Universität Konstanz
Cloudinary
HUAWEI
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Subjective Visual Quality Assessment for High-Fidelity Learning-Based Image Compression

Mohsen Jenadeleh, Jon Sneyers, Panqi Jia, Shima Mohammadi, João Ascenso, Dietmar Saupe

VQEG
Video Quality Experts Group (VQEG)

05 May 2025

Motivation

- Deep learning-based image compression is emerging as a powerful alternative to traditional codecs, offering superior rate-distortion performance and perceptual quality.
- JPEG AI is a new compression standard that leverages neural networks for high-fidelity image reconstruction.
- However, the artifacts introduced by these methods differ from those in classical codecs and are often misjudged by traditional quality metrics such as PSNR or SSIM.
- Robust subjective quality assessment is essential to properly evaluate and benchmark these advanced compression techniques.

Objectives

- Extend the JPEG AIC-3 IQA dataset* by including JPEG AI compressed images alongside existing codecs: JPEG 1, JPEG 2000, AVIF, JPEG XL, and VVC.
- Evaluate the perceptual quality of JPEG AI through large-scale subjective testing.
- Benchmark the performance of objective image quality metrics on the updated dataset.
- Analyze statistical differences in metric performance to assess reliability.

* *Testolina, Michela, Mohsen Jenadeleh, Shima Mohammadi, Shaolin Su, Joao Ascenso, Touradj Ebrahimi, Jon Sneyers, and Dietmar Saupe. "Fine-grained subjective visual quality assessment for high-fidelity compressed images". The Data Compression Conference (DCC), 2025.*

Source Images and Crops from JPEG AIC-3 dataset

00002 00006 00007 00009 00010

Top: Full-resolution images. **Bottom:** Crops (620×800 px) used in the subjective test..

JPEG AI Encoding Setup

- Each source image was compressed using JPEG AI at 10 distortion levels.
- Bitrates ranged from 0.3 to 1.65 bpp in steps of 0.15 bpp.
- Range visually verified to match the distortion levels of other codecs.
- JPEG AI VM 7.0 used with three highest-rate models.

Command used:

```
python -m src.reco.scripts.eval --cfg ./cfg/tools_on.json
    ./cfg/oper_point/hop.json ./cfg/BRM/regen_list.json
    --coding_type enc_dec -target_bpps [BPP*100]
```

- Dataset: 50 JPEG AI images (5 sources $\times$ 10 bitrates).

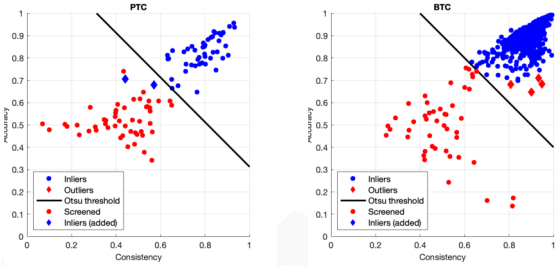
Subjective Evaluation Framework

- JPEG AIC-3 protocol* using plain (PTC) and boosted triplet (BTC) formats.
- Boosting includes: zoomin and artifact amplification (factor of 2), flicker (10 Hz).
- BTC is evaluated at all 10 bitrates; PTC at 5 bitrates (2, 4, 6, 8, 10).
- For the five source images, the BTC method included a total of 660 triplets, divided into 5 batches of 132 questions each.
- The PTC method consisted of 180 triplets, split into two batches of 90 questions each.
- Crowdsourcing experiment on MTurk using web interfaces of JPEG AIC -3.
- Per-triplet responses: 49 (PTC), 120 (BTC)
- 96,200 triplet comparisons from 459 participants were collected (Participants: 73 (PTC), 386 (BTC)).

**ISO/IEC DIS 29170-3, "Information technology — advanced image coding and evaluation — Part 3: Subjective quality assessment of high-fidelity images", 2025.*

