Registration-based Distortion and Binocular Representation for Blind Quality Assessment of Multiply-Distorted Stereoscopic Image

Yiqing Shi
College of Photonic and Electronic Engineering,
Fujian Normal University
Fuzhou, China
417shelly@gmail.com

Wenzhong Guo
College of Computer and Data Science,
Fuzhou University
Fuzhou, China
fzugwz@163.com

Yuzhen Niu*
College of Computer and Data Science,
Fuzhou University
Fuzhou, China
yuzhenniu@gmail.com

Yi Wu
College of Photonic and Electronic Engineering,
Fujian Normal University
Fuzhou, China
wuyi@fjnu.edu.cn

Abstract

Multiply-distorted stereoscopic images are common in real-world applications. The mixture of multiple distortions leads to complex binocular visual behavior of multiply-distorted stereoscopic images, so the existing blind singly-distorted stereoscopic image quality assessment (IQA) methods cannot obtain satisfactory results on multiply-distorted stereoscopic images. Because binocular rivalry caused by different distortions in the left and right views greatly influences the final stereoscopic image quality, we present a registration-based distortion and binocular representation for blind quality assessment of multiply-distorted stereoscopic image in this paper. We first use a registration-based distortion representation to represent the distortion in the stereoscopic image. Then we represent the binocular rivalry by merging the left and right views into a cyclopean image. Considering that the color and intensity of pixels in the RGB image can better reflect the information of the distorted image, then a grayscale cyclopean image is further converted to the color binocular representation through tone mapping. Finally, a multiply-distorted stereoscopic IQA method based on a double-stream convolutional neural network is proposed. The two subnetworks are used to extract quality features from the registration-based distortion representation and color binocular representation, respectively. Experimental results demonstrate that the proposed model outperforms the state-of-the-art models on the multiply-distorted stereoscopic image databases.

Keywords: ***Blind stereoscopic image quality assessment, multiply-distorted stereoscopic image, binocular rivalry, registration-based distortion representation, color binocular representation.***

1. Introduction

With the advent of the 5G era, visual information is changing from two-dimensional to stereoscopic. Users' demand for stereoscopic content with deep visual perception has driven the rapid development of stereoscopic display technology, especially in cinema and television. As an important part of the stereoscopic acquisition system, stereoscopic image quality assessment (SIQA) aims to determine whether the perceptual quality of stereoscopic images meets the requirements. Stereoscopic distorted images can be divided into singly-distorted stereoscopic image (SDSI) and multiply-distorted stereoscopic image (MDSI), where the quality of SDSI is only related to the perception of a certain distortion type, while MDSI is affected by the interaction between different types of distortion.

Although SIQA has attracted a lot of attention, only a small number of studies focus on MDSI [26, 10, 23, 32]. Actually, the stereoscopic images will undergo different stages of acquisition, compression, and transmission, the stereoscopic image may be contaminated with multiple types of distortion, and the left and right views of the stereoscopic image are also subjected to different degrees and types of distortion symmetrically or asymmetrically during the processing stage. It poses a great challenge to the binocular combination of stereoscopic vision [33, 36], and binocular rivalry [6] and other unpredictable visual behaviors [25] occur during the process. To this end, the mixture of multiple distortions causes the problem of binocular quality predic-

tion more complex and challenging.

Because of the scene discrepancy between the left and right views, in existing works [33, 34, 16, 35], the difference image between the left and right views of the distorted stereoscopic image is not a good representation of the distortion. To resolve the negative influence of the inaccurate distortion representation on SIQA, the monocular model based on the registration-based distortion representation is built to represent the distortion in the stereoscopic image more accurately.

Furthermore, the left and right views of an MDSI have been distorted by different types and degrees of distortion symmetrically or asymmetrically, it is necessary to consider the influence of image content information and binocular visual behavior on the stereoscopic image [10]. Since MDSI is more complicated than SDSI, the image presented to the human eye is strongly affected by binocular rivalry during the subjective assessment. Therefore, the binocular representation of MDSI is calculated to simulate the imaging of the stereoscopic image in the brain based on binocular rivalry, and then represent the image content information and the effect of binocular rivalry on image quality.

Compared with grayscale images, color images show the color and intensity of their pixels and can represent the information of distorted images more completely [20]. Hence, the grayscale cyclopean image is converted to the RGB image by tone mapping. After that, we design a double-stream convolutional neural network (CNN) model that learns from the registration-based distortion representation and color binocular representation, respectively.

In this paper, we proposed a blind/no-reference (NR) SIQA framework for MDSI, a double-stream CNN model designed to fuse the monocular and binocular features. The experimental results on the LIVE 3D [6, 18] and NBU-MDSID [26, 23] databases demonstrate the effectiveness of the proposed model for complex multiple distortion cases. The main contributions of this paper are summarized as follows.

(1) Based on the observation that the scene discrepancy causes the inaccurate distortion representation, a registration-based distortion representation, is proposed to better represent the distortion situation of the stereoscopic image.

(2) The color binocular representation which merges left and right views into a cyclopean view, is introduced to incorporate the influence of the binocular rivalry on stereoscopic imaging.

(3) A unified blind SIQA metric is proposed to evaluate both SDSI and MDSI, using a double-stream CNN architecture, which outperforms the state-of-the-art SIQA metrics.

2. Related Work

Based on the availability of the reference image, SIQA metrics can be divided into full-reference (FR), reduced-reference (RR), and blind/NR metrics. Most existing SIQA metrics for SDSI are of the FR or blind/NR type, while existing works for MDSI are limited and focus on the blind/NR type. We present related works on blind SIQA for SISD and MDSI in Subsection 2.1 and 2.2, respectively.

2.1. Blind SIQA for singly-distorted image

SDSI means that the information of the stereoscopic image is corrupted by a single distortion type, so that its quality is only related to the perception of the corresponding single distortion type. Blind SIQA metrics do not use any information from the original reference image; therefore, the application prospects of blind IQA are more practical than those of FR-IQA and RR-IQA metrics. Based on the information that they use, blind SIQA can be further classified into three categories: binocular perception-based, depth perception-based, and difference perception-based metrics.

