AIC-3 Contribution from Cloudinary: CID22

Jon Sneyers, Elad Ben Baruch, Yaron Vaxman
Cloudinary

***Abstract*—We propose a new methodology for large-scale subjective quality assessment of compressed still images in the high fidelity range (medium quality to visually lossless). It combines two different assessment protocols, one based on pairwise comparisons, the other aimed at obtaining absolute opinions. The methodology is designed to allow assessing the range of qualities not well-covered by previous methodologies, and to be suitable for large-scale crowd-sourced experiment setups. This new methodology was applied to create CID22 — the Cloudinary Image Dataset '22. It is a large set of 22,153 annotated images, mostly in the medium-high quality to near visually lossless range, originating from 250 pristine images compressed using JPEG, JPEG 2000, JPEG XL, HEIC, WebP, and AVIF, with a dense sampling of encoder settings. The quality scores are based on 1.4 million opinions. Using this data, we evaluate the performance of the various image encoders as well as various objective metrics.**

■ BETTER DISPLAY and camera technology, increased storage capabilities, broadband Internet, and advances in image coding have created the conditions for high fidelity images to become increasingly feasible, desirable and widespread. However, quality assessment standards such as ITU-R Rec. BT.500 [1] and AIC-1 [2] are not very suitable for the range between high quality and visually lossless, as the quality scores obtained in this way tend to saturate at relatively low fidelity. The AIC-2 Annex B [3] methodology approaches the problem from the other end: it is based on a very sensitive flicker test that will catch even the slightest visual distortion. It is a protocol that leads to binary results: an image is either visually lossless or it is not. Therefore, for somewhat lower fidelity targets and more economical bitrates, the test might not be adequate.

In this context, the JPEG Committee has launched a new activity aiming at developing a new standard for subjective and objective image quality assessment (IQA), known as AIC-3, which is sensitive and discriminative in the currently uncovered range from high quality to near visually lossless quality (cf. **Figure 1**). A call for contributions on subjective IQA methods [4] was announced in October 2022.

In response to this AIC-3 call for contributions, we propose a new subjective IQA methodology. It is based on a dual approach combining the results of two different assessment protocols, making it feasible to perform a large-scale assessment task within a reasonable budget and time frame. We present the Cloudinary Image Dataset (CID22), a set of 22,153 images, originating from 250 pristine reference images. We discuss how this dataset was constructed, as well as the methods we used to obtain accurate quality annotations — mean bias-corrected opinion scores (MCOS scores). Besides the assessment protocols and experiment setup, we discuss outlier detection and bias correction procedures, as well as the analysis required to combine the results of the two different experiment types. Then we present results on encoders for various image codecs (JPEG, JPEG XL, WebP, and AVIF), both in terms of

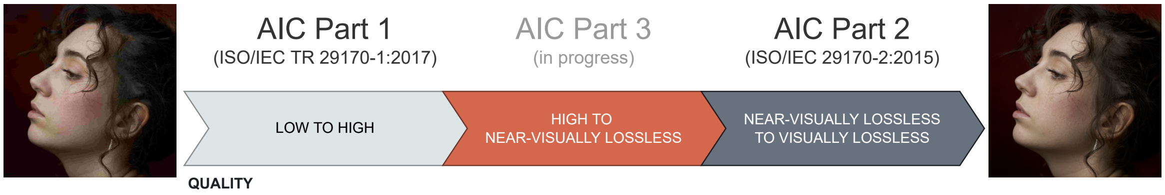

AIC Part 1
(ISO/IEC TR 29170-1:2017)
AIC Part 3
(in progress)
AIC Part 2
(ISO/IEC 29170-2:2015)
LOW TO HIGH
HIGH TO NEAR-VISUALLY LOSSLESS
NEAR-VISUALLY LOSSLESS TO VISUALLY LOSSLESS
QUALITY

Figure 1. Image quality ranges covered by existing and upcoming AIC standards.

compression (bitrate-distortion) and in terms of visual consistency of encoder settings; we believe this latter aspect has not been investigated in the IQA literature before. Finally we evaluate objective metrics in terms of their correlation with the subjective results, and propose a new objective metric called SSIMULACRA 2.

RELATED WORK

Compared to lab-based IQA datasets, the number of images in CID22 — 250 pristine images, 21,903 distorted images — is rather large: for example, the LIVE IQA database [5], [6] has 29 pristine and 779 distorted images, and the Tampere Image Database TID2013 [7] has 25 pristine and 3000 distorted images.

Crowd-sourced IQA datasets like KADID10k [8] (81 pristine, 10,125 distorted images) and the PieApp dataset [9] (200 pristine, 20,280 distorted images) are larger. The main difference between CID22 and these existing datasets (both the lab-based ones and the crowd-sourced ones) is the types and amplitudes of the distortions: CID22 covers only image compression and a specific range of qualities (medium quality to near visually lossless), as opposed to the wide range of distortions and qualities contained in existing datasets. For example, in the KADID10k set, only two out of 25 distortion types correspond to image compression (JPEG and JPEG 2000), and out of the 5 distortion levels, only 2 or 3 are within the quality range that would be typically used in practice for still images (the remaining distortion levels are too strong). In other words, out of the 10,125 distorted images, less than 5%, perhaps 400 images, are (directly) relevant for practical image compression use cases. TID2013 similarly contains relatively few distorted images relevant to image compression. The PieAPP dataset contains a wide variety of distortion types, but again only few distortions relevant to image compression (again JPEG and JPEG 2000), and mostly in the extremely low quality range.

As a result, CID22 is possibly less relevant than these existing datasets for research into modeling the human visual system and subjective quality perception *in general*, but more relevant to study the range of qualities typically used in practical image compression applications.

The KonJND-1k database [10] contains data on a large number of pristine images (1008) and the distortions are relevant to image compression (JPEG and BPG, which is similar to HEIC as it is based on HEVC). This dataset provides data related to the picturewise just noticeable difference, i.e. the threshold of distortion where an average observer can notice (or object to) compression artifacts. While relevant for practical image compression applications, a limitation of this dataset is that it does not allow comparing different image codecs (every pristine image was only compressed with one codec) and only provides information on a specific quality point, rather than a range of qualities.

IQA protocols

In [11], an overview is presented of the various image quality assessment protocols described in AIC-1 [2] and AIC-2 [3]. Single stimulus approaches like ACR and ACR-HR are suitable for assessing the appeal of a distorted image, but not the fidelity, since the test subjects cannot compare a stimulus to a reference image. The DSCQS and DSCS protocols [1], even though they are double stimulus approaches, are also more suited for assessing appeal rather than fidelity: the test subject does not know which stimulus is the 'correct' reference image, so it is possible that a distorted image will get a score that is 'better than the original'. This typically happens when the reference image is noisy or grainy, and compression artifacts act like a denoising filter. The

DSIS protocol is suitable for assessing fidelity, but since the stimuli are presented side-by-side, it is not discriminative in the high fidelity range. Comparing two very similar images side-by-side is, after all, a rather hard task. There is even a genre of puzzles ("spot the 7 differences") devoted to specifically this task. Hence it is not surprising that when using the DSIS protocol, MOS confidence intervals overlap with those of the reference image at relatively low bitrates.

In-place image comparison (as opposed to side-by-side) makes it considerably easier to spot the differences. An extreme example of this is the flicker test described in AIC-2 Annex B. This protocol allows test subjects to notice even the tiniest visual difference, and can be used to assess whether a codec can achieve full visually lossless compression. However at somewhat lower qualities, the outcome of this test will simply be "no, not visually lossless". So it is not discriminative in the range below visually lossless.

By amplifying the visibility of distortions, *boosted* triplet comparison [12] improves the discriminative power of pairwise comparisons (PC). Additionally, by presenting three stimuli (reference and two distorted images), it can assess fidelity, unlike double stimulus pairwise comparison protocols which are effectively assessing only appeal. One of our proposed protocols is effectively a variant of boosted triplet comparison.

Pairwise comparisons between distorted images derived from the same reference image, with the simple question "which image has the highest quality?" (with a binary or ternary answer) allow making a detailed ranking of distorted images by perceived quality, without necessarily requiring a large amount of opinions. While the number of pairs is $O(n^2)$, it is not necessary to exhaustively obtain opinions on all of these pairs. The main problem with this approach is that it only leads to a ranking, i.e. relative mean opinion scores (RMOS), where the lowest ranked image gets a score of 0 and the highest ranked image gets a score of 1. Such RMOS scores cannot meaningfully be compared across images originating from different reference images.

A Thurstonian analysis of pairwise comparisons allows expressing RMOS values in absolute units of just noticeable difference (JND), or rather, just objectionable difference [13]. These are comparable across images originating from different reference images. However, in practice this approach tends to only be reliable for low JND values. At higher distortion levels, comparing JND values across different reference images becomes problematic. For example, one distorted image X may be 3 JND units removed from reference image A, while another distorted image Y may be 4 JND units removed from reference image B, but that does not necessarily imply that the MOS score of X is higher than the MOS score of B when both the image content and the distortion types are different. The number of noticeable 'steps' a distorted image is removed from a reference image is not necessarily inversely proportional to its perceived quality. For example, one can imagine that if the distortion is a color shift from green-blue to a purple-blue, there might be many intermediate steps of distortion amplitude that are noticeable different, while if the distortion is a Gaussian blur, only few intermediate steps might be noticeable. Still, the image with the shifted color might receive a higher MOS score than the blurred one.

It is therefore hard to accurately convert scores obtained through pairwise comparisons to absolute quality categories across a range of image content and distortion types. While accurate rankings can be obtained, the rankings are separate per reference image and interpreting them in a consistent absolute way is hard when the distance from the reference grows. Boosting allows improving the accuracy of the rankings, but it does not solve the problem of divergence at higher JND (or lower RMOS) values.

Impairment scale methodologies like DSIS [1] have the advantage that they lead to mean opinion scores (MOS) on an absolute scale, which can be directly compared across different reference images. They do however require collecting many opinions in order to obtain MOS scores that are sufficiently accurate, and even then, confidence intervals tend to be too large to allow accurately ranking distorted images. This is a problem when the number of distorted images to be tested is large or the range of qualities is relatively narrow.

A hybrid approach combining MOS and PC tests was previously proposed [14]. We did not apply active sampling to maximize the expected information gain of paired comparisons. Avoiding the dependency on previous test rounds in practice simplifies crowd-sourcing workflows, since a single batch of tasks can be used.

ASSESSMENT PROTOCOLS

We propose the following hybrid approach:

- pairwise comparisons (Triple Stimulus Boosted Pairwise Comparison, TSBPC) are performed in order to produce RMOS scores, covering all distorted images; the comparisons are mostly done between different distortion types (different encoders) at not-too-different distortion amplitudes;
- absolute grading (Double Stimulus Boosted Quality Scale, DSBQS) is performed to obtain MOS scores for a subset of the distorted images ("anchors");
- MOS scores for all non-anchor images are interpolated based on the RMOS scores and the anchor MOS scores.

Triple Stimulus Boosted Pairwise Comparison

The TSBPC protocol consists of displaying three stimuli: a reference image R, distorted image A, and distorted image B. The reference image is displayed on the left side of the screen and the participant knows this is the reference image. On the right side of the screen, one distorted image is displayed, and the participant can freely switch between image A and image B by pressing a key or clicking a button; this toggles between the two images, replacing them in-place and instantaneously. Half of the participants see A first, the other half sees B first. There is no time limit and no limit on how often and how quickly the distorted images are switched. Additionally, the images are displayed with upscaling in order to fill the screen height minus the space needed for the interface. After switching at least two times, the participant can submit a ternary response: "A is best", "B is best", or "I can't choose".

The purpose of this approach is to apply some amount of boosting [12] to the triplet comparison, in order to obtain a ranking that is as precise as possible, i.e., avoiding "I can't choose" cases that are only due to the differences being too small to notice in a more superficial comparison. The images are scaled up to ensure that the physical dimensions are large enough also on high density displays; the lack of switching restrictions allows participants to effectively perform a "manual flickering effect". However, the images are not altered to exaggerate pixel-wise differences.

Double Stimulus Boosted Quality Scale

The DSBQS protocol is similar to the well-known DSIS protocol [1] with one major difference: instead of displaying the reference image and the distorted image in a side-by-side way, only one image is shown, and the participant can freely switch between the reference image and the distorted image. The interface marks which of those two (reference or distorted) is currently being displayed. Again, there is no time limit and no limit on how often and how quickly the images are switched. The images are not displayed with scaling to fit the screen, but at 'dpr1' resolution, i.e. how the image gets displayed by default in a web browser when the image is in a simple <img> tag without additional layout — in case of normal-density screens, this means one image pixel corresponds to one display pixel (1:1); in case of high-density ('retina') screens, this means one image pixel corresponds to 2x2 display pixels (2:1). In other words, one image pixel corresponds to one CSS pixel[15], which theoretically corresponds to a visual angle of 0.0213 degrees (though in practice this may only be an approximation). The aim is to make the viewing conditions as uniform as possible between test subjects — although in a crowd-sourced setup, large differences in viewing conditions will inevitably remain.

After switching at least two times, the participant can submit a numerical response on a semi-continuous scale from 0 to 10, which is described to the participants as follows:
1: very low quality; very annoying artifacts
3: low quality; mildly annoying artifacts
5: medium quality; no annoying artifacts
7: high quality; no visible artifacts
9: very high quality; no visible difference at all.

Compared to the the five-grade impairment scale of the DSIS method [1], the quality scale of DSBQS has more resolution in the high fidelity range: DSIS scale 4 ("perceptible, but not annoying") corresponds to DSBQS scale 5 ("medium quality; no annoying artifacts"). In this sense, not only does the methodology apply boosting in the viewing conditions (by allowing in-place switching), it also uses a 'boosted' quality scale.

Responses can be registered by adjusting a slider which is initially in the middle (5) and which can be moved using the mouse in increments of 0.5, or using the keyboard arrow keys in increments of 1.

In this protocol, while "manual flickering" is still allowed, the images are displayed without additional upscaling (only adjustment for high-density displays), in order to make the conditions more consistent across participants and to limit the visibility of artifacts to a relevant level. Effectively, the experimental setup is similar to opening the reference image and the distorted image in two browser tabs and manually switching between the two tabs to see the differences before making a judgement on the image quality.

EXPERIMENT SETUP

The goal of our experiment was to create a large dataset of quality-annotated images, covering various types of image content. The distortions of interest are the compression artifacts of image codecs. We focus on codecs relevant to web delivery and choices of encoders and encoder configurations that are relevant for a production environment (i.e., reasonable encode speeds).

Selection of Reference Images

An overview of the set of reference images is given in **Figure 2**. All the images have pixel dimensions of 512×512. Most were obtained by cropping and downscaling high-resolution photos sourced from the royalty-free stock photography service Pexels. The reference images are clustered into 15 categories according to image content.

