



No-reference stereoscopic image quality assessment using 3D visual saliency maps fused with three-channel convolutional neural network

Chaofeng Li¹ · Lixia Yun¹ · Hui Chen² · Shoukun Xu³

Received: 13 September 2020 / Revised: 24 May 2021 / Accepted: 10 July 2021 / Published online: 29 July 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

In this paper, we present a depth-perceived 3D visual saliency map and propose a no-reference stereoscopic image quality assessment (NR SIQA) algorithm using 3D visual saliency maps and convolutional neural network (CNN). Firstly, the 2D salient region of stereoscopic image is generated, and the depth saliency map is calculated, and then, they are combined to compute 3D visual saliency map by linear weighted method, which can better use depth and disparity information of 3D image. Finally, 3D visual saliency map, together with distorted stereoscopic pairs, is fed into a three-channel CNN to learn human subjective perception. We call proposed depth perception and CNN-based SIQA method DPCNN. The performances of DPCNN are evaluated over the popular LIVE 3D Phase I and LIVE 3D Phase II databases, which demonstrates to be competitive with the state-of-the-art NR SIQA algorithms.

Keywords No-reference stereoscopic image quality assessment · 3D visual saliency maps · convolutional neural network · Depth saliency map

1 Introduction

Image and video will inevitably lead to distortion during acquisition, synthesis, compression, and transmission, etc. In order to address this issue, effective image quality assessment (IQA) has long been an active research field [1–3]. In the real world, ideal reference image is hard to gain, so no-reference IQA methods are more practical and have wider application prospects. As more and more image processing operations have been specifically designed for stereoscopic images, no-reference (NR) stereoscopic IQA (SIQA) shows more important.

Comparing to 2D IQA, SIQA is much more sophisticated, because it is affected by many factors, such as 2D image quality, depth perception, visual comfort, and other factors [2]. It is particularly challenging, especially under

the condition, that the stereoscopic image pair consists of two views with different quality levels. Therefore, how to understand the binocular vision perception is a very important issue in 3D IQA.

During the past few decades, binocular vision perception for 3D IQA has made significant progress. Häkkinen et al. [4] studied the difference in eye movement patterns between 2 and 3D versions by viewing the same video content. They found that eye movements were more widely distributed in 3D content. Their results show that depth information from binocular depth cues provides viewers with additional information, creating new salient regions in the scene. In building a stereoscopic saliency model, depth (binocular disparity or disparity contrast) is considered as an attentional cue that is similar to other features such as color and orientation [5, 6]. Jansen et al. [5] suggested that when building a 3D salient model, in addition to considering the disparity factor, it is also necessary to extend features of monocular image, such as luminance and texture contrast. The authors of [6] learned color characteristic and proposed a NR-SIQA method using gradient dictionary. Maki et al. [7] suggested that a target closer in depth has higher priority for 3D saliency model. Wang et al. [8] presented a stereo saliency model integrating edges, disparity boundaries, and a saliency bias computed from stereoscopic perception. Fang et al. [9] proposed to

✉ Chaofeng Li
wxlichaocheng@126.com

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

² School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China

³ School of Information Science and Engineering, Changzhou University, Changzhou 213164, China

incorporate the color, luminance, texture, and depth cues to generate the saliency map for 3D images. Nevertheless, the benefit of using the spatial variance [10] of a color image as a depth map is not always accurately estimated. Li et al. [11] built a saliency dictionary by selecting a group of potential foreground objects, pruning the outliers in the dictionary, and running iterative tests on the remaining superpixels to refine the dictionary. For stereoscopic saliency detection, disparity was appended directly to the feature vector. Liu et al. [12] used the binocular combined behaviors and human 3D visual saliency characteristics to evaluate the quality of stereoscopic images. Li et al. [13] designed a NR-SIQA method based on visual attention and perception, which used saliency and just noticeable difference (JND) to weight the global and local features extracted from the left and right images. Meanwhile, the global structural features are extracted from the cyclopean map.

Since the human eye is the ultimate receiver of stereoscopic images, HVS characteristics, such as binocular rivalry and suppression mechanisms and depth perceptive characteristics, need to be considered for designing SIQA. Some researchers combined depth (or disparity) information with 2D image information together to analyze the quality of stereoscopic image. Benoit et al. [14] proposed a stereo image quality assessment method using some 2D full-reference quality assessment algorithms such as C4 and SSIM to calculate the disparity map between the distorted image and reference images. You et al. [15] applied various methods to calculate disparity maps and used a variety of 2D IQA methods to evaluate image pairs and disparity maps of stereoscopic images. They proved that disparity is an important factor in stereoscopic vision. Chen et al. [16] extracted 2D natural scene statistical features from the cyclopean of stereo images and stereo features from disparity maps and uncertainty maps and then learn and predict quality scores. Akhter et al. [17] extracted features from stereoscopic image pairs and disparity maps and used logistic regression models to predict quality scores.

