Wien and Jung
EURASIP Journal on Image and Video Processing (2024) 2024:16
https://doi.org/10.1186/s13640-024-00630-7

EURASIP Journal on Image
and Video Processing

RESEARCH Open Access

Remote expert viewing, laboratory tests or objective metrics: which one(s) to trust?

Mathias Wien[1]* and Joel Jung[2]

*Correspondence:
wien@lfb.rwth-aachen.de

[1] Lehrstuhl für Bildverarbeitung, RWTH Aachen University, Aachen, Germany
[2] Tencent Media Lab, Palo Alto, CA, USA

Abstract

We present a study on the validity of quality assessment in the context of the development of visual media coding schemes. The work is motivated by the need for reliable means for decision-taking in standardization efforts of MPEG and JVET, i.e., the adoption or rejection of coding tools during the development process of the coding standard. The study includes results considering three means: objective quality metrics, remote expert viewing, which is a method designed in the context of MPEG standardization, and formal laboratory visual evaluation. The focus of this work is on the comparison of pairs of coded video sequences, e.g., a proposed change and an anchor scheme at a given rate point. An aggregation of performance measurements across multiple rate points, such as the Bjøntegaard Delta rate, is out of the scope of this paper. The paper details the test setup for the subjective assessment methods and the objective quality metrics under consideration. The results of the three approaches are reviewed, analyzed, and compared with respect to their suitability for the decision-taking task. The study indicates that, subject to the chosen test content and test protocols, the results of remote expert viewing using a forced-choice scale can be considered more discriminatory than the results of naïve viewers in the laboratory tests. The results further that, in general, the well-established quality metrics, such as PSNR, SSIM, or MS-SSIM, exhibit a high rate of correct decision-making when their results are compared with both types of viewing tests. Among the learning-based metrics, VMAF and AVQT appear to be most robust. For the development process of a coding standard, the selection of the most suitable means must be guided by the context, where a small number of carefully selected objective metrics, in combination with viewing tests for unclear cases, appears recommendable.

Keywords: Visual quality assessment, Quality metrics, Remote expert viewing, Visual media coding

1 Introduction

In the context of the development of compression algorithms for visual media, the determination of compression efficiency and quality improvements plays a crucial role. For conventional 2D video, the standardization groups MPEG, VCEG, and JVET have developed common testing procedures to allow for a fair comparison between a so-called "anchor", which represents the performance of a test model, which in turn implements the draft standard at a certain point in time, and the so-called "proposal", which attempts to improve the performance of this test model. These testing procedures include defined test sets and encoder configurations for assessing the compression performance, known as Common Test Conditions, e.g., [1]. Since these testing conditions typically remain stable over a long period in the development process, this also enables the groups to assess the improvement in performance over time. It is remarkable that for 2D video, while being systematically questioned by the expert community, the assessment typically relies on the Peak Signal-to-Noise Ratio (PSNR), i.e., the pixel-based Euclidian distance between the source and the reconstructed compressed signal, with elaborate methods and procedures for evaluation [2]. While this approach does not necessarily imply decisions towards best visual performance, it has been observed that verification testing consistently indicates substantial visual gains, see, e.g., the verifications tests for Versatile Video Coding (VVC) [3–5]. Such verification testing is usually performed at the end of the standardization process, with laboratories conducting a formal subjective assessment of the compression performance of the coding [6]. Similarly, in the contexts of 3D and immersive video standardization, objective metrics are considered. In MPEG Immersive Video (MIV) [7], which deals with the coding of multiple views (textures and depth maps) to enable free navigation in the scene with six degrees of freedom (called 6DoF), common test conditions are similarly derived to guide the adoption process [8]. Given the challenge of quality assessment of immersive video, no unique metric was considered suitable for the development process. Instead, a combination of metrics is employed. The list of considered metrics has evolved over time: from an initial set of five metrics (Video Multimethod Assessment Fusion (VMAF), Structural Similarity Index Measure (SSIM), PSNR, Weighted Spherical PSNR (WS-PSNR), Immersive Video PSNR (IV-PSNR)), only two of them are currently reported (PSNR, IV-PSNR). In case of debatable or contradicting results, visual checks or remote expert viewing sessions are performed. In MPEG V-PCC and V-Mesh activities (video-based point cloud compression and video-based mesh compression) [9] the list of considered metrics includes dedicated metrics, such as point-to-point D1, point-to-plane D2, yuvPSNR applied on texture maps, and uvPSNR applied on the uv-coordinates, i.e., the position of each texture coordinate vertex [10].

In the context of standards development for visual media, frequent subjective evaluation would be very helpful in assessing the progress in terms of visual quality improvement. At the same time, the related effort is quite high, both in terms of human resources and time. Furthermore, the question remains unresolved as to what extent the employed objective metrics may be considered reliable when it comes to the visual quality impact of specific coding tools. This question motivates the investigation and development of suitable means for decision-taking for any type of visual media (2D, 360°, immersive videos, point clouds, meshes, etc.). The focus in this effort is on reliable means for decision-taking, i.e., the assessment of often very tiny compression improvements of a proposed change to an anchor. The evaluation method to be established (by subjective testing and/or using objective metrics) must be reproducible, and understood and accepted by the standardization group to serve the intended decision-taking usage. The problem scales with the increase in the degree of freedom of user interaction and assessment. While the viewing conditions for conventional 2D video may be inherently defined by presenting the compressed video sequence on an suitable device, the coding of immersive video implies the choice of an individual viewing perspective by the user. The intended use of immersive visual media relies significantly on interactivity and may allow for an assessment from virtually any viewing directions and viewing paths. However, this increased degree of freedom further implies the use of an extended processing chain operating between the decoding of the compressed visual media signal and the chosen assessment device, such as conventional monoscopic or stereoscopic displays, head-mounted displays, or mobile devices which, e.g., allow for navigation in the scene by movement of the device. Hence, multiple additional aspects such as mono- or stereoscopic rendering or user interaction for view path selection arise and the testing task becomes even more challenging.

In ISO/IEC JTC 1/SC 29, the Advisory Group 5 MPEG Visual Quality Assessment (AG 5) is tasked with the investigation, development, and recommendation of tools and methods for this purpose [11]. AG 5 has developed a remote expert viewing (REV) protocol [12] with the goal of a) providing a reliable means for ranking the visual quality of the proposal and the anchor in the development process, and of b) testing objective metrics for their suitability for doing the same task. The method originally relied on remote assessments under the conditions of the COVID pandemic in the years 2020-2022. It is similarly applicable for on-site use, e.g., with experts attending a standardization meeting in presence. The REV scheme has been adopted by the MPEG working groups on Video and 3D Graphics, as well as JVET, for various purposes such as tool development or the preparation of calls for proposals or verification tests, e.g., [13].

Numerous studies have analyzed the correlation between objective metrics and MOS. Similarly, comparisons of subjective methods have also been widely researched. For instance, studies [14] and [15] evaluated subjective methods in the context of mobile video and 3D video, respectively, and more recently, in the context of virtual reality as seen in [16]. Due to the extensive literature on this topic, we focus our description on the most recent studies. [17] has explored the feasibility of performing subjective tests with a limited number of viewers but with repetitions. Study [18] compares the DCR (Double Stimulus Continuous Quality Scale) with the EVP (Enhanced Video Perception) rating scale to the traditional ACR-HR (Absolute Category Rating with Hidden Reference) approach. Finally, in 2024, a study was published comparing omnidirectional video and spatial audio conditions in terms of subjective quality and the corresponding impact on the resolving power of metrics [19].

In this paper, a study on the problem of quality assessment and decision-taking is presented considering three perspectives: objective metrics, remote expert viewing, and formal visual evaluation under laboratory conditions. The focus is on the evaluation of pairs of coded video sequences, which typically comprises a proposed change to a video coding scheme scheme and the unmodified version. The original and the changed version are compared at a one or more selected rate points. No further aggregation, e.g., as provided by the Bjøntegaard Delta rate, is regarded here. To lower the dimensionality of the general problem, the focus is on test material from 2D video compression which has been assessed by the REV method in JVET. Most of the test content represents tool on/off tests, i.e., the comparison of a proposed change in the coding scheme to the anchor represented by the unchanged reference software. An extension to immersive visual media compression is future work, in which aspects such as user interaction and different playout devices may be considered.

The main questions to be addressed can be phrased as follows: is the REV method functional? Are objective metrics able to indicate the correct decision? And, considering the standardization scope of this work, is it possible to detect difficult cases, e.g., by objective metrics, such that an indication of the need of some type of subjective quality assessment can be drawn?

The paper is organized as follows: in Sect. 2, the data set, the REV method, and laboratory test are presented. Section 3 details the assessment of the data set by objective metrics and presents the results of the objective evaluation of the data set. In Sect. 4, the results of the REV tests and the laboratory test are analyzed with respect to consistency, reliability, and potential challenges. In Sect. 5, an overall discussion of the objective and subjective results is presented and potential answers to the question in the title are provided. We conclude the paper in Sect. 6.

2 Data set and visual test methodologies

2.1 Content description

The data set used in this study for comparison of Remote Expert Viewing (REV) tests and laboratory viewing (LAB) tests comprises test results acquired in a series of six JVET meetings during the COVID period from 2021 to 2022. It comprises a total of 232 test points, including trapping sequences inserted into the test sessions for control purposes. All test points have been evaluated by JVET video coding experts. The results are reported in [20–25]. The REV tests were mostly conducted in the context of an exploration experiment called EE1, investigating the compression efficiency improvement of neural network-based (NN) coding tools. Furthermore, modifications to the deblocking filter applied to both JVET test models, the VVC Test Model (VTM) and the Enhanced Compression Model (ECM), were investigated, and new test sequences were evaluated. The full data set, called DS, is presented in Tables 14 and 15 in the Appendix of this paper. In the REV tests, each test point is compared to its corresponding anchor (VTM in the case of EE1, VTM and ECM, respectively, for the deblocking filter tests, and the HEVC test model (HM) for the exploration of new test sequences). For this study, the data set has been divided into six categories listed below:

- Loop filters (LF): This category includes all NN-based proposals for in-loop enhancement filters, NN-based deblocking filters, combinations of these, as well as tests for modifications of the conventional deblocking filter.
- DNN super-resolution (DNN-SR): This category includes all proposals for NN-based re-scaling, where the test sequence is coded at a lower resolution (subsampled by a factor of two in both horizontal and vertical directions), and subsequent up-sampling with a proposed NN-based method.
- DNN decomposition - compression - synthesis (DNN-DCS): This category includes proposals with a modified coding loop where texture detail is represented at full spatial resolution while temporal changes are encoded at a lower spatial resolution.
- Reference picture resampling (RPR): RPR is a coding tool in VVC enabling the change of picture resolution within a coded video sequence. It can be used for coding

a sequence at a lower resolution and upscaling it with the standardized RPR method. Since it is readily available with VVC, it is used as a reference point for proposals in the DNN super-resolution category in the context of the JVET exploration experiments.

- Comparing HM and VTM (HM-VTM): A comparison of the HM and the VTM was performed in JVET for studying properties of new test sequences which were considered for potential inclusion in the set of test sequences of the common test conditions.
- Trap (TRP): This category includes control test points inserted into the REV test sessions to verify the validity and consistency of the rating of the participants. Such traps could e.g., consist of a sequence coded at two clearly different qualities, or comparing a compressed sequence to an uncompressed one. In either case, the incorrect scoring of a participant (e.g., rating the compressed version over the uncompressed original) indicates either a lack of attention, or issues with the setup at the remote participants site, or other problems.

2.2 Content selection for laboratory tests

Due to the size of the full data set, a formal subjective evaluation of all test points in a formal laboratory test was not possible. Therefore, a subset was created, referred to as DS-LAB in the following. It was created by manually selecting test points according to the criteria listed below.

