Large-Scale Crowdsourcing Subjective Quality Evaluation of Learning-Based Image Coding

Evgeniy Upenik*, Michela Testolina*, João Ascenso†, Fernando Pereira† and Touradj Ebrahimi*
*Multimedia Signal Processing Group - Ecole Polytechnique Fédérale de Lausanne
†Instituto de Telecomunicações - Instituto Superior Técnico
Email: *firstname.lastname@epfl.ch, †firstname.lastname@lx.it.pt

***Abstract*—Learning-based image codecs produce different compression artifacts, when compared to the blocking and blurring degradation introduced by conventional image codecs, such as JPEG, JPEG 2000 and HEIC. In this paper, a crowdsourcing based subjective quality evaluation procedure was used to benchmark a representative set of end-to-end deep learning-based image codecs submitted to the MMSP'2020 Grand Challenge on Learning-Based Image Coding and the JPEG AI Call for Evidence. For the first time, a double stimulus methodology with a continuous quality scale was applied to evaluate this type of image codecs. The subjective experiment is one of the largest ever reported including more than 240 pair-comparisons evaluated by 118 naïve subjects. The results of the benchmarking of learning-based image coding solutions against conventional codecs are organized in a dataset of differential mean opinion scores along with the stimuli and made publicly available.**

***Index Terms*—deep learning, image coding, learning-based compression, subjective evaluation, visual quality, crowdsourcing**

I. Introduction

Image coding has a long history that goes back to the eighties of the last century. Until recently, the successful image coding solutions were entirely based on human-engineered architectures. In the last five years, however, new and competitive end-to-end learning-based image coding solutions have emerged as summarized in [1] and can be found in [2]–[6]. Such image coding solutions apply a non-linear transform, in contrast to the classical approach, where a linear transform, such as the Discrete Cosine Transform (DCT) or the Discrete Wavelet Transform (DWT), is used. Typically, the non-linear transform is implemented through a series of neural network layers, with parameters that are learned from a large amount of data and using a training procedure guided by a suitable loss function. Moreover, end-to-end learning-based image codecs still include steps such as quantization and entropy coding, where the quantization levels or probability models may be learned.

Due to their different nature, lossy image coding methods based on deep learning architectures lead to novel visual quality degradation and artifacts. In this context, the available objective visual quality metrics may not represent well the human perception of these artifacts [7]. Therefore, subjective quality evaluation studies are crucial to design a reliable image quality framework that accounts for the new coding artifacts.

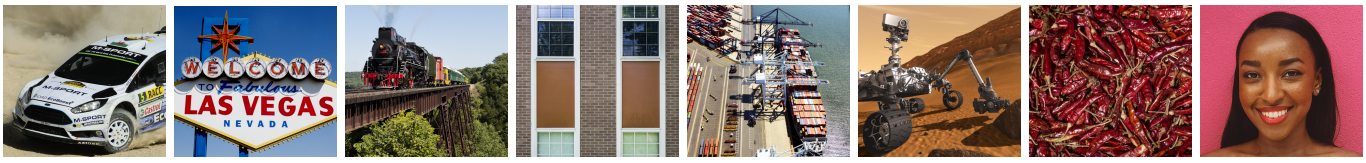
Nonetheless, only a few subjective studies on this topic can be found. In one of the first works [8], a subjective evaluation of two early learning-based image coding solutions is reported following a Double Stimulus Impairment Scale (DSIS) methodology. The main conclusion of this study is that the performance achieved by such early learning-based coding solutions is similar or sometimes better than JPEG 2000. In [9], a convolutional autoencoder optimized for MSE and MS-SSIM with two different tile arrangement strategies is compared to JPEG, JPEG 2000, and HEVC Intra. An Absolute Category Rating with Hidden Reference (ACR-HR) methodology was used, and conclusions show that the MS-SSIM optimization leads to larger perceptual quality gains in comparison to the MSE optimization. In this subjective study, HEVC Intra outperforms all the learning-based image coding configurations for many test images and target bitrates. In [10] a DSIS subjective quality assessment test is described, including learning-based image coding solutions using factorized or hyperprior entropy models to exploit the dependencies within the latent representation. The results show that learning-based image coding solutions have competitive compression efficiency in comparison to conventional image codecs such as HEVC Intra.

