Dual-Stream Interactive Networks for No-Reference Stereoscopic Image Quality Assessment

Wei Zhou, Zhibo Chen, *Senior Member, IEEE*, and Weiping Li, *Fellow, IEEE*

***Abstract*—The goal of objective stereoscopic image quality assessment (SIQA) is to predict the human perceptual quality of stereoscopic/3D images automatically and accurately. Compared with traditional 2D image quality assessment (2D IQA), the quality assessment of stereoscopic images is more challenging because of complex binocular vision mechanisms and multiple quality dimensions. In this paper, inspired by the hierarchical dual-stream interactive nature of the human visual system (HVS), we propose a Stereoscopic Image Quality Assessment Network (StereoQA-Net) for No-Reference stereoscopic image quality assessment (NR-SIQA). The proposed StereoQA-Net is an end-to-end dual-stream interactive network containing left and right view sub-networks, where the interaction of the two sub-networks exists in multiple layers. We evaluate our method on the LIVE stereoscopic image quality databases. Experimental results show that our proposed StereoQA-Net outperforms state-of-the-art algorithms on both symmetrically and asymmetrically distorted stereoscopic image pairs of various distortion types. And in a more general case, the proposed StereoQA-Net can effectively predict the perceptual quality of local regions. In addition, cross-dataset experiments also demonstrate the generalization ability of our algorithm.**

***Index Terms*—Stereoscopic image quality assessment, dual-stream, interactive network, human vision, end-to-end prediction.**

I. Introduction

SINCE Stereoscopic/3D media data undergo diverse quality degradations during various processing stages, predicting the perceptual quality of stereoscopic contents objectively is important [1]–[3]. Stereoscopic image quality assessment (SIQA) is different from 2D image quality assessment (2D IQA) due to the extended depth perception dimension and binocular vision mechanisms between left and right views. Moreover, the artifacts of stereoscopic images consist of two categories, namely symmetric distortion and asymmetric distortion. The symmetrically distorted stereoscopic image pairs have the same distortion in both left and right view images, while the asymmetrically distorted stereoscopic left and right view images have different degrees of distortion. Therefore, how to effectively evaluate the human perceptual quality of stereoscopic images, especially those with asymmetric distortions, still remains a challenging research problem.

In general, stereoscopic visual quality assessment is a kind of artifact measurement in distorted image pairs. When reference contents are accessible, some full-reference stereoscopic image quality assessment (FR-SIQA) algorithms have been proposed [4]–[8]. Early methods for FR-SIQA directly stemmed from metrics for 2D IQA. Several 2D IQA approaches were applied to left and right views separately, and depth information was then integrated to provide the ultimate 3D image quality assessment [4], [5]. Afterward, more sophisticated algorithms were proposed by incorporating the binocular vision properties of the human vision system (HVS) into 2D IQA metrics. Typically, psychological vision findings such as contrast masking effect [6], cyclopean image [7], and binocular integration behaviors [8] have been employed to develop various computational models of perceptual 3D image quality prediction.

However, since original images are not always available in most practical situations, it is increasingly required to develop no-reference stereoscopic image quality assessment (NR-SIQA) methods. These metrics exploit the discriminative features of distorted 3D images to assess the perceptual quality. Conventionally, a number of NR-SIQA methods [9]–[13] manually extract some hand-crafted features based on the HVS characteristics, natural scene statistics (NSS), etc. The extracted hand-crafted features are then fed into a regression learning model such as support vector regression (SVR) [14] to predict the perceptual quality of stereoscopic images. These NR-SIQA methods can be further divided into two categories including distortion-specific NR-SIQA [15] and general-purpose NR-SIQA [16]. Nevertheless, it is not robust enough to represent stereoscopic image distortions by using hand-crafted features in learning the regression model according to the pre-defined HVS and NSS models. Therefore, it is difficult to predict the perceptual quality of stereoscopic images accurately and to generalize these models to practical NR-SIQA scenarios.

Based on these observations, we explore using an end-to-end dual-stream interactive deep neural network (DNN) with multi-layer network interaction to predict stereoscopic image quality. Recently, deep learning techniques have been widely used and achieved great success in solving various image processing and computer vision problems [17]. Except for the image classification framework, the remarkable ability of DNN to learn discriminative features provides a promising method for addressing the NR-SIQA task. One of the advantages of applying DNN is that it can directly take raw image data as input and then combine feature learning with quality regression in the training process. In addition, the DNN can be more effectively and robustly trained with the specific domain knowledge. Therefore, this work focuses on the most challenging general-purpose NR-SIQA, which evaluates stereoscopic

The authors are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei, Anhui, 230027, China (e-mail: weichou@mail.ustc.edu.cn; chenzhibo@ustc.edu.cn; wpli@ustc.edu.cn).

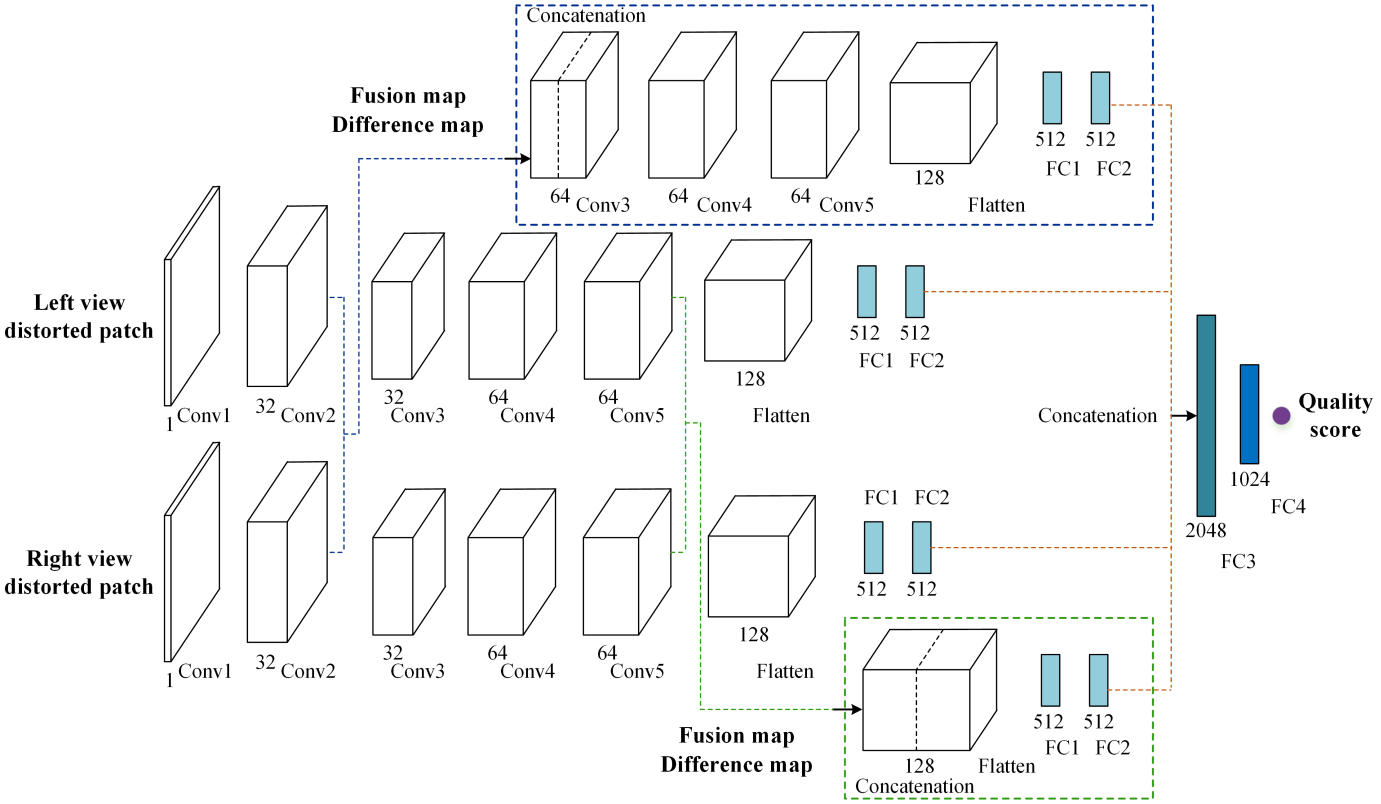


Fig. 1: The architecture of our proposed StereoQA-Net. The model takes as input left and right view distorted image patches, and conducts network interaction in multiple layers. The feature maps are subsampled by max pooling. After flattening convolutional layers and concatenating fully connected layers, the predicted quality score is regressed as a scalar output.

image quality without requiring reference data and knowing the distortion types. Specifically, this paper presents an end-to-end dual-stream interactive network called Stereoscopic Image Quality Assessment Network (StereoQA-Net) that can predict the perceptual quality of stereoscopic images effectively. The human visual cortex is a hierarchical structure with reciprocal cortico-cortical connections among the constituent cortical areas, which includes the low-level visual area, namely primary visual cortex (V1), and high-level visual areas from V2 to V5 [18]. Inspired by the HVS, our proposed StereoQA-Net involves multi-layer network interaction. In our architecture, except for the concatenation of a fully connected layer, the proposed StereoQA-Net integrates left and right view primary sub-networks at the convolutional layers by summation and subtraction of the corresponding feature maps for distorted patch pairs, in accord with the fusion and disparity information in the HVS.

