Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A multimodal dense convolution network for blind image quality assessment#

Nandhini CHOCKALINGAM, Brindha MURUGAN‡

Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli 620015, India

E-mail: cn.nandhini@gmail.com; brindham@nitt.edu

Received Nov. 2, 2022; Revision accepted Apr. 24, 2023; Crosschecked Sept. 20, 2023

Abstract: Technological advancements continue to expand the communications industry's potential. Images, which are an important component in strengthening communication, are widely available. Therefore, image quality assessment (IQA) is critical in improving content delivered to end users. Convolutional neural networks (CNNs) used in IQA face two common challenges. One issue is that these methods fail to provide the best representation of the image. The other issue is that the models have a large number of parameters, which easily leads to overfitting. To address these issues, the dense convolution network (DSC-Net), a deep learning model with fewer parameters, is proposed for no-reference image quality assessment (NR-IQA). Moreover, it is obvious that the use of multimodal data for deep learning has improved the performance of applications. As a result, multimodal dense convolution network (MDSC-Net) fuses the texture features extracted using the gray-level co-occurrence matrix (GLCM) method and spatial features extracted using DSC-Net and predicts the image quality. The performance of the proposed framework on the benchmark synthetic datasets LIVE, TID2013, and KADID-10k demonstrates that the MDSC-Net approach achieves good performance over state-of-the-art methods for the NR-IQA task.

Key words: No-reference image quality assessment (NR-IQA); Blind image quality assessment; Multimodal dense convolution network (MDSC-Net); Deep learning; Visual quality; Perceptual quality

https://doi.org/10.1631/FITEE.2200534 **CLC number:** TP39

1 Introduction

Information and communication technology (ICT) has permeated almost every aspect of human life, helping people to collaborate, exchange knowledge, and learn more effectively. Multimedia technologies for online education, surveillance, on-demand video streaming, high-definition televisions, social media, and video chat have gained popularity, resulting in vast numbers of digital images and videos. The rapid growth of multimedia applications demands a high quality of experience (QoE) for end users. The perceptual image quality is critical to the performance of these applications. On the other hand, digital images are subjected to distortion at different stages, including acquisition, digitization, transmission, noise processing, and display, which degrades the human visual experience. Therefore, predicting the image quality is essential to improve the image content delivered to end users. Image quality assessment (IQA) metrics aid in the identification of low-quality photographs and their removal from galleries and other sources. Furthermore, real-time monitoring during acquisition may help remove annoying distortions by setting optimum parameters.

Over the last decades, researchers have been developing various IQA metrics to predict perceptual image quality. Subjective assessment deals with

‡ Corresponding author

Electronic supplementary materials: the online version of this article (https://doi.org/10.1631/FITEE.2200534) contains supplementary materials, which are available to authorized users

ORCID: Nandhini CHOCKALINGAM, https://orcid.org/0000-0003-4767-9682; Brindha MURUGAN, https://orcid.org/0000-0002-3952-0674

images evaluated by human subjects, following a standard set of procedures. The mean opinion score (MOS) obtained from different participants over several viewing sessions seems to be the most reliable and accurate metric. However, subjective assessment is laborious, time-consuming, and expensive. Furthermore, it cannot be used directly as an optimization metric for real-time multimedia systems. Objective assessments, which are typically trained using subjective assessment data, can automatically estimate image quality and are ideal for real-time systems due to their performance evaluation and optimization. The three types of objective IQAs are as follows: full-reference IQA (FR-IQA) requires the full pristine image to evaluate the distorted image, reduced-reference IQA (RR-IQA) requires less detail from the reference image, and no-reference IQA (NR-IQA) does not need any reference image to access the image quality. Because reference images are typically unavailable in many scenarios, NR-IQA algorithms are useful for a variety of practical applications. As a result, the scientific community is interested in NR-IQA approaches.

Deep convolutional neural network (DCNN) approaches have been widely used in recent years and have achieved considerable success in many computer vision tasks, including image recognition (Krizhevsky et al., 2012; He et al., 2016; Ding et al., 2019), visual tracking (Ding et al., 2018), and social image comprehension (Li ZC et al., 2019). Traditional approaches for NR-IQA use natural scene statistics (NSS) based features in most of the successful approaches. NSS features are extracted in image transformation domains using wavelet transform or discrete cosine transform (DCT). These approaches are computationally expensive and too slow. Kang et al. (2014) were the first to apply convolutional neural networks (CNNs) for IQA and achieved excellent results, spawning a multitude of DCNN-based NR-IQA approaches (Mittal et al., 2012; Li QH et al., 2016; Bosse et al., 2018; Ma KD et al., 2018; Zhang WX et al., 2020). Inspired by the success of DenseNet, Huang et al. (2017) proposed a DenseNet-based network model to efficiently learn both the high- and low-level spatial features using a multi-layer network.

We analyze two different frameworks based on DenseNet in this paper. Deep networks can significantly increase the number of parameters to train. The increase in parameters leads to overfitting in the absence of massive data. As the dataset for an IQA task contains limited images, dense convolution network (DSC-Net), an efficient and simplified DenseNet framework with four dense blocks for semantic feature representation is leveraged for the IQA task. DSC-Net is much deeper compared to other shallow networks used for IQA tasks and is capable of generating a large number of reusable feature maps by dense connections with fewer parameters. DSC-Net uses a patch-based approach to augment the data, with each patch having equal importance and having the same MOS as the image.

Texture is essential in human visual perception because it helps object recognition. Identifying various textures is a simple task for human eyes, but it is more difficult when applied to computer vision tasks. To obtain a good feature representation, images with different textures are analyzed. Images with uniform and non-uniform textures are significantly different in their intensity distribution along the RGB channels. As shown in Fig. 1, the intensity values of non-uniform texture images are widely distributed compared to those of the uniform regions.

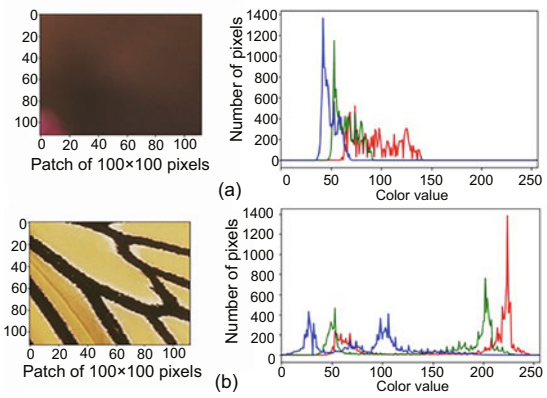

Fig. 1 Intensity distribution of uniform (a) and non-uniform (b) texture patches of 100×100 pixels

Moreover, the texture is stochastic in nature and is a function of the spatial variance of brightness and intensity of the pixels in an image. Hence, using their statistical properties is a reasonable choice for extracting the texture features. The gray-level co-occurrence matrix (GLCM) approach is one of the most commonly used statistical methods for extracting texture features. As a result, multimodal dense convolution network (MDSC-Net) introduces a feature fusion strategy that allows for a combined

regression of the multimodal features extracted from texture and spatial image representations. In summary, the main contributions are three-fold:

1. A coherent DSC-Net framework is proposed, which makes excellent use of feature reuse and significantly reduces the number of parameters to train.

2. The MDSC-Net model, which fuses texture features with spatial features from DSC-Net, is introduced to represent the structural information more efficiently.

3. The efficiency of the proposed system is analyzed on the LIVE, TID2013, and KADID-10k datasets, and a cross-dataset evaluation proves the generalization capability of the system.

2 Related works

The majority of the NR-IQA algorithms are concerned with NSS and deep learning based approaches. DIVINE (Moorthy and Bovik, 2011) and BLINDS-II (Saad et al., 2012) are two examples of popular NSS-based NR-IQA algorithms. The wavelet transform (Moorthy and Bovik, 2011) and the DCT (Saad et al., 2012) are often used to extract traditional NSS-based features in image transformation domains. Due to the use of image transformations, these techniques are typically slow and computationally expensive. In the absence of reference images, Sheikh et al. (2003) used statistical features derived from natural images to determine the image quality.