Many binocular perception-based metrics have been proposed for improving the performance of SIQA metrics by incorporating binocular perception. Ryu and Sohn [21] proposed a blind SIQA index that measures the extents of blurriness and blockiness for the left and right views and then combines these using a binocular perception model. Shao *et al.* [25] developed a phase-tuned quality lookup and a visual codebook from the binocular energy responses to achieve blind quality prediction by pooling. Zhou *et al.* [36] presented two binocular combinations of stimuli to extract quality features, and then adapted the extreme learning machine to predict image quality. Shao *et al.* [27] proposed a domain transfer framework that the information from the source feature domain is transferred to its target quality domain by means of dictionary learning.

Depth perception-based metrics assess the image quality based on the disparity map or synthesized cyclopean (human brain) image. Akhter *et al.* [1] proposed a blind SIQA index that first extracts image features from a stereoscopic image and its disparity map, and then uses a logistic regression model to predict the image quality. Chen *et al.* [5] proposed combining 2D cues in a cyclopean view and 3D cues in disparity information to estimate the perceptual quality of stereoscopic images. Jiang *et al.* [11] proposed an index based on a deep non-negativity constrained sparse autoencoder with the input of the cyclopean image and the left and right views. Shen *et al.* [28] proposed a blind SIQA that simulates the perception route of human visual system, and derives features from the fused view and single view. Liu *et al.* [15] proposed a two-stream interactive network model to simulate the process of human stereo visual perception.

Difference perception-based metrics assess the image quality based on the difference between the left and right

views. The difference image was first used in the FR-SIQA index presented in [33], which uses it to represent information differences between two views. Zhang *et al.* [34] proposed a CNN-based blind SIQA index, which considers the difference image to represent the depth and distortion of the stereoscopic image. Shen *et al.* [29] proposed combining the spatial frequency information and statistic feature extracted from the cyclopean and difference maps to represent the binocular characteristic and asymmetric information. Shi *et al.* [30] computed a registered distortion representation based on the left and registered right views to represent the distortion in the stereoscopic image, then designed a three-column model that learns from the registered distortion representation and the left and right views.

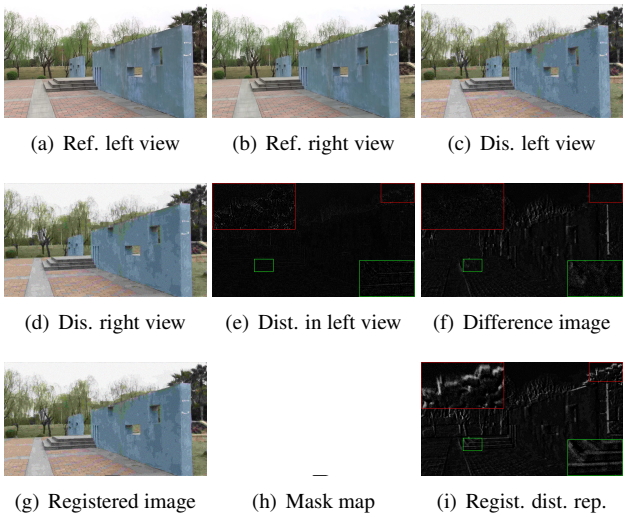
2.2. Blind SIQA for multiply-distorted image

The quality of SDSI is only related to the perception of the associated distortion type, while the interaction among different distortion types influences the perceived quality of MDSI. Due to some unpredictable visual behavior that may occur during image processing, various distortions are applied to the left and right views symmetrically or asymmetrically. Moreover, stereoscopic images are more likely to be polluted by multiple distortion types in the acquisition, processing, and transmission stages, thus bringing a greater challenge to research work on IQA.

Although stereoscopic images are prone to suffer from multiple distortions, there are few works on quality assessment for MDSI actually. Shao *et al.* [26] proposed a multi-modal joint sparse representation framework to learn a set of modality specific dictionaries, and then evaluates the quality of the image based on the reconstruction error to assess the image quality. Jiang *et al.* [10] presented a unified blind quality evaluator by learning monocular and binocular local visual primitives based on a task-driven and modality-specific sparse reconstruction errors. In the work [23], a multistage pooling model for asymmetric MDSI was proposed, which establishes a multimodal sparse representation framework for the phase and magnitude components and employs a multistage pooling strategy to simulate the pooling procedures. Wang *et al.* [32] proposed a sparse representation framework to learn the local and global quality perception functions and characterized the perceptual features of MDSI through five different channels.

3. Proposed method

The scene discrepancy causes the difference image between the left and right views to be inaccurate for distortion representation, especially in image regions where the depth changes. However, the problem of scene discrepancy between the left and right views of a stereoscopic image can be solved by image registration. Compared to the difference image, the registration-based distortion representation



(a) Ref. left view (b) Ref. right view (c) Dis. left view

(d) Dis. right view (e) Dist. in left view (f) Difference image

(g) Registered image (h) Mask map (i) Regist. dist. rep.

Figure 1. Example of proposed registration-based distortion representation.

can more accurately represent the distortion in the stereoscopic image, and has proven to be effective in the quality evaluation of SDSI [30].

Nevertheless, MDSI is more complex than SDSI, where the left and right views of the stereoscopic image are introduced to different types or degrees of distortion, and it is insufficient to only consider the distortion of the images, the effects of image content and binocular vision to the stereoscopic image also need to be taken into consideration. The imaging in the mind under the action of binocular vision can more accurately represent the image content information and binocular visual behavior than consider the content information of left and right views separately. To this end, the color binocular representation is proposed to simulate the actual imaging in the human brain based on binocular rivalry.

Motivated by [31], we design a double-stream CNN model which uses image patches from registration-based distortion representation and color binocular representation to train the monocular model and binocular model respectively, and the two subnetworks are used to extract feature information separately and then perform feature fusion, and finally mapping to obtain the image quality score. With the combination of distortion information and content information based on binocular rivalry, the proposed model can effectively predict the visual quality of MDSI.

