PERCEPTUAL QUALITY ASSESSMENT OF UHD-HDR-WCG VIDEOS

Shahrukh Athar[*1], *Thilan Costa*[*1], *Kai Zeng*[2] *and Zhou Wang*[1]

[1]Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
[2]SSIMWAVE Inc., Waterloo, ON, Canada

ABSTRACT

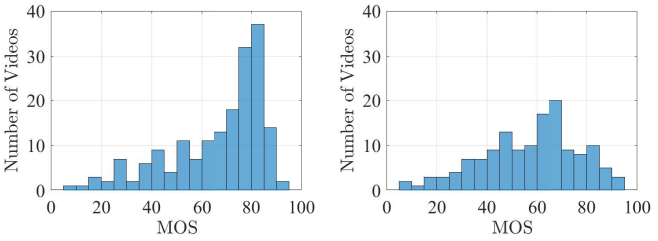
High Dynamic Range (HDR) Wide Color Gamut (WCG) Ultra High Definition (4K/UHD) content has become increasingly popular recently. Due to the increased data rate, novel video compression methods have been developed to maintain the quality of the videos being delivered to consumers under bandwidth constraints. This has led to new challenges for the development of objective Video Quality Assessment (VQA) models, which are traditionally designed without sufficient calibration and validation based on subjective quality assessment of UHD-HDR-WCG videos. The large performance variations between different consumer HDR TVs, and between consumer HDR TVs and professional HDR reference displays used for content production, further complicates the task of acquiring reliable subjective data that faithfully reflects the impact of compression on UHD-HDR-WCG videos. In this work, we construct a first-of-its-kind video database composed of PQ-encoded UHD-HDR-WCG content, which is subsequently compressed by H.264 and HEVC encoders. We carry out a subjective study on a professional 4K-HDR reference display in a controlled lab environment. We also benchmark representative Full Reference (FR) and No-Reference (NR) objective VQA models against the subjective data to evaluate their performance on compressed UHD-HDR-WCG video content. The database will be made available to the public, subject to content copyright constraints.

Index Terms— video quality assessment, high dynamic range, wide color gamut, ultra high definition, 4K, subjective testing, subjective data processing, objective analysis

1. INTRODUCTION

The last two decades have witnessed enormous gains in the development of perceptual objective Image and Video Quality Assessment (IQA/VQA) methods. However, most of these developments were made while working with visual content of low resolution (1080p or below), low dynamic range (8 bits per color) and constrained color gamut (ITU-R BT.709). Recent advances in the acquisition, transmission, display, and storage technologies have resulted in the widespread availability and adoption of High Dynamic Range (HDR) Wide Color Gamut (WCG) Ultra High Definition (4K/UHD) content and displays. This has led to new challenges for the development of objective Video Quality Assessment (VQA) models.

Subjective testing plays a critical role in the development and validation of objective VQA models. Although some recent subjective studies have been carried out on newly constructed HDR databases [1, 2, 3, 4, 5, 6, 7], they exhibit one or more of the following limitations: 1) The maximum spatial resolution of visual content is Full High Definition (FHD 1080p). UHD/4K content is lacking; 2) The color gamut of the content and/or the displays used in these studies is usually BT.709, despite the growing popularity of WCG, such as DCI-P3 and BT.2020; 3) Most studies use content with a maximum temporal resolution of 30 frames per second (fps); 4) Fixed distortion levels such as a given bitrate in video encoding are used to create these databases, regardless of content complexity variations, leading to inadequate perceptual separation across distortion levels and reduced overall effectiveness for objective benchmarking purposes; 5) Only a limited number of Full Reference (FR) objective methods were evaluated, while state-of-the-art FR and No-Reference (NR) models were missing from these tests.



(a) Preliminary Database (b) Waterloo UHD-HDR-WCG Database

Fig. 1: MOS distribution of Waterloo UHD-HDR-WCG database (b) in comparison with a preliminary database (a).

In this work we design a new dataset namely the Waterloo UHD-HDR-WCG Database. It includes PQ-encoded HDR content with UHD resolution, BT.2020 color gamut and two frame rates (24 fps and 59.94 fps). Adaptive bitrates are used to generate perceptually separated H.264 and HEVC compressed videos. We use a state-of-the-art professional 4K-HDR reference display, with a dedicated hardware pipeline, to construct the subjective experiment environment. A novel data processing procedure is used to generate the Mean Opinion Scores (MOS). Finally, we use the subjective data to evaluate the performance of eleven FR and seven NR representative objective quality assessment methods.

