J. Vis. Commun. Image R. 80 (2021) 103296

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Blind image quality assessment with channel attention based deep residual network and extended LargeVis dimensionality reduction☆

Han Han [a,b], Li Zhuo [a,b,*], Jiafeng Li [a,b], Jing Zhang [a,b], Meng Wang [a,b,c]

[a] Faculty of Information, Beijing University of Technology, Beijing 100124, China
[b] Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China
[c] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China

ARTICLE INFO

Keywords:
Blind image quality assessment
ResNet-50
Channel attention mechanism
LargeVis dimensionality reduction

ABSTRACT

Image Quality Assessment (IQA) is one of the fundamental problems in the fields of image processing, image/video coding and transmission, and so on. In this paper, a Blind Image Quality Assessment (BIQA) approach with channel attention based deep Residual Network (ResNet)and extended LargeVis dimensionality reduction is proposed. Firstly, ResNet50 with channel attention mechanism is used as the backbone network to extract the deep features from the image. In order to reduce the dimensionality of the deep features, LargeVis, which is originally designed for the visualization of large scale high-dimensional data, is extended by using Support Vector Regression (SVR) to perform on a single feature vector data. The extended LargeVis can remove the redundant information of the deep features so as to obtain a low-dimensional and discriminative feature representation. Finally, the quality prediction model is established by using SVR as the fitting method. The low-dimensional feature representation and quality score of the image form the pair-wise data samples to train the fitting model. Experimental results on authentic distortions datasets and synthetic distortions datasets show that our proposed method can achieve superior performance compared with the state-of-the-art methods.

1. Introduction

Image Quality Assessment (IQA) is one of the fundamental problems in the image processing field. In recent years, many achievements have been made [1–3], in which No-Reference (NR) or Blind IQA (BIQA) has become the hotspots since the reference information is often unavailable (or may not exist) in many practical applications. IQA has been developed from specific distortion to non-specific distortion, and from single distortion to hybrid distortions. IQA for specific distortion is designed for certain distortion types such as color, blur, noise, and JPEG compression effect [4,5], thereby, they need to know the distortion type of the image in advance. In contrast, IQA for non-specific distortion does not need to distinguish the type and the cause of the distortion [6–9], which enable it to have wider application than IQA for specific distortion.

Since 2012, deep learning has made tremendous breakthroughs in image classification [10], natural language processing [11], speech recognition [12] and other fields. The most representative algorithm for deep learning is Convolutional Neural Network (CNN), which can learn the implicit relationships from big data directly to obtain a hierarchical feature representation from low-level visual features to high-level semantic features by simulating the multi-layered cognitive mechanism of the human brain. Compared with the traditional handcrafted features, deep features have a very prominent advantage in extracting multi-level features and context information from the image, and has stronger representative and discriminative capability [13].

Recently, the researchers have applied deep learning to BIQA, and proposed many models [1,2]. Deep learning based BIQA has become the mainstream work. The existing models can be divided into two categories. The first is to use an end-to-end training manner to obtain the mapping model between the image and the quality score directly [14,15]. And the second is to adopt a framework of "feature extraction + regression " [16,17], that is, extract the deep features of the image first, and then use a regression method to establish a mapping model between the deep features and the quality score. The key of both categories is to use large scale pair-wise data to train the deep CNN networks for extracting the deep features of the image.

In this paper, a BIQA method has been proposed under the

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author at: Faculty of Information, Beijing University of Technology, Beijing 100124, China.
E-mail address: zhuoli@bjut.edu.cn (L. Zhuo).

https://doi.org/10.1016/j.jvcir.2021.103296
Received 7 December 2020; Received in revised form 20 August 2021; Accepted 22 August 2021
Available online 25 August 2021

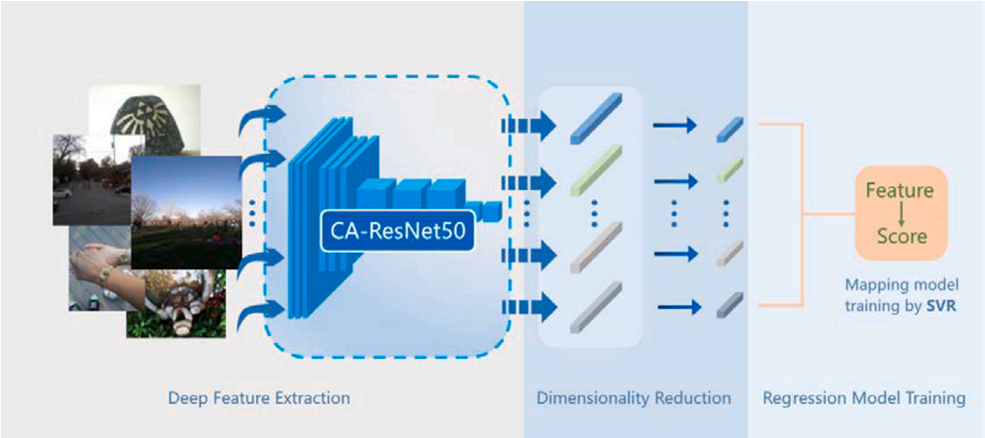

Fig. 1. The flowchart of the proposed method.

framework of "feature extraction + regression". First, channel attention based ResNet50 [10] is used as the backbone network to extract the deep features of the image [18]. By assigning different weights to the feature channels, the representative capability of the deep features can be improved significantly. In order to reduce the dimensionality of deep features, the LargeVis [19] dimensionality reduction method, which was originally designed for the visualization of big data, is extended to enable it to perform on a single feature vector data, Next, the dimensionality reduced features and quality score of the image form the pairwise data samples, which are used to train SVR to establish a mapping model between the low-dimensional feature representation and quality score. Extensive experiments demonstrate that the proposed method can achieve superior performance on both synthetic and authentic distortions datasets compared with the state-of-the-art methods.