3.1. Registration-based distortion representation

Existing works [34, 33, 16, 35, 29] use the difference image between the left and right views of a distorted stereoscopic image to represent the distortion, however, the scene discrepancy causes inaccurate distortion representation of it, especially in edge and contour regions with significan-

t depth changes. As indicated in Figure 1, the difference image (Figure 1(f)) between the left (Figure 1(c)) and right (Figure 1(d)) views is affected by the scene discrepancy and displays strong fake cues of distortion around the boundary of steps and twigs. Compared to the true distortion representation of the left view (the difference image between the reference left (Figure 1(a)) and distorted left (Figure 1(c)) views displayed in Figure 1(e)), there is no severe distortion around the boundary of steps and twigs. Therefore, we propose to first address scene discrepancy by image registration, and then compute a registration-based distortion representation to represent the distortion in the stereoscopic image more accurately.

In this paper, we first perform the image registration on the input left and right views by the SIFT flow algorithm [14]. The SIFT flow is a method of scene registration to its nearest image in a large image database containing various scenes according to the input images. Specifically, we register the right view I_r to the left view I_l of the stereoscopic image, distinguish the matching and no-matching regions according to the masking map (Figure 1(h)), and obtain the registered image, denoted as I_m. The proposed metric only uses the matched regions in the registered image because information in the no-matching regions is unavailable. As indicated in Figure 1(g), the pixels of the registered image are derived from the right view (Figure 1(d)) and the structure of the registered image is the same as that of the left view (Figure 1(c)). We compute the registration-based distortion representation as the difference of the registered image I_m and the left view I_l as follows,

$$I_d(x,y) = I_m^g(x,y) - I_l^g(x,y), \tag{1}$$

where (x,y) indicates the position of the pixel, I_m^g and I_l^g are the registered image and left view of the distorted stereoscopic image in grayscale, and I_d is the computed registration-based distortion representation.

Compared to the difference image displayed in Figure 1(f), the registration-based distortion representation displayed in Figure 1(i) is more similar to the distortion in the left view presented in Figure 1(e), especially in the edge and contour regions. Moreover, the left and right views of MDSI are distorted by different types or degrees of distortion symmetrically or asymmetrically. Although the registration-based distortion representation can better represent the distortion of the image than the difference image, the image registration cannot guarantee its validity due to the interaction of multiple distortion types, and the registration-based distortion representation of the MDSI (Figure 1(i)) still has some inaccurate distortion information around the boundary of steps and twigs compared with the actual distortion of the left view (Figure 1(e)).

3.2. Color binocular representation

The registration-based distortion representation can effectively represent the distortion between the left and right views on the SDSI, due to the interaction among different distortion types on MDSI, especially on asymmetrically stereoscopic distorted images, where the visual stimuli of the left and right views are prone to large differences. The registration-based distortion representation cannot ensure high accuracy for MDSI when only considering distortion information may have a negative impact on SIQA.

To address the effect of binocular rivalry on stereoscopic imaging, the left and right views are merged into a cyclopean view to simulate the actual imaging of the stereoscopic image in the brain. We calculate the cyclopean image based on the linear summation model [13] and Gabor filter [7], the obtained grayscale cyclopean image is then converted to an RGB image by tone mapping to further extract color and content information.

3.2.1 Cyclopean image synthesis

Binocular rivalry has a significant impact on the final imaging of MDSI, in which two distinct views compete for dominance, so that only one monocular input is visible and its contralateral input is suppressed. Specifically, the left and right views can be merged into a single cyclopean view to represent the results of binocular rivalry [6]. Since binocular perception has a tremendous impact on the visual perception of the stereoscopic image, synthesizing left and right views to simulate stereoscopic scene perception is the key to a successful SIQA model.

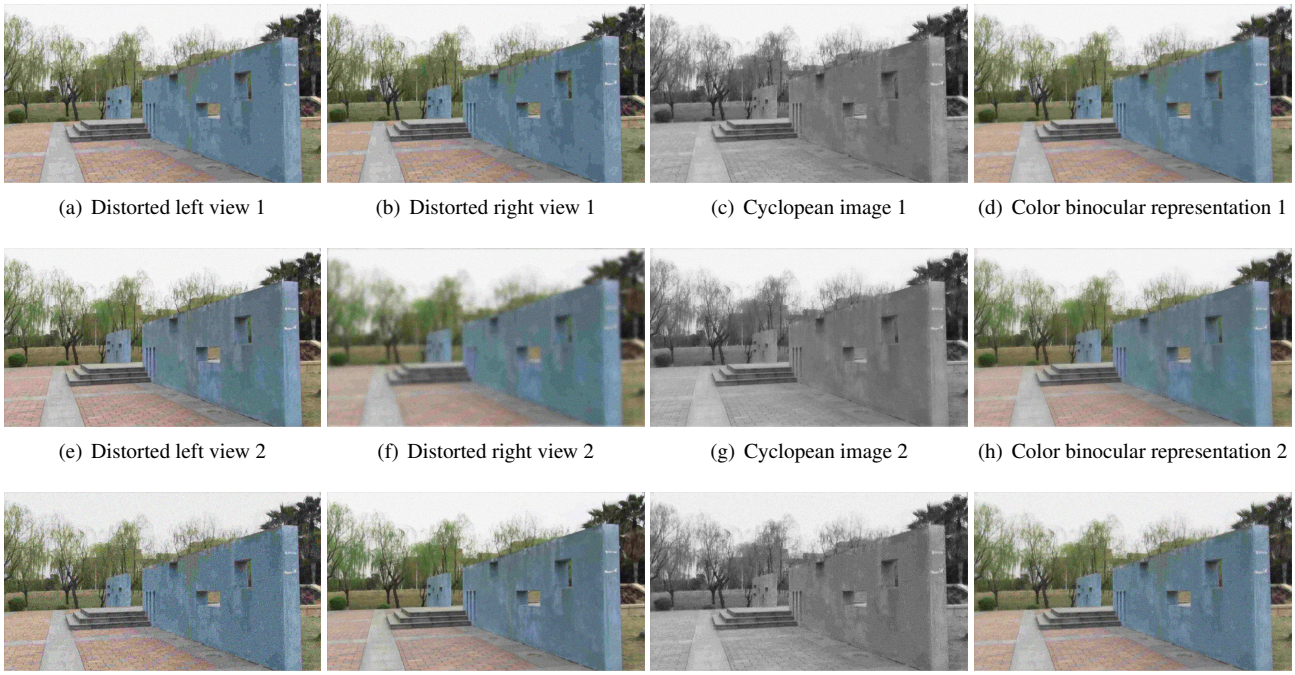
Inspired by the discovery of biological vision, the linear summation model [13] was proposed to explain the binocular combination process of visual information from left and right views. Although this model cannot fully characterize the complex binocular vision mechanism, it is regarded as a basic model for binocular vision because of its simplicity and reasonableness, and the model is as follows,