In recent years, deep learning shows great success on object detection, recognition, and semantic segmentation [18–22], and some researchers also begin to use deep learning for NR IQA. The authors of [23] combined the feature extraction and learning process, and simultaneously used the max pool and mean pool to reduce the dimensionality of feature maps, and employed a shallow CNN to assess image quality and got a relatively good performance. Bosse et al. [24] designed an in-depth CNN for 2D IQA and achieved relatively high results. Zhang et al. [25] designed a CNN model, which uses left view, right view, and disparity maps normalized by luminance and contrast as input to the network for SIQA, and also improve the results. Zhou et al. [26] devised a deep-fused CNN model for NR-SIQA, which consists of monocular feature encoded networks and binocular feature fused networks, followed by a quality prediction

layer. These studies show the tremendous potential of CNN for improving SIQA performance.

Based on above analysis, here, we consider visual saliency and CNN to design NR SIQA. A depth-perceived 3D visual saliency model is proposed by fusing 2D saliency map and depth saliency map, which can more realistically reflect the extraction procedure of the salient region in human vision system. Both the stereoscopic image pair and 3D visual saliency map are fed into CNN to train and learn stereoscopic image quality scores. The main contributions are as follows:

1. We propose a new NR-SIQA model based on three-channel CNN, which utilized both stereoscopic image pairs and 3D visual saliency map.
2. A depth-perceived 3D visual saliency map is constructed by weighting 2D saliency map and depth saliency map.

The rest of this paper is organized as follows. The proposed NR SIQA model is elaborated in Sect. 2. Experimental results and analyses are presented in Sect. 3. Conclusions are provided in Sect. 4.

2 Proposed NR SIQA method

The framework of proposed NR SIQA is shown in Fig. 1, which includes four steps. Firstly, a 2D saliency map is calculated by SDSP algorithm [23]. Secondly, a depth saliency map is calculated by the optical flow algorithm [27]. Thirdly, a 3D visual saliency map is gained by weighting the 2D saliency map and the depth saliency map. At last, the 3D saliency map, left view and right view are fed into a three-channel CNN to learn image quality prediction model.

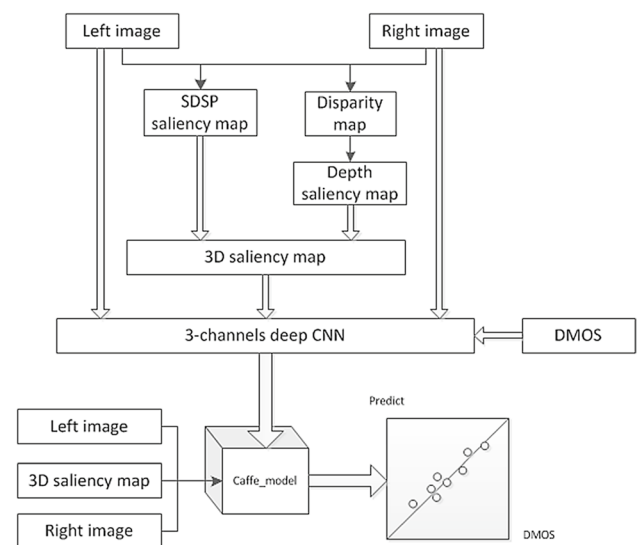


Fig. 1 The framework of proposed NR SIQA

2.1 2D saliency map by SDSP algorithm

The authors of [27] proposed a SDSP algorithm to calculate the salient region of the image by using three kinds of prior knowledge. In the SDSP model, a band-pass filtering saliency map $S_F(x)$, a location saliency map $S_D(x)$, and a color saliency map $S_C(x)$ are calculated based on the three prior knowledge as follows.

2.1.1 Bandpass filtering

The behavior of HVS detecting salient objects in a visual scene can be well modeled by band-pass filtering [28].

A log-Gabor filter was used to select the salient regions of the bandwidth. The filter conversion formula is shown as following:

$$G(u) = \exp\left(\frac{-\left(\log \frac{\|u\|_2}{\omega_0}\right)^2}{2\sigma_F^2}\right) \tag{1}$$

where $u = (u, v) \in R^2$ is the coordinates in the frequency domain, ω_0 is the center frequency of the filter, and σ_F controls the filtering bandwidth which can be obtained by inverse Fourier transform. For a RGB image $f(x)$, it is first converted to CIE L*a*b* space to get three channels, namely $f_L(x), f_a(x)$ and $f_b(x)$. The band-pass filtering saliency map $S_F(x)$ is defined as Eq. (2).

$$S_F(x) = ((f_L * g)^2 + (f_a * g)^2 + (f_b * g)^2)(x) \tag{2}$$

where “*” represents the convolution operation.

2.1.2 Position significance

Objects near the center of the image are more attractive to people [29]. This means that the position near the center of the image is more likely to be "significant" than the position far from the center. Assume that c is the center of the image.

