A genetic programming-based convolutional neural network for image quality evaluations

Kit Yan Chan1 • Hak-Keung Lam2 • Huimin Jiang3

Received: 27 August 2021 / Accepted: 29 March 2022 / Published online: 22 April 2022 © The Author(s) 2022

Abstract
Monitoring the perceptual quality of digital images is fundamentally important since digital image transmissions through the Internet continue to increase exponentially. Many automatic image quality evaluation (IQE) metrics have been developed based on image features correlated to image distortions; however, those metrics are only effective on particular image distortion types. In recent years, convolutional neural network (CNNs) have been developed for IQEs. These CNNs first capture image features from distorted images; image qualities are predicted based on the captured image features. Since the CNN weights are randomly initialized and are updated with respect to a loss function, image features which are strongly correlated to image quality are not guaranteed to be captured. In this paper, a hybrid deep neural network (DNN) is proposed by integrating image quality metrics to capture image features which are correlated to image quality; the approach guarantees that significant image features are included to predict image quality. Also, a tree-based classifier namely geometric semantic genetic programming is proposed to perform the overall predictions by incorporating CNN predictions and image features; the approach is simpler than the fully connected network but is able to model the nonlinear image qualities. The performance of the proposed hybrid DNN is evaluated by an image quality database with 3000 distorted images. The mean correlation achieved by the proposed hybrid DNN is 0.57 which is higher than the other tested methods. Experimental results with the t-test, F-test and Tukey’s range tests show that the proposed hybrid DNN achieves more accurate image predictions with a 99.9% confidence level, compared to the state-of-the-art IQE metrics and the most recently developed CNN for IQEs.

Keywords Wireless imaging • No-reference image quality metric • Objective perceptual image quality • Convolutionary neural networks • Genetic programming

1 Introduction
Digital images uploaded to social media such as facebook, Instagram are essential in our daily life for communications and entertainments. Those images are captured by a camera at the front stage and are watched by human at the end stage in a pipeline. Between the front and end stages, images digitization and compression are performed; processed images are transmitted through a communication channel such as Internet, wireless network; images decompression and reconstruction are performed before watching by human users. Within the pipeline, certain distortions are contaminated to the original images. Coding, decoding, capturing, storing and displaying images generate another distortion. Those processes further degrade the visual qualities. At the end of the pipeline,
human users watch the distorted versions. Quantifying visual quality can be conducted by subjective quality assessments with human judgements. However, subjective quality assessments are time-consuming and impractical for online applications. To automatically evaluate image qualities, image quality evaluation (IQE) models can be used [1, 2]. Those IQE models [3–6] have been developed to engage with many image processing tasks. The predictions of the IQE models can be used to optimize the multi-parameters in the pipeline with many image processing tasks, in order to improve the image qualities.

IQE models can be classified into three domains, namely the full-reference, reduced-reference and no-reference models. Full-reference models use the original images without distortion as a reference in order to determine qualities of distorted images; full-reference models predict image qualities when information of original images is available; reduced-reference models use image features captured from original images and compare those features with those of the distorted images, in order to perform image quality predictions [7, 8]. However, the original images in many applications are not available to compare with distorted images. As an example, an image is transmitted from Smartphone A to Smartphone B; the original image from the Smartphone A is an unknown to Smartphone B, and the original image is not available. Therefore, no-reference models are implemented in many applications since original images are generally not available. No-reference models are commonly used, although IQE is more challenging since no information of the original images is available. They automatically predict images qualities with respect to human perception, when distorted images are only given.

No-reference models can be catalyzed into two types namely the image feature models and machine learning models which are the explicit and implicit typed models, respectively. The image feature model captures an image feature from distorted images which are correlated to subject image qualities evaluated by people. The image feature indicates the quantity of a distortion type, such as blur, blocking, ringing artifacts. Based on the distortion quantity, the image quality can be estimated [9]. However, image feature models cannot be applied for general or multi-purposes; they can be used only when the distortion type is already known or the distortion type of the images is covered by the model. Another type of no-reference models, machine learning models, map a set of image features to the subjective image qualities [3]. They were developed based on a dataset which consists of images features and subjective image qualities scored by human judgement. The image features correlated to perceived image qualities can be extracted using statistical analysis [3]. The machine learning models have been developed based on the approaches of probabilistic prediction models [10, 11], support vector machines [12], fuzzy methods [13, 14], simple neural networks with shallow network architectures [15, 16]. The machine learning models do not rely on explicit information between image features and subjective image qualities. Motivated by the effectiveness of deep neural networks (DNNs) such as convolutional neural networks (CNNs) for object classifications and detections, CNNs have been developed to perform IQEs [5]. More recently Bosses et al. [17] have developed a CNN for predicting image qualities. Better results can be achieved compared with many recently developed no-reference IQE models [4, 5, 11, 18, 19] and other DNNs for IQEs [20, 21].

So far, standard CNNs with cascade convolutional and pooling operations have only been implemented for IQEs [17, 20, 21]. The CNN consists of two main stages namely feature extraction stage and classification stage. In the feature extraction stage, the CNN captures the image features using a set of cascade convolutional and pooling operations from the original distorted images. In the classification stage, those image features are fed into a fully connected neural network in order to quantify image qualities. Generally CNN consists of millions of network weights which are initialized randomly. The initialized weights are updated iteratively by minimizing a loss function which quantifies the differences between real image quality scores and predictions. Therefore, some features are significant to classifications, but some are insignificant; also some image features are strongly correlated to image qualities but are not captured by the cascade convolutional and pooling operations. When those excluded image features are included as an input to the classification stages, the performance of image quality evaluations can be improved.

In this paper, we propose a hybrid DNN in order to improve the feature extraction stage and the classification stage of the currently used CNNs for IQEs. In the feature extraction stage, the proposed hybrid DNN integrates significant image features which are generated by image feature models, and those significant features are fed to the classification stage. The proposed hybrid DNN ensures that significant image features are included to evaluate image qualities. Since significant image features are generally correlated to image qualities, they are likely to improve the IQEs when they are integrated in the CNN for IQE. In the classification stage, we propose another machine learning approach namely geometric semantic genetic programming which uses the DNN predictions and the image features to perform the IQEs. Similar to the fully connected neural networks which are commonly used in the DNNs to the classification stage, geometric semantic genetic programming represents as a tree which consists of branches and
nodes. Small trees are initialized and are grown iteratively to achieve a better classification. The size of the geometric semantic genetic programming is more flexible, unlike the fully connected neural network which has a fixed architecture and consists of a lot of links of which some of the links are not significant for classifications. Unnecessary branches of tree are not included in the geometric semantic genetic programming which is simpler than the cumbersome fully connected neural network.

The performance of the proposed hybrid DNN is evaluated using an image quality database which contains 3000 distorted images [22]. The backbone of the proposed hybrid DNN is implemented with the recently developed CNN [17] and is integrated with several commonly used simple image feature models namely blur [23], blocking [24], and two ringing artifacts [25]. Here we particularly select these four image features models which are computationally simpler than the modern IQE models. If these four computationally simple models are able to assist the hybrid DNN to perform better IQEs, better performance can be achieved when more accurate and computationally complex models are integrated. Experimental results show that the proposed hybrid DNN is capable to perform more accurate IQEs compared to the four image feature models and the recently developed CNN [17] which outperformed many IQA models [4, 5, 11, 18, 19] and neural network models [20, 21]. The rest of the paper is organized as follows: Sect. 2 discusses the commonly used IQE models and the metric for evaluating the models. Section 3 presents the mechanisms of the proposed DNN, and the motivations of why the hybrid DNN are proposed. Section 4 presents how the proposed DNN is implemented, and the performance is evaluated. Experimental results and analysis are also shown. A conclusion is drawn in Sect. 5. A novel hybrid DNN attempts to improve the currently used CNNs of which some significant image features cannot be included for image quality evaluations. This limitation is overcome by incorporating typical image features generated by classical models with the CNN. The hybrid DNN further improves the prediction capabilities of the CNN for IQEs.

