Soft Computing
https://doi.org/10.1007/s00500-021-06662-9

FOCUS

Deep ensembling for perceptual image quality assessment

Nisar Ahmed[1] · H. M. Shahzad Asif[2,3] · Abdul Rauf Bhatti[4] · Atif Khan[5]

Accepted: 7 December 2021

Abstract

Blind image quality assessment is a challenging task particularly due to the unavailability of reference information. Training a deep neural network requires a large amount of training data which is not readily available for image quality. Transfer learning is usually opted to overcome this limitation and different deep architectures are used for this purpose as they learn features differently. After extensive experiments, we have designed a deep architecture containing two CNN architectures as its sub-units. Moreover, a self-collected image database BIQ2021 is proposed with 12,000 images having natural distortions. The self-collected database is subjectively scored and is used for model training and validation. It is demonstrated that synthetic distortion databases cannot provide generalization beyond the distortion types used in the database and they are not ideal candidates for general-purpose image quality assessment. Moreover, a large-scale database of 18.75 million images with synthetic distortions is used to pretrain the model and then retrain it on benchmark databases for evaluation. Experiments are conducted on six benchmark databases three of which are synthetic distortion databases (LIVE, CSIQ and TID2013) and three are natural distortion databases (LIVE Challenge Database, CID2013 and KonIQ-10 k). The proposed approach has provided a Pearson correlation coefficient of 0.8992, 0.8472 and 0.9452 subsequently and Spearman correlation coefficient of 0.8863, 0.8408 and 0.9421. Moreover, the performance is demonstrated using perceptually weighted rank correlation to indicatethe perceptual superiority of the proposed approach. Multiple experiments are conducted to validate the generalization performance of the proposed model by training on different subsets of the databases and validating on the test subset of BIQ2021 database.

Keywords Image quality assessment · Perceptual quality assessment · Blind image quality assessment · No-reference image quality assessment · Deep learning · Deep ensemble · Ensemble learning · Convolutional neural networks · Natural distortion image database

Communicated by Irfan Uddin.

✉ Nisar Ahmed
nisarahmedrana@yahoo.com

H. M. Shahzad Asif
Shehzad@uet.edu.pk

Abdul Rauf Bhatti
bhatti_abdulrauf@gcuf.edu.pk

Atif Khan
atifkhan@icp.edu.pk

[1] Department of Computer Engineering, University of Engineering and Technology, Lahore 54890, Pakistan

[2] Department of Computer Science, University of Engineering and Technology (New Campus), Lahore 39020, Pakistan

[3] Department of Computer Science, University of Engineering and Technology, Lahore 54890, Pakistan

[4] Department of Electrical Engineering and Technology, Government College University Faisalabad, Faisalabad 38000, Pakistan

[5] Department of Computer Science, Islamia College Peshawar, Peshawar 25120, Pakistan

Published online: 30 January 2022

1 Introduction

Image Quality Assessment (IQA) is a crucial and challenging task and is required for many image processing applications (Chandler 2013). IQA systems try to learn the relationship between the image and its relative quality score provided by human observers in subjective quality assessment experiments. This quality score is a Mean Opinion Score (MOS) of human judgment. The relationship between image and its corresponding quality score is dependent on the human visual system which is a naively understood area and therefore modeling such a system is tough (Ahmed et al. 2019).

IQA has several applications as a quality metric (Wang et al. 2002; Prabha and Kumar 2017; Selva Nidhyanandhan et al. 2020). It can be used to benchmark an image processing algorithm. If we have to select an image processing algorithm among several choices, a quality metric can help us to identify the one with best-reproduced image quality. Similarly, it can be used to monitor image quality such as in a video transmission network; the metric can inspect the image/video quality and control the streaming. In a video or image acquisition system; a quality metrics can monitor and adjust the acquisition hardware parameters to obtain the best quality image/video. Moreover, such quality metrics can be embedded into an image processing system for optimization of an algorithm's parameters to obtain the best quality image, e.g., image enhancement or visual communication systems. In short, IQA is crucial in numerous application scenarios where it can be used as benchmarking, optimization, control, or monitoring setting.

The ultimate way to assess image quality is by visual inspection by humans (Ahmed and Asif 2019). It can be performed by subjective assessment and averaging the opinion of several subjects that is an expensive, cumbersome and difficult process. Objective quality assessment is sought for this scenario where an algorithm performs the task of IQA but it must correlate well with human judgment. The conventional image quality assessment algorithms such as PSNR or SSIM (Wang et al. 2004a) are regarded as full-reference IQA methods. These methods require a reference image to perform the quality assessment by learning some error/structural change to perform the task of quality assessment. These approaches have limited utility and are useful only when the algorithm has access to reference information. A more robust system, which can model the behavior of human visual system and can work in no-reference setting is desired.

Classical no-reference IQA approaches extract natural scene statistics, which are affected by degradations in an image and train a regression algorithm (Sheikh et al. 2005a, b; Saad et al. 2012; Ahmed et al. 2021; Khalid et al. 2021). On the contrary, deep learning approaches especially Convolutional Neural Networks (CNN) have shown some promising performance in image quality assessment (Ahmed and Asif 2019, 2020; Ma and Jiang 2019; Bianco et al. 2018; Bosse et al. 2017). The success of CNN in IQA is due to the fact that it is inspired by the human visual cortex and it can learn quality-aware features by itself provided with representative training data. A problem in training a CNN for image quality assessment is the lack of sufficient training data. This lack of sufficient data is tried to overcome by using transfer learning, data augmentation, or training a relatively shallow CNN. Despite a lot of success in IQA by deep learning, there is still room for improvement as newer architectures and approaches provide better quality assessment performance and unveil methods that can result in true human visual system representation.

In this work, we have designed two, closely related, deep ensemble-based architectures. Contrary to our previous work (Ahmed and Asif 2019) which uses multiple snapshots of training with a cyclic learning rate to construct an ensemble. The proposed deep ensemble can be regarded as a single model and trained end-to-end making it easier to train and it provides better quality assessment performance. The intuition of this architecture is based on the idea that image quality depends on microstructures such as pixel relations and macro structures such as objects of interest. Different CNN architectures learn these features differently and their features can be combined for better representation. The proposed architecture, therefore, contains two entirely different classes of CNNs as a subset. The features of these two networks are concatenated and passed through global average-pooling and a few fully connected layers. Moreover, an image database BIQ2021 with natural distortions having 12,000 images is proposed. Extensive experiments are conducted to train the proposed DeepEns and DeepEns-Lite models on different subsets of databases and perform validation on the test-set of the proposed database BIQ2021. The objective of the experimentations is to demonstrate the usefulness of the proposed end-to-end training approach for general-purpose image quality assessment. The proposed image database is used for a similar purpose to highlight that model trained on image databases with simulated distortions does not perform well on images with natural distortions. Specific contributions of this work are highlighted below:

1. A large-scale image database (12,000 images) having natural distortions and laboratory-controlled subjective scores are introduced.
2. Deep ensemble-based architectures are proposed for end-to-end training.

3. A deep CNN training approach is proposed using a synthetic distortion database with 18.75 million images.
4. End-to-end training of deep CNN architecture and comparison with existing approaches to demonstrate effectiveness.
5. Cross-validation of the trained model on natural distortion databases to highlight its generalizability.
6. Demonstration of the fact that image databases with simulated distortions are not suitable for the training of models for general-purpose image quality assessment.

2 Related work

This section provides a brief overview of recent work which is closely related to the proposed approach. The problem of blind image quality assessment is conventionally addressed based on Natural Scene Statistics (NSS) (Chandler 2013). The surprising performance of deep learning in visual recognition has directed the IQA research from NSS to deep learning. An important advantage of deep learning-based methods is that there is no need for handcrafted features as they are learned directly based on training images. The difficulty in deep learning-based methods lies in the availability of training data as the available training image for IQA are not very large. However, researchers have developed techniques that can address the problem of IQA using deep learning-based methods.

Kang et al. (2014, 2015) proposed a deep model which trains the CNN using spatially normalized image patches. The quality and distortion types are identified simultaneously using a multi-task CNN architecture. Bianco et al. (2018) have proposed DeepBIQ which is a blind IQA model based on CNN. They have used an AlexNet like architecture, which is pretrained on ImageNet. They have extracted features from this deep network by taking multiple crops of 224 × 224 and then average-pooled the features to train a support vector regression (SVR). Gao et al. (2018) have used a Vgg16 architecture for quality assessment and reasoned that different levels of convolution represent image quality differently. They trained an SVR on features extracted at different levels of convolution and then averaged pooled to provide image quality. Kim et al. (2018) proposed a two-stage image quality predictor. The first stage predicts an objective error map and the second stage predicts subjective quality score. Ravela et al. (2019) has also opted for a two-stage approach. The first stage predicts the distortion type and the second stage uses a specialized deep model to predict the subjective score and perform average-pooling with the prediction of other deep models. Ma et al. (2017a) also proposed a two-stage network: the first identifies the type of distortion and the second stage performs the subjective score prediction using a specialized quality prediction network for each distortion type.

Zhang et al. (2018b) also proposed a two-stage framework, the first stage identifies the distortion type as well as the level of distortion. The second stage performs quality assessment using a specialized CNN model. Their models have focused to perform quality assessment for both authentically distorted as well as inauthentically distorted images. Fan et al. (2018) also follow a two-stage approach by first predicting the distortion type and then performing subjective score prediction using multiple CNNs.

Deep features have provided unreasonable effectiveness for image quality assessment (Zhang et al. 2018a). The problem with such methods is that they cannot be trained end-to-end but they are simple and effective. The features extracted from ImageNet pretrained models such as VGG are used to train a regression algorithm for image quality assessment (Ma and Jiang 2019). Ahmed (2020) has experimented with feature extraction from different layers of ImageNet pretrained models. Moreover, they have retrained these models on image quality databases and then performed the feature extraction. It is demonstrated that ImageNet pretrained models are not a good candidate for perceptual quality assessment of digital images. Retraining of these models makes them learn quality-aware features and these features can be used for image quality assessment. Another approach is to train the regression algorithm with natural scene statistics as well as deep features for an enriched feature experience (Ahmed et al. 2019).

The proposed approach, on the other hand, follows a slightly different approach and proposes an architecture that uses two entirely different CNN architectures as its sub-units. The EfficientNet-B0 (Tan and Le 1905) is a lightweight CNN model which uses inverted bottleneck units for its construction followed by pooling, fully connected, and classification layers. NASNet-mobile (Zoph et al. 2018) on the other hand uses a cell structure that is learned using neural architecture search on CIFAR-10 dataset. Both of them learn different types of visual features, are lightweight and provide superior performance to a single computationally expensive CNN model. A novel training strategy is proposed for quality assessment tasks using synthetically distorted images. The proposed architecture is, therefore, more suitable for learning rich-feature sets and in turn, provides quality scores, which are comparable to the state-of-the-art.

