A no-Reference Stereoscopic Image Quality Assessment Network Based on Binocular Interaction and Fusion Mechanisms

Jianwei Si, Baoxiang Huang, *Member, IEEE*, Huan Yang, *Member, IEEE*, Weisi Lin, *Fellow, IEEE*, and Zhenkuan Pan, *Senior Member, IEEE*

***Abstract*— In contemporary society full of stereoscopic images, how to assess visual quality of 3D images has attracted an increasing attention in field of Stereoscopic Image Quality Assessment (SIQA). Compared with 2D-IQA, SIQA is more challenging because some complicated features of Human Visual System (HVS), such as binocular interaction and binocular fusion, must be considered. In this paper, considering both binocular interaction and fusion mechanisms of the HVS, a hierarchical no-reference stereoscopic image quality assessment network (StereoIF-Net) is proposed to simulate the whole quality perception of 3D visual signals in human cortex, including two key modules: BIM and BFM. In particular, Binocular Interaction Modules (BIMs) are constructed to simulate binocular interaction in V2-V5 visual cortex regions, in which a novel cross convolution is designed to explore the interaction details in each region. In the BIMs, different output channel numbers are designed to imitate various receptive fields in V2-V5. Furthermore, a Binocular Fusion Module (BFM) with automatic learned weights is proposed to model binocular fusion of the HVS in higher cortex layers. The verification experiments are conducted on the LIVE 3D, IVC and Waterloo-IVC SIQA databases and three indices including PLCC, SROCC and RMSE are employed to evaluate the assessment consistency between StereoIF-Net and the HVS. The proposed StereoIF-Net achieves almost the best results compared with advanced SIQA methods. Specifically, the metric values on LIVE 3D, IVC and WIVC-I are the best, and are the second-best on the WIVC-II.**

***Index Terms*— Stereoscopic image quality assessment, human visual system, binocular interaction, binocular fusion.**

I. Introduction

With the rapid development of imaging technology, a growing number of 3D images appear in various applications to help humans to perceive the real world, such as automatic driving, virtual reality, 3D TV/films and so on. However, due to inevitable distortions produced by acquisition, compression and transmission, how to assess and quantify diverse quality levels of 3D views accurately has drawn an increasing attention of researchers in the field of Stereoscopic Image Quality Assessment (SIQA) [1]. During recent years, in order to keep high consistency with the Human Visual System (HVS), a series of SIQA metrics have been proposed. Especially, no-reference (NR) methods that do not need any information of reference views have been prevalently studied [2], [3]. Compared with traditional 2D NR-IQA methods, 3D metrics are more challengeable because more information such as depth, disparity and binocular features should be considered to build SIQAs. Besides, as the improvement of calculating ability, Convolution Neutral Network (CNN) has made a major breakthrough in image processing such as image classification, speech recognition, semantic segmentation, target tracking and so on [4]–[7].

Some researchers has successfully introduced CNN into NR SIQAs [8]–[15]. For example, in [8], Oh *et al.* proposed a simple CNN structure with a series of convolution and dense layers is applied to extract features of left and right patches respectively. The obtained features are simply concatenated to produce the final quality score. Zhang *et al.* [9] proposed a three column CNN model, with patches cropped from left view, right view and their difference map as inputs. The three columns have the same network structure, and are finally concatenated by Multilayer Perception (MLP) module to obtain the quality assessment result. In these papers [8]–[10], [12], [14] and [15], the feature extraction of left and right views are wholly separated, without any interaction. Hence, this procedure only considers the monocular vision and ignores the binocular interaction mechanism of the HVS. Fang *et al.* [11] and Zhou *et al.* [13] take the mechanism into account. In other words, convolutional layers and summation/difference calculation between the input feature maps are adopted to extract interactive features respectively. However, due to the complexity for cross processing of visual signals, the simple design of cross procedures for the metrics can not obtain desirable SIQA results.

According to the human visual cortex responses to stereoscopic visual signals [16]–[21], the human visual perception processes can be summarized as follows. Stereoscopic visual signals are transmitted into Lateral

Manuscript received March 26, 2021; revised November 22, 2021 and March 5, 2022; accepted March 23, 2022. Date of publication April 8, 2022; date of current version April 14, 2022. This work was supported in part by the Key Research and Development Plan of Shandong Province under Grant 2019GGX101021 and in part by the Natural Science Foundation of Shandong Province under Grant ZR2021MD001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chaker Larabi. *(Corresponding author: Huan Yang.)*

Jianwei Si, Baoxiang Huang, Huan Yang, and Zhenkuan Pan are with the College of Computer Science and Technology, Qingdao University, Qingdao 266071, China (e-mail: jianweisi_1995@hotmail.com; hbx3726@163.com; cathy_huanyang@hotmail.com; zkpan@qdu.edu.cn).

Weisi Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Digital Object Identifier 10.1109/TIP.2022.3164537

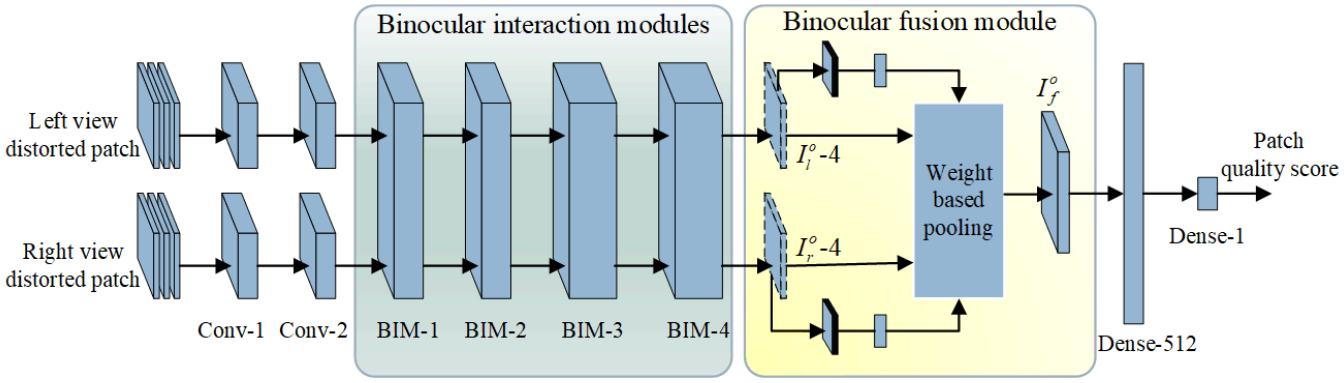


Fig. 1. The architecture of proposed StereoIF-Net.

Geniculate Nucleus (LGN) by opticnerve [17]. Then visual information arrives at V1 cortical area after the transmission of retina and LGN [19]. In V1 area, some low-level features such as size, direction, edge information are extracted. After V1 cortical area, visual signals are transmitted into higher areas such as V2, V3, V4 and V5. Higher level visual features are extracted in these areas, such as object parts in V2-V3 and primitive object models in V4-V5. For neural cells in V4 and V5, they have a larger receptive field than V2 and V3, which means in V4 and V5 areas, more neutral cells are activated by the visual signals. Finally the visual signals are transmitted to higher level areas than V2-V5 and produce the human vision in human's brain. During the entire visual signal perception, binocular visual mechanisms (i.e. binocular interaction, binocular fusion, etc.) exist. For example, visual signals of left and right views are processed cooperatively rather than independently, and have a complicated cross processing relationship, which is so-called binocular interaction. The interaction of left and right views goes through the whole hierarchical human visual cortex [17]. In other words, binocular interaction between left and right visual views always occurs at each high level visual cortical area.

Inspired by the mechanisms of the HVS mentioned above, in this paper, we propose a no-reference stereoscopic image quality assessment network, named StereoIF-Net, aiming to simulate the binocular interaction and fusion mechanisms of the HVS. The hierarchical structure of the HVS is taken into account, and the architecture of the StereoIF-Net is shown in Fig. 1. To simulate the V1 cortical area, two convolutional layers are constructed to extract low-level features in this region. In order to model the visual process of V2-V5 areas, four hierarchical Binocular Interaction Modules (BIM-i, i = 1, 2, 3, 4) based on cross convolution are designed. The proposed cross convolution module can explore binocular interaction more effectively due to the sufficient consideration of cross-over between left and right visual signals. Finally, a weighted Binocular Fusion Module (BFM) is developed to fuse the outputs of BIMs and get the local quality score by two dense layers. The proposed StereoIF-Net has been verified on the LIVE 3D, IVC and Waterloo-IVC SIQA databases, and outperforms state-of-the-art NR SIQA metrics. The software release of StereoIF-Net is available online: ***https://github.com/QDUAIGroup/StereoIF-Net***.

