Towards JPEG AIC part 3: visual quality assessment of high to visually-lossless image coding

Michela Testolina[a], Evgeniy Upenik[b], Jon Sneyers[c], and Touradj Ebrahimi[a]

[a]Multimedia Signal Processing Group (MMSPG), Ecole Polytechnique Fédérale de Lausanne
[b]Audiovisual Technology Laboratory, Munich Research Center, Huawei Technologies
[c]Media Technology Research, Cloudinary

ABSTRACT

Due to the increasing number of pictures captured and stored every day by and on digital devices, lossy image compression has become inevitable to limit the needed storage requirement. As a consequence, these compression methods might introduce some visual artifacts, whose visibility depends on the chosen bitrate. Modern applications target images with high to near-visually lossless quality, in order to maximize the visual quality while still reducing storage space consumption. In this context, subjective and objective image quality assessment are essential tools in order to develop compression methods able to generate images with high visual quality. While a large variety of subjective quality assessment protocols have been standardized in the past, they have been found to be imprecise in the quality interval from high to near-visually lossless. Similarly, an objective quality metric designed to work specifically in the mentioned range has not been designed yet. As current quality assessment methodologies have proven to be unreliable, a renewed activity on the Assessment of Image Coding, also referred to as JPEG AIC, was recently launched by the JPEG Committee. The goal of this activity is to extend previous standardization efforts, i.e. AIC Part 1 and AIC Part 2 (also know as AIC-1 and AIC-2), by developing a new standard, known as AIC Part 3 (or AIC-3). Notably, the goal of the activity is to standardize both subjective and objective visual quality assessment methods, specifically targeting images with quality in the range from high to near-visually lossless. Two Draft Calls for Contributions on Subjective Image Quality Assessment[1,2] were released, aiming at collecting contributions on new methods and best practices for subjective image quality assessment in the target quality range, while a Call for Proposals on Objective Image Quality Assessment is expected to be released at a later date. This paper aims at summarizing past JPEG AIC efforts and reviewing the main objectives of the future activities, outlining the scope of the activity, the main use cases and requirements, and call for contributions. Finally, conclusions on the activity are drawn.

Keywords: Image quality assessment, image compression, JPEG AIC, subjective quality assessment

1. INTRODUCTION

Figure 1: Previous AIC standards and covered image quality ranges.

Recent advances in digital cameras, broadband Internet, wider storage devices, and image coding technologies have made high-quality imaging feasible and desirable. Traditional image quality assessment standards, such as those described in ITU-T Rec. BT.500[3] and AIC-1,[4] are mainly suitable for evaluating the *visual appeal* of images (i.e. how obvious and/or annoying the artifacts are) rather than their *visual fidelity* (i.e. how true the

E-mails: michela.testolina@epfl.ch, evgeniy.upenik@huawei.com, jon@cloudinary.com, touradj.ebrahimi@epfl.ch

(a) Original (b) AVIF - ↑ appeal, ↓ fidelity (c) JPEG - ↓ appeal, ↑ fidelity

(d) Original (e) AVIF - ↑ appeal, ↓ fidelity (f) JPEG - ↓ appeal, ↑ fidelity

Figure 2: Example of images presenting different levels of visual fidelity and appeal. Notably, the image showing a detail of a human face presents a higher appeal when compressed with AVIF, as the smoothness of the skin's texture makes the image more appealing to the human eye. However, the same image compressed with JPEG presents a higher fidelity, as the details of the eyebrows are better represented. Likewise, the image representing a wall with bricks presents a higher fidelity when compressed with JPEG, as AVIF degrades the texture of the wall.

distorted image is to its original version).[5] As an example, subtle variations in the colors might be hard to detect in a traditional side-by-side quality assessment experiment, but might lead to discontinuities between a picture and the background or an adjacent image in web use cases. Moreover, in e-commerce websites, reducing the size of images is crucial to minimize the loading time, while the details of the products should be accurately reproduced in order to guarantee an experience as close as possible to the one in physical stores.

