Multimedia Tools and Applications (2022) 81:18437–18455
https://doi.org/10.1007/s11042-022-12707-4

Blind stereoscopic image quality assessment using 3D saliency selected binocular perception and 3D convolutional neural network

Chaofeng Li[1] · LiXia Yun[1] · Shoukun Xu[2]

Received: 27 November 2020 / Revised: 22 March 2021 / Accepted: 21 February 2022 /
Published online: 9 March 2022

Abstract
The purpose of stereoscopic image quality assessment (SIQA) is to design an objective evaluation algorithm to automatically evaluate the quality of stereoscopic image. In this paper, we propose a blind SIQA method via 3D saliency selected binocular perception and 3D convolutional neural network (CNN). Given a pair of stereoscopic images, we first generate 3D saliency map by weighted average of 2D saliency map and depth saliency map. Then, when the value of 3D saliency map patches is higher than the setting threshold, these patches from left and right images are selected to feed to 3D-CNN to predict the perceived quality. Finally, the score of the distorted stereoscopic image is computed by the weighted average of the quality scores of these saliency image patches. Experimental results on LIVE 3D Phase I and Phase II databases show that our proposed method is robust and competitive with the state-of-the-art NR SIQA methods.

Keywords Blind stereoscopic image quality assessment · Convolutional neural network · 3D saliency map · Summation and difference image · Cyclopean image

1 Introduction

In recent decades, with the rapid development of multimedia and network technology, the amount of digital image is explosive growth, which plays an increasingly important role in people's daily life. However, the current technology has many limitations, which lead to various distortions in the process of the collection, transmission, processing and display of images. Image quality assessment (IQA) is developed to evaluate and monitor image quality,

✉ Chaofeng Li
wxlichaofeng@126.com

[1] Institute of Logistics Science and Engineering, Shanghai Maritime University, Shanghai 201306, China

[2] School of Information Science and Engineering, Changzhou University, Changzhou 213164 Jiangsu, China

which shows great potential in controlling and improving the performance of image processing systems, such as compression [41], enhancement [15] and segmentation [16]. Recently, stereoscopic/3D multimedia applications have become increasingly popular, and greatly increasing the immersive experience, so stereoscopic image quality assessment (SIQA) has become one of the research hotspots.

Generally, quality assessment metrics can be divided into subjective method and objective method. Subjective method is effective and reliable, but it is more time-consuming, laborious and cannot be completed in real time. In contrast, objective method is more in line with the needs of reality, so it has been widely investigated. Broadly speaking, according to the participation of reference information, objective methods can be further divided into full reference (FR), reduced reference (RR) and no reference (NR). Now, there are many 2D IQA methods [35, 37, 47] that have achieved quite competitive results. In contrast, stereoscopic image quality assessment is more challenging, which needs to consider a variety of factors, such as binocular fusion, binocular competition and so on. The distortion of stereoscopic image pairs can be divided into symmetric distortion and asymmetric distortion. The asymmetric distortion may have different degrees or even different types, which also has a great impact on the quality of stereoscopic image. In addition, different from 2D IQA, SIQA has extended depth perception and binocular vision mechanism between left and right visual fields, which leads to the difficulty of current research.

Stereo visualization involves more and more application fields, such as distance education, medical treatment, robot navigation, and so on. Therefore, it is reasonable to believe that the number of 3D content will continue to grow in the next few years. SIQA is a key technology in stereo image and video processing, which can help image retrieval system to filter low-quality images by monitoring the quality of stereo image, so as to produce better subjective experience. In addition, SIQA will also promote research in other fields, such as stereoscopic video quality evaluation [7, 13, 40], stereo matching [9].

Yang et al. [42] uses 3D-CNN model to capture spatial-temporal features, and evaluates the quality of stereoscopic video. Inspired, we try to solve difficulty of SIQA with 3D-CNN, which has been applied in many research topics. For example, Zhang et al. [51] proposed a 3D-CNN structure for mental workload assessment, learning the EEG features from spatial–spectral–temporal dimensions. Considering the continuity of video in time, Yang et al. [44] uses the correlation between HIS spatial-spectral domain to design a multiscale wavelet 3D-CNN method for hyperspectral image super-resolution. As the work of these different directions proves, 3D-CNN can find the connection of different features, and it is a very effective solution to the problem. In this work, we extend 2D-CNN to 3D-CNN for SIQA, and omit the design of binocular fusion through CNN. In addition, we consider the "binocular summation/difference theory" [26], which is to convert the information obtained by the left and right eyes into uncorrelated sum and difference signals, and then transmit them forward, so that 3D-CNN can obtain multi-dimensional information.

In addition to the fusion mechanism of the brain, visual saliency is also important for image processing. The research of visual psychophysics has found that when the human eye looks at an image, it will unconsciously focus on certain areas, and give priority to the information of these areas [36], called them as salient areas. In reference [11], a Saliency-based DCNN (SDCNN) framework for NR-IQA is proposed. Inspired by this, we use saliency mapping to modify monocular image to highlight regions of interest and weaken insensitive parts, and propose a blind SIQA method via 3D saliency selected binocular perception and 3D-CNN.

Here we use 3D-CNN model that automatically simulate human vision, and build the relationship between subjective perception and predicted scores of stereoscopic image quality. Our main contributions are as follows:

(1) We propose a 3D-CNN based NR-SIQA method. To the best of our knowledge, we are the pioneers in using 3D-CNN to evaluate the quality of stereoscopic images.
(2) We propose a method to select salient image patches from 3D saliency map. By weighting 2D saliency map and depth saliency map, 3D saliency map of depth perception is obtained. Only when the value of 3D saliency patch is greater than the set threshold, the corresponding image patch can be selected from the left and right image for predicting image quality.
(3) For 3D-CNN model, we add simple summation and difference images to supplement the left and right images as input, providing more different and effective information for predicting image quality.

The rest of this paper is organized as follows. Section 2 describes related work on stereoscopic image quality assessment. In Section 3, proposed model is introduced in detail. The experimental results and analysis on multiple databases are presented in Section 4. Finally, we conclude this paper in Section 5.

2 Related works

Generally, according to whether to use reference image information, objective methods can be categorized into full reference (FR), reduced reference (RR) and no reference (NR, also called as blind reference) [22]. In the process of SIQA, if the original reference image is used, the quality of the distorted stereoscopic image pairs can be obtained more accurately by comparing the local similarity between the two groups of images. This method is called full-reference (FR) SIQA [4, 5, 12, 39]. Shao et al. [31] simulated simple and complex visual cell to obtain a feature encoding approach, and define a similarity measure approach between original image and distorted image. Li et al. [20] proposed an FR SIQA based on ensemble learning and an adaptive cyclopean image, which was modified by a salient map.