The main contributions of this work are summarized as follows:

1) Considering complicated binocular interaction in various visual regions in the HVS including V2, V3, V4 and V5, four cross convolution based BIMs are proposed to model the interactive procedures. Different output channel numbers of BIMs are designed, reflecting the distinct receptive fields of V2-V5;

2) The proposed cross convolution in BIMs including group division, subgroup summation, difference and concatenation, can explore interaction details in each cortex region, which can enable StereoIF-Net to extract more cross characteristics;

3) Considering the significance of binocular fusion, a BFM with automatic learned weights is proposed to represent the fusion process in higher level regions than V2-V5. The module can better balance weights of left and right views and therefore it can well model advanced visual processing of human's brain.

The rest of this paper is organized as follows. In Section II, we review related works on FR-SIQA and NR-SIQA. In Section III, the architecture and details of the proposed StereoIF-Net are introduced. We evaluate the performance of BIM, BFM and the proposed network in Section IV. Finally, the conclusion is depicted in Section V.

II. Related Works

According to accessible information of reference images, SIQA can be classified into three branches: full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. For the first branch, entire information of reference images can be obtained to build SIQA algorithms. Similarly, partial reference information can be applied for RR methods. However, for NR metrics, none of the contents of reference images can be acquired and only the information of distorted views is used to assess image qualities. For existing SIQA methods, most are FR and NR ones, therefore, in this section, we introduce FR and NR works in the following two subsections respectively.

A. FR Methods of Stereoscopic Image Quality Assessment

Because of the accessibility of reference images, the quality of distorted images can be reaped by comparing them with reference ones. Eminent FR SIQA methods can be mainly categorized into three types as follows: 1) methods based

on monocular vision; 2) methods considering disparity/ depth information; 3) methods including binocular characteristics.

For the methods of the first type, only left and right views are taken into account. In [22], Yasakethu *et al.* used existing well-known 2D IQA methods including Structural SIMilarity (SSIM) [23] and Peak Signal-to-Noise Ratio (PSNR) to assess left and right images. Then two scores are combined to gain the final quality using a weighted fusing strategy. Because of less consideration of stereoscopic features, the methods based on 2D IQA metrics can not achieve expected performance.

The measures of the second class also consider 2D features. Moreover, depth/disparity information related to stereoscopic visual perception is included. Khan and Channappayya [24] proposed a FR SIQA model based on 2D and 3D features. A series of maps including saliency maps, gradient maps and inner gradient maps can be obtained from reference and distorted images. Then these maps are combined with depth perception edges to produce the final quality score. In [25], Liu *et al.* obtained two depth maps and cyclopean images from reference and distorted views respectively. Then MS-SSIM [26] is employed to calculate similarities of depth maps and cyclopean views. The final quality score can be reaped by combining the similarity results. In [27], disparity maps are used by Fan *et al.* to produce cyclopean images. Then they use UQI [28] to attain the similarity of disparity maps and cyclopean images respectively. The ultimate quality score can be obtained by pooling the similarity maps and Just Noticeable Difference (JND) [29] values of cyclopean views.

As research progressed, the third category of the FR methods appears. They consider some binocular characteristics to boost the performance of SIQA methods. In [30], Bensalma and Larabi proposed a SIQA method based on binocular energy difference between reference and distorted images. Zhu *et al.* [31] proposed a SIQA method by importing 2D IQA methods into SIQA ones with a dual-weight model. Moreover, based on [29], Zhao *et al.* [32] proposed a Binocular JND (BJND) model considering some binocular features, for example, perceptible difference between left and right images. Based on both 2D and 3D perceptual factors, Fezza *et al.* [33] proposed a SIQA model based on both JND and BJND. For reference and distorted images, two regions named occluded (OC) area and non-occluded (NOC) area are segmented. In the OC region, JND is used to get a predicted score and BJND is employed to get another one in the NOC region. Finally, the SIQA score can be gained by aggregating both scores. Based on [33], Si *et al.* [34] proposed a FR SIQA method based on stable aggregation of monocular and binocular visual features. For distorted stereo pairs, OC and NOC region can be segmented by a novel region segmentation method. In OC and NOC region, JND and BJND maps can be obtained respectively. Then a Pooling Strategy based on Global Edge features (PSGE) is proposed to get the quality of global monocular and binocular visual characteristics.

B. NR Methods of Stereoscopic Image Quality Assessment

NR methods, owing to practicality, convenience and the advantage of assessing images' quality without any information of reference views, have aroused extensive attention in field of SIQA. So far, NR SIQA methods can be classified into three categories: 1) methods based on hand-crafted features; 2) methods based on deep neural networks; 3) others.

For conventional NR methods in SIQA, they usually extract features based on Natural Scene Statistics (NSS), HVS, etc. Then these artificial features are fed into learning models such as Support Vector Regression (SVR) [35] and k-nearest neighbors (KNN) [36] to get predicted quality score. In [37], Zhou *et al.* proposed a NR-SIQA model based on binocular combination and a learning machine. Firstly, two binocular combinations of stimuli are generated by different strategies. Various binocular quality-aware features of the combinations are extracted by local binary pattern operators. Then the features are mapped to subjective scores of distorted images by Extreme Learning Machine (ELM). Shao *et al.* [38] proposed a NR-SIQA model that can predict the quality score of distorted views by combining the feature prior and feature-distribution. SVR is used to characterize the feature-prior, and sparsity regularization is applied to present feature distribution. In [39], Liu *et al.* proposed a novel blind stereoscopic image quality assessment model with segmented monocular features and perceptual binocular features. A Simple Linear Iterative Clustering (SLIC) segmentation is applied to obtain superpixel of left and right view respectively. Binocular signals including summation signal, difference signal and cyclopean signal can be obtained from the views simultaneously. Finally, monocular and binocular features are extracted from the signals and SVR is adopted to get the final prediction result. Considering monocular visual properties and binocular interaction, Liu *et al.* [40] proposed a NR stereoscopic image quality assessment method based on SVR. Monocular and binocular features including color, luminance, summation and difference from left and right images are extracted. Then SVR is applied to combine the features and get final assessment results. In [41], Su *et al.* proposed a no-reference SIQA framework utilizing both univariate and generalized bivariate NSS models. A convergent cyclopean image can be obtained from both left and right views. Then univariate NSS features, bivariate and correlation NSS features are extracted from the cyclopean image. Finally, SVR is applied to regress the features and acquire the final predicted score. Fang *et al.* [42] proposed a no-reference quality evaluator of stereoscopic images based on monocular and binocular visual properties. Monocular and binocular characteristics including intensity, structure and depth can be obtained from left and right images. SVR is further employed to gain final assessment results. In [43], Zhou and Yu proposed a binocular response based no-reference SIQA method. Binocular responses including Binocular Energy Response (BER) and Binocular Rivalry Response (BRR) of distorted stereo pairs can be obtained and employed to produce quality-predictive features. Then the characteristics are applied to drive the final quality score by KNN. Messai *et al.* [44] proposed a no reference SIQA network using a neural network Adaptive Boosting (AdaBoost). Cyclopean views of distorted left and right images can be acquired to extract features including gradient magnitude,

relative gradient orientation, and relative gradient magnitude. Then AdaBoost is employed to produce the final quality score. However, attributed to the limitation of hand-crafted features, for example, high complexity and imprecision of regression models, these conventional NR methods can not achieve the best consistency with HVS and the improvements of the measures are rather slow.

