EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Benchmarking of objective quality metrics for HDR image quality assessment

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

Abstract

Recent advances in high dynamic range (HDR) capture and display technologies have attracted a lot of interest from scientific, professional, and artistic communities. As in any technology, the evaluation of HDR systems in terms of quality of experience is essential. Subjective evaluations are time consuming and expensive, and thus objective quality assessment tools are needed as well. In this paper, we report and analyze the results of an extensive benchmarking of objective quality metrics for HDR image quality assessment. In total, 35 objective metrics were benchmarked on a database of 20 HDR contents encoded with 3 compression algorithms at 4 bit rates, leading to a total of 240 compressed HDR images, using subjective quality scores as ground truth. Performance indexes were computed to assess the accuracy, monotonicity, and consistency of the metric estimation of subjective scores. Statistical analysis was performed on the performance indexes to discriminate small differences between metrics. Results demonstrated that metrics designed for HDR content, i.e., HDR-VDP-2 and HDR-VQM, are the most reliable predictors of perceived quality. Finally, our findings suggested that the performance of most full-reference metrics can be improved by considering non-linearities of the human visual system, while further efforts are necessary to improve performance of no-reference quality metrics for HDR content.

Keywords: Image quality assessment, Objective metrics, High dynamic range, JPEG XT

1 Introduction

Recently, the world of multimedia has been observing a growth in new imaging modalities aiming at improving user immersion capability, providing more realistic perception of content, and consequently, reaching new levels of quality of experience. This trend has began with the introduction of 3D capable devices in the consumer market, providing depth perception, followed by ultra high definition (UHD), focused on higher pixel resolutions beyond high definition, high frame rate (HFR), to provide more fluid motion, and, more recently, high dynamic range (HDR), intended to capture a wider range of luminance values. Moreover, academia, industry and service providers have proposed new models to further enrich content, such as plenoptic and holographic systems, although these latest modalities are still in very early stages. Many aspects need further improvement on such new trends. For instance, consumers still experience some lack of reliable 3D content and still suffer from discomfort caused by long exposure. UHD systems face a small number of appropriate content although they have been evolving into consumer markets.

HDR imaging systems are stepping into the multimedia technologies for consumer market. HDR pursues a more complete representation of information that the human eye can see, capturing all the brightness information of the visible range of a scene, even in extreme lighting conditions. Hence, it pursues the representation of the entire dynamic range and color gamut perceived by human visual system (HVS). 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].

HDR systems are becoming available for the general public. Acquisition systems are present in a large variety of photographic equipment and even in some mobile devices. Typically, computer rendering and merging of multiple low dynamic range (LDR) images taken at different exposure settings are the two methods used to generate HDR images [3]. Nowadays, HDR images can

*Correspondence: philippe.hanhart@epfl.ch
[1] Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland
Full list of author information is available at the end of the article

also be acquired using specific image sensors. HDR displays are also becoming increasingly available and enable representation of better contrasts, higher luminance, and wider color gamut [4]. Optionally, tone mapping operators (TMO) that map HDR content into the luminance range and color gamut of conventional displays can be used [5].

In addition to the acquisition and display technologies, JPEG has been standardizing new codecs for HDR content. JPEG XT is a recent standard for JPEG backward-compatible compression of HDR images [6]. Using this compression standard, HDR images are coded in two layers. A tone-mapped version of the HDR image is encoded using the legacy JPEG format in a base layer, and the extra HDR information is encoded in a residual layer. The advantage of this layered scheme 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 JPEG XT compliant decoder can use the residual layer to reconstruct a lossy or even lossless version of the HDR image. Currently, JPEG XT defines four profiles (A, B, C, and D) for HDR image compression, of which profile D is a very simple entry-level decoder that roughly uses the 12-bit mode of JPEG. Profiles A, B, and C all take into account the non-linearity of the human visual system. They essentially differ on the strategy used for creating the residual information and on the pre- and post-processing techniques. In profile A, the residual is represented as a ratio of the luminance of the HDR image and the tone-mapped image after inverse gamma correction. The residual is log-encoded and compressed as an 8-bit greyscale image [7]. In profile B, the image is split into "overexposed" areas and LDR areas. The extension image is represented as a ratio of the HDR image and the tone-mapped image, after inverse gamma correction. Note that instead of a ratio, profile B uses a difference of logarithms. Finally, rofile C computes the residual image as a ratio of the HDR image and the inverse tone-mapped image. Unlike the other profiles, the inverse TMO is not a simple inverse gamma, but rather a global approximation of the inverse of the (possibly local) TMO that was used to generate the base-layer image. Similarly to profile B, the ratio is implemented as a difference of logarithms. However, instead of using the exact mathematical log operation, profile C uses a piecewise linear approximation, defined by re-interpreting the bit-pattern of the half-logarithmic IEEE representation of floating-point numbers as integers, which is exactly invertible [8]. MPEG is also starting a new standardization effort on HDR video [9], revealing the growing importance of HDR technologies.

As for any technology, evaluation of HDR systems, in terms of quality of experience, is essential. Subjective evaluations are time consuming and expensive, thus objective quality assessment tools are needed as well. To the best of our knowledge, only three objective metrics have been developed so far for HDR content. The most relevant work on this domain is the HDR visual detection predictor (HDR-VDP) metric proposed by Mantiuk et al. [10], which is an extension of Daly's VDP [11] for the HDR domain. The second version of this metric, HDR-VDP-2 [12, 13], is considered as the state-of-the-art in HDR image quality assessment. The dynamic range independent metric (DRIM) proposed in [14] can also be used for HDR quality assessment. Nevertheless, this metric results in three distortion maps, which is difficult to interpret, as there is no pooling of the different values. Recently, the high dynamic range video quality metric (HDR-VQM) was proposed by Narwaria et al. [15]. The metric was designed for quality assessment of HDR video content, but can also be used for HDR still images.

To overcome the lack of HDR objective metrics, LDR metrics, e.g., PSNR, were also used to evaluate HDR quality, especially in early HDR studies. However, LDR metrics are designed for gamma encoded images, typically having luminance values in the range 0.1–100 cd/m^2, while HDR images have linear values and are meant to capture a much wider range of luminance. Originally, gamma encoding was developed to compensate for the characteristics of cathode ray tube (CRT) displays, but it also takes advantage of the non-linearity in HVS to optimize quantization when encoding an image [16]. Under common illumination conditions, the HVS is more sensitive to relative differences between darker and brighter tones. According to Weber's law, the HVS sensitivity approximately follows a logarithm function for light luminance values [17]. Therefore, in several studies, LDR metrics have been computed in the log domain to predict HDR quality. However, at the darkest levels, the HVS sensitivity is closer to a square-root behavior, according to Rose-DeVries law [18, 19]. To extend the range of LDR metrics and to consider the sensitivity of the HVS, Aydin et al. [20] have proposed the perceptually uniform (PU) encoding. Another approach to apply LDR metrics on HDR images was proposed in [21]. This technique consists in tone-mapping the HDR image to several LDR images with different exposure ranges and to take the average objective score computed on each exposure. However, this approach is more time consuming and requires more computational power, proportionally to the number of exposures.

For LDR content, extensive studies have shown that not all metrics can be considered as reliable predictors of perceived quality [22, 23], while only a few recent studies have benchmarked objective metrics for HDR quality assessment. The study of Valenzise et al. [24] compared the performance of PSNR and SSIM, computed in the logarithmic and PU [20] spaces, and HDR-VDP. The authors have concluded that non-uniformity must be corrected for

a proper metric application, as most have been designed for perceptual uniform scales. Another subjective study was reported by Mantel et al. in [25]. A comparison with objective metrics in physical domain and using a gamma correction to approximate perceptually uniform luminance is also presented, concluding that the mean relative squared error (MRSE) metric provides the best performance in predicting quality. The correlation between 13 well-known full-reference metrics and perceived quality of compressed HDR content is investigated in [26]. The metrics were applied on the linear domain, and results show that only HDR-VDP-2 and FSIM predicted visual quality reasonably well. Finally, Narwaria et al. [15] have reported that their HDR-VQM metric performs similar or slightly better than HDR-VDP-2 for HDR image quality assessment. Regarding HDR video quality assessment, four studies were also reported by Azimi et al. [27], Rerabek et al. [28], Hanhart et al. [9], and Narwaria et al. [15]. The authors of [15] found that HDR-VQM is the best metric, far beyond HDR-VDP-2, in contradiction to the findings of [9], which showed lower performance for HDR-VQM when compared to HDR-VDP-2. Also, the other two studies found that HDR-VDP-2 has the highest correlation. The divergence between these findings might be due to the contents and types of artifacts considered in the different studies. Indeed, the first three studies consider HDR video sequences captured using HDR video cameras, manually graded, and encoded with compression schemes based on AVC or HEVC, whereas Narwaria et al. have mostly used computer-generated contents, an automatic algorithm to adjust the luminance of the HDR video sequences, and their own backward compatible HDR compression algorithm.

The main limitation of these studies lies in the small number of images or video sequences used in their experiments, which was limited to five or six contents. Also, a proper adaptation of the contents to the HDR display and correction of the metrics for non-uniformity were not always considered. Therefore, in this paper, we report and analyze the results of an extensive benchmarking of objective quality metrics for HDR image quality assessment. In total, 35 objective metrics were benchmarked using subjective scores as ground truth. The database used in our experiments [29, 30] is composed of 20 different contents and a total of 240 compressed HDR images with corresponding subjective quality scores. The HDR images are adapted (resized, cropped, and tone-mapped using display-adaptive tone-mapping operator) to SIM2 HDR monitor. The objective metrics were computed in the linear, logarithmic, PU [20], and Dolby perceptual quantizer (PQ) [31] domains. Additionally, the metrics were computed both on the luminance channel alone and as the average quality score of the Y, C_b, and C_r channels. For each metric, objective scores were fitted to subjective scores using logistic fitting. Performance indexes were computed to assess the accuracy, monotonicity, and consistency of the metrics estimation of subjective scores. Finally, statistical analysis was performed on the performance indexes computed from 240 data points to discriminate small differences between two metrics. 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 2.1. The different metrics benchmarked in this study are defined in Section 2.2. In Section 2.3, the methodology used to evaluate the performance of the metrics is described. Section 3 provides a detailed analysis of the objective results and discusses the reliability of objective metrics. Finally, Section 4 concludes the paper.

2 Methodology

The results of subjective tests can be used as ground truth to evaluate how well objective metrics estimate perceived quality. In this paper, we use the publicly available dataset provided by Korshunov et al. [29, 30] to benchmark 35 objective metrics. This section describes in details the dataset, objective metrics, and performance analysis used in our benchmark.

