impact Assessment Rev.*, vol. 91, Nov. 2021, Art. no. 106661.
[5] T. O. Aydın, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full dynamic range images," *Proc. SPIE*, vol. 6806, Jan. 2008, Art. no. 68060B.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[7] M. Bertalmío, *Vision Models for High Dynamic Range and Wide Colour Gamut Imaging: Techniques and Applications*. New York, NY, USA: Academic, 2019.

[8] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, "HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images," *J. Electron. Imag.*, vol. 24, no. 1, 2015, Art. no. 010501.

[9] M. Narwaria, M. P. Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Process., Image Commun.*, vol. 35, pp. 46–60, Jul. 2015.

[10] R. K. Mantiuk *et al.*, "FovVideoVDP: A visible difference predictor for wide field-of-view video," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–19, Jul. 2021, doi: 10.1145/3450626.3459831.

[11] P. Andersson, J. Nilsson, P. Shirley, and T. Akenine-Möller, "Visualizing errors in rendered high dynamic range images," in *Proc. Conf. Eur. Assoc. Comput. Graph.*, May 2021, pp. 25–28.

[12] 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," in *Proc. Int. Conf. Multimedia Signal Process. (ICMSP)*, 2014, pp. 1–5.

[13] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. G. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for HDR image quality assessment," *EURASIP J. Image Video Process.*, vol. 2015, no. 1, pp. 1–18, Dec. 2015.

[14] T. Vigier, L. Krasula, A. Milliat, M. P. Da Silva, and P. Le Callet, "Performance and robustness of HDR objective quality metrics in the context of recent compression scenarios," in *Proc. Digit. Media Ind. Academic Forum (DMIAF)*, Jul. 2016, pp. 59–64.

[15] E. Zerman, G. Valenzise, and F. Dufaux, "An extensive performance evaluation of full-reference HDR image quality metrics," *Qual. User Exper.*, vol. 2, no. 1, pp. 1–5, Dec. 2017.

[16] R. K. Mantiuk, "Practicalities of predicting quality of high dynamic range images and video," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 904–908.

[17] R. K. Mantiuk and M. Azimi, "PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.

[18] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Preference of experience in image tone-mapping: Dataset and framework for objective measures comparison," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 64–74, Feb. 2017.

[19] P. Cyriac, T. Canham, D. Kane, and M. Bertalmío, "Vision models fine-tuned by cinema professionals for high dynamic range imaging in movies," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2537–2563, Jan. 2021.

[20] Y. Sugito and M. Bertalmio, "Performance evaluation of objective quality metrics on HLG-based HDR image coding," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 96–100.

[21] Y. Sugito and M. Bertalmio, "Practical use suggests a re-evaluation of HDR objective quality metrics," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.

[22] K. Wolski *et al.*, "Dataset and metrics for predicting local visible differences," *ACM Trans. Graph.*, vol. 37, no. 5, pp. 1–14, Oct. 2018.

[23] N. Ye, K. Wolski, and R. K. Mantiuk, "Predicting visible image differences under varying display brightness and viewing distance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5434–5442.

[24] *Parameter Values for the HDTV Standards for Production and International Programme Exchange*, Recommendation ITU-R BT.709-6, Int. Telecommun. Union, Geneva, Switzerland, Jun. 2015.

[25] *Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange*, Recommendation ITU-R BT.2020-2, Int. Telecommun. Union, Geneva, Switzerland, Oct. 2015.

[26] A. Choudhury, J. Pytlarz, and S. Daly, "HDR and WCG image quality assessment using color difference metrics," in *Proc. SMPTE*, 2019, pp. 1–22.

[27] M. Rousselot, O. Meur, R. Cozot, and X. Ducloux, "Quality assessment of HDR/WCG images using HDR uniform color spaces," *J. Imag.*, vol. 5, no. 1, p. 18, Jan. 2019.

[28] M. Rousselot, X. Ducloux, O. L. Meur, and R. Cozot, "Quality metric aggregation for HDR/WCG images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3786–3790.

[29] *Objective Metric for the Assessment of the Potential Visibility of Colour Differences in Television*, Recommendation ITU-R BT.2124-0, Int. Telecommun. Union, Geneva, Switzerland, Jan. 2019.

[30] Y. Sugito, T. Canham, J. Vazquez-Corral, and M. Bertalmio, "A study of objective quality metrics for HLG-based HDR/WCG image coding," *SMPTE Motion Imag. J.*, vol. 130, no. 4, pp. 53–65, May 2021.

[31] "Guidance for operational practices in HDR television production," Int. Telecommun. Union, Geneva, Switzerland, Report ITU-R BT.2408-3, Jul. 2019.

