JPEG AIC-3 Dataset: Towards Defining the High Quality to Nearly Visually Lossless Quality Range

Michela Testolina*, Vlad Hosu†, Mohsen Jenadeleh†, Davi Lazzarotto*, Dietmar Saupe†, Touradj Ebrahimi*

*Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

†Multimedia Signal Processing Group, Universität Konstanz, Konstanz, Germany

{michela.testolina, davi.nachtigalllazzarotto, touradj.ebrahimi}@epfl.ch

{vlad.hosu, mohsen.jenadeleh, dietmar.saupe}@uni-konstanz.de

Abstract—Visual data play a crucial role in modern society, and the rate at which images and videos are acquired, stored, and exchanged every day is rapidly increasing. Image compression is the key technology that enables storing and sharing of visual content in an efficient and cost-effective manner, by removing redundant and irrelevant information. On the other hand, image compression often introduces undesirable artifacts that reduce the perceived quality of the media. Subjective image quality assessment experiments allow for the collection of information on the visual quality of the media as perceived by human observers, and therefore quantifying the impact of such distortions. Nevertheless, the most commonly used subjective image quality assessment methodologies were designed to evaluate compressed images with visible distortions, and therefore are not accurate and reliable when evaluating images having higher visual qualities. In this paper, we present a dataset of compressed images with quality levels that range from high to nearly visually lossless, with associated quality scores in JND units. The images were subjectively evaluated by expert human observers, and the results were used to define the range from high to nearly visually lossless quality. The dataset is made publicly available to researchers, providing a valuable resource for the development of novel subjective quality assessment methodologies or compression methods that are more effective in this quality range.

Index Terms—subjective visual quality assessment, dataset, high visual quality, nearly visually lossless quality, JND, JPEG AIC-3

Fig. 1: Reference images in the JPEG AIC-3 dataset

I. INTRODUCTION

Advances in digital cameras, broadband internet, and display technologies have made high-quality imaging feasible, desirable, and more accessible than ever before, opening up new possibilities for creative expression and scientific discovery. Image compression is essential in order to limit the storage resources, which have been constantly increasing over the years [1]. Depending on the coding algorithm and desired compression ratio, image compression might introduce visible and undesirable artifacts to images, reducing their perceived visual quality. Recent image compression methods, e.g. JPEG XL [1], attempt to reach a high compression ratio without significantly compromising the visual quality of the reconstructed images. Moreover, the Joint Photographic Experts Group (JPEG) is working, in the context of JPEG AI [2], towards a method able to even enhance the quality of the reconstructed images thanks to embedded image processing operations such as denoising and super-resolution.

The impact of compression artifacts on the visual quality of images may be measured through subjective quality assessment experiments, which consist in collecting multiple subjective visual quality scores from a variety of human observers. While these experiments are able to precisely estimate the visual quality under specific conditions, different subjective quality assessment methodologies might provide different results when compared to each other. For this reason, the approach used for presenting images and the specific task or question asked to test subjects must be thoughtfully selected.

The JPEG AIC-3, initiated by the JPEG Committee in July 2021, is the result of observations that the most frequently used subjective image quality assessment methodologies, e.g. the methods presented in [2], are more suitable for evaluating compression artifacts with low to high visual qualities. Moreover, the methods standardized in the context of the JPEG AIC-2, e.g. flicker test [3], are suitable for evaluating images

---

[1] https://phototutorial.com/photos-statistics/, accessed on 2023.02.13

[2] https://jpeg.org/jpegai/index.html

TABLE I: Reference images in the JPEG AIC-3 dataset and crop information.

IMAGE NUMBER	CONTENT	REFERENCE RESOLUTION	CROP SIZE	(x,y)
00001	Object	1192×832	945×832	(54,0)
00002	Human portrait	853×945	853×880	(2,7)
00003	Food	945×840	945×840	(0,0)
00004	Computer generated	2000×2496	945×880	(165,1019)
00005	Animal	560×888	560×880	(0,3)
00006	Scene with water	2048×1536	945×880	(102,228)
00007	Night scene	1600×1200	945×880	(54,177)
00008	Fabric	1430×1834	945×880	(31,43)
00009	Landscape	2048×1536	945×880	(12,44)
00010	Buildings	2592×1946	945×880	(139,261)

with nearly lossless visual quality. It has also been observed that the quality range between high to nearly visually lossless still lacks a suitable and effective methodology for measuring the visual quality of images [4]. However, such a quality range was not rigorously defined yet.

