NNCD-IQA: A new neural networks based compressed database for image quality assessment.

Zohaib Amjad Khan, Tassnim Dardouri, Mounir Kaaniche, Gabriel Dauphin

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13842-8

NNCD-IQA: A new neural networks based compressed database for image quality assessment

Zohaib Amjad Khan[1] · Tassnim Dardouri[2] · Mounir Kaaniche[2] · Gabriel Dauphin[2]

Received: 24 December 2021 / Revised: 9 April 2022 / Accepted: 6 September 2022

Abstract
Objective and subjective quality assessment is still a challenging problem in various image processing tasks. For instance, in the context of image compression, most of the conducted studies have focused on image datasets encoded using standard algorithms such as JPEG and JPEG2000. In this paper, we propose to further investigate the quality assessment issue in the presence of neural networks-based compressed images. More precisely, a new database of compressed images has been firstly built using JPEG2000 standard as well as four recent neural networks based coding schemes. Then, subjective experiments are performed to obtain the mean opinion scores of the generated distorted images. Finally, an extensive evaluation and analysis of objective image quality assessment metrics is achieved. For instance, in addition to conventional and machine learning metrics, we have considered different deep learning based models, which have been trained on our database. The new subjective database with its associated mean opinion scores as well as the learned models are publicly available at https://github.com/zakopz/NNCD-IQA-Database. The obtained results show the interest of deep learning based metrics in the context of neural networks-based compressed images.

Keywords Image compression · Neural networks · Quality assessment · Subjective scores · Learning based metrics

Zohaib Amjad Khan and Tassnim Dardouri contributed equally.

✉ Tassnim Dardouri
tasnim.dardouri@edu.univ-paris13.fr

Zohaib Amjad Khan
zohaib.khan@centralesupelec.fr

Mounir Kaaniche
mounir.kaaniche@univ-paris13.fr

Gabriel Dauphin
gabriel.dauphin@univ-paris13.fr

[1] L2S, CentraleSupélec, Université Paris Saclay, 91190 Gif-sur-Yvette, France

[2] L2TI, UR 3043, Université Sorbonne Paris Nord, F-93430, Villetaneuse, France

Published online: 27 September 2022

1 Introduction

Due to the continuous advances of display and acquisition technologies, huge amounts of diverse visual data are generated every day, which constitute a major issue in terms of storage and transmission. In this respect, many research works have been dedicated to the design of efficient visual data compression methods [15, 19]. For instance, most of the developed algorithms are devoted to lossy compression and aim at minimizing the distortion of the reconstructed image at a given bitrate. In this context, it becomes necessary to find appropriate quality metrics to assess the quality of the reconstructed images resulting from the employed image compression method. However, most of the developed quality assessment studies, conducted in the context of compressed images, have considered conventional (i.e non-deep learning) coding techniques [61]. For this reason, the main objective of this paper is to perform an extensive analysis of quality assessment problem in the case of neural networks-based compressed images. Before summarizing our contributions, we will first review recent neural networks-based image compression techniques as well as image quality assessment related works.

1.1 Related works

Many research efforts have been devoted to image (resp. video) compression, and contributed to various standardization activities such as JPEG and JPEG2000 [50] (resp. HEVC, AV1 and VVC [14]). The above codecs rely on linear transforms which are either the Discrete Cosine Transform (DCT) or the Discrete Wavelet Transform (DWT). However, such linear transforms may not appear so efficient to process complex and non-linear data. For this reason, deep learning-based compression algorithms have attracted a great attention in the last years due to the advantages of Neural Networks (NN) in achieving accurate non-linear approximation and enabling high level data description. An overview of image and video compression with deep learning approaches is provided in [28, 31]. More precisely, the developed NN-based algorithms consist generally of three main steps. First, a NN-based analysis stage is performed to transform the input image into a compact representation. The latter is then quantized and encoded. Finally, the inverse transform is achieved to obtain the reconstructed image. This typical architecture is referred to as auto-encoder where the network parameters are trained in an end-to-end manner [1, 4, 5, 8, 27, 40, 54]. It should be noted here that the main differences between these methods are related to the employed NN architecture and/or the retained loss function in the training phase. For instance, among the existing architectures, the Convolutional Neural Network (CNN) and Fully Connected Neural Network (FCNN) have been recently investigated for intra prediction in the context of image and video coding [26, 42]. A hybrid method, where small (resp. large) blocks are predicted using an FCNN (resp. a CNN) model, is proposed in [13]. Moreover, motivated by the different advantages of transform coding schemes, other methods have been developed to improve DCT (Discrete Cosine Transform) and DWT-based coding schemes [2, 10, 29, 30]. Indeed, a DCT-based coding scheme using a CNN is used in [29]. In [2], a DWT is first applied to the input image, and then, the generated subbands are fed into a CNN to produce the final detail coefficients. In [30], the authors propose to design a separable lifting structure based wavelet transform using a CNN. While the latter method employs the CNN for only the prediction stage, a fully nonlinear transform where prediction and update stages are performed using an FCNN has been recently developed in [10]. This recent method has also been made adaptive by taking into account the input image to be encoded.

Since the developed methods are often dedicated to lossy compression, the quality assessment of the reconstructed (i.e. decoded) images becomes an important step to evaluate the performance of these deep learning based coding schemes. In fact, quality assessment is a crucial step in various image processing tasks such as compression, restoration [59], inpainting [3], etc. Most importantly, in the context of image compression, the quality assessment problem has been widely investigated in the case of JPEG and JPEG2000 encoded images. In this context, the full-reference PSNR and SSIM metrics have been extensively used to evaluate the performance of conventional image compression algorithms. Moreover, other works have also been developed to design no-reference quality assessment metrics [44, 57, 58]. For example, in [58], a no-reference metric for JPEG compressed images based on DCT coefficients distribution and PSNR estimation is proposed. Another method for JPEG2000 compressed images using natural scene statistics is developed in [44].

However, in the context of neural networks based image compression algorithms, the latter are often evaluated in terms of PSNR and SSIM (or MS-SSIM). In addition to this commonly used evaluation approach, there are very few works which have proposed to conduct some subjective experiments [7, 32, 53, 55]. Indeed, in [32], the authors achieve a pairwise comparison study with 10 observers to show the preference of their method compared to JPEG standard and a neural network baseline method [54]. In [53], a single-stimulus rating test with 25 observers is also performed to validate their coding approach and show its preference over JPEG and JPEG2000 compression methods as well as the neural network baseline method [54]. A similar subjective evaluation was conducted in [7] to show that MS-SSIM is better than MSE for optimizing an end-to-end learned compression method. However, the latter works do not investigate the correlation between the obtained Mean Opinion Scores (MOS) and the employed objective metrics (PSNR and MS-SSIM). For this reason, Valenzise et al. have proposed to study in [55] the accuracy of classical metrics in predicting MOS for deep learning based compression methods. More precisely, using a double stimulus rating test and 23 observers, they have considered 6 reference images and 4 image compression methods which are JPEG2000, BPG as well as two NN-based algorithms [4, 54]. It has been concluded that conventional PSNR and SSIM metrics are not appropriate to assess the quality of deep learning-based compressed images. A similar work using 8 reference images and focusing on traditional objective quality assessment metrics has also been presented in [52].

1.2 Limitations and contributions

Although great attention has been paid to Image Quality Assessment (IQA), there are still some issues that need to be addressed. For instance, in the context of compressed images, most of image quality assessment subjective studies have been conducted using JPEG and JPEG2000 coding schemes. Moreover, the few recent works dedicated to QA in the case of deep learning based compressed images (addressed in Section 1.1) have performed some analysis with mainly non-deep learning based metrics. It should be noted here that, in these recent works, the employed deep learning based compressed images are not publicly available. Finally, the PSNR and SSIM metrics, often considered to assess the quality of compressed images, were found to be less correlated with human opinion as discussed in recent deep learning based image coding works [4, 11, 54].

