ISO/IEC JTC 1/SC29/WG1 **N85013**
**85th JPEG Meeting, San Jose, USA, November 2019**

**ISO/IEC JTC 1/SC 29/WG 1**
**(ITU-T SG16)**

# Coding of Still Pictures

**JBIG**
Joint Bi-level Image Experts Group

**JPEG**
Joint Photographic Experts Group

| | |
|---|---|
| **TITLE:** | Performance Evaluation of Learning based Image Coding Solutions and Quality Metrics |
| **SOURCE:** | Requirements |
| **EDITORS:** | João Ascenso, Pinar Akayzi, Michela Testolina, Atanas Boev, Elena Alshina |
| **STATUS:** | Final |
| **REQUESTED ACTION:** | Distribute |
| **DISTRIBUTION:** | Public |

**Contact:**
ISO/IEC JTC 1/SC 29/WG 1 Convener – Prof. Touradj Ebrahimi
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
Tel: +41 21 693 2606, Fax: +41 21 693 7600, E-mail: Touradj.Ebrahimi@epfl.ch

# 1. Purpose of This Document

The JPEG AI database was constructed to (i) evaluate the performance of state-of-the-art learning-based image coding solutions and (ii) to be used for training, validation and testing of novel learning-based image coding solutions. To fulfill objective (i), subjective quality assessment experiments were conducted during the 84th JPEG meeting in Brussels, Belgium. This document outlines the framework in detail and reports the experimental results and analysis. Moreover, the correlation between the subjective assessment MOS scores and the objective metrics is reported with the aim to evaluate the performance of the metrics defined in the common test conditions document [1], for both classical and deep learning-based codecs.

# 2. Subjective Assessment of Deep Learning Based Codecs

In this Section, the subjective assessment experiments are described, namely the test images, coding conditions, subjective assessment methodology and the experimental results.

## 2.1. Test Material

Prior to the subjective tests, there are several important elements to select and define, such as the images. Eight contents from the test set of JPEG AI database were selected through expert viewing sessions, to be used for subjective quality assessment experiments. The selected contents are depicted in Figure 1.

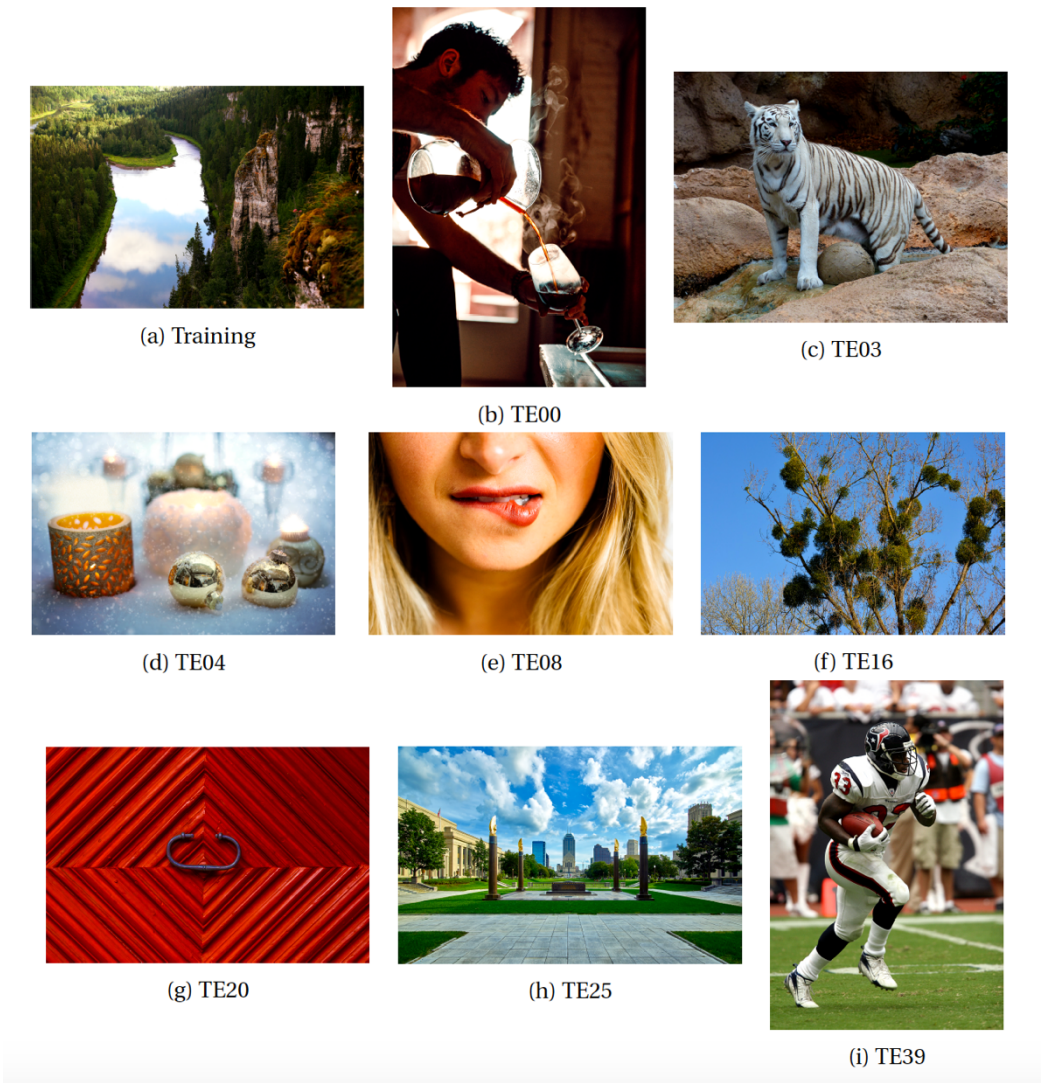



Figure 1. Thumbnails of JPEG AI contents selected for objective and subjective quality assessment.

## 2.2. Coding Solutions

Five learning-based image compression algorithms available online were selected for performance assessment against four anchors, i.e. HEVC, JPEG2000, WebP and JPEG. The list of learning-based image coding solutions is as follows:

- FRICwRNN[1] [2]: TensorFlow model for compressing and decompressing images using an already trained Residual Gated Recurrent Unit (GRU) model. The model is fully convolutional and recurrent and the input image size needs to be multiples of 32, therefore zero padding was applied when necessary.
- The following models using factorized entropy models (Fact-) or exploiting the dependencies within the latent representation through a scale hyperprior at the encoder (Hyper-), with Mean Squared Error (MSE) or Multi-scale Structural Similarity Index (MS-SSIM) loss as a distortion measure[2] [3]:
    - Factorized Entropy Model with Mean Squared Error Loss (FactMSE).
    - Factorized Entropy Model with Multi-scale Structural Similarity Index Loss (FactMS-SSIM).
    - EntropyModel with Scale Hyperprior using Mean Squared Error Loss (HyperMSE).
    - Entropy Model with Scale Hyperprior using Multi-scale Structural Similarity Index Loss (HyperMS-SSIM).

The software used for encoding anchors were JPEG XT reference software (v1.53) for JPEG, Kakadu (v7.10.2) for JPEG 2000, HM-16.18+SCM-8.7 for HEVC and cwebp 1.0.0 for WebP. Command lines used for encoding the anchors and learning-based solutions are given in Table 1.

Table 1. Selected parameters and settings for anchors and learning-based codecs.

| Codec | Input format | Command line |
|---|---|---|
| JPEG | YCbCr 4:4:4 8-bit | jpeg -qt 3 -h -v -c -q <qp> -s 1x1,1x1,1x1 <input> <output> |
| JPEG 2000 | YCbCr 4:4:4 8-bit | kdu_v_compress -i <input> -o <output> –rate <bpp> -precise -tolerance 0 |
| HEVC | YCbCr 4:4:4 8-bit | TAppEncoderStatic -c encoder_intra_main_scc.cfg -f 1 -fr 1 -q <qp> -wdt <width> -hgt <height> –InputChromaFormat=<chroma_format> –InternalBitDepth=<bit_depth> –InputBitDepth=<bit_depth> –OutputBitDepth=<bit_depth> –ConformanceWindowMode=1 -i <input> -b <output> -o /dev/null |
| WebP | YCbCr 4:2:0 8-bit | cwebp -m 6 -q <qp> -s <width> <height> <depth> <input> -o <output> |
| FRICwRNN | RGB 4:4:4 8-bit | python encoder.py –input_image=<input> –output_codes=<output> –iteration=<qp> –model=residual_gru.pb |
| FactMSE | RGB 4:4:4 8-bit | python tfci compress bmshj2018-factorized-mse-<qp> <input> <output> |
| FactMS-SSIM | RGB 4:4:4 8-bit | python tfci compress bmshj2018-factorized-msssim-<qp> <input> <output> |
| HyperMSE | RGB 4:4:4 8-bit | python tfci compress bmshj2018-hyperprior-mse-<qp> <input> <output> |
| HyperMS-SSIM | RGB 4:4:4 8-bit | python tfci compress bmshj2018-hyperprior-msssim-<qp> <input> <output> |

During the preparation of JPEG AI CTC, YCbCr 4:4:4 color space was preferred to avoid negative bias on anchor results. The learning-based codecs, on the other hand, all operated with RGB 4:4:4 inputs. The color space conversion for JPEG XT was handled inside the codec, so the files were not converted. Conversions from RGB 4:4:4 to YCbCr 4:4:4 for JPEG 2000 and HEVC-Intra and YCbCr 4:2:0 for WebP were conducted using FFmpeg 3.4.1 with the following command, with <pix_fmt> parameter set either to yuvj444p or yuvj420p:

ffmpeg -i <input> -s <width>x<height> -pix_fmt <pix_fmt> <output>

[1] https://github.com/tensorflow/models/tree/master/research/compression/image_encoder
[2] https://github.com/tensorflow/compression

## 2.3. Subjective Test Methodology and Subjective Scores Processing

The Double Stimulus Impairment Scale (DSIS) Variant I [4] was the test methodology selected for subjective quality assessment. The stimulus under assessment and the reference were presented simultaneously to the subject and the subject was then asked to rate the degree of annoyance of the visual distortions in the stimulus under assessment with respect to the reference. The degree of annoyance was divided into five different levels labeled as Very annoying, Annoying, Slightly annoying, Perceptible but not annoying and Imperceptible, corresponding to a quality scale ranging from 1 to 5, respectively.

The content and rate selection was carried out during expert viewing sessions prior to setting up the experiments. All contents in the test dataset depicted in Figure 1 were encoded using the anchor software at 8 rate points [0.06, 0.12, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00] bpp and the decoded images were viewed by experts. To obtain meaningful results from the experiments, the selected rate points needed to span a range that covers very low to high bitrates, corresponding to very low to transparent visual quality. The anchor with the best performance, i.e. HEVC, was used to select such rate points and the selection was verified using other anchors. Table 2 depicts the original resolutions of the test contents and the selected bitrates for subjective evaluation.

