HDR IMAGE COMPRESSION: A NEW CHALLENGE FOR OBJECTIVE QUALITY METRICS

Philippe Hanhart[1]*, Marco V. Bernardo*[2,3]*, Pavel Korshunov*[1]*, Manuela Pereira*[3]*,*
António M. G. Pinheiro[2]*, and Touradj Ebrahimi*[1]

[1] Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland
[2] Remote Sensing Unit/Optics Center, UBI, Covilhã, Portugal
[3] Instituto de Telecomunicações, UBI, Covilhã, Portugal

ABSTRACT

High Dynamic Range (HDR) imaging is able to capture a wide range of luminance values, closer to what the human visual system can perceive. It is believed by many that HDR is a technology that will revolutionize TV and cinema industry similar to how color television did. However, the complexity of HDR requires reinvention of the whole chain from capture to display. In this paper, HDR images compressed with the upcoming JPEG XT HDR image coding standard are used to investigate the correlation between thirteen well known full-reference metrics and perceived quality of HDR content. The metrics are benchmarked using ground truth subjective scores collected during quality evaluations performed on a Dolby Pulsar HDR monitor. Results demonstrate that objective quality assessment of HDR image compression is challenging. Most of the tested metrics, with exceptions of HDR-VDP-2 and FSIM computed for luma component, poorly predict human perception of visual quality.

Index Terms— Image quality assessment, objective metrics, High Dynamic Range, JPEG XT

1. INTRODUCTION

High Dynamic Range (HDR) imaging systems pursue the acquisition of images where all the brightness information of the visible range of a scene is represented. Hence, they can capture the whole dynamic range and color gamut perceived by the human visual system (HVS). Thus, many applications can greatly benefit from the adoption of HDR imaging. For example, HDR imaging can be exploited to improve quality of experience in multimedia applications [1] and to enhance intelligibility in security applications where lighting conditions cannot be controlled [2].

There are different methods to obtain HDR images. Computer rendering and merging multiple low dynamic range (LDR) images taken at different exposure settings are the two methods initially used to generate HDR images. Nowadays, HDR images can also be acquired using specific image sensors. There are two forms of visualization in HDR images. The first and the best solution is to use a specific HDR display that has the ability of representing a wider luminance range and color gamut. The second solution is to map the HDR image to a LDR display luminance range and color gamut, using a tone mapping operator (TMO).

JPEG XT is an upcoming standard for JPEG backward-compatible compression of HDR images [3]. Using this compression standard, HDR images are coded in two layers. The base layer, where a tone mapped version of the HDR image is encoded in the normal JPEG format, and a residual layer, where the extra HDR information is encoded. The advantage of this method is that any conventional JPEG decoder can extract the tone mapped image, keeping backward compatibility and allowing for display on a conventional LDR monitor. Furthermore, a specific JPEG XT decoder can use the residual layer to reconstruct a lossy version of the HDR image.

In this paper, HDR images encoded with JPEG XT profile A and corresponding ground truth subjective scores are used. During the subjective quality assessment, HDR images compressed at four different bit rates were displayed side-by-side on a Dolby Research HDR RGB backlight dual modulation display (aka Pulsar). The black level was held constant, so the luminance dynamic range was solely determined by the maximum luminance. The paired comparison evaluation methodology was selected for its high accuracy and reliability in constructing a scale of perceptual preferences. The subjects participating in the evaluation experiment were naïve viewers.

This paper investigates the performance of state-of-the-art objective metrics in predicting perceived quality of compressed HDR images. A good objective metric should take the psychophysical process of the human vision and perception system into account. The main characteristics of

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1), EC funded Network of Excellence VideoSense, Portuguese "FCT – Fundação para a Ciência e a Tecnologia" (projects PTDC/EIA-EIA/119004/2010, PEst-OE/EEI/LA0008/2013, and PEst-OE-FIS/UI0524/2014), and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET. The authors would like to thank Dolby Laboratories Inc. staff for providing Dolby Research HDR RGB backlight dual modulation display (aka Pulsar).

