Blind Image Quality Assessment Based on Classification Guidance and Feature Aggregation

WeiPeng Cai
Wuhan University Electronic Information School https://orcid.org/0000-0001-5934-5327
Cien Fan (fce@whu.edu.cn)
Wuhan University Electronic Information School https://orcid.org/0000-0002-4973-6444
Lian Zou
Wuhan University Electronic Information School
Yifeng Liu
Wuhan University Electronic Information School
Yang Ma
Wuhan University Electronic Information School
MinYuan Wu
Wuhan University Electronic Information School

Research

Keywords: blind image quality assessment, deep neutral networks, feature aggregation

Posted Date: September 3rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-67554/v1

Version of Record: A version of this preprint was published at Electronics on November 2nd, 2020. See the published version at https://doi.org/10.3390/electronics9111811.

RESEARCH

Blind Image Quality Assessment Based on Classification Guidance and Feature Aggregation

WeiPeng Cai[1], Cien Fan[1*], Lian Zou[1], Yifeng Liu[2], Yang Ma[1] and Minyuan Wu[1]

[*]Correspondence: fce@whu.edu.cn
[1]Electronic Information School, Wuhan University, Wuhan, 430072 Wuhan, China
Full list of author information is available at the end of the article

Abstract

In this work, we present a convolutional neural network (CNN) named CGFA-CNN for blind image quality assessment (BIQA). A unique 2-stage strategy is utilized which firstly identifies the distortion type in an image using Sub-network I and then quantifies this distortion using Sub-network II. And different from most deep neural networks, we extract hierarchical features as descriptors to enhance the image representation and design a feature aggregation layer in an end-to-end training manner applying Fisher encoding to visual vocabularies modeled by Gaussian mixture models (GMMs). Considering the authentic distortions and synthetic distortions, the hierarchical feature contains the characteristics of a CNN trained on the self-built dataset and a CNN trained on ImageNet. We evaluate our algorithm on the four publicly available databases, and results demonstrate that our CGFA-CNN has superior performance over other methods both on synthetic and authentic databases.

Keywords: blind image quality assessment; deep neutral networks; feature aggregation

1 Introduction

Digital pictures may occur different distortions in the procedure of acquisition, transmission, and compression, leading to an unsatisfactory perceived visual quality or a certain level of annoyance. Thus, it is crucial to predict the quality of digital pictures in many applications, such as compression, communication, printing, display, analysis, registration, restoration, and enhancement [1–3]. Generally, image quality assessment approaches can be classified into three kinds according to the additional information needed. Specifically, full-reference image quality assessment (FR-IQA) [4–7] and reduced-reference image quality assessment (RR-IQA) [8–10] need full and partial information of reference images respectively, while blind image quality assessment (BIQA) [11–14] performs quality measure without any information from the reference image. Thus BIQA methods are more attractive in many practical applications because the reference image usually is not available or hard to drive.

Early researches mainly focus on one or more specific distortion types, such as Gaussian blur [15], blockiness from JPEG compression [16], or ringing arising from JPEG2000 compression [17]. However, images may be affected by unknown distortion in many practical scenarios. In contrast, general BIQA methods aim to work well for arbitrary distortion, which can be classified into two categories according to the features extracted, i.e., Natural Scene Statistics (NSS) based methods and training based methods.

NSS based methods [18] assume that the natural image with non-distorted obeys certain perceptually relevant statistical laws that are violated by the presence of common image distortions, and they attempt to describe an image utilizing its scene statistics from different domains. For example, BIRSQUE [19] derives features from the locally normalized luminance coefficients in the spatial domain. M3 [20] utilizes the joint local contrast features from the gradient magnitude (GM) map and the Laplacian of Gaussian (LOG) response. And later a perceptually motivated and feature-driven model is deployed in FRIQUEE [21], in which a large collection of features defined in various complementary, perceptually relevant color and transform-domain spaces are drawn from among the most successful BIQA models produced to date.

However, the knowledge-driven feature extraction and data-driven quality prediction are separated in the above methods. And it has been demonstrated that training based methods outperform the NSS based methods by a large margin because a fully data-driven BIQA solution becomes possible. For example, CORNIA [22] constructs a codebook in an unsupervised manner, using raw image patches as local descriptors and using soft-assignment for encoding. Considering that the feature set generally adopted in previous methods are from zero-order statistics and insufficient for BIQA, HOSA [23] constructs a much smaller codebook using K-means clustering [24] and introducing higher-order statistics. In contrast, the above methods capture spatially normalized coefficients and codebook-based features which are learned automatically from beginning to end by using CNNs. For example, TCSN [25] aims to learn the complicated relationship between visual appearance and perceived quality via a two-stream convolutional neural network. DIQA [26] defines two separated CNN branches to learn objective distortion and human visual sensitivity, respectively.

In this work, we propose an end-to-end BIQA based on classification guidance and feature aggregation, which is accomplished by two sub-networks with shared features in the early layers. Due to the lack of training data, we construct a large-scale dataset by means of synthesizing distortions and pre-train Sub-network I to identify an image into a specific distortion type from a set of pre-defined categories. We find the proposed method will be much harder to achieve high accuracies on authentic images if only it is exposed to synthetic distortions during training. Then we extract hierarchical features from the shared layers of two-subnetworks and another CNN (VGG-16 [27]) pre-trained on ImageNet [28] in which pictures occur as a natural consequence of photography and a unified feature group is formed.

Sub-network II takes the hierarchical features and the classification information as inputs to predict the perceptual quality. The combination of two sub-networks enables the learning framework to have the probability of favorable quality perception and proper parameter initialization in an end-to-end training manner. We design a feature aggregation layer that could convert arbitrary input seizes to a fixed-length representation. Then fully connected layer is exploited as a linear regression model to map the high-dimensional features into the quality scores. This allows the proposed CGFA-CNN to accept an image of any size as the input, thus there is no need to perform any transformation of images (including cropping, scaling, etc.), which would affect perceptual quality scores.

The paper is structured as follows. In Sec. 2, previous work on CNN-based BIQA related to our work is briefly reviewed. In Sec. 3, details of the proposed method are described. In Sec. 4, experimental results on the public IQA databases and the corresponding analysis are presented. In Sec. 5, the work of this paper is concluded.

2 Related Work

In this section, we provide a brief survey about the major solutions to the lack of training data in BIQA, and a review of recent studies related to our work.

Due to the number of parameters to be trained on CNN is usually very large, the training set needs containing sufficient data to avoid over-fitting. However, the number of samples and image contents in the public quality-annotated image databases are rather limited, which cannot meet the need for end-to-end training of a deep network. Currently, there are two main methods to tackle this challenge.

Methods of the first are to train the model based on the image patches. For example, deepIQA [29] randomly samples image patches from the entire image as inputs and predict the quality score on local regions by assigning the subjective mean score (MOS) of the pristine image to all patches within it. Although taking small patches as inputs for data augmentation is superior to use the whole image on a given dataset, this method still suffers from limitations because local image quality with context varies across spatial locations even the distortion is homogeneous. To resolve this problem, BIECON [30] makes use of the existing FR-IQA algorithms to assign quality labels for sampled image patches, but the performance of such a network depends highly on that of FR-IQA models. Other methods such as dipIQ [31] attempting to generate discriminable image pairs by involving FR-IQA models may suffer from similar problems.

The second method is to pre-train a network with large-scale datasets in other fields. And for each pre-trained architectures, two types of back-end training strategies are available: replacing the last layer of the pre-trained CNN model with the regression layer and fine-tuning it with the IQA database to conduct image quality prediction or using SVR to regress the extracted features through the pre-trained networks onto subjective scores. For instance, DeepBIQ [32] reports on the use of different features extracted from pre-trained CNNs for different image classification tasks via ImageNet [28] and Places365 [33] as a generic image description. Kim et al. [34] select the well-known deep CNN models AlexNet [35] and ResNet50 [36] as the architectures of the baseline models, which have been pre-trained for the task of image classification on ImageNet [28]. These methods directly inheriting the weights from the pre-trained models for general image classification tasks have a defect of low relevance to BIQA but unnecessary complexity.

To better address the training data shortage problem, MEON [37] proposes a cascaded multi-task framework, which firstly trains a distortion type identification network by large-scale pre-defined samples. Then a quality prediction network is trained subsequently, taking advantage of distortion information obtained from the first stage. Furthermore, DB-CNN [38] not only constructs a pre-training set based on Waterloo Exploration Database [39] and PASCAL VOC [40] for synthetic distortions, but also uses ImageNet [28] to pre-trained another CNN for authentic distortions. Motivated by the previous studies MEON [37] and DB-CNN [38],

we construct a pre-training set based on Waterloo Exploration Database [39] and PASCAL VOC [40] for synthetic distortions. Besides, both distortion type and distortion level are considered at the same time, which results in better quality-aware initializations and distortion information.

