Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

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

No-reference stereoscopic image quality assessment based on global and local content characteristics

Lili Shen [a], Xiongfei Chen [a], Zhaoqing Pan [a,b,*], Kefeng Fan [c], Fei Li [d], Jianjun Lei [a]

[a] School of Electrical and Information Engineering, Tianjin University, Weijin Road, Tianjin 300072, China
[b] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[c] China Electronics Standardization Institute, Beijing 100007, China
[d] Guangzhou SequoiaDB Co., Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 15 August 2020
Revised 29 September 2020
Accepted 10 October 2020
Available online xxxx
Communicated by Steven Hoi

Keywords:
Stereoscopic images quality assessment
Cross-fusion
Multi-scales pooling
Asymmetric convolution block
Weighted average

ABSTRACT

No-reference stereoscopic images quality assessment (NR-SIQA) via deep learning has gained increasing attention. In this paper, we propose a no-reference stereoscopic image quality assessment method based on global and local content characteristics. The proposed method simulates the perception route of human visual system, and derives features from the fused view and single view through the global feature fusion sub-network and local feature enhancement sub-network. As for the fused view, a cross-fusion strategy is applied to model the process in the V1 visual cortex, and the multi-scales pooling (MSP) is utilized to integrate context information under different sub-regions for effective global feature extraction. As for the single view, the asymmetric convolution block (ACB) is introduced to strengthen the local information description. By jointly considering the fused view and single view, the proposed network can efficiently extract the features for quality assessment. Finally, a weighted average strategy is applied to estimate the visual quality of stereoscopic image. Experimental results on 3D quality databases demonstrate that the proposed network is superior to the state-of-the-art metrics, and achieves an excellent performance.

1. Introduction

In the past decades, 3D imaging technology and multimedia applications have developed rapidly, and more and more 3D productions are applied to people's life. However, the stereoscopic content may be contaminated by various distortions during the processes of acquisition, processing, transmission and display, which will affect human visual experience [1–3]. In addition, the study on stereoscopic image quality assessment (SIQA) is still immature due to the binocular visual characteristics. Therefore, it is necessary to find an efficient stereoscopic image quality assessment (SIQA) algorithm.

SIQA can be classified into two categories: subjective SIQA and objective SIQA. Subjective SIQA algorithms require observers to evaluate the image quality directly. As human being is the final receiver of an image, subjective SIQA algorithms are more reliable. However, subjective SIQA is time-consuming, expensive, and cannot be applied to practical scenarios [4]. Therefore, objective SIQA has gained great attentions in recent years. Objective SIQA algorithms are generally classified into three categories, full-reference (FR) [5,6], reduced-reference (RR) [7,8] and no-reference (NR) [9–11] algorithms, depending on how much reference image information is accessible. In practice, the reference images are often unavailable, so no-reference stereoscopic image quality assessment (NR-SIQA) is preferred by multimedia applications. NR-SIQA algorithms extract the discriminative features from the distorted images to predict the image quality without any reference data [12]. Hand-crafted NR-SIQA algorithms mainly extract the features based on human visual system (HVS) [13] or natural scene statistics (NSS) [14], and utilize machine learning methods such as support vector regression (SVR) for feature training. Feature extraction based on HVS refers to establishing a model that can simulate the binocular vision characteristics. Feature extraction of NSS is mainly based on statistical models. It is supposed that natural images have certain statistical characteristics, which can be changed by distortions. However, these methods are not robust enough to achieve better image quality prediction, and their application scenarios are limited.

* Corresponding author at: School of Electrical and Information Engineering, Tianjin University, Weijin Road, Tianjin 300072, China.
E-mail address: zqpan3-c@my.cityu.edu.hk (Z. Pan).

https://doi.org/10.1016/j.neucom.2020.10.024
0925-2312/

In recent years, convolutional neural network (CNN) has shown remarkable power in the field of computer vision, such as image classification [15,16], scene segmentation [17], and image retrieval [18]. Inspired by this, CNN is gradually used in image quality assessment (IQA). It can automatically learn features and regress them together to build quality prediction models by using raw images as input. This kind of feature extraction method greatly improves the robustness of image quality prediction accuracy. The performance of a CNN-based IQA algorithm is mainly determined by special network structure. Kang et al. [19] utilized one convolutional layer and one pooling layer to extract useful features for 2D image quality assessment (2D IQA), and evaluated the whole image quality by averaging the predicted scores of all image patches. However, the network structure is too shallow to capture the advanced information. Therefore, to increase the learning ability of the network, researchers extended the depth of networks. Bosse et al. [20] proposed a deep CNN-based NR-IQA algorithm, which consists of 10 convolution layers and 5 pooling layers. Moreover, a weight branch is used in the regression module for learning a weight for each local quality.

Compared with 2D IQA, SIQA needs to consider the binocular visual characteristics, which increases the difficulty of designing SIQA algorithms. To address this problem, most methods concatenate the left and right features directly, and use the fused features for the quality regression. For example, Fang et al. [21] designed a siamese network to learn the high-level semantic information from the left and right views, and combined them for final quality regression. Li et al. [22] proposed a two-step training network, and utilized the fused features for image quality estimation. However, these raise questions whether fused features suffice, and whether this concatenation method can model the binocular fusion process in the V1 cortical region. It is well known that the HVS is a hierarchical structure [23]. There are multiple interactions between the left and right views in human brain for 3D imaging process [12]. In other words, the image quality perception needs to consider both the fused view and single view information. In addition, the human brain fuses corresponding points of left and right views into a 3D image, while concatenating the feature maps directly can separate them away in these existed networks. Hence, in this paper, the single view features are introduced into the final quality regression to make up for the lack of information, and concatenate these feature maps by cross-fusion to model the 3D imaging effectively.

In the HVS, image signals go through visual areas with different roles, and their perceived qualities are determined in the frontal lobe [24]. Therefore, different structures are utilized to simulate these visual areas. As for the fused view, the global feature extraction is enhanced to represent the cyclopean information formed in the visual cortex regions. Only utilizing single pooling will lose some information inevitably, while multiple poolings can perceive context information with different scales through changing the size of receptive field [25]. Thus, the multi-scales pooling (MSP) is applied to the global feature fusion sub-network. The MSP has two parallel pooling layers with different receptive fields, and this can provide more context information for global feature extraction. As for the single view, strengthening local features can describe planar content in details relative to the fused view. Based on the fact that a square convolution kernel distributes learned knowledge in an imbalance manner, and aggravating this imbalance can acquire more efficient information [26], the asymmetric convolution block (ACB) is introduced in the local feature enhancement sub-network. An ACB includes three parallel convolution kernels with compatible sizes. It can learn more local information from single view.

Our contributions can be summarized as follows:

(1) In order to model the fusion process in the V1 cortex, the cross-fusion is applied to the left- and right-view feature maps. Moreover, the MSP structure is added into the global feature fusion sub-network to integrate features under different receptive fields, which can help this sub-network to learn complete global information from the fused riew.
(2) To address the problem of information loss, and learn more local features from the single view, the ACB is used to enrich the feature space for the single view. It pays more attention to the local information which is in accordance with the human visual characteristics.
(3) The proposed CNNs takes into account the global information of the fused view and the local information of the single view. It can deeply model the visual information processing mechanism in HVS, and enhance the ability of feature representation.