Selection of Distorted Images

The following codecs and encoders were used to produce distorted images:

- JPEG: mozjpeg 4.1.0 (3 Mpx/s)
- JPEG 2000: Kakadu 8.2.2 (8 Mpx/s)
- JPEG XL: libjxl 0.6.1 (3 Mpx/s)
- HEIC: libheif / x265 2.8.0 (2 Mpx/s)
- WebP: libwebp 1.0.3 (6 Mpx/s)
- AVIF (aom s7): libaom 3.1.2 (2 Mpx/s)
- AVIF (aom s1): libaom 3.1.2 (0.1 Mpx/s)
- AVIF (aurora): wzav1 1.0.2 (1 Mpx/s)
- AVIF (aurora slow): wzav1 1.0.2 (0.3 Mpx/s)

For each of these encoders, between 8 and 11 quality settings were used, relatively densily sampling the medium to high fidelity range. For example, for mozjpeg, a default `cjpeg` command line was used, using the following values for the `-quality` parameter: 30, 40, 50, 60, 65, 70, 75, 80, 85, 90, 95. We used fixed encoder settings (as opposed to fixed bit rates) in order to better match typical usage patterns, as well as to be able to assess encoder consistency.

Some of the modern encoders can typically be configured to reach different trade-offs between speed and compression. We mostly used default configurations (which are most likely to be used in practice). While encoder speed typically varies depending on the image content and the quality setting, as an indication, the approximate encode speed in megapixels per second (Mpx/s) is indicated in the list above (single-threaded encoding on an Intel Core i7-9750H CPU at 2.60GHz). For the two AVIF encoders with a significantly slower configuration (aom s1 and aurora slow) only partial data was collected.

Selection of Stimuli

For the TSBPC (RMOS) experiment, we conceptually considered all triplets of the form (R, A, B) where both A and B are derived from reference image R, and eliminated 'trivial' triplets based on bits per pixel of the compressed image and some weak prior assumptions about codec performance. For example, a 0.5 bpp JPEG image versus an AVIF image at more than 1.5 bpp was considered a trivial comparison (likely the AVIF would be better), while a 0.5 bpp AVIF versus a 1.5 bpp JPEG image was not considered trivial, though a 0.3 bpp AVIF versus a JPEG image at more than 2 bpp would be. This filtering step helps to avoid collecting opinions expected to bring little new information. From the remaining triplets, we randomly sampled 105,155 triplets and then aimed to collect 10 opinions per triplet.

For the DSBQS (MOS) experiment, we used 10 "anchor" distorted images per reference image plus the reference image itself (presented as a distorted image). The following encoder settings were used as anchors: mozjpeg q30, q50, q70,

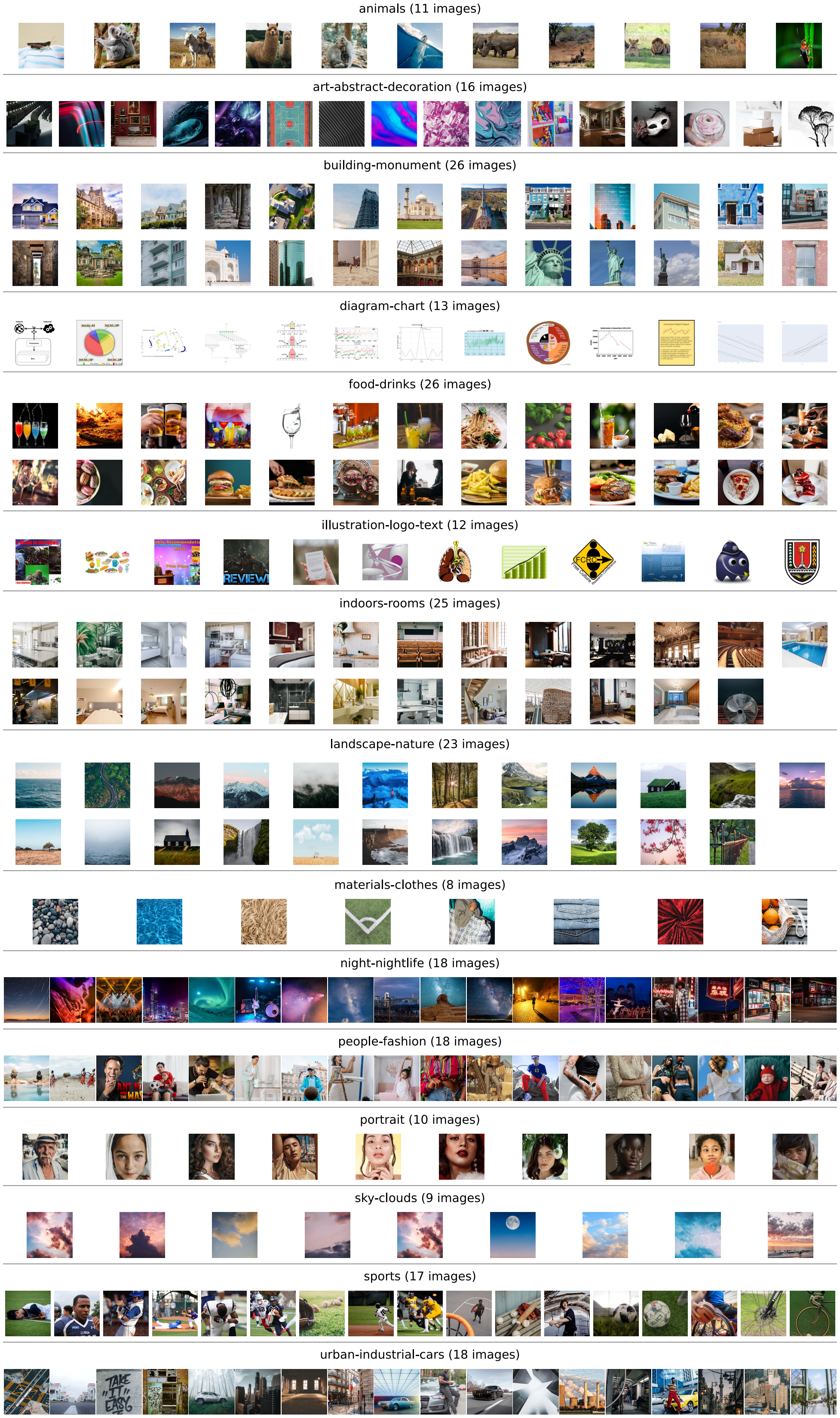

animals (11 images)
art-abstract-decoration (16 images)
building-monument (26 images)
diagram-chart (13 images)
food-drinks (26 images)
illustration-logo-text (12 images)
indoors-rooms (25 images)
landscape-nature (23 images)
materials-clothes (8 images)
night-nightlife (18 images)
people-fashion (18 images)
portrait (10 images)
sky-clouds (9 images)
sports (17 images)
urban-industrial-cars (18 images)

Figure 2. Thumbnails of all reference images in the CID22 set, clustered into 15 categories.

q90; libjxl q30, q60, q85; avif aurora quantizer settings 37, 32, 28. For each of the 2750 stimuli, we aimed to collect at least 100 opinions. Each test session started with 4 training images to expose the participant to examples of very low and very high quality, before the actual test started.

In both experiments, each test session consisted of 30 questions plus 2 additional 'honeypot' questions which were inserted randomly and used for verification. In the case of triplet comparisons, these were ‘obvious’ comparisons (A is clearly best) where a wrong answer (B is best, or “I cannot choose”) would cause the session to be discarded. For the absolute grading, these were one near-lossless image (where a score below 5 would lead to disqualification) and one very poor image (where a score above 5 would lead to disqualification). Participants could engage in up to 4 sessions, but only with a 24-hour break between the sessions in order to prevent fatigue. They were instructed to use a desktop or laptop for the experiment, and this was checked during recruitment.

PARTICIPANT SCREENING

The crowd-sourcing platform Subjectify was used to perform this experiment. In total, 1,071,300 TSBPC opinions were collected in 35,710 individual test sessions, as well as 334,920 DSBQS opinions collected in 11,164 individual test sessions. These numbers do not include participants who failed the initial ‘honeypot’ screening. The experiments were conducted in the first half of 2022.

Inevitably, a fraction of the participants in a crowd-sourced experiment is not providing high-quality responses. In order to reduce the noise introduced by such responses and to improve the accuracy of our dataset, a further screening step beyond the 'honeypot' questions was applied.

In the DSBQS experiment, sessions were discarded when one or more of these conditions were true: 1) a reference image (presented as a distorted image) received a score below 5; 2) more than 20 percent of the responses of the session (including the training and verification questions) was exactly the score of 5, which corresponds to the initial position of the slider; 3) the participant had switched to a mobile device (phone or tablet) between recruitment and actually performing the test, despite the instruction to use a desktop or laptop. This extra screening step reduced the average number of opinions per anchor image from 122 to 101.

Outlier detection

Outliers were detected and discarded as follows in the TSBPC experiment. First, for each compared triplet (R, A, B), the average opinion $MO(A, B)$ was computed by counting $A > B$ opinions as 1, $A < B$ opinions as -1, and “I cannot choose” as 0, and then taking the arithmetic mean. Next, for each participant, the agreement of their submitted opinions with average opinions was computed as follows: if the participant answered $A > B$ and $MO(A, B) > 0.3$ or if they answered “I cannot choose” and $|MO(A, B)| \leq 0.3$, then it counts as 1 (agreement); if the participant answered $A > B$ and $MO(A, B) < -0.3$ or if they answered “I cannot choose” and $|MO(A, B)| > 0.5$, then it counts as -1 (disagreement); otherwise it counts as 0 (neutral). If the average agreement (over all 30 triplets evaluated in a session) was below 0.25, then that session was discarded. In total, 5257 sessions (14.7% of all TSBPC sessions) were discarded in this way. The result is that participants answering randomly or carelessly are not included in the RMOS computation.

In the DSBQS experiment, outlier participants who frequently disagreed with the general opinion were detected as follows. For each submitted score S, the difference between S and the average score A for that stimulus was divided by the standard deviation in the set of all scores for that stimulus in order to compute a normalized difference (how many standard deviations removed from the mean). If the mean of the normalized differences in a session was greater than 1 or less than -1 (indicating very biased scoring, either very lenient or very strict), or if the standard deviation of the mean of the *absolute* normalized differences was greater than 1 (indicating random or very polarized scoring), then the session was discarded. Finally, the first three scores of each session were also discarded (effectively considering them as part of training).

After outlier removal, in the TSBPC experiment, every distorted image was on average compared to 9 other images, with 8.7 opinions per comparison (in total 395 triplets per reference image). In the DSBQS experiment, after outlier removal (and discarding the first three scores of each session), between 43 and 94 opinions were left per image (mean: 63.6).

Bias Correction

Since in the DSBQS experiment every image ends up getting scored by a different set of participants, and every participant has their own interpretation of the quality scale, it is useful to apply a bias correction before computing mean opinion scores. Scores were adjusted by shifting all the scores of a session by an additive constant, chosen per session so to make the mean normalized difference, as computed for outlier detection and in the range [-1,1], become equal to zero. The resulting adjusted scores are clamped to the $[0, 10]$ range. For example, a 'pessimistic' participant who gave scores which are on average 0.8 standard deviations below the (tentative) MOS of the images they rated, would have a mean normalized difference of -0.8. Say the average standard deviation is 2, then the adjustment constant would be +1.6, so if the original scores are (2, 3, 4.5, 10, 1.5, $\ldots$) then the adjusted scores would be equal to (3.6, 4.6, 6.1, 10, 3.1, $\ldots$).

SCORE ANALYSIS

After bias correction, the mean corrected opinion score (MCOS) was calculated for each anchor image as ten times the average of the bias-corrected scores for that stimulus. The resulting values are on a scale from 0 to 100. The MCOS score for the reference images was between 82.5 and 92.6 (mean: 88.3).

RMOS Computation

Our TSBPC experiment has an incomplete and imbalanced design by necessity, since the number of stimuli (let alone the number of pairs) is much larger than the number of comparisons done per participant. To compute relative mean opinion scores (RMOS) from the TSBPC results, we used the Elo rating system. Ratings are computed independently per reference image. The procedure we used is as follows.

All distorted images derived from a particular reference image are treated as players in a tournament. Every opinion of the form $A > B$ was counted as two wins of A against B, every "I can't choose" opinion was counted one win for each image. Since it can happen that the worst image loses against all other images, or the best image wins against all other images, which would lead to Elo scores of $-\infty$ or $+\infty$, we add 10% of a tie (0.1 win for each) between all pairs A and B (where $A \neq B$).

Based on the win rates derived from the actual opinions and the dummy opinions to enforce monotonicity and to reflect the anchor MCOS scores, converged Elo ratings can then be computed. This is the limit of the Elo ratings as the number of games played goes to infinity. Finally these ratings are normalized to the interval $[0, 1]$ to obtain the RMOS scores, so 0 corresponds to the image with the lowest Elo score (typically the q30 JPEG image) and 1 corresponds to the image with the highest Elo score (typically the q95 JPEG or JPEG XL image).

Monotonicity constraint

Furthermore, besides the actual pairwise opinions, additional information is taken into consideration in the Elo computation. While in principle (due to encoder bugs or strange phenomena) encoders do not necessarily behave monotonically, we made the assumption that all encoders we tested do in fact behave monotonically. A compressed image with a larger file size (higher quality setting) is assumed to be better than (or at least as good as) an image with a smaller file size if it is encoded with the exact same encoder (and at the same speed setting). Without this "forced monotonicity", it can happen that e.g. a q40 JPEG gets a lower score than a q30 JPEG due to the incomplete and imbalanced sampling. We add 200 dummy opinions for each pair of same-codec images to enforce the monotonicity constraint.

MCOS disagreement mitigation

Finally, for pairs of anchor images, additional opinions are added as follows. If the 90% confidence intervals of the MCOS scores of both images do not overlap, then the image with the higher MCOS score is considered to be better a number of times (we used the arbitrary constant 20) proportional to the gap between the confidence intervals — e.g. if image A has MCOS score 60 ± 4 and image B has MCOS score 83 ± 5, then the gap between the confidence intervals is $78 - 64 = 14$, so $14 \times 20 = 280$ additional opinions "$B > A$" are added. If the confidence intervals do overlap, then 200 dummy opinions are added, consisting of "I can't choose" opinions proportional to the amount of overlap, and "one is better than the other" opinions proportional to the amount of non-overlap — e.g. if image C has MCOS score 70 ± 5 and image D has MCOS score 74 ± 3, then the union of the intervals is $[65, 77]$ and has size 12, the region of overlap is $[71, 75]$ with size 4, so $200 \times 4/12$ opinions "I can't choose between C and D" and $200 \times 8/12$ opinions "$D > C$" are added.