Table 2. Original resolutions and selected bitrates for subjective quality assessment of JPEG AI contents.

| Name | Resolution | Bitrates |
|---|---|---|
| TE00 | 1486 × 2230 | [0.06, 0.12, 0.25, 0.50] |
| TE03 | 4000 × 3000 | [0.06, 0.12, 0.25, 0.50] |
| TE04 | 5000 × 3332 | [0.06, 0.12, 0.25, 0.50] |
| TE08 | 3400 × 2266 | [0.06, 0.12, 1.00, 2.00] |
| TE16 | 1280 × 852 | [0.06, 0.12, 0.25, 0.50] |
| TE20 | 4000 × 2666 | [0.06, 0.12, 0.25, 0.50] |
| TE25 | 2200 × 1392 | [0.06, 0.12, 0.50, 0.75] |
| TE39 | 2336 × 3504 | [0.06, 0.12, 0.25, 0.50] |

Selected contents were processed according to the DSIS framework. A 30 inch Eizo 10bit ColorEdge CG301W monitor with a resolution of 4096 x 2160 was used. Stimuli were cropped using FFmpeg[3] to fit the screen resolutions. The region to be cropped for each stimulus was determined during expert viewing. Each decoded stimulus was placed side by side with its reference, with a 20 pixel mid-gray colored separation in between. The side-by-side stimuli were then displayed in front of the same mid-gray colored background, and were randomized such that the same content was never presented consecutively [4]. Two dummy sequences were included in each test, about which the subjects were not informed. A training session was conducted for each subject prior to the experiment, during which three stimuli were presented as examples for the two extremes of the voting scale, i.e. Very annoying (1) and Imperceptible (5), along with an example in the middle, i.e. Slightly annoying (3). For half of the subjects the reference was placed at the right side of the screen, whereas for the other half it was placed on the left to avoid position bias. Each experiment was conducted in two sessions to prevent subject fatigue. The monitors were calibrated using an i1 DisplayPro color calibration device according to the guidelines described in [4, 5]. Same guidelines were followed to set up the controlled environment for viewing with a mid gray level background inside the test rooms.

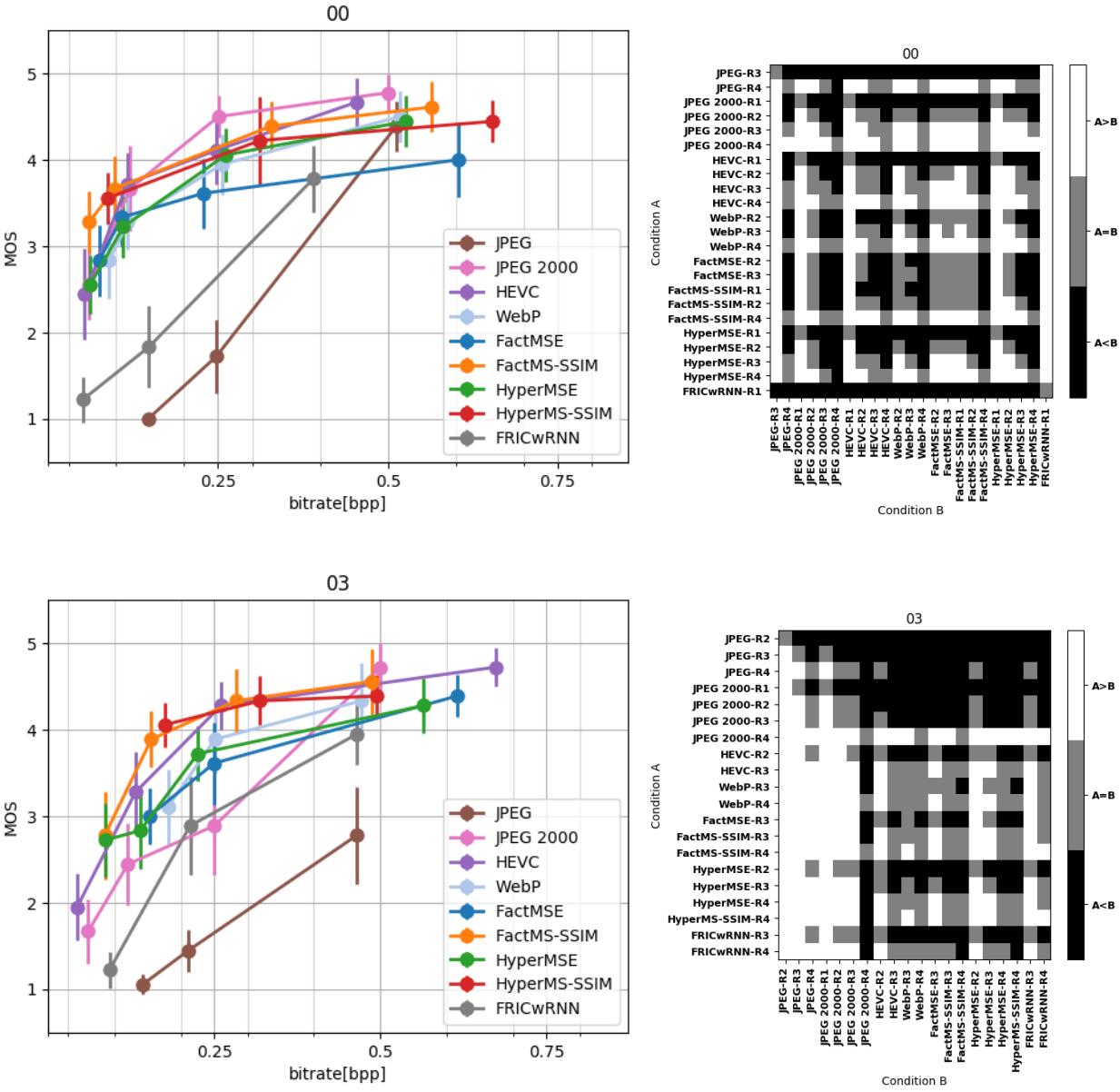
The experiments were conducted in Vrije Universiteit Brussel (VUB) with the participation of 18 volunteering subjects. Viewing time was not restricted during the experiments. Subjects, however, were instructed to vote within reasonable time for the experiments to proceed smoothly. No viewing distance or position was specified.

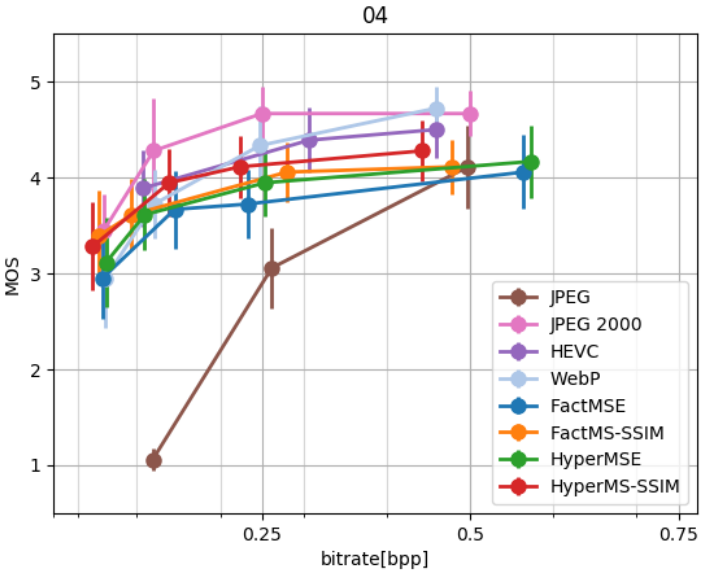
After the collection of results, a standard outlier detection was performed on all sets of raw scores to remove subjects whose ratings deviated strongly from others [6]. None of the subjects were identified as outliers in the experiments. The Mean Opinion Scores (MOS) and 95\% Confidence Intervals (CIs) assuming a Student's t-distribution of the scores were computed for each test condition [7]. To determine and compare the differences among MOS obtained for different codecs and bitrates, a one-sided Welch test at 5% significance level was performed on the scores. Bitrates that deviated more than 20% from the target rates were excluded from statistical significance tests.

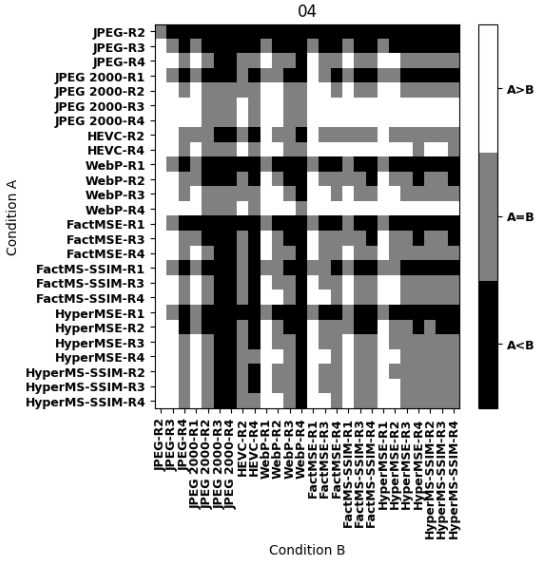
[3] http://ffmpeg.org

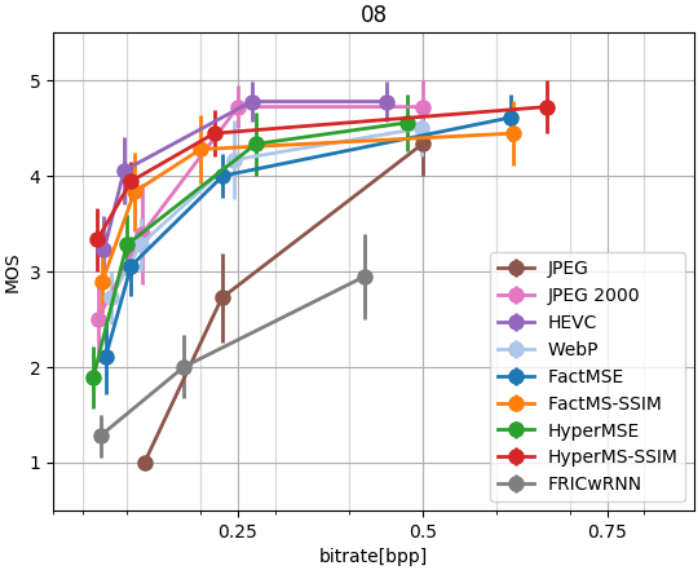
## 2.4. Experimental Results

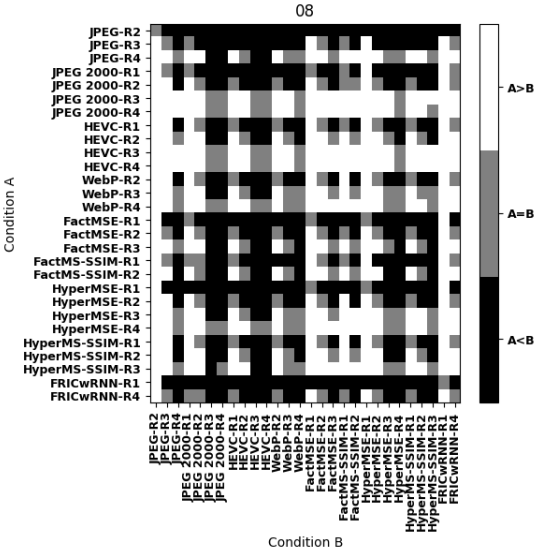
Subjective quality assessment was performed on the selected contents at the screened out bitrates given in Table 2, for all proponents and anchors. The MOS vs. bitrate plots and comparisons between pairwise conditions are presented in Figure 2.