With development of artificial intelligence, CNN gradually emerges into researchers' sight. Considering the advantages of CNN, such as extracting feature automatically, processing high-dimensional data conveniently and so on, some researches devote tremendous efforts to apply CNN into SIQA. In [8], Oh *et al.* proposed a deep no-reference image quality assessment model in terms of local to global feature aggregation. For stereoscopic pairs, they are divided into patch pairs with the same size of 32×32 at first. These pairs are fed into a network to obtain a series of local features. Then the features are aggregated to produce global features and predict the final SIQA score. Zhang *et al.* [9] proposed two SIQA models based on CNNs, one-column and three-column CNN. Based on left and right views, difference map can be acquired. Then the two views and the difference map are segmented into a series of patches with the same size as [8]. For one-column CNN model, only patches of the difference map are taken as input. Multilayer Perception (MLP) is further employed to summarize the learned characteristics into a final SIQA quality score of the stereo pair. Different from one-column model, the three-column model takes patches of both stereo pairs and difference map as input. The next steps are similar to one-column model. In [10], Shen *et al.* proposed a NR SIQA method based on global and local content characteristics. A primary sub-network is employed to extract low-level features of left and right patches firstly. Then two local feature enhancement sub-networks and a global feature fusion sub-network are applied to extract various characteristics. Finally, the outputs of sub-networks are concatenated and mapped to the final quality score via two FC layers. Fang *et al.* [11] proposed a no-reference quality assessment method for stereoscopic images by deep convolutional neural network. For left and right views, a Siamese Network is employed to extract high-level semantic features. Then the features are concatenated to produce the final quality score of the stereo pair. In [12], Shi *et al.* proposed a three-column multi-mask CNN model, taking the left view, right view and the registered distortion representation as inputs independently, to predict image quality and distortion types simultaneously. Zhou *et al.* [13] proposed a NR SIQA network StereoQA Net. The inputs of the network are multiple patches as [9]. The network includes four parts, two primary networks and two aided sub networks. The latter sub networks are introduced to extract features from the difference maps and fusion maps, which are calculated based on the second and the fifth convolution layers in the primary networks. Through the sub networks, a series of features can be obtained. The features are then concatenated to generate patch based qualities. Finally, the scores are applied to produce the final SIQA quality via a average pooling strategy. In [14], Ding *et al.* proposed a two-column dense CNN for SIQA. Cyclopean views and disparity images of distorted stereo pairs are obtained and taken as inputs of the CNN. Sim *et al.* [15] proposed a blind stereoscopic image quality evaluator based on binocular semantic and quality channels. Distorted stereoscopic pairs are transformed into patch pairs with the same size 224×224 at first. A cross fusion strategy is applied to model the process in the V1 visual cortex, and the Multi-scales Pooling (MSP) is utilized to integrate context information of different sub-regions for effective global feature extraction. Finally, local and global features are pooled into the final assessment score by a weighted pooling strategy.

TABLE I
CHARACTERISTICS OF MAINSTREAM CNN-BASED SIQA METHODS

Method	Data insufficiency	Training samples*	Binocular interaction	Fusion with learned weights	Cross convolution
Oh [8]	1	0	0	0	0
Zhang [9]	1	1	0	0	0
Shi [12]	1	1	0	0	0
Sim [15]	1	2	0	0	0
Fang [11]	1	1	1	0	0
Zhou [13]	1	1	1	0	0
Shen [10]	1	1	0	1	1
StereoIF-Net	1	1	1	1	1

* 0: patches with scores computed via FR methods; 1: patches with the same scores as related image; 2: entire image

In addition, Deep Brief Network (DBN) and dilation convolution are also considered in NR SIQAs. In [45], Yang *et al.* proposed a DBN based NR SIQA method. Three DBNs are employed to combine the features extracted from left image, right image and Depth Perception Map (DPM) respectively, and three quality scores can be obtained. Then the scores are aggregated to the final quality assessment result. Shao *et al.* [46] proposed a Blind deep Quality Evaluator (DQE) for stereoscopic images (denoted by 3D-DQE) through DBNs. In [47], Zhao *et al.* proposed a NR SIQA network based on Dilation CNN (DCNN). Cyclopean images from left and right views can be obtained. Then these cyclopean images are fed into three hierarchical multi-scale units and a convolutional layer to get the final quality score.

In order to clearly present the characteristics of some mainstream CNN-based SIQA methods, five critical factors (i.e. data insufficiency, homogeneous patch score assignment, binocular interaction, fusion with learned weights and cross convolution) are taken into account, as shown in Table I. In the second column, training samples in these methods are illustrated, where "0" means patches with computed scores via FR methods, "1" indicates patches with the same scores as related image; "2" represents entire image. In other columns, the symbol "1" indicates the factor has been considered in the method, otherwise "0." From the table, it can be found that the factor of data insufficiency is included in all compared methods, and different styles of training samples are adopted. Fang *et al.*'s [11] and Zhou *et al.*'s metrics [13] take binocular interaction into account. The former extracts interactive features by convolutional layers while the latter through summation and difference calculation between the input feature maps, but cross convolution is still not considered in these two methods. In the feature extraction stage, Shen's network does not take the binocular interaction into account, that is, it processes left and right views separately in

two local feature enhancement sub-networks. Shen's network also considers the cross convolution, in which the left and right feature maps generated in the primary sub-network are equally divided into two parts and then simply concatenated together. This cross convolution strategy is much simpler than the cross convolution module in the proposed StereoIF-Net. Cross convolution in the StereoIF-Net includes group division, subgroup summation/difference and subgroup concatenation, and is implemented in four BIMs to simulate the binocular interaction of the HVS.

III. Proposed StereoIF-Net for no-Reference Stereoscopic Image Quality Assessment

In this paper, we propose a no-reference stereoscopic image quality assessment network (i.e. StereoIF-Net) based on binocular visual characteristics of the HVS. The proposed StereoIF-Net is a hierarchical network aiming to model the complicated human visual cortex structure in V1-V5. As shown in Fig. 1, the proposed StereoIF-Net is composed of two convolution layers Conv-1 and Conv-2, four binocular interaction modules BIMs, a Binocular Fusion Module (BFM) and two Dense layers (Dense-512 and Dense-1). For Conv-1 and Conv-2, 64 kernels with the size of 3×3 and the stride of 1 are applied to filter both left and right patches, which can extract low-level visual features of input signals. **The selection of kernel sizes is on empirical, which is simple and effective as used in [10] and VGG network [48]**. Then four hierarchical BIMs are employed to simulate the binocular interaction of HVS in V2-V5. After BIM-4, a BFM based on automatic weight assignment is used to produce fusion characteristics of BIM-4. Finally, the final quality regression of each input patch pair can be achieved by two dense layers including Dense-512 and Dense-1. Through the proposed StereoIF-Net, a series of patch quality scores can be obtained and used to produce the final quality score of stereo pair via average pooling.

A. Pre-Processing of Stereoscopic Images

For SIQAs based on CNN, the insufficiency of training data is a very critical problem. The common method to solve the problem is patch based preprocessing of stereoscopic images [11]–[13]. In this paper, we use the identical way to divide stereoscopic pairs into a series of patch pairs with the size of 32×32 before they are fed into the proposed StereoIF-Net. We conduct experiments on LIVE 3D database Phase I [49], Phase II [50], IVC SIQA dataset [51] and Waterloo-IVC SIQA database [52]. Finally, we get 80, 300 stereo patch pairs for LIVE Phase I, 79, 200 stereo patch pairs for LIVE Phase II and 23, 040 stereo patch pairs for IVC SIQA dataset. For Waterloo-IVC SIQA database, 446, 490 and 910, 800 stereo pairs can be obtained on Phase I and Phase II respectively. Then we normalize the patch pairs for the training and testing processes [12].

B. Binocular Interaction Module (BIM)

In this paper, we propose a Binocular Interaction Module (BIM) based on cross convolution to simulate complicated cross transmission of visual signals in different cortex regions. As shown in Fig. 1, four BIMs with different channel numbers, e.g. BIM-1, BIM-2, BIM-3 and BIM-4, are designed to simulate receptive processes in various regions. They have a hierarchical structure and each has two inputs and two outputs. In this paper, for BIM-1 and BIM-2, 768 output channels are assigned to model receptive fields in V2 and V3. In V4 and V5 region, neural cells have a larger receptive field than that in V2 and V3, therefore, for BIM-3 and BIM-4, 1536 output channels are set to imitate the phenomenon.

The architecture of each BIM can be seen in Fig. 2. The fundamental structure of proposed BIMs includes ConvNac module, ConvAc module, cross convolution and feature concatenation. The input of BIMs can be denoted as I_m-i (where $m \in \{l, r\}$ and i represents the index of BIM). In order to retain original perceptual information during the transmission of visual signals in different regions, the input I_m-i is followed by a ConvNac module, in which the inputs are just filtered and normalized, without any activation transformation. As shown in Fig. 3a, the ConvNac module is composed of a convolutional layer (marked as **Conv1**) with the kernel size of 3×3 and a batch normalization operation, but without activation operation. In the four BIMs, the number of convolutional kernels in ConvNac is distinct. In this paper, the kernel number of ConvNac in BIM-1 to BIM-4 is assigned to 384, 384, 768 and 768 respectively. Through the ConvNac module, the output I_m^{conv1} $(m \in \{l, r\})$ can be obtained for subsequent aggregation.

