EGB: Image Quality Assessment based on Ensemble of Gradient Boosting

Dounia Hammou[1], Sid Ahmed Fezza[1], Wassim Hamidouche[2]
[1]National Institute of Telecommunications and ICT, Oran, Algeria
[2]Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
dhammou@inttic.dz, sfezza@inttic.dz, wassim.hamidouche@insa-rennes.fr

Abstract

Multimedia services are constantly trying to deliver better image quality to users. To meet this need, they must have an effective and reliable tool to assess the perceptual image quality. This is particularly true for image restoration (IR) algorithms, where the image quality assessment (IQA) metric plays a key role in the development of these latter. For instance, the recent advances in IR algorithms, which are mainly due to the adoption of generative adversarial network (GAN)-based methods, have clearly shown the need for a reliable IQA metric highly correlated with human judgment. In this paper, we propose an ensemble of gradient boosting (EGB) metric based on selected features similarity and ensemble learning. First, we analyzed the capability of features extracted by different layers of deep convolutional neural network (CNN) to characterize the perceptual quality distance between the reference and distorted/processed images. We observed that a subset of these layers is more relevant to the IQA task. Accordingly, we exploited these selected layers to compute the features similarity, which are then used as input to a regression network to predict the image quality score. The regression network consists of three gradient boosting regression models that are combined to derive the final quality score. Experiments were performed on the perceptual image processing algorithms (PIPAL) dataset, which has been used in the NTIRE 2021 perceptual image quality assessment challenge. The results show that the proposed metric significantly outperforms the state-of-the-art methods for IQA task. The source code is available at: https://github.com/Dounia18/EGB.

1. Introduction

In the last few years, driven by the continuous technological advances in computer and internet, multimedia technologies have evolved considerably at a faster rate than ever. Multimedia data has become integral parts of our daily life, this is particularly true with the explosion in the number of smartphones and similar devices. These allow users to take pictures anytime, anywhere and share them on social networks. However, digital images before they reach the end user go through a series of processing steps that can introduce different distortions, reducing the perceptual quality. Therefore, it is important to have an effective tool to reliably assess, control and ensure high quality.

Image quality assessment (IQA) can be carried-out in two ways, either subjectively or objectively. Subjective image quality assessment consists in asking a group of people to give their opinion about the quality of each image in a given dataset. The evaluation can take several forms, the observers may be asked to rate the quality on either a category scale containing one of the five categories (excellent, good, fair, poor, or bad) or on a continuous scale to avoid quantization artifacts. They can also be asked to compare two or more images. Subjective IQA is the most accurate and reliable way for assessing image quality, since human observers are in most cases the ultimate end users. Subjective experiments are controlled and influenced by many factors which make them time consuming, expensive and impractical in real world applications. To avoid such limitations, objective quality assessment metrics aim to define quality measures that predict quality scores highly correlated with those given by a set of observers.

Until recently, the task of assessing image quality was addressed in two separate steps: 1) designing methods for extracting relevant visual features, then 2) designing regression algorithms that derives a quality score from a set of visual features. The first step consists in using mathematical formulas to extract visual characteristics such as contrast, edge, texture, etc. The second step consists in learning the relationship between the features and the quality score. Even if the extracted features are fed into the learning algorithm, the above two steps are still independent and require some hand-engineering. They are then called hand-crafted metrics in the literature [45], [29], [26].

In recent years, convolutional neural networks (CNNs) [35] have shown to outperform traditional approaches and have been extremely useful and successful in many computer vision tasks. Motivated by the great success of CNNs

on numerous applications, recent IQA works have adopted the use of CNNs on the image quality prediction problem which have proven to be more efficient than hand-crafted metrics [23], [4], [20].

Despite these valuable works, the existing metrics show high performance on IQA datasets including classical distortions, such as blur, noise, compression, etc. While until now, there is no commonly accepted metric that ensures reliable evaluation of image restoration (IR) algorithms [10, 16]. This is even more remarkable for images resulting from the generative adversarial network (GAN)-based IR methods [15, 37], where there is a large inconsistency between the subjective and objective quality assessments.

Consequently, in this paper, we propose an IQA metric for perceptual IR algorithms based on the similarities between the deep features and ensemble learning, called ensemble of gradient boosting (EGB) metric. Using a pre-trained CNN model for the object classification task, we performed an analysis on features extracted from different layers of deep CNN to determine which ones are the most relevant for IQA task. On the basis of this study, the features of the layers providing a high correlation with human assessment of image quality are considered. Specifically, we measure the distance between the selected features extracted from both the reference and distorted images, and use this distance as an input to the regression network to provide the image quality score. The regression network consists of three gradient boosting regression models that are combined to derive the final quality score.