BRISQUE (Mittal et al., 2012) and CORNIA (Ye et al., 2012) encouraged the extraction of features from the spatial domain, which reduced computation time significantly. CORNIA (Ye et al., 2012) and HOSA (Xu et al., 2016) showed that instead of using handcrafted features, it is possible to learn discriminant image features directly from the raw image pixels using codebooks to estimate the visual quality scores. However, the codebook is created using complex K-means clustering on local features. Liu LX et al. (2014) used spectral and spatial entropies to access the image quality. Gu et al. (2018) extracted 17 features from image data sets by analyzing contrast, sharpness, brightness, and other factors, and then used a regression module to produce a measure of visual quality using big-data training samples. In LBIQ (Tang et al., 2011), a set of low-level features from NSS and texture statistics are combined and a kernel support vector machine (SVM) is used to pre-train a deep belief network that works well as a predictor of image quality. The aforementioned approaches depend on a subset of features that are expected to capture important factors affecting image quality, but determining which features are better for IQA tasks is difficult and requires domain knowledge.

According to human visual system (HVS) modeling (Lu et al., 2005), certain areas of an image have more importance than other areas. Taking this into consideration, some works (Wang and Shang, 2006; Zhang P et al., 2015) explored different weighting strategies for local quality estimation. Zhang P et al. (2015) suggested a semantic information-based NR-IQA algorithm. The features based on the image's semantic obviousness and the local characteristics are combined to access the image quality. However, it has an additional overhead of using the binarized normed gradients (BING) object detector to extract the object-like features which have poor object proposal quality. NRVPD (Wu et al., 2019) used an orientation similarity pattern based modeling that extracts both the visual content and quality degradation patterns using a support vector regression (SVR) for NR-IQA.

Nowadays, deep learning has been successfully applied to a variety of applications (Song et al., 2016; Qiu et al., 2018; Zhang SQ et al., 2018; Gu et al., 2020, 2021a, 2021b) and achieved remarkable success. Kang et al. (2014) were the first to apply a CNN to the NR-IQA. The CNN was trained with normalized small image patches and the whole image quality was predicted by averaging the scores of all the patches. This method assumes that each local patch quality is the same as the global image quality, which is not true, and their network is also too shallow. Kang et al. (2015) further extended their work to reduce the number of parameters and simultaneously estimated the distortion type. Bosse et al. (2016) fine-tuned a CNN network to improve the prediction score by weighting the patches based on image saliency. Kim and Lee (2017) trained a CNN model based on patch extraction, in which the quality scores were estimated from an FR-IQA model and then the network was fine-tuned using mean and standard deviation statistics of the local features. Despite encouraging performance, these approaches have complex architectures

and the evaluation is based on smaller subsets from the IQA dataset. Cheng et al. (2017) trained a CNN and an adaptive strategy using the saliency map, which assigns higher weights to salient patches. RankIQA (Liu XL et al., 2017) trained a siamese network with reference and distorted images and using that knowledge, the network is fine-tuned for single image quality prediction. Multitask end-to-end optimized deep neural network (MEON) (Ma KD et al., 2018) is a multitask framework that identifies the type of distortion, and using that pre-trained layer and distortion type, a quality prediction network is trained. Moreover, the generalized divisive normalization layer was used as the activation function instead of rectified linear unit (ReLU), which significantly reduced model parameters. However, it used the weights from synthetically generated images for different distortion types. Bianco et al. (2018) used different CNN design options, and the CNN features that had previously been trained on the image classification task were taken as inputs to train a quality predictor using SVR. However, their method involves manual parameter settings which are not optimized.

DB-CNN (Zhang WX et al., 2020) models synthetic and authentic distortions used two features, an ImageNet (Deng et al., 2009) trained VGG-16 (Simonyan and Zisserman, 2014) model and bi-linear pooling (Lin TY et al., 2015) for quality prediction. It also used distortion type and distortion level information for training to obtain better results. The ImageNet dataset includes high-quality photographs intended for object classification and contains different forms of distortions from IQA task.

VIDGIQA (Wu et al., 2019) made use of the distortion information and multitask learning procedure to learn features. In addition to estimating the visual quality, with the help of the distortion classification task, the accuracy of visual quality estimation was improved. EI-IQA (Yang et al., 2020) learned the quality score using a dual-stream network that used wavelet kurtosis features and VGG-16-based spatial features. However, the computation complexity for extracting wavelet features is significantly high. Po et al. (2019) proposed a framework, in which homogenous patches were ignored based on a variance threshold during training and the quality score was based only on the patches with complex structures. NCMQA (Zhou et al., 2020) introduced a neighborhood co-occurrence matrix to extract statistical features and the SVR was trained to predict image quality scores. The dataset amalgamation method (Zhang WX et al., 2021) has been shown to be effective in handling generalization, but it is limited in terms of adaptability to new datasets and computation scalability. RankIQA (Liu XL et al., 2017) used a two-step approach that generates a vast number of images by introducing different levels/types of distortions and ranks them according to their distortion type and level of distortion. Then, the network was trained to learn these rankings using synthetically generated images. Using pre-trained weights, the system was fine-tuned on the NR-IQA dataset. VCRNet (Pan et al., 2022) used the non-adversarial model, which relied on the relationship between the distorted image and its restored image to accurately predict the image quality.

Many works try to model the HVS (Lu et al., 2005; Zhang P et al., 2015), and others (Kang et al., 2014; Bosse et al., 2016; Chockalingam and Murugan, 2023) try to make the best use of the limited number of images in the IQA datasets by learning from small patches of the images as individual samples. Some approaches (Bianco et al., 2018; Zhang SQ et al., 2018; Li ZC et al., 2019) try to fine-tune a pre-trained network on a large dataset for other tasks to transfer the knowledge to the IQA tasks, while other frameworks (Liu XL et al., 2017; Lin HH et al., 2019; Wu et al., 2019) augment the dataset, thereby increasing the labeled training samples. In this paper, we adapt learning from patches and make use of multimodal features to model the HVS to improve the performance of NR-IQA tasks.

3 Proposed work

The proposed approach, which employs a DSC-Net convolutional network to predict the visual image quality, is discussed in this section. It begins by laying a general framework for the proposed approach, and then describes the functionality of each module. This section begins by describing how the DSC-Net architecture can be used to learn from image patches, and also explores how multimodal features affect the performance of the IQA task.

The proposed scheme is shown in Fig. 2. The non-overlapping image patches are generated for a given image. The spatial features are extracted using DSC-Net, which is fused with texture features

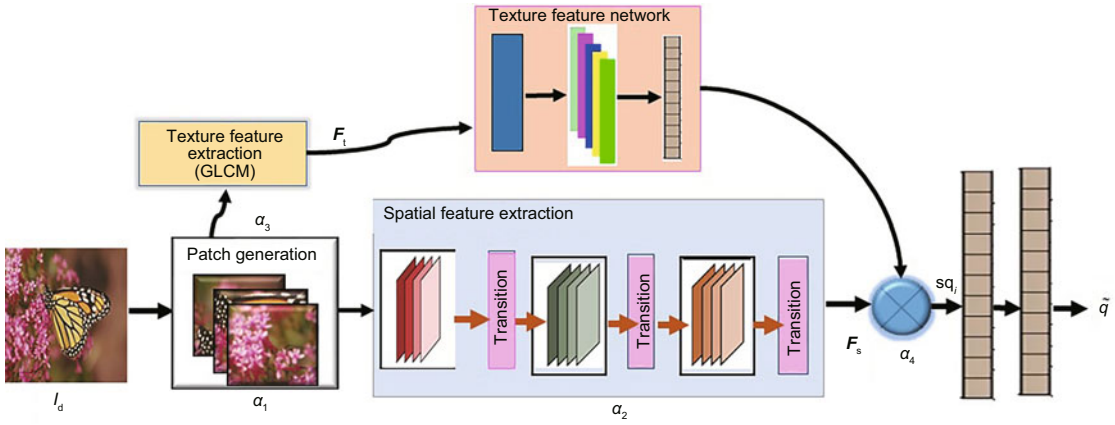

Fig. 2 General framework of the proposed work

extracted using GLCM. The fully connected layers at the end are used as regressors to estimate the image quality.