An example of images showing different levels of visual fidelity and appeal is presented in Figure 2. Notably, the image representing a detail of a human face presents higher appeal when compressed with AVIF, as the smoothness of the skin is more appealing to the human eye. On the other hand, the same image compressed with JPEG presents a higher fidelity as the details of the eyebrows are better reconstructed by this codec. Similarly, the image showing a wall with bricks compressed with JPEG presents a higher fidelity, as the smoothing artifact caused by the AVIF compression degrades the fine textures of the image. At the same time, blocking artifacts introduced by the JPEG compression decreases the visual appeal of the image compared to the AVIF-compressed one.

As will be reported later in this paper, the grading scale of methods described in BT.500,[3] e.g. in the double stimulus impairment scale (DSIS) assessment method, are designed to assess how *annoying* a visual artifact is, and imply that images which present artifacts that are noticeable but not 'annoying' do not justify a low rate. Moreover, methods in BT.500[3] often compare images in a side-by-side manner, which makes it hard to spot even significant differences. This makes the methods in BT.500[3] more suitable to evaluate the *appeal* of images, rather than their fidelity to the original.

(a) Original (b) HEVC Intra Q25 (c) Difference between Y components

Figure 3: Example of two images with a small difference, indistinguishable for average non-expert viewers in a side-by-side comparison, but detectable by a flicker test.

Historically, the difference between fidelity and appeal was a relatively minor concern, as for most types of degradations low fidelity implied low appeal. This is no longer the case for modern codecs like HEIC or AVIF which are much better capable of 'hiding' compression artifacts even when severely degrading the image. The upcoming learning-based codecs can be expected to be even more capable of producing appealing images with low fidelity. Therefore, more sensitive assessment methodologies are required in order to evaluate the quality range where the appeal is high enough to saturate when measured with the methods described in AIC-1,[4] yet with lower fidelity.

The methods described in AIC-2[6] are based on an extremely sensitive flicker test that is able to detect even the slightest visual distortion, making them highly sensitive and suitable for evaluating the visual *fidelity* of images. An example of images whose differences are hardly noticeable in a side-by-side test, but detectable when subjected to a flicker test, is presented in Figure 3.

In this regard, the JPEG Committee has launched an activity with the goal of developing new standards for subjective and objective image quality assessment, known as AIC-3, sensitive and discriminative in the range from high quality to near-visually lossless quality as illustrated in Figure 1. A "First Draft Call for Contributions on Subjective Image Quality Assessment Methods"[1] and successively a "Second Draft Call for Contributions on Subjective Image Quality Assessment Methods"[2] were released during the 95$^{\text{th}}$ and the 96$^{\text{th}}$ JPEG meetings respectively, while a "Final Call for Contributions on Subjective Image Quality Assessment Methods" is expected to be released in October 2022. The goal of this call is to collect new methods and best practices for subjective quality assessment of images in the range from high quality to near visually lossless quality.

2. RELATED WORK

Previous efforts in the JPEG AIC activity allowed the development and release of two specifications, namely JPEG AIC-1 and JPEG AIC-2:

- **ISO/IEC TR 29170-1:2017 (JPEG AIC-1)**[4] defines guidelines for the assessment of image coding systems, both subjectively and objectively. Particular attention was devoted to the works conducted by the International Telecommunication Union, where documents ITU-R BT.500[3] and ITU-R BT.1082[7] are reviewed in the context of image quality assessment.
- **ISO/IEC 29170-2:2015 (JPEG AIC-2)**[6] proposes new evaluation procedures for subjective visual quality assessment of images with near-visually lossless to visually lossless qualities. Notably, two different methods have been proposed in Annex A and Annex B of the standard.

In the next sections, a wider review of the JPEG AIC-1 and JPEG AIC-2 specifications, as well as other relevant subjective image quality assessment methods, is proposed.

2.1 Overview of ITU-R BT.500

Recommendation ITU-R BT.500[8] proposes a number of methods for the subjective evaluation of image quality. The methods described in the document are mainly double stimulus (DS), meaning that the quality is assessed by considering two stimuli at the time. This is the case for assessing the artifacts introduced by compression, as the original uncompressed image is generally available. The predominant and most commonly used methods proposed in the standard can be summarized as follows:

Double Stimulus Impairment Scale (DSIS): where two stimuli are presented to a number of observers in a simultaneous or sequential fashion, asking them to rate the level of impairment of the compressed image with respect to the original. Notably, the position of the original stimulus is always disclosed to the observers. The grading scale is discrete, with values corresponding to (1) very annoying, (2) annoying, (3) slightly annoying, (4) perceptible but not annoying, and (5) imperceptible. This method is widely used to evaluate subjectively the quality of compressed images, e.g. Ascenso et al.[9] where a number of learning-based compression techniques where assessed and compared to each other.