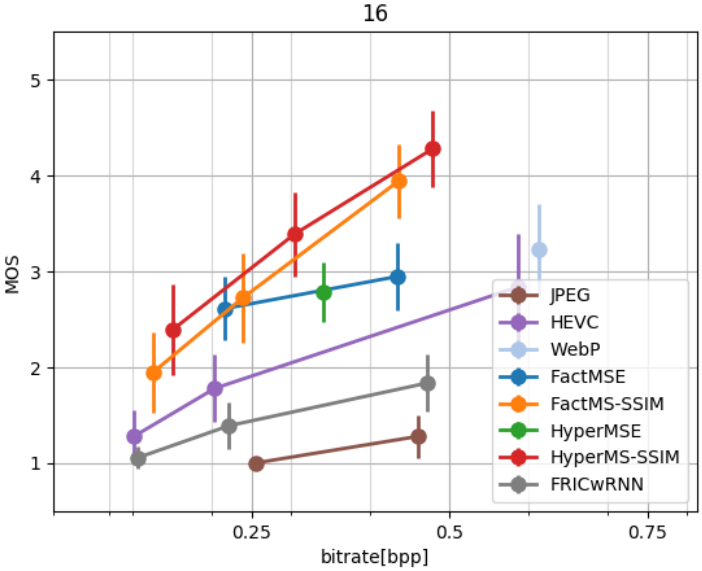


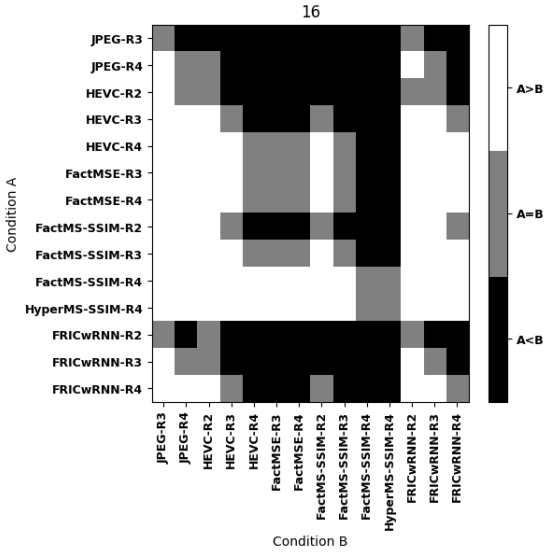

04
MOS
bitrate[bpp]
JPEG
JPEG 2000
HEVC
WebP
FactMSE
FactMS-SSIM
HyperMSE
HyperMS-SSIM


04
Condition A
Condition B
A>B
A=B
A<B


08
MOS
bitrate[bpp]
JPEG
JPEG 2000
HEVC
WebP
FactMSE
FactMS-SSIM
HyperMSE
HyperMS-SSIM
FRICwRNN


08
Condition A
Condition B
A>B
A=B
A<B


16
MOS
bitrate[bpp]
JPEG
HEVC
WebP
FactMSE
FactMS-SSIM
HyperMSE
HyperMS-SSIM
FRICwRNN


16
Condition A
Condition B
A>B
A=B
A<B

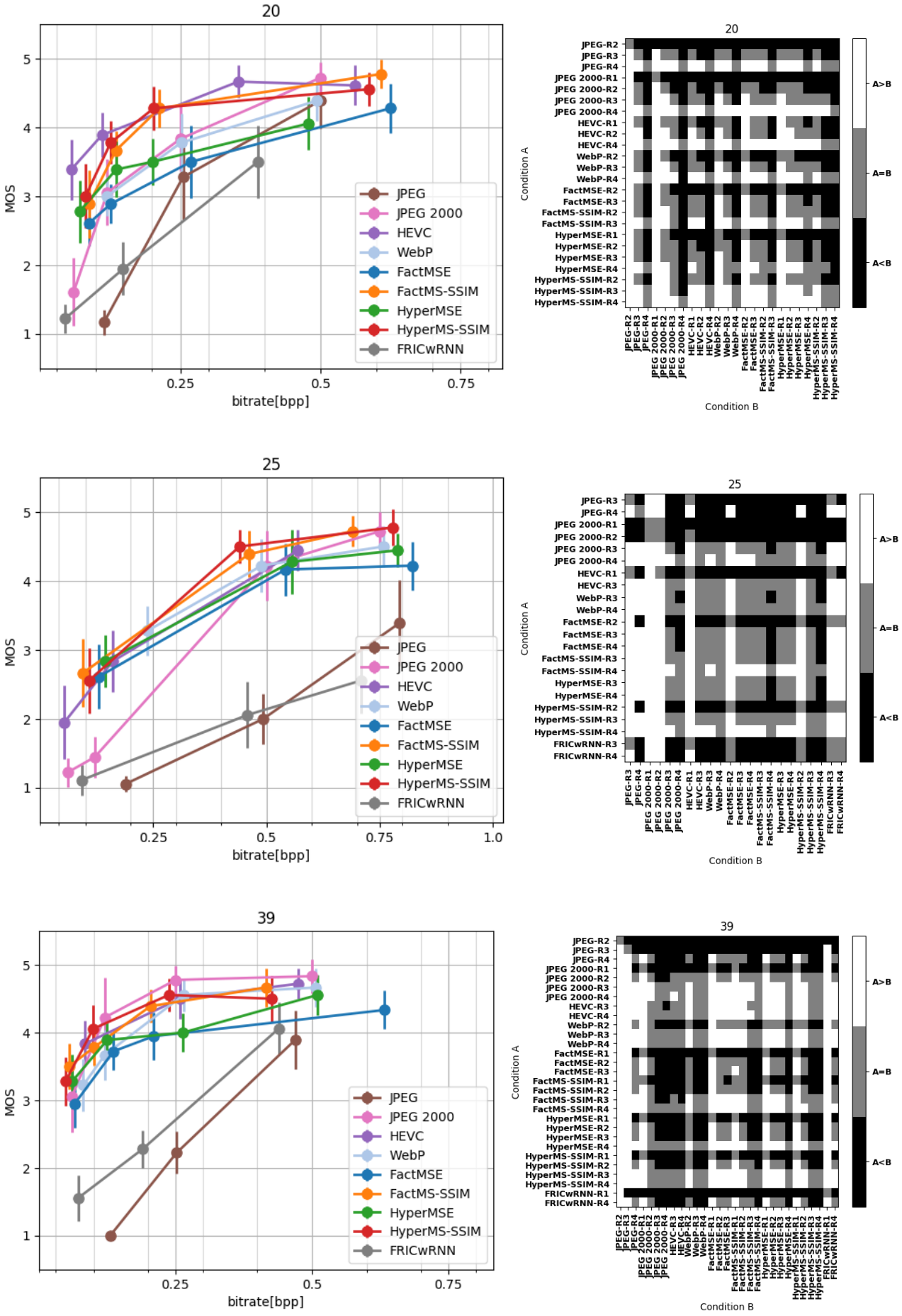



Figure 2. Subjective results for contents TE00, TE03 and TE04, TE08, TE16, TE20, TE25, TE39.

Throughout all contents, the performance of JPEG is mostly inferior to all other codecs at the lowest three bitrates. A very similar trend is observed when the performance of FRICwRNN is compared to other codecs except JPEG. The interpolated MOS curves suggest that the subjective ratings of FRICwRNN were higher than JPEG for contents 00, 03 and 39. Comparisons between pairwise conditions presented on the right columns in Figure 2 indicate statistically significant differences at comparable rate points. FRICwRNN was superior to JPEG for two contents at R3 and R4 namely, TE03 and TE16.

The comparison between the learning-based codec performances and anchors based on the results of Figure 2 is performed next. An initial observation is that FactMSE usually performs inferior to the remaining codecs (i.e. excluding JPEG and FRICwRNN), with exceptions at target rate points R2 and R3 when compared to JPEG 2000, and at target rate point R3 when compared to HEVC. More specifically, FactMSE was rated statistically significantly higher than JPEG 2000 at R2 for content TE25 and at R3 for content TE03. FactMSE was performing statistically significantly better than HEVC at R3 only for content TE16. The results on content TE16 are particularly interesting, indicating an exceptionally low performance on anchors and leading performances of learning-based codecs optimized using MS-SSIM metric. A closer examination on the visuals is provided in Figure 3. The abrupt patterns generated by FRICwRNN at R2 are evident, followed by the clear blocking artifacts of HEVC. HyperMS-SSIM is able to preserve the details better than the other codecs in comparison, with slightly better performance than that of FactMS-SSIM, yet without any statistically significant advantage at comparable rates. It is worth noting that FactMS-SSIM and HyperMS-SSIM are the only codecs for content TE16 that are able to reach high qualities.


Figure 3. Section of test content TE16 compressed using HEVC, FRICwRNN, FactMS-SSIM and HyperMS-SSIM from top to bottom.

(a) Reference (b) JPEG 2000, R1

(c) JPEG, R2 (d) FRICwRNN, R1

(e) FactMSE, R1 (f) FactMS-SSIM, R1

(g) HyperMSE, R1 (h) HyperMS-SSIM, R1

Figure 4. Test content TE08 compressed using JPEG at target bitrate 0.12bpp, FRICwRNN, FactMSE, FactMS-SSIM, HyperMSE and HyperMS-SSIM at target bitrate 0.06bpp.

Contents TE04 and TE39 exhibit similar plots, with all codecs except FRICwRNN and JPEG attaining ratings of minimum 3. TE04 is an image from still-life category while TE39 includes a person, yet both images have focused objects in the foreground and an out-of-focus background.

TE08 displays a more versatile range of MOS values. All codecs except FRICwRNN reach transparent quality at R4 for TE08, yet codecs FactMSE and HyperMSE have much lower ratings at low bitrates compared to their counterparts optimized using MS-SSIM. The apparent artifacts in TE08 at the lower bitrates are depicted in Figure 4.

FRICwRNN possesses artifacts similar to blocks, yet they have a regular pattern that is visually more pleasant compared to JPEG for many images and bitrates. An interesting pattern that is generally observed at low bitrates with codecs optimized using MS-SSIM is the contrast change. The RD optimization favors preserving structural information at the expense of less fidelity to color components. MSE optimization, on the other hand, is more inclined to introduce blur and introduces only local contrast changes in the form of emphasized colors. In contents like TE20 that exhibit clear structural patterns, MS-SSIM optimized learning-based codecs are performing better than other solutions and are on-par with HEVC that is superior to other anchors due to its Intra directional modes. In urban contents like TE25, however, with several elements of different characteristics scattered all around the image, it is more difficult for both learning-based and transform-based codecs to not introduce perceptible artifacts during the encoding process.

# 3. Objective Assessment of Deep Learning Based Codecs

Objective quality assessment was carried out at the bitrates selected for subjective quality evaluation for all codecs, in RGB color space.

## 3.1. Objective Metrics

Selected metrics for objective quality assessment were Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), MS-SSIM, Visual Information Fidelity (VIF) and Video Multimethod Assessment Fusion (VMAF). All metrics were computed using FFmpeg with command lines as provided in Table 3.

