Performance Evaluation of Objective Image Quality Metrics on Conventional and Learning-Based Compression Artifacts

Michela Testolina
Multimedia Signal Processing Group
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
michela.testolina@epfl.ch

Evgeniy Upenik
Multimedia Signal Processing Group
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
evgeniy.upenik@epfl.ch

João Ascenso
Instituto Superior Técnico, Universidade de Lisboa - Instituto de Telecomunicações
Lisbon, Portugal
joao.ascenso@lx.it.pt

Fernando Pereira
Instituto Superior Técnico, Universidade de Lisboa - Instituto de Telecomunicações
Lisbon, Portugal
fp@lx.it.pt

Touradj Ebrahimi
Multimedia Signal Processing Group
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
touradj.ebrahimi@epfl.ch

***Abstract*—Lossy image compression is a popular, simple and effective solution to reduce the amount of data representing digital pictures. In most lossy compression methods, the reduced volume of data in bits is achieved at the expense of introducing visual artifacts in the picture. The perceptual quality impact of such artifacts can be assessed with expensive and time-consuming subjective image quality experiments or through objective image quality metrics. However, the faster and less resource demanding objective quality metrics are not always able to reliably predict the quality as perceived by human observers. In this paper, the performance of 14 objective image quality metrics is benchmarked against a dataset of compressed images labeled with their subjective quality scores. Moreover, the performance of the above objective quality metrics in predicting the subjective quality of images distorted by both conventional and learning-based lossy compression artifacts is assessed and conclusions are drawn.**

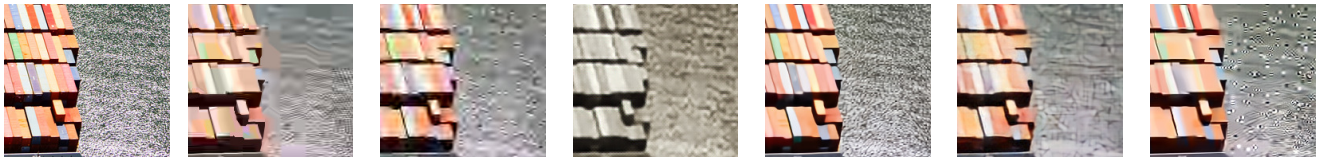
***Index Terms*—image compression, perceptual visual quality, objective quality metrics, objective-subjective correlation, deep learning**

I. Introduction

In the last decades, the number of images captured on a daily basis using digital devices has grown exponentially. For this reason, more efficient solutions in the field of image compression are constantly under research. Recently, learning-based image compression methods have demonstrated competitive, if not superior, compression performance when compared to conventional methods in terms of compression efficiency and perceived visual quality. The compression performance may be evaluated through expensive and costly subjective experiments, or objectively through metrics that aim at predicting the visual quality as perceived by a human observer. However, the objective quality metrics don't always match the human perception for a wide range of distortions.

In the past, multiple studies have been conducted to evaluate the correlation performance of objective image quality metrics with the corresponding subjective quality scores. As an example, Zhang et al. [1] proposed an extensive correlation study on 7 different subjective quality datasets, i.e., image datasets with subjective quality scores, and 11 full-reference objective quality metrics. In [2], Ma et al. proposed a new subjective quality dataset, which was used for assessing the correlation of 20 image quality metrics, both full-reference and non-reference. Furthermore, Hanhart et al. [3] proposed a correlation experiment for High Dynamic Range (HDR) images. Recently, Valenzise et al. [4] evaluated both subjectively and objectively the performance of two deep-learning based compression methods, observing that the objective metrics are, in general, less accurate in predicting the quality of learning-based methods compared to JPEG 2000 or Better Portable Graphics (BPG). However, a comprehensive study on the objective-subjective correlation performance between conventional and learning-based compression approaches has not been conducted yet. Also, the state-of-the-art lacks a study addressing a wide range of objective quality metrics and a large number of learning-based codecs.