To assist neural cells to further extract visual features, a ConvAc module is applied to filter I_m-i. The architecture of the module is depicted in Fig. 3b. Compared with ConvNac module, the ConvAc module also includes a convolutional layer (denoted as **Conv2**) and batch normalization. But the normalized features are then transformed by Rectified Linear Units (ReLU) [53], and the results can be denoted as I_m^{conv2} $(m \in \{l, r\})$. The number of convolutional kernels in ConvAc is also diverse in different BIMs. In this paper, the kernel number of **Conv2** in BIM-1 to BIM-4 is assigned to 192, 192, 384 and 384 respectively.

To simulate binocular interaction of left and right visual signals, **cross convolution** including group division, subgroup summation, difference and concatenation is then proposed. The strategy divides I_m^{conv2} (including I_l^{conv2} and I_r^{conv2}) into different subgroups, and extracts summation and difference features in these subgroups, therefore more interactive information can be further obtained after subgroup convolution.

Group division: After the ConvAc module, the size of I_m^{conv2} is $b * w * h * c$ (where b, w, h, and c denote batch size, width, height and channels of I_m^{conv2} respectively.). We divide the channels of I_l^{conv2} into six subgroups with the same size $b * w * h * \frac{c}{6}$, and each one can be denoted as $I_l^t (t = 1, 2, \ldots, 6)$. Similarly, the same operations are adopted to I_r^{conv2} and $I_r^t (t = 1, 2, \ldots, 6)$ can be obtained. As shown in Fig. 2, I_l^t and I_r^t are marked by red rectangles and blue rectangles respectively.

Subgroup summation and difference: For each subgroup, cross features such as summation and difference are then

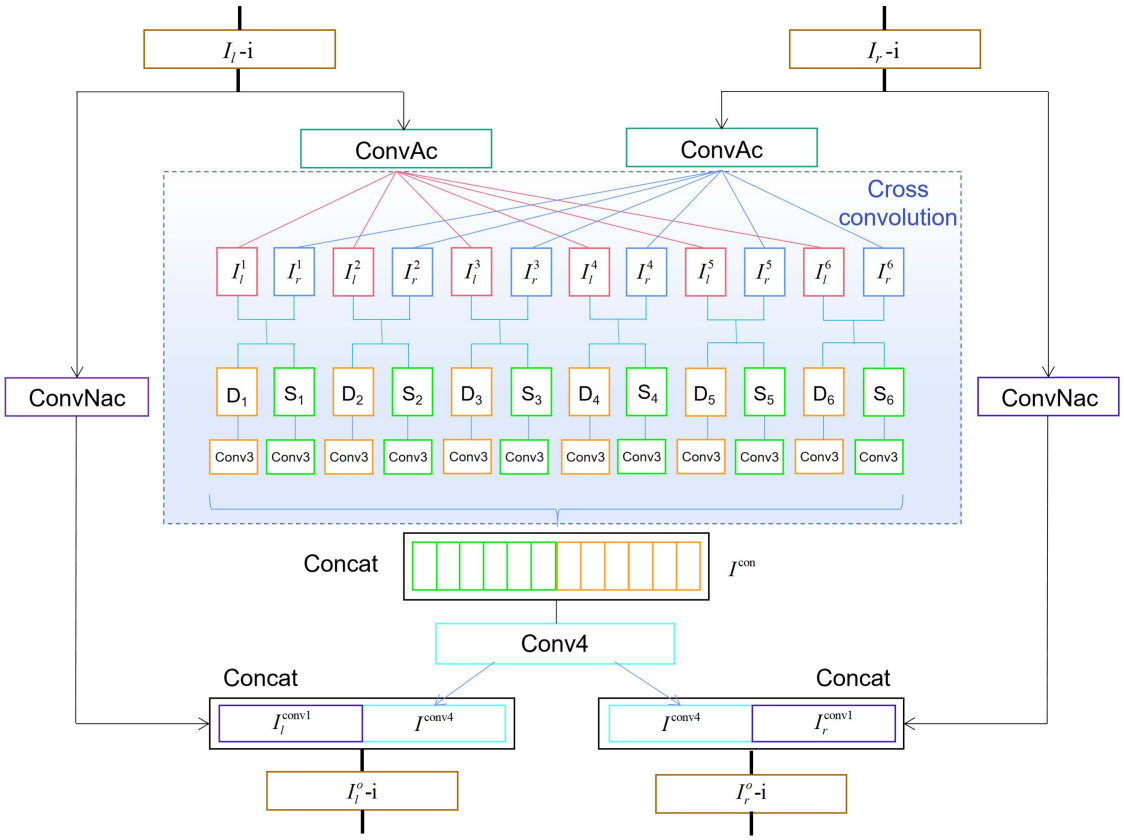
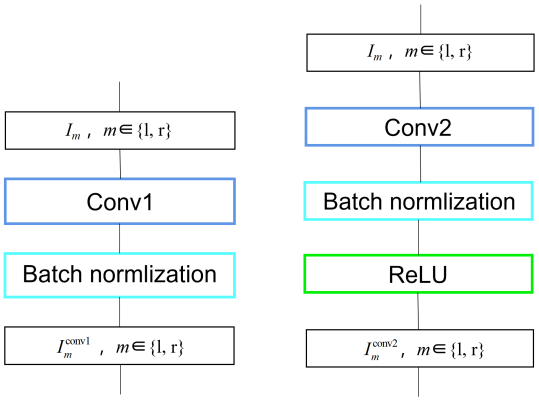


Fig. 2. The proposed binocular interaction module (BIM).



(a) ConvNac module (b) ConvAc module

Fig. 3. The ConvAc and ConvNac module.

extracted to acquire more interactive information. The summation and difference of each sub-group can be obtained as follows:

$$S_t = I_l^t + I_r^t \ (t = 1, \ldots, 6) \quad (1)$$

$$D_t = I_l^t - I_r^t \ (t = 1, \ldots, 6) \quad (2)$$

In Fig. 2, S_t and D_t are marked by green and orange rectangles respectively.

Subgroup concatenation: S_t and D_t are filtered by 1×1 kernels (named as **Conv3**) with a assigned amount and the results can be denoted as S_t^{conv3} and $D_t^{conv3} (t = 1, \ldots, 6)$. The kernel number of **Conv3** in BIM-1 to BIM-4 is set to 32, 32, 64 and 64 respectively. S_t^{conv3} and D_t^{conv3} are then concatenated as follows:

$$I^{con} = \text{concat}(S_1^{conv3}, \ldots, S_6^{conv3}, D_1^{conv3}, \ldots, D_6^{conv3}) \quad (3)$$

where concat$(\cdot)$ represents concatenation operation.

After the operation, I^{con} is convoluted by **Conv4** module and I^{conv4} can be acquired. The kernel number of **Conv4** in BIM-1 to BIM-4 is assigned to 384, 384, 768 and 768 respectively. Finally, the result is concatenated with I_l^{conv1} and I_r^{conv1} respectively, and the outputs of i-th BIM can be generated as follows:

$$I_l^o\text{-}i = \text{concat}(I_l^{conv1}, I^{conv4}) \quad (4)$$

$$I_r^o\text{-}i = \text{concat}(I^{conv4}, I_r^{conv1}) \quad (5)$$

C. Binocular Fusion Module (BFM)

The visual signal process of human has been illustrated explicitly in Section I. In this subsection, we mainly focus on the final parts of the process, that is the transmission of visual signals in the regions higher than V1-V5.

In SIQA, binocular fusion is an essential factor to be studied. Some FR SIQAs have taken it into account to simulate the visual process of human's brain [25], [27]. Fusion methods in these works mainly adopt averaging or manually assigned weights to fuse binocular visual signals. However, the man-made assignment of weights is time-consuming and inaccurate, which may lead to unexpected fusion results. Therefore, in this

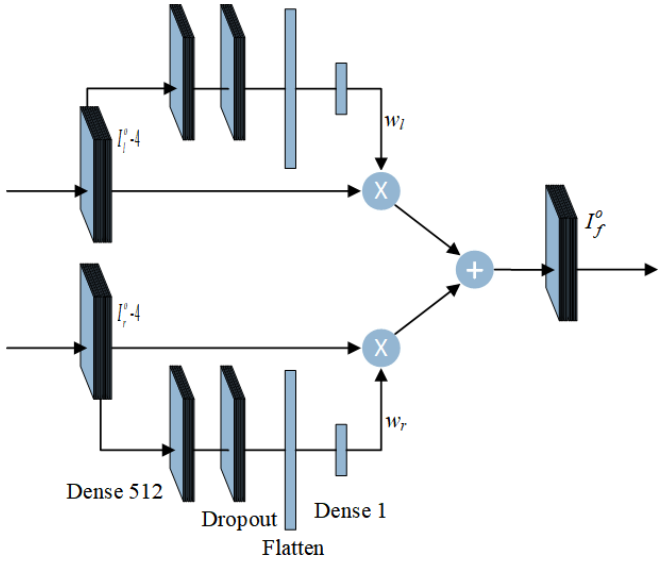


Fig. 4. The proposed binocular fusion module (BFM).

work, a Binocular Fusion Module (BFM) with automatic learned weights is proposed to model binocular fusion of the HVS in higher cortex layers, as shown in Fig. 4. In the BFM, weights can be automatically learned by the constructed network in the training stage, which can better simulate binocular fusion mechanism.