2 Image quality evaluation

When an image, \( \hat{I} \in \mathbb{R}^{n \times m} \), with the dimension \( n \times m \), is given, an IQE model, \( F_{\text{IQE}} \), can be used to estimate the image quality, \( q \) for \( \hat{I} \):

\[
q = F_{\text{IQE}}(\hat{I})
\]

To evaluate the performance of \( F_{\text{IQE}} \), we can use a set of \( N_I \) images, \( \hat{I}_i \), with \( i = 1, 2, ..., N_I \), which are contaminated with different types of noise and different levels of distortion. Image qualities are scored by many people, and \( \text{MOS} \) is the corresponding mean opinion score (MOS) for \( \hat{I}_i \). Based on \( \text{MOS}_i \) and \( \hat{I}_i \), Pearson linear correlation, \( r \) in (2), is commonly used to evaluate the performance of \( F_{\text{IQE}} \), which is widely used to evaluate IQE models [2]:

\[
r = \frac{\sum_{i=1}^{N_I} (F_{\text{IQE}}(\hat{I}_i) - \bar{q})(\text{MOS}_i - \bar{\text{MOS}})}{\sqrt{\sum_{i=1}^{N_I} (F_{\text{IQE}}(\hat{I}_i) - \bar{q})^2 \sum_{i=1}^{N_I} (\text{MOS}_i - \bar{\text{MOS}})^2}}
\]

where \( F_{\text{IQE}}(\hat{I}_i) \) is the predicted image quality for the image \( \hat{I}_i \); \( \bar{q} \) and \( \bar{\text{MOS}} \) are the means of all \( F_{\text{IQE}}(\hat{I}_i) \) and all \( \text{MOS}_i \), respectively; \( r \) is the covariance between the predictions of \( F_{\text{IQE}} \) and the actual mean opinion scores divided by the product of the two standard deviations. The performance of \( F_{\text{IQE}} \) is good when the correlation (2) between all \( \hat{I}_i \) and their corresponding \( \text{MOS}_i \) is high. The correlation is strong when \( 0.5 < r \leq 1 \) or \( -1 \leq r < -0.5 \); the correlation is moderate when \( 0.3 < r \leq 0.5 \) or \( -0.5 \leq r < -0.3 \); the correlation is weak when \( 0.1 \leq r < 0.3 \) or \( -0.3 \leq r < -0.1 \); the correlation is very weak when \( -0.1 \leq r \leq 0.1 \).

2.1 Image feature models for IQE

Some distortion types such as blurring, blocking, ringing, are correlated to image qualities. If the distortion type contaminated on the image is available, one can quantify the distortion level using an image feature model which is particularly developed to quantify this distortion type. Based on the quantified distortion, one can estimate the image quality. Given a distorted image, \( F_{\text{IQE}} \) can be developed as an image feature model to quantify its image quality when its distortion type is available. Three image features, namely blur, blocking and ringing are commonly used to quantify image qualities [2].

1. **Blur artifact** is caused by camera movements, long exposure times, object movement or improper camera focus. Blur in an image can be observed by loss of semantic information such as object shapes. The amount of blur is correlated to the loss of spatial detail in an image. Hence, blur in an image can be quantified by the spatial-frequency domain [26]. Blur artifact also affects object edges and fine details on an image. Blur can also be measured by edge spreads [27], edge gradients [28] and edge widths [29] which are correlated to the magnitude of blur.

2. **Blocking artifact** is caused by the block-based image coding in low bit-rate rates, packet loss in image transmissions or block-based image compressions. The blocking artifact can be observed as artificial horizontal and vertical contours which are in block edges. The
blocking artifact can also be caused by quantization of pixels at image blocks, and it causes image discontinuity at block boundaries. To quantify the magnitude of blocking artifact, ones can quantify the edge strength at block boundaries. Those Blocking magnitudes can be quantified by measuring the energy level of the blocky signal [23], detecting step edges with low amplitudes [30] and discrete Fourier transform (DFT)-based measure of blocking artifact [31].

3. **Ringing artifact** is caused by the coarse quantization of discrete wavelet transform or improper truncation of high-frequency components. It can also be generated in high-frequency irregularities during the reconstruction. Ringing distortion can be observed in high contrast edges from smooth texture regions. Ringing distortion can be quantified by edge-detection techniques which measure the overall magnitudes of edge-spread on an image. Those techniques include principal components analysis [32], quantifying pixel and edge distortion [33], and changes in statistic regularities of DWT/DCT coefficients [34].

However, this is ineffective to implement \( F_{IQE} \) based on the quantity of a unique distortion type, since information about the distortion type is not known in most practical applications. When a single distortion type is only used to develop \( F_{IQE} \), other distortion types cannot be quantified. The \( F_{IQE} \) is only sensitive to a particular distortion type, although an image can be contaminated by many distortion types. Therefore, developing \( F_{IQE} \) based on a unique distortion type is impractical.

### 2.2 Machine learning for IQE

To overcome the limitation of image feature models, \( F_{IQE} \) can be developed by machine learning which is trained automatically using distorted images and their MOSs. Machine learning is involved with two main steps [35]. First, image features corresponding to different distortion types are captured from the distorted image; second, \( F_{IQE} \) is developed by learning the relationship between captured image features and MOSs. \( F_{IQE} \) is capable to predict visual image qualities across different distortion types and image contents, when images contaminated with more than one distortion types are used to develop \( F_{IQE} \) [4, 5, 11, 36].

Recent approaches based on the machine learning have been developed to evaluate image qualities when training data with images as input and MOSs as labels is available. Artusi et al. [37] proposed a DNN to evaluate image qualities. However, this approach is not a no-reference image quality metric, where the original undistorted image is used as another input to the DNN. The image quality of the distorted image is compared to that of the original undistorted image. The approach is not practice for many applications such as image transmissions or compressions since the original undistorted images are generally not available. Despite this full-reference image quality metric, no-reference image quality metrics based on DNNs have been developed to perform image quality evaluations for particular object types such as magnetic resonance images [38], sonar images [39], images for liquid crystal displays [40]. Also an approach based on the genetic programming was developed to evaluate qualities of fish images [41]. Those approaches attempt to evaluate image qualities before performing further image processing or object recognitions. However, those approaches were only focused on a particular object type, where image features of those objects are only captured. Those approaches are not developed to evaluate general images which are captured by cameras, where the images are contaminated by common distortion types such as image digitization and compression, internet image transmission, wireless network, image decompression and reconstruction. Although Bi et al. [42] has developed a genetic programming-based model to evaluate qualities for general images, the model is only effective to evaluate image qualities of which the images are distorted with either blurring, lowering contrast or gaussian noise. The model is not developed for multi-image distortion types.

A more robust metric based on convolutional neural network CNN has been used to develop \( F_{IQE} \) where the CNN is trained by a set of distorted images and MOSs of the corresponding distorted images [17]. The performance of CNN has been evaluated by the TID database with 24 image distortion types and better performance can be achieved by the CNN compared with many no-reference IQE models [4, 5, 11, 18, 19] and other deep neural networks [20, 21]. The CNN was developed since the CNNs are capable and effective to perform object classification and detection applications involved with images. A distorted image is the input of the CNN, and the corresponding MOS is the label. The fundamental building blocks of a CNN are illustrated in Fig. 1. The topology of CNN consists of many convolution and pooling layers. The image patch is the inputted to the first layer. The CNN uses multi-layers with pooling and convolution kernels to generate a set of features \( \bar{f}_{CNN} = \{ f_{CNN}^1, f_{CNN}^2, ..., f_{CNN}^N \} \). The final layer uses \( f_{CNN} \) to predict the image quality, \( q_{CNN} \); a fully connected neural network is generally used as the final layer.

CNN weights of the pooling and convolution kernels and those of the fully connected neural network are generally determined by the back-propagation algorithm. First, the CNN weights are randomly initialized. The back-propagation algorithm uses the loss function to determine
the prediction error between the actual MOSs and the estimated image qualities. In each iteration, the CNN weights are updated based on the prediction error. The generalization capability of CNN can be improved through the iterations. When the CNN weights are properly fine-tuned iteratively, a lower prediction error can be obtained. After running the back-propagation algorithm with a certain number of iterations, satisfactory predictions can be achieved by the CNN.

