On the Performance of Subjective Visual Quality Assessment Protocols for Nearly Visually Lossless Image Compression

Michela Testolina*
Davi Lazzarotto*
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
michela.testolina@epfl.ch
davi.nachtigalllazzarotto@epfl.ch

Rafael Rodrigues
Universidade da Beira Interior & Instituto de Telecomunicações
Covilhã, Portugal
rafael.rodrigues@ubi.pt

Shima Mohammadi
Instituto Superior Técnico & Instituto de Telecomunicações
Lisbon, Portugal
shima.mohammadi@lx.it.pt

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

António M. G. Pinheiro
Universidade da Beira Interior & Instituto de Telecomunicações
Covilhã, Portugal
pinheiro@ubi.pt

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

ABSTRACT

The past decades have witnessed rapid growth in imaging as a major form of communication between individuals. Due to recent advances in capture, storage, delivery and display technologies, consumers demand improved perceptual quality while requiring reduced storage. In this context, research and innovation in lossy image compression have steered towards methods capable of achieving high compression ratios without compromising the perceived visual quality of images, and in some cases even enhancing the latter. Subjective visual quality assessment of images plays a fundamental role in defining quality as perceived by human observers. Although the field of image compression is constantly evolving towards efficient solutions for higher visual qualities, standardized subjective visual quality assessment protocols are still limited to those proposed in ITU-R Recommendation BT.500 and JPEG AIC standards. The number of comprehensive and in-depth studies where different protocols are compared is still insufficient. Moreover, previous works have not investigated the effectiveness of these methods on higher quality ranges, using recent image compression methods. In this paper, subjective visual scores collected from three subjective image quality assessment protocols, namely the Double Stimulus Continuous Quality Scale (DSCQS) and two test methods described in the JPEG AIC Part 2 standard, are compared between different laboratories under similar controlled conditions. The analysis of the experimental results has revealed that the DSCQS protocol is highly influenced by the quality of the reference images and experience of the subjects, while the JPEG AIC Part 2 specifications produce more stable results but are expensive and only suitable for a limited range of qualities. These emphasize the need for new robust subjective image quality assessment methodologies able to discriminate in the range of qualities generally demanded by consumers, i.e. from high to nearly visually lossless.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → *Image compression.*

KEYWORDS

Subjective image quality assessment, visual quality, quality of experience, BT.500, JPEG AIC

ACM Reference Format:
Michela Testolina, Davi Lazzarotto, Rafael Rodrigues, Shima Mohammadi, João Ascenso, António M. G. Pinheiro, and Touradj Ebrahimi. 2023. On the Performance of Subjective Visual Quality Assessment Protocols for Nearly Visually Lossless Image Compression. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3613835

1 INTRODUCTION

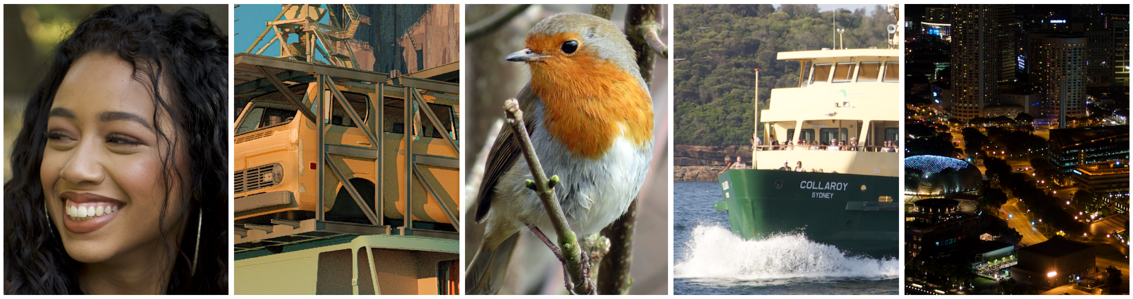
With proliferation of digital images and increasing demand for high-quality content, image compression has become a critical aspect in visual communication. The ability to transmit and store large volumes of digital images efficiently is a key requirement in a wide range of applications, including social media, cloud storage, visual surveillance, and medical imaging, among others.

Image compression techniques have been developed in the past decades to reduce the storage and bandwidth requirements of digital images, while maintaining acceptable levels of visual quality. However, users have become much more demanding in the last years concerning image quality, and thus often express their desire to have compressed images where artifacts are not perceptible even after close inspection. There are two main types of image compression:

[*] Both authors contributed equally to this research.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613835



(a) Face **(b) Artificial** **(c) Bird** **(d) Boat** **(e) Night**

Figure 1: Crops of the reference images used for the experiment.

lossy and lossless. Lossy image compression techniques, both conventional [1, 6, 26, 31, 36] and learning-based [4, 5, 7, 9, 14, 15, 37], exploit the limitations of human visual system to eliminate redundant or irrelevant image information and are able to achieve high compression ratios, while lossless compression [10, 19, 26, 28] can achieve mathematically bit-exact reconstruction but are less efficient. Visually lossless or nearly visually lossless methods are between lossy and lossless compression and thus introduce artifacts, but ensure that those artifacts are unlikely to be visible, i.e. observers cannot easily differentiate between compressed, and original or reference content.