2. Related work

BIQA methods can be roughly divided into traditional methods and deep learning-based methods. Traditional methods generally adopt a typical framework of "feature extraction + modeling", that is, extract the features of the image first, and then use a mathematical method to establish a mapping model between the features and the quality score of the image [20]. One of the most representative work is Distortion Identification-based Image Verity and INtegrity Evaluation(DIIVINE) [21], proposed by A. K. Moorthy et al. in 2011. DIIVINE first identifies the distortion type that degraded the image quality, and then uses the regression strategies to estimate the image quality. Blind Referenceless Image Spatial Quality Evaluator (BRISQUE) [7] proposed by A. Mittal in 2012 uses an asymmetric generalized Gaussian distribution to model the images in the spatial domain. In follows, Mean Subtracted Contrast Normalized (MSCN) coefficients was proposed to quantify the distorted degree. In 2013, Gao et al. proposed a heterogeneous property based on multiple kernel learning to assess image quality [22]. Their method combines the nonGaussianity, local dependence and exponential decay characteristics of natural images. This approach overcomes the one-sidedness of using only one type to assess the image quality. In 2015, Wu et al. proposed a quality prediction model based on KNN and new feature fusion method [23]. They combined the statistics results of multiple domains and multiple color channels as image features, and then assessed the quality of image based on non-parametric models. In 2016, Wang et al. proposed an image quality assessment method based on non-negative matrix factorization and extreme learning machine [24]. They used features based on non-negative matrix factorization to express image quality. Then they use a brand new extreme learning machine to learn the correct mapping between features and quality scores. In 2016, D.Ghadiyaram et al. proposed Referenceless Image Quality Evaluation Engine (FRIQUEE) [25] by capturing the consistency of the statistical information of the real-world distorted image and then used SVR to perform regression training on the calculated feature maps, achieving the state-of-the-art results on the authentic distortions datasets.

Since 2012, deep learning is gradually applied in BIQA by using large-scale training samples [1,2,14,15], which are mainly based on two kinds of frameworks. One is an end-to-end training manner [14,15,23] by integrating feature extraction and regression into a unified framework to establish a BIQA model with the mapping relationship between the input image and the quality score learned and trained by a CNN network [7,14]. Driven by this idea, Kang et al. proposed an IQA model in 2014 [6] that was also one of the earlier research work by using CNN for IQA. In order to solve the problem of too small data volume, this method crops the image into 32×32 blocks to improve the volume of training samples so as to achieve the stat-of-the-art performance, which also proved that CNN has great potential in IQA.

In 2018, Bosses et al. proposed DeepIQA [15] by expanding the IQA dataset through randomly dicing the image. In addition, at the training stage, the weights of each patch block are simultaneously returned. Then each patch and their weights are used to operate spatial pooling and predict the final image quality. The experimental results demonstrated that DeepIQA can perform IQA for synthetic and authentic distortions images.

Kede Ma et al. proposed a Multi-task End-to-end Optimized Deep Neural Network (MEON) [14] based on multi-task learning strategy. The network can realize two subtasks of distortion type prediction and IQA through sharing the network parameters between the two subtasks. This method uses synthetic distortions datasets to train the network model. So it can achieve an excellent performance on synthetic distortions datasets. However, the model fails if applied to the authentic distortions images. To solve this problem, Kede Ma et al. further proposed DB-CNN in 2019 [26] including two different CNN networks trained by synthetic and authentic distortions datasets respectively. The fully connected layer features of the two networks are fused using bilinear pooling to extract the deep features, and then the mapping relationship between the deep features and the quality score is fitted. The performance of this method on synthetic and authentic distortions datasets has been greatly improved, which brings a new idea for the current IQA research work.

Another kind of deep learning based IQA method is implemented under the framework of "deep features + regression", which is similar with the traditional methods. The difference is that the features are extracted by using CNNs. One of the representative work is deepBIQ proposed by Bianco et al. in 2017 [16] by adopting AlexNet [27] as the backbone network to extract fully connected layer features, and then using SVR to learn the mapping model between the deep features and quality score of the image. This method achieved superior results on the authentic distortions datasets. DeepRN proposed by Domonkos [17]

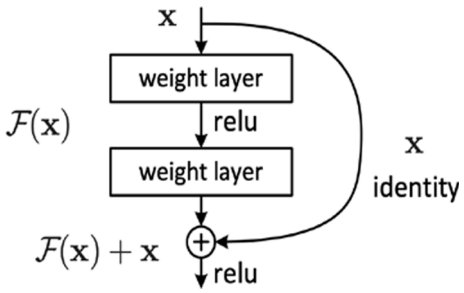

Fig. 2. structure of shortcut connections.

et al. in 2018 also adopted a similar framework by using ResNet101 as the backbone network for feature extraction, and then using SVR for fitting. In addition, the method used the spatial pyramid pooling module to enable the network to train the images with the original size, avoid losing the visual information contained in the image, obtaining a higher prediction accuracy.

The experimental results demonstrate that, these two kinds of deep learning based methods can obtain a comparable performance. The former is simple to implement, but it is also difficult to combine with other technical means. In contrast, the latter is more flexible and can take advantage of deep learning and the framework of "feature extraction + regression". Various technical means, such as feature dimensionality reduction, feature selection, and so on, can be easily integrated into it. The core is how to obtain the features with powerful representative and discriminative ability. In this paper, we adopt ResNet50 as the backbone network to extract the deep features of images, and introduce the channel attention mechanism into it to enhance the important features for IQA tasks and suppress irrelevant ones, so as to improve the representative ability of the features. Further, we use LargeVis to enlarge the inter-class distance and minimize intra-class, thus, improving the discriminative ability of the features. The experimental results show that channel attention and LargeVis can effectively improve the performance, yield the state-of-the-art prediction accuracy.

3. Proposed BIQA method

Based on the framework of "feature extraction + regression", a BIQA method is proposed in this paper, as shown in Fig. 1. At first, channel attention mechanism is integrated into ResNet50 [10] to enhance the representative capability of deep features by assigning different weights to the feature channels. The constructed network is denoted as CA-ResNet50, which is used as the backbone network to extract the features of the image. The fully connected layer features of CA-ResNet50 are extracted, whose dimension is 2048. Next, LargeVis [19] algorithm is extended to reduce the dimensionality of these deep features, which is originally designed for the visualization of big data, obtaining more compact and less redundant low-dimensional feature representation. On the one hand, it helps to reduce the computational complexity and storage space of the subsequent processing. On the other hand, LargeVis can enlarge the inter-class distance and minimize the intra-class distance to improve the discriminative ability of the deep features. Finally, SVR is utilized as a fitting method to establish the mapping model between the dimensionality reduced features and the MOS value of the image for predicting the score of a new image.

3.1. Feature extraction

3.1.1. ResNet50

ResNet was proposed by He et al. in 2015, the main contribution of which is to solve the problem that the classification accuracy decreases with the depth of the convolutional neural network. In this paper, the idea of residual learning was put forward, which can not only accelerate the CNN training process, but also effectively avoid the problem of gradient disappearance and gradient explosion.