CNN weights for the pooling and convolution kernels are optimized with respect to the loss function and the training data set. The pooling and convolution kernels in CNNs only generates a limited set of features, \( f_{\text{CNN}} \). The features in \( f_{\text{CNN}} \) only cover distortion types in the training dataset. When the image quality is evaluated based on an image feature model, the corresponding distortion type is guaranteed to be quantified on the image. For an example, when a model is particularly developed to quantify image blur, blur distortion can be fully quantify on the image. The approach of the image feature models do not have the limitation of the CNN since the CNN only relies on the training dataset, and some image features cannot be fully quantified.

3 Proposed hybrid deep neural network

In this paper, we propose a hybrid deep neural network namely hybrid DNN which integrates CNN evaluations and image features captured from distortion metrics, in order to predict image qualities. The hybrid DNN attempts to overcome the limitations of the CNN approach and the unique distortion metrics. The hybrid DNN is illustrated in Fig. 2 which uses the geometric semantic genetic programming (GSGP) [43, 44] to predict image qualities based on CNN predictions and image features captured from image features models. Compared to the neural networks and the regression models, the GSGP is proposed since (1)
the GSGP is a heuristic algorithm of which better solutions can be explored when keep running the algorithm. The coefficients of regression model and the weights of neural network are determined based on the least square method and the backpropagation method, respectively, which only reaches local optima. (2) When using the regression model, the model structures including the interaction and orders have to be predefined. Also, in the neural network, the network configuration has to be predefined based on users’ experience. The models generated by the GSGP have more variants, and the model structures can be optimized automatically when keep running the algorithm. Also, the models generated by the GSGP are more transparent, compared to the neural networks.

The hybrid DNN model, $F_{\text{DNNGP}}$ in (3), integrates the image features and CNN prediction by using the GSGP, in order to determine the image quality:

$$q_{\text{DNNGP}} = F_{\text{DNNGP}}(\overline{f_{\text{Dis}}}, q_{\text{CNN}})$$

(3)

where $\overline{f_{\text{Dis}}} = \{f_{\text{Dis}}^1, f_{\text{Dis}}^2, ..., f_{\text{Dis}}^{Nd}\}$ is a set of $Nd$ image features, and $q_{\text{CNN}}$ is the CNN prediction. $F_{\text{DNNGP}}$ attempts to include the image features which are generated by the image feature models and are not included by the CNN.

### 3.1 Algorithmic flow

The DNNGP-Algorithm in Algorithm 1 is proposed to generate $F_{\text{DNNGP}}$. The flow of DNNGP-Algorithm is illustrated in Fig. 3. To develop $F_{\text{DNNGP}}$, a set of $Nd$ IQA samples, $\{\overline{f_{\text{Dis}}}, q_{\text{DNN}}^j, \text{MOS}^j\}$ with $j = 1, 2, ..., Nd$, is collected, where $q_{\text{DNN}}^j$ and $f_{\text{Dis}}^j$ are the $j$th data argument, and MOS$^j$ is the $j$th data label; $q_{\text{DNN}}^j$ and $f_{\text{Dis}}^j$ are the DNN prediction and image features to the $j$th image sample, respectively; MOS$^j$ is the mean opinion score to the $j$th image sample. $F_{\text{DNNGP}}$ attempts to correlate the data argument and the data label. In the DNNGP-Algorithm, the
predefined number of generations is denoted as $\Gamma_{\text{max}}$. A population of $N_{\text{POP}}$ models namely

$$F_{\text{DNNGP}}^i = \{F_{\text{DNNGP},1}^i, F_{\text{DNNGP},2}^i, \ldots, F_{\text{DNNGP},N_{\text{POP}}}^i\}$$

is initialized randomly, where $i$ is the generation number of the genetic process. When $i = 0$, the genetic process is at the first generation. The representation of $F_{\text{DNNGP}}^i$ with $k = 1, 2, \ldots, N_{\text{POP}}$ is formatted as a geometric semantic tree, where the arithmetic operations, $\{+,-,\times,\div\}$ are used as the tree nodes. The image features, $f_{\text{Dis}}^1, f_{\text{Dis}}^2, \ldots, f_{\text{Dis}}^{N_{\text{Di}}}$, and DNN prediction, $q_{\text{DNN}}$, are used as the tree terminals. The nonlinear functions, such as exponential function, sinusoid function, can be used in the nodes. In the proposed DNNGP-Algorithm, the arithmetic operations are used since the execution time of arithmetic operations is shorter than that of the nonlinear functions (Fig. 3).

The proposed DNNGP-Algorithm reproduces new models based on the two geometric semantic operators, namely geometric semantic crossover and geometric semantic mutation [45], where the two geometric semantic operators are discussed in Sect. 3.4. An empty set, namely $\mathcal{S}$, is generated to store the generalization capabilities of all models in the current $F_{\text{DNNGP}}^i$ generation. The generalization capabilities of models are evaluated by the proposed fitness function, namely $\mathcal{F}$, which is discussed in Sect. 3.3. $\mathcal{F}$ evaluates the correlation between the predictions of $F_{\text{DNNGP}}^i$ and the actual MOSs of the images. The commonly used tournament selection method is used to select the good models from the current $i$th generation into the $(i + 1)$th generation.
3.2 Model representation for image quality predictions

The model, which is generated by the DNNGP-Algorithm, predicts the image quality, $y_{DNNGP}$, when the image features, $f_{Dis}^1, f_{Dis}^2, \ldots, f_{Dis}^N_D$, and the CNN prediction, $q_{CNN}$, are given. In the algorithm, the model, $F_{DNNGP,k}^i$, is represented as the following regular expression:

$$y_{DNNGP} = F_{DNNGP,k}^i(f_{Dis}^1, f_{Dis}^2, \ldots, f_{Dis}^N_D, q_{CNN}) = (A \lor string) \cdot a \cdot (A \lor string)$$

where
- $A = \text{string} \cdot a \cdot \text{string}$;
- $a \in \{+, -, \times, \div\}$;
- $\text{string} \in \{f_{Dis}^1, f_{Dis}^2, \ldots, f_{Dis}^N_D, q_{CNN}\}$;
- $\cdot$ is the concatenation;
- $\lor$ is the OR operation; $\ast$ is the Kleen star which is the unary operation.

(4) is incorporated with the model arguments, $f_{Dis}^1, f_{Dis}^2, \ldots, f_{Dis}^N_D$, and $q_{CNN}$, in order to predict $y_{DNNGP}$. The manipulations between the model arguments...