In the design of (nearly) visually lossless compression solutions it is fundamental to measure the perceptual quality accurately as perceived by human observers. Therefore, subjective quality assessment is a critical component in the design, development, and optimization of image compression solutions. In this context, small differences in image quality need to be accurately measured, providing discriminative scores between decoded images that have high or (nearly) visually lossless quality. The JPEG Committee has recently launched a renewed activity on Assessment of Image Coding (AIC), also referred to as JPEG AIC, to develop a new standard (AIC-3) [33] *focusing on the methodologies for assessment of images with the quality levels in between the range where ITU-R Rec. BT.500 [24] is suitable and the range where AIC-2 [11] is suitable.*

The ITU-R Recommendation BT.500 [24] and the JPEG AIC Part 2 (AIC-2) [11, 12] are the most widely used subjective image quality assessment protocols. These protocols define standardized test procedures for evaluating the perceptual quality of compressed images based on the judgment of human observers. Nowadays, double stimulus subjective assessment methodologies (such as those defined in BT.500) are often used, where subjects are asked to score the stimuli using either a numerical (categorical) or continuous quality scale. However, these methods also exhibit disadvantages: *i)* the interpretation of the quality scale may change from observer to observer (even with training procedure); *ii)* they are not well suited for high or nearly visually lossless scenarios, since small differences between stimuli cannot be perceived by the observers. When coding solutions are evaluated with this type of methodology, it is difficult to conclude, from the obtained experimental results, which solution provides better performance at high to visually lossless qualities. AIC-2 was developed for this case and provides two alternative forced-choice methodologies for subjective image quality assessment of (nearly) visually lossless qualities. Despite their availability, subjective assessment methodologies such as AIC-2 were not evaluated with recent image compression solutions (such as JPEG XL and AVIF), nor evaluated between them in a statistically meaningful way. Moreover, the effectiveness of the specifications in assessing distortions in high-quality contents or nearly visually lossless image coding has not been evaluated.

The main objective of this paper is to evaluate and compare the accuracy and reliability of three subjective image quality assessment protocols: the Double Stimulus Continuous Quality Scale (DSCQS) and two subjective assessment methods described in the JPEG AIC Part 2 standard. The study compares the subjective visual scores collected from different laboratories under similar controlled conditions. To obtain test images, several image compression techniques were used, including the latest JPEG XL and AVIF image coding standards. The results of this study will allow to identify the strengths and weaknesses of each protocol, with a particular focus on their effectiveness in discriminating between high-quality to visually lossless compressed images. This is very important to better understand the *status quo* in subjective quality assessment with respect to the aforementioned quality ranges, and thus guide further developments in the area of subjective quality assessment protocols.

The collected subjective quality scores as well as the employed graphical user interface are made publicly available to facilitate further research on the topic [1].

2 RELATED WORK

ITU-R Rec. BT.500 [24] establishes recommendations and best practices for subjective image quality assessment, with the definition of both single and double stimulus test methodologies. Among the most widely used methodologies, the DSCQS protocol prompts the subjects to evaluate the quality of two stimuli presented side-by-side, using a continuous quality rating scale. The reference stimulus is placed randomly, and the subjects are not explicitly instructed about its presence. This protocol is particularly effective in the evaluation of compression methods that employ processing algorithms that may improve the visual appeal of a given image. As an

[1] https://www.epfl.ch/labs/mmspg/downloads/nearly-visually-lossless-subjective-iqa/

example, the authors in [34] used the DSCQS method for the subjective evaluation of learning-based coding solutions in the context of the JPEG AI Call for Evidence. Recent studies reported that the methodologies in ITU-R Rec. BT.500 [24] are primarily suitable for evaluating the *visual appeal*, and for applications where the bitrate reduction is a major concern [33], e.g. web distributions. Therefore, their use to assess nearly visually lossless qualities is not optimal.

The JPEG committee has designed and reported two additional methodologies included in its JPEG AIC specifications, i.e. ISO/IEC TR 29170-1:2017 (AIC-1) [12] and ISO/IEC 29170-2:2015 (AIC-2) [11]. The first focuses on defining common vocabulary and guidelines for subjective, objective, and computational evaluation of image coding systems, including a review of ITU-R Rec. BT.500 [24] and ITU-R Rec. BT.1082 [25]. In contrast, the AIC-2 specifications present two novel methodologies for the assessment of visually lossless image compression, namely AIC-2 A and AIC-2 B (or Flicker test).

The AIC-2 A test protocol is a forced-choice methodology where two test images are presented side-by-side along with a reference. One of the two test images is a copy of the reference. The observer is prompted to choose the test image that is the closest match to the reference. To the best of the authors' knowledge, this methodology was never independently evaluated in previous studies.

The AIC-2 B test protocol, or Flicker test, implements a forced-choice comparison using interleaved images. In each trial, two test samples are presented side-by-side. One of these is the reference image, whereas the other is a compressed and reconstructed image, which is temporally interleaved with the reference image. The observer is requested to choose the non-flickering stimulus. As both the AIC-2 A and the AIC-2 B protocols are forced-choice methodologies, the subjects will submit random answers if the reconstructed image has visually lossless distortions. The Flicker test was often adopted in previous works for evaluating image codecs which target the range from nearly visually lossless to visually lossless qualities [29]. For example, this methodology was used for assessing the performance of submissions to the JPEG XS Call for Proposals [17] and for the assessment of the VESA Display Stream Compression (DSC) codec with Standard Dynamic Range (SDR) images [3], High Dynamic Range (HDR) images [30], and stereoscopic images [22]. Moreover, a large-scale dataset including subjective visual scores collected using the Flicker methodology was presented in [8].

Although both AIC-2 A and Flicker test are suitable for images compressed with light coding solutions, the Flicker test is considered more sensitive than AIC-2 A, as the in-place switching, or toggling, allows for easier identification of subtle artifacts [2], making them suitable for applications such as the visual assessment of photography.

Recently, a number of studies explored alternative methodologies for subjective visual quality assessment. As an example, the authors in [18] proposed to use boosting techniques to improve the visibility of subtle artifacts. In addition, a hybrid approach that combines a pairwise comparison experiment with absolute grading scores was proposed in [27]. The JPEG AIC-3 Dataset [32] was collected in the context of the JPEG AIC Part 3 activity using a variation of the Pair Comparison (PC) protocol. It includes 10 reference images compressed with several codecs, i.e. JPEG, JPEG 2000, HEVC Intra, VVC Intra, JPEG XL, and AVIF, at 10 distortion levels in the range from high to nearly visually lossless quality, corresponding to 0.25 to 2.5 Just Noticeable Difference (JND) units.

