



HDRC: a subjective quality assessment database for compressed high dynamic range image

Yue Liu¹ · Zhangkai Ni² · Peilin Chen¹ · Shiqi Wang¹ · Sam Kwong³

Received: 9 November 2023 / Accepted: 17 March 2024 / Published online: 6 May 2024
© The Author(s) 2024

Abstract

Evaluating the quality of high dynamic range (HDR) images has emerged as a challenging and contemporary topic with the proliferation of HDR content. In this work, a new HDR compression (HDRC) database is proposed, aiming to provide a benchmark for the development of full-reference HDR image quality assessment (IQA) algorithms when facing the latest HDR compression distortions. In particular, the proposed HDRC database is the first HDR-IQA database to incorporate Versatile Video Coding (VVC) compression distortions, closely associated with practical application scenarios. Furthermore, the proposed HDRC database is currently the largest HDR-IQA database, including 80 reference images and 400 distorted images. Extensive experiments are conducted by studying the performance compared to existing HDR-IQA databases when evaluating three HDR-specific IQA models and nine IQA models prevalent for low dynamic range (LDR) content, revealing the challenges the proposed HDRC database brings. The results indicate that the existing IQA models demonstrate noticeable decreases in accuracy when assessing new compression distortions, underscoring the need for the development of novel HDR-IQA models. Consequently, the suggested HDRC database can serve as a potential database for HDR-IQA research, fostering a comprehensive exploration of the associated fields.

Keywords Image quality assessment · High dynamic range · Subjective test

1 Introduction

In nature scenes, the human visual system (HVS) is capable of perceiving the luminance range from 5×10^{-3} nits to 10×8 nits [1], comprising a rich spectrum of color and brightness information. However, the low dynamic range (LDR) image captured by the traditional camera can only represent

a limited dynamic range due to hardware limitations. Consequently, perceptual disparities in terms of both color and brightness inevitably appear when humans watch the LDR content. This divergence becomes particularly pronounced in content captured in scenes with a wider luminance range. In recent years, driven by technological advancements, HDR imaging has emerged as a viable solution to mitigate these constraints. Compared with LDR images, HDR images offer an extended range of luminance levels, thus yielding more vibrant color reproduction, greater intricacy in visual detail, and heightened fidelity in photographic tonality. Due to their ability to significantly enhance user visual experiences, display devices like smartphones, televisions, and monitors compatible with HDR content are increasingly prevalent.

The proliferation of HDR content is accompanied by the ongoing development of compression algorithms tailored to HDR content. In the early stage, HDR images are first converted to LDR images using tone mapping algorithms. After compression and transmission via either JPEG or JPEG2000, an inverse tone mapping algorithm is applied to reconstruct the HDR images. This methodology not only heavily relies on precise tone mapping operations

✉ Sam Kwong
samkwong@ln.edu.hk

Yue Liu
yliu724-c@my.cityu.edu.hk

Zhangkai Ni
zkni@tongji.edu.cn

Peilin Chen
plchen3@cityu.edu.hk

¹ Department of Computer Science, City University of Hong Kong, Hong Kong, China

² Computer Science and Technology, Tongji University, Shanghai 200092, China

³ Lingnan University, Hong Kong, China

but also consumes significant computational resources during the conversion process. To address these issues, JPEG-XT is proposed [2], wherein the LDR version of the HDR image is compressed into an 8-bit base layer, while the residuals of the original HDR image and reconstructed HDR image are compressed in the 10-bit enhancement layer. For HDR video compression, the latest VVC [3] proposes the luma mapping with chroma scaling operation [4] and the luma-adaptive deblocking algorithm [5] for HDR content. Amid the continuous evolution of compression algorithms, the effectiveness of HDR image quality metrics holds paramount significance in the design of codecs for HDR content.

In contrast to the abundance of subjective quality assessment databases related to LDR content [6–8], there is a notable scarcity of subjective HDR-IQA databases. In particular, Narwaria et al. [9] constructed an HDR Images JPEG Compression database (HIJC), including 10 reference HDR images (RHIs) and 140 distorted HDR images (DHIs) compressed by JPEG with iCAM06 [10] as the tone-map operator (TMO). Valenzise et al. first proposed a database (HIFM) [11] to evaluate the performance of HDR image fidelity metrics, which contains 5 RHIs and 45 DHIs compressed by JPEG, JPEG2000, and JPEG-XT with Mai [12] as the TMO. They further expand the database (eHIFM) by applying the perceptual quantizer (PQ) [13] for tonemapping in [14]. To further expand the size of the subjective HDR quality database, Mikhailiuk et al. established a unified photometric image quality database (UPIQ) [15], combining two existing HDR-IQA databases [9, 16] by introducing a psychometric scaling strategy to align the mean opinion scores (MOS) of these two databases.

Existing HDR-IQA databases have shown the following limitations: (1) limited content diversity due to a restricted number of RHIs; (2) insufficient research on the latest HDR compression distortion types, particularly JPEG-XT and VVC; These inherent limitations in existing databases can impact the assessment of current HDR-IQA algorithms and affect the accuracy when developing new HDR-IQA algorithms. Motivated by the limitations above, an HDR image compression (HDRC) database is proposed in this paper to provide a more comprehensive and diverse collection of HDR images with compression distortions introduced by the latest codecs, which aims to facilitate the evaluation of state-of-the-art HDR-IQA metrics. Specifically, the HDRC database contains a total of 80 RHIs and 400 DHIs, where two types of compression distortion (i.e., JPEG-XT and VVC) are utilized in the degraded HDR image generations. The Double Stimulus Impairment Scale (DSIS) [17] methodology is employed to obtain the MOS value for each distorted HDR image in the subjective experiment. The contributions of this paper are summarized as follows:

- We perform an extensive subjective experiment to evaluate the perceptual quality of the HDR images compressed by JPEG-XT and VVC. To the best of our knowledge, this is the first database to study the impact of VVC compression on HDR images.
- We construct a new database, HDRC, with 80 reference HDR images and 400 distorted HDR images, which is the largest existing HDR-IQA database in terms of the number of both RHIs and DHIs.
- We evaluate the performance of three HDR-specific IQA algorithms and nine prevalent LDR metrics on the HDRC database and four existing databases, aiming to provide a comprehensive benchmark for the development of HDR-IQA models.

The remainder of this paper is organized as follows. Section 2 summarizes existing full-reference HDR-IQA metrics and the subjective databases. Section 3 elaborates on the construction details of the HDRC database and the process of the subjective experiment. Section 4 analyzes the performance of the state-of-the-art HDR-IQA metrics on the HDRC database. The conclusion and future work are discussed in Section 5.

2 Related work

2.1 Full-reference HDR-IQA metrics

With the ongoing development of research in the IQA task, the accuracy of both full-reference (FR) [18–20] and no-reference (NR) [21–23] IQA algorithms for LDR images has significantly improved. However, designing IQA models for HDR images that are highly relevant to the HVS remains a challenging task. In this paper, we focus on existing commonly used FR IQA metrics for HDR images, which can be classified into two categories: (1) HDR-specific IQA metrics and (2) perceptual-mapping-based LDR metrics.

3 HDR-specific IQA metrics

The predominant FR IQA metric for HDR images is HDR visual difference predictor (HDR-VDP) [24, 25] proposed by Mantiuk et al. where the HVS is modeled by the optical transfer function, just noticeable differences (JND), and contrast sensitive function (CSF). In HDR-VDP-2.2 [26], the metric is improved by applying new models for the CSF, sensitive function for cones and rods, contrast masking, and image decomposition method. Recently, the HDR-VDP-3 [27] is proposed to adapt to multiple quality-related tasks. The above works proposed by Mantiuk et al. are foundational to the HDR-IQA task, providing essential insights and

methodologies. Recently, a local-global frequency feature-based model (LGFM) [28] has been proposed to measure the quality of HDR images by integrating the local and global frequency features. For HDR videos, Narwaria et al. introduced an HDR video quality metric (HDR-VQM) [29], where the Gabor filter is applied to extract the quality of the HDR images in the perceptual domain.

4 Perceptual-mapping-based LDR metrics

The main differences between the LDR and HDR images lie in their encoding way for captured light and the luminance range they can provide on the display side. Precisely, HDR images can effectively capture a broader range of luminance values, which are linear with real-world luminance levels. In contrast, LDR images are constrained by storage limitations, thus accommodating a narrower luminance range and usually employing perceptual mapping techniques to maximize perceptual information. Therefore, perceptual mapping methods need to be adopted when applying the LDR metrics for HDR-IQA. One popular perceptual mapping method is Gamma correction, which has been well-studied in adjusting contrast for better LDR image visualization. The perceptually uniform (PU) encoding [30] proposed by Aydin et al. encodes the HDR image into the perceptual domain based on the CSF. Recently, Mantiuk and Azimi improved the PU encoder by modeling the coding artifacts to find the visibility thresholds, denoted as PU21 [31]. Moreover, the PQ function has been widely applied in image and video compression tasks and has become a basis of various HDR video compression standards.

5 Subjective HDR IQA databases

Innovative and effective databases [34, 35] play a crucial role in leading the trajectory of the corresponding research and serve as the foundation of plenty of robust analyses. For instance, the MSVD database first describes the emotions of the video and inspires the work [36] to incorporate the logical expression with emotions. The database proposed in this paper is also motivated by prior research and aims to provide a border and innovative perspective on HDR IQA. With the development of HDR imaging and compression technologies, investigating the effects of various distortion types on the HVS in HDR images holds significant importance. In this section, four publicly available HDR IQA databases are introduced in detail, where the summary of these databases is shown in Table 1.