Previous research works have attempted to exploit deep neural networks for no-reference image quality assessment (NR-IQA). Different from other tasks, image quality databases generally lack large-scale training images with subjective quality scores. In addition to the insufficient training data, existing data augmentation and image preprocessing techniques are not suitable for NR-IQA [19]. According to the training strategies, deep learning approaches on NR-IQA can be classified into two categories, namely patch-wise training and image-wise training.

The patch-wise training strategy partitions raw images into patches, and then predicts the perceptual quality of each patch by DNN regression learning. Kang *et al.* [20] applied convolutional neural network (CNN) to NR-IQA. They proposed a new shallow network architecture containing a single convolutional layer with a pooling layer, which extracts discriminative quality-related features from image patches. The whole image quality is then obtained by averaging the predicted quality of image patches. Li *et al.* [21] developed a general-purpose NR-IQA algorithm by using the shearlet transform [22]–[24] to extract the primary features of image patches, and then evolving the features through stacked auto-encoders [25]. They also utilized a Network in Network (NIN) model [26] pre-trained on ImageNet. The last several layers of the model are modified and the image patch qualities are then regressed through fine-tuning.

In contrast, the image-wise training strategy obtains the perceptual image quality by aggregating and pooling patch features or the predicted score of each patch. Bosse *et al.* [27], [28] used two training methods for NR-IQA. One of the training methods is the patch-wise training which is similar to [20]. The other takes both patch and image into account, i.e. adding a patch weighted average aggregation layer to learn the importance of each image patch, and then optimizing the loss based on the patch and the image jointly. Hou *et al.* [29] developed a classification-based deep model to learn qualitative quality grades. The newly designed quality pooling

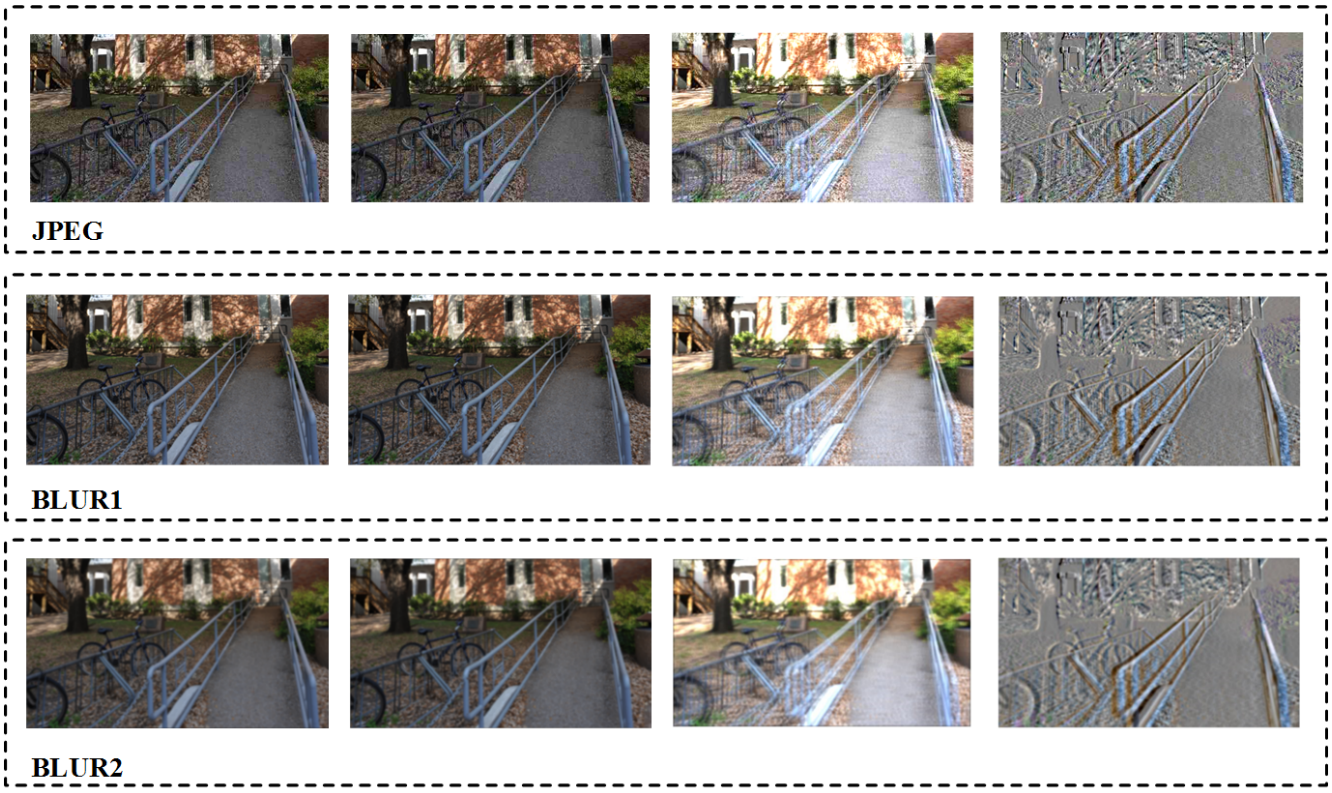


Fig. 2: Examples of distorted stereoscopic images with different distortion types/levels and the corresponding fusion as well as difference maps. JPEG distorted stereoscopic image (top), low-degree BLUR (i.e. BLUR1) distorted stereoscopic image (middle) and high-degree BLUR (i.e. BLUR2) distorted stereoscopic image (bottom). The first to the last columns show left view images, right view images, fusion maps and difference maps, respectively.

approach is then applied to convert the qualitative grades to numerical quality scores. Kim *et al.* [30] presented a two-stage NR-IQA model based on CNN. The model first generates local quality scores as proxy patch targets. The feature vectors obtained from image patches are then aggregated by statistical moments and regressed onto subjective image quality scores. Ma *et al.* [31] proposed a multi-task learning framework by decomposing the NR-IQA task into two subtasks with dependent loss functions. However, these deep learning methods for NR-IQA are not appropriate for assessing the perceptual quality of stereoscopic images due to the complex binocular vision mechanisms in 3D vision.

In addition, several NR-SIQA methods using CNN and deep belief network (DBN) have been proposed. For example, Oh *et al.* [32] presented a deep no-reference stereoscopic image quality evaluator by extracting local abstractions, and then aggregating these local representations into global features. As for DBN-based methods, Yang *et al.* [33] considered the deep perception map and binocular weight model with DBN to predict stereoscopic image quality.

In this paper, we propose a generic network architecture called Stereoscopic Image Quality Assessment Network (StereoQA-Net), which is an end-to-end dual-stream interactive network for NR-SIQA. Additionally, to the best of our knowledge, the proposed StereoQA-Net is the first study of applying the dual-stream network architecture to the challenging NR-SIQA task. Our StereoQA-Net is inspired by the recent works [34], [35] and the human visual cortex responses to stereoscopic visual signals [36]–[39]. Specifically, during the 3D visual stimuli processing, binocular fusion and disparity responses are primitively formed in the V1 cortical area which refers to a low-level visual area. Moreover, the visual signals from the binocular summation and subtraction channels are multiplexed, and then each neuron in V1 receives a weighted sum of the visual stimuli from these two channels [36]. Then, the output of V2 visual area is used for the processing of two streams, namely the dorsal stream and the ventral stream. It is generally assumed that the dorsal stream focuses on the coarse stereopsis while the ventral stream manages the fine stereopsis [37]. Further, the binocular fusion and disparity are enhanced through the high-level cortical areas from V2 to V5 [37]–[39]. Thus the neuron responses to the binocular fusion as well as the disparity present in both low-level and high-level visual areas. In other words, the interaction of left and right views goes through the whole hierarchical human visual cortex. Therefore, inspired by the dual-stream interaction mechanism of the HVS, our proposed StereoQA-Net is an end-to-end dual-stream network involving multi-layer network interaction between left and right view sub-networks. The software release of StereoQA-Net is available online: http://staff.ustc.edu.cn/~chenzhibo/resources.html for public research usage.