Interpolating and extrapolating MCOS

MCOS scores of the anchor images can then be used to interpolate MCOS scores for the other images using the RMOS scores. Simple linear interpolation is used based on the nearest neighbors. There is one caveat: there are still some (rare) cases where RMOS scores and anchor MCOS scores disagree on the order of a pair. If $\text{MCOS}(A) > \text{MCOS}(B)$ while $\text{RMOS}(A) < \text{RMOS}(B)$, then the MCOS scores of A and B are slightly adjusted by moving the score of A from the mean opinion towards the 20th percentile and the score of B towards the 80th percentile until the disagreement is resolved. There were 39 such cases; a typical example would be a high-bitrate AVIF anchor getting a slightly lower MCOS score than a lower-bitrate AVIF anchor.

At the extremes, extrapolation is done as follows. The maximum RMOS score of 1 is assumed to correspond to the MCOS score of the reference image — while the reference image was not presented as part of the pairs to be compared in the TSBPC, it is a reasonable assumption that the least distorted stimulus is indistinguishable from the reference. In fact, the q90 JPEG anchor has an average MCOS of 86.7, which is quite close already to the average MCOS score of the reference image (88.3). There are several encoder settings (e.g. q95 JPEG and q95 JPEG XL) that achieve better RMOS scores than this, so it can be expected that the image with the highest RMOS score is in fact visually lossless. So arguably, no actual extrapolation is done at this end. In case a distorted anchor obtained a (slightly) higher MCOS score than the corresponding reference image, both MCOS scores are again adjusted as described above, moving the score for the distorted image towards the 20th percentile and the score for the reference image towards the 80th percentile until the scores are in the expected order. There were 28 cases where the q90 JPEG anchor image had a higher MCOS score than the reference image, 4 cases where this happened for the q85 JPEG XL anchor, and once for the highest bitrate AVIF anchor. **Table 1** gives an overview of all the adjustments done to the MCOS scores, either to resolve remaining rank-order disagreements, or to ensure that no distorted image gets a higher score than the reference image. About 5% of the scores were adjusted in this way. The amplitude of the changes was small: the largest difference is 2.59 MCOS points (on a scale from 0 to 100), the average absolute change amongst the adjusted anchor scores was 0.72 MCOS points (and 95% of the anchor scores were not adjusted).

Table 1. MCOS adjustments during interpolation.

anchor	count	min	mean	max
avif cq37	5	-1.98	-1.3121	0.00
avif cq32	11	-2.19	-0.1170	1.11
avif cq28	23	-1.06	0.8176	2.56
libjxl q30	0			
libjxl q60	12	-0.85	-0.4291	0.27
libjxl q85	6	-0.67	-0.2845	0.00
mozjpeg q30	0			
mozjpeg q50	3	-0.41	-0.2500	0.00
mozjpeg q70	12	-1.12	-0.3562	0.28
mozjpeg q90	33	-1.66	-0.5777	0.42
Reference	33	0.00	0.8864	2.59
Total	138	-2.19	0.0671	2.59

The minimum RMOS score of 0 corresponds to an anchor image in 97% of the cases; in 87% of the cases, it was specifically the q30 JPEG image that was the distorted image with the minimum RMOS score. In these cases, no extrapolation at all is required. In the remaining 3% of the cases (8 out of 250 reference images), where RMOS score 0 does not correspond to an anchor image, we extrapolate by arbitrarily assuming the lowest score to correspond to 0.75 times the mean plus 0.25 times the 20th percentile opinion for the worst anchor image. This ends up assigning an extrapolated score to the worst image of 3 to 4 MCOS points lower (on a scale from 0 to 100) than the worst anchor image for that reference image.

Final MCOS scores

The bulk (91.7%) of the images in the CID22 dataset have an MCOS score of at least 50, i.e. they correspond to "medium quality" or better. **Figure 3** shows the distribution of MCOS scores. Most images are in the range between medium-high quality (MCOS score 60) and visually lossless (MCOS score around 88). All of the tested encoders are represented well acrosss this range.

Confidence Intervals

Bootstrapping was applied on the entire score analysis process to obtain confidence intervals on the MCOS scores: 200 iterations of resampling-with-replacement were done on both sets of opinions (TSBPC and DSBQS, resampling was done after participant screening and bias correction), re-calculating the MCOS scores, Elo rankings and MCOS interpolation in every iteration. Across all MCOS scores, the width of the 90% confidence intervals obtained in this way is 4.457 on average, with standard deviation $\sigma = 1.254$. We consider these confidence intervals to be small and the accuracy of the MCOS scores to be quite high.

METHODOLOGY DISCUSSION

After describing the experiment protocols, screening procedures and score analysis, we now discuss some aspects of the proposed assessment methodology. In particular, we evaluate the effect of participant screening and bias correction, the assumption of monotonic encoder behavior and the impact of imposing it as a constraint in RMOS computation, and the relationship between sample size and accuracy. Finally we propose some potential improvements to the protocols.

Effect of screening and bias correction

The overall impact of bias correction in computing MCOS scores is relatively mild. If this step is skipped, and instead the uncorrected scores are computed, performing interpolation from anchors to the other images in the same way as before, we get MOS scores that are close to the bias-corrected MCOS scores: the Kendall rank correlation coefficient is 0.9411, Spearman rank correlation is 0.9955, Pearson correlation is 0.9953, mean absolute error is 1.006, mean square error is 1.795, and peak error is 10.82.

When additionally also not doing any participant screening, outlier removal (besides the 'honeypot question' screening that was already done by the platform Subjectify), and not discarding the first three opinions of each DSBQS session, the additional accuracy drop is also mild: KRCC is 0.9361, SRCC is 0.9947, PCC is 0.9947, MAE is 1.139, MSE is 2.197, and peak error is 10.04.

It is in our opinion still useful to perform these steps in order to improve the accuracy of the aggregated quality annotations. The effect is overall rather subtle though.

Monotonicity constraint in RMOS computation

By contrast, removing or relaxing the monotonicity constraint in the computation of RMOS scores does have a significant impact on the scores that are obtained. Removing the constraint entirely leads to much noisier MCOS scores: KRCC is 0.5559, SRCC is 0.7417, PCC is 0.7755, MAE is 7.222, MSE is 88.03, and peak error is 38.07 (compared to the MCOS scores computed with the monotonicity constraint). Relaxing the constraint by reducing the number of dummy "higher bitrate is better" opinions per same-encoder pair (from 200 to 20), so monotonicity is encouraged but no enforced, still results in scores that deviate quite strongly (KRCC is 0.8699, SRCC is 0.9773, PCC is 0.9750, MAE is 2.773, MSE is 12.78, and peak error is 15.88).

Removing the monotonicity constraint leads to scores with a more concentrated distribution: with the monotonicity constraint, the range from the 5th percentile to the 95th percentile of the MCOS scores (not including the reference images) is [45.7, 88.2], while without the monotonicity constraint, that range is [51.5, 81.4]. Still, even though the score range is narrower, the 90% confidence interval of scores obtained without the monotonicity constraint has an average width of 5.267 ($\sigma = 1.462$), as opposed to an average width of 4.457 ($\sigma = 1.254$) for scores obtained with the monotonicity constraint. So the monotonicity constraint does contribute substantially to more accurate and discriminative annotations.

We speculate that three factors are at play to explain this substantial difference. Firstly, we excluded most same-codec pairs from the TSBPC experiment, as well as any pairs involving different codecs with a large difference in bitrate.

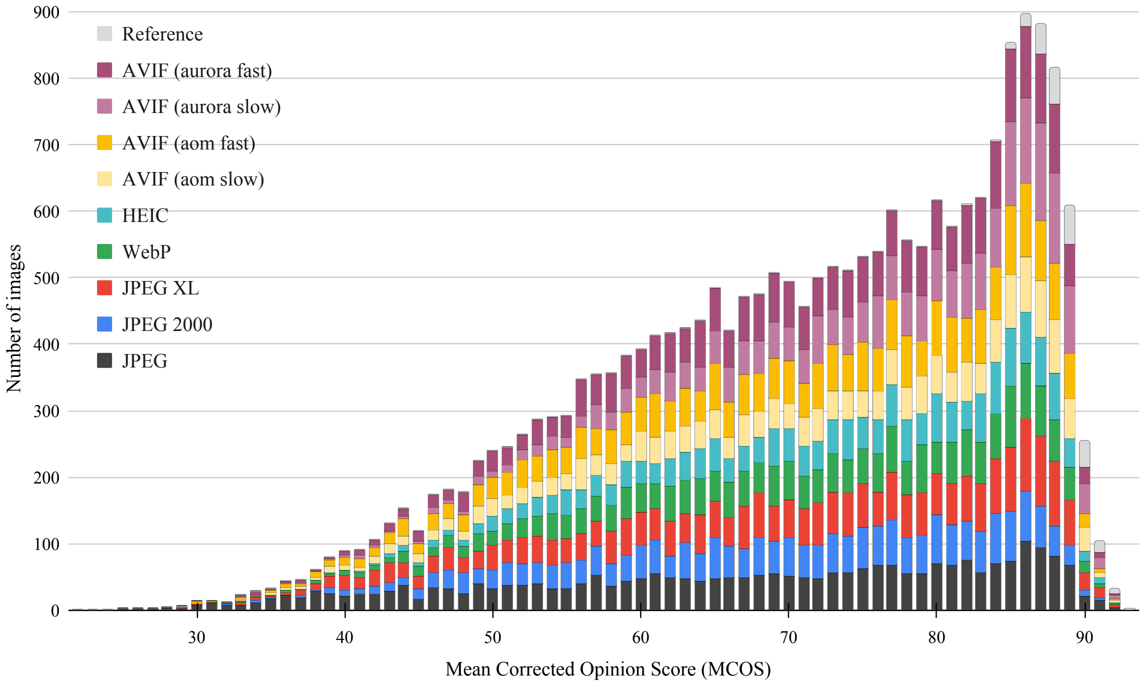


Figure 3. Distribution of MCOS scores in the CID22 dataset, by encoder.

This biased sampling made it hard to reconstruct monotonic results. Secondly, while images were on average compared to 9 other images, due to the random and biased sampling, some images were compared to more images while others were compared to perhaps only 2 other images, leading to a large confidence interval since only 20 opinions are available. In such cases, the monotonicity constraint helps, since neighboring settings of the same encoder may have been involved in more comparisons, providing additional information. Thirdly, the monotonicity constraint effectively helps to improve the interconnections between the opinions about various distorted images. For example, if for one image the TSBPC opinions indicate that webp q60 $<$ mozjpeg q60 and also that mozjpeg q65 $<$ jxl q65, then the monotonicity constraint allows deducing that webp q60 $<$ jxl q65 while without the monotonicity constraint, this would not necessarily be the case. In this way, assuming monotonicity allows constructing a more robust RMOS ranking based on a sparse, biased and noisy sampling of pairwise opinions.

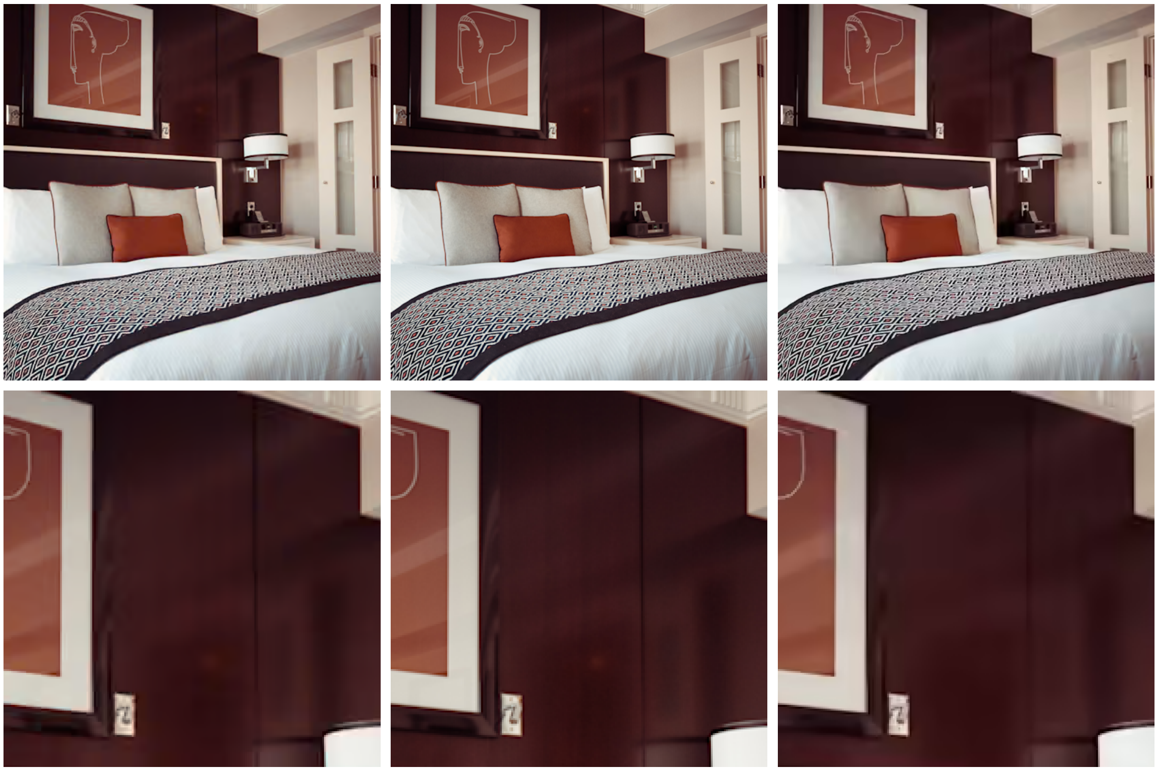
The main downsides of enforcing monotonicity are that it does not allow discovering erratic nonmonotonic encoder behavior, and that adding dummy opinions complicates a JND interpretation of the pairwise comparisons.

Effect of disagreement mitigation

The purpose of adding the MCOS-based dummy PC opinions in the RMOS computation is twofold: it complements the incomplete TSBPC data, and it helps to mitigate disagreements between TSBPC and DSBQS opinions.

Skipping this mitigation step results in scores with the following difference: KRCC is 0.7139, SRCC is 0.8868, PCC is 0.9013, MAE is 4.4717, MSE is 39.13, and peak error is 38.72.

Figure 4 shows an example of such a disagreement. The DSBQS MCOS score for the JPEG XL image on the left is higher than that of the AVIF image on the right, but according to the TSBPC data, the AVIF image is better (has a higher Elo/RMOS score). Without mitigation, this would lead to both images (and any other image with an intermediate Elo score) getting the same interpolated MCOS score. The disagreement mitigation effectively causes the DSBQS MCOS scores to prevail in case of disagreements.