Table 3. Command lines for objective metric computations for JPEG AI experiments.

| Metric | Command line |
|---|---|
| PSNR | ffmpeg -s:v <width>x<height> -i <decoded> -s:v <width>x<height> -i <reference> -lavfi psnr=stats_file=<log_file> -f null - |
| SSIM, MS-SSIM VIF, VMAF | ffmpeg -s:v <width>x<height> -i <decoded> -s:v <width>x<height> -i <reference> -lavfi libvmaf=ssim=true:ms_ssim=true:log_fmt=json:log_path=<log_file> -f null - |

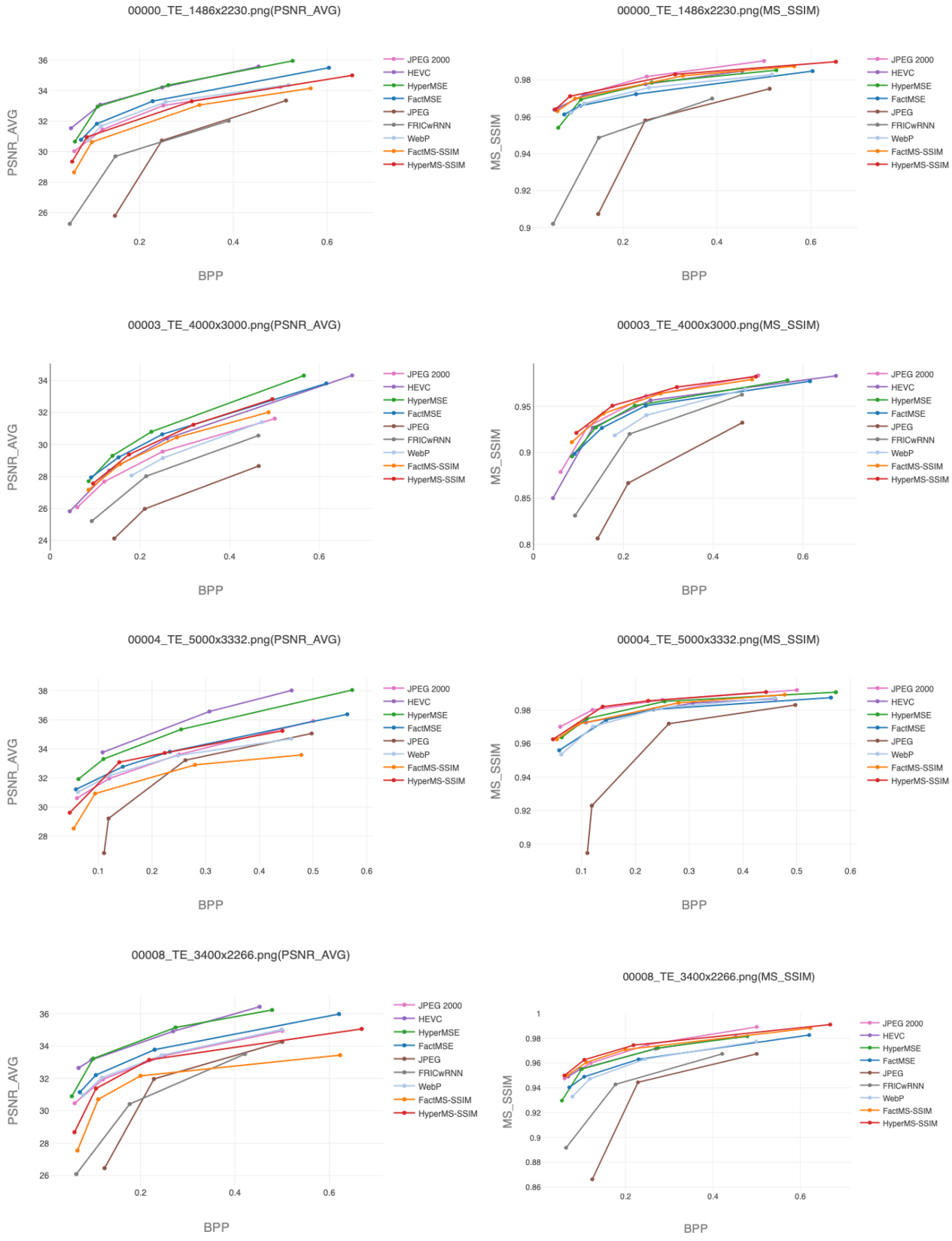
Objective quality assessment was performed on all 8 contents, at selected bitrates depicted in Table 2, for all proponents and anchors, and interactive plots were generated shown in the next Section.

## 3.2. Experimental Results

The results for all codecs and images are shown in Figure 5 for the PNSR and MS-SSIM objective quality metrics (Annex A includes for SSIM, VIF and VMAF). For content TE00, the leading codecs according to the PSNR metric are HEVC and HyperMSE, followed by FactMSE. The leading codecs according to the MS-SSIM metric are JPEG2000, HyperMS-SSIM and FactMSSIM.

For content TE03, the leading codec according to the PSNR metric is HyperMSE, followed by FactMSE, HEVC and HyperMS-SSIM performing on par. However, the leading codec according to the MS-SSIM is HyperMS-SSIM closely followed by FactMS-SSIM and JPEG 2000. Similar behavior is observed in the rest of objective results, with MS-SSIM-optimized codecs performing better in terms of MS-SSIM and MSE-optimized codecs performing well for PSNR metrics. Moreover, the performance of all codecs except FRICwRNN and JPEG are very close at all bitrates, which indicates that learning-based codecs are indeed able to reach the performance of their transform-based counterparts for several objective metrics.

A comparison between the objective and subjective results indicate that both assessments follow similar trends. Just as a codec optimized for MSE metric yields to higher PSNR, the codec which is perceptual optimized according to HVS (MS-SSIM based) yields to higher MOS; naturally, the rate-MOS results do not always match with each tested metric. Despite these differences, both subjective and objective metrics prove that many deep learning-based codecs, such as HyperMS-SSIM, are performing as well as the state-of-the-art anchors.



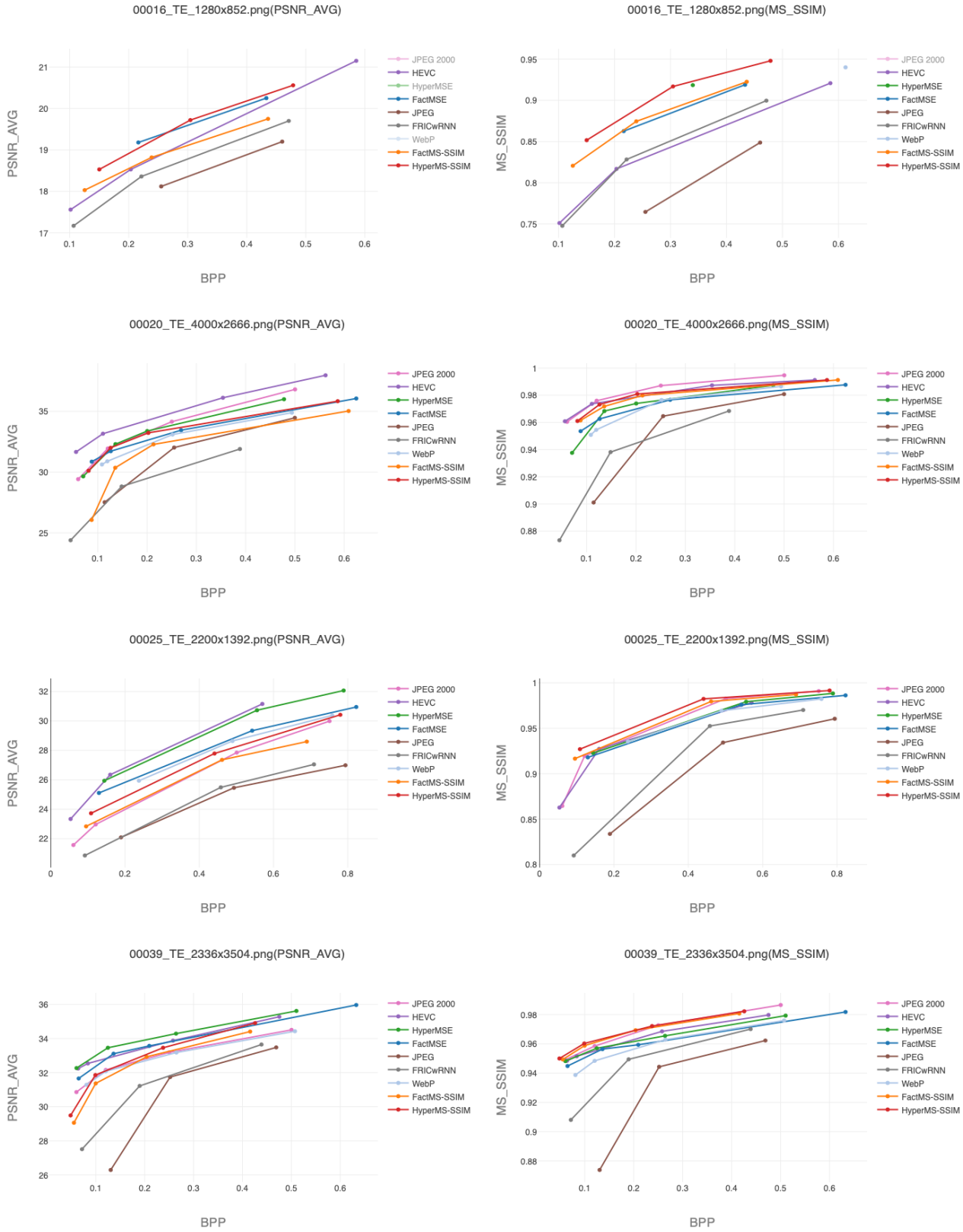



Figure 5. RD performance evaluation using PSNR and MS-SSIM metrics for contents TE0, TE3 and TE4, TE8, TE16, TE20, TE25, TE39.

# 4. Objective-Subjective Correlation Study

This section describes the experiments regarding the correlation between the objective quality metrics and the MOS scores given by the users. The main objectives of this section are: 1) summarise the objective quality metrics and 2) describe the way that the objective-subjective study was performed, present and analyze the results obtained.

## 4.1. Objective Quality Metrics

First, the objective full-reference quality metrics evaluated are enumerated and summarized. A few details about the implementation are also included, e.g. in which color space the metrics has been tested.

Since all input images are in RGB 8 bit, for single-channel metrics (SSIM, MS-SSIM, VIF(P), FSIM and NLPD), the RGB score is computed as the average over the 3 channel scores separately. Regarding metrics computed over the Y component, the conversion is done using the function rgb2YCbCr of MATLAB. In particular the Y component is computed as Y = 16 + (65.481*R + 128.553*G + 24.966*B) with R,G, B ∈[0,1]. For the PSNR, the conversion is done using the built-in PIL library convert("YCbCr"). In this case, the Y component is computed as Y = R * 76.245 + G * 149.685 + B * 29.07 with R,G, B ∈[0,1].

### 4.1.1. PSNR

The PSNR is a well-known quality metric in terms of the ratio between the maximum image value and the mean square error. Given and image I and the compressed I', the PSNR is computed as:

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) = 20\log_{10}\left(\frac{MAX}{\sqrt{MSE}}\right)$$

where the mean square error MSE is computed as:

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(I(i,j) - I'(i,j)\right)^2$$

M and N are respectively the width and the height of the image. The PSNR has been computed over the RGB images, over the Y channel in the YCbCr colour space and over the weighted average in the YCbCr color space as:

$$PSNR_{YCbCr} = \frac{6\times PSNR_Y + PSNR_{Cb} + PSNR_{Cr}}{8}$$

The metric has been tested using a self-implemented python script. A high metric score expresses high image quality.

### 4.1.2. SSIM

The Structural Similarity Index (SSIM) [8] express the difference between two images by analyzing of the structural information. The tested reference code as well as the paper can be found at the following link: https://www.cns.nyu.edu/~lcv/ssim/ . Since the SSIM metric has been designed for monochromatic images, it is computed over the R, G and B channel separately, over the grayscale-converted input image (converted using the function rgb2gray) and on the Y component in the YCbCr colour space (converted using the function rgb2ycbcr). It is important to notice that the SSIM grayscale and the SSIM Y are different, because they are generated from a different conversion tables). The SSIM over the RGB image as the average of the R,G and B scores. The original source code is available in MATLAB. The metric takes values in range 0-1 and a high score express better quality.