In this paper, the performance of several objective image quality metrics is assessed in the context of lossy image compression, including conventional as well as multiple emerging learning-based image codecs. The objective-subjective correlation performance of 14 full-reference objective quality metrics is assessed based on a subjective study including 8 test images and 8 image codecs, both conventional and learning-based, at multiple bitrates. In fact, these two image compression approaches introduce different types of visual artifacts/distortions, and therefore it is important to assess



(a) Original (b) HEVC (c) JPEG 2000 (d) LB 1 (e) LB 4 (f) LB 5 (g) LB 6

Fig. 1: Visual comparison of conventional and learning-based (LB) compression artifacts in the image "port" at 0.12bpp.

how the objective quality metrics perform against the human perception, expressed by subjective quality scores, for each type of compression solution and associated artifacts. The reported subjective experiment follows a *crowdsourcing* approach and was organized in the context of the JPEG AI Call for Evidence, co-organized with the IEEE MMSP'2020 Challenge on Learning-Based Image Coding. To the best of the authors' knowledge, no other objective-subjective correlation experiments have been conducted on an image compression specific dataset with such a large variety of objective quality metrics and codecs, including both conventional and learning-based compression artifacts, and therefore on a wide set of possible degradation. Moreover, only a few studies in the field have been conducted utilizing subjective visual quality scores collected through a *crowdsourcing* subjective experiment, and therefore including a large variety of subjects and viewing conditions. The conclusions and insights will be useful for future objective image quality assessment efforts, as they will highlight which objective quality metrics perform better with learning-based image codecs, and the main strengths and limitations of the available objective quality metrics for different types of lossy image compression.

II. Subjective Quality Assessment Experiment

The subjective quality assessment experiment has been conducted on a novel image dataset that includes subjective quality scores, which targeted the evaluation of the compression efficiency performance of both conventional and learning-based lossy image compression methods.

Test Material: The dataset includes 8 uncompressed images with different contents including faces, signs, buildings, natural scenery, repetitive patterns and artificially generated images. The raw images were compressed with eight different compression methods, two conventional and six learning-based, at four different target bitrates.

Image Codecs: The conventional codecs include HEVC Intra and JPEG 2000 visually optimized, while the six learning-based codecs correspond to the submissions made to the JPEG AI Call for Evidence in conjunction with the IEEE MMSP'2020 Challenge on Learning-Based Image Coding. In particular, five compression methods use a novel end-to-end approach based on convolutional neural networks, and another is a hybrid codec extending the Versatile Video Coding (VVC Intra) [5] standard with deep learning-based tools. Since not all the teams participating in the JPEG AI Call for Evidence/MMSP'2020 Challenge agreed to reveal the information about their compression method, only the details about the accessible compression methods are mentioned in this paper [6]–[9].

Distortion analysis: The dataset includes both conventional and learning-based image compression artifacts. In particular, HEVC is known for introducing blocking artifacts, blurriness and color bleeding, while JPEG 2000 mainly introduces blurriness and ringing artifacts. On the new learning-based codecs, it is possible to observe that the reconstruction of the colors is not always accurate, and in few cases, the color component is completely lost at the lowest bitrates. Furthermore, the learning-based approaches also introduce blurring artifacts and degradation of the texture areas, as well as blocking artifacts in some cases. Some visual examples of compression artifacts are shown in Figure 1.

Subjective Quality Assessment Protocol: The subjective scores were collected through a Double Stimulus Continuous Quality Scale (DSCQS) experiment [10] with hidden references, in which the distorted/decoded and original images were presented side-by-side and subjects were asked to rate the visual quality of both stimuli on a continuous scale between 0 and 100. The choice of this quality assessment protocol derives from the fact that learning-based compression methods may improve the quality of decoded images, especially at higher bitrates, and thus, might exhibit a better visual subjective quality compared to the original stimulus. As not all the codecs achieved the target bitrates, the dataset includes, in total, 240 compressed images with the respective subjective quality scores.