The rest of this paper is organized as follows. Section 2 first presents a review of IQA metrics. Section 3 describes our EGB-based IQA metric. The performance of the proposed metric is assessed in terms of correlation with subjective scores in Section 4. Finally, Section 5 concludes this paper.

2. Related Works

In this section, we will give a brief overview on image quality metrics (IQMs). For more exhaustive description of the cited metrics and other non-cited metrics, the reader may also refer to the following overview papers [40], [28], [21].

The most widely used traditional IQMs are mean squared error (MSE) and peak signal to noise ratio (PSNR). MSE is a signal based metric that represents the cumulative squared error between the distorted and the reference image. PSNR is the most popular pixel based metric which represents a measure of the peak error. These mathematical based metrics exhibit weak performance and have been widely criticized for not involving any perceptual information [39]. Other metrics were then developed to include human visual system (HVS) properties to IQA such as structural similarity index (SSIM) [41]. It was then extended to multi-scale structural similarity index (MS-SSIM) [42], which is based on modeling of image luminance, contrast and structure at multiple scales. Image fidelity criterion (IFC) [34] and visual information fidelity (VIF) [33] are natural scene statistics (NSS) based metrics that model natural images in the wavelet domain using Gaussian scale mixtures (GSMs). Natural image quality elevator (NIQE) [27] is based on constructing a collection of quality aware features and fitting them to a multivariate Gaussian (MVG) model. The quality aware features are derived from NSS model, then quality is expressed as the distance between MVG fit of the NSS features extracted from the test image and those extracted from the corpus of natural images. Most of the above described image quality prediction models work by performing regression on the extracted visual feature vector to obtain quality scores. These models usually rely on machine learning algorithms such as support vector regressions (SVRs), random forests (RFs), general regression neural networks (GRNNs), etc.

The first application of a CNN model to the problem of IQA was investigated by the authors in [18]. They applied a CNN to the no reference image quality assessment (NR-IQA) framework by regressing images on the target subjective scores without hand-crafted features. The CNN directly learns discriminant features from normalized raw image pixels to achieve much better performance than hand-crafted metrics. Zhang *et al.* [48], showed that networks trained to solve challenging visual prediction and modeling tasks end up learning a representation of the world that matches perceptual judgments well, meaning that features extracted from deep architectures are as good as hand-crafted features.

CNN-based IQA models can be divided into two categories : 1) CNNs used for features extraction and regression. 2) CNNs used to only extract features. CNN-based models of the first category are trained in an end-to-end manner to predict quality scores. Like in [36], where authors explored few different classifier architectures for IQA task. The weights of the baseline CNN are initialized by training on the ImageNet dataset [8], and then an end-to-end training on quality assessment is performed. CNN-based models of the second category are followed by a SVR to perform regression on the features. Authors in [2], proposed a deep blind image quality (DeepBIQ) model that consists of a CNN originally trained to discriminate visual categories, which is then fine-tuned for category-based IQA. The CNN is used to extract features that are then fed to a SVR model to predict image quality scores.

3. Proposed Approach

The use of convolutional layers of deep neural network (DNN) for features extraction has demonstrated its effectiveness in various applications. For instance, it has been shown that CNN features can be utilized in the applicati-

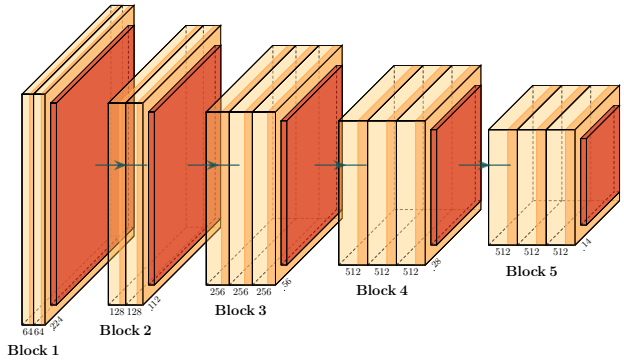

Figure 1. VGG16 network architecture [35].

ons related to human perception, such as IQA [1, 12, 43] and visual similarity [17, 48]. However, most of the works mainly rely on the transfer learning technique, which focuses on transferring the knowledge across domains, i.e., from the source to the target domains [49]. This is usually done using the last deep layers of CNN, without looking which layer is most relevant for the new task.

Unlike using systematically deep features extracted from fully-connected (FC) layers of DNN, the intuition behind our approach is that intermediate convolutional layers may be better candidates for quantifying perceptual image quality. In particular, the layers related to the attributes of human perception. In order to evaluate that, we firstly conducted a study on the applicability of deep CNN features extracted from different layers of CNN model to quantify the image quality. Then, based on this study, we selected the most relevant features, which are subsequently used as input into the regression network that maps them to the subjective image quality scores.