- “INC”, which shows inconsistent or unclear results for the objective metrics under evaluation. Test points in this class, e.g., show diverging results among objective metrics or in comparison to the subjective scores from the REV tests.
- “SIG”, where the REV revealed a Differential Mean Opinion Score (DMOS) close to zero and a confidence interval overlapping or touching the zero line, i.e., not or almost not significant.
- “LC”, which shows a large confidence interval in the REV but is clearly removed from DMOS $= 0$. The large size of the CI is taken as an indication that participating experts scored differently, which indicates the potential occurrence of artifacts which might be difficult to rate, either subjectively or objectively. Such cases may occur, e.g., if a proposal shows more details, yet also more artifacts than the anchor.
- “OPP”, which shows opposite results at two tested rate points, e.g., the proposal is better at the low rate but worse at the high rate.
- “DIF”, that shows a clear difference between the anchor and the proposal under test in terms of DMOS for the REV. The confidence interval does not include DMOS$=0$. These test points are considered clear cases.

Based on these criteria 52 pairs of test sequences were identified as candidates for the LAB test. The selected points are marked in bold font in Tables 14 and 15 in the Appendix. They are further tagged with bold “INC”, “SIG”, “LC”, “OPP” or “DIF” to indicate the applicable criterion for the respective test points. By design, the 52 pairs of test sequences in DS-LAB are considered to be difficult for the expert viewers to score.

Hence, they also are expected to be difficult to score for the naïve viewers, and for the objective metrics.

2.3 REV methodology

ISO/IEC JTC 1/SC 29 Advisory Group 5 MPEG Visual Quality Assessment has developed guidelines for Remote expert viewing (REV) [12]. The guidelines have been developed for the purpose of enabling visual quality assessment during online standardization meetings. They are based on established test protocols, such as ITU-R BT.500, ITU-T P.808, and ITU-T P.910 [26–28], and provide recommended steps in terms of the preparation of the video sequences to be tested, as well as preparation and implementation procedures. The REV method has been applied in the context of JVET exploration experiments for 2D video, for core experiments in the development of MPEG Immersive Video (MIV), and in the preparation of verification tests for video-based point cloud coding (V-PCC). In most cases, the REV is used for the comparison of a proposed technology to the previously established anchor. Due to its high discriminatory power, Comparison Category Rating (CCR) [27] is recommended for this purpose, and used in the presented study. Other protocols can also be applied. REV sessions using Absolute Category Rating (ACR) and Degradation Category Rating (DCR) methods [28] were also conducted [13, 29]. To enable easy application, the guidelines rely on the use of widely available open-source software (ffmpeg [30], VideoLAN VLC [31]) for preparation and viewing. The REV method is briefly presented in the following.

2.3.1 REV procedure

For conducting a REV session, the group appoints a test coordinator and selects the content to be visually evaluated. The coordinator leads the test effort and reports the results to the group. For immersive video content, one or several camera paths (sometimes called viewport or pose trace) are defined for each sequence under test. The decoded video sequences, or the rendered camera paths, of all rate points are converted into mp4 files for playout with VLC on the computers of the participants. The conversion is made via ffmpeg with a high-quality setting (constant rate factor parameter) to prevent the introduction of visible artifacts. Their duration is recommended to be in the range of 5 s to 10 s. The group appoints one or more cross-checkers to verify that the converted mp4 files match the intended content under test. The verified set of test sequences is provided to the test coordinator.

Volunteers from the group are selected as viewers for the REV sessions. They are expected to report any potential issues with visual acuity or color vision to the test coordinator. If the REV is conducted for the purpose of decision-taking in the adoption process for a proposal, then experts from the proposing institution should not participate to avoid potential bias. The viewers must confirm having suitable equipment and setups available as defined by the test coordinator. This includes a computer capable of smooth playout for the test sequences shown in the REV sessions, a suitable display, and reasonable viewing conditions, such as a calm room with indirect light not reflecting on the screen and a setup with the recommended viewing distance. The suitability of the technical setup is defined by the recommendations of the test coordinator and tested by the participating experts using a demonstration playlist with high-bit rate mp4-files.

Within a REV session, the viewers may be presented with multiple test sessions. The test coordinator provides the anonymized test sequences and the playlists in a zipped and password-protected package to the viewers. The viewers are required to have this data set downloaded and available at the time for the REV session. Further, the test coordinator provides the viewers with scoring sheets formatted to support the voting during the test sessions. The viewers are then instructed to note their scores on printouts of these sheets to minimize their effort during voting. In the REV session, the test coordinator first provides final instructions to the viewers. Furthermore, a training session for the viewers is conducted to verify that the test procedure and the rating scale are properly understood. Based on the training session scores provided to the test coordinator, the password of the package is disclosed, and the participants run the test sessions. The viewers must run each session without interruption and execute their voting during the voting periods of the Basic Test Cells (BTCs). Any operations such as pausing, re-play, or speed manipulation are not permitted. To avoid making the test twice, the viewers are requested to immediately provide their scores to the test coordinator after finishing the sessions. In practice, this is handled via a web interface with personalized access for each viewer.

2.3.2 REV rating scales

In the CCR scenario, two rating scales are suggested, depending on the tested material. For MIV, the 7-grade scale of ITU-T P.808 is employed. In JVET, a 4-grade scale has been established for expert viewing in the development phase of the VVC standard [32–34]. This scale applies a forced choice. The two scales are presented in Fig. 1. In this paper, results from REV sessions using the 4-grade scale are further studied.

The choice of a 4-grade scale is motivated by two considerations: (a) the two variants of the coded sequence under test are actually different, and (b) the experts participating in the tests as viewers are expected to be able to express an opinion on the observed differences.

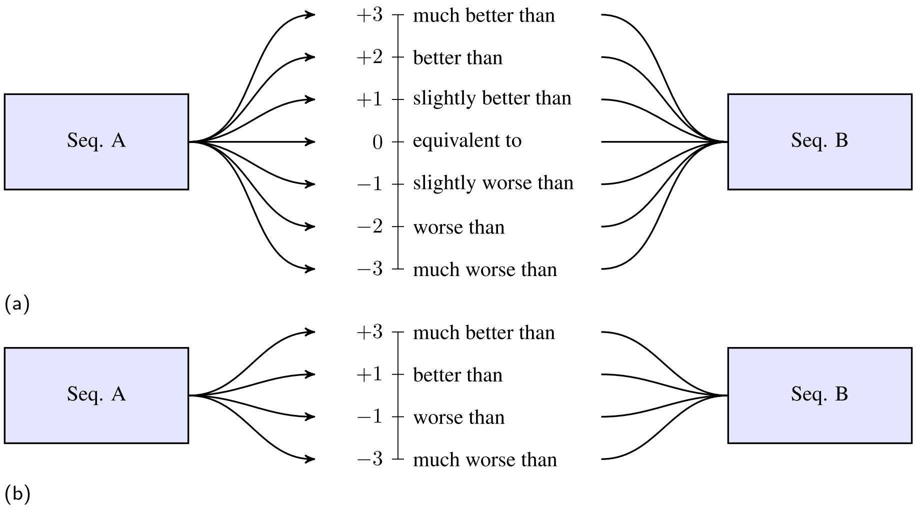


Fig. 1 REV rating scales. **a** 7-grade scale, **b** 4-grade scale with forced choice

2.3.3 Test session design

The test coordinator takes the complete set of provided test sequences and splits them into multiple test sessions. Each session is constructed out of a series of BTCs containing test sequences of the same resolution. They should not exceed a duration of 15 min. The test sequences are renamed for the purpose of anonymity. For the CCR method, each BTC consists of the uncompressed source sequence, a consecutive playout of versions A and B of the sequence, and a 5 s voting time. The presentation of the uncompressed source may be skipped in cases where no original sequence is available, such as immersive video. To increase the discrimination of small impairements, the A/B pair is shown twice, as suggested in variant II of double stimulus tests in ITU-R BT.500 [26].

Each test session includes a stabilization phase of two to four BTCs to allow for an adaptation phase for the viewers. In each BTC, the A/B playing order of the anchor and the proposal is randomized such that the viewers cannot guess the variant from the playing position. Furthermore, each test session includes at least one trapping BTC where the attention of the viewers is tested. This may include displaying the same sequence as variant A and B within a BTC or displaying two variants with a clearly known quality relation (e.g., two different quantizer settings of the same coding scheme).

2.3.4 Processing of REV results

After the results of the viewers have been received, they are processed by the test coordinator. The A/B randomization of the BTCs is reverted, leading to a consistent mapping of negative scores for the anchor and positive scores for the proposal. The scores of participants failing to vote correctly on the trapping BTCs of a single session are discarded for that session. If a participant fails for multiple sessions, the scores for all test sessions are discarded completely. In the case of a low correlation of the viewerâ€™s scores with the overall Comparison MOS (CMOS), further participants scores may be discarded. The applied criteria and the number of affected viewers is reported by the test coordinator in the REV report documents [20–25]. The number of participants and viewer rejections in the test sessions are reported in Table 1.

A challenging aspect of remote tests is the qualification of viewers based on equipment and overall environment setup conformity. It is advised to ensure correct viewing conditions, encompassing factors such as lighting sources, display contrast, viewing distance, etc., akin to traditional LAB tests. It is noted that these conditions are not

Table 1 Viewer rejection in the REV tests

		Rejection due to	
Test	**Participants**	**Trap**	**Correlation**
[20] JVET-U	13	0	0
[21] JVET-V	19	0	1
[22] JVET-W	22	0	4
[23] JVET-X	17	1	0
[24] JVET-Y	25	5	0
[25] JVET-Z	18	3*)	0

*): Maximum three rejected viewers per session, overall 7 different viewers

explicitly controlled or verified in our REV. Precise instructions were provided by the test coordinator to the viewers including details on appropriate equipment and setup. The only method for discarding scores is by outlier detection; there is no mechanism for detecting an incorrect setup. Nonetheless, the authors believe that this approach is valid, as, in essence, in a REV scenario, all participants are experts who are expected to be committed to providing reliable scores. This is different from regular crowd-sourcing where viewers may be motivated by financial incentives only. Consequently, adhering to and implementing the instructions provided by the test coordinator is deemed sufficient.

2.4 LAB test methodology

Laboratory tests following the recommended setup specified in ITU-R BT.500 [26] were used for studying the DS-LAB subset of the test points. Since both the LAB and the REV test methodologies are based on this Recommendation, some of the design elements and the setup are similar to the REV methodology. The test setup was implemented at RWTH Aachen University in a dedicated quiet black room. The configuration comprises four displays separated by black mobile walls which are operated simultaneously by the playout server. The LAB test has involved naïve viewers only, including 9 females and 22 males of age 19 to 28. The viewers were all students at RWTH Aachen University. By positioning two viewers in front of each display, viewing can be simultaneously performed with eight viewers in parallel. The specification of the test setup is provided in Table 2.

The tests were conducted following the Degradation Category rating (DCR) protocol of ITU-T P.910 [28] using the 11-grade impairment scale defined in ITU-R BT.500-14 Tab. 2-4 [26] (see Tables 3 and 4). The scale ranges from “imperceptible” impairments (score 10) to “severely annoying, everywhere” (score 0). This scale is widely used in the context of MPEG visual quality assessment work and has shown to provide the viewers with a scoring range of reasonable granularity [6]. It is noted that the scale is originally recommended to be used for expert viewing sessions. The naïve viewers did not report problems with using this scale and the results of the tests are considered reliable. In general, the scale is suggested to be used with great care. The set of test points was arranged in six sessions, arranged according to the different resolutions of the test sequences (3×UHD, 2×1080p, 1×720p). The maximum number of

Table 2 Setup for the laboratory tests at RWTH Aachen University

	Setting
Display, size (resolution setting), connection	1× Sony 55″ PVM X550 (3840×2160), Quad-SDI 1× LG OLED65CX (3840×2160), HDMI 2× LG OLED55G19LA (3840×2160), HDMI The Sony display is driven by a PC with a DeckLink 4K Extreme 12G video board via Quad-SDI, the SDI signal is converted to HDMI by an AJA Hi5-4K-Plus converter and sent in parallel to the three LG displays via an HDMI splitter
Viewing distance	2 viewers at 1.5H per display, the HD video signal was displayed centered to the UHD screen with a mid-grey padding for the unused area.
Viewing angle	70°
Total number of naïve viewers	31 (9 females, 22 males; age 19-28), all screened for visual acuity and normal color vision.