2.1 Dataset and subjective scores

The dataset is composed of 20 HDR images with a resolution of 944×1080 pixels. The dataset contains scenes with architecture, landscapes, and portraits and is composed of HDR images fused from multiple exposure pictures, frames extracted from HDR video, and computer-generated images. Since publicly available HDR images are usually not graded, the images are adjusted for a SIM2 HDR monitor using a display-adaptive TMO [32] to map the relative radiance representation of the images to an absolute radiance and color space of the HDR monitor. These display-adapted images are then considered as original images and compressed with JPEG XT using profiles A, B, and C. The base and extension layers chroma-subsampling are set to 4:2:0 and 4:4:4, respectively, while optimized Huffman coding is enabled for all implementations. For each content and profile, four different bit rates were selected, leading to a total of 240 compressed HDR images. Figure 1 shows tone-mapped versions of the images in the dataset, and Table 1 reports the dynamic range and key [33] characteristics of these images. The key is in the range $[0, 1]$ and gives a measure of the overall brightness

$$key = \frac{\log L_{\text{avg}} - \log L_{\text{min}}}{\log L_{\text{max}} - \log L_{\text{min}}} \tag{1}$$

Fig. 1 Display-adapted images of the dataset. The *reinhard02* TMO was used for images from **a**–**g** and the *mantiuk06* TMO was used for the remaining images. Copyrights: *2006-2007 Mark D. Fairchild, †Blender Foundation | www.sintel.org, under Creative Commons BY, #Mark Evans, under Creative Commons BY

where $L_{\min}$, $L_{\max}$, and L_{avg} are the minimum, maximum, and average luminance values, respectively, computed after excluding 1 % of the darkest and lightest pixels.

The evaluation was performed using a full HD 47" SIM2 HDR monitor with individually controlled LED backlight modulation, capable of displaying content with luminance values ranging from 0.001 to 4000 cd/m^2. In every session, three subjects assessed the displayed test images simultaneously. They were seated in an arc configuration, at a constant distance of 3.2 times the picture height as recommended in [34], which corresponds to 1.87 m and a visual resolution of 60 pixels per degree. The laboratory was equipped with a controlled lighting system with a 6500 K color temperature, while a mid gray color was used for all background walls and curtains. The background luminance behind the monitor was set to 20 cd/m^2 and did not directly reflect off of the monitor.

The double-stimulus impairment scale (DSIS) Variant I methodology [35] was used for the evaluation. For scoring, a five-grade impairment scale (1: *very annoying*, 2: *annoying*, 3: *slightly annoying*, 4: *perceptible, but not annoying*, 5: *imperceptible*) was used. Two images were presented in side-by-side fashion with 32 pixels of black border separating the two images: one of the two images was always the reference (unimpaired) image, while the other was the test image. To reduce the effect of order of images on the screen, the participants were divided into two groups: the left image was always the reference image for the first group, whereas the right image was always the reference image for the second group. After the presentation of each pair of images, a six-second voting time followed. Subjects were asked to rate the impairments of the test images in relation to the reference image.

Before the experiment, oral instructions were provided to the subjects to explain their tasks. Additionally, a training session was organized, allowing subjects to familiarize themselves with the test procedure. For this purpose two images outside of the dataset were used. Five samples were manually selected by expert viewers for each image so that the quality of samples was representative of the rating scale.

Since the total number of test samples was too large for a single test session, the overall experiment was split into three sessions of approximately 16 min each. Between the sessions, subjects took a 15-min break. The test material was randomly distributed over the test sessions. To reduce contextual effects, the order of displayed stimuli was randomized applying different permutations for each group of subjects, whereas the same content was never shown consecutively.

Table 1 Characteristics of HDR images from the dataset

	Dynamic range	Key
507	4.097	0.743
AirBellowsGap	4.311	0.768
BloomingGorse2	2.336	0.748
CanadianFalls	2.175	0.729
DevilsBathtub	2.886	0.621
dragon	4.386	0.766
HancockKitchenInside	4.263	0.697
LabTypewriter	4.316	0.733
LasVegasStore	4.131	0.636
McKeesPub	3.943	0.713
MtRushmore2	4.082	0.713
PaulBunyan	2.458	0.702
set18	4.376	0.724
set22	3.162	0.766
set23	3.359	0.764
set24	3.862	0.778
set31	4.118	0.678
set33	4.344	0.698
set70	3.441	0.735
showgirl	4.369	0.723
sintel	3.195	0.781
WillyDesk	4.284	0.777
Min	2.175	0.621
Max	4.386	0.781
Mean	3.722	0.727
Median	4.089	0.731

A total of 24 naïve subjects (12 females and 12 males) took part in the experiments. Subjects were aged between 18 and 30 years old with an average of 22.1. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

The subjective scores were processed by first detecting and removing subjects whose scores deviated strongly from others. The outlier detection was applied to the set of results obtained from the 24 subjects and performed according to the guidelines described in Section 2.3.1 of Annex 2 of [35]. Two outliers were detected. Then, the mean opinion score (MOS) was computed for each test stimulus as the mean score across the 22 valid subjects, as well as the associated 95 % confidence interval (CI), assuming a Student's t-distribution of the scores. As it can be observed in Fig. 2, MOS values reflect the subjects perception fairly with enough MOS samples for each meaningful value range. More details about the dataset and subjective evaluations can be found in [29].

2.2 Objective quality metrics

Depending on the amount of information required about the reference image, objective metrics can be classified into three categories:

i) Full-reference (FR) metrics, which compare the test image with a reference image
ii) Reduced-reference (RR) metrics, which have access to a number of features from the reference image, extract the same features from the test image and compare them
iii) No-reference (NR) metrics, which do not use any information about the reference image

In this study, only FR and NR metrics were considered.

2.2.1 Full-reference metrics

To the best of our knowledge, there are only two metrics for HDR quality assessment that have a publicly available implementation: (1) HDR-VDP: high dynamic range visible difference predictor [10, 12, 13] and (2) HDR-VQM: an objective quality measure for high dynamic range video [15].

The original HDR-VDP metric [10] was the first metric designed for HDR content. It is an extension of the VDP model [11] that considers a light-adaptive contrast sensitivity function (CSF), which is necessary for HDR content as the ranges of light adaptation can vary substantially. The metric was further extended [12] with different features, including a specific model of the point spread function (PSF) of the optics of the eye, as human optical lens flare can be very strong in high contrast HDR content. The front-end amplitude non-linearity is based on integration of the Weber-Fechner law. HDR-VDP is a calibrated metric and takes into account the angular resolution. The metric uses a multi-scale decomposition. A neural noise block is defined to calculate per-pixel probabilities maps of visibility and the predicted quality metric. In this study, we used the latest version of HDR-VDP, i.e., version 2.2.1 [13], referred to as HDR-VDP-2 in this paper.

HDR-VQM was designed for quality assessment of HDR video content. The metric is computed in the PU space and relies on a multi-scale and multi-orientations analysis, similarly to HDR-VDP, based on a subband decomposition using log-Gabor filters to estimate the subband errors. The subband errors are pooled over non-overlapping spatiotemporal tubes to account for short-term memory effects. Further spatial and long-term temporal poolings are performed to compute the overall quality score. In the case of still images, only spatial pooling is performed.

The remaining FR metrics considered in this study are all designed for LDR content and can be divided into different categories: difference measures and statistical-oriented metrics, structural similarity measures, visual

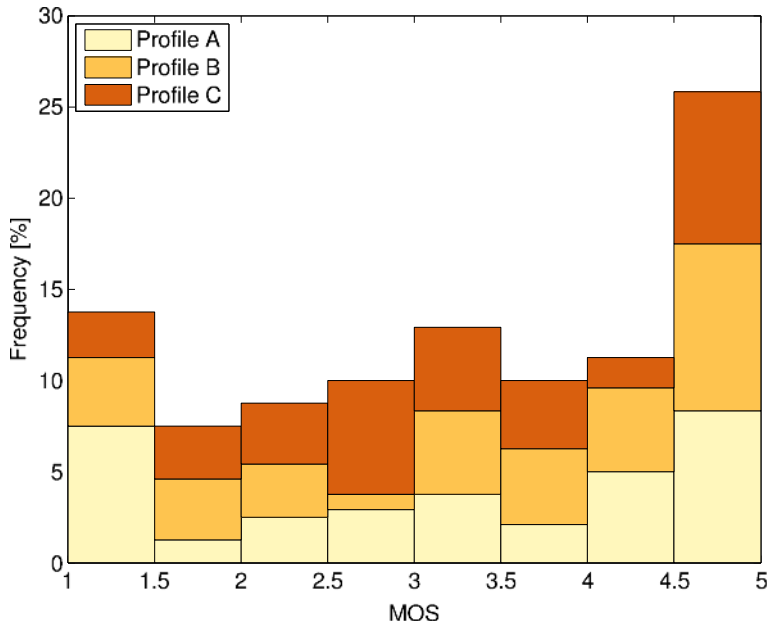

Fig. 2 MOS values distribution

information measures, information weighted metrics, HVS inspired metrics, and objective color difference measures (studied in the vision science). The complete list of considered FR metrics is provided in the following:

- Difference measures and statistical-oriented metrics
  These metrics are based on pixel color differences and provide a measure of the difference between the reference image and the distorted image. The following metrics of this category were considered: (3) MSE: mean squared error, (4) PSNR: peak signal-to-noise ratio, and (5) SNR: signal-to-noise ratio.
- Structural similarity measures
  These metrics model the quality based on pixel statistics to model the luminance (using the mean), the contrast (variance), and the structure (cross-correlation) [36]. The metrics considered in this category are the following: (6) UQI: universal quality index [37], (7) SSIM: structural similarity index [38], (8) MS-SSIM: multiscale SSIM index [39], (9) M-SVD: measure - singular value decomposition [40], and (10) QILV: quality index on local variance [41].
  The MS-SSIM index is a multiscale extension of SSIM, which has a higher correlation with perceived quality when compared to SSIM. It is a perceptual metric based on the content features extraction and abstraction. This quality metric considers that the HVS uses the structural information from a scene [38]. The structure of objects in the scene can be represented by their attributes, which are independent of both contrast and average luminance. Hence, the changes in the structural information from the reference and distorted images can be perceived as a measure of distortion. The MS-SSIM algorithm calculates multiple SSIM values at multiple image scales. By running the algorithm at different scales, the quality of the image is evaluated for different viewing distances. MS-SSIM also puts less emphasis on the luminance component when compared to contrast and structure components [39].
- Visual information measures
  These metrics aim at measuring the image information by modeling the psycho-visual features of the HVS or by measuring the information fidelity. Then, the models are applied to the reference and distorted images, resulting in a measure of the difference between them. The following metrics on this category were considered: (11) IFC: image fidelity criterion [42], (12) VIF: visual information fidelity [43], (13) VIFp: VIF pixel-based version [43], and (14) FSIM: feature similarity index [44].
  The VIF criterion analyses the natural scene statistics, using an image degradation model and the HVS model. This FR metric is based on the quantification of the Shannon information present in both the reference and the distorted images. VIFP is derived from the VIF criterion.

FSIM is a perceptual metric that results from SSIM. FSIM adds the comparison of low-level feature sets between the reference and the distorted images [44]. Hence, FSIM analyzes the high phase congruency extracting highly informative features and the gradient magnitude, to encode the contrast information. This analysis is complementary and reflects different aspects of the HVS in assessing the local quality of an image.