For a given distorted input image I_{d}, its perceptual quality $\tilde{q}$ can be predicted by a quality estimation model ψ:

$$\tilde{q} = \psi(I_{\mathrm{d}}; \beta), \tag{1}$$

where β is the model parameter. The proposed framework ψ consists of four major elements, including a patch generation module (α_1), spatial feature extraction module (α_2), texture feature extraction module (α_3), and feature fusion and regression module (α_4).

Multiple non-overlapping image patches are generated for the input image I_{d} by the patch generation module. This module also applies local contrast normalization (LCN) to all the patches as a preprocessing step. This module (α_1) can be formulated as

$$I_{\mathrm{d}} \rightarrow \{P_1, P_2, \ldots, P_N\}. \tag{2}$$

The spatial feature extraction module extracts spatial features $\boldsymbol{F}_{\mathrm{s}} \in \mathbb{R}^{256}$ for each image patch P_i. This module is denoted as α_2 with parameter $\beta_1 \in \beta$.

$$\boldsymbol{F}_{\mathrm{s}} = \alpha_2(P_i; \beta_1),\ i = 1, 2, \ldots, N. \tag{3}$$

The DSC-Net model learns the perceptual quality sq_i of the image, using the spatial features. The MDSC-Net model extends the DSC-Net model with the addition of two more modules. The texture feature extraction module extracts texture features using GLCM $\boldsymbol{F}_{\mathrm{t}} \in \mathbb{R}^{60}$ for each image patch P_i. This module is used for all image patches and denoted as α_3 with parameter $\beta_2 \in \beta$.

$$\boldsymbol{F}_{\mathrm{t}} = \alpha_3(P_i; \beta_2),\ i = 1, 2, \ldots, N. \tag{4}$$

The feature fusion and regression module concatenates the spatial and texture features and regresses the quality score sq_i for each patch P_i. This module is denoted as α_4 with the network parameters $\beta_1, \beta_2 \in \beta$ from the spatial feature extraction and texture feature extraction modules, respectively.

$$[\mathrm{sq}_i] = \alpha_4(\boldsymbol{F}_{\mathrm{s}}, \boldsymbol{F}_{\mathrm{t}}; \beta_1, \beta_2),\ i = 1, 2, \ldots, N. \tag{5}$$

3.1 Patch generation (α_1)

The patch generation module is used to increase the number of training samples in this study. This module also deals with LCN as a pre-processing step. To train a deep learning network, the data available in publicly accessible image quality datasets are inadequate. As a result, cropping is used to increase the amount of training data in the proposed study. Cropping, rather than resizing, retains the image's perceptual consistency. As a result, a large number of patches are generated from different spatial positions to cover the object's local visual details.

The non-overlapping cropped patches can be of any size. However, the neural network typically accepts fixed size input, and the proposed implementation uses a patch size of 56×56. The total number of local image patches in an image will vary depending on the image size.

Therefore, the total number of non-overlapping cropped patches for each image is given by

$$N_{\mathrm{p}} = \left\lfloor \frac{w_{\mathrm{i}}}{w_{\mathrm{p}}} \right\rfloor \times \left\lfloor \frac{h_{\mathrm{i}}}{h_{\mathrm{p}}} \right\rfloor, \tag{6}$$

where $w_{\mathrm{i}} \times h_{\mathrm{i}}$ is the resolution of the image and $w_{\mathrm{p}} \times h_{\mathrm{p}}$ is the resolution of the local patches. $\lfloor * \rfloor$ means floor function, which inputs a real number x and gives the highest integer less than or equal to x. Once the patches are generated, LCN is applied to all the patches, as described in the supplementary materials, to make the network robust to illumination and contrast changes.

3.2 Spatial feature extraction ($\boldsymbol{\alpha_2}$)

A good set of features plays a significant role in any deep learning task, like IQA. The DenseNet architecture with more layers has shown excellent accuracy in various image classification tasks, and has good generalization ability. To overcome the vanishing gradient descent problem, DenseNet directly connects all layers with each other by concatenating its features, which reduces the computation. Hence, to extract the spatial features, DSC-Net, the network based on DenseNet, is explored for the first time for the IQA task.

The architecture of DSC-Net is shown in Fig. 3a and has components like DenseBlock and a transition layer. The DenseBlock contains a fixed number of dense layers with bottleneck connections to reduce the depth of feature maps, which significantly reduces the model parameters. The transition layer aggregates the output from the DenseBlock and helps reduce the size of the feature maps using average pooling.

The architecture of the DenseBlock is as shown in Fig. 3b. A DenseBlock contains a sequence of dense layers. A dense layer has feature maps from all its preceding layers in that block, which contributes to diversified features. Each dense layer also adds its own feature maps, thus building collective knowledge.

DSC-Net concatenates the feature maps between the subsequent layers. Let x_l and x_{l-1} denote the output of the l^{th} and $(l-1)^{\mathrm{th}}$ layers, respectively. Then produces the composite function $H(*)$. x_l is given as follows:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]). \tag{7}$$

The amount of information contributed by each layer is regulated by the growth rate (k) of the network. Hence, if each dense layer produces k feature maps, the l^{th} layer has $k_0 + k(l-1)$ feature maps, where k_0 is the number of channels in the input image. As the number of layers increases, the number of feature maps also increases, to reduce the depth of the feature maps, and a bottleneck layer with a 1×1 filter is introduced, which outputs $4 \times k$ feature maps. So, each dense layer in Fig. 3c has batch normalization (BN), ReLU, and convolutional filters with kernel size 1×1 followed by a BN–ReLU 3×3 convolution. In the proposed framework, four DenseBlocks are used, each with three layers, and the growth rate k of each layer is 8. Transition layers are introduced after the DenseBlock to down-sample the feature maps. The architecture of each transition layer is as shown in Fig. 3d, which contains a BN layer, ReLU activation followed by a 1×1 convolution filter, and a 2×2 average pooling layer.

In the proposed framework, DSC-Net is used to extract spatial features and it has the following differences compared with the original DenseNet architecture:

1. Input size: the input layer of DSC-Net accepts 56×56 patches, whereas in the original DenseNet implementation it is 224×224. This size is fixed due to limited data in the IQA dataset, and patch-based training is done.

2. DenseBlocks with fewer dense layers: as the input size is small, the number of layers is reduced in DSC-Net. In the DenseNet implementation, the number of layers in each DenseBlock is 6, 12, 24, and 16, and hence the total number of convolution filters sums to more than 100. However, in the proposed implementation, the number of layers in all the dense blocks is 3, reducing the number of dense layers to 12 with a total of 28 convolution filters.

3. Reduced growth rate: each dense layer produces the number of feature maps based on the growth rate. In DenseNet implementation, the growth rate is 32, while in the proposed work it is set to 8.

4. Fewer parameters to train: because the number of layers is reduced in DSC-Net, the number of parameters is significantly reduced.

The DSC-Net generates the feature map $\boldsymbol{F}_{\mathrm{s}} \in \mathbb{R}^{256}$ with parameters β_1, which is fed into a fully connected layer for quality score sq_i estimation.

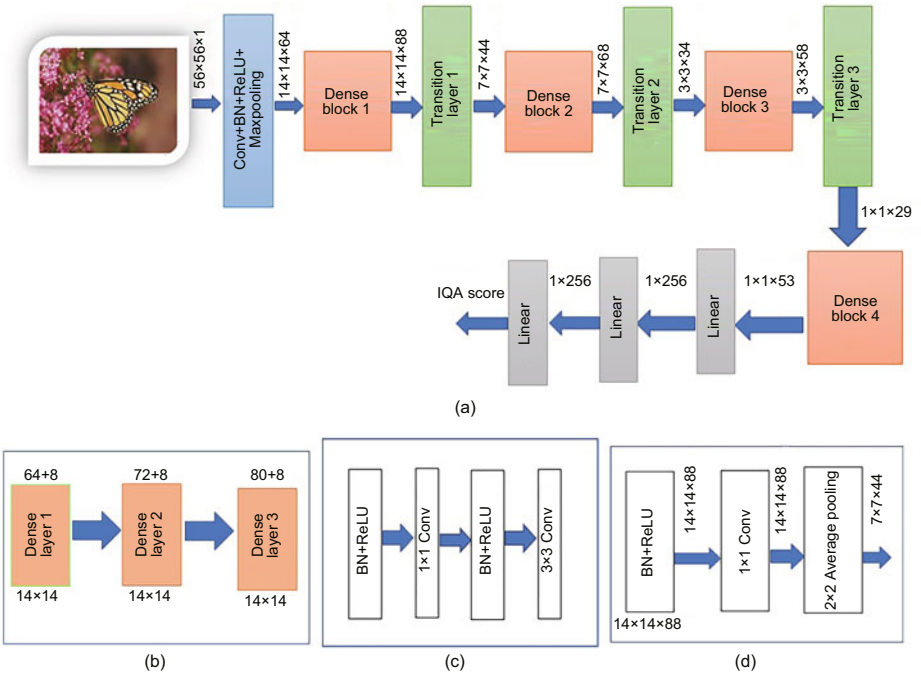

Fig. 3 Spatial feature extraction using DSC-Net: (a) DSC-Net architecture; (b) structure of DenseBlock; (c) structure of dense layer; (d) structure of the transition layer

3.3 Texture feature extraction (α_3)

In addition to the existing DSC-Net implementation, texture feature extraction and feature fusion are used in the MDSC-Net model.