---

Algorithm 1 Pseudocode of the DNNGP-Algorithm

```
Input: $\{f_{Dis}^1, q_{CNN}\}^{N_D}$ with $j = 1, 2, \ldots, N_D$  \>
IQI image samples
Output: $F_{DNNGP}$  \>
IQA model generated by the DNNGP-Algorithm
1: procedure DNNGP-Algorithm($\{f_{Dis}^1, q_{CNN}\}^{N_D}$ with $j = 1, 2, \ldots, N_D$) $\rightarrow M$
2: \hspace{1em} $i = 0$;  \>
$\triangleright$ i is the generation number
3: \hspace{1em} Generate $N_{rep}$ models, $F_{DNNGP,1}^i, F_{DNNGP,2}^i, \ldots, F_{DNNGP,N_{rep}}^i$, and keep all the models, into $F_{DNNGP}^i$;
4: \hspace{1em} Evaluate the capabilities of predicting image qualities for all $F_{DNNGP,k}^i \in F_{DNNGP}^i$ : $f_{fit} = \Psi(f_{DNNGP,k}^i)$, and keep all $f_{fit}$ into $FIT$;
\hspace{1em} $\triangleright$ $\Psi$ is the fitness function defined in Section 3.3 and $f_{fit}$ is the fitness of $F_{DNNGP,k}^i$;
5: \hspace{1em} while $i < \Gamma_{max}$ do
6: \hspace{1em} \hspace{1em} Set $i_{rep} = 0$  \>
$\triangleright$ $i_{rep}$ is the counter for reproductions
7: \hspace{1em} \hspace{1em} while $i_{rep} < N_{rep}$ do
8: \hspace{1em} \hspace{1em} \hspace{1em} if $Rnd() \leq p_c$ then
9: \hspace{1em} \hspace{1em} \hspace{1em} \hspace{1em} Generate a random model $R$;
10: \hspace{1em} \hspace{1em} \hspace{1em} \hspace{1em} Select two models, $F_{DNNGP,k_1}^i$ and $F_{DNNGP,k_2}^i$, randomly from $F_{DNNGP}^i$, with $k_1 \neq k_2 \in \{1, 2, \ldots, N_{rep}\}$;
11: \hspace{1em} \hspace{1em} \hspace{1em} Perform CRO by (7): $CRO(F_{DNNGP,k_1}^i, F_{DNNGP,k_2}^i, R) \rightarrow \text{offspring}$;
12: \hspace{1em} \hspace{1em} \hspace{1em} Store offspring into $F_{DNNGP}^{i+1}$; $i_{rep} + +$;
13: \hspace{1em} \hspace{1em} else
14: \hspace{1em} \hspace{1em} \hspace{1em} if $Rnd() \leq p_m$ then
15: \hspace{1em} \hspace{1em} \hspace{1em} \hspace{1em} Generate two random models $R_1$ and $R_2$;
16: \hspace{1em} \hspace{1em} \hspace{1em} \hspace{1em} Pick a model, $F_{DNNGP,k}^i$, randomly from $F_{DNNGP}^i$ with $k \in \{1, 2, \ldots, N_{rep}\}$;
17: \hspace{1em} \hspace{1em} \hspace{1em} Perform MUT by (8): $MUT(F_{DNNGP,k}^i, R_1, R_2) \rightarrow \text{offspring}$;
18: \hspace{1em} \hspace{1em} \hspace{1em} Store offspring into $F_{DNNGP}^{i+1}$; $i_{rep} + +$;
19: \hspace{1em} \hspace{1em} end if
20: \hspace{1em} \hspace{1em} end if
21: \hspace{1em} \hspace{1em} Create $F_{DNNGP,k}^{i+1}$ by selecting models from $F_{DNNGP}^i$ into $F_{DNNGP}^{i+1}$ using tournament;
22: \hspace{1em} \hspace{1em} Evaluate the capabilities of predicting image qualities for all $F_{DNNGP,k}^{i+1} \in F_{DNNGP}^{i+1}$ : $f_{fit} = \Psi(f_{DNNGP,k}^{i+1})$ and keep $f_{fit}$ into $FIT$;
23: \hspace{1em} \hspace{1em} end while
24: \hspace{1em} \hspace{1em} Put the best model in $F_{DNNGP}^{i+1}$ into $F_{DNNGP}^i$ if the best model is better than $F_{DNNGP}^i$;
25: \hspace{1em} \hspace{1em} $i + +$;
26: \hspace{1em} \hspace{1em} end while
27: return $F_{DNNGP}^i$
28: end procedure
```
are performed by the arithmetical operations, ‘+’, ‘−’, ‘×’ and ‘÷’. The model includes both the CNN prediction and the image features which are correlated to the distorted image. The model overcomes the limitation of the CNN, which may not include those image features correlated to image distortions [17].

Here an example of regular expression is shown in (5) and Fig. 4. In this example, the final prediction is correlated to the CNN prediction, \( q_{\text{CNN}} \), and also it is correlated to the image features, \( f_{10}^{\text{Dis}}, f_8^{\text{Dis}} \) and \( f_2^{\text{Dis}} \). The final predictions are more robust to estimate image qualities where the images are contaminated with different distortion types.

\[
\gamma_{\text{DNNGP}} = \left( \frac{f_2^{\text{Dis}} \times f_{10}^{\text{Dis}}}{q_{\text{CNN}}} + (f_{10}^{\text{Dis}} - f_8^{\text{Dis}}) \right)
\]

\[
= \frac{f_2^{\text{Dis}} \times f_{10}^{\text{Dis}}}{q_{\text{CNN}}} + (f_{10}^{\text{Dis}} - f_8^{\text{Dis}})
\]  

(5)

3.3 Fitness evaluations

The correlation function in (2) is reformulated as the fitness function of the proposed DNNGP-Algorithm, \( \Upsilon_\text{Fitness} \). In order to evaluate the performance of the image quality prediction model, \( F_i^{\text{DNNGP}} \) in (6) is defined as \( \Upsilon_\text{Fitness} \) which is particularly incorporated with the real MOS and the image quality predictions obtained by \( F_i^{\text{DNNGP}} \) with respect to the \( j \)th IQE image sample, \( \{ f_{\text{Dis}}^{\text{i}}, q_{\text{CNN}}^{\text{j}} | \text{MOS} \} \) with \( j = 1, 2, ..., N_D \).

\[
\Upsilon_\text{Fitness}(F_{i}^{\text{DNNGP}}) = \frac{\sum_{j=1}^{N_D} (F_{i}^{\text{DNNGP}}(f_{\text{Dis}}^{\text{j}}, q_{\text{DNN}}^{\text{j}}) - F_{i}^{\text{DNNGP}}(f_{\text{Dis}}^{\text{j}}, q_{\text{DNN}}^{\text{j}}))(\text{MOS}_j - \text{MOS})}{\sqrt{\sum_{j=1}^{N_D} (F_{i}^{\text{DNNGP}}(f_{\text{Dis}}^{\text{j}}, q_{\text{DNN}}^{\text{j}}) - F_{i}^{\text{DNNGP}}(f_{\text{Dis}}^{\text{j}}, q_{\text{DNN}}^{\text{j}}))^2 \sum_{j=1}^{N_D} (\text{MOS}_j - \text{MOS})^2}}
\]

(6)

3.4 Reproductions of image prediction models

After evaluating the fitness of each model \( F_i^{\text{DNNGP}} \) using (6), some models in the current generation are selected to perform reproductions for the new generation. The new models are generated as candidates which have potential to have better capabilities to predict image qualities. The new model is reproduced by incorporating an old model and a random model in the form of (4). The components of the new model are created by exchanging some components of the random model. In the DNNGP-Algorithm, the following two operators, namely geometric semantic crossover and geometric semantic mutation [45], are used to reproduce new models.

a. **Geometric semantic crossover** namely CRO incorporates the components of two current models,
\[ F_{i}^{\text{DNNGP},k_{j}} \text{ and } F_{i}^{\text{DNNGP},k_{k}} \text{ in order to reproduce a new model. Compare to } F_{i}^{\text{DNNGP},k_{1}} \text{ and } F_{i}^{\text{DNNGP},k_{2}}, \text{ the new model has potential to generate better predictions of image qualities. CRO first generates a random model, } R. \text{ CRO then performs a mapping from the 2D-dimension } n \times n \text{ to the 1D-dimension } n. \text{ The new model, } \text{newmodel}, \text{ is reproduced as:}
\]
\[
\text{newmodel} = \text{CRO}(F_{i}^{\text{DNNGP},k_{1}}, F_{i}^{\text{DNNGP},k_{2}}, R) \\
= (F_{i}^{\text{DNNGP},k_{1}} \cdot R) + ((1 - R') \cdot F_{i}^{\text{DNNGP},k_{2}})
\]

(7)

where \( R' \) is given by \( R' = \frac{1}{1 + \text{e}^{-R}} \).

b. \textbf{Geometric semantic mutation} namely MUT creates a new model based on a single model, \( F_{i}^{\text{DNNGP},k_{1}} \). The new model involves new components which have potential to generate better image quality predictions. First, MUT creates two random models, \( R_{1} \) and \( R_{2} \). MUT then performs a map in the 1D-dimension \( n \). A new model, namely \( \text{newmodel} \), is created as:
\[
\text{newmodel} = \text{MUT}(F_{i}^{\text{DNNGP},k_{1}}, R_{1}, R_{2}) \\
= (F_{i}^{\text{DNNGP},k_{1}} + c_{m} \cdot (R_{1}^{'} - R_{2}^{'}))
\]

(8)

where \( c_{m} \) is a constant; \( R_{1}^{'} = \frac{1}{1 + \text{e}^{-R_{1}}} \text{ and } R_{2}^{'} = \frac{1}{1 + \text{e}^{-R_{2}}} \) are the functional values of \( R_{1} \) and \( R_{2} \), respectively.