(a) *BloomingGorse2* (b) *CanadianFalls* (c) *McKeesPub*

(d) *MtRushmore2* (e) *WillyDesk*

Fig. 1: HDR images used in the experiments.

Table 1: HDR images information.

Image	Resolution [pixels]	Dynamic range [dB] (cropped part)	Encoding parameters (q, Q) q: base layer, Q: residual layer
BloomingGorse2	4288×2848	42	$(11, 12), (20, 13), (32, 15), (62, 15)$
CanadianFalls	4312×2868	41	$(16, 29), (30, 30), (65, 30), (80, 33)$
McKeesPub	4258×2829	60	$(5, 64), (15, 91), (48, 88), (83, 91)$
MtRushmore2	4312×2868	50	$(5, 20), (24, 82), (67, 80), (89, 78)$
WillyDesk	4288×2848	70	$(5, 63), (15, 79), (57, 90), (85, 91)$

the HVS include contrast and orientation sensitivity, frequency selection, spatial and temporal pattern masking, and color perception [4]. In total, 36 metrics developed for image quality assessment were benchmarked using subjective scores as ground truth. Out of all metrics, only HDR-VDP-2 metric was specifically developed for HDR images. Out of the 36 metrics, thirteen full-reference metrics were selected for the detailed evaluation and analysis, including Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Signal-to-Noise Ratio (SNR), Weighted Signal-to-Noise Ratio (WSNR), Structural Similarity index (SSIM), Multiscale SSIM index (MS-SSIM), Visual Information Fidelity (VIF), Visual Information Fidelity pixel-based (VIFp), Universal Quality Index (UQI), Image Fidelity Criterion (IFC), Feature Similarity Index (FSIM), High Dynamic Range Visible Difference Predictor (HDR-VDP-2), and CIEDE2000 color difference. For each metric, their objective scores were fitted to subjective scores using logistic fitting. Several performance indexes, such as Pearson and Spearman correlation coefficients and root-mean-square-error, were computed to compare the metrics estimation of subjective scores. Hence, with this study we expect to produce a valid contribution for future objective quality studies on HDR imaging.

The remainder of the paper is organized as follows. The dataset and corresponding subjective scores used as ground truth are described in Section 2. The different metrics benchmarked in this study are defined in Section 3. In Section 4, the methodology used to evaluate the performance of the metrics is described. Section 5 provides a detailed analysis of the objective results and discusses the reliability of objective metrics. Finally, Section 6 concludes the paper.

2. DATASET AND SUBJECTIVE EVALUATIONS

2.1. Dataset

Five HDR images[1] of different dynamic ranges (computed using Banterle's HDR toolbox for MATLAB[2]), representing different typical scenes, were used in the experiments (see Figure 1 and Table 1 for details). Originally, these images were selected by JPEG for the verification tests of JPEG XT standard. JPEG also provided LDR versions of these images that were manually tone-mapped using Adobe Photoshop from the original HDR.

[1] http://www.cis.rit.edu/fairchild/HDR.html
[2] http://www.github.com/banterle/HDR_Toolbox

To prepare images for subjective experiments, both HDR and LDR versions were first downscaled by a factor of two with bicubic interpolation. The resulted images were first compressed using JPEG XT Profile A to four different bit rate values, ranging from a minimum of 0.3 bpp to a maximum of 2.2 bpp for different images. The bit rate values were selected for each content separately (see Table 1) in such a way that there is a noticeable visual difference between images with different bit rates when they are displayed on the HDR monitor.

Compressed images were then cropped to 950×1080 pixels regions for side by side subjective experiments (see Section 2.2 for details). The regions to crop were selected by expert viewers in such a way that cropped versions are representative of the quality and the dynamic range of the original images. Red rectangles in Figure 1 show the corresponding cropped regions. Downscaling together with cropping approach was selected as a compromise, so that a meaningful part of an image can be shown on the HDR monitor. Objective quality metrics were computed on the cropped versions of the images.