3.1. Deep Feature Space Analysis

A deep CNN consists of multiple layers of neurons, each layer provides different types of features. For instance, the early layers provide low-level features (edges, gradients), while the latest deeper layers provide high-level features (semantic information, texture). Thus, we investigated the best discriminating features of these layers for IQA task. For this purpose, we observed the differences between the reference and distorted/processed images as they flow through the CNN model.

In this work, we adopted the well-known VGG16 network [35] pre-trained on the ImageNet classification dataset [8]. This network architecture contains five blocks, as shown in Figure 1. Each block consists of two to three convolution layers with a different number of filters.

As illustrated in Figure 2, for each layer l of the VGG16 model, a distance between the reference and distorted/processed images is computed using extracted feature maps (also called activation maps) as follows:

$$d(l) = \frac{1}{C_l H_l W_l} \sum_{c,h,w} \left| f_l^{(ref)}(c,h,w) - f_l^{(dist)}(c,h,w) \right|, \tag{1}$$

where $f_l^{(ref)}$ and $f_l^{(dist)}$ are the feature maps of reference and distorted/processed images, respectively, extracted from the layer l. H_l, W_l and C_l are the height, width, and number of channels of the feature map of the layer l, respectively. The reference images and their processed versions are taken from the public training set of perceptual image processing algorithms (PIPAL) database [16] described in Section 4.1.

Finally, for each layer l, the correlation between the calculated distances and the subjective scores is computed using SROCC, as shown in Figure 3. This allows us to observe if a particular layer of the model provides more relevant feature maps for IQA task. As seen in Figure 2, some feature maps are more altered by blur distortion than others, which confirms our intuition.

By using a specific threshold, we found that the best correlations are achieved with three intermediate layers, which are the block 4 convolution layer 2, block 4 convolution layer 3 and block 5 convolution layer 1. These latter are the most altered by distortions and processing operations, and therefore we considered them to be the best candidates for the evaluation of the perceptual image restoration.

3.2. Proposed EGB Metric

The proposed EGB metric is composed of two blocks, namely feature extraction and quality estimation, as illustrated in Figure 4 and explained in the following sections.

3.2.1 Feature Extraction

First, the reference I_{ref} and distorted I_{dist} images are fed to the VGG16 network to derive the feature vector $\boldsymbol{v}$. We extract the features from the three selected intermediate convolution layers, namely block 4 convolution layer 2, block 4 convolution layer 3 and block 5 convolution layer 1, denoted as l_1, l_2 and l_3, respectively. Then, global average-pooling (GAP) is performed to reduce the feature maps into manageable size. The obtained three feature vectors $f_{l_1}^{(k)}$, $f_{l_2}^{(k)}$ and $f_{l_3}^{(k)}$ with $k \in \{ref, dist\}$, are fused using concatenation $f^{(k)} = \texttt{concat}(f_{l_1}^{(k)}, f_{l_2}^{(k)}, f_{l_3}^{(k)})$.

Finally, to provide the feature vector $\boldsymbol{v}$ to the quality estimation part of the framework, we calculate the absolute difference between the extracted features of the reference I_{ref} and distorted I_{dist} images as follows:

$$\boldsymbol{v} = \mid f^{(ref)} - f^{(dist)} \mid . \tag{2}$$

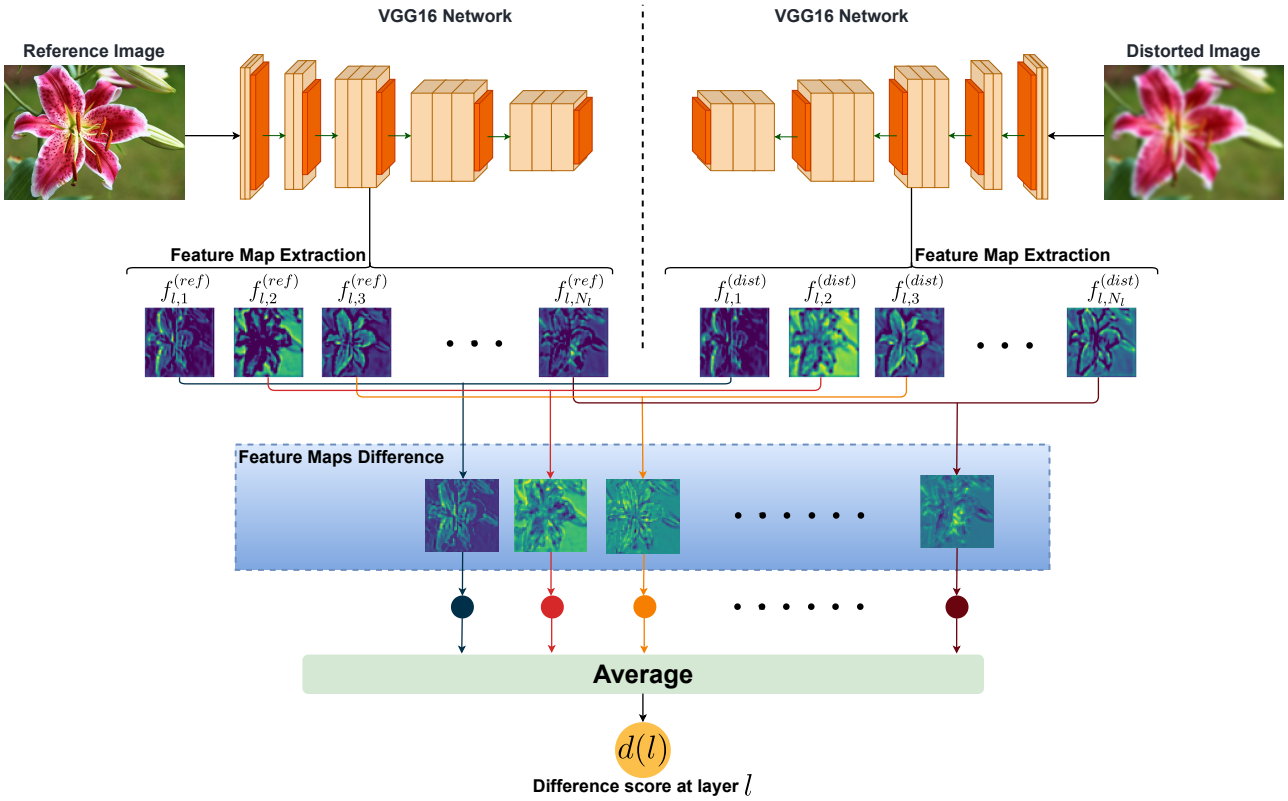

Figure 2. An example of calculation of the distance score $d(l)$, at each layer l of the VGG16 network, between the reference and distorted/processed images. $f_{l,i}^{(ref)}$ and $f_{l,i}^{(dist)}$ are the ith feature maps of the reference and distorted/processed images extracted from the layer l, respectively, and N_l is the number of feature map at layer l.

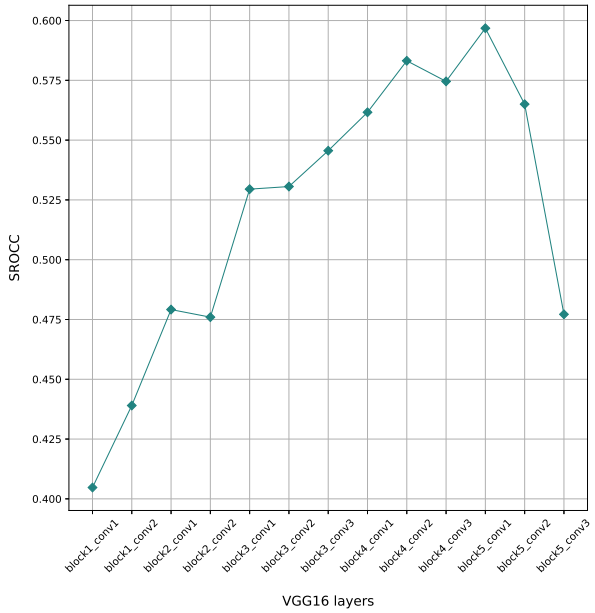

Figure 3. The SROCC correlation between the distance scores and the subjective scores for each layer l of the VGG16 network.

3.2.2 Quality Estimation

The feature vector $\boldsymbol{v}$ is then fed into the three $K = 3$ regression models, which form an ensemble of diverse predictors, to be regressed to predict an image quality score. The goal of using an ensemble methods is to combine their predictions in order to improve generalizability and robustness over a single predictor.

Thus, the quality estimation part of the framework is composed of three $K = 3$ regression models (base models) build independently. Based on the feature vector $\boldsymbol{v}$, each of the regression model predicts an image quality score $\hat{q}_i$ of an image index i as follows:

$$\hat{q}_{i,k} = h_{\theta_k}(\boldsymbol{v}), \tag{3}$$

where $h_{\theta_k}(\boldsymbol{v})$ is the parametric function of the regression model k with training parameters θ_k.

As regression models, we considered three gradient boosting regression models, which are XGBoost (eXtreme Gradient Boosting) [6], LightGBM (Light Gradient Boosting Machine) [19] and CatBoost (categorical boosting) [11]. The three models are ensemble models using decision trees.