### 4.1.3. MS-SSIM

The Multi-Scale Structural Similarity Index (MS-SSIM) [9] computes the SSIM over different resolutions to represents the quality at different resolutions and viewing conditions. The source code of this metric can be seen at this link: https://ece.uwaterloo.ca/~z70wang/research/iwssim/ . As for the SSIM metric, the MS-SSIM metric was designed for single-channel images, so I compute is over the R, G and B channels separately and over the grayscale-converted image (converted using the function rgb2gray, note that ). The SSIM over the whole RGB image will be the average of the R,G and B scores. The original source code is available in MATLAB. The metric takes values in range 0-1 and a high score express better image quality.

### 4.1.4. VIF(P)

The Visual Information Fidelity (VIF or VIFP) [10] express the quality of the image using Natural Image Statistics exploiting also the Human Visual System characteristics. The source code can be downloaded from the author's website at the following link: https://live.ece.utexas.edu/research/Quality/VIF.htm . The original code works with single-channel images, for this reason I tested the code on the R, G and B channel separately and on the grayscale image (converted using the function rgb2gray). The final RGB score is computed as the average over the 3 channels. The metric takes values in range 0-1 and a high score express better image quality.

### 4.1.5. VDP2

The HDR-VDP-2 [11], also known as VDP2, claims to be robust to different luminance conditions, performing better on low-light images. This metric predicts the visibility of differences between original and reference images for an average observer but also the quality degradation with the respect to the reference image, expressed as a mean-opinion-score.

More information about this metric can be found on https://www.cs.ubc.ca/labs/imager/tr/2011/Mantiuk_HDR-VDP-2/, while the source code can be downloaded from https://sourceforge.net/projects/hdrvdp/files/hdrvdp/2.2.1/. This metric was evaluated for the RGB color space. The original source code is available in MATLAB. The metric has range 0-100 and a high score express better image quality.

### 4.1.6. FSIM

The Feature-Similarity Index Metric (FSIM) [12] uses the Phase Congruency (PC) and Gradient Magnitude (GM) to assess image local quality. More information about this metric as well as the source code are available at: https://www4.comp.polyu.edu.hk/~cslzhang/IQA/FSIM/FSIM.htm . Since this metric allows as input only monochromatic images, the score has been computed over the R, G and B channel separately as well as on the grayscale version of the original image (converted using the function rgb2gray). The FSIM for the RGB image is computed as the average of the one on the R, G and B results. The original source code is available in MATLAB. A high metric value express better image quality.

### 4.1.7. NLPD

In the Normalized Laplacian Pyramid (NLPD) [13] the quality is computed using the Laplacian Pyramid. More details about the metric as well as the source code can be found here: https://www.cns.nyu.edu/~lcv/NLPyr/ . This metric is computed on every channel of the image separately as well as on the grayscale version; the RGB score is computed as the mean of the three channels. The original source code is available in MATLAB. The metric has range 0-1 and a lower score express better image quality.

### 4.1.8. CIEDE2000

The CIEDE2000 [14] is not a quality metric itself, but is color-difference metric between two colours in the CIELab colour space (the image was converted using the function rgb2lab). The full description as well as the source code are available at the following link: http://www2.ece.rochester.edu/~gsharma/ciede2000/. Since this metric returns a difference value for every pixel of the image, it is computed the score for the whole image as the average over all the pixel scores. The original source code is available in MATLAB. Because of it represents a difference, a lower score express better image quality.

### 4.1.9. Butteraugli

The Butteraugli metric [15] computes the psycho-visual difference between two images. This metric was developed by Google. Butteraugli does not consider visually imperceptible differences and outputs a score that considers only the parts of the degraded image with perceived artifacts. This metric not only outputs a quality metric, but also a heatmap describing the differences between two images. The full code can be download from the following link:

https://github.com/google/butteraugli. The input images are in the RGB format. The reference code is in c++. A lower metric score express better image quality.

#### 4.1.10. WaDIQaM

The Weighted Average Deep Image QuAlity Measure for FR IQA (WaDIQaM) [16] is a deep-neural network based full reference quality metric. The network is trained end-to-end on the LIVE and TID2013 datasets. The reference code can be downloaded from the following link: https://github.com/dmaniry/deepIQA. This metric takes as input RGB images; moreover, the metric has been computed over all available pre-trained models. The reference code is available in python. The metric has range 0-100 and a lower score express better image quality.

#### 4.1.11. VMAF

Video Multimethod Assessment Fusion (VMAF) [17] is an objective full-reference video quality metric developed by Netflix in collaboration with the University of Southern California and the University of Texas at Austin. This metric is suitable for quality evaluation of different video codecs, encoders and encoding configurations. It relies on the fusion of several video quality metrics using support vector machines (SVM), i.e. using a machine learning approach. The reference code can be downloaded from the following link: https://github.com/dmaniry/deepIQA.

#### 4.1.12. LPIPS

LPIPS exploits the fact that deep network activations can be employed as a perceptual similarity metric, even for different neural network architectures. This metric provides quality scores by linearly "calibrating" networks - adding a linear layer on top of off-the-shelf classification networks. The reference code can be downloaded from the following link: https://github.com/richzhang/PerceptualSimilarity.

### 4.2. Subjective Scores Processing

Before analyzing the correlation between the MOS and the objective metrics, it is necessary to check the results of the subjective experiment to identify any outlier in the answers. In particular the distribution of the answers is expected to be normally distributed. In MATLAB is possible to verify if a set of measures follow the normal distribution using the following built-in functions: 1) Kolmogorov-Smirnov test (kstest); 2) Lilliefors test (lillietest). According to both metrics, our measures are normally distributed. This means that the MOS can be computed as the average over all subjective scores. The correlation is computed according to 3 different correlation metrics: the Pearson, the Spearman and the Kendall. The Pearson is the most popular metric for correlation computation. Its formulation is:

$$rho(a,b) = \frac{\sum_{i=1}^{n}(X_{a,i}-\overline{X}_a)(Y_{b,i}-\overline{Y}_b)}{\left\{\sum_{i=1}^{n}(X_{a,i}-\overline{X}_a)^2\sum_{j=1}^{n}(Y_{b,j}-\overline{Y}_b)^2\right\}^{1/2}},$$

where $X_a$ and $Y_b$ are column vectors, $X_a$ and $Y_b$ are the average over the whole column and $n$ is the length of each column vector. The Pearson correlation can be computed using the built in Matlab function as:

```
value = corr(mos,current_metric, 'Type', 'Pearson')
```

The Spearman metric is another popular way to compute the correlation. It is computed as:

$$rho(a,b) = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where $d$ is the difference of the rank of the two column vectors, and n is the length of the column vectors. The Spearman correlation can be computed using the built in Matlab function as:

```
value = corr(mos,current_metric, 'Type', 'Spearman')
```

The Kendall correlation is another popular metric to find the correlation coefficients between two measures; in particular this express the strength of the dependence between two variables. It is computed as:

$$\tau = \frac{2K}{n(n-1)},$$

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \xi^*(X_{a,i}, X_{a,j}, Y_{b,i}, Y_{b,j}),$$

$$\xi^*(X_{a,i}, X_{a,j}, Y_{b,i}, Y_{b,j}) = \begin{cases} 1 & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) > 0 \\ 0 & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) = 0 \\ -1 & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) < 0 \end{cases}$$

where $X_a$ and $Y_b$ are column vectors, $X_a$ and $Y_b$ are the average over the whole column and $n$ is the length of each column vector. The Kendall correlation can be computed using the built in Matlab function as:

```
value = corr(mos,current_metric, 'Type', 'Kendall')
```

## 4.3. Objective Metrics Performance Evaluation

The correlation between the MOS and the objective metrics was assessed with the Pearson, Spearman and Kendall metrics described in Section 4.2. The test was performed over the whole dataset as well as dividing it in two parts according to the metric type: classical and AI. Those two metrics are in fact introducing different types distortion into the image, so it is interesting to understand which quality metric performs better for each distortion type.

### 4.3.1. Experimental results: all codecs

The experimental results are shown in Figure 6. The x axis represents the listed the quality metrics, while on the y axis the correlation value. Values marked with the red cross represents the Pearson, the values with the blue star the Spearman and the green circle the Kendall correlation. The correlation has values between 0 and 1, where 0 express no correlation while 1 express perfect correlation.

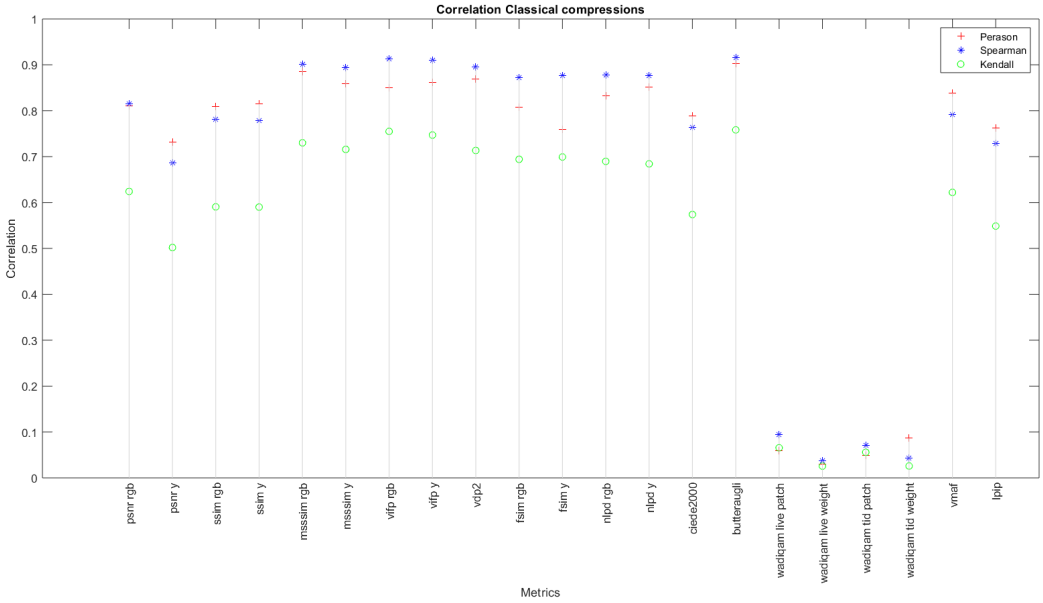



Figure 6. Correlation results between the MOS and objective metrics for all codecs.