The contributions of this paper are the following:

- We present the JPEG AIC-3 dataset, i.e. a dataset of 500 images with different contents and resolutions, in both their original and decoded forms, in the range from high quality to nearly visually lossless quality.
- We provide the associated subjective visual scores in terms of JND units obtained through a subjective quality assessment experiment conducted in a crowdsourcing environment by expert viewers.
- We propose a more definite description of the high to nearly visually lossless quality range, in terms of JND units, to facilitate the research towards novel subjective quality assessment methodologies or compression methods effective in this range.

The JPEG AIC-3 dataset, as well as the associated scores in JND units, are made publicly available to facilitate further research on the topic [3].

II. RELATED WORK

The impact of the distortions introduced by image compression algorithms can be assessed through subjective image quality assessment experiments, which aim at measuring the quality of an image as perceived by a large number of human observers. Multiple subjective image quality assessment experiments have been proposed and standardized in the past, e.g. in ITU-R Rec. BT.500 [2]. Notably, the specifications include both single stimulus protocols, e.g. Absolute Category Rating (ACR), and double stimulus, e.g, Double Stimulus Impairment Scale (DSIS), Double Stimulus Continuous Quality-Scale(DSCQS) and Pair Comparison (PC). A wider review of the methodologies presented in BT.500 may be found in [5]. The methodologies reported in BT.500 [2] were previously used in a number of studies assessing the performance of image compression. For example, the ACR protocol was used to evaluate early learning-based compression algorithms in [6]. Afterward, the DSCQS protocol was used to assess the performance of more recent learning-based compression

algorithms in the context of the JPEG AI Call for Evidence [7] in a crowdsourcing environment. Crowdsourcing approaches to subjective visual quality assessment are, in fact, becoming increasingly widespread over the years, offering the possibility of collecting a large number of subjective quality scores at a contained cost. Nevertheless, as these experiments are conducted in an uncontrolled environment, a number of best practices are defined to obtain trustworthy results [8].

In recent years, the JPEG Committee has standardized additional subjective quality assessment methodologies in the context of the JPEG AIC [3], [9] targeting visually lossless qualities:

- **JPEG AIC Part 2 Annex A** is a triple-stimulus experiment that consists in asking a number of human observers, or test subjects, to select the closest match to the reference image among two choices, where one is the decoded image and the other the reference itself.
- **JPEG AIC Part 2 Annex B** presents two stimuli to the test subjects, the first being the distorted image interleaved with the reference, and the second being the reference image interleaved with itself, at a specific frequency. The position of the stimuli is randomized, and test subjects are asked to select the non-flickering stimulus.

As reported in [10], [11], the methods proposed in BT.500 are more suitable for evaluating the *visual appeal* of images, while the methods proposed in JPEG AIC Part 2 are extremely sensitive and therefore more suitable for evaluating their *visual fidelity* to a reference.

A number of alternative methodologies to the standards have been proposed in the state of the art. The Just Noticeable Difference (JND) represents an alternative to the computation of the MOS. Notably, the JND indicates the smallest change in image quality that the human visual system (HVS) is able to discriminate. In this paper, 1 JND is defined at a 50% detection rate, meaning that two images are 1 JND apart if the probability of an observer being able to discriminate between two images is estimated at about 50%.