The rest of this paper is organized as follows. Section 2 briefly reviews the related works on image quality assessment. Section 3 introduces the proposed NR-SIQA algorithm in details. Section 4 presents the experimental results and analysis to demonstrate the advantages of the proposed network. Finally, Section 5 concludes this paper.

2. Related work

Machine learning has developed from traditional method to deep learning [27,28]. Based on this fact, the objective IQA algorithms have been developed rapidly. In this section, we briefly review the related works including hand-crafted-based NR-IQA algorithms and deep neural network-based NR-IQA algorithms.

2.1. Hand-crafted-based NR-IQA algorithms

The hand-crafted features are derived from empirical observations, which can be mapped to an image quality score. Saad et al. [29] extracted the features through NSS model which is based on the image discrete cosine transform coefficients, and utilized a Bayesian inference method to predict the image quality score. Mittal et al. [30] learnt the features from the empirical distribution of locally normalized luminance under a spatial NSS model. The gradient information of an image is a potentially predictive feature for IQA. Xue et al. [31] designed a NR-IQA model that utilized the joint statistics of gradient magnitude map and Laplacian of Gaussian response. Liu et al. [32] explored the gradient statistical information, including gradient magnitude and gradient orientation, and estimated an image quality score with the AdaBoosting back-propagation neural network. The perceptual qualities of stereoscopic images mainly depend on the characteristics of binocular vision, such as binocular fusion, binocular suppression and binocular energy response (BER). Therefore, researchers simulated the binocular visual characteristics to design NR-SIQA models. Chen et al. [33] designed a cyclopean image to simulate the fused image formed in the human brain. Liu et al. [11] presented a NR-SIQA algorithm by joinly considering the meaningful superpixel and the significant binocular perception models. Zhou et al. [34] proposed a NR-SIQA algorithm based on the complementary local patterns of BER and binocular rivalry response (BRR). Zhou et al. [9] extracted features from the self-similarity of BRR and binocular orientation selectivity in the 3D image. Finally, the SVR is used to drive the image quality score. However, the prediction accuracy of these hand-crafted methods depends on the

validity of features, which reduces the robustness of prediction performance.

2.2. Deep neural network-based NR-IQA algorithms

Over the past years, deep neural network-based NR-IQA algorithms have shown great success in image quality score prediction. Gao et al. [35] suggested to constitute multi-level representations from a pretrained VGGnet model. Gu et al. [36] proposed an attention-based pooling network to address the pooling problem, which can automatically generate local weight while estimating local quality. Po et al. [37] improved the conventional CNN-based NR-IQA algorithm by avoiding homogenous patches appearing in the network training and quality estimation stages. Ji et al. [38] measured the image quality by structural semantics and spatial semantics. Shen et al. [39] proposed a feature-segmentation strategy to train CNN model without any pre-processing. In [40,41], generative adversarial network (GAN) provided a novel direction for quality assessment. However, these algorithms are not suitable for SIQA because the binocular visual characteristics are not considered.

By fully considering the characteristics of stereoscopic images, deep neural network-based NR-SIQA algorithms have been proposed. Yang et al. [24] firstly utilized segmented stacked auto-encoder to simulate the complex structure of the visual cortex based on the visual perception route from eyes to the frontal lobe. Karimi et al. [42] extracted NSS features from two synthesized images, and utilized a stacked neural network model to estimate the stereoscopic image quality. Yang et al. [43] introduced the 2D features from the monocular images and the 3D features from the depth perception map, and utilized three deep belief network models to map these features to the final score. Messai et al. [44] proposed a NR-SIAQ algorithm based on HVS, which utilizes the cyclopean image patches to train four CNN prediction models. Oh et al. [45] presented a two-step network structure-based NR-SIQA algorithm. In the first step, they utilized the structure similarity (SSIM) to provide a ground-truth for each image patch to train local prediction model. In the second step, they aggregated these local features into global abstractions, and updated local model parameters by using subjective mean opinion score (MOS). Zhang et al. [46] took the image patches from the left view, right view and difference image as the input, and built a three-column CNN to better model human perception. Zhou et al. [12] proposed a dual-stream interaction network, where the interaction process between the left and right sub-networks occurs multiple times. To further improve the stereoscopic image quality prediction accuracy, a NR-SIQA algorithm is proposed by considering the fused and single views content characteristics.

3. Proposed method

In the process of visual processing, the retina first perceives the original information, and then transmits them to multiple visual areas with different roles, such as V1-V5. In the V1, the left and right views are fused, while the other visual cortexes can extract high-level information. To simulate the process of visual perception effectively, and address the lack of information during feature extraction, a NR-SIQA algorithm is proposed based on the global and local content characteristics, in which the global feature fusion sub-network and local feature enhancement sub-network are utilized to achieve the roles of different visual areas. In the global feature fusion sub-network, a cross-fusion strategy is proposed to fuse the left and right views which occurs in the V1 region. Besides, the MSP is introduced to integrate context information under different receptive fields for complete global information extraction. In the local feature enhancement sub-network, the ACB is utilized to enhance the ability of the local feature extraction. Combining the global and local content, the image quality score can be predicted accurately.

Fig. 1 shows the architecture of proposed network. To extract useful low-level features from the left and right views, the primary sub-network is used at the beginning of the proposed network, which is composed of two groups of convolution operations. And each group includes two convolution layers. Inspired by the visual processing mechanism in HVS, the extraction of high-level features consists of two types of sub-networks, namely global feature fusion sub-network and local feature enhencement sub-network. In the global feature fusion sub-network, the cross-fusion is conducted to form fused feature maps. And the MSP structure is made up of two convolution layers and three pooling layers. Finally, a convolution layer and a pooling layer are used to get the fused view features after multi-scales concatenation. For each local feature enhancement sub-network, it has two ACBs and three pooling layers. The ACB consists of three types of convolution operation with kernel sizes of 3×3, 3×1 and 1×3. The input of the network is preprocessed image patch with the size of 32×32. The regression module consists of two branches: a weight branch and a regression branch. The final output is the weighted average of image patches quality scores. The detailed configurations are shown in Table 1, where only the structure of the left view is listed. "Pool_2" and "Pool_4" represent two kinds of pooling with the window sizes of 2×2 and 4×4. In the regression module, only the weight branch is listed.

3.1. Proposed MSP-based global feature fusion method

When eyes receive the external images information, the optic nerve transmits the input visual signals to the lateral geniculate nucleus (LGN) which is an important relay center for the visual information to the human cerebral cortex [12]. Then, the complex interaction occurs between the left- and right-view signals in the V1 cortex. Moreover, the visual processing mechanism can be modeled by DCNN. However, most SIQA methods fuse these signals via a simple concatenation, which leads to the separation of corresponding features maps from the left and right views, and cannot model the fusion process well in the V1 cortex. Therefore, in order to integrate the left- and right-view feature maps closely, the left- and right-view feature maps are divided according to their channels and concatenated, namely cross-fusion. The formula of fused feature maps FM_C is defined as follows:

$$FM_C = H\Big(FM_l^{(1:\frac{c}{n})}, FM_r^{(1:\frac{c}{n})}, FM_l^{(\frac{c}{n}+1:\frac{2c}{n})}, FM_r^{(\frac{c}{n}+1:\frac{2c}{n})}, \cdots FM_l^{(\frac{c(n-1)}{n+1}:c)}, FM_r^{(\frac{c(n-1)}{n+1}:c)}\Big), \tag{1}$$

where H(.) is the concatenation function, FM_l and FM_r denote the left- and right-view feature maps with a spatial resolution of $H \times W$ and C channels, n is the slice number. To determine the optimal n, the performance of different slice numbers are tested, and the results are listed in Table 2. In Table 2, "Number 0" denotes the original feature maps. "Number 2" means that the left and right feature maps are equally divided into two parts and then fused. Other cases are similar. It can be seen that the left- and right-view features can be better fused when the slice number is 2.