Although previous DNN-based BIQA methods have achieved significant performance, all of these methods usually comprise convolutional layers and pooling layers for feature extraction and employ fully connected layers for regression, which would suffer three limitations. First, such techniques such as averaging or maximum pooling are too simple to be accurate for long sequences. Second, a fully connected layer is destructive to the high-dimensional disorder and spatial invariance of the local feature. Third, such CNNs typically require a fixed image size. To feed into the network, images have to be resized or cropped to a fixed size, and either scaling or cropping would cause the perceptual difference with the assigned quality labels. To tackle these challenges, we explore more sophisticated pooling techniques based on clustering approaches such as Bag-of-visual-words (BOW) [41], Vector of Locally aggregated Descriptors (VLAD) [42] and Fisher Vectors [43]. Studies have shown that integrating VLAD as a differentiable module in a neural network can significantly improve the aggregated representation for the task of place recognition [44] and video classification [45]. Our proposed feature aggregation layer acts as a pooling layer on top of the convolutional layers, which converts arbitrary input seizes to a fixed-length representation. Afterward, using a fully connected layer for regression does not require any preprocessing of the input image.

3 The Proposed Method

The framework CGFA-CNN is illustrated in Fig. 1. Sub-network I aims to classify an image into a specific distortion type and initialize the shared layers for a further learning process, which is firstly pre-trained on a self-built dataset. Sub-network II predicts the perceptual quality of the same image, which is fine-tuned with the IQA databases and takes advantage of distortion information obtained from Subnetwork I. The feature aggregation layer (FV layer) and classification-guided gating unit (CGU) will be described in Sec. 3.3 and Sec. 3.4.

3.1 Distortion Type Identification

Construction of the Pre-training Dataset. Due to the deficiency of the available quality-annotated samples, we firstly construct a large-scale dataset based on Waterloo Database [39] and PASCAL VOC Database [40]. The former contains 4,744 images and can be loosely categorized into 7 classes. The latter contains 17,125 images covering 20 categories. In this paper, we merge the two databases and obtain 21,869 pristine images with various contents. Then 9 types of distortion are introduced—JPEG compression, JPEG2000 compression, Gaussian blur, white Gaussian noise, contrast stretching, pink noise, image quantization with color dithering, over-exposure, and under-exposure. We synthesize each image with 5 distortion levels following [39] except for over-exposure and under-exposure, where only three levels are generated according to [46]. The constructed dataset consists of 896,629 images, which are organized into 41 subcategories according to the distortion type and degradation level. We label these images by the subcategory they belong to.

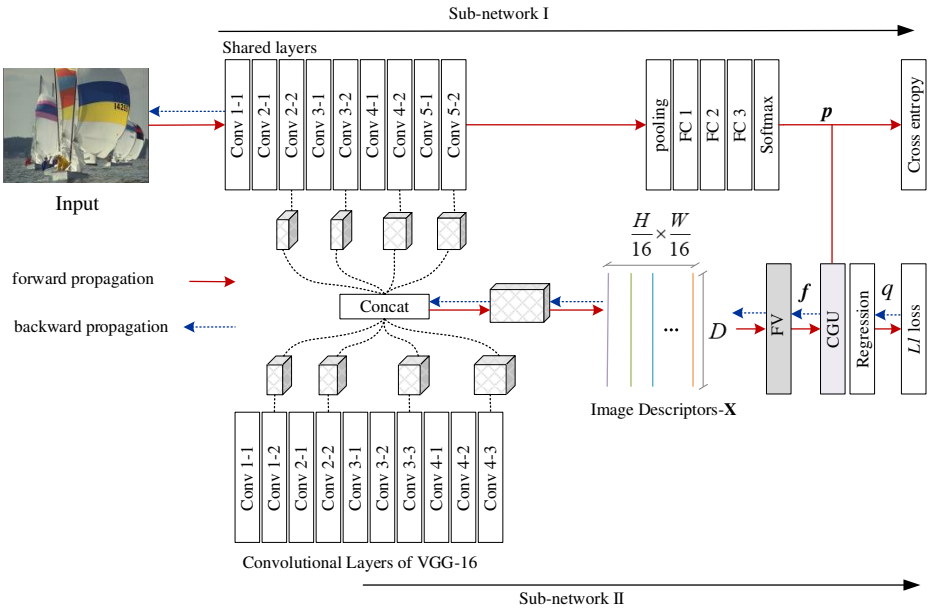

Figure 1 Illustration of CGFA-CNN configurations for BIQA, highlighting the feature aggregation layer (denoted as FV layer) and classification-guided gating unit (denoted as CGU). Features are extracted from the distorted image by Sub-network I.

Sub-network I Architecture. Inspired by the VGG-16 network architecture [27], we design a similar structure subject to some modifications identifying the distortion type of the input image. Details are given in Table 1. The tailored VGG-16 network comprises a stack of convolutions (Conv) for feature extraction, one maximum pooling (MaxPool) for feature fusion, three fully connected layers (FC) for feature regression. All hidden layers are equipped with the Rectified Linear Unit (ReLU) [35] and Batch Normalization (BN) [47]. We denote the input mini-batch training data by $\left\{\left(\mathbf{X}^{(n)}, \mathbf{p}^{(\mathbf{n})}\right)\right\}_{n=i}^{N}$, where $\mathbf{X}^{(n)}$ is the n-th input image, $\mathbf{p}^{(n)}$ is a multi-class indicator vector of the ground truth distortion type. We append the soft-max layer at the end and define the soft-max function as

$$\hat{p}_i^{(n)}(\mathbf{X}^{(n)}; \mathbf{W}) = \frac{\exp\left(y_i^{(n)}(\mathbf{X}^{(n)}; \mathbf{W})\right)}{\sum_{j=1}^{C} \exp\left(y_i^{(n)}(\mathbf{X}^{(n)}; \mathbf{W})\right)}, \tag{1}$$

where $\hat{\mathbf{p}}^{(n)} = \left[\hat{p}_1^{(n)}, \cdots, \hat{p}_C^{(n)}\right]^T$ is a C-dimensional probability vector of the n-th input in a mini-batch, indicating the probability of each distortion type. Model parameters of Sub-network I are collectively denoted as $\mathbf{W}$. A cross-entropy loss is used to train this sub-network

$$\ell_s(\{\mathbf{X}^{(n)}\}; \mathbf{W}) = -\sum_{n=1}^{N} \sum_{i=1}^{C} p_i^{(n)} \log \hat{p}_i^{(n)}(\mathbf{X}^{(n)}; \mathbf{W}). \tag{2}$$

Table 1 Architecture of Sub-network I.

layer name	type	patch size	stride	output size
Conv 1-1	Conv+ReLU+BN	$3 \times 3 \times 48$	1	$H \times W \times 48$
Conv 2-1	Conv+ReLU+BN	$3 \times 3 \times 48$	2	$\frac{H}{2} \times \frac{W}{2} \times 48$
Conv 2-2	Conv+ReLU+BN	$3 \times 3 \times 64$	1	$\frac{H}{2} \times \frac{W}{2} \times 64$
Conv 3-1	Conv+ReLU+BN	$3 \times 3 \times 64$	2	$\frac{H}{4} \times \frac{W}{4} \times 64$
Conv 3-2	Conv+ReLU+BN	$3 \times 3 \times 64$	1	$\frac{H}{4} \times \frac{W}{4} \times 64$
Conv 4-1	Conv+ReLU+BN	$3 \times 3 \times 64$	2	$\frac{H}{8} \times \frac{W}{8} \times 64$
Conv 4-2	Conv+ReLU+BN	$3 \times 3 \times 128$	1	$\frac{H}{8} \times \frac{W}{8} \times 128$
Conv 5-1	Conv+ReLU+BN	$3 \times 3 \times 128$	2	$\frac{H}{16} \times \frac{W}{16} \times 128$
Conv 5-2	Conv+ReLU+BN	$3 \times 3 \times 128$	1	$\frac{H}{16} \times \frac{W}{16} \times 128$
Pool	MaxPool	$1 \times 1 \times 128$	1	$1 \times 1 \times 128$
FC-1	FC+ReLU	$1 \times 1 \times 256$	1	$1 \times 1 \times 256$
FC-2	FC+ReLU	$1 \times 1 \times 256$	1	$1 \times 1 \times 256$
FC-3	FC	$1 \times 1 \times 41$	1	$1 \times 1 \times 41$
Classifier	Soft-max	$1 \times 1 \times 41$	1	$1 \times 1 \times 41$

Notably, in the fine-tuning phase, except for the shared layers, the rest of the Sub-network I only participates in the forward propagation and the parameters are fixed.