6 HIJC proposed by Narwaria et al.

In the HIJC database, Narwaria et al. introduces an optimization strategy to improve the quality of compressed HDR images. In the proposed framework, the mean square error (MSE) or structural similarity index measure (SSIM) [37] is used as criteria to minimize the difference between μ -Law encoded HDR image and tone-mapped image. The HIJC database is constructed to evaluate the quality of decompressed HDR images. The database is composed of 10 RHIs, each with 7 distortion levels and 2 optimization metrics, resulting in a total of 140 DHIs. In the subjective experiment, 26 observers are invented to collect the subjective scores using the absolute category rating with hidden reference (ACR-HR) methodology [38]. From the experimental results, the HDR-VDP-2 performs better on individual content compared with the overall results of all content. Furthermore, the HDR-VDP-2 outperforms MSE and SSIM.

Table 1 Summary of HDR-IQA databases. The number of observers, distortion type, and subjective methodology are denoted as Obs., Dist., and Meth., respectively.

Database	#Obs.	#RHIs	#DHIs	Dist.	TMO	Meth.
HIJC [9]	26	10	140	JPEG	iCAM06	ACR-HR
HIFM [11]	15	5	50	JPEG	Mai	DSIS
				JPEG2000		
				JPEG-XT		
eHIFM [14]	15	10	100	JPEG	Mai, PQ	DSIS
				JPEG2000		
UPIQ [15] (HDR part)	20	30	380	JPEG	iCAM06	
				JPEG-XT	RG [32]	–
					MT [33]	
HDRC (Ours)	20	80	400	JPEG-XT	RG	DSIS
				VVC	PQ	

7 HIFM proposed by Valenzise et al.

HIFM proposed by Valenzise et al. is the first database constructed to evaluate the performance of the HDR IQA metrics. The database contains 5 RHIs and 50 DHIs. The subjective scores are obtained from 15 observers with DSIS methodology. The performance of HDR-VDP, peak signal-to-noise ratio (PSNR), and SSIM are evaluated, where the latter two metrics are adopted with PU encoding and logarithmic operation as perceptual mapping. From the results, the perceptual-mapped LDR metrics achieve similar performance compared with HDR-VDP.

8 eHIFM proposed by Zerman et al.

The eHIFM database is an extension version of the HIFM database, where another 5 RHIs are added, including 10 RHIs and 100 DHIs in total with the same methodology to get the subjective scores of the images. In the experiments, the performance of two HDR-specific metrics, two color difference metrics, and seven LDR metrics are extensively studied, where the LDR metrics are adopted with linear mapping, PU, and logarithmic mapping. Moreover, a PQ-based tPSNR-YUV is also been considered. From the results, PU mapping outperforms the other two mapping methods.

Additionally, SSIM and multiscale structure similarity (MSSIM) [39] perform better among the LDR metrics, while HDR-VDP-2.2 and HDR-VQM demonstrate superior performance compared with all other metrics.

9 UPIQ proposed by Mikhailiuk et al.

In the work proposed by Mikhailiuk et al., a psychometric scale strategy is proposed to combine the quality scores of different databases, including 3779 LDR images and 380 HDR images. Besides, a PU-encoded LDR model is retrained on the proposed database. The experiment takes PU-encoded full reference LDR metrics, no reference LDR metrics, and two HDR-specific metrics for comparison. The results show that the retrained model outperforms other

models, followed by the PU-encoded feature similarity (FSIM) [40].

10 Construction details of the HDRC database

10.1 Reference HDR Image Selection

Reference images often play a crucial role within databases. They not only serve as the benchmarks for the training and testing of the metrics, but also contribute to data enrichment by providing a diverse range of visual content. Herein, the following three elements are seriously considered when selecting reliable reference HDR images: image quality, the size of images, and content diversity.

The high-quality candidate image set With thorough investigation, the Fairchild database is constructed by Fairchild in the HDR photographic survey [41], including 106 4k HDR images with detailed colorimetric data and absolute luminance calibration. The HDREye Database [42] is conducted to assess the impact of HDR images on the human visual system (HVS), which provides 46 2k high-quality HDR images. Therefore, the Fairchild database and reference images in the HDREye Database are used as the candidate image set, which consists of 152 high-quality HDR images in total.

The size of the reference images The size of the images is primarily constrained by the resolution of the display and the experimental methodology. We employ a 4K monitor for our evaluations. During the subjective testing process, it's necessary to simultaneously present two images on the monitor: one as the reference image and the other as the distorted image. To ensure that the images can be displayed fully on the monitor, all images are set to a size of 1920×1080 pixels. Fairchild's images have a resolution of 4K. Therefore, we crop each image to an appropriate region of 1920×1080 pixels. The images in the HDREye dataset already meet these size requirements, obviating the need for further cropping.

Content diversity of the reference images In order to choose images with various content from the candidate image set, three parameters are applied for reference:

(1) *Dynamic Range*, which measures the overall contrast of the images:

$$DR = L_{max}/L_{min}, \quad (1)$$

where the L_{max} and L_{min} denote the maximum and minimum value of the luminance map of the image. Note that the top and bottom 0.1% of the brightest and darkest pixels are excluded to mitigate the impact of outliers before calculation.

(2) *Key* [43], which measures the subjective feeling of the overall brightness of the images:

$$\log L_{avg} = \frac{1}{N} \sum \log[\sigma + L_i], \quad (2)$$

$$\text{Key} = \frac{\log L_{avg} - \log L_{min}}{\log L_{max} - \log L_{min}}, \quad (3)$$

where N , L_i , and σ refer to the number of pixels, the value of i -th pixel in the luminance map, and the small positive number (*i.e.*, $\sigma = 1e^{-20}$) to avoid L_i being equal to 0.

(3) *Spatial Information* [38], which measures the overall structure of the images:

$$\text{SI} = \text{std}[\text{Sobel}(I)], \quad (4)$$

after applying the Sobel filter (Sobel) to the input LDR image I , the standard deviation (std) is calculated across the pixels within the filtered image. Herein, Reinhard's tone-mapping [32] operator is employed to generate LDR images from HDR images.

Considering the above three characters, 80 high-quality HDR images are selected from the candidate image set with rich textures and scene information. As shown in Fig. 1, the scenes included in this database contain a variety of categories, such as daytime, nighttime, architecture, natural landscapes, human elements, indoor settings, and outdoor environments. Furthermore, the three parameters of the

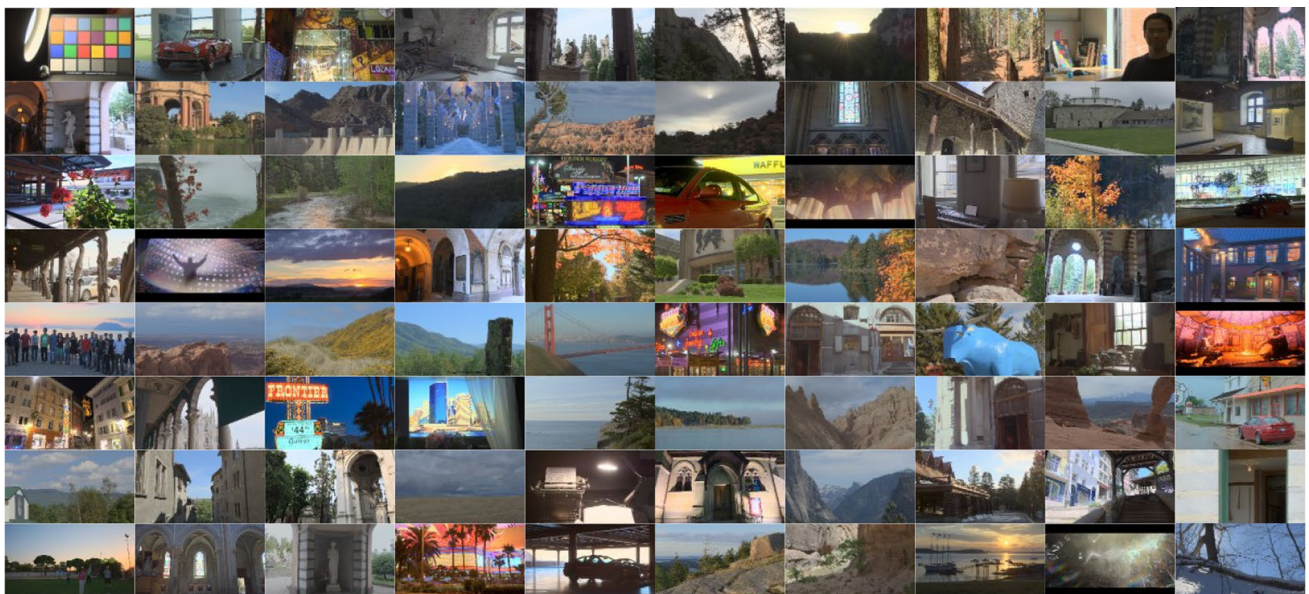


Fig. 1 The reference HDR images in HDRC database. Note that all the HDR images are tone-mapped by Reinhard's TMO for better visualization

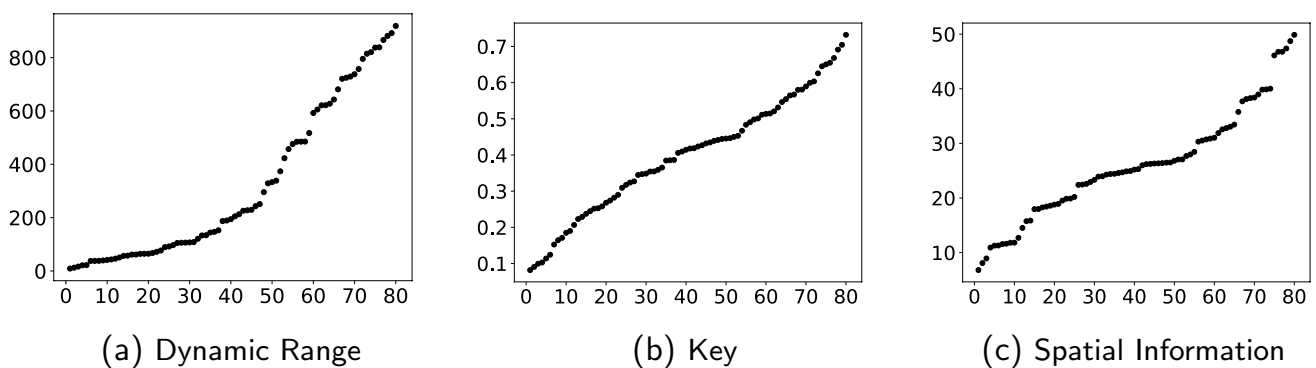


Fig. 2 The three content diversity parameters of the selected reference HDR images. Note that the values in the three figures are arranged in ascending order

reference images in HDRC are presented in Fig. 2, which illustrates the content diversity by a large range of DR, Key, and SI. Note that all the HDR images in this paper are tone-mapped by Reinhard's tone-mapping operator for better visualization.