Furthermore, the global feature enhancement operation is performed on the fusion map. When applying CNN for feature extraction, the size of receptive field can roughly indicate how much context information we receive [25]. The larger receptive field is, the more global information we can perceive. The pooling operation is commonly used to change the receptive field and

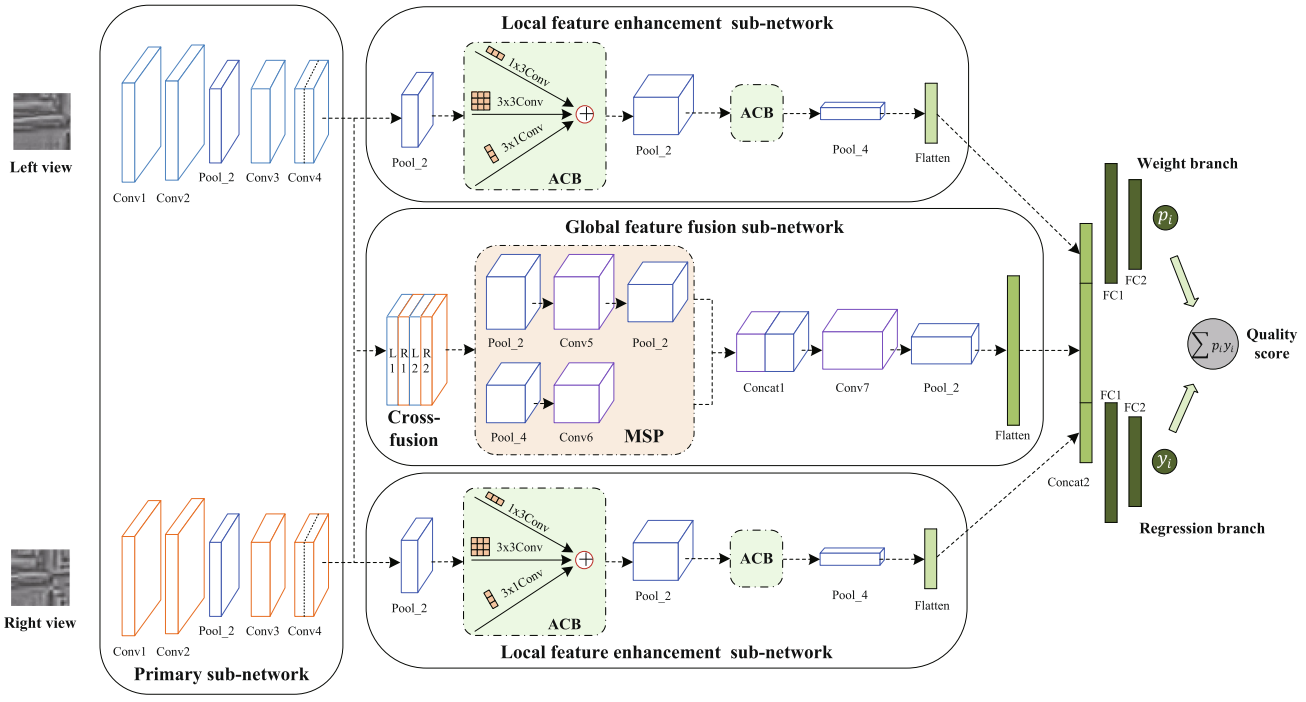


Fig. 1. The architecture of proposed network.

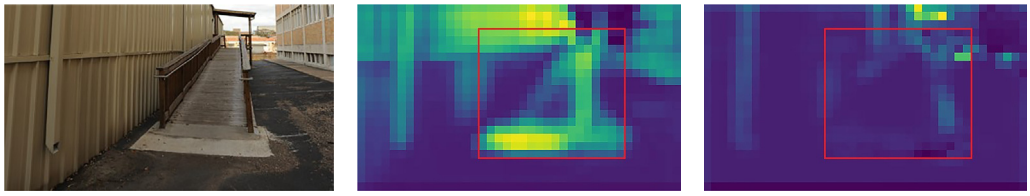
Table 1
The detailed configurations of proposed method.

	Layer	Feature map	Parameters
Primary sub-network	Input	32 × 32×1	-
	Conv1	32×32×32	3×3×32
	Conv2	32×32×32	3×3×32
	Pool_2	16×16×32	2×2
	Conv3	16×16×64	3×3×64
	Conv4	16×16×64	3×3×64
Local feature enhancement sub-network	Pool_2	8×8×64	2×2
	ACB1	8×8×128	(3×3 + 3×1 + 1×3)× 128
	Pool_2	4×4×128	2×2
	ACB2	4×4×256	(3×3 + 3×1 + 1×3)× 256
	Pool_4	1×1×256	4×4
	Faltten	256	-
Global feature fusion sub-network	Cross-fusion	16×16×128	-
	Pool_2	8×8×128	2×2
	Pool_4	4×4×128	4×4
	Conv5	8×8×128	3×3×128
	Conv6	4×4×128	3×3×128
	Pool_2	4×4×128	2×2
	Concat1	4×4×256	-
	Conv7	4×4×256	3×3×256
	Pool_2	2×2×256	2×2
	Flatten	1024	-
Regression module	Concat2	1536	-
	FC1	512	512
	FC2	256	256
	output	1	-

reduce the data dimension. The existed SIQA algorithms use a single pooling to obtain the receptive field. However, it may loss some context information during down-sampling. To address this problem, the multiple-pooling operation is proposed, which can generate context information with different receptive fields. Specifically, the cross-fusion result is taken as the input of MSP. The MSP starts with two parallel maximum pooling layers, where the window sizes are 2 × 2 and 4 × 4 for different pooled representations. With the use of different poolings, the convolution operations can extract features from multiple sub-regions. To concatenate these feature maps with two kinds of sizes, the 2 × 2 pooling is conducted on the larger one. Finally, the output of MSP which contains feature maps under different receptive fields are further learned to form the complete global representation. To verify the effectiveness of proposed key module, the corresponding feature maps are visualized in Fig. 2, in which the whole stereoscopic image is taken as the input for observation, and the feature maps from all channels are added to observe whether the global information of the image is enhanced. Fig. 2 (a)-(c) represent the left-view image, and its feature maps from proposed model, and the model without MSP (replacing MSP with single pooling) under JP2K distortion. As we can see, the staircase is the main object with a higher saliency in the image. And the image perception quality is mainly affected by the staircase. (b) preserves most of contour feature, while a lot of useful information is lost in (c). When these features are transmitted to the final image quality estimation, complete contour information leads to higher prediction accuracy. Generally, the MSP module provides more context information to form the complete global features.

Table 2
Performance under different slice numbers on LIVE databases.