4 Experimental results and analysis

In this section, we discuss how we implement the proposed hybrid DNN, and also we evaluate the performance of the proposed hybrid DNN. Section 4.1 discusses the database of which we use to evaluate the algorithmic performance. Section 4.2 discusses how the proposed hybrid DNN is implemented. Section 4.3 presents the experimental and comparison results with other methods including the CNN, and the commonly used image feature models for IQEs.

4.1 Image quality assessment database

The prediction capability of the proposed hybrid DNN is evaluated by predicting image qualities of the commonly used image quality assessment database namely TID2013 [22] which is developed for academic research\(^1\). TID2013 database is an extended version of the old version TID2008 database [46]. TID2013 contains 3000 distorted images which were developed based on 25 reference images. The distorted images are contaminated with 25 distortion types in 5 levels. Each distorted image is created by contaminating a reference image with a distortion type in a single distortion level. Those distortions are caused by camera operations, image transmissions, image compressions, and conventional image processing. TID2013 database also contains images with exotic distortions, which do not exist in general applications of image processing but are challenging for image quality prediction algorithms. The image qualities of the distorted images in the TID2013 were evaluated by 985 subjective experiments, which were involved with human observers from 5 countries, Finland, France, Italy, Ukraine and USA. MOSs from 0 to 100 were scored and were rescaled to the range from 1 to 5. The images contaminated with the 25 distortion types are shown in Fig. 5. These 25 distortion types cover many image processing applications. TID2013 is a commonly used image quality assessment database to assess the performance of image quality prediction models [22]. We attempt to evaluate the prediction capability of the proposed hybrid DNN based on (2) which is the correlation between algorithmic predictions and true visual qualities to images.

Despite the TID2013, another image quality assessment database namely LIVE database\(^2\) is used to develop the hybrid DNN which estimates image quality when a distorted image is given [47, 48]. The distorted images in LIVE database are created based on 29 reference images without distortion. These 29 images were contaminated with 5 distortion types namely additive white Gaussian noise, Gaussian blur, a simulated fast fading Rayleigh channel, JP2K compression and JPEG compression. The images were contaminated with many distortion levels. The image qualities of the distorted images were evaluated by more than 25,000 human image quality judgements. Quality difference scores of distorted images were evaluated by comparing the reference images and the distorted images. The quality difference scores are in the range between 0 to 100. When the quality difference score is lower, the image quality is better. Hence, the quality difference scores are different to the MOS. MOS is the higher the better, while the quality difference score is the lower the better.

To evaluate the performance of the proposed DNNGPGP-Algorithm, the proposed DNNGPGP-Algorithm uses the LIVE database to train the hybrid DNN, and TID2013 database is used to validate the generalization capabilities of the trained hybrid DNN. Since the number of distortion types in TID2013 database is 24 and the number of distortion types in LIVE database is 6, the number of distortion types of TID2013 is larger than that of LIVE. When the LIVE database is used to train the hybrid DNN, some

\(^{1}\) TID2013 can be downloaded from \text{http://www.ponomarenko.info/tid2013.htm}.

\(^{2}\) LIVE can be downloaded from \text{https://live.ece.utexas.edu/research/Quality/subjective.htm}.
image distortions including in the TID2013 are not covered. Performing these experiments attempt to evaluate whether the performance of the proposed hybrid DNN is better than the CNN [17] and the other compared algorithms [23–25]. The proposed hybrid DNN uses classical image feature models to capture distortion types which are correlated to the image qualities. The approach overcomes the limitations of the CNN of which the training is only relied on the training database. If the image distortion types are not included in the training database, the CNN is unlikely to estimate image qualities of those distorted images since the CNN is fully relied on training samples. Also the four image feature models are only developed for a particularly distortion type. We attempt to verify whether the performance difference between the proposed hybrid DNN and the other tested methods is significant.

4.2 Algorithmic implementation of proposed hybrid DNN

The implementation of the proposed hybrid DNN in Fig. 2 consists of three main components, GSGP-Algorithm, CNN and image feature models which are illustrated in Fig. 6. The GSGP-Algorithm [44] is implemented since the algorithm is computationally simpler than the classical genetic programming models and is more capable to generate nonlinear models compared to the statistical regression. The CNN is implemented with a convolutional neural
network namely Boses-CNN [17] which is recently developed for predicting image qualities; experimental results showed that the Boses-CNN is significantly better than the commonly used image quality metrics. In the proposed hybrid DNN, the image feature models are selected as Blocking artifact, Blur artifact, Ringing artifact (edge magnitude) and Ringing artifact (edge gradient) [49] which are correlated to the image qualities and have small computational costs. Here we do intentionally not implement the most modern and computationally complex models. We particularly select these four state-of-the-art models which are simple and are not most modern. If these four computationally simple models are able to perform better image predictions and achieve improvement, much better performance can be achieved by the proposed hybrid DNN when the more modern and computationally complex models are integrated.

The implementation details of the GSGP-Algorithm are given in Sect. 4.2.1. The implementations of CNN and distortion metrics are described in Sects. 4.2.2 and 4.2.3, respectively.

4.2.1 GSGP-Algorithm

The GSGP-Algorithm integrates the four image feature models and the Boses-CNN, in order to develop the hybrid DNN for IQEs. The following algorithmic parameters are implemented in the GSGP-Algorithm: (Population size) = 200; (Maximum number of generations) = 50; (Generation gap) = 0.9; (Crossover probability) = 0.5; (Mutation probability) = 0.5; Tournament selection with the size of 4 is used. These parameters were determined by experiments where the hybrid DNN with convincing prediction capabilities can be generated by the GSGP-Algorithm.

The image quality samples in the TID database were used to generate the proposed hybrid DNN. Since the TID database is created by 25 reference images, 25-fold cross was used to evaluate the prediction capability of the proposed hybrid DNN. The image quality samples were divided into 25 folds, where each fold contains the contaminated images which were distorted on one of the reference images. In each validation, 24 out of the 25 folds were used to develop the model, and the remaining data fold were used to validate the prediction capability of the model. The performance of the models was evaluated based on the Pearson linear correlation in (2) which indicates the correlation between the actual MOSs and the model predictions. The proposed GSGP-Algorithm was coded based on the C++ framework for the public source of GSGP³. The GSGP-Algorithm was implemented by a P510 Xeon E5-1630 v4 machine with 32 GB memory and two xGTX1080 GPU cards.

4.2.2 CNN

Boses-CNN [17] was implemented on the proposed hybrid DNN. The backbone of Boses-CNN is based on the architecture of VGGnet which is embedded with cascaded convolution kernels with the size of 3×3 [50]. The VGGnet consists of eight convolution layers and four maxpool layers. Since the VGGnet was only developed for images with the size of 224×224 pixels, extra image resizing was performed between the images and the input of the VGGnet by two convolution layers and a single maxpool layer. The output of VGGnet is a set of image features which are fed into two cascaded fully connected neural networks and a single pooling layer. In order to avoid

³ The C++ framework for geometric semantic genetic programming can be downloaded from http://gsgp.sourceforge.net/
overtraining the Boses-CNNs, the approach of dropout regularization is used to determine the network parameters. Based on [17], the total number of parameters in Boses-CNN is about 5.2 million. The detailed description of the Boses-CNN can be referred to [17]. Also the implementable Boses-CNNs are available for public use; they have been trained by two image quality databases namely TID and LIVE, and they can be downloaded4.

The proposed hybrid model is integrated with the Boses-CNN which is trained by the LIVE database. In [17], experiments have been conducted to evaluate the performance of the Boses-CNNs for predicting the image qualities. Their image prediction performance was evaluated based on the TID2013 database. The experiments attempt to evaluate the prediction capabilities of the Boses-CNNs while the training images are not duplicate to the test images. Experimental results showed that more accurate image quality predictions can be achieved by the Boses-CNNs compared to other recently developed no-reference image quality metrics namely Saad-BLIINDS-II [11], Mittal-DIVINE [4], Mittal-BRISQUE [5], Ghadiyaram-NIQE [18], Ye-CORN-A [19] and Zhang-SOM [51], and other deep neural networks namely Kang-CNN [20] and Kim-BIECON [21].