JPEG XL, q60 (0.62 bpp)
DSBQS MCOS: 69.0
'raw' TSBPC Elo: 1552
'raw' MCOS: 65.6
CID22 MCOS: 67.1

reference

AVIF aurora, cq37 (0.51 bpp)
DSBQS MCOS: 59.3
'raw' TSBPC Elo: 1695
'raw' MCOS: 65.6
CID22 MCOS: 56.7

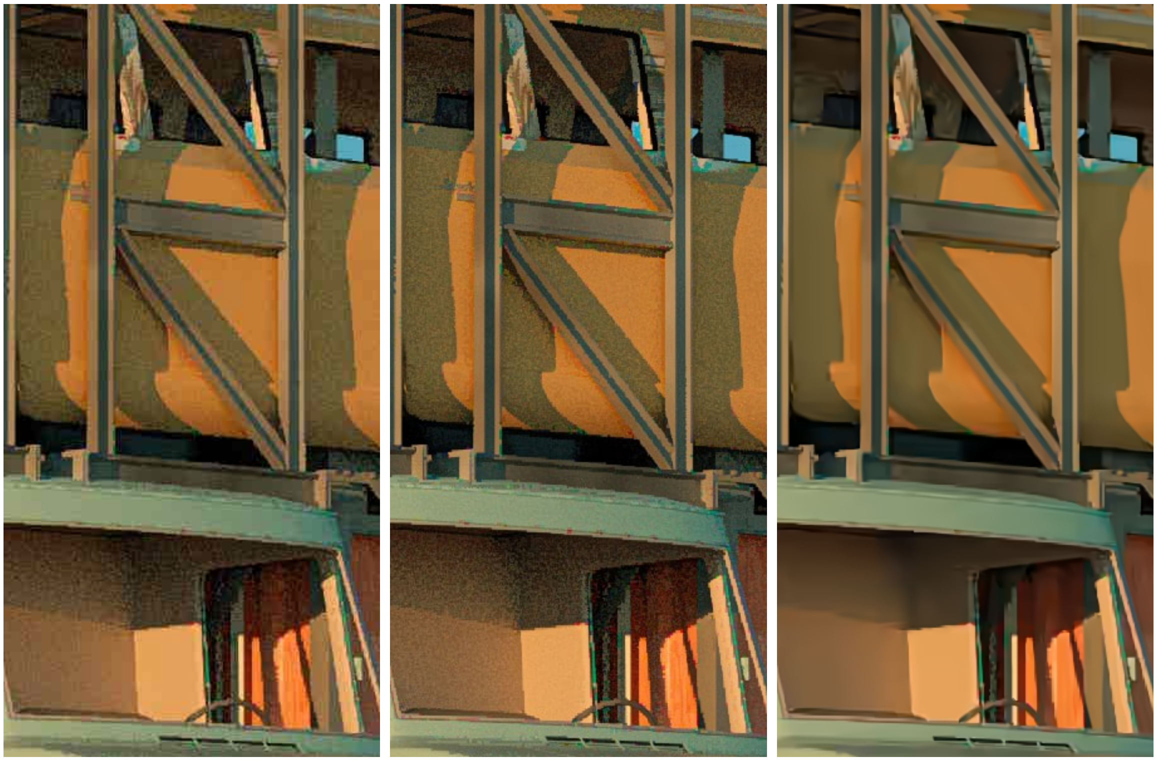
Figure 4. Example of a disagreement between DSBQS and TSBPC data.

Fidelity versus appeal

When there are disagreements like this that are not just the result of sparse sampling, we speculate that it is often caused by the difference between fidelity and appeal. For example, when comparing the JPEG XL image in Figure 4 to the AVIF image, some participants may prefer the smoother AVIF image, even if it has a lower fidelity compared to the reference image. In the DSBQS experiment, the detail loss in the AVIF image cause it to get a lower MCOS score — fidelity to the reference image is the only criterion. In the TSBPC experiment though, participants may not necessarily carefully compare the distorted images to the reference, and rather focus on the difference between the two distorted images, where the smoother one can be the preferred one.

In the experiment that was performed to create the JPEG AIC-3 CTC dataset [16], pairwise comparisons were done (in a side-by-side way) without presenting the reference image. Participants were asked to "select the image with the highest visual quality". Since participants did not see the reference image, effectively they could not assess fidelity, only appeal.

This posed a problem in particular for image number 4, a digital artwork that contains intentional (artistic) noise. **Figure 5** shows a detail from this image, compressed with JPEG and with VVC. For VVC, the Thurstonian reconstruction of the quality scale was erratic for this image. Our conjecture is that this is caused by the difference between fidelity and appeal: the VVC image effectively denoises the reference image, which is good for appeal but bad for fidelity. Since participants could only assess appeal, this effectively causes distorted images to be considered "higher quality" than the reference image.



JPEG at 2 JND | reference | VVC at 2 JND

Figure 5. AIC-3 image number 4 (detail): example of fidelity versus appeal.

Preservation of TSBPC preferences

Table 2 and **Figure** 6 show the effect of the monotonicity constraint and the DSBQS disagreement mitigation on the agreement between raw TSBPC comparison results and the scores. TSBPC delta (ΔTSBPC) corresponds to the difference between the number of $A > B$ opinions and $B > A$ opinions; the higher this number, the clearer the consensus that image A is better than image B. For example, if 1 participant said $A = B$, 6 participants said $A > B$ and 3 participants said $A < B$, then ΔTSBPC is $6 - 3 = 3$. For a single pairwise comparison where the overall opinion is not a tie, we say image scores agree with the comparison if the preferred image has a higher score.

Converting TSBPC results to RMOS scores, even when using only the raw TSBPC data without any mitigations, does not lead to perfect agreement. Numerical scores induce a total order, while TSBPC results include non-transitive preferences and sampling error so it does not correspond to any preorder. Inevitably, the mitigations to enforce monotonicity and reduce disagreement with DSBQS further reduce the agreement with raw TSBPC results. Still, even with both mitigations, the MCOS scores arguably agree well with the TSBPC opinions, especially when there is a clear consensus.

The monotonicity constraint does not significantly affect pairwise comparisons with a clear consensus; it mostly affects triplets like a q60 JPEG image A versus a q65 JPEG B where 3 opinions are $A > B$, 1 opinion is $A < B$, and 6 opinions are $A = B$. Its main effect is to reduce such 'noise' and to 'fill in the gaps' where the triplet selection was too sparse. The DSBQS disagreement mitigation however also affects some of the pairwise comparisons with clear consensus (say, $\Delta\text{TSBPC} > 5$). In those cases we situations where (probably somewhat appeal-oriented) TSBPC results (based on 10 opinions per triplet) contradict (fidelity-oriented) DSBQS results (based on 100 opinions per image). The effect of this mitigation is to let the DSBQS opinions take precedence in such cases, to the extent that they do not violate the monotonicity constraint; we wanted the quality

Table 2. Agreement between scores and TSBPC.

	mitigations			avg	
ΔTSBPC	none	monoton.	both	ΔMCOS	pairs
1	64.6%	54.1%	56.3%	1.98	12997
2	76.7%	67.4%	66.4%	3.76	11957
3	86.9%	79.6%	75.3%	5.76	11168
4	93.5%	87.8%	82.6%	7.76	11026
5	96.8%	93.5%	88.9%	10.08	11388
6	98.7%	96.6%	93.5%	12.59	11225
7	99.4%	98.6%	96.0%	14.83	10169
8	99.9%	99.4%	97.6%	17.11	8500
9	100.0%	99.8%	98.6%	19.26	5884
10	100.0%	99.8%	99.4%	21.19	3241
11	100.0%	100.0%	99.1%	20.60	454
12	100.0%	100.0%	100.0%	19.38	119

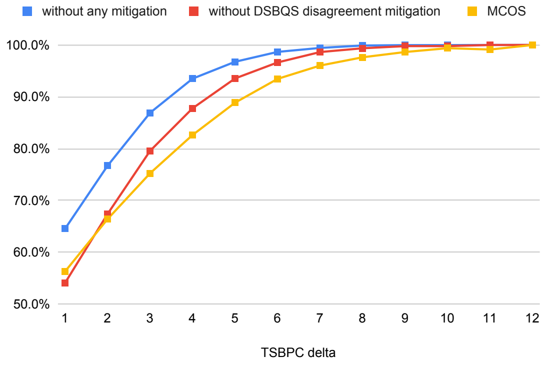


Figure 6. Agreement between scores and TSBPC.

annotations in CID22 to be more fidelity-based than appeal-based.

Interesting to note is that on average, ΔMCOS is approximately equal to ΔTSBPC times two. If the difference in MCOS scores is 20 or more, one can expect pairwise comparisons to be unanimous; if MCOS$(A)-$MCOS(B) is 10, then it can be expected that ΔTSBPC is 5, so there is a clear majority (e.g. 6 saying $A > B$, 1 saying $A < B$, 2 saying $A = B$). Smaller MCOS differences correspond to increasingly narrow majorities in pairwise comparisons.

Sample size

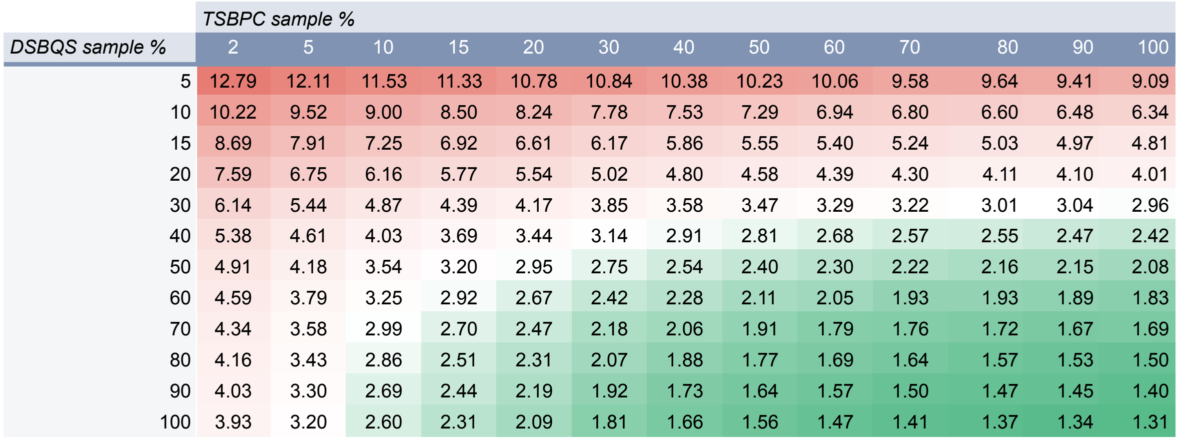
The cost of subjective testing is proportional to the number of opinions that are gathered. We can estimate the sample size that is required to reach a reasonable accuracy by simulating smaller sample sizes and measuring the accuracy of the scores obtained from smaller samplings compared to the scores obtained from the full sampling. **Figure** 7 shows the RMSE between MOS scores computed from random smaller subsets of the opinion sampling (without applying the screening and bias correction steps) and the MCOS scores computed from the full set of opinions (after screening and bias correction).

As a reminder: about 11% of the distorted images were anchors for which we did a DSBQS experiment, and we obtained about 122 opinions before screening. For the full set of distorted images, we sampled pairs (selected randomly from the set of all pairs after pruning pairs with a low expected information gain based on a priori considerations) and did a TSBPC experiment, comparing every image on average to 10 others and gathering about 10 opinions per pair before screening. Based on the tables in Figure 7, a possible recommendation for future experiments could be to aim for at least 80 DSBQS opinions per anchor image and 5 TSBPC opinions per pair (before screening), assuming a similar experiment setup. Such a reduced sample size can still be expected to produce annotations that are within the 90% confidence interval of scores obtained by a larger experiment.

Potential protocol improvements

In the TSBPC experiment, the reference image could be included in the pairwise comparison — i.e., in addition to triplets of the form (R, A, B), triplets of the form (R, R, A) or (R, A, R) could be included. This would be particularly useful when either doing a stronger form of boosting, or when the selection of encoder settings does not allow assuming that the least distorted image is effectively indistinguishable from the reference image.

Variants of the TSBPC protocol could be considered. The protocol as we implemented it presents a triplet (R, A, B) by showing R on the left side of the screen and allowing the participant to switch between A and B on the right side, so switching toggles between $R|A$ and $R|B$. In practice, this might cause participants to focus mostly on the right hand side of the screen, which could lead to a predominantly appeal-oriented judgement rather than a fidelity-oriented one. It could be better to let switching toggle between $A|B$ and $R|R$. This way the fidelity of the distorted images to the reference can be assessed directly (without eye travel). However, it might require more time though for participants to compare a pair in this way. Also, it could



	TSBPC sample %												
DSBQS sample %	2	5	10	15	20	30	40	50	60	70	80	90	100
5	12.79	12.11	11.53	11.33	10.78	10.84	10.38	10.23	10.06	9.58	9.64	9.41	9.09
10	10.22	9.52	9.00	8.50	8.24	7.78	7.53	7.29	6.94	6.80	6.60	6.48	6.34
15	8.69	7.91	7.25	6.92	6.61	6.17	5.86	5.55	5.40	5.24	5.03	4.97	4.81
20	7.59	6.75	6.16	5.77	5.54	5.02	4.80	4.58	4.39	4.30	4.11	4.10	4.01
30	6.14	5.44	4.87	4.39	4.17	3.85	3.58	3.47	3.29	3.22	3.01	3.04	2.96
40	5.38	4.61	4.03	3.69	3.44	3.14	2.91	2.81	2.68	2.57	2.55	2.47	2.42
50	4.91	4.18	3.54	3.20	2.95	2.75	2.54	2.40	2.30	2.22	2.16	2.15	2.08
60	4.59	3.79	3.25	2.92	2.67	2.42	2.28	2.11	2.05	1.93	1.93	1.89	1.83
70	4.34	3.58	2.99	2.70	2.47	2.18	2.06	1.91	1.79	1.76	1.72	1.67	1.69
80	4.16	3.43	2.86	2.51	2.31	2.07	1.88	1.77	1.69	1.64	1.57	1.53	1.50
90	4.03	3.30	2.69	2.44	2.19	1.92	1.73	1.64	1.57	1.50	1.47	1.45	1.40
100	3.93	3.20	2.60	2.31	2.09	1.81	1.66	1.56	1.47	1.41	1.37	1.34	1.31

Figure 7. An estimation of the relationship between accuracy and sample size. The table lists the root mean square error (RMSE) between MOS scores obtained from a random subset of the samples (without screening and bias correction steps) and the MCOS scores obtained from the full sample.

effectively undo the boosting effect of in-place switching, since a participant might make decisions based mostly on the $A|B$ view, ignoring the reference image and doing a side-by-side pairwise comparison. This could be avoided by presenting the stimuli as $A|R$ switchable to $R|B$, although that would likely be more confusing. Another option could be to show only a single image at a time (R, A or B), with three buttons to toggle between them (e.g., keyboard keys 1, 2, 3). This way, a participant can switch between stimuli in various ways (e.g., first alternating between A and B, then alternating between R and A and then alternating between R and B).