3 Materials and methods

IQA is a challenging task and numerous datasets have been published for full-reference as well as no-reference quality assessment experiments. We have selected six popular databases to perform experiments and benchmark the results. LIVE and TID are image quality databases released in three and two versions subsequently whereas CSIQ, LIVE in the wild Challenge Database (LiveCD), CID2013 and KonIQ-10 k are released in a single version. Table 1 provides the eight versions of six databases with the number of images, scoring method, and range. We have linearly transformed the range of all the databases from 0 to 1 so they can be used for cross-dataset evaluation. It is highlighted that CSIQ, LIVE-I, LIVE-II, TID2008, and TID2013 are datasets generated from reference images by simulating distortions and are called synthetic distortion databases. LiveCD, CID2013 and KonIQ-10 k are different types of datasets with images having naturally occurring distortions. LiveCD contains images that are captured by random devices and have some sort of distortions whereas CID2013 has used a fixed set of image acquisition devices with a set of parameters to introduce distortions intentionally. They have captured images in 36 different indoor and outdoor locations with different cameras and settings to capture the distortions occurring naturally. KonIQ-10 k is the largest database introduced recently. It contains 10,073 images collected from online sources and subjectively scored using crowdsourcing. Our proposed dataset contains all three types of images having natural distortions. The details of this database are described further in the next section.

3.1 BIQ2021

BIQ202 is the proposed database, which contains images having natural distortions. It is a large-scale database of 12,000 images. The collection of the images for the dataset is performed in three different subsets: (i) The first subset contains 2000 images which are captured with different camera settings (i.e., ISO, shutter speed, focal length, and motion) to introduce distortions. This subset contains images captured from the same scene having degradations ranging from just noticeable to severe as depicted in Fig. 1. These images contain a lack of focus, motion blur, non-uniform illumination, sensor noise, and a mix of different degradations. (ii) The second subset contains 2000 images with natural distortions introduced during the process of acquisition or storage. These images are not captured with intentions to be used for image quality research and are the author's collection. These images, apart from the first subset contain compression, processing, and storage-related degradations. Figure 2 provides a subset of images for demonstration. (iii) The third subset contains 8000 images that are manually selected from Unsplash.com having varying image quality and content. These images contain all the distortions occurring in the earlier subsets and they may involve degradations resulting due to pot-processing and trends of the photographic community. These images are added to introduce content diversity by downloading images having different tags such as animals, people, babies, sports, architecture, nature, etc. Moreover, this subset provides a representation of images captured by a typical photography community. Figure 3 provides a depiction of the subset of images in this category.

Subjective scoring of these images is performed in laboratory settings during multiple sessions. Each session consisted of the half to one hour. The session time, viewing distance and viewing angle were at the liberty of the subject, so the quality scores can be obtained naturally. The ambient conditions, display device and scoring mechanism are kept constant. The display device used during the subjective scoring is HP 24 inches LED, full-HD display, 16:9 aspect ratio, and 200 cd/m^2 brightness. The scoring is

Table 1 Benchmark datasets with number of images, scoring method and range

Dataset name	Number of reference images	Number of distorted images	Scoring method	Range
CSIQ (Larson 2010)	30	900	DMOS	0–1
LIVE-I (Sheikh and LIVE, 2005)	29	460	DMOS	0–100
LIVE-II (Sheikh and LIVE, 2005)	29	982	DMOS	0–100
TID2008 (Ponomarenko et al. 2009)	25	1700	MOS	1–10
TID2013 (Ponomarenko et al. 2015)	25	3000	MOS	1–10
CID2013 (Virtanen et al. 2014)	Nil	474	MOS	0–100
LiveCD (Ghadiyaram and Bovik 2015)	Nil	1169	MOS	0–100
KonIQ-10 k	Nil	10,073	MOS	1–5
BIQ2021	Nil	12,000	MOS	0–1

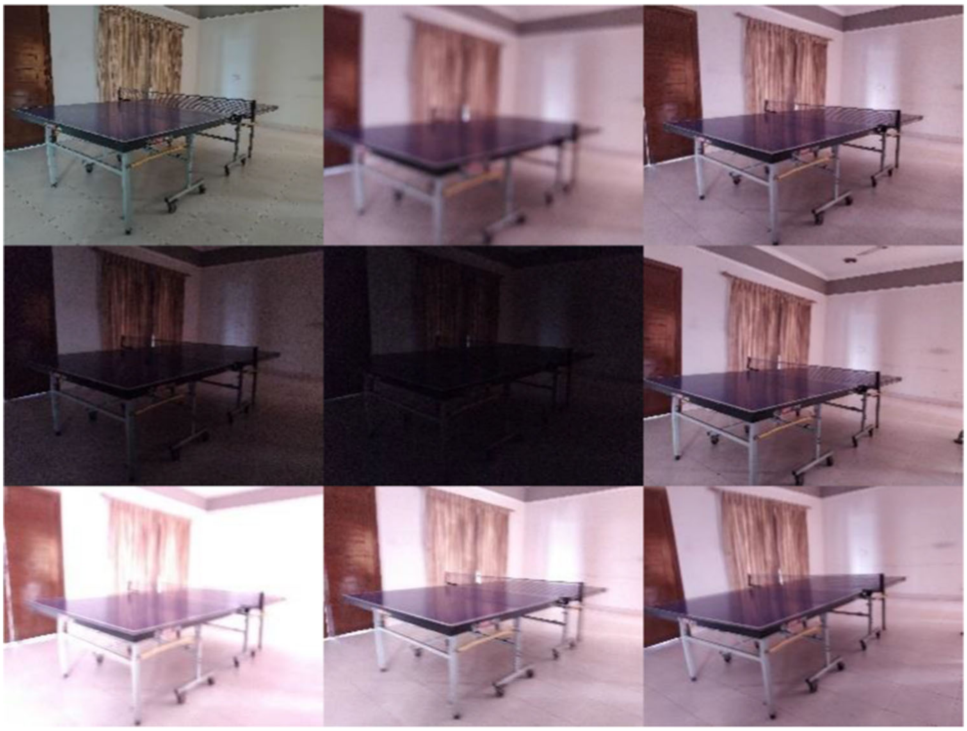


Fig. 1 Images captured with different camera settings

done through absolute category rating with five quality levels. The observer was free to use a continuous value through a slider or select a discrete rating. Figure 4 provides the GUI of the desktop application used for subjective scoring. Moreover, the score history was provided to encourage the observer to use the full range of the scoring scale. The observer was free to leave the experiment if he/she felt withdrawn. Most of the observers were graduate or undergraduate students of computer science and engineering and had no specific expertise in the domain of image quality assessment. The experiments were terminated after 30 observers completed the subjective scoring. MOS is calculated by averaging the quality scores of 30 observers and then scaled to a range of 0–1. The variance of scores provided by 30 observers along with MOS are provided in the dataset. Figure 5 provides the histogram of MOS with 100 bins.

3.2 Deep ensemble

CNNs are a class of neural networks designed specifically for visual recognition tasks. There are several pretrained models which are originally trained on ImageNet visual recognition challenge to classify images into 1000 categories. Image quality assessment researchers either use these models for feature extraction or perform transfer learning for image quality assessment. These pretrained models have different representational power and learn different types of deep features. Figures 6 and 7 provide DeepDream visualization of the last fully connected layer for six channels (114, 293, 341, 484, 563, and 950) of two of these architectures with the same initialization. It can be seen that these features are visually different from each other. Moreover, it is demonstrated experimentally that they provide different quality assessment performances (Ahmed and Asif 2020).

We have provided two versions of the proposed DeepEns architecture. The lighter version named DeepEns-Lite contains EfficientNet-B0 and NASNet-mobile as base architectures as they are small and have good visual recognition performance. The selection of these two base architectures is based on their lighter weight and good quality assessment performance. The full-version named DeepEns contains InceptionResNet-V2 and EfficientNet-B7 as base architectures as they have provided the highest quality assessment performances among 19 popular CNN models. The architectures of DeepEns and DeepEns-Lite are provided in Figs. 8 and 9. We have repurposed these models for our problem and taken the CNN base to construct our architecture which extracts features by both and then concatenates them. Although there are different global pooling operations, max-pooling and average-pooling are the most common ones. Guo et al. (2019) have presented a study to analyze the effect of nine pooling strategies on fine-grained visual recognition. They have concluded that max-pooling learns the discriminative details by learning smaller and important distinguishing parts of the image, whereas the average-pooling provides an averaging effect

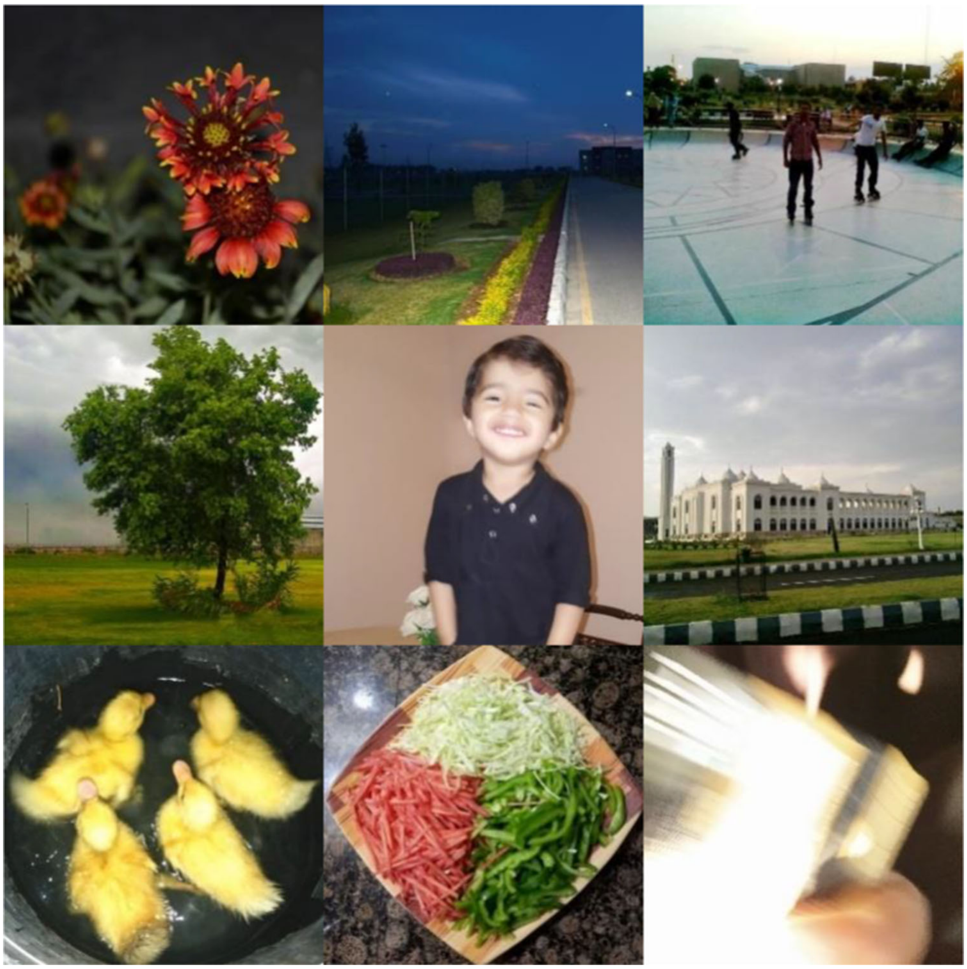


Fig. 2 Images selected from random image collections

and learn global image parameter and does not learn specific details. As the image quality is a global parameter and is not specific to shapes or areas, we have therefore used global average-pooling. Although a thorough study can be conducted to study the effect of different pooling strategies on image quality assessment, but it is not done here and left as future work.

Three fully connected layers with 1024, 256 and 1 neurons that preceded the global average-pooling layer are used in the lighter version of DeepEns-Lite. In the full version, global average-pooling is applied on both base architectures prior to concatenation. The concatenation is followed by 4096, 1024, 256, and 1 neurons. The dropout of 25% is used in earlier layers and 50% in the layer with 256 neurons which provides additional regularization. The rectified linear unit (ReLU) is used as an activation function with all the fully connected layers. The final fully connected layer contains one neuron and provides the activations to the regression layer.