3.2 Feature Extraction and Fusion

From Fig. 2 we can see that the representation of different distortion types varies in each convolution. Therefore, only using features extracted from the last convolution is not enough to predict the quality of an image. And inspired by the idea of combining the complementary features and hierarchical feature extraction strategy in our previous work [48], we resort to extracting features from low-level, middle-level and high-level convolutional layers as descriptors by rescaling and concatenating them. We design Sub-network I to identify a given image's distortion type pre-trained on a synthesized dataset. We find this takes advantage of synthetic images but fails to handle those authentically distorted. More details can be found in Sec 4.5. Then we model synthetic and authentic distortions by two separated CNNs and fuse the two feature sets into a unified representation for final quality prediction. The tailored VGG-16 pre-trained on ImageNet that contains many realistic natural images of different perceptual quality is added to extract relevant features for authentic images. In the proposed CGFA-CNN index, we take a raw image of $H \times W \times 3$ as input and predict its perceptual quality. Then the fused feature group acquired is with the size of $\frac{H}{16} \times \frac{W}{16} \times D$. Here D is the channel of hierarchical features. Sub-network II takes the fused feature group and the estimated probability vector $\hat{\mathbf{p}}^{(n)}$ as inputs.

3.3 Feature Aggregation Layer and Encoding

In this paper, we design a feature aggregation layer that employs the Fisher Vectors (FV) [43] to perform the feature aggregation and encoding procedures. Because GMM [49] and FV are non-differentiable and fail to achieve theoretically valid backpropagation, we define a FV layer to yield a quality-aware feature vector $\boldsymbol{f}$. The implementation is shown in Fig. 3.

As illustrated in Fig. 1, the fused feature group is a $\frac{H}{16} \times \frac{W}{16} \times D$ map which can be considered as a set of D-dimensional descriptors extracted at $\frac{H}{16} \times \frac{W}{16}$ spatial

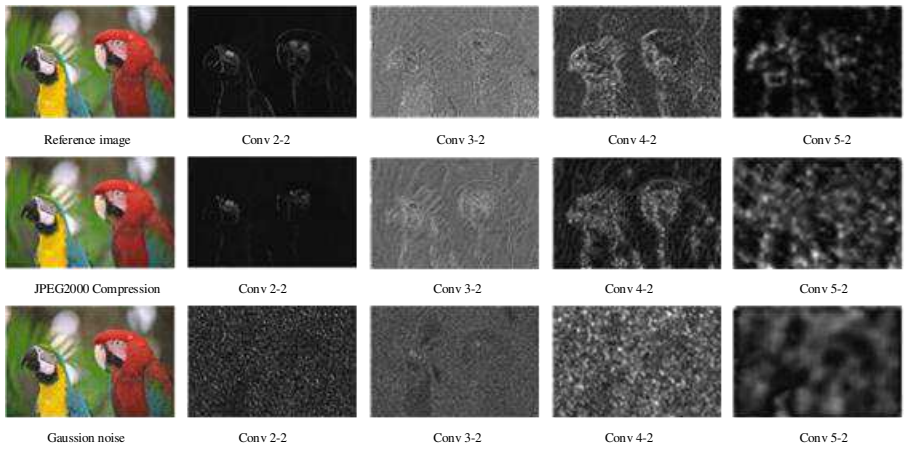

Figure 2 A comparison of several distortion types identified by Sub-network I.

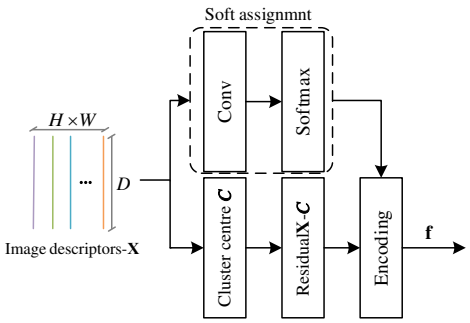

Figure 3 The configurations of the proposed FV layer. Convolution kernel size is 1×1.

locations. Then we utilize GMM to obtain the cluster centres $\boldsymbol{C}$ of K components and encoding vector $\boldsymbol{f}$ of the image descriptors-$\mathbf{X}$.

GMM clustering. A Gaussian mixture model $p(\mathbf{x}|\theta)$ is a mixture of K multivariate Gaussian distributions [49], which can be formulated as

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} p(\mathbf{x}|\mu_k, \sum\nolimits_k)\pi_k, \tag{3}$$

$$p(\mathbf{x}|\mu_k, \sum\nolimits_k) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \sum_k^{-1}(\mathbf{x}-\mu_k)}}{\sqrt{(2\pi)^D \det \sum_k}}, \tag{4}$$

$$\theta = (\pi_1, \mu_1, \sum\nolimits_1, \cdots, \pi_K, \mu_K, \sum\nolimits_K), \tag{5}$$

where θ is the vector of parameters of the model. For each Gaussian component, π_k is the prior probability value, μ_k is the means, and $\sum_K$ is the diagonal covariance matrices. The parameters are learnt from a training set of descriptors $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The GMM defines the assignments $q_{ki}(k = 1, \cdots, K, i = 1, \cdots, N)$ of the N de-

scriptors to the K Gaussian components

$$q_{ki} = \frac{p(\mathbf{x}_i|\mu_k, \sum_k)\pi_k}{\sum_{j=1}^{K} p(\mathbf{x}_i|\mu_j, \sum_j)\pi_j}, k = 1, \cdots, K. \tag{6}$$

Fisher encoding. Fisher encoding captures both the 1st order and 2nd order differences between the image descriptors and the centres of a GMM. The construction of the encoding begins by learning a GMM model θ. For each $k = 1, \cdots, K$, define the vectors

$$\mathbf{u}_k = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ki} \sum_{k}^{-\frac{1}{2}} (\mathbf{x}_i - \mu_k), \tag{7}$$

$$\mathbf{v}_k = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{ki} [(\mathbf{x}_i - \mu_k) \sum_{k}^{-1} (\mathbf{x}_i - \mu_k) - 1]. \tag{8}$$

The Fisher encoding of the set of local descriptors is then given by the concatenation of μ_k and $\mathbf{v}_k$ for all K components, giving an encoding of size $2 \times D \times K$

$$\mathbf{f}_{Fisher} = [\mathbf{u}_1^T, \cdots \mathbf{u}_K^T, \mathbf{v}_1^T, \cdots, \mathbf{v}_K^T]^T. \tag{9}$$

In order to integrate Fisher vector as a differentiable module in a neural network, we write the descriptor x_i hard assignment to the cluster k as a soft assignment

$$a_k(x_i) = \frac{e^{-\alpha \|x_i - c_k\|^2}}{\sum_{j=1}^{K} e^{-\alpha \|x_i - c_j\|^2}}. \tag{10}$$

Then we can write the FV representation as

$$FV1_{j,k} = \sum_{i=1}^{N} a_k(x_i) \left(\frac{x_i(j) - c_k(j)}{\sigma_k(j)} \right), \tag{11}$$

$$FV2_{j,k} = \sum_{i=1}^{N} a_k(x_i) \left(\left(\frac{x_i(j) - c_k(j)}{\sigma_k(j)} \right)^2 - 1 \right), \tag{12}$$

where $FV1$ and $FV2$ are capturing the 1st order and 2nd order statistics respectively. $x_i(j)$ is the j-th dimensions of the i-th descriptor and $c_k(j)$ is the k-th cluster centres. c_k and σ_k ($k \in [1, K]$) are the learnable clusters and the clusters' diagonal covariance. We define α as positive ranging between 0 and 1.

Let $\omega_k = 2\alpha c_k$, $b_k = -\alpha \|c_k\|^2$, Equ. (10) can then be written as

$$a_k(x_i) = \frac{e^{\omega_k^T x_i + b_k}}{\sum_{j=1}^{K} e^{\omega_j^T x_i + b_j}}. \tag{13}$$

where $\{\omega_k\}$, $\{b_k\}$, $\{c_k\}$ are sets of trainable parameters for each cluster k,

Beyond the FV aggregation. The source of discontinuities in the traditional Bag-of-visual-words (BOW) [41] and Vector of Locally aggregated Descriptors (VLAD) [42] are the hard assignments q_{ki} of descriptors $\mathbf{x}$ to cluster centres c_k. To make this operation differentiable, we replace it with the descriptor x_i hard assignment to the cluster as a soft assignment, and reuse the same soft assignment established in Equ. (12) to obtain differentiable representation. We denote them as VLAD layer and BOW layer, respectively. The differentiable BOW representation and VLAD representation can be written as

$$BOW_k = \sum_{i=1}^{N} a_k(x_i), \tag{14}$$

$$VLAD_{j,k} = \sum_{i=1}^{N} a_k(x_i)\left(x_i(j) - c_k(j)\right), \tag{15}$$

where $a_k(x_i)$ denotes the membership of the descriptor x_i to cluster k. BOW is the histogram of the number of image descriptors assigned to each visual word. Therefore, it produces a K-dimensional vector, while VLAD is a simplified non-probabilistic version of the FV and produces a $D \times K$-dimensional vector.