The number of studies which compare different subjective image quality assessment methodologies is still limited. A preliminary analysis in the context of video compression was conducted in [23], where the authors compared three different subjective quality assessment methodologies. A similar study was conducted in the context of image compression [16], where the authors compared four different subjective image quality assessment protocols. Nevertheless, the state of the art still lacks an extensive analysis on the performance of subjective quality assessment protocols targeting the assessment of recent image compression solutions, and in the range from high to nearly visually lossless quality. Moreover, the two methodologies presented in AIC-2 have not been thoroughly evaluated and mutually compared.

3 SUBJECTIVE EXPERIMENTS

Subjective visual scores have been collected and evaluated using three different subjective protocols, namely DSCQS, AIC-2 A, and Flicker test, by three different institutions, namely EPFL, UBI, and IST.

3.1 Test material: JPEG AIC-3 dataset

The JPEG AIC-3 dataset [32] was employed for the study. To limit the cost and complexity of the experiment, following the recommendations in ITU-T P.910 [13], only images *00002*, *00004*, *00005*, *00006*, and *00007* were adopted and cropped to a size of 620x800. A preview of the reference crops is provided in Figure 1. Moreover, only the images encoded with JPEG, JPEG 2000, VVC Intra, JPEG XL, and AVIF, and with only a limited number of quality levels were considered. Notably, for the DSCQS and AIC-2 A experiments, quality levels 2, 5, 8, and 10 of the dataset were considered, corresponding to JND values of -0.5, -1.25, -2, and -2.5 computed in a PC experiment [32]. For the Flicker experiment, given its increased sensitivity, only quality levels 1 and 2 were considered, corresponding to JND values of -0.25 and -0.5. It is important to emphasize that the JND values were obtained with a different protocol and therefore do not necessarily apply to the experiments conducted in this paper. In the rest of this paper, the images will be referred to as *Face*, *Artificial*, *Bird*, *Boat*, and *Night*.

According to the AIC-2 specifications, the experiments should include a number of *control images*, i.e. images unambiguously compressed and presenting artifacts that can be easily detected by an average viewer in a side-by-side experiment. These images are used to detect the subjects who misunderstood the objective of the experiment or were inattentive. Notably, five *control images* were included in the experiments, compressed with JPEG at quality 10 and presenting severe blocking artifacts and color distortions.

3.2 Graphical user interface

The graphical user interface (GUI) for the three experiments was implemented in MATLAB. Depending on the protocol, two or three images were presented side-by-side in the central part of the screen with a 1:1 pixel ratio. A pseudo-random order of presentation was applied, and care was devoted to avoid displaying the same content consecutively. Under the images, the voting mechanism was displayed according to the specific protocol. For the AIC-2 methodologies, the subjects had the possibility to select the desired image through a radio button or by clicking directly on the images. For the DSCQS protocol, two slide bars were placed under the left and right images in a vertical position. Five markers on the slide bars reported the labels *Excellent*, *Good*, *Fair*, *Poor*, and *Bad*, corresponding to proportional values from 100 to 0. The sliders were initialized to the value 0 (*Bad*), and the system did not allow moving to the subsequent step until at least one click was registered on each slide bar. For all the experiments, the subjects were required to press the button *'Next'* in order to proceed to the next stimulus. This button was hidden during the first four seconds in order to impose a minimum viewing time, but no upper limit was set. The background was set to a mid-dark gray tone (HEX #333333), and a blank interface with only the background color was displayed for 0.25 seconds between two consecutive steps.

Prior to the beginning of the experiments, the subjects were requested to conduct a training session in order to get accustomed to the goal of the experiment and grading scale. For the AIC-2 A and Flicker protocols, the training session included 6 examples and, in order to prevent misinterpretations of the task, feedback was provided to the subjects if an incorrect answer was given. For the DSCQS protocol, three examples with qualities *Excellent-Excellent*, *Excellent-Fair* and *Excellent-Bad* were presented during the training session.

3.3 Experimental setup

The experiments were conducted in the three institutions in similar environments with ambient light set to approximately 15 lux. Monitors of sizes 31.5" and 32" were adopted for the experiments, notably a Dell UltraSharp 32 4K U3219Q at EPFL, an Asus ProArt PA32UC at IST, and an EIZO ColorEdge CG318 at UBI. All monitors were configured to work with a Full HD resolution (1920x1080). Moreover, the monitors were calibrated by setting a D65 white point, and $120cd/m^2$ monitor brightness. The viewing distance was set to 62 cm. Written and oral instructions were provided prior to the beginning of the experiment. At each institution, 20 subjects participated in the experiments. All subjects passed a Snellen visual acuity test and Ishihara color vision test prior to the experiment.

At EPFL, the average age of the subjects was 22, with maximum age 25 and minimum age 18. At UBI, the average age of the subjects was 21.7, with maximum age 26 and minimum age 19. Finally, at IST, the average age of the subjects was 29.35, with maximum age 47 and minimum age 24.

The order of the experiments varied between the different institutions, notably: at EPFL and IST the adopted order was (1) DSCQS (2) AIC-2 A (3) Flicker test, while at UBI the order was (1) AIC-2 A (2) Flicker test (3) DSCQS. This variation was introduced to evaluate the impact of the experience of the subjects on the results.

4 STATISTICAL ANALYSIS

4.1 Results processing

The subjective scores were separately screened for outliers for each protocol and each laboratory. Outlier detection from ITU-R Rec. BT.500 [24] was applied to the subjective scores from the DSCQS experiment. For the AIC-2 methodologies, a subject was considered an outlier in case of wrong detection of one or more control images [11]. Two outliers were identified following this procedure, one for the DSCQS experiment at EPFL and one for the AIC-2 A experiment at UBI, and their scores were removed from the analysis.