The BFM module includes a series of layers such as Dense, Dropout and Flatten layers. Firstly, the weight of left input can be obtained through a structure including Dense 512, Dropout, Flatten and Dense 1. So as the right one. Therefore, two weights can be acquired and denoted as w_l and w_r. Let I_l^o-4 and I_r^o-4 represent the outputs of BIM-4, which can also be seen as the inputs of the proposed BFM. The output of BFM can be acquired as follows:

$$I_f^o = I_l^o\text{-}4 \times w_l + I_r^o\text{-}4 \times w_r \tag{6}$$

As shown in Fig. 1, based on the obtained fusion result I_f^o, two dense layers Dense-512 and Dense-1 are adopted to produce the final patch quality regression scores.

D. Local Quality Pooling

As shown in Table I, characteristics of some mainstream CNN-based SIQA methods are investigated, and most of them adopt the homogeneous patch score assignment. According to this homogeneous assignment, average pooling is usually adopted to fuse the predicted quality scores of patches, as done in our manuscript and Zhou's metric [13]. For local quality scores of a special stereo pair, the average quality score can be obtained as follows:

$$Q = \frac{1}{N_p} \sum_{j=1}^{N_p} f(x_j; \omega) \tag{7}$$

where $f(x_j; \omega)$ represents the predicted quality score of j-th patch x_j in the stereo pair and N_p is the number of patches. ω denotes the weights of network. In this paper, we divide every stereo pair into patches with the size of 32×32, therefore N_p can be obtained as follows:

$$N_p = \lfloor \frac{M}{32} \rfloor \times \lfloor \frac{N}{32} \rfloor \tag{8}$$

where M and N represent the width and height of the stereo pairs. In our experiments, $M = 640$, $N = 360$ on LIVE 3D database and therefore $N_p = 220$ while $M = N = 512$ on IVC SIQA database and therefore $N_p = 256$.

There are also some other strategies on the subjective score assignment and patch pooling, for example, based on entropy [34], saliency [54], activity [55], texture [56], etc. For the sake of simplicity, we just adopt the homogeneous assignment and pooling, and focus on the improvement of the network construction.

E. Patch Pair Based Training of StereoIF-Net

In our experiments, a patch pair based training strategy is applied to solve the problem of data insufficiency. We divide stereoscopic pairs into a series of patch pairs with the same size of 32×32. Inspired by [9]–[13], we assign the ground truth score of source images as the label of corresponding patch pairs since images with a assigned noise are distorted homogeneously for each pixel.

To train the proposed StereoIF-Net, as done in most of CNN-based SIQA methods, such as Shen *et al.* [10], Shi *et al.* [12], Zhou *et al.* [13], Sim *et al.* [15] and Liu *et al.* [39], on each database such as LIVE 3D Phase I, Phase II, IVC SIQA database, WaterlooIVC database Phase I and Phase II, 80% of distorted stereoscopic image pairs are randomly selected as training set and the remaining 20% as testing set in our experiments. The k-fold cross validation (k = 10 in our experiments.) is used over these four databases and the mean value is adopted. We train the proposed network for 100 epochs to obtain the experimental results.

During the training process, a well-performed loss function can optimize backpropagation, accelerate the convergence of the network, and prevent over-fitting. Similar to [11], in this paper, a L1 loss function is employed to assess the performance of the proposed network. Suppose that x_k represents k-th patch pair obtained from the training dataset, the loss function can be described as follows:

$$\min_\omega \frac{1}{N} \sum_{k=1}^{N} |f(x_k; \omega) - y_k| \tag{9}$$

where $f(x_k; \omega)$ is the predicted quality score of the k-th input x_k, and y_k is the ground truth of the k-th pair. N is the number of training patch pairs.

In order to accelerate network's convergence during training process, we optimize the regression object by Adam [57] with initial learning rate 1×10^{-4} and a mini-batch of 64. Similar to ConvAC module, ReLU is also adopted as the activation function for other layers. Compared with conventional activation functions such as sigmoid and tanh, ReLU can not only solve the problems such as gradient explosion and gradient disappearance but also make gradient descent and

backpropagation more effectively. Besides, before the dense layers of the proposed StereoIF-Net, a dropout strategy is used to prevent overfitting. In our experiments, we randomly assigned the outputs of the neurons with 0.5 or 0.35. By using the technology, the network can avoid overfitting effectively and have a better generalization.

IV. Experiments and Analysis

A. SIQA Databases

A set of databases such as LIVE 3D Phase I [49], LIVE 3D Phase II [50], IVC SIQA database [51] and Waterloo-IVC SIQA database [52] have been available for SIQA. In this paper, we adopt these four datasets to assess the performance of the proposed StereoIF-Net and conduct cross test experiments.

1) LIVE 3D Phase I: 20 reference stereo pairs are used. Each pair is distorted by five different distorted types including JPEG, JPEG2000 (JP2K), Gaussian White Noise (WN), Gaussian blur (Blur) and Raleigh Fast Fading (FF) at different extent. Left and right patches have symmetric distortion. Here, there are 365 distorted pairs in the database including 80 pairs each for JPEG, JP2K, WN, FF and 45 for Blur. Moreover, the database also provides the corresponding Differential Mean Opinion Score (DMOS) value for each pair. The DMOS values are limited in [0, 80] and higher value means lower perceptual quality.

2) LIVE 3D Phase II: The database consists of 8 reference images and 360 distorted images with co-registered human scores in terms of DMOS. The distorted types are the same to Phase I. But for each type there are nine different distorted levels, where one third of the levels are symmetric and the remaining is asymmetric. In summary, there are 120 symmetric stereo pairs and 240 asymmetric stereo pairs with different distorted degrees.

3) IVC SIQA Database: 6 different reference stereoscopic pairs are considered in the database. Three distortions including JPEG, JP2K and Blur are applied to distort each pair and produce 15 distortion versions. Therefore, the database is composed of 96 stereoscopic pairs including 6 reference stereoscopic pairs and 90 distorted pairs. Similar to LIVE 3D Phase I, left and right images of each pair have symmetric distortion. In addition, the corresponding DMOS values of stereoscopic pairs on the database are also equipped. The range of DMOS values is [0, 80] and lower value symbolizes a better perceptual quality.

4) WaterlooIVC Database: The database is composed of two phases including Phase I and Phase II. Phase I is created from 6 pristine stereoscopic image pairs. Each reference image is altered by three types of distortions (i.e. WN, Blur and JPEG). Totally, there are 78 symmetrical and 252 asymmetrical distorted stereo pairs with corresponding DMOS values. Phase II is originated from 10 pristine stereoscopic image pairs and distorted by the same distortions. There are totally 460 distorted stereo pairs with individual subjective DMOS values.

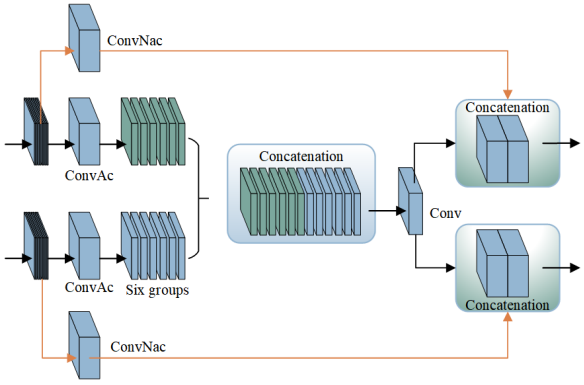


Fig. 5. The framework with BCMs.

B. Performance Index

In this paper, we apply three widely used indicators including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC) and Root Mean Squared Error (RMSE) between subjective and objective scores to assess the performance of different methods.