For these reasons, the objective of this paper is to further investigate quality assessment issue in the context of deep learning based image compression. More precisely, we first propose to build a new Neural Networks-based Compressed image Database referred to as NNCD-IQA. The distorted images, obtained with the standard JPEG2000 coding standard

and some recent neural networks based compression algorithms, as well as their associated MOS are made publicly available. Moreover, compared to the previous subjective studies [52, 55], the proposed database includes more reference images with recent NN compression algorithms resulting in a larger dataset. It is worth pointing out that such new subjective dataset presents a great interest to the research communities working both on the development of IQA algorithms as well as the design of deep learning based compression algorithms. Based on this database, we achieve an extensive evaluation of image quality assessment metrics and analyse their correlation with the subjective quality scores. In this respect, and unlike recent studies [52, 55] focusing on traditional IQA algorithms, we propose here to investigate new emerging IQA methods based on deep learning approaches. It is important to note here that the corresponding models have been trained on our database, which allows us to learn new models more adapted to the quality assessment of neural networks-based compressed images.

The remainder of this paper is organized as follows. In Section 2, an overview of the retained neural networks based image compression algorithms is provided. Then, the subjective test methodology is described in Section 3. Finally, the objective IQA metrics as well as the experimental results are discussed in Section 4, and some conclusions are drawn in Section 5.

2 Retained neural networks-based image compression algorithms

Since the focus of this paper is on the quality assessment of deep learning compressed images, we propose to select one conventional image compression method, which is often used as a comparison method, and four neural networks based compression methods. The conventional method is the standard JPEG2000 coding scheme [51] whereas the neural networks based ones are described in the following.

2.1 End-to-end learned image compression models

This method, developed by Ballé et al. [4], is among the first developed end-to-end image compression methods based on deep learning. Its block diagram is shown in Fig. 1.

Thus, the method consists of a nonlinear analysis transform g_a, a uniform quantizer q and a nonlinear synthesis transform g_s. For the analysis and synthesis transforms, they are performed using three convolutional layers and nonlinear activation functions. More precisely, for the analysis (resp. synthesis) stage, each convolutional layer is followed by downsampling (resp. upsampling) and generalized divisive normalization GDN (resp. inverse of the generalized divisive normalization IGDN) operations. To optimize their network and find the optimal parameters of the analysis and synthesis transforms, the authors use a Rate-Distortion (R-D) loss function $\mathcal{L}$ given by

$$\begin{aligned} \mathcal{L} &= R + \lambda D \\ &= -\mathsf{E}[\log_2(P_{\mathbf{q}})] + \lambda \mathsf{E}[d(\mathbf{x}, \hat{\mathbf{x}})] \end{aligned} \tag{1}$$

where $P_{\mathbf{q}}$ is the discrete probability distribution of the quantized vector, $d(\mathbf{x}, \hat{\mathbf{x}})$ is the distortion (typically the Mean Square Error) between the original $\mathbf{x}$ and reconstructed images $\hat{\mathbf{x}}$, λ controls the trade-off between the rate and distortion terms, and $\mathsf{E}[\cdot]$ represents the expectation operation approximated by average over a given training set of images. For optimization purpose via gradient descent algorithm, $P_{\mathbf{q}}$ will be approximated by the density

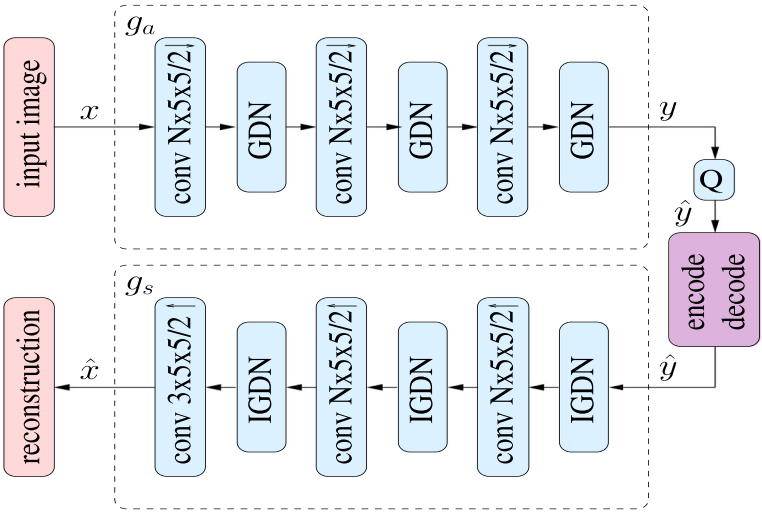


Fig. 1 Block diagram of the baseline end-to-end image compression algorithm [4]

function of $\tilde{\mathbf{y}}$ (denoted by $p_{\tilde{\mathbf{y}}}$) which is obtained by replacing the quantizer with an additive i.i.d uniform noise source. Based on this approximation of the quantized coefficient distribution, and given a probability model $p_{\tilde{\mathbf{y}}}$, the loss function becomes suitable for stochastic optimization.

The above architecture [4] has been then considered as a reference model in many other deep learning based image compression algorithms [5, 8, 25, 47]. Among them, we retain here the method proposed in [5]. The latter aims to extend the first model [4] by integrating a hyperprior h that captures the spatial dependencies in the latent representation $\mathbf{y}$. The block diagram of this architecture is shown in Fig. 2.

In this extended architecture, h_a and h_s can be seen as an auxiliary autoencoder that aims to estimate the probability distribution $p_{\tilde{\mathbf{y}}}$ after decoding $\hat{\mathbf{z}}$. In this respect, different

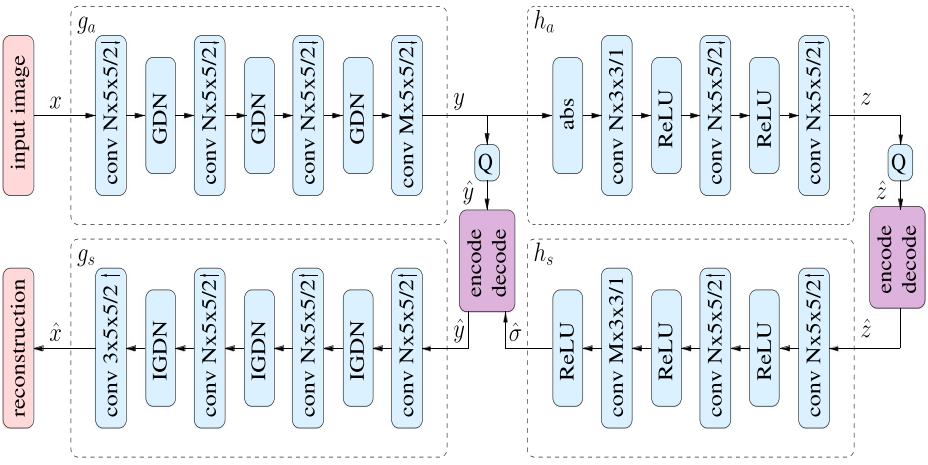


Fig. 2 Block diagram of the end-to-end image compression algorithm with a scale hyperprior [5]

methods have been developed to model this distribution. More precisely, *three variants* of this generic autoencoder (AE) architecture have been retained in this work. The first approach uses a non-adaptive distribution model based on piecewise linear functions and is referred to as *factorized*-prior model [5]. The second one, designated by *hyperprior* model in [5], assumes a zero mean Gaussian distribution with standard deviation parameters σ^2. The third approach corresponds to a more recent work where authors resort to a *Gaussian mixture model* [8]. In what follows, these three NN based coding schemes will be designated by AE-Factor [5], AE-Hyp-GM [5] and AE-Hyp-GMM [8], respectively.

2.2 Fully connected network for lifting based image coders

While the previous approaches as well as most of the developed neural networks based compression methods are not suitable for *lossy-to-lossless* coding applications, a novel method based on lifting schemes [49] has been recently developed in [10]. In this architecture, shown in Fig. 3, the conventional predictors and update linear operators are replaced by fully connected neural network (FCNN) models.