The “Position Significance” at pixel x can be represented by Eq. (3):

$$S_D(x) = \exp\left(-\frac{\|x - c\|_2^2}{\sigma_D^2}\right) \tag{3}$$

2.1.3 Color saliency map

Compared to cool colors, the human eye is more likely to be attracted by warm colors [30]. In order to distinguish cool and warm colors in the image, we convert it

to CIEL*a*b* space. The a* channel represents green-red color information and b* represents blue-yellow color information, and the larger value of a pixel means that the point is more similar to red and yellow in color. Based on the above theory, the color saliency map $S_C(x)$ is defined by following Eq. (4).

$$S_C(x) = 1 - \exp\left(-\frac{f_{an}^2 + f_{bn}^2(x)}{\sigma_c^2}\right) \tag{4}$$

where σ_c is the parameter, $f_{an}(x)$ is the min-max normalized a* channel, and $f_{bn}(x)$ is the min-max normalized b* channel.

Finally, the calculation of the saliency map is shown as follows:

$$SDSP(x) = S_C(x) \times S_D(x) \times S_F(x) \tag{5}$$

Figure 2 shows the calculation procedure of SDSP saliency map.

2.2 Depth saliency map

In general, when the target object and the background image have different depth values, the target object can attract more human visual attention. In this paper, the optical flow algorithm [31] is used to calculate the disparity map of a stereoscopic image. The optical flow estimation-based algorithm is developed by the classical optical flow objective function as following Eq. (6).

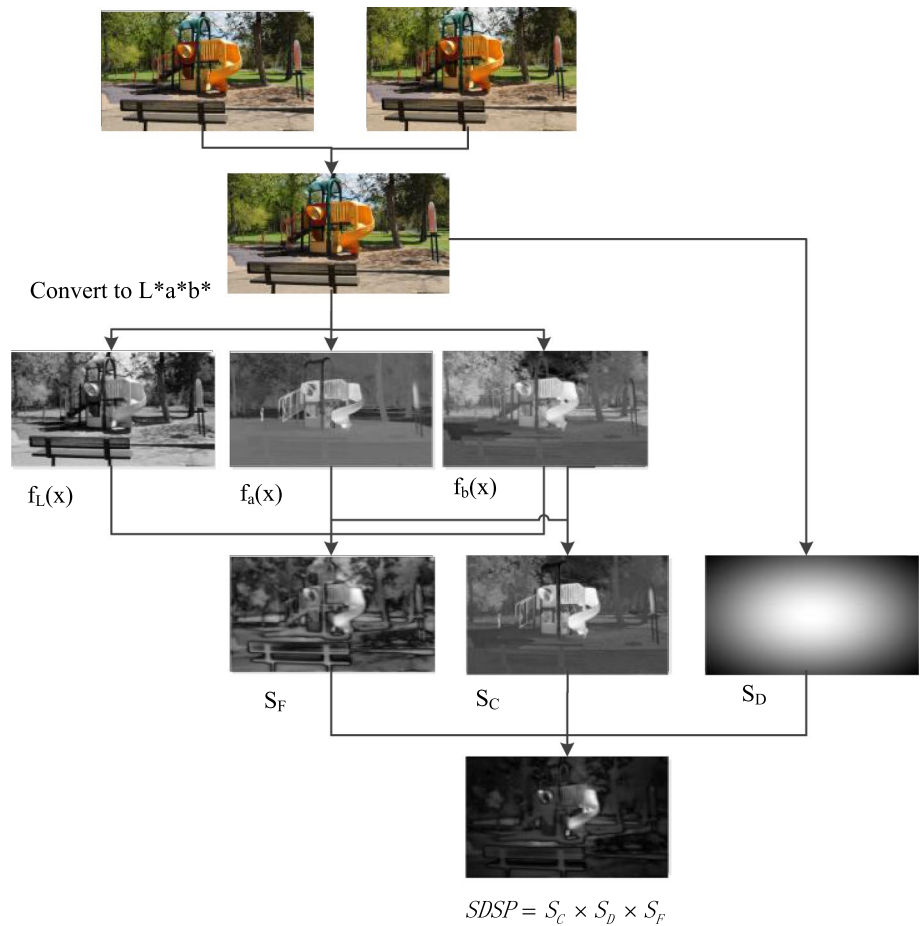
$$E(u, v) = \sum_{ij} \{ \rho_D(I_L(i, j) - I_R(i + u_{ij}, j + v_{ij})) + \lambda[\rho_s(u_{ij} - u_{i+1,j}) + \rho_s(u_{ij} - u_{i,j+1}) + \rho_s(v_{ij} - v_{i+1,j}) + \rho_s(v_{ij} - v_{i,j+1})] \} \tag{6}$$

where u and v are the horizontal and vertical components of the optical flow field obtained from the left and right images. λ is a regularization parameter, ρ_D denotes a data penalty function, and ρ_s is a space penalty function. Since the horizontal disparities on the retina are related to the depth perception, here, only the horizontal component of the computed motion vectors u is chosen as the horizontal disparities. The depth feature map can be directly obtained by the disparity map as following Eq. (7).

$$f_D(x, y) = 1 - \frac{D(x, y) - D_{\min}}{D_{\max} - D_{\min}} \tag{7}$$

where D_{\min} and D_{\max} are the minimum and maximum values in the disparity map, $D(x, y)$ is the value at the location (x, y) in the disparity map and subjects to $D_{\min} \leq D(x, y) \leq D_{\max}$. Objects closer to the camera have smaller depth values because they produce larger disparity values, and distant

Fig.2 The flowchart of calculating SDSP saliency map



objects have larger depth values because the far objects produce smaller values.