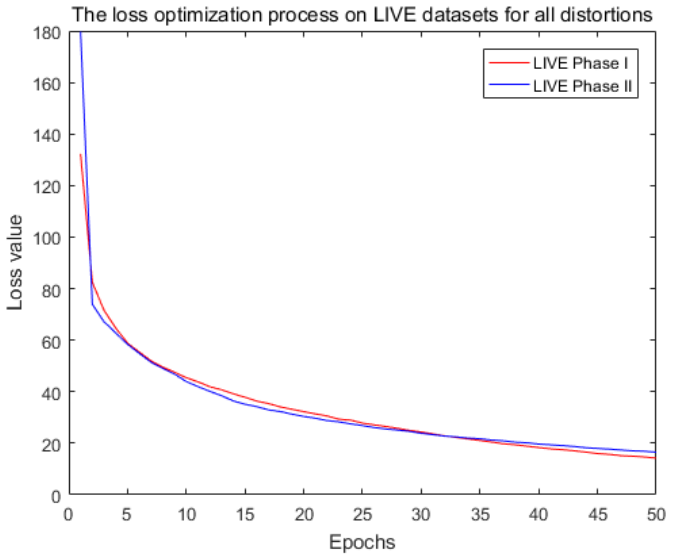


Fig. 3: Loss optimization process of the first 50 epochs for the training set on the two LIVE SIQA datasets.

The main contribution of the proposed method is the end-to-end dual-stream interactive network for no-reference stereoscopic image quality assessment, which contains dense interaction in multiple layers inspired by the hierarchical dual-stream interaction of the human visual system. The results show that:

- The visual quality predictions by the proposed network are highly correlated with subjective quality judgments for both symmetrically and asymmetrically distorted image pairs of various distortion types.
- The proposed method can effectively estimate the perceptual visual quality of local regions and has a promising generalization ability.
- For perceptual quality prediction, the experimental results demonstrate that end-to-end deep learning related features are more effective than traditional features.
- Different from conventional SIQA, the proposed method saves the complex computation of disparity map, which is verified to have lower time complexity.

The remainder of this paper is organized as follows. Section II introduces our proposed Stereoscopic Image Quality Assessment Network (StereoQA-Net) for NR-SIQA in details. In Section III, we present the experimental results and analysis. We then conclude the paper in Section IV.

II. Proposed StereoQA-Net

Our proposed framework of using DNN for no-reference stereoscopic image quality estimation is presented as follows. We name the proposed end-to-end dual-stream interactive network as StereoQA-Net. Given an RGB distorted stereoscopic image containing left and right view images, we first sample multiple patches from both left and right view image pairs. We then use the StereoQA-Net to predict the perceptual quality score of each input patch pair for distorted stereoscopic images. After that, the average local quality pooling method between left and right views is conducted to obtain a quality estimation for the whole stereoscopic image.

A. Architecture

Fig. 1 illustrates the architecture of our proposed network. The proposed StereoQA-Net consists of two primary sub-network streams that represent left and right view paths. Inspired by the existence of the binocular fusion and disparity between left and right views in both low-level and high-level human visual areas [36]–[39], the concatenations between these two primary sub-networks exist at different layers. Additionally, the inputs of the proposed StereoQA-Net are multiple 32×32 patch pairs which are sampled from distorted left and right view images.

As shown in Fig. 1, each primary sub-network includes five convolutional layers and two fully connected layers, i.e. $Conv1 - Conv5$ and $FC1 - FC2$. The identical structure of each primary sub-network is $32 \times 32 - 16 \times 16 \times 32 - 8 \times 8 \times 32 - 8 \times 8 \times 64 - 8 \times 8 \times 64 - 4 \times 4 \times 128 - 512 - 512$. The first and the second layers are convolutional layers that filter the image patches with 32 kernels. Then, the third and the fourth layers are convolutional layers with 64 kernels. Besides, the fifth layer is a convolutional layer with 128 kernels. Finally, two fully connected layers of both 512 nodes come after the flattening operation. Deep convolutional networks with each kernel size of 3×3 and a stride of 1 pixel are used for the model inspired by the recent work [40], where padding is applied to maintain the patch size unchanged. This small receptive field for convolutional layers can effectively capture the notion of five orientations containing up, down, left, right, and center. Moreover, the max pooling layers are used for subsampling the image patches that come after $Conv1$, $Conv2$ and $Conv5$ layers. Therefore, the final output feature maps of these convolutional and max pooling layers are 1/8 of the original input image patches.

In addition, the concatenations between these two sub-networks are performed after $Conv2$, $Conv5$ and $FC2$. For the concatenations of $Conv2$ and $Conv5$, the fusion and difference maps of the corresponding feature maps for distorted left and right patch pairs are concatenated after the convolutional layers. Further, each $FC2$ is concatenated, and then the fully connected structure $2048 - 1024 - 1$ is adopted to conduct quality regression for each input patch pair. The final layer of our proposed StereoQA-Net is one dimensional scalar output that provides the perceptual quality score.

B. Stereoscopic Image Preprocessing

The distorted stereoscopic images are simply partitioned into multiple $m \times n$ image patches for both left and right view images before they are fed into the proposed StereoQA-Net. Let I_l be a distorted left view image, and I_r be a distorted right view image. The partitioned image patch number is then given by:

$$p = \left\lfloor \frac{M}{m} \right\rfloor \times \left\lfloor \frac{N}{n} \right\rfloor, \tag{1}$$

where $M \times N$ is the resolution of each view ($M > m$, $N > n$), and $m = n = 32$ for the experiments.

TABLE I: SROCC, PLCC, AND RMSE COMPARISON ON THE TWO LIVE SIQA DATABASES. FR (NR) INDICATES FULL-REFERENCE (NO-REFERENCE) MODELS, AND THE BEST EXPERIMENTAL RESULTS ARE IN BOLD

Type	Metrics	LIVE Phase I			LIVE Phase II			Weighted Avg.		
		SROCC	PLCC	RMSE	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
FR	Gorley [6]	0.142	0.451	14.635	0.146	0.515	9.675	0.144	0.483	12.172
	You [5]	0.878	0.881	7.746	0.786	0.800	6.772	0.832	0.841	7.262
	Benoit [4]	0.899	0.902	7.061	0.728	0.748	7.490	0.814	0.826	7.274
	Lin [8]	0.856	0.784	-	0.638	0.642	-	0.748	0.714	-
	Chen (MS-SSIM) [7]	0.916	0.917	6.533	0.889	0.900	4.987	0.903	0.909	5.765
NR	Akhter [9]	0.383	0.626	14.827	0.543	0.568	9.294	0.463	0.597	12.079
	Sazzad [15]	0.624	0.624	-	0.648	0.669	-	0.636	0.646	-
	Chen [16]	0.891	0.895	7.247	0.880	0.895	5.102	0.886	0.895	6.182
	S3D-BLINQ [41]	-	-	-	0.905	0.913	4.657	-	-	-
	CNN [20]	0.896	0.933	5.948	0.633	0.634	8.632	0.765	0.785	7.281
	DNR-S3DIQE [32]	0.935	0.943	-	0.871	0.863	-	0.903	0.903	-
	DBN [33]	0.944	0.956	4.917	0.921	0.934	4.005	0.933	0.945	4.464
	FC-1024 + SVR	0.935	0.953	4.711	0.942	0.950	3.521	0.939	0.952	4.120
	Proposed StereoQA-Net	**0.965**	**0.973**	**3.682**	**0.947**	**0.957**	**3.270**	**0.956**	**0.965**	**3.477**

TABLE II: OR COMPARISON ON LIVE PHASE II DATABASE: FR (NR) INDICATES FULL-REFERENCE (NO-REFERENCE) MODELS

Type	Metrics	JP2K	JPEG	WN	BLUR	FF	All
FR	Gorley [6]	0.028	0	0	0	0.028	0.044
	You [5]	0	0	0	0.042	0	0.008
	Benoit [4]	0	0	0	0.125	0.014	0.028
	Chen (MS-SSIM) [7]	0	0	0	0	0	0
NR	Akhter [9]	0	0	0	0.056	0.069	0.039
	Chen [16]	0	0	0	0	0	0
	CNN [20]	0.1	0.923	0	0.111	0.083	0.056
	Proposed StereoQA-Net	0	0	0	0	0	0

We rescale each image patch to the range $[0, 1]$ and conduct local normalization [20]. The normalized patch pairs for left and right views are denoted by P_{li} and P_{ri}, where $i = 1, 2, ..., p$. It should be noted that the patch based stereoscopic image preprocessing is similar to many research works such as [20], which is a common method to effectively resolve the lack of training data problem. Therefore, we can obtain a lot of image patch pairs in this way, which provides sufficient image patch pairs for training.