PLCC indicates the linear dependence between the predicted scores and the ground truth targets while SROCC measure the monotonicity of two quantities. The ranges of them are both [0 1]. Higher PLCC and SROCC value represents better correlation between subjective and predicted scores. For RMSE, it is the distance between subjective scores and predictions. A lower value symbolizes a better performance of the model.

C. Verification of BIM

In order to illustrate the performance of the proposed BIM, the comparison experiments are conducted by introducing a Binocular Concatenation Module (BCM) in the proposed StereoIF-Net, instead of the BIMs. The architecture with BCMs can be found in Fig. 5. Different from BIM, BCMs based structure concatenates the groups directly rather than adopting the operations such as difference and summation. Apart from the replacement of BIM, the training process and other settings are the same to the proposed StereoIF-Net. The network training is conducted on LIVE 3D dataset. The comparison results and analysis of the experiments can be seen in Sec. IV-E.

To test the relationship between group numbers in BIM and the performance of StereoIF-Net, a variety of comparison experiments are conducted. In the group division module, the channels of I_l^{conv2} and I_r^{conv2} can not be divided with no remainder if group number is set to odd number, therefore in this paper, even group number including 2, 4, 6, 8, 12 are adopted for performance comparison. The results can be seen in Table II. From the table, we can find that when the group number is set to 6 the network can achieve a better performance than other settings. We can also find that when group number is more than 6, StereoIF-Net's performance is steady and fluctuate slightly. For the networks with the group number more than 6, they have reached their own best performance in extracting various visual characteristics.

TABLE II
PERFORMANCE COMPARISON WITH DIFFERENT GROUPS IN BIMS

Group number	2	4	6	8	12
PLCC	0.9388	0.9500	**0.9779**	0.9750	0.9745
SROCC	0.9323	0.9435	**0.9656**	0.9649	0.9626
RMSE	5.6151	5.0904	**2.6077**	4.0306	4.3656

TABLE III
PERFORMANCE COMPARISON WITH DIFFERENT NUMBER OF BIMS

BIM number	1	2	3	4	5	6	7
PLCC	0.9398	0.9441	0.9593	**0.9779**	0.9710	0.9650	0.9679
SROCC	0.9349	0.9410	0.9539	**0.9656**	0.9505	0.9620	0.9629
RMSE	5.5713	5.3128	5.1217	**2.6077**	3.9098	4.1325	4.1921

TABLE IV
PERFORMANCE COMPARISON WITH DIFFERENT KERNEL NUMBER SCHEMES

Scheme	Scheme-1	Scheme-2	StereoIF-Net
PLCC	0.9611	0.9656	**0.9779**
SROCC	0.9576	0.9601	**0.9656**
RMSE	4.6436	4.6031	**2.6077**

The number of BIMs is also a vital factor to be considered. As is shown in Table III, in order to explore how the number of BIMs affect networks' performance, various BIM's numbers such as 1, 2, 3, 4, 5, 6, 7 are applied to establish the comparison experiments. When the number of BIMs is less than 4, the performance of the StereoIF-Net is increasing with the BIM number. However, when the number of the modules is more than 4, the performance of the models tends to be worse, which can be ascribed to the superfluity of BIMs. The phenomenon shows that the selection of 4 BIMs enables the proposed StereoIF-Net to achieve the best performance.

In addition, to prove the effectiveness in simulating various receptive fields, a series of comparison experiments are also conducted. In our experiments, three schemes with different channel assignment in BIM-1 to BIM-4 are compared. For **Scheme-1**, the channels for the outputs of BIM-1 to BIM-4 are all designed to 768, while assigning 384 channels to Conv1 and Conv4 separately. For **Scheme-2**, 768 channels are assigned to Conv1 and Conv4 and therefore there are 1536 output channels in BIM-1 to BIM-4. The experimental results can be found in Table IV. From the table, it can be found that due to proper channel designing, the proposed StereoIF-Net has a better performance than other schemes, which can prove the effectiveness of StereoIF-Net's channel designing scheme.

D. Verification of BFM

In this subsection, a diversity of comparison experiments are set up on LIVE 3D database to verify the effectiveness of both BFM and automatic weight assignment. The comparison experiments are designed as follows. To prove the performance of BFM, we concatenate two outputs of BIM-4 directly rather than using BFM before the last Dense-512. Further, to verify the effectiveness of automatically trained weights, the weights of left and right input are both set to 0.5 artificially. The experimental results can be seen in Table V. From the table, it can be found that StereoIF-Net with learned weight based BFM achieves the best performance on the entire LIVE 3D database.

TABLE V
PERFORMANCE COMPARISON FOR BFM ON LIVE 3D DATABASE

Method	LIVE 3D Phase I		LIVE 3D Phase II	
	PLCC	SROCC	PLCC	SROCC
With direct concatenation	0.9200	0.9123	0.9210	0.9201
With artificial weights	0.9479	0.9412	0.9435	0.9399
With learned weights	**0.9779**	**0.9656**	**0.9717**	**0.9529**

TABLE VI
THE RESULTS OF ABLATION EXPERIMENTS ON LIVE 3D DATABASE

Method	LIVE 3D Phase I		LIVE 3D Phase II	
	PLCC	SROCC	PLCC	SROCC
Base	0.8500	0.8446	0.8333	0.8301
Base+BCMs	0.8955	0.8910	0.8800	0.8776
Base+BCMs+BFM	0.9312	0.9240	0.9232	0.9190
Base+BIMs+BFM	**0.9779**	**0.9656**	**0.9717**	**0.9529**

E. Ablation Experiments

In order to verify the effectiveness of proposed StereoIF-Net, in this subsection, the ablation experiments on LIVE 3D database are conducted. In Table VI, several ablation results of the proposed metric are listed. In the table, the "Base" means the structure consisted of Conv-1, Conv-2, the layer concatenated by the outputs of Conv-1 and Conv-2, Dense-512 and Dense-1. From the table, it can be found that by adding various parts such as BCMs, BIMs and BFM, the algorithm performance keeps improving, which can reflect the effectiveness of the proposed BIMs and BFM. Especially, from the third and forth lines in the table, it can be found that the proposed StereoIF-Net with cross convolution based BIMs achieves better performance than the network with BCMs. We can draw the conclusion that the cross convolution with subgroup summation and difference strategy plays a significant role in modeling the binocular interaction phenomenon in different regions. Further, due to the introduce of cross convolution, the PLCC and SROCC values of the StereoIF-Net have been improved more than 0.04.

F. Effects of Training Parameters

During the training and testing process, different parameters may influence the performance of proposed network. Therefore, a diversity of experiments are set up to examine how they affect the performance of StereoIF-Net.

1) Kernel Size: To explore the effect of kernel size, in our experiments, we take the Conv-1 and Conv-2 layers as the comparison objects and use different kernel sizes such as 1×1, 3×3, 5×5, 7×7 to conduct comparison experiments. The results can be seen in Table VII. From the table, it can be found that proposed StereoIF-Net is not sensitive to kernel size. Hence, we use 3×3 convolution kernels in the network, considering both performance and computation complexity.

2) Learning Rate: During the training process, learning rate is an important parameter affecting the performance of the proposed StereoIF-Net. In our experiments, we test a series of learning rates such as 10^{-2}, 10^{-3}, 10^{-4} and 10^{-5}. From Table VIII, we can also conclude that the proposed model

TABLE VII

PERFORMANCE COMPARISON OF DIFFERENT KERNEL SIZES IN CONV-1 AND CONV-2

Kernel size	LIVE 3D Phase I		LIVE 3D Phase II	
	PLCC	SROCC	PLCC	SROCC
1x1	0.9745	0.9610	0.9700	0.9490
3x3	**0.9779**	**0.9656**	**0.9717**	**0.9529**
5x5	0.9762	0.9660	0.9710	0.9500
7x7	0.9770	0.9646	0.9705	0.9510

TABLE VIII

PERFORMANCE COMPARISON OF DIFFERENT LEARNING RATES

Learning rate	LIVE 3D Phase I		LIVE 3D Phase II	
	PLCC	SROCC	PLCC	SROCC
10^{-2}	0.9710	0.9630	0.9690	0.9493
10^{-3}	0.9736	0.9599	0.9695	0.9506
10^{-4}	**0.9779**	**0.9656**	**0.9717**	**0.9529**
10^{-5}	0.9760	0.9618	0.9700	0.9498

TABLE IX

PERFORMANCE COMPARISON OF DIFFERENT PATCH SIZES

Patch size	LIVE 3D Phase I		LIVE 3D Phase II	
	PLCC	SROCC	PLCC	SROCC
24×24	0.9656	0.9550	0.9600	0.9490
32×32	**0.9779**	**0.9656**	**0.9717**	**0.9529**
48×48	0.9763	0.9660	0.9695	0.9510

is not sensitive to learning rate. The setting of 10^{-4} can enable the proposed StereoIF-Net to achieve best regression performance.