2.3 3D visual saliency map

Here, we propose a fused 3D visual saliency map, by combining the 2D saliency map with the depth saliency map. Linear weighting method is used to combine the 2D saliency map and the depth saliency map as following Eq. (8).

$$SVMap(x, y) = w_1 \times VS_{2D}(x, y) + w_2 \times D(x, y) \tag{8}$$

where w_1 and w_2 are the weight parameter, which they are simply set to 0.5 in this paper.

Figure 3 shows the left view in a stereo image pair and its corresponding 2D saliency map, depth saliency map, and 3D saliency map.

2.4 Structure of designed CNN for NR SIQA

In recent years, convolutional neural networks have achieved great success in the representation and classification of image features, language recognition, etc. In this paper, a three-channel CNN is constructed for NR SIQA, as shown in Fig. 4. The network parameters are configured as listed in Table 1.

The inputs of the network are a left-view and right-view image of a stereoscopic image and their 3D saliency

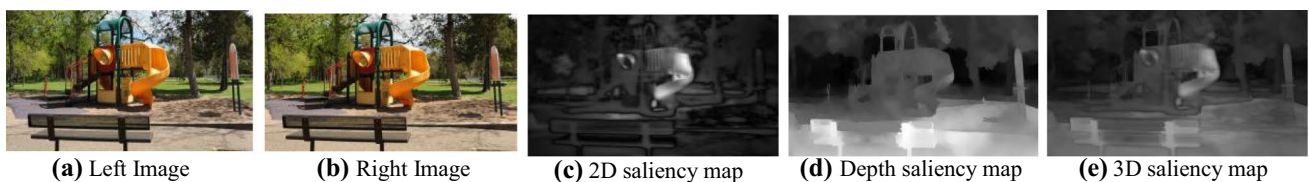


Fig.3 Left–Right image and their saliency map

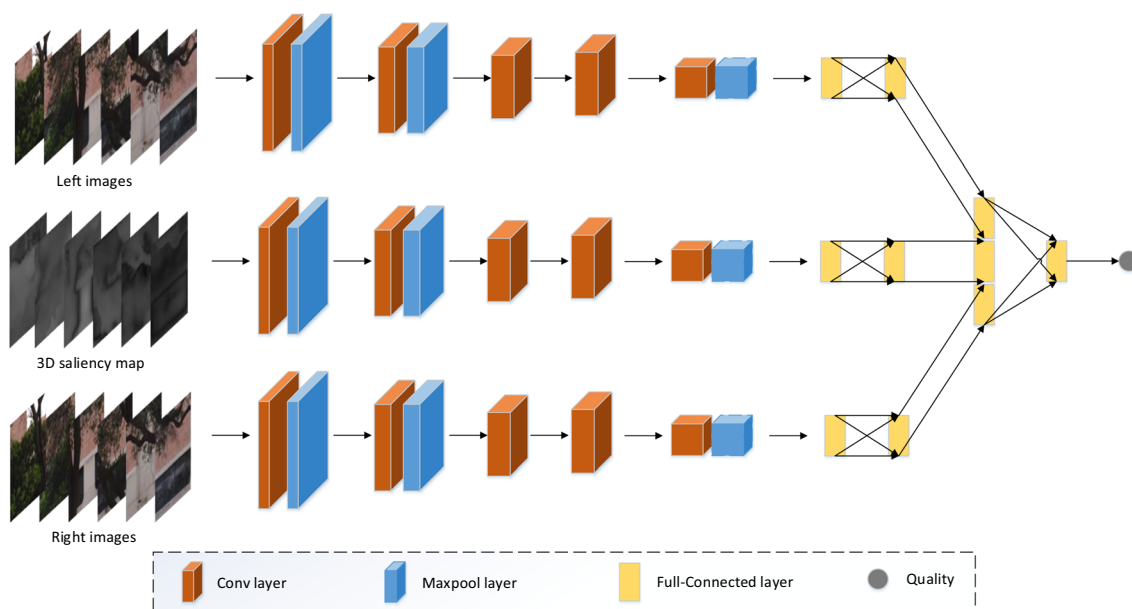


Fig.4 Structure of proposed three-channel CNN for SIQA

Table 1 Parameters of CNN

Left image channel	Right image channel	Saliency map channel
Input: 32 × 32 patches		
Conv3-32	Conv3-32	Conv3-32
maxpool	maxpool	maxpool
Conv3-32	Conv3-32	Conv3-32
maxpool	maxpool	maxpool
Conv3-64	Conv3-64	Conv3-64
Conv3-64	Conv3-64	Conv3-64
Conv3-128	Conv3-128	Conv3-128
maxpool	maxpool	maxpool
FC-512	FC-512	FC-512
FC-512	FC-512	FC-512
Three channels of linearly spliced vectors (FC-3 × 512)		
FC-512		
output		

map. Since the neural network input is usually a fixed size, but the image size in the image database is not always the same, we cut the original color image into a 32 × 32 image patches as input. Because the picture distortion in the LIVE 3D image database is uniformly distorted, each input block is given the same quality score as the original image. The last predicted image quality score is the average value of all image block quality scores of an image.