3.3 Loss function

The regression layer normally uses mean squared error (MSE) as a loss function for training. The advantage of MSE is that it provides a smooth convex function, which is easier to train due to ease in gradient computation. Although, mean squared error seems fine as a loss function but keeping in view the trend in literature, we have experimented with mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared logarithmic error (MSLE), LogCosh loss, and Huber loss along with MSE. It is observed that mean squared error provides faster training and the performance obtained with a model trained using MSE is slightly better than the others. Table 1 provides the validation RMSE-values obtained after training the DeepEns-Lite for 30 epochs. The experiments are conducted using LiveCD. The details of these loss functions along with their training progress are provided in Appendix. It is evident from the training progress of Figs. 15, 16, 17, 18, 19, and 20 and Table 2 that MSE is the best candidate loss function among the six loss functions. The better performance of MSE is possibly due to the fact

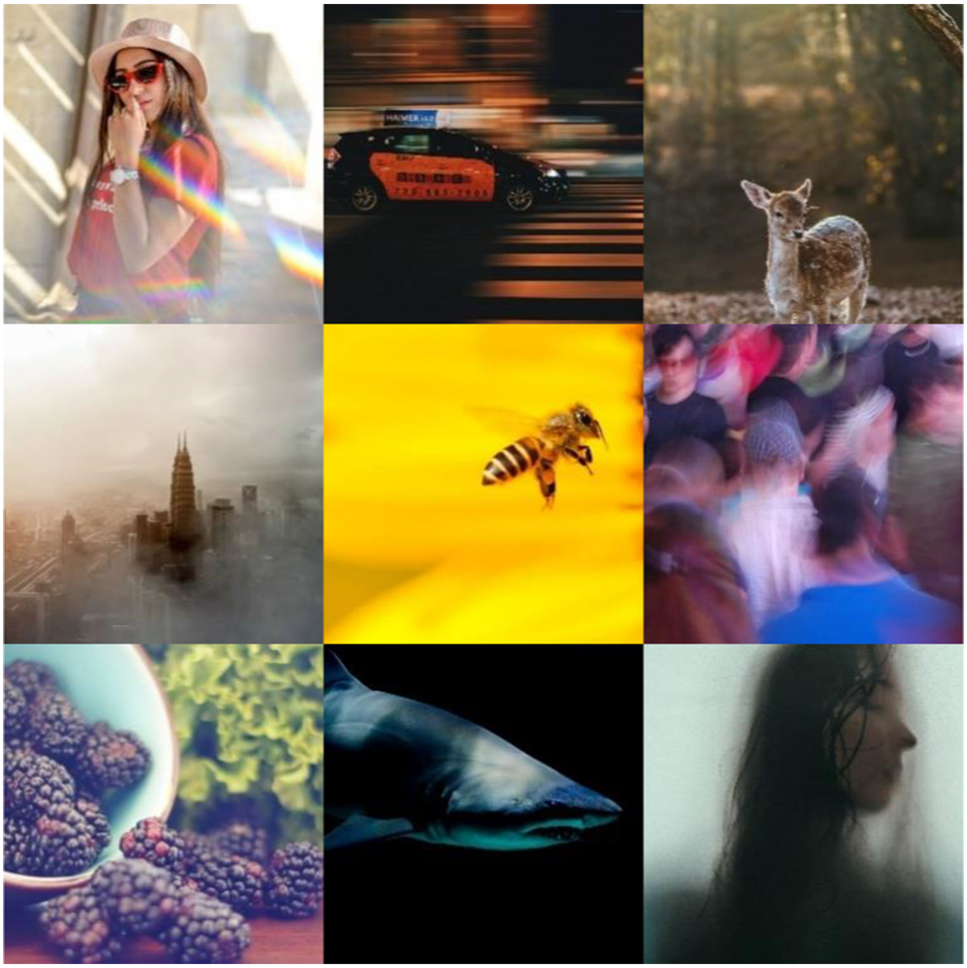
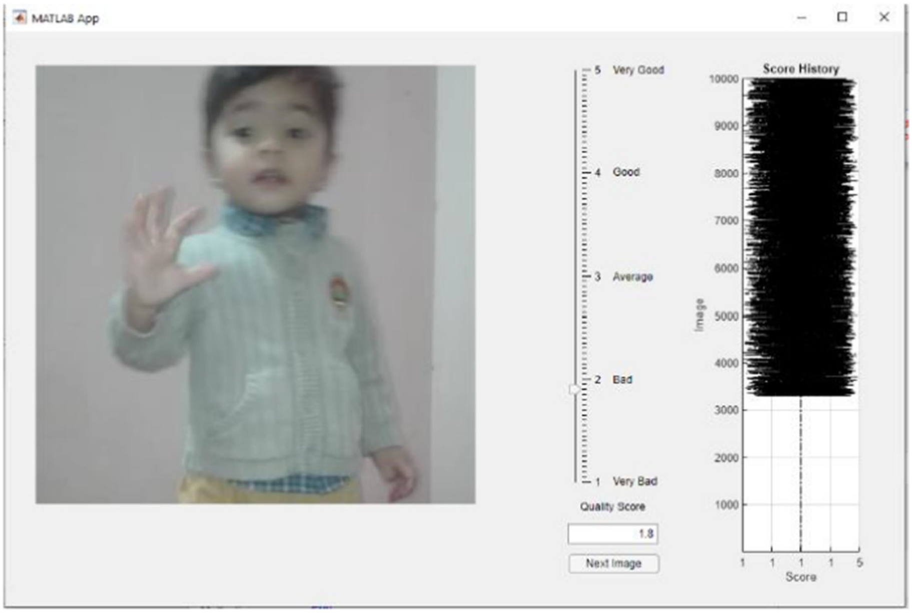


Fig. 3 Images with diverse content downloaded from Unsplash.com



Quality Score 1.8

Fig. 4 GUI used for subjective quality scoring

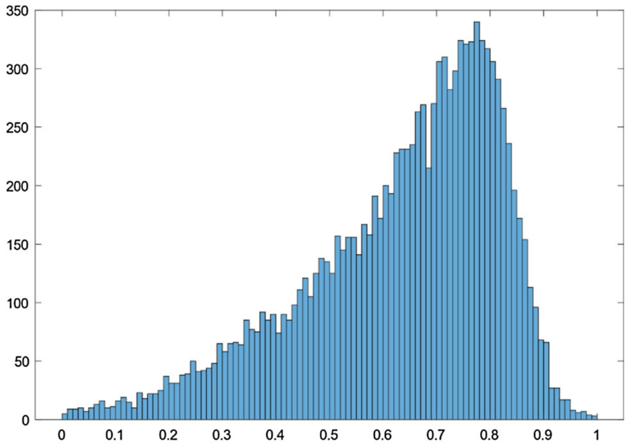


Fig. 5 MOS distribution of BIQ2021

that MOS is obtained through controlled experiments with outlier rejection and is average of a number of observers. It is, therefore, less possible to experience outliers and other abnormalities in training data and, therefore, quadratic nature of MSE makes it a better candidate. Although, visual perception-based loss functions, such as the one based on visual saliency may perform better, but they are not explored in this study and left as a future work.

3.4 Model pretraining

Pretraining of deep CNN models is required as the benchmark database sizes are not sufficient to train a model from scratch. The largest benchmark database has 10,000 images. It is the practice in the literature to use models pretrained on ImageNet and repurpose them to be used for image quality assessment through transfer learning. The authors have used a different strategy to train the models for quality assessment apart from ImageNet. We have used 150,000 images provided by Kadis-700 K (Lin et al. 2019) to generate distorted images. We have generated 25 different types of distortions in five distinct levels and obtained 18.75 million distorted images and named Kadid-19 M. We have assigned distortion type and level as the label of each image such as '23_3' for the distortion number 23 with a distortion level of 3. This way, we had a labeled database of 18.75 million images with 125 categories. This is a large database and is different from ImageNet. The models trained on ImageNet are specialized for visual recognition and can be transferred to other related tasks. Image quality assessment is an entirely different problem as images with entirely different content may have the same quality score or images of the same scene may have different quality scores. It is, therefore, proposed to use a database, which is related although having less reliable labels. This proposed strategy is found to be more useful in performing transfer learning. The performance of the proposed strategy is demonstrated using EfficientNet-B0 on TID2013 image database in Table 3. The training was performed with a piecewise learning rate until convergence.