4.2.3 Image feature models

The following image features Fig. 75, namely blocking, blur and ringing artifacts which commonly exist in images, are integrated to the proposed hybrid DNN. CNN-Bosses in the hybrid DNN is trained by the LIVE database which contains the distortion types of blur and image compressions. Although those distortion types are similar to those of the following image features, CNN-Bosses is only trained by a limited number of images contaminated by those distortion types. The CNN-Bosses may not cover all the levels of those distortions. The proposed hybrid DNN attempts to integrate CNN predictions and the four image feature models, in order to compensate those uncovered levels. It attempts to improve the performance of image predictions.

- **Blocking artifacts**: The metric of block artifacts was developed by Wang et al. [23]. The blocking artifacts are usually caused by image compressions such as JPEG, JPEG2000. The artifacts appear continuously at block boundaries and are caused individually by the quantization of those blocks. Three quantity parameters indicate the blocking measures. The first parameter quantifies the blocks by measuring overall differences between block boundaries. The other two parameters quantify image blurs in horizontal and vertical directions based on differences between blocks and a zero-crossing rate, respectively. A high order polynomial function is used to combine the three parameters in order to quantify the blocking artifacts.

- **Blur artifacts**: The metric of blur artifacts was developed by Marziliano et al. [24]. Blur reduces spatial detail and object shapes in an image. Blur is caused by the overall increasing of edge smoothness or reducing of edge sharpness. The metric quantifies the blur by measuring the widths of vertical edges in the image. Detections of vertical edges are applied since the required computations are less, comparing to the inclusion of both vertical and horizontal edges. The metric first uses the Sobel filter to detect vertical edges, where the Sobel filter is commonly used for edge detections [52]. Blur measures are quantified in each detected edge. The overall blur is quantified based on the average of all those blur measures.

- **Ringing (edge magnitude and gradient)** - The metric of ringing artifacts was developed by Saha et al. [25]. The ringing artifacts appear on high contrast edges which are in smooth textures. The ringing artifacts are caused by image processing or transmissions, where high-frequency components exist in the images. In the metric of ringing artifacts, two parameters are quantified. First, Sobel edge detection [52] is used to generate the edge image from the original one. Based on the edge image, the first parameter is measured as the overall edge magnitude. The second parameter indicates the overall edge gradients in both vertical and horizontal directions. These two parameters are used in order to quantify the image activity which is correlated to ringing artifacts (Fig. 7).

4.3 Experimental results and statistical tests

Since the GSGP-Algorithm is a heuristic algorithm, the random operations such as the population initialization, mutation and crossover are involved. Different hybrid DNNs are generated in different runs, although the same parameters are used in the GSGP-Algorithm. Therefore, the GSGP-Algorithm was run for 30 times. 25-fold cross validations were conducted in which each trial was corresponded to a reference image. The averages for 30 runs for each reference image were recorded and are shown in Table 1. Section 4.3.1 shows the correlation performance for the proposed hybrid DNN and the other tested methods. Section 4.3.2 shows the statistical tests including the t-test, F-test and Tueky’s range test which are achieved by the proposed hybrid DNN.

4 The implementable versions of Boses-CNNs can be downloaded from https://github.com/dmaniry/deepIQA.

5 The four image feature models in MATLAB scripts were developed by the second author of the paper [13]. Please contact either the first or second authors of the paper if the source codes are requested.
4.3.1 Correlation performance

Table 1 also shows the cross validation results for the Boses-CNN and the proposed hybrid DNNs. For Image 1, the models were developed based on the distorted images and the MOSs of Image 2–25. After the models were developed, they were used to predict the MOSs of the untrained distorted images. The Pearson linear correlations of the predicted image qualities and the actual qualities are calculated in order to evaluate whether the predictions of the models are linearly correlated. If the correlation is high, the model is capable to explain the linear relationship between the actual MOSs and the predictions. The table also shows the correlations for the four commonly used image quality metrics for blocking artifacts, blur artifacts, and ringing artifacts (edge magnitude and gradient). In the first row for Image 1, the means of the Pearson linear correlations obtained by the proposed hybrid DNNs are generally larger than those obtained by Boses-CNN which are larger than the four commonly used image quality metrics. Similar results can be found on the second main row regarding for Image 2 to Image 25. The means of Pearson linear correlations obtained by the proposed hybrid DNNs are higher than those of the Boses-CNN which are larger than those of the four commonly used image quality metrics. Therefore, the proposed hybrid DNNs are generally better than the Boses-CNN models which can generate more accurate image quality predictions than the four commonly used image quality metrics.

All the four commonly used image quality metrics achieved poorer results compared to the proposed hybrid DNNs and the Boses-CNN, the four metrics only address one of the four artifacts namely blocking artifacts, blur artifacts and two ringing artifacts (edge magnitude and gradient), respectively. If the image quality metric is used to measure the images with other distortion types, those distortions cannot be quantified. For example, the metric for quantifying blur artifacts is not effective to quantify the images with blocking artifacts. If this metric is used to measure images contaminated with blocking artifacts, predictions of images qualities are likely to be poor. Therefore, solely quantifying one distortion type is not enough to quantify the overall image distortion since the image database is involved with distorted images which are contaminated by other distortion types. Since both Boses-CNN and the proposed hybrid DNNs are developed by images contaminated with many distortion types, the overall performance of image quality predictions is generally better than that of the four image quality metrics.

When the performance of the Boses-CNN and the proposed hybrid DNNs is compared, the correlations in Table 1 show that the proposed hybrid DNN outperforms the Boses-CNN. The image quality predictions of the proposed hybrid DNN are integrated with both the predictions from the Boses-CNN and the image quality metrics. The exact quantity of a distortion type can be evaluated by the corresponding distortion metric. The image quality metric does not have the limitation of the Boses-CNN which is trained by a limited number of...
features from a database. When images contaminated with some distortion types are not included for training, quality predictions for those image types can be poorer. Hence, the overall predictions of Boses-CNN are also likely to be poorer. The proposed hybrid DNN is generally better than the Boses-CNN.

### 4.3.2 Statistical tests

To further compare the prediction performance of the four commonly image quality metrics, Boses-CNN and the proposed hybrid DNN, the t-test [53] was used. The t-test evaluated the significance of the hypothesis that the Pearson linear correlations achieved by the proposed hybrid DNN are larger than those of the four commonly image