In contrast, the no-reference (NR) SIQA method does not need reference image, which is more in line with the actual need, more promising in practical applications, but more challenging [3, 32, 50]. Akhter et al. [2] proposed a NR SIQA method, which combined the extraction of manually designed segmented local features and estimated parallax information from stereoscopic image pairs. Subsequently, by exploring the interaction of two views in HVS, many researches began to focus on the binocular behavior of simple and complex cells in human brain to generate “cyclopean” images from two views [5]. The algorithm proposed by Chen et al. [6] used the 2D features of the synthesized cyclopean image and the 3D features of the corresponding depth map to predict the perceived quality of the stereoscopic image pairs. Zhou et al. [52] describe a blind SIQA based on binocular combination and an extreme learning machine (ELM). Yang et al. [45] used depth perception map to quantify the depth features of stereoscopic images, and also considered binocular features. In addition, deep belief network is used to evaluate content quality. Li et al. [21] proposed an NR-SIQA method based on visual attention and perception. The model combined saliency and just noticeable difference (JND), and weight the global and local features extracted from the left and right views.

Finally, a support vector regression (SVR) is learned to evaluate the quality of stereoscopic images. Liu et al. [23] extracted the monocular color and luminance features and binocular summation / difference features and proposed a blind SIQA model by SVR.

In recent years, deep learning technology has been widely used in solving various image processing and computer vision problems [29], and has achieved great success. Convolutional neural network (CNN) has shown outstanding performance in many applications of computer vision and image processing. Compared with the traditional image processing methods, CNN can automatically learn the deep visual features that closely related to the target by optimizing the network parameters rather than using hand-made features. The main advantage of CNN is that it can directly input the images, and then combine feature learning with quality regression in the training process. When CNN is directly used for NR SIQA, there will be a key obstacle: the training data is not enough because of the limited number of subjective perception images. The existing data enhancement and image preprocessing technology is also not suitable for the NR SIQA [17]. In order to solve these problems, most scholars adopt the strategy of image segmentation: cut the image into patches of the same size, input patches into CNN model to predict the quality score of each image patch, and then get the image perception quality according to the established rules. Kang et al. [14] proposed a CNN model for 2D images, input image patches, learn the quality characteristics of image patches and obtain their visual quality. Finally, the quality of all image patches is weighted and averaged to calculate the objective score of the image. Li et al. [19] transferred the structure and weights of a model pre-trained on ImageNet. Then, they modified the last several layers to directly output the quality score, and fine-tuned the network to regress the objective image quality. Lv et al. [24] established a depth neural network model to predict monocular distortion of stereoscopic images, and evaluated binocular features considering binocular competition. Finally, the two features were fused to comprehensively evaluate stereoscopic images from various aspects. Sun et al. [34] use CNNs to learn deeper local quality-aware structures, and remove related features on non-salient patches. Then, the reserved features are aggregated into a final quality score in an end-to-end manner.

3 Proposed method

Our proposed blind SIQA framework is shown in Fig. 1. Given left and right images, we first compute the 3D saliency map by combining 2D saliency map with depth saliency map. Then saliency left and right image patches are selected by using the saliency of 3D saliency map patches, and saliency summation and difference images are generated, which are fed to 3D-CNN to predict the perceived quality after performing local normalization. Finally, the score of the distorted stereoscopic image is obtained by the weighted average of the quality scores of the saliency image patches. Next, our proposed blind SIQA is divided into six parts (A, B, C, D, E and F) for detailed introduction.

3.1 3D saliency map

Zhang et al. [49] found that the visual saliency of an image varies with the change of image quality, so we design 3D saliency map and select the image patches with high saliency to train the network, so as to reduce the adverse effect of low saliency image patches on the result. In [38], they compared the results of several methods, including no-depth, depth-weighting (DW)

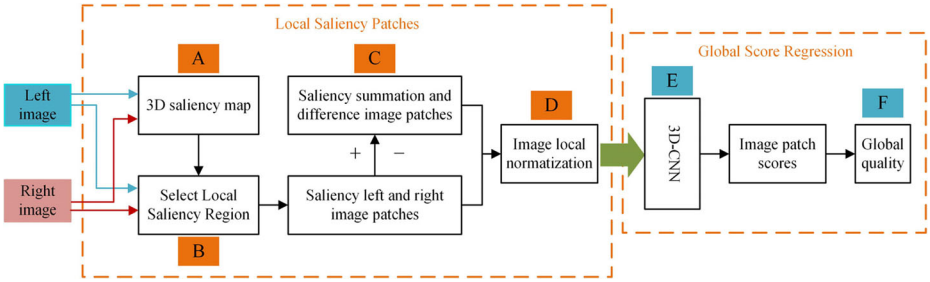

Fig. 1 Structure of our proposed blind SIQA

and depth-saliency (DS) method, which show that the depth-saliency is the best, and indicates the importance of depth saliency map in the modeling of 3D visual attention. Reasonable fusion of depth saliency map and 2D visual feature detection results can better predict saliency region. We use the depth-saliency model from [38]. As shown in Fig. 2, firstly cyclopean image and depth saliency map are calculated from left and right images, then 2D saliency map is obtained from cyclopean image, and 3D saliency map is computed by combining 2D saliency map with depth saliency map.

Figure 3 shows the left image of a stereoscopic image and its cyclopean image and saliency map. Compared with cyclopean image and 2D saliency map, 3D saliency map emphasizes the contour and edge information of the object, and the region closer to the observer, which is in line with our daily experience. Therefore, 3D saliency map not only reflects the saliency of 2D image, but also emphasizes the depth information of stereoscopic image. Next, the details of cyclopean image, 2D saliency map, depth saliency map and saliency maps combination are

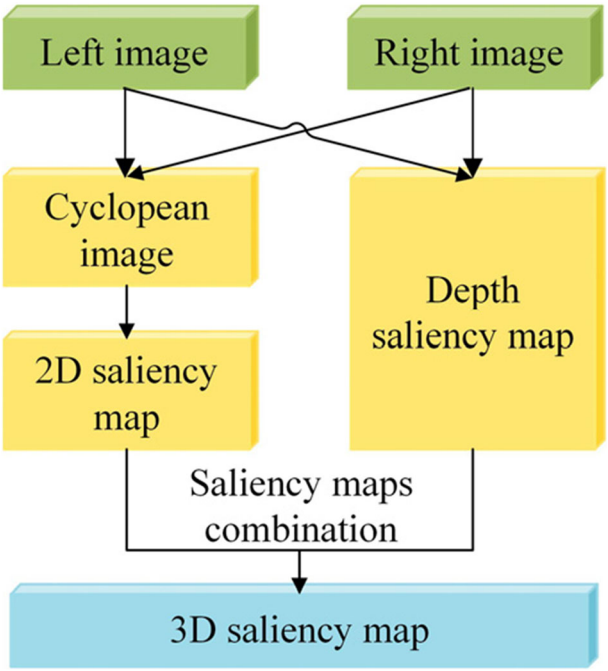

Fig. 2 Calculation process of 3D saliency map

(a) Left image (b) Cyclopean image (c) 2D saliency map (d) Depth saliency map (e) 3D saliency map

Fig. 3 Stereoscopic image and its cyclopean image and saliency map. **a** Left image. **b** Cyclopean image. **c** 2D saliency map. **d** Depth saliency map. **e** 3D saliency map

introduced.

1) Cyclopean image

We adopt the method of calculating cyclopean image in reference [5].

$$CI(x,y) = W_L(x,y) \times I_L(x,y) + W_R((x+d),y) \times I_R((x+d),y) \tag{1}$$

where CI denotes the cyclopean image, I_L and I_R are the left and right views, d is the parallax obtained from the left image, and W_L and W_R are calculated from the normalized Gabor filter amplitude response.