In all convolutional layers, the convolution kernel size used was all 3 × 3. Except for the last fully connected layer, all layers are activated through the ReLU [32] activation function. The convolutions are applied with zero-padding, so their output has the same spatial dimensions

as their input. The window size for all largest pools is 2 × 2. We apply dropout regularization [33] with a ratio 0.35 to the fully connected layers. Given an image represented by N_p randomly sampled patches and a ground truth quality label of q_t . The quality prediction q is calculated by averaging the CNN output y_i for each patch:

$$q = \frac{1}{N_p} \sum_i^{N_p} y_i \tag{8}$$

The definition of the objective function in the model is shown as follows:

$$E_{patchwise} = \frac{1}{N_p} \sum_i^{N_p} \|y_i - q_t\|_2^2 \tag{9}$$

Note that, the optimization is done through the adaptive learning rate optimizer ADAM [34] with $\alpha = 0.0001$.

3 Experimental result and analysis

3.1 Databases and evaluation metrics

The LIVE 3D Phase I [2] database includes 20 reference images, five distortion types, and a total of 365 distorted images, including 45 groups of Gaussian blur (Blur) distortion, 80 groups of JPEG2000 compression (JP2K), JPEG compression (JPEG), white noise (WN), and fast fading (FF).

distortion, and the DMOS value of each distorted stereo image. The LIVE 3D Phase II [16] database consists of eight pairs of reference stereo image scenes and 360 pairs of distorted stereo image pairs. The distortion type is the same as LIVE 3D Phase I.

Two commonly used indicators are employed to evaluate the performances, the Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SROCC). These coefficients measure the correlation

between the objective scores predicted by the objective stereoscopic IQA models and the subjective mean.

opinion scores (MOS) provided by the database. The PLCC is used to evaluate prediction accuracy, and the SROCC evaluates prediction monotonicity. A better model is expected to have higher PLCC and SROCC.

In our experiments, all distorted images are selected 80% as training set and the remaining 20% as testing set. The results are obtained after 100 epochs.

Table 2 SROCC comparison of several metrics on the LIVE 3D Phase I database

Metrics	JP2K	JPEG	WN	Blur	FF	ALL
Akhter [17]	0.866	0.675	0.914	0.555	0.640	0.383
Chen [16]	0.863	0.617	0.919	0.877	0.652	0.891
Zhang [25]	0.931	0.693	0.946	0.909	0.834	0.943
Liu [12]	0.930	0.738	0.939	0.944	0.774	0.939
Shao [36]	–	0.693	0.899	0.853	–	0.885
Oh [37]	0.885	0.765	0.921	0.930	0.944	0.935
Wang [38]	0.881	0.513	0.944	0.931	0.686	0.916
Li [13]	0.910	0.760	0.930	0.864	0.789	0.953
Liu [39]	0.901	0.962	0.926	0.891	0.834	0.936
Yang [40]	0.897	0.768	0.929	0.917	0.685	0.944
Proposed	0.947	0.871	0.949	0.933	0.884	0.961

Bold font denotes the best result

Table 3 PLCC comparison of several metrics on the LIVE 3D Phase I database

Metrics	JP2K	JPEG	WN	Blur	FF	ALL
Akhter [17]	0.905	0.729	0.904	0.617	0.503	0.626
Chen [16]	0.907	0.695	0.917	0.968	0.937	0.895
Zhang [25]	0.926	0.740	0.944	0.930	0.883	0.947
Liu [12]	0.945	0.753	0.933	0.959	0.838	0.937
Shao [36]	–	0.703	0.896	0.862	–	0.891
Oh [37]	0.913	0.767	0.910	0.950	0.954	0.943
Wang [38]	0.909	0.617	0.949	0.951	0.778	0.924
Liu [39]	0.925	0.713	0.929	0.948	0.857	0.941
Yang [40]	0.942	0.824	0.954	0.963	0.789	0.956
Proposed	0.966	0.891	0.965	0.968	0.915	0.963

Bold font denotes the best result

Table 4 SROCC comparison of several metrics on the LIVE 3D Phase II database

Metrics	JP2K	JPEG	WN	Blur	FF	ALL
Akhter [17]	0.724	0.649	0.714	0.682	0.559	0.543
Chen [16]	0.867	0.867	0.950	0.900	0.933	0.880
Zhang [25]	0.793	0.581	0.780	0.864	–	0.708
Shao [36]	–	0.622	0.803	0.713	–	0.805
Oh [37]	0.853	0.822	0.833	0.889	0.878	0.871
Li [13]	0.918	0.834	0.950	0.951	0.929	0.946
Liu [39]	0.923	0.856	0.954	0.947	0.924	0.949
Yang [40]	0.859	0.806	0.864	0.834	0.877	0.921
Proposed	0.928	0.854	0.955	0.936	0.946	0.951

Bold font denotes the best result

3.2 Performance comparison

Overall SROCC and PLCC on the LIVE 3D Phase I and Phase II database are given in Table 2, 3, 4, and 5, for comparison with current reported representative NR SIQA methods including Akhter [17], Chen [16], Zhang [25], Liu [12], Shao [36], Oh [37], Wang [38], Li [13], Liu [39], Yang [40], and their results are also listed. All in Table 2, 3, 4, and 5 indicates that the test image contains five distortion types.