Double-Stimulus Continuous Quality-Scale (DSCQS): in which, the overall quality of a pair of displayed stimuli is rated by a group of observers using a continuous quality rating scale. The reference stimulus is displayed in a random position, unknown to the subjects. This method is slower than DSIS, as the observers are requested to rate the quality of two stimuli at each step. However, this method is effective in evaluating learning-based compression methods where, at the highest bitrates, the compressed image might present a higher visual appeal than the original thanks to possible pre- and post-processing algorithms. As an example, the DSCQS method was utilized in the subjective visual quality assessment experiment conducted to evaluate the submissions to the JPEG AI Call for Evidence, co-organized in conjunction with the IEEE MMSP'2020 Challenge on Learning-Based Image Coding.[10,11]

Pair Comparison (PC): consists in comparing, at each step, the visual quality of one stimulus the other , used as a reference, where all combinations of the stimuli must be covered. The grading scale is discrete and the possible grades are (-3) much worse, (-2) worse, (-1) slightly worse, (0) the same, (1) slightly better, (2) better, (3) much better. However, variants with grading scales including three (Better, The same, Worse) or even two (Better, Worse) levels are possible. The PC method is the approach, among those reviewed in this paper, that has the largest number of steps, and therefore is the most time-consuming. While this method is the most accurate in evaluating the performance of different compression algorithms, the bitrates of the test stimuli are required to be as close as possible in order to guarantee a fair comparison. Partial comparison among stimuli is possible in order to reduce the duration of the tests.

2.2 Overview of ITU-R BT.1082

ITU-R BT.1082[12] reviews additional subjective quality assessment methods, with the goal of covering certain shortcomings observed in ITU-R BT.500. In particular, the main observed weaknesses are the following:

- The methods in BT.500 can be used only if a high-quality reference image is available.
- The difference between the various values of the quality scales is not necessarily uniform.
- Different subjects may associate different meanings to the descriptors of the quality levels.
- Double stimulus methods can be too time-consuming.

While ITU-R BT.1082 reviews six additional subjective methods, only one, i.e. ratio scaling, is summarized in this paper.

Method	Advantages	Disadvantages
DSIS	Fast and intuitive	Does not cover a wide range of qualities
DSCQS	More accurate	Time consuming
PC	Intuitive	Highly time consuming
Ratio scaling	More accurately represents the subjective opinion	Does not use the information of the original, time consuming

Table 1: Advantages and disadvantages of the main subjective quality methods in ITU-R BT.500[8] and ITU-R BT.1082.[12]

Ratio Scaling: consists in presenting a sequence of stimuli to the observers, asking them to numerically grade each image based on the previous one. The numerical scale can be freely decided by the subjects and with no limit to the range. Each stimulus is presented twice to the subjects. The advantage of this method is that it allows not only to determine the ranking of the compressed stimuli, but also answers the question of "how much better" one stimulus is when compared to another. Additionally, this method is still applicable even when a reference is not available.

A summary of the methods introduced in this section is available in Table 1, presenting the main advantages and disadvantages of each method.

2.3 Overview of JPEG AIC-2

The subjective quality assessment methods and standards presented above were mainly designed for web-quality applications, i.e. applications with a limited or variable bitrate requirement, where visual artifacts caused by compression are usually easily perceived by an average viewer. In recent years, the number of applications that target high visual qualities is increasing.[13] As an example, limited memory and bandwidth consumption is no longer the primary requirement thanks to cheap and large storage devices, wide band communication channels and cloud storage services. Accordingly, users demand compression methods that maximize the visual quality of images and do not tolerate compromises to reach lower bitrates at the expense of lower quality. In this context, the standard subjective quality assessment methods presented in Section 2.1 and 2.2 are not suitable anymore, due to the difficulties in their ability to assess subtle artifacts. For instance, it is usually difficult to detect slight shifts in the colors when images are presented alone or even side-by-side.

To address this issue, the JPEG committee has released the standard ISO/IEC 29170-2 (AIC Part-2),[6,14] which includes specification for subjective visual quality assessment of images in the range spanning from nearly lossless to lossless visual qualities. In particular, two different methods have been proposed in Annex A and Annex B of the standard, known as AIC-2 A and AIC-2 B.