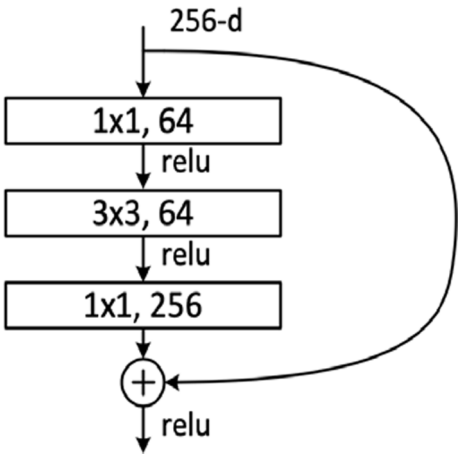

Fig. 3. Schematic diagram of Bottleneck.

Based on the idea of residual learning, He et al. designed a shortcut connection structure of identity mapping as shown in Fig. 2, where x is the input, $F(x)$ is the residual map, $H(x)$ is the ideal map, $H(x) = F(x) + x$. By converting $H(x)$ to $F(x)$, the output can be turned into a superposition of the input and residual maps, making the network more sensitive to the change between the input x and output $H(x)$. This design can enable the network to achieve the role of identity mapping in the process of forward propagation. In addition, there are more convenient paths for gradient conduction during back propagation.

In order to build a deeper network structure, He et al. also designed the Bottleneck structure shown in Fig. 3. They added a 1×1 convolution to the bottleneck structure, which can also reduce the dimension of the input. The Bottleneck structure is adopted in ResNet-50/101/152 networks.

In recent years, ResNet has been widely used in various computer vision tasks [28–30]. In this paper, ResNet50 is used for IQA. The experimental results show that ResNet50 can significantly improve the performance compared with the existing CNN networks.

3.1.2. Channel attention mechanism

Since the general CNN framework is regarded that each feature channel has the equal contributions to the tasks, the dependence among the feature channels is often ignored. The basic idea of the attention mechanism is to enable the network to selectively enhance the features contributing greatly to the task while suppressing the irrelevant features, so as to improve the representative capability of the deep features. Recently, various attention mechanisms have been proposed [18,31–34], such as channel attention, spatial attention, and so on, in which the channel attention is widely used in various computer vision tasks and has achieved outstanding performance. In this paper, channel attention mechanism is introduced into ResNet50. The diagram of channel attention module is shown in Fig. 4 [18].

Given the convolutional layer feature maps (fm) as input, average pooling operation is conducted on fm to generate a descriptor F_{AP}.

$$F_{AP} = AveragePool(fm) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} fm(i,j) \tag{1}$$

where H and W are the height and width of feature map, $fm(i, j)$ denotes the feature value at (i, j) location of feature map.

Then, CA module will learn the channel attention map $CNN(F_{AP})$. In order to ensure that this module can express the non-linear relationship among the channels, and its parameters can be learned, it is necessary for the CNN module to have at least two or more layers [35–37]. We built a module with two fully connected layers and an excitation

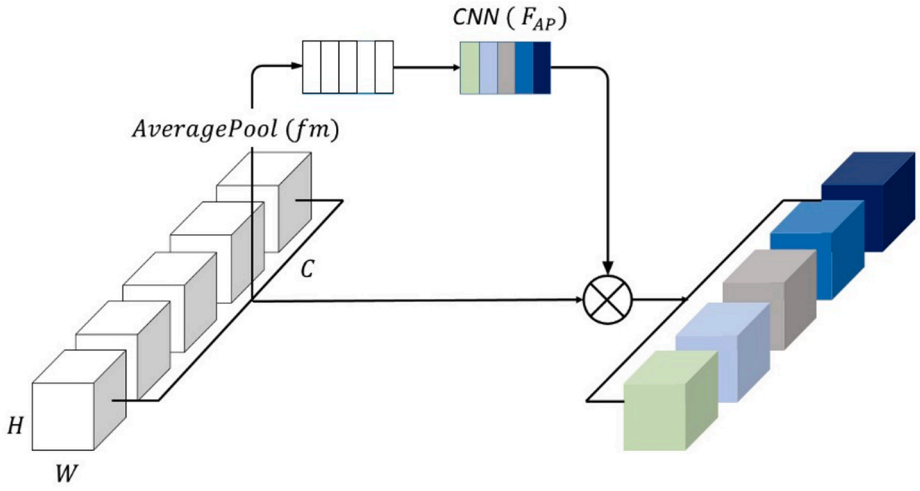

Fig. 4. Diagram of Channel Attention Module.

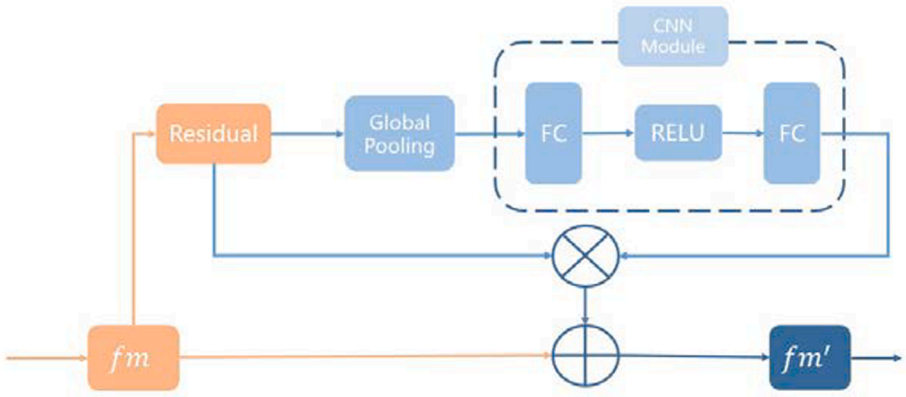

Fig. 5. The network structure of CA-ResNet50.

function RELU [38] in the middle layer as shown in Fig. 5 to ensure that the network scale will be small. F_{AP} is input to CA module, and the output is $CNN(F_{AP})$.

$$CNN(F_{AP}) = \sigma(W_2(RELU(W_1(F_{AP}))))) \tag{2}$$

where $W_1 \in R^{\frac{c}{r}\times c} W_2 \in R^{c\times\frac{c}{r}}$ There are two fully connected layers in this module. The first one is to reduce the dimensionality of *fm* to *r*, here, *r* is a hyper-parameter. The second one is to restore the dimensionality of the feature maps to the same as that of the original feature maps.

