CID22: Large-Scale Subjective Quality Assessment for High Fidelity Image Compression

Jon Sneyers, Elad Ben Baruch, and Yaron Vaxman, *Cloudinary, Petah Tikva, Israel*

Abstract—We propose a new methodology for large-scale subjective quality assessment of compressed still images in the high fidelity range. Combining two different assessment protocols, one based on pairwise comparisons, the other on absolute opinions, it is designed to assess this range of qualities not well-covered by previous methodologies. The methodology was applied to create the Cloudinary Image Dataset ’22 (CID22), consisting of 22,153 annotated images (with scores based on 1.4 million opinions), originating from 250 pristine images compressed using JPEG, JPEG 2000, JPEG XL, HEIC, WebP, and AVIF at high fidelity settings. Using this data, we evaluate various image encoders and objective metrics.

Better display and camera technology, increased storage capacity, broadband Internet, and advances in image coding caused high fidelity images to become increasingly feasible, desirable and widespread. However, quality assessment standards like ITU-R Rec. BT.500 [1] and AIC-1 [2] are not suitable for the range between high quality and visually lossless, as quality scores obtained in this way saturate at relatively low fidelity. AIC-2 [3] approaches the problem from the other end: based on a very sensitive flicker test catching even the slightest visual distortion, it leads to binary results: an image is either visually lossless or not. For somewhat lower fidelity targets, this test is not suitable. In this context, the JPEG Committee launched a new activity to create 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.

We propose a subjective IQA methodology based on a combination of two protocols, which enables large-scale IQA within a reasonable budget and time frame. We present the Cloudinary Image Dataset (CID22), a set of 22,153 quality-annotated images, originating from 250 pristine images. We describe the dataset construction and analysis to combine results of both experiment types. Next, we evaluate image encoders in terms of compression (bitrate-distortion) and visual consistency of encoder settings; we believe the latter has not been investigated before. Finally we evaluate correlation of objective metrics with subjective results, and propose a new metric called SSIMULACRA 2.

RELATED WORK

Compared to lab-based IQA datasets, CID22 is rather large: for example, the LIVE IQA database [4] has 29 pristine and 779 distorted images, and TID2013 [5] has 25 pristine and 3000 distorted images. Crowd-sourced IQA datasets like KADID10k [6] (81 pristine, 10,125 distorted) and the PieApp dataset [7] (200 pristine, 20,280 distorted) are larger. The main difference between CID22 and these existing datasets is the types and amplitudes of distortions: CID22 covers only image compression and a specific range of qualities (medium quality to near visually lossless), as opposed to the wider range of distortions and qualities contained in existing datasets. For example, in KADID10k and TID2013, only 2 out of 25 distortion types correspond to image compression (JPEG and JPEG 2000), and only 2 or 3 distortion levels are within the quality range that would be typically used for still images (the remaining distortion levels are too strong). Therefore CID22 is possibly less relevant than existing datasets for research into the human visual system and subjective quality perception in general, but more relevant for practical image compression applications.

The KonJND-1k database [8] contains many pristine images (1008), compressed with JPEG and BPG. It provides data on the picturewise just noticeable difference, i.e. the distortion threshold where an average observer notices artifacts. While relevant for practical image compression, it does not allow comparing different codecs (every pristine image was only compressed with one codec) and only provides information on one specific quality point, rather than a range of qualities.

IQA protocols

Various image quality assessment protocols described in AIC-1 [2] and AIC-2 [3] are reviewed in [9]. Single stimulus approaches like ACR and ACR-HR are suitable for assessing the appeal of a distorted image, but not fidelity, since test subjects cannot compare 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: participants do not know which stimulus is the ‘correct’ reference image, so it is possible that a distorted image gets a score ‘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 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 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 comparison (as opposed to side-by-side) makes it easier to spot differences. An extreme example is the AIC-2 flicker test. While useful to assess whether a codec achieves visually lossless compression, it is not discriminative in the range below visually lossless.

By amplifying the visibility of distortions, *boosted* triplet comparison [10] improves the discriminative power of pairwise comparisons (PC). By presenting three stimuli (reference and two distorted images), it assesses fidelity, unlike double stimulus PC protocols which effectively assess only appeal. One of our proposed protocols is a variant of boosted triplet comparison.

Pairwise comparisons between distorted images derived from the same reference image allow constructing a detailed ranking. The problem with this approach is that it only leads to relative scores, i.e. relative mean opinion scores (RMOS), where the lowest ranked image gets score 0 and the highest ranked image score 1. Such scores cannot meaningfully be compared across images originating from different reference images.

Impairment scale methodologies like DSIS [1] use an absolute scale, which can be compared across different reference images. They do however require collecting many opinions in order to obtain accurate mean opinion scores (MOS). Even then, confidence intervals tend to be too large to accurately rank distorted images, especially when the number of images to be tested is large or the range of qualities is relatively narrow.

ASSESSMENT PROTOCOLS

We propose the following hybrid approach:

- pairwise comparisons (TSBPC) resulting in RMOS scores, covering all distorted images;
- absolute grading (DSBQS) resulting in MOS scores for a subset of the images (“anchors");
- remaining MOS scores are interpolated based on RMOS scores and anchor MOS scores.

Triple Stimulus Boosted Pair 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 participants know this is the reference. On the right side of the screen, one distorted image is displayed, and participants 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, participants can submit a ternary response: “*A* is best", “*B* is best", or “I can’t choose".

Boosting [10] is applied to obtain a maximally accurate ranking: 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 perform “manual flickering". However, images are not altered to exaggerate pixel-wise differences.