Table 3 Comparative overview of REV and LAB test methodologies

Methodology	LAB	REV
Viewers	Naïve (students)	Expert (MPEG members)
Number of viewers	31	Varying (see Table 1)
Viewer's qualification	Understanding of instructions validated on site by test coordinator	Training session with score controlled by the test coordinator
Location	RWTH Aachen University	Viewer's location
Display, viewing conditions	See Table 2	Uncontrolled (instructions shared by test coordinator)
Methodology	DCR*)	CCR
Rating scale	11 grade	Forced-choice 4 grade**)
Max length of a session	11.5 minutes	15 minutes
Stabilization phase	3 BTCs	2 to 4 BTCs
Trapping per session	1 trapping BTC	At least 1
Outlier rejection	Based on trapping BTC and statistical analysis	Based on trapping BTC and statistical analysis

*): Stimuli shown twice per variant II of double stimulus tests in ITU-R BT.500 [26]

**): For REV tests with MIV, the 7-grade scale was used. These are not studied this paper

Table 4 11-grade scale of the expert viewing protocol from ITU-R BT.500 [26] used in the laboratory tests

Score	Impairment item
10	Imperceptible
9	Slightly perceptible somewhere
8	Slightly perceptible everywhere
7	Perceptible somewhere
6	Perceptible everywhere
5	Clearly perceptible somewhere
4	Clearly perceptible everywhere
3	Annoying somewhere
2	Annoying everywhere
1	Severely annoying somewhere
0	Severely annoying everywhere

BTCs in a session was 23, resulting in a maximum duration of about 11.5 min per session. The tests were conducted in an alternating fashion with two alternating groups of viewers at a time, where one group performed the test session while the second group was resting.

The BTCs were designed according to A1-3 Variant II of ITU-R BT.500-14 with the pair of the original and the processed video sequences being shown twice, followed by a voting time of 5 s where a grey screen was shown with the number of the current vote indicated. Between the video sequences, a grey screen was shown for 1 s indicating if the original or the processed video sequence would follow.

The processed v5

deo sequences were presented in a randomized order. The session design prohibited the consecutive presentation of two processed versions of a test sequence. Furthermore, the order was arranged to avoid the presentation of similar impairment levels in subsequent BTCs. Each test session was initiated with a stabilization phase of three BTCs. The results of these BTCs were not considered in the evaluation. Each test session included a trapping BTC, where the original sequence was displayed and scored.

For training, the viewers were introduced to the procedure and example BTCs were presented covering the range of impairments presented in the test sessions. The viewers were advised to make use of the full scale according to the observed impairments. They were not informed of the presence and characteristics of the trapping BTCs.

After completion of the test sessions with all viewers, the scores were screened. First, the viewers scoring a trapping BTCs below 7 were excluded. One viewer failed for the trapping BTC in 5 out of 6 test sessions. The results of this viewer were completely discarded. The results were then screened applying the outlier detection mechanism specified in ITU-R BT.500-14 A1-2.3 [26]. The screening resulted in no further exclusion of viewers.

A comparative overview highlighting the LAB and the REV test setups is provided in Table 3.

3 Objective metrics evaluation

In quality evaluation studies, experimental results are generally reported according to three classical indicators: the Root Mean Square Error (RMSE) and the Pearson Linear Correlation Coefficient (PCC) reflect the prediction accuracy, i.e., the ability to predict the subjective quality with low error. The Spearman Rank Correlation Coefficient (SRCC) reflects the prediction monotonicity, i.e., the degree to which the predictions of the metric agree with the relative magnitudes of subjective quality ratings. These indicators typically evaluate the correlation with a Mean Opinion Score (MOS). This MOS may result from test sessions performed with an Absolute Category Rating (ACR) methodology, or with a Differential MOS (DMOS), which result from test sessions performed with a Degradation Category Rating (DCR) methodology. Another possible methodology is the Comparison Category Rating (CCR) which produces Comparison MOS (CMOS).

In this paper, the LAB tests provided DMOS, while the remote tests provided CMOS. RMSE, PCC and SRCC are not suitable for CMOS results analysis [35]. To get an insight into the performance of objective metrics in comparison to the collected DMOS and CMOS scores, the analysis is carried out through two different and complementary approaches. In the first analysis, the objective metrics are evaluated against the DMOS results measured on the DS-LAB data set. The method of [36] is used here. This method provides a comprehensive analysis of the pair-wise comparison of the assessed subjective scores and the same pair-wise comparison of the corresponding values for each objective metric. Hence, relation of all scores for all PVS are set into relation. In the second analysis, the focus is on the pair-wise comparisons which have been performed in the REV tests. These thereby represent a subset of the first analysis. Since the number of test points in the full data set (DS) is much larger than in the DS-LAB data set, additional analyses on subset of DS, e.g., according to different classes can advantageously be provided.

3.1 Objective metrics

Thirteen full-reference metrics are evaluated. Among them, the PSNR, SSIM and MS-SSIM (Multi-Scale SSIM) can be considered as simple low-complexity metrics. Another three are more complex: 3SSIM (Three Component SSIM) [37] computes a weighted average of SSIM for three different categories (edges, textures and smooth regions). VQM (Video Quality Metric) [38] is a video quality metric based on the Discrete Cosine Transform (DCT): an 8 × 8 DCT is used for the spectral representation of the image. The DCT coefficients are used to compute local contrast and just noticeable differences to derive the score. VIF (Visual Information Fidelity) [39] is based on natural scene statistics and the notion of image information extracted by the human visual system.

The five remaining metrics are learning-based, trained on different 2D data sets. AVQT (Advanced Video Quality Tool) [40] is a command line tool that estimates the perceptual quality of compressed videos. VMAF [41] is the Video-Multi-Method Assessment Fusion quality assessment algorithm: Support Vector Regression is used to fuse three metrics to obtain a single quality score. WaDiQAM (Weighted Average Deep Image QuAlity Measure) [42] is a deep neural network-based approach for image quality assessment. The network is trained end-to-end and comprises ten convolutional layers and five pooling layers for feature extraction, and two fully connected layers for regression.

LPIPS (Learned Perceptual Image Patch Similarity) [43] belongs to the family of deep learning based objective metrics that evaluate the similarity of an image with a reference image, rather than evaluating the quality itself. Here, the default version employing the alexNet network as well as two variants, i.e. squeezeNet and Visual Geometry Group (VGG) network (vggNet), are included. DISTS (Deep Image Structure and Texture Similarity) [44] is a full-reference image quality assessment model. It is based on a variant of VGG that combines texture similarity, spatial averages, and structure similarity. It is designed to provide explicit tolerance to texture resampling, and insensitivity to geometric transformations.

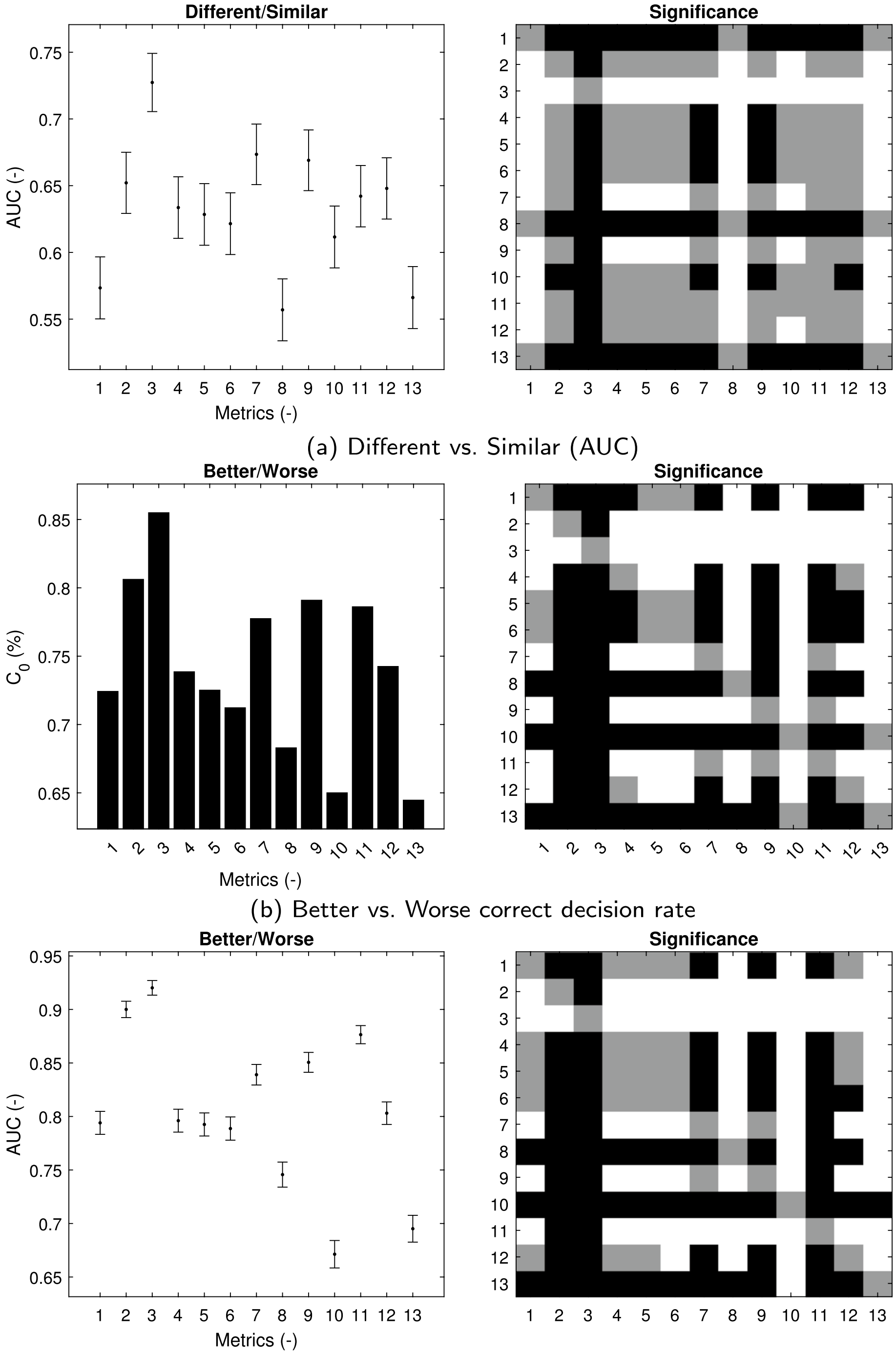
3.2 Full pairwise metric evaluation on the LAB test results

3.2.1 Methodology

The results of the LAB test provide an individual MOS score for each PVS hich can be compared to the individual scores received for the objective metrics under investigation. In contrast to the REV tests where the CMOS results are obtained for predefined pairs of PVS, a full analysis of the scores is enabled here. We apply the method of Krasula et al. [36] for this purpose. The method studies two aspects: The "different vs. similar" analysis evaluates the ability of a metric to distinguish between significantly different and similar pairs. The area under the receiver operating characteristics curve (AUC) is computed taking into account the significance of the observed MOS difference. The "ybetter vs. worse" analysis only considers the pairs which were identified as significantly different. For these pairs, the correct decision rate is computed (referred to as C_0 here) and again, the AUC is computed for this subset. Based on the significance of these measures, a comparison plot can be drawn, indicating the significance level for each pair of metrics. Thereby, the metrics can be ranked by the number of significantly worse performing metrics.

3.2.2 Evaluation of the correct decision rate

The results of the described two analyses are provided in Fig. 2. For compact presentation reasons, the 13 objective metrics under study are referred to by corresponding



(a) Different vs. Similar (AUC)

(b) Better vs. Worse correct decision rate

(c) Better vs. Worse (AUC)

Fig. 2 Analysis of the objective metrics against the LAB results, indices defined in Table 5. The significance plots show that the performance of the method in the row is either significantly better (white), lower (black), or none of the previous (gray) [36]

Table 5 Indices associated to the objective metrics in the AUC evaluation in Fig. 2

3SSIM	AVQT	DISTS	LPIPS (alexNet)	LPIPS (squeezeNet)	LPIPS (vggNet)	
1	2	3	4	5	6	
MS-SSIM	PSNR	SSIM	VIF	VMAF	VQM	WADIQAM
7	8	9	10	11	12	13

indices as defined in Table 5. Subplot (a) of Figure 2 reveals that the DISTS metric performs significantly better than all other metrics with respect to the “different vs. similar” comparison. It is followed by the SSIM and the MS-SSIM with seven significantly inferior metrics, and the AVQT and VQM which have four inferior metrics each. Remarkably, the PSNR performs at the worst ranking position in this analysis. In the “better vs. worse” analysis, DISTS again is ranked top over all other metrics. This study is presented in subplots (b) and (c) of Fig. 2. It only considers the cases with significantly different pairs of PVS. Here, AVQT performs second best, followed by VMAF and then SSIM and MS-SSIM. The PSNR performs third but last in this analysis.