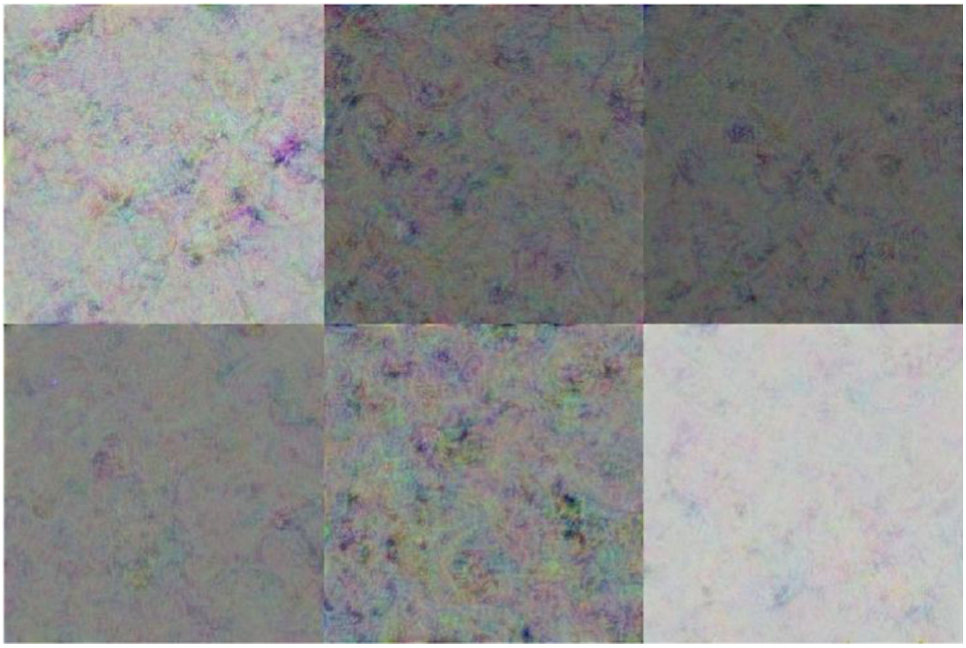


Fig. 6 Deepdream visualization of features learned by NASNet-Mobile

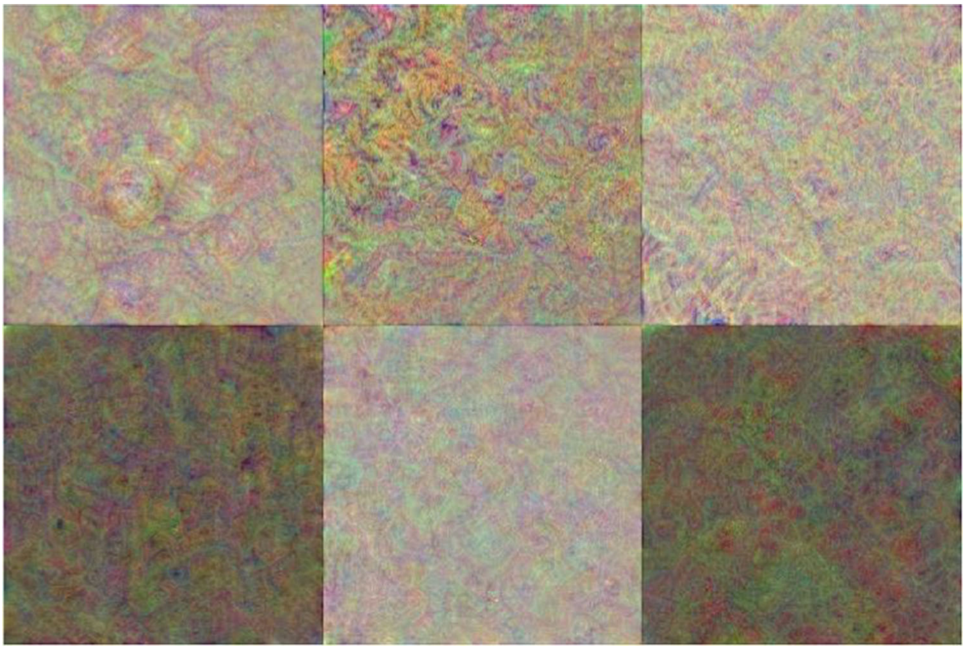


Fig. 7 Deepdream visualization of features learned by EfficientNet-B0

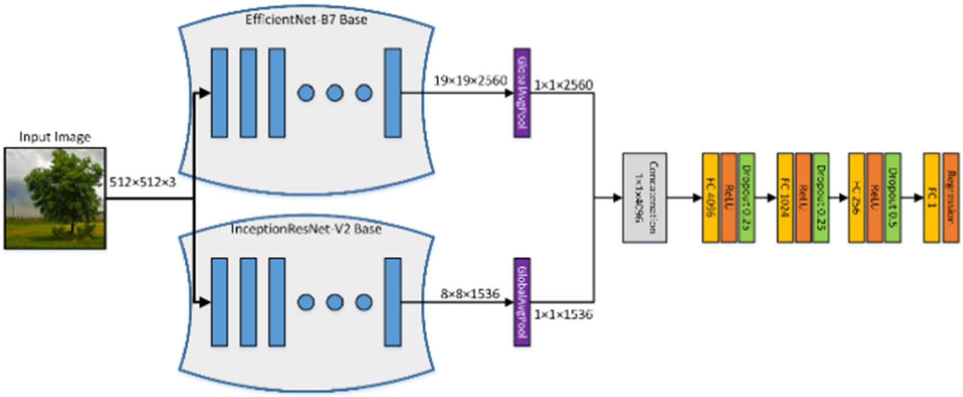


Fig. 8 Architecture of DeepEns

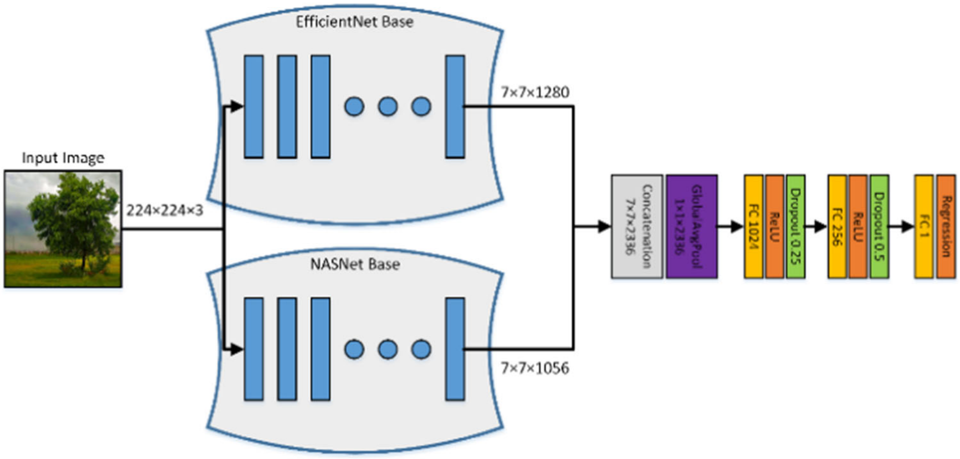


Fig. 9 Architecture of DeepEns-Lite

Table 2 Loss function performance for 30 epochs on LiveCD database

Sr	Loss function	RMSE
1	Mean squared error	1.1403
2	Mean absolute error	1.9873
3	Mean absolute percentage error	2.1201
4	Mean squared logarithmic error	1.5501
5	LogCosh loss	1.2553
6	Huber loss	1.9315

Table 3 Transfer learning for image quality

Metric	ImageNet	Kadid-19 M
RMSE	0.4561	0.4426
PLCC	0.9471	0.9490
SROCC	0.9427	0.9488

3.5 Training strategy

Same training options are used for training of pretrained model and the transfer learning. The only difference was the choice of learning rate and epochs. The initial training using a synthetic database used adam optimizer with an initial learning rate of $1e^{-2}$ with a piecewise learning rate with a drop factor of 50% and a drop period of 10 epochs. The training was performed for 100 epochs with a batch size of 16.

Regularizations are important to improve the generalization of deep CNN models. We have incorporated dropout in the fully connected layers to reduce the overfitting. Similarly, we have incorporated data augmentation to reduce overfitting and improve generalization. We have used two image augmentation strategies, namely, random horizontal reflection and random rotation. Some image augmentation strategies such as scaling, shear, contrast, equalization, color, and brightness variation are not used as the perceptual quality is sensitive to these augmentations. Random cropping also serves the purpose of regularization, and therefore, horizontal or vertical translation is not necessary.

As the input size of the proposed architecture is 224 × 224 incase of DeepEns-Lite and 512 × 2 in the case of DeepEns but the image size in the different databases is different. One method to match the input resolution is to resize but it affects the perceptual quality of the image as demonstrated by Wang et al. 2004b. The second approach is center cropping which does not provide a regularization effect and may be less useful in case of non-uniform illumination, narrow depth of field, and varying information in different image regions. The third approach is random cropping in which a randomly selected image patch equal to input resolution is cropped in each epoch and therefore resolves the issue of input size inconsistency as well as provides regularization. This will also provide averaging effect over different epochs and will be a close representation of the image. The training progress of DeepEns-Lite is provided in Fig. 10 for demonstration.

In the testing phase, the trained network performs predictions on N randomly cropped regions of each image, and their scores are averaged. Given an input image divided into N_c random crops and each having a quality score q_i. The predicted quality score is calculated by averaging the individual quality score of each crop by using Eq. (1).

$$q = \frac{1}{N_c}\sum_{i}^{N_c} q_i \tag{1}$$

4 Experiments

This section describes the experimental setup used for the evaluation of the proposed architectures. The experiments are conducted by taking 224 × 224 image patches in the case of DeepEns-Lite and 512 × 512 in the case of DeepEns. The predicted quality scores reported in this section are average of 10 crops calculated using the formula in Eq. (1). The experiments are performed on Dell T5600 workstation with two Intel E5-2687 W CPUs, 32 GB RAM, and RTX 2070 GPU. The dataset is stored on SATA SSD to reduce the preprocessing and mini-batch loading latency.

4.1 Evaluation metrics

The performance of the proposed approach is assessed using four different metrics which measure correlation or error between subjective and predicted quality scores.

Root mean squared error (RMSE) measures the error between the subjective score y_S and the predicted score y_P. It measures how accurately the model has predicted the quality score. A smaller value of RMSE means the model provides predictions with less average deviation from subjective scores. Equation (2) provides the formula for RMSE calculation.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i (y_P - y_S)^2}{n-1}} \tag{2}$$

Pearson linear correlation coefficient (PLCC) measures the linear relation between subjective score y_S and the

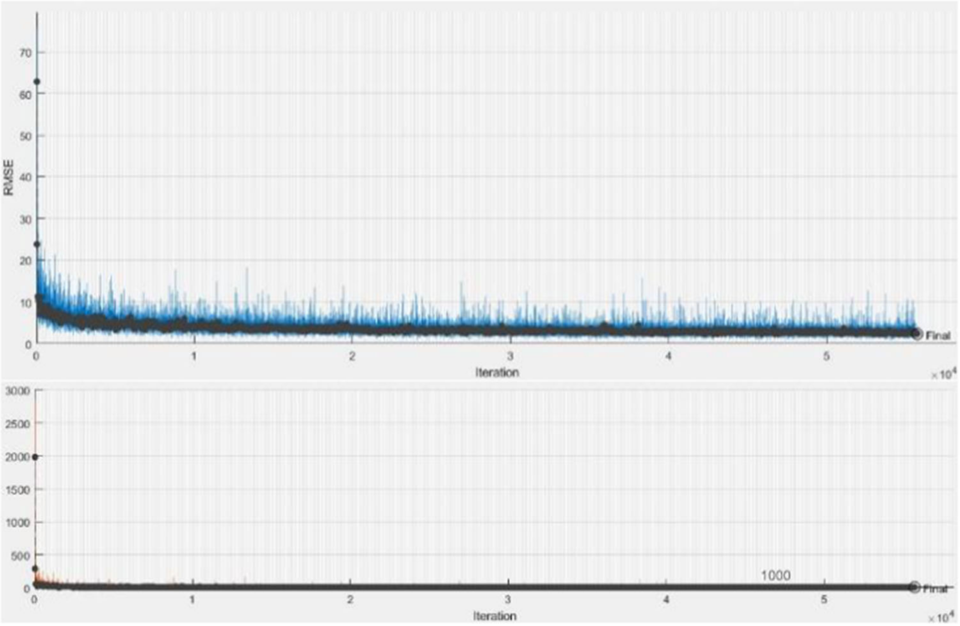


Fig. 10 Training Progress of the DeepEns-Lite architecture

predicted score y_P. A higher value of PLCC indicates that predicted scores are consistent with human judgment. Equation (3) provides the formula for PLCC calculation.

$$\text{PLCC} = 1 - \frac{\sum_i (y_P - y_S)(\hat{y}_P - \hat{y}_S)}{\sqrt{\sum_i (y_P - y_S)^2}\sqrt{(\hat{y}_P - \hat{y}_S)^2}} \tag{3}$$

Spearman rank order correlation coefficient (SROCC) measures the correlation with ranked subjective scores. A higher value of SROCC indicates that the trend of the subjective score is correlated with that of the predicted score which indicates consistency with human judgment. Equation (4) provides the formula for SROCC computation.

$$\text{SROCC} = 1 - \frac{6\sum_n d_n^2}{N(N^2 - 1)} \tag{4}$$

Perceptually weighted rank correlation (PWRC) (Wu et al. 2018) is a perceptual approach for the comparison of objective image quality assessment algorithms. It works by rewarding the capability of correct ranking of high-quality images and suppressing the attention towards insensitive rank mistakes. The PWRC calculates the area under the curve for $S(x, y, T)$ which is a combination function that fuses three rank correlation components. The components of the combination function are perceptually weighted activation function, outlier detection function, and importance measurement function.