AIC-2 A: where a total of three stimuli, i.e. two distorted test images along with the original image, are presented to a group of viewers. The subjects are asked to select the closest stimulus to the original within 4 seconds.

AIC-2 B, or "Flicker" test: where two stimuli, corresponding to an original and a distorted version constructed from the test image, are presented to subjects in a side-by-side fashion. Notably, the distorted version is obtained by interleaving the test image with the original image at a certain frequency, while the original stimulus is obtained by interleaving the original image with itself, creating a static sequence. In the case where the test image presents perceivable degradation, the distorted version will appear "flickering". The position of the original and "flickering" stimuli is random selected and unknown to subjects. The subjects are asked to choose which of the two stimuli presents flickering. If the image presents visible distortions, the flickering will be visible by the human eye and the subjects can detect the flickering image correctly. If the distortions in the test image are not perceivable by the human eye, the subjects are not able to correctly detect the flickering image and will answer randomly. Figure 4 depicts the data preparation process for the flicker test.

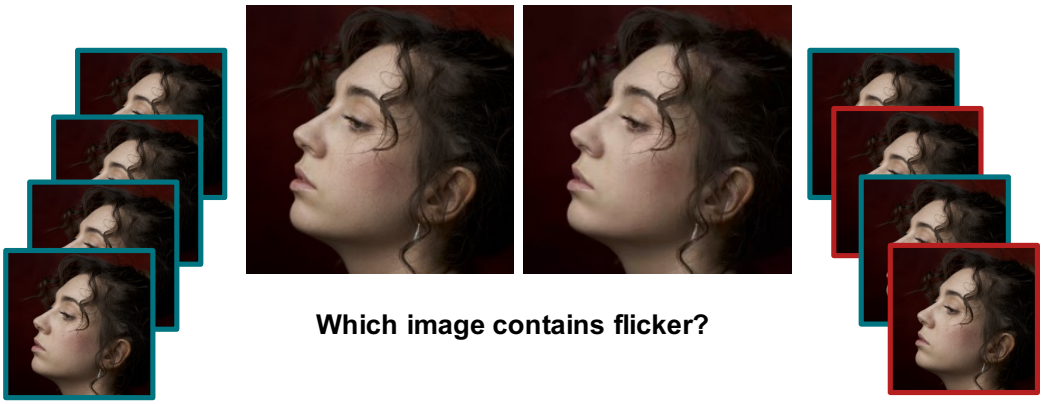

Figure 4: Visual representation of the data preparation for the JPEG AIC-2 flicker test. Notably, one sequence is created by interleaving the original image with itself (creating a static sequence), while the second is created by interleaving the original and the test image. If the quality of the distroted image is not visually lossless, the second sequence will flicker when compared to the other stimulus.

3. OVERVIEW OF JPEG AIC-3

The scope of AIC activities is to specify standards or best practices with respect to subjective and objective image quality assessment methodologies that cover a range from high quality to near-visually lossless quality. In this activity, high quality is defined as the lowest visual quality level where artifacts are not noticeable by an average non-expert viewer in a side-by-side comparison. Similarly, near-visually lossless quality is defined as the quality with the smallest amount of artifacts where a flicker test[6] decides that images are not visually lossless. Moreover, the new standard shall target specifically distortions due to compression and not necessarily other types of distortions caused by capture, sensor, or rendering artifacts to mention a few.

3.1 Use Cases and Requirements

The use cases and requirements identified in the context of the JEPG AIC-3 activity are available in the document entitled "Use Cases and Requirements for Image Quality Assessment v3.0",[15] which is the latest release at the time of the writing of this paper.

3.1.1 Use Cases

In this section, the use cases addressed by the JPEG AIC-3 standard are presented.

Assessment of End-User Web Delivery Image Coding: On the web, visual fidelity targets are relatively low, in order to maximally save bandwidth. Both high-appeal encoding (low bitrate, but "the image looks good" without comparing to a reference) and high-fidelity encoding (higher bitrate, accurate color, and detail preservation when compared to a reference) are relevant. Various image codecs are now (becoming) available in browsers, including JPEG, PNG, JPEG 2000, WebP, AVIF, and JPEG XL. There is still limited knowledge about the compression performance of different codecs and their encoders (e.g. `mozjpeg`, various AVIF encoders like `libaom`, `rav1e`, `svt-av1` and `aurora`), as well as the type of image content they are most suitable for. Subjective evaluation is needed to get more insights. Objective metrics are valuable to make automated per-image encoder configuration decisions.