The differential mean opinion scores (DMOS) and confidence intervals (CI) were then obtained for the DSCQS experiment. Considering that (D_i, R_i) denotes a stimulus pair where D_i is a distorted image and R_i is its corresponding reference, the differential score δ_{ij} for each subject j is computed through $\delta_{ij} = d_{ij} - r_{ij} + 100$, with d_{ij} and r_{ij} being the scores attributed by subject j to D_i and R_i, respectively. The DMOS value and CI are then obtained from Equations 1 and 2, where S_i corresponds to the standard deviation of the differential scores across all subjects.

$$DMOS_i = \frac{1}{N}\sum_{j=1}^{N}\delta_{ij} \tag{1}$$

$$CI_i = 1.96\frac{S_i}{\sqrt{N}} \tag{2}$$

The obtained DMOS values represent the difference in visual quality between the reference and distorted stimuli. It is also possible to ignore the score given to the reference and compute only the mean opinion score (MOS), which represents the visual appeal of the distorted stimulus only. This value is given by Equation 3 and can be related to the DMOS through Equation 4.

$$MOS_i = \frac{1}{N}\sum_{j=1}^{N}d_{ij} \tag{3}$$

$$DMOS_i = 100 + MOS_i - \frac{1}{N}\sum_{j=1}^{N}r_{ij} \tag{4}$$

For the protocols based on the AIC-2 specifications, the correct detection rate (CDR) is computed for each stimulus pair. In particular, the detection score c_{ij} receives value 1 if subject j correctly detects the reference R_i, and 0 otherwise. The CDR is then computed through Equation 5.

$$CDR_i = \frac{1}{N}\sum_{j=1}^{N}c_{ij} \tag{5}$$

The CDR evaluates the *visual fidelity* of the distorted stimuli. High CDR values indicate that most subjects were able to differentiate the distorted stimulus from the reference, implying low visual fidelity, and vice-versa. In the extreme case where no subject was able to detect any difference, the CDR should approach 0.5, since all individual detection scores are random.

4.2 AIC-2 probability analysis

The AIC-2 standard defines that all subjects should visualize each stimulus multiple times, resulting in individual CDR values for each subject. A distorted stimulus is then considered visually lossless if it differs from its corresponding references at 1 JND for all subjects. At this threshold level, a subject would correctly detect the reference half of the time, and answer randomly at the other half, resulting in a CDR approaching 0.75. In this experiment, although the subjects

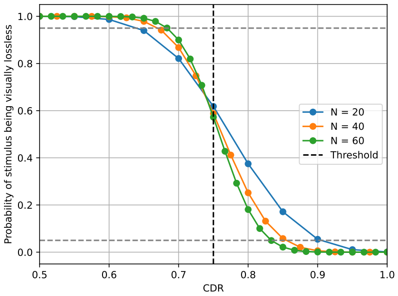


Figure 2: Probability of a stimulus being visually lossless

rate each stimulus only once, the threshold CDR value is kept at 0.75. This threshold should correspond to the level of 1 JND for the entire subjects pool rather than for individuals. However, it is also possible for this threshold to be reached even if the difference is higher than 1 JND, in the case that the amount of subjects that randomly assign false detection scores is higher than the amount of subjects with correct detection scores. Therefore, an analysis of the probability of a stimulus actually being visually lossless at a CDR of 0.75 is needed.

Let us consider that each stimulus pair (D_i, R_i) is evaluated by a total of N subjects, where N_a^i subjects are aware of the distortions and can fully differentiate both images, resulting in an individual detection score of $c_{ij} = 1$. The remaining N_b^i subjects are unaware of the impairments and are here denominated as *distortion-blind subjects*. The value of c_{ij} for each distortion-blind subject is a random variable with a 0.5 probability of being either 0 or 1. Assuming the threshold of 1 JND, a stimulus pair (D_i, R_i) can be considered as visually lossless if $N_b^i > N/2$.

However, the value for N_b^i cannot be directly derived from CDR_i. Instead, the total number of subjects who correctly detected the reference $N_{cd}^i = CDR_i \times N$ is the sum between N_a^i and the amount of distortion-blind subjects who randomly assigned the correct score. This second term follows a binomial distribution $B(N_b^i, 0.5)$. The probability of N_{cd}^i, assuming a given N_b^i, can therefore be obtained by Equation 6, with $n = N_b^i$, $k = N_{cd}^i - N_a^i$ and $p = 0.5$:

$$P(N_{cd}^i | N_b^i) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6}$$

In order to derive the probability $P(N_b^i | N_{cd}^i)$ of the amount of distortion-blind subjects being N_b^i given the observed N_{cd}^i from the experiment, the Equation 7 derived from Bayes' theorem is employed. The prior probability $P(N_b^i)$ is set to a constant value of $\frac{1}{N+1}$, since all possible values of N_b^i are initially considered as equally likely.

$$P(N_b^i | N_{cd}^i) = \frac{P(N_{cd}^i | N_b^i) P(N_b^i)}{\sum_{x=0}^{N} P(N_{cd}^i | N_b^i = x) P(N_b^i = x)} \tag{7}$$

Finally, the probability p_{VL}^i that the stimulus pair (D_i, R_i) is visually lossless is given by $P(N_b^i > N/2)$, which is expressed in Equation 8.