Experiment setup: Due to the COVID-19 pandemic, the subjective assessment experiment was conducted through *crowdsourcing* on Amazon Mechanical Turk (MTurk). A total of 118 naïve subjects took part in the experiment but only 116 successfully completed the evaluation. Among them, 32 were females and 84 males, with age between 18 and 70, and mean age 34.72. The subjects took part remotely in the experiments from 10 different countries, where the United States, India, and Brazil registered the highest number of participants. All the subjects had a monitor with 1920x1080 resolution or higher, with the HiDPI/Retina mode disabled. The original images were cut into tiles of size 945×880 to fit the screen side-by-side. No specific instructions about the distance to the screen were given to the subjects.

Subjective Data Processing: After collecting the subjective scores, two outliers were detected using the methodology specified in ITU Rec. BT-500 [10], and therefore their scores

were discarded. Then, the Mean Opinion Score (MOS) was computed among the valid subjective scores. Subsequently, the Differential Mean Opinion Score (DMOS) has been computed from the MOS of the source reference (SR) and processed stimuli (PS).

The subjective quality scores collected through this procedure were correlated to the objective quality scores for several metrics, following the procedure described in the next section.

III. Objective Quality Metrics and Data Processing

A. Full-Reference Objective Quality Metrics

Although subjective image quality assessment is more reliable, it is also more costly and time-consuming, and in many scenarios, for example, perceptual quality optimizations in image compression, it is not even feasible. To avoid subjective image quality assessment, several objective image quality metrics have been proposed over the years. In particular, three main approaches to this problem are possible: full-reference, reduced-reference and no-reference. In the context of image compression, the full-reference approach is the most popular as it expresses the fidelity of the compression process; however, it requires both the original and the decoded images to be simultaneously available.

One of the most popular full-reference objective image quality metrics is the Peak Signal to Noise Ratio (PSNR), which measures the mathematical dissimilarity between two images through the Mean Squared Error (MSE) between the original and the degraded/decoded images. A more advanced version of the PSNR, computed in the DCT domain, is the PSNR-HVS-M [11], which considers the characteristics of the Human Visual System (HVS) and it was shown in the past to correlate better with human perception than the standard PSNR.

A turning point in the development of objective image quality metrics was the Structural Similarity Index Measure (SSIM) [12], inspired by the HVS and claimed as a better solution than PSNR. The objective quality score is computed through a combination of three different comparisons related to luminance, contrast and structure. Successively, different variants of this metric were proposed, notably, the Multi-Scale Structural Similarity Index (MS-SSIM) [13], which assesses the SSIM at multiple resolutions and viewing conditions, and the Information Content Weighted Structural Similarity Measure (IW-SSIM) [14], which weights the SSIM by perceptual coefficients, proportional to the amount of local information in the image.

Many other objective image quality metrics have also been proposed over the years. Among them, the Visual Information Fidelity (VIFp) [15] measures the loss of information between a reference and a distorted image, taking into account the natural image statistics and the HVS characteristics. This metric was adopted as part of the Video Multi-Method Assessment Fusion (VMAF) metric [16], an objective quality metric proposed by Netflix. Although this metric was designed for video quality, it can also be computed for single frames/images. The intra-frame quality estimation fuses VIFp with the so-called Detail Loss Metric (DLM), which estimates the loss of detail in the decoded image. These two metrics are then combined through a Support Vector Machine (SVM) regressor trained on subjective quality data to obtain the final quality score.

Metric	Ref	Color space	Remarks
CIEDE2000	[20]	Lab	The final score corresponds to the average difference for all pixels in the image.
FID	[21]	RGB	The default inception model was used.
FSIM	[18]	RGB	The color version of the metric was used.
IW-SSIM	[14]	Y	The default parameters were used.
LPIPS	[22]	RGB	The pre-trained model was used.
MS-SSIM	[13]	Y	The default parameters were used.
NLPD	[19]	Y	The default parameters and the output scores in the normalized domain were used.
PSNR	-	Y	The Matlab version was used.
PSNR-HVS-M	[11]	Y	The default window size was used.
SSIM	[12]	Y	The default parameters were used.
VDP2	[17]	RGB	The LDR mode was used.
VIFp	[15]	Y	The default parameters were used.
VMAF	[16]	YUV	*libvmaf* version V.2.1.1 was used.
WaDIQaM	[23]	RGB	The weighted approach pre-trained on both the LIVE and TID2013 datasets was used.