Finally, the learnt $CNN(F_{AP})$ is multiplied correspondingly to the original feature channels to obtain the weighted feature maps *fm'*.

$$fm' = CNN(F_{AP}) \cdot fm \tag{3}$$

Except for the hyper-parameter *r*, the remaining parameters of CA module are all learned during the training process. In the training process, the feature channels that are more important for the task will be assigned higher weights, which means that the representative capability of these "important features" has been enhanced. CA module is easy to implement and can be integrated into many existing CNN networks.

In this paper, the channel attention module is combined with ResNet as the backbone network to extract the deep features of the image, namely CA-ResNet50. The network structure of CA-ResNet50 is shown in Fig. 5.

3.1.3. Network training

In this paper, "pre-training + fine-tuning" strategy is exploited to train the CA-ResNet50 network. It is the most commonly used means in various computer vision tasks that pre-train the network model on ImageNet and then fine-tune the model parameters on the dataset of the specific task. Plenty of research results show that, compared with no pre-training, the strategy of "pre-training + fine-tuning" can usually achieve better performance. The existing DeepBIQ, DB-CNN and other deep learning based IQA methods have proved that pre-training on ImageNet is also effective for IQA tasks. So, we pre-trained ResNet50 on ImageNet dataset to obtain the initial network parameters. Then IQA datasets are utilized to fine-tune the network parameters of CA-ResNet50. ResNet50 was originally designed for image classification task. In order to perform fine-tuning, it requires to assign a category label for each image in IQA datasets. We quantized the image quality into 5 levels based on the MOS values as listed in Table 1, in which each image is assigned a category corresponding to the quality level and then relabeled with one of the five categories.

Table 1
The relation between the MOS range and the quality level.

MOS range	Quality Level
0–20%	very poor
20%–40%	poor
40%–60%	medium
60%–80%	good
80%–100%	very good

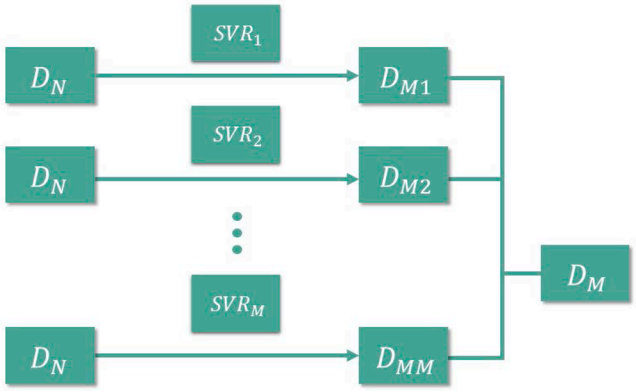

Fig. 6. Diagram of the extended LargeVis.

Finally, the features extracted from the last fully connected layer of the fine-tuned CA-ResNet50 network form a 2048-dimensional feature vector, which is used to build the IQA model. In the pre-training stage, Softmax is adopted as the excitation function. Cross entropy function is used as the loss function, shown as follows:

$$Loss = -\sum_i t_i ln y_i \tag{4}$$

where t_i represents the ground truth, and y_i represents the calculated Softmax value.

3.2. Extended LargeVis dimension reduction

LargeVis is a dimensionality reduction algorithm proposed by Tang Jian et al. in 2016 [19]. In addition to removing the redundancy of high-dimensional data, it can effectively enlarge the inter-class distance and minimize the intra-class distance in the high-dimensional space, thereby improving the discriminative capability of the features.

LargeVis was originally designed for the visualization of large scale high-dimensional data. It exploits the distance relationship among different clusters to perform dimensionality reduction, so it cannot conduct on a single vector. To solve this problem, we extended LargeVis to enable it suitable for processing a single feature vector [39].

As we all know, the essence of dimensionality reduction is to learn a mapping function f: $x \rightarrow y$, where x is the original high-dimensional data, and y is the low-dimensional data. Usually the dimension of y is smaller than that of x. How to determine the mapping function f is the key of dimensionality reduction algorithms. Different algorithms construct f according to different criteria. Compared with other fitting methods, SVR [40] can produce better fitting accuracy. Therefore, we explored SVR as the fitting method to obtain the mapping function f of LargeVis.

The specific implementation diagram using SVR is shown in Fig. 6. The whole process can be divided into a training stage and a dimensionality reduction stage.

In the training stage, we firstly used LargeVis to reduce the dimensionality of the entire training set to produce the corresponding low-dimensional data. High-dimensional data and low-dimensional data are used as pair-wise training samples. The samples are then used to train SVR to generate the mapping model f. It should be noted that it demands to build different SVR fitting model for each component of low-dimensional data. For example, if the dimensions of high-dimensional and low-dimensional data are N and M ($N > M$) respectively, M fitting models are needed to build. Each component of low-dimensional data needs one SVR fitting model.

In the dimensionality reduction stage, the appropriately trained SVR fitting models can be selected according to the required dimensionality. The low-dimensional data is finally obtained by sequentially combining all the individual fitting results.

3.3. Quality prediction model

In this paper, the dimensionality reduced feature vector and the image quality score form the pair-wise training samples. SVR is used to establish the mapping model, as shown in Fig. 7. The process is divided into two parts, *i.e* training phase and test phase. In the training phase, the paired data is used to train SVR that has the ability to characterize the mapping relationship between the sample feature vector and quality score. In the prediction stage, the low-dimensional feature vector of the image is input into SVR model. And the output of SVR model is the predicted quality score.

4. Experimental results and analysis

In order to verify the effectiveness of the proposed BIQA method in this paper, extensive experiments are conducted on authentic distortions datasets and synthetic distortions datasets respectively. And it is also compared with state-of-the-art BIQA methods.

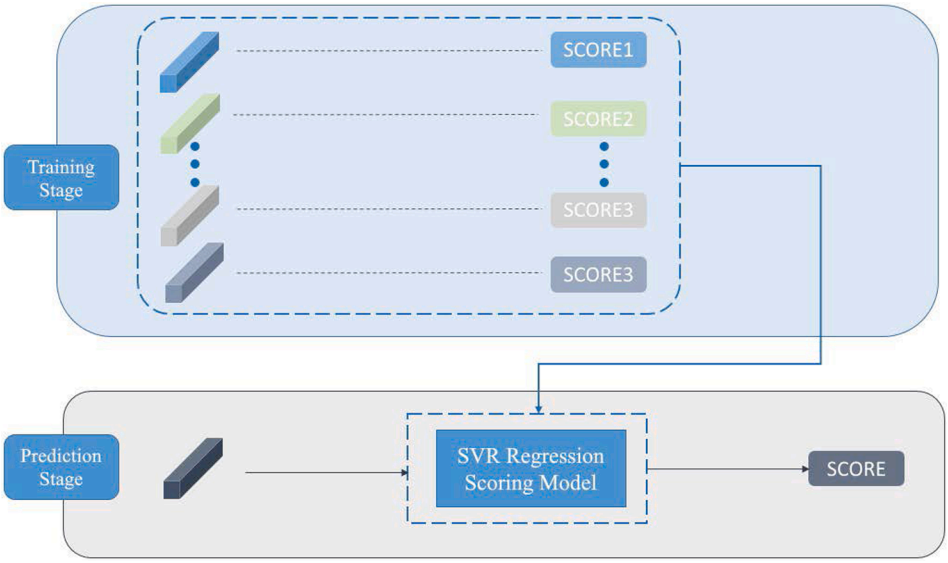

Fig. 7. Fitting the mapping relationship model between quality scores and feature vectors through SVR.