2. DATABASE AND HARDWARE SETUP

The Waterloo UHD-HDR-WCG database is created from 14 ten-second high-quality reference videos, all of which have Ultra High Definition (UHD) resolution (3840×2160), bit depth of 10 bits (Luma), YUV 4:2:0 chroma format, SMPTE ST 2084 (PQ) transfer function, and BT.2020 color primaries to ensure Wide Color Gamut (WCG) content. The frame rate is 59.94 fps and 24 fps for nine and five reference videos, respectively. The focus of this work is to study the impact of compression on UHD-HDR-WCG content.

* Corresponding authors. Email: shahrukh.athar@uwaterloo.ca, tcosta@uwaterloo.ca

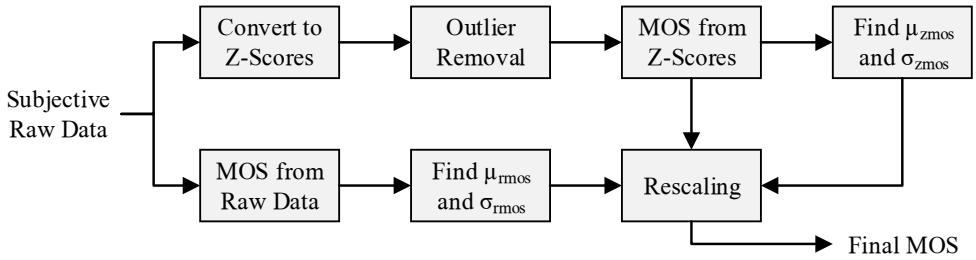


Fig. 2: Process of Mean Opinion Scores (MOS) generation.

Therefore, the reference videos are compressed by two encoders (H.264 and HEVC) at five bitrates each, resulting in 140 distorted videos. One way to construct VQA databases is to distort reference content at predefined distortion levels, that is, by using fixed bitrates for all contents. While this is a convenient approach, it does not lead to a uniform distribution of distorted content in the visual quality range. In order to uncover such issues, we first constructed a test FHD-HDR database that had fixed distortion levels (bitrates) regardless of content and carried out a preliminary subjective test. The MOS histogram of this test database is shown in Fig. 1(a), where it can be seen that this database has a highly non-uniform distribution of distorted content in the quality range. To address this issue, we selected the distortion levels for the Waterloo UHD-HDR-WCG database in a content-adaptive manner. Considering the visual quality range to be [0,100], where 100 is the highest quality, we first encoded the reference videos at multiple bitrates and used a state-of-the-art FR VQA method, SSIMplus [8], to select bitrates that led to distorted videos closest to predefined quality levels (94, 74, 54, 36, 18), followed by manual observation and bitrate adjustment to obtain perceptually separated distorted videos for each reference. The MOS histogram of the Waterloo UHD-HDR-WCG database is shown in Fig. 1(b), where it can be seen that the distorted content is more evenly distributed in the visual quality range for better perceptual quality separation.

Subjective experiments were carried out on a Canon DP-V2420 4K/UHD HDR Reference Display [9] which is a mastering monitor that is compatible with the Academy Color Encoding System (ACES) [10] and supports both the SMPTE ST 2084 (PQ) and Hybrid Log Gamma (HLG) transfer functions. The display's peak luminance is 1000 cd/m^2, minimum black level is 0.005 cd/m^2, and screen size is 24 inch. To preserve the integrity of the content, we used the display's Quad 3G Serial Digital Interface (SDI) which supports a throughput of 12 Gbits/s. This fulfills the maximum throughput requirement of the high frame rate (59.94 fps) content of the database, which stands at around 11.12 Gbits/s. The workstation holding the database was connected through a Blackmagicdesign PCI Express Cable Kit to a Blackmagicdesign Ultrastudio 4K Extreme 3 [11] playback device. Here the single data stream is split into four streams that are connected to the Ultrastudio's SDI output interface, which is connected to the Reference Display. For smooth operation, the compressed videos were decoded and the entire database was stored in the YUV file format. It was ensured that all components in the video playback pipeline were capable of handling the high throughput requirements. Thus, the entire database (around 1.64 TBytes) was stored in a Samsung 2 TByte 960 Pro M.2 PCIe NVMe Solid State Drive (SSD) which is capable of sequential read speeds of up to 3.5 GBytes/s. The workstation is equipped with 32 GBytes of 3000 MHz DDR4 RAM to allow for storing an entire video in memory for quick transmission to the display. For optimal operation, customized video playback software was written by using the Blackmagicdesign Software Development Kit (SDK) which was invoked through MATLAB during subjective testing.