More precisely, three FCNN based prediction models, denoted by $f_j^{(HH)}$, $f_j^{(LH)}$, and $f_j^{(HL)}$, are employed to generate the three detail wavelet subbands oriented diagonally, vertically and horizontally. These FCNN models are learned by minimizing the energy (i.e the ℓ_2-norm) of the detail coefficients. Then, an FCNN based update model, designated by $f_j^{(LL)}$, is used to generate the approximation subband. The latter model is optimized by minimizing the quadratic error between the approximation coefficients and those obtained using an ideal low pass filter. More details regarding this approach can be found in [11].

3 Subjective study

In this section, we describe the conducted subjective experiment to build a new database with MOS for quality assessment of neural networks based compressed images.

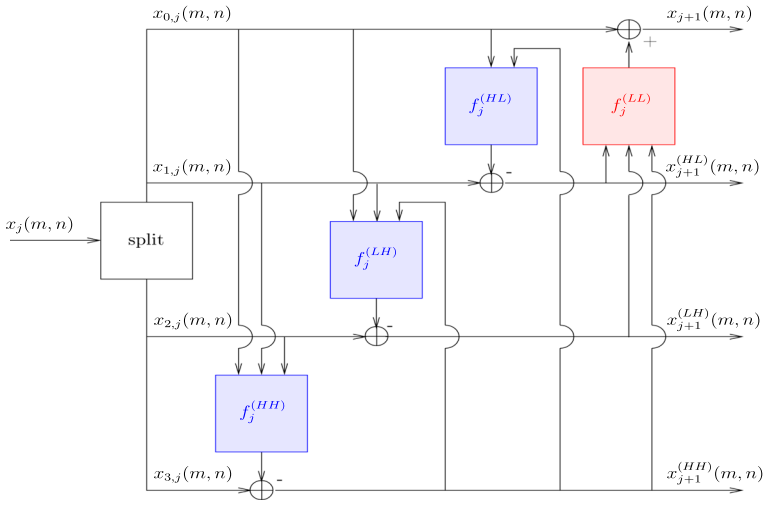


Fig. 3 Block diagram of the fully connected network based lifting coding scheme [11]

3.1 Image database

Our database has been derived from 16 reference (i.e. uncompressed) images taken from the standard Kodak PhotoCD dataset. The latter is composed of color pictures of size 768×512 with different foreground/background contents as it can be seen from some samples shown in Fig. 4.

While other dataset images (such as CLIC and DIV2K) exist for training neural networks based compression models, it should be noted that the popular Kodak dataset has been selected since it is often used to validate recent deep learning-based image coding algorithms.

Based on these reference images, we applied the standard JPEG2000 image compression algorithm as well as the four deep learning-based ones, described in Section 2, to generate the different distorted images. Moreover, and in order to generate different distortion levels, the retained compression methods are performed at four bitrates (i.e four quality levels). Since distortions are more visible at low and middle bitrates, the latter are set to 0.1,

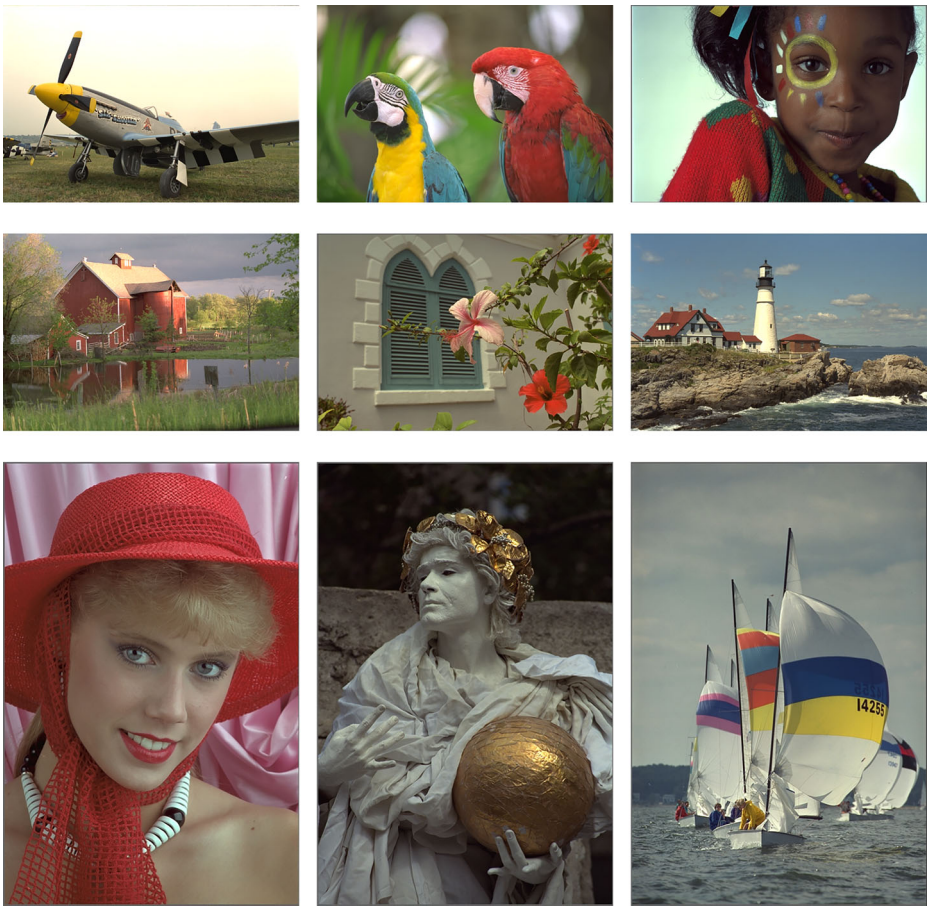


Fig. 4 Some reference images taken from Kodak and used in our study

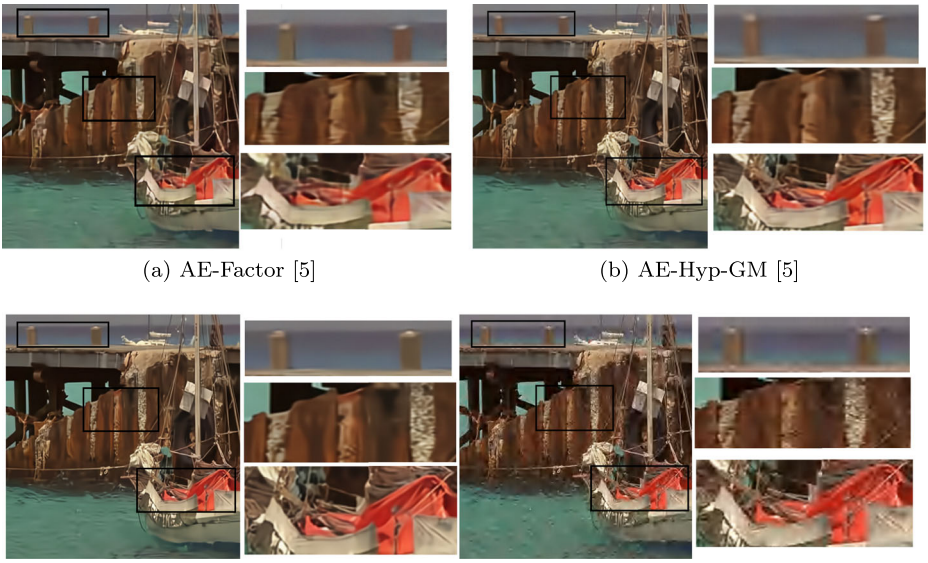
0.2, 0.3, 0.4 bits per pixel (bpp). Some examples of compressed images are displayed in Fig. 5. To further illustrate the distortion types obtained with conventional and recent neural networks-based compression methods, Fig. 6 shows an example of decoded images at 0.2 bpp. As it can be seen from the cropped windows (shown on the right side of each decoded image), JPEG2000 results in visual artefacts like blurring and ringing. While NN-based compression methods allow to suppress the ringing artefacts, they may suffer from blurring and smoothing effects.