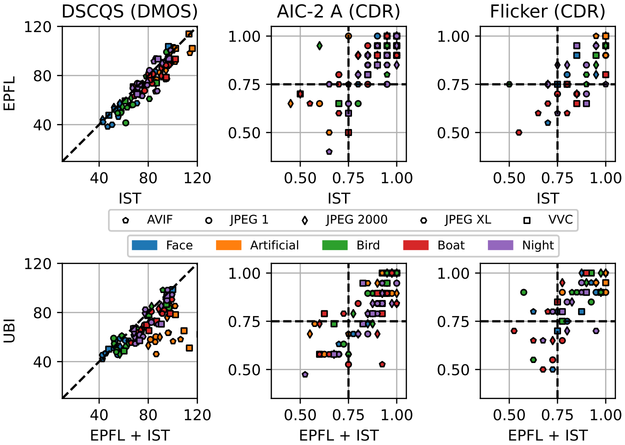


Figure 3: Comparison of results between all laboratories.

$$p_{VL}^i = \sum_{x=\lfloor \frac{N}{2} \rfloor + 1}^{N} P(N_b^i = x | N_{cd}^i) \tag{8}$$

In order to evaluate the reliability of experiments following the AIC-2 standard, the values of p_{VL} given all possible observed CDR values for different total amounts of subjects N is displayed in Figure 2. The proposed probability model results in a p_{VL} value higher than 0.5 at the adopted CDR threshold of 0.75 for all evaluated values of N, which decrease rapidly for higher CDR values. However, the obtained $p_{VL}(CDR = 0.75)$ is still far from 1, leaving about a 40% chance that the stimulus is actually not visually lossless. In order to define a 5% confidence interval for the classification of the stimulus, two dashed lines are displayed at $p_{VL} = 0.05$ and $p_{VL} = 0.95$. With $N = 40$, corresponding to the number of subjects from two out of the three laboratories of the present experiment, a CDR as low as 0.675 is needed to achieve p_{VL} next to 95%, and a value of approximately 0.85 is needed to affirm that a stimulus has only 5% chance of being visually lossless. If the outcome of the experiment lies between these values, then it is not possible to affirm at a 5% confidence interval whether the stimulus is visually lossless or not. If however the results from only one laboratory are used individually, then this interval would be enlarged from 0.65 to 0.9. On the other hand, having 20 more subjects would reduce the confidence interval from 0.683 to 0.833. In general terms, more subjects should be added to allow for higher reliability of the test, which results however in higher costs. In subsequent analysis, the CDR threshold of 0.75 is kept, always considering that a significant distance from this value is needed for robust conclusions.

5 RESULTS AND DISCUSSION

5.1 Comparison between laboratories

A comparison between the results obtained from each laboratory is conducted to determine if the results are strongly correlated. Since the experiments were conducted using the same order at EPFL and IST, their results are compared first.

Scatter plots comparing the DSCQS DMOS and AIC-2 A and Flicker test CDR are presented in Figure 3 (upper). Results show that the DMOS values in the DSCQS experiment from both laboratories agree, with a Pearson linear correlation coefficient value of 0.939, despite the slight tendency of IST subjects to accord higher differential scores than EPFL. The correlation values were however lower for the remaining experiments, with 0.750 in AIC-2 A and 0.635 in the Flicker test. Differently from the DMOS, the CDR values can be considered as being quantized with a step value of $\frac{1}{N}$, where N is the total number of subjects. Since only 20 subjects were employed in each laboratory, the CDR can only vary with a step of 0.05, which corresponds to 10% of the total range given that all values are higher than 0.5. Therefore, random variations in the data have a higher impact on the correlation between laboratories. This implies that, in order to maintain the same correlation values between different experiments, either a larger number of subjects should participate in the experiment, or each subject should rate each stimulus multiple times. Any of the alternatives would increase the time duration of the experiment, and therefore the associated cost. These results suggest therefore that the protocols based on the AIC-2 standard are more expensive than the DSCQS.

Since the experiments were conducted in the same order at EPFL and IST, and a high correlation was achieved for the DSCQS experiment, the subjective scores from both laboratories were merged for all subsequent analyses. Figure 3 (lower) compares the result from merged scores between EPFL and IST against UBI. Contrary to what was observed in the previous comparison, the correlation coefficient of DSCQS results reaches a low value of only 0.684. The scatter plot reveals that the DMOS values are mainly different for *Artificial*, where UBI subjects tend to attribute much lower DMOS scores. Moreover, many distorted stimuli from *Bird* and *Face* received higher scores at UBI, which was not the case for *Night* and *Boat*. It is clear that the order of the experiments had a high impact on the DMOS scores obtained from the DSCQS experiment, and this is however not the case for the remaining protocols, which achieved a correlation of 0.817 for AIC-2 A and 0.671 for the Flicker test. The fact that these values are actually higher than the previous comparison between EPFL and IST corroborates the conclusion that the correlation is negatively impacted by coarser quantization steps of the CDR. Here, since the results from EPFL and IST are joined, the total number of subjects is doubled and its corresponding quantization step for the CDR of $\frac{1}{N}$ is halved. These results indicate that the amount of subjects affects the consistency of the CDR values.

Considering the impact that the order of experiments had on DSCQS results, scores obtained at UBI are not included in the subsequent analysis.

5.2 DSCQS experiment

Figure 4 depicts the DMOS values with CI obtained in the experiment separately for each original content. In most cases, the behavior of the DMOS is monotonically decreasing with the quality value for each codec, which is expected given the way that the quality levels were defined. *Artificial* is an outlier since the DMOS at quality 10 (highest distortions) is higher than quality 2 (lowest distortions) for both AVIF and VVC Intra. More surprisingly, these stimuli receive DMOS higher than 100, which indicates that the distorted images are perceived as having higher quality than the original. This observation can be explained by the denoising mechanisms of such codecs which exert particularly strong effects at lower bitrates. Since the reference image contains a high level of noise, subjects associated noise attenuation with an increase in *visual appeal*. However, the CIs of such stimuli are among the largest of the dataset, indicating that the level of agreement on how this denoising effect impacts the visual appeal is low.