3. SUBJECTIVE STUDY AND DATA PROCESSING

The subjective study was conducted in the Image and Vision Computing (IVC) laboratory at the University of Waterloo in a dark room environment. A total of 51 subjects, including 29 males and 22 females aged between 18 and 35, took part in the study. Further, eight subjects were regarded as experts since they worked in the area of VQA, while the remaining 43 were considered as naïve subjects. All the subjects had normal or corrected-to-normal vision, and were not color-blind. The single-stimulus methodology with hidden reference [12] was used to carry out the study. The subjects were asked to evaluate the content at a viewing distance of approximately twice the screen height. The length of the study was around 80 minutes for each subject, which included two 30 minute rating sessions with a mandatory break in-between to reduce visual fatigue effect. After subjects viewed a 10-second content, the test display went black and a scoring GUI appeared on a secondary screen where subjects recorded their scores by using a sliding bar. The score range of 0 to 100 was divided into intervals of 20 and labeled respectively as Bad, Poor, Fair, Good, Excellent, and subjects could select any integer value in this range. A higher score indicates better visual quality. To help orient the subjects with the test environment and to familiarize them with the quality range, a training session was carried out before the actual test which was composed of five distorted videos with varying distortion levels. The training videos had no overlap with the formal test videos and no instructions were given about which video should get what score.

Raw subjective scores are processed into final Mean Opinion Scores (MOS) by using the procedure shown in Fig. 2, where the goal is to take into account the variations in individual subject quality scales while maintaining the overall mean and variance of the raw scores. Since subjects may use the quality scale variably with respect to each other, raw scores per subject are first converted to Z-scores to account for these variations:

$$z_{ij} = \frac{s_{ij} - \mu_i}{\sigma_i} \tag{1}$$

where s_{ij} denotes the raw score assigned by subject i to video j, z_{ij} denotes the corresponding Z-score, μ_i and σ_i are respectively the mean and standard deviation of all the raw scores assigned by subject i in the test. Next, outlier detection and removal is performed as

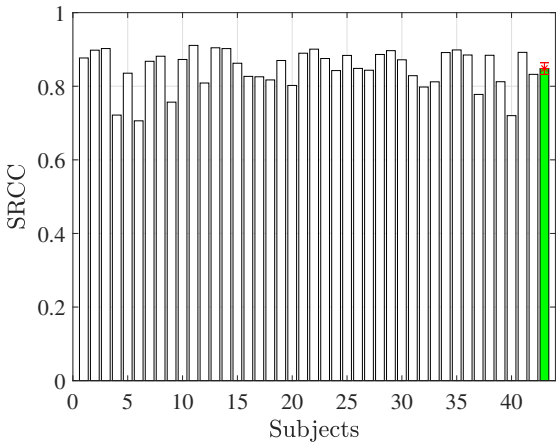


Fig. 3: SRCC between MOS and individual subject scores. The right-most bar shows average subject performance with error bar.

suggested in [12], which leads to the rejection of nine subjects. The mean of the Z-scores ($\hat{z_{ij}}$) of the remaining subjects ($N = 42$) for each video j is computed which leads to the MOS in the Z domain (MOS_{z_j}), given as:

$$MOS_{z_j} = \frac{1}{N}\sum_{i=1}^{N} \hat{z_{ij}} \tag{2}$$

The vector of all MOS_{z_j} values (MOS_z) has the range [-2.27, 1.43] and needs to be rescaled. Although minmax normalization has been used to perform such rescaling [13, 14], we avoid using this technique since it can alter the distribution of data. Instead we use the following approach to generate the final MOS:

$$MOS = \sigma_{rmos}\left[\frac{MOS_z - \mu_{zmos}}{\sigma_{zmos}}\right] + \mu_{rmos} \tag{3}$$

where μ_{zmos} and σ_{zmos} are respectively the mean and standard deviation of MOS_z, whereas μ_{rmos} and σ_{rmos} are respectively the mean and standard deviation of the Mean Opinion Scores obtained from the raw subjective ratings. To evaluate the effectiveness of the final MOS, we compute its correlation with individual subjects' scores. Fig. 3 shows the Spearman Rank Correlation Coefficient (SRCC) of each valid subject with respect to MOS, where the right-most column shows the performance of an average subject with error bar. It can be observed that there is a good degree of agreement between individual subjects and MOS.