Screening of Subject Responses

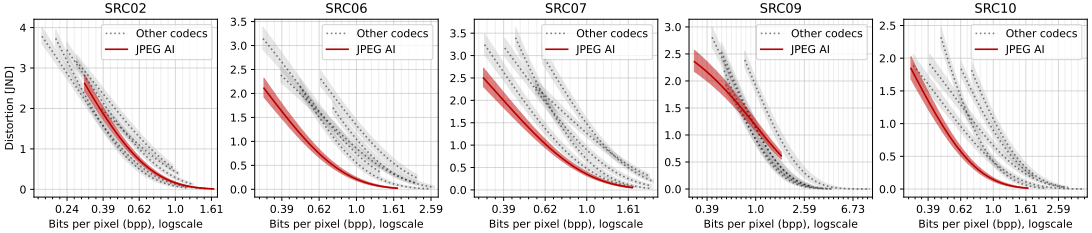
- Batch-level screening was performed for quality control using JPEG AIC-3 criteria.
- Responses were evaluated using two indicators:
  - Accuracy: Distortion-weighted proportion of correct responses.
  - Consistency: Agreement between mirrored triplet responses.
- Batches with an average score (accuracy + consistency) below a threshold were excluded.
  - Thresholds: 0.6563 (PTC), 0.6992 (BTC)
  - Excluded Batches: 51 of 98 (PTC), 46 of 600 (BTC)



Accuracy and consistency scores for observer batches in PTC (left) and BTC (right)..

Bitrate–Distortion Analysis

- A perceptual quality scale in JND units was derived using the JPEG AIC-3 reconstruction method.
- A unified model was fitted using maximum likelihood estimation (MLER), using both BTC and PTC triplet responses to estimate an exponential rate–distortion (RD) function.
- RD curves for JPEG AI are shown in red; other codecs (JPEG 1, JPEG 2000, AVIF, JPEG XL, VVC) are shown in gray.
- JPEG AI delivers good perceptual quality at low bitrates.



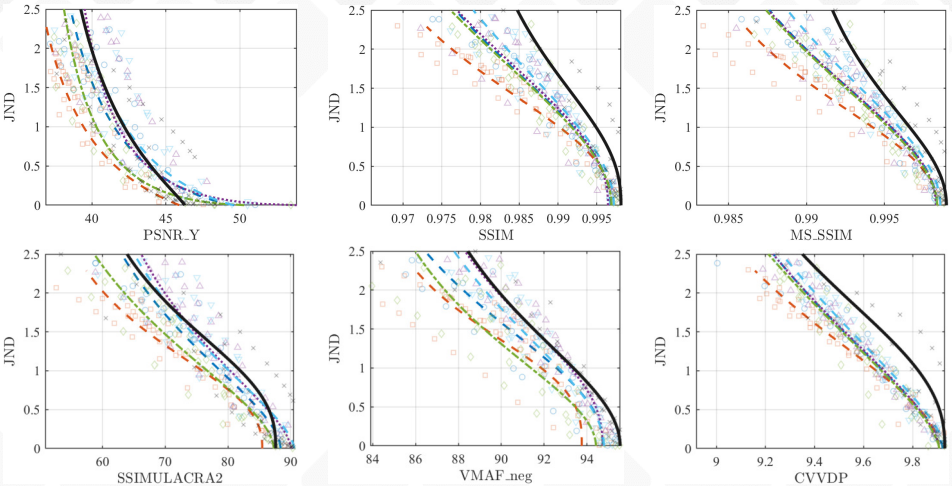
Distortion in JND units versus bitrate, with 95% confidence intervals..

Objective Metrics Correlation Results

Metric	Overall		Per-source		Per-codec		JPEG AI	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
PSNR-Y	0.807	0.816	0.883	0.885	0.851	0.832	0.849	0.815
SSIM	0.894	0.905	0.908	0.904	0.936	0.932	0.926	0.904
MS-SSIM	0.931	0.941	0.943	0.942	0.952	0.951	0.927	0.918
PSNR-HVS	0.867	0.877	0.941	0.948	0.882	0.870	0.846	0.813
IW-SSIM	0.950	0.951	0.947	0.947	**0.968**	**0.962**	0.919	0.889
NLPD	0.900	0.913	0.927	0.927	0.936	0.935	0.923	0.905
GMSD	0.894	0.903	0.953	0.957	0.903	0.892	0.901	0.874
Butteraugli	0.883	0.897	0.933	0.937	0.908	0.900	0.858	0.842
SSIMULACRA1	0.898	0.908	0.907	0.902	0.955	0.957	0.923	0.904
SSIMULACRA2	0.900	0.913	0.940	0.941	0.926	0.924	0.874	0.841
VMAF	0.882	0.891	0.903	0.900	0.912	0.925	0.855	0.899
VMAF-neg	0.909	0.921	0.940	0.937	0.932	0.936	**0.958**	**0.963**
HDR-VDP-2	0.919	0.929	0.944	0.943	0.929	0.919	0.865	0.814
HDR-VDP-3	0.919	0.933	0.948	0.950	0.937	0.939	0.931	0.953
CVVDP	**0.960**	**0.961**	**0.962**	**0.962**	0.963	0.958	0.880	0.843

Top 5 values are underlined; **Best overall** *is in bold.*

Objective Metric Evaluation (subset)



Selected IQA metrics vs. JND values. Logistic curve fits shown per codec.