Double Stimulus Boosted Quality Scale

The DSBQS protocol is similar to DSIS [1] with one major difference: instead of displaying reference and distorted images side-by-side, only one image is shown, and participants can freely switch between the reference and distorted image. The interface marks which image (reference or distorted) is currently being displayed. There is no time limit and no limit on how often and how quickly images are switched. Images are displayed at 'dpr1' resolution: on normal-density screens, one image pixel corresponds to one display pixel (1:1); on high-density ('retina') screens, one image pixel corresponds to 2x2 display pixels (2:1). In other words, we use CSS pixels [11], which theoretically span a visual angle of 0.0213 degrees, though in practice this is only an approximation. After switching at least twice, participants submit a response on a semi-continuous scale from 0 to 10, described 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.

Responses are registered by adjusting a slider which is initially in the middle (5) and which can be moved in increments of 0.5.

While 'manual flickering' is allowed, images are displayed without additional upscaling (only adjusted for display density), in order to make conditions more consistent across participants and to limit visibility of artifacts to a relevant level. Compared to the the five-grade DSIS impairment scale, 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. In this sense, both the viewing conditions and quality scale are 'boosted'.

EXPERIMENT SETUP

The goal was to create a large dataset of quality-annotated images covering various types of image content. Distortions of interest are compression artifacts, focusing on encoders relevant to web delivery and a production environment.

Reference Images

All images are 512×512 pixels. Most are cropped and downscaled high-resolution photos sourced from stock photography service Pexels. Images are clustered into 15 categories: 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), and urban-industrial-cars (18 images).

Distorted Images

The following codecs and encoders were used:

- 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 encoder, 8 to 11 quality settings were used, densily sampling the high fidelity range. For example, for mozjpeg, we used `-quality` parameter values 30, 40, 50, 60, 65, 70, 75, 80, 85, 90, 95. We used fixed encoder settings (as opposed to fixed bit rates) to match typical usage patterns and to assess encoder consistency.

Modern encoders can typically be configured to reach different trade-offs between speed and compression. We mostly used default configurations. Approximate encode speed in megapixels per second is indicated in the list above (single-threaded, Intel Core i7-9750H). For the slower AVIF encoder configurations (aom s1, aurora slow) only partial data was collected.

Selection of Stimuli

For the TSBPC 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 and prior assumptions about codec performance. For example, a 0.5 bpp JPEG image versus a 1.5 bpp AVIF image 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. This filtering step helps to avoid collecting opinions expected to bring little information. From the remaining triplets, we randomly sampled 105,155 triplets. We aimed to collect 10 opinions per triplet.

For the DSBQS experiment, we used 10 distorted anchor 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, 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, exposing participants to examples of very low and very high quality before the actual test started.

In both experiments, test sessions consist of 30 questions plus 2 additional 'honeypot' questions, inserted randomly and used for verification. For TSBPC, 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 DSBQS, these were one near-lossless image (where a score below 5 led to disqualification) and one very poor image (where a score above 5 led to disqualification). Participants could engage in up to 4 sessions, with a 24-hour break between sessions to prevent fatigue. They were instructed to use a desktop or laptop. This was checked during recruitment.

PARTICIPANT SCREENING

The crowd-sourcing platform Subjectify was used to conduct the experiment in the first half of 2022. Excluding participants who failed the initial ‘honeypot’ screening, 1,071,300 TSBPC opinions were collected in 35,710 test sessions and 334,920 DSBQS opinions in 11,164 sessions.

Inevitably, in crowd-sourced experiments some participants provide poor responses. To reduce the noise introduced by such responses, additional screening was applied. In the DSBQS experiment, sessions were discarded when one or more of these conditions were true: 1) a reference image received a score below 5; 2) more than 20 percent of the responses of the session 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), despite the instruction to use a desktop or laptop. This extra screening reduced the average number of opinions per anchor image from 122 to 101.

Outlier detection

Outliers (participants answering randomly or carelessly) were detected in the TSBPC experiment based on average agreement with other participants on all triplets evaluated in a session. In total, 5257 sessions (14.7% of TSBPC sessions) were discarded.

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), or if the standard deviation 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.

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 the DSBQS experiment, 43 to 94 opinions remained per image (mean: 63.6).

Bias Correction

In DSBQS, every image is scored by different participants, each with their own interpretation of the quality scale. We applied bias correction, adjusting scores by shifting all scores of a session by an additive constant, chosen per session to reduce the mean normalized difference to zero. Adjusted scores are clamped to $[0, 10]$. For example, scores would be adjusted upwards for a ‘pessimistic’ participant who systematically rated images lower than the (tentative) MOS.

SCORE ANALYSIS

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

RMOS Computation

The 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 per participant. To compute relative mean opinion scores (RMOS), we used the Elo rating system, independently per reference image. All distorted images derived from a particular reference are treated as players in a tournament. Opinions of the form $A > B$ count as two wins of A against B; “I can’t choose” counts as one win for each. To stabilize the computation, we add 10% of a tie (0.1 win for each) between all pairs $A \neq B$. Converged Elo ratings are then computed, i.e. the limit of the Elo ratings as the number of games played goes to infinity. These ratings are normalized to $[0, 1]$ to obtain RMOS scores, so 0 corresponds to the image with the lowest Elo rating (typically q30 JPEG) and 1 to the highest rated image (typically q95 JPEG or JPEG XL).

Monotonicity constraint Besides actual pairwise opinions, additional information is taken into consideration in the Elo computation. While in principle (due to bugs or strange phenomena) encoders can behave non-monotonically, we assumed that all tested encoders do in fact behave monotonically. A compressed image with a larger file size (higher quality setting) is assumed to be at least as good as an image with a smaller file size encoded with the exact same encoder (at the same speed setting). Without this monotonicity constraint, e.g. a q40 JPEG can get a lower score than a q30 JPEG due to incomplete sampling. We add dummy opinions to enforce monotonicity.