- Information weighted metrics
  The metrics in this category are based on the modeling of relative local importance of the image information. As not all regions of the image have the same importance in the perception of distortion, the image differences computed by any metrics have allocated local weights resulting in a more perceptual measure of quality. The following metrics were computed: (15) IW-MSE: information content weighting MSE [45], (16) IW-PSNR: information content weighting PSNR [45], and (17) IW-SSIM: information content weighting SSIM [45].
- HVS-inspired metrics
  These metrics try to model empirically the human perception of images from natural scenes. The following metrics were considered: (18) JND_st: just noticeable distortion [46], (19) WSNR: weighted SNR [47, 48], and (20) DN: divisive normalization [36].
- Objective color difference measures
  The color difference metrics were developed because the CIE1976 color difference [49] magnitude in different regions of the color space did not appear correlated with perceived colors. These metrics were designed to compensate the non-linearities of the HVS present on the CIE1976 model. The following CIE metrics were computed: (21) CIE1976 [49], (22) CIE94 [50], (23) CMC [51], and (24) CIEDE2000 [52]. The CIEDE2000 metric is a color difference measure that includes not only weighting factors for lightness, chroma, and hue but also factors to handle the relationship between chroma and hue. The CIEDE2000 computation is not reliable in all color spaces. However, in this case, it can be used because the tested images are represented in the CIELAB color space that allows a precise computation.

2.2.2 No-reference metrics

These metrics are based on the analysis of a set of well-known sharpness measures. The following NR metrics were considered: (25) JND: just noticeable distortion [46], (26) VAR: variance [53], (27) LAP: laplacian [54], (28) GRAD: gradient [54], (29) FTM: frequency threshold metric [55], (30) HPM: HP metric [56], (31) Marziliano: Marziliano blurring metric [55], (32) KurtZhang: kurtosis-based metric [57], (33) KurtWav: kurtosis of wavelet coefficients [58], (34) AutoCorr: auto correlation [54], and (35) RTBM: Riemannian tensor-based metric [59].

2.2.3 Metrics computation and transform domains

LDR metrics are designed for gamma encoded images, typically having luminance values in the range 0.1–100 cd/m^2, while HDR images have linear values and are meant to capture a much wider range of luminance. Therefore, in this study, metrics were computed not only in the linear space but also in transformed spaces that provide a more perceptual uniformity. This space transformation was not applied to HDR-VDP-2 and HDR-VQM, which are calibrated metrics and require absolute luminance values as input. The color difference metrics, i.e., CIE1976, CIE94, CMC, and CIEDE2000, were also not computed in transformed spaces. These color difference measures require a conversion from the RGB representation to the CIELAB color space, considering a D65 100 cd/m^2 reflective white point as reference white point.

Before any metric was computed, images were clipped to the range [0.001,4000] cd/m^2 (theoretical range of luminance values that the HDR monitor used in the subjective tests can render) to mimic the physical clipping performed by the HDR display. To compute the metrics in the linear domain, these luminance values were normalized to the interval $[0, 1]$. This normalization was not applied to HDR metrics and to color difference metrics.

The remaining metrics were computed in three transform domains: the log domain, the PU domain [20], and the PQ domain [31]. The PU transform is derived using the threshold-integration method [60]. The transform is constrained such that luminance values in the range 0.1–80 cd/m^2, as produced by a typical CRT display, are mapped to the range 0–255 to mimic the sRGB non-linearity. The PQ transform is derived from the Barten contrast sensitivity function [61]. The PQ curve has a square-root and log behavior at the darkest and highest light levels, respectively, while it exhibits a slope similar to the gamma non-linearities between those extreme luminance regions. Figure 3 depicts the normalized response of the log, PU, and PQ responses in the range [0,4000] cd/m^2.

These transformations were applied before any normalization and only after their application the resulting color components were normalized to the interval $[0, 1]$. After the normalizations, the values considered to be in the RGB color space were transformed to the YC_bC_r color space [62]. The exception is the DN metric, which uses directly these RGB components. The metrics were computed on each of these components separately and two final metrics were considered: the quality score computed on the luminance channel alone and the average quality score of the Y, C_b, and C_r channels.

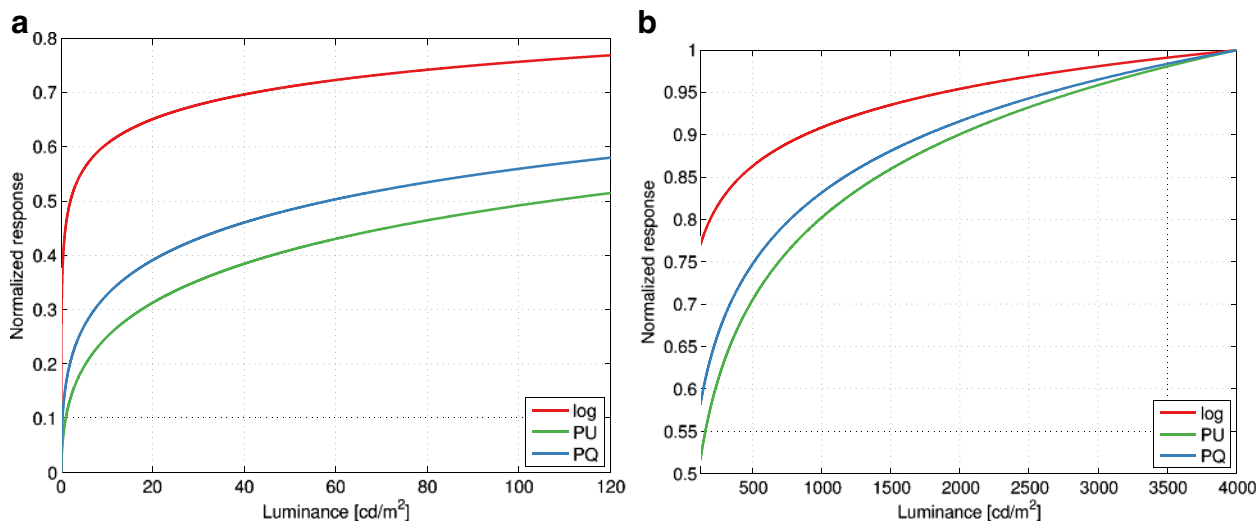

Fig. 3 Comparison of responses. Comparison of responses for the transformation functions in two different luminance ranges (a: [0, 120] cd/m^2, b: [120, 4000] cd/m^2)

2.3 Benchmarking of quality metrics

To evaluate how well an objective metric is able to estimate perceived quality, the MOS obtained from subjective experiments are taken as ground truth and compared to predicted MOS values obtained from objective metrics. To compute the predicted MOS $\tilde{M}$, a regression analysis on each objective metric results O was performed using a logistic function as a regression model:

$$\tilde{M} = a + \frac{b}{1 + \exp\left(-c \cdot (O - d)\right)} \tag{2}$$

where a, b, c, and d are the parameters that define the shape of the logistic fitting function and were determined using a least squares method.

2.3.1 Performance indexes

Performance indexes to assess the accuracy of objective metrics were computed following the same procedure as in [63]. In particular, the Pearson linear correlation coefficient (PLCC) and the unbiased estimator of the root-mean-square error (RMSE) were used. The Spearman rank order correlation (SROCC) coefficient and the outlier ratio (OR) were also used to estimate respectively the monotonicity and the consistency of the objective metric as compared with the ground truth subjective data. The OR is the ratio of points for which the error between the predicted and actual MOS values exceeds the 95 % confidence interval of MOS values.

2.3.2 Statistical analysis

To determine whether the difference between two performance index values corresponding to two different metrics is statistically significant, two-sample statistical tests were performed on all four performance indexes. In particular, for the PLCC and SROCC, a Z-test was performed using Fisher z-transformation. For the RMSE, an F-test was performed, whereas a Z-test for the equality of two proportions was performed for the OR. No processing was applied to correct for the multiple comparisons. The statistical tests were performed according to the guidelines of recommendation ITU-T P.1401 [64].

3 Results

Figures 4, 5, 6, and 7 report the accuracy, monotonicity, and consistency indexes, as defined in Section 2.3, for the metrics computed in the different domains. The metrics are sorted from best (top) to least (bottom) performing, based on the different performance indexes (higher PLCC/SROCC and lower RMSE/OR values indicate better performance). As HDR-VDP-2 and HDR-VQM require absolute luminance values as input, these metrics were computed neither on the chrominance channels nor in the transform domains. Similarly, the different color difference metrics were computed only in the linear domain, after converting the absolute RGB values to the CIELAB color space. The DN metric was computed on the RGB components, considering all three channels together. Finally, the remaining 28 metrics were computed both on the luminance channel alone (_Y suffix) and as the average quality score of the luminance, blue-difference, and red-difference channels (_M suffix). The statistical analysis results are reported in the same tables. This analysis was performed on the performance indexes computed from 240 data points to discriminate small differences between two metrics. Metrics whose performance indexes are connected by a line are considered statistically not significantly different. For example, in the linear domain,