The soft assignment $a_k(x_i)$ can be regarded as a two-step process: (i) performing a 1×1 convolution with a set of K filters ω_k and bias b_i. Then the output produced is $\omega_k^T x_i + b_k$; (ii) following a soft-max function to obtain soft assignment of the descriptor x_i to the cluster k. Notably, for BOW encoding, there is no need to store the sum of residuals for each visual word, which is the difference vector between the descriptor and its corresponding cluster centre.

The advantage of the BOW aggregation is that it aggregates the descriptor into a more compact representation, and fewer parameters are trained in a discriminative manner only including $\{\omega_k\}$ and $\{b_k\}$. The drawback is that significantly more clusters are needed to obtain a rich representation. The VLAD computes the 1st order residuals between the descriptors and the cluster centres, making the richness of representation relatively sufficient, and parameters to be learned are moderate, including $\{\omega_k\}$, $\{b_k\}$ and $\{c_k\}$. In contrast, the FV aggregation concatenates both the 1st order and 2nd order aggregated residuals, but too many parameters need to be learned, including $\{\omega_k\}$, $\{b_k\}$, $\{c_k\}$ and $\{\sigma_k\}$.

In Sec. 4.5, we also experiment with averaging and maximum pooling of the image descriptor-$\mathbf{X}$. Results will show that FV proves itself to be superior to the reference BOW and VLAD approach. Additionally, simply using averaging or maximum pooling will result in poor performance.

3.4 Classification-guided Gating Unit and Quality Prediction

We have pre-trained Sub-network I to identify the distortion type of the input, and Sub-network II takes the estimated probability vector $\hat{\mathbf{p}}$ from Sub-network I as partial input. In order to introduce this prior information of the classification, a Classification-guided Gating Unit (CGU) is utilized to emphasize informative features and suppress less useful ones. The CGU combines $\hat{\mathbf{p}}$ and $\mathbf{f}$ to produce a

score vector $\hat{\mathbf{f}}$

$$\hat{\mathbf{f}} = \hat{\mathbf{p}} \cdot \sigma \left(W \cdot \mathbf{f}_{\mathbf{Fisher}} + b \right), \tag{16}$$

where σ is a linear mapping, (W, b) the learnable parameters. Then a linear mapping is followed to yield an overall quality score q. To increase nonlinearity, two fully connected layers are applied as the linear mapping.

For Sub-network II, the L_1 function is used as the empirical loss

$$\ell = \frac{1}{N} \sum_{i=1}^{N} \| q_i - \hat{q}_i \|, \tag{17}$$

where q_i is the MOS of the i-th image in a mini-batch and $\hat{q}_i$ is the predicted quality score by CGFA-CNN.

4 Experimental Results and Discussions

4.1 Database Description and Experimental Protocol

1)*IQA databases*: These experiments are confirmed on three singly distorted synthetic IQA databases, i.e., LIVE [50], CSIQ [51], and TID2013 [52], and an authentic LIVE Challenge database [53]. LIVE contains 5 distortion types—JPEG compression (JPEG), JPEG-2000 compression (JP2K), White noise (GN), Gaussian blurring (GB), and Fast-fading error (FF) at 7 to 8 degradation levels. CSIQ contains 6 distortion types—JPEG compression (JPEG), JPEG-2000 compression (JP2K), global contrast decrements (GC), additive pink Gaussian noise (PN), additive white Gaussian noise (WN), and Gaussian blurring (GB) at 3 to 5 degradation levels. TID2013 contains 24 sceptic distortion types—additive Gaussian noise, additive noise in color components, spatially correlated noise, masked noise, high-frequency noise, impulse noise, quantization noise, Gaussian blur, image denoising, JPEG compression, JPEG2000 compression, JPEG transmission errors, non-eccentricity pattern errors, local bock-wise distortions, mean shift, contrast change, change of color saturation, multiplicative Gaussian noise, comfort noise, lossy compression of noisy images, color quantization with dither, chromatic aberrations, sparse sampling and reconstruction, which denote as #01 to #24 respectively.

2)*Evaluation Criteria*: Two evaluation criteria are adopted as follows to benchmark BIQA models:

- Spearman's rank-order correlation coefficient (SRCC) is a nonparametric measure

$$\mathrm{SRCC} = 1 - \frac{6 \sum_i d_i^2}{I \left(I^2 - 1 \right)}, \tag{18}$$

  where I is the test image number and d_i is the rank difference between the MOS and the model prediction of the i-th image.
- Pearson linear correlation coefficient (PLCC) is a nonparametric measure of the linear correlation

$$\mathrm{PLCC} = \frac{\sum_i \left(q_i - q_m \right) \left(\hat{q}_i - \hat{q}_m \right)}{\sqrt{\sum_i \left(q_i - q_m \right)^2} \sqrt{\sum_i \left(\hat{q}_i - \hat{q}_m \right)^2}}, \tag{19}$$

where q_i and $\hat{q}_i$ stand for the MOS and the model prediction of the i-th image, respectively.

For synthetic databases LIVE, CSIQ and TID2013, we divide the distorted images into two splits of non-overlapping content—80% of which are used as fine-tuning samples and the rest 20% are left as testing samples. For the LIVE Challenge database, the distorted images are divided into two groups—80% for training and 20% for testing. This random process repeats ten times, and the average SRCC and PLCC are reported as the final results. Besides, the three synthetic databases are selected for cross-database experiments, using one database as the training sets while the other as the testing.

We compare the proposed CGFA-CNN against several state-of-the-art BIQA methods, including three based on NSS (BRISQUE [19], M3 [20], FRIQUEE [21]), two based on manual feature learning (CORNIA [22], HOSA [23]) and eight based on CNN (BIECON [30], dipIQ [31], deepIQA [29], ResNet50+ft [34], MEON [48], DIQA [26], TSCN[25], and DB-CNN[38]). Due to the source codes of some methods are not available to the public, we only copy the metrics from the corresponding papers.

4.2 Experimental Settings

Table 2 Demographic prediction performance comparison by two evaluation metrics.

Method	LIVE		CSIQ		TID2013		LIVE Challenge	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE[19]	0.940	0.945	0.777	0.817	0.573	0.651	0.603	0.641
M3[20]	0.950	0.954	0.804	0.835	0.679	0.705	0.595	0.620
FRIQUEE[21]	0.948	0.955	0.844	0.889	0.668	0.705	0.694	0.710
CORNIA[22]	0.943	0.946	0.730	0.800	0.550	0.613	0.618	0.665
BIECON[30]	0.958	0.960	0.815	0.823	0.717	0.762	0.595	0.613
deepIQA[29]	0.960	**0.972**	—	—	0.803	0.821	0.671	0.680
ResNet50+ft[34]	0.950	0.954	0.876	0.905	0.712	0.756	**0.819**	**0.849**
DIQA[26]	**0.975**	**0.977**	**0.884**	**0.915**	**0.825**	**0.850**	0.703	0.704
TSCN[25]	**0.969**	**0.972**	—	—	—	—	—	—
DB-CNN[38]	0.968	0.971	**0.946**	**0.959**	**0.816**	**0.865**	**0.851**	**0.869**
CGFA-CNN	**0.971**	**0.973**	**0.953**	**0.965**	**0.841**	**0.858**	**0.837**	**0.846**

Parameters in Sub-network I are initialized by He's method [54], and Adam is adopted as optimizer with the default parameters with a mini-batch of 64. The learning rate is initialized as a decay logarithmically from $\left[10^{-4}, 10^{-6}\right]$ in 30 epochs. The construction details of the pre-training dataset have been described in Section 3.1, which is randomly divided into two subsets, 80% for training and 20% for testing. All images are firstly scaled to $256 \times 256 \times 3$ and then cropped to $224 \times 224 \times 3$ as inputs. The top-1 and top-5 errors are 3.842% and 0.026% respectively.

In the fine-tuning phase, the shared layers are directly initialized with the parameters of Sub-network I. Adam is used as optimizer with the default parameters for 20 epochs and the learning rate is set to 10^{-5}. Except for the LIVE database, images are input without any pre-processing during training with a mini-batch of

Table 3 Average SRCC and PLCC results of individual distortion types across ten sessions on LIVE database.