$$\text{PWRC} = \int_{T_{\min}}^{T_{\max}} S(x, y, T)dT$$

where

$$S(x, y, T) = f[A(x, T), D(p, q), M(p, q)]$$

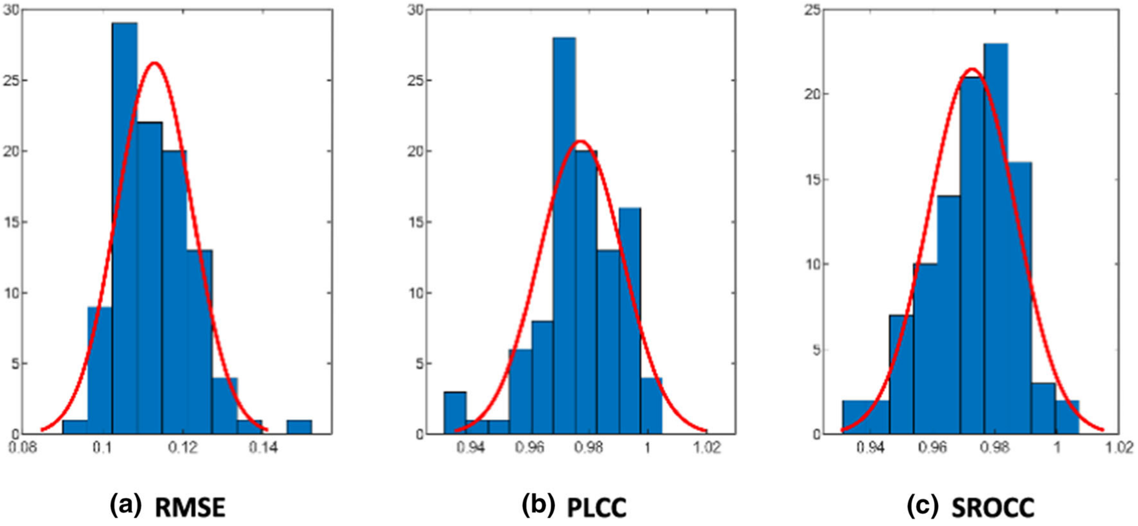
4.2 Performance assessment on individual datasets

The performance assessment test is performed on the 20% hold-out dataset which is not exposed during training. The model is trained on 80% dataset and the testing results are evaluated using three performance metrics described earlier. Table 4 provides the results of experimental evaluation based on defined metrics on six datasets. It is worth noting that LIVE is the combination of LIVE release-1 and LIVE release-2 as they have followed the same scoring methods but different distortion categories. TID2008 and TID2013 cannot be combined this way as they have overlapping distortions and are independently reported in the literature. It can be noted from Table 4 that the synthetically distorted image databases can be easily modeled due to the limited number of distortions simulated on a set of reference images. The naturally distorted image databases are difficult to model as they contain distortions introduced during the process of image acquisition either intentionally or unintentionally.

It is worth mentioning, as the benchmark image databases do not have a train/test split, it is therefore not possible to make a fair comparison. We have, therefore, followed the literature in which 80% of data is randomly selected for training and 20% for testing. The experiment is repeated several times to make the performance independent of the split as much as possible. We have repeated this experiment 10 times as the training of a CNN is a

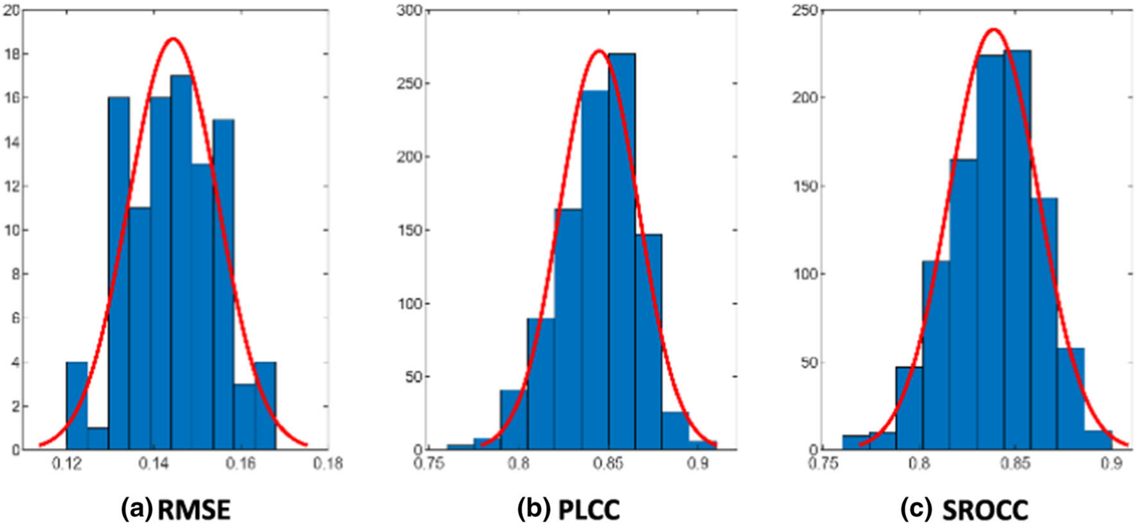
Table 4 Experimental evaluation on individual datasets

	LIVE	TID2013	CSIQ	LiveCD	CID2013	KonIQ-10 K
DeepEns-Lite						
RMES	0.0825	0.4212	0.1117	0.1095	0.1414	0.1011
PLCC	0.9813	0.9738	0.9784	0.8992	0.8472	0.9452
SROCC	0.9781	0.9761	0.9748	0.8863	0.8408	0.9421
PWRC	8.2121	7.9931	10.2141	11.0214	13.7723	12.1641
DeepEns						
RMES	0.0569	0.4170	0.1074	0.0834	0.1268	0.0859
PLCC	0.9860	0.9874	0.9846	0.9135	0.8620	0.9478
SROCC	0.9887	0.9818	0.9817	0.8985	0.8421	0.9442
PWRC	8.2136	8.3944	10.2157	12.0242	13.7752	14.1769



(a) RMSE (b) PLCC (c) SROCC

Fig. 11 CSIQ: Histograms of **a** RMSE, **b** PLCC and **c** SROCC for 100 random train-test splits



(a) RMSE (b) PLCC (c) SROCC

Fig. 12 CID2013: Histograms of **a** RMSE, **b** PLCC and **c** SROCC for 100 random train-test splits

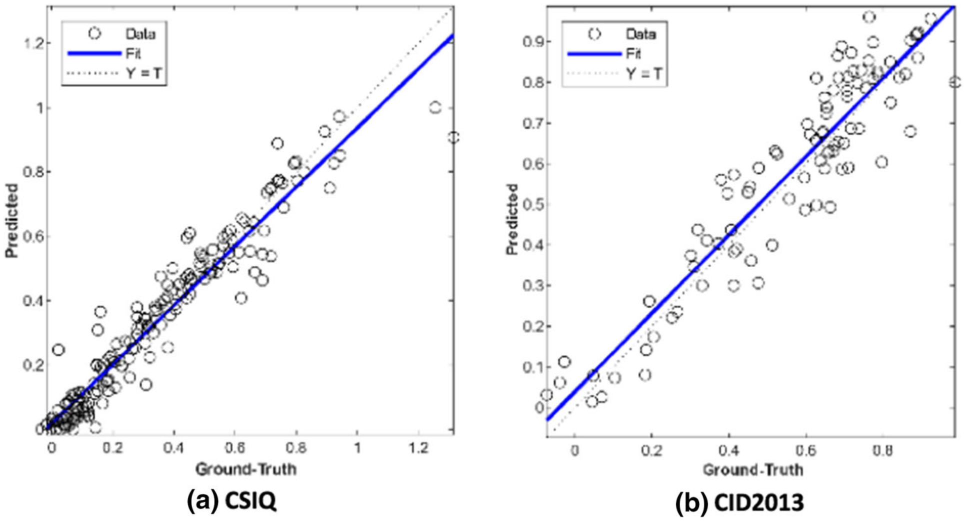


Fig. 13 Scatter-plot between ground-truth versus predicted values along with regression line for **a** CSIQ database and **b** CID2013 database

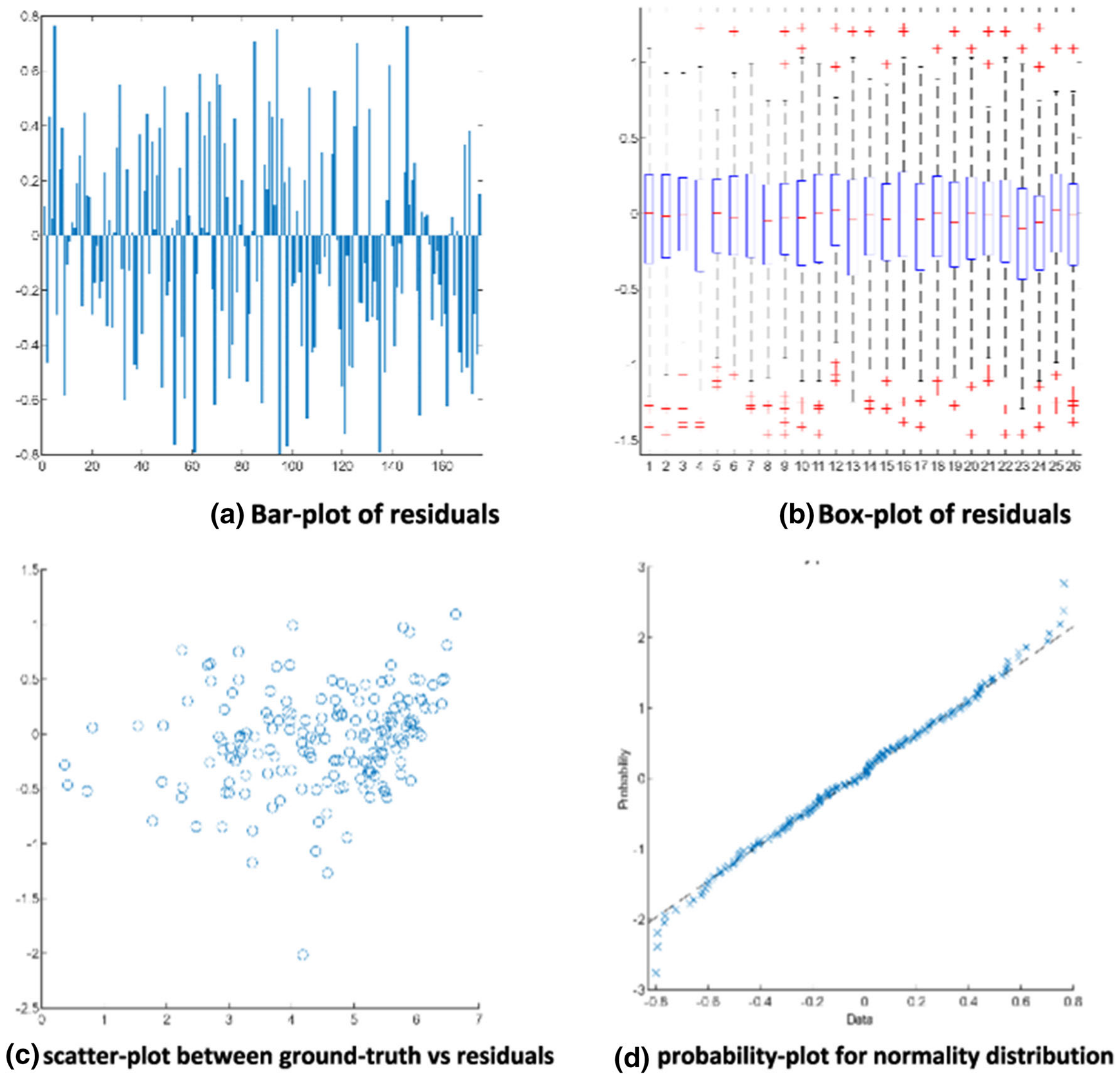
computationally expensive task and repeating it for a larger number of times is not feasible for us. The histograms of three of these performance parameters are provided for CSIQ and CID2013 in Figs. 11 and 12, respectively. It can be noted from Fig.s that the performance metrics are close to the median value and follow nearly a normal distribution.