11 Distorted HDR image generation

To provide a comprehensive evaluation of how the compression algorithms affect the perception of the HVS, three degradation levels are set for VVC distortion and two for JPEG-XT distortion in the proposed HDRC database. The detailed explication of distortion typologies, alongside their corresponding parameter configurations, is expounded as follows:

- *JPEG-XT* The HDR images with JPEG-XT compression are obtained by libjpeg [44]. In JPEG-XT, the HDR image is separated into two parts, where the tone-mapped LDR images are encoded in the base layer, and the residuals are encoded in the enhancement layer. In the experiment, the quality factors for the base layer are set to [15, 50], and for the enhancement layer is set to [4, 9]. The tone-mapping operator is set to global Reinhard as default.
- *VVC* The HDR images with VVC compression are obtained by the VTM reference software [45]. In the implementation, each HDR image is converted to the YUV format and then encoded as a single frame with the input bit-depth set to 10. Additionally, the PQ transform is used during the compression by setting the parameter *LMCSSignalType* to 1. The quantization parameter (QP) values for the compression under three levels are set to [5, 15, 25].

12 Subjective experiment

Environment setup The environment of the subjective experiment follows the suggestions in ITU-R BT. 500 [17] with a constant background luminance setting. According to the recommendation, the viewing distance is set to 1.6 times the height of the monitor. The HDR images are displayed on a 55-inch Samsung QN90A Neo QLED 4K Smart TV, which supports the HDR content with a maximum luminance of 3200 cd/m^2 and resolution of 3840×2160 . To properly display the HDR contents on

the monitor, all the testing materials are converted to HDR videos by x265 in FFmpeg with yuv444p10le format, BT. 2020 color space [46], and PQ transfer function. Specifically, the inverse PQ function first converts the raw

HDR image to the non-linear space. Afterwards, the color format of the image is converted from RGB to YCbCr representation. Finally, the floating point data is converted to 10 bits for encoding.

Testing Methodology The subjective experiment is conducted following the DSIS methodology. To be specific, two images are simultaneously displayed on the monitor. Participants will be informed in advance that the left image is the reference image (of the highest quality) and they are required to evaluate the quality of the image on the right based on the reference image. Specifically, two images are simultaneously displayed on the monitor. Participants are informed in advance that the image on the left is the reference image (without distortion), and they need to judge the quality of the image on the right based on the reference image. After observation, the participants need to select the quality level of the distorted one based on their perception. In the experiment, one participant will be involved at one time, seated at the center position in front of the monitor. The images will be displayed for 10 seconds, after which the participant selects the quality level on an electronic spreadsheet.

A comprehensive experimental procedure explanation is provided before the experiment, along with a brief 5-minute test. This preliminary test is intended to familiarize participants with the experimental process and the visual effects of different image quality levels. Note that the images used in this brief test did not overlap with those in the following formal evaluation procedure. Furthermore, the display order of images was randomized during the formal phase. During the experiment, a supervisor was available to provide assistance to participants as needed without causing any disruption to the experimental process. During the experiment, the database was randomized and divided into 4 groups, each containing 100 pairs of RHI and DHI. The average completion time for one group of data by each participant was approximately 25 mins. If the subjective evaluation time exceeded 30 mins, the supervisor would remind the participant to take a break to prevent visual fatigue. There are a total of 20 participants took part in this experiment, including 12 females and 8 males. The database contains a total of 80 reference HDR images and 400 degraded HDR images. After the experiment, each distorted image received evaluations from 20 participants, resulting in a total of 8×10^3 subjective assessment scores.

Statistical property analysis Herein, two characters of the obtained MOS values are analyzed below to demonstrate the property of the subjective assessment scores. First, to measure the distribution of the quality ratings, the histogram of the MOS values of the distorted HDR images in the HDRC database is provided in Fig. 3. It can be observed that the distribution of MOS values is relatively uniform across different distortion levels, which

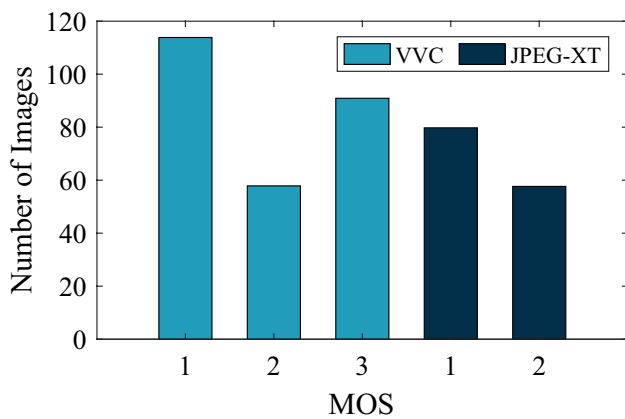


Fig. 3 Histogram of the MOS values of distorted HDR images in HDRC

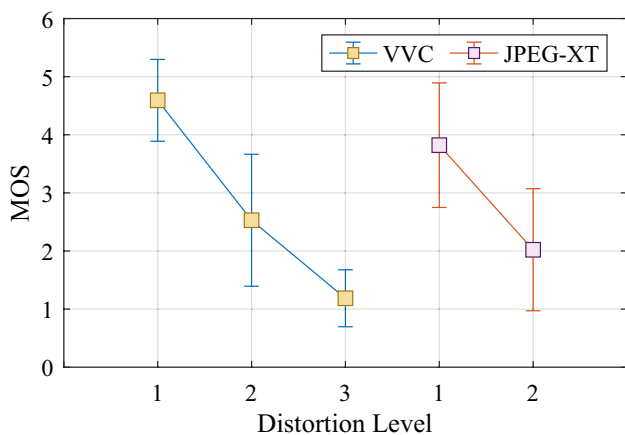


Fig. 4 Dependence of MOS on distortion type and level

indicates that the quality range of distorted images within the constructed database is wide, and they exhibit good discrimination. Additionally, Fig. 4 provides the dependence of MOS on distortion type and level, where distinct colors differentiate the two distortion types. In each distortion type, the error bars indicate the average value and confidence intervals of the distortion levels. The result shows that distorted HDR images with the same distortion type and level tend to receive similar perceptual quality scores. In contrast, the MOS averages of images under different levels decrease as the quality decreases, which suggests the perceptual similarity in human perception of HDR images.

13 Experiment

In this section, the experiment setting is first introduced, followed by the performance comparison of state-of-the-art full reference IQA metrics on existing HDR-IQA databases and the proposed HDRC.

14 Experiment protocol

14.1 IQA metrics for HDR images

To comprehensively evaluate IQA metrics for HDR images, three dedicated HDR content quality assessment algorithms are adopted in the comparison, including HDR-VDP-3 [47], HDR-VQM [29], and LGFM [28]. Additionally, nine commonly used and state-of-the-art LDR-IQA metrics are selected to evaluate the quality of perceptual-mapped HDR image, including PSNR, SSIM, MSSIM, FSIM, gradient magnitude similarity deviation (GMSD) [48], visual information fidelity (VIF) [49], edge similarity (ESIM) [50], Gabor feature-based model (GFM) [51], and deep image structure and texture similarity (DISTS) [52]. To broaden the evaluation scope, four perceptual mapping functions are adopted in the experiments:

- *Linear* mapping, where the HDR images are converted to luminance images with the value of the pixels linearly related to the input image.
- *Gamma* correction, where the HDR images are compressed with the Gamma encoder with the Gamma value set to 1/2.2.
- *PQ* function, which is used in the JPEG-XT and VVC for data compression.
- *PU21* function, which is proposed for HDR image and video quality assessment by Mantiuk et al..

15 Databases

In addition to the proposed HDRC database, four publicly available HDR-IQA databases, including HIJC, HIFM, eHIFM, and UPIQ, are incorporated into our experiments to assess the performance of HDR-IQA metrics.

16 Evaluation criterias

Initially, the objective scores predicted by the IQA metrics are first transformed to a common space by a logistic function as suggested in [53],

$$Q(q) = \gamma_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\gamma_2(q - \gamma_3)}} \right) + \gamma_4 q + \gamma_5, \quad (5)$$

where q and $Q(q)$ represent the objective score and corresponding transformed results. The parameters γ_1 , γ_2 , γ_3 , γ_4 , and γ_5 are determined by fitting the regression model.

Subsequently, three well-established evaluation metrics, namely the Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Order Correlation Coefficient (KROCC), and the Pearson Linear Correlation Coefficient (PLCC), are employed to assess the performance of the HDR-IQA metrics. Furthermore, the average scores for each metric over the five databases are computed for the overall performance evaluation. Herein, the direct average score (AVG_d) is computed, giving the same weight for each database, while the weighted average score (AVG_w) is obtained given the number of DHIs in the database as the weighting factors.