Table 2
Characteristics of the datasets.

Category	Dataset	Years	Reference images	Distorted images	Image size	MOS value range
Synthetic Distortions	LIVE	2006	29	779	640 × 512	[0,100]
	TID2013	2013	25	3000	512 × 384	[0,9]
Authentic Distortions	LIVE In the Wild Image Quality Challenge Dataset	2016	—	1169	640 × 512	[0,100]
	KonIQ-10 K	2017	—	10,073	1024 × 768	[0,5]

4.1. Datasets and performance evaluation criteria

There are two kinds of datasets commonly used in the IQA tasks. One kind is of synthetic distortions datasets, which generate multi-level distortions in a simulated manner according to a certain type of distortion, such as blur, compression, and noise. The subjective evaluation score is marked according to the degree of controllable distortions. The representative datasets include LIVE [41], TID2008 [42], TID2013 [43] and so on. The other kind is of authentic distortions datasets captured by using image acquisition equipments in the real world, which is hard to label. The representative datasets include LIVE In the Wild Image Quality Challenge Dataset (CLIVE) [44], KonIQ-10K [45] and so on. The characteristics of these datasets are shown in Table 2.

There are many assessment indicators used in the IQA tasks. Among them, Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Correlation Coefficient (SROCC) are the most commonly used. PLCC is used to measure the prediction accuracy, which is defined as:

$$PLCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma_x}\right)\left(\frac{y_i-\overline{y}}{\sigma_y}\right) \tag{5}$$

where $\overline{x}$ and $\overline{y}$ are the mean values of x_i and y_i, and σ_i is the corresponding standard deviation respectively. And SROCC is used to measure the monotonicity, which is defined as:

$$SROCC = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(r_{xi}-r_{yi})^2 \tag{6}$$

where r_{xi} and r_{yi} are the ranking positions in the individual data sequences of the two sets of scores.

The newly proposed Perceptually Weighted Rank Correlation (PWRC) [46] is also an assessment index designed based on the perceptual characteristics of the human eyes. Although PWRC is more reliable in evaluating the accuracy of the algorithm, in order to fairly compare with other methods, in this paper, we used PLCC and SROCC as the metrics, which are also the most commonly used in the IQA tasks.

4.2. Implementation details

In the experiments, the IQA datasets are randomly divided into training set and test set. We randomly chose 80% reference images along with their corresponding distorted images as the training set to fine-tune the CA-ResNet50 and the remaining parts are used as test set. The experiments are repeated 10 times and the median value is taken as the final experimental results.

To improve the volume of the training samples, each image was randomly cropped into 224 × 224 × 3 patches, which were used as training samples. The proposed method was implemented on PyTorch platform. In terms of computing hardware platform, Titan XP 12 GB graphics card was used for training and testing. In addition, the CPU model is i5-6400, and the memory size is 8 GB. The number of training iterations was set as 300 rounds, the batch size was 64, and the learning rate was 0.001. We used momentum and weight decay methods to optimize the training process to prevent overfitting. The weight decay rate was set as 1e-5, and the momentum parameter was 0.9.

Table 4
Comparison results by using different methods on LIVE In the wild image quality challenge dataset.

Method	PLCC	SROCC
FRIQUEE[25]	0.7060	0.6820
DIIVINE[21]	0.5577	0.5094
BRISQUE[7]	0.6100	0.6020
DeepIQA[15]	0.6800	0.6710
DB-CNN[26]	0.8690	0.8510
DeepBIQ[16]	0.9082	0.8894
DeepRN[17]	0.9300	0.9100
MetaIQA[48]	0.8350	0.8020
SAHNIQA[49]	0.8820	0.8590
OURS	**0.9459**	**0.9266**

Table 5
Comparison results by using different methods on KonIQ-10K dataset.

Method	PLCC	SROCC
BRISQUE[7]	0.70	0.70
DIIVINE[21]	0.62	0.58
KangCNN[1]	0.67	0.63
DeepRN[17]	0.95	0.90
DeepBIQ[16]	0.92	0.90
TL-Xception[50]	0.71	0.70
MetaIQA[48]	0.88	0.85
SAHNIQA[49]	0.91	0.90
OURS	**0.95**	**0.91**

4.3. Impact of CNN network architecture on performance

In order to verify the impact of CNN network architecture on the performance, various CNN networks are compared, including AlexNet [27], VGGNet [13], GoogleNet [47] and ResNet [10]. All of the networks were pre-trained on ImageNet, and then fine-tuned with IQA datasets. Afterwards, the fully connected layer features of each network were extracted to establish the quality predict model by SVR. Table 3 shows the dimensionality of each feature vector and the performance comparison results obtained by using eight CNN networks on the LIVE In the Wild Image Quality Challenge Dataset.

It can be seen from Table 3 that ResNet50 can achieve the highest PLCC and SROCC values compared with other CNN networks. Further, ResNet50 outperforms ResNet101, ResNet34, and ResNet18, which indicates that the depth of network layer is not necessarily high. Based on

Table 3
Comparison results by using different CNN networks on the LIVE In the wild image quality challenge dataset.

	AlexNet	VGG16	VGG19	VGG16bn	GoogleNet	ResNet18	ResNet34	ResNet50	ResNet101
FC-Vector Dimension	4096	4096	4096	4096	1024	2048	2048	2048	2048
PLCC	0.9082	0.9318	0.9300	0.9292	0.9188	0.9333	0.9325	**0.9335**	0.9294
SROCC	0.8894	0.9019	0.9175	0.8910	0.8852	0.9196	0.9225	**0.9238**	0.9176

Table 6
Comparison results by using different methods on LIVE dataset.