Finally, in the past studies several weaknesses have been observed: (1) a limited number of learning-based image codecs that do not represent well the recent advances; (2) the involvement of a small number of subjects and thus subjective judgements; (3) a limited variability of the image content; (4) limited or no public availability of the database; and (5) the lack a fine-grained, continuous scale rating. The issues above show the need for improved subjective quality evaluation studies involving emerging learning-based image codecs.

In this paper, we report the results and analysis of a large-scale subjective quality evaluation study following a double stimulus methodology, never used in the context of image quality assessment that is applied to recent learning-based image codecs. The obtained experimental data are organized in a dataset with corresponding stimuli and made publicly available[1] to the research community.

This work was funded by the H2020 Innovation Action "Advanced Mixed Realities (AdMiRe)" under grant agreement 952027 and the Swiss National Foundation for Scientific Research project "Advanced Visual Representation and Coding in Augmented and Virtual Reality" under grant number 200021_178854.

[1]https://www.epfl.ch/labs/mmspg/jpeg-ai-subjective-dataset/



(a) Racing car (b) Las Vegas sign (c) Train (d) Windows (e) Port (f) Curiosity rover (g) Bell pepper (h) Woman

Fig. 1. Cropped images used for the subjective quality evaluation experiment.

II. Subjective Quality Evaluation Platform: Crowdsourcing

Cambridge Business English Dictionary defines *crowdsourcing* as "the act of giving tasks to a large group of people or to the general public, for example, by asking for help on the internet, rather than having tasks done within a company by employees"[2]. A specific case of crowdsourcing is microtasking. It is characterized by an even larger number of persons performing the same simple task.

A. Crowdsourcing platform

As a software platform, the subjective experiment reported in this paper used the *QualityCrowd* framework [11] that was designed to perform subjective quality assessment with crowdsourcing. *QualityCrowd* allows codec independent quality assessment with a simple web interface, usable with common web browsers. For the purpose of the subjective quality evaluation, *QualityCrowd* was extended to be able to enforce screen size restrictions on the participants. The source code and instructions to run the modified software are publicly available[3].

The subjective quality evaluation experiment was performed with participants recruited from Amazon Mechanical Turk, a crowdsourcing website for businesses (called Requesters) to hire remotely located *crowdworkers* to perform discrete on-demand tasks that computers are currently unable to do. Employers post jobs known as Human Intelligence Tasks (HITs), e.g., identifying specific content in an image or video, writing product descriptions, or answering questions.

III. Subjective Quality Evaluation Experiment

Learning-based image codecs were subjectively evaluated for the first time using a crowdsourcing approach, involving more than 100 participants. The subjective evaluation experiment followed a Double Stimulus Continuous Quality Scale (DSCQS) methodology able to account for cases where the decoded image quality may be higher than the reference image quality (in theory not impossible with learning-based codecs).

A. Codecs and anchors

Six end-to-end learning-based image codecs submitted to the MMSP'2020 Grand Challenge on Learning-Based Image Coding and the JPEG AI Call for Evidence [12] were investigated. These coding solutions are representative of the state-of-the-art in learning-based image coding, following different

[2]https://dictionary.cambridge.org/dictionary/english/crowdsourcing
[3]https://github.com/mmspg/qualitycrowd2.1

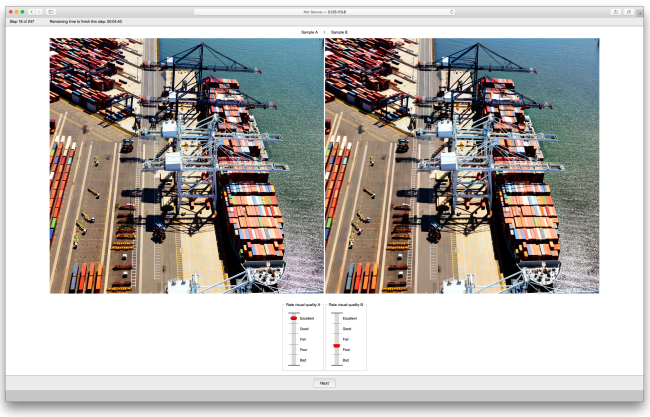


Fig. 2. Layout of the voting step in QualityCrowd2.1 implementing the DSCQS methodology.