MCOS disagreement mitigation Finally, for pairs of anchor images, additional dummy opinions are added. If the 90% MCOS confidence intervals of both images do not overlap, then the image with the higher MCOS score is considered to be better a number of times proportional to the gap between confidence intervals. If confidence intervals overlap, dummy $A = B$ opinions are added proportional to the amount of overlap and $A > B$ opinions proportional to the amount of non-overlap.

Interpolating and extrapolating MCOS

MCOS scores of anchor images are then used to linearly interpolate MCOS scores for the other images using RMOS scores. There is one caveat: there are still (rare) cases where RMOS scores and anchor MCOS scores disagree on the order of a pair. If MCOS(A) $>$ MCOS(B) while RMOS(A) $<$ 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 is a high-bitrate AVIF anchor with slightly lower MCOS score than a lower-bitrate AVIF anchor.

At the extremes, we extrapolate as follows. The maximum RMOS score 1 is assumed to correspond to the MCOS score of the reference image. While the reference image was not compared in 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 close already to the average reference MCOS (88.3). Several encoder settings (e.g. q95 JPEG) achieve better RMOS scores than this, so it can be expected that the image with the highest RMOS score is visually lossless. So arguably, no actual extrapolation is done at this end. In case a distorted anchor obtained a higher MCOS score than the corresponding reference, both scores are adjusted as described above, moving scores towards the 20th and 80th percentile, respectively. There were 33 such cases of which 28 were a q90 JPEG with a higher MCOS than the reference.

About 5% of the anchor scores were adjusted in this way to resolve remaining rank-order disagreements and to ensure that no distorted image scores higher than the reference. The amplitude of changes was small: the largest difference is 2.59 MCOS points, the average absolute change amongst adjusted scores was 0.72 MCOS points (and 95% of the anchor scores were not adjusted).

RMOS score 0 corresponds to an anchor in 97% of the cases. For 8 images, we extrapolated by arbitrarily assuming the lowest RMOS score to correspond to 0.75 times the mean plus 0.25 times the 20th percentile opinion for the worst anchor image, assigning extrapolated scores at most 4 MCOS points below the worst anchor score.

Final MCOS scores

The bulk (91.7%) of the images in the CID22 dataset have MCOS $\geq$ 50, i.e. “medium quality” or better. Figure 1 shows the distribution of scores. Most images range from medium-high quality (MCOS 60) to visually lossless (MCOS around 88). All tested encoders are represented well acrosss this range.

Preservation of TSBPC preferences Table 1 shows the effect of the monotonicity constraint and the DSBQS disagreement mitigation on the agreement between raw TSBPC comparison results and the scores. We define ΔPC as the difference between the number of $A > B$ opinions and $B > A$ opinions; the higher this number, the clearer the consensus. For example, if 7 participants said $A > B$ and 3 participants said $A < B$, then ΔPC is $7 - 3 = 4$. Scores agree with TSBPC if the preferred image has a higher score.

Converting TSBPC results to RMOS, even when using 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 do not correspond to a preorder. Mitigations inevitably further reduce agreement with raw TSBPC results. Still, even with both mitigations, MCOS scores arguably agree well with TSBPC, especially when consensus is clear.

Confidence Intervals Bootstrapping was applied: 200 iterations of resampling-with-replacement were done on both sets of opinions (TSBPC and DSBQS, after participant screening and bias correction), recalculating MCOS scores, Elo rankings and MCOS interpolation in every iteration. The mean width of the 90% confidence intervals is 4.457.

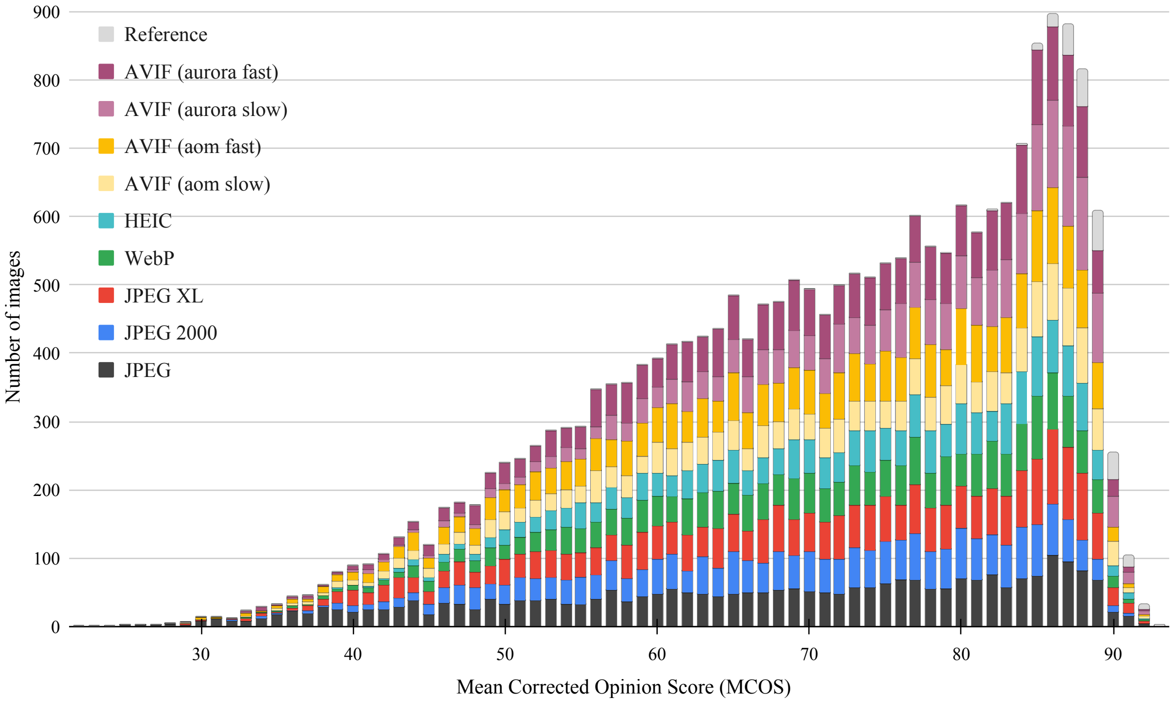

FIGURE 1. Distribution of MCOS scores in the CID22 dataset, by encoder.