An example of cyclopean image is shown as Fig. 4. Figure 4a is a Left and right view with white noise (WN) asymmetry distortion, and Fig. 4b is Left and right view with Gaussian blur (BLUR) asymmetric distortion. Figure 4c and d are cyclopean images separately generated from Figure 4a and b. Obviously noise can be found from Fig. 4c, and the image quality is significantly reduced due to noise. But in Fig. 4d blur is hardly seen, which indicates that binocular suppression exists. These are consistent with those in reference [1] (binocular suppression is found in BLUR and JP2JK images, but not in WN and JPEG). Cyclopean image can well reflect the phenomenon of binocular suppression, but it is difficult to perceive the depth information of stereoscopic image.

2) 2D saliency map

We use the SDSP method [48] to calculate the 2D saliency map, which combines the prior knowledge that people are more interested in the object with warm color and middle. The saliency map calculated by SDSP focuses on the object, from which the shape and boundary of the object can be clearly seen.

When viewing stereoscopic images, due to the addition of new depth information, depth features and their combination or conflict with other single eye cues, it is unreasonable and ineffective to directly use the 2D visual saliency model for 3D saliency calculation.

3) Depth saliency map

We consider not only the 2D saliency, but also the depth saliency of stereoscopic images. The purpose of the optical flow algorithm [33] is to calculate the velocity vector of each pixel. If we regard the left view as the first frame and the right view as the second frame, the objects close to the human eye will have large parallax, which will show high speed in the streamer field. Therefore, we use optical flow algorithm to calculate the parallax map of stereoscopic image, and the formula is as follows:

(a) Left and right view with white noise asymmetry distortion

(b) Left and right view with Gaussian blur asymmetric distortion.

(c) Cyclopean image via (a) (d) Cyclopean image via (b)

Fig. 4 An example of cyclopean image generated. **a** Left and right view with white noise asymmetry distortion. **b** Left and right view with Gaussian blur asymmetric distortion. **c** Cyclopean image view (**a**). **d** Cyclopean image view (**b**)

$$E(u,v)=\sum_{i,j}\begin{Bmatrix}\rho_D\left(I_L(i,j)-I_R\left(i+u_{i,j},j+v_{i,j}\right)\right)\\+\lambda\left[\rho_s\left(u_{i,j}-u_{i+1,j}\right)+\rho_s\left(u_{i.j}-u_{i,j+1}\right)+\rho_s\left(v_{i,j}-v_{i+1,j}\right)+\rho_s\left(v_{i,j}-v_{i,j+1}\right)\right]\end{Bmatrix} \quad (2)$$

In the optical flow field, u is the horizontal component, v is the vertical component, λ is the regularization parameter, ρ_D is the data penalty function and ρ_S is the spatial penalty function.

For the visual system, the horizontal difference is much larger than the vertical difference, and the depth perception is more effective. Therefore, only the horizontal component of the calculated motion vector is selected as the horizontal difference. Disparity map D was formed by horizontal difference. The calculated formula of depth saliency map is defined as

$$S_D(x,y)=1-\frac{D(x,y)-D_{\min}}{D_{\max}-D_{\min}} \quad (3)$$

where $D_{\min}$ and $D_{\max}$ are the minimum and maximum values in the parallax map D, respectively.

4) Saliency maps combination

The purpose of saliency maps combination is to fuse the features of different dimensions, including saliency map of 2D visual attention feature and depth saliency maps, so as to obtain 3D saliency information of stereoscopic image. At present, although many scholars have proposed depth-saliency model, there is still no standard and widely used fusion method. So, considering the different importance of 2D saliency and depth saliency, we use linear pooling strategy to fuse the 2D saliency map and depth saliency map obtained by SDSP algorithm to synthesize the final 3D saliency map. The linear pooling strategy is the same as the approach of [30]:

$$S = \gamma S_{SDSP} + (1-\gamma) S_D \tag{4}$$

where S_{SDSP} denotes the 2D saliency map obtained by SDSP algorithm, and S_D is the depth saliency map, γ is the weighted coefficient. In this study, we use a linear pooling strategy, which weights the 2D saliency map and the depth saliency map averagely. At present, how the two saliency maps interact and ultimately affect 3D saliency are not completely clear, and we consider both content saliency and depth perception have a great impact on stereoscopic image, so we think they are equally important and set γ to 0.5.

3.2 Local saliency region

As shown in Fig. 5, the left and right images and 3D saliency map are divided into image patches, and saliency image patches are chosen via 3D saliency map. If the saliency value of an image patch of the 3D saliency map is bigger than a given threshold, the corresponding image patches of left and right images are selected to feed to 3D-CNN.

3.3 Summation / difference image

The authors of [25] suggest that the vision system has a separate adaptive binocular summation and difference channel to achieve efficient transmission of binocular information. At the physiological level, it is explained that the signals of the summation and difference channels are multiplexed, and each V1 neuron receives the weighted sum of the signals in these two channels [18]. In order to clearly show the effect of summation/difference theory, a sample of

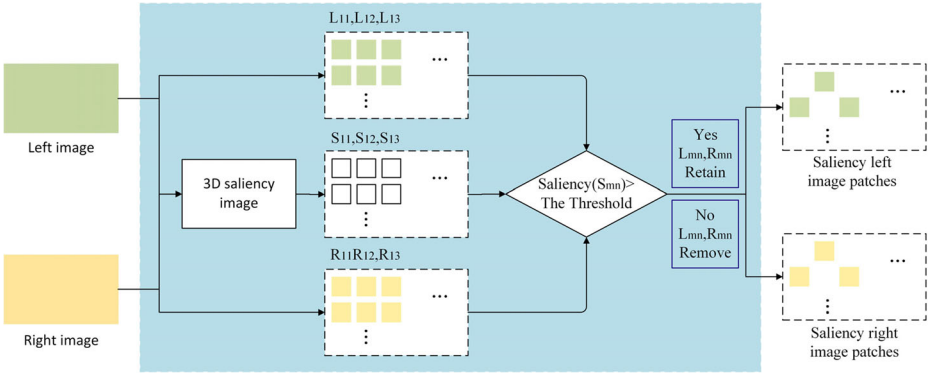

Fig. 5 Selecting procedure of saliency judgment image patches

reference stereoscopic images and corresponding summation and difference images are shown in Fig. 6. The image is like a plane image with ghosting, which is also caused by the parallax of the stereo image. We believe that the human visual system can sense the parallax and convert it into depth information, because the images reflected in the brain are clear and three-dimensional, while the difference images mainly show the depth and contour information of objects. Therefore, we extract the quality features from summation and difference images of stereoscopic images to predict image quality. According to reference [8, 25], the binocular summation/difference image can be simply calculated as follows:

$$\begin{aligned} I_S(i,j) &= I_L(i,j) + I_R(i,j) \\ I_D(i,j) &= I_L(i,j) - I_R(i,j) \end{aligned} \tag{5}$$

where I_S is the summation image, and I_D is the difference image.