$$C = \omega_l E_l + \omega_r E_r, \tag{2}$$

where E_l and E_r denote the visual signals of the left and right views, respectively, ω_l and ω_r represent the corresponding weights of the two views, w_l and w_r satisfy $\omega_l + \omega_r = 1$. Eq. 2 can explain the experience of binocular rivalry in the perceiving cyclopean image when stereoscopic stimuli are presented. Since binocular rivalry is locally independent [8], the local linear model for synthesizing cyclopean image according to [6] is as follows,

$$\begin{aligned} I_c^g(x,y) =& \omega_l(x,y) I_l^g(x,y) \\ &+ \omega_r((x+d),y) I_r^g((x+d),y), \end{aligned} \tag{3}$$

where I_c^g is the synthetic grayscale cyclopean image, I_l^g and I_r^g are the lefe and right views in grayscale, and d is the



(a) Distorted left view 1 (b) Distorted right view 1 (c) Cyclopean image 1 (d) Color binocular representation 1

(e) Distorted left view 2 (f) Distorted right view 2 (g) Cyclopean image 2 (h) Color binocular representation 2

(i) Distorted left view 3 (j) Distorted right view 3 (k) Cyclopean image 3 (l) Color binocular representation 3

Figure 2. Examples of proposed color binocular representation. (Distortion degree for (a)-(d) is 131 in the left view and 131 in the right view; for (e)-(h) is 121 in the left view and 321 in the right view; for (i)-(l) is 132 in the left view and 111 in the right view. Corresponding distortion types are Gaussian blur, JPEG and Gaussian noise.)

disparity of the left view corresponding to the relevant pixel on the right view, ω_l and ω_r represent the corresponding weights of the two views, respectively.

Since the experience of binocular rivalry is independent of the absolute stimulus intensity of each view and is related to the relative stimulus intensity of two views, the local energy of the Gabor filter response is used to weight the left and right views stimuli [7]. The ω_l and ω_r can be obtained as follows,

$$\omega_l(x,y) = \frac{G_l(x,y)}{G_l(x+y) + G_r((x+d),y)}, \quad (4)$$

$$\omega_r((x+d),y) = \frac{G_r((x+d),y)}{G_l(x+y) + G_r((x+d),y)}, \quad (5)$$

where G_l and G_r represent the Gabor filter responses for the left and right views, respectively.

As mentioned in [4], when the left eye sees an undistorted image and the right eye sees a distorted image with blur, the eye that sees the undistorted image will dominate because blur reduces visual stimulation. Conversely, when the undistorted image is presented on the retina of one eye and the JPEG distorted image is presented on another eye, the eye that sees the JPEG distorted image will dominate because JPEG increases visual stimulation.

As shown in Figure 2, the left and right views are symmetrically distorted in the MDSI 1, the distortion in the grayscale cyclopean image 1 (Figure 2(c)) is also balanced, but the distorted details of the JPEG on the wall are less obvious. In the MDSI 2, the distortion degree of blur in the left view is lower than that of the right view. Therefore, the blur distortion in the grayscale cyclopean image 2 (Figure 2(g)) is less pronounced than in the right view, which is consistent with the principle that blur reduces visual stimulation. The JPEG distortion degree of the left view is higher than that of the right view in the MDSI 3. However, because JPEG distortion involves color mode conversion, the human eye is relatively insensitive to the JPEG distortion of the generated grayscale cyclopean image 3 (Figure 2(k)).

3.2.2 Tone mapping

Information such as distortion, brightness, and contrast can be more easily extracted from RGB images than from grayscale images, and the color and intensity of pixels in RGB images can reflect the original information of distorted images well. In order to further simulate the final imaging of the brain for the cyclopean image, we converted the grayscale cyclopean image to the RGB image. Movitated by the tone mapping method in [19, 2], assuming that the three channels of the color image are R, G, and B, and the grayscale map is Gs, the left and right views of the stereoscopic image are first converted into grayscale maps to ob-