SROCC	JPEG	JP2K	WN	GB	FF	ALL
DIIVINE[21]	0.910	0.913	0.984	0.921	0.863	0.916
BRISQUE[7]	0.965	0.929	0.982	0.964	0.828	0.940
CORINIA[51]	0.947	0.924	0.958	0.951	0.921	0.942
FRIQUEE[25]	0.947	0.919	0.983	0.937	0.884	0.957
RankIQA[52]	0.978	**0.970**	**0.991**	**0.988**	0.954	0.981
DB-CNN[26]	0.972	0.955	0.980	0.935	0.930	–
SAHNIQA[49]	0.961	0.949	0.982	0.926	0.934	0.968
OURS	**0.981**	0.963	0.987	0.982	**0.974**	**0.982**
PLCC	**JPEG**	**JP2K**	**WN**	**GB**	**FF**	**ALL**
DIIVINE[21]	0.921	0.922	0.988	0.923	0.888	0.917
BRISQUE[7]	0.971	0.940	0.989	0.965	0.894	0.942
CORINIA[51]	0.962	0.944	0.974	0.961	0.943	0.935
FRIQUEE[25]	0.955	0.935	0.991	0.949	0.936	0.959
RankIQA[52]	0.986	0.975	**0.994**	**0.988**	0.960	0.982
DB-CNN[26]	**0.986**	0.967	0.988	0.956	0.961	–
SAHNIQA[49]	–	–	–	–	–	0.966
OURS	0.983	**0.976**	0.991	0.977	**0.974**	**0.984**

Table 7
Performance comparison results by using different methods on TID2013 dataset.

SROCC	JPEG	JP2K	GN	GB	ALL
DIIVINE[21]	0680	0.857	0.879	0.859	0.795
BRISQUE[7]	0.894	0.906	0.889	0.886	0.883
CORINIA[51]	0.912	0.907	0.798	0.934	0.893
FRIQUEE[25]	0.895	0.849	0.812	0.861	0.835
DeepIQA[15]	0.921	**0.948**	**0.938**	0.910	0.885
MEON[14]	0.919	0.911	0.908	0.891	0.912
DB-CNN[26]	0.894	0.916	0.813	0.859	0.851
OURS	**0.929**	0.937	0.921	**0.911**	**0.915**
PLCC	**JPEG**	**JP2K**	**GN**	**GB**	**ALL**
DIIVINE[21]	0.696	0.901	0.882	0.860	0.794
BRISQUE[7]	0.950	0.919	0.886	0.884	0.900
CORINIA[51]	0.960	0.928	0.778	0.934	0.904
FRIQUEE[25]	0.870	0.870	0.821	0.856	0.852
DeepIQA[15]	0.960	0.963	0.943	0.897	0.913
MEON[14]	**0.969**	0.924	0.911	**0.899**	0.912
DB-CNN[26]	–	–	–	–	0.869
OURS	0.955	**0.971**	**0.944**	0.889	**0.919**

the results, ResNet50 is adopted in this paper.

4.4. Comparison with the-state-of-art methods

In order to verify the effectiveness of our proposed BIQA method, we compared it with the existing IQA methods, including the traditional methods, such as FRIQUEE, BRISUQE, DIIVINE, and the deep learning-based methods, such as WaDIQaM-NR, DB-CNN, DeepBIQ, DeepRN. The experimental data of these methods were all referenced from the literatures.

Firstly, we conducted experiments on the two Authentic Distortions datasets, namely LIVE In the Wild Image Quality Challenge Dataset and KonIQ-10 K Dataset. The comparison results on the two Datasets are shown in Table 4 and Table 5 respectively.

It can be seen from Table 4 and Table 5 that, compared with the other methods, the proposed method can achieve a superior prediction accuracy on both of the two datasets. To further demonstrate the effectiveness of the proposed method, we conducted experiments on the two synthetic distortions datasets of LIVE and TID2013 and tested each distortion type individually.

Table 6 shows the comparison results on LIVE dataset. The distortion types of this dataset include JPEG compression (JPEG), JEPG2000 compression (JP2K), White Noise (WN), Gaussian blur (GB) and Fast Fading (FF). It can be seen that, the overall SROCC and PLCC values of the proposed method both exceed those of the state-of-the-art methods. We further found that the proposed method and RankIQA have their own advantages on different types of distortions. For example, for PLCC indicator, the proposed method can achieve better performance on two distortion types of JPEG and FF, but worse on JP2K WN and GB. RankIQA was originally designed for the synthetic distortions datasets by comparing the distortion degree between the images for ranking training. Therefore, it performs well on the synthetic distortion datasets. But the overall performance of the proposed method is still better than that of RankIQA, which fully proves the effectiveness of the proposed method on both authentic distortion datasets and synthetic distortion datasets.

Table 7 shows the comparison results by suing different methods on TID2013 dataset. TID2013 has a total of 24 distortion types. The common distortion types with other datasets include JPEG, JP2K, Gaussian Noise (GN), and GB. We chose these four representative distortion types for comparison. It can be seen that the proposed method can achieve a superior overall performance, though the SROCCs of some distortion types are not optimal.

Table 8
Comparison results of SROCC cross datasets.

Training	LIVE		TID2013		CLIVE	
Testing	TID2013	CLIVE	LIVE	CLIVE	LIVE	TID2013
BRISQUE[7]	0.358	0.337	0.790	0.254	0.238	0.280
FRIQUEE[25]	0.461	0.411	0.755	0.181	0.644	0.424
CORNIA[51]	0.360	0.443	0.846	0.293	0.588	0.403
DeepIQA[15]	0.462	—	—	—	—	—
DB-CNN[26]	0.524	**0.567**	**0.891**	0.457	0.746	0.424
OURS	**0.587**	0.541	0.882	**0.490**	**0.783**	**0.471**

Table 9
Comparison results of ablation experiments on different datasets.

PLCC	LIVE	TID2013	CLIVE
ResNet50	0.9837	0.9156	0.9335
ResNet50 + CANet	0.9842	0.9162	0.9395
ResNet50 + CANet + LargeVis	**0.9835**	**0.9189**	**0.9459**
SROCC	**LIVE**	**TID2013**	**CLIVE**
ResNet50	0.9803	0.9135	0.9238
ResNet50 + CANet	0.9829	0.9139	0.9241
ResNet50 + CANet + LargeVis	**0.9822**	**0.9154**	**0.9266**

4.5. Performance across different datasets

In order to verify the generalization ability of the proposed method, we conducted cross-dataset experiments. Three datasets, *i.e* LIVE, TID2013, and LIVE In the Wild Image Quality Challenge, are involved in the experiments. The experimental strategy is to train on one dataset and test on the other two datasets. The comparison results of SROCC are shown in Table 8. It can be seen that, compared with the other five methods, our method has obvious advantages.