3.4 Image local normalization

We divide the M × N stereoscopic image into m × n patches without overlapping, and then reduce the image patch to the range of [0,1] by local normalization [27]. The local normalized image $I'(x, y)$ is calculated as follows:

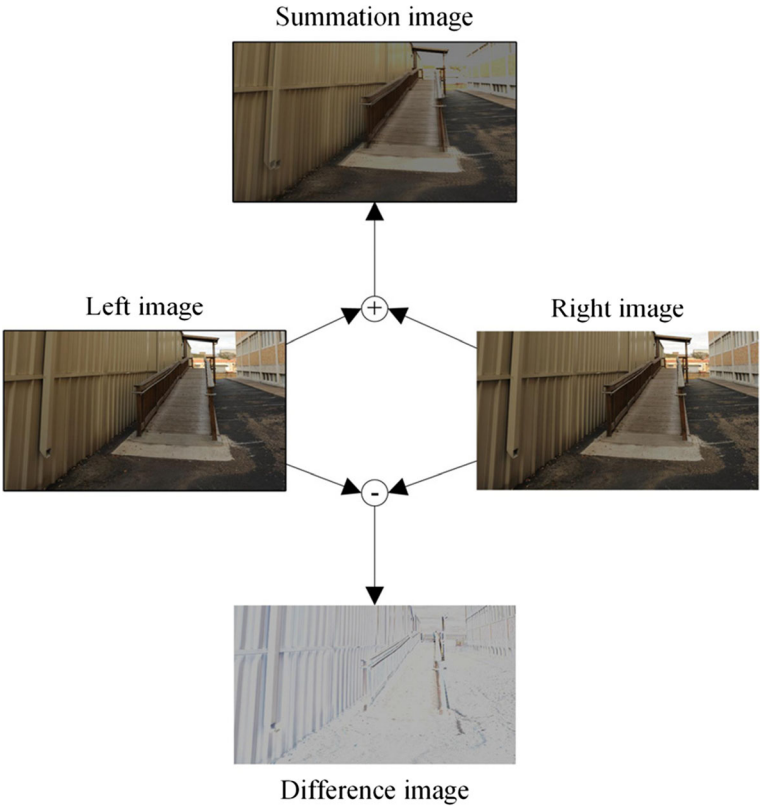

Fig. 6 Reference stereoscopic image and corresponding summation and difference image from LIVE 3D Phase I database

$$I'(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \qquad \sigma(i,j) = \sqrt{\sum_{k,l} \omega_{k,l} \left(I_{k,l}(i,j) - \mu(i,j)\right)^2} \tag{6}$$
$$\mu(i,j) = \sum_{k,l} \omega_{k,l} I_{k,l}(i,j)$$

where I is the image before local normalization, μ and σ are the average and standard deviation of I respectively, and $\omega_{k,\,l}$ is the two-dimensional circularly symmetric Gaussian weighting function.

3.5 3D convolutional neural networks

In general, the typical CNN model structure uses convolution layer and pooling layer alternately to process input information, and then uses full-connected layer to obtain the mapping relationship between features and targets. In 2D-CNN, convolution layer and pooling layer can only extract the features of 2D image, but cannot automatically obtain the interaction information between stereoscopic images. 3D convolution and 3D pooling can extract features between different images, which is exactly what stereoscopic images need. Therefore, we employ 3D-CNN to complete the NR SIQA task.

1) 3D Convolution

Convolution in CNN is a special linear operation between input data and multiple kernel functions, which is used to generate feature maps. For SIQA, 3D convolution adds the features of depth information on the basis of 2D convolution. According to formula (7), the value of (x, y, z) position of the $j - th$ feature map of the $i - th$ layer can be written as [10]:

$$v_{ij}^{xyz} = g\left(b_{ij} + \sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-0} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{7}$$

where $g(\cdot)$ is a non-linear activation function such as hyperbolic tangent function (tanh) or Rectified Linear Unit (ReLU), b_{ij} is the bias for the current feature map, m means the set of feature maps in the $(i - 1) - th$ layer connected to $i - th$ layer, the size of the 3D kernel is $P \times Q \times R$, and w_{ijm}^{pqr} is the value at the position (p, q, r) of the kernel connected to the m feature maps.

2) Structure of 3D-CNN

Based on 3D convolution, we design a 3D-CNN model to automatically learn the quality aware features of stereoscopic image, as shown in Fig. 7. In theory, the more levels of the model structure, the stronger the expression ability, but the more training data required. However, the number of images in stereoscopic image database is limited, which makes the complex model easy to fall into overfitting. Therefore, we designed an effective model according to the size of the database, which consists of five convolution layers, three pooling layers and two fully connected layers, namely Conv1-Conv5, Maxpooling1-Maxpooling3 and FC1-FC2. The input part contains RGB color channels of left and right images, and their summation and difference images, which make the network accept more different features. Therefore, a cube of 4 × m × n (for example, n = 32, M = 32) is taken as the input of 3D-

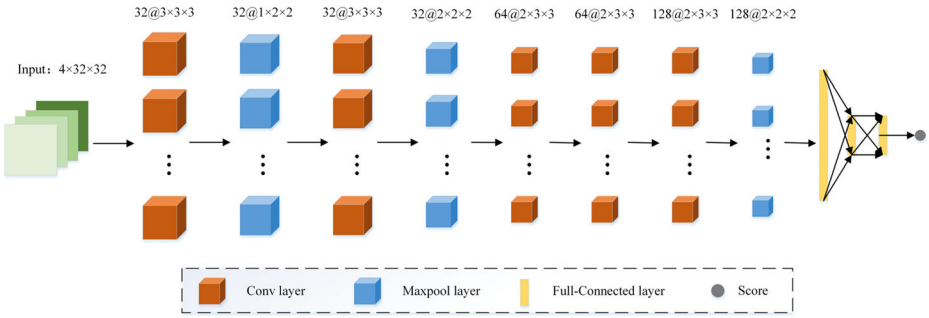

Fig. 7 Structure of 3D-CNN model

CNN model, and it has three feature maps. In particular, Conv1–2 has 32 convolution kernels, Conv3–4 has 64 convolution kernels, and Conv5 has 128 convolution kernels. The size and stride of convolution kernels are shown in Table 1. Filling is used to keep the size of input and output of containment layer consistent. In addition, there are Maxpooling layers after Conv1, Conv2 and Conv5. Finally, the two full connection layers have 512 nodes. The whole parameters of network are shown in Table 1.

We use adaptive moment estimation optimizer (Adam) and back-propagation method to train the network, and the learning rate is set to 0.0001.The minibatch size is set to 32 input data, and the optimal parameters of the model are updated after each iteration. The activation function of all convolution layers and fully connected layers use Rectified Linear Unit (ReLU), which can simplify the back-propagation and enhance the optimization effect by setting a threshold for the input. In order to avoid overfitting, dropout is used in the fully connected layer. The output of neurons is randomly set to 0 with a probability of 0.5. As an effective approximation, dropout can prevent the training network from overfitting in the case of sharing weights.

3.6 Global quality

The input of the network is salient image patches cut from left and right images, and the quality of each patch is predicted by 3D-CNN. The quality of the whole stereoscopic image is calculated by weighted averaging the local quality of each salient image patch as following (8).