neural network architectures and having high compression efficiency measured objectively [13]. Five of these codecs are end-to-end learning-based and include recent and novel tools and methods, namely, training methodologies and loss functions, new convolutional layers and non-local attention modules as well as improved probability models for entropy coding. Moreover, one hybrid codec combining the Versatile Video Coding (VVC Intra) standard [14] with learning-based tools was proposed. Some of these learning-based image coding submissions are briefly described:

1) Team NJU-VISION [15]: This image codec uses a variational autoencoder structure, with non-local attention modules at both main and hyperprior encoder-decoder pairs following a similar architecture to [5]. A masked 3D prediction based on a convolutional neural network is used to obtain accurate conditional statistics for the entropy coding engine.
2) Team Four-leaf Clover [16]: This image codec uses generalized octave convolutions and transpose convolutions to factorize (or divide) the latent representation into high and low frequency components. The low frequency components have a lower resolution that leads to improved rate-distortion performance.
3) Team Nokia [17]: This image codec is based on the concept of meta-learning where the encoder and decoder models are optimized for each input image to overfit (or highly adapt) the latents to the image content. A new probability model based on multi-scale progressive statistical models is also proposed.
4) Team NCTU [18]: This image codec features a frame-

work with VVC Intra at a base layer and a learning-based residual codec at an enhancement layer. The enhancement layer *per se* includes a local attention block for enhancing critical high-frequency areas. Additionally, multi-rate encoding is supported with a single programmable autoencoder.

Furthermore, two anchor image codecs were used for benchmarking, namely JPEG 2000 and HEVC Intra. Information about the encoding configurations for the anchors is included in Table I.

B. Image Dataset

The eight images selected for the subjective quality assessment experiment, shown in Figure 1, are a part of the JPEG AI test dataset[4], and have spatial resolutions between 1336x872 and 2144x1424 pixels. To fit the side-by-side placement in the screen of resolution 1920x1080 pixels or higher, the images have been cropped into tiles of size 945x880 pixels. The cropping area was selected while attempting to maintain the most salient area of the images intact. The cropping parameters can be found in Table II. The original images are in PNG format and RGB24 colorspace with 8 bit per channel.

Since some of the image codecs could only receive input images in YUV colorspace, an RGB to YUV colorspace conversion was performed following ITU Recommendation BT.709 [19] using FFMPEG software.

In addition, the ICC color profiles from originals were copied to all corresponding decoded images to avoid visual changes in color between the original and the assessed (decoded) pictures when displayed for subjective evaluation. This was performed with the ImageMagick software.

C. Subjective evaluation methodology

The DSCQS methodology was used for subjective evaluation. It allows evaluating pairs of images where the reference image may be of lower quality than the decoded image. In some cases, in fact, the learning-based image codecs may perceptually enhance an image, e.g., by reducing the noise, improving the contrast, or increasing the sharpness. According to DSCQS, the subjects look at the original and decoded images that are presented side by side on the screen, without being aware of which image is the reference; the position (left or right) of the reference image is changed in a pseudo-random order. Both the original and decoded images are evaluated using a continuous scale. The subjects assess the overall quality of the original and decoded images by moving the mark on the vertical scale, which is divided into five equal-length intervals corresponding to the standard ITU-R five-point quality scale, namely, *Excellent*, *Good*, *Fair*, *Poor*, and *Bad*. The vertical scales are printed in pairs to reflect the side-by-side presentation of the two images under evaluation. Figure 2 shows the layout of a voting step as used in the experimental setup.

The subjective evaluation workflow follows the ITU-R Rec. BT.500 [20]. A randomized presentation order for the stimuli, as in ITU-R P.910 [21] was used, i.e., the same content was never consecutively shown to the subjects. There was no presentation or voting time limit. Prior to the experiment, the display devices of all the subjects were required to successfully pass the screen size test, thus restricting the screen resolution to be equal or larger than 1920x1080 pixels. At the beginning of the experiment, subjects were presented with written instructions explaining the purpose of the experiment, the voting interface, and the meaning of the quality rating. Each session started with three training examples (bad, excellent and fair) in order to familiarize the subjects with the graphical interface and the range of visual quality. The scores from the training examples were discarded. The subjects were requested to evaluate 240 image pairs to complete their tasks.

[4]https://jpeg.org/jpegai/dataset.html

IV. Data Analysis

A. Population

Subjects of different age, gender, and country of residence were recruited at the Amazon MTurk crowdsourcing platform. In total, 118 participants finished the experiment. The data from two outlier subjects were eliminated leading to a final number of 116 people of which 32 were females and 84 were males. The age of the participants spanned from 18 to 70 years old with the mean of 34.32 and the median of 32.50. The subjects remotely connected from ten different countries with the United States, India, and Brazil being the top three (Table III right). Most of the display devices used by the participants had a screen resolution of 1920x1080 pixels (Table III left).