TABLE I: Summary of the objective image quality metrics benchmarked in this paper.

Other popular objective image quality metrics are the HDR-VDP-2 (also know as VDP2) [17], designed to be robust to different lighting conditions, the Feature-Similarity Index Metric (FSIM) [18], which assesses the quality through the Phase Congruency (PC) and Gradient Magnitude (GM), the Normalized Laplacian Pyramid Distance (NLPD) metric [19], which decomposes images using the Laplacian pyramid, and CIEDE2000 [20], which computes the difference between two colors in the CIELab color space.

Recently, a few learning-based objective image quality metrics have been developed. As an example, the Weighted Average Deep Image QuAlity Measure (WaDIQaM) [23] is an end-to-end deep neural network metric able to assess the visual quality of images both with full-reference and no-reference approaches. This metric computes the score in a patch-wise fashion, with the final score being a weighted combination of the patches' scores considering their saliency. Another learning-based approach to objective image quality assessment is the Learned Perceptual Image Patch Similarity (LPIPS) [22], where the perceptual similarity is measured between two images using deep neural network activations, in this case using a VGG (Visual Geometry Group) neural network trained for image classification on ImageNet. Moreover, the Fréchet Inception Distance (FID) [21] was proposed to assess the quality of images generated with Generative Adversarial Networks (GANs).

In this paper, the objective-subjective correlation perfor-

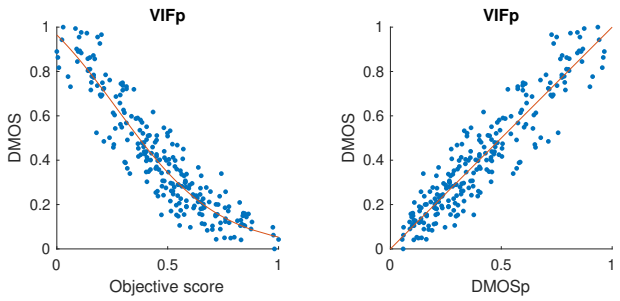


Fig. 2: DMOS versus VIFp scores (left) and DMOSp (right) scatter plots.

mance of all the objective image quality metrics introduced in this section will be evaluated against the subjective dataset presented in Section II. Table I summarizes the studied objective image quality metrics.

B. Color Transformations

Since not all the objective image quality metrics were designed to work in the RGB color space, some color transformations are sometimes necessary. Table I indicates the color spaces in which the quality metrics are assessed in this paper. In particular, the PSNR, SSIM, MS-SSIM, IW-SSIM, VIFp, NLPD and PSNR-HSV-M are designed to work with gray-scale images, so the input RGB images were used to obtain the luminance component following ITU-R Rec. BT.709-6 [24]. Y was computed as a weighted combination of the red, green and blue channels as:

$$Y = 0.2126 * R + 0.7152 * G + 0.0722 * B \qquad (1)$$

Moreover, for VMAF, the images were converted to the YUV color space. To do so, the *FFmpeg* libraries were used to convert the images to the *yuv444* format.

Lastly, since CIEDE2000 computes the color difference in the CIELab color space, the images have been converted into this space using the Matlab function `rgb2lab`.

C. Statistical Analysis

The performance of the objective image quality metrics in Table I has been evaluated through their objective-subjective correlation scores. This was performed by computing the Pearson Linear Correlation Coefficient (PLCC), Spearman's Rank Correlation Coefficient (SROCC) and Kendall's Rank Correlation Coefficient (KROCC), which are different methods to assess the relationship between two variables.