Database	LIVE Phase I				LIVE Phase II			
Number	0	2	4	8	0	2	4	8
PLCC	0.967	**0.972**	0.965	0.967	0.950	**0.953**	0.943	0.946
SROCC	0.956	**0.962**	0.954	0.957	0.949	**0.951**	0.943	0.940



(a) (b) (c)

Fig. 2. Examples of feature maps in global feature fusion sub-network. (a) Left-view image distorted by JP2K. (b) Feature map in the last convolution layer of fusion sub-network based on MSP. (c) Corresponding feature map without MSP.

3.2. Proposed ACB-based local feature enhancement method

For enhancing the local features in the single view, the ACB is used to learn more local information. The core of feature extraction for CNN is the validity of kernel parameters. The parameters of standard square convolution kernel located in the central criss-cross positions can extract more effective features [26]. However, in the training of neural network, the learned parameters of convolution kernels cannot be changed artificially. In order to enhancing the feature extraction ability, three compatible convolution kernels are applied to accomplish this task instead of changing parameters directly based on the additivity [26]:

$$I*K^1 + I*K^2 = I*(K^1 \oplus K^2), \tag{2}$$

where I is a matrix, K^1 and K^2 are convolution kernels with compatible sizes, $*$ is the convolution operation, and $\oplus$ is the element-wise addition. In other words, if three kernels are utilized to operate on one input and the outputs are summarized, then the same output can be got when adding up these three kernels and convoluting the input. In addition, [47] has proved that the asymmetric convolution cannot work well in the low-level layers. Therefore, the ACB is adopted to extract the significant features in the local feature enhancement sub-network. Since different kernel sizes of ACB could have different network learning ability, a group of kernel sizes, including 3×3, 5×5, and 7×7, are tested to determine the optimal kernel size, and the results are listed in Table 3. It can be seen that when the 3×3 kernel size is used, the network achieves the best performance in terms of Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC). Hence, the kernels of 3×3, 3×1 and 1×3 are used to build the ACB. To intuitively show the feature extraction ability of the adopted ACB, the extracted feature maps are visualized in Fig. 3. Figs. 3 (a)-(c) are the left-view image, and its feature maps from proposed model and the model without ACB (replacing ACBs with standard square convolution kernels) under BLUR distortion. In (b) and (c), they both have some contour information, like house. In addition, (b) presents some details clearly, while in (c), some local information from the image is lost, such as the trash can and door. These objects have a high contrast to the background, and they are easy to attract human's attention. From these figures, it can be observed that the adopted ACB can efficiently extract the local useful features for increasing the learning ability of the proposed network.

3.3. Regression and training

In the quality regression stage, a weight branch is integrated into the regression module [20] (Fig. 1). Specifically, the weight branch outputs an α_i for each patch. By applying a Rectified Linear Unit (ReLU) to α_i and adding a small stability term ϵ, we can get α_i^*. Then the normalization is applied to α_i^* to get the weight p_i. With the quality score y_i and normalized weight p_i of each patch, the whole image quality $\hat{q}_i$ is predicted as follow:

$$\hat{q}_i = \sum_{i=1}^{N_p} p_i y_i = \frac{\sum_{i=1}^{N_p} \alpha_i^* y_i}{\sum_{i=1}^{N_p} \alpha_i^*}, \tag{3}$$

where N_P is 66 denoting the number of patches over each image.

Our network is trained iteratively by using backpropagation. After a number of epochs, we can get our trained model. One epoch refers to the period during which each image patches from the training set can be trained once. In every epoch, the optimization is based on mini-batch which is commonly used. It is worth noting that patches from one image must be distributed over the same mini-batch. Therefore, the batch size is set based on the whole image. Specifically, each mini-batch contains one stereoscopic image, leading to the batch size of 66 patch pairs. Furthermore, supposed that q_i denotes the DMOS value of the stereoscopic image i, and we can get the loss function which need to be minimized as follows:

$$Loss = |q_i - \hat{q}_i|^2. \tag{4}$$

The ADAM method [48] is used to train the network. The initial learning rate is set to 0.001, and the exponential decay method is used to reduce the learning rate. The learning rate decay factor is set to 0.8. Each convolution layer and fully connected layer is followed by the ReLU [49]. Furthermore, the moving average model is used to control the updating speed of network parameters which can improve the performance of the testing data. The dropout is also applied to each fully connected layer and randomly set the output of neurons to zero with a probability of 0.5 to reduce overfitting.

4. Experimental results and analysis

4.1. Datasets and indicators

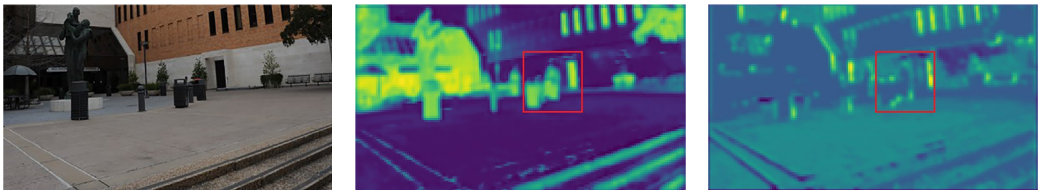
To validate the performance of proposed algorithm, the experiments are performed on two 3D image databases: LIVE 3D Phase I [50] and LIVE 3D Phase II [51]. In our experiments, 80% of the distorted stereoscopic images are randomly selected as the training sets and the remaining 20% are selected as the testing sets. For each database, the network training is repeated 10 times and the median is recorded.

LIVE Phase Databases: LIVE Phase I database consists of 20 reference stereoscopic images and 365 distorted stereopairs. It is composed of five common distortion types, including JPEG2000 compression (JP2K), JPEG compression (JPEG), additive white Gaussian noise (WN), Raleigh fast fading channel distortion (FF) and Gaussian blur (BLUR). They are all symmetrical distortion. While LIVE Phase II database consists of 8 reference stereoscopic images and 360 distorted stereopairs, where 120 stereopairs are symmetrical and 240 stereopairs are asymmetrical. The distortion types are the same as that in LIVE Phase I database. Besides, these two databases both have DMOS values.

Evaluation metrics: In order to evaluate the performance of SIQA algorithm quantitatively, two effective indexes are adopted: SROCC and PLCC. They range from 0 to 1, where a bigger value of

Table 3
Performance under different kernel size on LIVE databases.

Database	LIVE Phase I			LIVE Phase II		
Kernel Size	3X3	5X5	7X7	3X3	5X5	7X7
PLCC	**0.972**	0.962	0.963	**0.953**	0.950	0.951
SROCC	**0.962**	0.961	0.959	**0.951**	0.942	0.945



(a) (b) (c)

Fig. 3. Examples of feature maps in local feature extraction sub-network. (a) Left-view image distorted by BLUR. (b) Feature map in the first ACB. (c) Corresponding feature map without ACB.

SROCC and PLCC indicates the method has better image quality prediction accuracy.