In the DSBQS experiment, we did not attempt to model the effect of viewing conditions. We tried to make the conditions as uniform as possible by only allowing the use of desktop or laptop computers (not mobile phones) and by displaying images at similar angular dimensions (to the extent that this can be done in a crowd-sourced experiment). Of course viewing conditions do play an important role, and it would be interesting for future experiments to gather more data on this aspect. In particular, as mobile devices and high-density displays are becoming increasingly ubiquitous, it could make sense to perform an experiment where opinions are gathered for three different types of participants: normal-density, high-density, and mobile, displaying images at native screen resolution in each group. This could bring insight into the effect of viewing conditions on perceived quality — an effect that is likely not simply a matter of rescaling the scores globally. After all, while distortions are generally harder to notice as the viewing distance increases (or equivalently, the screen density), different types of distortions will have different behavior. For example, long-range color banding might remain visible even at a larger viewing distance, while blockiness might more rapidly become unnoticeable as the viewing distance increases.

We only used images in the sRGB color space. It would be interesting to investigate the effect of wide color gamut and high dynamic range, although this is currently challenging to do in a crowd-sourced approach.

Potential methodology improvements

When combining the DSBQS and TSBPC results, effectively disagreements were resolved by letting the more fidelity-oriented DSBQS results take precedence. Both protocols aimed to assess fidelity, but the TSBPC protocol nevertheless led at least some participants to make an appeal-based decision. An interesting alternative could be to use double stimulus pairwise comparisons (not presenting the reference image) in order to explicitly assess appeal. Disagreements between appeal-based rankings and fidelity-based MCOS scores could then be studied, in order to get a better understanding of the difference between fidelity and appeal. This would also require the reference image itself to be included in the pairwise comparisons, and it should no longer be the assumed that the reference image is the image with the highest appeal (while by definition it has the highest fidelity).

Another potential methodology improvement could be to organize the experiment in two rounds. After the first round, confidence intervals can be computed for all scores, and images for which the size of the interval is above a threshold can be selected for a more focused second round to gather additional opinions on specifically these images.

ENCODER RESULTS

Figure 8 shows the overall performance of four encoders relevant for web delivery. To aggregate results over multiple images, we consider the average bits per pixel and the median MCOS score per encoder setting. This aggregation hides significant image-dependent variation in the quality obtained using a given encoder setting, as seen in the box plots that provide an indication of the spread. Obviously the bpp is also strongly image-dependent; the average does however indicate the total compressed size of the corpus.

In practice, encoder settings are often chosen using a "set it and forget it" approach where a fixed setting is used for encoding many images, and what matters is not that the median result reaches a certain fidelity target, but that all (or at least almost all) images reach a minimum fidelity target. In other words, the worst-case performance perhaps matters more than the median performance when selecting an encoder setting. **Figure 9** shows the 5th percentile MCOS scores per encoder setting.

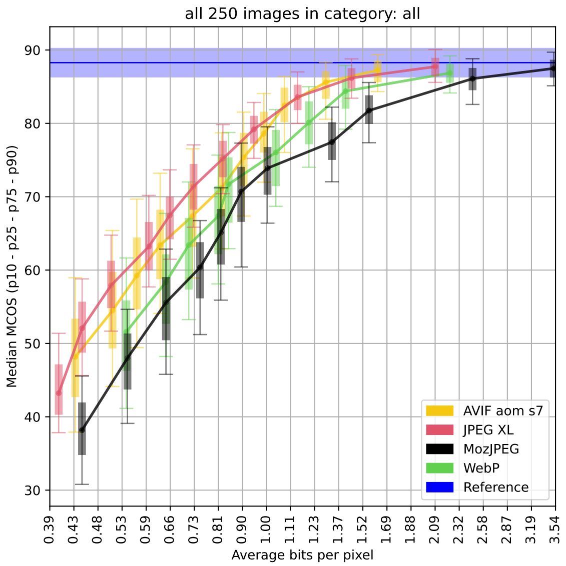


Figure 8. Median performance of selected encoders.

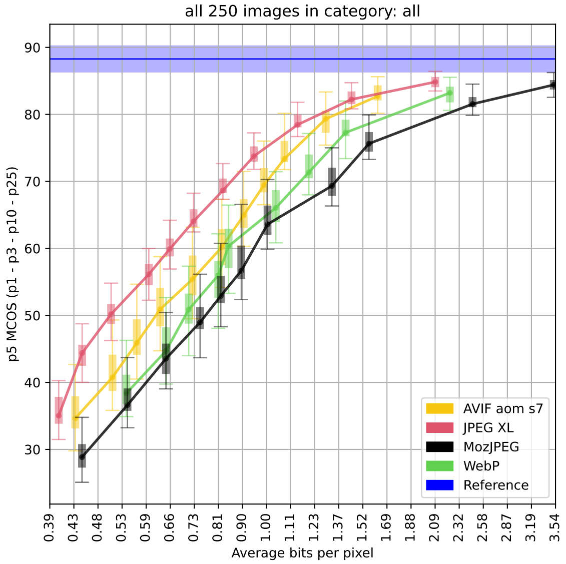


Figure 9. 5th percentile (worst-case) performance.

Encoder consistency

Consistency is a desirable encoder feature since it reduces the likelihood of 'surprising' results — in particular, 'bad' surprises where a compressed image has a noticeably worse visual quality than most other images encoded with the same setting. To investigate the visual consistency of the various encoders, we compute the standard deviation of the MCOS scores obtained for each encoder setting. **Figure 10** shows these results.

All encoders exhibit the behavior that higher quality settings produce more consistent results, which is to be expected: as the MCOS scores get closer to the highest possible value corresponding to visually lossless, variance naturally diminishes. At lower quality settings, consistency decreases for all encoders, but there are clear differences: JPEG XL is more consistent, while AVIF and WebP are less consistent.

One way to interpret the results shown in Figure 10 is as follows: if the goal is to reach a certain minimum MCOS score M_{min} for most encoded images, it is advisable to aim use an

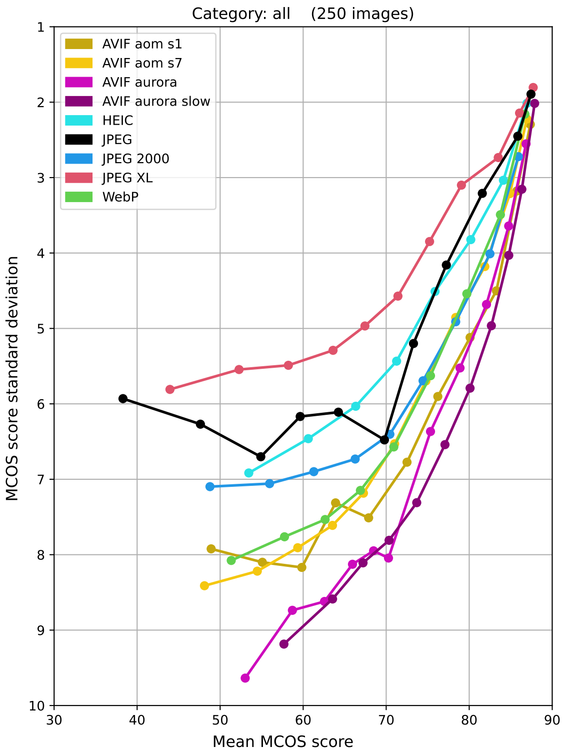


Figure 10. Visual consistency of encoder settings, as indicated by standard deviation of MCOS scores. Note: the vertical axis is flipped, so higher is better (more consistent, lower standard deviation).

encoder setting that leads to a higher average MCOS score M_{avg} with a standard deviation σ such that $M_{avg} - m\sigma \geq M_{min}$, where the choice of m determines the amount of 'risk' (of not obtaining the minimum score) one is willing to take: assuming a Gaussian distribution, $m = 1$ implies that about one sixth of the images would fall below the minimum score; $m = 2$ implies that 1 out of 50 images would be below the threshold; with $m = 3$ it would be around 1 in a 1000. For example, with $M_{min} = 60$ and $m = 2$, for JPEG XL an encoder setting could be used that will result in an average MCOS score of 70 (or even slightly lower), since $\sigma < 5$ for such a setting. For AVIF or WebP, an encoder setting would have to be used that results in a higher average MCOS score (about 73).

Traditionally, encoder assessment results are often presented as bitrate-distortion curves where the various codecs are aligned on bitrate. This obfuscates the aspect of encoder consistency and hides the practical need for a safety margin in the encoder settings.

Results by image category

Figure 11 shows plots of the median MCOS aggregated separately for each of the 15 image categories. There are some notable differences between the categories: e.g. in the two non-photographic categories ('diagram-chart' and 'illustration-logo-text'), AVIF is clearly outperforming the other codecs, while in categories like 'landscape-nature' and 'materials-clothes', AVIF is not performing better than MozJPEG.

Within each category, the relative performance of the various encoders is generally similar, though there is still image-dependent variation. By means of example, **Figure 12** shows the per-image results for the ten images in the 'portrait' category. In these plots, the anchor images are marked with a star.

OBJECTIVE METRICS

Generally the purpose of objective metrics is to predict subjective quality scores in order to easily and quickly assess image quality — algorithmically rather than involving human test subjects. They can be a valuable tool during encoder development, or even be used internally by an encoder to guide encoder choices. Objective metrics are only useful to the extent that they indeed correlate well with subjective results. The following objective metrics were computed:

- PSNR, as implemented in ImageMagick 6.9.11 (`compare -metric psnr`, clamped to 60)
  https://imagemagick.org
- VMAF [17], SSIM [6], MS-SSIM [18], PSNR-Y, PSNR-HVS [19], CIEDE2000 [20]:
  as implemented in libvmaf v2.3.0
  https://github.com/Netflix/vmaf
- Butteraugli, SSIMULACRA (1 and 2): as implemented in libjxl 0.8
  https://github.com/libjxl/libjxl
- LPIPS v0.1.4 [21]
  https://richzhang.github.io/PerceptualSimilarity
- DSSIM v3.2.0
  https://github.com/kornelski/dssim
- FSIM v0.3.5 [22], [23]
  https://github.com/up42/image-similarity-measures

Table 3 lists the Kendall rank-order, Spear-

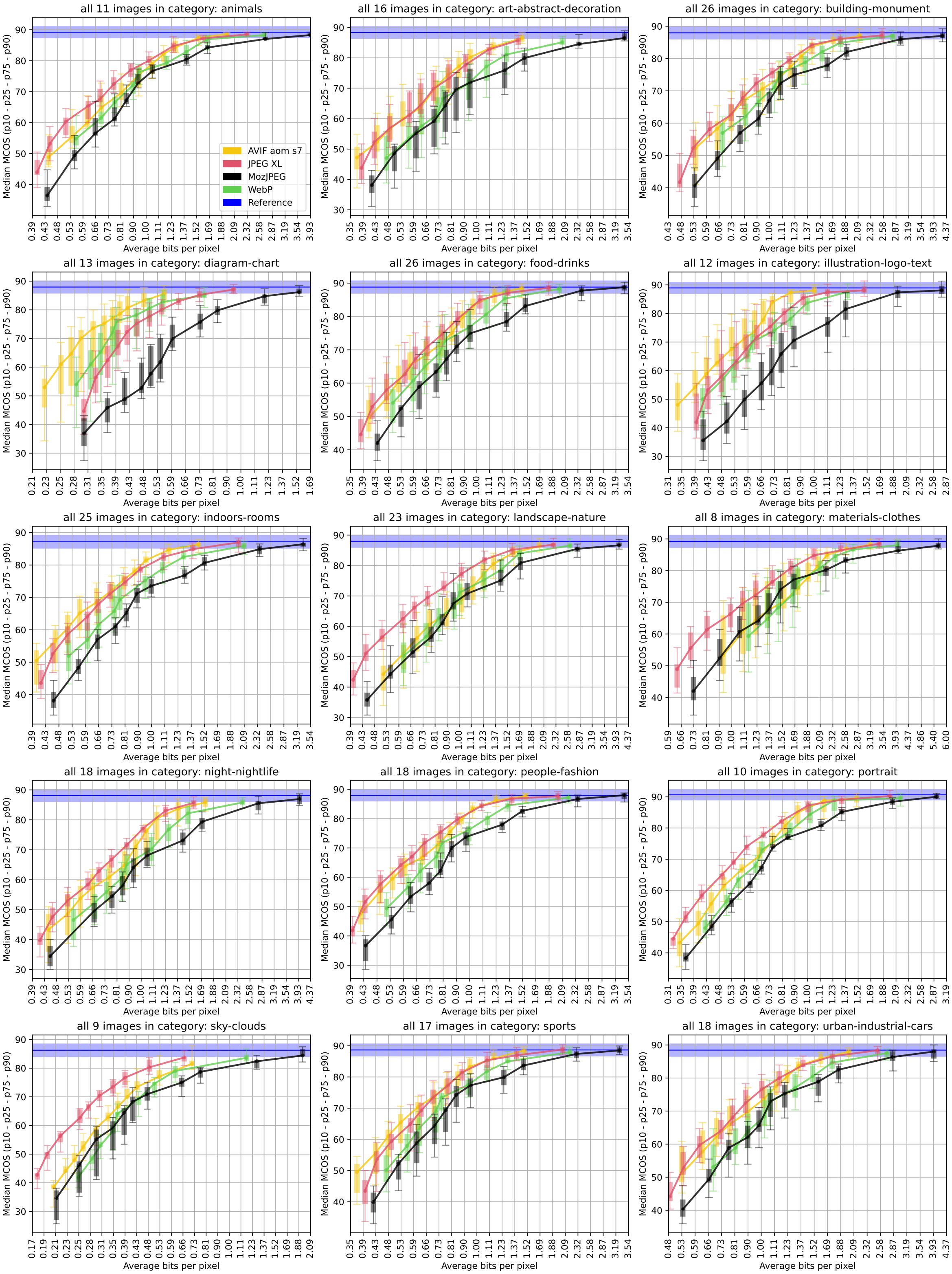


Figure 11. Median performance of selected encoders, per category.