4.6. Ablation study

To verify the impact of different components on the performance, we conducted the ablation study. The detailed results are presented in the following section.

4.6.1. Impact of different components on performance

The ablation experimental results on different datasets are shown in Table 9.

It can be seen that, compared with ResNet50, CA-ResNet50 can obtain higher performance on three datasets, especially on the LIVE In the Wild Image Quality Challenge Dataset. It means that the channel attention module can effectively improve the representative capability of the deep features, and thus, improving the performance. When using the extended LargeVis on the basis of CA-ResNet50, PLCC and SROCC

Table 10
Comparison results of different dimensions on the LIVE in the wild image quality challenge, TID2013 and KonIQ-10K datasets.

		2048	1024	512	256	128	64	32	16	8
LIVE in the Wild	**PLCC**	0.9395	0.9419	0.9457	0.9459	0.9456	**0.9461**	0.9432	0.9429	0.9428
	SROCC	0.9241	0.9218	0.9234	**0.9266**	0.9225	0.9264	0.9246	0.9217	0.9167
TID2013	**PLCC**	0.9162	0.9175	0.9186	0.9189	**0.9194**	0.9179	0.9159	0.9143	0.9138
	SROCC	0.9139	0.9148	0.9153	**0.9154**	0.9151	0.9144	0.9149	0.9130	0.9127
KonIQ-10 K	**PLCC**	0.9451	0.9447	0.9440	**0.9501**	0.9448	0.9399	0.9326	0.9390	0.9301
	SROCC	0.9106	**0.9124**	0.9119	0.9090	0.9063	0.9101	0.9056	0.9019	0.8990

Table 11
Comparison results by using different dimensionality reduction methods on KonIQ-10K dataset.

PLCC	KonIQ-10 K						
	1024	512	256	128	64	32	16
PCA	0.9285	0.9271	0.9179	0.9263	0.9312	**0.9356**	0.9143
LPP	0.6995	0.7545	0.7869	0.8106	0.8079	0.8561	0.8753
LLE	0.8939	0.8767	0.8785	0.8643	0.8512	0.8531	0.8311
LargeVis	**0.9447**	**0.9440**	**0.9501**	**0.9448**	**0.9399**	0.9326	**0.9390**
SROCC	**KonIQ-10K**						
	1024	**512**	**256**	**128**	**64**	**32**	**16**
PCA	0.9086	0.9023	0.9017	0.8951	0.9012	0.9041	0.8954
LPP	0.6978	0.6745	0.7341	0.7592	0.7974	0.8134	0.8099
LLE	0.8534	0.8165	0.8242	0.7941	0.7569	0.7479	0.7185
LargeVis	**0.9124**	**0.9119**	**0.9090**	**0.9063**	**0.9101**	**0.9056**	**0.9019**

Table 12
Comparison results by using different dimensionality reduction methods on TID2013 dataset.

PLCC	TLD2013						
	1024	512	256	128	64	32	16
PCA	0.9107	0.9132	0.9099	0.9141	0.9033	0.8975	0.9018
LPP	0.4613	0.5412	0.5763	0.6120	0.6781	0.7284	0.7961
LLE	0.8675	0.8297	0.8021	0.7584	0.7864	0.7491	0.7612
LargeVis	**0.9148**	**0.9153**	**0.9154**	**0.9151**	**0.9144**	**0.9149**	**0.9130**
SROCC	**TLD2013**						
	1024	**512**	**256**	**128**	**64**	**32**	**16**
PCA	0.9104	0.9093	0.9141	0.9127	0.9064	0.9171	0.8861
LPP	0.4961	0.5473	0.5964	0.6529	0.6977	0.7695	0.7864
LLE	0.8343	0.8427	0.8390	0.7846	0.7962	0.7461	0.7379
LargeVis	**0.9175**	**0.9186**	**0.9189**	**0.9194**	**0.9179**	**0.9159**	**0.9143**

have been further improved. This also fully proves the effectiveness of the various components in the proposed method.

4.6.2. Impact of extended LargeVis dimensionality reduction on performance

The dimension of the fully connected layer features extracted from CA-ResNet50 is 2048. We used the extended LargeVis method to reduce the dimensionality of the feature vector. Table 10 shows the impact of different dimensions on the performance on the LIVE in the Wild Image Quality Challenge, TID2013 and KonIQ-10K datasets.

From Table 10, it can be seen that, for all the datasets, by using the extended LargeVis, the optimal performance can be obtained. Since the type distortions of the images in each dataset are different, the optimal dimensionalities and performance are also different. On the whole, when the dimension is reduced to 256, in most cases, the optimal results can be obtained. Therefore, in this paper, we set the dimension of the extended LargeVis method as 256.

In order to further verify the performance of the extended LargeVis, we compared the extended LargeVis with other dimensionality reduction methods, including Principal Component Analysis (PCA), Locality Preserving Projection (LPP) and Locally linear embedding (LLE). The comparison results on the KonIQ-10 K and TID2013 datasets are shown in Table 11 and Table 12 respectively.

From Table 11 and Table 12, it can be seen that, compared with the other three dimensionality reduction methods, the extended LargeVis yields a higher performance with a large margin. Except for 32 dimension using PCA, these three dimensionality reduction methods usually have a negative effect on the performance. However, the extended LargeVis can improve the performance when the dimensions are 1024, 512, 256,128 and 64. This is because the extended LargeVis can enlarge the inter-class distance and minimize the intra-class distance, thus improving the discriminative ability of the deep features.

5. Conclusions

Based on a framework of “feature extraction + regression”, we proposed a BIQA method based on the channel attention mechanism and extended LargeVis dimensionality reduction. Channel attention mechanism is embedded into ResNet50 to extract the deep features with powerful representative capability. Then, the extended LargeVis is used to reduce the dimensionality of the deep features. The quality assessment model is finally established by training SVR with the dimensionality reduced low-dimensional features. We compared the proposed method with several mainstream IQA methods. Experimental results on synthetic distortions datasets and authentic distortions datasets showed that our method can obtain better performance.