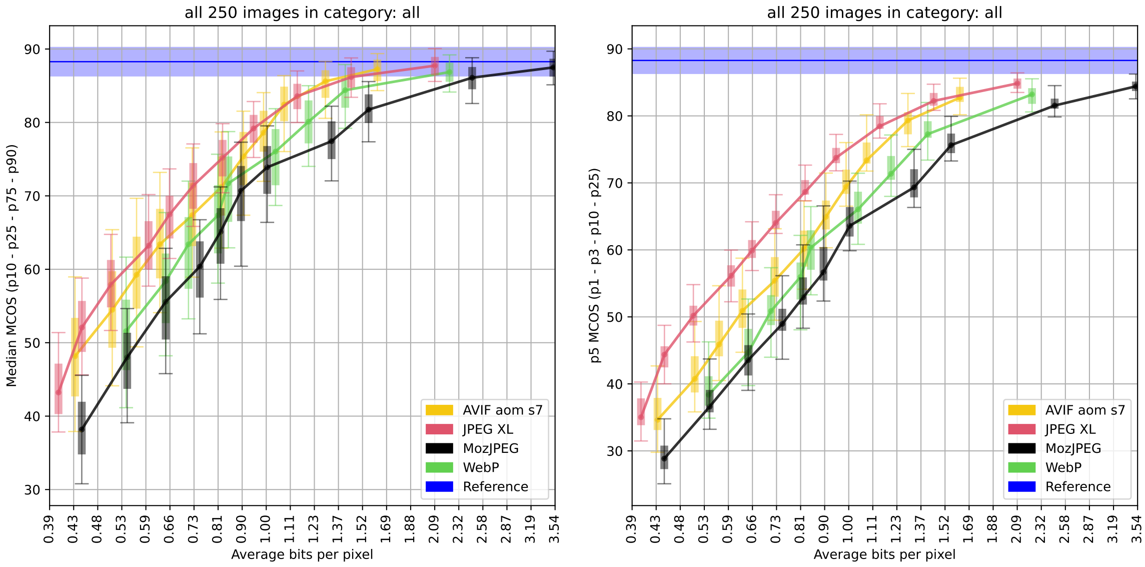

FIGURE 2. Median and 5th percentile (worst-case) performance of selected encoders.

TABLE 1. Agreement between scores and TSBPC.

	mitigations applied			mean	
ΔPC	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

ENCODER RESULTS

Figure 2 shows the performance of four encoders relevant for web delivery. To aggregate results over multiple images, we consider average bpp and median MCOS score per encoder setting. This aggregation hides image-dependent variation in the quality obtained using a given encoder setting, as seen in the box plots indicating spread. Obviously bpp is also image-dependent; averages do however indicate total compressed corpus size.

Encoder settings are often chosen using a “set it and forget it” approach: a fixed setting is used for many images. What matters is not the median result, but that (almost) all images reach a minimum fidelity. In other words, worst-case performance is what matters. For this reason, Figure 2 also shows 5th percentile scores.

Encoder consistency

Consistency is a desirable encoder feature as it reduces the likelihood of ‘surprising’ results — in particular, compressed images with noticeably worse quality than most other images encoded with the same setting. To investigate encoder consistency, we compute the standard deviation of MCOS scores per encoder setting (see Figure 3).

For all encoders, high quality settings produce consistent results; as MCOS scores approach the highest possible value (visually lossless), variance naturally diminishes. At lower quality settings, consistency decreases for all encoders, but there are differences: JPEG XL is more consistent than AVIF and WebP.

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

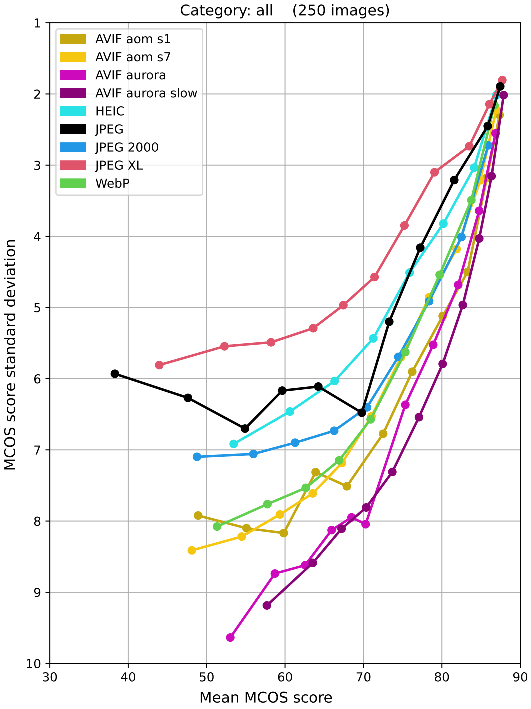

FIGURE 3. Visual consistency of encoder settings, as indicated by standard deviation of MCOS scores. Note: the vertical axis is flipped, so higher is better (more consistent, lower standard deviation).

Results by image category

Figure 4 shows MCOS plots aggregated per category. There are notable differences between categories: e.g. in the non-photographic categories (diagram-chart and illustration-logo-text), AVIF outperforms other codecs, while for landscape-nature and materials-clothes, it does not perform well.

Within each category, relative performance of the various encoders is generally similar, though there is still image-dependent variation. By means of example, Figure 5 shows per-image results for the portrait category. In these plots, anchors are marked with stars.