4. PERFORMANCE OF OBJECTIVE VQA MODELS

We tested the performance of representative VQA methods on the Waterloo UHD-HDR-WCG database. These include the FR methods: DSS [15], ESSIM [16], FSIM [17], GMSD [18], GSIM [19], HDRVDP2 [20], HDRVQM [7], IWSSIM [21], PSNR, SRSIM [22], and VIFDWT [23], and NR methods: BRISQUE [24], CORNIA [25], dipIQ [26], HOSA [27], LPSI [28], NIQE [29], and VMEON [30]. All but HDRVQM and VMEON are designed for objective Image Quality Assessment (IQA). These methods are applied to videos in a frame-by-frame manner and a final quality score is obtained by averaging across all frames. Among these methods, only HDRVDP2 and HDRVQM are designed specifically for HDR content, whereas all other methods have been designed and validated for Low Dynamic Range (LDR) content. PQ-encoding was substituted for the mapping used by HDRVQM (PU-encoding) to convert the linear light data into a perceptually uniform space. While PQ is one of the specifications recommended by ITU for mapping HDR data [31], PU is an older mapping that is not included in the recommendation. The performance of these methods was evaluated by using three evaluation metrics: Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE) to assess prediction accuracy and SRCC to assess prediction monotonicity [32]. A five-parameter logistic function [33] was used to perform non-linear mapping of objective scores to MOS before the computation of PLCC and RMSE. A better objective method should have higher PLCC and SRCC, and lower RMSE values. Table 1 shows the database-wide performance of the objective methods in terms of the three evaluation metrics. To draw statistically sound inferences about the performance of these methods, we carried out hypothesis testing on model prediction residuals (after non-linear mapping). Through the Jarque-Bera test [34] at the 5% significance level, we determined that the prediction residuals of all methods (except PSNR) are likely normally distributed. This enabled us to compare the model residuals through statistical significance testing by using the F-test [35]. The results are shown in Table 2, where "1", "–", or "0" mean that the method in the row is statistically (with 95% confidence) better, indistinguishable, or worse than the method in the column respectively.

Table 1: Performance of Quality Assessment Algorithms. Best performing results in each category are in bold.

Category	Method	PLCC	SRCC	RMSE
FR	DSS [15]	0.7685	0.7456	12.3718
	ESSIM [16]	0.8512	0.8389	10.1485
	FSIM [17]	0.8693	0.8564	9.5568
	GMSD [18]	0.7366	0.7045	13.0781
	GSIM [19]	0.8596	0.8453	9.8812
	HDRVDP2 [20]	0.7035	0.6703	13.7423
	HDRVQM [7]	0.7783	0.7759	12.1428
	IWSSIM [21]	0.8088	0.7861	11.3730
	PSNR	0.5113	0.4615	16.6185
	SRSIM [22]	**0.8726**	**0.8630**	**9.4462**
	VIFDWT [23]	0.6809	0.6748	14.1612
NR	BRISQUE [24]	0.3622	0.3271	18.0241
	CORNIA [25]	**0.6497**	**0.6296**	**14.7003**
	dipIQ [26]	0.6192	0.5560	15.1845
	HOSA [27]	0.5379	0.5138	16.3015
	LPSI [28]	0.3941	0.3820	17.7718
	NIQE [29]	0.5286	0.4922	16.4152
	VMEON [30]	0.5776	0.5308	15.7845

The LDR FR method SRSIM is found to be the top performer in terms of PLCC, SRCC and RMSE (Table 1). Other high performing FR methods include ESSIM, GSIM, and FSIM, where it can be seen from Table 2 that their performance is statistically indistinguishable from SRSIM. All of them inherit a similar formulation of signal fidelity measurement from SSIM [36]. Somewhat surprisingly, the HDR specific FR methods HDRVDP2 and HDRVQM do not offer superior performance on the Waterloo UHD-HDR-WCG database. This analysis suggests that LDR FR methods may be extended for HDR VQA, at least as far as compression is concerned, and potential further enhancement is possible by making HDR specific adjustments. Tables 1 and 2 also indicate that all NR methods under testing perform rather inadequately. All these methods were designed for LDR content and most of them required some form of