3.3 Decision rate evaluation on the REV test pairs

3.3.1 Methodology

The analysis presented in this paper departs from the comparison of sets of two differently encoded video sequences in the REV tests. We now focus the analysis on the pair-wise relation of the available MOS scores and the deltas computed between the corresponding objective metric scores. For this purpose, we are interested in the correct decision rate on the test pairs defined in the REV tests. Thus, we have used a Receiver Operating Characteristic (ROC) analysis, which is adapted to binary classifiers.

We have arbitrarily decided that a positive case corresponds to the situation in which the proposal (P) is better than the anchor (A). We then derived true positive, true negative, false positive and false negative cases as follows:

- TP (True Positive): CMOS says $A < P$, metric says $A < P$.
- TN (True Negative): CMOS says $A > P$, metric says $A > P$.
- FP (False positive): CMOS says $A < P$, metric says $A > P$.
- FN (False negative): CMOS says $A > P$, metric says $A < P$.

We have also computed the percentage of correct decisions, CD:

$$\mathrm{CD} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

This correct decision rate was also computed for the laboratory tests, after having subtracted the DMOS of the proposal and the DMOS of the anchor.

3.3.2 Evaluation of the correct decision rate

Evaluation on the full data set: Table 6 reports the true positive, true negative, false positive and false negative over the entire data set, relative to the CMOS obtained by the remote expert viewing. Given the arbitrary choice of associating "positive" to the event "the proposal is better than the anchor", we observe that, since proposals are generally better than anchors, TP gathers most of the cases. In addition, we remark that TP + TN is far above FP + FN, indicating that, overall, the metrics tend to make the correct choice.

For a deeper analysis, the percentage of correct decisions, computed from TP, TN, FP, and FN, is provided in Table 7 for two different scenarios:

- "CD-ALL" computes the correct decision rate on the full data set.
- "CD-CI+CMOS" corresponds to all test cells for which a solid decision is taken by the viewers to select the best method among A and P. It keeps data for which: a) the confidence interval (CI) of CMOS does not cross the x-axis, and b) the absolute value of CMOS is above or equal to T_{CMOS}=0.4. The rationale for applying this filtering, with this specific threshold, is that the scale used for the remote expert viewing is a forced-choice scale, i.e., it does not include a "0" value with an "equivalent" grade. Our subjective tests have revealed that, when similar content is shown as A and P, the CMOS can approach 0.4. Therefore, it is unfair to consider scores between 0 and 0.4 as clear decisions taken by humans. Note that the effect of varying the threshold TCMOS is studied in Fig. 3, see below.

It is noted that, by its definition, CD-ALL does not regard any uncertainty. For this reason, this scenario has to be interpreted with caution. To leverage the uncertainty information provided by the confidence intervals, the study is extended by proposing correct decision rates defined in the CD-CI+CMOS scenario. This scenario further addresses the uncertainty induced by the forced-choice scale used in the REV tests by filtering the results based on a confidence treshold in the CD-MOS scenario.

Table 6 True positive, true negative, false positive and false negative count, versus remote expert viewing CMOS on the full data set

Full reference metric	TP	TN	FP	FN
3SSIM	161	14	37	20
AVQT	168	14	37	13
DISTS	139	13	38	42
LPIPS (alexNet)	146	15	36	35
LPIPS (squeezeNet)	160	16	35	21
LPIPS (vggNet)	162	17	34	19
MS-SSIM	173	11	40	8
PSNR	169	12	39	12
SSIM	172	13	38	9
VIF	156	24	27	25
VMAF	165	17	34	16
VQM	156	20	30	24
WADIQAM	157	14	37	24

Table 7 Correct decision rate for the objective metrics, in two different scenarios

Full reference metric	**CD-ALL**	**CD-CI+CMOS**
# of cells	**232**	**111**
3SSIM	75.4	87.4
AVQT	**78.4**	89.2
DISTS	65.5	72.1
LPIPS (alexNet)	69.4	81.1
LPIPS (squeezeNet)	75.9	85.6
LPIPS (vggNet)	77.1	88.3
MS-SSIM	**79.3**	**92.8**
PSNR	78.0	**91.0**
SSIM	**79.7**	**92.8**
VIF	77.6	88.3
VMAF	**78.4**	**91.0**
VQM	76.5	79.8
WaDiQAM	73.7	85.6

The best performing metrics for each CD variant are marked bold

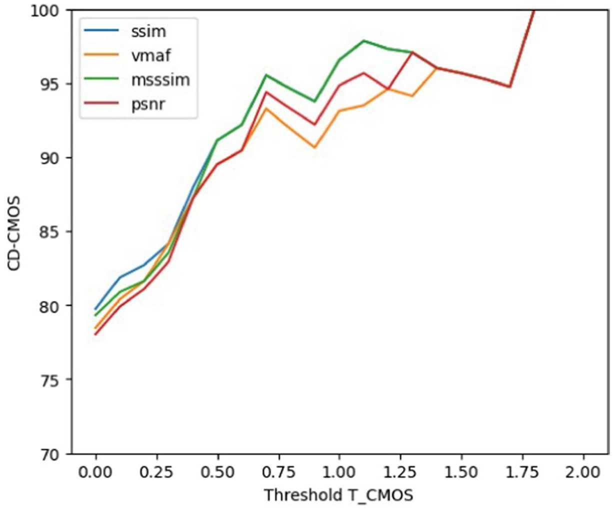


Fig. 3 Evolution of the correct decision rate for four metrics based on T_{CMOS}

The two criteria, CI and CMOS, have an overlapping effect on the acquired data. It is observed that the data points removed based on the the CMOS constraint are a subset of the data points removed by the CI constraint. This observation suggests that the threshold to accommodate potential effects of the forced-choice scale do not overrule the decision based on the confidence interval. The number of test cells remaining after applying the different filtering approaches is reported in the first row of the table. The total number of test cells is 232.

From Table 7, we can derive that, no matter which scenario is considered, the group of best performing metrics remains the same. SSIM, MS-SSIM, VMAF, PSNR and AVQT are among the most accurate metrics for taking correct decisions. It is noteworthy that none of the learning-based metrics (AVQT, DISTS, LPIPS, VMAF, WaDiQAM) were retrained specifically on JVET contents. Thus, among these, VMAF and AVQT represent the most robust metrics.

While the result for “CD-ALL” may be misleading, because it considers test cells where the viewers were not able to make clear decisions, the results for “CD-CI+CMOS” show an excellent ability of the metrics to take correct decisions (above 90%), when the confidence interval is considered.

As mentioned previously, “CD-CI+CMOS” relies on a T_{CMOS}=0.4 threshold that has been empirically selected. Fig. 3 exemplifies the evolution of the CD rate for four metrics over T_{CMOS}. A T_{CMOS} value of 0 corresponds to the CD-ALL case. The higher T_{CMOS}, the higher the number of test cells removed from the data set, assuming that the viewers were not able to decide. And, as expected, the higher T_{CMOS} the higher the correct decision rate for each metric.

Evaluation per coding tool category: As indicated in Section 2.1, five categories of proposals were evaluated through the test sessions, as well as â€˜trapping cellsâ€™ intended to discard unqualified viewers. The number of associated cells is listed in Table 8.

The analysis provided in this section discards the categories for which the number of samples is too low to guarantee reliable conclusions. Thus, Tables 9 and 10 report the percentages of correct decisions for the “DNN Inloop Filter + Deblocking” and the “DNN Super-resolution” categories, for the CD-ALL and CD-CI+CMOS configurations, respectively.

By studying the overall result (CD-ALL), it can be observed that the deblocking category has a lower correct decision rate than the super-resolution category. Objective metrics tend to be less reliable in evaluating the efficiency of deblocking algorithms. For the deblocking category, none of the metrics manages to reach a rate of 80% correct decisions, whereas six metrics reach this threshold in the super-resolution category.

However, when considering the CD-CI+CMOS correct decision rates, we observe that the gap between the two categories decreases. The percentage of correct decisions is higher for super-resolution than for deblocking for 85% of the metrics (11/13) for CD-ALL. This figure declines to for 46% of the metrics (6/13) for CD-CI+CMOS.

When comparing the metrics to each other, there is little difference for the deblocking category: apart from the two exceptions (DISTS and LPIPS-alexNet), they all provide similar level of correct decisions. For the super-resolution category, a group of five metrics perform better than others: SSIM, MS-SSIM, VMAF, PSNR and AVQT. LPIPS-vggNet is more reliable for deblocking than for super-resolution, highlighting the importance of the training set definition for learning-based approaches.

Table 8 Coding tools categories and number of corresponding test cells

Category	# of test cells
DNN Decomposition Compression Synthesis	5
DNN Inloop Filter + Deblocking	149
Reference Picture Resampling	12
DNN Super-resolution	43
Comparison of HM and VTM	14
‘Trapping cells’	9

Table 9 Correct decision rate for DNN inloop filter + deblocking

Full reference metric **# of test cells**	**CD-ALL** **149**	**CD-CI+CMOS** **57**
3SSIM	73.8	**89.4**
AVQT	**74.5**	84.2
DISTS	61.1	61.4
LPIPS (alexNet)	68.5	78.9
LPIPS (squeezeNet)	**74.5**	87.7
LPIPS (with vggNet)	**75.2**	**89.4**
MS-SSIM	73.8	**89.4**
PSNR	73.8	87.7
SSIM	**74.5**	**89.4**
VIF	**74.5**	87.7
VMAF	73.8	**89.4**
VQM	**75.5**	82.4
WaDiQAM	72.5	85.9

The best performing metrics for each CD variant are marked bold

Table 10 Correct decision rate for DNN super-resolution

Full reference metric **# of test cells**	**CD-ALL** **43**	**CD-CI+CMOS** **28**
3SSIM	74.4	78.5
AVQT	86.0	89.2
DISTS	65.1	71.4
LPIPS (alexNet)	69.8	78.5
LPIPS (squeezeNet)	76.7	78.5
LPIPS (vggNet)	79.1	82.1
MS-SSIM	**90.7**	**92.8**
PSNR	83.7	89.2
SSIM	**90.7**	**92.8**
VIF	81.4	82.1
VMAF	**88.4**	**92.8**
VQM	72.1	82.1
WaDiQAM	65.1	75.0

The best performing metrics for each CD variant are marked bold

When analyzing the metrics that can be considered as reliable, the conclusion is that most of standard metrics (SSIM, VMAF, MS-SSIM, AVQT, PSNR) are reliable for both categories with correct decision rates above 85% and which sometimes reach 90%.

Evaluation on the DS-LAB data set: In this section, the ability of the objective metrics to provide an accurate decision is investigated on the challenging DS-LAB data set.

First, the correction decision rates relative to the REV results are computed and reported in Table 11. It appears that the most efficient metrics are SSIM, VMAF, MS-SSIM and the three LPIPS variants. The CD-rate of all metrics is significantly lower on DS-LAB than on the full set, which clearly highlights the challenging character of this data set. With a maximum CD-rate of 74.1%, their overall ability to take correct decisions is not proven.

Table 11 Correct decision rate on DS-LAB compared to REV results

Full reference metric	**CD-ALL**	**CD-CI+CMOS**
# of cells	**51**	**27**
3SSIM	54.9	66.7
AVQT	60.8	59.2
DISTS	33.3	25.9
LPIPS (with alexNet)	**64.7**	70.4
LPIPS (with squeezeNet)	62.7	70.4
LPIPS (with vggNet)	60.8	70.4
MS-SSIM	**66.7**	**74.1**
PSNR	60.8	66.7
SSIM	**68.6**	**74.1**
VIF	**64.7**	66.7
VMAF	62.7	**74.1**
VQM	52.9	48.1
WaDiQAM	58.9	63.0

The best performing metrics for each CD variant are marked bold

In Table 12, the correct decision rate is reported when using the LAB results. Only the CD-ALL information is provided: in the other scenario, the number of concerned test cells is too low to provide meaningful results. Here again, the LPIPS family provides the best results, in particular the vggNet and squeezeNet approaches.