Assessment of High-Quality Consumer-Grade Image Coding: In contrast to the web delivery use case, for consumer-grade photography, the visual fidelity target for lossy encoding is higher. At the same time, mathematically lossless or visually lossless encoding is typically not required.

Encoder Perceptual Optimization: In order to improve the trade-offs between perceptual quality, compression ratio, and encoding speed, various approaches can be proposed, e.g., to make better use of coding tools or to prune the search space. Such proposals require validation in the form of subjective and/or objective evaluation. Encoders can also internally implement objective metrics for the purpose of optimization. In this use case, objective metrics are generally preferred in order to allow for faster iterations of development and finetuning of their parameters. The computational complexity of the metric itself is important if the metric is used internally by an encoder.

Quality Assessment under Various Viewing Conditions: Subjective image quality depends on the viewing conditions, including display device technology, brightness, ambient light, pixel density, viewing distance, etc. Besides the traditional desktop or laptop computer monitors, images are also displayed using a variety of different display devices: including smartphones, smart watches, HDR televisions, beamers, etc. In many practical use cases, what matters is the way images are realistically and commonly viewed on consumer-grade display devices. This could be rather different from the way they are viewed on professional-grade and calibrated display devices in a lab-controlled environment as specified in current standards.

3.1.2 Requirements

AIC-3 has determined both *core requirements*, identified with a 'shall', and *complementary requirements*, identified with a 'should', which have been reported in the "Use Cases and Requirements for Image Quality Assessment" document.[15] In this paper, only the core requirement are presented.

Notably, the core requirements for the subjective quality assessment and score screening are the following:

- Quality range: The standard shall provide discriminative scores in the quality range where both ITU Rec. BT-500[3] and AIC-2 (flicker test)[6] are not discriminative between two candidate coding technologies;
- Suitable to evaluate compression artifacts: The standard shall be suitable for evaluating quality degradations introduced by compression;
- Score screening: The standard shall specify effective score screening methods in order to identify outliers;
- Reliability: The standard shall ensure that after screening, scores are reliable and consistent (non-self-contradictory);
- Reproducibility: The standard shall ensure that the scores obtained from one experiment shall be reproducible and confirmable by another experiment that follows the same methodology but performed independently from the first one;
- Scalability: The standard shall provide a possibility to be efficiently usable at a large scale (hundreds or thousands of test subjects);
- Controllability: The standard shall provide a possibility to be usable in scenarios with limited control over the experimental setup (e.g., crowdsourcing experiments);
- Variety of image content: The standard shall be reliable when applied to a variety of image content, i.e. not only photographic images but also synthetic images (graphics, screenshots, etc.);
- Variety of consumption environments: The standard shall be applicable in different consumption environments, i.e. display devices (TVs, desktops, laptops, smartphones, etc.) and viewing conditions (ambient light, viewing distance, etc.);
- Flexibility: The standard shall be able to obtain multiple trade-offs between competing requirements;
- Machine-readable scores: The standard shall specify a format that provides a machine-readable representation of subjective scores in order to facilitate the validation, development, and training of objective quality metrics, and the aggregation and consolidation across multiple experiments.

Timeline	
96$^{\text{th}}$ JPEG Meeting, July 2022	Second Draft Call for Contributions on Subjective Image Quality Assessment
97$^{\text{th}}$ JPEG Meeting, October 2022	**Final** Call for Contributions on Subjective Image Quality Assessment
1st April 2023	**Deadline** for the submission of the contributions
99$^{\text{th}}$ JPEG Meeting, April 2023	Start of the **collaborative process** for Subjective Image Quality Assessment
101$^{\text{th}}$ JPEG Meeting, October 2023	Working Draft for Subjective Image Quality Assessment
103$^{\text{th}}$ JPEG Meeting, April 2024	Committee Draft for Subjective Image Quality Assessment

Table 2: Timeline of the JPEG AIC-3 Call for Contributions on Subjective Quality Assessment.