It is worth noting that the created database is made publicly available. For instance, all of the publicly available IQA databases (shown in Table 1) are generated using JPEG and/or JPEG2000 compression methods (in addition to other distortion types such as noise, blur, etc). However, the recent image quality assessment works devoted to deep learning based compressed images (addressed in Section 1.1) are not public. Moreover, compared to these recent subjective studies which are conducted with small datasets (6 reference images and 113 compressed ones in [55]; 8 reference images and 240 compressed ones in [52]), our new database is larger and contains 336 images (16 reference and 320 compressed ones).

3.2 Test methodology

While single and double stimulus procedures have been widely used in the literature [22], we propose in this paper to follow the second one during the subjective quality assessment step. More specifically, we have selected Double Stimulus Impairment Scale (DSIS) methodology [17] where a pair of images are displayed side-by-side with one of them being the reference and the other one is the distorted. During the subjective experiments, the observers are asked to evaluate the distorted image in comparison to the reference one using a continuous linear impairment scale from 0 to 100. The scale is also marked with five equally-spaced adjectives corresponding to the level of impairment. These five levels from best (i.e 100) to worst (i.e 0) were imperceptible, perceptible but not annoying, slightly annoying, annoying and very annoying. In our study, it has been observed that 7-8 seconds are sufficient to provide a score to a given test image, resulting in a duration of around 40 minutes for the whole subjective evaluation process. The total number of observers participating in the subjective study is 21. The latter correspond mainly to naive subjects and few

Fig. 5 Examples of compressed images (at different quality levels)



(a) AE-Factor [5] (b) AE-Hyp-GM [5]

(c) AE-Hyp-GMM [8] (d) JPEG2000

Fig. 6 Image distortions with JPEG2000 and NN based compression methods (at bitrate of 0.2 bits per pixel)

researchers with age ranging from 24 to 44 years. Before starting the subjective tests, a brief introduction of the goal of the study followed by a short training session (of about 5 minutes) is presented to each observer. Note that the images used in the training are different from those of the constructed subjective database. Moreover, the different images are randomly displayed while ensuring that two consecutive test images do not correspond to the same content. For each single test image, the observer provides an opinion score of the picture visual quality by moving a slider along a continuous graded quality scale. A screenshot of the interface used in our experiments is shown in Fig. 7.

Table 1 Summary of publicly available IQA databases with compression distortions. Note that the new dataset is highlighted in bold

Database	No. of reference / compressed images	Resolution	No. of compression methods/levels	NN-based compression methods
LIVE [45]	29 / 344	768×512	2 / 8	0
TID2008 [38]	25 / 400	512×384	2 / 4	0
TID2013 [37]	25 / 500	512×384	2 / 5	0
CSIQ [24]	30 / 300	512×512	2 / 5	0
MICT [9]	14 / 196	768×512	2 / 7	0
MDID [48]	20 / 160	512×384	2 / 4	0
MCL-JCI [18]	50 / 5000	1920×1080	1 / 100	0
FG-IQA [60]	100 / 1200	723×480	1 / 3	0
NNCD-IQA	**16 / 320**	**768×512**	**5 / 4**	**4**

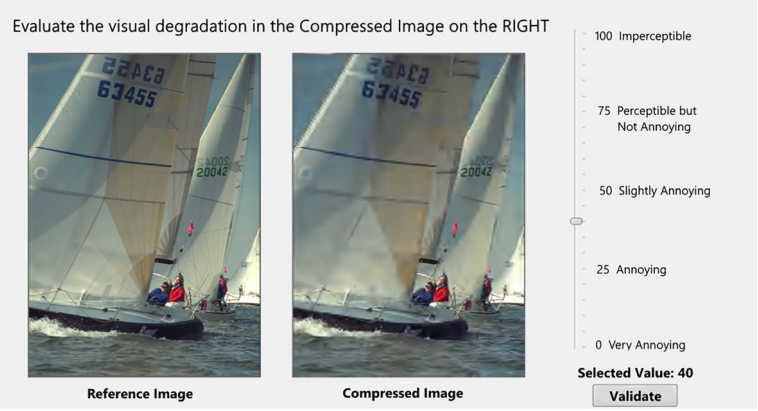


Fig. 7 Subjective experiment interface

3.3 Subjective scores

Let us denote by $S_{i,j}$ the raw score assigned by the i-th observer to the j-th image. Thus, by taking the average of the scores given by the $N = 21$ observers, we obtain the Mean Opinion Score MOS_j for each image with index number j:

$$\mathrm{MOS}_j = \frac{1}{N}\sum_{i=1}^{N} S_{i,j} \tag{2}$$

Once the subjective scores are collected, a screening of observers was firstly performed for outlier detection using the method described in the ITU-R-REC-BT.500-13 [17]. Following this procedure, no outlier was detected.

Figure 8 illustrates the MOS distribution for all the tested images. Firstly, it can be seen that the MOS histogram is close to a uniform distribution. Moreover, there are no values at the extremities of the MOS scale especially at the higher end, depicting that none of the compressed images is perceptually similar to the pristine one.

4 Performance evaluation

In this section, we evaluate the performance of various objective quality assessment metrics when applied to the new neural networks based compressed image database.

4.1 Objective image quality assessment metrics

Unlike the recent works devoted to quality assessment of NN-based compressed images which considered only conventional IQA metrics [52, 55], we propose here to cover a wide range of metrics by investigating machine learning as well as recent deep learning based metrics. In what follows, we will briefly describe these metrics.

Conventional metrics They include some popular metrics often used in IQA as well as some recent ones. In addition to the most commonly used Peak Signal-to-Noise ratio (PSNR) metric, we have considered the following ones:

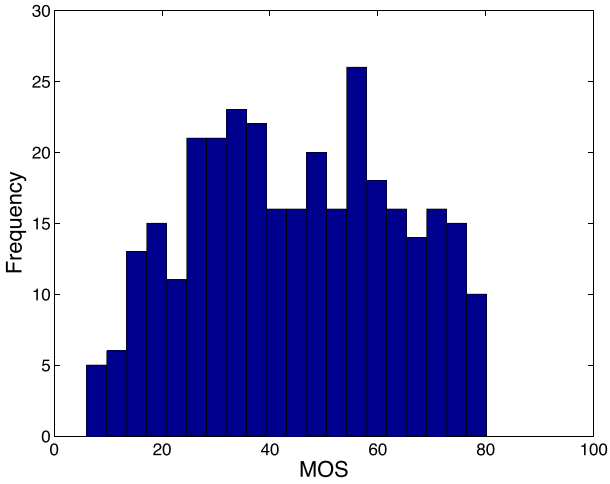


Fig. 8 Distribution of all mean opinion scores (MOS)

- Structure SIMilarity (SSIM) [56]: It is a widely used metric that measures the similarity between a reference image and the tested one based on their structural information.
- Visual Information Fidelity (VIF) [43]: This measure is considered as a Human Visual System (HVS) based method and aims to quantify the loss of image information to the distortion process.
- Gabor Features-based Model (GFM) [36]: It is based on the exploitation of imaginary part of Gabor filter to extract features from luminance components for both the reference and distorted images. The local similarity between these features for the two images along with the chrominance components are then used in the prediction of the final quality score.
- Perceptual image quality assessment using a Normalized Laplacian Pyramid (PIQA-NLP) [23]: It is based on local luminance subtraction and local gain control obtained after applying the Laplacian pyramid decomposition to images.

  It should be noted here that these metrics belong to the class of Full-Reference (FR) metrics which use the reference image to assess the quality of the tested one.

Machine Learning (ML)-based metrics Natural Scene Statistics (NSS) models followed by training are among the ML-based metrics which are often used in IQA studies. These metrics include:

- DIIVINE [35]: It aims to extract statistical features using Discrete Wavelet Transform (DWT). Then, Support Vector Machine (SVM) followed by Support Vector Regression (SVR) stages are used to predict the quality score of the tested image.
- BLIINDS-II [41]: It relies on a statistical model of local discrete cosine transform (DCT) coefficients and employs a probabilistic predictive model to train the features and predict the image quality.
- BRISQUE [33]: Unlike the two previous metrics where statistical features are extracted from DCT and DWT domains, BRISQUE operates in the spatial domain, and then uses SVM and SVR to predict the image quality score.
- NIQE [34]: Based on spatial domain NSS features, it consists in evaluating the image quality based on a multivariate Gaussian (MVG) fitting model. While this method does

not involve SVM and SVR modules, it requires a training on pristine images to generate the parameters of the MVG model.