When comparing the decision rates achieved in the two tables, i.e. REV vs. LAB, we observe that the CD-rate is lower with the LAB than with the REV test, except for LPIPS (vggNet and squeezeNet), 3SSIM and DISTS.

Table 12 Correct decision rate on DS-LAB compared to LAB results

Full reference metric	**CD-ALL**
# of cells	**51**
3SSIM	**62.7**
AVQT	52.9
DISTS	49.0
LPIPS (with alexNet)	60.8
LPIPS (with squeezeNet)	**66.7**
LPIPS (with vggNet)	**64.7**
MS-SSIM	**62.7**
PSNR	60.8
SSIM	60.8
VIF	60.8
VMAF	58.8
VQM	52.9
WaDiQAM	49.0

The best performing metrics are marked bold

3.4 Discussion of full-set vs. REV pair results

It is noted that the results presented in Sections 3.2 and 3.3 are computed with somewhat diverging objectives: the full pair-wise metric evaluation in Section 3.2 reveals the overall performance ranking of the metrics based on the possible pair combinations considering all available PVS under test. These results provide an indication of the global performance of the objective metrics on the challenging DS-LAB data set. The decision rate evaluation on the REV test pairs in Section 3.3 strictly focuses on those pairs which have been requested for study in the standardization context of JVET. Hence, these results give an impression of the performance of the metrics under different conditions: the differences between the evaluated pair of PVS are tied to always be evaluated for the same test sequence. Given this pair-wise sequence-bound evaluation, the performance ranking of the metrics is modified. While from the global perspective, DISTS appears to be the best-performing metric on the evaluated DS-LAB data set, the established simple metrics, such as PSNR, SSIM, or MS-SSIM, raise in the performance ranking on the sequence-bound setting. This phenomenon may indicate that the simple metrics have more difficulty correlating with humans when the entire set of content is considered, but prove sufficiently effective for evaluating content independently. It is important to remember that the latter is the primary goal for a metric in a standardization context. This observation, obtained from the analysis of Section 3.3, is crucial in the context of this paper as the goal here is to provide insight related to suitability of metrics for decision-making.

4 REV CMOS vs. Laboratory DMOS

The results of the REV tests and the LAB tests are produced with two different rating scales. Since REV employs a CCR method, the output is a comparative MOS (CMOS). The CMOS is accompanied by a confidence interval (CI), resulting from the comparative measurement. If the range CMOS±CI includes the zero value, then the two test points of the comparison are noted as equivalent. This case is denoted as "$A = P$" in the following, with "A" and "P" corresponding to the anchor and proposal, respectively. Otherwise, either the anchor or the proposal is considered superior, which is then denoted as "$A < P$" or "$A > P$", respectively.

In the LAB tests, the DCR method is employed. Here, the impairments of the PVS are scored relative to the provided uncompressed reference, resulting in a DMOS. Again, the measurement relative to the uncompressed reference comes with a confidence interval. To compare the MOS values of two test points, such as the anchor and the proposal, two approaches are considered. First, the distance of the MOS values, taking the CI into consideration, is evaluated. If MOS+CI of the lower MOS value does not overlap with MOS-CI of the higher MOS value, the two points can be considered as significantly different. In this case, a conclusion of the proposal being superior to the anchor (or vice versa) can be drawn. If the confidence intervals overlap, such a conclusion cannot be drawn and the two test points are noted as equivalent. Second, the difference of the MOS values, based on a one-way ANOVA (ANalysis Of VAriance), is considered [45]. Here, the confidence interval of the $\mathrm{DMOS} = \mathrm{MOS_p} - \mathrm{MOS_a}$ is computed as

$$\mathrm{CI} = t_{1-\frac{\alpha}{2},N-k}\sqrt{\sigma_{\epsilon}^{2}\left(\frac{1}{n_{\mathrm{a}}}+\frac{1}{n_{\mathrm{p}}}\right)} \tag{2}$$

where σ_{ϵ}^{2} denotes the mean squared error, $t_{1-\frac{\alpha}{2},N-k}$ denotes the value of the Studentâ€™s t-distribution for a significance level α with $N-k$ degrees of freedom, where $N = n_{\mathrm{a}} + n_{\mathrm{p}}$ indicates the number of test points for the anchor and the proposal, and $k = 2$ is the number of treatments to be considered. For the resulting DMOS value, the same considerations apply as for the CMOS.

In general, the first approach may be considered more conservative than the second. The second one provides a statistically motivated computation of the DMOS values. Both are deemed suitable for the intended comparison.

4.1 Results and analysis

Reporting the complete set of results of the REV and LAB tests would require excessive space and is, therefore, omitted here. The detailed results of the REV tests are available in [20–25]. A set of example results for the LAB and REV tests is provided in Fig. 4. For comparison purposes, the DMOS and CMOS values are plotted on the same axis. The meaning of the scale, however, is different for the two sets of results, and, therefore, a comparison of the values must be taken with care.

In the following, the results are assessed by means of confusion matrices, which present the proportion of similar and dissimilar conclusions to be drawn from the two different assessment methods. The computation of correlation coefficients, such as the PCC or the SRCC, are not deemed to be suitable for the intended analysis. The PCC relies on the assumption of a comparison of linearly scaled metrics. This condition is not met especially for the scores recorded in the REV tests. The computation of the SRCC for these results may also be considered unsuitable: due to the coarse nature of the CMOS scale with only four selectable values, results from multiple test cases may share the same rank, thereby limiting the interpretability as an ordered set. Since this effect is observed for the results of the reported tests, both the PCC and the SRCC are omitted here.

For further analysis, the categorization "$A > P$", "$A = P$", and "$A < P$" as discussed in the previous subsection, is applied to the DS-LAB test set. These are presented as confusion matrices for both the ANOVA-based and the CI-based classification in Table 13a, b.

Based on these numbers, it can be concluded that REV and LAB provided a consistent conclusion in 57.7% and 51.9% of the test cases for the categorization based on ANOVA and the overlap of the confidence intervals, respectively. In the case of the ANOVA-based method, a single case of contradicting conclusions occurs (1.9%). This case (EE1-1.2.1 Kimono QP34) is also shown in Fig. 4. Since, in all other cases, the confidence interval of either the REV or the LAB test overlap with the x-axis while it does not for the other test, these are considered to be borderline and therefore only slightly wrong.

In general, the tables are somewhat unbalanced on the main diagonal since, apparently, none of the test cases consistently indicated the anchor to be superior to the proposal for both visual test methods. This may be attributed to the fact that, by expectation, a proposal submitted for evaluation provides an improvement according to the intention of the proponent. This implicit bias is supported by the inspected test cases.

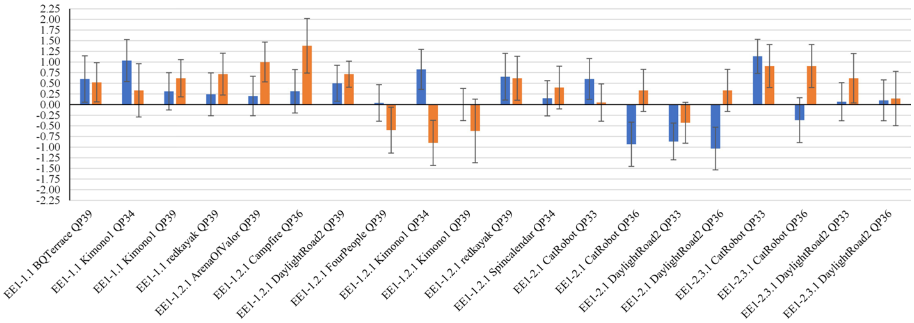


Fig. 4 Comparison of DMOS results computed from the LAB tests with the confidence interval using the ANOVA method (blue) with the corresponding CMOS results recorded in the REV sessions (orange). Note: although the value ranges of both are comparable, the meaning of the scale is different in the two sets of results

When neglecting confidence intervals and inspecting only the positive or negative direction of the DMOS and CMOS scores, about 65% of the test cases show the same tendency while about 31% diverge. Although this result has to be treated with caution, it gives an indication that the two groups of viewers, expert and naïve, tend, on average, to score similarly. Nonetheless, the confidence intervals must be considered. It is, therefore, noted that the result "$A = P$" is most frequent in the tables. For this set of results, none of the test methods provides a reliable indication of the relation between the proposal and the anchor. Since the tests generally concern PVSs with potentially small differences, it may be concluded that such cases are not reliably judged by any of the presented methods.

As a further perspective, the results for the DIF subset of the LAB test are inspected. As indicated in section 2.2, the data set DS-LAB includes this category, i.e., test points where the CI does not include the zero value. Table 13 (c) and (d) present the confusion matrices for this subset. The results emphasize the fact that the naïve viewers may not have voted as clearly as the expert viewers did. Specifically, the contradictory case mentioned above is one of these clear cases. Moreover, the cases considered to have a clear outcome in the REV are not as clear in the LAB tests. Instead, many, or even majority fall into the $A = P$ category, i.e., the naïve viewers did not express a clear decision (4̃3% for the ANOVA-based and 9̃3%for the CI-based evaluation).

When inspecting the "slightly wrong" cases it becomes obvious that for the LAB test the largest portion of the cases are observed as "$A = P$". Whereas, the REV test results in "$A < P$" or "$A > P$" most of the time (about 67% based on ANOVA and about 88% based on the CI overlap). It can be argued that the experts involved in the REV tests may have a more pronounced ability in distinguishing the quality of the inspected test sequences. Further, it must be taken into account that the experts were exposed to a forced-choice scale in their evaluation. This may push their evaluations in a certain direction but, also, it may impose an unrevealed bias if the compared PVSs are very close.

The observed single case of contradiction between the results of the experts and the naïve viewers may point to another aspect to be considered when comparing the results of these two groups of viewers. The more specialized view of the experts involved in

Table 13 Confusion matrices comparing the rating of the proposal against the anchor of the LAB and REV results, for the full set (a) ANOVA-based, (b) CI-based. for the "DIF" subset (c) ANOVA-based, (d) CI-based

REV\LAB	$A > P$	$A = P$	$A < P$
(a)			
$A > P$	0	3	1
$A = P$	3	19	4
$A < P$	0	11	11
(b)			
$A > P$	0	4	0
$A = P$	2	23	1
$A < P$	0	18	4
(c)			
$A > P$	0	1	1
$A = P$	0	0	0
$A < P$	0	5	7
(d)			
$A > P$	0	2	0
$A = P$	0	0	0
$A < P$	0	11	1

the development of the investigated tools and their standardization process may induce a biased view of artifacts induced by compression. This view may differ from a naïve person who, by definition, is not aware of the development process of a specific coding tool. This aspect must be carefully considered when interpreting the results of experts in comparison to naïve viewers.

To gain further insight into the non-congruent entries of the confusion matrices, the authors performed a personal visual assessment of the corresponding test cases, i.e., the entries in the confusion matrix where one test suggests "$A = P$" and the other test suggests "$A > P$" or "$A < P$". For this purpose, both inspected the pair of PVSs of each test case according to their own preference (such as split-view, or repeatedly viewing the sequences one after the other). For each test case, they independently recorded their personal impression. For this purpose, the established categorization "$A > P$", "$A = P$", and "$A < P$" was applied with the option of selecting several categories to express borderline cases. For instance, if the selection is not obvious, a selected impression could be "$A > P$" or "$A = P$". Yet it could also be "$A > P$" or "$A < P$", as for instance, in the situation where different artifacts with different nature appear and some viewers could sincerely score either the one or the other. However, this exercise must be considered highly subjective since only the two authors were involved. It can and will be assumed that, in no way whatsoever, these responses are ground truth. This additional viewing can only be considered as a deep investigation of the difference in qualities for some test cells; namely, trying to understand why experts or naïve viewers score them in one way or another. Or, if expert or naïve viewers have commonly failed on some test cells, trying to understand if this is due to some side-effect like the category of content, the rating scale, etc.

After completing the individual categorization task, the results were compared to each other. The first observation from this exercise is that no contradictions of the two votes were observed. In about 78% of the test cases, both votes were consistent, while in the remaining 22%, the vote of one expert was embraced by the wider vote of the other expert. The results were merged by choosing the wider vote in these cases for the further analysis.