The texture of an image is an attribute that represents the grouping of pixels based on their similarity. It also exhibits the structural arrangement of an image and can be used to recognize different objects. The texture characteristics are extracted using the GLCM method in the proposed research. GLCM is a second-order statistical approach that calculates the co-occurrence matrix by looking at how often pairs of pixels with the same intensity value appear in an image. Once the matrix is created, several statistical features related to textures can be derived.

The GLCM $\boldsymbol{M}$ will be of size $L \times L$, where L is the number of gray levels in an image I irrespective of the image size. The element $M_\phi(i,j)$ of the GLCM is expressed using Eq. (8):

$$M_\phi(i,j) = \frac{p_{ij}}{\sum_{i,j=1}^{L} p_{ij}}, \quad i,j = 1,2,\ldots,L, \qquad (8)$$

where p_{ij} represents the number of occurrences of gray level i together with gray level j in nearest neighbor pixels either horizontally, vertically, or diagonally. ϕ defines the position relation between pixels i and j (ϕ =(d, θ), d decides the interpixel distance), and θ defines the orientation of the neighbor pixels in the four directions (0°, 45°, 90°, and 135°).

In the proposed work, the GLCM is constructed for each gray-scale image patch P_i. Then, the five textural features, contrast, correlation, energy, homogeneity, and dissimilarity, as defined in Table 1 are derived from the GLCM for the four directions with interpixel distance $d = 1$, 2, and 3, and used for training.

Table 1 Texture features extracted using GLCM and their definition

Statistics feature	Definition
Contrast	$\sum_{i,j=1}^{L} M_{ij}(i-j)^2$
Correlation	$\sum_{i,j=1}^{L} \frac{(i-\mu_i)(j-\mu_j)}{\sigma_i \sigma_j} M_{ij}$
Dissimilarity	$\sum_{i,j=1}^{L} M_{ij}\lvert i-j\rvert$
Energy	$\sum_{i,j=1}^{L} M_{ij}^2$
Homogeneity	$\sum_{i,j=1}^{L} \frac{1}{1+\lvert i-j\rvert} M_{ij}$

In Table 1, M_{ij} refers to the element in the normalized co-occurrence matrix, and μ_i and μ_j are the averages calculated along rows and columns

of matrix $\boldsymbol{M}$, respectively. Similarly, σ_i and σ_j are standard deviations calculated along rows and columns, respectively. The detailed analysis of the texture features mentioned in Table 1 is discussed in the supplementary materials to show that the features extracted using GLCM can influence the performance.

Once the texture features $\mathbf{TF}_i \in \mathbb{R}^{60}$ are extracted for each image patch P_i, it is trained using one-dimensional convolution (1D-conv) to boost the performance of the image quality prediction as shown in Fig. 4. It contains stacking two layers of 1D-conv filters, followed by a pooling and fully connected layer, which produces the output vector of dimension $\boldsymbol{F}_\mathrm{t} \in \mathbb{R}^{256}$ with parameters β_2.

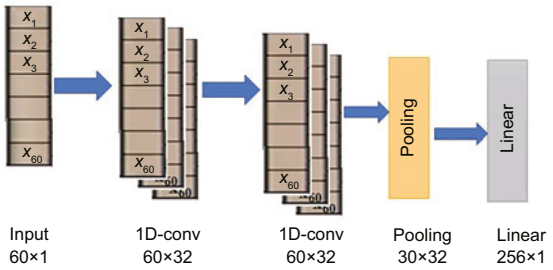

Fig. 4 Texture feature extraction network

3.4 Feature fusion and regression (α_4)

The feature fusion and regression layer is implemented as part of the MDSC-Net model. Feature fusion is one of the major research components in modern deep learning architecture and is applied in various computer vision tasks like image recognition. The model becomes more robust by learning rich features from different forms of input. Spatial features $\boldsymbol{F}_\mathrm{s} \in \mathbb{R}^{256}$ and texture features $\boldsymbol{F}_\mathrm{t} \in \mathbb{R}^{256}$ are fused together in the proposed MDSC-Net model. A concatenation operation is used in the proposed feature fusion to build a more complete representation of the image. The spatial feature provides the pixel-level information, like edges, location, and the orientation of objects. The smoothness and roughness characteristics can then be retrieved using the texture features. These two sets of features $\boldsymbol{F}_\mathrm{s}, \boldsymbol{F}_\mathrm{t} \in \mathbb{R}^{512}$ have different characteristics and give precise details about the image, which improves the image quality prediction accuracy. The fused features are connected using a fully connected layer with 256 neurons. The previous layer is coupled with a regression layer that outputs the predicted visual quality score sq_i. The input features are translated into a continuous target variable by the regression layer with a single-node neuron. The regression network is trained using the mean absolute error as the loss function that seeks to minimize the difference between the predicted and target variables.

4 Experiments

This section reports on the experiments designed to evaluate the proposed approach and compare it to other existing state-of-the-art approaches for NR-IQA. The experiments are conducted using Intel Core i7-9700K CPU with the NVIDIA GeForce RTX 2080Ti with 16 GB RAM and a 3.60 GHz processor, and the implementation is done using Python. The proposed framework uses the three large-scale benchmark IQA datasets, LIVE (Sheikh, 2003), TID2013 (Ponomarenko et al., 2015), and KADID-10k (Lin HH et al., 2019). These three datasets contain images along with their MOS or differential MOS (DMOS) predicted by the observers.

4.1 Experimental protocols

The two different proposed network architectures have been tested, and the performance is evaluated using standard evaluation metrics on the three benchmark datasets described in Section 4.

1. DSC-Net: the network uses spatial features derived using the simplified DenseNet architecture as described in Section 3.2. The input to the model is P_i and the predicted visual quality score is represented using $\tilde{q} = \psi(I_\mathrm{d}; \beta_1)$.

2. MDSC-Net: the network fuses texture features extracted using GLCM as described in Section 3.3, and features extracted from DSC-Net are concatenated and trained to predict the quality score $\tilde{q}$. The model takes input as $(P_i, \mathbf{TF}_i)$ and the output of the DSC-Net feature $\boldsymbol{F}_\mathrm{s}$, and the predicted quality score can be formulated as $\tilde{q} = \psi(I_\mathrm{d}; \beta_1, \beta_2)$

The proposed work uses the Spearman rank-order correlation coefficient (SROCC) and Pearson linear correlation coefficient (PLCC) to evaluate the performance of the model.

4.2 Training and testing strategy

The proposed network architectures mentioned in Section 4.1 are implemented using the PyTorch library. The entire dataset is split into training and testing by the ratio of 80:20. The training and testing sets are unrelated subsets of the dataset. All the different versions of the distorted images generated from a pristine image belong to either training or testing. During training, all patches are assigned the same quality score as their image MOS. The training phase of quality score estimation is given in Algorithm 1. The proposed network is trained to minimize the objective function so that the input image is expected to output a quality score that is as close to its MOS/DMOS values as possible.