#### Table 1 Correlation to MOS using different methods

| Image number | Block artifact metric | Blur artifact metric | Ringing artifact metric (edge magnitude) | Ringing artifact metric (edge gradient) | Boses-CNN | Hybrid DNN |
|--------------|-----------------------|----------------------|------------------------------------------|----------------------------------------|-----------|------------|
| Image 1      | 0.173                 | 0.187                | 0.218                                    | 0.199                                  | 0.547     | 0.603      |
| Image 2      | 0.172                 | 0.154                | 0.159                                    | 0.050                                  | 0.574     | 0.694      |
| Image 3      | 0.280                 | 0.046                | 0.273                                    | 0.283                                  | 0.522     | 0.625      |
| Image 4      | 0.088                 | 0.115                | 0.135                                    | 0.112                                  | 0.491     | 0.609      |
| Image 5      | 0.250                 | 0.038                | 0.253                                    | 0.252                                  | 0.512     | 0.579      |
| Image 6      | 0.410                 | 0.126                | 0.225                                    | 0.319                                  | 0.628     | 0.684      |
| Image 7      | 0.304                 | 0.084                | 0.115                                    | 0.019                                  | 0.447     | 0.503      |
| Image 8      | 0.064                 | 0.105                | 0.005                                    | 0.038                                  | 0.378     | 0.499      |
| Image 9      | 0.034                 | 0.012                | 0.198                                    | 0.061                                  | 0.374     | 0.521      |
| Image 10     | 0.458                 | 0.151                | 0.336                                    | 0.405                                  | 0.446     | 0.576      |
| Image 11     | 0.305                 | 0.075                | 0.347                                    | 0.329                                  | 0.354     | 0.381      |
| Image 12     | 0.277                 | 0.063                | 0.058                                    | 0.191                                  | 0.496     | 0.578      |
| Image 13     | 0.152                 | 0.392                | 0.205                                    | 0.1803                                 | 0.337     | 0.330      |
| Image 14     | 0.396                 | 0.136                | 0.362                                    | 0.377                                  | 0.557     | 0.654      |
| Image 15     | 0.293                 | 0.033                | 0.326                                    | 0.312                                  | 0.471     | 0.547      |
| Image 16     | 0.018                 | 0.338                | 0.133                                    | 0.069                                  | 0.566     | 0.653      |
| Image 17     | 0.353                 | 0.018                | 0.252                                    | 0.306                                  | 0.397     | 0.517      |
| Image 18     | 0.236                 | 0.063                | 0.222                                    | 0.230                                  | 0.493     | 0.585      |
| Image 19     | 0.195                 | 0.104                | 0.100                                    | 0.146                                  | 0.480     | 0.584      |
| Image 20     | 0.369                 | 0.004                | 0.353                                    | 0.362                                  | 0.482     | 0.577      |
| Image 21     | 0.327                 | 0.170                | 0.282                                    | 0.305                                  | 0.460     | 0.541      |
| Image 22     | 0.130                 | 0.140                | 0.219                                    | 0.180                                  | 0.489     | 0.597      |
| Image 23     | 0.272                 | 0.067                | 0.244                                    | 0.263                                  | 0.457     | 0.560      |
| Image 24     | 0.217                 | 0.083                | 0.172                                    | 0.195                                  | 0.555     | 0.631      |
| Image 25     | 0.151                 | 0.107                | 0.137                                    | 0.144                                  | 0.536     | 0.690      |
| Mean         | 0.237                 | 0.112                | 0.213                                    | 0.213                                  | 0.482     | 0.573      |

#### Table 2 Correlation to MOS using different methods

| Comparisons | Block artifact metric vs Hybrid DNN | Blur artifact metric vs Hybrid DNN | Ringing artifact metric (edge magnitude) vs Hybrid DNN | Ringing artifact metric (edge gradient) vs Hybrid DNN | Boses-CNN vs Hybrid DNN |
|-------------|-------------------------------------|-----------------------------------|-------------------------------------------------------|-------------------------------------------------------|------------------------|
| t values    | 11.460                              | 18.373                            | 14.112                                                | 12.656                                                | 4.025                  |
| p values    | $0.000 \times 10^{-3}$               | $0.000 \times 10^{-3}$            | $0.000 \times 10^{-3}$                                | $0.000 \times 10^{-3}$                                | $0.201 \times 10^{-3}$ |
quality metrics and the Boses-CNN. In the t-distribution table, the hypothesis has 99.9% confidence level when the t-value is higher than 3.99. Hence, the Pearson linear correlations of the hybrid DNN are higher than those of the other methods with a 99.9% confidence level. Table 2 shows the t-values for the proposed hybrid DNN, compared to the four commonly used image quality metrics and the Boses-CNN. The second row shows the t-values for the Pearson linear correlations. The t-distribution table shows that the Pearson linear correlations achieved by the proposed hybrid DNN are significantly larger than those of the other five tested methods with a very high confidence level of 99.9%, since the t-values are higher than 3.9. Therefore, the hybrid DNN is able to obtain better Pearson linear correlations in terms of image quality predictions than the other five tested methods with a very high confidence level. Also we have conducted the F-test for the Pearson linear correlations achieved by the proposed hybrid DNN which was developed by the proposed DNNGA-Algorithm, compared to the other five methods which were all developed to predict the image qualities for the distorted images. Table 2 shows the P-values which are all less than 0.0000.
for the hybrid DNN compared to the four image quality metrics. Also the P-value is 0.0002 for the hybrid DNN compared to the Boses-CNN. Therefore, we have significant confidence to claim that the hybrid DNN generated by the DNNGA-Algorithm is able to generate better results than the other four tested methods for predicting image qualities.

Since the p-values are all zeros or close to zero, the post-hoc analysis namely Tukey’s range test is used to further validate whether the performance of the proposed hybrid DNN generated by the proposed DNNGA-Algorithm and the other five tested methods is significant difference. The results in Fig. 8a–e show that the Pearson linear correlation mean achieved by the proposed hybrid DNNs is better than those achieved by the four image quality metrics and the Boses-CNN, respectively. These results further validate the performance of the proposed hybrid DNN.

These validations further demonstrate that the proposed hybrid DNN which is a novel version of the CNN and is better than the recently developed Boses-CNN. The prediction of hybrid DNN is integrated with the Boses-CNN prediction and the four image quality metrics. The integration is performed by the proposed DNNGA-Algorithm which uses the five predictions in order to achieve better predictions with better Pearson linear correlations. Therefore, better results can be achieved by the proposed hybrid DNN.

These experimental results demonstrate that the proposed hybrid-DNN achieves higher correlations to the real MOSs in the TID database, compared to the five tested methods including block artifact metric, blur artifact metric, ringing artifact metric, ringing artifact metric, Boses-CNN. Also the validation results are shown by the T-test, F-test and Tukey’s range test. They showed that the proposed approach is able to achieve significantly more accurate quality predictions, compared to the five tested methods. These results show that the proposed hybrid-DNN outperforms the CNN-Bosses of which those distortion types are not included for training. Better results are achieved by the hybrid-DNN since the hybrid-DNN integrates with image features generated by the four image feature models, block artifact metric, blur artifact metric, ringing artifact metric, and ringing artifact metric. Some distortion types, which have not been covered by the CNN-Bosses, are included in the hybrid-DNN. Also the hybrid-DNN is better than the individual image feature model which is only robust on a single distortion type. Therefore, better results can be achieved by the proposed hybrid-DNN.

5 Conclusion

In this paper, a novel hybrid DNN was developed to perform automatic IQEs. The proposed hybrid DNN consists of two stages, namely feature extraction stage and classification stage. In the feature extraction stage, the proposed approach integrates image features captured from IQE models in order to predict image qualities. The proposed approach ensures that significant features correlated to image qualities are integrated by image feature models. It overcomes the limitation of the recently developed CNN that image features are only captured randomly, and significant one cannot be guaranteed to be included. In the classification stage, the tree-based model namely geometric semantic genetic programming integrates image features to perform the final image quality predictions. The approach is simpler than the cumbersome fully connected neural networks.

The performance of the proposed approach was evaluated by the TID image quality database which is commonly used to evaluate the performance of image quality metrics. The database consists of 3000 distorted images which were contaminated with 25 distortion types in 5 levels. We have compared the proposed approach with the four state-of-art IQE metrics and a powerful CNN which outperformed many IQE models. The mean correlation achieved by the proposed hybrid DNN is 0.57 which is higher than the tested methods including the state-of-the-art IQE metrics and the powerful CNN for IQEs. The correlation results in terms of t-test, F-test and Tukey’s range tests showed that the proposed approach is able to achieve significantly more accurate quality predictions with a 99.9% confidence level, compared to the tested methods.