4.2. Stereoscopic image preprocessing

In the process of image preprocessing, the patch sampling strategy is modified to handle the homogeneous patches problem. Given a distorted stereoscopic image, the cyclopean image [33] is synthesized and the local contrast normalization is performed on left and right views [30]. Then the cyclopean image is divided into 220 non-overlapping image patches with the size of 32×32, and these patches are sorted by variances. The performance with different screening ratio on LIVE databases is shown in Table 4. In Table 4, five ratios are tested, including 10%, 20%, 30%, 40% and 50%. According to the results, if the ratio is too small, parts of information will be lost, resulting in performance degradation. However, when the ratio is too large, the redundant information can also reduce performance of the algorithm. Therefore, the middle 30% of patches are selected with positions recorded in the whole cyclopean image. Based on these positions, the patches can be sampled from normalized left and right images. Finally, one stereoscopic image is made up of 66 image patch pairs to train the model.

4.3. Performance comparison

To prove the efficiency of the proposed NR-SIQA method, the proposed method is compared with two FR-SIQA algorithms, namely Md [5] and Wang [6], three NR-SIQA algorithms developed by Zhou [9], Shen [10] and Liu [11], three CNN-based algorithms including Zhang [46], Oh [45] and Fang [21], and a DBN-based algorithm Yang [43]. It should be noted that the results of compared SIQA methods are from their corresponding original papers. The PLCC and SROCC metrics testing on LIVE Phase I database are shown in Tables 5 and 6. In Tables 7 and 8, the performance of PLCC and SROCC is compared based on LIVE Phase II database. The best results are in bold. As we can see from Tables 5–8, our algorithm outperforms the state-of-art FR-SIQA and NR-SIQA algorithms. It is noted that our proposed method doesn't achieve the best performance on some distortion types, such as WN and FF. These two distortion types will be discussed in the Section 4.5. At the same time, it can be found that the deep structure-based SIQA algorithms are more effective than the traditional hand-crafted algorithms among the whole database. One possible reason is that these algorithms can automatically learn effective features from images. However, we find that Liu [11] is competitive with the deep structure-based algorithms, which is due to the consideration of both monocular and binocular visual features.

4.4. Ablation experiments

In order to verify the validity of the proposed algorithm, the ablation experiments are conducted. The experimental results of several baseline models are listed in Table 9. The first row shows the performance of proposed method without local feature enhancement (LFE) sub-network and MSP module (using single 2×2 pooling). The second row shows the results of training without local feature enhancement sub-network. The third row represents that there is no MSP. And the fourth row indicates that the ACB modules are replaced by the standard square convolution kernels. Some conclusions can be made from these results. The local feature enhancement sub-network can make up for the loss in the fusion process. By adding the single-view features, the algorithm performance is improved. Meanwhile, compared with the traditional convolution, more information can be learned from the ACB. The features under different receptive fields can be obtained by applying the MSP to the fused view. Compared with the baseline model w/o LFE-MSP, in our algorithm, the PLCC is improved by 0.025, and the SROCC is improved by 0.025 on LIVE Phase I database. Besides, the PLCC is improved by 0.028, and the SROCC is improved by 0.026 on LIVE Phase II database. These results prove that our proposed module can bring improvement for the baseline models.

Table 4
Performance under different ratios on LIVE databases.

Database	LIVE Phase I					LIVE Phase II				
Ratio	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
PLCC	0.932	0.959	**0.972**	0.952	0.967	0.932	0.941	**0.953**	0.951	0.949
SROCC	0.932	0.950	**0.962**	0.951	0.955	0.933	0.934	**0.951**	0.936	0.944

Table 5
PLCC comparison on the LIVE Phase I database.

	SIQA	JP2K	JPEG	WN	BLUR	FF	ALL
FR	Md [5]	0.935	0.735	0.931	0.963	0.832	0.928
	Wang [6]	0.909	0.617	0.949	0.951	0.778	0.924
NR	Zhou [9]	0.945	0.748	0.953	0.975	0.838	0.934
	Shen [10]	0.947	0.784	0.945	0.953	0.849	0.936
	Liu [11]	0.938	0.810	0.966	0.956	0.855	0.958
	Zhang [46]	0.926	0.740	0.944	0.930	0.883	0.947
	Oh [45]	0.913	0.767	0.910	0.950	**0.954**	0.943
	Fang [21]	0.975	0.753	**0.973**	0.953	0.868	0.957
	Yang [43]	0.942	0.824	0.954	0.963	0.789	0.956
	Proposed	**0.984**	**0.906**	0.947	**0.988**	0.939	**0.972**

Table 6
SROCC comparison on the LIVE Phase I database.

	SIQA	JP2K	JPEG	WN	BLUR	FF	ALL
FR	Md [5]	0.907	0.676	0.929	0.843	0.735	0.922
	Wang [6]	0.881	0.513	0.944	0.931	0.686	0.916
NR	Zhou [9]	0.856	0.670	0.918	0.929	0.768	0.904
	Shen [10]	0.902	0.757	0.919	0.909	0.796	0.932
	Liu [11]	0.888	0.785	**0.951**	0.917	0.821	0.949
	Zhang [46]	0.931	0.693	0.946	0.909	0.834	0.943
	Oh [45]	0.885	0.765	0.921	0.930	**0.944**	0.935
	Fang [21]	-	-	-	-	-	0.946
	Yang [43]	0.897	0.768	0.929	0.917	0.685	0.944
	Proposed	**0.965**	**0.879**	0.921	**0.945**	0.900	**0.962**

Table 7
PLCC comparison on the LIVE Phase II database.

	SIQA	JP2K	JPEG	WN	BLUR	FF	ALL
FR	Md [5]	0.877	0.823	0.925	0.976	0.895	0.902
	Wang [6]	0.957	0.790	0.829	0.784	0.868	0.745
NR	Zhou [9]	0.835	0.860	**0.973**	0.977	0.915	0.905
	Shen [10]	0.887	0.874	0.874	0.981	0.943	0.932
	Liu [11]	0.936	0.867	0.969	0.987	0.959	0.935
	Zhang [46]	-	-	-	-	-	-
	Oh [45]	0.865	0.821	0.836	0.934	0.815	0.863
	Fang [21]	**0.975**	**0.952**	0.972	0.983	0.929	0.946
	Yang [43]	0.886	0.867	0.887	**0.988**	0.916	0.934
	Proposed	0.956	0.825	0.954	**0.988**	**0.964**	**0.953**

Table 8
SROCC comparison on the LIVE Phase II database.

	SIQA	JP2K	JPEG	WN	BLUR	FF	ALL
FR	Md [5]	0.873	0.790	0.912	0.931	0.885	0.892
	Wang [6]	0.953	0.773	0.826	0.770	0.831	0.714
NR	Zhou [9]	0.812	0.773	**0.946**	0.901	0.860	0.890
	Shen [10]	0.848	0.822	0.942	0.913	0.910	0.927
	Liu [11]	0.909	**0.825**	**0.946**	0.936	0.938	0.933
	Zhang [46]	-	-	-	-	-	-
	Oh [45]	0.853	0.822	0.833	0.889	0.878	0.871
	Fang [21]	-	-	-	-	-	0.934
	Yang [43]	0.859	0.806	0.864	0.834	0.877	0.921
	Proposed	**0.954**	0.816	0.923	**0.951**	**0.969**	**0.951**

Table 9
The results of ablation experiments.