SRCC	JPEG	JP2K	WN	GB	FF
BRISQUE[19]	0.965	0.929	0.982	**0.964**	0.828
M3[20]	0.966	0.930	**0.986**	0.935	0.902
FRIQUEE[21]	0.947	0.919	**0.983**	0.937	0.884
CORNIA[22]	0.947	0.924	0.958	0.951	**0.921**
HOSA[23]	0.954	0.935	0.975	0.954	**0.954**
dipIQ[31]	**0.969**	**0.956**	0.975	0.940	—
DIQA [26]	0.961	**0.976**	**0.988**	0.962	0.912
TCSN [25]	0.966	0.950	0.979	**0.963**	0.911
DB-CNN[38]	**0.972**	0.955	0.980	0.935	**0.930**
CGFA-CNN	**0.973**	**0.975**	**0.986**	**0.968**	0.912
PLCC	JPEG	JP2K	WN	GB	FF
BRISQUE[19]	0.971	0.940	0.989	**0.965**	0.894
M3[20]	**0.977**	0.945	**0.992**	0.947	0.920
FRIQUEE[21]	0.955	0.935	**0.991**	0.949	0.943
CORNIA[22]	0.962	0.944	0.974	0.961	0.943
HOSA[23]	0.967	0.949	0.983	**0.967**	**0.967**
dipIQ[31]	**0.980**	**0.964**	0.983	0.948	—
DIQA [26]	—	—	—	—	—
TCSN [25]	0.966	0.963	**0.995**	0.950	**0.949**
DB-CNN[38]	**0.986**	**0.967**	0.988	0.956	**0.961**
CGFA-CNN	0.972	**0.976**	0.981	**0.974**	0.947

8. Since the LIVE database contains images in different size, images are randomly cropped to 320×320 during training in a mini-batch, whose quality annotated are assigned from the corresponding image. And all of the images are input without any preprocessing during testing. We implemented all of our models using PyTorch 0.4.1 deep learning framework and the numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

4.3 Consistency Experiment

We investigate the effectiveness of CGFA-CNN on LIVE, TID2013, CSIQ and LIVE Challenge databases and the results are presented in Table 2. The results of each specific distortion type on LIVE, CSIQ and TID2013 databases are reported in Tables 3, 4, and 5. The top three SRCC and PLCC results are highlighted in red, green and blue, respectively.

From Table 2, we can have the following observations. First, DIQA [26] achieves state-of-the-art accuracies which surpasses CGFA-CNN about 0.004 in SRCC and PLCC, and most methods take great advantages in indexes on LIVE. However, their results on CSIQ and TID2013 are rather diverse. Second, CGFA-CNN achieves comparable accuracies on LIVE Challenge comparing with DB-CNN [38] and ResNet50+ft [34], which are pre-trained on ImageNet [28] databases. This suggests that CNNs pre-trained on ImageNet [28] could extract relevant features for authentically distorted images.

Performance on individual distortion types on LIVE, CSIQ, and TID2013 are shown in Tables 3, 4 and 5. On LIVE, we also find that CGFA-CNN is superior to

Table 4 Average SRCC and PLCC results of individual distortion types across ten sessions on CSIQ database.

SRCC	JPEG	JP2K	WN	GB	PN	CC
BRISQUE[19]	0.806	0.840	0.732	0.820	0.378	0.804
M3[20]	0.740	0.911	0.741	0.868	0.663	0.770
FRIQUEE[21]	0.869	0.846	0.748	0.870	0.753	**0.838**
CORNIA[22]	0.513	0.831	0.664	0.836	0.493	0.462
HOSA[23]	0.733	0.818	0.604	0.841	0.500	0.716
dipIQ[31]	0.936	**0.944**	0.904	**0.932**	—	—
MEON[48]	**0.948**	0.898	**0.951**	**0.918**	—	—
DIQA [26]	0.835	0.931	0.927	0.893	**0.870**	0.718
DB-CNN [38]	**0.940**	**0.953**	**0.948**	**0.947**	**0.940**	**0.870**
CGFA-CNN	**0.950**	**0.939**	**0.956**	**0.941**	**0.952**	**0.897**

PLCC	JPEG	JP2K	WN	GB	PN	CC
BRISQUE[19]	0.828	0.887	0.742	0.891	0.496	0.835
M3[20]	0.768	0.928	0.728	0.917	0.717	0.787
FRIQUEE[21]	0.885	0.883	0.778	0.905	**0.769**	**0.864**
CORNIA[22]	0.563	0.883	0.778	0.905	0.632	0.543
HOSA[23]	0.759	0.899	0.656	0.912	0.601	0.744
dipIQ[31]	**0.975**	**0.959**	0.927	**0.958**	—	—
MEON[48]	**0.979**	0.925	**0.958**	0.846	—	—
DIQA [26]	—	—	—	—	—	—
DB-CNN[38]	**0.982**	**0.971**	**0.956**	**0.969**	**0.950**	**0.895**
CGFA-CNN	0.972	**0.953**	**0.969**	**0.955**	**0.942**	**0.893**

Table 5 Average SRCC results of individual distortion types across ten sessions on TID2013 database.

Method	#01	#02	#03	#04	#05	#06	#07	#08	#09	#10	#11	#12
BRISQUE[19]	**0.852**	0.709	0.491	0.575	0.753	0.630	0.798	0.813	0.586	0.852	0.893	0.315
M3[20]	0.748	0.591	0.769	0.491	0.875	0.693	0.833	0.878	0.721	0.823	0.872	0.400
FRIQUEE[21]	0.730	0.573	**0.866**	0.345	0.345	**0.847**	0.730	0.764	**0.881**	0.839	0.813	0.498
CORNIA[22]	0.756	**0.750**	0.7i 27	0.726	0.769	0.767	0.016	**0.921**	0.832	0.874	0.910	0.686
HOSA[23]	0.833	0.551	0.842	0.468	0.897	0.809	0.815	0.883	0.854	0.891	0.730	0.710
MEON[48]	0.813	0.722	**0.926**	**0.728**	**0.911**	**0.901**	**0.888**	0.887	0.797	0.860	0.891	0.746
DIQA [26]	**0.915**	**0.755**	**0.878**	**0.734**	**0.939**	**0.843**	**0.858**	**0.920**	0.788	**0.892**	**0.912**	**0.861**
DB-CNN [38]	0.790	0.700	0.826	0.646	0.879	0.708	0.825	0.859	**0.865**	**0.894**	**0.916**	**0.772**
CGFA-CNN	**0.812**	**0.804**	0.851	**0.845**	**0.910**	0.794	**0.867**	**0.933**	**0.866**	**0.914**	**0.922**	**0.763**

Method	#13	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24
BRISQUE[19]	0.359	0.145	0.224	0.124	0.040	0.109	0.724	0.008	0.685	0.764	0.616	0.784
M3[20]	0.731	0.190	0.318	0.119	0.224	-0.121	0.701	0.202	0.664	**0.886**	0.648	**0.915**
FRIQUEE[21]	0.660	0.076	0.032	0.254	**0.585**	0.589	0.704	0.318	0.641	0.768	0.737	0.891
CORNIA[22]	**0.805**	**0.286**	0.219	0.065	0.182	0.081	0.644	0.534	**0.862**	0.272	**0.792**	0.862
MEON[48]	0.716	0.116	**0.500**	0.177	0.252	**0.684**	**0.849**	0.406	0.772	0.857	**0.779**	0.855
DIQA [26]	**0.812**	**0.659**	0.407	**0.299**	**0.687**	-0.151	**0.904**	**0.655**	**0.930**	**0.936**	**0.756**	**0.909**
DB-CNN [38]	**0.773**	0.270	**0.444**	**0.646**	0.548	**0.631**	0.711	**0.752**	0.860	0.833	0.732	0.902
CGFA-CNN	0.757	**0.335**	**0.649**	**0.441**	**0.573**	**0.657**	**0.819**	**0.785**	**0.897**	**0.940**	0.711	**0.938**

Table 6 SRCC comparison on cross-database.

Method	CSIQ		TID2013	
	LIVE	TID2013	LIVE	CSIQ
BRISQUE[19]	0.847	0.454	0.790	0.590
M3[20]	0.797	0.328	**0.873**	0.605
FRIQUEE[21]	**0.879**	**0.463**	0.755	0.635
CORNIA[22]	0.853	0.312	0.846	**0.672**
HOSA[23]	0.773	0.329	0.594	0.462
DB-CNN [38]	**0.877**	**0.540**	**0.891**	**0.807**
CGFA-CNN	**0.891**	**0.533**	**0.898**	**0.774**

other methods in most distortions, except Fast-fading error which is not introduced into the pre-training dataset because there is no open-source or detailed description of it. On CSIQ, CGFA-CNN has obvious advantages compared with other methods, especially in contrast change and pink noise. On TID2013, CGFA-CNN achieves state-of-the-art performance in 10 of the 24 distortions and the whole effect standout accuracies other methods. Also, we find that CGFA-CNN performs well when the distortion shares similar artifacts with the distortion synthesized in the pre-training dataset. For example, additive Gaussian noise, additive noise in color components, and high-frequency noise are all grainy noise; quantization noise, and image color quantization with dither exhibit similar appearances; Gaussian blur, image denoising, and sparse sampling and reconstruction all introduce blur effects on the image. Therefore, although the pre-training dataset constructed in this paper does not cover all distortion types, CGFA-CNN still achieves impressive gains in performance.