The network is trained to minimize the loss function, which gives equal importance to each patch:

$$\text{Loss}_{I_\text{d}} = \frac{1}{N_\text{p}} \sum_{i=1}^{N_\text{p}} \mid q_{\text{sub}i} - \tilde{q}_i \mid \times \epsilon, \tag{9}$$

where ϵ is 1. The error calculated using the loss function is backpropagated, and gradient descent is used to update the model parameters β_1 and β_2 based on the magnitude of the loss. Adam, an adaptive optimization algorithm, is used in the proposed work to update the weights of the network with a learning rate of 1e-3. The first- and second-order moments equivalent to 0.9 and 0.999, respectively, are used to enhance the generalization ability of the network.

During testing, the image is cropped into non-overlapping patches of size 56×56. For each patch, LCN is applied, and the texture features are extracted using the GLCM approach and given as the input to the trained model, which outputs the predicted quality score of that patch. The scalar average value is the final quality score of the input image.

4.3 Performance analysis

To evaluate the performance of the proposed work, experiments are conducted on LIVE, TID2013, and KADID-10k datasets. The IQA dataset contains images of varying resolution and the number of images in each dataset may vary. Each dataset will have different types of distortions. Because rescaling or resizing the image will affect the image quality, the proposed implementation augments the dataset by cropping the images with small patches. The network is trained 10 times with different samples each time and the result is reported by taking an average of those values to avoid bias. The evaluation results are grouped into traditional and DCNN-based approaches, and the results are reported in their respective papers. The PLCC and SROCC values for the proposed networks on the LIVE and TID2013 datasets are as shown in Table 2, and the top two results are in bold. The GLCM features with dimension $\mathbb{R}^{60}$ performed well in terms of PLCC and SROCC in both datasets. MDSC-Net has shown good performance compared to other models, which shows that GLCM features play a significant role in NR-IQA approaches. The proposed approach achieves comparable performance on both LIVE and TID2013 datasets. It is better than the DB-CNN (Zhang WX et al., 2020), which uses the pre-trained features and GAN-based RAN4IQA (Ren et al., 2018). The performance of the proposed method is slightly lower than those of AIGQA (Ma JP et al., 2021) and VCRNet (Pan et al., 2022) on the TID2013 dataset. Generally, the GAN-based models RAN4IQA (Ren et al., 2018), AIGQA (Ma JP et al., 2021), and VCRNet (Pan et al., 2022) have a large number of training parameters, so they take more training time and memory overhead. Overall, the proposed model can achieve state-of-the-art

Algorithm 1 Quality estimation learning

$\text{Img}_\text{D} \leftarrow$ distorted training image dataset
$M_\text{qs} \leftarrow$ MDSC-Net model
loop
  do forward algorithm, yielding estimate M_qs
  for each $I_{\text{d}i} \in \text{Img}_\text{D}$ **do**
    $P_\text{nop} \leftarrow$ extract patches($I_{\text{d}i}$)
    for each patch $P_i \in P_\text{nop}$ **do**
      $F_{\text{p}i} \leftarrow$ LCN(patch)
      $F_{\text{glcm}i} \leftarrow$ glcmFeat(patch)
    end for
  end for
  $M_\text{qs} =$ customizeNet(DenseNet, GLCMNet)
  $\theta = [w, b]$ of M_qs
  while $i \leq$ maxepochs **do**
    for $i = 0$ **to** $\text{len}_{\text{Img}_\text{D}}$ **do**
      $\hat{\text{qs}}_i = M_\text{qs}(F_{\text{p}i}, F_{\text{glcm}i}, \text{mos}_i)$
      compute loss $L_i = (\hat{\text{qs}}_i, \text{qs}_i)$
      compute $\theta_i = \text{Adam}(L_i, \theta)$
      update θ
    end for
  end while
end loop
return M_qs

Table 2 PLCC and SROCC performance on the LIVE and TID2013 datasets

Method	PLCC		SROCC	
	LIVE	TID2013	LIVE	TID2013
BLIINDS-II	0.916	0.628	0.912	0.536
DIIVINE	0.923	0.654	0.925	0.549
BRISQUE	0.935	0.651	0.939	0.573
CORNIA	0.943	0.613	0.942	0.549
CNN	0.953	0.653	0.956	0.558
BIECON	0.960	0.762	0.961	0.717
WaDIQaM-NR	0.963	0.787	0.954	0.761
TS-CNN	0.965	0.824	0.969	0.783
RAN4IQA	0.961	0.859	0.962	0.820
DB-CNN	0.971	0.865	0.968	0.816
AIGQA	0.957	**0.893**	0.960	**0.871**
VCRNet	**0.974**	**0.875**	**0.973**	0.846
HyperIQA	0.966	0.873	0.962	0.846
DSC-Net	0.971	0.868	0.967	0.857
MDSC-Net	**0.976**	0.870	**0.971**	**0.867**

PLCC: Pearson linear correlation coefficient; SROCC: Spearman rankorder correlation coefficient. Top two results are in bold

Table 3 Performance comparison on the KADID-10k dataset

Method	PLCC	SROCC
BLIINDS-II	0.548	0.530
DIIVINE	0.423	0.428
BRISQUE	0.383	0.386
CNN	0.619	0.603
WaDIQaM-NR	0.868	0.865
MultiGAP-GPR	0.820	0.814
AIGQA	0.863	0.864
DSC-Net	0.872	0.867
MDSC-Net	**0.876**	**0.873**

PLCC: Pearson linear correlation coefficient; SROCC: Spearman rankorder correlation coefficient. Best results are in bold

performance in the LIVE and TID2013 datasets with less memory and less computation overhead.

Because KADID-10k is a recent dataset, most of the approaches did not report their performances with that dataset. The performance of MDSC-Net on the KADID-10k dataset is given in Table 3. All the compared results are taken from Ma JP et al. (2021). The proposed method performs better than the GAN-based approach AIGQA (Ma JP et al., 2021) on the KADID-10k dataset, which uses active inference to learn the primary content and multi-stream prior information to predict the IQA score. In general, the proposed method works well on all synthetically distorted datasets, which verifies the effectiveness of the proposed approach. This performance improvement is mainly due to many images in the KADID-10k dataset, which helps the model effectively learn the features.

The objective of distortion specific experiments is to determine how well the proposed algorithm performs on images of a particular type of distortion.

The performance of SROCC and PLCC in the LIVE dataset is evaluated using the model that performs better during the training session. During distortion specific evaluation, all the images of that particular distortion are considered in the LIVE dataset. The resulting analysis for various distortion types is shown in Table 4. The proposed approach is compared with four traditional NR-IQA methods, DIVINE (Moorthy and Bovik, 2011), BLIINDS-II (Saad et al., 2012), BRISQUE (Mittal et al., 2012), and CORNIA (Ye et al., 2012), and deep learning based approaches like CNN (Kang et al., 2014), SOM (Zhang P et al., 2015), MEON (Ma KD et al., 2018), BIECON (Kim and Lee, 2017) and DB-CNN (Zhang WX et al., 2020). It is also compared with the recently introduced NR-IQA method, NRVPD (Wu et al., 2019), which uses neighborhood correlation among the pixels to facilitate the quality prediction. The top three results from each distortion type are marked in bold. The performance of VCRNet (Pan et al., 2022) is relatively good, but the model size is larger, which makes it unsuitable for embedded devices like cameras. DSC-Net performs well in terms of SROCC for JPEG compression and WN distortions. It also has significant performance for WN in terms of PLCC. MDSC-Net achieves state-of-the-art SROCC and PLCC values for JP2K, JPEG, and WN. SOM (Zhang P et al., 2015) performs well for BLUR and FF distortions in terms of both SROCC and PLCC. SOM augments the dataset based on the objectness measure and is trained with a large dataset. In summary, the proposed method performs well on distortions due to compression, which clearly shows that it is sensitive to capturing distortions related to compression artifacts.