The concept of JND is fairly established in the context of image quality assessment. Among the first works in the context of subjective visual quality assessment, [12] presented a novel approach for measuring the video quality in terms of JND scale values, i.e. a two-alternative forced-choice method

Sample A | Sample B

Please select the image with the highest visual quality

○ Sample A   ○ Not sure   ○ Sample B

Next

Fig. 2: Interface of the conducted subjective experiment.

where test subjects were asked to select the most impaired video. Nevertheless, the research in this field has developed only recently. In [13], the authors conducted a JND test for both compressed images and videos, where the media were shown side-by-side and test subjects were asked to assess whether the differences were noticeable. In [14], the authors combined two different approaches, namely a two-alternative forced choice (2AFC) test and a JND test, utilizing small patches of size 64x64 in order to reduce the influence of the high-level semantics and allow test subjects to focus on lower-level aspects of similarity. Recently, a subjective image quality assessment methodology using boosted triplet comparisons has been presented in [15], introducing different boosting techniques and improving the sensitivity to smaller distortion levels. In [16], a variation of the flicker test was used to collect crowdsourcing-based subjective picture-wise just noticeable difference (PJND) scores, where only one flickering stimulus at the time was presented and where test subjects were asked to adjust a slider until they were able to perceive flicker.

A number of databases of images and associated JND scores are available in the state of the art. The MCL-JCI dataset [17] presents raw JND data along with their processed stair quality function (SQF) for JPEG-coded images. Following, a dataset including JND scores collected over images compressed with VVC [18], [19] was presented. The KonJND-1k [16] is the largest-scale dataset of images and associated JND scores, including a total of 1,008 source images, each having an average of 42 samples and associated PJND scores, collected in an extensive crowdsourcing-based experiment. While the mentioned datasets include a large variety of test stimuli and JND samples, they do not cover a large variety of compression artifacts caused by different codecs and are not specifically designed to cover the range of visual qualities from high to nearly visually lossless.

III. SUBJECTIVE EXPERIMENT

A. Test material

The JPEG AIC-3 dataset consists of 10 uncompressed reference images with different resolutions and content. The images belong to different categories, namely: objects, human portraits, food, computer-generated images, animals, scenes with water, night scenes, fabric/fine texture, landscapes, and

buildings. A preview of the source images in the JPEG AIC-3 dataset is provided in Figure 1, while the information on the content and resolution of the images is available in Table I.

The images were compressed with multiple compression methods, namely JPEG, JPEG 2000, HEVC Intra (HM), VVC Intra, and JPEG XL, with all the available quality levels provided by the codecs. In order to select the quality levels in the interval between high to nearly visually lossless quality, subjective scores were collected on a preliminary subset of five distorted images selected by visual inspection to cover a large number of JND units. Statistical analysis and interpolation were applied to the obtained JND scores in order to refine the initial selection and extract the final images to be included in the dataset.

B. Experiment setup

In order to select the compressed images for the JPEG AIC-3, their visual quality was assessed subjectively. The experiments were conducted in a crowdsourcing environment with expert viewers. The selected platform for the experiment was "QualityCrowd 2" [4] [20]. Prior to the beginning of the experiment, a preliminary screen-check was conducted, and only subjects with a screen of size 1920×1080 or larger, with retina mode disabled (device pixel ratio equal to 1), were able to proceed to the experiment.

Image cropping to a size of 945×880 was necessary in order to fit two stimuli side-by-side on the target screen size. The cropping was performed by selecting the salient area of each image or the area where the artifacts are the most visible, according to prior visual inspection by the authors. The information on the cropping area is reported in Table I.

The protocol adopted for the experiment is a variation of the pair comparison (PC) experiment, where the subjects were asked to select the stimulus presenting the highest visual quality between two options, displayed side-by-side. The order of the pairs and the position of each stimulus were selected randomly. In addition, care was devoted to not displaying the same content consecutively. The question presented to test subjects was "Please select the image with the highest visual quality", where the subjects could choose between (a) "Sample A" (b) "Not sure" and (c) "Sample B". The interface presented to the subjects during the experiment is shown in Figure 2.