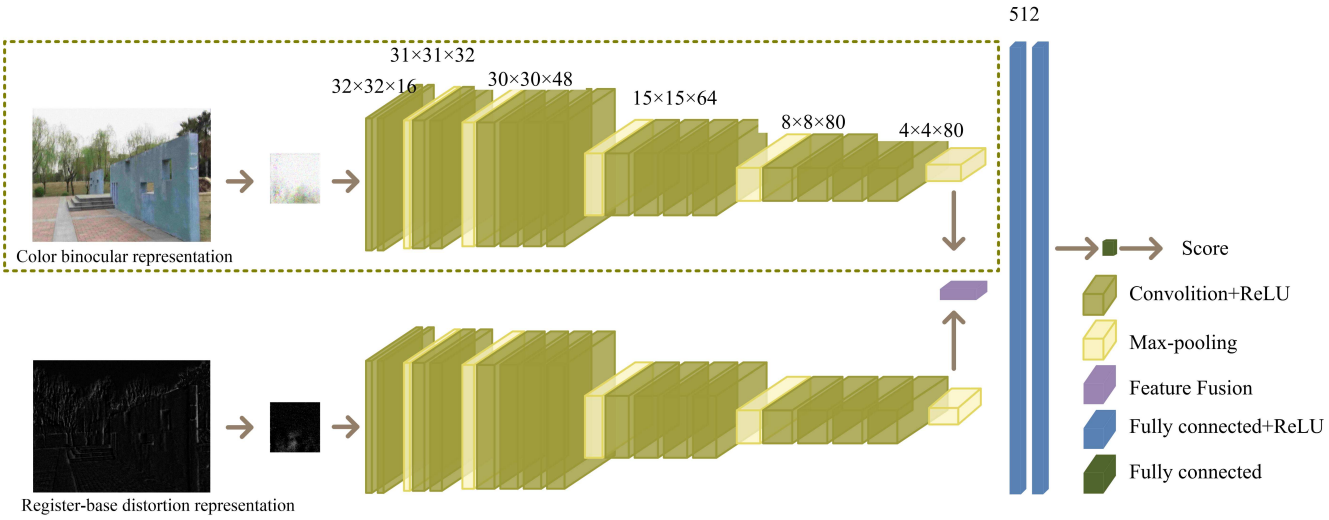


Figure 3. The architecture of proposed double-stream CNN model.

tain the three-dimensional scale coefficients of the left and right views, and then the RGB and grayscale scale coefficients are averaged as follows,

$$r_l = \frac{(R_l, G_l, B_l)}{(Gs_l, Gs_l, Gs_l)}, r_r = \frac{(R_r, G_r, B_r)}{(Gs_r, Gs_r, Gs_r)}, \quad (6)$$

The grayscale channel (Gs_c, Gs_c, Gs_c) of the grayscale cyclopean image is synthesized with the scale coefficients $r = (r_l + r_r)/2$ to obtain the color binocular representation I_c as follows,

$$I_c = (Gs_c, Gs_c, Gs_c) * r, \quad (7)$$

As shown in Figure 2, compared to the grayscale cyclopean image 1 (Figure 2(c)), the JPEG distortion detail on the wall of the color binocular representation 1 (Figure 2(d)) is more pronounced. Consistent with the grayscale cyclopean image 2 (Figure 2(g)), the blur distortion remains less obvious in the color binocular representation 2 (Figure 2(h)) than in the distorted right view 2 (Figure 2(f)). In the MDSI 3, the JPEG distortion degree is higher in the left view than in the right view, and the JPEG distortion is more visible in the color binocular representation 3 (Figure 2(l)) than in the distorted right view 3 (Figure 2(j)), in line with JPEG distortion increases the visual stimulus. In summary, color binocular representation can improve the sensitivity of distorted information and make stereoscopic imaging based on binocular competition more accurate.

3.3. Double-stream convolutional neural network model

Figure 3 illustrates a double-stream CNN architecture, which learns from registration-based distortion representation and color binocular representation. The double-stream CNN model uses five cascaded convolutional layers (consisting of 16 convolutional layers and 5 pooling layers) for feature extraction and two fully connected layers for regression. The two sub-networks extract the feature information respectively and fuse them as the input of the fully connected layer to achieve the final image quality evaluation.

The registration-based distortion representation and color binocular representation are both divided into a number of k×k image patches with overlaps to increase the scale of training data; the patches that have overlaps with no-matching regions are discarded. Meanwhile, two image patches from the same position with the same size of the registration-based distortion representation and color binocular representation are used as the inputs of the double-stream CNN model.

The proposed architecture of the double-stream CNN model has five cascaded convolutional layers and two fully connected layers. Each of the first two cascaded convolutional layers consists of the repeated application of two 3×3 convolutional layers, followed by a 2×2 max pooling operation with stride "1". Then, each of the following three cascaded convolutional layers consists of the repeated application of four 3×3 convolutional layers, followed by a 2×2 max pooling operation with stride "2" for downsampling. All the convolutional layers are applied with zero padding and stride "1"to obtain an output of equal size to the input. ReLU is used in all convolutional layers and the first two fully connected layers because ReLU can effectively reduce the likelihood of the gradient vanishing and accelerate the convergence of optimization [9].

During the training stage, we use the Euclidean distance as a loss function. The optimal weights of the proposed double-stream CNN model can be learned via adaptive moment estimation (Adam) [12] and back-propagation. The initial learning rate is set to be 10^{-4} and we reduce it every 20000 iterations by a gamma of 0.7. During the testing stage, the global image score is obtained by calculating the

average score of the patches belonging to the same image.

4. Experiment

In this section, we present the experimental results of the proposed model and the performance comparisons with some state-of-the-art SIQA metrics on four widely used 3D IQA databases.

4.1. Databases and performance indicators

In the experiments, we used four 3D IQA databases.

LIVE 3D IQA Database Phase-I [18] consists of 20 reference stereoscopic images and 365 distorted stereoscopic images, including 80 images for JPEG, JP2K, FF, and WN, and 45 images for BLUR. Each image in the database is symmetrically distorted on its left and right views.

LIVE 3D IQA Database Phase-II [6] consists of 8 reference images and 360 symmetrically or asymmetrically distorted stereoscopic images. The distortion types include BLUR, JPEG, JP2K, FF, and WN. For each distortion type, a reference image pair generates three symmetrically distorted images and six asymmetrically distorted images.

NBU-MDSID Phase-I [26] consists of 10 reference stereoscopic images, 270 MDSIs, and 90 SDSIs. MDSI is corrupted by JPEG, WN, and BLUR. Each image in the database is symmetrically distorted on its left and right views.

NBU-MDSID Phase-II [23] consists of 10 reference images and 300 asymmetrically MDSIs distorted by JPEG, WN, and BLUR. For each distorted image in the database, one or two types of distortion are applied asymmetrically on the left and right views.

In this paper, three widely used performance indicators are used to evaluate the performance of SIQA: 1) Spearmans Rank Order Correlation Coefficient (SRCC); 2) Pearsons Linear Correlation Coefficient (PLCC); 3) Root Mean Squared Error (RMSE). Greater PLCC and SROCC values indicate a closer relation with the human subjective evaluation, smaller RMSE values indicate superior correlation with human perception.

We report the median results obtained from train-test iterations of 20. Specifically, distorted images corresponding to 80% of the reference images were used as the training set and distorted images corresponding to the remaining 20% of the reference images were used as the testing set, such that there was no overlap between the training and testing sets. In the tables in this section, we use a symbol "-" to indicate that the performance value was not provided in the corresponding paper and we could not obtain the corresponding source code.