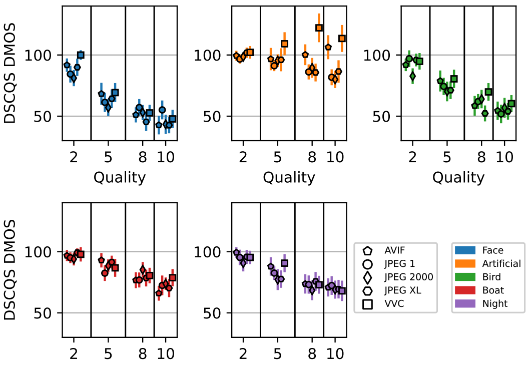


Figure 4: DSCQS DMOS results

Table 1: No reference metrics computed on reference images

	Face	*Artificial*	*Bird*	*Boat*	*Night*
BRISQUE	19.33	43.46	11.38	29.74	31.29
NIQE	2.91	5.49	3.01	3.20	4.24
PIQE	20.64	29.78	15.03	29.09	40.86

The plots also indicate that a higher range of DMOS values is achieved by using *Face* and *Bird*, allowing to better discriminate between different quality levels, with maximum values going from around 100 to less than 50. This range is moderately reduced for *Night* and *Boat*, and even more for *Artificial* without considering the enhancing effect of VVC Intra and AVIF. Visual inspection of the reference images in Figure 1 suggests that *Face* and *Bird* have the highest visual quality, where fine details in skin, hair, and feathers can be discerned. In *Boat* and *Night* however, the distinction of finer details is hindered due to motion blur and acquisition noise. *Artificial* is heavily different from the remaining, not only because it is not a natural image, but also due to the high level of added noise and edges being perceived as jagged lines. Such inspection suggests that *Face* and *Bird* have higher quality than the remaining reference images.

In order to further investigate this assumption, three widely used non-reference quality metrics, namely BRISQUE [20], NIQE [21] and PIQE [35], were computed directly on the reference images and reported in Table 1. While both *Face* and *Bird* present the lowest values for all three evaluators, two out of three metrics give the highest value for *Artificial*, as it presents statistics far from those

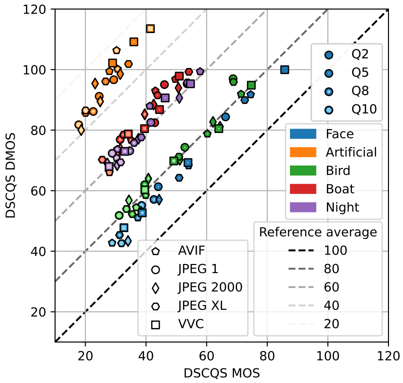


Figure 5: Scatter plot of DSCQS DMOS against MOS

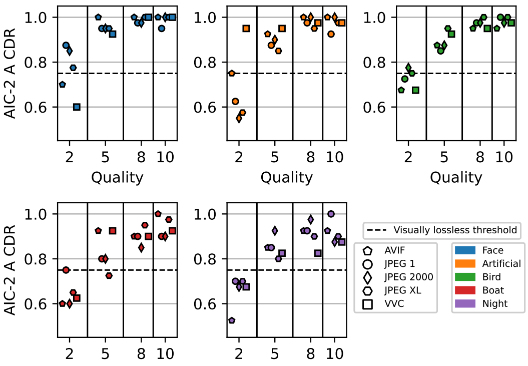


Figure 6: AIC-2 A CDR results

in natural images. Such results corroborate previous conclusions derived from visual inspection.

The effect of the quality of the reference image on the DSCQS results is more clearly observed in Figure 5, where the DMOS of each stimulus is plotted against the MOS obtained by disregarding the scores given to the reference images in the same experiment. According to Equation 4, the MOS and DMOS are related to the average score attributed to the reference image of the stimulus pair. Figure 5 takes that into account by including dashed lines that correspond to constant reference average scores. The plot indicates that rating behavior is divided between three distinct clusters. The references *Face* and *Bird* received similar average scores of approximately 80, with the former achieving slightly higher quality. *Boat* and *Night* received lower average scores lying mostly between 50 and 60, and *Artificial* has an average reference score of less than 40, with some stimuli even approaching 20.

The plot in Figure 5 strongly suggests that the selection of the reference images plays a crucial role in the outcome of DSCQS experiments. This is due to the fact that this experiment targets the assessment of the *visual appeal* of the distorted images, rather than their *fidelity* to the reference. It is observed that the distorted stimuli from references with higher visual quality achieve higher differentiation between quality levels, which is a desirable feature when the performances of different codecs are being evaluated at similar bitrates. Therefore, these results indicate that images with high visual quality should always be included in datasets comparing the rate-distortion performance of coding methods with the DSCQS experiment.

Moreover, Figure 5 reveals that the high DMOS values for *Artificial* were only possible since the average reference scores were low. In fact, the stimulus with the highest DMOS value has a MOS value of only 40, indicating that its absolute quality is low despite its DMOS being above 100.

5.3 AIC-2 experiments results

The CDR results obtained in the AIC-2 A experiment are presented in Figure 6. It is observed that the majority of stimuli under the CDR threshold were compressed with quality level 2. Moreover, no clear difference between CDR values is observed at higher quality levels, especially between levels 8 and 10. These results are in line with the experiment in [32] where these quality levels were defined since all levels higher than 4 should differ from the original by 1 JND or more.

However, some stimuli from quality 2 are not labeled as visually lossless with a significant margin from the threshold, even if their difference to the reference image was measured as being only 0.5 JND in [32]. One possible reason for the difference in the results between both experiments is that the original JND values were obtained through the interpolation of a curve fitted to the observed data, relying on the precision of values that are susceptible to errors. Moreover, the experiment setup from [32] and from the current study are not equivalent: while in [32] each subject was asked to select the image with the highest visual quality, in the AIC-2 A experiment they are asked to select the closest match to the original.

Therefore, while [32] assessed if one image is more visually appealing than the other, the experiment in this paper evaluated the visual fidelity of the distorted images. It is possible that the same subject could identify an objective difference between both images of the stimulus pair in the latter case, but not be able to select an image with higher visual quality in the former. Therefore, given the way that the JND levels were defined, subjects are expected to be more strict in the AIC-2 A experiment than in [32].