It should also be noted here that these metrics belong to the class of No-Reference (NR) metrics where the quality of a tested image is evaluated without using the reference image.

Deep Learning (DL)-based metrics Recently, and motivated by the success of neural networks, deep learning based metrics have been developed. Among them, we have considered the following ones.

- Blind Image Evaluator based on a Convolutional Neural Network (BIECON) [21]: A CNN is used to estimate a local quality map followed by one hidden layer to regress the extracted features into a subjective quality score.
- Region-Adaptive Deformable Network (RADN) [46]: It consists of a modified residual block, a patch-level attention one and a reference-oriented deformable convolution block. The latter is performed on different non-overlapping patches and the final quality score is obtained using a weighted average operation.
- Perceptual Image-Error Assessment through Pairwise Preference (PieAPP) [39]: A pairwise-learning approach is developed to predict the perceptual error between an original image and a tested one. A deep CNN is used to train an error estimation function and produce the perceptual error score.
- Deep Image QuAlity Measure for FR IQA (DIQaM) [6]: The method uses a Convolutional Neural Network (CNN) for feature extraction followed by a Fully Connected Neural Network (FCNN) for regression, yielding the quality score prediction.
- Deep Image Quality Assessment Model (DeepQA) [20]: In this method, an error map with the compressed image are fed into a deep convolutional network to generate a sensitivity map. The product of this sensitivity map and the error map are then regressed onto the subjective score.
- Ensemble of Gradient Boosting (EGB) based metric [16]: It is composed of two main blocks. The first one uses VGG16 network to extract feature vector from the reference and distorted images. Then, three gradient boosting regression models are considered to produce the final quality score.
- Deep Image Structure and Texture Similarity (DISTS) index [12]: It firstly consists in generating new representations of the reference and distorted images using CNN. Then, a set of measurements that captures the appearance of different visual textures and structural details are combined to produce an IQA score.

While BIECON is a NR-metric, the remaining DL based metrics are FR ones.

4.2 Analysis of objective quality metrics

Experimental setup While the conventional IQA metrics as well as the ML-based ones can be easily tested since their implementations are publicly available, those of DL-based metrics need more care since only the pre-trained models and the code for the test phase are available. However, such models may not be appropriate for neural networks based compressed image dataset and so should be fine-tuned. In this respect, significant efforts have been made to perform the training phase and obtain the new models. To this end, and for each DL-based metric, we have used the default setting parameters (number of epochs, learning rate, optimizer) provided in its respective reference paper. It is important to note

here that, in addition to the new database and MOS, the obtained trained models will also be provided.

Moreover, for the ML and DL based metrics, 75% of the dataset is used for training while the rest is used for testing. More precisely, since our new database is built from 16 reference images, we have chosen to use 4-fold cross validation. This is achieved by dividing them into four non-overlapping subsets where each test subset is composed of 4 reference images (i.e 80 distorted images) and the remaining 12 reference images (i.e 240 distorted images) are used for training.

Evaluation criteria In order to judge the performance of the objective metrics against our benchmark subjective scores, we have used three different criteria. They are the Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC) and Kendall Rank-Order Correlation Coefficient (KROCC). Before evaluating the correlation coefficients, a five-parameter logistic function, given by (3), is applied to the predicted scores to take intro account for non-linear relation between MOS and the predicted scores

$$f(x) = \beta_1 (\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}}) + \beta_4 x + \beta_5 \tag{3}$$

where β_1, β_2, β_3, β_4 and β_5 are the five model parameters which are obtained by minimizing the mean square error between the MOS and the predicted scores.

Correlation results Tables 2, 3 and 4 provide the correlation values between the objective scores and the subjective ones in terms of PLCC, SROCC and KROCC, respectively. It should be noted that the two best metric values are highlighted in bold.

Thus, different observations can be made from these tables. First, regarding the conventional IQA metrics, it can be seen that the widely used PSNR metric leads to the lowest

Table 2 Pearson Linear Correlation Coefficient of different IQA metrics with MOS

Category	Metric	JPEG2000	AE-Factor [5]	AE-Hyp-GM [5]	AE-Hyp GMM [8]	FCNN-LS [10]	Overall
Classical	**PSNR**	0.7856	0.7943	0.8206	0.8063	0.8478	0.8409
	SSIM [56]	0.8127	0.9009	0.8971	0.9134	0.8444	0.8953
	VIF [43]	0.8482	0.9018	0.8951	0.9016	0.8737	0.9061
	GFM [36]	0.8749	0.9081	0.9292	0.9214	0.8992	0.9096
	PIQA_NLP [23]	0.8206	0.8777	0.8814	0.8939	0.8710	0.9012
ML-based	**BRISQUE** [33]	0.7035	0.8418	0.8299	0.5683	0.7787	0.7862
	DIIVINE [35]	0.5986	0.8126	0.8864	0.7162	0.8301	0.7271
	BLIINDS-II [41]	0.8469	0.7664	0.7973	0.6543	0.7678	0.7313
	NIQE [34]	0.9070	0.7552	0.7908	0.6912	0.5503	0.7712
DL-based	**RADN** [46]	0.8690	0.8688	0.8767	0.9082	0.9395	0.8777
	EGB [16]	0.7499	0.7563	0.7802	0.6712	0.8074	0.7152
	BIECON [21]	0.8180	0.8189	0.8464	0.7566	0.7900	0.8494
	DISTS [12]	0.8696	0.8288	0.8635	0.8615	0.8689	0.8475
	PieApp [39]	**0.9103**	0.8635	0.8875	0.8727	**0.9418**	0.8958
	DIQaM [6]	0.8720	**0.9425**	**0.9388**	**0.9472**	0.9338	**0.9347**
	DeepQA [20]	**0.9361**	**0.9365**	**0.9441**	**0.9635**	**0.9526**	**0.9461**

Table 3 Spearman Rank-order Correlation Coefficient of different IQA metrics with MOS

Category	Metric	JPEG2000	AE-Factor [5]	AE-Hyp-GM [5]	AE-Hyp GMM [8]	FCNN-LS [10]	Overall
Classical	**PSNR**	0.7758	0.8040	0.8058	0.8284	0.8355	0.8322
	SSIM [56]	0.8091	0.8983	0.8903	0.9133	0.8496	0.8908
	VIF [43]	0.8478	0.8938	0.8855	0.9049	0.8609	0.9013
	GFM [36]	0.8824	0.9067	**0.9222**	**0.9278**	0.8987	0.9062
	PIQA_NLP [23]	0.8258	0.8783	0.8750	0.8922	0.8610	0.8966
ML-based	**BRISQUE** [33]	0.6843	0.8272	0.7382	0.5037	0.7603	0.7780
	DIIVINE [35]	0.6320	0.7978	0.8596	0.7088	0.8206	0.7211
	BLIINDS-II [41]	0.8653	0.6794	0.7544	0.5419	0.7809	0.7224
	NIQE [34]	0.9005	0.6875	0.7478	0.6456	0.4963	0.7608
DL-based	**RADN** [46]	0.8487	0.8618	0.8478	0.8794	0.9199	0.8696
	EGB [16]	0.7281	0.7610	0.7066	0.6441	0.7853	0.6923
	BIECON [21]	0.8178	0.7963	0.7868	0.6816	0.7904	0.8372
	DISTS [12]	0.8586	0.8154	0.8221	0.8419	0.8119	0.8412
	PieApp [39]	**0.9042**	0.8699	0.8801	0.8507	**0.9338**	0.8956
	DIQaM [6]	0.8483	**0.9346**	0.9191	0.9184	0.9206	**0.9343**
	DeepQA [20]	**0.9303**	**0.9265**	**0.9522**	**0.9463**	**0.9419**	**0.9451**

correlation values while the recent GFM one presents higher correlations. Moreover, the ML-based metrics have lower correlation values overall. Indeed, while some of them (especially BLIINDS-II and NIQE) can outperform some conventional metrics (like PSNR and