C. Network Interaction

In our proposed StereoQA-Net, the concatenations are at various layers because the interaction between left and right views exists in the whole hierarchical human visual cortex. Specifically, we perform concatenation after the second and the fifth convolutional layers as well as the last fully connected layer of each sub-network. In other words, the network interaction is adopted in multiple layers.

We design our dual-stream network with inspiration from the relationship between deep neural networks and hierarchical human visual cortical areas. The deep neural networks model the responses of brain activity across the hierarchical visual pathway [42]. In addition, the processing scheme in the primate visual system is presented in the way of deep hierarchies rather than flat processing [43]. More specifically, the optic nerve transmits the input stereoscopic visual signals to the lateral geniculate nucleus (LGN) which relays the information to the human visual cortex [44]. Further, the performance optimization of DNN models enables the output layer to resemble inferior temporal (IT) cortex and the feature representation of DNN intermediate layers is similar to that of functional areas in HVS, such as V4 cortex [45]. Also, quantitative experiments show that the primate ventral visual pathway encodes increasingly complex stimulus features corresponding to the DNN layers [46]. Moreover, the visual brain representation can be modeled by DNN, which verifies the consistency between the DNN and HVS [47]. Therefore, the lower layer $Conv2$ and the higher layer $Conv5$ (i.e. the last convolutional layer of each primary sub-network) are exploited to generate two concatenated sub-networks.

For the interactions between the convolutional layers such as $Conv2$ and $Conv5$, we first compute fusion and difference maps by the summation and subtraction operations [48]–[50] as follows:

$$S^+ = F_l + F_r, \tag{2}$$

$$S^- = F_l - F_r, \tag{3}$$

where F_l and F_r are the corresponding feature maps of input patches for left and right view primary sub-networks.

To demonstrate the effects of summation and subtraction operations, the fusion and difference maps of distorted stereoscopic images are shown in Fig. 2. As we can see from Fig. 2, the fusion as well as difference maps of left and right view images with different distortion types/levels are discriminative and can be trained to learn effective quality features. In addition, the fusion map reveals the fusion ability of left and right stereo-halves, while the difference map reflects the disparity information [48], [51].

For fully connected layers, we also perform the network interaction of the last fully connected layer for each sub-

network. Then, the output feature vector of this network interaction for fully connected layers is denoted as:

$$V = [V_l, V_r, V_{Concate2}, V_{Concate5}], \tag{4}$$

where V_l and V_r represent the output feature vectors of left and right view primary sub-networks, respectively. Besides, the output feature vectors of concatenated sub-networks are $V_{Concate2}$ and $V_{Concate5}$, respectively.

D. Patch Pair-Wise Training

To train the proposed StereoQA-Net on GPU, we need to obtain large-scale training data. Since we adopt fully connected layers in the proposed network, the sizes of input images need to be fixed. Additionally, the feature correspondence between left and right views is required to predict the perceptual quality for distorted stereoscopic images, especially for those with asymmetrical artifacts. Therefore, we train our network on multiple 32×32 patch pairs taken from relatively larger left and right view images separately.

During the training stage, we assign each patch pair a target quality score as the ground truth score of the corresponding source stereoscopic image. This is because in real applications, compression distortions for stereoscopic images such as JPEG are widely used and have been included in the existing stereoscopic image quality databases. Due to the homogeneous characteristic of these distortions (i.e. state-of-the-art distortions in all stereoscopic image quality databases), we assign the whole distorted image score to cropped patches inspired by the work in [20]. Suppose that (P_{li}, P_{ri}) denotes the input patch pair, and y_i is the corresponding ground truth quality score. Our learning objective function is defined by Euclidean Loss as follows:

$$\begin{aligned} L &= ||f(P_{li}, P_{ri}; w) - y_i||_F^2 \\ w' &= \min_w L \end{aligned}, \tag{5}$$

where $f(P_{li}, P_{ri}; w)$ represents the predicted quality score of the patch pair (P_{li}, P_{ri}) with network weights w.

For training, stochastic gradient descent (SGD) with momentum as well as backpropagation is applied to train the network. Also, the learning rate is initially set to 10^{-2}. Here, each convolutional layer and the fully connected layers of sub-networks are followed by Rectified Linear Units (ReLUs) [52] instead of conventional sigmoid or tanh neurons. Let w_i and a_i denote the weights of the ReLU and the outputs of the previous layer, respectively. Then, the ReLU can be represented by $ReLU = \max(0, \sum_i w_i a_i)$. Therefore, the ReLU is a nonlinear activation function by employing a threshold value to the input, which can effectively simplify back-propagation, enhance optimization, etc.

In addition to the ReLU, we also adopt dropout technique to avoid network overfitting. Specifically, we apply dropout at each fully connected layer of sub-networks. By randomly setting the outputs of neurons to zero with a probability of 0.5 or 0.35 in our experiments, dropout acts as an effective approximation and prevents overfitting for training networks with shared weights.

TABLE III: PERFORMANCE COMPARISON OF SYMMETRICALLY AND ASYMMETRICALLY DISTORTED STEREOSCOPIC IMAGE PAIRS SEPARATELY ON LIVE PHASE II DATABASE (SROCC), AND THE BEST EXPERIMENTAL RESULTS ARE IN BOLD

Metrics	Symmetric	Asymmetric
Chen (MS-SSIM) [7]	0.923	0.842
Chen [16]	0.918	0.834
CNN [20]	0.590	0.633
Proposed StereoQA-Net	**0.979**	**0.927**

E. Local Quality Pooling

In the testing stage, we adopt a spatial local quality pooling method to estimate the whole stereoscopic image quality. According to the predicted local quality scores of distorted patch pairs and the homogeneous artifacts in symmetrically and asymmetrically distorted stereoscopic images, the final perceptual stereoscopic image quality is derived by averaging the predicted local quality for each patch pair as follows:

$$Q = \frac{1}{p} \sum_{i=1}^{p} f(P_{li}, P_{ri}), \tag{6}$$

where $i = 1, 2, ..., p$ denote p patch pairs for each distorted stereoscopic image. Then, we can obtain the global human visual based perceptual quality for stereoscopic images by the local quality pooling.

III. Experiment Results and Analysis

In this section, we first describe the databases and criteria used for our experiments. Second, the performance comparison results for both symmetrically and asymmetrically distorted stereoscopic images on the LIVE stereoscopic image quality databases are given. Then, we show the evaluation results on individual distortion type. Also, we examine the effects of several network parameters. Moreover, local quality estimation further demonstrates the effectiveness of our algorithm. In addition, we conduct statistical significance tests to verify the statistical superiority of the proposed method. Finally, cross database and time complexity tests are performed in our experiments.

A. Databases and Criteria

In order to evaluate our algorithm, we conduct experiments on the following two SIQA benchmark databases.

LIVE Phase I [53]: This database consists of 365 distorted stereopairs with five various distortion types derived from 20 reference stereoscopic images. Among the total of 365 distorted stereoscopic image pairs, 80 pairs contain each of JPEG2000 compression (JP2K), JPEG compression (JPEG), additive white Gaussian noise (WN), and Raleigh fast fading channel distortion (FF), and 45 pairs are related to Gaussian blur (BLUR). All of the distortions are symmetric in nature representing the same distortion degree in both left and right view images. The corresponding differential Mean Opinion

TABLE IV: SROCC COMPARISON FOR INDIVIDUAL DISTORTION TYPE ON THE TWO LIVE SIQA DATABASES. FR (NR) INDICATES FULL-REFERENCE (NO-REFERENCE) MODELS, AND THE BEST EXPERIMENTAL RESULTS ARE IN BOLD