17 Performance of IQA metrics

17.1 Performance of the IQA metrics on multiple HDR-IQA databases

To evaluate the performance of existing IQA metrics, the SROCC, KROCC, and PLCC on four databases and the proposed HDRC are shown in Tables 2, 3, and 4, respectively. For each database, the first-ranked, second-ranked, and third-ranked metrics are marked in red, blue, and black in bold. For each IQA metric, the poorest performance in the five databases is marked with an underline.

The performance of the 39 metrics can be evaluated by counting the number of times they appear in the top three rankings (denoted as N_3) on five datasets in Table 2, 3, and 4. First, considering the five individual databases, the LGFM, PU21-DISTS, and PQ-DISTS achieve much better performances than other metrics with N_3 equal to 10, 10, and 8, indicating that the compression features extracted

Table 2 Performance evaluation of the state-of-the-art IQA metrics in terms of SROCC $\uparrow$ on four public databases and the proposed HDRC

Metric	HIFM	eHIFM	HIJC	UPIQ	HDRC	AVG_d	AVG_w
Linear-PSNR	0.5253	0.5097	0.5740	0.5780	<u>0.4568</u>	0.5288	0.5238
Linear-SSIM	0.5622	<u>0.5193</u>	0.5502	0.6173	0.5410	0.5580	0.5681
Linear-MSSIM	0.6464	0.5938	<u>0.5779</u>	0.7184	0.7299	0.6533	0.6883
Linear-GMSD	0.7820	0.7412	<u>0.5896</u>	0.7805	0.8111	0.7409	0.7618
Linear-FSIM	0.8515	0.7330	<u>0.6192</u>	0.7625	0.8335	0.7599	0.7703
Linear-ESIM	0.8230	0.7400	<u>0.6290</u>	0.8157	0.8513	0.7718	0.7963
Linear-VIF	0.7893	0.7136	<u>0.5370</u>	0.7373	0.8264	0.7207	0.7427
Linear-GFM	0.7331	0.7494	<u>0.6045</u>	0.7875	0.6972	0.7143	0.7226
Linear-DISTS	0.8167	0.7842	<u>0.6297</u>	0.8187	0.8599	0.7818	0.8044
Gamma-PSNR	0.5689	0.5810	0.5761	0.6135	<u>0.3358</u>	0.5351	0.5004
Gamma-SSIM	0.7118	0.6217	<u>0.6174</u>	0.7081	0.6807	0.6679	0.6775
Gamma-MSSIM	0.7888	<u>0.6921</u>	0.7237	0.8352	0.8327	0.7745	0.8034
Gamma-GMSD	0.8754	0.8261	<u>0.7662</u>	0.8763	0.8601	0.8408	0.8503
Gamma-FSIM	0.8774	0.8135	<u>0.7335</u>	0.8414	0.8732	0.8278	0.8373
Gamma-ESIM	0.8650	0.8291	<u>0.7514</u>	0.8793	0.8621	0.8374	0.8499
Gamma-VIF	0.8219	0.7539	<u>0.6355</u>	0.7784	0.8529	0.7685	0.7859
Gamma-GFM	0.7860	0.8103	<u>0.6802</u>	0.7942	0.7513	0.7644	0.7636
Gamma-DISTS	0.8559	0.8213	<u>0.8094</u>	0.8684	0.8770	0.8464	0.8585
PQ-PSNR	0.6681	0.6739	0.6198	0.6305	<u>0.1084</u>	0.5401	0.4414
PQ-SSIM	0.8843	0.7736	<u>0.6667</u>	0.7451	0.6760	0.7491	0.7177
PQ-MSSIM	0.9037	0.8205	0.8392	0.8232	<u>0.8016</u>	0.8376	0.8209
PQ-GMSD	0.8846	0.8723	0.8035	<u>0.7730</u>	0.7925	0.8252	0.7988
PQ-FSIM	0.9064	0.8358	0.8213	<u>0.7648</u>	0.7828	0.8222	0.7924
PQ-ESIM	0.9139	0.8836	<u>0.7685</u>	0.8529	0.8380	0.8514	0.8413
PQ-VIF	0.9158	0.8283	<u>0.6631</u>	0.7568	0.8380	0.8004	0.7878
PQ-GFM	0.8984	0.9088	0.7119	0.7693	<u>0.6960</u>	0.7969	0.7531
PQ-DISTS	0.9300	0.8711	0.8910	<u>0.8526</u>	0.8713	0.8832	0.8702
PU21-PSNR	0.6402	0.6519	0.6058	0.6343	<u>0.1491</u>	0.5363	0.4525
PU21-SSIM	0.8469	0.7328	<u>0.6736</u>	0.7539	0.6761	0.7367	0.7163
PU21-MSSIM	0.8971	<u>0.8001</u>	0.8305	0.8452	0.8220	0.8390	0.8328
PU21-GMSD	0.8890	0.8667	<u>0.7960</u>	0.8155	0.8154	0.8365	0.8209
PU21-FSIM	0.9051	0.8401	0.8116	<u>0.7993</u>	0.8061	0.8324	0.8122
PU21-ESIM	0.9086	0.8736	<u>0.7703</u>	0.8620	0.8458	0.8521	0.8465
PU21-VIF	0.9069	0.8197	<u>0.6641</u>	0.7696	0.8474	0.8015	0.7948
PU21-GFM	0.8859	0.8990	<u>0.7120</u>	0.7935	0.7292	0.8039	0.7724
PU21-DISTS	0.9283	0.8611	0.8799	<u>0.8554</u>	0.8874	0.8824	0.8746
HDR-VQM	0.9041	0.8239	<u>0.8024</u>	0.8175	0.8296	0.8355	0.8245
HDR-VDP-3	0.9149	<u>0.7629</u>	0.8465	0.8465	0.8396	0.8421	0.8394
LGFM	0.9200	0.9322	0.8539	0.9032	<u>0.8069</u>	0.8832	0.8641

Table 3 Performance evaluation of the state-of-the-art IQA metrics in terms of KROCC $\uparrow$ on four public databases and the proposed HDRC

Metric	HIFM	eHIFM	HIJC	UPIQ	HDRC	AVG_d	AVG_w
Linear-PSNR	0.3877	0.3491	0.4211	0.4092	<u>0.3105</u>	0.3755	0.3677
Linear-SSIM	0.4222	0.3757	0.4032	0.4355	<u>0.3682</u>	0.4010	0.3999
Linear-MSSIM	0.4928	0.4275	<u>0.4208</u>	0.5259	0.5320	0.4798	0.5029
Linear-GMSD	0.5947	0.5508	<u>0.4366</u>	0.5911	0.6131	0.5573	0.5742
Linear-FSIM	0.6801	0.5467	<u>0.4526</u>	0.5665	0.6386	0.5769	0.5808
Linear-ESIM	0.6325	0.5601	<u>0.4640</u>	0.6266	0.6604	0.5887	0.6107
Linear-VIF	0.5852	0.5374	<u>0.3911</u>	0.5535	0.6439	0.5422	0.5644
Linear-GFM	0.5667	0.5811	<u>0.4535</u>	0.6078	0.5045	0.5427	0.5437
Linear-DISTS	0.6091	0.5819	<u>0.4605</u>	0.6341	0.6733	0.5918	0.6185
Gamma-PSNR	0.4255	0.3968	0.4266	0.4427	<u>0.2206</u>	0.3824	0.3532
Gamma-SSIM	0.5454	<u>0.4554</u>	0.4564	0.5229	0.4863	0.4933	0.4949
Gamma-MSSIM	0.6029	<u>0.5071</u>	0.5401	0.6472	0.6311	0.5857	0.6113
Gamma-GMSD	0.6817	0.6276	<u>0.5860</u>	0.6912	0.6634	0.6500	0.6600
Gamma-FSIM	0.7097	0.6450	<u>0.5498</u>	0.6524	0.6839	0.6482	0.6518
Gamma-ESIM	0.6702	0.6482	<u>0.5685</u>	0.7002	0.6687	0.6512	0.6640
Gamma-VIF	0.6114	0.5722	<u>0.4663</u>	0.5900	0.6635	0.5807	0.5994
Gamma-GFM	0.5881	0.6264	<u>0.5117</u>	0.6038	0.5489	0.5758	0.5720
Gamma-DISTS	0.6423	<u>0.6167</u>	0.6265	0.6855	0.6920	0.6526	0.6713
PQ-PSNR	0.4797	0.4910	0.4571	0.4565	<u>0.0685</u>	0.3906	0.3171
PQ-SSIM	0.6932	0.5823	0.4996	0.5598	<u>0.4747</u>	0.5619	0.5282
PQ-MSSIM	0.7228	0.6345	0.6619	0.6300	<u>0.5866</u>	0.6472	0.6231
PQ-GMSD	0.6900	0.6898	0.6213	0.5755	<u>0.5741</u>	0.6301	0.5972
PQ-FSIM	0.7392	0.6583	0.6401	0.5669	<u>0.5655</u>	0.6340	0.5930
PQ-ESIM	0.7392	0.7088	<u>0.5863</u>	0.6720	0.6349	0.6682	0.6529
PQ-VIF	0.7307	0.6753	<u>0.4977</u>	0.5761	0.6395	0.6239	0.6050
PQ-GFM	0.7113	0.7359	0.5380	0.5738	<u>0.4945</u>	0.6107	0.5608
PQ-DISTS	0.7589	0.7020	0.7262	<u>0.6625</u>	0.6804	0.7060	0.6861
PU21-PSNR	0.4633	0.4643	0.4461	0.4622	<u>0.0954</u>	0.3863	0.3244
PU21-SSIM	0.6555	0.5484	0.5074	0.5685	<u>0.4748</u>	0.5509	0.5275
PU21-MSSIM	0.7064	0.6171	0.6539	0.6574	<u>0.6120</u>	0.6494	0.6386
PU21-GMSD	0.7015	0.6781	0.6141	0.6238	<u>0.6034</u>	0.6442	0.6235
PU21-FSIM	0.7277	0.6640	0.6301	0.6039	<u>0.5958</u>	0.6443	0.6158
PU21-ESIM	0.7277	0.6975	<u>0.5887</u>	0.6861	0.6445	0.6689	0.6601
PU21-VIF	0.7160	0.6591	<u>0.5039</u>	0.5886	0.6509	0.6237	0.6123
PU21-GFM	0.6932	0.7206	0.5368	0.6017	<u>0.5262</u>	0.6157	0.5800
PU21-DISTS	0.7540	0.6870	0.7121	<u>0.6678</u>	0.7037	0.7049	0.6930
HDR-VQM	0.7359	<u>0.6292</u>	0.6610	0.6397	0.6309	0.6593	0.6429
HDR-VDP-3	0.7474	<u>0.5714</u>	0.6712	0.6613	0.6346	0.6572	0.6484
LGFM	0.7491	0.7707	0.6824	0.7236	<u>0.6045</u>	0.7061	0.6793