Database	LIVE Phase I		LIVE Phase II	
Criteria	PLCC	SROCC	PLCC	SROCC
w/o LFE-MSP	0.947	0.937	0.925	0.925
w/o LFE	0.956	0.947	0.944	0.938
w/o MSP	0.964	0.950	0.944	0.943
w/o ACB	0.962	0.953	0.949	0.943
Proposed	**0.972**	**0.962**	**0.953**	**0.951**

4.5. Visualization of feature maps

In this section, the feature maps are visualized to understand what network can learn. In Fig. 4, each row shows the distorted stereoscopic image (left view), the feature maps from the first ACB, and the feature maps from the last convolution layer in global feature fusion sub-network. Each column shows the results according to JP2K, JPEG, WN, BLUR and FF. These feature maps are from nine channels. It can be found that the feature maps of different

channels extract specific image features. Meanwhile, it can be seen that the output of ACB (middle column) mainly includes the local features, such as texture and other details, while global feature fusion (right column) is mainly the global information, such as contour. These feature maps reflect that our two types of sub-networks can extract effective information from the distorted image. It explains why we can get superior performance in experiments.

In Tables 5 and 6, it can be seen that our algorithm cannot achieve the optimal performance on WN and FF distortion types. The reason is that the WN and FF distortion types destroy the structure of the original image. As shown in Fig. 5, we can see that the feature maps fail to capture the meaningful information, which reduces the prediction accuracy. Nevertheless, our results are still competitive with other algorithms.

4.6. Cross database and time complexity tests

The cross-database experiments are conducted to verify the generalization ability of our proposed network. One of LIVE Phase I and Phase II databases is utilized as the training set and the other as the testing set. The results of cross-database test are shown in Table 10. It is worth noting that only Shen [10], Liu[11], Fang [21] and Yang [43] are given for comparison because the source codes of other algorithms are not public. As can be seen from Table 10, the performance of training in LIVE Phase II is better than in LIVE Phase I. It is because that that LIVE Phase II contains both symmetrical and asymmetrical distorted 3D images. Training on LIVE Phase II can capture the features from both symmetrically and asymmetrically distortions. While LIVE Phase I only contains the symmetrical distortion types, so the testing results on LIVE

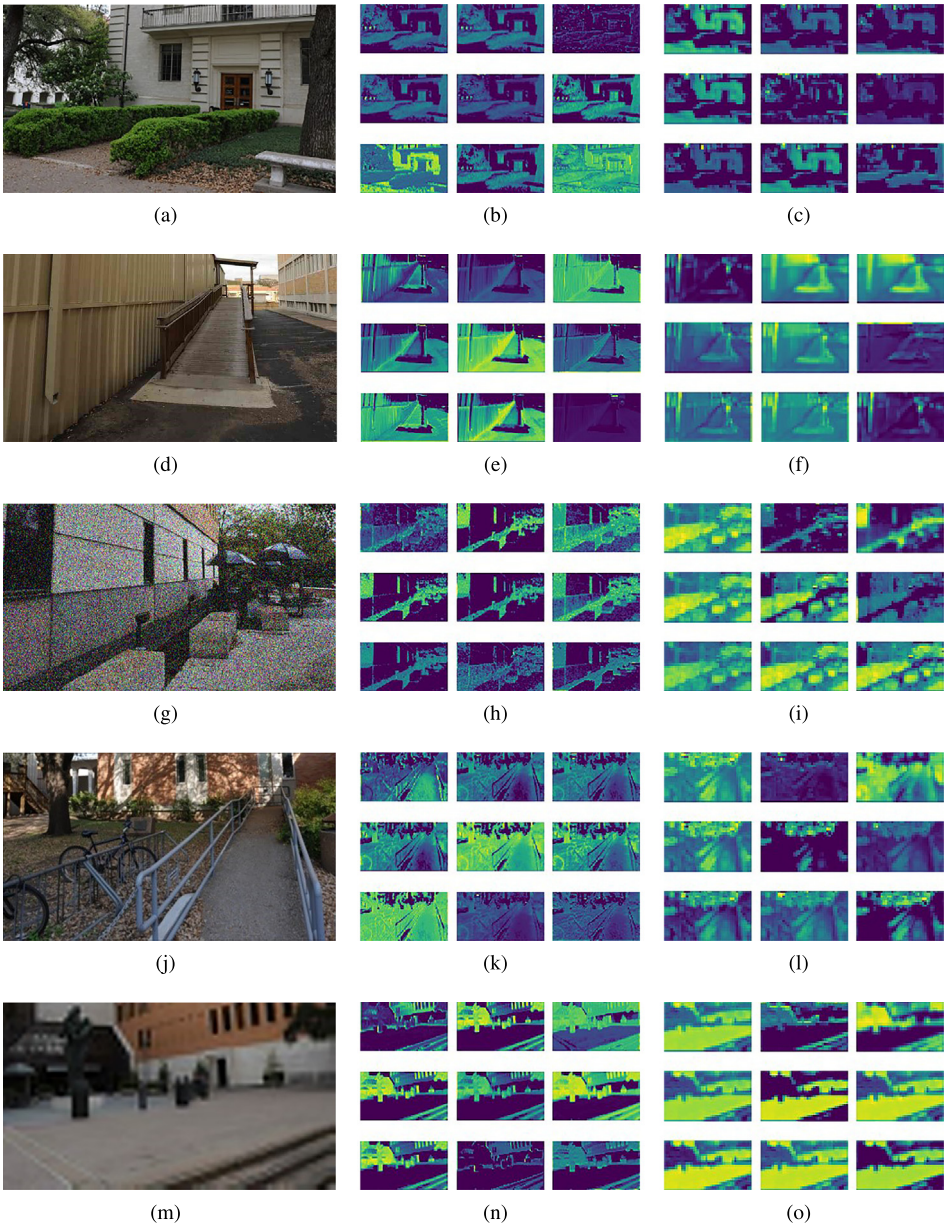


Fig. 4. Examples of feature maps in two types of sub-network: (a), (d), (g), (j), (m) are images distorted by JP2K, JPEG, WN, BLUR and FF. (b), (e), (h), (k), (n) are feature maps in the first ACB. (c), (f), (i), (l), (o) are feature maps from global feature fusion sub-network.

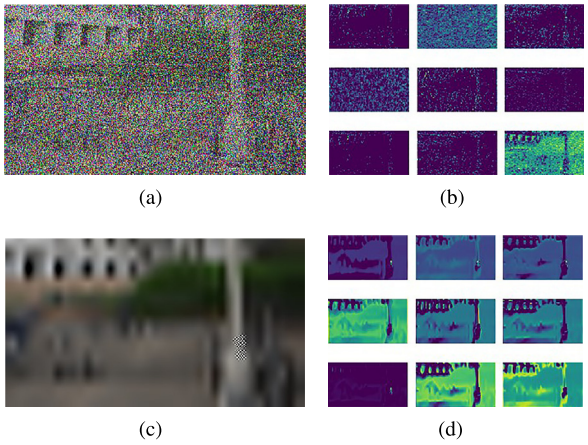


Fig. 5. Examples of feature maps under severe WN and FF distortion. (a) and (c) are left-view images distorted by WN and FF. (b) and (d) are feature maps extracted from first ACB under WN and FF.

Table 10
The results of cross-database test on LIVE databases.