Table 2: Statistical Significance Testing results for competing objective models on the Waterloo UHD-HDR-WCG database. A "1", "–", or "0" means that the method in the row is statistically (with 95% confidence) better, indistinguishable, or worse than the method in the column respectively. Legend: BRISQUE (m1), LPSI (m2), PSNR (m3), NIQE (m4), HOSA (m5), VMEON (m6), dipIQ (m7), CORNIA (m8), VIFDWT (m9), HDRVDP2 (m10), GMSD (m11), DSS (m12), HDRVQM (m13), IWSSIM (m14), ESSIM (m15), GSIM (m16), FSIM (m17), SRSIM (m18). FR methods are marked in **bold** while NR methods are marked in *italic*.

	m1	*m2*	**m3**	*m4*	*m5*	*m6*	*m7*	*m8*	**m9**	**m10**	**m11**	**m12**	**m13**	**m14**	**m15**	**m16**	**m17**	**m18**
m1	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0
m2	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0
m3	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0
m4	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0
m5	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0
m6	–	–	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0
m7	1	1	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0
m8	1	1	–	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0
m9	1	1	1	1	1	–	–	–	–	–	–	–	0	0	0	0	0	0
m10	1	1	1	1	1	–	–	–	–	–	–	–	–	0	0	0	0	0
m11	1	1	1	1	1	1	1	–	–	–	–	–	–	–	0	0	0	0
m12	1	1	1	1	1	1	1	1	–	–	–	–	–	–	0	0	0	0
m13	1	1	1	1	1	1	1	1	1	–	–	–	–	–	0	0	0	0
m14	1	1	1	1	1	1	1	1	1	1	–	–	–	–	–	0	0	0
m15	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–	–	–
m16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–	–
m17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–	–
m18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–	–

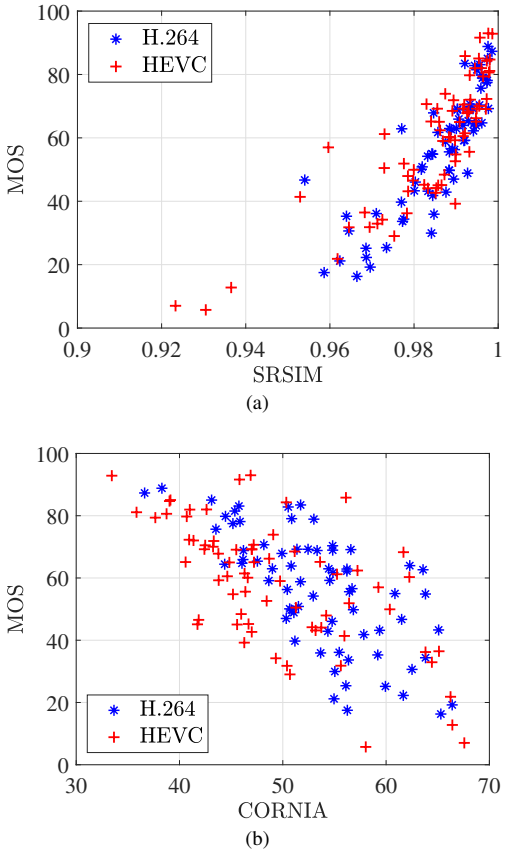


Fig. 4: Scatter plots of best performing FR method SRSIM (a) and NR method CORNIA (b).

training, that was also done on LDR content. Clearly, there is significant room for improvement in HDR specific design innovations.

Similar performance evaluation results on FR- and NR-VQA models are observed when H.264 and HEVC compressed videos are evaluated separately. Fig. 4 shows the scatter plots for the top performing FR and NR methods, where the H.264 and HEVC compressed videos are separately identified.

5. CONCLUSION

In this work, we have constructed a first-of-its-kind Waterloo UHD-HDR-WCG database composed of PQ-encoded UHD-HDR-WCG content compressed by H.264 and HEVC encoders in a content adaptive manner. We have carried out a subjective study on a professional Canon DP-V2420 4K/UHD HDR Reference Display. To the best of our knowledge, such an endeavor has not been attempted before, and no database of its kind is available to the research community. We have also proposed a novel method to process subjective data into MOS that accounts for subject quality scale variations while keeping the overall mean and standard deviation of subjective scores unchanged. Finally, we have evaluated the performance of eleven FR and seven NR representative objective quality assessment methods on the new database. Our analysis indicates that FR methods developed for LDR content are promising to serve as the basis for the development of highly effective FR-VQA models for UHD-HDR-WCG videos. On the other hand, there is substantial room for improvement when it comes to NR-VQA of UHD-HDR-WCG content.