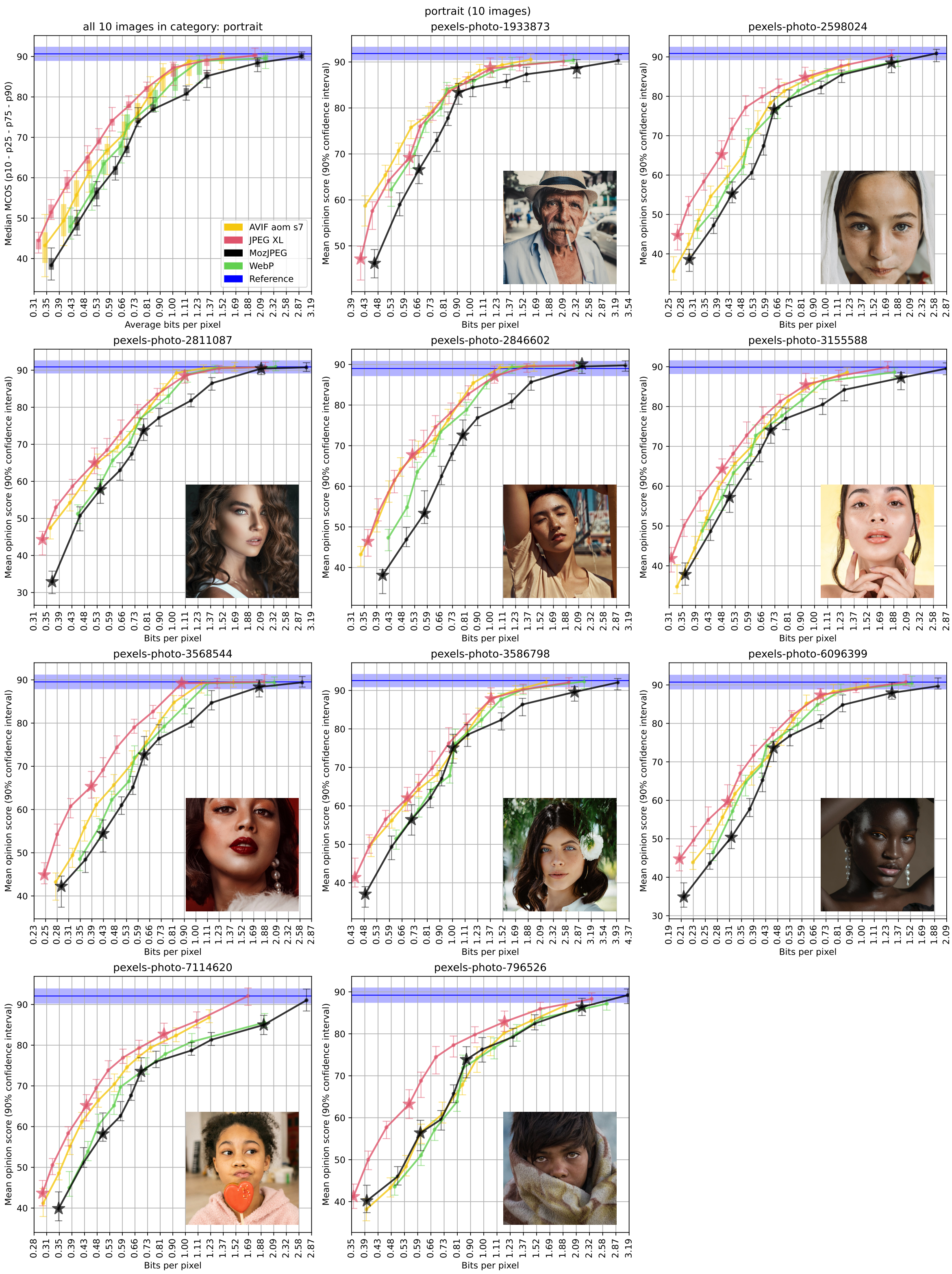


Figure 12. Per-image performance of selected encoders, for the specific category of portrait photos.

Table 3. Metric correlation with CID22 MCOS.

Metric	KRCC	SRCC	PCC
(SSIMULACRA 2)	0.6934	0.882	0.8601
Butteraugli 2-norm	**-0.6575**	**-0.8455**	**-0.8089**
Butteraugli 3-norm	-0.6547	-0.8387	-0.7903
DSSIM	-0.6428	-0.8399	-0.7813
VMAF	0.6176	0.8163	0.7799
FSIM	0.6089	0.8005	0.7676
PSNR-HVS	0.6076	0.8100	0.7559
Butteraugli max-norm	-0.5843	-0.7738	-0.7074
SSIM	0.5628	0.7577	0.7005
MS-SSIM	0.5596	0.7551	0.7035
LPIPS	-0.5417	-0.7316	-0.6932
SSIMULACRA 1	-0.5255	-0.7175	-0.6940
PSNR-Y	0.4452	0.6246	0.5901
PSNR (ImageMagick)	0.3472	0.5002	0.4817
CIEDE2000	0.3154	0.4584	0.4096

Table 4. Metric scores for KonJND-1k (mean ± stdev).

Metric	BPG images	JPEG images
PSNR-Y	39.61 ± 2.98	36.70 ± 3.79
PSNR-HVS	40.31 ± 1.78	39.96 ± 1.79
SSIM (×100)	98.55 ± 0.76	98.54 ± 0.81
MS-SSIM (×100)	99.21 ± 0.40	99.22 ± 0.38
VMAF	90.05 ± 2.25	91.56 ± 1.90
SSIMULACRA 2	65.38 ± 5.10	63.10 ± 4.65
DSSIM (×1000)	3.357 ± 1.267	3.817 ± 1.297
Butteraugli 3-norm	1.528 ± 0.192	1.699 ± 0.229
PSNR (ImageMagick)	35.17 ± 2.69	32.70 ± 3.32

Table 5. Approximate alignment of quality scales.

	medium	high	visually
Dataset / metric	quality	quality	lossless
CID22 (MCOS)	50	65	90
TID2013 (MOS)	4.5	5.5	6
KADID10k (DMOS)	3.7	4.3	4.5
KonFiG-IQA (F-JND)	1.5	0.7	0
AIC-3 (JND)	3	1.7	0
KonJND-1k (PJND)		1	
PSNR-HVS	35	40	50
MS-SSIM (×100)	98	99.2	99.8
VMAF	83	91	96
DSSIM (×1000)	8	3.5	1
Butteraugli 3-norm	2.5	1.6	0.5
SSIMULACRA 2	50	65	90

man rank-order and Pearson correlation coefficients between the CID22 dataset MCOS scores (not including the reference images) and these metrics. SSIMULACRA 2 was tuned using (part of) the CID22 dataset, so these results are likely overestimating its performance. Amongst the other metrics, the best-performing ones are indicated in bold.

Alignment to other datasets

The KonJND-1k database [10] provides a calibration point to help compare the various objective metrics and subjective IQA datasets. Metric scores can be computed for the distorted images at the (mean) PJND threshold. **Table 4** lists the mean metric scores for the two subsets (the overall mean is the average of the two numbers,; the two subsets have the same size).

Figure 13 visualizes the correlations between a selection of objective metrics and the MCOS scores of CID22 (excluding the reference images), using two-dimensional histograms. The horizontal axis corresponds to the subjective scores, the vertical axis to the metric value, and the color indicates the number of images. The region shaded in purple corresponds to the range of metric scores within one standard deviation of the mean in the KonJND-1k dataset (the purple line indicates the mean metric score corresponding to the PJND). The black curve indicates the mean MCOS score for a given metric score; the dashed black lines indicate the 25th and 75th percentiles, the dotted black lines indicate the 5th and 95th percentiles. The horizontal spread between these lines indicates the variation in subjective scores for a given metric score; a smaller spread makes a metric more reliable.

For comparison, in **Figures 14, and 15**, a similar visualization is provided for the TID2013 [7] and KADID10k [8] datasets. Note that both the amplitude and the type of distortions are quite different in these datasets: they include low and very low quality images, and mostly artificial distortions like applying blur or noise. They both only include JPEG and JPEG 2000 compression artifacts, no other encoders. **Figure 16** shows plots for the KonFiG-IQA dataset [12], specifically for the data from Experiment I with F boosting (flicker). **Figure 17** shows plots for the AIC-3 CTC [16] dataset, including only the subjective data (not the estimated data) and excluding the atypical image number 4. Based on the best-correlating metrics, the quality scales from these datasets can be approximately aligned as indicated in **Table 5**.

Pairwise correlation

For some use cases, metrics do not need to accurately predict absolute MOS scores consistently across images originating from different reference images, but it is sufficient to predict the result of pairwise comparisons of images

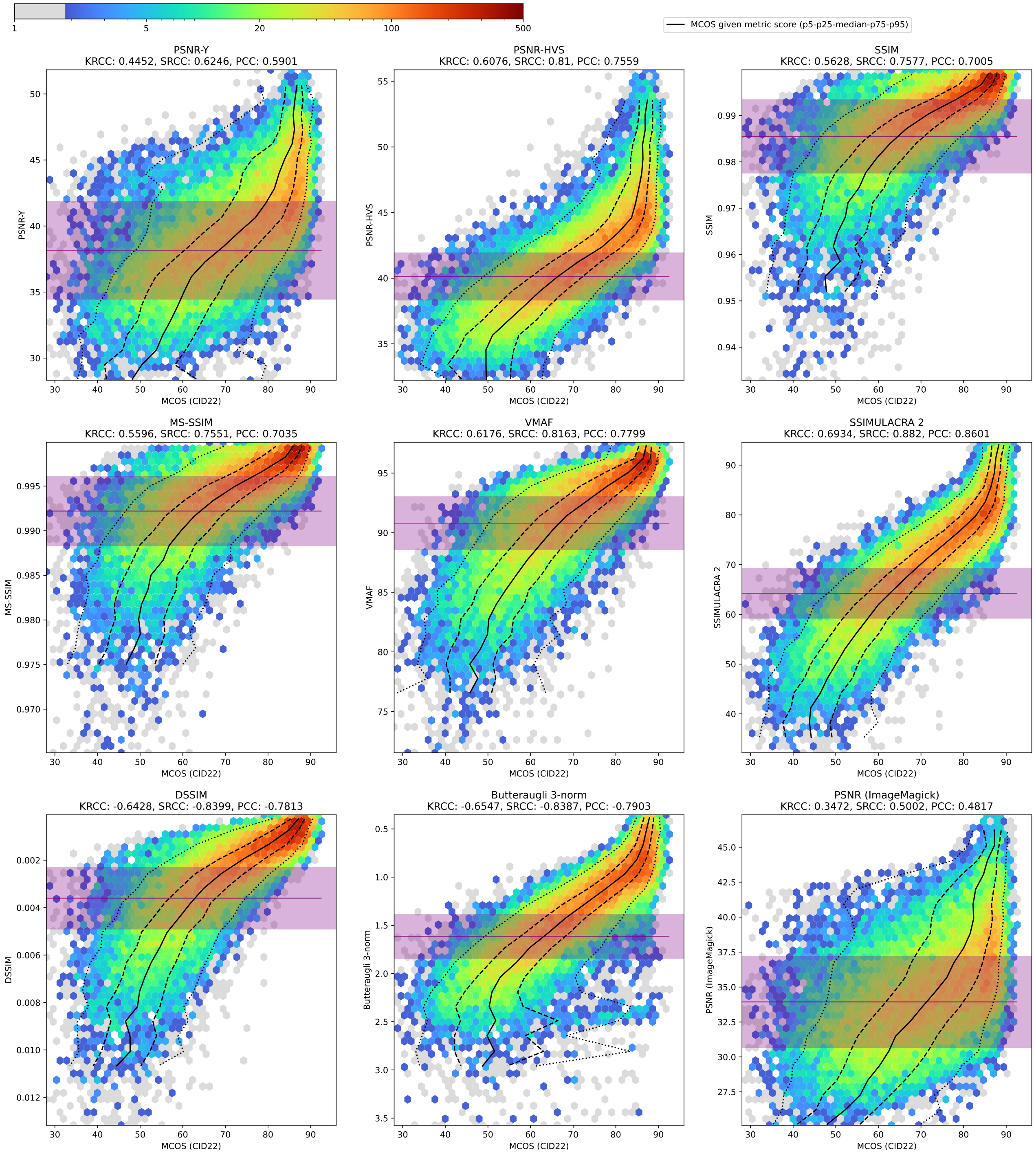


Figure 13. Correlation between objective metrics and the CID22 dataset.

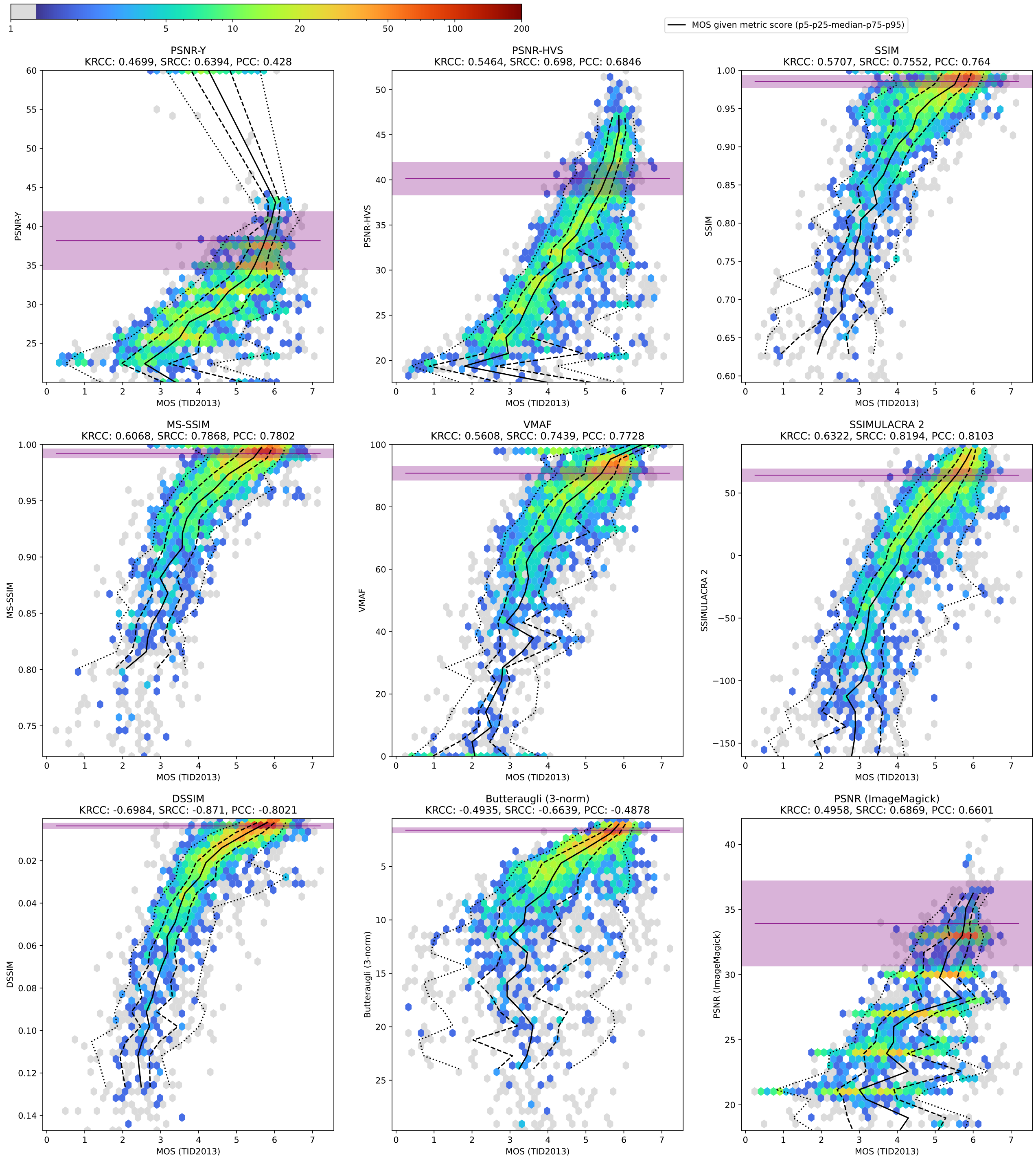


Figure 14. Correlation between objective metrics and the TID2013 dataset.