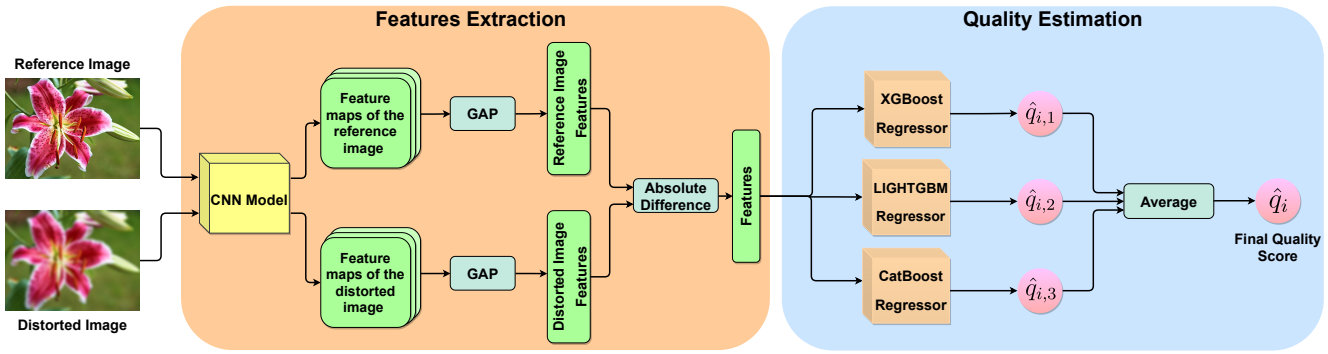

Figure 4. The framework of our proposed EGB IQA metric.

Finally, the predictions of the three trained regression models are averaged to obtain the final prediction quality score

$$\hat{q}_i = \sum_{k=1}^{K} \omega_k \, \hat{q}_{i,k}, \tag{4}$$

where ω_k is the weight assigned to each regression model h_{θ_k}. Thus, by averaging the predictions of the three sub-optimal models, we obtain a more effective model which offers a high correlation with human image quality judgment.

4. Experiments

4.1. Datasets

PIPAL dataset[1] [16] contains 250 reference images of size 288×288 that were distorted using 40 distortion types, resulting in 23k distorted images that was assessed by more than one million human ratings. The set of distortions that was considered can be divided into four sub-types. The first sub-type of distortion includes some traditional distortions (e.g., blur, noise, and compression), which are commune to other datasets. The second sub-type includes the result of super-resolution (SR) methods, including interpolation method, traditional methods, SR with kernel mismatch, PSNR-oriented methods, GAN-based methods. The third sub-type contains the results of several denoising algorithms, including both hand-crafted and deep learning-based methods. Finally, the four sub-type contains the restoration results of images with multi-distortion.

Actually this dataset is unique because it contains a lot of new distortion types, for instance, the results of diverse IR algorithms, especially the GAN-base methods.

During the NTIRE 2021 challenge [14], the dataset was divided into: 1) training data with 200 reference images and 23,200 distorted images, where mean opinion score (MOS) for each distorted image was provided, 2) validation data including 25 reference images with 1000 distorted images, 3) test data containing 25 reference images and 1650 distorted images. For the validation and test sets, the ground truth label, i.e., MOS score, is not provided.

[1]This dateset has been used in the NTIRE 2021 Perceptual Image Quality Assessment Challenge [14].

4.2. Hyperparameters

Each gradient boosting regression model was trained using the following specific parameters:

- The XGBoost regressor was optimized using a 0.09 learning rate. The regression model has a max depth of 5 and min child weight of 4, a 0.7 subsample and 0.01 colsample bytree. To avoid overfitting, a λ and α factors were used with values of 0.02 and 0.01, respectively.

- The LightGBM regressor was optimized using a 0.1 learning rate. The regression model has a 0.7 factor for feature fraction and 140 trees, each tree with 32 leaves and the minimum data in each leaf was 15. To avoid overfitting, λ_{L1} and λ_{L2} factors were used with values of 0.02 and 0.08, respectively.

- The CatBoost regressor was optimized using a 0.1 learning rate. The regression model has 256 estimators, each estimator with a depth factor of 6 and a subsample of 1. To avoid overfitting, L2 regularization was used with a factor of 10.

4.3. Experimental Setup

We compared our EGB metric with state-of-the-art IQA methods, including hand-crafted- and deep learning-based metrics: peak signal to noise ratio (PSNR), noise quality measure (NQM) [7], universal quality index (UQI) [38], structural similarity index measure (SSIM) [41], multiscale structural similarity index measure (MS-SSIM) [42], information fidelity criterion (IFC) [34], visual information fidelity (VIF) [33], visible signal-to-noise ratio (vsnr) [5], Riesz-transform based Feature SIMilarity metric (RFSIM) [46], gradient similarity (GSM) [24], spectral residual based similarity (SRSIM) [44], feature similarity index (FSIM)

Table 1. Evaluation of the three gradient boosting regression models and their combination on the validation set of PIPAL dataset.