6. REFERENCES

[1] A. Banitalebi-Dehkordi, M. Azimi, M. T. Pourazad, and P. Nasiopoulos, "Compression of High Dynamic Range Video using the HEVC and H.264/AVC Standards," in *Int. Conf. Het. Netw. Quality, Rel., Security, Robustness*, Aug. 2014, pp. 8–12.

[2] M. Narwaria, M. P. Da Silva, and P. Le Callet, "Study of High Dynamic Range Video Quality Assessment," in *Proc. SPIE Opt. Eng. Appl.*, Sept. 2015, vol. 9599, pp. 95990V:1–13.

[3] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective and Objective Evaluation of HDR Video Compression," in *Int. Workshop Video Process., Quality Metrics Consum. Electron. (VPQM)*, 2015.

[4] K. Minoo, Z. Gu, D. Baylon, and A. Luthra, "On metrics for objective and subjective evaluation of high dynamic range video," in *Proc. SPIE Opt. Eng. Appl.*, Sept. 2015, vol. 9599, pp. 95990F:1–14.

[5] R. Mukherjee, K. Debattista, T. Bashford-Rogers, P. Vangorp, R. Mantiuk, M. Bessa, B. Waterfield, and A. Chalmers, "Objective and subjective evaluation of High Dynamic Range video compression," *Signal Process.: Image Commun.*, vol. 47, pp. 426–437, 2016.

[6] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the Performance of Existing Full-Reference Quality Metrics on High Dynamic Range (HDR) Video Content," *arXiv preprint arXiv:1803.04815*, 2018.

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

[8] A. Rehman, K. Zeng, and Z. Wang, "Display Device-Adapted Video Quality-of-Experience Assessment," in *Proc. SPIE Electron. Imag.*, Mar. 2015, vol. 9394, p. 939406:1–11.

[9] "Canon DP-V2420 Reference Display product page," https://www.usa.canon.com/internet/portal/us/home/products/details/reference-displays/4k-uhd-reference-displays/dp-v2420.

[10] The Academy of Motion Picture Arts and Sciences, "Academy Color Encoding System (ACES)," https://www.oscars.org/science-technology/sci-tech-projects/aces.

[11] Blackmagicdesign, "Product Technical Specifications: UltraStudio 4K Extreme 3," https://www.blackmagicdesign.com/ca/products/ultrastudio/techspecs/W-DLUS-09.

[12] Rec. ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," Jan. 2012.

[13] W. Sun, F. Zhou, and Q. Liao, "MDID: A Multiply Distorted Image Database for Image Quality Assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, 2017.

[14] A. Zarić, N. Tatalović, N. Brajković, H. Hlevnjak, M. Lončarić, E. Dumić, and S. Grgić, "VCL@ FER Image Quality Assessment Database," *Automatika*, vol. 53, no. 4, pp. 344–354, 2012.

[15] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, "Image quality assessment based on DCT subband similarity," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sept. 2015, pp. 2105–2109.

[16] X. Zhang, X. Feng, W. Wang, and W. Xue, "Edge Strength Similarity for Image Quality Assessment," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 319–322, Apr. 2013.

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

[18] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[19] A. Liu, W. Lin, and M. Narwaria, "Image Quality Assessment Based on Gradient Similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[20] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graphics*, vol. 30, no. 4, pp. 40:1–40:14, 2011.

[21] Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[22] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sept. 2012, pp. 1473–1476.

[23] S. Rezazadeh and S. Coulombe, "A novel discrete wavelet transform framework for full reference image quality assessment," *Signal, Image, Video Process. (SIViP)*, vol. 7, no. 3, pp. 559–573, 2013.

[24] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[25] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2012, pp. 1098–1105.

[26] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[27] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind Image Quality Assessment Based on High Order Statistics Aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sept. 2016.

[28] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sept. 2015, pp. 339–343.

[29] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[30] W. Liu, Z. Duanmu, and Z. Wang, "End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks," in *ACM Int. Conf. Multimedia*, 2018, pp. 546–554.

[31] Rec. ITU-R BT.2100-2, "Image parameter values for high dynamic range television for use in production and international programme exchange," July 2018.

[32] Video Quality Experts Group, "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment," Mar. 2000.

[33] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[34] C. M. Jarque and A. K. Bera, "A Test for Normality of Observations and Regression Residuals," *Int. Statist. Rev.*, vol. 55, no. 2, pp. 163–172, Aug. 1987.

[35] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, 2011.

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