Table 1. Experimental results on NBU-MDSID Phase-I and NBU-MDSID Phase-II. Best performance values on each database are indicated in boldface.

Type	Metric	NBU-MDSID Phase-I			NBU-MDSID Phase-II		
		SRCC	PLCC	RMSE	SRCC	PLCC	RMSE
FR	Chen	0.877	0.885	4.385	0.749	0.763	7.560
	Bensalma	0.834	0.856	4.943	0.780	0.819	7.110
	Shao	0.905	0.919	3.687	**0.862**	0.802	7.212
NR	BLIINDS-II	0.919	0.921	3.543	0.746	0.763	7.763
	BRISQUE	0.889	0.910	3.967	0.750	0.766	7.723
	MUMBLIM	0.882	0.878	4.570	0.627	0.606	9.586
	MUSF	0.922	0.916	3.836	0.765	0.785	7.442
	Wang	0.936	0.940	3.804	0.819	0.845	7.020
	Shi	0.921	0.910	3.279	0.831	0.836	3.749
	Shen	-	-	-	-	-	-
	Proposed	**0.939**	**0.944**	**2.835**	0.861	**0.869**	**3.565**

4.2. Performance on multiply-distorted stereoscopic image databases

To validate the performance of the proposed model on MDSI, the proposed model was first performed on the NBU-MDSID Phase-I and NBU-MDSID Phase-II, which were compared with ten existing IQA metrics. Among the comparison IQA metrics, Chen [6], Bensalma [3] and Shao [24] are FR-IQA metrics; BLIINDS-II [22] and BRISQUE [17] were initially presented for 2D images. For these 2D IQA metrics, we first used them to compute the quality scores of the left and right views individually, and then averaged them as the 3D quality score of the stereoscopic image; MUMBLIM [26], MUSF [23], and Wang [32] are blind IQA metrics designed for MDSI; Shi [30] and Shen [28] are blind IQA metric designed for SDSI.

As indicated in Table 1, the proposed model achieved better performance than all comparison metrics in terms of all performance indicators except SRCC on the NBU-MDSID Phase-II, which ranked the proposed model inferior to the FR-IQA metrics presented by Shao [24]. From the results, there are following observations: 1) on symmetrically MDSI (NBU-MDSID Phase-I), Shao [24], BLIINDS-II [22], MUSF [23], Wang [32], Shi [30], and the proposed model achieved the competitive performance; 2) on asymmetrically MDSI (NBU-MDSID Phase-II), the performance of other metrics is significantly decreased except for Shao [24] and the proposed model. In summary, the proposed model outperformed most existing IQA metrics.

4.3. Performance on singly-distorted stereoscopic image databases

To verify the scalability of the proposed model and further evaluate the performance on SDSI, comparative experiments were conducted on two SDSI databases(LIVE 3D Phase-I database and LIVE 3D Phase-II database) with ten existing IQA metrics. The experimental results on the two databases are presented in Table 2 and Table 3, which include the experimental results of SRCC and PLCC values

Table 2. Experimental results on LIVE 3D Phase-I. Best performance values are indicated in boldface.

Type	Metric	JPEG		WN		BLUR		LIVE 3D Phase-I		
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	RMSE
FR	Chen	0.530	0.603	0.948	0.942	0.925	0.942	0.916	0.917	6.533
	Bensalma	0.328	0.380	0.906	0.915	0.916	0.937	0.875	0.887	7.559
	Shao	0.615	0.656	**0.943**	0.941	0.938	0.951	-	-	-
NR	BLIINDS-II	0.496	0.525	0.726	0.835	0.786	0.871	0.910	0.917	6.553
	BRISQUE	0.490	0.529	0.479	0.446	0.764	0.774	0.901	0.910	6.793
	MUMBLIM	0.693	0.703	0.899	0.896	0.853	0.862	0.885	0.8914	-
	MUSF	0.696	-	0.914	-	0.875	-	0.896	-	-
	Wang	0.633	0.762	0.920	0.951	0.903	0.958	0.868	0.938	-
	Shi	0.681	0.780	0.938	0.970	0.910	0.974	0.936	0.963	4.161
	Shen	**0.879**	**0.906**	0.921	0.947	**0.945**	**0.988**	**0.962**	**0.972**	-
	Proposed	0.755	0.874	**0.943**	**0.972**	0.908	0.985	0.938	0.962	**3.872**

Table 3. Experimental results on LIVE 3D Phase-II. Best performance values are indicated in boldface.

Type	Metric	JPEG		WN		BLUR		LIVE 3D Phase-I		
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	RMSE
FR	Chen	0.843	0.862	0.940	0.957	0.908	0.963	0.889	0.900	4.987
	Bensalma	0.846	0.858	0.939	0.944	0.884	0.908	0.751	0.770	7.204
	Shao	0.720	0.750	0.846	0.850	0.801	0.827	-	-	-
NR	BLIINDS-II	0.516	0.576	0.904	0.900	0.677	0.708	0.910	0.917	6.553
	BRISQUE	0.736	0.760	0.831	0.758	0.743	0.823	0.901	0.910	6.793
	MUMBLIM	0.622	0.583	0.803	0.824	0.713	0.755	0.805	0.784	-
	MUSF	0.653	-	0.836	-	0.733	-	0.875	-	-
	Wang	0.788	0.846	0.929	0.957	0.909	0.984	0.831	0.851	-
	Shi	0.945	0.967	**0.967**	0.972	0.933	0.991	0.948	**0.961**	2.675
	Shen	0.816	0.825	0.923	0.954	**0.951**	0.988	**0.951**	0.953	-
	Proposed	**0.947**	**0.975**	0.952	**0.978**	0.933	**0.993**	0.941	0.954	**2.492**

on individual distortion type (JPEG, WN, and BLYR) and SRCC, PLCC, and RMSE values for all SISDs. In Table 2 and Table 3, the metrics with the best performance are indicated in bold. From the experimental results, we can conclude that proposed model still has high performance on most individual distortion types and also performs well on all distorted stereoscopic images.