3) Patch Size: In order to explore how different patch sizes influence the performance of the proposed network, as shown in Table IX, different patch sizes such as 24×24, 32×32 and 48×48 are considered in our experiments. The table shows that the patch size with 32×32 has a better performance than other models and therefore it is applied to build the proposed StereoIF-Net.

4) Iterative Epoch: To explore how StereoIF-Net's performance is affected by iterative epochs, we conduct plenty of experiments on LIVE 3D, IVC and WIVC SIQA datasets under the premise of changing the iterative epoch and fixing the rest of proposed architecture. The visualized results can be seen in Fig. 6. From the figure, we can observe that the performance of StereoIF-Net tends to be better with increasing of iterative epochs. Further, when the iterative epoch is more than 90, the performance improves slightly. As shown in Fig. 6, the training performance of StereoIF-Net on these three datasets is illustrated. For example, the PLCC and SROCC values of LIVE 3D Phase I is about 0.9800 and 0.9660 in the training stage. From Table X, it can be found that in the testing the PLCC and SROCC values of LIVE 3D Phase I is 0.9779 and 0.9656. These results can demonstrate that the network is not overfitting. Such training and testing result comparison can also be found from Table XIV and XV.

G. Comparison of StereoIF-Net With Existing Methods

All experiments are conducted on LIVE 3D, WaterlooIVC and IVC SIQA databases, and three indices including PLCC, SROCC and RMSE are applied to assess the performance of various methods.

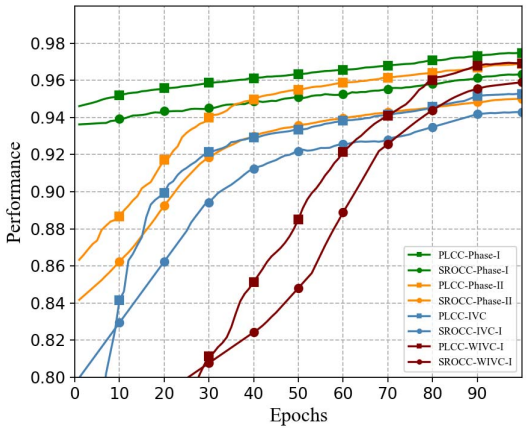


Fig. 6. The training performance visualization on different databases.

In Table X, in order to verify the performance of the proposed StereoIF-Net, seventeen existing prevalent models including four FR models (Khan and Channappayya [24], Liu *et al.* [25], Chen [58], and Chen and Zhao [59]), and thirteen NR models (Oh *et al.* [8], Zhang *et al.* [9], Shen *et al.* [10], Fang *et al.* [11], Shi *et al.* [12], Zhou *et al.* [13], Sim *et al.* [15], Liu *et al.* [39], Su *et al.* [41], Fang *et al.* [42], Zhou and Yu [43], and Messai *et al.* [44]) are taken into account as comparison objects. From Table X, it can be found that most of FR models are worse than NR ones and the StereoIF-Net. The phenomenon can be attributed to the facts: 1) these FR metrics are all based on artificial feature extraction, which may be not as effective; 2) for FR models, all visual features related to the HVS can not be comprehensively considered. Therefore, the limitations of feature extraction can lead to poor assessment results of FR models.

In contrary, with the CNN advantages such as automatic feature extraction, NR models have better performance than FR ones. But for different NR SIQAs, they have shown their own particularities, which is bound up to various network structures, pre-processing, optimizers, parameter settings and so on. For Yang *et al.*'s method [45], though DBN network is adopted, a set of features are still extracted artificially, therefore it may not achieve the best performance. Shi *et al.*'s network [12] can extract features automatically, but some significant visual features such as binocular interaction are not considered. Compared with Liu's *et al.* [39], Shen *et al.*'s [10], Oh *et al.*'s [8], Zhang *et al.*'s [9], Fang *et al.*'s [11], Su *et al.*'s [41], Fang *et al.*'s [42], Zhou *et al.*'s [43], and Messai *et al.*'s method [44], due to the automatic feature extraction and the comprehensive consideration of complicated visual features such as binocular interaction, StereoIF-Net has a better performance than these metrics. From Table X, it can be found that apart from SROCC on LIVE 3D Phase II, other indices of the StereoIF-Net are all higher than other compared NR and FR models. Further, compared with Zhou's method [13] (the second best on Phase I), PLCC has been improved from 0.9730 to 0.9779 on Phase I and 0.9570 to 0.9717 on Phase II. Compared with Sim's method [15] (the

TABLE X
OVERALL PERFORMANCE ON LIVE 3D DATABASE

	Methods	LIVE 3D Phase I			LIVE 3D Phase II		
		PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
FR models	Chen [58]	0.9161	0.9153	6.5740	0.9067	0.9068	4.7587
	Khan [24]	0.9270	0.9160	-	0.9320	0.9220	-
	Liu [25]	0.9430	0.9402	5.4238	0.8417	0.8317	6.0946
	Chen [59]	0.9390	0.9290	6.1580	0.8650	0.8590	6.2630
NR models	Liu [39]	0.9580	0.9490	5.0690	0.9350	0.9330	4.0140
	Shen [10]	0.9720	0.9620	-	0.9530	0.9510	-
	Zhang [9]	0.9470	0.9430	5.3360	0.9110	0.9010	-
	Oh [8]	0.9430	0.9350	-	0.8630	0.8710	-
	Fang [11]	0.9570	0.9460	-	0.9460	0.9340	-
	Yang [45]	0.9556	0.9437	4.9171	0.9335	0.9206	4.0053
	Shi [12]	0.9630	0.9360	4.1610	0.9610	0.9480	2.6750
	Su [41]	-	-	-	0.9050	0.9130	4.5670
	Fang [42]	0.9510	0.9320	-	0.9310	0.9190	-
	Zhou [43]	0.9280	0.8870	6.0250	0.8610	0.8230	5.7790
	Messai [44]	0.9390	0.9300	5.6050	0.9220	0.9130	4.3520
	Zhou [13]	0.9730	0.9650	3.6820	0.9570	0.9470	3.2700
	Sim [15]	0.9697	0.9622	3.9441	0.9619	**0.9550**	3.0422
Proposed model	StereoIF-Net	**0.9779**	**0.9656**	**2.6077**	**0.9717**	0.9529	**2.2771**

TABLE XI
PLCC PERFORMANCE OF DIFFERENT DISTORTIONS ON LIVE 3D DATABASE

	Criteria	Chen [58]	Khan [24]	Liu [25]	Chen [59]	Liu [39]	Shen [10]	Zhang [9]	Oh [8]	Fang [11]	Yang [45]	Su [41]	Fang [42]	Zhou [43]	Messai [44]	Zhou [13]	Sim [15]	Proposed
Phase I	JP2K	0.9166	0.9510	0.9423	0.8260	0.9380	0.9840	0.9260	0.9130	0.9750	0.9424	-	0.9630	-	0.9260	0.9880	0.9939	**0.9980**
	JPEG	0.6356	0.7110	0.7315	0.5210	0.8100	0.9060	0.7400	0.7670	0.7530	0.8243	-	0.7670	-	0.6680	0.9160	0.9574	**0.9603**
	WN	0.9353	0.9470	0.9463	0.9520	0.9660	0.9470	0.9440	0.9100	0.9730	0.9536	-	0.9620	-	0.9410	0.9880	0.9943	**0.9965**
	FF	0.7583	0.8580	0.8658	0.9410	0.8550	0.9390	0.8830	0.9540	0.8680	0.7893	-	0.8940	-	0.8450	0.9650	0.9874	**0.9962**
	Blur	0.9418	0.9590	0.9530	0.9390	0.9560	0.9880	0.9300	0.9500	0.9530	0.9634	-	0.9820	-	0.9350	0.9740	0.9946	**0.9956**
Phase II	JP2K	0.8426	0.9270	0.8701	0.8360	0.9360	0.9560	-	0.8650	0.9750	0.8855	0.8470	0.9230	-	0.8420	0.9050	0.9021	**0.9999**
	JPEG	0.8422	0.8930	0.8758	0.8580	0.8670	0.8250	-	0.8210	0.9520	0.8671	0.8880	0.8950	-	0.8370	0.9330	0.4090	**0.9697**
	WN	0.9602	0.9700	0.9325	0.9270	0.9690	0.9540	-	0.8360	0.9720	0.8873	0.9530	0.9760	-	0.9430	0.9720	0.9082	**0.9933**
	FF	0.9097	0.8990	0.9218	0.8920	0.9590	0.9640	-	0.8150	0.9290	0.9162	0.9440	0.9320	-	0.9250	0.9940	0.8670	**0.9955**
	Blur	0.9650	0.9780	0.9430	0.9460	0.9870	0.9880	-	0.9340	0.9830	0.9877	0.9680	0.9890	-	0.9130	0.9550	0.9437	**0.9963**