Type	Metrics	LIVE Phase I					LIVE Phase II				
		JP2K	JPEG	WN	BLUR	FF	JP2K	JPEG	WN	BLUR	FF
FR	Gorley [6]	0.015	0.569	0.741	0.750	0.366	0.110	0.027	0.875	0.770	0.601
	You [5]	0.860	0.439	0.940	0.882	0.588	**0.894**	0.795	0.909	0.813	0.891
	Benoit [4]	0.910	0.603	0.930	0.931	0.699	0.751	**0.867**	0.923	0.455	0.773
	Lin [8]	0.839	0.207	0.928	**0.935**	0.658	0.718	0.613	0.907	0.711	0.701
	Chen (MS-SSIM) [7]	0.888	0.530	0.948	0.925	0.707	0.814	0.843	0.940	**0.908**	0.884
NR	Akhter [9]	0.866	0.675	0.914	0.555	0.640	0.724	0.649	0.714	0.682	0.559
	Sazzad [15]	0.721	0.526	0.807	0.597	0.705	0.625	0.479	0.647	0.775	0.725
	Chen [16]	0.863	0.617	0.919	0.878	0.652	0.867	**0.867**	**0.950**	0.900	0.933
	S3D-BLINQ [41]	-	-	-	-	-	0.845	0.818	0.946	0.903	0.899
	CNN [20]	0.857	0.447	0.874	0.782	0.670	0.660	0.598	0.769	0.317	0.476
	DNR-S3DIQE [32]	0.885	0.765	0.921	0.930	**0.944**	0.853	0.822	0.833	0.889	0.878
	DBN [33]	0.897	0.768	0.929	0.917	0.685	0.859	0.806	0.864	0.834	0.877
	FC-1024 + SVR	0.932	0.668	0.920	0.896	0.865	0.873	0.808	0.931	0.660	0.935
	Proposed StereoQA-Net	**0.961**	**0.912**	**0.965**	0.855	0.917	0.874	0.747	0.942	0.600	**0.951**

TABLE V: PLCC COMPARISON FOR INDIVIDUAL DISTORTION TYPE ON THE TWO LIVE SIQA DATABASES. FR (NR) INDICATES FULL-REFERENCE (NO-REFERENCE) MODELS, AND THE BEST EXPERIMENTAL RESULTS ARE IN BOLD

Type	Metrics	LIVE Phase I					LIVE Phase II				
		JP2K	JPEG	WN	BLUR	FF	JP2K	JPEG	WN	BLUR	FF
FR	Gorley [6]	0.485	0.312	0.796	0.852	0.364	0.372	0.322	0.874	0.934	0.706
	You [5]	0.877	0.487	0.941	0.919	0.730	0.905	0.830	0.912	0.784	0.915
	Benoit [4]	0.939	0.640	0.925	0.948	0.747	0.784	0.853	0.926	0.535	0.807
	Lin [8]	0.799	0.196	0.925	0.811	0.700	0.744	0.583	0.909	0.671	0.699
	Chen (MS-SSIM) [7]	0.912	0.603	0.942	0.942	0.776	0.834	0.862	0.957	0.963	0.901
NR	Akhter [9]	0.905	0.729	0.904	0.617	0.503	0.776	0.786	0.722	0.795	0.674
	Sazzad [15]	0.774	0.565	0.803	0.628	0.694	0.645	0.531	0.657	0.721	0.727
	Chen [16]	0.907	0.695	0.917	0.917	0.735	0.899	0.901	0.947	0.941	0.932
	S3D-BLINQ [41]	-	-	-	-	-	0.847	0.888	0.953	0.968	0.944
	CNN [20]	0.956	0.630	0.983	0.862	0.846	0.685	0.567	0.855	0.455	0.662
	DNR-S3DIQE [32]	0.913	0.767	0.910	0.950	0.954	0.865	0.821	0.836	0.934	0.815
	DBN [33]	0.942	0.824	0.954	0.963	0.789	0.886	0.867	0.887	**0.988**	0.916
	FC-1024 + SVR	0.966	0.700	0.971	0.947	0.926	0.887	0.910	0.969	0.959	0.984
	Proposed StereoQA-Net	**0.988**	**0.916**	**0.988**	**0.974**	**0.965**	**0.905**	**0.933**	**0.972**	0.955	**0.994**

Score (DMOS) value is provided for each distorted stereoscopic image pair. Besides, the DMOS values are roughly in the range $[0, 80]$, where higher DMOS values indicate lower visual quality.

LIVE Phase II [7], [16]: It contains 8 reference stereoscopic images and 360 distorted stereopairs with five different distortion types. The distortion types are the same as that of the LIVE Phase I database. Each distortion type has nine levels, where one third of the levels are symmetric in nature. In other words, this database includes 120 symmetrically distorted stereoscopic images and 240 asymmetrically distorted stereoscopic images that have different degrees of distortion in left and right view images. Each distorted stereoscopic image is also associated with a DMOS value, which is similar to that of the LIVE Phase I database.

Evaluation Method: By following previous methods [54], four general measures are adopted to evaluate the performance of SIQA algorithms: 1) Spearman Rank Order Correlation Coefficient (SROCC), 2) Pearson Linear Correlation Coefficient (PLCC), 3) Root Mean Squared Error (RMSE), and 4) Outlier Ratio (OR). Among these four metrics, the SROCC measures the monotonicity of two quantities, while the PLCC measures the linear dependence between the predicted quality scores and the ground truth targets. Apart from the PLCC, the RMSE also measures the prediction accuracy, which represents the distance between the subjective scores and predicted scores. Moreover, the OR measures the prediction consistency. For PLCC, RMSE and OR, the five-parameter logistic function is applied to fit predicted quality scores and DMOS values using nonlinear least-squares optimization [55]. Higher SROCC and PLCC values represent good correlation (monotonicity and accuracy) with human perceptual quality judgments, while the lower values of RMSE and OR indicate better performance. In our experiments, the distorted stereoscopic images are randomly selected 80% as training set and the remaining 20% as testing set, and the experimental results are obtained after 100 epochs.

B. Performance Comparison

We train the network in a non-distortion-specific manner to evaluate the performance of our proposed StereoQA-Net. This way, the distorted stereoscopic images with five various distortion types are trained and tested simultaneously without needing a specific distortion type.

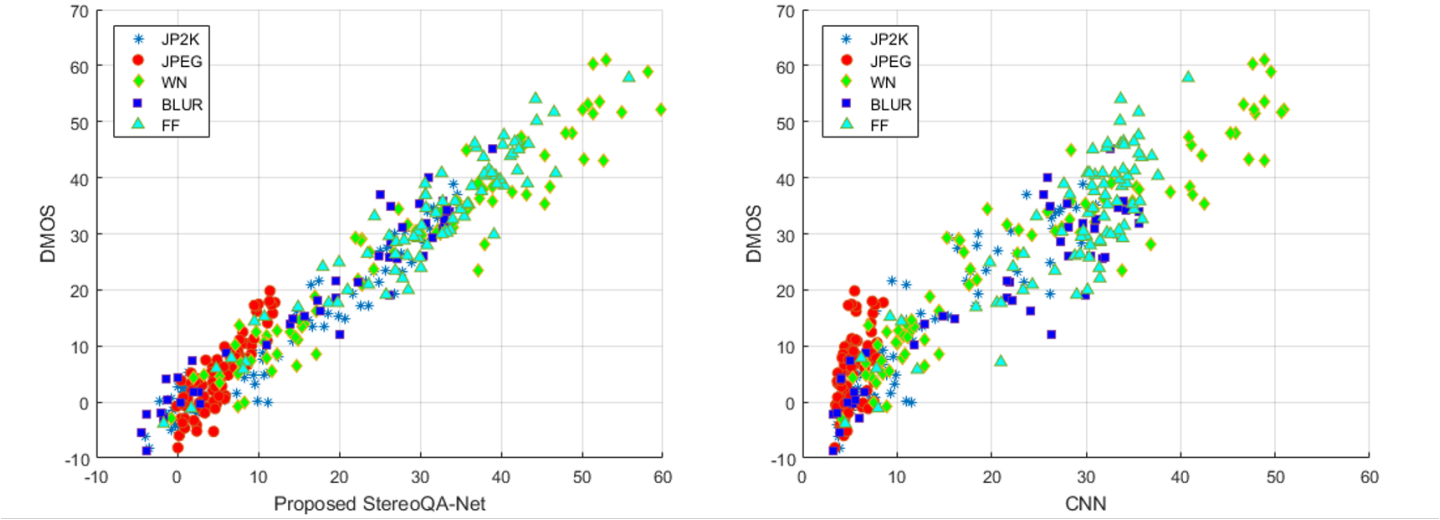
Table I shows the SROCC, PLCC and RMSE performance results on the LIVE Phase I and LIVE Phase II databases