Each compressed image was compared to its reference and to all the other images compressed with the same codec, excluding comparisons between different codecs. Following this approach, 750 pairs were rated during the experiment. Nevertheless, in order to avoid fatigue, the test was divided into two parts including 375 comparisons each. Prior to the beginning of the experiment, a short training session was conducted to get test subjects acquainted with the experiment and the grading scale. A total of 31 subjects completed the experiment, where 15 participants completed part 1, and the remaining 16 participants completed part 2. The average age

---

[4]The implementation provided at https://github.com/mmspg/qualitycrowd2.1 was used for the experiment.
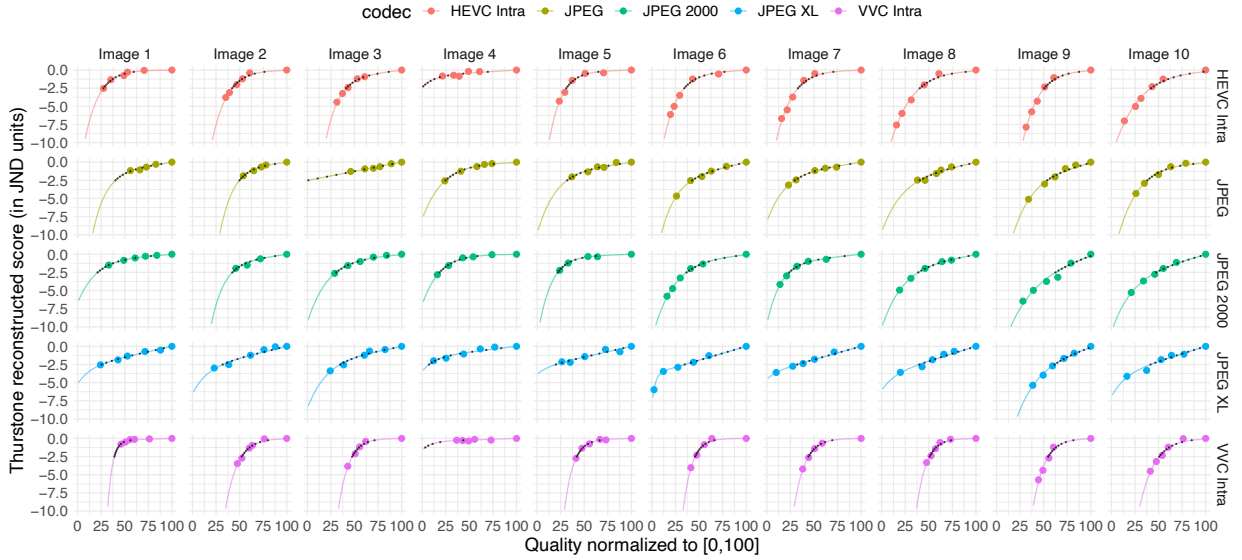
Fig. 3: Result of the expert viewing. Columns correspond to source images, and rows to codecs. The horizontal axis corresponds to the (normalized) codec quality settings, from 0 (worst) to 100 (best quality). The vertical axis gives the reconstructed quality scale in JND units from the subjective ratings. Each large colored dot represents one image. The small black dots on the fitting curves indicate the interpolated images selected for the JPEG AIC-3 dataset (best viewed when the figure is enlarged).

of participants was 35.7, with a minimum age of 22 and maximum age of 68. 9 subjects (approx. 29%) identified as female, 21 subjects (approx. 68%) identified as male, and one subject preferred not to disclose their gender. All subjects were directly recruited by the authors of this paper.

IV. STATISTICAL ANALYSIS

Statistical analysis was conducted on the collected scores with the goal of providing perceptual image quality scale values for all the selected preliminary compressed images. A similar procedure to [15] was adopted to process the results collected during the subjective quality assessment experiment. Notably, standard reconstruction was applied by maximum likelihood estimation according to the *Thurstonian* probabilistic model (Case V), where the "eba" R-Package [21] containing the "thurstone()" function was used. This model assumes that the quality of each stimulus is a random variable on a latent perceptual quality scale, having a normal distribution with a variance of $1/2$ and mean values that need to be estimated. Under this model, a paired comparison of two stimuli proceeds by considering the difference between the two associated (independent) random variables, which is normally distributed with variance 1 and mean equal to the difference of the means. Accordingly, in a paired comparison, the stimulus with a larger magnitude will be chosen as the better one with probability $\Phi(\Delta)$, where Φ is the normal cumulative distribution function and Δ is the difference of the two means. In the reconstruction algorithm, the means are determined so that the yielded probabilities best match with those estimated from the responses in the pair comparisons. See [22] for a broad introduction to scale reconstruction from pair comparisons.