by these metrics exhibit a higher consistency with human visual perception. Furthermore, Gamma-DISTS and Gamma-ESIM also show competitiveness across multiple databases with both N_3 equal to 3. Additionally, an analysis of the average and weighted average scores across the five databases is conducted. Notably, PQ-DISTS, PU21-DISTS, and LGFM consistently maintained high rankings in most cases, underscoring their exceptional overall performance.

When we further focus on the proposed HDRC database, it becomes evident that the top-ranked score is notably lower in comparison to the scores obtained in the other

four existing databases in terms of both SROCC, KROCC, and PLCC. In addition, most of the metrics provided the lowest SROCC score in the HIJC database with 23 times, followed by the proposed HDRC database with 7 times. When it comes to the KROCC, the number of the metrics that perform worst in the HIJC and HDRC comes to 16 and 15, which emphasizes the greater challenges posed by the latest HDR compression distortions to HDR-IQA metrics and the necessity of introducing this database.

18 Performance of the IQA metrics on specific distortion types within the HDRC database

For better visualization and further analysis, the SROCC of the IQA metrics on JPEG-XT distortion and VVC distortion within the HDRC database are performed in Figs. 5 and 6, respectively.

Table 4 Performance evaluation of the state-of-the-art IQA metrics in terms of PLCC $\uparrow$ on four public databases and the proposed HDRC

Metric	HIFM	eHIFM	HIJC	UPIQ	HDRC	AVG_d	AVG_w
Linear-PSNR	0.5610	0.4857	0.5744	0.5345	<u>0.4098</u>	0.5131	0.4906
Linear-SSIM	0.4410	0.4082	<u>0.3762</u>	0.4825	0.5324	0.4481	0.4774
Linear-MSSIM	0.5189	0.4613	<u>0.4296</u>	0.5915	0.6083	0.5219	0.5598
Linear-GMSD	0.7984	0.7288	<u>0.5231</u>	0.7533	0.7366	0.7080	0.7150
Linear-FSIM	0.7244	0.5069	<u>0.4643</u>	0.5982	0.6722	0.5932	0.6044
Linear-ESIM	0.8119	0.7122	<u>0.6759</u>	0.7957	0.8261	0.7644	0.7833
Linear-VIF	0.7722	<u>0.3006</u>	0.5541	0.7263	0.7926	0.6292	0.6896
Linear-GFM	0.6103	0.5894	<u>0.5893</u>	0.7878	0.6701	0.6494	0.6901
Linear-DISTS	0.8279	0.7741	<u>0.6378</u>	0.8164	0.8260	0.7764	0.7918
Gamma-PSNR	0.5891	0.5788	0.5779	0.6188	<u>0.2877</u>	0.5305	0.4854
Gamma-SSIM	0.5998	<u>0.4923</u>	0.5427	0.6258	0.6313	0.5784	0.6027
Gamma-MSSIM	0.6417	<u>0.4698</u>	0.6273	0.7290	0.6777	0.6291	0.6678
Gamma-GMSD	0.8913	0.8301	<u>0.7032</u>	0.8371	0.7937	0.8111	0.8043
Gamma-FSIM	0.8931	<u>0.3860</u>	0.6346	0.6933	0.7120	0.6638	0.6729
Gamma-ESIM	0.8807	0.8169	<u>0.7731</u>	0.8619	0.8515	0.8368	0.8424
Gamma-VIF	0.8383	<u>0.5347</u>	0.6883	0.7836	0.8338	0.7357	0.7684
Gamma-GFM	0.7965	0.6982	<u>0.6500</u>	0.7769	0.7131	0.7269	0.7293
Gamma-DISTS	0.8710	0.8296	<u>0.7984</u>	0.8570	0.8587	0.8429	0.8476
PQ-PSNR	0.6342	0.6826	0.5828	0.6390	<u>0.1008</u>	0.5279	0.4357
PQ-SSIM	0.7882	0.6380	0.6189	0.6539	<u>0.5796</u>	0.6557	0.6263
PQ-MSSIM	0.5694	<u>0.4358</u>	0.7202	0.6912	0.6452	0.6124	0.6489
PQ-GMSD	0.9157	0.8881	0.7209	<u>0.6374</u>	0.7106	0.7745	0.7122
PQ-FSIM	0.4043	<u>0.3137</u>	0.6652	0.5714	0.6705	0.5250	0.5895
PQ-ESIM	0.9120	0.8856	<u>0.7967</u>	0.8437	0.8407	0.8557	0.8431
PQ-VIF	0.9256	0.7902	<u>0.7197</u>	0.7532	0.8309	0.8039	0.7887
PQ-GFM	0.7901	0.7998	0.6736	0.6858	<u>0.6420</u>	0.7183	0.6833
PQ-DISTS	0.9501	0.8877	0.8959	<u>0.8555</u>	0.8606	0.8900	0.8704
PU21-PSNR	0.6954	0.6588	0.5797	0.6426	<u>0.1296</u>	0.5412	0.4478
PU21-SSIM	0.7150	<u>0.5467</u>	0.6273	0.6701	0.5999	0.6318	0.6288
PU21-MSSIM	0.5575	<u>0.4266</u>	0.7152	0.7222	0.6712	0.6185	0.6673
PU21-GMSD	0.9154	0.8832	0.7190	<u>0.6981</u>	0.7483	0.7928	0.7468
PU21-FSIM	0.4076	<u>0.3141</u>	0.6698	0.6223	0.6867	0.5401	0.6143
PU21-ESIM	0.9035	0.8733	<u>0.7990</u>	0.8491	0.8469	0.8544	0.8461
PU21-VIF	0.9096	0.7596	<u>0.7231</u>	0.7708	0.8378	0.8002	0.7944
PU21-GFM	0.7786	0.8038	<u>0.6753</u>	0.7372	0.6863	0.7362	0.7178
PU21-DISTS	0.9431	0.8732	0.8870	0.8579	0.8795	0.8881	0.8753
HDR-VQM	0.9150	0.8227	0.8066	<u>0.8017</u>	0.8106	0.8313	0.8129
HDR-VDP-3	0.9386	<u>0.7537</u>	0.8674	0.8390	0.8429	0.8483	0.8411
LGFM	0.9279	0.9345	0.8644	0.8813	<u>0.8197</u>	0.8856	0.8632

From the results, three key observations emerge. Firstly, it is evident that IQA metrics exhibit a higher consistency with HVS in the case of VVC distortions as compared to JPEG-Xt distortions. Secondly, when considering individual distortion types, FSIM, DISTS, and HDR-VQM demonstrate superior performance relative to other metrics. Furthermore, for LDR IQA metrics, the Gamma correction and PU21 function significantly outperform PQ and linear mapping. Additionally, DISTS outperforms other metrics on both the single distortion type and overall evaluation, which implies the importance of the texture information in establishing the correlation between image features and human visual perception.

The scatter plots of the MOS value and predicted quality score are shown in Fig. 7. The fitted scores are generated from Equation 5. For a single distortion type, a denser clustering of points around the diagonal indicates a higher correlation between the predicted quality scores and MOS values. It can be observed that most HDR-IQA metrics exhibit

a noticeable deviation from the fitted line, suggesting that existing IQA models are ineffective in evaluating distortions resulting from JPEG-Xt and VVC compression.