4.4 Cross-database Experiment

To analyze the generalization ability of the proposed method, we train CGFA-CNN on one full database and evaluate on another database. Specifically, a model is trained on CSIQ and evaluated either on LIVE or TID2013. Results are reported in Table 6. It can be concluded that CGFA-CNN can be easier to generalize to distortions that have not been seen during training.

4.5 Comparison Among Different Experimental Settings

In this section, we first work with the performance of different feature aggregation layers investigated in this paper and number of GMM components K. Experiments are conducted on LIVE and results are shown in Fig. 4. We observe that SRCC gradually increases and eventually keeps stability as K increases. Besides, CGFA-CNN FV, CGFA-CNN VLAD, and CGFA-CNN BOW attain highly competitive prediction accuracy when K is set to 32, 64 and 1024, respectively. By contrast, CGFA-CNN FV is superior to CGFA-CNN VLAD and CGFA-CNN BOW.

Additionally, we report ablation studies to evaluate the design rationality of CGFA-CNN and the following comparative set of experiments are conducted: (1) to evaluate the effectiveness of the proposed FV layer, we use the maximum pooling (denoted as CGFA-CNN (MaxPool)) and average pooling (denoted as CGFA-CNN

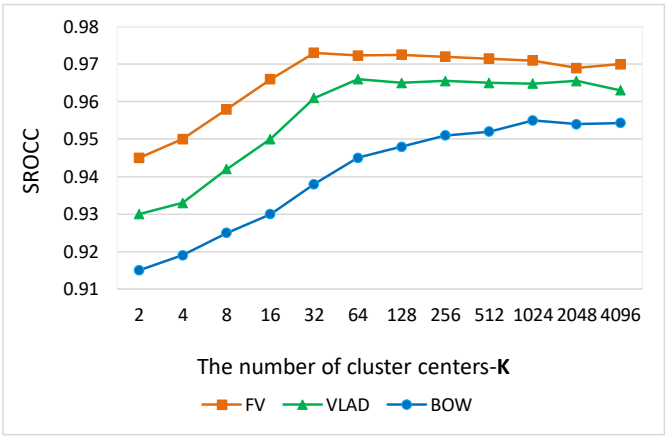

Figure 4 Relationship between the SRCC on LIVE of different feature aggregation layer and the number of K.

Table 7 SRCC with different settings.

Method	LIVE	CSIQ	TID2013	LIVE Challenge
CGFA-CNN (MaxPool)	0.915	0.893	0.778	0.766
CGFA-CNN (AvgPool)	0.909	0.876	0.755	0.761
CGFA-CNN (w/o CGU)	0.948	0.919	0.783	0.799
CGFA-CNN (single feature)	0.931	0.890	0.757	0.765
CGFA-CNN (BOW layer (K=1024))	0.955	0.936	0.808	0.791
CGFA-CNN (VLAD layer (K=64))	0.966	0.945	0.819	0.810
CGFA-CNN (w/o VGG-16)	0.970	0.950	0.836	0.672
CGFA-CNN (proposed)	0.973	0.953	0.841	0.837

(AvgPool)) instead; (2) to examine the validity of the CGU described in this work, we predict the quality score directly by regressing the output feature vector without CGU (denoted as CGFA-CNN (w/o CGU)); (3) to verify the necessity of the hierarchical feature extraction, we extract features only from high-level (Conv 5-2 of shared layers and Conv 4-3 of VGG-16) convolutional layers as descriptors (denoted as CGFA-CNN (single feature)); (4) to discuss the optimal settings of for feature aggregation layer, we set the BOW with $K = 1024$ (denoted as CGFA-CNN (BOW layer ($K = 1024$))), VLAD with $K = 64$ (denoted as CGFA-CNN (VLAD layer ($K = 1024$))) and FV with $K = 32$ (denoted as CGFA-CNN (proposed)); (5) to demonstrate the prediction accuracies on authentic distortions by involving VGG-16, we only include Sub-network I pre-trained on self-built dataset to extract features (denoted as CGFA-CNN (w/o VGG-16)). The results are demonstrated in Table 7. We empirically find that the proposed CGFA-CNN could achieve state-of-the-art prediction accuracies both on the synthetic and authentic distortion image quality databases. Besides, CGFA-CNN (w/o VGG-16) can only deliver promising performance on synthetic databases and its results on LIVE Challenge is inferior to CGFA-CNN (proposed), suggesting that authentic distortions cannot be fully fitted by synthetic distortions.

5 CONCLUSION

In this work, we propose an end-to-end learning framework for BIQA based on classification guidance and feature aggregation, which is named as CGFA-CNN. In the fine-tuning phase, except for the shared convolutional layers, the rest of the Sub-network I only participates in the forward propagation, and the parameters are fixed. The fused feature group is aggregated and encoded by the FV layer to obtain a fisher vector. Then the fisher vector is corrected by the CGU and obtain a quality-ware feature, which will be mapped to a quality score by the regression model. In the test phase, only forward propagation is required to obtain the quality score. The results on the four publicly IQA databases demonstrate that the proposed method indeed benefited image quality assessment. However, CGFA-CNN is not a unified learning framework because it takes two steps to pre-train and fine-tune. The promising future direction is to optimize CGFA-CNN for both distortion identification and quality prediction at the same time.

Abbreviations

NSS:Natural scene statistics;GMMS:Gaussian mixture models;PLCC:Pearson linear correlation coefficient; MOS:Mean opinion score;SRCC:Spearman's rank-order correlation coefficient.

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

About the authors

WeiPeng Cai received the B.E. degree in electronic information engineering from WuYi University, Guangdong, China, in 2018. He is currently pursuing the M.E. degree from Wuhan University, Wuhan, China. His current research interest is image quality assessment.

Cien Fan received the B.S degree in electronic instruments and measurement technology from Wuhan University, Wuhan, China, in 1998, the M.S. degree in measurement technology and instruments from Wuhan University, Wuhan, China, in 2001, and Ph.D. degree in radio physics from Wuhan University, Wuhan, China, in 2012. She is currently an associate professor in Electronic Information School, Wuhan University, Wuhan, China. Her research interest includes artificial intelligent, machine learning and image processing.

Zou Lian received the B.S degree in wireless communication from Nanjing University of Posts and Telecommunications in 1998 and the M.S. and Ph.D. degree in Wuhan University in 2000 and 2004.From 2005 to 2010, he was a lecturer in Electronic Information School of Wuhan University and from 2011 to 2016, he became an associate professor. Since 2017, he has been a professor with the Digital Signal Processing Lab in Electronic Information School of Wuhan University. His research interest includes image analysis and understanding, object detection, image super resolution and denoising.

Yifeng Liu received the Ph.D. degrees in Electronic Engineering from Wuhan University, Wuhan, China, in July 2016. He is currently the principal investigator of machine intelligence for the Innovation Center, China Academy of Electronics and Information Technology, Beijing, China. His current research interests include around deep learning, machine learning, computer vision, and knowledge engineering.

Yang Ma received the B.E. degree in electronic information engineering from Central China Normal University, Wuhan, China, in 2016. the M.S. degree in electronic information engineering from Wuhan University, Wuhan, China, in 2018.Her current research interest is image quality assessment.

Minyuan Wu received the B.S degree in optical instruments from Wuhan University, Wuhan, China, in 1984, the MS. degree in optical instruments from Wuhan University, Wuhan China,in 1989. He is currently an associate professor in Electronic Information School, Wuhan University, Wuhan,China. His research interest includes image identification, machine learning and machine vision.

Funding

This research was funded partly by the National Key R&D Program of China (Project No. 2017YFC0821603)

Availability of data and materials

We can provide the data.

Author's contributions

Weipeng Cai, Cien Fan and Yang Ma conceived the idea; Weipeng Cai, Cien Fan and Lian Zou performed the experiments, analyzed the data, and wrote the paper; Minyuan Wu developed the proofs. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Electronic Information School, Wuhan University, Wuhan, 430072 Wuhan, China. [2]National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (NEL-PSRPC), Beijing, 100041 Beijing, China.