Model	Main Score ↑	SROCC ↑	PLCC ↑
XGBoost	1.5119	0.7572	0.7546
LIGHTGBM	1.5252	0.7626	0.7626
CatBoost	1.5402	0.7713	0.7689
Average	**1.5511**	**0.7758**	**0.7752**

[47] and its color version, visual saliency-based index (VSI) [45], most apparent distortion (MAD) [22], natural image auality evaluator (NIQE) [27], MA [25], PI [3], learned perceptual image patch similarity (LPIPS)-Alex [48], LPIPS-VGG [48], perceptual image-error assessment through pairwise preference (PieAPP) [31], weighted average deep image quality measure (WaDIQaM) [4], deep image structure and texture similarity (DISTS) [9] and space warping difference (SWD) [13].

Two indices are used as performance criteria, namely the Spearman rank order correlation coefficient (SROCC), to evaluate the prediction monotonicity, and the Pearson linear correlation coefficient (PLCC) after non-linear regression, to quantify the deviation between the predicted and subjective scores. For the two correlations, values that are close to 1 indicate good performance in terms of correlation with human judgment. The main score is simply the addition between the results of the two correlations, as was done during the challenge.

4.4. Results

First, the three gradient boosting regression models are evaluated separately on the validation set of PIPAL dataset, in addition to their combination using the average, as reported in Table 1. For the average, as preliminary results, we only considered the case of uniformly weighted models, i.e., $\forall i \in \{1, \cdots, K\},\ \omega_i = 1/K$.

From this table, we can see that the CatBoost model obtained the best results compared to the remaining two models. However, the combination of the three models using the average provides a performance improvement over a single model, thus illustrating the add value of the proposed ensemble gradient boosting. The behavior of the proposed metric is illustrated via the scatter distributions provided in Figure 5, in which each data point represents an image from the validation set of PIPAL dataset. This figure shows the scatter distributions of MOS scores versus the predicted ones obtained by the EGB metric, in addition to the non-linear fitted curve obtained by interpolating the objective scores. This figure shows a good agreement between subjective and objective scores and the scatter distributions are consistent, which demonstrates a good prediction of the human judgment.

The results on validation and test sets of PIPAL dataset are reported in Tables 2 and 3, respectively. From both ta-

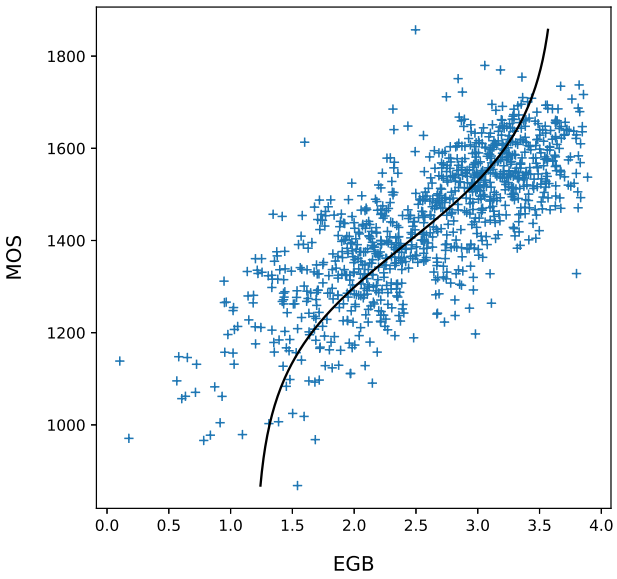

Figure 5. Scatter plot of predicted quality scores against subjective scores (MOS) on the validation set of PIPAL dataset. The black line is a curve fitted with logistic function.

Table 2. Performance comparison on validation set of PIPAL dataset. The last part of the table includes deep learning-based metrics. The **2** best results are highlighted in **bold**.

Metric	Main Score ↑	SROCC ↑	PLCC ↑
PSNR	0.5464	0.2547	0.2916
NQM [7]	0.7621	0.3457	0.4163
UQI [38]	1.0334	0.4858	0.5475
SSIM [41]	0.7383	0.3399	0.3984
MS-SSIM [42]	1.0496	0.4863	0.5632
IFC [34]	1.2703	0.5936	0.6766
VIF [33]	0.9570	0.4334	0.5235
VSNR [5]	0.6962	0.3212	0.3750
RFSIM [46]	0.5700	0.2655	0.3044
GSM [24]	0.8869	0.4181	0.4688
SRSIM [44]	1.2199	0.5658	0.6541
FSIM [47]	1.0277	0.4671	0.5605
FSIMc [47]	1.0265	0.4678	0.5586
VSI [45]	0.9662	0.4500	0.5161
MAD [22]	1.2340	0.6077	0.6262
NIQE [27]	0.1661	0.0643	0.1017
MA [25]	0.4039	0.2005	0.2034
PI [3]	0.3352	0.1690	0.1662
LPIPS-Alex [48]	1.2738	0.6275	0.6462
LPIPS-VGG [48]	1.2385	0.5914	0.6471
PieAPP [31]	**1.4034**	**0.7062**	**0.6972**
WaDIQaM [4]	1.3322	0.6779	0.6543
DISTS [9]	1.3600	0.6742	0.6858
SWD [13]	1.3291	0.6611	0.6680
EGB (Our)	**1.5511**	**0.7758**	**0.7752**

Table 3. Performance comparison on test set of PIPAL dataset. The last part of the table includes deep learning-based metrics. The **2** best results are highlighted in **bold**.