19 Performance of various perceptual mapping functions

In this subsection, the effectiveness of the perceptual mapping methods is thoroughly evaluated. Herein, for each mapping method, the average SROCC, KROCC, and PLCC score over the nine LDR metrics on each database is calculated, respectively. The results, as depicted in Table 5, reveal an intriguing observation. It is noteworthy that in the HIFM, eHIFM, and HIJC databases, the PQ function outperforms other mapping methods. However, the performance of the PQ function declines in the UPIQ database and is even inferior to the linear mapping algorithm in the HDRC database. Likewise, the PU21 function exhibits subpar performance in

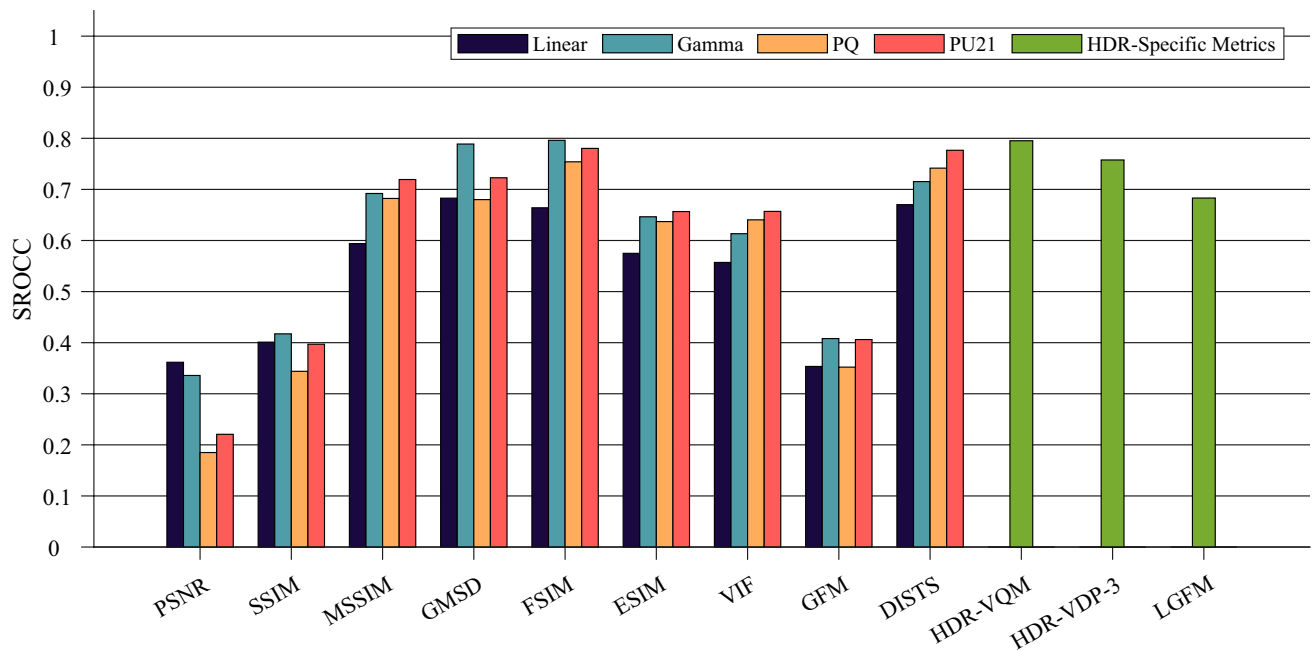


Fig. 5 Performance evaluation of the state-of-the-art IQA metrics in terms of SROCC on the JPEG-XT compression distortion type in the proposed HDRC database

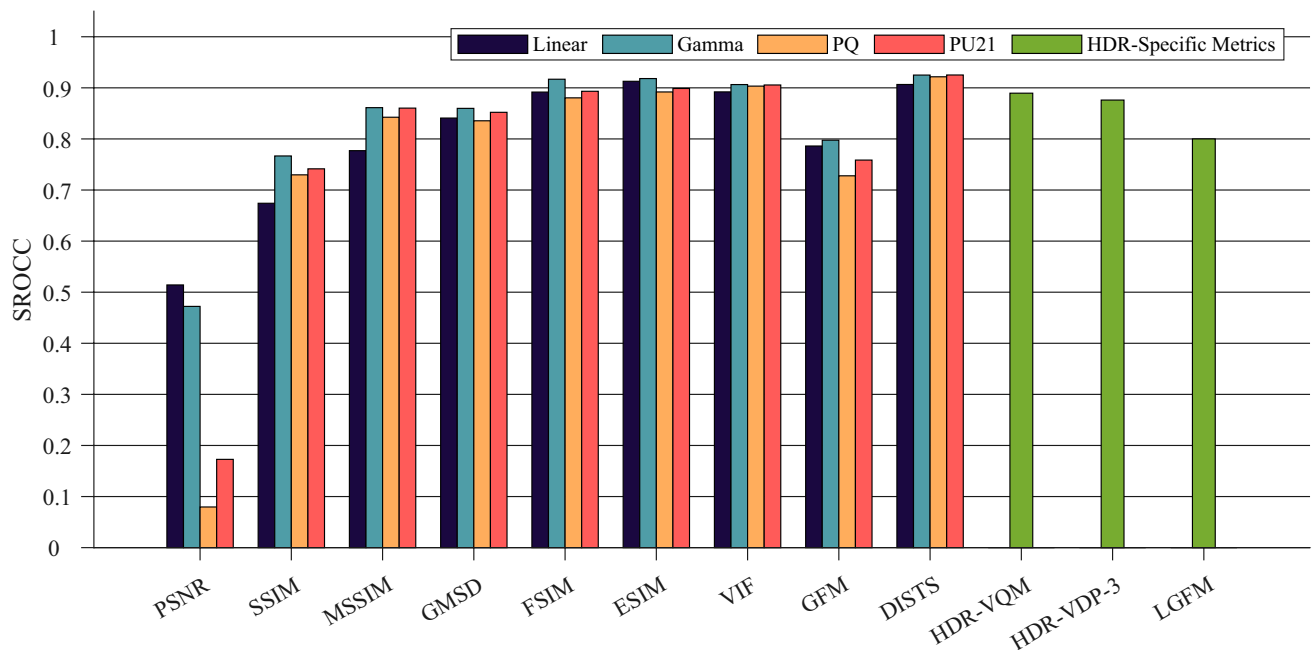


Fig. 6 Performance evaluation of the state-of-the-art IQA metrics in terms of SROCC on the VVC compression distortion type in the proposed HDRC database

the HDRC database compared to the linear mapping method. This observation is consistent with the findings presented in Figs. 5 and 6. The evaluation results for JPEG-XT distortions indicate that among the nine LDR metrics, four

algorithms exhibit better performance with PU21 mapping, while five algorithms perform better with Gamma mapping. However, regarding the VVC distortion, it is evident that for eight of the algorithms, Gamma mapping outperforms

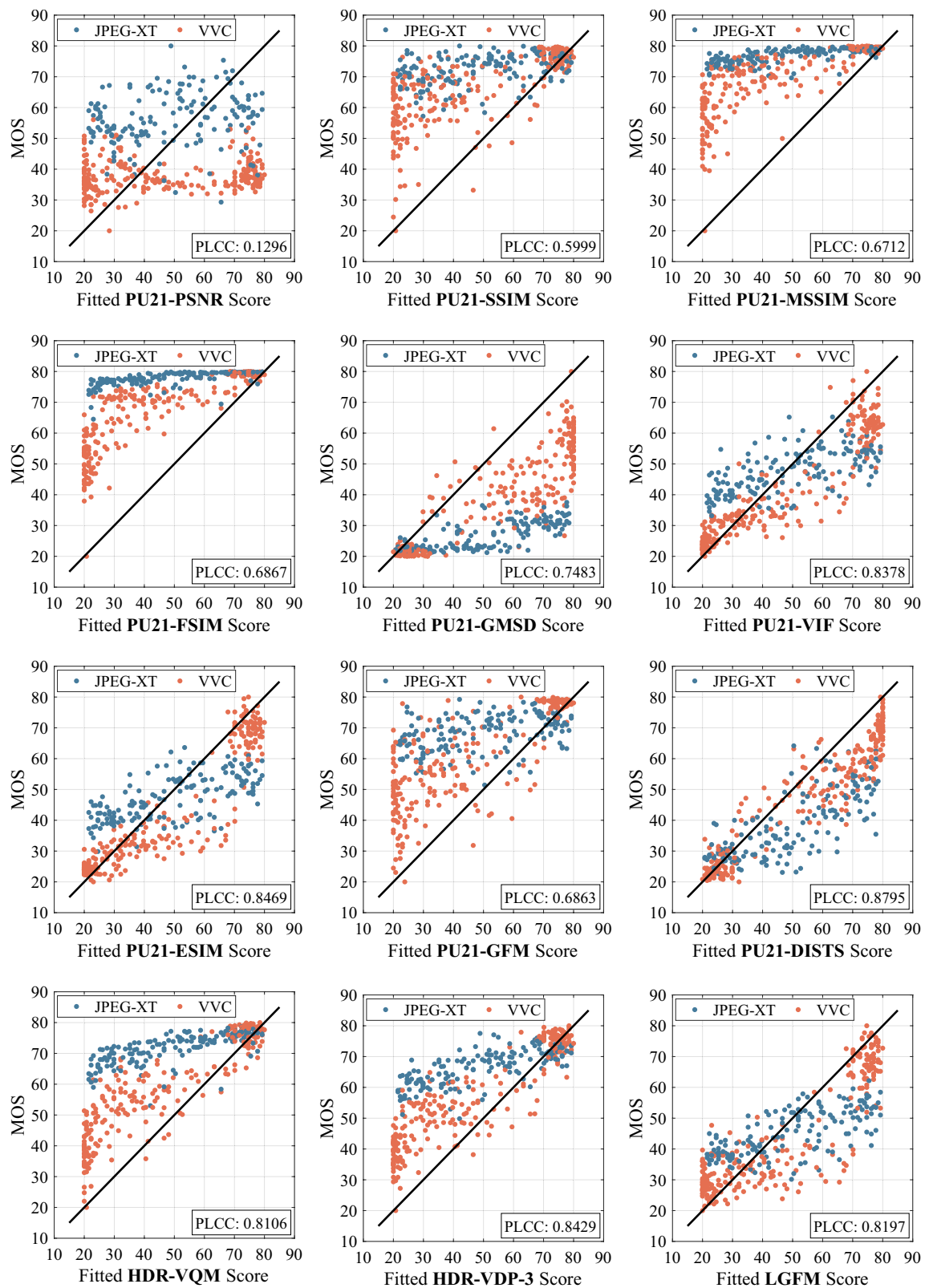


Fig. 7 A set of scatter plots showing the MOS values versus the objective scores computed by the IQA models. A denser clustering of points around the diagonal (black line) indicates a higher correlation between the predicted quality scores and MOS values

Table 5 Performance evaluation of the perceptual mapping functions applied for LDR Metrics

Mapping Function	HIFM	eHIFM	HIJC	UPIQ	HDRC
SROCC ↑					
Linear	0.7255	0.6760	0.5901	0.7351	0.7341
Gamma	0.7946	0.7499	0.6993	0.7994	0.7695
PQ	0.8784	0.8298	0.7539	0.7742	0.7116
PU21	0.8676	0.8161	0.7493	0.7921	0.7309
KROCC ↑					
Linear	0.5523	0.5011	0.4337	0.5500	0.5494
Gamma	0.6086	0.5662	0.5258	0.6151	0.5843
PQ	0.6961	0.6531	0.5809	0.5859	0.5243
PU21	0.6828	0.6373	0.5770	0.6067	0.5452
PLCC ↑					
Linear	0.6740	0.5519	0.5361	0.6762	0.6749
Gamma	0.7779	0.6263	0.6662	0.7537	0.7066
PQ	0.7655	0.7024	0.7104	0.7035	0.6534
PU21	0.7584	0.6821	0.7106	0.7300	0.6762

PU21 mapping, and linear mapping surpasses PQ mapping on five IQA metrics.