Table 4 Kendall Rank-Order Correlation Coefficient of different IQA metrics with MOS

Category	Metric	JPEG2000	AE-Factor [5]	AE-Hyp-GM [5]	AE-Hyp GMM [8]	FCNN-LS [10]	Overall
Classical	**PSNR**	0.5816	0.6071	0.6114	0.6190	0.6455	0.6384
	SSIM [56]	0.5995	0.7192	0.7067	0.7411	0.6514	0.7091
	VIF [43]	0.6481	0.7093	0.7156	0.7321	0.6723	0.7267
	GFM [36]	0.6988	0.7232	0.7682	0.7589	0.7219	0.7325
	PIQA_NLP [23]	0.6313	0.6954	0.6908	0.7054	0.6683	0.7162
ML-based	**BRISQUE** [33]	0.5317	0.6500	0.5625	0.3917	0.5958	0.5854
	DIIVINE [35]	0.4773	0.6250	0.6917	0.5584	0.6125	0.5447
	BLIINDS-II [41]	0.7030	0.5125	0.5792	0.4042	0.6458	0.5458
	NIQE [34]	0.7612	0.5625	0.6208	0.4917	0.3958	0.5698
DL-based	**RADN** [46]	0.6862	0.7000	0.6875	0.7417	0.7917	0.7051
	EGB [16]	0.5778	0.5750	0.5375	0.4792	0.6042	0.5180
	BIECON [21]	0.6487	0.6208	0.6250	0.5375	0.6417	0.6590
	DISTS [12]	0.7320	0.6500	0.6917	0.6708	0.7000	0.6707
	PieApp [39]	**0.7670**	0.7042	0.7208	0.6750	**0.8250**	0.7231
	DIQaM [6]	0.6737	**0.8000**	**0.7792**	**0.7750**	0.7917	**0.7821**
	DeepQA [20]	**0.8070**	**0.7958**	**0.8458**	**0.8375**	**0.8208**	**0.7975**

SSIM) for JPEG2000 compressed images, ML-based metrics are generally less performant than the conventional metrics for the DL compressed images. This suggests that this category of metrics are not suitable for quality assessment in the context of neural networks based image compression. This may be explained by the fact that these ML based metrics belong to the class of NR IQA methods and so they are trained using only the distorted images. Finally, using the DL based metrics, better correlation values are obtained. For instance, DIQaM and DeepQA metrics outperform all the other metrics and yield the highest correlation values in overall.

In addition, Fig. 9 illustrates the scatter plots of all the considered metrics versus the MOS. Ideally, for a good IQA method, the scatter plot should show good linearity, tight clustering and a relatively uniform density along both axes. Our results show that the conventional GFM metric as well as the two FR DL based metrics DIQaM and DeepQA exhibit high correlations with MOS. Moreover, DIQaM and DeepQA show better linear relationship with respect to the MOS compared to the remaining metrics.

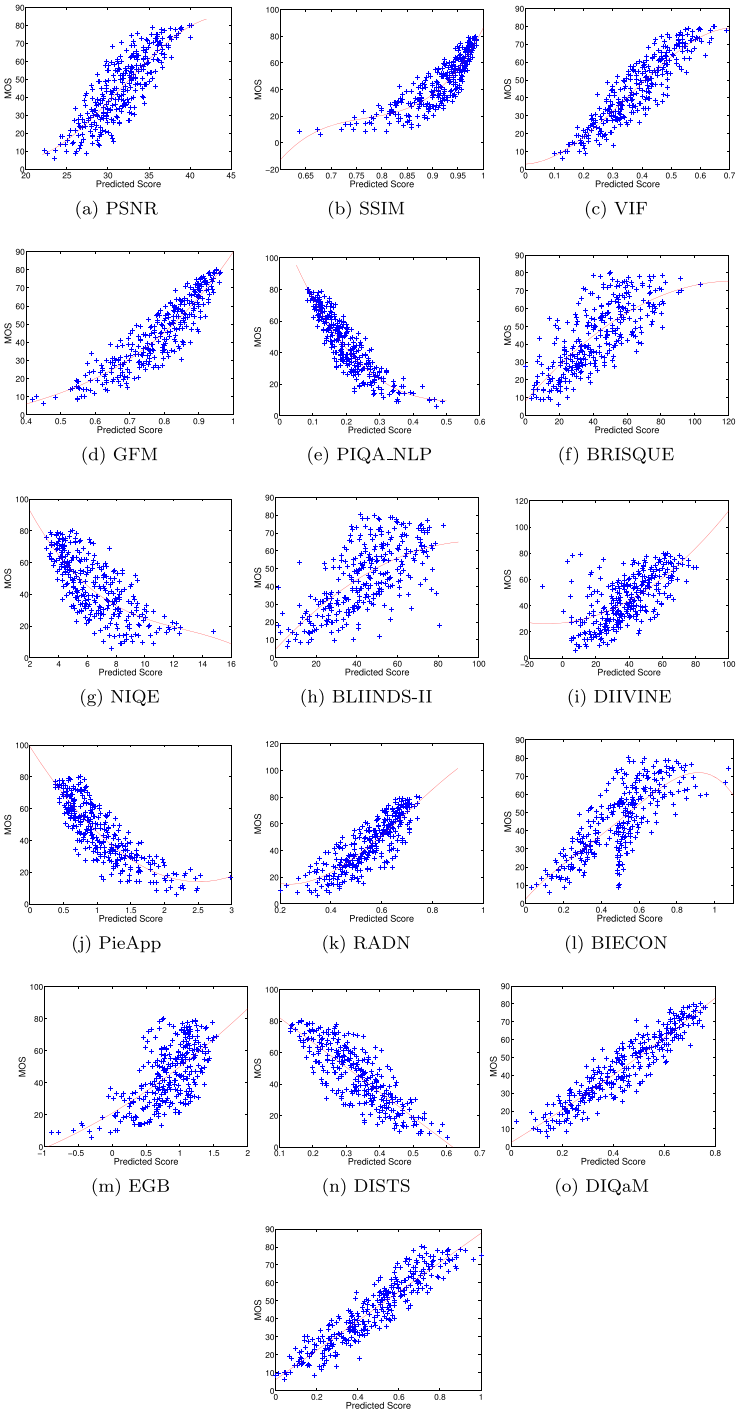
Finally, the best IQA metrics have been retained to evaluate their performance with respect to the different quality levels of distorted images. More precisely, in addition to the PSNR and SSIM metrics often used to assess the quality of deep learning based image compression algorithms, we have considered VIF, GFM, PieAPP, DIQaM and DeepQA. Figure 10 shows the correlation values of these metrics with respect to the four quality levels. These plots confirm the results provided in the previous tables and show that DIQaM and DeepQA outperform the other IQA metrics. Moreover, two main important observations can be made from this figure. First, while some of the classical FR metrics have led to good correlation values in overall (around 0.9 as shown in Tables 2 and 3), it can be seen from Fig. 10 that the DL-based metrics DeepQA, DIQaM and PieApp are more suitable than the classical ones at low birtates (i.e higher distortion levels). Moreover, the curves obtained with DIQaM and DeepQA show small variations of correlation values compared to other curves such as those corresponding to PSNR and SSIM. This indicates that the performance of the above DL based metrics are less sensitive to the coding rate (i.e quality level) and so have the advantage to be more consistent.