PLCC		SROCC		RMSE		OR	
HDRVDP2	0.9604	HDRVDP2	0.9564	HDRVDP2	0.3498	HDRVDP2	0.3500
HDRVQM	0.9602	HDRVQM	0.9564	HDRVQM	0.3506	HDRVQM	0.4083
IFC_Y	0.9140	IFC_Y	0.9205	IFC_Y	0.5109	IFC_Y	0.5458
FSIM_Y	0.8938	FSIM_Y	0.9160	FSIM_Y	0.5643	UQI_Y	0.5667
UQI_Y	0.8873	IFC_M	0.8952	UQI_Y	0.5792	IWPSNR_Y	0.6167
IFC_M	0.8855	DN	0.8926	IFC_M	0.5845	IWSSIM_Y	0.6167
DN	0.8814	WSNR_Y	0.8791	DN	0.5936	IWPSNR_M	0.6417
WSNR_Y	0.8786	MSSSIM_Y	0.8776	WSNR_Y	0.5995	IWSSIM_M	0.6417
FSIM_M	0.8571	FSIM_M	0.8768	FSIM_M	0.6476	IFC_M	0.6458
MSSSIM_Y	0.8545	UQI_Y	0.8737	VIF_Y	0.6521	WSNR_Y	0.6500
VIF_Y	0.8545	VIF_Y	0.8617	MSSSIM_Y	0.6527	VIF_Y	0.6583
IWPSNR_Y	0.8352	IWMSE_Y	0.8374	IWPSNR_Y	0.6907	WSNR_M	0.6667
IWSSIM_Y	0.8352	IWPSNR_Y	0.8374	IWSSIM_Y	0.6907	VIFP_Y	0.6667
VIFP_Y	0.8275	IWSSIM_Y	0.8374	VIFP_Y	0.7050	DN	0.6667
WSNR_M	0.8178	VIFP_Y	0.8344	WSNR_M	0.7225	FSIM_M	0.6750
UQI_M	0.8082	IWMSE_M	0.8298	UQI_M	0.7396	FSIM_Y	0.6833
IWPSNR_M	0.8052	WSNR_M	0.8273	IWPSNR_M	0.7446	UQI_M	0.6917
IWSSIM_M	0.8052	IWPSNR_M	0.8145	IWSSIM_M	0.7446	CMC	0.6917
CMC	0.8045	IWSSIM_M	0.8145	CMC	0.7458	VIF_M	0.7000
CIE94	0.7987	UQI_M	0.8112	CIE94	0.7558	IWMSE_M	0.7042
CIEDE00	0.7951	CIE94	0.8031	CIEDE00	0.7615	MSSSIM_M	0.7042
IWMSE_Y	0.7951	CMC	0.8019	IWMSE_Y	0.7622	CIEDE00	0.7083
IWMSE_M	0.7907	VIF_M	0.7946	IWMSE_M	0.7689	CIE94	0.7083
VIF_M	0.7813	CIEDE00	0.7908	VIF_M	0.7836	SNR_M	0.7208
MSVD_Y	0.7629	MSSSIM_M	0.7877	MSVD_Y	0.8120	SNR_Y	0.7333
MSSSIM_M	0.7610	MSVD_Y	0.7758	MSSSIM_M	0.8146	PSNR_M	0.7375
SNR_Y	0.7535	VIFP_M	0.7657	SNR_Y	0.8253	CIE1976	0.7458
VIFP_M	0.7520	MSVD_M	0.7600	VIFP_M	0.8275	VIFP_M	0.7500
MSVD_M	0.7466	SNR_Y	0.7574	MSVD_M	0.8355	IWMSE_Y	0.7542
SSIM_Y	0.7374	SSIM_Y	0.7559	SSIM_Y	0.8482	MSSSIM_Y	0.7583
CIE1976	0.7254	MSE_Y	0.7262	CIE1976	0.8642	MSVD_M	0.7667
PSNR_Y	0.7152	PSNR_Y	0.7262	PSNR_Y	0.8774	SSIM_M	0.7708
MSE_Y	0.7050	CIE1976	0.7256	MSE_Y	0.8924	PSNR_Y	0.7792
SNR_M	0.6872	MSE_M	0.6978	SNR_M	0.9120	QILV_M	0.7792
MSE_M	0.6777	SNR_M	0.6969	MSE_M	0.9232	MSVD_Y	0.7833
PSNR_M	0.6525	PSNR_M	0.6628	PSNR_M	0.9513	MSE_M	0.8000
QILV_Y	0.6183	QILV_Y	0.6615	QILV_Y	0.9867	QILV_Y	0.8042
QILV_M	0.6164	QILV_M	0.6535	QILV_M	0.9886	SSIM_Y	0.8125
SSIM_M	0.5904	SSIM_M	0.5948	SSIM_M	1.0132	MSE_Y	0.8167
Marziliano_Y	0.4551	Marziliano_Y	0.4160	Marziliano_Y	1.1178	Marziliano_M	0.8292
Marziliano_M	0.3669	HPM_M	0.3341	Marziliano_M	1.1678	Marziliano_Y	0.8333
HPM_Y	0.3625	HPM_Y	0.3211	HPM_Y	1.1700	JND_St_M	0.8375
HPM_M	0.3503	Marziliano_M	0.3093	HPM_M	1.1758	RTBM_Y	0.8375
JND_St_Y	0.2913	JND_St_Y	0.2393	JND_St_Y	1.2009	RTBM_M	0.8375
JND_St_M	0.2509	JND_St_M	0.1599	JND_St_M	1.2152	JND_St_Y	0.8500
GRAD_M	0.1060	LAP_M	0.1197	GRAD_M	1.2483	VAR_Y	0.8500
GRAD_Y	0.1012	LAP_Y	0.1030	GRAD_Y	1.2489	VAR_M	0.8500
KurtZhang_M	0.0829	GRAD_M	0.0784	KurtZhang_M	1.2513	LAP_M	0.8500
KurtWav_M	0.0709	GRAD_Y	0.0742	KurtWav_M	1.2522	GRAD_M	0.8500
RTBM_Y	0.0709	KurtWav_M	0.0607	RTBM_Y	1.2522	FTM_Y	0.8500
AutoCorr_Y	0.0675	KurtZhang_Y	0.0518	AutoCorr_Y	1.2525	HPM_Y	0.8500
AutoCorr_M	0.0664	KurtZhang_M	0.0366	AutoCorr_M	1.2526	KurtZhang_Y	0.8500
RTBM_M	0.0623	AutoCorr_Y	0.0301	RTBM_M	1.2529	KurtWav_Y	0.8500
LAP_M	0.0537	AutoCorr_M	0.0297	LAP_M	1.2535	KurtWav_M	0.8500
LAP_Y	0.0458	KurtWav_Y	0.0265	LAP_Y	1.2540	JND_Y	0.8542
FTM_Y	0.0324	JND_M	0.0138	FTM_Y	1.2547	JND_M	0.8542
KurtWav_Y	0.0290	JND_Y	0.0110	KurtWav_Y	1.2548	LAP_Y	0.8542
JND_Y	0.0200	RTBM_Y	−0.0095	JND_Y	1.2551	FTM_M	0.8542
JND_M	0.0198	VAR_Y	−0.0221	JND_M	1.2551	KurtZhang_M	0.8542
KurtZhang_Y	0.0194	VAR_M	−0.0354	KurtZhang_Y	1.2551	AutoCorr_Y	0.8542
FTM_M	0.0082	FTM_Y	−0.0396	FTM_M	1.2553	AutoCorr_M	0.8542
VAR_M	0.0068	RTBM_M	−0.0421	VAR_M	1.2553	GRAD_Y	0.8625
VAR_Y	0.0067	FTM_M	−0.0748	VAR_Y	1.2553	HPM_M	0.8625

Fig. 4 Accuracy, consistency, and monotonicity indexes for each objective metric computed in the linear space

PLCC		SROCC		RMSE		OR	
VIFP_Y	0.9230	VIFP_Y	0.9200	VIFP_Y	0.4832	VIFP_Y	0.4833
VIF_Y	0.9185	VIF_Y	0.9174	VIF_Y	0.4974	IFC_Y	0.5500
IFC_Y	0.9051	IFC_Y	0.9112	IFC_Y	0.5355	VIF_Y	0.5583
MSSSIM_Y	0.8971	MSSSIM_Y	0.9091	MSSSIM_Y	0.5560	UQI_Y	0.5917
IFC_M	0.8928	IFC_M	0.9037	IFC_M	0.5672	SSIM_Y	0.6125
SSIM_Y	0.8900	SSIM_Y	0.8952	SSIM_Y	0.5727	PSNR_Y	0.6208
UQI_Y	0.8780	FSIM_Y	0.8817	UQI_Y	0.6009	IFC_M	0.6250
FSIM_Y	0.8553	UQI_Y	0.8603	FSIM_Y	0.6516	MSSSIM_Y	0.6292
WSNR_Y	0.8404	UQI_M	0.8441	WSNR_Y	0.6803	MSE_Y	0.6375
UQI_M	0.8373	WSNR_Y	0.8416	UQI_M	0.6870	MSVD_Y	0.6500
PSNR_Y	0.8348	MSE_Y	0.8399	PSNR_Y	0.6911	IWPSNR_Y	0.6542
MSVD_Y	0.8316	PSNR_Y	0.8399	MSVD_Y	0.6975	SNR_Y	0.6542
MSE_Y	0.8272	MSVD_Y	0.8370	MSE_Y	0.7055	IWSSIM_Y	0.6542
SNR_Y	0.8269	SNR_Y	0.8333	SNR_Y	0.7060	UQI_M	0.6583
IWPSNR_Y	0.8160	IWPSNR_Y	0.8165	IWPSNR_Y	0.7256	WSNR_Y	0.6625
IWSSIM_Y	0.8160	IWSSIM_Y	0.8165	IWSSIM_Y	0.7256	DN	0.6625
VIF_M	0.8079	IWMSE_Y	0.8165	VIF_M	0.7401	FSIM_Y	0.6708
VIFP_M	0.7986	VIF_M	0.8152	VIFP_M	0.7556	MSE_M	0.6958
IWMSE_Y	0.7912	VIFP_M	0.8082	IWMSE_Y	0.7680	VIF_M	0.7000
DN	0.7877	DN	0.7993	DN	0.7737	IWPSNR_M	0.7083
MSSSIM_M	0.7482	FSIM_M	0.7603	MSSSIM_M	0.8330	IWSSIM_M	0.7083
FSIM_M	0.7363	MSSSIM_M	0.7584	FSIM_M	0.8498	VIFP_M	0.7125
WSNR_M	0.7252	WSNR_M	0.7324	WSNR_M	0.8644	MSSSIM_M	0.7125
QILV_Y	0.6918	QILV_Y	0.6913	QILV_Y	0.9088	WSNR_M	0.7208
SSIM_M	0.6855	SSIM_M	0.6847	SSIM_M	0.9139	PSNR_M	0.7250
MSE_M	0.6785	MSVD_M	0.6772	MSE_M	0.9222	SSIM_M	0.7250
MSVD_M	0.6779	IWPSNR_M	0.6580	MSVD_M	0.9229	IWMSE_Y	0.7333
IWPSNR_M	0.6646	IWSSIM_M	0.6580	IWPSNR_M	0.9380	SNR_M	0.7417
IWSSIM_M	0.6646	PSNR_M	0.6409	IWSSIM_M	0.9380	IWMSE_M	0.7458
SNR_M	0.6415	SNR_M	0.6394	SNR_M	0.9630	MSVD_M	0.7583
PSNR_M	0.6412	MSE_M	0.6360	PSNR_M	0.9633	QILV_Y	0.7583
IWMSE_M	0.6162	IWMSE_M	0.5931	IWMSE_M	0.9890	FSIM_M	0.8125
HPM_Y	0.4900	QILV_M	0.5265	HPM_Y	1.0944	KurtWav_M	0.8333
Marziliano_Y	0.4855	HPM_Y	0.4874	Marziliano_Y	1.0975	HPM_Y	0.8417
Marziliano_M	0.4059	Marziliano_Y	0.4303	Marziliano_M	1.1473	AutoCorr_Y	0.8458
GRAD_Y	0.3736	HPM_M	0.3518	GRAD_Y	1.1645	JND_St_Y	0.8542
GRAD_M	0.2844	Marziliano_M	0.3056	GRAD_M	1.2035	JND_St_M	0.8542
JND_St_M	0.2591	GRAD_Y	0.2570	JND_St_M	1.2125	JND_Y	0.8542
LAP_Y	0.2153	GRAD_M	0.1915	LAP_Y	1.2270	JND_M	0.8542
VAR_M	0.1654	LAP_M	0.1736	VAR_M	1.2381	VAR_Y	0.8542
QILV_M	0.1427	LAP_Y	0.1642	KurtWav_M	1.2425	LAP_M	0.8542
KurtWav_M	0.1425	VAR_M	0.1548	VAR_Y	1.2426	FTM_Y	0.8542
VAR_Y	0.1423	FTM_M	0.1314	QILV_M	1.2427	KurtZhang_M	0.8542
KurtZhang_Y	0.1053	KurtZhang_Y	0.1268	KurtZhang_Y	1.2484	KurtWav_Y	0.8542
FTM_M	0.0736	VAR_Y	0.1226	FTM_M	1.2520	AutoCorr_M	0.8542
HPM_M	0.0599	KurtWav_M	0.0961	HPM_M	1.2531	RTBM_Y	0.8542
AutoCorr_M	0.0560	AutoCorr_Y	0.0754	AutoCorr_M	1.2534	RTBM_M	0.8542
KurtZhang_M	0.0429	JND_St_M	0.0752	KurtZhang_M	1.2542	QILV_M	0.8583
JND_M	0.0402	KurtZhang_M	0.0726	JND_M	1.2543	FTM_M	0.8583
JND_Y	0.0401	KurtWav_Y	0.0543	JND_Y	1.2543	HPM_M	0.8583
AutoCorr_Y	0.0379	JND_M	0.0516	AutoCorr_Y	1.2545	Marziliano_Y	0.8583
RTBM_M	0.0315	JND_Y	0.0498	RTBM_M	1.2547	KurtZhang_Y	0.8583
KurtWav_Y	0.0312	RTBM_M	0.0429	KurtWav_Y	1.2547	LAP_Y	0.8625
LAP_M	0.0306	FTM_Y	0.0230	LAP_M	1.2548	GRAD_Y	0.8625
FTM_Y	0.0227	AutoCorr_M	−0.0019	FTM_Y	1.2551	VAR_M	0.8667
RTBM_Y	0.0173	JND_St_Y	−0.0482	RTBM_Y	1.2553	Marziliano_M	0.8750
JND_St_Y	0.0038	RTBM_Y	−0.0494	JND_St_Y	1.2553	GRAD_M	0.8792