The scatter-plot of ground-truth versus predicted values along with regression line fitting is provided in Fig. 13. The scatter plots are provided for the train-test split with median value of PLCC. In CSIQ database Fig. 13a, 175 samples are used for testing with most of the samples very closely predicted to their ground-truth scores. In CID2013 database Fig. 13b 93 samples are used for testing with samples evenly distributed along the regression line.

The residuals play an important role in the analysis of a regression model. We have chosen CSIQ image database to perform the residual analysis, whereas similar results were obtained using other databases. The bar-plot as depicted in Fig. 14a is a useful method to visualize the magnitude and direction of the residuals and can be used for small to medium numbers of samples. The box-plot can provide the mean and the distribution of residual values in different iterations. Figure 14b provides the box-plot for 25 iterations drawn randomly for better visualization. It can be seen that the variation in the mean value among different iterations does not vary largely, and the absolute magnitude goes up to 1.5, which is even less in the positive direction. The scatter-plot between ground-truth on horizontal axes and residual magnitude on vertical axes is a very important graph that helps in understanding the model's behavior. A skewness in this distribution such as the points following a certain shape such as a cone or parabola or skewed to the lower or the upper extreme, indicates some limitations in the modeling process. A random distribution that does not follow any clear pattern indicates the model is reasonable, and the same is indicated in Fig. 14c. Another important measure in the residual analysis is that the residuals should follow a normal distribution. We have plotted a probability-plot for normal distribution on the residuals in Fig. 14d and it can be observed that most of the points fall on the line except for very few deviations. It indicates that the residuals follow a normal distribution for the proposed model.

4.3 Comparison with existing approaches

This section provides a comparison with the existing approaches. We have performed comparison with seventeen existing approaches and some of them are discussed in related work due to similarity of their methodology with the DeepEns. It is worth mentioning that most of the top-performing approaches are based on deep learning and their contribution lies in architecture or training methodology. Moreover, we have performed comparisons on four benchmark datasets, but not all the techniques have reported results on these datasets. Moreover, the reported results are as per the respective author's claim and they have not been computed again due to constraints of time and availability of source code. The reported score is provided for up to four decimal places for the DeepEns models, but some of the approaches have reported scores up to two decimal places, and they are mentioned as it is. RMSE is not reported by most of the approaches, therefore, it is eliminated from the comparison and only PLCC and SROCC are mentioned in Table 5. Scores for TID2008 are reported by only Kang et al. (2015) with values of 0.903 for PLCC and 0.999 for SROCC and, therefore, not mentioned in the table. It can be seen that DeepEns has provided the highest performance in all the categories, whereas



(a) Bar-plot of residuals

(b) Box-plot of residuals

(c) scatter-plot between ground-truth vs residuals

(d) probability-plot for normality distribution

Fig. 14 **a** CSIQ Database: Bar-plot of residuals, **b** box-plot of residuals for 25 iterations, **c** scatter-plot between ground-truth versus residuals and probability-plot for normal distribution

DeepBIQ (Bianco et al. 2018) stands as the second-best approach. The results for PWRC and comparison of existing approaches are provided in Table 6, and the scores of existing approaches are taken from Wu et al. (2018).

4.4 Statistical significance test

The comparison of DeepEns with existing approaches is performed in Table 7. To avoid the confusion of superiority of one approach with the other, we have performed a statistical significance test with one variable *t*-test. The hypothesis testing is performed with a 95% confidence interval with the null hypothesis stated as “the mean value of correlation coefficient of row algorithm is greater than the value of column algorithm. In Table 7, a value of ‘0’ indicates an indistinguishable scenario, whereas a ‘1’ indicates that the row algorithm is superior to the column algorithm and ‘− 1’ otherwise.

4.5 Many-vs-one cross-dataset evaluation

Evaluation of an approach through cross-validation is a good measure of its performance. Image quality assessment databases are relatively small-sized, have followed constrained subjective scoring experiments and, therefore, exhibit a lot of variation. Moreover, the databases with simulated distortion such as LIVE, TID2013 and CSIQ contain specific distortion categories with a discrete level of degradation on a set of reference images and, therefore, are not the ideal candidate for evaluation of image quality assessment. We have therefore proposed to train the IQA algorithm on simulated distortion databases (i.e., LIVE,

Table 5 Comparison of the DeepEns with existing approaches

Sr	Dataset	LIVE		TID2013		CSIQ	
	Metric » → ▶	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
1	Kang et al. (Kang, et al. 2014)	0.953	0.956	–	–	–	–
2	DeepBIQ (Bianco et al. 2018)	0.98	0.97	0.96	0.96	0.97	0.96
3	BLINDER (Gao et al. 2018)	0.959	0.966	0.838	0.819	0.968	0.961
4	DIQA (Kim et al. 2018)	0.977	0.975	0.850	0.825	0.915	0.884
5	Ravela et al. (Ravela et al. 2019)	0.9492	0.9492	–	–	0.9445	0.9445
6	Meon (Ma et al. 2017a)	–	–	0.912	0.912	0.944	0.932
7	DB-CNN (Zhang et al. 2018b)	0.971	0.968	0.865	0.816	0.959	0.946
8	MCNN-IQA (Fan et al. 2018)	0.957	0.9531	–	–	0.894	0.8766
9	DIQaM-NR (Bosse et al. 2018)	0.972	0.960	0.855	0.835	–	–
10	WaDIQaM-NR (Bosse et al. 2018)	0.963	0.954	0.787	0.761	–	–
11	OG-IQA (Liu et al. 2016)	0.952	0.950	–	–	–	–
12	VIDGIQA (Guan et al. 2017)	0.973	0.969	–	–	–	–
13	Bosse et al. (Bosse et al. 2016)	0.972	0.960	–	–	–	–
14	Bare et al. (Bare et al. 2017)	0.974	0.971	–	–	–	–
15	dipIQ (Ma et al. 2017b)	0.958	0.957	0.894	0.877	0.949	0.93
16	HOSA (Xu et al. 2016)	0.950	0.952	0.952	0.959	0.930	0.948
17	BIECON (Kim and Lee 2017)	0.962	0.958	0.765	0.721	0.838	0.825
	DeepEns-Lite	0.9858	0.9812	0.9781	0.9774	0.9831	0.9803
	DeepEns	0.9860	0.9887	0.9874	0.9818	0.9846	0.9817

Table 6 Comparison of the DeepEns with existing approaches

	LIVE		TID2013		LiveCD	
	SROCC	PWRC	SROCC	PWRC	SROCC	PWRC
BIQI	0.784	4.195	0.674	3.701	0.522	9.193
BLIINDS II	0.925	5.362	0.839	5.04	0.489	8.593
BRISQUE	0.925	5.374	0.865	5.303	0.592	10.55
DIIVINE	0.9	5.17	0.775	4.495	0.576	10.413
NFERM	0.933	5.472	0.888	5.509	0.579	10.165
M3	0.951	5.659	0.89	5.491	0.598	10.479
TCLT	0.942	5.642	0.932	5.964	0.561	9.601
DeepEns-Lite	0.9781	6.8121	0.9761	6.4931	0.8863	11.0214
DeepEns	0.9887	8.2136	0.9818	8.3944	0.8985	12.0242

TID2013, CSIQ) and cross-validate them on image databases with natural distortion. It is to be noted that different image databases use different scales for subjective scoring; therefore, we have normalized the subjective score of all the training and validation databases from 0 to 1. We have three image databases with natural distortions. Two are benchmark databases and one is a self-collected image database. The results are reported in Table 8.

4.6 Experiments on BIQ2021

We have conducted three experiments for BIQ2021 image quality assessment database. In the first experiment, the synthetic distortion databases (LIVE, TID2013 and CSIQ) are used for training and the 2000 test images of BIQ2021 are used for testing. In the second experiment, naturally distorted databases (LIVE CD, CID2013 and KonIQ-10 k) are used for training and BIQ2021 test-set of 2000 images is used for testing. Whereas, the third experiment both synthetic and naturally distorted databases LIVE, CSIQ, TID2013, LIVE CD, CID2013, KonIQ-10 K, and 10,000 images of BIQ2021 are used for training and 2000 images of BIQ2021 are used for testing. The fourth experiment uses 10,000 images of BIQ2021 for training and 2000 images of BIQ2021 for validation. The results of these four experiments are reported in Table 9.

Table 7 One variable t-test to check statistical significance

	Kang, et al. 2014)	Bianco et al. 2018)	Gao et al. 2018)	Kim et al. 2018)	Ravela et al. 2019)	Ma et al. 2017a)	Zhang et al. (2018b)	Fan et al. 2018)	Bosse et al. 2018)	Bosse et al. 2018)	Liu et al. 2016)	Guan et al. 2017)	Bosse, et al. 2016)	Bare et al. 2017)	Ma et al. 2017b)	Xu et al. 2016)	Kim and Lee 2017)	Our
Kang, et al. (2014)	0	−1	−1	−1	1	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	−1	−1
Bianco et al. (2018)	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	−1
Gao et al. (2018)	1	−1	0	−1	1	−1	−1	1	−1	1	1	−1	1	−1	1	1	1	−1
Kim et al. (2018)	1	−1	1	0	1	−1	−1	1	−1	1	1	1	1	1	−1	−1	1	−1
Ravela et al. (2019)	−1	−1	−1	−1	0	1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	1	−1
Ma et al. (2017a)	0	−1	1	1	−1	0	−1	1	1	1	0	0	0	0	1	−1	1	−1
Zhang et al. (2018b)	1	−1	1	1	1	1	0	1	1	1	1	−1	1	1	1	1	1	−1
Fan et al. (2018)	1	−1	−1	−1	−1	−1	−1	0	−1	−1	1	−1	−1	−1	−1	−1	1	−1
Bosse et al. (2018)	1	−1	1	1	1	−1	−1	1	0	1	1	−1	0	−1	−1	−1	1	−1
Bosse et al. (2018)	1	−1	−1	−1	1	−1	−1	1	−1	0	1	−1	−1	−1	−1	−1	1	−1
Liu et al. (2016)	−1	−1	−1	−1	1	0	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	−1	−1
Guan et al. (2017)	1	−1	1	−1	1	0	1	1	1	1	1	0	1	−1	1	1	1	−1
Bosse, et al. (2016)	1	−1	−1	−1	1	0	−1	1	0	1	1	−1	0	−1	1	1	1	−1

Table 7 continued

	Kang, et al. 2014)	Bianco et al. 2018)	Gao et al. 2018)	Kim et al. 2018)	Ravela et al. 2019)	Ma et al. 2017a)	Zhang et al. (2018b)	Fan et al. 2018)	Bosse et al. 2018)	Bosse et al. 2018)	Liu et al. 2016)	Guan et al. 2017)	Bosse, et al. 2016)	Bare et al. 2017)	Ma et al. 2017b)	Xu et al. 2016)	Kim and Lee 2017)	Our
Bare et al. (2017)	1	−1	1	−1	1	0	−1	1	1	1	1	1	1	0	1	1	1	−1
Ma et al. (2017b)	1	−1	−1	1	1	−1	−1	1	1	1	1	−1	−1	−1	0	−1	1	−1
Xu et al. (2016)	−1	−1	−1	1	1	1	−1	1	1	1	1	−1	−1	−1	1	0	1	−1
Kim and Lee (2017)	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	−1	−1	−1	−1	−1	0	−1
Our	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

A. Ablation study

The ablation experiments are extensively incorporated into neuroscience to check these complex systems. In the area of artificial neural networks (ANNs), these experiments are done to check if all the parts of the architecture are necessary for the required performance. As the ANN are complex architectures and they do not have a set of rules for their construction, rather they follow heuristics for their construction. It is, therefore, necessary to verify if the designed architecture as a whole is necessary or if some part of the architecture can be removed with no decrease in the intended performance of the ANN. We have done the experimentations during the construction of the architecture with low epochs to select the most suitable architecture for the problem at hand; however, the final architecture is pruned at some bottleneck points to verify the efficacy of its key components. We have used CSIQ database for ablation study and tested three different subsets of the architecture for validation. The pruned architecture is retrained for 100 epochs with CSIQ database, and the experiment is repeated ten times to report the median values of PLCC and SROCC provided in Table 10.