The performance of the distortion specific experiments conducted with 24 types of synthetic distortions available in the TID2013 dataset is shown in Table 5, and the top two results are in bold. The results with respect to the challenging TID2013 dataset are diverse compared to those on the LIVE dataset. DSC-Net is able to capture the distortions related to additive gaussian noise (AGN), Gaussian blur (GB), image denoising (DEN), and transmission

Table 4 Distortion specific comparison of SROCC and PLCC on the LIVE dataset

Method	SROCC						PLCC					
	JP2K	JPEG	WN	BLUR	FF	All*	JP2K	JPEG	WN	BLUR	FF	All*
DIVINE	0.913	0.910	**0.984**	0.921	0.863	0.916	0.922	0.921	0.988	0.923	0.888	0.917
BLIINDS-II	0.929	0.942	0.969	0.923	0.889	0.931	0.935	0.968	0.980	0.938	0.896	0.930
BRISQUE	0.914	0.965	0.979	0.951	0.887	0.940	0.923	0.973	0.985	0.951	0.903	0.942
CORNIA	0.943	0.955	0.976	**0.969**	0.906	0.942	0.951	0.965	0.987	**0.968**	0.917	0.935
CNN	0.952	**0.977**	0.978	0.962	0.908	0.956	0.953	0.981	0.984	0.953	**0.933**	0.953
SOM	0.947	0.952	**0.984**	**0.976**	**0.937**	0.964	0.952	0.961	**0.991**	**0.974**	**0.954**	0.962
MEON	0.914	0.951	0.972	0.944	0.926	0.943	0.923	0.968	0.982	0.929	**0.936**	0.954
BIECON	0.952	0.974	0.980	0.956	0.923	0.961	**0.965**	**0.987**	0.970	0.945	0.931	0.962
NRVPD	0.941	0.953	0.968	0.910	0.917	0.943	0.955	0.971	0.975	0.938	**0.933**	0.947
DB-CNN	**0.955**	0.972	0.983	0.935	**0.930**	**0.968**	–	–	–	–	–	**0.971**
VCRNet	**0.975**	**0.979**	**0.988**	**0.978**	**0.962**	**0.973**	–	–	–	–	–	**0.974**
DSC-Net	0.951	**0.977**	**0.990**	0.963	0.921	0.967	**0.961**	**0.986**	**0.992**	0.952	0.926	**0.971**
MDSC-Net	**0.953**	**0.976**	**0.988**	0.968	0.918	**0.971**	**0.963**	**0.987**	**0.992**	**0.963**	0.931	**0.976**

* With all the distortions. PLCC: Pearson linear correlation coefficient; SROCC: Spearman rankorder correlation coefficient. Top three results are in bold

Table 5 SROCC comparison among different methods under different distortion types on the TID2013 dataset

Type	SROCC							
	BLIINDS-II	BRISQUE	CORNIA	RankIQA	DB-CNN	VCRNet	DSC-Net	MDSC-Net
AGN	0.714	0.630	0.341	0.891	0.790	0.844	**0.947**	**0.925**
ANC	0.728	0.424	–0.196	**0.799**	0.700	**0.785**	0.765	0.769
SCN	0.825	0.727	0.689	**0.911**	0.826	0.787	0.825	**0.842**
MN	0.358	0.321	0.184	0.644	0.646	**0.795**	0.765	**0.769**
HFN	0.852	0.775	0.607	0.873	0.879	**0.942**	0.878	**0.885**
IN	0.664	0.669	–0.014	**0.869**	0.708	0.826	0.857	**0.868**
QN	0.780	0.592	0.673	**0.910**	0.825	0.847	0.884	**0.894**
GB	0.852	0.845	0.896	0.835	0.859	**0.906**	**0.955**	0.886
DEN	0.754	0.553	0.787	0.894	0.865	**0.937**	**0.941**	0.934
JPEG	0.808	0.742	0.875	**0.902**	0.894	**0.934**	0.880	0.888
JP2K	0.862	0.799	0.911	**0.923**	**0.916**	0.906	0.876	0.894
JGTE	0.251	0.301	0.310	0.579	0.772	0.762	**0.852**	**0.857**
J2TE	0.755	0.672	0.625	0.431	0.773	**0.865**	**0.855**	0.835
NEPN	0.081	0.175	0.161	0.463	0.270	0.457	**0.734**	**0.745**
Block	0.371	0.184	0.096	**0.693**	0.444	0.601	0.672	**0.683**
MS	0.159	0.155	0.008	0.321	–0.009	0.509	**0.631**	**0.645**
CTC	–0.082	0.125	0.423	0.657	0.548	**0.595**	0.572	**0.576**
CCS	0.109	0.032	–0.055	0.622	0.631	**0.855**	0.798	**0.821**
MGN	0.699	0.560	0.259	**0.845**	**0.711**	**0.845**	0.664	0.672
CN	0.222	0.282	0.606	0.609	0.752	0.819	**0.820**	**0.828**
LCNI	0.451	0.680	0.555	**0.891**	0.860	**0.895**	0.878	0.883
ICQD	0.815	0.804	0.592	0.788	**0.833**	0.822	0.831	**0.842**
CHA	0.568	0.715	0.759	0.727	0.732	0.762	**0.827**	**0.816**
SSR	0.856	0.800	0.903	0.768	0.902	0.714	**0.913**	**0.906**

SROCC: Spearman rankorder correlation coefficient. Top two results are in bold

errors such as JGTE and J2TE well in TID2013 dataset. Moreover, MDSC-Net outperforms existing state-of-the-art RankIQA (Liu XL et al., 2017) and DB-CNN (Zhang WX et al., 2020) models, and uses pre-trained weights for different distortions in the TID2013 dataset. Particularly, the proposed method MDSC-Net is able to perform well on distortions introduced by spatially correlated noise (SCN), quantization noise (QN), and compression related distortions such as JPEG, sparse sampling and reconstruction (SSR), and change of color saturation (CCS), which clearly shows that it is sensitive to capture distortions related to spatial frequency. Most of the blind image quality assessment (BIQA) models

fail to achieve reasonable performance on distortion types like non-eccentricity pattern noise (NEPN), local block-wise distortions (Block), and mean shift (MS). These results demonstrate that DenseNet-based architecture can improve the deep learning based NR-IQA performance, and GLCM features play a significant role in NR-IQA approaches. The performance of MDSC-Net on KADID-10k with different distortion types is shown in Fig. 5, and the proposed approach performs well for all distortion types except color saturation 1 (CSA1), color shift (CS), color block (CB), and non-eccentric patch (NEP). The performance of the proposed model on TID2013 and KADID-10k with up to 25 distortion types is superior to other state-of-the-art solutions, which shows that the proposed MDSC-Net is capable of learning multiple distortions effectively. In general, the proposed method works well on all synthetically distorted datasets.

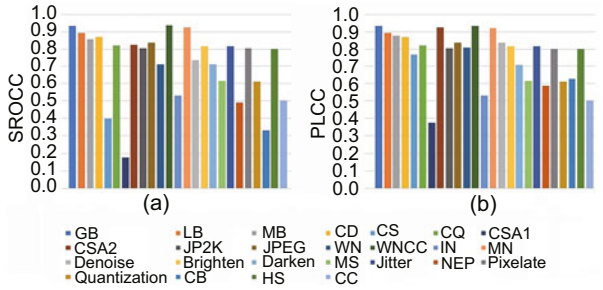

Fig. 5 Distortion specific performance on the KADID-10k dataset: (a) Spearman rankorder correlation coefficient (SROCC); (b) Pearson linear correlation coefficient (PLCC)

4.4 Cross-dataset evaluation

To evaluate the generalization performance, the model trained on one dataset is used to test on the remaining datasets as shown in Table 6. The MDSC-Net model trained on the LIVE dataset is evaluated on the TID2013/KADID-10k dataset. The five distortion types in the LIVE dataset are only parts of the 24/25 distortion types in another dataset. The remaining distortion types are not included in the training data. The proposed model is tested on the full set. Moreover, both MOS and DMOS have different ranges and meanings, and hence logistic regression is added to match the quality scores predicted by the network. The cross-dataset evaluation is repeated for other datasets in a similar manner. It is observed from Table 6 that the cross-dataset evaluation on the LIVE dataset, which is trained on TID2013/KADID-10k, is superior, because it learns from known distortion types. The generalization capability of the KADID-10k model is superior to those of the other models due to the larger size of the dataset.