Fig. 5 Accuracy, consistency, and monotonicity indexes for each objective metric computed in the logarithm space

according to PLCC, there is no statistical evidence to show performance differences between IFC and FSIM computed on the luminance channel, but they are statistically different from HDR-VDP-2 (see Fig. 4).

3.1 Best performing metrics

As expected, HDR-VDP-2 and HDR-VQM, which are the only true HDR quality metrics considered in this study, computed on absolute luminance values, are the

PLCC		SROCC		RMSE		OR	
MSSSIM_Y	0.9447	MSSSIM_Y	0.9501	MSSSIM_Y	0.4132	VIF_Y	0.4833
FSIM_Y	0.9376	FSIM_Y	0.9470	FSIM_Y	0.4377	IWPSNR_Y	0.5167
VIF_Y	0.9291	VIF_Y	0.9276	VIF_Y	0.4649	IWSSIM_Y	0.5167
VIFP_Y	0.9288	VIFP_Y	0.9228	VIFP_Y	0.4656	VIFP_Y	0.5208
IWPSNR_Y	0.9130	IFC_Y	0.9170	IWPSNR_Y	0.5121	IFC_Y	0.5375
IWSSIM_Y	0.9130	IWMSE_Y	0.9109	IWSSIM_Y	0.5121	MSSSIM_Y	0.5417
IFC_Y	0.9110	IWPSNR_Y	0.9109	IFC_Y	0.5196	WSNR_Y	0.5583
DN	0.9078	IWSSIM_Y	0.9109	DN	0.5275	FSIM_Y	0.5625
SSIM_Y	0.9060	DN	0.9090	SSIM_Y	0.5316	SSIM_Y	0.5792
WSNR_Y	0.8959	SSIM_Y	0.9072	WSNR_Y	0.5577	UQI_Y	0.5833
IFC_M	0.8928	IFC_M	0.9043	IFC_M	0.5670	DN	0.5875
IWMSE_Y	0.8841	WSNR_Y	0.8950	IWMSE_Y	0.5878	PSNR_Y	0.5917
UQI_Y	0.8777	MSVD_Y	0.8638	UQI_Y	0.6016	SNR_Y	0.5958
MSVD_Y	0.8612	UQI_Y	0.8610	MSVD_Y	0.6392	IWMSE_Y	0.6250
PSNR_Y	0.8526	MSE_Y	0.8564	PSNR_Y	0.6562	IFC_M	0.6375
SNR_Y	0.8472	PSNR_Y	0.8564	SNR_Y	0.6669	MSVD_Y	0.6375
MSE_Y	0.8352	SNR_Y	0.8556	MSE_Y	0.6915	VIF_M	0.6625
FSIM_M	0.8310	FSIM_M	0.8484	FSIM_M	0.6991	VIFP_M	0.6667
UQI_M	0.8278	MSSSIM_M	0.8442	UQI_M	0.7049	UQI_M	0.6833
VIF_M	0.8275	VIF_M	0.8373	VIF_M	0.7053	IWPSNR_M	0.6917
MSSSIM_M	0.8273	UQI_M	0.8335	MSSSIM_M	0.7059	WSNR_M	0.6917
VIFP_M	0.8242	VIFP_M	0.8327	VIFP_M	0.7109	IWSSIM_M	0.6917
WSNR_M	0.8163	WSNR_M	0.8233	WSNR_M	0.7254	MSE_Y	0.7000
IWPSNR_M	0.7848	QILV_Y	0.8047	IWPSNR_M	0.7781	QILV_M	0.7042
IWSSIM_M	0.7848	IWPSNR_M	0.7937	IWSSIM_M	0.7781	SSIM_M	0.7083
MSVD_M	0.7837	IWSSIM_M	0.7937	MSVD_M	0.7804	MSSSIM_M	0.7083
QILV_Y	0.7779	MSVD_M	0.7862	QILV_Y	0.7922	IWMSE_M	0.7458
IWMSE_M	0.7414	IWMSE_M	0.7700	IWMSE_M	0.8426	SNR_M	0.7458
MSE_M	0.7280	MSE_M	0.7444	MSE_M	0.8651	FSIM_M	0.7458
SSIM_M	0.7194	SSIM_M	0.7324	SSIM_M	0.8719	PSNR_M	0.7542
SNR_M	0.7085	SNR_M	0.7147	SNR_M	0.8859	MSVD_M	0.7542
PSNR_M	0.7033	PSNR_M	0.7088	PSNR_M	0.8925	QILV_Y	0.7542
QILV_M	0.6789	QILV_M	0.6739	QILV_M	0.9218	MSE_M	0.7708
Marziliano_Y	0.5114	HPM_Y	0.4442	Marziliano_Y	1.0788	HPM_Y	0.8375
HPM_Y	0.4548	Marziliano_Y	0.4179	HPM_Y	1.1181	Marziliano_M	0.8375
Marziliano_M	0.4217	HPM_M	0.3679	Marziliano_M	1.1383	JND_St_Y	0.8417
HPM_M	0.4004	Marziliano_M	0.3378	HPM_M	1.1503	LAP_Y	0.8417
JND_St_Y	0.2975	GRAD_Y	0.2040	JND_St_Y	1.1985	AutoCorr_M	0.8458
LAP_Y	0.1824	GRAD_M	0.1869	LAP_Y	1.2343	RTBM_M	0.8458
VAR_Y	0.1736	VAR_M	0.1387	VAR_Y	1.2363	VAR_Y	0.8500
GRAD_M	0.1618	VAR_Y	0.1258	GRAD_M	1.2389	GRAD_M	0.8500
GRAD_Y	0.1599	RTBM_Y	0.1223	GRAD_Y	1.2397	FTM_Y	0.8500
VAR_M	0.1031	KurtZhang_Y	0.1044	VAR_M	1.2487	FTM_M	0.8500
LAP_M	0.0948	LAP_M	0.0858	LAP_M	1.2497	AutoCorr_Y	0.8500
RTBM_Y	0.0946	RTBM_M	0.0744	RTBM_Y	1.2498	JND_St_M	0.8542
AutoCorr_Y	0.0860	LAP_Y	0.0713	AutoCorr_Y	1.2507	JND_Y	0.8542
KurtZhang_Y	0.0803	KurtWav_M	0.0634	KurtZhang_Y	1.2513	JND_M	0.8542
AutoCorr_M	0.0609	KurtWav_Y	0.0596	AutoCorr_M	1.2530	GRAD_Y	0.8542
RTBM_M	0.0577	FTM_Y	0.0578	RTBM_M	1.2533	HPM_M	0.8542
FTM_Y	0.0560	AutoCorr_Y	0.0518	FTM_Y	1.2534	KurtZhang_M	0.8542
JND_St_M	0.0545	JND_M	0.0515	JND_St_M	1.2538	KurtWav_Y	0.8542
KurtWav_M	0.0422	JND_Y	0.0499	JND_M	1.2545	KurtWav_M	0.8542
JND_M	0.0361	KurtZhang_M	0.0357	JND_Y	1.2545	RTBM_Y	0.8542
JND_Y	0.0360	AutoCorr_M	0.0356	KurtWav_Y	1.2552	LAP_M	0.8625
KurtWav_Y	0.0143	JND_St_M	0.0321	KurtZhang_M	1.2553	KurtZhang_Y	0.8625
KurtZhang_M	0.0093	JND_St_Y	0.0313	FTM_M	1.2553	VAR_M	0.8667
FTM_M	0.0090	FTM_M	−0.0193	KurtWav_M	1.2553	Marziliano_Y	0.8708

Fig. 6 Accuracy, consistency, and monotonicity indexes for each objective metric computed in the PU space

best performing metrics when compared to all other metrics and domains. Both metrics have a correlation above 0.95 and a particularly low RMSE (around 0.35) and low OR, whereas all other metrics have an OR above 0.48. HDR-VDP-2 (OR = 0.35) has a slightly lower OR than HDR-VQM (OR = 0.4083), but there is no statistical evidence to show a significant difference. However, HDR-VQM is over three times faster than