According to this result, the metrics have been ordered from the best to the worst performing. The results are shown in Figure 7.

| Pearson | | |
|---|---|---|
| 1 | msssim rgb | 0.86747 |
| 2 | msssim y | 0.84058 |
| 3 | vifp y | 0.80823 |
| 4 | vdp2 | 0.80296 |
| 5 | vifp rgb | 0.80251 |
| 6 | nlpd y | 0.78759 |
| 7 | nlpd rgb | 0.77709 |
| 8 | fsim rgb | 0.76626 |
| 9 | fsim y | 0.72919 |
| 10 | vmaf | 0.72196 |
| 11 | lpip | 0.68589 |
| 12 | ssim y | 0.68418 |
| 13 | ciede2000 | 0.67893 |
| 14 | ssim rgb | 0.65134 |
| 15 | psnr rgb | 0.62327 |
| 16 | psnr y | 0.54439 |
| 17 | butteraugli | 0.53793 |
| 18 | wadiqam tid weight | 0.085961 |
| 19 | wadiqam live weight | 0.068955 |
| 20 | wadiqam live patch | 0.056057 |
| 21 | wadiqam tid patch | 0.031987 |

| Spearman | | |
|---|---|---|
| 1 | msssim rgb | 0.87685 |
| 2 | vifp rgb | 0.8713 |
| 3 | vifp y | 0.86878 |
| 4 | msssim y | 0.86042 |
| 5 | nlpd rgb | 0.82812 |
| 6 | vdp2 | 0.81963 |
| 7 | nlpd y | 0.79007 |
| 8 | fsim rgb | 0.75218 |
| 9 | fsim y | 0.72752 |
| 10 | vmaf | 0.68292 |
| 11 | lpip | 0.66133 |
| 12 | ssim y | 0.6475 |
| 13 | ssim rgb | 0.64033 |
| 14 | ciede2000 | 0.62332 |
| 15 | psnr rgb | 0.62287 |
| 16 | psnr y | 0.54987 |
| 17 | butteraugli | 0.51609 |
| 18 | wadiqam tid weight | 0.063183 |
| 19 | wadiqam live weight | 0.055247 |
| 20 | wadiqam live patch | 0.052199 |
| 21 | wadiqam tid patch | 0.036194 |

Figure 7. Ranking of the quality metrics for all codecs, from the highest to the lowest correlation.

In general the MS-SSIM and the VIF(P) metrics has the highest correlation with the MOS. On the contrary, the PSNR correlates only around 60% to the subjective score.

### 4.3.2. Experimental results: classical vs deep-learning compression methods

An additional experiment was performed where the decoded images for which subjective scores were obtained have been split according to the type of compression: classical (HEVC, JPEG2000, JPEG, WebP) or AI (remaining codecs). In fact these codecs are so different that introduce different artefacts into the images. Thus, the metric performance was assessed for each type of compression. The results for the classical-compressed images can be seen in Figure 8, while the ranking is presented in Figure 9. As shown, the Butteraugli metric correlates very well with the MOS, together with the MS-SSIM and VDP2/VIF(P).

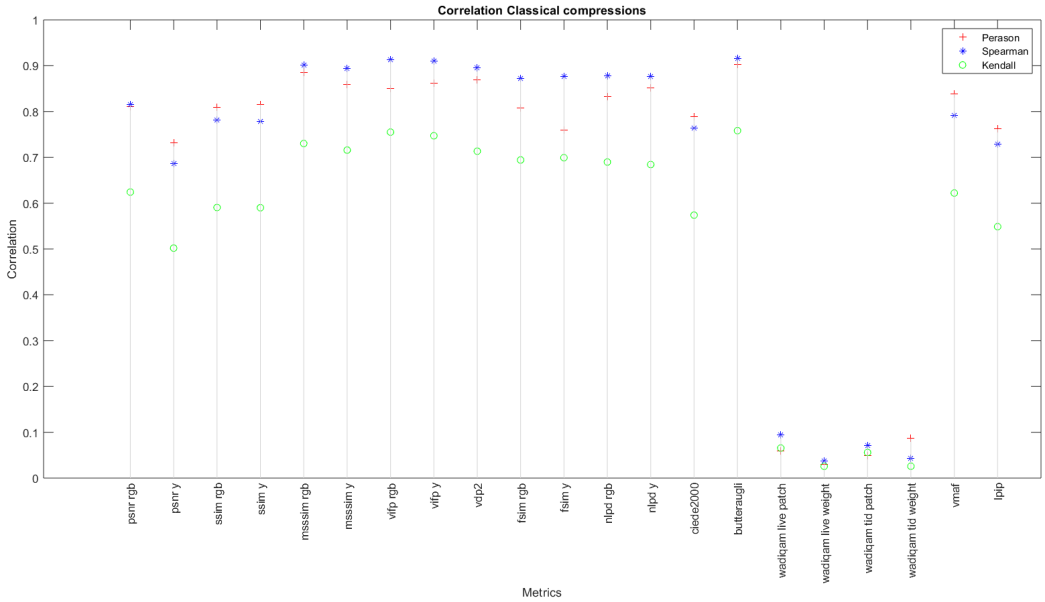



Figure 8. Correlation results between the MOS and objective metrics for classical codecs: HEVC, WebP, JPEG2000 and JPEG.

| Pearson | | |
|---|---|---|
| 1 | butteraugli | 0.90319 |
| 2 | msssim rgb | 0.88539 |
| 3 | vdp2 | 0.86986 |
| 4 | vifp y | 0.86165 |
| 5 | msssim y | 0.85928 |
| 6 | nlpd y | 0.85106 |
| 7 | vifp rgb | 0.85014 |
| 8 | vmaf | 0.83851 |
| 9 | nlpd rgb | 0.83225 |
| 10 | ssim y | 0.81523 |
| 11 | psnr rgb | 0.81114 |
| 12 | ssim rgb | 0.80881 |
| 13 | fsim rgb | 0.80745 |
| 14 | ciede2000 | 0.78862 |
| 15 | lpip | 0.76189 |
| 16 | fsim y | 0.75878 |
| 17 | psnr y | 0.73147 |
| 18 | wadiqam tid weight | 0.086368 |
| 19 | wadiqam live patch | 0.059524 |
| 20 | wadiqam tid patch | 0.048268 |
| 21 | wadiqam live weight | 0.030653 |

| Spearman | | |
|---|---|---|
| 1 | butteraugli | 0.91609 |
| 2 | vifp rgb | 0.91297 |
| 3 | vifp y | 0.9108 |
| 4 | msssim rgb | 0.9019 |
| 5 | vdp2 | 0.89574 |
| 6 | msssim y | 0.89398 |
| 7 | nlpd rgb | 0.87824 |
| 8 | nlpd y | 0.87615 |
| 9 | fsim y | 0.87606 |
| 10 | fsim rgb | 0.87195 |
| 11 | psnr rgb | 0.81508 |
| 12 | vmaf | 0.79128 |
| 13 | ssim rgb | 0.78076 |
| 14 | ssim y | 0.77894 |
| 15 | ciede2000 | 0.76362 |
| 16 | lpip | 0.72875 |
| 17 | psnr y | 0.68707 |
| 18 | wadiqam live patch | 0.09449 |
| 19 | wadiqam tid patch | 0.070268 |
| 20 | wadiqam tid weight | 0.043825 |
| 21 | wadiqam live weight | 0.037956 |

Figure 9. Ranking of the quality metrics for classical codecs, from the highest to the lowest correlation.

On the other hand, the results for the deep-learning compressed images are shown in Figure 10, while the ranking is shown in Figure 11.

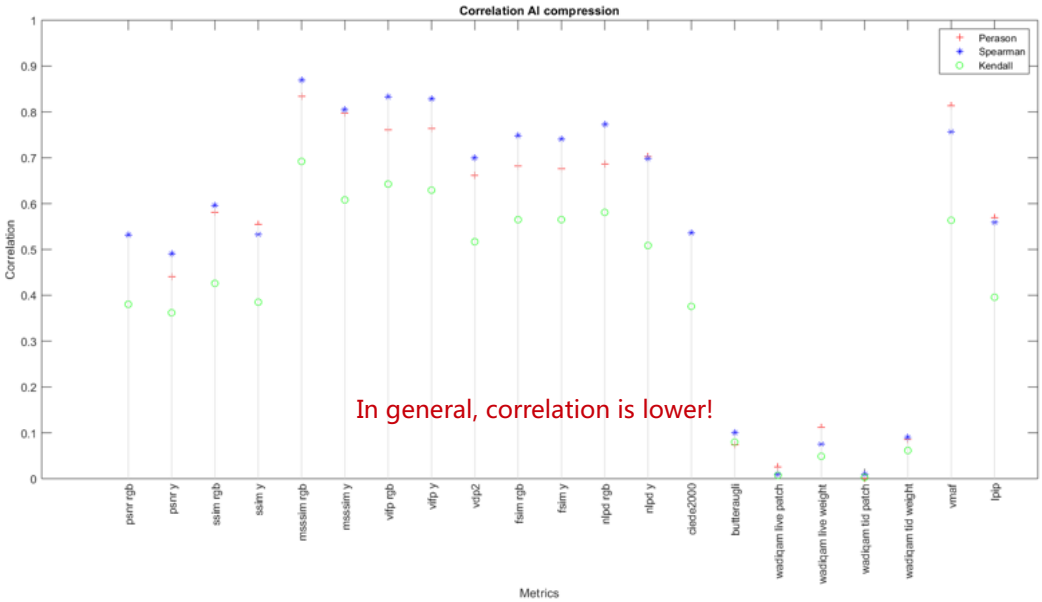



Figure 10. Correlation results between the MOS and objective metrics for deep learning codecs.

| Pearson | | |
|---|---|---|
| 1 | msssim rgb | 0.83431 |
| 2 | vmaf | 0.81395 |
| 3 | msssim y | 0.79723 |
| 4 | vifp y | 0.76336 |
| 5 | vifp rgb | 0.7613 |
| 6 | nlpd y | 0.70289 |
| 7 | nlpd rgb | 0.68654 |
| 8 | fsim rgb | 0.68215 |
| 9 | fsim y | 0.6755 |
| 10 | vdp2 | 0.66208 |
| 11 | ssim rgb | 0.58148 |
| 12 | lpip | 0.56922 |
| 13 | ssim y | 0.55443 |
| 14 | ciede2000 | 0.5354 |
| 15 | psnr rgb | 0.53113 |
| 16 | psnr y | 0.441 |
| 17 | wadiqam live weight | 0.11244 |
| 18 | wadiqam tid weight | 0.085947 |
| 19 | butteraugli | 0.074473 |
| 20 | wadiqam live patch | 0.026234 |
| 21 | wadiqam tid patch | 0.001075 |

| Spearman | | |
|---|---|---|
| 1 | msssim rgb | 0.86962 |
| 2 | vifp rgb | 0.83317 |
| 3 | vifp y | 0.82763 |
| 4 | msssim y | 0.80553 |
| 5 | nlpd rgb | 0.77336 |
| 6 | vmaf | 0.75724 |
| 7 | fsim rgb | 0.74774 |
| 8 | fsim y | 0.74074 |
| 9 | vdp2 | 0.70001 |
| 10 | nlpd y | 0.69872 |
| 11 | ssim rgb | 0.59527 |
| 12 | lpip | 0.55888 |
| 13 | ciede2000 | 0.53544 |
| 14 | ssim y | 0.53303 |
| 15 | psnr rgb | 0.53122 |
| 16 | psnr y | 0.48982 |
| 17 | butteraugli | 0.1007 |
| 18 | wadiqam tid weight | 0.09007 |
| 19 | wadiqam live weight | 0.075662 |
| 20 | wadiqam tid patch | 0.0095786 |
| 21 | wadiqam live patch | 0.0089276 |

Figure 11. Ranking of the quality metrics for deep learning codecs, from the highest to the lowest correlation.

As shown, the correlation with the MOS is lower in general, and the Butteraugli metric now performs poorly. Still in this case, the MS-SSIM is the best quality metric, even if with a lower performance of around 7% less compared to the result obtained for the classical compression methods. Also, it is important to underline that only 8 images were selected for subjective assessment and thus, to confirm the results it is necessary to repeat the experiment with a larger dataset, including a larger variety of image types and distortions.

From the results over all the images, the metrics that correlates better (according to Pearson) with the MOS follow this order: MS-SSIM, VIF(P) and NLDP. If only classical classical compression metrics are considered, the scores with the highest correlation, in order, are: Butteraugli, MS-SSIM and VDP2. On the contrary, for AI compressed images the best performing metrics are: MS-SSIM, VMAP and VIF(P); however, for this last case, VMAF underperforms for the Spearman correlation. In general, the MS-SSIM is clearly the winner since it predicts well the image quality for the different compression artefacts.