Table 1 Configurations of the proposed 3D-CNN structure

Layer	Kernel	Stride	Output size	Feature maps
Input	–	–	4×32×32	3
Conv1	3×3×3	1	4×32×32	32
Maxpooling1	1×2×2	1×2×2	4×16×16	32
Conv2	3×3×3	1	4×16×16	32
Maxpooling2	2×2×2	2	2×8×8	32
Conv3	2×3×3	1	2×8×8	64
Conv4	2×3×3	1	2×8×8	64
Conv5	2×3×3	1	2×8×8	128
Maxpooling3	2×2×2	2	1×4×4	128
FC1	–	–	512	–
FC2	–	–	512	–
Output	–	–	1	–

$$Q = \frac{1}{N_p} \sum_i^{N_p} q_i \tag{8}$$

where $i = 1, 2, \ldots, N_p$, N_p is the number of salient image patches, and q_i is the predicted quality of salient image patches.

4 Experimental results and analysis

In this section, the LIVE 3D databases and evaluation metrics are introduced, and the performance of our proposed blind SIQA has been evaluated comprehensively, including performance comparison on overall database, single distorted types, cross-database validation, saliency threshold analysis, and on symmetrically and asymmetrically distorted images.

4.1 Databases and evaluation metrics

We use two public LIVE 3D Phase I and Phase II IQA database to verify the effectiveness of the algorithm. LIVE 3D Phase I [28] contains 20 reference images, and each reference image has five types of distorted images. There are 80 distorted images in JP2K, JPEG, WN and FF, and 45 distorted images in BLUR, a total of 365. Each pair of stereoscopic images is symmetric distortion, that is, the distortion type and degree of left and right views are the same. In addition, the database also contains the corresponding differential mean opinion score (DMOS) of all stereoscopic images. The lower the DMOS value, the better the image quality. The DMOS value of 20 reference images is 0. LIVE 3D Phase II [5] contains 8 reference images and 360 distorted images. The distortion type is the same as Phase I. Each pair of stereoscopic images has a corresponding DMOS value. The difference is that only 120 of them are symmetrical, and the remaining 240 are asymmetric.

In our experiments, we use the same three performance indicators as in most literatures: SROCC, PLCC and RMSE. When SROCC and PLCC are close to 1 and RMSE is close to 0, the objective evaluation effect is the better. In our experiment, 80% of the images were randomly selected as the training set, and the remaining 20% as the test set. The median of 100 random experiments was the final result.

4.2 Overall performance comparison

To evaluate the effectiveness of the proposed model, we compared the results with the four most advanced FR-SIQA methods (Chen2013 [5], Shao2017 [31], Jiang2018 [12] and Li2019 [20]) and nine NR-SIQA methods (Chen2013 [6], Appina2016 [3], Zhou2017 [52], Yang2018 [43], Yue2018 [46], Yang2019 [45], Li2019 [21], Liu2020 [23] and Sun2020 [34]). The comparison results of SROCC, PLCC and RMSE on LIVE 3D Phase I and II databases are summarized as Table 2, and the best two results are displayed in bold type. It can be seen from Table 2 that the proposed model has competitive advantages in both databases, which proves that the model can effectively predict the quality of stereoscopic images. In particular, SROCC in the Phase II is 0.954, which is 0.008 higher than the best result (0.946 in Li2019 [21]) of the other thirteen methods. Based on the above analysis, our model can simulate the human visual system well for both symmetric and asymmetric stereoscopic images.

Table 2 Comparison of overall performance of different methods

Type	Methods	LIVE 3D Phase I			LIVE 3D Phase II		
		SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
FR	Chen2013 [5]	0.916	0.917	6.550	0.893	0.901	4.987
	Shao2017 [31]	0.931	0.939	5.646	0.928	0.926	4.199
	Jiang2018 [12]	0.938	0.946	5.316	0.926	0.926	4.268
	Li2019 [20]	0.947	0.958	**4.248**	0.940	0.948	**2.986**
NR	Chen2013 [6]	0.891	0.895	7.247	0.880	0.895	5.102
	Appina2016 [3]	0.911	0.917	6.598	0.880	0.845	7.279
	Zhou2017 [52]	0.921	0.941	5.540	0.919	0.923	4.262
	Yang2018 [43]	0.946	0.954	4.874	0.923	0.934	3.999
	Yue2018 [46]	0.914	0.937	5.652	0.906	0.914	4.449
	Yang2019 [45]	0.944	0.956	4.917	0.921	0.934	4.005
	Li2019 [21]	0.953	**0.965**	–	**0.946**	**0.955**	–
	Liu2020 [23]	0.949	0.959	4.574	0.933	0.936	3.804
	Sun2020 [34]	**0.959**	0.951	4.573	0.918	0.938	3.809
	Proposed	**0.962**	**0.966**	**4.470**	**0.954**	**0.957**	**3.520**

Figure 8 shows the change of the loss during the training process on the two databases. It can be seen that the MSE loss value decreases rapidly in the first few iterations and tends to be stable after 30 iterations, which shows that the model can converge quickly, and reduce the time cost in the learning process.

4.3 Performance comparison of single distorted types

In practical application, the distorted types of images in the process of acquisition, transmission and display are different. Therefore SIQA method should not only have a good overall performance, but also achieve satisfactory results in various distortion types. We list the SROCC and PLCC results of five distortion types in LIVE 3D Phase I and II databases as Tables 3 and 4 respectively, and the best two results are displayed in bold.

According to Tables 3 and 4, the proposed model shows the highest predicted accuracy for most of distorted categories, only SROCC and PLCC of Blur do not achieve satisfactory results. In addition, the FR Jiang2018 [12] also shows good performance in WN, Blur and FF