The cross-dataset comparison is done by training on LIVE and testing on the full set of TID2013 and vice versa, and the proposed MDSC-Net has state-of-the-art generalization ability. The cross-dataset results are shown in Table 7, and the proposed method achieves competitive results and performs above state-of-the-art results. It can be concluded that the proposed model is more robust and able to perform well on unknown images. The proposed method performs significantly better, because it extracts all the salient patches and learns more significant features to predict the image quality.

Table 6 SROCC results of cross dataset evaluation

Training dataset	SROCC		
	LIVE*	TID2013*	KADID-10k*
LIVE	*0.971*	0.794	0.802
TID2013	0.881	*0.867*	0.724
KADID-10k	0.872	0.621	*0.873*

* Testing dataset. SROCC: Spearman rankorder correlation coefficient. The intra-dataset evaluations are in italics

Table 7 Cross-dataset comparison of SROCC

Method	SROCC	
	LIVE[1], TID2013[2]	TID2013[1], LIVE[2]
BLIINDS-II	0.393	0.842
DIIVINE	0.355	0.796
BRISQUE	0.358	0.790
CORNIA	0.360	0.846
HOSA	0.361	0.846
CNN	0.407	0.532
WaDIQaM	0.392	–
DB-CNN	0.524	0.872
TS-CNN	0.431	0.566
RAN4IQA	0.466	0.811
VCRNet	0.502	0.822
DSC-Net	0.679	0.880
MDSC-Net	**0.794**	**0.881**

[1] Training dataset; [2] testing dataset. SROCC: Spearman rankorder correlation coefficient. Best results are in bold

4.5 Algorithm complexity

The informal analysis, to determine the speed of the proposed system prediction, is discussed in this subsection. Let image I_{d} be cropped into N_{p} patches of size $w_{\mathrm{p}} \times h_{\mathrm{p}}$. T_{lcn} is the time taken to obtain the LCN image for a patch of size $w_{\mathrm{p}} \times h_{\mathrm{p}}$, T_{glcm} is the time taken to extract the texture features using a GLCM for a patch, and T_{pred} is the time taken to predict the image quality for all N_{p} patches using the proposed trained model. The time taken to predict the quality of image I_{d} is $O((N_{\mathrm{p}} \times (T_{\mathrm{lcn}} + T_{\mathrm{glcm}})) + T_{\mathrm{pred}})$. An image of resolution 512×384 is cropped into 54 patches, each of size 56×56. The time taken to extract features for MDSC-Net is about 1.04 s including T_{lcn} and T_{glcm} for all patches and for DSC-Net with T_{lcn} of 0.06 s, which is negligible. DSC-Net is faster than other models, which implies that the GLCM feature extraction step is time consuming. For the above analysis, the time used for feature extraction is the only time considered, beause the time for loading the pre-trained model and prediction time are negligible.

4.6 Effect of trainable parameters

The model parameters are computed for the proposed work, and the comparison with state-of-the-art methods is shown in Table 8. DSC-Net has fewer parameters compared to MDSC-Net. The added GLCM feature network adds more parameters to MDSC-Net due to the fully connected layers. Due to dense connections and bottleneck layers in the proposed work, the number of parameters is reduced significantly, which avoids overfitting and reduces the training time. The BIECON (Kim and Lee, 2017) model has slightly lower SROCC, but the model size is almost 7 times that of MDSC-Net. The VCRNet (Pan et al., 2022) model has a slightly higher SROCC value, but there are significantly more parameters to train than for DSC-Net. CNN++ (Kang et al., 2015) and MEON (Ma KD et al., 2018) have fewer parameters, but the SROCC value is lower compared to that of DSC-Net. Hence, the proposed MDSC-Net achieves state-of-the-art performance with fewer parameters.

Table 8 Comparison of trainable parameters and SROCC on the LIVE dataset

Method	Model size (KB)	SROCC
CNN	72	0.956
CNN++	**7.9**	0.950
DeepIQA	523	0.920
BIECON	330	0.961
MEON	**10.6**	0.943
WaDIQaM	5200	0.954
DBCNN	58410	0.968
RAN4IQA	158140	0.962
VCRNet	473590	**0.973**
DSC-Net	14.06	0.967
MDSC-Net	45.54	**0.971**

SROCC: Spearman rankorder correlation coefficient. Top two results are in bold

5 Conclusions

A multimodal simplified DenseNet-based model is proposed for the NR-IQA problem, which demonstrates good performance. The proposed DSC-Net network with more layers promotes feature reuse with fewer parameters, which makes the model more appealing for spatial feature extraction. The GLCM features fused with DSC-Net significantly increase the model performance. The proposed algorithm has shown good performance on the LIVE, TID2013, and KADID-10k datasets and demonstrates high consistency with human-perceived quality. The cross-dataset evaluation has shown good generalization capability of the proposed approach and is found to be efficient over state-of-the-art NR-IQA algorithms. The current work deals with IQA for synthetic distortion. The possible future direction will be designing a unified NR-IQA method that deals with both authentic and synthetically distorted images. Furthermore, the MDSC-Net based approach may be used in other image processing tasks that deal with the perceptual attributes of the images, such as image restoration and image enhancement. In the future, hardware implementation of the proposed model will be realized to integrate it in real-time applications such as surveillance cameras. The other possible future direction of this work can be a video quality assessment to address issues related to motion, frame rate, and compression artifacts. It can also be extended to focus on applications such as medical quality assessment or social media image quality. Further research is needed to enhance the knowledge behind IQA and its relationship to neuroscience, which may provide new insights.

Contributors

Nandhini CHOCKALINGAM designed the research, processed the data, and drafted the paper. Brindha MURUGAN helped organize and revise the paper. Nandhini CHOCKALINGAM and Brindha MURUGAN finalized the paper.

Compliance with ethics guidelines

Nandhini CHOCKALINGAM and Brindha MURUGAN declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Bianco S, Celona L, Napoletano P, et al., 2018. On the use of deep learning for blind image quality assessment. *Signal Image Video Process*, 12(2):355-362. https://doi.org/10.1007/s11760-017-1166-8

Bosse S, Maniry D, Wiegand T, et al., 2016. A deep neural network for image quality assessment. Proc IEEE Int Conf on Image Processing, p.3773-3777. https://doi.org/10.1109/ICIP.2016.7533065

Bosse S, Maniry D, Müller KR, et al., 2018. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans Image Process*, 27(1):206-219. https://doi.org/10.1109/TIP.2017.2760518

Chockalingam N, Murugan B, 2023. Hierarchical patch selection: an improved patch sampling for no reference image quality assessment. *IEEE Trans Artif Intell*, early access. https://doi.org/10.1109/TAI.2023.3262623

Cheng ZX, Takeuchi M, Katto J, 2017. A pre-saliency map based blind image quality assessment via convolutional neural networks. Proc IEEE Int Symp on Multimedia, p.77-82. https://doi.org/10.1109/ISM.2017.21

Deng J, Dong W, Socher R, et al., 2009. ImageNet: a large-scale hierarchical image database. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.248-255. https://doi.org/10.1109/CVPR.2009.5206848

Ding GG, Chen WS, Zhao SC, et al., 2018. Real-time scalable visual tracking via quadrangle kernelized correlation filters. *IEEE Trans Intell Transp Syst*, 19(1):140-150. https://doi.org/10.1109/CVPR.2009.5206848

Ding GG, Guo YC, Chen K, et al., 2019. DECODE: deep confidence network for robust image classification. *IEEE Trans Image Process*, 28(8):3752-3765. https://doi.org/10.1109/TIP.2019.2902115

Gu K, Tao DC, Qiao JF, et al., 2018. Learning a no-reference quality assessment model of enhanced images with big data. *IEEE Trans Neur Netw Learn Syst*, 29(4):1301-1313. https://doi.org/10.1109/TNNLS.2017.2649101

Gu K, Xia ZF, Qiao JF, et al., 2020. Deep dual-channel neural network for image-based smoke detection. *IEEE Trans Multimed*, 22(2):311-323. https://doi.org/10.1109/TMM.2019.2929009

Gu K, Zhang YH, Qiao JF, 2021a. Ensemble meta-learning for few-shot soot density recognition. *IEEE Trans Industr Inform*, 17(3):2261-2270. https://doi.org/10.1109/TII.2020.2991208

Gu K, Liu HY, Xia ZF, et al., 2021b. $PM_{2.5}$ monitoring: use information abundance measurement and wide and deep learning. *IEEE Trans Neur Netw Learn Syst*, 32(10):4278-4290. https://doi.org/10.1109/TNNLS.2021.3105394

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Huang G, Liu Z, Van Der Maaten L, et al., 2017. Densely connected convolutional networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.4700-4708. https://doi.org/10.1109/CVPR.2017.243

Kang L, Ye P, Li Y, et al., 2014. Convolutional neural networks for no-reference image quality assessment. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1733-1740. https://doi.org/10.1109/CVPR.2014.224

Kang L, Ye P, Li Y, et al., 2015. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. Proc IEEE Int Conf on Image Processing, p.2791-2795. https://doi.org/10.1109/ICIP.2015.7351311

Kim J, Lee S, 2017. Fully deep blind image quality predictor. *IEEE J Sel Top Signal Process*, 11(1):206-220. https://doi.org/10.1109/JSTSP.2016.2639328

Krizhevsky A, Sutskever I, Hinton GE, 2012. ImageNet classification with deep convolutional neural networks. Proc 25th Int Conf on Neural Information Processing Systems, p.1097-1105.