Train/Test	Phase I/ Phase II		Phase II/ Phase I	
Criteria	PLCC	SROCC	PLCC	SROCC
Shen [10]	0.812	0.806	0.822	0.811
Liu [11]	**0.862**	0.832	0.888	0.874
Fang [21]	0.811	0.797	0.899	0.898
Yang [43]	0.852	**0.849**	0.869	0.860
Proposed	0.848	0.833	**0.915**	**0.921**

Table 11
The results of time complexity test on LIVE databases.

Database	LIVE Phase I	LIVE Phase II
Testing time (s)	7.961	8.822

Phase II are slightly inferior. Overall, our proposed algorithm performs well, and can be extended to different databases.

In addition, the testing time is used to evaluate the computational complexity of our proposed network. From Table 11, we can find it has very low computational complexity. A possible explanation is that there are only selected 66 patches for a stereoscopic image pair. Therefore, our proposed method has both low computational complexity and high accuracy.

5. Conclusion

In this paper, we present a novel DCNN for NR-SIQA that combines the global and local content characteristics. First, a sampling indicator with variance-based ratio is utilized to screen homogenous patches. Considering there is a complex inner fusion process in the brain, a cross-fusion strategy is proposed to concatenate the left- and right-view features in the global feature fusion subnetwork. Then, the MSP is applied to provide more context information for learning effective global features. While for the single view, the ACB is introduced to strengthen the local feature extraction capability. Finally, the weighted average is used to estimate the whole stereoscopic image quality. Experimental results on multiple 3D databases demonstrate that our proposed model is consistent with human visual perception.

CRediT authorship contribution statement

Lili Shen: Conceptualization, Methodology, Supervision, Writing - review & editing, Funding acquisition. **Xiongfei Chen:** Formal analysis, Software, Writing - original draft. **Zhaoqing Pan:** Writing - review & editing, Funding acquisition. **Kefeng Fan:** Supervision, Writing - review & editing, Funding acquisition. **Fei Li:** Validation, Writing - review & editing, Investigation.

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.

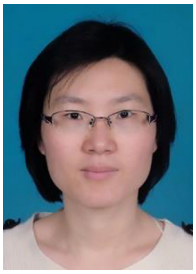
Acknowledgements

This work is supported by National Key Research and Development Program of China (No. 2019YFB1405503) and National Natural Science Foundation of China (No. 615-20106002 and No. 61971232) and 2019 Public Service Platform of Industrial Technology Foundation (No. 2019-00895-2-1).

References

[1] Q. Yan, D. Gong, Y. Zhang, Two-Stream Convolutional Networks for Blind Image Quality Assessment, IEEE Trans. Image Process. 28 (5) (2019) 2200–2211.
[2] X. Yang, F. Li, H. Liu, Deep feature importance awareness based no-reference image quality prediction, Neurocomputing 401 (2020) 209–223.
[3] M. Zhou, X. Wei, S. Wang, S. Kwong, C.-K. Fong, P.H.W. Wong, W.Y.F. Yuen, W. Gao, SSIM-Based Global Optimization for CTU-Level Rate Control in HEVC, IEEE Trans. Multimedia 21 (8) (2019) 1921–1933.
[4] J. Gu, G. Meng, J.A. Redi, S. Xiang, C. Pan, Blind Image Quality Assessment via Vector Regression and Object Oriented Pooling, IEEE Trans. Multimedia 20 (5) (2018) 1140–1153.
[5] S.K. Md, B. Appina, S.S. Channappayya, Full-Reference Stereo Image Quality Assessment Using Natural Stereo Scene Statistics, IEEE Signal Process. Lett. 22 (11) (2015) 1985–1989.
[6] X. Wang, L. Ma, S. Kwong, Y. Zhou, Quaternion representation based visual saliency for stereoscopic image quality assessment, Signal Processing 145 (2018) 202–213.
[7] X. Wang, Q. Liu, R. Wang, Z. Chen, Natural image statistics based 3d reduced reference image quality assessment in contourlet domain, Neurocomputing 151 (2) (2015) 683–691.
[8] L. Ma, X. Wang, Q. Liu, K.N. Ngan, Reorganized dct-based image representation for reduced reference stereoscopic image quality assessment, Neurocomputing 215 (2016) 21–31.
[9] W. Zhou, S. Zhang, T. Pan, L. Yu, W. Qiu, Y. Zhou, T. Luo, Blind 3D image quality assessment based on self-similarity of binocular features, Neurocomputing 224 (2017) 128–134.
[10] L. Shen, R. Fang, Y. Yao, X. Geng, D. Wu, No-Reference Stereoscopic Image Quality Assessment Based on Image Distortion and Stereo Perceptual Information, IEEE Trans. Emerging Topics Comput. Intell. 3 (1) (2019) 59–72.
[11] Y. Liu, C. Tang, Z. Zheng, L. Lin, No-reference stereoscopic image quality evaluator with segmented monocular features and perceptual binocular features, Neurocomputing 405 (2020) 126–137.
[12] W. Zhou, Z. Chen, W. Li, Dual-Stream Interactive Networks for No-Reference Stereoscopic Image Quality Assessment, IEEE Trans. Image Process. 28 (8) (2019) 3946–3958.
[13] S. Khan, S.S. Channappayya, Estimating Depth-Salient Edges and Its Application to Stereoscopic Image Quality Assessment, IEEE Trans. Image Process. 27 (12) (2018) 5892–5903.
[14] J. Yang, H. Xu, Y. Zhao, H. Liu, W. Lu, Stereoscopic image quality assessment combining statistical features and binocular theory, Pattern Recogn. Lett. 127 (2019) 48–55.
[15] F. Zhu, Z. Ma, X. Li, G. Chen, J.-T. Chien, J.-H. Xue, J. Guo, Image-text dual neural network with decision strategy for small-sample image classification, Neurocomputing 328 (2019) 182–188.
[16] Z. Ma, D. Chang, J. Xie, Y. Ding, S. Wen, X. Li, Z. Si, J. Guo, Fine-Grained Vehicle Classification With Channel Max Pooling Modified CNNs, IEEE Trans. Veh. Technol. 68 (4) (2019) 3224–3233.
[17] G. Lin, A. Milan, C. Shen, I. Reid, RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5168–5177..
[18] J. Lei, Y. Song, B. Peng, Z. Ma, L. Shao, Y.-Z. Song, Semi-Heterogeneous Three-Way Joint Embedding Network for Sketch-Based Image Retrieval, IEEE Trans. Circuits Syst. Video Technol. 30 (9) (2020) 3226–3237.
[19] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional Neural Networks for No-Reference Image Quality Assessment, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1733–1740..
[20] S. Bosse, D. Maniry, K.-R. Mueller, T. Wiegand, W. Samek, Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment, IEEE Trans. Image Process. 27 (1) (2018) 206–219.