As a next step, these results were compared to the categorization of the REV and the LAB results. Again, both variants from the LAB results, namely, the ANOVA-based and the CI-based categorizations, were considered. When comparing these results to the categorization by the authors, the (not necessarily matching) results of the REV and LAB tests were found to be included in the extended vote by the authors in about 70% of the cases. In two cases for the LAB tests and in three cases for the REV tests, the results were found to be matching with those of the authors. For the test case of EE1-1.2.1 Kimono QP34, which was marked as contradicting when applying the ANOVA-based categorization, the votes by the authors were found to be matching the result of the LAB test. Interestingly in one case, the authors unanimously voted "$A < P$" while the REV and LAB tests suggested "$A > P$" and "$A = P$" (with slight tendency towards "$A > P$"), respectively. This result may indicate the proximity of the possible results. It may also indicate that the application of the different approach in scoring leads to different conclusions.

5 Discussion

This section attempts to aggregate the most important previously reported observations and to analyze them from a more general perspective.

As a first notion, the results seem to suggest that the remote expert viewing can be considered reliable with respect to the results of the laboratory tests with naïve viewers under the applied test protocols (i.e., A/B comparison with forced-choice 4-grade scale for REV tests and DCR with the 11-grade scale for the LAB test). The observed REV results are more discriminatory than the LAB tests, which may be attributed to the fact that the participants were rating under a forced-choice regime in the REV design and/or to the fact that they are experts. We observe, however, that the opinions of the experts may include some bias, e.g., when dealing with loop filtering artifacts. An illustrative example for differences with respect to loop filtering artifacts is provided in Fig. 5 where a detail area of a test sequence with two different deblocking filter configurations is shown. It has to be noted that the example cannot illustrate the impact of overlaying temporal artifacts. Here, the focus of the experts on removing artifacts may influence their perception, and, consequently, the results compared to the naïve viewers who do not consider any particular artifact, but only the overall impression. The observed disagreement between experts and naïve viewers (as well as a further potential disagreement with the opinion of the authors) may be considered as a predictable effect. Ultimately, recorded scores reflect "personal opinions".

Another finding emerges from the two analytical approaches discussed in Sections 3.2 and 3.3. These approaches differ significantly: the former assesses the overall performance ranking of the metrics based on all possible pairwise comparisons, mixing all content, while the latter focuses exclusively on comparing pairs of similar content. The

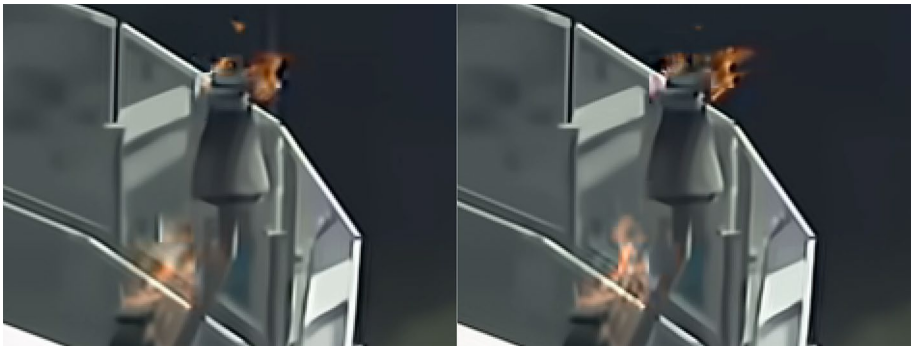


Fig. 5 Illustrative example of loop filtering artifacts: detail of the sequence BQTerrace compressed with the ECM software at QP42 and two different deblocking filter configurations (left: stronger deblocking, right: weaker deblocking)

analysis reveals that simple SSIM and MS-SSIM metrics struggle more with predicting overall quality scores compared to DISTS, or AVQT. However, they perform quite well when analyzing individual contents under different encoding configurations and rates. Notably, PSNR also demonstrates this ability. In the context of this paper, which aims to provide insights into the suitability of metrics for decision-making, the approach presented in Section 3.3 appears particularly well-suited.

However, the results for loop filter or deblocking filter proposals suggest a generally weaker match of the metrics with the opinion scores. This may specifically be true for the learning-based metrics. On the one hand, it can be argued that specific training of the metrics could compensate for this issue and, hence, a better performance of some metrics may be achieved. On the other hand, it must be considered that such training may still not encompass new types of artifacts induced by newly proposed coding tools in the codec development process. Hence, unexpected behavior of metrics may occur in such cases. We emphasize that this aspect can be considered as a specific issue of the development process for new coding tools. In the context of encoder optimization for an established and stable coding scheme, metrics learned on the features of the scheme may function in an excellent way. For the development process of coding tools, it may be advisable to include an (ideally too large) number of different metrics which are employed as a warning system: if the metrics disagree in the suggested benefit or drawback of a tool, subjective assessment may be indicated.

We highlight that the discussion above is focused on the results achieved with conventional 2D video. In the case of immersive video, the task of decision-taking, based on objective or subjective results, becomes even more challenging. This aspect is largely driven by the fact that, in the case of immersive video, the processing chain includes additional mandatory building blocks which have significant impact on the perceived quality. These include the rendering process (adapted to the intended assessment, e.g., by a conventional 2D video screen, a head-mounted display, or a mobile device), and, if applicable, the consideration of illumination. It is, therefore, advisable to generate a very clear common understanding among the involved parties, of what reference scenario(s) for assessment should be, as well as what be the agreed metrics to base the decision-taking on should be. Remote expert viewing may be considered as a suitable tool for analyzing the results of objective metrics, and the outcome of such REV tests in the course of the development may help to develop a stable set of metrics suitable for this purpose.

Some results highlight the fact that the correct decision rate is higher for the REV test than for the LAB test. Assuming the LAB test is more accurate, this means that the metrics fail. Conversely, assuming that the REV test is more accurate, this means that the metrics are heading in the right direction. More generally, in difficult situations humans apparently have different judgments. Let us, once again, review the practical example of deblocking. The viewers are classified into two categories: some of them will give higher scores to a proposal that reduces a blocking effect, while others prefer an anchor that still exhibits some blocking artifacts and, generally, better preserves the sharpness of textured areas. In this situation, the scores of the viewers are averaged. In contrast to this effect, the metric may have been designed or trained to match with one category or another of viewers. Instead of reporting a single score, an ideal metric reports a distribution of opinions, such as: for 40% of the viewers the score will be x, while for 60% of the viewers, the score will be y. Since this kind of metric does not exist, it is problematic to report that metrics are failing in situations where humans disagree on the evaluation of the video.

Due to the difficulty of deciding on small changes in a coding scheme, employing frequent LAB tests in the standardization development process may not be advisable. Nonetheless, a confirmation of the successive decisions, through so-called “verification tests”, at the end of the standardization process, appears advisable to confirm the overall suitability of the developed design.

From our findings in this study, we cannot give a definite answer to the question which process, among REV, LAB or objective metrics, to trust most. It became evident that none of them fully fails or fully succeeds. The selection of the most suitable means must be guided by the context. Some insights were given with this goal in mind. In general, for decision-taking in video coding tools development, it would be helpful to have a small number of criteria to consider, such as a single quality metric (besides other aspects such as encoding/decoding complexity). Unfortunately, this tends to over-simplify the complex nature of the impact a coding scheme may have on the visual media signal. Hence, the consideration of multiple metrics, and the usage of (remote) expert viewing in cases of doubt or contradiction of metrics is recommended.

6 Conclusions

This paper discusses the task of quality assessment for decision-taking in the development of visual media coding schemes, considering 2D videos. A comprehensive study of the relation between remote expert viewing tests, laboratory tests, and objective quality metrics is provided. The results reveal that, in this study, the well-established quality metrics such as PSNR, SSIM, or MS-SSIM perform at a high rate of correct decisions when comparing their results with both types of viewing tests. Among the learning-based metrics, VMAF and AVQT appear to be the most robust. In particular, no specific training of the metrics was applied which may have had an impact on the performance of the metrics. Such training may be a difficult task in the context of tool development (in contrast to other applications, e.g., with established coding schemes).

Remote expert viewing is found to be reliable with respect to the results of the laboratory tests with naïve viewers in our study. Their viewing is accompanied by a tendency to be more discriminatory, but also to be potentially biased concerning specific artifacts resulting from coding tools such as deblocking or loop filters. This observation suggests REV to be a suitable tool for decision-taking.

The work in this field is still wide open with many aspects to be addressed. In a next step, a study inverting the perspective of this paper may be considered, i.e., investigating on the distance of metric values and the potential indication of reliability with respect to a visual test. Other aspects include a careful assessment of the REV design, including the suitability of the scale, with or without forced choice, and with or without equidistant allocation of the grade values. It is pointed out that the remote aspect of the presented scheme is not a requirement for the method but stems from the need for a viewing method when a physical meeting is not possible. The authors believe that the findings are transferable to on-site expert viewing sessions which may have the benefit of a more controlled environment, especially if conducted in a laboratory environment.

Appendix

See Tables 14 and 15

The full data set used for the evaluation presented in this paper is presented in Tables 14 and 15. For each point, the experiment, encoder configuration, test sequence, quantization parameter QP, and the duration of the sequence are provided.

Table 14 Test points in the DNN in-loop filter category (DIL) and the deblocking (DBL). EE1-1.7 is a combination of DNN in-loop filters and deblocking

Test point	Cat.	Res.	Report
JVET-U0074 RA BasketballDrive QP37, 42; 10 s	dil	1080p	JVET-U0142
JVET-U0074 RA BQTerrace QP37, 42; 10 s	dil	1080p	JVET-U0142
JVET-U0061 RA Campfire QP37, 42; 10 s; **INC, SIG**	dil	UHD	JVET-V0173
JVET-U0104 RA Campfire QP37, 42; 10 s	dil	UHD	JVET-V0173
JVET-U0104 LD DaylightRoad2 QP42; 3 s	dil	UHD	JVET-V0173
JVET-U0104 LD Foodmarket QP42; 3 s; **LC**	dil	UHD	JVET-V0173
JVET-U0115 RA Campfire QP37, 42; 10 s	dil	UHD	JVET-V0173
JVET-U0115 LD DaylightRoad2 QP42; 3 s	dil	UHD	JVET-V0173
JVET-U0115 LD Foodmarket QP42; 3 s	dil	UHD	JVET-V0173
JVET-W0130 RA Campfire QP42; 10 s; **INC, DIF**	dil	UHD	JVET-W0186
JVET-W0130 RA DaylightRoad QP42; 5 s; **SIG**	dil	UHD	JVET-W0186
JVET-X0066 RA CatRobot QP37, 42; 5 s; **SIG, DIF**	dil	UHD	JVET-X0209
JVET-X0066 RA DaylightRoad2 QP37, 42; 5 s; **INC, SIG**	dil	UHD	JVET-X0209
JVET-X0066 RA Tango QP37, 42; 5 s; **SIG, DIF**	dil	UHD	JVET-X0209
EE1-1.1 RA FourPeople QP34, 39; 5 s	dil	720p	JVET-Y0212
EE1-1.1 RA Jets QP34, 39; 5 s	dil	720p	JVET-Y0212
EE1-1.1 RA Johnny QP34, 39; 5 s	dil	720p	JVET-Y0212
EE1-1.1 RA KristenAndSara QP34, 39; 5 s	dil	720p	JVET-Y0212

Table 14 (continued)