Li QH, Lin WS, Xu JT, et al., 2016. Blind image quality assessment using statistical structural and luminance features. *IEEE Trans Multimed*, 18(12):2457-2469. https://doi.org/10.1109/TMM.2016.2601028

Li ZC, Tang JH, Mei T, 2019. Deep collaborative embedding for social image understanding. *IEEE Trans Patt Anal Mach Intell*, 41(9):2070-2083. https://doi.org/10.1109/TPAMI.2018.2852750

Lin HH, Hosu V, Saupe D, 2019. KADID-10k: a large-scale artificially distorted IQA data-base. Proc 11th Int Conf on Quality of Multimedia Experience, p.1-3. https://doi.org/10.1109/QoMEX.2019.8743252

Lin TY, RoyChowdhury A, Maji S, 2015. Bilinear CNN models for fine-grained visual recognition. Proc IEEE Int Conf on Computer Vision, p.1449-1457. https://doi.org/10.1109/ICCV.2015.170

Liu LX, Liu B, Huang H, et al., 2014. No-reference image quality assessment based on spatial and spectral entropies. *Signal Process Image Commun*, 29(8):856-863. https://doi.org/10.1016/j.image.2014.06.006

Liu XL, Van De Weijer J, Bagdanov AD, 2017. RankIQA: learning from rankings for no-reference image quality assessment. Proc IEEE Int Conf on Computer Vision, p.1040-1049. https://doi.org/10.1109/ICCV.2017.118

Lu ZK, Lin W, Yang X, et al., 2005. Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *IEEE Trans Image Process*, 14(11):1928-1942. https://doi.org/10.1109/TIP.2005.854478

Ma JP, Wu JJ, Li LD, et al., 2021. Blind image quality assessment with active inference. *IEEE Trans Image Process*, 30:3650-3663. https://doi.org/10.1109/TIP.2021.3064195

Ma KD, Liu WT, Zhang K, et al., 2018. End-to-end blind image quality assessment using deep neural networks. *IEEE Trans Image Process*, 27(3):1202-1213. https://doi.org/10.1109/TIP.2017.2774045

Mittal A, Moorthy AK, Bovik AC, 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans Image Process*, 21(12):4695-4708. https://doi.org/10.1109/TIP.2012.2214050

Moorthy AK, Bovik AC, 2011. Blind image quality assessment: from natural scene statistics to perceptual quality. *IEEE Trans Image Process*, 20(12):3350-3364. https://doi.org/10.1109/TIP.2011.2147325

Pan ZQ, Yuan F, Lei JJ, et al., 2022. VCRNet: visual compensation restoration network for no-reference image quality assessment. *IEEE Trans Image Process*, 31:1613-1627. https://doi.org/10.1109/TIP.2022.3144892

Po LM, Liu MY, Yuen WYF, et al., 2019. A novel patch variance biased convolutional neural network for no-reference image quality assessment. *IEEE Trans Circ Syst Video Technol*, 29(4):1223-1229. https://doi.org/10.1109/TCSVT.2019.2891159

Ponomarenko N, Jin LN, Ieremeiev O, et al., 2015. Image database TID2013: peculiarities, results and perspectives. *Signal Process Image Commun*, 30:57-77. https://doi.org/10.1016/j.image.2014.10.009

Qiu ZF, Yao T, Mei T, 2018. Learning deep spatio-temporal dependence for semantic video segmentation. *IEEE Trans Multimed*, 20(4):939-949. https://doi.org/10.1109/tmm.2017.2759504

Ren HY, Chen DQ, Wang YZ, 2018. RAN4IQA: restorative adversarial nets for no-reference image quality assessment. Proc 32nd AAAI Conf on Artificial Intelligence, p.7308-7314. https://doi.org/10.1609/aaai.v32i1.12258

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. https://doi.org/10.1109/TIP.2012.2191563

Sheikh HR, 2003. Image and Video Quality Assessment Research at Live. http://liveeceutexasedu/research/quality [Accessed on Oct. 30, 2022].

Sheikh HR, Bovik AC, Cormack L, 2003. Blind quality assessment of JEPG2000 compressed images using natural scene statistics. Proc 37th Asilomar Conf on Signals, Systems & Computers, p.1403-1407. https://doi.org/10.1109/ACSSC.2003.1292217

Simonyan K, Zisserman A, 2014. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

Song GH, Jin XG, Chen GL, et al., 2016. Two-level hierarchical feature learning for image classification. *Front Inform Technol Electron Eng*, 17(9):897-906. https://doi.org/10.1631/FITEE.1500346

Tang HX, Joshi N, Kapoor A, 2011. Learning a blind measure of perceptual image quality. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.305-312. https://doi.org/10.1109/CVPR.2011.5995446

Wang Z, Shang XL, 2006. Spatial pooling strategies for perceptual image quality assessment. Proc Int Conf on Image Processing, p.2945-2948. https://doi.org/10.1109/ICIP.2006.313136

Wu JJ, Zhang M, Li LD, et al., 2019. No-reference image quality assessment with visual pattern degradation. *Inform Sci*, 504:487-500. https://doi.org/10.1016/j.ins.2019.07.061

Xu JT, Ye P, Li QH, et al., 2016. Blind image quality assessment based on high order statistics aggregation. *IEEE Trans Image Process*, 25(9):4444-4457. https://doi.org/10.1109/TIP.2016.2585880

Yang GY, Ding XY, Huang T, et al., 2020. Explicit-implicit dual stream network for image quality assessment. *EURASIP J Image Video Process*, 2020(1):48. https://doi.org/10.1186/s13640-020-00538-y

Ye P, Kumar J, Kang L, et al., 2012. Unsupervised feature learning framework for no-reference image quality assessment. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1098-1105. https://doi.org/10.1109/CVPR.2012.6247789

Zhang P, Zhou WG, Wu L, et al., 2015. SOM: semantic obviousness metric for image quality assessment. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2394-2402. https://doi.org/10.1109/CVPR.2015.7298853

Zhang SQ, Zhang SL, Huang TJ, et al., 2018. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans Multimed*, 20(6):1576-1590. https://doi.org/10.1109/TMM.2017.2766843

Zhang WX, Ma KD, Yan J, et al., 2020. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans Circ Syst Video Technol*, 30(1):36-47. https://doi.org/10.1109/TCSVT.2018.2886771

Zhang WX, Ma KD, Zhai GT, et al., 2021. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Trans Image Process*, 30:3474-3486. https://doi.org/10.1109/TIP.2021.3061932

Zhou ZH, Lu W, Yang JC, et al., 2020. No-reference image quality assessment based on neighborhood co-occurrence matrix. *Signal Process Image Commun*, 81:115680. https://doi.org/10.1016/j.image.2019.115680

List of supplementary materials

1 Description of the dataset
2 Details of evaluation metrics
3 Local contrast normalization (LCN)
4 Importance of GLCM features for influential performance
5 Visualization of distortion specific prediction
6 Effect of patch size
7 Performance of pre-trained DenseNet
8 Visualization of patchwise training strategy
9 Convergence analysis