TABLE VI: PERFORMANCE UNDER DIFFERENT KERNEL SIZES ON THE TWO LIVE STEREOSCOPIC IMAGE QUALITY DATABASES

Kernel Size	LIVE Phase I				LIVE Phase II			
	3×3	5×5	7×7	9×9	3×3	5×5	7×7	9×9
SROCC	0.965	0.958	0.956	0.959	0.947	0.949	0.942	0.947
PLCC	0.973	0.971	0.969	0.972	0.957	0.957	0.952	0.954

TABLE VII: PERFORMANCE UNDER DIFFERENT PATCH SIZES ON THE TWO LIVE STEREOSCOPIC IMAGE QUALITY DATABASES

Patch Size	LIVE Phase I				LIVE Phase II			
	24×24	32×32	40×40	48×48	24×24	32×32	40×40	48×48
SROCC	0.967	0.965	0.962	0.962	0.946	0.947	0.934	0.964
PLCC	0.960	0.973	0.969	0.976	0.958	0.957	0.949	0.970

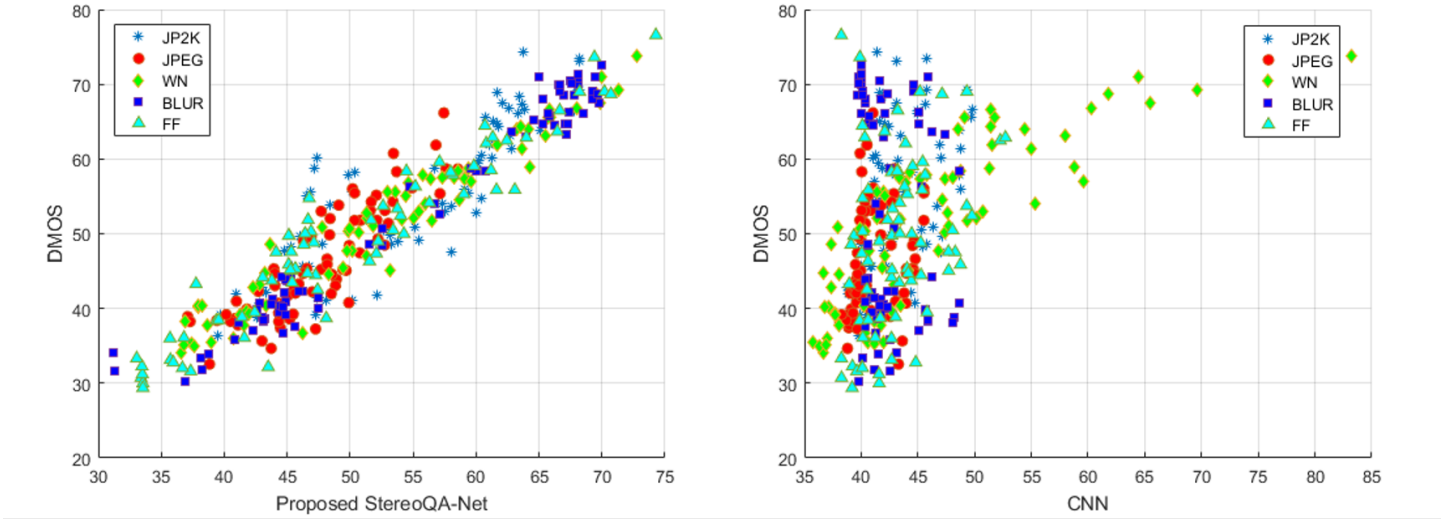


(a) (b)

Fig. 4: Scatter plots of predicted quality scores against the subjective scores (DMOS) of the proposed algorithm and CNN method [20] for individual distortion type on LIVE Phase I [53].

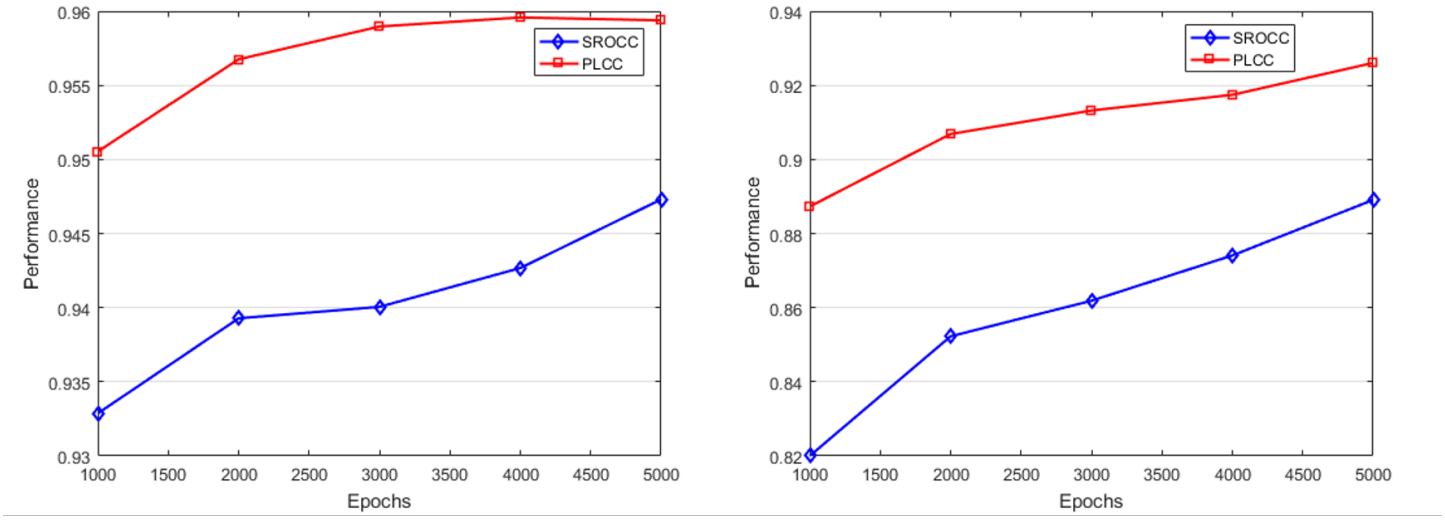
compared with five state-of-the-art FR-SIQA methods, namely those of Gorley [6], You [5], Benoit [4], Lin [8], and Chen [7]. For [7], multi-scale structure similarity index (MS-SSIM) [56] is used for the performance comparison. We also compare with four previous NR-SIQA metrics developed by Akhter [9], Sazzad [15], Chen [16], and Su (i.e. S3D-BLINQ) [41]. Also, two deep CNN-based methods, namely the NR-IQA model and the NR-SIQA approach, are benchmarked: CNN [20] and DNR-S3DIQE [32]. The NR-IQA model named CNN [20] is employed to train the left and right view networks separately. The predicted quality scores of distorted stereoscopic images are then obtained by averaging left and right view qualities. Moreover, a deep belief network based method called DBN [33] is also compared. In addition, to compare the end-to-end deep learning related features with traditional features, we conduct the experiments by combining the last fully connected layer feature denoted by FC-1024 with the well-known support vector regression (SVR) model. In the last column, the weighted average of SROCC, PLCC and RMSE over the two databases are reported, we adopt the same method as [57] to produce the weights. In Table II, it should be noted that since the standard deviations of DMOS scores of the LIVE Phase I database are not available [16], we only show OR numbers on the LIVE Phase II database, as reported in [16]. From Table I and Table II, we can find that our proposed algorithm outperforms existing state-of-the-art NR as well as FR methods. The end-to-end deep learning related features are more effective than traditional features for the perceptual quality prediction of stereoscopic images, which further demonstrates the superiority of deep features [58]. One possible explanation may be that the proposed network directly learns the effective feature representations by training in an end-to-end manner. Note that this network is not overfitting since the training/testing method is the same as other learning-based 2D IQA algorithms such as [20] and the high performance is obtained on testing set, not training set.

Furthermore, we conduct the ablation study of the proposed network interaction which demonstrates that the experimental results are improved with the addition of the multi-layer network interaction, especially for those asymmetrically distorted



(a) (b)

Fig. 5: Scatter plots of predicted quality scores against the subjective scores (DMOS) of the proposed algorithm and CNN method [20] for individual distortion type on LIVE Phase II [7], [16].



(a) (b)

Fig. 6: SROCC and PLCC performance with respect to training epochs. (a) Run on LIVE Phase I [53]; (b) Run on LIVE Phase II [7], [16].

stereoscopic image pairs. Specifically, the SROCC can be improved from 0.890 to 0.947 for the LIVE Phase II database.

The training loss optimization process on LIVE Phase I and LIVE Phase II from the first 50 epochs is shown in Fig. 3. It can be seen that the training process converges well. Further, we also validate the performance of our proposed algorithm on symmetrically and asymmetrically distorted stereoscopic image pairs separately. As can be seen in Table III, our proposed StereoQA-Net outperforms other state-of-the-art FR and NR methods for both symmetrically and asymmetrically distorted stereoscopic images.