Statistical Comparison of Metrics

- Meng–Rosenthal–Rubin (MRR) test* evaluates correlation differences.
  Null hypothesis: Equal correlation with subjective scores.

$$H_0 : r_{XZ} = r_{YZ}$$

  where r_{XZ} and r_{YZ} are correlations of metrics X and Y with the subjective scores Z.
- The test statistic is:

$$Z = (z_1 - z_2)\left(\frac{2(1 - r_{XY})h}{n - 3}\right)^{-0.5}$$

  where:

$$z_1 = \tanh^{-1}(r_{XZ}), \quad z_2 = \tanh^{-1}(r_{YZ})$$

$$h = \frac{1 - f \cdot \bar{r}^2}{1 - \bar{r}^2}, \quad f = \frac{1 - r_{XY}}{2(1 - \bar{r}^2)}, \quad \bar{r}^2 = \frac{r_{XZ}^2 + r_{YZ}^2}{2}$$

- We use a two-tailed test with significance level $\alpha = 0.05$ to determine if differences are statistically meaningful.

Xiao-Li Meng, Robert Rosenthal, and Donald B Rubin, "Comparing correlated correlation coefficients.," Psychological Bulletin, vol. 111, no. 1, pp. 172, 1992.

MRR Statistical Test Result

Metric	PSNR-Y	SSIM	MS-SSIM	PSNR-HVS	IW-SSIM	NLPD	GMSD	Butteraugli	SSIMU1	SSIMU2	VMAF	VMAF-neg	HDR-VDP2	HDR-VDP3	CVVDP
PSNR-Y	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
SSIM	+1	0	0	+1	-1	+1	+1	+1	+1	+1	+1	+1	+1	0	-1
MS-SSIM	+1	0	0	+1	-1	+1	+1	+1	+1	+1	+1	+1	+1	+1	-1
PSNR-HVS	+1	-1	-1	0	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1
IW-SSIM	+1	+1	+1	+1	0	+1	+1	+1	+1	+1	+1	+1	+1	+1	-1
NLPD	+1	-1	-1	+1	-1	0	0	+1	0	0	+1	0	-1	-1	-1
GMSD	+1	-1	-1	+1	-1	0	0	0	0	0	0	0	-1	-1	-1
Butteraugli	+1	-1	-1	+1	-1	-1	0	0	0	-1	0	-1	-1	-1	-1
SSIMULACRA1	+1	-1	-1	+1	-1	0	0	0	0	0	+1	0	-1	-1	-1
SSIMULACRA2	+1	-1	-1	+1	-1	0	0	+1	0	0	+1	0	-1	-1	-1
VMAF	+1	-1	-1	0	-1	-1	0	0	-1	-1	0	-1	-1	-1	-1
VMAF-neg	+1	-1	-1	+1	-1	0	0	+1	0	0	+1	0	0	-1	-1
HDR-VDP2	+1	-1	-1	+1	-1	+1	+1	+1	+1	+1	+1	0	0	0	-1
HDR-VDP3	+1	0	-1	+1	-1	+1	+1	+1	+1	+1	+1	+1	0	0	-1
CVVDP	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	0

MRR Test: +1 = row is significantly better; -1 = significantly worse; 0 = no significant difference.

Conclusion

- 15 full-reference IQA metrics evaluated.
- Most metrics overestimate quality for JPEG AI images.
- Best performers overall:

  CVVDP (highest overall correlation)
  VMAF-neg (best for JPEG AI specifically)
- JPEG AI delivers good perceptual quality at low bitrates.
- Subjective testing remains essential in high-fidelity ranges.
- Use statistical testing for comparing objective metrics performance.

Resources and Dataset

- **Dataset and code:**
  `https://github.com/jpeg-aic/dataset-JPEG-AI-SDR25`
- Includes:
  - Source and compressed images
  - Triplet responses (plain and boosted)
  - JND reconstructions and metric scores
  - Statistical analysis scripts

Scan to access dataset and code

Thank you!