Outlier detection was performed according to ITU-R P.910 [20], by taking into account the number of far-off votes of a subject, as well as how balanced were their votes. Out of the 118 naïve subjects who completed the experiment, two outliers were detected and their data were excluded.

B. Differential Mean Opinion Scores

The Mean Opinion Scores (MOS) were computed independently for each test condition according to the following equation:

$$MOS_i = \frac{1}{N} \sum s_{ij} \tag{1}$$

where N is the number of valid subjects and s_{ij} is the score by subject j for the test condition i.

Because in DSCQS the reference image must also be graded by the subjects, instead of reporting the MOS for both the source reference (SR) and processed stimuli (PS), the Differential Mean Opinion Scores (DMOS) were computed as:

$$DMOS_{PS} = MOS_{PS} - MOS_{SR} + \max(\text{RatingScale}) \tag{2}$$

DMOS were computed for each of the 240 stimuli with a scale from 0 to 100. Additionally, a confidence interval of 95% was calculated for each MOS_{PS} assuming a Student's t-distribution.

TABLE I
ANCHOR CODING CONFIGURATIONS

Codec	Version	Command line or configuration file
JPEG 2000	Kakadu V8.0.5	`kdu_compress -i <ppm> -o <j2k> -rate   Qstep=0.001 -tolerance 0 -full -precise`
HEVC Intra	HM-16.20+SCM-8.8	`https://jpegai.github.io/public/encoder_intra_main_scc_10.cfg`

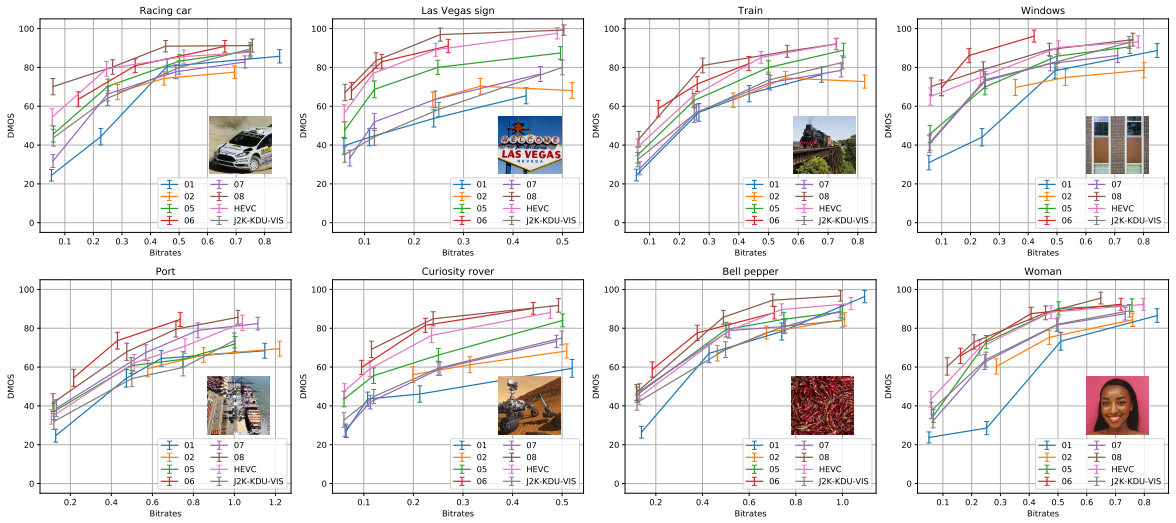


Fig. 3. Differential mean opinion scores (DMOS) versus bitrates with 95% confidence intervals.