[32] M. D. Fairchild, "The HDR photographic survey," in *Proc. 15th Color Imag. Conf.*, 2007, pp. 233–238.

[33] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," *Proc. SPIE*, vol. 9023, pp. 279–288, Mar. 2014.

[34] *Methodologies for the Subjective Assessment of the Quality of Television Images*, Recommendation ITU-R BT.500-14, Int. Telecommun. Union, Geneva, Switzerland, Oct. 2019.

[35] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Proc. SPIE*, vol. 5007, pp. 87–95, Jun. 2003.

[36] *High Efficiency Video Coding*, Recommendation ITU-T H.265 (V7), Int. Telecommun. Union, Geneva, Switzerland, Nov. 2019.

[37] *Versatile Video Coding*, Recommendation ITU-T H.266, Int. Telecommun. Union, Geneva, Switzerland, Aug. 2020.

[38] E. François, J. Sole, J. Ström, and P. Yin, *Common Test Conditions for HDR/WCG Video Coding Experiments*, document JCTVC-Z1020, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 26th Meeting, Geneva, Switzerland, Jan. 2017.

[39] *High Efficiency Video Coding (HEVC) | JCT-VC*. Accessed: May 14, 2021. [Online]. Available: https://hevc.hhi.fraunhofer.de/

[40] *Versatile Video Coding (VVC) | JVET*. Accessed: May 14, 2021. [Online]. Available: https://jvet.hhi.fraunhofer.de/

[41] *D-Cinema Quality—Reference Projector and Environment*, Standard SMPTE RP 431-2:2011, Apr. 2011.

[42] M. Kleiner, D. Brainard, D. Pelli, R. Murray, and C. Broussard, "What's new in psychtoolbox-3?" *Perception*, vol. 36, no. 14, pp. 1–16, 2007.

[43] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

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

[45] 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, Aug. 2011.

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

[47] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl.*, vol. 30, no. 1, pp. 21–30, 2005.

[48] X. Zhang and B. A. Wandell, "A spatial extension of CIELAB for digital color-image reproduction," *J. Soc. Inf. Display*, vol. 5, no. 1, pp. 61–63, 1997.

[49] E. Pieri and J. Pytlarz, "Hitting the mark—A new color difference metric for HDR and WCG imagery," in *Proc. Annu. Tech. Conf. Exhib.*, Oct. 2017, pp. 1–13.

[50] M. Safdar, G. Cui, Y. J. Kim, and M. R. Luo, "Perceptually uniform color space for image signals including high dynamic range and wide gamut," *Opt. Exp.*, vol. 25, no. 13, pp. 15131–15151, Jun. 2017.

[51] A. Segall, E. François, W. Husak, S. Iwamura, and D. Rusanovskyy, *JVET Common Test Conditions and Evaluation Procedures for HDR/WCG Video*, document JVET-T2011, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 20th Meeting, by Teleconference, Oct. 2020.

[52] M. Bertalmío, A. Gomez-Villa, A. Martín, J. Vazquez-Corral, D. Kane, and J. Malo, "Evidence for the intrinsically nonlinear nature of receptive fields in vision," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, Dec. 2020.

Yasuko Sugito (Member, IEEE) is a Principal Research Engineer with the Japan Broadcasting Corporation (NHK) Science and Technology Research Laboratories (STRL), Tokyo, Japan, researching video compression algorithms and image processing on 8K. Her current research interests include image quality assessment, both subjectively and objectively, for 8K videos with high-frame-rate (HFR) 120-Hz, high dynamic range (HDR), and wide color gamut (WCG).

Javier Vazquez-Corral is an Associate Professor with the Universitat Autonoma de Barcelona (UAB), Bellaterra, Spain. Prior to that, he was a Postdoctoral Researcher at the Universitat Pompeu Fabra (UPF) and at the University of East Anglia (UEA). His research interests include the use of color in image processing, computer vision problems, bridging the gap between color in the human brain, and its use in computer-vision applications.

Trevor Canham was born and raised in Rochester, NY, USA. He received the B.Sc. degree in motion picture science from the Rochester Institute of Technology in 2018. Since 2018, he has been researching perceptual problems in color management and color grading tools, and is currently with the Spanish National Research Council (CSIC), Instituto de Óptica. His interest includes the interaction between human visual perception and aesthetic imaging systems.

Marcelo Bertalmío received the B.Sc. and M.Sc. degrees in electrical engineering from the Universidad de la República, Uruguay, and the Ph.D. degree in electrical and computer engineering from the University of Minnesota, USA, in 2001. He is a Scientific Researcher with the Spanish National Research Council (CSIC), Madrid, Spain. His current research interests include developing image processing algorithms for cinema that mimic neural and perceptual processes in the visual system, and to investigate new vision models based on the efficient representation principle.