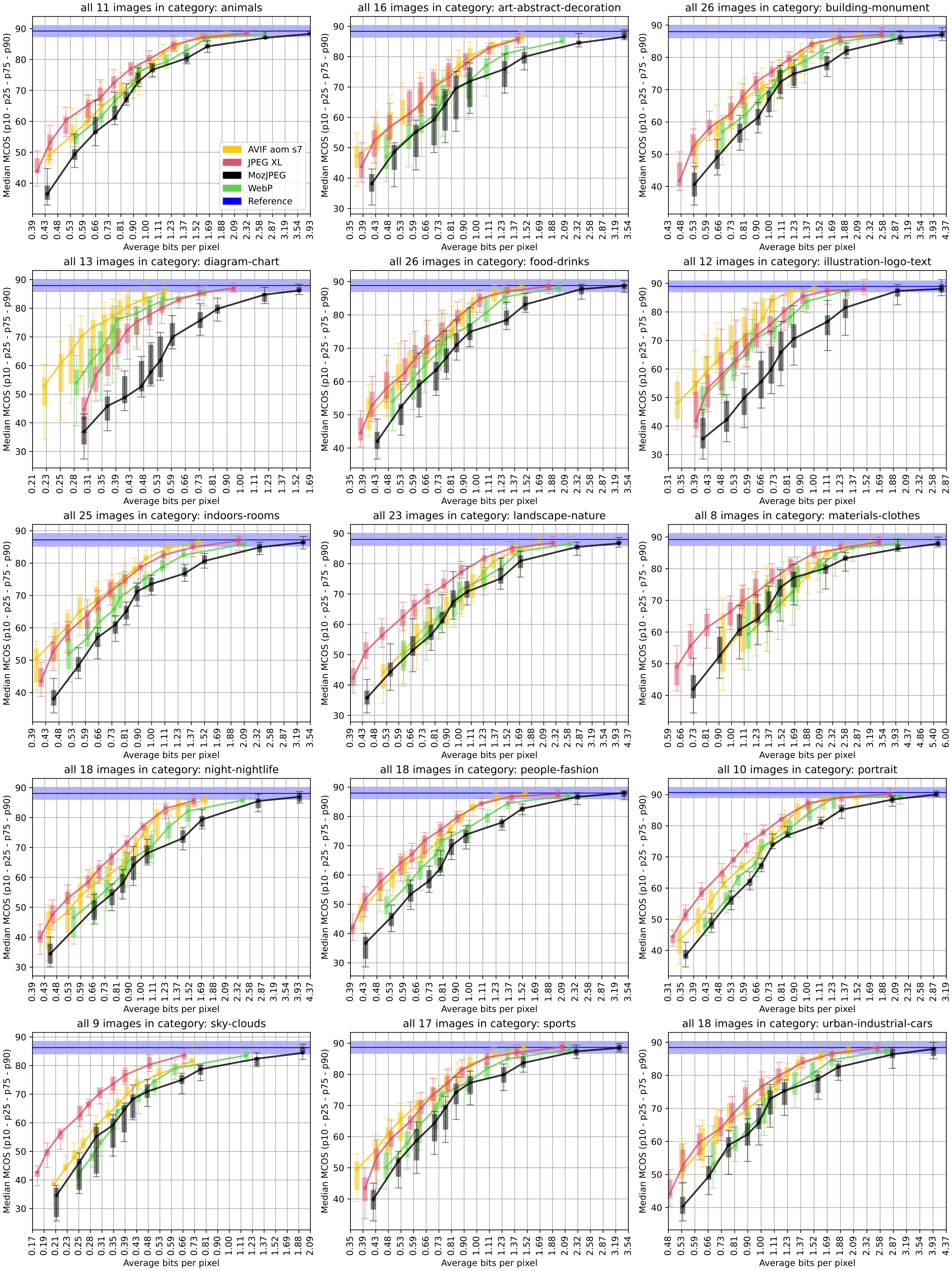

FIGURE 4. Median performance of selected encoders, per category.