The results were shifted so that the score of the source image was set at zero. Moreover, the results were scaled to JND units by dividing all scale values by $\Phi^{-1}(0.75) = 0.6745$. Accordingly, if two images are scaled and 1 JND unit apart, then the model predicts a 50% probability for the detection of the difference by a random observer.

For some of the contents, the worst-quality image won hardly any of the comparisons, leading to the zero-frequency problem. In order to regularize the outcome, the pair comparison matrix was initialized with a small constant at each entry. In this experiment, an initialization value of 0.1 was used, i.e. for each paired comparison, a virtual *"not sure"* vote was introduced, weighted with a factor of 0.2.

Figure 3 shows the results for each source image and codec as functions of the codec quality levels, normalized to $[0, 100]$ for all the codecs. For example, in VVC Intra, the quantization parameter $QP \in [0, 63]$ determines the quality, and consequently 0 was mapped to 100 and 63 to 0 linearly.

In each of the plots, a parametric curve was fitted to the collected subjective quality scores. In particular, the parametric curve $f(x)$ was estimated as the sum of a linear part with a slope parameter (a) and a logistic part with two parameters (b and c), where $x \in [0, 100]$ is the standardized quality,

$$f(x) = -a\left(1 - \frac{x}{100}\right) + \frac{100}{1 + e^{-100b(\frac{x}{100} - c)}} - 100. \quad (1)$$

The curves show how the perceived image quality monotonically increases as the codec quality level increases to its maximum. After a visual assessment of the images and associated JND values, the minimum scale value of -2.5 JND was selected for the images included in the dataset. Consequently, the scale interval of [-2.5, 0] was subdivided into 10 sub-

(a) Reference   (b) Reconstructed

Fig. 4: Crop of reference image 00004 and its reconstructed version with VVC Intra.

intervals of length equal to 0.25 JND. The corresponding selected quality levels, shown in Figure 3 as small black dots, were then mapped back to the closest integer for each of the five codecs.

V. RESULTS AND DISCUSSION

The fitting curves in Figure 3 present, overall, a close-to-logarithmic behavior, where the visual difference between two consecutive quality levels is greater for lower-quality values and minor for higher-quality values. An atypical behavior occurs for image 00004 (artificially generated content), mainly visible for the HEVC Intra and VVC Intra codecs. The reason can be better appreciated after visual inspection of the images: Figure 4 reports a crop of (a) reference image and (b) its compressed version with VVC Intra with quality parameter 32. It can be observed that the reference image presents some noisy areas, possibly introduced for artistic purposes. Both HEVC Intra and VVC Intra generate smoothing and loss of details as artifacts, resulting in a reduction of the noise. This effect was rated by many viewers as visually more appealing. This reveals that, in the performed experiment, subjects were more inclined to rate the *visual appeal* of images as opposed to their *visual fidelity*. This was eventually emphasized as the reference image was kept hidden.

Four objective quality metrics, i.e. PSNR, SSIM, VMAF [23] and LPIPS [14], were computed and their behavior was compared to the estimated JND values. The results are shown in Figure 5, where the average objective metrics over all the images in the dataset are plotted against the JND values. It may be observed that the HEVC Intra and VVC Intra are, overall, the codecs with the highest scores for PSNR, SSIM, and LPIPS metrics at equal JND values. By contrast, the JPEG codec has the lowest score at equal JND. Moreover, at parity of the objective score, images compressed with JPEG-associated codecs (i.e. JPEG, JPEG 2000, and JPEG XL) have higher visual quality. This might happen because better visual quality is not always linked to higher objective metric values. Regardless, this trend does not occur for VMAF, possibly because the metric was designed to evaluate the quality of compressed videos rather than images. Further research in this area, therefore, is essential to bring more insights.
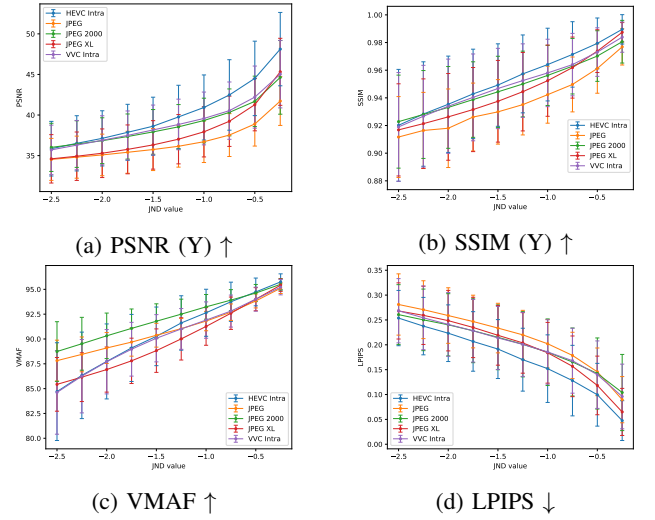