References

1. Z. Wang, H. R. Sheikh, A. C. Bovik *et al.*, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*. CRC press Boca Raton, Florida, 2003, vol. 41, pp. 1041–1078.
2. K. Panetta, A. Samani, and S. Agaian, "A robust no-reference, no-parameter, transform domain image quality metric for evaluating the quality of color images," *IEEE Access*, vol. 6, pp. 10 979–10 985, 2018.
3. M. Jian, A. Ping, L. Shen, and L. Kai, "Reduced-reference stereoscopic image quality assessment using natural scene statistics and structural degradation," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2017.
4. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.
5. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2004, pp. iii–709.
6. Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
7. D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.
8. Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 202–211, 2009.
9. D. Liu, F. Li, and H. Song, "Image quality assessment using regularity of color distribution," *IEEE Access*, vol. 4, pp. 4478–4483.
10. J. Wu, Y. Liu, L. Li, and G. Shi, "Attended visual content degradation based reduced reference image quality assessment," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2018.
11. T. Brandão and M. P. Queluz, "No-reference image quality assessment based on dct domain statistics," *Signal Processing*, vol. 88, no. 4, pp. 822–833, 2008.
12. M. A. Saad, A. C. Bovik, and C. Charrier, "A dct statistics-based blind image quality index," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 583–586, 2010.
13. A. Moorthy and A. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, may 2010.
14. J. Li, J. Yan, D. Deng, W. Shi, and S. Deng, "No-reference image quality assessment based on hybrid model," *Signal, Image and Video Processing*, vol. 11, no. 6, pp. 985–992, 2017.
15. R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 717–728, 2009.
16. Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proceedings. International Conference on Image Processing*, vol. 1. IEEE, 2002, pp. I–I.
17. P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to jpeg2000," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, 2004.
18. A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008–2024, 2013.
19. A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
20. W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.
21. D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32–32, 2017.
22. P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1098–1105.
23. J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.
24. S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
25. Q. Yan, D. Gong, and Y. Zhang, "Two-stream convolutional networks for blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2200–2211, 2018.
26. J. Kim, A.-D. Nguyen, and S. Lee, "Deep cnn-based blind image quality predictor," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 11–24, 2018.
27. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
28. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.
29. S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.
30. J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Slected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, 2017.
31. K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, Aug 2017.

32. S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, 2018.
33. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.
34. J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 2017.
35. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
36. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
37. K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.
38. W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2018.
39. K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2017.
40. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.
41. H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society, 2010, pp. 3304–3311.
42. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.
43. F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.
44. R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.
45. A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.
46. K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, Nov 2015.
47. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
48. Y. Ma, W. Zhang, J. Yan, C. Fan, and W. Shi, "Blind image quality assessment in multiple bandpass and redundancy domains," *Digital Signal Processing*, vol. 80, pp. 37–47, 2018.
49. K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods." in *BMVC*, vol. 2, no. 4, 2011, p. 8.
50. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.
51. E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.
52. N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Color image database tid2013: Peculiarities and preliminary results," in *European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2013, pp. 106–111.
53. D. Ghadiyaram and A. C. Bovik, "Crowdsourced study of subjective image quality," in *2014 48th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2014, pp. 84–88.
54. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

Figures

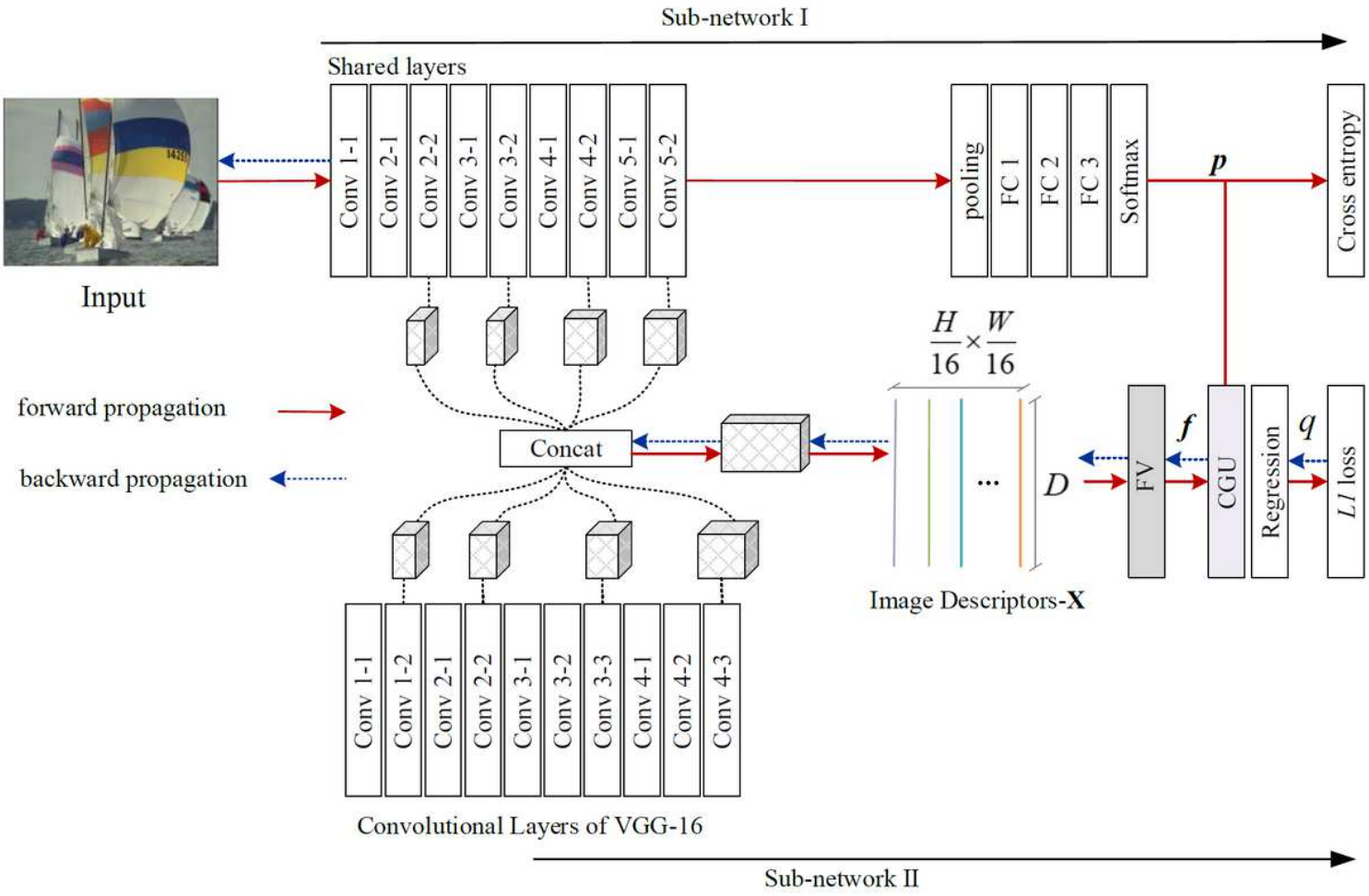

Figure 1

Illustration of CGFA-CNN congurations for BIQA, highlighting the feature aggregation layer (denoted as FV layer) and classication-guided gating unit (denoted as CGU). Features are extracted from the distorted image by Sub-network I.

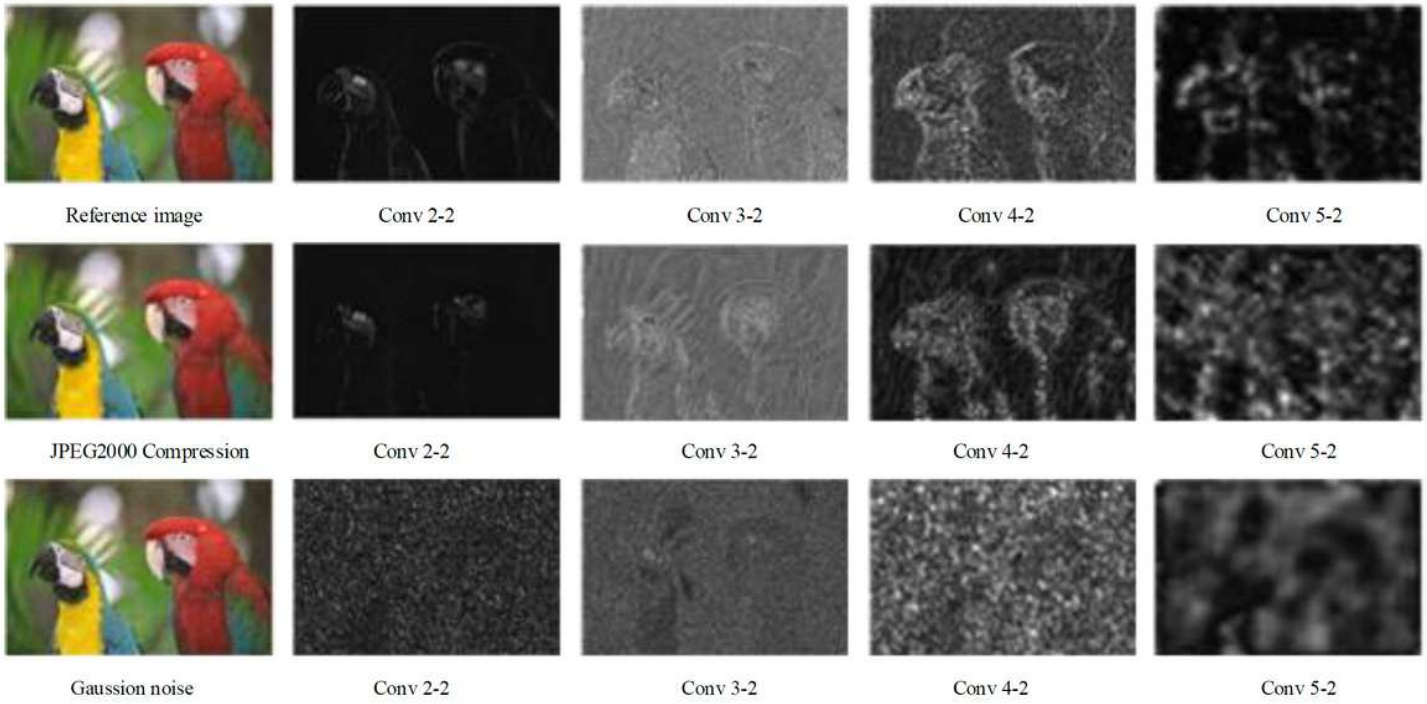

Figure 2

A comparison of several distortion types identified by Sub-network I.