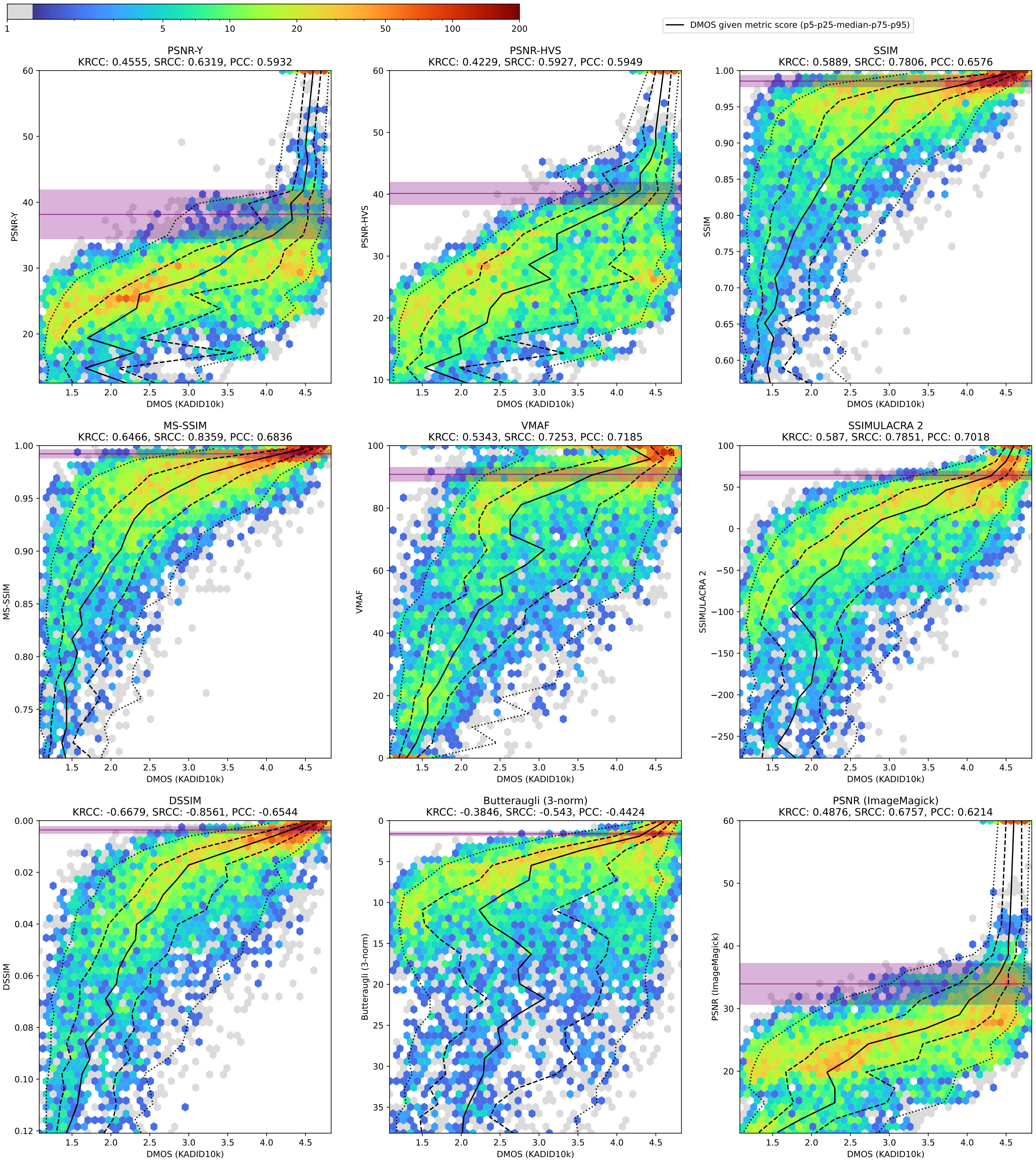


Figure 15. Correlation between objective metrics and the KADID10k dataset.

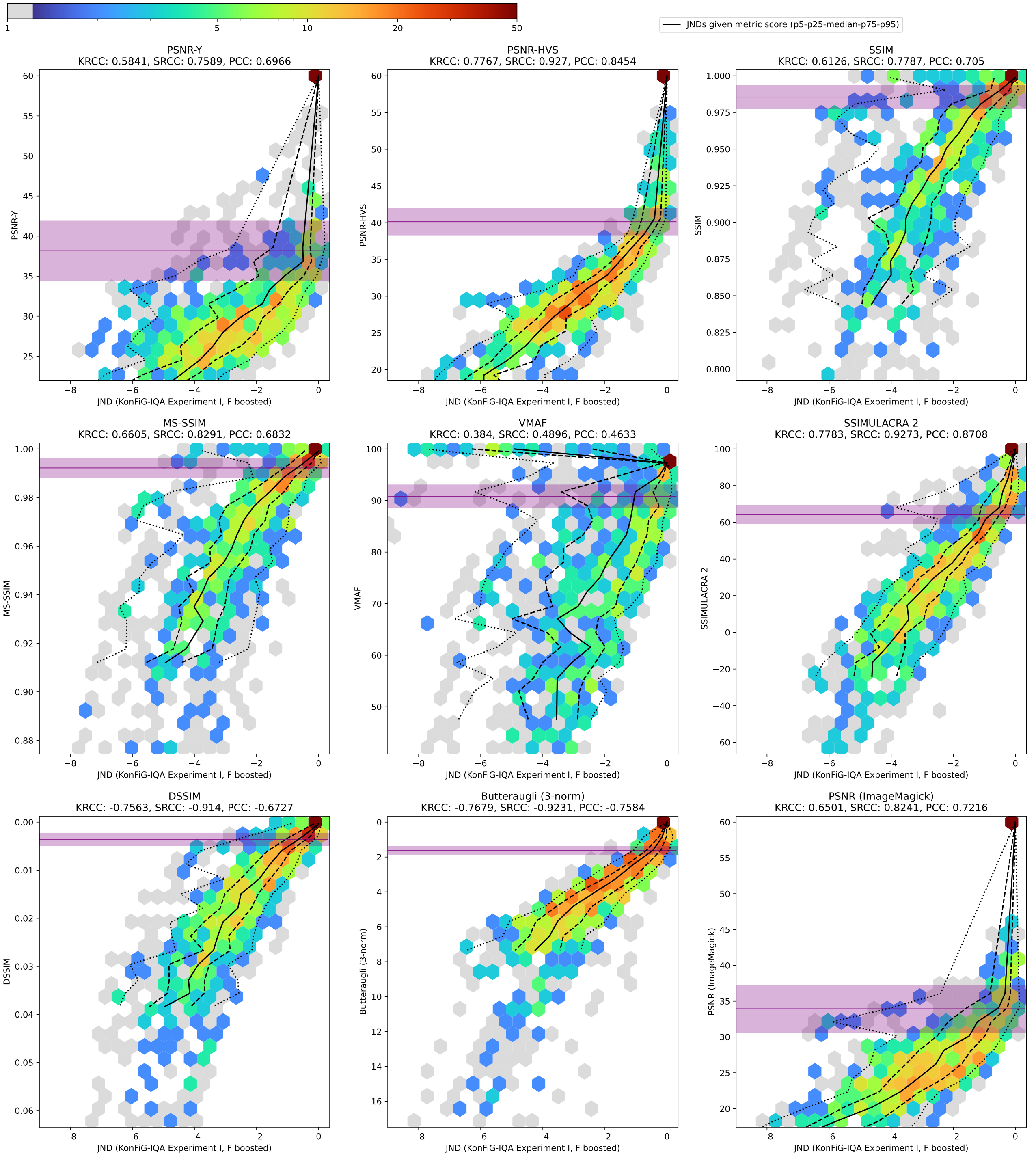


Figure 16. Correlation between objective metrics and the KonFiG-IQA dataset.

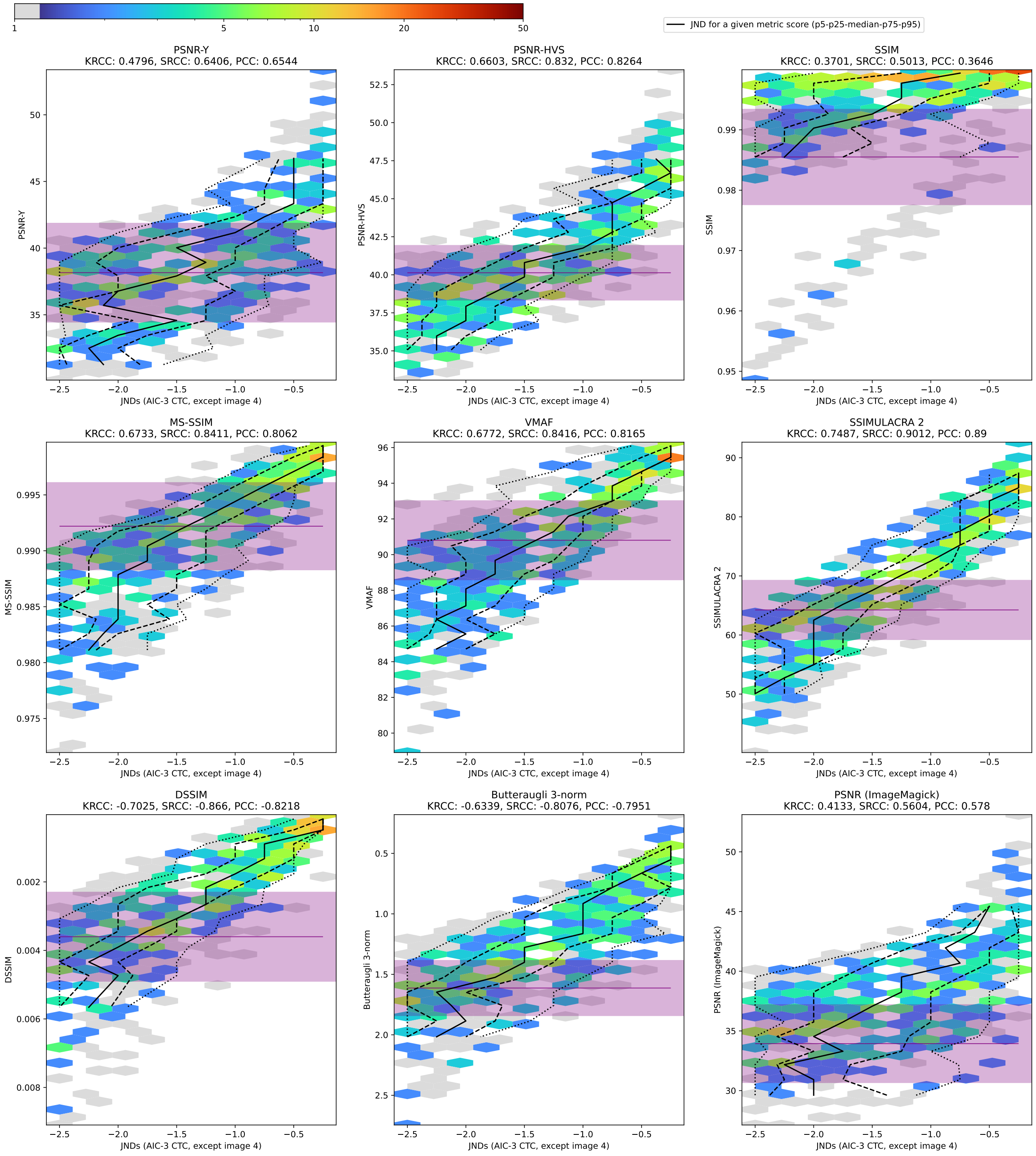


Figure 17. Correlation between objective metrics and the AIC-3 CTC dataset (excluding image number 4)

Table 6. Metric correlation with MCOS differences.

Metric	KRCC	SRCC	PCC
(SSIMULACRA 2)	0.7536	0.9210	0.9085
DSSIM	**-0.7203**	**-0.9019**	-0.8352
SSIMULACRA 1	-0.7059	-0.8915	**-0.8399**
Butteraugli 2-norm	-0.6852	-0.8688	-0.8422
FSIM	0.6828	0.8656	0.8411
Butteraugli 3-norm	-0.6787	-0.8610	-0.8252
LPIPS	-0.6711	-0.8612	-0.7901
CIEDE2000	0.6576	0.8482	0.7690
SSIM	0.6487	0.8426	0.7703
PSNR-HVS	0.6440	0.8365	0.7992
PSNR-Y	0.6264	0.8259	0.7888
PSNR (ImageMagick)	0.6214	0.8197	0.7745
MS-SSIM	0.6039	0.7967	0.7367
VMAF	0.6018	0.7894	0.7784
Butteraugli max-norm	-0.5877	-0.7773	-0.7351

originating from the same reference image, i.e., MOS differences. For example when assessing a potential encoder change, typically the aim is to improve the visual quality when keeping the bitrate constant. We can compute the correlation between the differences in scores (MCOS(A) − MCOS(B)) and the differences in metric results (metric(R, A) − metric(R, B), for all triplets (R, A, B) which were evaluated in the TSBPC experiment (this excludes 'trivial' pairs). **Table 6** lists these correlations.

Predicting pairwise comparisons (between two distorted images derived from the same reference image) is generally an easier task for an objective metric than predicting absolute quality in a way that is consistent between images derived from different reference images. For almost all metrics, the correlations in Table 6 are higher than those in Table 3. A notable exception is VMAF, which seems to be (slightly) better at absolute quality assessment than at predicting pairwise comparisons. Tables 3 and 6 are sorted from highest correlations (best metrics) to lowest correlations (worst metrics). There are large differences between these two metric rankings: e.g. SSIMULACRA 1 performs rather poorly at absolute quality assessment but is one of the best metrics for relative quality assessment, while with VMAF it is the other way around. Interestingly, PSNR performs rather well at predicting relative opinions, outperforming some of the more advanced metrics like MS-SSIM and VMAF. For absolute quality assessment however, PSNR shows only a very weak correlation.

Figure 18 shows 2D histograms of differences in MCOS (horizontally) compared to differences in metric score (vertically). The area between the two horizontal blue lines contains half of the pairs where the metric difference is small, the area between the two vertical blue lines contains half of the pairs where the MCOS difference is small. In the areas on the bottom left and top right, the metric correctly predicts the pairwise preference. In the areas on the top left and bottom right, the metric is wrong. The central area contains pairs where both the metric difference and the MCOS difference is small, so arguably the metric is correct (even if it has the sign wrong). The remaining areas are cases where the MCOS difference is large but the metric difference is small, or the other way around. The percentages of pairs in each of these regions are indicated.