4.3 Qualitative results

To confirm again the limitations of standards PSNR and SSIM metrics, often used in the evaluation of deep learning based image compression algorithms, Fig. 11 illustrates some reconstructed images with their associated PSNR, SSIM, GFM, DIQaM and DeepQA metrics as well as the MOS. For example, from the first row of Fig. 11, it can be observed that the AE-Factor [5] method leads to better subjective reconstructed quality compared to JPEG2000. However, the conventional PSNR and SSIM metrics obtained with JPEG2000 are higher than those obtained with the AE-Factor [5] method. Thus, these metrics are not appropriate to show the relevance of the DL based compression method. Unlike PSNR and SSIM, DIQaM and DeepQA show more coherent results well correlated with the human perception.

4.4 Overall discussion

Based on the above results, we summarize here the main observations of the conducted study and analysis. In fact, while the PSNR has poor correlation scores with all compression methods, the SSIM leads to better results. However, the correlation coefficient of SSIM



(a) PSNR (b) SSIM (c) VIF

(d) GFM (e) PIQA_NLP (f) BRISQUE

(g) NIQE (h) BLIINDS-II (i) DIIVINE

(j) PieApp (k) RADN (l) BIECON

(m) EGB (n) DISTS (o) DIQaM

(p) DeepQA

Fig. 9 Scatter plots of the different IQA metrics versus the MOS

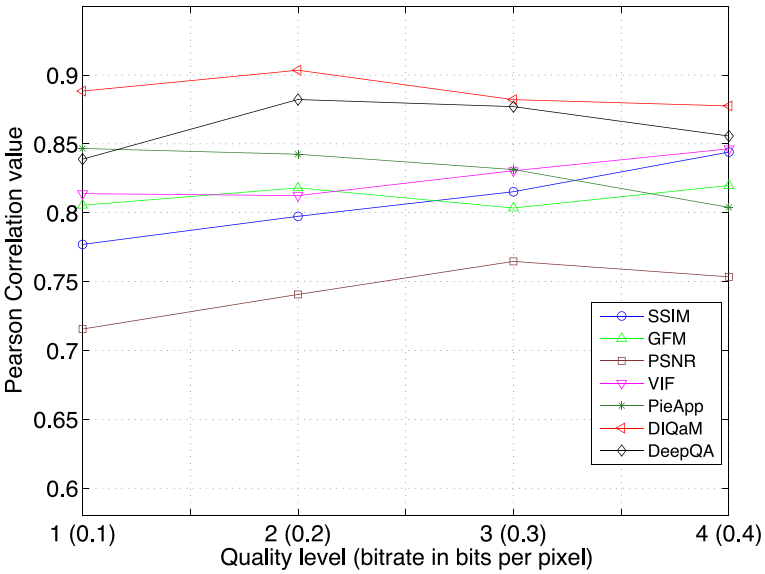


Fig. 10 PLCC Performance of IQA metrics with respect to the compression quality level

obtained with the JPEG2000 coding standard is much lower to those obtained with neural networks-based compressed images. This gap is significantly reduced by the deep learning based metrics (such as DeepQA) which leads to high correlation scores with JPEG2000 standard compression method as well as recent neural networks-based coding methods. This confirms the interest of such deep learning based metric for assessing both traditional as well as NN-based compressed images.

Moreover, it must be emphasized that the interest of deep learning based metrics with respect to the conventional quality measures is much important at low compression quality level. This suggests the strong need for deep learning based metrics for very low bitrate coding application.

5 Conclusion and perspectives

In this paper, a new database of deep learning-based compressed images is built for quality assessment purpose. In this respect, in addition to the JPEG2000 compression standard, four recent neural networks based coding methods have been considered while using different coding rates. Then, after performing the subjective experiments, different categories of IQA metrics, including conventional, ML and DL based metrics, have been evaluated. Our experiments confirm that the standard PSNR and SSIM metrics, often used in the context of image and video coding, are not suitable for neural networks based compressed images, and promising results are obtained with recent DL based metrics like DIQaM and DeepQA.

It is worth pointing out that this new database with the subjective scores will allow to advance the future research works of IQA community. Moreover, the trained models, obtained with our neural networks based compressed database, will be of great interest to

(a) JPEG2000: **PSNR = 30.35 dB, SSIM = 0.533, GFM = 0.81**, DIQaM = 0.39, DeepQA = 0.48, MOS = 43.24

(b) AE-Factor [5]: PSNR = 29.94 dB, SSIM = 0.532, GFM =0.80, **DIQaM=0.43, DeepQA = 0.66, MOS = 54.81**

(c) JPEG2000: **PSNR = 34.50 dB, SSIM = 0.52**, GFM = 0.76, DIQaM = 0.51, DeepQA = 0.57, MOS = 36.57

(d) AE-Hyp GMM [8]: PSNR = 34.47 dB, SSIM = 0.50, **GFM =0.83, DIQaM=0.56, DeepQA = 0.57, MOS = 47.43**

Fig. 11 Examples of reconstructed images with their associated MOS as well as some evaluated IQA metrics

researchers working on the design of neural networks-based image compression methods and requiring to evaluate their compression methods using the DL based metrics retained in this paper. As a future work, it would be interesting to develop a new deep learning based metric for the quality assessment of neural networks-based compressed images.

Funding No funds, grants, or other support was received.

Declarations

Conflict of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agustsson E, Tschannen M, Mentzer F, Timofte R, Luc VG (2019) Generative adversarial networks for extreme learned image compression. In: International conference on learning representations. Louisiana, USA, pp 1–31
2. Ahanonu E, Marcellin M, Bilgin A (2018) Lossless image compression using reversible integer wavelet transforms and convolutional neural networks. In: Data compression conference. UT, USA, p 1
3. Amirkhani D, Bastanfard A (2021) An objective method to evaluate exemplar-based inpainted images quality using Jaccard index. Multimed Tools Appl 80:26199–26212
4. Ballé J, Laparra V, Simoncelli EP (2017) End-to-end optimized image compression. In: International conference on learning representations. Toulon, France, pp 1–27
5. Ballé J, Minnen D, Singh S, Hwang SJ, Johnston N (2018) Variational image compression with a scale hyperprior. In: International conference on learning representations. Vancouver, Canada, pp 1–47
6. Bosse S, Maniry D, Muller KR, Wiegand T, Samek W (2018) Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans Image Process 27(1):206–219
7. Cheng Z, Akyazi P, Sun H, Katto J, Ebrahimi T (2019) Perceptual quality study on deep learning based image compression. In: IEEE International conference on image processing. Taipei, Taiwan, pp 719–723
8. Cheng Z, Sun H, Takeuchi M, Katto J (2020) Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: IEEE International conference on computer vision and pattern recognition, pp 7936–7945
9. Corchs S, Gasparini F, Schettini R (2014) No reference image quality classification for JPEG-distorted images. Digit Signal Process 30:86–100
10. Dardouri T, Kaaniche M, Benazza-Benyahia A, Pesquet JC (2020) Optimized lifting scheme based on a dynamical fully connected network for image coding. In: International conference on image processing. Abu Dhabi, United Arab Emirates, pp 1–5
11. Dardouri T, Kaaniche M, Benazza-Benyahia A, Pesquet JC (2021) Dynamic neural network for lossy-to-lossless image coding. IEEE Trans Image Process 31:569–584
12. Ding K, Ma K, Wang S, Simoncelli EP (2022) Image quality assessment: Unifying structure and texture similarity. IEEE Trans Pattern Anal Mach Intell 44(5):2567–2581
13. Dumas T, Roumy A, Guillemot C (2019) Context-adaptive neural network-based prediction for image compression. IEEE Trans Image Process 29(1):679–693
14. Garcia-Lucas D, Cebrian-Márquez G, Cuenca P (2020) Rate-distortion/complexity analysis of HEVC, VVC and AV1 video codecs. Multimed Tools Appl 79:29621–29638
15. Hajihashemi V, Najafabadi HE, Gharahbagh AA, Leung H, Yousefan M, Tavares JMRS (2021) A novel high-efficiency holography image compression method, based on HEVC, wavelet, and nearest-neighbor interpolation. Multimed Tools Appl 80:31953–31966
16. Hammou D, Fezza SA, Hamidouche W (2021) EGB: Image Quality assessment based on ensemble of gradient boosting. In: IEEE International conference on computer vision and pattern recognition, pp 541–549
17. Methodology for the subjective assessment of the quality of television pictures. Recommendation, ITU-R BT. 500–13 (2012)
18. Jin L, Lin JY, Hu S, Wang H, Wang P, Katsavounidis I, Aaron A, Kuo CCJ (2016) Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis. Electron Imaging 2016(13):1–9
19. Kaaniche M, Benazza-Benyahia A, Pesquet-Popescu B, Pesquet JC (2011) Non separable lifting scheme with adaptive update step for still and stereo image coding. Elsevier Signal Processing: Special issue on Advances in Multirate Filter Bank Structures and Multiscale Representations 91(12):2767–2782
20. Kim J, Lee S (2017) Deep learning of human visual sensitivity in image quality assessment framework. In: IEEE Conference on computer vision and pattern recognition, pp 1676–1684
21. Kim J, Lee S (2017) Fully deep blind image quality predictor. IEEE J Select Top Signal Process 11(1):206–220
22. Korshunov P, Hanhart P, Richter T, Artusi A, Mantiuk R, Ebrahimi T (2015) Subjective quality assessment database of HDR images compressed with JPEG XT. In: International conference on quality of multimedia experience (qoMEX), pp 1–6
23. Laparra V, Ballé J, Berardino A, Simoncelli EP (2016) Perceptual image quality assessment using a normalized laplacian pyramid. Electron Imaging 2016(16):1–6
24. Larson EC, Chandler DM (2010) Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electron Imaging 19(1):011006