While the subjective scores usually have a non-linear behaviour, PLCC looks for a linear correlation. For this reason, a least-squares regression procedure was applied to the data, following the method proposed in the ITU Tutorial [25]. To remove the non-linearity, a non-linear regression was fitted to the objective quality scores through the logistic function proposed in [26]:

$$DMOS_p = \frac{\beta_1}{1 + \exp(-\beta_2 * (IQS - \beta_3))} \qquad (2)$$

Where IQS represents the image quality scores obtained from the objective image quality metrics. The non-linear regression was performed using the Matlab function `nlinfit`, while the final score prediction was done through the Matlab function `nlpredci`. As an example, Figure 2 shows the scatter plots of $DMOS$ against the VIFp scores and the $DMOS_p$ (predicted DMOS) data obtained with the process described above.

It is important to underline that, while the PLCC looks for a linear correlation, the SROCC and the KROCC are based on the ranking of the data, and therefore are not influenced by the fitting procedure.

The correlation scores presented above have values in the range ±1, where +1 denotes perfect positive correlation, -1 perfect negative correlation and 0 no correlation. Since we are not interested in the trend of the correlation, only its absolute value will be considered.

Besides the correlation scores, the objective quality metrics performance against the subjective quality scores can also be assessed with two other metrics, i.e., the Root Mean Square Error (RMSE) and the Outlier Ratio (OR). To compute them, the guidelines in ITU-T Rec. P.1401 [27] were followed. Contrary to the previous correlations measures, a lower RMSE or OR value expresses a better performance by the metrics.

IV. Performance and Analysis

The PLCC, SROCC, KROCC, RMSE and OR scores are computed between the subjective DMOS and predicted DMOSp scores, to identify the best performing objective image quality metrics. To compare the performance on the different compression artifacts, the subjective dataset was split into two subsets: the first contains only the images coded with the conventional codecs, i.e., HEVC Intra and JPEG 2000, while the second subset includes only the images coded with the learning-based codecs. The two types of compression artifacts are, as previously explained, different from each other, and this division allows to assess the objective image quality metrics for both cases. Thus, the PLCC, SROCC, KROCC, RMSE and OR scores are separately computed for the conventional and learning-based (LB) subsets, as well as for the entire dataset.

Table II shows the PLCC, SROCC, KROCC correlation results for the two subsets and the entire dataset, and Table III shows the results for the RMSE and OR metrics. The results show that the quality metrics with the highest correlation scores and the lowest RMSE and OR are VMAF, VIFp, PSNR-HVS-M and FSIM. Moreover, the following observations can be formulated:

- Among the best correlating metrics, VMAF and FSIM consider the color information, while VIFp and PSNR-HVS-M are computed only for the luminance component. The color components do not seem to impact much the overall objective image quality metrics performance.
- The metrics with the lowest correlation scores are WaDIQaM and CIEDE2000. A possible explanation for

TABLE II: PLCC, SROCC and KROOC correlation scores for the benchmarked objective image quality metrics.

Metrics	PLCC			SROCC			KROCC		
	Conventional	***LB***	***Overall***	***Conventional***	***LB***	***Overall***	***Conventional***	***LB***	***Overall***
CIEDE2000	0.4973	0.6698	0.6251	0.4785	0.6021	0.5720	0.3373	0.4409	0.4143
FID	0.8029	0.8395	0.8281	0.7964	0.7922	0.7967	0.6012	0.5999	0.6001
FSIM	**0.9283**	**0.9177**	**0.9192**	**0.9004**	**0.8955**	**0.8992**	**0.7341**	**0.7177**	**0.7204**
IW-SSIM	0.8719	0.8574	0.8611	0.8427	0.8314	0.8359	0.6637	0.6428	0.6487
LPIPS	0.8292	0.8232	0.8243	0.7888	0.7570	0.7660	0.6011	0.5773	0.5850
MS-SSIM	0.8719	0.8574	0.8611	0.8427	0.8314	0.8359	0.6637	0.6428	0.6487
NLPD	0.7861	0.6807	0.7007	0.7825	0.6861	0.7090	0.5784	0.4962	0.5147
PSNR	0.8473	0.8062	0.8080	0.8697	0.8070	0.8168	0.6766	0.6123	0.6194
PSNR-HVS-M	**0.9445**	**0.9162**	**0.9137**	**0.9381**	**0.9000**	**0.9028**	**0.7927**	**0.7256**	**0.7276**
SSIM	0.8156	0.7810	0.7890	0.7959	0.7420	0.7569	0.6181	0.5609	0.5755
VDP2	0.8823	0.8306	0.8370	0.8739	0.8120	0.8209	0.6865	0.6134	0.6192
VIFp	**0.9460**	**0.9269**	**0.9299**	**0.9346**	**0.9171**	**0.9208**	**0.7936**	**0.7495**	**0.7534**
VMAF	**0.9522**	**0.9046**	**0.9087**	**0.9275**	**0.8746**	**0.8826**	**0.7778**	**0.7042**	**0.7065**
WaD LW	0.7157	0.5832	0.6087	0.7169	0.6367	0.6531	0.5079	0.4658	0.4722
WaD TW	0.8029	0.7442	0.7461	0.8122	0.7714	0.7708	0.6091	0.5831	0.5735