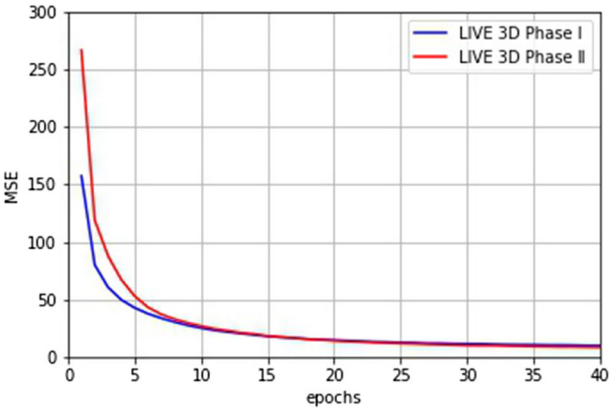

Fig. 8 Optimization process of training loss on LIVE 3D Phase I and II databases

Table 3 Performance comparison of SROCC with five kinds of distortion

Type	Methods	LIVE 3D Phase I					LIVE 3D Phase II				
		JP2K	JPEG	WN	BLUR	FF	JP2K	JPEG	WN	BLUR	FF
FR	Chen2013 [5]	0.896	0.558	0.948	0.926	0.688	0.833	0.840	0.955	0.910	0.889
	Shao2017 [31]	0.900	0.634	0.943	0.925	0.781	0.875	0.834	0.933	0.924	0.941
	Jiang2018 [12]	0.903	0.663	**0.953**	0.936	0.808	0.850	0.855	0.956	0.938	**0.956**
	Li2019 [20]	0.941	0.762	0.936	**0.979**	**0.869**	0.920	0.843	**0.975**	0.936	**0.976**
NR	Chen2013 [6]	0.863	0.617	0.919	0.878	0.652	0.867	**0.867**	0.950	0.900	0.933
	Appina2016 [3]	0.917	**0.782**	0.910	0.865	0.666	0.864	0.839	0.932	0.846	0.860
	Zhou2017 [52]	0.871	0.687	0.941	0.905	0.785	0.897	0.802	0.942	0.907	0.920
	Yang2018 [43]	0.903	0.739	0.927	0.862	0.799	0.908	0.793	0.920	0.892	0.908
	Yue2018 [46]	0.832	0.595	0.932	0.857	0.779	**0.959**	0.769	0.959	0.868	0.913
	Yang2019 [45]	0.897	0.768	0.929	0.917	0.685	0.859	0.806	0.864	0.834	0.877
	Li2019 [21]	0.910	0.760	0.930	0.864	0.789	0.918	0.834	0.950	**0.951**	0.929
	Liu2020 [23]	0.912	0.743	**0.953**	0.901	0.845	**0.921**	0.788	0.954	0.936	0.939
	Sun2020 [34]	**0.970**	0.687	0.893	**0.979**	0.853	0.897	0.579	0.933	**0.964**	0.918
	Proposed	**0.950**	**0.827**	**0.961**	0.917	**0.909**	0.864	**0.882**	**0.966**	0.925	0.948

distortion category, but its performance in JP2K and JPEG distortion on the two databases are not good, which leads to its overall performance is poor. In general, the proposed model performs well, which proves its robustness and effectiveness.

In order to show the predicted effect of the proposed model more intuitively, Fig. 9 gives the scatter plots of predicted DMOS against subjective DMOS on the LIVE 3D Phase I and 3D Phase II. The horizontal axis represents the DMOS predicted by the proposed model, and the vertical axis represents the subjective DMOS. The more the scatter points converge, the closer the fitting curve is to the straight line, which indicates that the model is better. From Fig. 9 it can be seen that the scatter distribution of various distortion types shows straight line, and the fitting curve is very close to the diagonal line of the first quadrant, which further shows that the proposed algorithm is linearly consistent with the subjective perception.

Table 4 Performance comparison of PLCC with five kinds of distortion

Type	Methods	LIVE 3D Phase I					LIVE 3D Phase II				
		JP2K	JPEG	WN	BLUR	FF	JP2K	JPEG	WN	BLUR	FF
FR	Chen2013 [5]	0.916	0.634	0.944	0.942	0.758	0.843	0.842	0.960	0.965	0.910
	Shao2017 [31]	0.936	0.665	0.944	0.954	0.830	0.877	0.851	0.934	0.945	0.933
	Jiang2018 [12]	0.941	0.698	0.952	0.958	0.855	0.846	0.877	0.955	0.985	0.960
	Li2019 [20]	**0.977**	**0.927**	0.956	**0.987**	**0.928**	0.940	0.855	**0.986**	0.990	**0.966**
NR	Chen2013 [6]	0.907	0.695	0.917	0.917	0.735	0.899	**0.901**	0.947	0.941	0.932
	Appina2016 [3]	0.938	0.806	0.919	0.881	0.758	0.867	0.829	0.920	0.878	0.836
	Yang2018 [43]	0.947	0.820	0.957	0.952	0.876	0.936	0.862	0.952	0.975	0.935
	Yue2018 [46]	0.934	0.744	0.962	**0.971**	0.854	**0.986**	0.843	**0.986**	0.973	0.923
	Yang2019 [45]	0.942	0.824	0.954	0.963	0.789	0.886	0.867	0.887	0.988	0.916
	Li2019 [21]	0.957	0.812	0.958	0.948	0.846	**0.950**	0.879	0.974	**0.991**	0.956
	Liu2020 [23]	0.948	0.768	**0.967**	0.910	0.852	0.942	0.788	0.973	0.986	0.949
	Sun2020 [34]	0.948	0.806	0.956	0.960	0.890	0.900	0.823	0.956	**0.996**	0.901
	Proposed	**0.973**	**0.842**	**0.970**	0.966	**0.916**	0.885	**0.897**	0.973	0.989	**0.970**

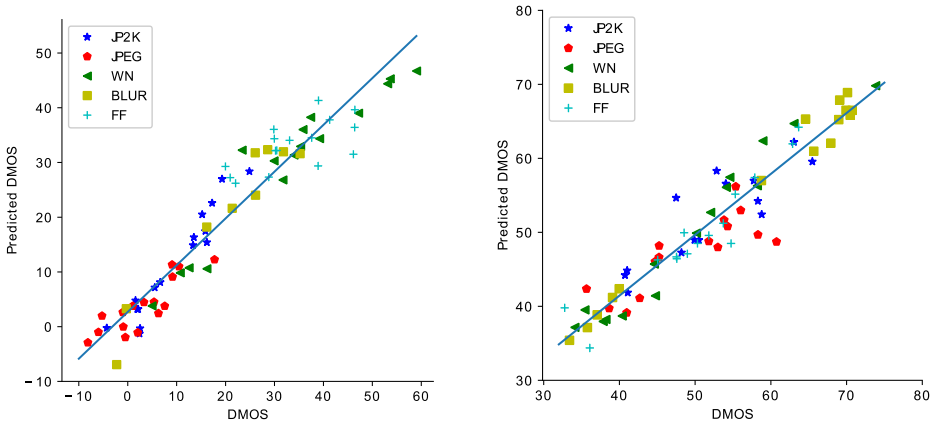

Fig. 9 Scatter plots of predicted DMOS of the proposed method against subjective DMOS. **a** On the LIVE 3D Phase I. **b** On the LIVE 3D Phase II

4.4 Cross-database validation

In order to further illustrate that the proposed model is not limited to the samples in the database, we conduct cross-database validation, training and learning the model on one database, and testing it with a different database. Here, two cross-test are conducted: (1) LIVE I/LIVE II, which means that the experiment is trained on LIVE 3D PHASE I and tested on LIVE 3D PHASE II, (2) LIVE II/LIVE I, which denotes that the experiment is trained on LIVE 3D PHASE II and tested on LIVE 3D PHASE I. The results are given in Table 5, and it can be seen that the results of cross-database test are significantly lower than those of the same training-testing database, because the distortion types and degree of the two database samples are quite different. Moreover, the performances of LIVE I/LIVE II are apparently the worse than that of LIVE II/LIVEI, and it may be that LIVE II contains both symmetric and asymmetric stereoscopic images, while LIVE I contains only symmetric distorted images, resulting in asymmetric distorted images not learned for LIVE I/LIVE II. Compared with other recently six methods, the proposed model shows competitive performances, which further suggest the proposed model is effective for SIQA, insensitive to image content, and has good universality and stability.