- Efficiency: Given a targeted statistical sensitivity, the standard shall strive to reach it with a minimum number of test subjects and conditions.
- Complexity: Given a targeted statistical sensitivity, the standard shall minimize the required test setup complexity.

Additional complementary requirements are available in the "Use Cases and Requirements for Image Quality Assessment" document.[15]

3.2 Call for Contributions on Subjective Image Quality Assessment

In order to investigate, develop, and validate new subjective image quality assessment methodologies, the JPEG Committee will be issuing a call for contributions on subjective image quality assessment. Two Draft Calls for Contributions on Subjective Image Quality Assessment[1,2] have been already released during past JPEG meetings, with the goal of providing information to the interested parties and reaching out to new contributors. The Final Call for Contributions on Subjective Image Quality Assessment is planned to be issued at the 97$^{\text{th}}$ JPEG meeting in October 2022.

In the context of the future call, the process will be collaborative from the very beginning and all received contributions will be considered in developing the standard by consensus among the JPEG experts. During the collaborative phase, elements of complementary contributions may be combined into a single coherent specification. Table 2 reports the current timeline of the AIC-3 activity until creation of a first Committee Draft.

4. CONCLUSIONS

This paper summarized and reviewed the activities of the JPEG AIC standardization project, including a discussion about its Use Cases and Requirements[15] and a future Call for Contributions in Subjective Image Quality Assessment under AIC-3.

ACKNOWLEDGMENTS

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

REFERENCES

[1] ISO/IEC JTC1/SC29/WG1 N100165, REQ, "Draft call for contributions on subjective image quality assessment." https://jpeg.org/aic/documentation.html.

[2] ISO/IEC JTC1/SC29/WG1 N100259, REQ, "Second draft call for contributions on subjective image quality assessment." https://jpeg.org/aic/documentation.html.

[3] ITU-R Rec. BT.500, "Methodologies for the subjective assessment of the quality of television images," (2012).
[4] ISO/IEC TR 29170-1:2017, "Information technology — advanced image coding and evaluation — part 1: Guidelines for image coding system evaluation."
[5] "What to focus on in image compression: Fidelity or appeal," (accessed: 16.08.2022). "https://cloudinary.com/blog/what_to_focus_on_in_image_compression_fidelity_or_appeal".
[6] ISO/IEC 29170-2:2015, "Information technology — advanced image coding and evaluation — part 2: Evaluation procedure for nearly lossless coding."
[7] ITU-R Report BT.1082, "Studies toward the unification of picture assessment methodology," (1990).
[8] Recommendation ITU-R BT.500-14, "Methodologies for the subjective assessment of the quality of television images," *International Telecommunication Union* (2019).
[9] Ascenso, J., Akyazi, P., Pereira, F., and Ebrahimi, T., "Learning-based image coding: early solutions reviewing and subjective quality evaluation," in [*Optics, Photonics and Digital Technologies for Imaging Applications VI*], **11353**, 164–176, SPIE (2020).
[10] Testolina, M., Upenik, E., Ascenso, J., Pereira, F., and Ebrahimi, T., "Performance evaluation of objective image quality metrics on conventional and learning-based compression artifacts," in [*2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*], 109–114, IEEE (2021).
[11] Upenik, E., Testolina, M., Ascenso, J., Pereira, F., and Ebrahimi, T., "Large-scale crowdsourcing subjective quality evaluation of learning-based image coding," in [*2021 International Conference on Visual Communications and Image Processing (VCIP)*], 1–5, IEEE (2021).
[12] Report BT.1082, "Studies toward the unification of picture assessment methodology," *International Telecommunication Union* (1990).
[13] Testolina, M., Upenik, E., and Ebrahimi, T., "Comprehensive assessment of image compression algorithms," in [*Applications of Digital Image Processing XLIII*], **11510**, 469–485, SPIE (2020).
[14] Stolitzka, D. F., Schelkens, P., and Bruylants, T., "New procedures to evaluate visually lossless compression for display systems," in [*Applications of Digital Image Processing XL*], **10396**, 98–108, SPIE (2017).
[15] ISO/IEC JTC 1/SC29/WG1 N100258, REQ, "Use cases and requirements for image quality assessment v3.0." https://jpeg.org/aic/documentation.html.