2.2. Subjective evaluations

The experiments were conducted at the MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU-R [5]. The test room is equipped with a controlled lighting system with a 6500 K color temperature, whereas the color of all the background walls and curtains present in the test area were mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors.

To display the test stimuli, a full HD (1920×1080p) 42" Dolby Research HDR RGB backlight dual modulation display (aka Pulsar) was used. The monitor has the following specifications: full Rec. 709 color gamut, 4000 $\mathrm{cd/m^2}$ peak luminance, low black level (0.005 $\mathrm{cd/m^2}$), 12 bits/color input with accurate and reliable reproduction of color and luminance. In the experiments, the luminance of the background behind the monitor was about 20 $\mathrm{cd/m^2}$. The ambient illumination did not directly reflect off of the display.

In every session, three subjects were assessing the displayed images simultaneously. They were seated in one row, aligned with the center of the monitor, at a distance of 3.2 times the picture height, as suggested in [6].

The paired comparison evaluation methodology was selected for its high accuracy and reliability in constructing a scale of perceptual preferences. The image pairs were presented in side-by-side fashion to minimize visual working memory limitations. Since only one full HD 1920×1080 HDR monitor was available, each image was cropped to 950×1080 pixels (for details see Section 2.1) with 20 pixels of black border separating the two images. Subjects were asked to judge which image in a pair ('left' or 'right') has the best overall quality. The option 'same' was also included to avoid random preference selections. For each of the 5 contents, all the possible combinations of the 4 bit rates were considered, i.e., 6 pairs for each content, leading to a total of $5 \times 6 = 30$ paired comparisons for all contents.

Before the experiment, a consent form was handed to subjects for signature and oral instructions were provided to explain their tasks. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively. A training session was organized using additional contents to allow subjects to familiarize with the assessment procedure.

To reduce contextual effects, the stimuli orders of display were randomized applying different permutations for each group of subjects and special care was taken for the same content not to be shown consecutively.

A total of 20 naïve subjects (13 females and 7 males) took part in the evaluation. They were between 20 and 34 years old with an average of 25.3 years of age.

The Thurstone Case V model [7] was used to convert the ratings from the ternary scale to continuous-scale quality score values, which are equivalent to mean opinion scores (MOS), considering ties as being half way between the two preference options. For each content, the quality score values were converted to the range $[1, 5]$ by mapping the lowest and highest quality score values to 1 and 5, respectively, as the lower and upper bit rates were selected to be representative of the lowest and best quality (see Section 2.1), respectively. The intermediate values were scaled proportionally.

3. OBJECTIVE QUALITY METRICS

In this study, the performance of a set of 13 full-reference objective metrics in predicting HDR image quality was assessed:

1. MSE: Mean Squared Error,
2. PSNR: Peak Signal-to-Noise Ratio,
3. SNR: Signal-to-Noise Ratio,
4. WSNR: Weighted Signal-to-Noise Ratio [8, 9],
5. SSIM: Structural Similarity index [10],
6. MS-SSIM: Multiscale SSIM index [10],
7. VIF: Visual Information Fidelity [11],
8. VIFp: Visual Information Fidelity pixel-based [11],
9. UQI: Universal Quality Index [12],
10. IFC: Image Fidelity Criterion [13],
11. FSIM: Feature Similarity Index [14],
12. HDR-VDP-2: High Dynamic Range Visible Difference Predictor [15],
13. CIEDE2000 color difference [16].

Table 2: Accuracy and monotonicity indexes for the different metrics.