Results from the Flicker test are presented in Figure 7. If quality level 2 was enough to generate CDR values under the threshold for most stimuli in AIC-2 A, the same distortion is much more easily perceived on the Flicker test. Even for quality level 1, which corresponds to a difference of only 0.25 JND in [32], the subjects were able to differentiate the stimulus pair at high CDR for some stimuli. A higher variability between different content is also observed. For *Artificial*, no stimulus was considered as visually lossless, mainly due to the presence of high levels of noise in the original image. Due to the high spatial frequencies of the added noise, this image is extremely hard to compress with high fidelity levels. Moreover, while slight modifications may pass undetected in a side-by-side comparison, they are much more easily noticed between two interleaved images. Among the remaining stimuli, *Face* and *Bird* have on average larger CDR values, suggesting that subjects are in general

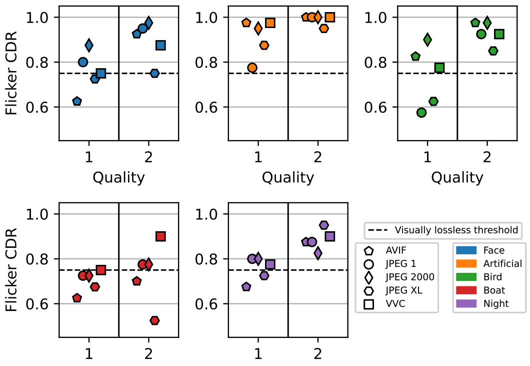


Figure 7: Flicker CDR results

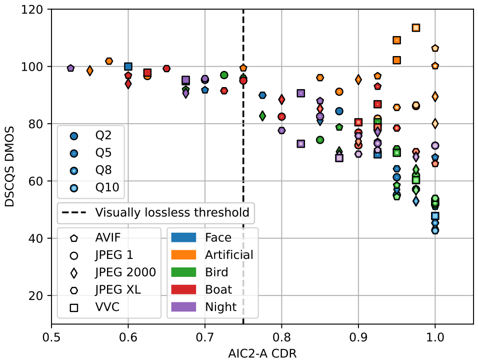


Figure 8: Result comparison between DSCQS and AIC-2 A

more able to detect impairments in images with higher qualities. Such variability between content also indicates JND levels defined through side-by-side comparison do not translate easily to the Flicker test.

5.4 Protocol comparison

Since the Flicker test has a higher discerning ability and may detect perceptual differences even in stimuli with excellent perceptual quality, it is difficult to compare the outcome of this experiment to DSCQS and AIC-2 A. On the other hand, the same quality levels were used in the datasets evaluated by the two latter protocols, allowing a straightforward comparison.

A scatter plot comparing the CDR values obtained in the AIC-2 A experiment against the DMOS values from DSCQS is displayed in Figure 8. There is an evident difference between the points on the left of the CDR threshold and the points on its right. All stimuli considered as visually lossless received high DMOS, with minimum values ranging around 90. Assuming that the number of distortion-blind subjects for a given stimulus is the same for the DSCQS and AIC-2 A protocols, and assuming that such subjects attribute the same score for both images of the stimulus pair during the DSCQS experiment, then more than half of the differential scores of that stimulus should be equal to 100. Considering that the increased *visual fidelity* of such stimuli also corresponds to high *visual appeal* for the remaining distortion-aware subjects, the overall DMOS should remain near 100, which is observed here.

The stimuli with CDR higher than the threshold present however a different behavior. While, in general, the DMOS decreases with higher CDR for most images, stimuli from *Artificial* received high differential scores even when differences between both stimuli could be perceived. These findings indicate that, although there is a correlation between *visual fidelity* and *visual appeal* for most of the test images, the DSCQS protocol cannot be used to determine if a distorted stimulus is visually lossless.

Overall, the analysis conducted in this paper shows different weaknesses of the considered standardized protocols:

- The DSCQS methodology is highly influenced by the quality of the reference image and is unable to differentiate between images with slight differences in visual quality,
- The AIC-2 A methodology is not able to discriminate between images with visual quality lower than (nearly) visually lossless,
- The Flicker test is too sensitive and does not provide any meaningful result in the quality range of interest of this paper.

6 CONCLUSIONS

In this study, three different subjective assessment protocols, namely DSCQS and the two variants of the AIC-2 standard, are comprehensively assessed and compared. A dataset of compressed images distorted using multiple codecs was evaluated on experiments conducted in three separate laboratories. The outcome from different laboratories was first analyzed, revealing that the DSCQS experiment was heavily affected by the order in which the experiments were conducted, or more specifically by the subject experience. Moreover, the quality of the reference image was observed to strongly impact the range of obtained DMOS scores. The two experiments following the AIC-2 protocols evidence that the definition of visually lossless is largely dependent on the visualization conditions. The comparison between protocols also indicates that, while there is a correlation between *visual appeal* measured by DSCQS and *visual fidelity* measured by AIC-2 protocols, the two concepts are not interchangeable, and stimuli with high appeal scores may present low fidelity. This study, therefore, suggests the need for novel subjective image quality assessment protocols robust in quality range from high to nearly visually lossless.

ACKNOWLEDGMENTS

EPFL-affiliated authors acknowledge support from the Swiss National Scientific Research project under grant number 200020_207918. UBI/IT-affiliated authors acknowledge support from FCT/MCTES through national funds and when applicable co-funded EU funds under the projects UIDB/50008/2020. IST/IT-affiliated authors acknowledge support from the FCT project DARING with reference PTDC/EEI-COM/7775/2020.

REFERENCES

[1] Jyrki Alakuijala, Ruud Van Asseldonk, Sami Boukortt, Martin Bruse, Iulia-Maria Comșa, Moritz Firsching, Thomas Fischbacher, Evgenii Kliuchnikov, Sebastian Gomez, Robert Obryk, et al. 2019. JPEG XL next-generation image compression architecture and coding tools. In *Applications of Digital Image Processing XLII*, Vol. 11137. SPIE, 112–124.