Metric	Main Score ↑	SROCC ↑	PLCC ↑
PSNR	0.5262	0.2493	0.2769
NQM [7]	0.7598	0.3644	0.3953
UQI [38]	0.8695	0.4195	0.4500
SSIM [41]	0.7549	0.3613	0.3935
MS-SSIM [42]	0.9624	0.4617	0.5006
IFC [34]	1.0400	0.4851	0.5548
VIF [33]	0.8765	0.3970	0.4794
VSNR [5]	0.7789	0.3682	0.4107
RFSIM [46]	0.6321	0.3037	0.3284
GSM [24]	0.8740	0.4093	0.4646
SRSIM [44]	1.2087	0.5728	0.6359
FSIM [47]	1.0747	0.5038	0.5709
FSIMc [47]	1.0783	0.5057	0.5726
VSI [45]	0.9752	0.4583	0.5168
MAD [22]	1.1237	0.5433	0.5804
NIQE [27]	0.1658	0.0340	0.1317
MA [25]	0.2873	0.1404	0.1468
PI [3]	0.2490	0.1036	0.1454
LPIPS-Alex [48]	1.1368	0.5658	0.5710
LPIPS-VGG [48]	1.2277	0.5947	0.6330
PieAPP [31]	1.2048	0.6074	0.5974
WaDIQaM [4]	1.1012	0.5532	0.5480
DISTS [9]	**1.3421**	**0.6548**	**0.6873**
SWD [13]	1.2584	0.6242	0.6341
EGB (Our)	**1.3774**	**0.7003**	**0.6771**

bles, we can observe that hand-crafted-based methods provide very low performance, thus demonstrating their inadequacy for assessing IR algorithms. This may be mainly due to the integration of distortions resulting from GAN-based methods for which these methods are not designed.

On the other hand, metrics based on deep learning show a fairly high correlation compared to the first category of methods analyzed. For instance, PieAPP [31] metric provides a high correlation on the validation set, while DISTS [9] metric achieves one the highest performance on the test set. These latter are well designed for the evaluation of IR algorithms. However, all the methods considered are surpassed by the three proposed models on the validation set. Moreover, when combined, we obtain the highest correlation on the validation and test sets, thus outperforming the considered IQA metrics. Finally, the obtained results show the efficiency of the proposed EGB metric for the evaluation of IR algorithms, including classical and GAN-based methods.

4.5. Cross Database Evaluations

In order to evaluate the generalization ability of EGB metric, a cross-database evaluation was carried out using the PIPAL as a training datatest and testing the proposed metric on TID2013 [30] and LIVE [32] datasets. As shown in Table 4, the proposed method outperforms all considered metrics on TID2013 dataset and also achieves comparable results on LIVE dataset. This demonstrates that the proposed EGB metric is not limited by the database on which it was trained and shows a strong capacity for generalization.

Table 4. Cross database evaluations of the proposed metric when training on PIPAL dataset and tested on TID2013 and LIVE datasets. The best result is highlighted in **bold**.

	TID2013		LIVE	
	SROCC ↑	PLCC ↑	SROCC ↑	PLCC ↑
PSNR	0.6684	0.6679	0.7379	0.7295
SSIM	0.6273	0.6211	0.6917	0.5639
MS-SSIM	0.5637	0.5907	0.7467	0.6296
VIF	0.6757	0.7325	**0.9525**	**0.9570**
GSM	0.8083	0.8771	0.8954	0.8669
FSIM	0.8078	0.8765	0.9099	0.8450
FSIMc	0.8535	0.8869	0.9056	0.8322
MAD	0.7921	0.8241	0.9073	0.8812
NIQE	0.2955	0.3475	0.7939	0.5959
EGB	**0.9034**	**0.9155**	0.8561	0.8543

Table 5. Evaluation of the three gradient boosting regression models and their combination when we consider different number of convolution layers.