25. Lee J, Cho S, Beack SK (2019) Context adaptive entropy model for end-to-end optimized image compression. In: International conference on learning representations. Louisiana, USA, pp 1–20
26. Li J, Li B, Xu J, Xiong R, Gao W (2018) Fully connected network-based intra prediction for image coding. IEEE Trans Image Process 27(7):3236–3247
27. Li M, Zuo W, Gu S, Zhao D, Zhang D (2018) Learning convolutional networks for content-weighted image compression. In: IEEE International conference on computer vision and pattern recognition. UT, USA, pp 3214–3223
28. Liu D, Li Y, Lin J, Li H, Wu F (2020) Deep learning-based video coding: a review and a case study. ACM Comput Surv 53(1):1–35
29. Liu D, Ma H, Xiong Z, Wu F (2018) CNN-Based DCT-like transform for image compression. In: International conference on multimedia modeling. Bangkok, Thailand, pp 61–72
30. Ma H, Liu D, Xiong R, Wu F (2020) iWave: CNN-based wavelet-like transform for image compression. IEEE Trans Multimedia 22(7):1667–1679
31. Ma S, Zhang X, Jia C, Zhao Z, Wang S, Wang S (2020) Image and video compression with neural networks: a review. IEEE Trans Circuits Syst Video Technol 30(6):1683–1698
32. Minnen D, Toderici G, Covell M, Chinen T, Johnston N, Shor J, Hwang SJ, Vincent D, Singh S (2017) Spatially adaptive image compression using a tiled deep network. In: IEEE International conference on image processing. Beijing, China, pp 1–5
33. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessmentin the spatial domain. IEEE Trans Image Process 21(12):4695–4708
34. Mittal A, Soundararajan R, Bovik AC (2013) Making a completely blind image quality analyzer. IEEE Signal Process Lett 20(3):909–912
35. Moorthy AK, Bovik AC (2011) Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Trans Image Process 20(12):3350–3364
36. Ni Z, Zeng H, Ma L, Hou J, Chen J, Ma KK (2018) A gabor feature-based quality assessment model for the screen content images. IEEE Trans Image Process 27(9):4516–4528
37. Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, Vozel B, Chehdi K, Carli M, Battisti F et al (2015) Image database TID2013: peculiarities, results and perspectives. Signal Processing: Image communication 30:57–77
38. Ponomarenko N, Lukin V, Zelensky A, Egiazarian K, Carli M, Battisti F (2009) TID2008-A database for evaluation of full-reference visual quality assessment metrics. Adv Mod Radioelectron 10(4):30–45
39. Prashnani E, Cai H, Mostofi Y, Sen P (2018) PieAPP: Perceptual image-error assessment through pairwise preference. In: IEEE Conference on computer vision and pattern recognition. UT, USA, pp 1808–1817
40. Rippel O, Bourdev L (2017) Real-time adaptive image compression. In: International conference on machine learning. Sydney, Australia, pp 1–9
41. Saad MA, Bovik AC, Charrier C (2012) Blind image quality assessment: a natural scene statistics approach in the DCT domain. IEEE Trans Image Process 21(8):3339–3352
42. Schiopu I, Munteanu A (2018) Macro-pixel prediction based on convolutional neural networks for lossless compression of light field images. In: International conference on image processing. Athens, Greece, pp 445–449
43. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–444
44. Sheikh HR, Bovik AC, Cormack L (2005) No-reference quality assessment using natural scene statistics: JPEG2000. IEEE Trans Image Process 14(11):1918–1927
45. Sheikh HR, Sabir MF, Bovik AC (2006) A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process 15(11):3440–3451
46. Shi S, Bai Q, Cao M, Xia W, Wang J, Chen Y, Yang Y (2018) Region-adaptive deformable network for image quality assessment. In: IEEE Conference on computer vision and pattern recognition, pp 324–333
47. Sun H, Cheng Z, Takeuchi M, Katto J (2020) End-to-end learned image compression with fixed point weight quantization. In: International conference on image processing. Abu Dhabi, United Arab Emirates, pp 1–5
48. Sun W, Zhou F, Liao Q (2017) MDID: A multiply distorted image database for image quality assessment. Pattern Recogn 61:153–168
49. Sweldens W (1996) The lifting scheme: a custom-design construction of biorthogonal wavelets. Appl Comput Harmon Anal 3(2):186–200
50. Taubman D (2000) High performance scalable image compression with EBCOT. IEEE Trans Image Process 9(7):1158–1170
51. Taubman D, Marcellin M (2002) JPEG2000: Image Compression fundamentals, standards and practice. Kluwer academic publishers, norwell, MA USA

52. Testolina M, Upenik E, Ascenso J, Pereira F, Ebrahimi T (2021) Performance evaluation of objective image quality metrics on conventional and learning-based compression artifacts. In: International conference on quality of multimedia experience (qoMEX), pp 1–6
53. Theis L, Shi W, Cunningham A, Huszar F (2017) Lossy image compression with compressive autoencoders. In: International conference on learning representation. Toulon, France, pp 1–19
54. Toderici G, Vincent D, Johnston N, Hwang SJ, Minnen D, Shor J, Covell M (2017) Full resolution image compression with recurrent neural networks. In: IEEE International conference on computer vision and pattern recognition. Hawai, USA, pp 5306–5314
55. Valenzise G, Purica A, Hulusic V, Cagnazzo M (2018) Quality assessment of deep learning-based image compression. In: International workshop on multimedia signal processing. Vancouver, Canada, pp 1–6
56. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
57. Wang Z, Sheikh HR, Bovik AC (2002) No-reference perceptual quality assessment of JPEG compressed images. In: IEEE International conference on image processing. Rochester, USA, pp 477–480
58. Wang Z, Wang W, Li Z, Wang J, Lin W (2012) No-reference image quality assessment for compressed images based on DCT coefficient distribution and PSNR estimation. In: Asia pacific signal and information processing association annual summit and conference. CA, USA, pp 1–4
59. Zhang C, Cheng W, Hirakawa K (2019) Corrupted reference image quality assessment of denoised images. IEEE Trans Image Process 28(4):1732–1747
60. Zhang X, Lin W, Wang S, Liu J, Ma S, Gao W (2018) Fine-grained quality assessment for compressed images. IEEE Trans Image Process 28(3):1163–1175
61. Zhang X, Lin W, Wang S, Liu J, Ma S, Gao W (2019) Fine-grained quality assessment for compressed images. IEEE Trans Image Process 23(3):1163–1175