PLCC		SROCC		RMSE		OR	
MSSSIM_Y	0.9380	MSSSIM_Y	0.9435	MSSSIM_Y	0.4366	VIF_Y	0.4917
VIFP_Y	0.9301	FSIM_Y	0.9361	VIFP_Y	0.4613	VIFP_Y	0.4958
VIF_Y	0.9292	VIF_Y	0.9272	VIF_Y	0.4646	IFC_Y	0.5333
FSIM_Y	0.9240	VIFP_Y	0.9242	FSIM_Y	0.4812	IWPSNR_Y	0.5458
SSIM_Y	0.9107	IFC_Y	0.9151	SSIM_Y	0.5188	IWSSIM_Y	0.5458
IFC_Y	0.9093	SSIM_Y	0.9117	IFC_Y	0.5243	MSSSIM_Y	0.5542
IWPSNR_Y	0.9025	IFC_M	0.9039	IWPSNR_Y	0.5407	WSNR_Y	0.5750
IWSSIM_Y	0.9025	IWPSNR_Y	0.9024	IWSSIM_Y	0.5407	SNR_Y	0.5792
IFC_M	0.8930	IWSSIM_Y	0.9024	IFC_M	0.5667	SSIM_Y	0.5792
WSNR_Y	0.8893	IWMSE_Y	0.9022	WSNR_Y	0.5743	UQI_Y	0.5875
DN	0.8887	DN	0.8917	DN	0.5768	PSNR_Y	0.5917
UQI_Y	0.8767	WSNR_Y	0.8890	UQI_Y	0.6039	FSIM_Y	0.6042
IWMSE_Y	0.8730	MSE_Y	0.8656	IWMSE_Y	0.6132	IWMSE_Y	0.6250
MSVD_Y	0.8604	PSNR_Y	0.8656	PSNR_Y	0.6400	DN	0.6375
PSNR_Y	0.8603	MSVD_Y	0.8646	MSVD_Y	0.6409	MSVD_Y	0.6417
SNR_Y	0.8511	UQI_Y	0.8603	SNR_Y	0.6592	IFC_M	0.6500
MSE_Y	0.8451	SNR_Y	0.8589	MSE_Y	0.6721	VIF_M	0.6542
UQI_M	0.8285	MSSSIM_M	0.8359	UQI_M	0.7037	VIFP_M	0.6583
VIF_M	0.8282	VIF_M	0.8358	VIF_M	0.7039	WSNR_M	0.6833
VIFP_M	0.8224	UQI_M	0.8344	VIFP_M	0.7143	UQI_M	0.6875
MSSSIM_M	0.8181	FSIM_M	0.8315	MSSSIM_M	0.7224	QILV_Y	0.6917
FSIM_M	0.8129	VIFP_M	0.8302	FSIM_M	0.7318	QILV_M	0.6917
WSNR_M	0.8047	WSNR_M	0.8127	WSNR_M	0.7455	MSE_M	0.7083
QILV_Y	0.7744	QILV_Y	0.7964	QILV_Y	0.7946	IWPSNR_M	0.7083
MSVD_M	0.7671	IWPSNR_M	0.7734	MSVD_M	0.8058	SSIM_M	0.7083
IWPSNR_M	0.7653	IWSSIM_M	0.7734	IWPSNR_M	0.8081	IWSSIM_M	0.7083
IWSSIM_M	0.7653	MSVD_M	0.7690	IWSSIM_M	0.8081	MSE_Y	0.7167
SSIM_M	0.7243	IWMSE_M	0.7368	SSIM_M	0.8655	MSSSIM_M	0.7292
MSE_M	0.7219	MSE_M	0.7362	MSE_M	0.8688	PSNR_M	0.7417
IWMSE_M	0.7125	SSIM_M	0.7359	IWMSE_M	0.8810	SNR_M	0.7417
SNR_M	0.7041	SNR_M	0.7117	SNR_M	0.8914	IWMSE_M	0.7458
PSNR_M	0.7007	PSNR_M	0.7088	PSNR_M	0.8956	MSVD_M	0.7500
QILV_M	0.6601	QILV_M	0.6411	QILV_M	0.9430	FSIM_M	0.7708
Marziliano_Y	0.5065	HPM_Y	0.4685	Marziliano_Y	1.0824	LAP_Y	0.8375
HPM_Y	0.4717	Marziliano_Y	0.4199	HPM_Y	1.1069	HPM_Y	0.8375
Marziliano_M	0.4213	HPM_M	0.3486	Marziliano_M	1.1385	Marziliano_M	0.8417
HPM_M	0.4108	Marziliano_M	0.3267	HPM_M	1.1445	AutoCorr_M	0.8458
LAP_Y	0.1929	GRAD_Y	0.2241	LAP_Y	1.2318	RTBM_M	0.8458
VAR_M	0.1797	GRAD_M	0.1917	VAR_M	1.2349	JND_St_Y	0.8500
GRAD_Y	0.1733	VAR_M	0.1483	GRAD_Y	1.2368	FTM_Y	0.8500
JND_St_Y	0.1603	VAR_Y	0.1273	GRAD_M	1.2403	HPM_M	0.8500
GRAD_M	0.1556	RTBM_Y	0.1177	JND_St_Y	1.2411	JND_St_M	0.8542
VAR_Y	0.1048	KurtZhang_Y	0.1062	VAR_Y	1.2485	JND_Y	0.8542
LAP_M	0.0978	LAP_M	0.1004	LAP_M	1.2493	JND_M	0.8542
KurtZhang_Y	0.0867	LAP_Y	0.0846	KurtZhang_Y	1.2506	GRAD_Y	0.8542
AutoCorr_Y	0.0724	KurtWav_M	0.0780	AutoCorr_Y	1.2521	GRAD_M	0.8542
RTBM_Y	0.0651	FTM_Y	0.0642	RTBM_Y	1.2527	FTM_M	0.8542
KurtWav_M	0.0587	RTBM_M	0.0639	KurtWav_M	1.2532	KurtZhang_M	0.8542
JND_Y	0.0373	AutoCorr_Y	0.0578	JND_Y	1.2545	KurtWav_Y	0.8542
FTM_Y	0.0372	KurtWav_Y	0.0546	FTM_Y	1.2545	KurtWav_M	0.8542
JND_M	0.0362	JND_M	0.0511	JND_M	1.2545	AutoCorr_Y	0.8542
AutoCorr_M	0.0333	JND_Y	0.0506	AutoCorr_M	1.2547	RTBM_Y	0.8542
RTBM_M	0.0324	KurtZhang_M	0.0358	RTBM_M	1.2547	VAR_M	0.8583
KurtWav_Y	0.0166	JND_St_Y	0.0310	KurtWav_Y	1.2552	Marziliano_Y	0.8583
JND_St_M	0.0128	AutoCorr_M	0.0264	JND_St_M	1.2552	KurtZhang_Y	0.8625
KurtZhang_M	0.0119	JND_St_M	−0.0238	KurtZhang_M	1.2553	LAP_M	0.8667
FTM_M	−0.0086	FTM_M	−0.0521	FTM_M	1.2553	VAR_Y	0.8708

Fig. 7 Accuracy, consistency, and monotonicity indexes for each objective metric computed in the PQ space

HDR-VDP-2 [15], which makes it a suitable alternative to HDR-VDP-2.

The results for HDR-VDP-2 are in line with the finding of [26], slightly better than that of Valenzise et al. [24], but in contradiction with Mantel et al. [25], who reported a much lower correlation. However, Mantel et al. used unusual combinations of parameters for the base and extension layers, especially for content *BloomingGorse*.

Narwaria et al. [15] found that HDR-VQM was performing significantly better than HDR-VDP-2 for both video and still image content. However, our results show that both metrics have similar performance, while it was reported in [9] that HDR-VQM performs lower than HDR-VDP-2 for HDR video compression. The divergence between these findings might be due to the contents and types of artifacts considered in the different studies.

In contrast to the HDR metrics, the NR metrics show the worst performance with PLCC and SROCC values below 0.5 and RMSE and OR values above 1 and 0.8, respectively, independently of the domain in which the metric was computed. These results show that NR metrics are not sufficient to reach satisfactory prediction accuracy considering a perceptual domain and that specific NR metrics should be designed for HDR image quality assessment.

3.2 Difference measures and statistical-oriented metrics

Results show that MSE-based metrics, i.e., MSE, SNR, and PSNR, are not very reliable predictors of perceived quality when computed in the linear domain, with correlation between 0.65 and 0.75. Higher PLCC values were reported in [26] for MSE and SNR (PLCC = 0.88), but the study was performed considering only five contents. These metrics are know to be very content dependent [65], which might explain the drop in performance when considering 20 images. The correlation of MSE-based metrics computed on the luminance channel alone can be improved by about 0.1 by considering a more perceptual domain than the linear domain, which does not take into account the contrast sensitivity response of the HVS. In the log and PU domains, the correlation is about 0.83 and 0.84, respectively, which is in line with the results from [24]. Nevertheless, the performance of the MSE-based metrics computed as the average quality score of the Y, C_b, and C_r channels did not improve when considering perceptual domains. These observations indicate that the log, PU, and PQ domains can better represent the luminance sensitivity of the HVS than the linear domain, but they might not be optimal for the chrominance sensitivity.

3.3 Objective color difference measures

In the linear domain, the color difference metrics, with the exception of the original CIE1976 color difference metric, are the best performing pixel-based metrics. They outperform the MSE-based metrics, but there is no statistical evidence to show a significant improvement over SNR computed on the luminance alone. Nevertheless, their correlation with perceived visual quality is only about 80 %, with an OR above 69 %, which cannot be considered as reliable prediction. Since the release of the CIE1976 color difference metric, two extensions have been developed in 1994 and 2000 to better address perceptual non-uniformities of the HVS. But, according to the benchmarking results, further improvements might be necessary for HDR images to handle non-uniformities in low and high luminance ranges, outside of the typical range of LDR displays. The color difference metrics are computed in the CIELAB color space, which considers relative luminance values with respect to a reference white point, typically a reflective D65 white point about 100–120 cd/m^2. This reference white point is similar to the targeted peak luminance that is typically considered when calibrating LDR reference monitors. Therefore, for HDR images, one would be tempted to set the luminance of the reference white point considered in the color conversion equal to the peak luminance of the HDR monitor. However, this leads to lower performance of the color difference metrics and the reflective white point should also be used for HDR content instead.

3.4 Structural similarity and visual information measures

The performance of SSIM and its multiscale extension, MS-SSIM, is improved by considering logarithm instead of linear values and is even further improved by considering the PU or PQ transform. In particular, on the luminance channel, the correlation of SSIM is increased by about 0.15 from linear to logarithm, while MS-SSIM improved by only about 0.03. From log to PU/PQ, improvements are relatively low for SSIM, whereas MS-SSIM exhibits a gain of about 0.04. Results show that MS-SSIM (luminance only) performs the best in PU and PQ spaces according to the PLCC, SROCC,

PLCC		SROCC		RMSE		OR	
HDRVDP2	0.9604	HDRVDP2	0.9564	HDRVDP2	0.3498	HDRVDP2	0.3500
HDRVQM	0.9602	HDRVQM	0.9564	HDRVQM	0.3506	HDRVQM	0.4083
PU2MSSSIM_Y	0.9447	PU2MSSSIM_Y	0.9501	PU2MSSSIM_Y	0.4132	LOG_VIFP_Y	0.4833
PQ2MSSSIM_Y	0.9380	PQ2MSSSIM_Y	0.9435	PQ2MSSSIM_Y	0.4366	PU2MSSSIM_Y	0.5417
LOG_VIFP_Y	0.9230	LOG_VIFP_Y	0.9200	LOG_VIFP_Y	0.4832	PQ2MSSSIM_Y	0.5542
LOG_PSNR_Y	0.8348	LOG_PSNR_Y	0.8399	LOG_PSNR_Y	0.6911	LOG_PSNR_Y	0.6208

Fig. 8 Statistical analysis comparing the HDR metrics and best performing metric of each domain