#Layers	Model	SROCC ↑	PLCC ↑
3	XGBoost	0.7572	0.7546
	LIGHTGBM	0.7626	0.7626
	CatBoost	0.7713	0.7689
	Average	0.7758	0.7752
5	XGBoost	0.7391	0.7385
	LIGHTGBM	0.7549	0.7530
	CatBoost	0.7596	0.7633
	Average	0.7632	0.7612
All	XGBoost	0.7167	0.6865
	LIGHTGBM	0.7104	0.7056
	CatBoost	0.7393	0.7219
	Average	0.7356	0.7251

4.6. Ablation Study

To evaluate the efficiency of our proposed metric, we conduct an ablation study. We therefore evaluate the three gradient boosting regression models and their combination when we consider different number of convolution layers. Based on the Figure 3, three cases are analyzed. First, when three intermediate layers are used, which are the block 4 convolution layer 2, block 4 convolution layer 3 and block 5 convolution layer 1. Second, when five intermediate layers are used, the previous three layers in addition to the block 4 convolution layer 1 and block 5 convolution layer 2. Finally, when we consider all the layers of VGG16 architecture. The results are shown in Table 5, where the best per-

formance are achieved with the three layers, which proofs that our selection strategy provides the best result.

5. Conclusion

In this paper, we have proposed EGB metric for accurate prediction of perceptual image quality, in particular to address the evaluation of IR algorithms. Our contribution mainly focuses in finding perceptual quality distance in VGG16 feature space that correlates well with subjective quality scores, in addition to the adoption of ensemble gradient boosting approach. The proposed metric was compared with different state-of-the-art methods. The obtained results demonstrate clearly the relevance of the selected deep features for this specific task and the efficiency of our proposed metric.

Two possible extensions are foreseen: the consideration of weighting voting strategy in the combination of the three models, as well as taking into account several/other regression models in the quality estimation part.

References

[1] S. A. Amirshahi, M. Pedersen, and S. X. Yu. Image quality assessment by comparing cnn features between images. *Journal of Imaging Science and Technology*, 60(6):60410–1, 2016. 3

[2] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, 2018. 2

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 6, 7

[4] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017. 2, 6, 7

[5] Damon M Chandler and Sheila S Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE transactions on image processing*, 16(9):2284–2298, 2007. 5, 6, 7

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 4

[7] Niranjan Damera-Venkata, Thomas D Kite, Wilson S Geisler, Brian L Evans, and Alan C Bovik. Image quality assessment based on a degradation model. *IEEE transactions on image processing*, 9(4):636–650, 2000. 5, 6, 7

[8] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, 2009. 2, 3

[9] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. 6, 7

[10] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4):1258–1281, 2021. 2

[11] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018. 4

[12] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114, 2017. 3

[13] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric. *arXiv preprint arXiv:2011.15002*, 2020. 6, 7

[14] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 5

[15] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3012–3021, 2020. 2

[16] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020. 2, 3, 5

[17] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[18] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 2

[19] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017. 4

[20] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee. Deep cnn-based blind image quality predictor. *IEEE transactions on neural networks and learning systems*, 30(1):11–24, 2018. 2

[21] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal processing magazine*, 34(6):130–141, 2017. 2

[22] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010. 6, 7

[23] Yuming Li, Lai-Man Po, Litong Feng, and Fang Yuan. No-reference image quality assessment with deep convolutional neural networks. In *2016 IEEE International Conference on Digital Signal Processing (DSP)*, pages 685–689. IEEE, 2016. 2

[24] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2011. 5, 6, 7

[25] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 6, 7

[26] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 723–727. IEEE, 2011. 1

[27] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2, 6, 7

[28] Pedram Mohammadi, Abbas Ebrahimi-Moghadam, and Shahram Shirani. Subjective and objective quality assessment of image: A survey. *arXiv preprint arXiv:1406.7799*, 2014. 2

[29] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 1

[30] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015. 7

[31] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 6, 7

[32] HR Sheikh. Live image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*, 2005. 7

[33] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006. 2, 5, 6, 7

[34] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005. 2, 5, 6, 7

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 3

[36] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. 2

[37] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[38] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002. 5, 6, 7

[39] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009. 2

[40] Zhou Wang and Alan C Bovik. Reduced-and no-reference image quality assessment. *IEEE Signal Processing Magazine*, 28(6):29–40, 2011. 2

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 5, 6, 7

[42] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 2, 5, 6, 7

[43] X. Yang, F. Li, and H. Liu. Deep feature importance awareness based no-reference image quality prediction. *Neurocomputing*, 401:209–223, 2020. 3

[44] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE international conference on image processing*, pages 1473–1476. IEEE, 2012. 5, 6, 7

[45] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10):4270–4281, 2014. 1, 6, 7

[46] Lin Zhang, Lei Zhang, and Xuanqin Mou. Rfsim: A feature based image quality assessment metric using riesz transforms. In *2010 IEEE International Conference on Image Processing*, pages 321–324. IEEE, 2010. 5, 6, 7

[47] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 6, 7

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 3, 6, 7

[49] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proc. of the IEEE*, 2020. 3