These results further underscore the variability of mapping methods' reliability across different types of distortions. Our newly established HDRC database includes the VVC distortion type that is not present in previous assessments, leading to distinctive outcomes.

This emphasizes the necessity of creating this database and conducting extensive experiments.

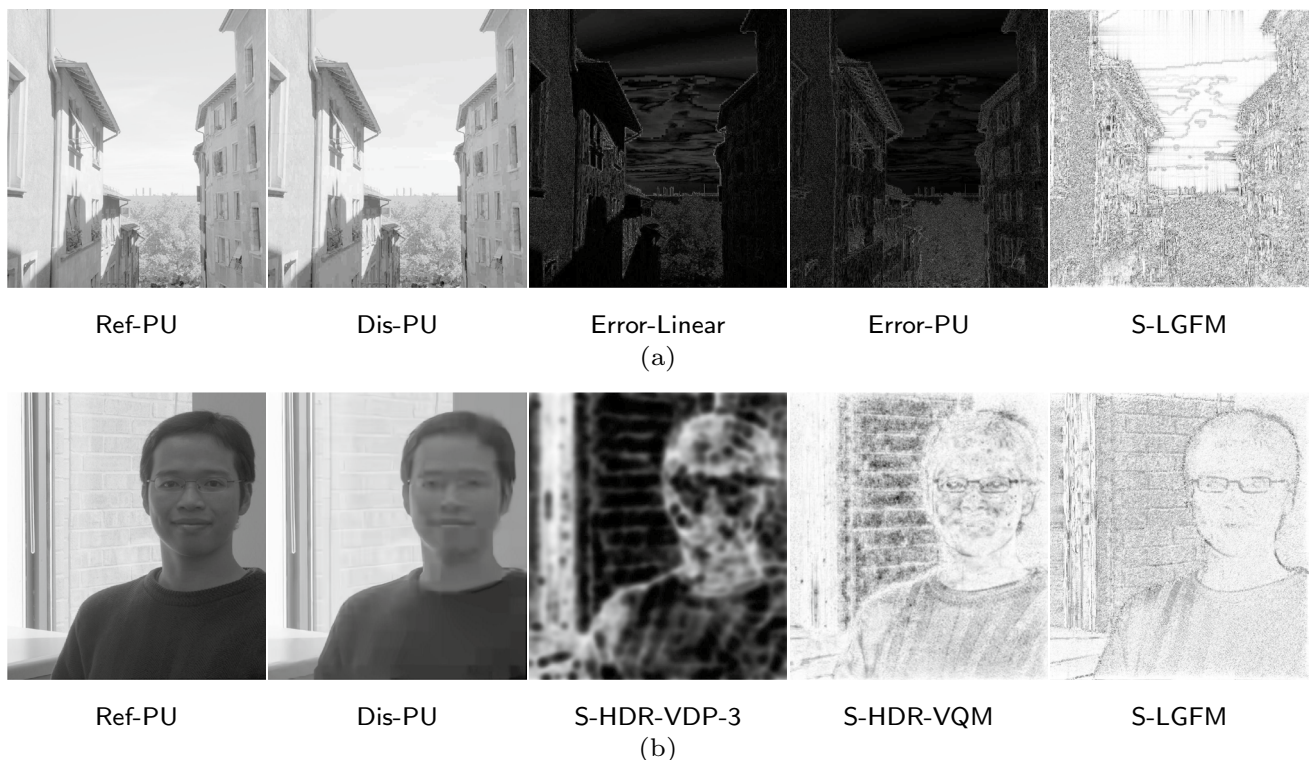


Fig. 8 Example of failure cases. The DHIs in (a) (b) are compressed by the JPEG-Xt and VVC, correspondingly. Ref-PU and Dis-PU signify the PU-encoded reference and distorted HDR images in the luminance domain. Error-Linear and Error-PU denote pixel-wise

error maps in the linear domain and PU domain, respectively. S-HDR-VDP-3, S-HDR-VQM, and S-LGFM represent similarity maps derived from the HDR-VDP-3, HDR-VQM, and LGFM, where a higher pixel value means higher similarity between the RHI and DHI

20 Limitations

In this subsection, three limitations of existing HDR-specific IQA metrics are discussed, aiming to provide a potential design strategy for the HDR IQA models. Firstly, existing HDR IQA methods overlook the significance of the color information. HDR-VQM and LGFM fall into this category, where only the luminance map of the input image is considered. Secondly, the use of PU coding can over-compress high-luminance regions. This issue becomes pronounced with HDR images, where the desire to preserve detail in bright areas clashes with the limitations of PU coding. As shown in Fig. 8a, the block distortions in the sky are visible in the original pixel-wise error map. After the PU coding, such difference becomes invisible compared to the distortions in the low luminance regions. A more meticulous approach to perceptual mapping functions is imperative to fit the HVS. Lastly, the current HDR IQA metrics lack the content-adaptive module, leading to misjudgments in perceived distortions. As shown in Fig. 8b, where facial distortions are more conspicuous compared to those on the wall and sweater. However, due to the intricate textures of the latter, existing HDR IQA metrics erroneously prioritize them as regions of severely distorted areas. This misalignment with HVS illustrates the necessity for HDR IQA methods to incorporate content adaptability for more accurate assessments.

21 Conclusion and future work

This paper proposes a comprehensive and extensive HDR-IQA database, referred to as HDRC, aiming to provide a benchmark for the development of HDR-IQA algorithms regarding the latest HDR compression distortions brought by VVC and JPEG-XT. The HDRC database contains 80 RHIs and 400 DHIs, with MOS values obtained through subjective experiments using the DSIS methodology. In the experiments, 12 existing IQA models are compared on the proposed HDRC database and four existing databases. The results indicate a decrease in the efficacy of current IQA models when applied to the HDRC database in contrast to other databases. This not only underscores the necessity of constructing the HDRC database but also emphasizes the requirement for new HDR-IQA models dedicated to assessing the latest HDR image compression distortions.

In the realm of HDR image quality assessment, there exist several promising avenues for future research. Firstly, expanding the subjective database with a more extensive collection of distorted HDR image samples and corresponding subjective ratings could enhance the statistical significance of the findings. Additionally, the improvement

and development of dedicated IQA algorithms tailored specifically for HDR images warrant continued attention. This might entail the exploration of novel perceptual mapping methods and content-adaptive attention modules [54, 55] to enhance the accuracy of HDR image quality evaluation. Furthermore, the proposed HDRC database can be leveraged for a range of HDR-related low-level tasks, such as denoising, where its comprehensive HDR content and perceptual data can inform the development of advanced denoising algorithms tailored to distorted HDR images. By annotating specific scenes within the database in the future, it can offer an invaluable platform for investigating compressed HDR image segmentation and edge detection, where some efficient clustering methods can be applied [56, 57]. These future directions represent an exciting frontier in HDR research, harnessing the potential of the HDRC database to drive advancements in various low-level tasks and further enrich the field of HDR-related applications.

Author contributions Author Contributions Statement Yue Liu constructed the database and designed the subjective experiment. She also analyzed the results and wrote the main manuscript text. Zhangkai Ni assisted in generating the distorted images and conducted a thorough analysis of the effectiveness of the IQA methods employed in the study. Peilin Chen prepared figures 5–7 and provided valuable insights during the results analysis. He also helps to enhance the language and clarity of the paper. Shiqi Wang provided essential technical expertise during the database generation phase and offered valuable insights during the analysis of the results. He is one of the project leaders. Sam Kwong served as the chief leader of the project and initiated the research work. He provided supervision to the first author throughout the work and let the team in deriving the research findings.

Funding Open Access Publishing Support Fund provided by Lingnan University. This work in part supported by the Hong Kong GRFRGC General Research Fund under Grant 11203820 and Grant 11209819.