TABLE III: RMSE and OR scores for the benchmarked objective image quality metrics.

	RMSE			OR		
	Conventional	***LB***	***Overall***	***Conventional***	***LB***	***Overall***
CIEDE2000	0.2201	0.1860	0.1958	0.8437	0.7898	0.8083
FID	0.1513	0.1361	0.1407	0.8125	0.7898	0.8000
FSIM	**0.0944**	**0.0995**	**0.0988**	**0.6250**	**0.6591**	**0.6708**
IW-SSIM	0.1243	0.1289	0.1276	0.6562	0.7273	0.7125
LPIPS	0.1418	0.1422	0.1420	0.7969	0.7443	0.7583
MS-SSIM	0.1242	0.1290	0.1276	0.6562	0.7273	0.7125
NLPD	0.1569	0.1835	0.1790	0.8125	0.7500	0.7583
PSNR	0.1424	0.1392	0.1418	0.7187	0.7159	0.7208
PSNR-HVS-M	**0.0834**	**0.1004**	**0.1020**	**0.5312**	**0.6023**	**0.6375**
SSIM	0.1472	0.1568	0.1545	0.7656	0.7273	0.7375
VDP2	0.1194	0.1395	0.1373	0.6562	0.7614	0.7583
VIFp	**0.0823**	**0.0940**	**0.0923**	**0.4375**	**0.6250**	**0.5958**
VMAF	**0.0775**	**0.1068**	**0.1047**	**0.6094**	**0.5739**	**0.6375**
WaD LW	0.1772	0.2044	0.1996	0.7969	0.7954	0.8125
WaD TW	0.1512	0.1673	0.1670	0.7812	0.7216	0.7583

this might be the fact that WaDIQaM, a deep learning-based objective image quality metric, was trained on datasets including only a small percentage of conventional compression artifacts and no learning-based compression artifacts at all. Regarding CIEDE2000, this metric does not include any information about the image structure and only considers the difference between the colors.

- As expected, the PSNR appears in the lower range of the correlation scores ranking, since simple pixel-wise error measures do not correlate very well with human perception. However, the PSNR-HVS-M variant, which considers the distortion visibility in the frequency domain and thus some HVS characteristics, is well correlated with the subjective mean opinion scores.
- Surprisingly, MS-SSIM and IW-SSIM do not show outstanding correlation performance as they are in the middle of the ranking even for the conventional coding subset. This finding is in contrast with what was observed in [1] and [2], and could be caused by the fact that the subjective experiment reported in this paper has been performed with a crowdsourcing approach, thus closer to the way humans consume images nowadays.

From Table III, it is possible to observe that, on average, the PLCC, SROCC and KROCC correlation scores are lower, while the RMSE and OR scores are higher on the learning-based compression subset compared to the conventional compression subset. This confirms the hypothesis that all quality metrics, excluding the CIEDE2000, perform better on the conventional subset. In fact, most of the reviewed objective image quality metrics were designed before the emergence of learning-based codecs, and therefore were naturally not designed taking into account the types of compression artifacts the latter exhibit. This conclusion, however, doesn't hold in the case of LPIPS and FID: they were, in fact, designed to specifically deal with images generated with learning-based approaches, so they exhibit comparable or, in some cases, slightly higher performance on the learning-based subset compared to the conventional one. Anyway, these metrics don't present remarkable performance in the global ranking of the quality metrics.