SSIMULACRA 2

SSIMULACRA 2 is a new objective metric for image quality assessment, developed based on the CID22 dataset. It is technically not a distance metric since it does not respect symmetry: in general, $\text{SSIMULACRA2}(a, b) \neq \text{SSIMULACRA2}(b, a)$. In particular, if the distorted image is smooth in a region where the original image has edges, this can get penalized differently than if the distorted image has edges in a region where the original image is smooth, i.e. smoothing artifacts are weighted differently than ringing, banding or blockiness artifacts.

The metric is based on multiscale SSIM [18]. The computation is done in the XYB color space, while the downsampling between scales is done in linear RGB. SSIM error maps are computed at six scales (1:1 to 1:32) for each component. Two additional error maps are computed in order to explicitly model ringing and smoothing artifacts. For each of the resulting $6 \times 3 \times 3$ error maps, two aggregation methods are used (L_1 and L_4 norms). The final score is based on a weighted sum of the resulting 108 sub-scores, where the weights were optimized to correlate with a subset of the CID22 dataset, corresponding to 201 out of the 250 reference images. It was validated on the images derived from the remaining 49 reference images (which were not used in the weight tuning). For this validation set, the KRCC is 0.7033, SRCC is 0.88541, PCC is 0.87448 and the mean absolute error is 4.97.

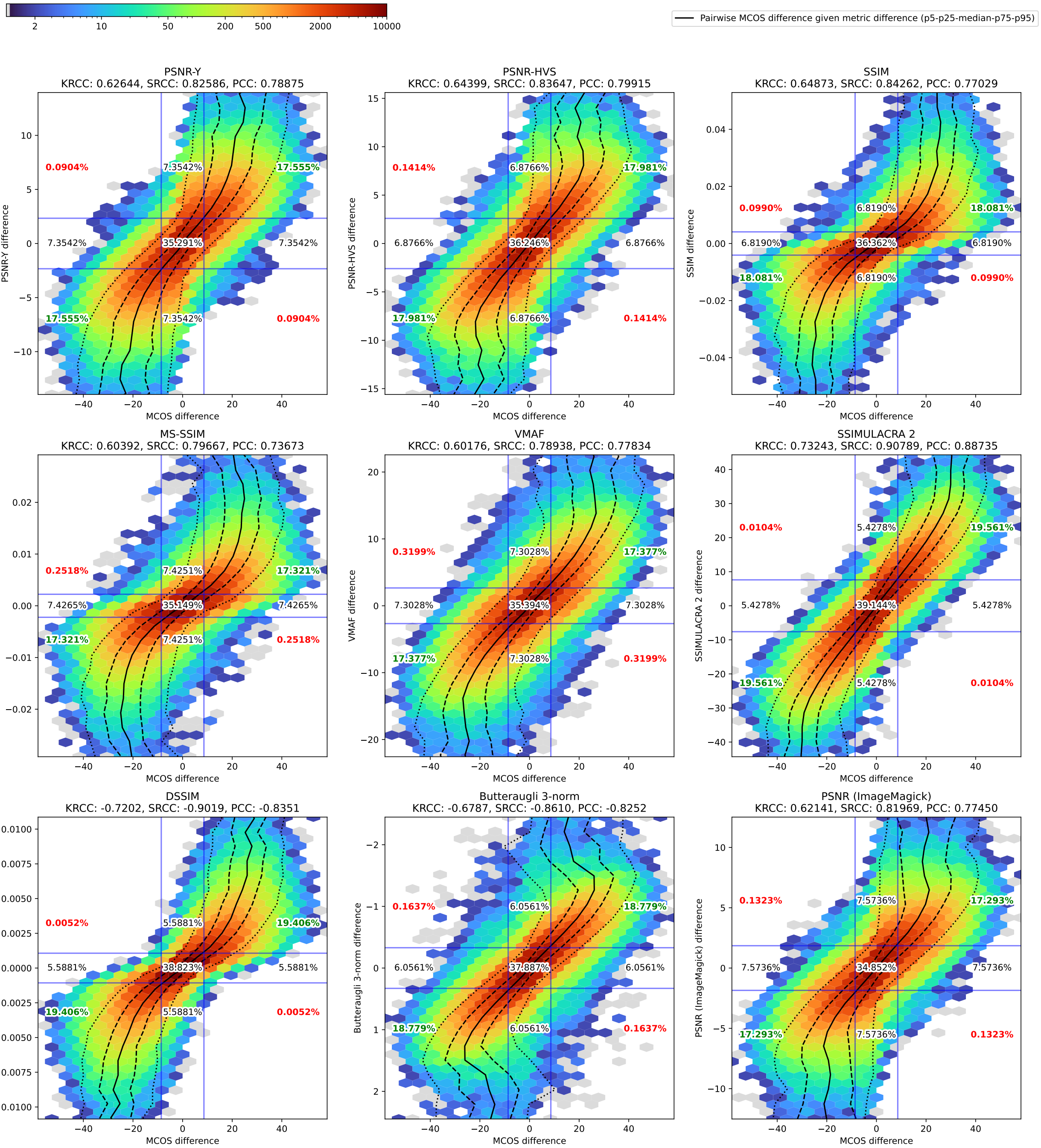


Figure 18. Correlation between objective metrics and pairwise comparisons.

Although no other IQA datasets were used in the weight tuning, SSIMULACRA 2 also correlates well with other datasets.

An open-source software implementation of the SSIMULACRA 2 metric is available[1]. It is also part of the benchmark tools available at the JPEG XL reference software repository.

Recommended application ranges

Correlation coefficients give a good indication of the overall performance of a metric, but the 2D histograms in Figures 13 to 17 provide more detailed information. For example, there can be an asymmetry between 'false positives' (the metric predicting a high quality while the subjective score is low) and 'false negatives' (the metric predicting a low quality while the subjective score is high). Butteraugli suffers from false negatives but avoids false positives, while MS-SSIM suffers from false positives and avoids false negatives. It may be application-dependent which of these asymmetries is preferred.

Metrics also perform differently in different regions of the quality spectrum. Some metrics are designed to be used specifically for high fidelity still image compression (e.g., Butteraugli) while others are designed for lower quality ranges and are mostly used to evaluate video compression (e.g. VMAF, PSNR-Y, SSIM). In order to improve our understanding of when to 'trust' which objective metrics, we can summarize the 2D histograms in a way that allows comparing the various metrics directly. The plots in **Figures 19 and 20** were created as follows. For each metric, we sorted all of the distorted images in the dataset by metric score, and then for each bucket of about 1000 images with a similar metric score, the median and standard deviation of the subjective scores is computed. These points are then connected linearly in order of increasing metric score.

When the resulting curves are nonmonotonic or otherwise erratic, this can indicate either poor correlation between the metric and the subjective scores, or noise in the dataset. Higher curves indicate a lower standard deviation, so better consistency of the metric in that quality range. **Table 7** indicates recommended quality ranges for each metric. This table can be used to select a suitable objective metric for a given application. For example, if the range of qualities to be covered is very broad, DSSIM is a good choice, while if the range is more narrow around visually lossless quality, Butteraugli is a good choice. For medium quality or better, SSIMULACRA 2 is a good choice. For very low to medium quality, MS-SSIM is a good choice.

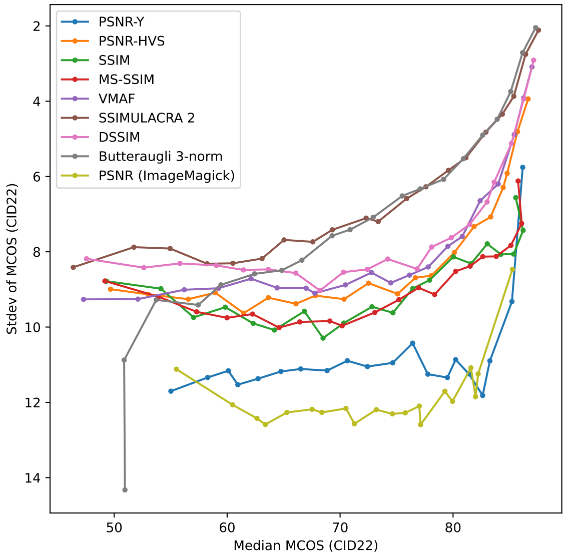


Figure 19. Metric consistency w.r.t CID22

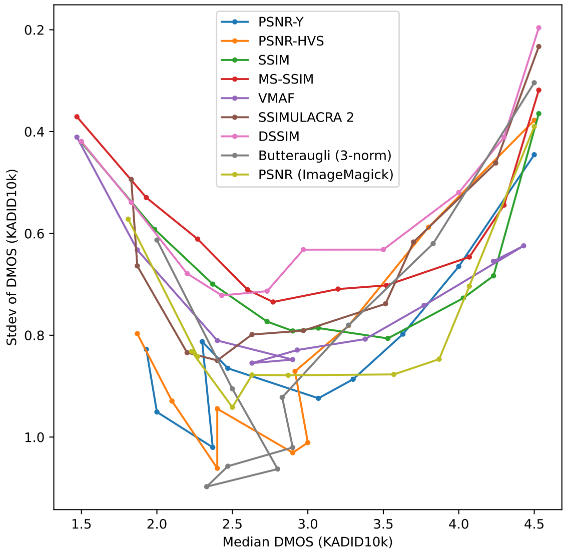


Figure 20. Metric consistency w.r.t KADID10k

[1] Repository: https://github.com/cloudinary/ssimulacra2

Table 7. Recommended quality ranges for various objective metrics.

	very low	low	medium	high	very high quality	visually lossless
CID22 MCOS		25	50	65	80	90
KADID10k DMOS	2	3	3.7	4.3	4.4	4.5
PSNR-Y	very poor	very poor	very poor	very poor	very poor	very poor
PSNR-HVS	poor	mediocre	good	mediocre	good	good
SSIM	good	mediocre	good	mediocre	poor	poor
MS-SSIM	very good	good	good	mediocre	poor	poor
VMAF	good	mediocre	good	good	mediocre	mediocre
SSIMULACRA 2	mediocre	good	very good	very good	very good	very good
DSSIM	good	very good	very good	good	good	good
Butteraugli 3-norm	very poor	poor	mediocre	good	very good	very good
PSNR (ImageMagick)	very poor	very poor	very poor	very poor	very poor	poor

CONCLUSION

We described a new subjective image quality assessment methodology based on a combination of two experiment protocols suitable for crowdsourcing: Triple Stimulus Boosted Pairwise Comparison (TSBPC) and Double Stimulus Boosted Quality Scale (DSBQS). We discussed our experiment setup, participant screening procedures, bias correction, and an analysis method to combine the scores obtained using both protocols. This led to the CID22 dataset, a large dataset of over 22,153 images, which was annotated in 2022 with accurate subjective quality scores based on 1.4 million human opinions. Compared to other datasets, it is both larger and more focused, covering specifically distortions caused by image compression in the range from medium quality to visually lossless. Using the CID22 dataset, we investigated the compression performance and visual consistency of different image encoders. We evaluated various objective metrics in terms of their correlation with the subjective scores, both in terms of absolute quality assessment (correlation with MCOS) and in terms of relative quality assessment (correlation with MCOS differences). We introduced a new metric (SSIMULACRA 2), and formulated recommendations on the application range of this and other metrics.

ACKNOWLEDGMENTS

We would like to thank Michela Testolina and Touradj Ebrahimi for their valuable feedback.

■ REFERENCES

1. ITU-R Rec. BT.500, “Methodologies for the subjective assessment of the quality of television images,” 2012.
2. ISO/IEC TR 29170-1:2017, “Information technology — advanced image coding and evaluation — part 1: Guidelines for image coding system evaluation.”
3. ISO/IEC 29170-2:2015, “Information technology — advanced image coding and evaluation — part 2: Evaluation procedure for nearly lossless coding.”
4. ISO/IEC JTC1/SC29/WG1 N100311, REQ, “Final call for contributions on subjective image quality assessment,” October 2022. [Online]. Available: https://jpeg.org/aic/documentation.html
5. H. Sheikh, M. Sabir, and A. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.
6. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
7. N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, “Image database TID2013: Peculiarities, results and perspectives,” *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596514001490
8. H. Lin, V. Hosu, and D. Saupe, “KADID-10k: A large-scale artificially distorted IQA database,” in *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.
9. E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, “Pieapp: Perceptual image-error assessment through pairwise preference,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
10. H. Lin, G. Chen, M. Jenadeleh, V. Hosu, U.-D. Reips, R. Hamzaoui, and D. Saupe, “Large-scale crowdsourced subjective assessment of picturewise just noticeable difference,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5859–

5873, 2022.

11. ISO/IEC JTC 1/SC29/WG1 N100163, REQ, "Review of the state of the art on subjective image quality assessment." [Online]. Available: https://jpeg.org/aic/documentation.html
12. H. Men, H. Lin, M. Jenadeleh, and D. Saupe, "Subjective image quality assessment with boosted triplet comparisons," 2021. [Online]. Available: https://arxiv.org/abs/2108.00201
13. M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," 2017. [Online]. Available: https://arxiv.org/abs/1712.03686
14. P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4249–4256.
15. B. Bos, T. Çelik, and I. H. H. W. Lie, "Cascading style sheets level 2 revision 1 (CSS2.1) specification," W3C, W3C Recommendation, Jun. 2008. [Online]. Available: https://www.w3.org/TR/CSS2/
16. ISO/IEC JTC 1/SC29/WG1 N100334, ICQ, "JPEG AIC common test conditions for subjective quality assessment." [Online]. Available: https://jpeg.org/aic/documentation.html
17. R. Rassool, "VMAF reproducibility: Validating a perceptual practical video quality metric," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2017, pp. 1–2.
18. Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402.
19. N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM 2007, Scottsdale, Arizona, USA, 25-26 January*, 2007.
20. Y. Yang, J. Ming, and N. Yu, "Color image quality assessment based on CIEDE2000," *Adv. MultiMedia*, vol. 2012, jan 2012. [Online]. Available: https://doi.org/10.1155/2012/273723
21. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
22. L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
23. M. U. Müller, N. Ekhtiari, R. M. Almeida, and C. Rieke, "Super-resolution of multispectral satellite images using convolutional neural networks," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-1-2020, pp. 33–40, 2020. [Online]. Available: https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/V-1-2020/33/2020/

Jon Sneyers is currently an image researcher at the Media Technology Research Group of Cloudinary in Petah Tikvah, Israel. His research interests include image processing, compression, and quality assessment. Sneyers received a Ph.D. degree from KU Leuven, Belgium. Contact him at jon@cloudinary.com.

Elad Ben Baruch is with the AI Research Group of Cloudinary in Petah Tikvah, Israel. Contact him at elad.benbaruch@cloudinary.com.

Yaron Vaxman is with the AI Research Group of Cloudinary in Petah Tikvah, Israel. Contact him at yaron@cloudinary.com.