Test point	Cat.	Res.	Report
EE1-1.1 RA SpinCalendar QP34, 39; 5 s	dil	720p	JVET-Y0212
EE1-1.1 RA ArenaOfValour QP34, 39; 5 s	dil	1080p	JVET-Y0212
EE1-1.1 RA BQTerrace QP34, **39**; 5 s; **INC, SIG**	dil	1080p	JVET-Y0212
EE1-1.1 RA Kimono1 QP34, 39; 5 s; SIG, SIG	dil	1080p	JVET-Y0212
EE1-1.1 RA RedKayak QP34, **39**; 5 s; **DIF**	dil	1080p	JVET-Y0212
EE1-1.2.1 RA FourPeople QP34, **39**; 5 s; **INC, SIG**	dil	720p	JVET-Y0212
EE1-1.2.1 RA Jets QP34, 39; 5 s	dil	720p	JVET-Y0212
EE1-1.2.1 RA Johnny QP34, 39; 5 s	dil	720p	JVET-Y0212
EE1-1.2.1 RA KristenAndSara QP34, 39; 5 s	dil	720p	JVET-Y0212
EE1-1.2.1 RA SpinCalendar QP34, 39; 5 s; **INC, SIG**	dil	720p	JVET-Y0212
EE1-1.2.1 RA ArenaOfValour QP34, **39**; 5 s; **INC, DIF**	dil	1080p	JVET-Y0212
EE1-1.2.1 RA BQTerrace QP34, 39; 5 s	dil	1080p	JVET-Y0212
EE1-1.2.1 RA Kimono1 QP34, 39; 5 s; **INC, DIF, LC**	dil	1080p	JVET-Y0212
EE1-1.2.1 RA RedKayak QP34, 39; 5 s; **INC, DIF**	dil	1080p	JVET-Y0212
EE1-1.2.1 RA Campfire QP36, 39; 5 s; **INC, DIF**	dil	UHD	JVET-Y0212
EE1-1.2.1 RA CatRobot QP39, 42; 5 s	dil	UHD	JVET-Y0212
EE1-1.2.1 RA DaylightRoad2 QP39, 42; 5 s	dil	UHD	JVET-Y0212
EE1-1.2.1 RA Tango QP39, 42; 5 s	dil	UHD	JVET-Y0212
EE1-1.2, EE1-1.7 RA FourPeople QP34, 39; 5 s	dil	720p	JVET-Z0053
EE1-1.2, EE1-1.7 RA Johnny QP34, 39; 5 s	dil	720p	JVET-Z0053
EE1-1.2, EE1-1.7 RA KristenAndSara QP34, 39; 5 s	dil	720p	JVET-Z0053
EE1-1.2, EE1-1.7 RA ArenaOfValour QP34, 39; 5 s	dil	1080p	JVET-Z0053
EE1-1.2, EE1-1.7 RA BQTerrace QP34, 39; 5 s	dil	1080p	JVET-Z0053
EE1-1.2, EE1-1.7 RA Kimono1 QP34, 39; 5 s	dil	1080p	JVET-Z0053
EE1-1.2, EE1-1.7 RA RedKayak QP34, 39; 5 s	dil	1080p	JVET-Z0053
EE1-1.2, EE1-1.7 RA SpinCalendar QP34, 39; 5 s	dil	1080p	JVET-Z0053
EE1-1.2, EE1-1.7 RA Campfire QP34, 37; 5 s	dil	UHD	JVET-Z0053
EE1-1.2, EE1-1.7 RA CatRobot QP39, 42; 5 s	dil	UHD	JVET-Z0053
EE1-1.2, EE1-1.7 RA DaylightRoad2 QP39, 42; 5 s	dil	UHD	JVET-Z0053
EE1-1.2, EE1-1.7 RA Tango QP39, 42; 5 s	dil	UHD	JVET-Z0053
VTM, ECM **RA FourPeople** QP37, **42**; 5 s	dbl	720p	JVET-Z0053
VTM, ECM RA Jets QP37, **42**; 5 s; **INC, SIG**	dbl	720p	JVET-Z0053
VTM, ECM RA Johnny QP37, 42; 5 s	dbl	720p	JVET-Z0053
VTM, ECM RA KristenAndSara QP37, 42; 5 s	dbl	720p	JVET-Z0053
VTM, ECM RA SpinCalendar QP37, 42; 5 s	dbl	720p	JVET-Z0053
VTM, ECM RA ArenaOfValour QP37, 42 INC; 5 s; **OPP**	dbl	1080p	JVET-Z0053
VTM, ECM RA BQTerrace QP37, 42; 5 s; **SIG**	dbl	1080p	JVET-Z0053
VTM, ECM RA Kimono1 QP37, 42; 5 s; **OPP**	dbl	1080p	JVET-Z0053
VTM, ECM RA RedKayak QP37, 42; 5 s	dbl	1080p	JVET-Z0053

Test points evaluated in the LAB test are marked bold

Table 15 Test points in the DNN super-resolution (sr), DNN decomposition-compression-synthesis (dcs), reference picture resampling (rpr), HM2VTM (h2v), and Trap (trp) categories

Test point	Cat.	Res.	Report
JVET-U0096 LD Cactus QP37 ; 10 s	dcs	1080p	JVET-U0142
JVET-U0096 LD SlideEditing QP37; 10 s	dcs	720p	JVET-U0142
JVET-U0053 RA CatRobot QP37; 5 s	sr	UHD	JVET-V0173
JVET-U0053 RA Tango QP37; 5 s	sr	UHD	JVET-V0173
JVET-U0096 RA Campfire QP37; 10 s	dcs	UHD	JVET-V0173
JVET-U0096 RA DaylightRoad QP37; 5 s	dcs	UHD	JVET-V0173
JVET-U0096 RA Tango QP37; 4.9 s	dcs	UHD	JVET-V0173
JVET-U0099 RA Campfire QP37; 10 s	sr	UHD	JVET-V0173
JVET-U0099 RA DaylightRoad QP37; 5 s; **SIG**	sr	UHD	JVET-V0173
JVET-U0099 RA Tango QP37; 4.9 s	sr	UHD	JVET-V0173
VTM RA Campfire QP37; 5 s	rpr	UHD	JVET-V0173
VTM RA CatRobot QP37; 5 s; **DIF**	rpr	UHD	JVET-V0173
VTM RA Foodmarket QP37; 5 s	rpr	UHD	JVET-V0173
VTM RA Tango QP37; 4.9 s	rpr	UHD	JVET-V0173
JVET-Q0105 RA Campfire QP37; 5 s	sr	UHD	JVET-W0186
JVET-Q0105 RA Tango QP37; 4.9 s	sr	UHD	JVET-W0186
VTM11 RA Campfire QP37; 5 s	rpr	UHD	JVET-W0186
VTM11 RA Tango QP37; 4.9 s	rpr	UHD	JVET-W0186
JVET-X0064 RA Campfire QP37, **42**; 5 s; **INC, DIF**	sr	UHD	JVET-X0209
JVET-X0064 RA ParkRunning QP37 INC 42; 5 s; **DIF**	sr	UHD	JVET-X0209
JVET-X0064 RA Tango QP37, **42**; 5 s; **LC**	sr	UHD	JVET-X0209
JVET-X0117 RA Campfire QP37,42; 5 s; **INC, DIF**	rpr	UHD	JVET-X0209
JVET-X0117 RA ParkRunning QP37, 42; 5 s; **SIG**	rpr	UHD	JVET-X0209
JVET-X0117 RA Tango QP37, 42; 5 s	rpr	UHD	JVET-X0209
JVET-X0117 RA CatRobot QP42-QP42 ; 5 s	trp	UHD	JVET-X0209
EE1-2.3.1 RA Campfire QP28, 32; 5 s	sr	UHD	JVET-Y0212
EE1-2.3.1 RA CatRobot QP33, 36; 5 s; **SIG**	sr	UHD	JVET-Y0212
EE1-2.3.1 RA DaylightRoad QP33, 36; 5 s; **SIG**	sr	UHD	JVET-Y0212
EE1-2.3.1 RA Tango QP32, 35; 4.9 s	sr	UHD	JVET-Y0212
EE1-2.1 RA Campfire QP28, 32; 5 s	rpr	UHD	JVET-Y0212
EE1-2.1 RA CatRobot QP33, 36; 5 s; **SIG**	rpr	UHD	JVET-Y0212
EE1-2.1 RA Foodmarket QP33, 36; 5 s	rpr	UHD	JVET-Y0212
EE1-2.1 RA Tango QP32, 35; 4.9 s	rpr	UHD	JVET-Y0212
EE1-2.4 RA Campfire QP34, 39; 5 s	sr	UHD	JVET-Z0053
EE1-2.4 RA CatRobot QP34, 39; 5 s	sr	UHD	JVET-Z0053
EE1-2.4 RA DaylightRoad QP34, 39; 5 s	sr	UHD	JVET-Z0053
EE1-2.4 RA Tango QP34, 39 ; 4.9 s	sr	UHD	JVET-Z0053
EE1-2.1 RA Campfire QP34, 37; 5 s	rpr	UHD	JVET-Z0053
EE1-2.1 RA CatRobot QP39, 42; 5 s	rpr	UHD	JVET-Z0053
EE1-2.1 RA DaylightRoad QP39, 42; 5 s; **SIG**	rpr	UHD	JVET-Z0053
EE1-2.1 RA Tango QP39, 42; 4.9 s	rpr	UHD	JVET-Z0053
TRP RA FourPeople Orig-QP34; 5 s	trp	720p	JVET-Z0053
TRP RA Johnny Orig-QP34; 5 s	trp	720p	JVET-Z0053
TRP RA ArenaOfValour Orig-QP37; 5 s	trp	1080p	JVET-Z0053
TRP RA BQTerrace Orig-QP37; 5 s	trp	1080p	JVET-Z0053
TRP RA DaylightRoad2 Orig-QP42; 5 s	trp	UHD	JVET-Z0053
H2V RA Darktree QP32, 37, 42	h2v	1080p	JVET-Z0053
H2V RA FontainebleauCinematicS QP32, 37, 42	h2v	1080p	JVET-Z0053

Table 15 (continued)

Test point	Cat.	Res.	Report
H2V RA FontainebleauFPV QP32, 37, 42	h2v	1080p	JVET-Z0053
H2V RA FontainebleauCinematic QP32, 37, 42	h2v	UHD	JVET-Z0053
H2V RA Racing QP32, 37, 42	h2v	UHD	JVET-Z0053
TRP RA KristenAndSara Orig-QP34; 5 s	trp	720p	JVET-Z0053
TRP RA FontainbleauFPV QP32-QP37; 5 s	trp	1080p	JVET-Z0053
TRP RA Kimono QP34-QP39; 5 s	trp	1080p	JVET-Z0053
TRP RA CatRobot QP39-QP42; 5 s	trp	UHD	JVET-Z0053
TRP RA Racing QP37-QP42; 5 s	trp	UHD	JVET-Z0053
TRP RA Tango QP39-QP42 ; 4.9 s	trp	UHD	JVET-Z0053

Test points evaluated in the LAB test are marked bold

Abbreviations

2D, 3D	Two dimensional, three dimensional
3SSIM	Three Component SSIM
6DoF	Six degrees of freedom
A	Anchor
ACR	Absolute Category Rating
AG	Advisory Group
ANOVA	Analysis of Variance
AVQT	Advanced Video Quality Tool
BTC	Basic Test Cell
CD	(Percentage of) Correct Decisions
CD-ALL	Correct decision rate on the full data set
CD-CI+CMOS	Correct decision rate when confidence interval of CMOS does not cross the x-axis and and the absolute value of CMOS is above threshold
CI	Confidence Interval
CCR	Comparison Category Rating
CMOS	Comparison MOS
DCR	Degradation Category Rating
DCS	Decomposition - Compression - Synthesis
DCT	Discrete Cosine Transform
DIF	(Selection criterion for test sequences) Clear difference between anchor and proposal
DISTS	Deep Image Structure and Texture Similarity
DMOS	Differential MOS
DNN	Deep Neural Network
DS-LAB	Data set selected for evaluation in Laboratory
ECM	Enhanced Compression Model
EE	Exploration Experiment
FN	False negative
FP	False positive
HM	HEVC test Model
INC	(Selection criterion for test sequences) Inconsistent or unclear results for the objective metrics under evaluation
ISO/IEC	International Standardization Organization/International Electrotechnical Commission
ITU-T SG16	International Telecommunications Union, Telecommunications Sector, Study Group 16
IV-PSNR	Immersive Video PSNR
JTC 1/SC 29	ISO/IEC/Joint Technical Committee 1/Sub Committee 29
JVET	Joint Video Experts Team of ISO/IEC JTC 1/SC 29 and ITU-T SG16
LAB	laboratory viewing
LC	(Selection criterion for test sequences) Large Confidence interval
LF	Loop Filter
LPIPS	Learned Perceptual Image Patch Similarity
MIV	MPEG Immersive Video
MOS	Mean Opinion Score
MPEG	Moving Pictures Experts Group (WGs and AGs in ISO/IEC JTC 1/SC 29)
MS-SSIM	Multi-scale SSIM
NN	Neural Network
OPP	(Selection criterion for test sequences) OPPosite results at two rate points
P	Proposal
PCC	Pearson Linear Correlation Coefficient
PSNR	Peak Signal-to-Noise Ratio
PVS	Processed Video Sequence

QP	Quantization Parameter
REV	Remote Expert Viewing
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
RPR	Reference Picture Resampling
SIG	(Selection criterion for test sequences) no or very low significance
SR	Super Resolution
SRCC	Spearman Rank Correlation Coefficient
SSIM	Structural Similarity Metric
TN	True Negative
TP	True Positive
TRP	Trap
VCEG	Video Coding Experts Group
VIF	Visual Information Fidelity
VMAF	Video Multi-method Assessment Fusion
V-Mesh	Video-based Mesh coding
V-PCC	Video-based Point Cloud Coding
VTM	VVC Test model
WaDiQAM	Weighted Average Deep Image QuAlity Measure
WG	Working Group
WS-PSNR	Weighted spherical PSNR
YUV	Y=Luma, UV=Chroma components of a color video signal

Acknowledgements
The authors would like to express their great thanks to Vittorio Baroncini for his significant contribution to the subjective tests used in this paper. The authors would further like to thank the anonymous reviewers for their insightful comments and suggestions. These helped a lot to improve the quality of the paper.