TABLE VIII: THE T-TEST RESULTS ON LIVE PHASE I AND LIVE PHASE II DATABASES

Method	LIVE Phase I	LIVE Phase II
CNN [20]	1	1

C. Evaluation on Individual Distortion Type

In order to illustrate the performance comparison for individual distortion type on the hybrid distortion database, we give the SROCC and PLCC performance comparison according to various distortion types on both LIVE Phase I and

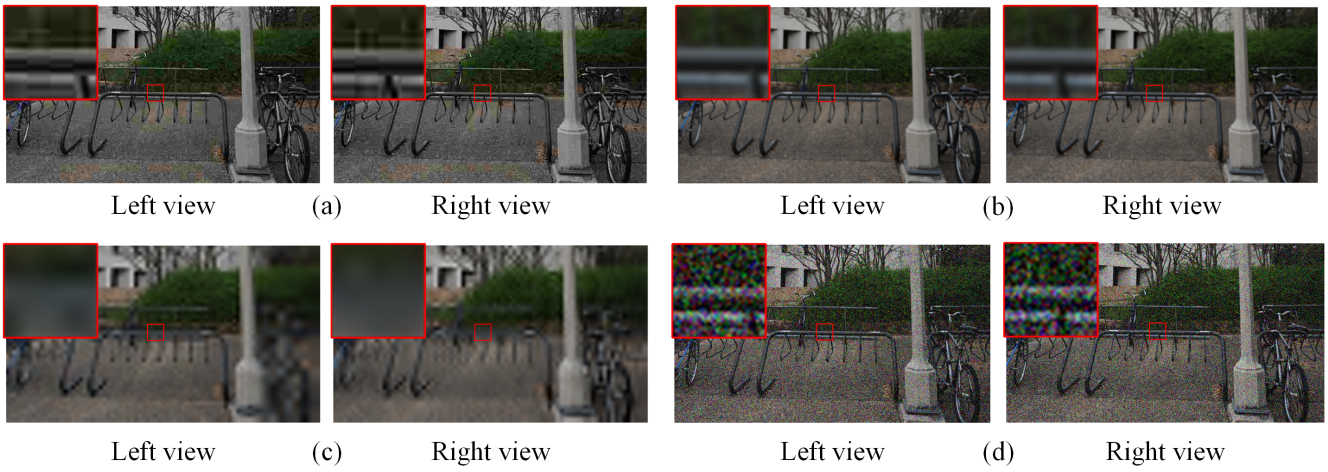


Fig. 7: Examples of local quality estimation for different distorted patch pairs (denoted by red bounding boxes) from LIVE Phase I database. (a) JPEG; (b) BLUR; (c) FF; (d) WN. The corresponding quality scores predicted by the proposed StereoQA-Net are 46.128, 53.495, 66.804, and 74.013, respectively. Note that higher predicted quality values represent lower visual quality.

Phase II, as shown in Table IV and Table V respectively. We can see that even when each distortion type is tested separately, the proposed StereoQA-Net generally achieves the competitive performance on each of the five distortion types.

Moreover, Fig. 4 and Fig. 5 depict the scatter plots of D-MOS versus predictions for individual distortion type on LIVE Phase I and LIVE Phase II, respectively. Since a straight lined distribution of the points is better than other arbitrary shapes, our proposed StereoQA-Net shows much better linearity and monotonicity than the CNN method [20] for each distortion type on the two databases.

D. Effects of Parameters

Since several network parameters are involved in the proposed StereoQA-Net design, we examine how these parameters affect the performance of the network on the LIVE Phase I and LIVE Phase II databases as follows.

Kernel Size: We train and test the proposed StereoQA-Net with different kernel sizes while fixing the rest of architecture. Table VI shows how the performance changes with the kernel size on LIVE SIQA databases. As we can see that our proposed StereoQA-Net is not sensitive to kernel size.

Patch Size: In our experiments, the quality score of the whole distorted stereoscopic image pair is obtained by averaging the predicted scores of all sampled patch pairs. Therefore, we examine how the patch sampling strategy affects the performance of our method. Specifically, we train and test the proposed StereoQA-Net with different patch sizes while fixing the rest of architecture. The performance changes with the patch size on LIVE SIQA databases are shown in Table VII. From Table VII, we can see that our proposed StereoQA-Net is also not sensitive to patch size.

Iterative Epoch: Furthermore, in order to discover how the iterative epoch of training affects the performance of our proposed StereoQA-Net, we vary the iterative epoch of training to plot the performance variation for LIVE Phase I and LIVE Phase II databases. Specifically, we train the proposed StereoQA-Net by changing the iterative epoch while fixing the rest of architecture. Note that we initially set the learning rate to 10^{-4} for better visualization. Fig. 6 shows the SROCC and PLCC performance with respect to the iterative epoch on the two databases. We can observe that a large number of iterative epoch brings about the increase of performance for our proposed StereoQA-Net. Further, we increase the epochs from 5000 to 6000. We find that the performance improves slightly. For example, as for LIVE Phase I, the SROCC changes from 0.947 to 0.948. Therefore, our proposed method can reach a promising result without too many epochs, which turns out to be convergent.

E. Local Quality Estimation

Since our proposed StereoQA-Net measures the perceptual quality of small stereoscopic image patch pairs, it can be applied to detect local regions with different quality degrees as well as give a global quality score for the whole stereoscopic image. We select distorted stereoscopic patch pairs with different distortion types at the same spatial location from LIVE Phase I, and then perform local quality estimation using our model trained on LIVE Phase II.

Fig. 7 shows the examples of the predicted quality scores for the corresponding stereoscopic patch pairs with each of different distortion levels and types including JPEG, BLUR, FF, and WN. From Fig. 7, we can see that the proposed StereoQA-Net can effectively differentiate various distortion levels and types.

F. Statistical Significance

Further, we conduct statistical significance tests to quantify whether the comparison results are statistically significant. Specifically, on each stereoscopic image quality database, a two-sample t-test is carried out, which is at 1% significance level using the SROCC value pairs of 100 runs. Note that

TABLE IX: CROSS DATABASE TEST RESULTS OBTAINED BY TRAINING ON LIVE PHASE II AND TESTING ON LIVE PHASE I (PLCC)

Metrics	JP2K	JPEG	WN	BLUR	FF	ALL
CNN [20]	0.646	0.562	0.968	0.848	0.803	0.713
Proposed StereoQA-Net	0.981	0.678	0.988	0.947	0.839	0.932

TABLE X: CROSS DATABASE TEST RESULTS OBTAINED BY TRAINING ON LIVE PHASE I AND TESTING ON LIVE PHASE II (PLCC)

Metrics	JP2K	JPEG	WN	BLUR	FF	ALL
CNN [20]	0.577	0.834	0.739	0.711	0.921	0.656
Proposed StereoQA-Net	0.909	0.661	0.679	0.942	0.968	0.710

TABLE XI: PERFORMANCE COMPARISON OF THE TIME COMPLEXITY ON THE LIVE PHASE I DATABASE

Metrics	CNN [20]	Proposed StereQA-Net
Total time (s)	131.291	37.566

the source codes of other state-of-the-art algorithms are not publicly available, we only compare our method with CNN [20] method. Table VIII shows the statistical significance testing results, where 1 indicates that our proposed StereoQA-Net is statistically superior to the compared CNN [20] method.

G. Cross Database and Time Complexity Tests

As shown in Table IX, to evaluate the generalization ability of the proposed StereoQA-Net, we use the stereoscopic image patch pairs from LIVE Phase II to train the network, and then test on LIVE Phase I. This is because the LIVE Phase II database consists of both symmetric and asymmetric distortions, while the LIVE Phase I database only contains symmetrically distorted stereoscopic images. Furthermore, as can be seen in Table X, we also conduct the experiments of training on LIVE Phase I and then test on LIVE Phase II. It should be noted that the source codes of other state-of-the-art algorithms are not publicly available, only CNN [20] is used for comparison. Experimental results of the cross database test are shown in Table IX and Table X. As can be seen from these two tables, our proposed algorithm performs well and is generalized to different databases.

In addition, we compare the proposed method with CNN [20] on the LIVE Phase I database to show the lower computing complexity of our proposed StereoQA-Net. From Table XI, we can find that the proposed algorithm is verified to have lower time complexity. One possible explanation is that the CNN [20] method computes two times of quality prediction for left and right views. In other words, the proposed StereoQA-Net is specifically designed for predicting stereoscopic image quality. Moreover, different from conventional SIQA, the proposed metric can relieve the complex computation of disparity map.