Compared with Shi [30] and Shen [28] designed for SDSI, the proposed model achieved comparative performance on SDSI and better performance on MDSI. In terms of computational complexity, the proposed model is more complex and time-consuming to train because it needs to generate color binocular representation as input. The frameworks in Shi [30] and Shen [28] are more efficient than the proposed model in terms of input image generation and model training efficiency. Therefore, for SDSI, the frameworks in Shi [30] and Shen [28] can maintain the best balance between training efficiency and performance. For MDSI, although the proposed model has a higher computational and training complexity, its performance is greatly improved compared with them. In summary, the proposed model is more suitable for quality evaluation task of MDSI.

4.4. Validation on color binocular representation

In this subsection, we evaluate the effectiveness of the proposed color binocular representation. As described Subsection 3.2, the obtained color binocular representation is initially in grayscale, and we convert the grayscale cyclopean image into a RGB image by the tone mapping. To validate the effectiveness of the color binocular representation, We replaced the patch from the color binocular representation with a patch from the grayscale cyclopean image and obtained a variation of the double-stream CNN model, denoted by Gray-cy. The experimental results on four databases are presented in Table 4. The performance of the double-stream CNN model with color binocular representation is better than that of the variation with the grayscale cyclopean image. We can conclude that the proposed color binocular representation can better represent the results of binocular rivalry and the content information of stereoscopic image than the grayscale cyclopean image.

Table 4. Experimental results on different binocular representations on four databases. Best performance values across all models are indicated in boldface.

		SRCC	PLCC	RMSE
LIVE 3D IQA Phase-I	Gray-cy	0.928	0.955	4.195
	Left-right	0.929	0.956	4.285
	Proposed	**0.938**	**0.962**	**3.872**
LIVE 3D IQA Phase-II	Gray-cy	0.933	0.951	2.768
	Left-rightk	0.930	0.953	2.754
	Proposed	**0.941**	**0.954**	**2.492**
NBU-MDSID Phase-I	Gray-cy	0.932	0.939	2.899
	Left-right	0.921	0.910	3.279
	Proposed	**0.938**	**0.944**	**2.835**
NBU-MDSID Phase-II	Gray-cy	0.852	0.850	3.867
	Left-right	0.831	0.836	3.749
	Proposed	**0.861**	**0.869**	**3.565**

As described Subsection 3.3, the proposed model uses three image patches from the same position of the registration-based distortion representation and color binocular representation as inputs. The reason for using color binocular representation is that it can synthesize the stereoscopic image in the mind more accurately by considering binocular rivalry, however, the left and right views of the stereoscopic image also represent the content information. To verify the effectiveness of color binocular representation, we extended the double-stream CNN into a three-channel structure, useing three image patches from the same position of the registration-based distortion representation, and the left and right views as inputs, denoted as Left-right. As shown in Table 4, the performance of the proposed double-stream CNN network is better than that of the three-channel structure, especially on MDSI. The proposed double-stream CNN model that uses the color binocular representation as input indicates significant superiority. Therefore, it can be concluded that MDSI is more affected by binocular rivalry than SDSI, and the color binocular representation considers the influence of binocular rivalry on imaging, which effectively represents the actual imaging of the stereoscopic image in the brain and also represents the content information of the image.

4.5. Effects of patch size

In this subsection, we investigate e the extent to which the patch size affects the performance of the proposed double-stream CNN model. As indicated in Table 5, three different patch sizes (32×32, 48×48, and 64×64) were

Table 5. SRCC, PLCC, RMSE, and parameters for different patch sizes on NBU-MDSID Phase-I. Best performance values across all sizes are indicated in boldface.

SIZE	32×32	48×48	64×64
SRCC	**0.939**	0.934	0.938
PLCC	**0.944**	0.932	0.937
RMSE	**2.835**	3.107	3.088
Params.($\times 10^5$)	**97.5**	221.9	375.6

compared using the NBU-MDSID Phase-I to observe the performance change of the proposed model. As the patch size increased from 32×32 to 48×48, the performance decreased slightly. Then the performance increased gradually with the patch size increased from 48×48 to 64×64. The differences in terms of SRCC, PLCC, and RMSE are less than 0.005, 0.012, and 0.272, respectively, the patch size of 32×32 achieved the best performance. Because an image patch of 32×32 not only includes local information but also retains global structures for quality assessment and the training parameters are only 26% of those of 64×64. Above all, based on the consideration of training complexity and performance, the patch size of 32×32 was selected as the default patch size in all the experiments for the proposed double-stream CNN model.

5. Conclusion

In this paper, we presented a registration-based distortion and binocular representation for blind quality assessment of MDSI. Since the left and right views of MDSI are symmetrically or asymmetrically imposed with different types and degrees of distortion, the imaging inside human eye is influenced by binocular rivalry. The registration-based distortion representation was computed to represent the distortion in the stereoscopic image. Then we merged the left and right views into a cyclopean image to present the binocular rivalry, and further converted it to the color binocular representation through tone mapping. Finally, a double-stream CNN model was used to predict image quality of MDSI, and two subnetworks extracted quality features from the registration-based distortion representation and color binocular representation, respectively. The experimental results demonstrate the superiority of the proposed model over the state-of-the-art SIQA metrics.

Acknowledgement

This research was supported in part by the National Key Research and Development Plan of China under Grant 2021YFB3600503 and the in part by Natural Science Foundation of Fujian Province under Grant No. 2022J05043.