Funding

This work is supported by the National Natural Science Foundation of China (Grant Number 61871006, 61971016), and Beijing Municipal Education Commission Cooperation Beijing Natural Science Foundation (No. KZ201810005002, No. KZ201910005007).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional Neural Networks for No-Reference Image Quality Assessment, 2014.
[2] L. Kang, P. Ye, Y. Li, D. Doermann, Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 2791–2795.
[3] Y. Liang, J. Wang, X. Wan, Image quality assessment using similar scene as reference. European Conference on Computer Vision, Springer, Cham, 2016.
[4] Q. Yan, Y. Xu, X. Yang, No-reference image blur assessment based on gradient profile sharpness, in: 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2013, pp. 1–4.
[5] S.A. Golestaneh, D.M. Chandler, No-reference quality assessment of JPEG images via a quality relevance map, IEEE Signal Process Lett. 21 (2) (2014) 155–158.
[6] L. Liu, H. Dong, H. Huang, A.C. Bovik, No-reference image quality assessment in curvelet domain, Signal Process. Image Commun. 29 (2014) 494–505.
[7] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (2012) 4695–4708.
[8] A.K. Moorthy, A.C. Bovik, A two-step framework for constructing blind image quality indices, IEEE Signal Process Lett. 17 (2010) 513–516.
[9] M.A. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (2012) 3339–3352.
[10] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv e-prints, 2015, pp. arXiv:1512.03385.
[11] S. Ji, S.V.N. Vishwanathan, N. Satish, M.J. Anderson, P. Dubey, BlackOut: Speeding up Recurrent Neural Network Language Models With Very Large Vocabularies, arXiv e-prints, 2015, pp. arXiv:1511.06909.
[12] Haşim Sak, A. Senior, K. Rao, et al., Fast and accurate recurrent neural network acoustic models for speech recognition, Comput. Sci. (2015).
[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Computational and Biological Learning Society, 2015, pp. 1–14.
[14] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-end blind image quality assessment using deep neural networks, IEEE Trans. Image Process. 27 (2018) 1202–1213.
[15] S. Bosse, D. Maniry, K. Müller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Trans. Image Process. 27 (2018) 206–219.
[16] S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, Signal, Image Video Process. 12 (2018) 355–362.
[17] D. Varga, D. Saupe, T. Szirányi, DeepRN: A content preserving deep architecture for blind image quality assessment, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), 2018, pp. 1–6.
[18] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1.
[19] J. Tang, J. Liu, M. Zhang, Q. Mei, Visualizing large-scale and high-dimensional data, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, 2016, pp. 287–297.
[20] A.C. Bovik, Automatic prediction of perceptual image and video quality, Proc. IEEE 101 (2013) 2008–2024.
[21] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: from natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (2011) 3350–3364.
[22] X. Gao, F. Gao, D. Tao, Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning, IEEE Trans. Neural Networks Learn. Syst. 24 (12) (2013) 2013–2026.
[23] Q. Wu, H. Li, F. Meng, K.N. Ngan, B. Luo, C. Huang, B. Zeng, Blind image quality assessment based on multichannel feature fusion and label transfer, IEEE Trans. Circuits Syst. Video Technol. 26 (3) (2016) 425–440.
[24] S. Wang, C. Deng, W. Lin, G.-B. Huang, B. Zhao, NMF-based image quality assessment using extreme learning machine, IEEE Trans. Cybern. 47 (1) (2017) 232–243.
[25] D. Ghadiyaram, A.C. Bovik, Perceptual quality prediction on authentically distorted images using a bag of features approach, J. Vision 17 (2017) 32.
[26] W. Zhang, K. Ma, J. Yan, et al., Blind image quality assessment using a deep bilinear convolutional neural network, IEEE Transactions on Circuits and Systems for Video Technology 30 (1) (2018) 36–47.
[27] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[28] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (2015) 211–252.
[30] T. Lin, M. Maire, S.J. Belongie, L.D. Bourdev, R.B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context. CoRR abs/1405.0312 (2014). arXiv preprint arXiv:1405.0312 (2014).
[31] J. Ba, V. Mnih, K. Kavukcuoglu, Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755 (2014).
[32] K. Gregor, I. Danihelka, A. Graves, D.J. Rezende, D. Wierstra, Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623 (2015).
[33] V. Mnih, N. Heess, A. Graves, Recurrent models of visual attention, in: Advances in Neural Information Processing Systems, 2014, pp. 2204–2212.
[34] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[35] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794–1801.
[36] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: theory and practice, Int. J. Comput. Vision 105 (3) (2013) 222–245.
[37] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning, International Machine Learning Society, Haifa, 2010, pp. 807–814.
[38] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.
[39] Z. Zhuo, Z.J.S. Zhou, Low dimensional discriminative representation of fully connected layer features using extended LargeVis method for high-resolution remote sensing image retrieval, Sensors (Basel) 20 (2020) 4718.
[40] H. Drucker, C.J. Burges, L. Kaufman, A.J. Smola, V. Vapnik, Support vector regression machines, in: Advances in Neural Information Processing Systems, 1997, pp. 155–161.
[41] H.R. Sheikh, Z. Wang, L. Cormack, A.C. Bovik, LIVE image quality assessment database release 2 (2005). URL http://live.ece.utexas.edu/research/quality (2005).
[42] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, TID2008-a database for evaluation of full-reference visual quality assessment metrics, Adv. Modern Radioelectronics 10 (2009) 30–45.
[43] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, Color image database TID2013: Peculiarities and preliminary results, in: European Workshop on Visual Information Processing (EUVIP), IEEE, 2013, pp. 106–111.
[44] D. Ghadiyaram, A. Bovik, Massive online crowdsourced study of subjective and objective picture quality, IEEE Trans. Image Process. 25 (2015) 1.
[45] H. Lin, V. Hosu, D. Saupe, KonIQ-10K: Towards an ecologically valid and large-scale IQA database. arXiv preprint arXiv:1803.08489 (2018).
[46] W. Qingbo, et al., A perceptually weighted rank correlation indicator for objective image quality assessment, IEEE Trans. Image Process. (2018).
[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[48] Zhu H, Li L, Wu J, et al. MetaIQA: Deep Meta-learning for No-Reference Image Quality Assessment. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[49] S. Su, Q. Yan, Y. Zhu, et al., Blindly assess image quality in the wild guided by a self-adaptive hyper network. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[50] H. Otroshi-Shahreza, A. Amini, H. Behroozi, No-reference image quality assessment using transfer learning, in: 2018 9th International Symposium on Telecommunications (IST), IEEE, 2018, pp. 637–640.
[51] P. Ye, J. Kumar, L. Kang, 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.
[52] X. Liu, J. Weijer, A.D. Bagdanov, RankIQA: Learning From rankings for No-reference Image Quality Assessment, IEEE Computer Society, 2017.