From Table 2 and 3, it can be seen that our proposed method has achieved the best performance against currently reported methods on single JP2K and WN distortion types, and all distortion, and only a little inferior to Oh [37] on the FF distortion type.

From Table 4 and 5 on the LIVE 3D Phase II database, our proposed NR SIQA shows the best performance on single JP2K, WN, FF distortion types and all distortion, but SROCC is inferior to Liu [39] in JPEG and Blur distortion types, which verify our proposed method is more reliable.

Figure 5 gives scatter points between predicted DMOS via proposed NR SIQA and DMOS on the LIVE 3D Phase I and Phase II databases, which further suggests proposed algorithm is linear consistent with subjective perception.

3.3 Effect of 2D, depth, and 3D saliency map

In order to illustrate the importance of our proposed 3D saliency map, we input the 2D and depth saliency map into the CNN model instead of the 3D saliency map for training and

Table 5 PLCC comparison of several metrics on the LIVE 3D Phase II database

Metrics	JP2K	JPEG	WN	Blur	FF	ALL
Akhter [17]	0.766	0.786	0.722	0.795	0.674	0.568
Chen [16]	0.899	0.901	0.947	0.941	0.932	0.895
Zhang [25]	0.782	0.583	0.796	0.924	–	0.763
Shao [36]	–	0.583	0.822	0.755	–	0.784
Oh [37]	0.865	0.821	0.836	0.934	0.815	0.863
Liu [39]	0.919	0.868	0.966	0.973	0.941	0.950
Yang [40]	0.886	0.867	0.887	0.988	0.916	0.934
Proposed	0.935	0.873	0.977	0.987	0.962	0.954

Bold font denotes the best result

Fig.5 Scatter points between predicted DMOS and DMOS

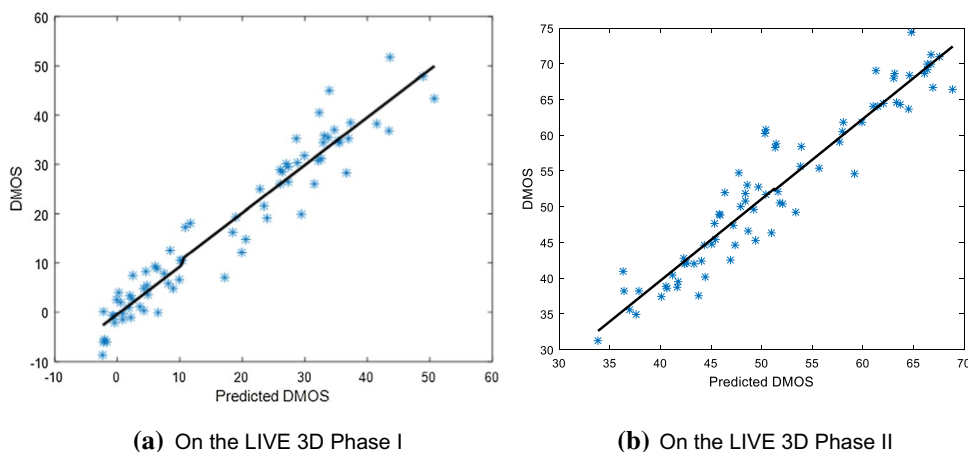


Table 6 Performance comparisons of different features on LIVE 3D Phase I database

	Saliency map	JP2K	JPEG	WN	Blur	FF	All
SROCC	2D	0.944	0.762	0.979	0.913	0.877	0.944
	depth	0.943	0.859	0.946	0.958	0.874	0.954
	3D	0.947	0.871	0.949	0.933	0.884	0.961
PLCC	2D	0.952	0.831	0.947	0.956	0.909	0.950
	depth	0.958	0.871	0.964	0.975	0.901	0.956
	3D	0.966	0.891	0.965	0.968	0.915	0.963

Bold font denotes the best result

Table 7 Performance comparisons of different features on LIVE 3D Phase II database

	Saliency map	JP2K	JPEG	WN	Blur	FF	All
SROCC	2D	0.928	0.930	0.990	0.932	0.942	0.906
	depth	0.861	0.846	0.936	0.928	0.939	0.949
	3D	0.928	0.854	0.955	0.936	0.946	0.951
PLCC	2D	0.898	0.914	0.953	0.951	0.939	0.875
	depth	0.902	0.824	0.972	0.977	0.955	0.950
	3D	0.935	0.873	0.977	0.987	0.962	0.954