Another observation is that the decoded images have different statistics from natural images (e.g. TE20), and this could explain why the WaDIQaM metric (AI) performs poorly; moreover, the selected original images have characteristics very different (e.g. spatial resolution) from the training-set used to create the WaDIQaM model. It could be interesting to retrain the network using images more similar and check if improvement on the correlation values are obtained.

# 5. Conclusions and Future Work

JPEG AI image quality assessment experiments evaluated the performance of five learning-based image coding solutions against four traditional image codecs, on 8 SD to UHD natural images, at four different bitrates. Results show that subjective and objective qualities of state-of-the-art learning-based image coding algorithms were competitive to transform-based codecs. Thorough inspection on the visual results revealed the typical artifacts encountered in the learning-based codecs. Moreover, several full-reference objective quality metrics were evaluated to find which metric correlates better with human opinion scores, for different types of coding solutions, i.e. for traditional and learning based image codecs.

Future work defined in JPEG AI CTC document suggests carrying out SS tests that are expected to reveal different characteristics of the learning-based solutions in the absence of reference images. For example, the contrast changes in MS-SSIM-optimized codecs may be perceived less as artifacts when not presented side-by-side with the references. Similarly, a variant of DSIS test that measures the level of "naturalness" of learning-based solutions perceived by subjects is proposed. Using the same DSIS methodology and changing the rating scale by asking the subjects to rate the naturalness of the images is expected to provide insight into the integration of learning-based features into human vision. These experiments are to involve state-of-the-art codecs, as well as solutions being currently developed.

# 6. References

*[1] J. Ascenso and P. Akayzi, "JPEG AI Image Coding Common Test Conditions", ISO/IEC JTC 1/SC 29/WG 1 N84035, 84th Meeting, Brussels, Belgium, July 2019.*

*[2]G. Toderici, S.M. O'Malley, S.J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, R. Sukthankar, "Variable Rate Image Compression with Recurrent Neural Networks," International Conference on Learning Representations, San Juan, Puerto Rico, May 2016.*

*[3] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, N. Johnston, "Variational Image Compression With a Scale Hyperprior," International Conference on Learning Representations, Vancouver, Canada, April 2018.*

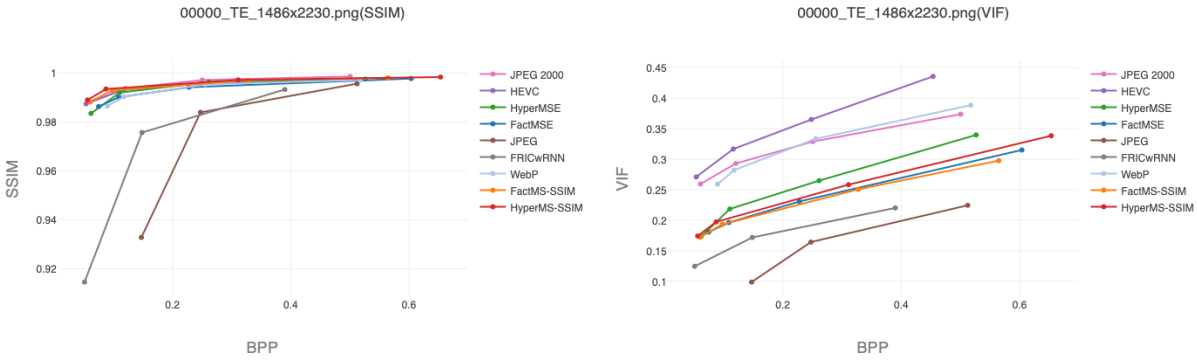
*[4] ITU-R BT.2022. "General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays". International Telecommunication Union, August 2012.*

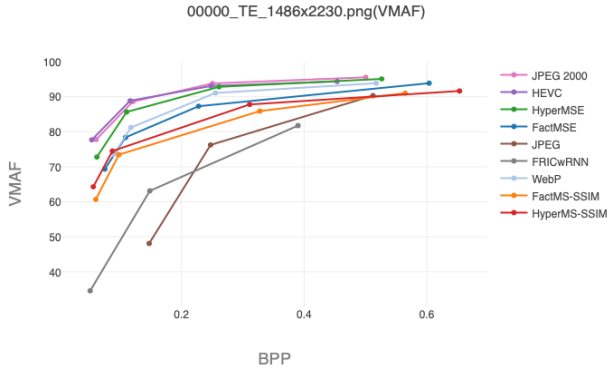
*[5] ITU-R BT.2100. "Image parameter values for high dynamic range television for use in production and international programme exchange", International Telecommunication Union, July 2018.*

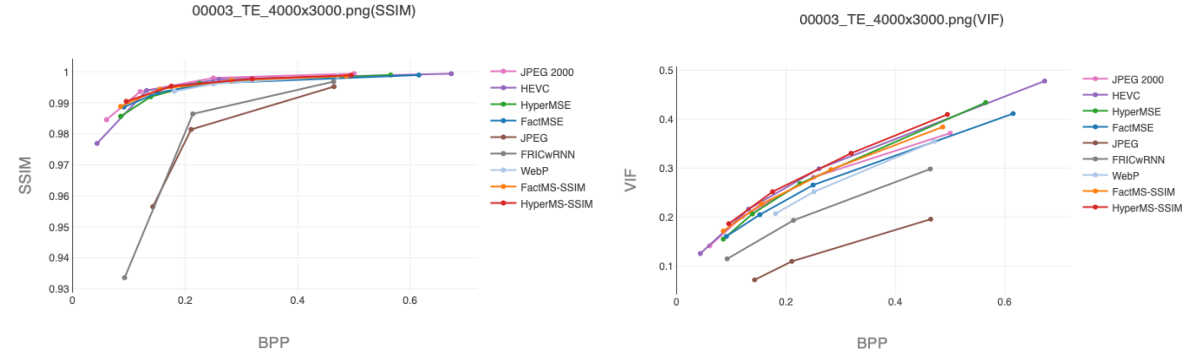
*[6] ITU-R BT.500-13. "Methodology for the subjective assessment of the quality of television pictures". International Telecommunication Union, January 2012.*

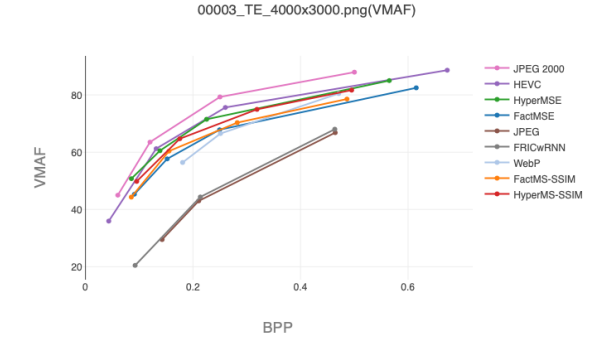
*[7] F. De Simone, L. Goldmann, J.-S. Lee, and T. Ebrahimi. "Towards high efficiency video coding: Subjective evaluation of potential coding technologies". Journal of Visual Communication and Image Representation, 22(8):734–748, 2011.*
*[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004.*
*[9] Z. Wang, E. P. Simoncelli1 and A. C. Bovik, "Multi-scale Structural Similarity for Image Quality Assessment", 37th IEEE Asilomar Conference on Signals, Systems and Computers, Nov. 2003.*
*[10] H.R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, August 2004.*
*[11] R. Mantiuk, K. J. Kim, A. G. Rempel, W. Heidrich, "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions", ACM Transactions on Graphics (TOG), vol. 30, no. 4, July 2011.*
*[12] L. Zhang, L. Zhang, X. Mou, D. Zhang, "FSIM: a feature similarity index for image quality assessment," IEEE Transactions on Image Processing, vol. 20, no. 8, pp. 2378-2386, August 2011.*
*[13] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized laplacian pyramid", Electronic Imaging, vol. 16, pp. 1-6, 2016.*
*[14] G. Sharma, W. Wu, E. N. Dalal, "The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations", Color Research and Application, vol. 30. No. 1, pp. 21-30, February 2005.*
*[15] J. Alakuijala, R. Obryk, Z. Szabadka, J. Wassenberg, "Users prefer Guetzli JPEG over same-sized libjpeg", arXiv:1703.04416*
*[16] S. Bosse, D. Maniry, K.R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," IEEE Transactions on Image Processing, vol. 27, no. 1, pp. 206-219, January 2018.*
*[17] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward A Practical Perceptual Video Quality Metric," https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652, Accessed: 2019-11-21.*
*[18] R. Zhang, P. Isola, A. Efros, E. Shechtman, O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric", IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018.*

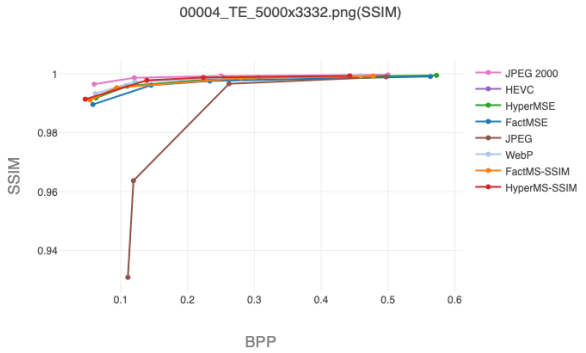
# Annex A – Objective quality assessment

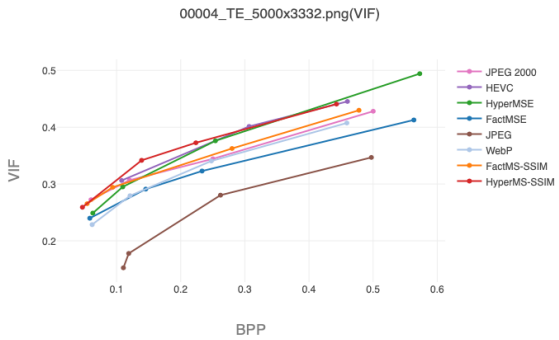


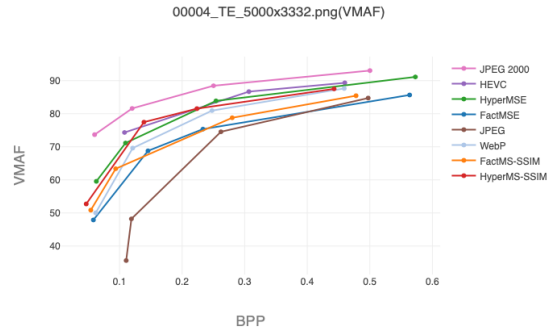



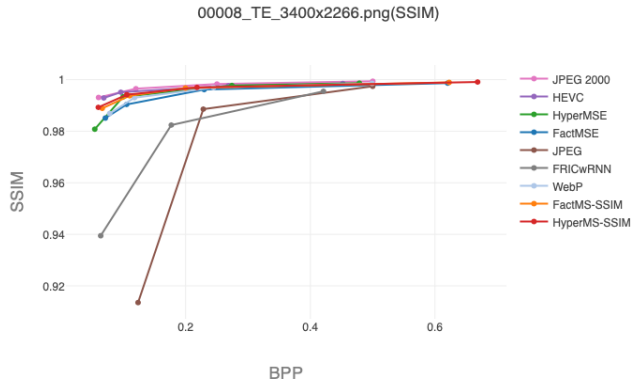



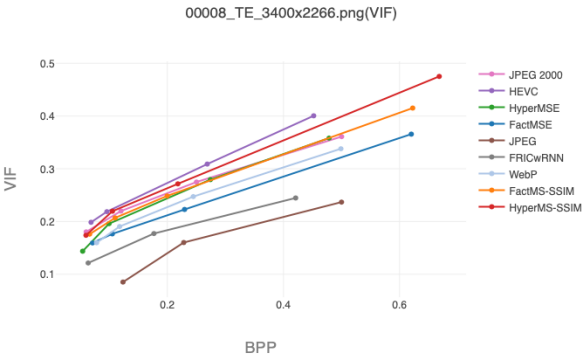



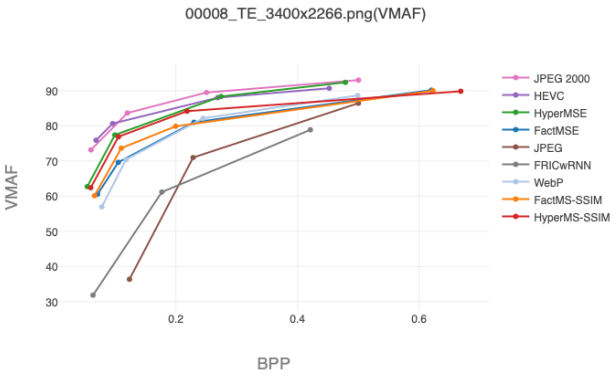

00004_TE_5000x3332.png(SSIM)
SSIM
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
WebP
FactMS-SSIM
HyperMS-SSIM