[2] Robert S Allison, Kjell Brunnström, Damon M Chandler, Hannah R Colett, Philip J Corriveau, Scott Daly, James Goel, Juliana Y Long, Laurie M Wilcox, Yusizwan M Yaacob, et al. 2018. Perspectives on the definition of visually lossless quality for mobile and large format displays. *Journal of Electronic Imaging* 27, 5 (2018), 053035–053035.

[3] Robert S Allison, Laurie M Wilcox, Wei Wang, David M Hoffman, Yuqian Hou, James Goel, Lesley Deas, and Dale Stolitzka. 2017. 75-2: Invited paper: Large scale subjective evaluation of display stream compression. In *SID Symposium Digest of Technical Papers*, Vol. 48. Wiley Online Library, 1101–1104.

[4] Sharon Ayzik and Shai Avidan. 2020. Deep image compression using decoder side information. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 699–714.

[5] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. 2016. End-to-end Optimized Image Compression. In *International Conference on Learning Representations*.

[6] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. 2021. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 10 (2021), 3736–3764.

[7] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. 2020. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7939–7948.

[8] David M Hoffman and Dale Stolitzka. 2014. A new standard method of subjective assessment of barely visible image artifacts and a new public database. *Journal of the Society for Information Display* 22, 12 (2014), 631–643.

[9] Yueyu Hu, Wenhan Yang, Zhan Ma, and Jiaying Liu. 2021. Learning end-to-end lossy image compression: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4194–4211.

[10] ISO/IEC 15948:2004. 2004. Information technology — Computer graphics and image processing — Portable Network Graphics (PNG): Functional specification.

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

[12] ISO/IEC TR 29170-1:2017. 2017. Information technology — Advanced image coding and evaluation — Part 1: Guidelines for image coding system evaluation.

[13] ITU-T Recommendations P.910. 2022. Subjective video quality assessment methods for multimedia applications.

[14] Haojie Liu, Tong Chen, Peiyao Guo, Qiu Shen, Xun Cao, Yao Wang, and Zhan Ma. 2019. Non-local attention optimized deep image compression. *arXiv preprint arXiv:1904.09757* (2019).

[15] Jinming Liu, Heming Sun, and Jiro Katto. 2023. Learned image compression with mixed transformer-cnn architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14388–14397.

[16] Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. 2012. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, Vol. 31. Wiley Online Library, 2478–2491.

[17] David McNally, Tim Bruylants, Alexandre Willème, Touradj Ebrahimi, Peter Schelkens, and Benoit Macq. 2017. JPEG XS call for proposals subjective evaluations. In *Applications of Digital Image Processing XL*, Vol. 10396. SPIE, 109–119.

[18] Hui Men, Hanhe Lin, Mohsen Jenadeleh, and Dietmar Saupe. 2021. Subjective image quality assessment with boosted triplet comparisons. *IEEE Access* 9 (2021), 138939–138975.

[19] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. 2019. Practical full resolution learned lossless image compression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10629–10638.

[20] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* 21, 12 (2012), 4695–4708.

[21] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* 20, 3 (2012), 209–212.

[22] Sanjida Sharmin Mohona, Domenic Au, Onoise Gerald Kio, Richard Robinson, Yuqian Hou, Laurie M Wilcox, and Robert S Allison. 2020. Subjective assessment of stereoscopic image quality: the impact of visually lossless compression. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[23] Margaret H Pinson and Stephen Wolf. 2003. Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing 2003*, Vol. 5150. SPIE, 573–582.

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

[25] Report ITU-R BT.1082-1. 1990. Studies toward the unification of picture assessment methodology. (1990).

[26] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi. 2001. The JPEG 2000 still image compression standard. *IEEE Signal processing magazine* 18, 5 (2001), 36–58.

[27] Jon Sneyers, Elad Ben Baruch, and Yaron Vaxman. 2023. CID22: Large-Scale Subjective Quality Assessment for High Fidelity Image Compression. *IEEE MultiMedia* (2023). Submitted manuscript. Online: https://cloudinary.com/labs/cid22 (accessed: April 2023).

[28] Jon Sneyers and Pieter Wuille. 2016. FLIF: Free lossless image format based on MANIAC compression. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 66–70.

[29] Dale F Stolitzka, Peter Schelkens, and Tim Bruylants. 2017. New procedures to evaluate visually lossless compression for display systems. In *Applications of Digital Image Processing XL*, Vol. 10396. SPIE, 98–108.

[30] Aishwarya Sudhama, Matthew D Cutone, Yuqian Hou, James Goel, Dale Stolitzka, Natan Jacobson, Robert S Allison, and Laurie M Wilcox. 2018. 85-1: Visually Lossless Compression of High Dynamic Range Images: A Large-Scale Evaluation. In *SID Symposium Digest of Technical Papers*, Vol. 49. Wiley Online Library, 1151–1154.

[31] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. 2012. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology* 22, 12 (2012), 1649–1668.

[32] Michela Testolina, Vlad Hosu, Mohsen Jenadeleh, Davi Lazzarotto, Dietmar Saupe, and Touradj Ebrahimi. 2023. JPEG AIC-3 Dataset: Towards Defining the High Quality to Nearly Visually Lossless Quality Range. In *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 55–60.

[33] Michela Testolina, Evgeniy Upenik, Jon Sneyer, and Touradj Ebrahimi. 2022. 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, 90–98.

[34] Evgeniy Upenik, Michela Testolina, João Ascenso, Fernando Pereira, and Touradj Ebrahimi. 2021. Large-scale crowdsourcing subjective quality evaluation of learning-based image coding. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 1–5.

[35] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. 2015. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*. IEEE, 1–6.

[36] Gregory K Wallace. 1992. The JPEG still picture compression standard. *IEEE transactions on consumer electronics* 38, 1 (1992), xviii–xxxiv.

[37] Yueqi Xie, Ka Leong Cheng, and Qifeng Chen. 2021. Enhanced invertible encoding for learned image compression. In *Proceedings of the 29th ACM international conference on multimedia*. 162–170.