In the future, we will incorporate the proposed hybrid DNN with more modern models for IQEs. This is expected that further improvement can be achieved but long execution time is required when the more computationally complex models are integrated. We will find the tradeoff between the prediction accuracy and computational time.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. We fully understand and fully comply the ethical responsibilities and rules for Authors for this journal.
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Kamble V, Bhurchandi K (2015) No-reference image quality assessment algorithms: a survey. Optik 126(11–12):1090–1097
2. Xu SP, Jiang SL, Min WD (2017) No-reference/blind image quality assessment: a survey. IETE Tech Rev 34(3):223–245
3. Bovik AC (2013) Automatic prediction of image quality and video quality. Proc IEEE 101(9):2008–2024
4. Mittal A, Moorby AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
5. Mittal A, Soundararajan R, Bovik AC (2013) Making a ‘completely blind’ image quality analyzer. IEEE Signal Process Lett 20(3):209–212
6. Reibman AR, Suthaharan S (2008) A no-reference spatial aliasing measure for digital image resizing. In: IEEE international conference on image processing, pp 1184–1187
7. Qian J, Wu D, Li L, Cheng D, Wang X (2014) Image quality assessment based on multi-scale representation of structure. Digit Signal Process 33:123–133
8. Guha T, Nezhadary E, Ward RK (2014) Sparse representation-based image quality assessment. Signal Process: Image Commun 29(10):1138–1148
9. Corchs S, Gasparini F, Schettini R (2014) No reference image quality classification for jpeg-distorted images. Digit Signal Process 30:86–100
10. Saad MA, Bovik AC, Charrrier C (2010) A dct statistics-based blind image quality index. IEEE Signal Process Lett 17(6):583–586
11. Saad MA, Bovik AC, Charrrier C (2012) Blind image quality assessment: a natural scene statistics approach in the dct domain. IEEE Trans Image Process 21(8):3339–3352
12. Chapelle O, Haffner P, Vapnik V (1999) Support vector machines for histogram-based image classification. IEEE Trans Neural Netw 10(5):1055–1064
13. Chan KY, Engelke U (2015) Fuzzy regression for perceptual image quality assessment. Eng Appl Artif Intell 43:102–110
14. Chan KY, Lam HK, Yiu KFC, Dillon TS (2017) A flexible fuzzy regression for addressing uncertainty on aesthetic quality assessments. IEEE Trans Syst Man Cybern: Systems 47(12):2363–2377
15. Babu RV, Sureshb S, Perkis A (2007) No-reference jpeg-image quality assessment using gap-rbf. Signal Process 87(6):1493–1503
16. Engelke U, Zepernick HJ (2007) An artificial neural network for quality assessment in wireless imaging based on extraction of structural information. In: IEEE international conference on acoustics, speech, and signal processing, pp 1249–1252
17. Bosse S, Maniry D, Muller TWKR, Samek W (2018) Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans Image Process 27(1):206–219
18. Ghadiyaram D, Bovik AC (2016) Massive online crowdsourced study of subjective and objective picture quality. IEEE Trans Image Process 25(1):372–397
19. Ye P, Kumar J, Kang L, Doermann D (2012) Unsupervised feature learning framework for no-reference image quality assessment. In: IEEE proceedings of conference on computer vision, pp 1098–1105
20. Kang L, Ye P, Li Y, Doermann D (2014) Convolutional neural networks for no-reference image quality assessment. In: Proceedings of IEEE international conference of conference on computer vision and pattern recognition, IEEE, pp 1733–1740
21. Kim J, Lee S (2017) Fully deep blind image quality predictor. IEEE J Sel Topics Signal Process 11(1):206–220
22. Ponomarenko N, Jinh L, Jeremeiev O, Lukina V, Egiazarian K, Astolab J, Vozecl B, Chehdic K, Carlid M, Battistid F, Kuo CCJ (2015) Image database tid2013: peculiarities, resultsand perspectives. Signal Process: Image Commun 30:57–77
23. Wang Z, Sheikh HR, Bovik AC (2002) No reference perceptual quality assessment of jpeg compressed images. In: IEEE Proceedings of the 15th international conference on image processing, pp 477–480
24. Moraglio A, Krawiec K, Johnson CG (2002) A no-reference perceptual blur metric. In: Proceedings of IEEE international conference of image processing, pp 57–60
25. Saha S, Vermuri R (2000) An analysis on the effect of image features on lossy coding performance. IEEE Signal Process Lett 7(5):104–107
26. Guan J, Zhang W, Gu J, Ren H (2015) No-reference blur assessment based on edge modeling. J Vis Commun Image Represent 29:1–7
27. Sang Q, Qi H, Wua X, Li C, Bovik AC (2014) Noreference image blur index based on singular value curve. J Vis Commun Image Represent 25(7):1625–1630
28. Varadarajan S, Karam L (2008) An improved perceptionbased no reference objective image sharpness metric using iterative edge refinement. In: IEEE Proceedings of the 15th international conference on image processing, pp 401–404
29. Ong E, Lin W, Lu Z (2003) A no-reference quality metric for measuring image blur. In: Proceedings of the 7th international symposium on signal processing and its applications, Vol. 1, pp 469–472
30. Meesters L, Martens JB (2002) A single-ended blockiness measuring image blur index based on singular value curve. J Vis Commun Image Represent 25(7):1625–1630
31. Chen CH, Bloom JA (2010) A blind reference-free blockiness measure. In: Proceedings of the 11th pacific rim conference on multimedia, pp 112–123
32. Tong H, Li M, Zhang H, Zhang C (2004) No-reference quality assessment for jpeg2000 compressed images. In: Proceedings of international conference on image processing, Vol. 5, pp 3539–3542
33. Sazzad ZMP, Kawayo Y, Horita Y (2007) Spatial features based no reference image quality assessment for jpeg2000. In: IEEE Proceedings of international conference on image processing, pp 517–520
34. Zhang J, Ong SH, Le TM (2011) Kurtosis-based no-reference quality assessment of jpeg2000 images. Signal Process: Image Commun 26(1):13–23
35. Yuan Y, Guo Q, Lu X (2015) Image quality assessment: a sparse learning way. Neurocomputing 159:227–241
36. Moorby AK, Bovik AC (2011) Blind image quality assessment: from natural scene statistics to perceptual quality. IEEE Trans Image Process 20(12):3350–3364
37. Artusi A, Banterle F, Carrara F, Moreo A (2020) Efficient evaluation of image quality via deep-learning approximation of perceptual metrics. IEEE Trans Image Process 29:1843–1854
38. Sujit SJ, Coronado I, Kamali A, Narayana PA, Gabr RE (2019) Automated image quality evaluation of structural brain mri using an ensemble of deep learning networks. J Magn Reson Imaging 50(4):1260–1267
39. Zhang HQ, Li S, Li DH (2020) A dual-path deep neural network for sonar image quality evaluation. In: IEEE Proceedings of conference on networking, sensing and control
40. Campo FAD, Villeas OOV, Sanchez VGC, Dominguez HDJO, Nandayapa M (2021) Radial basis function neural network for the evaluation of image color quality shown on liquid crystal displays. IEEE Access 9:21694–21707
41. Yan ZC, Bi Y, Xue B, Zhang MJ (2021) Automatically extracting features using genetic programming for low-quality fish image classification. In: IEEE congress on evolutionary computation, pp 2015–2022
42. Bi Y, Xue B, Zhang MJ (2021) Genetic programming-based discriminative feature learning for low-quality image classification. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2021.3049778
43. Castelli M, Vanneschi L, Silva S (2014) Semantic search-based genetic programming and the effect of intron deletion. IEEE Trans Cybern 44(1):103–113
44. Castelli M, Silva S, Vanneschi L (2015) A c++ framework for geometric semantic genetic programming. Genet Program Evolvable Mach 16:73–81
45. Moraglio A, Krawiec K, Johnson CG (2012) Geometric semantic genetic programming. In: Parallel Problem Solving From Nature, Vol. 7491 of Lecture notes in computer science, pp 21–31
46. Ponomarenko N, Lakin V, Zelensky A, Egiazarian K, Carli M, Battisti F (2009) Tid 2008 - a database for evaluation of full-reference visual quality assessment metrics. Adv Mod Radioelectron 10:30–45
47. Sheikh H, Wang Z, Cormack L, Bovik A (2005) Live image quality assessment database. https://live.ece.utexas.edu/research/Quality/subjective.htm
48. 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
49. Engelkea U, Maederb A, Zepernic HJ (2012) Human observer confidence in image quality assessment. IEEE Trans Image Process 27:935–947
50. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. In: IEEE Proceedings of imagenet challenge, IEEE, pp 1–10
51. Zhang P, Zhou W, Wu L, Li H (2015) Som: Semantic obviousness metric for image quality assessment. In: IEEE Proceedings of conference on computer vision and pattern recognition, pp 2394–2402
52. Lee SJ, Haralick RM, Shapiro LG (1987) Morphologic edge detection. IEEE J Robot Autom RA- 3(2):142–156
53. Box GEP, Hunter J, Hunter W (2005) Statistics for Experiments: Design, Innovation, and Discovery, 2nd edn. Wiley, New York

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.