Table 5 Performances of cross-database validations

train/test	LIVEI/LIVEII		LIVEII/LIVEI	
	SROCC	PLCC	SROCC	PLCC
Zhou2017 [52]	0.827	–	0.899	–
Yang2018 [43]	0.817	0.829	0.905	0.910
Yang2019 [45]	**0.852**	0.849	0.869	0.860
Li2019 [21]	0.818	0.826	0.852	0.861
Liu2020 [23]	0.832	**0.862**	0.874	0.888
Sun2020 [34]	**0.870**	**0.899**	**0.918**	**0.919**
Proposed	0.831	0.851	**0.910**	**0.911**

4.5 Saliency threshold analysis

Saliency threshold determines how many saliency patches are selected, the bigger the threshold, the less the selected saliency patches; conversely, the smaller the threshold, the more the selected saliency region. Here we did a comparative experiment under different saliency thresholds. The results are listed in Tables 6, 7, 8 and 9 when threshold are set to 0, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, and the best results are shown in bold. On the whole, when the saliency threshold is 0.2, the results are the best, so we set the threshold to 0.2 in this paper.

4.6 Performance comparison on symmetrically and asymmetrically distorted images

We further carried out comparative experiment of symmetric and asymmetric distortions on LIVE 3D Phase II. We compared the proposed method with three FR (Benoit2008 [4],

Table 6 Overall performance Comparison of different significance thresholds on LIVE 3D Phase I

Threshold	0	0.1	0.2	0.3	0.4	0.6	0.8
SROCC	0.959	0.961	**0.962**	0.956	0.957	0.952	0.948
PLCC	0.961	0.964	**0.966**	0.962	0.962	0.960	0.955
RMSE	4.824	4.598	**4.470**	4.763	4.856	4.840	4.998

Table 7 Overall performance Comparison of different significance thresholds on LIVE 3D Phase II

Threshold	0	0.1	0.2	0.3	0.4	0.6	0.8
SROCC	0.944	0.949	**0.954**	0.952	0.951	0.947	0.934
PLCC	0.954	0.956	**0.957**	0.956	0.955	0.954	0.944
RMSE	3.636	**3.494**	3.520	3.499	3.561	3.746	4.202

Table 8 SROCC of individual distortion type of different significance thresholds on LIVE 3D Phase I

Threshold	0	0.1	0.2	0.3	0.4	0.6	0.8
JP2K	0.943	0.949	**0.950**	0.938	0.931	0.926	0.919
JPEG	0.835	**0.840**	0.827	0.825	0.836	0.812	0.820
WN	0.960	0.940	**0.961**	0.957	0.960	0.950	0.957
BLUR	0.932	0.908	0.917	0.917	**0.933**	0.917	0.883
FF	0.829	0.865	**0.909**	0.841	0.852	0.859	0.823

Table 9 SROCC of individual distortion type of different significance thresholds on LIVE 3D Phase II

Threshold	0	0.1	0.2	0.3	0.4	0.6	0.8
JP2K	0.900	0.896	0.864	**0.929**	0.921	0.896	0.891
JPEG	0.877	0.877	**0.882**	0.881	0.863	0.846	0.838
WN	0.954	0.961	**0.966**	0.957	0.950	0.950	0.950
BLUR	0.930	0.904	0.925	0.936	**0.954**	0.946	0.930
FF	0.945	0.946	**0.948**	0.946	0.946	0.921	0.936

Table 10 Performance Comparison on symmetrically and asymmetrically distorted images

Type	Methods	Symmetric		Asymmetric	
		SROCC	PLCC	SROCC	PLCC
FR	Benoit2008 [4]	0.696	0.734	0.747	0.770
	Chen2013 [5]	**0.925**	**0.938**	0.854	0.875
	Wang2015 [39]	0.923	0.937	**0.902**	**0.898**
NR	Mittal2012 [27]	0.872	0.868	0.559	0.575
	Chen2013 [6]	0.918	–	0.834	–
	Zhang2016 [50]	0.915	0.912	0.708	0.763
	Shao2018 [32]	–	–	0.838	0.894
	Proposed	**0.927**	**0.939**	**0.931**	**0.940**

Chen2013 [5], and Wang2015 [39]) and four NR (Mittal2012 [27], Chen2013 [6], Zhang2016 [50], and Shao2018 [32]) IQA methods. Their results have been reported in related papers, or the source code of the methods has been made public. The comparison results of SROCC and PLCC are summarized as Table 10, and the best two results are displayed in bold type. It can be seen that most of the other methods have achieved good results in symmetric distortion, indicating that they can accurately evaluate the perceptual quality of symmetrically distorted stereoscopic images, but the effect is poor in asymmetric distortion. However, the results of the proposed method on asymmetric distortion are better than those on symmetric distortion, which shows that our method is also suitable for asymmetric distortion. In general, the proposed method has the best performance in both symmetrically and asymmetrically distorted images.

5 Conclusion and future work

In this paper, we proposed a blind SIQA model via 3D saliency selected stereoscopic images and 3D-CNN. 3D saliency map is used to select salient image patches more suitable for SIQA. Finally, the objective quality score of stereoscopic image is obtained by weighted average method. The experimental results show that the SROCC of our proposed method on LIVE 3D Phase I and Phase II is 0.962 and 0.954, respectively. In cross-database validation, the SROCC of LIVE II/LIVEI is 0.910. Compared with the state-of-the-art NR SIQAs, our metric has higher performance, which shows its superiority and robustness.

In this method we select the salient local regions of the image to train the 3D-CNN model, and use trained scores from DMOS of the whole image, which are not certainly the real quality score, and maybe lead to limited subjective relationships. In the future research, it is necessary to get real quality scores of local image patches for nonuniform distortion.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 62176150, No. 61771223).

References

1. Agrawal P, Stansbury D, Malik J, Gallant J (2014) Pixels to voxels: modeling visual representation in the human brain. Eprint Arxiv 1(1):1–15