Bold font denotes the best result

Table 8 Performance comparison of different features

Input features	LIVE 3D Phase I		LIVE 3D Phase II	
	SROCC	PLCC	SROCC	PLCC
3D saliency map	0.938	0.949	0.863	0.876
Left and right images	0.958	0.962	0.946	0.949
3D saliency map, Left and right images	0.961	0.963	0.951	0.954

Bold font denotes the best result

Table 9 Performances of cross-database validations

Train/test	LIVE I/LIVE II		LIVE II/LIVE I	
	SROCC	PLCC	SROCC	PLCC
Akater [17]	0.543	0.568	0.383	0.626
Chen [16]	0.543	0.895	0.891	0.626
Yang [41]	0.817	0.829	0.905	0.910
Li [13]	0.818	0.826	0.852	0.861
Proposed	0.788	0.808	0.910	0.916

Bold font denotes the best result

learning, and the test results are listed in Table 6 and 7. All in Table 6 and 7 indicates that the test image contains five distortion types.

It can be seen from Table 6 that using a 3D saliency map shows good performance for the single JP2K, JPEG, WN, FF distortion types, and all distortions on Phase I, while using a 2D saliency map only shows the best performance on Blur. In addition, from Table 7, it can be seen using a 3D saliency map has satisfactory performance for the single JP2K, WN, Blur, FF distortion types, and all distortions. Overall, the 3D saliency map proposed by us has obvious advantages and is effective for stereoscopic image quality evaluation.

3.4 Influence of 3D saliency map and image pairs

In order to further analyze the role of stereoscopic image and 3D saliency map in the model, the ablation experiments are carried out. In Table 8, we give the results of inputting only 3D saliency map, the left and right images, and combining 3D saliency map and the left and right images. It can be seen that the proposed method achieves

the best performance, which shows that the features extracted from left and right images and 3D saliency map are complementary.

3.5 Cross-database validation

Furthermore, we conducted cross-database validation to test the stability of the proposed model. The results are shown in Table 9, where LIVE I/LIVE II represents that the experiment is trained on LIVE 3D Phase I and tested on Phase II; similarly, LIVE II/LIVE I denotes that the experiment is trained on LIVE 3D Phase II and tested on Phase I. The best results are highlighted in bold. It can be noticed that the proposed model shows competitive performances, which suggest the proposed model is reliable for measuring the quality of the stereoscopic image, insensitive to image content, and has good universality and stability.

4 Conclusion

In this paper, we have proposed a NR SIQA algorithm based on 3D visual saliency maps and deep convolutional neural network (DCNN). 3D visual saliency map is generated by combining the depth saliency map of a stereoscopic image with its 2D saliency information. Our experiments suggest that proposed 3D visual saliency map is better than only 2D saliency map. At last, the 3D saliency map and left–right views are fed into a designed DCNN to predict the stereoscopic image quality score. Experimental results on the LIVE 3D Phase I and Phase II databases show our proposed method gets better performance than other NR-SIQA methods.

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

References

1. Appina, B., Khan, S., Channappayya, S.S.: No-reference stereoscopic image quality assessment using natural scene statistics. *Signal Process. Image Commun.* **43**, 1–14 (2016)
2. Moorthy, A.K., Su, C.C., Mittal, A., et al.: Subjective evaluation of stereoscopic image quality. *Signal Process. Image Commun.* **28**(8), 870–883 (2013)