References

[1] R. Akhter, Z. P. Sazzad, Y. Horita, and J. Baltes. No-reference stereoscopic image quality assessment. In *SPIE - International Society for Optics and Photonics*, volume 7524, pages 1–12, 2010. 2

[2] M. Aubry, S. Paris, S. W. Hasinoff, J. Kautz, and F. Durand. Fast and robust pyramid-based image processing. *MIT-CSAIL-TR-2011-049.*, 2011. 5

[3] R. Bensalma and M.-C. Larabi. A perceptual metric for stereoscopic image quality assessment based on the binocular energy. *Multidimensional Systems and Signal Processing*, 24(2):281–316, 2013. 7

[4] M.-J. Chen, A. C. Bovik, and L. K. Cormack. Study on distortion conspicuity in stereoscopically viewed 3d images. In *2011 IEEE 10th IVMSP Workshop: Perception and Visual Signal Analysis*, pages 24–29. IEEE, 2011. 5

[5] M.-J. Chen, L. K. Cormack, and A. C. Bovik. No-reference quality assessment of natural stereopairs. *IEEE Transactions on Image Processing*, 22(9):3379–3391, 2013. 2

[6] M. J. Chen, C. C. Su, D. K. Kwon, L. K. Cormack, and A. C. Bovik. Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Processing: Image Communication*, 28(9):1143–1155, 2013. 1, 2, 4, 7

[7] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987. 4, 5

[8] D. J. Field, A. Hayes, and R. F. Hess. Contour integration by the human visual system: evidence for a local association field. *Vision research*, 33(2):173–193, 1993. 4

[9] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 6

[10] Q. Jiang, F. Shao, W. Gao, Z. Chen, G. Jiang, and Y.-S. Ho. Unified no-reference quality assessment of singly and multiply distorted stereoscopic images. *IEEE Transactions on Image Processing*, 28(4):1866–1881, 2018. 1, 2, 3

[11] Q. Jiang, F. Shao, W. Lin, and G. Jiang. Learning a referenceless stereopair quality engine with deep nonnegativity constrained sparse autoencoder. *Pattern Recognition*, 76:242–255, 2018. 2

[12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[13] W. J. Levelt. *On binocular rivalry*. PhD thesis, Van Gorcum Assen, 1965. 4

[14] C. Liu, J. Yuen, and A. Torralba. SIFT flow: dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 4

[15] Y. Liu, B. Huang, G. Yue, J. Wu, X. Wang, and Z. Zheng. Two-stream interactive network based on local and global information for no-reference stereoscopic image quality assessment. *Journal of Visual Communication and Image Representation*, 87:103586, 2022. 2

[16] L. Ma, X. Wang, Q. Liu, and K. N. Ngan. Reorganized DCT-based image representation for reduced reference stereoscopic image quality assessment. *Neurocomputing*, 215:21–31, 2016. 2, 3

[17] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 7

[18] A. K. Moorthy, C. C. Su, A. Mittal, and A. C. Bovik. Subjective evaluation of stereoscopic image quality. *Signal Processing: Image Communication*, 28(8):870–883, 2013. 2, 7

[19] S. Paris, S. W. Hasinoff, and J. Kautz. Local laplacian filters: edge-aware image processing with a laplacian pyramid. *ACM Transactions on Graphics*, 30(4):68, 2011. 5

[20] E. Reinhard, E. A. Khan, A. O. Akyuz, and G. Johnson. *Color imaging: fundamentals and applications*. CRC Press, 2008. 2

[21] S. Ryu and K. Sohn. No-reference quality assessment for stereoscopic images based on binocular quality perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(4):591–602, 2013. 2

[22] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 7

[23] F. Shao, Y. Gao, Q. Jiang, G. Jiang, and Y.-S. Ho. Multistage pooling for blind quality prediction of asymmetric multiply-distorted stereoscopic images. *IEEE Transactions on Multimedia*, 20(10):2605–2619, 2018. 1, 2, 3, 7

[24] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan. Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics. *IEEE Transactions on Image Processing*, 22(5):1940–1953, 2013. 7

[25] F. Shao, W. Lin, S. Wang, G. Jiang, and M. Yu. Blind image quality assessment for stereoscopic images using binocular guided quality lookup and visual codebook. *IEEE Transactions on Broadcasting*, 61(2):154–165, 2015. 1, 2

[26] F. Shao, W. Tian, W. Lin, G. Jiang, and Q. Dai. Learning sparse representation for no-reference quality assessment of multiply distorted stereoscopic images. *IEEE Transactions on Multimedia*, 19(8):1821–1836, 2017. 1, 2, 3, 7

[27] F. Shao, Z. Zhang, Q. Jiang, W. Lin, and G. Jiang. Towards domain transfer for no-reference quality prediction of asymmetrically distorted stereoscopic images. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(3):573–585, 2016. 2

[28] L. Shen, X. Chen, Z. Pan, K. Fan, F. Li, and J. Lei. No-reference stereoscopic image quality assessment based on global and local content characteristics. *Neurocomputing*, 424:132–142, 2021. 2, 7, 8

[29] L. Shen, J. Lei, and C. Hou. No-reference stereoscopic 3D image quality assessment via combined model. *Multimedia Tools and Applications*, 77(7):8195–8212, 2018. 3

[30] Y. Shi, W. Guo, Y. Niu, and J. Zhan. No-reference stereoscopic image quality assessment using a multi-task cnn and registered distortion representation. *Pattern Recognition*, 100:107168, 2020. 3, 7, 8

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[32] X. Wang, M. Qi, F. Shao, Q. Jiang, and X. Meng. Blind quality assessment for multiply distorted stereoscopic images towards iot-based 3d capture systems. *Journal of Visual Communication and Image Representation*, 71:102868, 2020. 1, 3, 7

[33] J. Yang, C. Hou, Y. Zhou, Z. Zhang, and J. Guo. Objective quality assessment method of stereo images. In *2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*, pages 1–4. IEEE, 2009. 1, 2, 3

[34] W. Zhang, L. Ma, L. Ma, J. Guan, and R. Huang. Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. *Pattern Recognition*, 59:176–187, 2016. 2, 3

[35] W. Zhou, G. Jiang, M. Yu, Z. Wang, Z. Peng, and F. Shao. Reduced reference stereoscopic image quality assessment using digital watermarking. *Computers & Electrical Engineering*, 40(8):104–116, 2014. 2, 3

[36] W. Zhou, L. Yu, Y. Zhou, W. Qiu, M. W. Wu, and T. Luo. Blind quality estimator for 3D images based on binocular combination and extreme learning machine. *Pattern Recognition*, 71:207–217, 2017. 1, 2