Metric	Luma component only			All components		
	PCC	SROCC	RMSE	PCC	SROCC	RMSE
MSE	0.8794	0.6935	0.7866	0.8778	0.6655	0.7909
PSNR	0.6591	0.5167	1.2369	0.6164	0.5533	1.2950
SNR	0.8794	0.7375	0.7829	0.7355	0.6352	1.1143
WSNR	0.8099	0.7589	0.9647	0.8785	0.7672	0.7858
SSIM	0.7580	0.7375	1.1185	0.8091	0.8352	1.0448
MS-SSIM	0.8651	0.7131	0.8311	0.8157	0.7176	0.9657
VIF	0.6740	0.5588	1.2163	0.4820	0.1346	1.4468
VIFp	0.7533	0.6871	1.0817	0.3504	0.2611	1.5408
UQI	0.8068	0.8077	0.9725	0.7851	0.7864	1.0189
IFC	0.8833	0.8032	0.7709	0.8256	0.8337	0.9281
FSIM	0.9043	0.8245	0.7021	0.7692	0.7818	1.0513
HDR-VDP-2	0.9337	0.8657	0.5912	0.9241	0.7866	0.6284
CIEDE2000				0.5096	0.5191	1.4174

Almost all the objective metric that were analyzed, except for CIEDE2000, are typically computed on the luma component only. In this study, all HDR images were converted to the $Y'C_bC_r$ color space [17] and these metrics were applied to the components Y', C_b, and C_r separately. In this paper, the results of the metrics were computed in two different ways: on the luma component only and on all components, considering the average value computed on Y', C_b, and C_r. Regarding the PSNR metric, the maximum value of the image after conversion to $Y'C_bC_r$ was considered for the peak value. For the HDR-VDP-2 metric, the parameters were set according to the setup of the subjective evaluations (see Section 2.2) and only the quality value was used. To compute the CIEDE2000 color difference, all HDR images were converted to the CIELAB color space using Banterle's HDR toolbox for MATLAB[2].

4. PERFORMANCE INDEXES

The results of the subjective tests can be used as ground truth to evaluate how well the objective metrics estimate perceived quality. The result of execution of a particular objective metric is an image quality rating (IQR), which is expected to be the estimation of the MOS corresponding to the compressed HDR image. To be compliant with the standard procedure for evaluating the performance of objective metrics [18], the following properties of the IQR estimation of MOS should be considered: accuracy, monotonicity, and consistency. Consistency estimation is based on the confidence intervals, which are computed assuming a standard distribution of the subjective scores. In this study, the Thurstone Case V model was used to convert the paired comparison ratings to equivalent MOS values (see Section 2.2). Confidence intervals can be estimated from the paired comparison ratings, but their nature is different from that of confidence intervals computed directly on a discrete or continuous ratings scale. Therefore, only accuracy and monotonicity were considered.

First, a regression was fitted to each [IQR, DMOS] data set using logistic fitting:

$$MOS_p(IQR) = a + \frac{b}{1 + \exp\left[-c\left(IQR - d\right)\right]}$$

where a, b, c, and d are the parameters of the fitting function.

Then, the Pearson linear correlation coefficient (PCC) and the root-mean-square error (RMSE) were computed between MOS_p and MOS to estimate accuracy of the IQR. To estimate monotonicity, the Spearman rank order correlation coefficient (SROCC) was computed between MOS_p and MOS.

The RMSE is defined as follow:

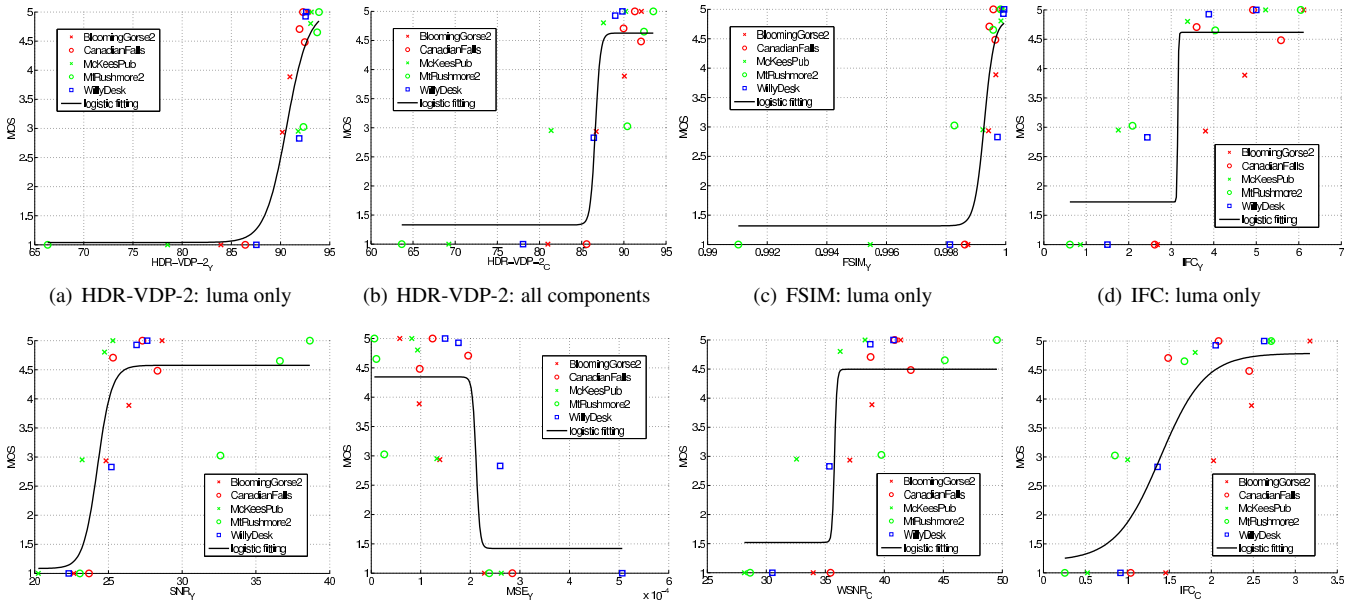
$$\text{RMSE} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(MOS_i - MOS_{pi}\right)^2}$$

where N is the total number of points.

To determine whether the difference between two performance index values corresponding to two different metrics is statistically significant, a statistical test was performed according to [19].

5. RESULTS

Table 2 reports the accuracy and monotonicity indexes, as defined in Section 4, for the different metrics computed on the luma component only and on all components. The fitting was applied on all contents at once. Results show that HDR-VDP-2, FSIM (luma only), IFC (luma only), SNR (luma only), MSE (luma only), and WSNR (all components) are among the best metrics, with a PCC above 0.87 and a RMSE below 0.79. On the other hand, results indicate that VIF, VIFp, and CIEDE2000 computed on all components perform the worst, with a PCC and SROCC below 0.51 and RMSE



(a) HDR-VDP-2: luma only (b) HDR-VDP-2: all components (c) FSIM: luma only (d) IFC: luma only

(e) SNR: luma only (f) MSE: luma only (g) WSNR: all components (h) IFC: all components

Fig. 2: Subjective versus objective results.

above 1.4. In many benchmarking performed on LDR content, VIF(p) is often among the best metrics and shows lower content dependency when compared to other metrics [19]. However, in this study, VIF(p) showed quite strong content dependency, which explains the low performance when considering all contents at once. As it can be observed, PSNR also shows quite poor performance, with a PCC between 0.6-0.66 and a MSE around 1.25. The low performance of PSNR may be due to the maximum possible pixel value, which is not well defined in the case of HDR content, used for computing PSNR.

Even though SSIM and MS-SSIM often have a good correlation with perceived quality, they are criticized by many researchers as it is hard to interpret their output values when compared to PSNR values. In most cases, the SSIM and MS-SSIM values only cover a very limited range, typically $[0.8, 1]$, when compared to the theoretical $[0, 1]$ range. In this study, the SSIM and MS-SSIM values are in the range $[0.99997, 1]$ and $[0.999997, 1]$, respectively. Therefore, the relative change between the worst and best qualities for SSIM and MS-SSIM is less than 0.003% and 0.0003%, respectively, which is almost imperceptible, especially for MS-SSIM. These findings suggest that SSIM and MS-SSIM should be adapted to cope with HDR images.