Data availability Our database is available at <https://github.com/Yliu724/HDRC>

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Francois E, Fogg C, He Y, Li X, Luthra A, Segall A (2015) High dynamic range and wide color gamut video coding in hevc: Status and potential future enhancements. *IEEE Trans Circuits Syst Video Technol* 26(1):63–75
2. Artusi A, Mantiuk RK, Richter T, Hanhart P, Korshunov P, Agostinelli M, Ten A, Ebrahimi T (2019) Overview and evaluation of the jpeg xt hdr image compression standard. *J Real-Time Image Proc* 16:413–428
3. Bross B, Chen J, Ohm J-R, Sullivan GJ, Wang Y-K (2021) Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc). *Proc IEEE* 109(9):1463–1493
4. Lu T, Pu F, Yin P, McCarthy S, Husak W, Chen T, Francois E, Chevance C, Hiron F, Chen J, et al (2020) Luma mapping with chroma scaling in versatile video coding. In: 2020 Data Compression Conference (DCC), pp 193–202. IEEE
5. Karczewicz M, Hu N, Taquet J, Chen C-Y, Misra K, Andersson K, Yin P, Lu T, Francois E, Chen J (2021) Vvc in-loop filters. *IEEE Trans Circuits Syst Video Technol* 31(10):3907–3925
6. Wang H, Katsavounidis I, Zhou J, Park J, Lei S, Zhou X, Pun M-O, Jin X, Wang R, Wang X et al (2017) VideoSet: a large-scale compressed video quality dataset based on jnd measurement. *J Vis Commun Image Represent* 46:292–302
7. Jiang Y, Xu Q, Zhang W, Huang Q (2018) Active sampling for subjective video quality assessment. In: 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), pp. 1–5. IEEE
8. Xu Q, Huang Q, Jiang T, Yan B, Lin W, Yao Y (2012) Hodgerank on random graphs for subjective video quality assessment. *IEEE Trans Multimed* 14(3):844–857
9. Narwaria M, Da Silva MP, Le Callet P, Pepion R (2013) Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality. *Opt Eng* 52(10):102008–102008
10. Kuang J, Johnson GM, Fairchild MD (2007) icam06: A refined image appearance model for hdr image rendering. *J Vis Commun Image Represent* 18(5):406–414
11. Valenzise G, De Simone F, Lauga P, Dufaux F (2014) Performance evaluation of objective quality metrics for hdr image compression. In: Applications of Digital Image Processing XXXVII, vol. 9217, pp 78–87. SPIE
12. Mai Z, Mansour H, Mantiuk R, Nasiopoulos P, Ward R, Heidrich W (2010) Optimizing a tone curve for backward-compatible high dynamic range image and video compression. *IEEE Trans Image Process* 20(6):1558–1571
13. ST 2084:2014 SMPTE standard high dynamic range electro-optical transfer function of mastering reference displays. ST 2084:2014, 1–14 (2014) <https://doi.org/10.5594/SMPTE.ST2084.2014>
14. Zerman E, Valenzise G, Dufaux F (2017) An extensive performance evaluation of full-reference hdr image quality metrics. *Qual User Exp* 2(1):1–16
15. Mikhailiuk A, P'erez-Ortiz M, Yue D, Suen W, Mantiuk RK (2021) Consolidated dataset and metrics for high-dynamic-range image quality. *IEEE Trans Multimed* 24:2125–2138
16. Korshunov P, Hanhart P, Richter T, Artusi A, Mantiuk R, Ebrahimi T (2015) Subjective quality assessment database of hdr images compressed with jpeg xt. In: 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), pp 1–6. IEEE
17. Rec. ITU-R BT.500–15: Methodologies for the subjective assessment of the quality of television images. Technical report, International Telecommunications Union (May 2023)
18. Liao X, Chen B, Zhu H, Wang S, Zhou M, Kwong S (2022) Deepwds: Projecting degradations in perceptual space to Wasserstein distance in deep feature space. In: Proceedings of the 30th ACM International Conference ON Multimedia, pp 970–978
19. Ding K, Liu Y, Zou X, Wang S, Ma K (2021) Locally adaptive structure and texture similarity for image quality assessment. In: Proceedings of the 29th ACM International Conference on multimedia, pp 2483–2491
20. Tian Y, Chen B, Wang S, Kwong S (2023) Towards thousands to one reference: can we trust the reference image for quality assessment? *IEEE Trans Multimed* (Early Access)
21. Pan Z, Yuan F, Lei J, Fang Y, Shao X, Kwong S (2022) Vcnet: Visual compensation restoration network for no-reference image quality assessment. *IEEE Trans Image Process* 31:1613–1627
22. Pan Z, Yuan F, Wang X, Xu L, Shao X, Kwong S (2022) No-reference image quality assessment via multibranch convolutional neural networks. *IEEE Trans Artif Intell* 4(1):148–160
23. Zeng C, Kwong S (2023) Combining cnn and transformers for full-reference and no-reference image quality assessment. *Neurocomputing* 549:126437
24. Mantiuk R., Daly, S.J., Myszkowski, K., Seidel, H.-P.: Predicting visible differences in high dynamic range images: model and its calibration. In: Human Vision and Electronic Imaging X, vol. 5666, pp. 204–214 (2005). SPIE
25. Mantiuk R, Myszkowski K, Seidel H-P (2004) Visible difference predictor for high dynamic range images. In: 2004 IEEE International Conference on systems, man and cybernetics (IEEE Cat. No. 04CH37583), vol. 3, pp. 2763–2769. IEEE
26. Mantiuk R, Kim KJ, Rempel AG, Heidrich W (2011) Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans Graph (TOG)* 30(4):1–14
27. Mantiuk RK, Hammou D, Hanji P (2023) Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625*
28. Liu Y, Ni Z, Wang S, Wang H, Kwong S (2023) High dynamic range image quality assessment based on frequency disparity. *IEEE Trans Circuits Syst Video Technol* 33(8):4435–4440
29. Narwaria M, Da Silva MP, Le Callet P (2015) Hdr-vqm: an objective quality measure for high dynamic range video. *Signal Process Image Commun* 35:46–60
30. Aydın TO, Mantiuk R, Seidel H-P (2008) Extending quality metrics to full luminance range images. In: Human Vision and Electronic Imaging XIII, vol. 6806, pp 109–118. SPIE
31. Mantiuk RK, Azimi M (2021) Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In: 2021 Picture Coding Symposium (PCS), pp 1–5. <https://doi.org/10.1109/PCS50896.2021.9477471>
32. Reinhard E, Stark M, Shirley P, Ferwerda J (2023) Photographic tone reproduction for digital images. In: Seminal graphics papers: pushing the boundaries, Volume 2, pp. 661–670
33. Mantiuk R, Myszkowski K, Seidel H-P (2006) A perceptual framework for contrast processing of high dynamic range images. *ACM Trans Appl Perception (TAP)* 3(3):286–308
34. Wang H, Tang P, Li Q, Cheng M (2021) Emotion expression with fact transfer for video description. *IEEE Trans Multimed* 24:715–727
35. Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, Vozel B, Chehdi K, Carli M, Battisti F et al (2015) Image database tid2013: Peculiarities, results and perspectives. *Signal Process Image Commun* 30:57–77
36. Li Q, Li T, Wang H, Chen CW (2022) Taking an emotional look at video paragraph captioning. *arXiv preprint arXiv:2203.06356*

37. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13(4):600–612
38. Rec (2022) ITU-R P.910: Subjective video quality assessment methods for multimedia applications. Technical report, International Telecommunications Union
39. Wang Z, Simoncelli EP, Bovik AC (20023) Multiscale structural similarity for image quality assessment. In: *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, vol. 2, pp 1398–1402. Ieee
40. Zhang L, Zhang L, Mou X, Zhang D (2011) Fsim: a feature similarity index for image quality assessment. *IEEE Trans Image Process* 20(8):2378–2386
41. Mark F The HDR Photographic Survey (2006) <http://markfairchild.org/HDR.html>
42. Korshunov P, Nemoto H, Skodras A, Ebrahimi T (2014) Crowdsourcing-based evaluation of privacy in hdr images. In: *Optics, photonics, and digital technologies for multimedia applications III*, vol. 9138, p. 913802. SPIE
43. Akyu..z AO, Reinhard E (2006) Color appearance in high-dynamic-range imaging. *J Electron Imaging* 15(3):033001–033001
44. Bob F, Lee H, Frank W libjpeg. <https://sourceforge.net/projects/libjpeg/>
45. VVCSoftware VTM Repository (2020) <https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftwareVTM/-/tree/VTM-8.0>
46. Rec. ITU-R BT.2020–2: Parameter values for ultra-high definition television systems for production and international programme exchange. Technical report, International Telecommunications Union (October 2015)
47. Narwaria M, Mantiuk R, Da Silva MP, Le Callet P (2015) Hdr-vdp-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. *J Electron Imaging* 24(1):010501
48. Xue W, Zhang L, Mou X, Bovik AC (2013) Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Trans Image Process* 23(2):684–695
49. Sheikh HR, Bovik AC, De Veciana G (2005) An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans Image Process* 14(12):2117–2128
50. Ni Z, Ma L, Zeng H, Chen J, Cai C, Ma K-K (2017) Esim: Edge similarity for 23 screen content image quality assessment. *IEEE Trans Image Process* 26(10):4818–4831
51. Ni Z, Zeng H, Ma L, Hou J, Chen J, Ma K-K (2018) A gabor feature-based quality assessment model for the screen content images. *IEEE Trans Image Process* 27(9):4516–4528
52. Ding K, Ma K, Wang S, Simoncelli EP (2020) Image quality assessment: Unifying structure and texture similarity. *IEEE Trans Pattern Anal Mach Intell* 44(5):2567–2581
53. Sheikh HR, Sabir MF, Bovik AC (2006) A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans Image Process* 15(11):3440–3451
54. Cong R, Qin Q, Zhang C, Jiang Q, Wang S, Zhao Y, Kwong S (2022) A weakly supervised learning framework for salient object detection via hybrid labels. *IEEE Trans Circuits Syst Video Technol* 33(2):534–548
55. Pan Z, Yu W, Lei J, Ling N, Kwong S (2021) Tsan: Synthesized view quality enhancement via two-stream attention network for 3d-hevc. *IEEE Trans Circuits Syst Video Technol* 32(1):345–358
56. Hong Y, Kwong S, Wang H, Ren Q (2009) Resampling-based selective clustering ensembles. *Pattern Recogn Lett* 30(3):298–305
57. Hong Y, Kwong S, Chang Y, Ren Q (2008) Consensus unsupervised feature ranking from multiple views. *Pattern Recogn Lett* 29(5):595–602

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.