IV. CONCLUSIONS

In this paper, we propose a novel general-purpose NR-SIQA architecture that contains the multi-layer network interaction between left and right view sub-networks inspired by the HVS. The discriminative feature extraction and regression learning are taken as an end-to-end optimization process. The predicted stereoscopic image quality of our metric correlates well with human visual perception. Moreover, our proposed StereoQA-Net achieves the state-of-the-art performance and has a promising generalization ability for both symmetrically and asymmetrically distorted stereoscopic images of various distortion types.

In future research, the way to design more effective networks for stereoscopic image quality assessment should be considered. More importantly, we will try to gain more theoretic insight into why the machine-learned features are better than hand-crafted features for evaluating the perceptual quality of stereoscopic images. Meanwhile, future work could also involve modeling the effects of different patches on the perceived quality for the entire distorted stereoscopic image. Besides, we plan to advance the proposed architecture for stereoscopic video quality assessment. Specifically, the spatiotemporal quality variation and more HVS characteristics should be taken into consideration. Furthermore, except for image quality, it is important to understand human perceptual opinions on other 3D dimensions such as depth perception and visual comfort, aiming to further improve the proposed method.

REFERENCES

[1] C.-C. Su, A. K. Moorthy, and A. C. Bovik, "Visual quality assessment of stereoscopic image and video: challenges, advances, and future trends," in *Visual Signal Quality Assessment*. Springer, 2015, pp. 185–212.

[2] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 2971–2983, 2015.

[3] F. Shao, W. Tian, W. Lin, G. Jiang, and Q. Dai, "Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2059–2074, 2016.

[4] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP journal on image and video processing*, vol. 2008, no. 1, p. 659024, 2009.

[5] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA*, 2010.

[6] P. Gorley and N. Holliman, "Stereoscopic image quality metrics and compression," in *Proc. SPIE*, vol. 6803, 2008, p. 680305.

[7] 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*, vol. 28, no. 9, pp. 1143–1155, 2013.

[8] Y.-H. Lin and J.-L. Wu, "Quality assessment of stereoscopic 3D image compression by binocular integration behaviors," *IEEE transactions on Image Processing*, vol. 23, no. 4, pp. 1527–1542, 2014.

[9] R. Akhter, Z. P. Sazzad, Y. Horita, and J. Baltes, "No-reference stereoscopic image quality assessment," in *Stereoscopic Displays and Applications XXI*, vol. 7524. International Society for Optics and Photonics, 2010, p. 75240T.

[10] F. Shao, K. Li, W. Lin, G. Jiang, and M. Yu, "Using binocular feature combination for blind quality assessment of stereoscopic images," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1548–1551, 2015.

[11] 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*, vol. 61, no. 2, pp. 154–165, 2015.

[12] F. Shao, W. Lin, S. Wang, G. Jiang, M. Yu, and Q. Dai, "Learning receptive fields and quality lookups for blind quality assessment of stereoscopic images," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 730–743, 2016.

[13] F. Shao, K. Li, W. Lin, G. Jiang, and Q. Dai, "Learning blind quality evaluator for stereoscopic images using joint sparse representation," *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 2104–2114, 2016.

[14] A. J. Smola and B. Schölkopf, *Learning with kernels*. GMD-Forschungszentrum Informationstechnik, 1998.

[15] Z. Sazzad, R. Akhter, J. Baltes, and Y. Horita, "Objective no-reference stereoscopic image quality prediction based on 2D image features and relative disparity," *Advances in Multimedia*, vol. 2012, p. 8, 2012.

[16] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-reference quality assessment of natural stereopairs," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3379–3391, 2013.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.

[18] K. Grill-Spector and R. Malach, "The human visual cortex," *Annu. Rev. Neurosci.*, vol. 27, pp. 649–677, 2004.

[19] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[20] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014, pp. 1733–1740.

[21] Y. Li, L.-M. Po, X. Xu, L. Feng, F. Yuan, C.-H. Cheung, and K.-W. Cheung, "No-reference image quality assessment with shearlet transform and deep neural networks," *Neurocomputing*, vol. 154, pp. 94–109, 2015.

[22] S. Yi, D. Labate, G. R. Easley, and H. Krim, "A shearlet approach to edge analysis and detection," *IEEE Transactions on Image Processing*, vol. 18, no. 5, pp. 929–941, 2009.

[23] G. Easley, D. Labate, and W.-Q. Lim, "Sparse directional image representations using the discrete shearlet transform," *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, pp. 25–46, 2008.

[24] G. Kutyniok, W.-Q. Lim, and X. Zhuang, "Digital shearlet transforms," *Shearlets*, pp. 239–282, 2012.

[25] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems (NIPS)*, 2007, pp. 153–160.

[26] Y. Li, L.-M. Po, L. Feng, and F. Yuan, "No-reference image quality assessment with deep convolutional neural networks," in *Digital Signal Processing (DSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 685–689.

[27] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3773–3777.

[28] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[29] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 6, pp. 1275–1286, 2015.

[30] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, 2017.

[31] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-End blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.

[32] H. Oh, S. Ahn, J. Kim, and S. Lee, "Blind deep S3D image quality evaluation via local to global feature aggregation," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4923–4936, 2017.

[33] J. Yang, Y. Zhao, Y. Zhu, H. Xu, W. Lu, and Q. Meng, "Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network," *Information Sciences*, 2018.

[34] A. Canziani and E. Culurciello, "CortexNet: a generic network family for robust visual temporal representations," *arXiv preprint arXiv:1706.02735*, 2017.

[35] Z. Chen, W. Zhou, and W. Li, "Blind stereoscopic video quality assessment: From depth perception to overall experience," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 721–734, 2018.

[36] K. A. May and L. Zhaoping, "Efficient coding theory predicts a tilt aftereffect from viewing untilted patterns," *Current Biology*, vol. 26, no. 12, pp. 1571–1576, 2016.

[37] A. J. Parker, "Binocular depth perception and the cerebral cortex," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 379–391, 2007.

[38] R. B. Tootell, J. D. Mendola, N. K. Hadjikhani, P. J. Ledden, A. K. Liu, J. B. Reppas, M. I. Sereno, and A. M. Dale, "Functional analysis of V3A and related areas in human visual cortex," *Journal of Neuroscience*, vol. 17, no. 18, pp. 7060–7078, 1997.

[39] A. W. Roe, L. Chelazzi, C. E. Connor, B. R. Conway, I. Fujita, J. L. Gallant, H. Lu, and W. Vanduffel, "Toward a unified theory of visual area V4," *Neuron*, vol. 74, no. 1, pp. 12–29, 2012.

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

[41] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation," *IEEE Transactions on image processing*, vol. 24, no. 5, pp. 1685–1699, 2015.

[42] P. Agrawal, D. Stansbury, J. Malik, and J. L. Gallant, "Pixels to voxels: modeling visual representation in the human brain," *arXiv preprint arXiv:1407.5104*, 2014.

[43] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1847–1871, 2013.

[44] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex." *Cerebral cortex (New York, NY: 1991)*, vol. 1, no. 1, pp. 1–47, 1991.

[45] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[46] U. Güçlü and M. A. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10 005–10 014, 2015.

[47] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Scientific reports*, vol. 6, p. 27755, 2016.

[48] S. Henriksen and J. C. Read, "Visual perception: a novel difference channel in binocular vision," *Current Biology*, vol. 26, no. 12, pp. R500–R503, 2016.

[49] J. Yang, Y. Liu, Z. Gao, R. Chu, and Z. Song, "A perceptual stereoscopic image quality assessment model accounting for binocular combination behavior," *Journal of Visual Communication and Image Representation*, vol. 31, pp. 138–145, 2015.

[50] J. Ma, P. An, L. Shen, K. Li, and J. Yang, "SSIM-based binocular perceptual model for quality assessment of stereoscopic images," in *Visual Communications and Image Processing (VCIP), 2017 IEEE*. IEEE, 2017, pp. 1–4.

[51] F. A. Kingdom, "Binocular vision: The eyes add and subtract," *Current Biology*, vol. 22, no. 1, pp. R22–R24, 2012.

[52] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[53] A. K. Moorthy, C.-C. Su, A. Mittal, and A. C. Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 870–883, 2013.

[54] "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II." *VQEG, 2003*.

[55] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.
[56] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1398–1402.
[57] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, 2017.
[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *arXiv preprint*, 2018.