TABLE II
IMAGE CROPPING PARAMETERS

Image	Original	Cropped	Top left corner
jpegai02-Racing car	2144x1424	945x880	(323,164)
jpegai05-Las Vegas sign	1336x872	945x872	(190,0)
jpegai06-Train	1544x1120	945x880	(235,248)
jpegai07-Windows	1472x976	945x880	(258,58)
jpegai09-Port	1976x1312	945x880	(259,299)
jpegai10-Curiosity rover	2000x1128	945x880	(547,185)
jpegai11-Bell pepper	1744x1160	945x880	(394,174)
jpegai12-Woman	1512x2016	945x880	(286,288)

TABLE III
SCREEN SIZE AND COUNTRY STATISTICS

Screen size	# Subj.
1920x1080	95
1920x1200	15
2560x1440	3
3440x1440	3
2048x1280	2
2560x1080	2
2560x1600	2
other >1920x1080	4

Country	# Subj.
United States	88
India	17
Brazil	8
United Kingdom	3
Honduras	2
Italy	2
other	5

V. RESULTS AND DISCUSSION

The subjective evaluation experiment results are shown in Figure 3. Since the learning-based codecs correspond to anonymous submissions to the MMSP'2020 Grand Challenge and the JPEG AI Call for Evidence, all codecs are identified with the team number instead of the codec name. The DMOS are plotted against the bitrates, including the 95% confidence intervals. The plots in Figure 3 show that the image codecs proposed by Team 08 and Team 06 can attain better compression performance compared to HEVC Intra, which was among the most competitive anchors at the time the experiment was performed. The gain for the learning-based image coding solutions compared to HEVC Intra depends on the specific image characteristics. It can be relatively large as shown for the *Windows* and *Port* images for Team 06 and *Curiosity rover* for Team 08. Moreover, Team 06 and Team 08 image codecs have an overall compression performance similar to HEVC Intra, and they compete for the best performing solution: indeed, Team 06 outperforms Team 08 for the *Windows* and *Port* images, whilst Team 08 outperforms Team 06 for the *Racing Car* and *Las Vegas* images. Moreover, despite its age, JPEG 2000 is able to outperform some learning-based image codecs for many images and bitrates.

VI. CONCLUSION

In this paper, a large-scale crowdsourcing subjective quality evaluation experiment was reported and its results and analysis for six learning-based image codecs and two anchor codecs were presented. The results show that the new end-to-end learning-based image coding approach can achieve promising compression performance, thus indicating that future developments in image coding will likely adopt this novel coding approach.

Moreover, a dataset of differential mean opinion scores together with compressed stimuli has been made publicity available for the community in order to facilitate further research on the subjective and objective quality evaluation and performance assessment in learning-based image coding.

REFERENCES

[1] M. Testolina, E. Upenik, and T. Ebrahimi, "Comprehensive assessment of image compression algorithms," in *SPIE Applications of Digital Image Processing XLIII*, vol. 11510, Aug. 2020.

[2] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable Rate Image Compression with Recurrent Neural Networks," *arXiv:1511.06085 [cs]*, Mar. 2016.

[3] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end Optimized Image Compression," *arXiv:1611.01704 [cs, math]*, 2017.

[4] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, and M. Covell, "Full Resolution Image Compression With Recurrent Neural Networks," in *CVPR*, 2017.

[5] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Int. Conf. on Learning Representations (ICLR)*, 2018.

[6] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, "High-Fidelity Generative Image Compression," *arXiv:2006.09965 [cs, eess]*, 2020.

[7] M. Testolina, E. Upenik, J. Ascenso, F. Pereira, and T. Ebrahimi, "Performance evaluation of objective image quality metrics on conventional and learning-based compression artifacts," in *13th International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2021.

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

[9] Z. Cheng, P. Akyazi, H. Sun, J. Katto, and T. Ebrahimi, "Perceptual quality study on deep learning based image compression," in *IEEE International Conference on Image Processing (ICIP)*, 2019.

[10] J. Ascenso, P. Akyazi, F. Pereira, and T. Ebrahimi, "Learning-based image coding: early solutions reviewing and subjective quality evaluation," in *Optics, Photonics and Digital Technologies for Imaging Applications VI*, vol. 11353. SPIE, 2020.

[11] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd - A framework for crowd-based quality evaluation," in *2012 Picture Coding Symposium*, 2012.

[12] "ISO/IEC JTC 1/SC29/WG1 N86018 Call for Evidence on Learning-based Image Coding Technologies (JPEG AI)," 2020.

[13] "ISO/IEC JTC 1/SC29/WG1 N89022 Report on the JPEG AI Call for Evidence Results," 2020.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] "ITU-R BT.709 : Parameter values for the HDTV standards for production and international programme exchange," 2015.

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

[21] "ITU-R P.910: Subjective video quality assessment methods for multimedia applications," 2008.