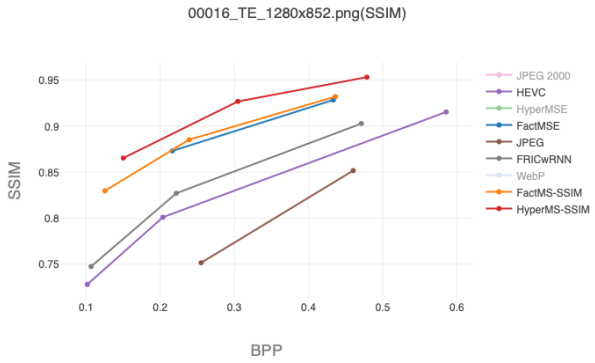

00004_TE_5000x3332.png(VIF)
VIF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
WebP
FactMS-SSIM
HyperMS-SSIM

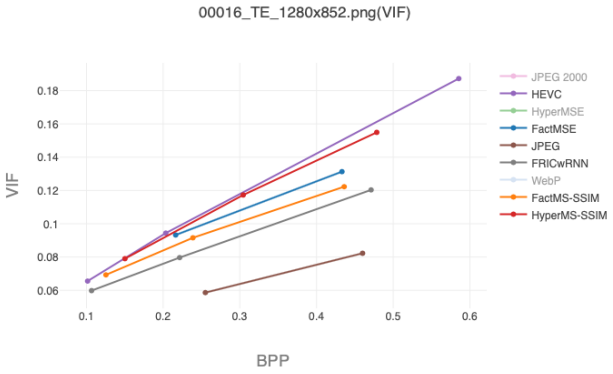

00004_TE_5000x3332.png(VMAF)
VMAF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
WebP
FactMS-SSIM
HyperMS-SSIM

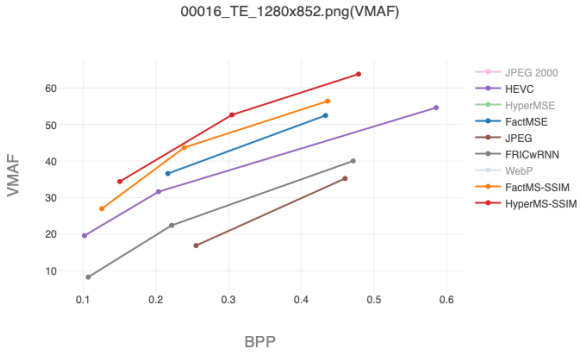

00008_TE_3400x2266.png(SSIM)
SSIM
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

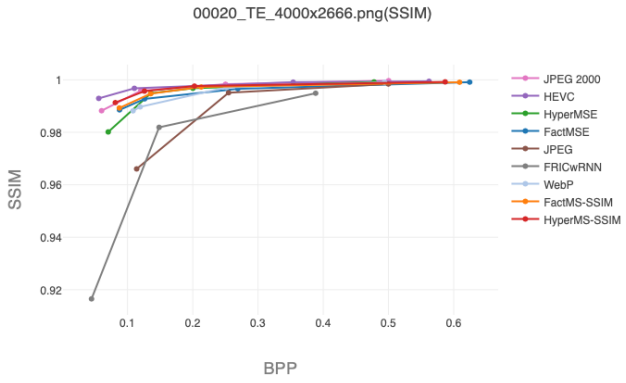

00008_TE_3400x2266.png(VIF)
VIF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

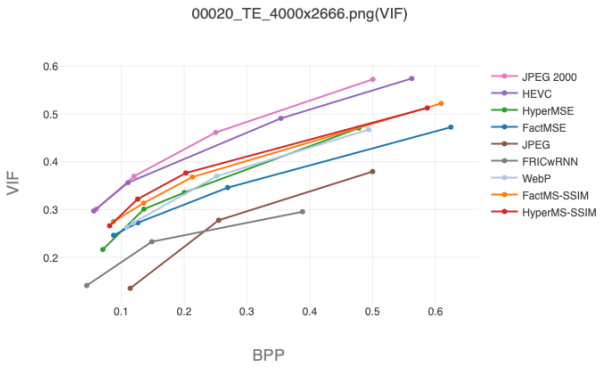

00008_TE_3400x2266.png(VMAF)
VMAF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

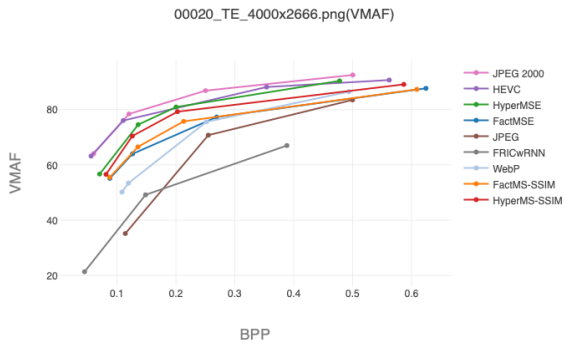

00016_TE_1280x852.png(SSIM)
SSIM
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

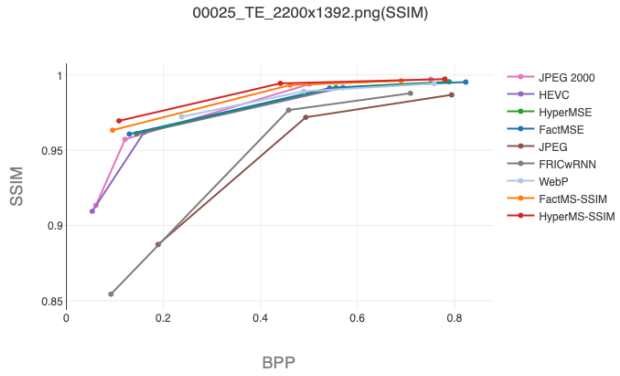

00016_TE_1280x852.png(VIF)
VIF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

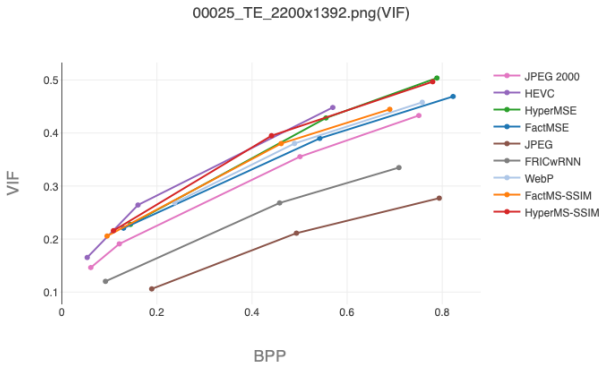

00016_TE_1280x852.png(VMAF)
VMAF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

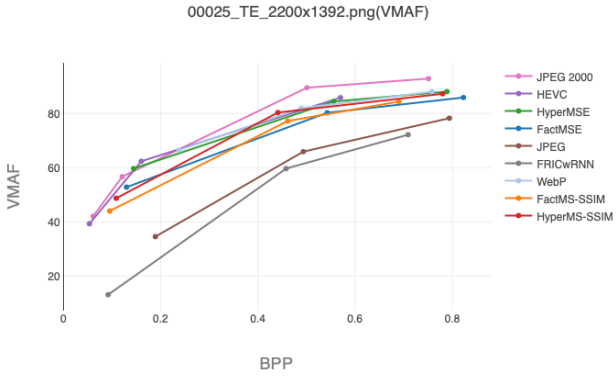

00020_TE_4000x2666.png(SSIM)
SSIM
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

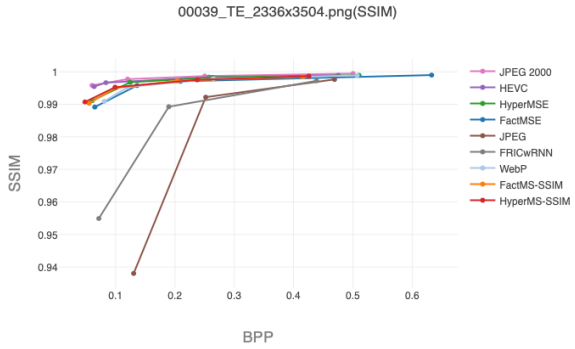

00020_TE_4000x2666.png(VIF)
VIF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

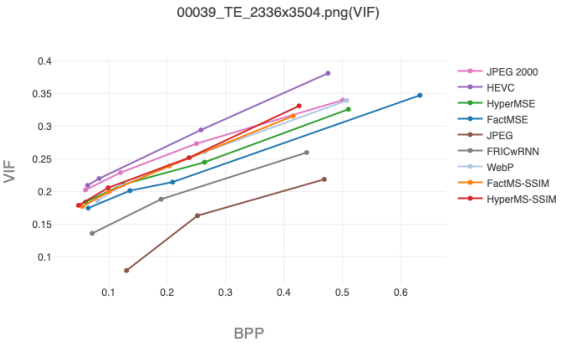

00020_TE_4000x2666.png(VMAF)
VMAF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM

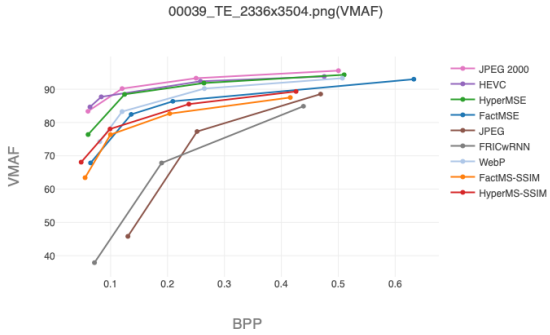

00025_TE_2200x1392.png(SSIM)
SSIM
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM


00025_TE_2200x1392.png(VIF)
VIF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM


00025_TE_2200x1392.png(VMAF)
VMAF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM


00039_TE_2336x3504.png(SSIM)
SSIM
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM


00039_TE_2336x3504.png(VIF)
VIF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM


00039_TE_2336x3504.png(VMAF)
VMAF
BPP
JPEG 2000
HEVC
HyperMSE
FactMSE
JPEG
FRICwRNN
WebP
FactMS-SSIM
HyperMS-SSIM