As it can be observed, the performance of VIF and especially VIFp drop drastically when considering all components. To further understand whether there is a statistically significant difference between the performance of each metric when computed on the luma component only and when computed on all components, a statistical analysis was performed on the different performance indexes. Results show that there is no significant different in the performance of the metric between the two approaches for any of the metrics. However, because of the relatively low number of conditions (20 stimuli), general conclusions should not be drawn from these results. As HDR is often considered in combination with wide color gamut, it is expected that the fidelity of color reproduction will play a more important role in the context of HDR when compared to LDR.

Figure 2 depicts the scatter plots of subjective versus objective results for some of the metrics considered in this study. The metrics that perform the best according to the performance indexes exhibit a very abrupt transition from low to high quality. Such binary behavior is not well suited for objective quality metrics, which are expected to discriminate between several granularities of distortions. This finding implies that these metrics do not correlate well with human perception of visual quality, as the response of the HVS is expected to be smoother and not abrupt, and that the performance indexes are not sufficient to select a good metric. On the other hand, IFC computed on all components performs lower but has a smoother transition between low and high quality. HDR-VDP-2 is the only metric considered in this study that was originally designed for HDR content. However, the performance of this metric is not significantly better than that of state-of-the-art metrics designed for LDR content. Overall, results show that there is a great room for improvement to better predict the perceived quality of HDR content.

6. CONCLUSION

In this paper, the correlation between thirteen well known full-reference metrics and perceived quality of compressed HDR content was investigated. Five HDR images were compressed with the upcoming JPEG XT HDR image coding standard to generate different distortion levels. The objective metrics were benchmarked using ground truth subjective scores collected during quality evaluations performed on a Dolby Pulsar HDR monitor. Results show that HDR images are challenging for objective metrics and that the most commonly used metrics, e.g., PSNR, SSIM, and MS-SSIM, predict perceived quality of HDR content unreliably. Only two metrics, HDR-VDP-2 metric designed for HDR content and FSIM metric computed on the luma component, predict human perception of visual quality reasonably well. However, they are complex and require heavy computational resources, which limits their use in many applications.

7. REFERENCES

[1] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective evaluation of higher dynamic range video," in *SPIE Applications of Digital Image Processing XXXVII*, San Diego, California, USA, Aug. 2014.

[2] P. Korshunov, H. Nemoto, A. Skodras, and T. Ebrahimi, "Crowdsourcing-based evaluation of privacy in HDR images," in *Proc. SPIE*, Brussels, Belgium, Apr. 2014, vol. 9138, pp. 913802–913802–11.

[3] T. Richter, "On the standardization of the JPEG XT image compression," in *Picture Coding Symposium (PCS)*, San Jose, CA, USA, Dec. 2013, pp. 37–40.

[4] S. Winkler, *Digital Video Quality: Vision Models and Metrics*, John Wiley & Sons, Mar. 2005.

[5] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Jan. 2012.

[6] ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays," International Telecommunication Union, Aug. 2012.

[7] L. L. Thurstone, "A law of comparative judgment," *Psychological review*, vol. 34, no. 4, pp. 273, 1927.

[8] J. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, 1974.

[9] T. Mitsa and K. L. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Minneapolis, Minnesota, USA, 1993, vol. 5, pp. 301–304.

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

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

[12] Z. Wang and A. C. Bovik, "A universal image quality index," *Signal Processing Letters, IEEE*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[13] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[14] L. Zhang, D. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[15] 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 Transactions on Graphics*, vol. 30, no. 4, pp. 40:1–40:14, July 2011.

[16] M. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Research & Application*, vol. 26, no. 5, pp. 340–350, Oct. 2001.

[17] ITU-R BT.601, "Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios," International Telecommunication Union, Mar. 2011.

[18] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, July 2012.

[19] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," in *18th International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, July 2013, pp. 1–8.