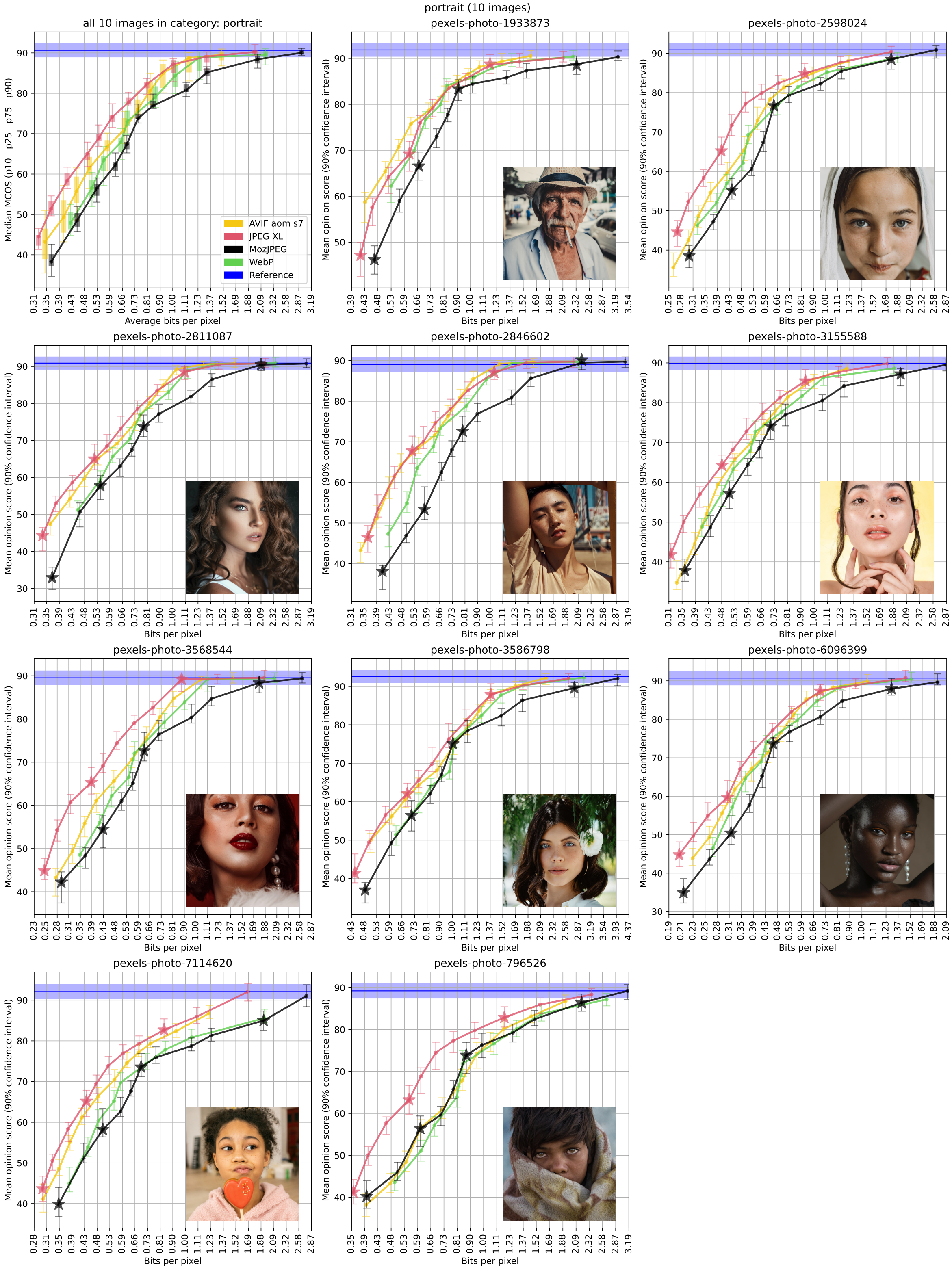

FIGURE 5. Per-image performance of selected encoders, for the specific category of portrait photos.

TABLE 2. Objective metric correlation with CID22 MCOS.

	correlation with MCOS (absolute quality)			corr. with MCOS differences (relative quality)		
Metric	KRCC	SRCC	PCC	KRCC	SRCC	PCC
(SSIMULACRA 2) [a]	0.6934	0.882	0.8601	0.7536	0.9210	0.9085
Butteraugli 2-norm [a]	**-0.6575**	**-0.8455**	**-0.8089**	-0.6852	-0.8688	**-0.8422**
Butteraugli 3-norm [a]	-0.6547	-0.8387	-0.7903	-0.6787	-0.8610	-0.8252
DSSIM (v3.2)	-0.6428	-0.8399	-0.7813	**-0.7203**	**-0.9019**	-0.8352
VMAF [12] [b]	0.6176	0.8163	0.7799	0.6018	0.7894	0.7784
FSIM (v0.3.5) [13]	0.6089	0.8005	0.7676	0.6828	0.8656	0.8411
PSNR-HVS [b]	0.6076	0.8100	0.7559	0.6440	0.8365	0.7992
Butteraugli max-norm [a]	-0.5843	-0.7738	-0.7074	-0.5877	-0.7773	-0.7351
SSIM [4] [b]	0.5628	0.7577	0.7005	0.6487	0.8426	0.7703
MS-SSIM [14] [b]	0.5596	0.7551	0.7035	0.6039	0.7967	0.7367
LPIPS (v0.1.4) [15]	-0.5417	-0.7316	-0.6932	-0.6711	-0.8612	-0.7901
SSIMULACRA 1 [a]	-0.5255	-0.7175	-0.6940	-0.7059	-0.8915	-0.8399
PSNR-Y [b]	0.4452	0.6246	0.5901	0.6264	0.8259	0.7888
PSNR (ImageMagick 6.9.11)	0.3472	0.5002	0.4817	0.6214	0.8197	0.7745
CIEDE2000 [b]	0.3154	0.4584	0.4096	0.6576	0.8482	0.7690

[a] Butteraugli, SSIMULACRA 1 and 2: as implemented in libjxl v0.8
[b] VMAF, PSNR-HVS, SSIM, MS-SSIM, PSNR-Y, CIEDE2000: as implemented in libvmaf v2.3.0

OBJECTIVE METRICS

Objective metrics assess image quality algorithmically rather than involving human test subjects. They are useful to the extent that they correlate with subjective results. Table 2 lists Kendall and Spearman rank-order and Pearson correlation coefficients between MCOS scores (excluding references) and various metrics.

Alignment to other datasets

Figure 6 visualizes correlations for a selection of objective metrics using 2D histograms. Horizontal axes correspond to subjective scores, vertical axes to metric values, and color indicates the number of images. Purple lines indicate the mean metric score of PJND images in KonJND-1k [8]; the purple shaded region indicates one standard deviation around the mean. Black curves indicate the mean MCOS for a given metric score; dashed and dotted lines indicate 25th/75th and 5th/95th percentiles, respectively. Horizontal spread between these lines shows variation in subjective scores for a given metric score. For comparison, Figure 7 shows a similar visualization for the KADID10k [6] dataset. Quality scales from this and other datasets can be approximately aligned as indicated in Table 3.

TABLE 3. Approximate alignment of quality scales.