In general, it may be observed that learning-based codecs lack of an accurate reconstruction of the colors, and that CIEDE2000, i.e., a metric specifically designed to evaluate

the color difference between two images, performs better on the learning-based subset rather than on the classical one. This might suggest that incorporating a more accurate evaluation of the colors on objective metrics might lead to an improvement in the objective-subjective correlation results. This observation might be taken into account in the design of a future objective image quality assessment metric able to deal with a wider range of distortions.

V. Conclusions

In this paper, the correlation performance of 14 full-reference objective image quality metrics was assessed on a subjective dataset labeled with a crowdsourcing-based subjective quality assessment experiment targeting at assessing the image compression performance of both conventional and learning-based compression approaches. It was found that the quality metrics correlating better with the subjective quality scores are VIFp, VMAF, FSIM and PSNR-HSV-M. Moreover, the correlation scores show that most objective quality metrics perform better for the conventional codecs than for the learning-based codecs, as expected. This observation motivates further research to seek better objective image quality metrics to more reliably predict the quality of images compressed with learning-based codecs, possibly introducing a better evaluation of the color accuracy.

Acknowledgments

EPFL affiliated authors would like to acknowledge support from the Swiss National Scientific Research project entitled "Advanced Visual Representation and Coding in Augmented and Virtual Reality" under grant number 200021_178854.

References

[1] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 1477–1480.

[2] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2016.

[3] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for hdr image quality assessment," *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[4] G. Valenzise, A. Purica, V. Hulusic, and M. Cagnazzo, "Quality assessment of deep-learning-based image compression," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2018, pp. 1–6.

[5] "ISO/IEC 23090-3:2021 Information technology - Coded representation of immersive media - Part 3: Versatile video coding," 2021.

[6] "ISO/IEC JTC 1/SC29/WG1 M89087 NJU-VISION Response to JPEG AI Call for Evidence," 2020.

[7] W.-C. Lee, C.-P. Chang, W.-H. Peng, and H.-M. Hang, "A hybrid layered image compressor with deep-learning technique," in *IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020.

[8] J. Lin, M. Akbari, H. Fu, Q. Zhang, S. Wang, J. Liang, D. Liu, F. Liang, G. Zhang, and C. Tu, "Variable-rate multi-frequency image compression using modulated generalized octave convolution," in *IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020.

[9] N. Zou, H. Zhang, F. Cricri, H. R. Tavakoli, J. Lainema, M. Hannuksela, E. Aksu, and E. Rahtu, "L2c - learning to learn to compress," in *IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020.

[10] Recommendation ITU-R BT.500-14, "Methodologies for the subjective assessment of the quality of television images," *International Telecommunication Union*, 2019.

[11] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in *Proceedings of the third international workshop on video processing and quality metrics*, vol. 4, 2007.

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

[13] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[14] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on image processing*, vol. 20, no. 5, pp. 1185–1198, 2010.

[15] H. Sheik and A. Bovik, "A visual information fidelity measure for image quality assessment," *IEEE T. Img. Proc*, vol. 15, no. 2, pp. 430–444, 2006.

[16] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, 2016.

[17] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on graphics (TOG)*, vol. 30, no. 4, pp. 1–14, 2011.

[18] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[19] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized laplacian pyramid," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–6, 2016.

[20] G. Sharma, W. Wu, and E. N. Dalal, "The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, vol. 30, no. 1, pp. 21–30, 2005.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems (NIPS)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[23] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on image processing*, vol. 27, no. 1, pp. 206–219, 2017.

[24] Recommendation ITU-R BT.709-6, "Parameter values for the hdtv standards for production and international programme exchange," *International Telecommunication Union*, 2006.

[25] ITU-R Tutorial, "Objective perceptual assessment of video quality: Full reference television," *International Telecommunication Union*, 2004.

[26] ITU-T SG09, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii (fr-tv2)," *International Telecommunication Union*, 2003.

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