(a) PSNR (Y) ↑   (b) SSIM (Y) ↑

(c) VMAF ↑   (d) LPIPS ↓

Fig. 5: Average PSNR, SSIM, VMAF and LPIPS over the images part of the JPEG AIC-3 dataset plotted against the JND values. ↑ indicates that higher metric values suggest better visual quality, while ↓ indicates that lower metric values suggest better visual quality. The error bar indicates the standard deviation of the objective values.

The meticulous visual inspection of the images and the analysis of the collected subjective results motivated further efforts in order to carefully define the range of quality between high and nearly visually lossless. In view of the definition of JND adopted in this paper, this range was defined between -1 and -2 JND values. Images presenting JND scores higher than -1, in fact, suggest that their difference with the reference is visible only by less than 50% of the population, defining the nearly visually lossless to visually lossless quality range. On the other hand, images presenting JND scores lower than -2 present perceivable artifacts that reduce their *visual appeal*. Figure 6 presents a visual example of images in the different quality ranges. While the reference and the compressed image at -0.5 JND present indiscernible visual quality for the majority of the population, the compressed image at -1.5 presents visual inconsistencies while still maintaining a high *visual appeal*. Emphasis must be put on the range between nearly visually lossless to visually lossless likewise. The methodologies presented in the context of the JPEG AIC-2, in fact, are only able to determine the threshold above which the quality becomes visually lossless, without discriminating between the subtle variations in this range. In further studies on the topic, the visually lossless threshold in terms of JND units may be investigated and subjective quality assessment methodologies able to discriminate images in both the reviewed visual quality ranges may be explored.

VI. CONCLUSIONS

In this paper, we presented a dataset including 500 images with 10 different contents and 10 different distortion levels from 5 different codecs, in the range of visual qualities from high to nearly visually lossless quality. The distorted images
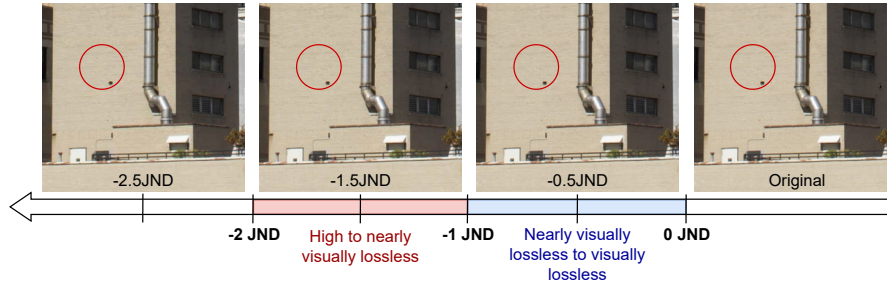
Fig. 6: Comparison of crops in the different quality ranges.

in the presented dataset were selected through a subjective image quality assessment experiment conducted by expert viewers, and the collected JND values are provided as part of the dataset. The results were also used to define the high quality to the nearly visually lossless quality range, which was defined as having JND values between -1 and -2. This study aspires to motivate further efforts on the subjective and objective assessment of images in the identified range of qualities and promote research on image compression methods with improved visual quality.