Dataset / metric	medium quality	high quality	visually lossless
CID22 (MCOS)	50	65	90
TID2013 (MOS)	4.5	5.5	6
KADID10k (DMOS)	3.7	4.3	4.5
KonJND-1k (PJND)		1	
KonFiG-IQA (F-JND)	1.5	0.7	0
AIC-3 (JND)	3	1.7	0
PSNR-HVS	35	40	50
MS-SSIM	0.98	0.992	0.998
VMAF	83	91	96
DSSIM	0.008	0.003	0.001
Butteraugli 3-norm	2.5	1.6	0.5
SSIMULACRA 2	50	65	90

Pairwise correlation

For some use cases, absolute scores are not needed and it suffices to compare images originating from the same reference. For example, potential encoder changes typically aim to improve quality while keeping bitrate constant. Table 2 also lists correlations between score differences $\mathit{MCOS}(A) - \mathit{MCOS}(B)$ and metric differences $\mathit{metric}(R, A) - \mathit{metric}(R, B)$ for triplets (R, A, B) of the TSBPC experiment.

Predicting pairwise comparisons is generally an easier task for an objective metric than predicting absolute quality consistently between images derived from different reference images. For most metrics, pairwise correlation is higher than absolute score correlation. A notable exception is VMAF, which is (slightly) better at absolute than at relative IQA. SSIMULACRA 1 performs rather poorly at absolute IQA but is one of the best metrics for relative IQA. Interestingly, PSNR outperforms MS-SSIM and VMAF at relative IQA. For absolute IQA however, PSNR performs very poorly.

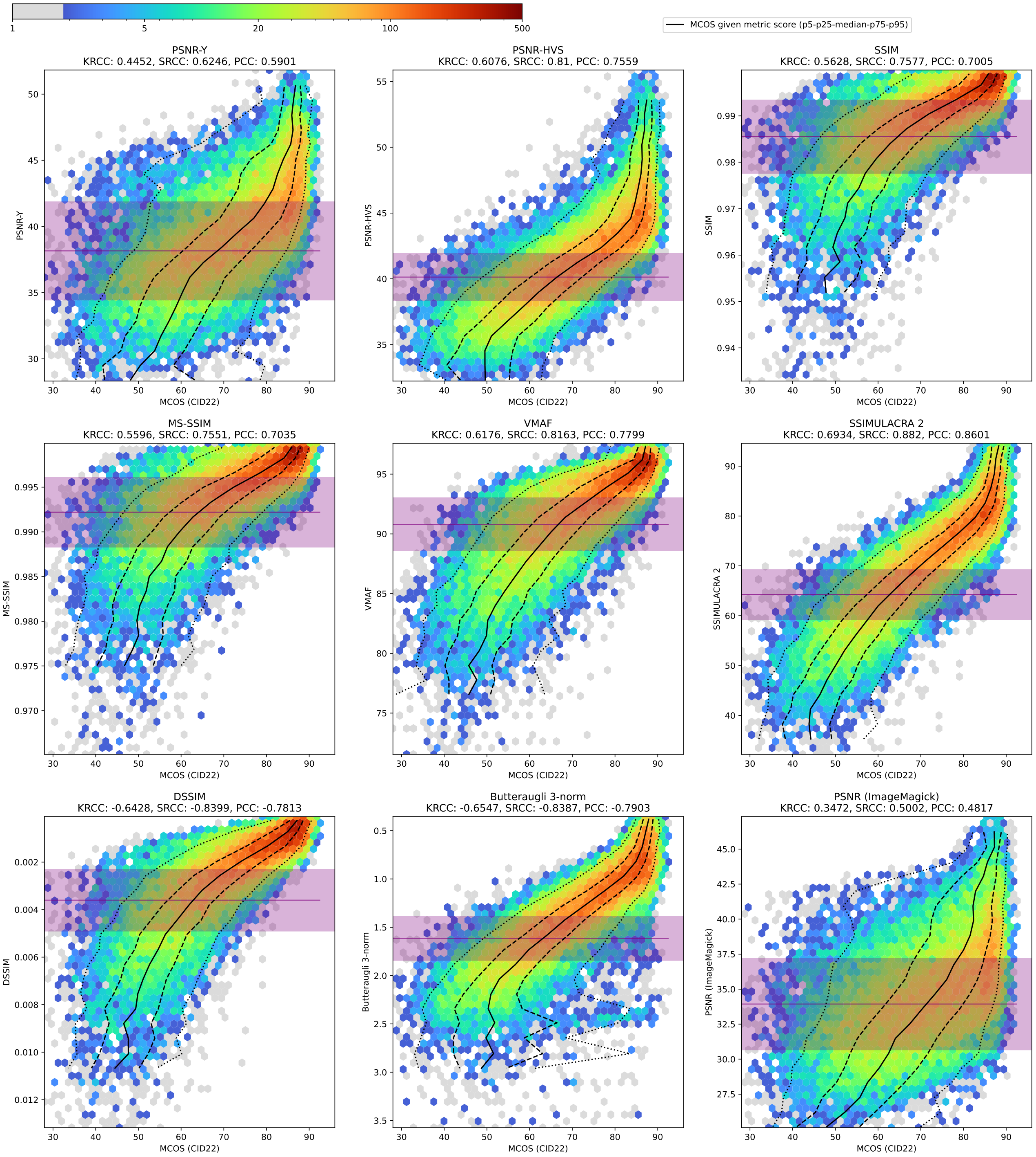

FIGURE 6. Correlation between objective metrics and the CID22 dataset.