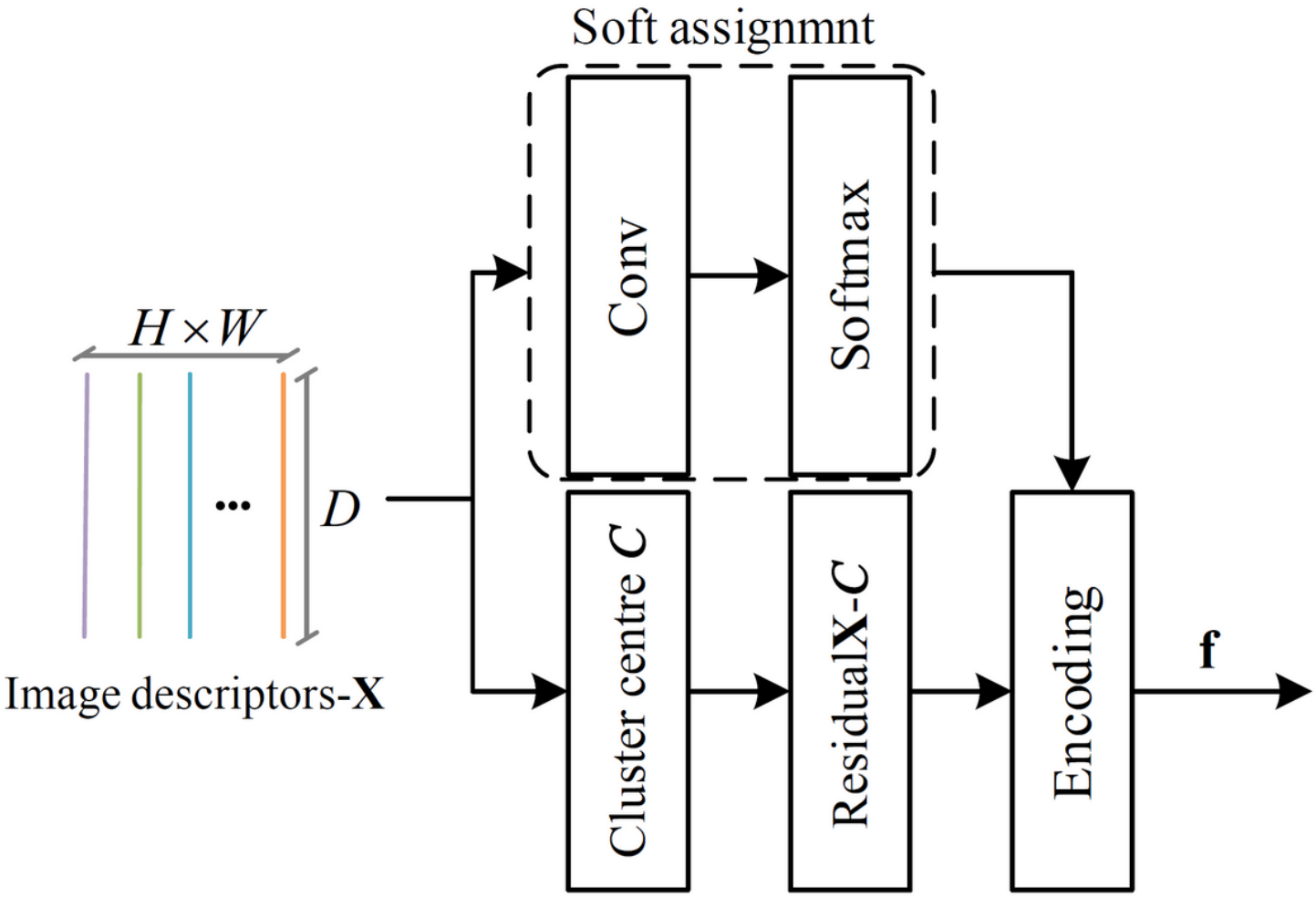

Figure 3

The configurations of the proposed FV layer. Convolution kernel size is 1 x 1.

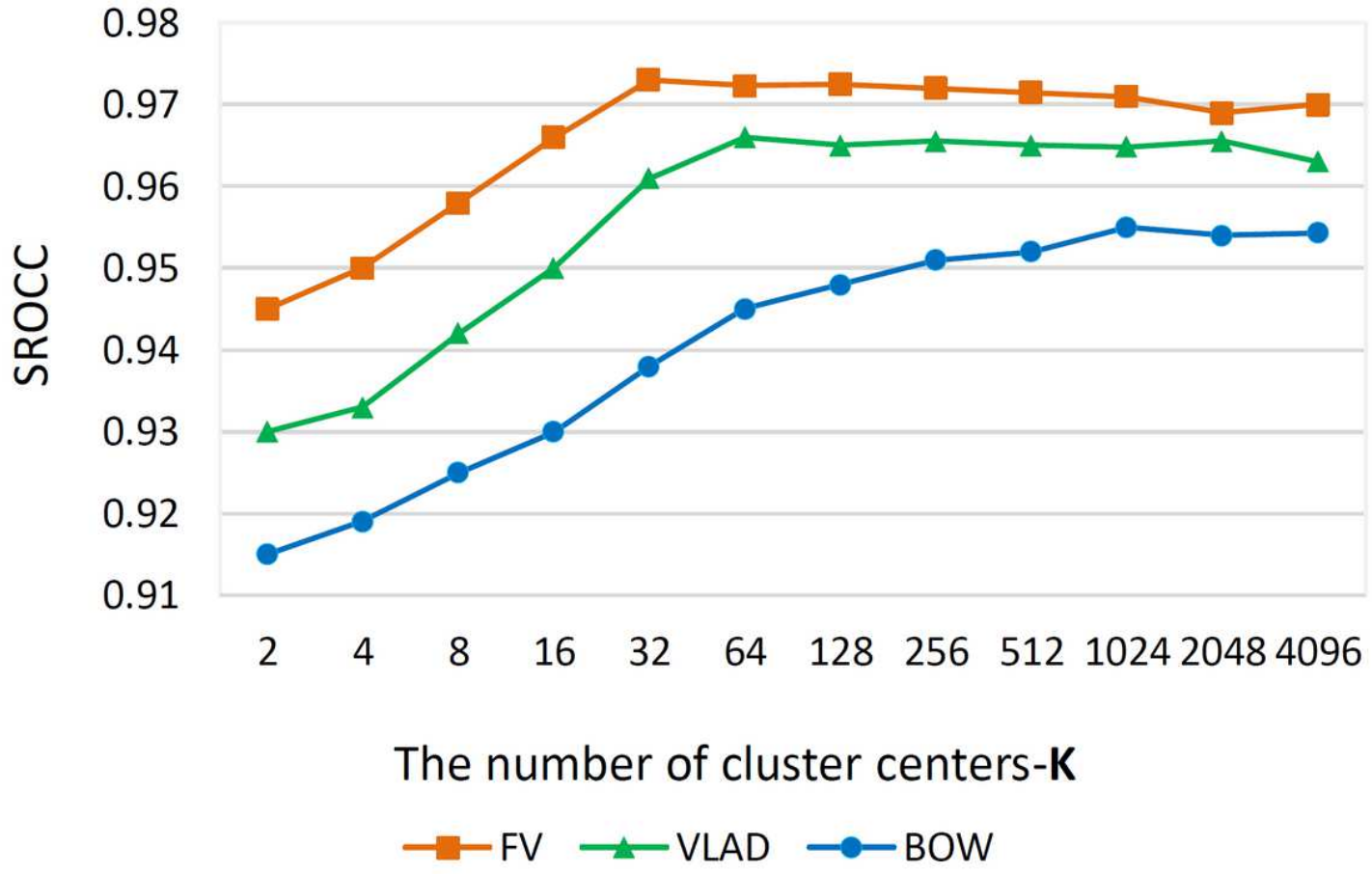

Figure 4

Relationship between the SRCC on LIVE of different feature aggregation layer and the number of K.