REFERENCES

[1] J. Alakuijala, S. Boukortt, T. Ebrahimi, E. Kliuchnikov, J. Sneyers, E. Upenik, L. Vandevenne, L. Versari, and J. Wassenberg, "Benchmarking JPEG XL image compression," in *Optics, Photonics and Digital Technologies for Imaging Applications VI*, vol. 11353. SPIE, 2020, pp. 187–206.

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

[3] ISO/IEC 29170-2:2015, "Information technology — Advanced image coding and evaluation — Part 2: Evaluation procedure for nearly lossless coding."

[4] ISO/IEC JTC1/SC29/WG1 N100311, "Final Call for Contributions on Subjective Image Quality Assessment," https://jpeg.org/aic/documentation.html.

[5] M. Testolina and T. Ebrahimi, "Review of subjective quality assessment methodologies and standards for compressed images evaluation," in *Applications of Digital Image Processing XLIV*, vol. 11842. SPIE, 2021, pp. 302–315.

[6] Z. Cheng, P. Akyazi, H. Sun, J. Katto, and T. Ebrahimi, "Perceptual quality study on deep learning based image compression," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 719–723.

[7] E. Upenik, M. Testolina, J. Ascenso, F. Pereira, and T. Ebrahimi, "Large-scale crowdsourcing subjective quality evaluation of learning-based image coding," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2021, pp. 1–5.

[8] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2013.

[9] D. F. Stolitzka, P. Schelkens, and T. Bruylants, "New procedures to evaluate visually lossless compression for display systems," in *Applications of Digital Image Processing XL*, vol. 10396. SPIE, 2017, pp. 98–108.

[10] M. Testolina, E. Upenik, J. Sneyer, and T. Ebrahimi, "Towards JPEG AIC part 3: visual quality assessment of high to visually-lossless image coding," in *Applications of Digital Image Processing XLV*, vol. 12226. SPIE, 2022, pp. 90–98.

[11] "What to Focus on in Image Compression: Fidelity Or Appeal," accessed: 2022.08.16, "https://cloudinary.com/blog/what_to_focus_on_in_image_compression_fidelity_or_appeal".

[12] A. B. Watson, "Proposal: Measurement of a JND scale for video quality," *IEEE G-2.1.6 Subcommittee Video Compression Measurements*, 2000.

[13] J. Y. Lin, L. Jin, S. Hu, I. Katsavounidis, Z. Li, A. Aaron, and C.-C. J. Kuo, "Experimental design and analysis of JND test on coded image/video," in *Applications of Digital Image Processing XXXVIII*, vol. 9599. SPIE, 2015, pp. 324–334.

[14] 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.

[15] H. Men, H. Lin, M. Jenadeleh, and D. Saupe, "Subjective image quality assessment with boosted triplet comparisons," *IEEE Access*, vol. 9, pp. 138 939–138 975, 2021.

[16] H. Lin, G. Chen, M. Jenadeleh, V. Hosu, U.-D. Reips, R. Hamzaoui, and D. Saupe, "Large-scale crowdsourced subjective assessment of picturewise just noticeable difference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5859–5873, 2022.

[17] L. Jin, J. Y. Lin, S. Hu, H. Wang, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis," *Electronic Imaging*, vol. 2016, no. 13, pp. 1–9, 2016.

[18] X. Shen, Z. Ni, W. Yang, X. Zhang, S. Wang, and S. Kwong, "A JND dataset based on VVC compressed images," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[19] ——, "Just noticeable distortion profile inference: A patch-level structural visibility learning approach," *IEEE Transactions on Image Processing*, vol. 30, pp. 26–38, 2020.

[20] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd—a framework for crowd-based quality evaluation," in *2012 Picture Coding Symposium*. IEEE, 2012, pp. 245–248.

[21] F. Wickelmaier, "Eba: Elimination-by-Aspects Models." [Online]. Available: https://CRAN.R-project.org/package=eba

[22] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," *arXiv preprint arXiv:1712.03686*, 2017.

[23] 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, p. 2, 2016.