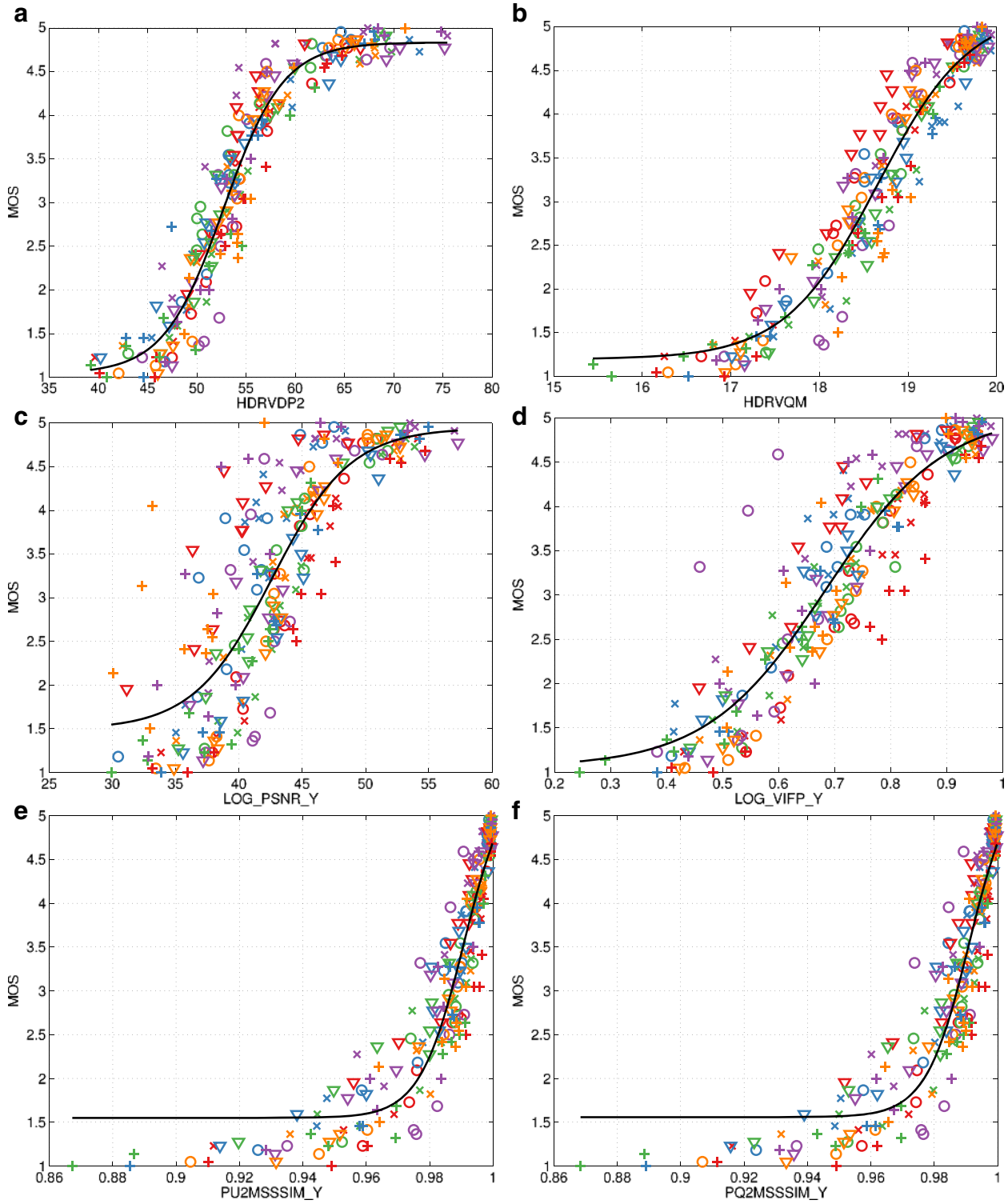

Fig. 9 Subjective versus objective results. Subjective versus objective evaluation results for the HDR metrics (**a**-**b**) and best performing metric of each domain (**c**-**f**). Each symbol, i.e., combination of marker and color, corresponds to a specific content

and RMSE indexes. The correlation obtained for SSIM in the log and PU domains is similar to the results of Valenzise et al. [24]. On the other hand, UQI, which corresponds to the special case of SSIM when the constants C_1 and C_2 are set to 0, does not perform better in the log, PU, or PQ space than in the linear domain. Similar correlation results for SSIM and MS-SSIM are reported in [26] as in this paper (for the linear domain). However, it is reported that the relative change between the worst and best qualities for SSIM and MS-SSIM was less than 0.003 and 0.0003 %, respectively. In this study, the average relative change computed over all domains is 16.5 and 11.5 % for SSIM and MS-SSIM, respectively. One major difference between the two works is the use of absolute luminance values in [26], whereas luminance values were linearly mapped from the theoretical display range to the range $[0, 1]$ in this paper. For LDR content, SSIM uses different values for C_1 and C_2 depending on whether the images are in the range $[0, 1]$ or $[0, 255]$. For HDR content, our findings suggest that the value of these constants should be adjusted according the luminance range and depending on whether scaling of the values is performed or not.

Metrics that quantify the loss of image information, i.e., VIF, its pixel-based version, VIFP, and its predecessor, IFC, also show good performance. In particular, IFC (luminance only) is the second best performing metric in the linear domain. While the performance of IFC is not influenced by the domain in which the metric is computed, the performance of VIF(P) is significantly improved when considering a more perceptual domain than the linear space. In the log domain, results show that VIF computed on the luminance alone is the best performing metric. Note that the correlation reported for VIF(P) in this paper is significantly better than the one reported in [26]. Similarly to (MS-)SSIM, the difference might be due to the scaling procedure. Among the other HVS-based metrics, FSIM also shows good performance, especially in the PU and PQ space (RMSE below 0.5). In the linear domain, results are similar to our previous work.

3.5 Statistical analyses

To determine how the best metrics of each domain compare to each other, a direct benchmarking of the two HDR metrics, which are the best performing metrics in the linear space, and the best performing metric of the log, PU, and PQ spaces was performed. The PSNR metric computed on the luminance channel in the log space was added to this comparison, as this metric is widely used in HDR compression studies. Figure 8 reports the results of the statistical analysis of the six metrics. To identify metrics computed in the log, PU, and PQ spaces, the *LOG_*, *PU2*, and *PQ2* prefixes are used, respectively. According to PLCC and SROCC, there is no statistical evidence to show performance differences between HDR-VDP-2, HDR-VQM, and MS-SSIM computed on the luminance channel in the PU space. However, HDR-VDP-2 and HDR-VQM have a significantly lower RMSE than all other metrics. Figure 9 depicts the scatter plots of subjective versus objective results for these metrics. As it can be observed, the data points are well concentrated near the fitting curve for HDR-VDP-2, as well as for HDR-VQM, while they are more scattered for the other metrics, especially in the case of LOG_PSNR_Y, which shows higher content dependency. These findings indicate that HDR-VDP-2 and HDR-VQM have a very high consistency when compared to the other metrics. Nevertheless, HDR-VDP-2 is complex and requires heavy computational resources, which limits its use in many applications. HDR-VQM and MS-SSIM computed in the PU space are lower complexity alternatives to HDR-VDP-2.

The statistical analysis was also used to understand whether there is a statistically significant difference between the performance of each metric when computed on the luminance component alone and when computed on all components. Only results from the analysis performed on the 28 metrics that were computed both on the Y channel alone and as the average quality score of the Y, C_b, and C_r channels were considered. Table 2 reports the number of metrics for which one approach was significantly better than the other one, as well as when no significant difference between the two approaches was observed. The analysis was performed individually for each performance index and domain. In the linear domain, there is no statistical evidence to show performance differences between the two approaches for about 80 % of the metrics. However, in the log, PU, and PQ space, roughly half of the metrics perform significantly better

Table 2 Comparison of the 28 metrics computed on the Y and YC_bC_r channels. Comparison of the metrics computed as the average quality score of the Y channel alone and as the average quality score of the YC_bC_r channels

	lin				log				PU				PQ			
	PLCC	SROCC	RMSE	OR	PLCC	SROCC	RMSE	OR	PLCC	SROCC	RMSE	OR	PLCC	SROCC	RMSE	OR
Y is better	6	7	8	3	16	14	16	8	16	14	15	14	14	14	15	14
Similar	22	21	20	25	11	14	12	20	12	14	13	14	14	14	13	14
YC_bC_r is better	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Table 3 Comparison of the 57 metrics computed on all domains. Results represent the number of times a metric computed in the domain *i* performs significantly better than when computed in the domain *j*, where *i* and *j* are the row and column of the table

	PLCC				SROCC				RMSE				OR			
	lin	log	PU	PQ	lin	log	PU	PQ	lin	log	PU	PQ	lin	log	PU	PQ
lin	0	9	5	6	0	9	6	8	0	8	5	5	0	2	1	1
log	10	0	2	2	10	0	4	4	8	0	0	0	8	0	3	3
PU	14	17	0	11	13	15	0	11	11	16	0	10	9	7	0	7
PQ	13	17	8	0	11	15	8	0	11	16	6	0	9	7	3	0

when computed on the luminance channel alone. According to PLCC, the JND metric, FR version, computed in the log domain, is the only case for which better performance is achieved when considering all channels. As HDR is often considered in combination with wide color gamut (WCG), it is expected that the fidelity of color reproduction will play a more important role in the context of HDR when compared to LDR. We believe that improvements can be achieved by considering different domains for computing the metrics on the chrominance channels and by using better pooling strategies.

Similarly, the statistical analysis was also used to understand whether there is a statistically significant difference between the performance of a particular metric computed in one domain and another domain. Only results from the analysis performed on the 57 metrics that were computed in all domains were considered. Table 3 reports the number of times a metric computed in the domain *i* performs significantly better than when computed in the domain *j*, where *i* and *j* are the row and column of the table. Results show that most metrics perform the best in the PU and PQ spaces when compared to the lin and log spaces, which is in line with our previous observations. Note that results based on PLCC, SROCC, and RMSE are in agreement, while the OR metric shows fewer cases where statistically significant differences are observed. Additionally, there are also metrics for which computations performed in the linear and logarithm domains perform better than in the PU and PQ space. Overall, there is no optimal domain that performs the best for all metrics. Instead, different metrics should use different domains to maximize the correlation with perceived quality.

4 Conclusions

In this paper, 35 objective metrics were benchmarked on a database of 240 compressed HDR images using subjective quality scores as ground truth. Additionally to the linear space, metrics were computed in the logarithm, PU, and PQ domains to mimic non-linearities of the HVS. Results showed that the performance of most full-reference metrics can be improved by considering perceptual transforms when compared to linear values. On the other hand, our findings suggested that a lot of work remains to be done for no-reference quality assessment of HDR content. Our benchmark demonstrated that HDR-VDP-2 and HDR-VQM are ultimately the most reliable predictors of perceived quality. Nevertheless, HDR-VDP-2 is complex and requires heavy computational resources, which limits its use in many applications. HDR-VQM is over three times faster, which makes it a suitable alternative to HDR-VDP-2. Alternatively, MS-SSIM computed in the PU space is another lower complexity substitute, as there is no statistical evidence to show performance differences between these metrics in terms of PLCC and SROCC. Even though the numbers of contents and compressed images considered in the experiments are quite large, different performance might be observed for other contents and types of artifacts.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1), Swiss SERI project Compression and Evaluation of High Dynamic Range Image and Video (C12.0081), Portuguese Instituto de Telecomunicações - "FCT – Fundação para a Ciência e a Tecnologia" (project UID/EEA/50008/2013), and COST IC1005 The digital capture, storage, transmission, and display of real-world lighting HDRi.