In the first experiment, we have pruned the Vgg16 subset of the architecture and verified the model with agreed parameters. In the second experiment, we have pruned the AlexNet subset and performed the validation. In the last experiment, the flatten layer is directly connected to the regression output layer. Table 10 can be viewed for the performance of these three ablation experiments.

B. Time complexity

The experimental setup is described at the start of this section. The training of DeepEns for fine-tuning on image quality database BIQ2021 is performed for 100 epochs for a batch size of 16 and it took almost 168 h. Whereas, in the testing phase, it takes 2.77 s per image. The testing is reported based on CPU only.

5 Discussion

The image quality assessment is a challenging task due to its relative nature. The perceptual quality of an image depends on several parameters and is largely influenced by the nature of the content in the image. A large number of researchers have proposed different statistical and other features which are affected by a change in image quality but they are unable to fully cover the complexities of the factors affecting the image quality. Natural scene statistics-based quality assessment algorithms designed for a certain

Table 8 Training on LIVE, CSIQ and TID2013 and testing on natural distortion databases

Dataset →	CID2013		Live CD		KonIQ-10 K		BIQ2021	
Metric →	PLCC	SROCC	PLCC	PLCC	PLCC	SROCC	PLCC	SROCC
DeepEns	0.6977	0.6762	0.6054	0.6878	0.6862	0.6917	0.6878	0.6724

Table 9 Cross-dataset testing on BIQ2021

Sr	Train set	Metric	BIQ2021 (Test set)
1	Synthetic distortion databases	PLCC	0.6878
		SROCC	0.6924
2	Natural distortion databases	PLCC	0.8118
		SROCC	0.8024
3	Synthetic distortion + natural distortion databases	PLCC	0.7427
		SROCC	0.7351
4	Train set of BIQ2021	PLCC	0.8098
		SROCC	0.7922

distortion type perform well only to those distortions and does not generalize to scenarios containing other types of distortions or a combination of distortions. Some datasets are designed with multiple distortions (Sun et al. 2017; Jayaraman et al. 2012) but they again have the same issue that the training algorithm learns specific distortions only. In the case of images with natural distortions such as (Virtanen et al. 2014), the algorithms trained on synthetic distortions do not perform well on these datasets as demonstrated in Sect. 4.5 and 4.7. Training and validation on the naturally occurring distortions may perform better but they are not the true representative of no-reference quality assessment scenario in general.

The proposed approach (DeepEns) has targeted the problem by training and testing it on five benchmark datasets. The results of individual dataset scenarios are reported and comparisons of synthetic distortion databases is made with the existing approaches. The comparison is made based on correlation coefficients. DeepBIQ (Bianco et al. 2018) is the second-best performing model which has used Caffe (Jia et al. 2014) architecture which is inspired from AlexNet as learning architecture and fine-tuned it on the dataset under test. Their architecture has performed exceptionally well on several benchmarking datasets. BLINDER (Gao et al. 2018) has used Vgg to extract deep activations at multiple levels and trained an ensemble to make a prediction and has reported second-highest scores for CSIQ dataset and comparable scores for other datasets. DIQA (Kim et al. 2018) have used a two-stage approach and reported second-highest scores for LIVE dataset and comparable scores for other datasets. One variable t-test is performed to check the statistical significance, and the DeepEns has provided superior performance. The DeepEns has used CNN base of two architectures performed feature pooling which is passed through a few fully connected layers with dropout. The proposed DeepEns architecture is retrained on the dataset under test until convergence and then validated. This deep ensembling strategy has provided a set of representations with diverse features. This architecture in a result provides a richer feature set that can learn quality-aware features. The pretraining of the proposed architecture on the synthetic database and then retraining on BIQ2021 with natural distortion make it a high-performing model.

Table 10 Results of ablation study

Sr	Ablation experiment	RMSE	PLCC	SROCC
1	Pruning EfficientNet-base	0.1419	0.8473	0.8176
2	Pruning NasNet-base	0.1628	0.8688	0.8294
3	Pruning fully connected layers	0.1383	0.9233	0.8559

The cross-dataset experiment is conducted to train the model on synthetic distortion databases and validate them on three natural distortion databases to check for generalization. The results of this experiment are encouraging and indicate good generalization. We again credit the better generalization to the ensembling behavior of the DeepEns architecture and its retraining providing quality-aware representations. The cross-dataset comparisons are not made as DIQA (Kim et al. 2018) has performed a cross-dataset experiment by taking a subset of four distortion types only from CSIQ and TID2013 dataset as LIVE dataset contains only these four types of distortions. It is to highlight that the purpose of the cross-dataset experiment is to check the generalizability of the model on other distortion types and scoring methods. Some of the approaches such as DeepBIQ (Bianco et al. 2018) and BLINDER (Gao et al. 2018) have not reported such results altogether.

Experiments on BIQ202 are conducted to further validate the generalization of the proposed model. Some interesting observations in these experiments are:

1. DeepEns trained on synthetic distortion databases generalize very poorly on BIQ2021 test-set as the training dataset contains a fixed number of simulated distortions which are not representative of natural distortions.
2. DeepEns trained on natural distortion databases performed well on the test-set of BIQ2021 as the distortions available in the training set are closer to the distortions occurring in the test set.
3. The experiment based on the combination of synthetic distortion as well as natural distortion database performed better than synthetic distortion database alone but worse than the natural distortion database.
4. The last experiment where the train set of BIQ2021 is used for training and the test-set is used for validation has provided the highest performance which is slightly superior to natural distortion databases. The reason could be that the train and test sets of this database are scored by the same people in the similar type of subjective scoring experiments.

It is, therefore, evident from these experiments that the synthetic distortion database is not representative of naturally occurring image distortions. Synthetic distortion databases, although widely available, can be used to assess image quality in scenarios where similar types of distortions are occurring, such as assessment of image quality of compression algorithms. Still, this type of comparison will be less suitable as the model trained on compressed images will be biased towards existing compression schemes used in mode training. Whereas the natural distortion database having distortions added during the process of image acquisition, storage and post-processing are better representatives. Distortions in these images occur due to hardware limitation, compression for storage purposes, or post-processing for enhancement, cropping, or adjustment. Therefore, large-scale image databases such as BIQ2021 are a better candidate to train or judge the performance of a general-purpose image quality assessment model.

6 Conclusion and future work

In this study, we have explored the problem of general-purpose image quality assessment. We have proposed a deep ensemble-based architecture DeepEns which is first trained on a synthetic database of 18.75 million images and then retrained on a natural distortion database of 12,000 subjectively scored images. The proposed approach has performed consistently better in comparison to the existing approaches in a single database experiment. In a cross-dataset experiment, we have trained the model on synthetic databases and validated it on two natural distortion benchmark databases and one self-collected natural distortion database. The results of these experiments indicate that the DeepEns is robust and performs well when trained on a representative image database. To make a general-purpose image quality assessment algorithm, the DeepEns is trained on three natural distortion databases and validated on an independent subset of natural distortion databases. The final trained model (DeepEns) and the self-collected database (BIQ2021) will be made publicly available for comparison and benchmarking.

6.1 Future work

In the future work we will address the following directions:

a. Improvement in the generalization of the trained model by training on a combination of different natural distortion databases after realignment and normalization of their subjective scores.
b. Making specific design decisions to further improve the generalization performance of end-to-end trained CNN architecture.
c. Use of quality-aware or visual perception-based loss function for model training.

Appendix

Loss functions

Choice of the loss function is important while training a regression algorithm. IQA is a regression problem, therefore, careful selection of suitable loss function is important. We have experimented with six loss functions. Although, the loss function can be chosen intuitively among the candidates but experimenting with different loss functions is a better way. The details of the six loss functions and the training progress are provided from Figs. 15, 16, 17, 18, 19 and 20. The training is performed by using the LiveCD database and trained for 30 epochs.

Mean squared error (MSE)

MSE is the most commonly used loss function for regression. It is a sum of the squared distances between the target and predicted values and is calculated by:

$$\mathrm{MSE} = \frac{1}{n}\sum_{n}(T - P)^2$$

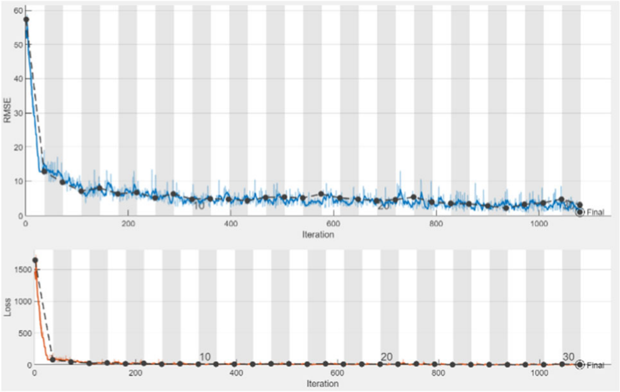


Fig. 15 Training progress using MSE as loss function

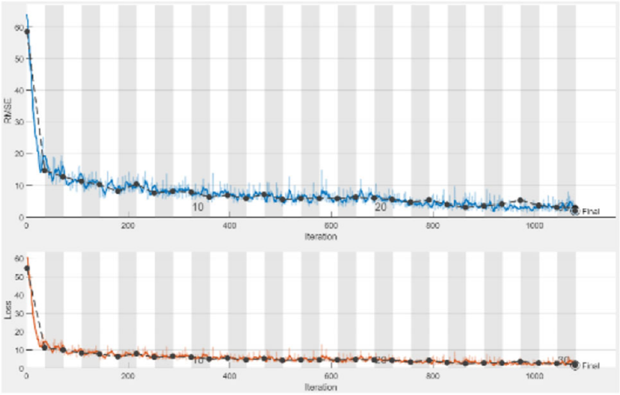


Fig. 16 Training progress using MAE as loss function

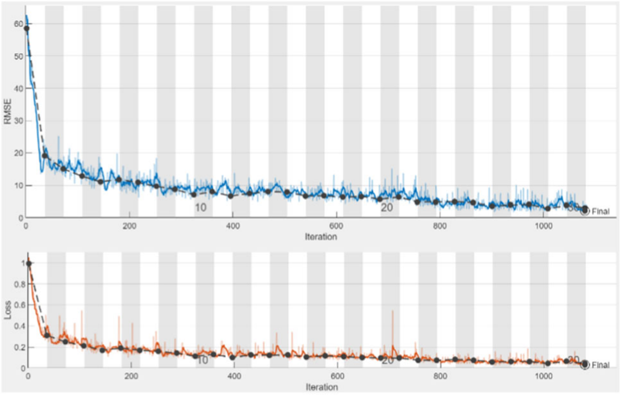


Fig. 17 Training progress using MAPE as loss function

Mean absolute error (MAE)