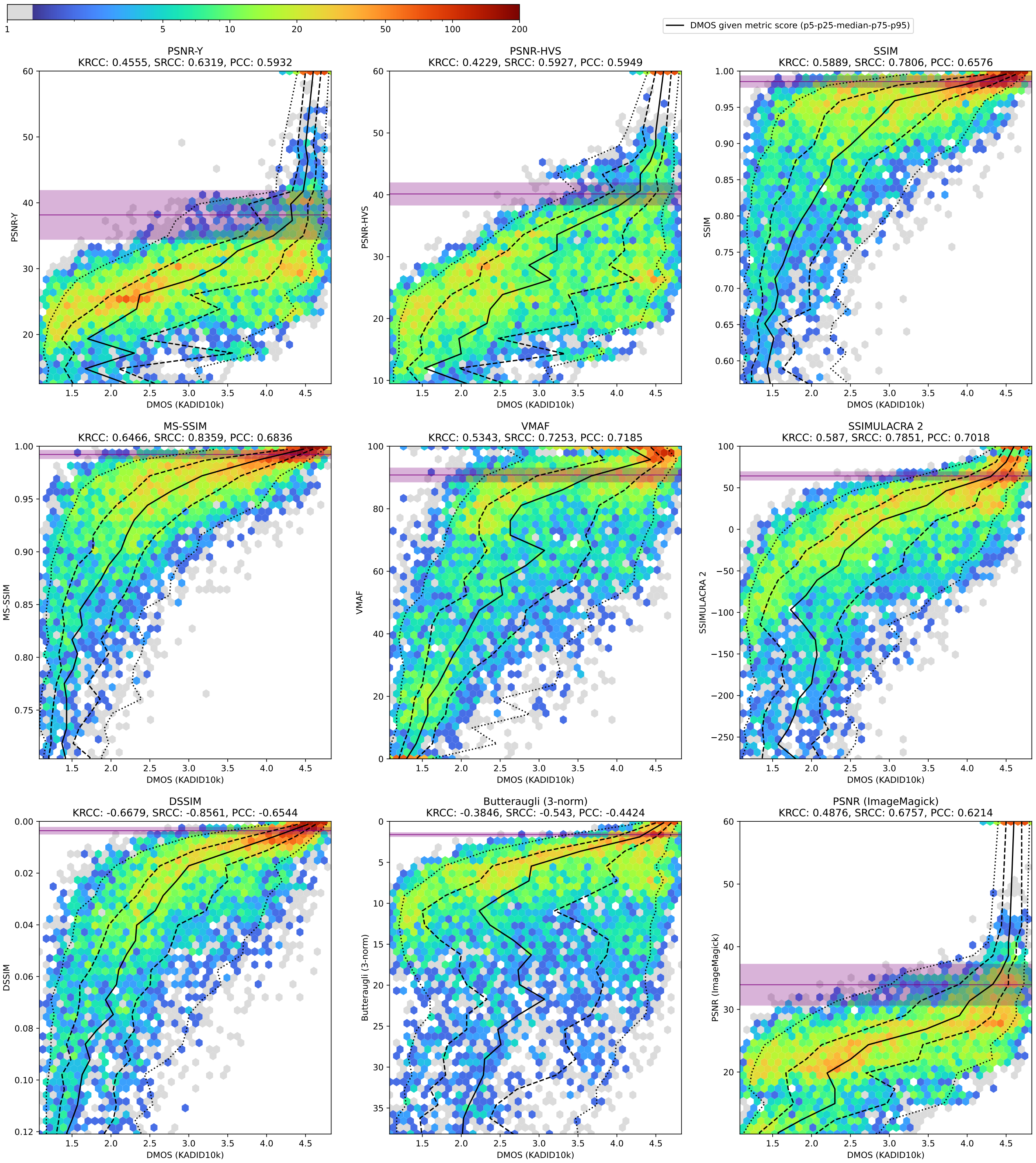

FIGURE 7. Correlation between objective metrics and the KADID10k dataset.

SSIMULACRA 2

Based on the CID22 dataset, we developed a new objective metric called SSIMULACRA 2, based on multiscale SSIM [14]. The computation is done in 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; they explicitly model ringing and smoothing artifacts. For each of the resulting $6\times3\times3$ error maps, L_1 and L_4 norms are computed. The final score is based on a weighted sum of the resulting 108 sub-scores. Weights were optimized to correlate with a subset of CID22 corresponding to 201 out of 250 references. For the remaining validation set (49 references), KRCC is 0.7033, SRCC is 0.8854, PCC is 0.8745 and mean absolute error is 4.97.

An open-source software implementation is available at github.com/cloudinary/ssimulacra2.

CONCLUSION

We described a new subjective image quality assessment methodology based on a combination of two experiment protocols suitable for crowd-sourcing: Triple Stimulus Boosted Pair Comparison (TSBPC) and Double Stimulus Boosted Quality Scale (DSBQS). We discussed our experiment setup, participant screening procedures, and a method to combine the scores obtained using both protocols. This led to the CID22 dataset of over 22,153 images. Compared to other datasets it is more focused, covering specifically distortions caused by image compression in the range from medium quality to visually lossless. Using this dataset, we investigated compression performance and visual consistency of different image encoders. We evaluated various objective metrics in terms of both absolute and relative quality assessment. Finally, we introduced the SSIMULACRA 2 metric.

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. 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.
5. N. Ponomarenko *et al.*, “Image database TID2013: Peculiarities, results and perspectives,” *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.
6. 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.
7. E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, “PieAPP: Perceptual image-error assessment through pairwise preference,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
8. H. Lin *et al.*, “Large-scale crowdsourced subjective assessment of picturewise just noticeable difference,” *IEEE Trans Circuits Syst Video Technol*, vol. 32, no. 9, pp. 5859–5873, 2022.
9. 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
10. H. Men, H. Lin, M. Jenadeleh, and D. Saupe, “Subjective image quality assessment with boosted triplet comparisons,” *IEEE Access*, vol. 9, 2021.
11. B. Bos *et al.*, “Cascading style sheets level 2 revision 1 (CSS2.1) specification,” W3C, Recommendation, 2016. [Online]. Available: https://www.w3.org/TR/CSS2/
12. 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.
13. 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.
14. 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.
15. R. Zhang *et al.*, “The unreasonable effectiveness of deep features as a perceptual metric,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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.