[21] Y. Fang, J. Yan, X. Liu, J. Wang, Stereoscopic image quality assessment by deep convolutional neural network, J. Vis. Commun. Image Represent. 58 (2019) 400–406.
[22] S. Li, J. Xue, Y. Han, No-reference stereoscopic image quality assessment based on local to global feature regression, in: IEEE International Conference on Multimedia and Expo, pp. 448–453..
[23] M. Zhou, X. Wei, S. Kwong, W. Jia, F. Bin, Just Noticeable Distortion-Based Perceptual Rate Controlin HEVC, IEEE Trans. Image Process. 99 (2020) 1.
[24] J. Yang, K. Sim, X. Gao, W. Lu, Q. Meng, B. Li, A Blind Stereoscopic Image Quality Evaluator With Segmented Stacked Autoencoders Considering the Whole Visual Perception Route, IEEE Trans. Image Process. 28 (3) (2019) 1314–1328.
[25] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6230–6239..
[26] X. Ding, Y. Guo, G. Ding, J. Han, ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks, in: IEEE International Conference on Computer Vision, pp. 1911–1920..
[27] M. Zhou, X. Wei, S. Kwong, W. Jia, B. Fang, Rate Control Method Based on Deep Reinforcement Learning for Dynamic Video Sequences in HEVC, IEEE Trans. Multimedia 99 (2020) 1.
[28] Z. Ma, J. Xie, H. Li, Q. Sun, F. Wallin, Z. Si, J. Guo, Deep Neural Network-Based Impacts Analysis of Multimodal Factors on Heat Demand Prediction, IEEE Trans. Big Data 6 (3) (2020) 594–605.
[29] 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 (8) (2012) 3339–3352.
[30] A. Mittal, A.K. Moorthy, A.C. Bovik, No-Reference Image Quality Assessment in the Spatial Domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.
[31] W. Xue, X. Mou, L. Zhang, A.C. Bovik, X. Feng, Blind Image Quality Assessment Using Joint Statistics of Gradient Magnitude and Laplacian Features, IEEE Trans. Image Process. 23 (11) (2014) 4850–4862.
[32] L. Liu, Y. Hua, Q. Zhao, H. Huang, A.C. Bovik, Blind image quality assessment by relative gradient statistics and adaboosting neural network, Signal Processing-Image Commun. 40 (2016) 1–15.
[33] M.-J. Chen, C.-C. Su, D.-K. Kwon, L.K. Cormack, A.C. Bovik, Full-reference quality assessment of stereopairs accounting for rivalry, Signal Processing-Image Commun. 28 (9) (2013) 1143–1155.
[34] W. Zhou, L. Yu, Binocular Responses for No-Reference 3D Image Quality Assessment, IEEE Trans. Multimedia 18 (6) (2016) 1077–1084.
[35] F. Gao, J. Yu, S. Zhu, Q. Huang, Q. Han, Blind image quality prediction by exploiting multi-level deep representations, Pattern Recogn. 81 (2018) 432–442.
[36] J. Gu, G. Meng, S. Xiang, C. Pan, Blind image quality assessment via learnable attention-based pooling, Pattern Recogn. 91 (2019) 332–344.
[37] L.-M. Po, M. Liu, W.Y.F. Yuen, Y. Li, X. Xu, C. Zhou, P.H.W. Wong, K.W. Lau, H.-T. Luk, A Novel Patch Variance Biased Convolutional Neural Network for No-Reference Image Quality Assessment, IEEE Trans. Circuits Syst. Video Technol. 29 (4) (2019) 1223–1229.
[38] W. Ji, J. Wu, G. Shi, W. Wan, X. Xie, Blind image quality assessment with semantic information, J. Vis. Commun. Image Represent. 58 (2019) 195–204.
[39] L. Shen, N. Hang, C. Hou, Feature-segmentation strategy based convolutional neural network for no-reference image quality assessment, Multimedia Tools Appl. 79 (17–18) (2020) 11891–11904.
[40] H. Ren, D. Chen, Y. Wang, RAN4IQA: Restorative Adversarial Nets for No-Reference Image Quality Assessment, in: AAAI Conference on Artificial Intelligence, pp. 7308–7314..
[41] K.-Y. Lin, G. Wang, Hallucinated-IQA: No-Reference Image Quality Assessment via Adversarial Learning, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 732–741..
[42] M. Karimi, N. Soltanian, S. Samavi, K. Najarian, N. Karimi, S.M.R. Soroushmehr, Blind stereo image quality assessment inspired by brain sensory-motor fusion, Digital Signal Process. 91 (2019) 91–104.
[43] J. Yang, Y. Zhao, Y. Zhu, H. Xu, W. Lu, Q. Meng, Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network, Inf. Sci. 474 (2019) 1–17.
[44] O. Messai, F. Hachouf, Z. A. Seghir, Deep learning and cyclopean view for no-reference stereoscopic image quality assessment, in: International Conference on Signal, Image, Vision and their Applications..
[45] H. Oh, S. Ahn, J. Kim, S. Lee, Blind Deep S3D Image Quality Evaluation via Local to Global Feature Aggregation, IEEE Trans. Image Process. 26 (10) (2017) 4923–4936.
[46] W. Zhang, C. Qu, L. Ma, J. Guan, R. Huang, Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network, Pattern Recogn. 59 (2016) 176–187.
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826..
[48] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations..
[49] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines vinod nair 27 (2010) 807–814..
[50] A.K. Moorthy, C.-C. Su, A. Mittal, A.C. Bovik, Subjective evaluation of stereoscopic image quality, Signal Processing: Image Commun. 28 (8) (2013) 870–883.
[51] M.-J. Chen, L.K. Cormack, A.C. Bovik, No-Reference Quality Assessment of Natural Stereopairs, IEEE Trans. Image Process. 22 (9) (2013) 3379–3391.



Lili Shen received the Ph.D. degree in Communication and information system from Tianjin University, Tianjin, China, in 2010. She was a visiting scholar at the Centre for Vision Research, York University, Canada, in 2014. She is currently an associate professor of the School of Electrical and Information Engineering, Tianjin University, China. Her research interests include 2D/3D image processing, computational vision and multimedia communications.

Xiongfei Chen received his B.S. degree from the School of Electronics and Information Engineering, Hebei University of Technology, Tianjin, China, in 2019. He is currently pursuing the M.S degree in the school of Electrical Automation and Information Engineering, Tianjin University, Tianjin, China. His research interests include stereoscopic image quality assessment and deep learning.

Zhaoqing Pan received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, China, in 2014. In 2013, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, for six months. He is currently a Full Professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China, and also with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include video coding, image quality assessment, and machine learning.

Kefeng Fan received Ph.D degree in test signal processing from Xidian University, Xi'an, China. From Feb., 2008 to Nov., 2010, he was doing the postdoctoral in State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China. From May, 2013 to Oct.,2018, he worked as the deputy director of the Research Center of Cyberspace Security, China Electronics Standardization Institute(CESI). From Dec.,2018 to now, He worked as the director of the Research Center of Digital Technology, CESI. His current research interests include data coding for machine, data security, and UHD signal processing. He is member of ISO/IEC JTC1/SC29&SC24. He was appointed as the Young Professional of IEC in 2011.

Fei Li received the M.S degree in management engineering from the Lingnan College of Sun Yat-sen University, Guangzhou, China, in 2009. In 2013, she was a Visiting Scholar with the Department of Electrical Information, MIIT, Beijing, China, for 24 months. She is currently a Senior Engineer with the Guangzhou SequoiaDB Company, Guangzhou, China. Her research interests include video coding, database, and artificial intelligence.

Jianjun Lei received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2007. He was a visiting researcher at the Department of Electrical Engineering, University of Washington, Seattle, WA, from August 2012 to August 2013. He is currently a Professor at Tianjin University, Tianjin, China. He is on the editorial boards of Neurocomputing and China Communications. His research interests include 3D video processing, virtual reality, and artificial intelligence.