2. Akhter R, Sazzad Z, Horita Y, Baltes J (2010) No-reference stereoscopic image quality assessment. Processing SPIE 7524:1–12
3. Appina B, Khan S, Channappayya S (2016) No-reference stereoscopic image quality assessment using natural scene statistics. Signal Process Image Commun 43(1):1–14
4. Benoit A, Callet P, Campisi P, Cousseau R (2008) Quality assessment of stereoscopic images. EURASIP J Image Video Process 2008(1):1–13
5. Chen M, Su C, Kwon D, Cormack L, Bovik A (2013) Full-reference quality assessment of stereopairs accounting for rivalry. Signal Process Image Commun 28(9):1143–1155
6. Chen M, Cormack L, Bovik A (2013) No-reference quality assessment of natural stereopairs. IEEE Trans Image Process 22(9):3379–3391
7. Han Y, Yuan Z, Muntean G (2016) An innovative no-reference metric for real-time 3D stereoscopic video quality assessment. IEEE Trans Broadcast 62(3):654–663
8. Henriksen S, Read J (2016) Visual perception: a novel difference channel in binocular vision. Curr Biol 26(12):500–503
9. Hong GS, Kim BG (2017) A local stereo matching algorithm based on weighted guided image filtering for improving the generation of depth range images. Displays 49(1):80–87
10. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
11. Jia S, Zhang Y (2018) Saliency-based deep convolutional neural network for no-reference image quality assessment. Multimed Tools Appl 77(12):14859–14872
12. Jiang G, He M, Yu M, Shao F, Peng Z (2018) Perceptual stereoscopic image quality assessment method with tensor decomposition and manifold learning. IET Image Process 12(5):810–818
13. Joveluro P, Malekmohamadi H, Fernando WAC, Kondoz AM (2010) Perceptual video quality metric for 3D video quality assessment. In: Processing 2010 3DTV-conference: the true vision - capture, transmission and display of 3D video, pp 1–4
14. Kang L, Ye P, Li Y, Doermann D (2014) Convolutional neural networks for no-reference image quality assessment. In: Processing IEEE conference on computer vision and pattern recognition, pp 1733–1740
15. Kim BG, Park DJ (2002) Adaptive image normalisation based on block processing for enhancement of fingerprint image. Electron Lett 38(14):697–698
16. Kim BG, Shim JI, Park DJ (2003) Fast image segmentation based on multi-resolution analysis and wavelets. Pattern Recogn Lett 24(16):2995–3006
17. Kim J, Zeng H, Ghadiyaram D, Lee S, Zhang L, Bovik A (2017) Deep convolutional neural models for picture-quality prediction: challenges and solutions to data-driven image quality assessment. IEEE Signal Process Mag 34(6):130–141
18. Li Z, Atick J (1994) Efficient stereo coding in the multiscale representation. Netw Comput Neural Syst 5(2): 157–174
19. Li Y, Po L, Feng L, Yuan F (2016) No-reference image quality assessment with deep convolutional neural networks. In: Processing. IEEE international conference on digital signal processing (DSP), pp 685–689
20. Li S, Han X, Chang Y (2019) Adaptive cyclopean image-based stereoscopic image-quality assessment using ensemble learning. IEEE Trans Multimed 21(10):2616–2624
21. Li Y, Yang F, Wan W, Wang J, Gao M, Zhang J, Sun J (2019) No-reference stereoscopic image quality assessment based on visual attention and perception. IEEE Access 7(1):46706–46716
22. Lin W, Kuo CCJ (2011) Perceptual visual quality metrics: a survey. J Vis Commun Image Represent 22(4): 297–312
23. Liu Y, Yan W, Zheng Z, Huang B, Yu H (2020) Blind stereoscopic image quality assessment accounting for human monocular visual properties and binocular interactions. IEEE Access 8(1):33666–33678
24. Lv Y, Yu M, Jiang G, Shao F, Peng Z, Chen F (2016) No-reference stereoscopic image quality assessment using binocular self-similarity and deep neural network. Signal Process Image Commun 47(1):346–357
25. May K, Li Z (2016) Efficient coding theory predicts a tilt aftereffect from viewing untilted patterns. Curr Biol 26(12):1571–1576
26. May K, Li Z, Hibbard P (2012) Perceived direction of motion determined by adaptation to static binocular images. Curr Biol 22(1):28–32
27. Mittal A, Moorthy A, Bovik A (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
28. Moorthy A, Su C, Mittal A, Bovik A (2013) Subjective evaluation of stereoscopic image quality. Signal Process Image Commun 28(8):870–883
29. Mukherjee P, Das A, Bhunia AK, Roy PP (2019) Cogni-net: cognitive feature learning through deep visual perception. 2019 IEEE International Conference on Image Processing (ICIP), vol 1, no 1, pp 4539–4543
30. Potapova E, Zillich M, Vincze M (2011) Learning what matters: combining probabilistic models of 2d and 3d saliency cues. Comput Vis Syst 6962:132–142

31. Shao F, Chen W, Jiang G, Ho Y (2017) Modeling the perceptual quality of stereoscopic images in the primary visual cortex. IEEE Access 5(1):15706–15716
32. Shao F, Zhang Z, Jiang Q, Lin W, Jiang G (2018) Toward domain transfer for no-reference quality prediction of asymmetrically distorted stereoscopic images. IEEE Trans Circuits Syst Video Technol 28(3):573–585
33. Sun D, Roth S, Black M (2010) Secrets of optical flow estimation and their principles. In: Processing IEEE computer society conference on computer vision and pattern recognition, pp 2432–2439
34. Sun G, Shi B, Chen X, Krylov AS, Ding Y (2020) Learning local quality-aware structures of salient regions for stereoscopic images via deep neural networks. IEEE Trans Multimed 22(11):2938–2949
35. Talebi H, Milanfar P (2018) NIMA: Neural Image Assessment. IEEE Trans Image Process 27(8):3998–4011
36. Tsotsos J, Culhane S, Wai W, Lai Y, Davis N, Nuflo F (1995) Modeling visual attention via selective tuning. Artif Intell 78(1–2):507–545
37. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
38. Wang J, Da Silva MP, Le Callet P, Ricordel V (2013) Computational model of stereoscopic 3D visual saliency. IEEE Trans Image Process 22(6):2151–2165
39. Wang J, Rehman A, Zeng K, Wang S, Wang Z (2015) Quality prediction of asymmetrically distorted stereoscopic 3D images. IEEE Trans Image Process 24(11):3400–3414
40. Wang YF, Shuai Y, Zhu Y, Zhang J, An P (2019) Jointly learning perceptually heterogeneous features for blind 3D video quality assessment. Neurocomputing 332(1):298–304
41. Weinberger MJ, Seroussi G, Sapiro G (2000) The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS. IEEE Trans Image Process 9(8):1309–1324
42. Yang J, Zhu Y, Ma C, Lu W, Meng Q (2018) Stereoscopic video quality assessment based on 3D convolutional neural networks. Neurocomputing 309(1):83–93
43. Yang J, Sim K, Jiang B, Lu W (2018) No-reference stereoscopic image quality assessment based on hue summation-difference mapping image and binocular joint mutual filtering. Appl Opt 57(14):3915–3926
44. Yang J, Zhao Y, Chan J, Xiao L (2019) A multi-scale wavelet 3D-CNN for hyperspectral image super-resolution. Remote Sens 11(13):1–22
45. Yang J, Zhao Y, Zhu Y, Xu H, Lu W, Meng Q (2019) Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network. Inf Sci 474(1):1–17
46. Yue G, Hou C, Jiang Q, Yang Y (2018) Blind stereoscopic 3D image quality assessment via analysis of naturalness, structure, and binocular asymmetry. Signal Process 150(1):204–214
47. Zhang L, Zhang L, Mou X, Zhang D (2011) FSIM: a feature similarity index for image quality assessment. IEEE Trans Image Process 20(8):2378–2386
48. Zhang L, Gu Z, Li H (2013) SDSP: a novel saliency detection method by combining simple priors. In: Processing IEEE international conference on image processing, pp 171–175
49. Zhang L, Shen Y, Li H (2014) VSI: a visual saliency-induced index for perceptual image quality assessment. IEEE Trans Image Process 23(10):4270–4281
50. Zhang W, Qu C, Ma L, Guan J, Huang R (2016) Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. Pattern Recogn 59(1):176–187
51. Zhang P, Wang X, Zhang W, Chen J (2019) Learning spatial–spectral–temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment. IEEE Trans Neural Syst Rehabil Eng 27(1):31–42
52. Zhou W, Yu L, Zhou Y, Qiu W, Wu M, Luo T (2017) Blind quality estimator for 3D images based on binocular combination and extreme learning machine. Pattern Recogn 71(1):207–217