3. Zhang, W., Borji, A., Wang, Z., et al.: The application of visual saliency models in objective image quality assessment: a statistical evaluation. *IEEE Trans. Neural Netw. Learn. Syst.* **27**(6), 1266–1278 (2016)
4. Häkkinen, A.J., Kawai, T., Takatalo, J., et al.: What do people look at when they watch stereoscopic movies? In: *Proceedings of SPIE—The International Society for Optical Engineering*, vol. 7524, pp. 75240E-75240E-10 (2010)
5. Jansen, L., Onat, S., König, P.: Influence of disparity on fixation and saccades in free viewing of natural scenes. *J. Vis.* **9**(1), 1–19 (2009)
6. Yang, J., An, P., Ma, J., Li, K., Shen, L.: No-reference stereo image quality assessment by learning gradient dictionary-based color visual characteristics. In: *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5 (2018)
7. Maki, A., Nordlund, P., Eklundh, J.-O.: Attentional scene segmentation: integrating depth and motion. *Comput. Vis. Image Underst.* **78**(3), 351–373 (2000)
8. Wang, W., Shen, J., Yu, Y., Ma, K.-L.: Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE Trans. Vis. Comput. Graphics* **23**(8), 2014–2027 (2017)
9. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W.: Saliency detection for stereoscopic images. *IEEE Trans. Image Process.* **23**(6), 2625–2636 (2014)
10. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: contrast based filtering for salient region detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–740 (2012).
11. Li, N., Sun, B., Yu, J.: A weighted sparse coding framework for saliency detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5216–5223 (2015)
12. Liu, Y., Yang, J.C., Meng, Q.G., et al.: Stereoscopic image quality assessment method based on binocular combination saliency model. *Signal. Process.* **125**, 237–248 (2016)
13. Li, Y.F., Yang, F., Wan, W.B., et al.: No-reference stereoscopic image quality assessment based on visual attention and perception. *IEEE Access* **7**, 46706–46716 (2019)
14. Benoit, A., Le Callet, P., Campisi, P., et al.: Quality assessment of stereoscopic images. *EURASIP J. Image Video Process.* **2008**, 659024 (2009). <https://doi.org/10.1155/2008/659024>
15. You, J., Xing, L., Perki, A., et al.: Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis. In: *Proceedings of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Arizona, U.S.A.*, pp. 1–6 (2010)
16. Chen, M.J., Cormack, L.K., Bovik, A.C.: No-reference quality assessment of natural stereopairs. *IEEE Trans. Image Process.* **22**(9), 3379–3391 (2013)
17. Akhter, R., Sazzad, Z.M.P., Horita, Y., et al.: No-reference stereoscopic image quality assessment. In: *Proceedings of Stereoscopic Displays and Applications XXI. International Society for Optics and Photonics, 75240T-75240T-12* (2010)
18. Zhou, W., Wu, J., Lei, J., et al.: Salient object detection in stereoscopic 3D images using a deep convolutional residual autoencoder. In: *IEEE Transactions on Multimedia*, pp. 1–12 (2020)
19. Zhang, W., Zhang, Y., Ma, L., et al.: Multimodal learning for facial expression recognition. *Pattern Recogn.* **48**(10), 3191–3202 (2015)
20. Zhou, W., Yuan, J., Lei, J., et al.: TSNet: three-stream self-attention network for RGB-D indoor semantic segmentation. In: *IEEE intelligent systems*, pp. 1–5 (2020).
21. Xu, Y., Du, J., Dai, L.-R., et al.: An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.* **21**, 65–68 (2014)
22. Zhou, W., Liu, W., Lei, J., et al.: Deep Binocular Fixation Prediction using a Hierarchical Multimodal Fusion Network. In: *IEEE Transactions on Cognitive and Developmental Systems*
23. Kang, L., Ye, P., Li, Y., et al.: Convolutional neural networks for no-reference image quality assessment. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733–1740 (2014)
24. Bosse, S., Maniry, D., Wiegand, T., et al.: A deep neural network for image quality assessment. In: *Proceedings of IEEE international conference on image processing*, pp. 3773–3777 (2016)
25. Zhang, W., Qu, C., Ma, L., et al.: Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. *Pattern Recognit.* **59**(C), 176–187 (2016)
26. Zhou, W., Lei, J.S., Jiang, Q.P.: Blind binocular visual quality predictor using deep fusion network. *IEEE Trans. Comput. Imaging.* **6**, 883–893 (2020)
27. Zhang, L., Gu, Z., Li, H.: SDSP: a novel saliency detection method by combining simple priors. In: *Proceedings of IEEE International Conference on Image Processing*, pp. 171–175 (2014)
28. Achanta, R., Hemami, S., Estrada, F., et al.: Frequency-tuned salient region detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604 (2009)
29. Judd, T., Ehinger, K., Durand, F., et al.: Learning to predict where humans look. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2106–2113 (2010)
30. Wu, Y., Shen, X.: A unified approach to salient object detection via low rank matrix recovery. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 853–860 (2012)
31. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2432–2439 (2010)
32. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814 (2010)
33. Hinton, G.E., Srivastava, N., Krizhevsky, A., et al.: Improving neural networks by preventing coadaptation of feature detectors. *Computer Science* **3**(4), 212–223 (2012)
34. Kingma, D., Ba, J.: Adam: a method for stochastic optimization (2014). [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
35. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
36. Shao, F., Tian, W., Lin, W., et al.: Learning sparse representation for no-reference quality assessment of multiply distorted stereoscopic images. *IEEE Image Qual. Assess.: error Vis. Struct. Similarity Multimed.* **19**(8), 1821–1836 (2017)
37. Oh, H., Ahn, S., Kim, J., et al.: Blind deep S3D image quality evaluation via local to global feature aggregation. *IEEE Trans. Image Process.* **26**(10), 4923–4936 (2017)
38. Wang, X., Ma, L., Kwong, S., et al.: Quaternion representation based visual saliency for stereoscopic image quality assessment. *Signal Process.* **145**, 202–213 (2018)
39. Liu, T.-J., Lin, C.-T., Liu, H.-H., et al.: Blind stereoscopic image quality assessment based on hierarchical learning. *IEEE Access* **7**, 8058–8069 (2019)
40. Yang, J., Zhao, Y., Zhu, Y., et al.: Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network. *Inf. Sci.* **474**, 1–17 (2019)
41. Yang, J., Sim, K., Jiang, B., et al.: No-reference stereoscopic image quality assessment based on hue summation-difference mapping image and binocular joint mutual filtering. *Appl. Opt.* **57**(14), 3915–3926 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.