Author details
[1]Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland. [2]Optics Center, UBI, Covilhã, Portugal. [3]Instituto de Telecomunicações, UBI, Covilhã, Portugal.

Received: 4 May 2015 Accepted: 3 November 2015
Published online: 02 December 2015

References

1. P Hanhart, P Korshunov, T Ebrahimi, in *SPIE Applications of Digital Image Processing XXXVII*. Subjective evaluation of higher dynamic range video, (2014)
2. P Korshunov, H Nemoto, A Skodras, T Ebrahimi, in *Proc. SPIE 9138, Optics, Photonics, and Digital Technologies for Multimedia Applications III*. Crowdsourcing-based Evaluation of Privacy in HDR Images, (2014)
3. E Reinhard, G Ward, S Pattanaik, P Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005)
4. H Seetzen, W Heidrich, W Stuerzlinger, G Ward, L Whitehead, M Trentacoste, A Ghosh, A Vorozcovs, High Dynamic Range Display Systems. ACM Trans. Graph. **23**(3), 760–768 (2004)
5. E Reinhard, M Stark, P Shirley, J Ferwerda, Photographic tone reproduction for digital images. ACM Trans. Graph. **21**(3), 267–276 (2002)

6. T Richter, in *Picture Coding Symposium*. On the standardization of the JPEG XT image compression, (2013)
7. G Ward, M Simmons, in *ACM SIGGRAPH 2006 Courses*. JPEG-HDR: a backwards-compatible, high dynamic range extension to JPEG, (2006)
8. T Richter, in *SPIE Applications Of Digital Image Processing XXXVII*. On the integer coding profile of JPEG XT, vol. 9217, (2014)
9. P Hanhart, M Rerabek, T Ebrahimi, in *Proc. SPIE, Applications of Digital Image Processing XXXVIII*. Towards high dynamic range extensions of HEVC: subjective evaluation of potential coding technologies, (2015)
10. R Mantiuk, S Daly, K Myszkowski, H-P Seidel, in *SPIE Human Vision and Electronic Imaging X,* vol. 5666. Predicting visible differences in high dynamic range images: model and its calibration, (2005)
11. SJ Daly, in *SPIE Human Vision, Visual Processing, and Digital Display III,* vol. 1666. Visible differences predictor: an algorithm for the assessment of image fidelity, (1992)
12. R Mantiuk, KJ Kim, AG Rempel, W Heidrich, HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graph. **30**(4), 40:1–40:14 (2011)
13. M Narwaria, RK Mantiuk, M Perreira Da Silva, P Le Callet, HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. J. Electron. Imaging. **24**(1), 010501 (2015)
14. TO Aydin, R Mantiuk, K Myszkowski, H-P Seidel, Dynamic Range Independent Image Quality Assessment. ACM Trans. Graph. **27**(3), 69:1–69:10 (2008)
15. M Narwaria, M Perreira Da Silva, P Le Callet, HDR-VQM: An objective quality measure for high dynamic range video. Signal Process. Image Commun. **35**, 46–60 (2015)
16. C Poynton, *Digital Video and HD: Algorithms and Interfaces*. (Elsevier/Morgan Kaufmann, Burlington, Vermont, USA, 2012)
17. SK Shevell, *The Science of Color*. (Elsevier, Boston, Massachusetts, USA, 2003)
18. A Rose, The sensitivity performance of the human eye on an absolute scale. J. Opt. Soc. Am. A. **38**(2), 196–208 (1948)
19. H De Vries, The quantum character of light and its bearing upon threshold of vision, the differential sensitivity and visual acuity of the eye. Physica. **10**(7), 553–564 (1943)
20. TO Aydın, R Mantiuk, K Myszkowski, H-P Seidel, in *SPIE Human Vision and Electronic Imaging XIII,* vol. 6806. Extending quality metrics to full luminance range images, (2008)
21. J Munkberg, P Clarberg, J Hasselgren, T Akenine-Möller, High Dynamic Range Texture Compression for Graphics Hardware. ACM Trans. Graph. **25**(3), 698–706 (2006)
22. HR Sheikh, MF Sabir, AC Bovik, A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. IEEE Trans. Image Process. **15**(11), 3440–3451 (2006)
23. K Seshadrinathan, R Soundararajan, AC Bovik, LK Cormack, Study of Subjective and Objective Quality Assessment of Video. IEEE Trans. Image Process. **19**(6), 1427–1441 (2010)
24. G Valenzise, F De Simone, P Lauga, F Dufaux, in *SPIE Applications of Digital Image Processing XXXVII,* vol. 9217. Performance evaluation of objective quality metrics for HDR image compression, (2014)
25. C Mantel, SC Ferchiu, S Forchhammer, in *16th International Workshop on Multimedia Signal Processing*. Comparing subjective and objective quality assessment of HDR images compressed with JPEG-XT, (2014)
26. P Hanhart, MV Bernardo, P Korshunov, M Pereira, AMG Pinheiro, T Ebrahimi, in *6th International Workshop on Quality of Multimedia Experience*. HDR image compression: A new challenge for objective quality metrics, (2014)
27. M Azimi, A Banitalebi-Dehkordi, Y Dong, MT Pourazad, P Nasiopoulos, in *International Conference on Multimedia Signal Processing*. Evaluating the Performance of Existing Full-Reference Quality Metrics on High Dynamic Range (HDR) Video Content, (2014)
28. M Rerabek, P Hanhart, P Korshunov, T Ebrahimi, in *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. Subjective and objective evaluation of HDR video compression, (2015)
29. P Korshunov, P Hanhart, T Richter, A Artusi, R Mantiuk, T Ebrahimi, in *7th International Workshop on Quality of Multimedia Experience*. Subjective quality assessment database of HDR images compressed with JPEG XT, (2015)
30. Subjective Quality Assessment Database of HDR Images Compressed with JPEG XT. http://mmspg.epfl.ch/jpegxt-hdr. Accessed 23 July 2015
31. S Miller, M Nezamabadi, S Daly, Perceptual signal coding for more efficient usage of bit codes. SMPTE Motion Imaging J. **122**(4), 52–59 (2013)
32. R Mantiuk, S Daly, L Kerofsky, Display adaptive tone mapping. ACM Trans. Graph. **27**(3), 68 (2008)
33. AO Akyüz, E Reinhard, Color appearance in high-dynamic-range imaging. SPIE J. Electron. Imaging. **15**(3) (2006)
34. 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 (2012)
35. ITU-R BT.500-13, Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union (2012)
36. V Laparra, JM noz-Marí, J Malo, Divisive normalization image quality metric revisited. J. Opt. Soc. Am. A. **27**(4), 852–864 (2010)
37. Z Wang, AC Bovik, A universal image quality index. IEEE Signal Process. Lett. **9**(3), 81–84 (2002)
38. Z Wang, AC Bovik, HR Sheikh, EP Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
39. Z Wang, EP Simoncelli, AC Bovik, in *37th Asilomar Conference on Signals, Systems and Computers*. Multiscale structural similarity for image quality assessment, (2003)
40. A Shnayderman, A Gusev, AM Eskicioglu, An SVD - based grayscale image quality measure for local and global assessment. IEEE Trans. Image Process. **15**(2), 422–429 (2006)
41. S Aja-Fernandéz, RSJ Estepar, C Alberola-Lopez, C-F Westin, in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Image Quality Assessment based on Local Variance, (2006)
42. HR Sheikh, AC Bovik, G de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. Image Process. **14**(12), 2117–2128 (2005)
43. HR Sheikh, AC Bovik, Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
44. L Zhang, D Zhang, X Mou, D Zhang, FSIM: A feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011)
45. Z Wang, Q Li, Information Content Weighting for Perceptual Image Quality Assessment. IEEE Trans. Image Process. **20**(5), 1185–1198 (2011)
46. XK Yang, WS Ling, ZK Lu, EP Ong, SS Yao, Just noticeable distortion model and its applications in video coding. Signal Process. Image Commun. **20**(7), 662–680 (2005)
47. J Mannos, DJ Sakrison, The effects of a visual fidelity criterion of the encoding of images. IEEE Trans. Inf. Theory. **20**(4), 525–536 (1974)
48. T Mitsa, KL Varkur, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms, (1993)
49. CIE, Colorimetry Official Recommendation of the International Commission on Illumination. CIE publication 15.2, CIE Central Bureau (1986)
50. CIE, Industrial Colour-Difference Evaluation. CIE publication 116, CIE Central Bureau (1995)
51. FJJ Clarke, R McDonald, B Rigg, Modification to the JPC79 Colour-difference Formula. J. Soc. Dye. Colour. **100**(4), 128–132 (1984)
52. M Luo, G Cui, B Rigg, The development of the CIE 2000 colour-difference formula: CIEDE2000. Color Res. Appl. **26**(5), 340–350 (2001)
53. SJ Erasmus, KCA Smith, An automatic focusing and astigmatism correction system for the SEM and CTEM. J. Microsc. **127**(2), 185–199 (1982)
54. CF Batten, *Autofocusing and astigmatism correction in the scanning electron microscope. Master's thesis*. (University of Cambridge, UK, 2000)
55. AV Murthy, LJ Karam, in *2nd International Workshop on Quality of Multimedia Experience*. A MATLAB-based framework for image and video quality evaluation, (2010)
56. D Shaked, I Tastl, in *IEEE International Conference on Image Processing*. Sharpness measure: towards automatic image enhancement, (2005)
57. N Zhang, A Vladar, M Postek, B Larrabee, in *Proceedings Section of Physical and Engineering Sciences of American Statistical Society*. A kurtosis-based statititcal measure for two-dimensional processes and its application to image sharpness, (2003)
58. R Ferzli, LJ Karam, J Caviedes, in *1st International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. A robust image sharpness metric based on kurtosis measurement of wavelet coefficients, (2005)

59. R Ferzli, LJ Karam, in *3rd International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. A no-reference objective sharpness metric using riemannian tensor, (2007)
60. H Wilson, A transducer function for threshold and suprathreshold human vision. Biol. Cybern. **38**(3), 171–178 (1980)
61. PG Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. (SPIE Optical Engineering Press, Bellingham, Washington, USA, 1999)
62. ITU-R BT.709, Parameter values for the HDTV standards for production and international programme exchange. International Telecommunication Union (2002)
63. P Hanhart, P Korshunov, T Ebrahimi, in *18th International Conference on Digital Signal Processing*. Benchmarking of quality metrics on ultra-high definition video sequences, (2013)
64. ITU-T P.1401, Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. International Telecommunication Union (2012)
65. Q Huynh-Thu, M Ghanbari, Scope of validity of PSNR in image/video quality assessment. Electron. Lett. **44**(13), 800–801 (2008)