MAE is another useful loss function and is robust to outliers. It measures the directionless magnitude of the errors by taking the sum of absolute differences between target and predictions. It is calculated by the following formula:

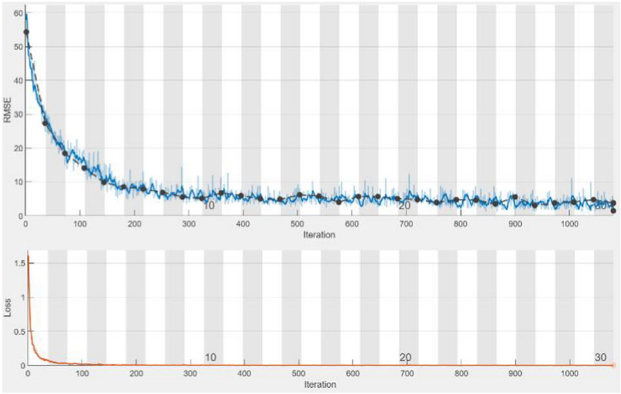


Fig. 18 Training progress using MSLE as loss function

$$\mathrm{MAE} = \frac{1}{n}\sum_{n}|T - P|$$

Mean absolute percentage error (MAPE)

MAPE provides the percentage magnitude of the error which is calculated by dividing the error calculated by MAE by the target value. Although it is simple and convincing, it has the drawback of not being useful when there are target values of zero. Moreover, it puts havier penality on negative errors where the forecasted value is higher than the actual value and, therefore, it provides outcomes that are less than the target value. The formula for calculation of MAPE is provided below:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{n}\left|\frac{T - P}{T}\right|$$

Mean squared logarithmic error (MSLE)

MSLE is a modified version of MSE and acts as a measure of the ratio between the target and the predicted values. As it is a ratio and cares more about the percentage difference between the target and predicted values, so it treats large and small differences similarly. Moreover, it is also asymmetric like MAPE but it favors larger predictions more than smaller predictions. The following formula is used to calculate MSLE:

$$\mathrm{MSLE} = \frac{1}{n}\sum_{n}(\log(T + 1) - \log(P + 1))^2$$

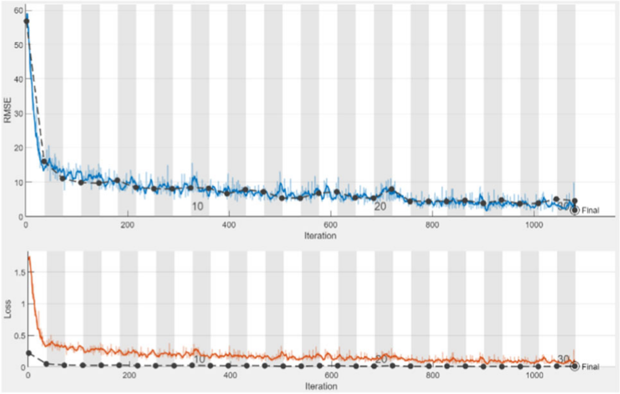


Fig. 19 Training progress using Hubber loss

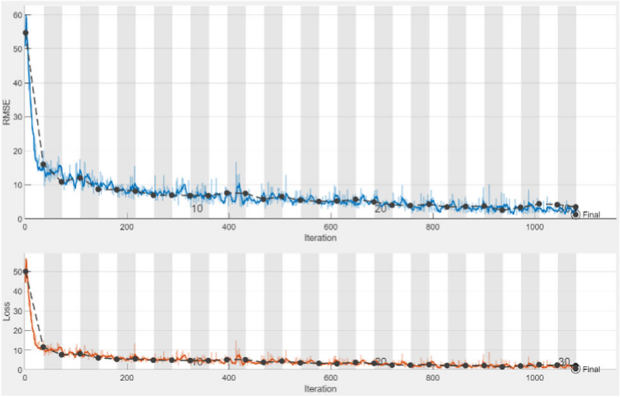


Fig. 20 Training progress using LogCosh loss

Huber loss

It is an alternative loss function to MSE and MAE which combines the good properties of both of them and is differentiable at 0. For small prediction error, it acts like MSE and for larger prediction error, it acts like MAE and therefore is robust to outliers and provides better convergence when the loss is near minima. The drawback with using hubber loss is that we have to tune the hyperparameter δ as its value will define the choice of piecewise function. The hubber loss can be calculated using the following formula, and it is to be noted that we have not tuned it for hyperparameter δ and used '1' as its value:

$$\text{Huber} = \begin{cases} \frac{1}{2}(T-P)^2 & \text{for}(T-P) \leq \delta \\ \delta|T-P| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

LogCosh loss

It measures the logarithm of the hyperbolic cosine of the prediction error and is smoother than MSE. It acts like MSE for smaller prediction errors and is not strongly affected by an intermittent large prediction error and is therefore advantageous to Hubber loss. The formula for calculation of LogCosh loss is provided below:

$$\text{LogCosh} = \sum_{n} \log(\cosh(P-T))$$

Author contributions The contributions of the authors are listed below: NA: Conception, coding, experimentations, and writeup. HMSA: Conception, supervision, Proof Read and review. ARB: Proof Read and review. AK: Proof Read and review.

Funding Not applicable.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Data availability The data is currently available in GitHub repository: https://github.com/nisarahmedrana/DeepEns

Code availability https://github.com/nisarahmedrana/DeepEns

References

Ahmed N, Asif HMS (2019) Ensembling convolutional neural networks for perceptual image quality assessment. In: 2019 13th international conference on mathematics, actuarial science, computer science and statistics (MACS). IEEE

Ahmed N, Asif HMS (2020) Perceptual quality assessment of digital images using deep features. Comput Inform 39(3):385–409

Ahmed N, Asif HMS, Khalid H (2019) Image quality assessment using a combination of hand-crafted and deep features. In: International conference on intelligent technologies and applications. Springer

Ahmed N, Asif HMS, Khalid H (2021) PIQI: perceptual image quality index based on ensemble of Gaussian process regression. Multimed Tools Appl. p. 1–24

Bare B, Li K, Yan B (2017) An accurate deep convolutional neural networks model for no-reference image quality assessment. In: 2017 IEEE international conference on multimedia and expo (ICME). IEEE

Bianco S et al (2018) On the use of deep learning for blind image quality assessment. SIViP 12(2):355–362

Bosse S et al (2017) Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans Image Process 27(1):206–219

Bosse S et al (2018) Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans Image Process 27(1):206–219

Bosse S, et al (2016) A deep neural network for image quality assessment. In: 2016 IEEE international conference on image processing (ICIP). IEEE

Chandler DM (2013) Seven challenges in image quality assessment: past, present, and future research. ISRN Signal Process

Fan C et al (2018) No reference image quality assessment based on multi-expert convolutional neural networks. IEEE Access 6:8934–8943

Gao F et al (2018) Blind image quality prediction by exploiting multi-level deep representations. Pattern Recogn 81:432–442

Ghadiyaram D, Bovik AC (2015) Massive online crowdsourced study of subjective and objective picture quality. IEEE Trans Image Process 25(1):372–387

Guan J et al (2017) Visual importance and distortion guided deep image quality assessment framework. IEEE Trans Multimed 19(11):2505–2520

Guo P, Anderson C, Farrell R (2019) On global feature pooling for fine-grained visual categorization

Jayaraman D, et al (2012) Objective quality assessment of multiply distorted images. In: 2012 Conference record of the forty sixth asilomar conference on signals, systems and computers (ASILOMAR). IEEE

Jia Y, et al (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia. ACM

Kang L, et al (2014) Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Kang L, et al Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE international conference on image processing (ICIP). IEEE

Khalid H, Ali M, Ahmed N (2021) Gaussian process-based feature-enriched blind image quality assessment. J vis Commun Image Represent 77:103092

Kim J, Lee S (2017) Fully deep blind image quality predictor. IEEE J Sel Top Signal Process 11(1):206–220

Kim J, Nguyen A-D, Lee S (2018) Deep CNN-based blind image quality predictor. IEEE Trans Neural Netw Learn Syst 30(1):11–24

Larson EC, Chandler D (2010) Categorical image quality (CSIQ) database

Lin H, Hosu V, Saupe D (2019) KADID-10k: a large-scale artificially distorted IQA database. In: 2019 eleventh international conference on quality of multimedia experience (QoMEX). IEEE

Liu L et al (2016) Blind image quality assessment by relative gradient statistics and adaboosting neural network. Signal Process Image Commun 40:1–15

Ma K et al (2017a) End-to-end blind image quality assessment using deep neural networks. IEEE Trans Image Process 27(3):1202–1213

Ma K et al (2017b) dipIQ: blind image quality assessment by learning-to-rank discriminable image pairs. IEEE Trans Image Process 26(8):3951–3964

Ma X, Jiang X (2019) Multimedia image quality assessment based on deep feature extraction. Multimed Tools Appl 1–12

Ponomarenko N et al (2009) TID2008-a database for evaluation of full-reference visual quality assessment metrics. Adv Mod Radioelectron 10(4):30–45

Ponomarenko N et al (2015) Image database TID2013: peculiarities, results and perspectives. Signal Process Image Commun 30:57–77

Prabha DS, Kumar JS (2017) An efficient image contrast enhancement algorithm using genetic algorithm and fuzzy intensification operator. Wirel Pers Commun 93(1):223–244

Ravela R, Shirvaikar M, Grecos C (2019) No-reference image quality assessment based on deep convolutional neural networks. In: Real-time image processing and deep learning. International Society for Optics and Photonics

Saad MA, Bovik AC, Charrier C (2012) Blind image quality assessment: a natural scene statistics approach in the DCT domain. IEEE Trans Image Process 21(8):3339–3352

Selva Nidhyanandhan S, Sindhuja R, Selva Kumari RS (2020) Double stage gaussian filter for better underwater image enhancement. Wirel Pers Commun 114:2909–2921

Sheikh HR, Bovik AC, Cormack L (2005a) No-reference quality assessment using natural scene statistics: JPEG2000. IEEE Trans Image Process 14(11):1918–1927

Sheikh HR, Bovik AC, De Veciana G (2005b) An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans Image Process 14(12):2117–2128

Sheikh H (2005) LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality

Sun W, Zhou F, Liao Q (2017) MDID: a multiply distorted image database for image quality assessment. Pattern Recogn 61:153–168

Tan M, QV Le (2019) Efficientnet: rethinking model scaling for convolutional neural networks. arXiv preprint http://arxiv.org/abs/1905.11946

Virtanen T et al (2014) CID2013: a database for evaluating no-reference image quality assessment algorithms. IEEE Trans Image Process 24(1):390–402

Wang Z et al (2004a) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612

Wang Z, Bovik AC, Lu L (2002) Why is image quality assessment so difficult?. In: 2002 IEEE international conference on acoustics, speech, and signal processing. IEEE

Wang Z, Simoncelli EP, Bovik AC(2003) Multiscale structural similarity for image quality assessment. In: Signals, systems and computers, 2004. Conference record of the thirty-seventh asilomar conference on. IEEE

Wu Q et al (2018) A perceptually weighted rank correlation indicator for objective image quality assessment. IEEE Trans Image Process 27(5):2499–2513

Xu J et al (2016) Blind image quality assessment based on high order statistics aggregation. IEEE Trans Image Process 25(9):4444–4457

Zhang R et al (2018a) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Zhang W et al (2018b) Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans Circuits Syst Video Technol 30(1):36–47

Zoph B et al (2018) Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition

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