Author contributions
Mathias Wien conducted the viewing tests, Joel Jung provided the assessment with the objective metrics. The other parts of the paper, including writing and review, are joint work of the authors.

Funding
Open Access funding enabled and organized by Projekt DEAL. No funding institutions other than the authors affiliations supported this research work.

Availability of data and materials
For access to the original video sequences as well as processed video sequences used in this study, membership in the JVET, i.e., ISO/IEC JTC 1/SC 29/WG 5 or ITU-T SG16/Q6 is required.

Declarations

Ethics approval and consent to participate
All viewers were volunteers consenting to participation. Volunteers participating as naïve viewers in the laboratory test sessions were recruited from students at RWTH Aachen University and were paid for participating in the tests. Viewers participating in expert viewing sessions were volunteers of organizations and companies participating in JVET and/or MPEG.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 22 May 2023 Accepted: 22 May 2024
Published online: 17 June 2024

References

1. F. Bossen, X. Li, V. Seregin, K. Sharman, K. Suehring, VTM and HM common test conditions and software reference configurations for SDR 4:2:0 10 bit video. Doc. JVET-AB2010, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Mainz, DE, 28th meeting (2022), https://jvet-experts.org/doc_end_user/documents/28_Mainz/wg11/JVET-AB2010-v1.zip. Accessed 12 May 2023
2. J. Ström, K. Andersson, R. Sjöberg, A. Segall, F. Bossen, G. Sullivan, J.-R. Ohm, A. Tourapis, Working practices using objective metrics forevaluation of video coding efficiency experiments. Doc. HSTP-VID-WPOM, International Telecommunication Union (2020), http://handle.itu.int/11.1002/pub/8160e8da-en . Accessed 12 May 2023
3. M. Wien, V. Baroncini, VVC verification test report for Ultra High Definition (UHD) Standard Dynamic Range (SDR) Video content. Doc. JVET-T2020, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Teleconference

20nd meeting (2020), https://jvet-experts.org/doc_end_user/documents/20_Teleconference/wg11/JVET-T2020-v1.zip. Accessed 12 May 2023

4. M. Wien, V. Baroncini, VC verification test report for hd sdr and 360° video content. Doc. JVET-V2020, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Teleconference 22nd meeting (2021), https://jvet-experts.org/doc_end_user/documents/22_Teleconference/wg11/JVET-V2020-v1.zip. Accessed 12 May 2023
5. M. Wien, V. Baroncini, VVC verification test report for high dynamic range video content. Doc. JVET-W2020, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Teleconference 23rd meeting (2021), https://jvet-experts.org/doc_end_user/documents/23_Teleconference/wg11/JVET-W2020-v1.zip. Accessed 12 May 2023
6. M. Wien, L. Yu, V. Baroncini, Guidelines for verification testing of visual media specifications. Doc. AG5N39, ISO/IEC JTC 1/SC 29/AG 5 MPEG Visual Quality Evaluation, Teleconference 5th meeting (2021), https://www.mpeg.org/wp-content/uploads/mpeg_meetings/136_OnLine/w20975.zip. Accessed 12 May 2023
7. J.M. Boyce, R. Dore, A. Dziembowski, J. Fleureau, J. Jung, B. Kroon, B. Salahieh, V.K.M. Vadakital, L. Yu, MPEG immersive video coding standard. Proc. IEEE **109**(9), 1521–1536 (2021). https://doi.org/10.1109/jproc.2021.3062590
8. J. Jung, B. Kroon, Common test conditions for mpeg immersive video. Doc. WG4N203, ISO/IEC JTC 1/SC 29/WG 4 MPEG Video Coding, online, 7th meeting (2022)
9. S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P.A. Chou, R.A. Cohen, M. Krivokuca, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A.M. Tourapis, V. Zakharchenko, Emerging MPEG standards for point cloud compression. IEEE J. Emerg. Select. Top. Circ. Syst. **9**(1), 133–148 (2019). https://doi.org/10.1109/jetcas.2018.2885981
10. MPEG 3D Graphics: CfP for dynamic mesh coding. Doc. WG7N231, ISO/IEC JTC 1/SC 29/WG 7 MPEG 3D Graphics, online, 5th meeting (2021), https://www.mpeg.org/wp-content/uploads/mpeg_meetings/136_OnLine/w21000.zip. Accessed 12 May 2023
11. ISO/IEC JTC 1/SC 29: MPEG visual quality assessment: terms of reference. Doc. SC29N19020, ISO/IEC JTC 1/SC 29 (2020)
12. J. Jung, M. Wien, V. Baroncini, Guidelines for remote experts viewing sessions. Doc. N40, ISO/IEC JTC 1/SC 29/AG 5 MPEG Visual Quality Evaluation, Teleconference 5th meeting (2021), https://www.mpeg.org/wp-content/uploads/mpeg_meetings/136_OnLine/w20976.zip . Accessed 12 May 2023
13. J. Jung, M. Wien, Analysis of the MIV verification test dry run and recommendations. Doc. m59791, ISO/IEC JTC 1/SC 29/AG 5 MPEG Visual Quality Evaluation, online, 7th meeting (2022)
14. T. Tominaga, T. Hayashi, J. Okamoto, A. Takahashi, Performance comparisons of subjective quality assessment methods for mobile video. In: Second International Workshop on Quality of Multimedia Experience (QoMEX). IEEE (2010). https://doi.org/10.1109/qomex.2010.5517948
15. T. Kawano, K. Yamagishi, T. Hayashi, Performance comparison of subjective assessment methods for 3d video quality. In, Fourth International Workshop on Quality of Multimedia Experience. IEEE (2012). https://doi.org/10.1109/qomex.2012.6263833
16. Y. Nehmé, J.-P. Farrugia, F. Dupont, P.L. Callet, G. Lavoué, Comparison of subjective methods for quality assessment of 3d graphics in virtual reality. ACM Trans Appl Percept **18**(1), 1–23 (2020). https://doi.org/10.1145/3427931
17. P. Perez, L. Janowski, N. Garcia, M. Pinson, Subjective assessment experiments that recruit few observers with repetitions (fowr). IEEE Trans. Multimedia **24**, 3442–3454 (2021). https://doi.org/10.1109/tmm.2021.3098450
18. A. Pastor, P. David, I. Katsavounidis, . Krasula, H. Tmar, P. Le Callet, 'Discriminability-Experimental Cost' Tradeoff in Subjective Video Quality Assessment of Codec: DCR with EVP Rating Scale Versus ACR-HR. https://hal.science/hal-04363990. Accessed 17 May 2024
19. A. Pastor, P. Lebreton, T. Vigier, P.L. Callet, Comparison of conditions for omnidirectional video with spatial audio in terms of subjective quality and impacts on objective metrics resolving power. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. https://doi.org/10.1109/icassp48485.2024.10448123
20. M. Wien, DNN viewing report. Doc. JVET-U0142, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, online, 21st meeting (2021)
21. M. Wien, EE1-related: Report on results of remote viewing session. Doc. JVET-V0173, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, online, 22nd meeting (2021)
22. M. Wien, V. Baroncini, A. Segall, EE1-related: Report on results of remote viewing session. Doc. JVET-W0186, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, online, 23rd meeting (2021)
23. M. Wien, EE1-related: Report on results of JVET-X remote viewing session. Doc. JVET-X0209, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, online, 24th meeting (2021)
24. M. Wien, AHG4: REV result for AHG11/EE1 and AHG10/Deblocking. Doc. JVET-Y0212, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, online, 25th meeting (2022)
25. M. Wien, [AHG4] REV result for AHG11/EE1 and AHG4 new test sequences. Doc. JVET-Z0053, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, online, 26th meeting (2022)
26. ITU-R BT.500-14: Methodologies for the Subjective Assessment of the Quality of Television Images. ITU-R. https://www.itu.int/rec/R-REC-BT.500-14-201910-I/en Accessed 2023-05-12
27. ITU-T P.808: Subjective Evaluation of Speech Quality with a Crowdsourcing Approach. ITU-T. https://www.itu.int/rec/T-REC-P.808/en. Accessed 12 May 2023
28. ITU-T P.910: Subjective Video Quality Assessment Methods for Multimedia Applications. ITU-T. http://www.itu.int/rec/T-REC-P.910-200804-I/en. Accessed 12 May 2023
29. M. Wien, V. Baroncini, Results of subjective testing of responses to the cfp for dynamic mesh coding. Doc. N57, ISO/IEC JTC 1/SC 29/AG 5 MPEG Visual Quality Evaluation, Teleconference 7th meeting (2022)
30. FFmpeg. https://ffmpeg.org/. Accessed 12 May 2023
31. VideoLan: VLC Media Player. https://www.videolan.org/vlc/index.html. Accessed 12 May 2023
32. V. Baroncini, A. Norkin, A.M. Kotra, K. Andersson, K. Misra, H. Jang, C.M. Tsai, D. Rusanovskyy, Subjective assessment of CE11 (deblocking filter) proposals. Doc. JVET-M0906, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Marrakech, 13th meeting (2019)
33. M. Wien, P. Hanhart, Core experiment viewing test procedure and results. Doc. JVET-N0835, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Geneva, 14th meeting (2019)

34. M. Wien, Core experiment viewing test procedure and results. Doc. JVET-O1118, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Gotenburg, 15th meeting (2019)
35. P. Hanhart, L. Krasula, P.L. Callet, T. Ebrahimi, How to benchmark objective quality metrics from paired comparison data? In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). IEEE. https://doi.org/10.1109/qomex.2016.7498960
36. L. Krasula, K. Fliegel, P. Le Callet, M. Klima, On the accuracy of objective image and video quality models: New methodology for performance evaluation. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). IEEE. https://doi.org/10.1109/qomex.2016.7498936
37. C. Li, A.C. Bovik, Content-weighted video quality assessment using a three-component image model. J. Electron. Imag. **19**(1), 011003 (2010)
38. F. Xiao, DCT-based video quality evaluation. Doc. Vol. 769, Final Project for EE392J
39. H.R. Sheikh, A.C. Bovik, Image information and visual quality **15**(2), 430–444. https://doi.org/10.1109/tip.2005.859378
40. AVQT. https://developer.apple.com/videos/play/wwdc2021/10145/. Accessed 12 May 2023
41. VMAF. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652. Accessed 12 May 2023
42. S. Bosse, D. Maniry, K.R. Muller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans. Image Process. 2017, 27(1), 206–219. https://doi.org/10.1109/tip.2017.2760518
43. R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. https://doi.org/10.1109/cvpr.2018.00068
44. K. Ding, K. Ma, S. Wang, E.P. Simoncelli, Image quality assessment: Unifying structure and texture similarity, 1–1. https://doi.org/10.1109/tpami.2020.3045810
45. NIST/SEMATECH: e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook/. Accessed 12 May 2023

Publisher's Note