TABLE XII
SROCC PERFORMANCE OF DIFFERENT DISTORTIONS ON LIVE 3D DATABASE

	Criteria	Chen [58]	Khan [24]	Liu [25]	Chen [59]	Liu [39]	Shen [10]	Zhang [9]	Oh [8]	Fang [11]	Yang [45]	Su [41]	Fang [42]	Zhou [43]	Messai [44]	Zhou [13]	Sim [15]	Proposed
Phase I	JP2K	0.8954	0.9070	0.9040	0.8350	0.8880	0.9650	0.9310	0.8850	-	0.8971	-	-	0.8240	0.8990	0.9610	0.9872	**0.9880**
	JPEG	0.5632	0.6060	0.6952	0.8300	0.7850	0.8790	0.6930	0.7650	-	0.7681	-	-	0.6140	0.6250	0.9120	**0.9531**	0.8943
	WN	0.9376	0.9380	0.9468	0.9230	0.9510	0.9210	0.9460	0.9210	-	0.9294	-	-	0.9150	0.9410	0.9650	0.9922	**0.9994**
	FF	0.6882	0.8090	0.8108	0.9070	0.8210	0.9000	0.8340	0.9440	-	0.6853	-	-	0.8670	0.7770	0.9170	**0.9800**	0.8996
	Blur	0.9283	0.9300	0.9294	0.9460	0.9170	0.9450	0.9090	0.9300	-	0.9167	-	-	0.9160	0.8870	0.8550	0.9928	**0.9988**
Phase II	JP2K	0.8334	0.9130	0.8642	0.8320	0.9090	0.9540	-	0.8530	-	0.8593	0.8450	-	0.7170	0.8420	0.8740	0.8148	**0.9756**
	JPEG	0.8396	0.8670	0.8640	0.8500	0.8250	0.8160	-	0.8220	-	0.8064	0.8180	-	0.5930	0.8370	0.7470	0.3150	**0.9312**
	WN	0.9554	0.9580	0.9230	0.9060	0.9460	0.9230	-	0.8330	-	0.8637	0.9460	-	0.8910	0.9430	0.9420	0.8876	**0.9719**
	FF	0.8890	0.8650	0.8977	0.8750	0.9380	0.9690	-	0.8780	-	0.8769	0.8990	-	0.8910	0.9250	0.9510	0.8164	**0.9562**
	Blur	0.9096	0.8850	0.9114	0.9490	0.9360	0.9510	-	0.8890	-	0.8341	0.9030	-	0.9030	0.9130	0.6000	0.9301	**0.9601**

second best on Phase II), the PLCC and RMSE of StereoIF-Net is higher, while the SROCC value is comparable.

In addition, in the field of SIQA, the KROCC criterion [60] is also used to reflect the monotonicity of subjective and objective scores. Therefore, we test the KROCC values of StereoIF-Net on LIVE 3D Phase I and Phase II, which are 0.8485 and 0.8158 respectively.

In Tables XI and XII, the comparison performances of different distortions on both LIVE 3D Phase I and Phase II are provided. From the tables, it can be found that in terms of the distortions such as JP2K, WN and Blur, the proposed StereoIF-Net has a better performance than other models. As to JPEG and FF distortion, PLCC of the proposed network is higher than other models. SROCCs of StereoIF-Net on JPEG and FF distortion is lower than Sim's method, which can be attributed to the fact that the extraction of semantic features leads to better feature representation of Sim's metric. Further, SROCC on JPEG distortion between the proposed model and Zhou*et al.*'s [13] is nearby. So as FF distortion. Comparing these five kinds of distortions, JPEG and FF has

TABLE XIII

COMPARISON PERFORMANCE OF SYMMETRIC AND ASYMMETRIC DISTORTION ON LIVE 3D PHASE II

Methods	Symmetric			Asymmetric		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Chen [58]	0.9380	0.9250	-	0.8750	0.8540	-
Zhang [9]	0.9121	0.9145	-	0.7625	0.7021	-
Zhou [13]	-	**0.9790**	-	-	0.9270	-
Shi [12]	**0.9700**	0.9280	-	**0.9530**	**0.9430**	-
Su [41]	-	**0.9370**	-	-	0.8490	-
Messai [44]	0.9300	0.9170	**4.6090**	0.9030	0.8980	**4.2160**
StereoIF-Net	**0.9833**	0.9217	**2.8656**	**0.9568**	**0.9312**	**2.7716**

TABLE XIV

COMPARISON PERFORMANCE ON IVC SIQA DATABASE

Methods	PLCC	SROCC	RMSE
Khan [24]	0.9177	0.9068	-
Chen [58]	0.6830	0.6760	-
Zhang [9]	0.7917	0.7644	-
Benoit [51]	0.5854	0.4523	-
Messai [44]	0.8450	0.8310	**11.7760**
Zhou [13]	**0.9450**	**0.9426**	-
StereoIF-Net	**0.9576**	**0.9458**	**2.9068**

more local degradation textures. Zhou's network has more dense layers than the SeteroIF-Net, which may lead to better local feature representation for images. Hence, the prediction to some specific patches are more accurate.

In the LIVE Phase II dataset, stereo pairs with both symmetric and asymmetric distortion are included. The StereoIF-Net is verified on symmetric and asymmetric distortios respectively, and the results are shown in Table XIII. In the table, the metric values in the first and second rank are marked in bold. From the table, it can be found that the proposed StereoIF-Net achieves comparatively better prediction results for both symmetric and asymmetric distortions, specifically, the PLCC and RMSE values are the best on the two parts, while SROCC is a little less than the best. Zhou's network [13] achieves the best SROCC value on symmetric set while Shi's method [12] obtains the best SROCC value on asymmetric set. The symmetric structure of Zhou's network may lead to better prediction results for symmetric distortion. In Shi's method, in order to solve the negative influence of difference image between left and right views, they proposed a registered distortion representation. This representation can address scene discrepancy (usually more serious in asymmetric pairs) and generate much better predictions especially for asymmetric pairs.

In Table XIV and XV, performance comparison of the StereoIF-Net and some prevalent methods on the IVC SIQA database and WaterlooIVC database is conducted. In the tables, the metric values in the first and second rank are marked in bold. From the Table XIV, it can been seen that the StereoIF-Net achieves the best results on the IVC dataset. From Table XV, it can be found that the StereoIF-Net has a better performance than compared methods on WIVC-I database but is a little less than Sim's method on WIVC-II database. By comparing the stereoscopic pairs in LIVE 3D and WaterlooIVC databases, it can be found that images in WaterlooIVC dataset have more apparent semantic meanings of scenes, especially in WaterlooIVC Phase II. Sim's method

TABLE XV

COMPARISON PERFORMANCE ON WATERLOO-IVC SIQA DATABASE

Methods	WIVC-I			WIVC-II		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Khan [24]	0.9344	0.9253	-	0.9097	0.9053	-
Fang [42]	0.9530	0.9500	-	0.9360	0.9220	-
Sim [15]	**0.9625**	**0.9566**	**4.1916**	**0.9698**	**0.9699**	**4.5981**
Liu [39]	0.9450	0.9280	5.2680	0.9130	0.9010	7.6580
Zhou [13]	0.9570	0.9400	-	0.9490	0.9500	-
StereoIF-Net	**0.9690**	**0.9599**	**2.9508**	**0.9580**	**0.9501**	**3.0506**

TABLE XVI

THE STATISTICAL SIGNIFICANCE TEST RESULTS ON LIVE 3D DATABASE

Method	LIVE 3D Phase I		LIVE 3D Phase II	
	PLCC	SROCC	PLCC	SROCC
Chen [58]	1	1	1	1
Zhou [13]	1	1	1	1

proposed a deep network including semantic and quality channels, by the hypothesis that people refer to semantic content of images when assessing perceptual quality of images. Semantic features can be regarded as indirect features on image quality. This maybe the reason why Sim's method can perform best on Waterloo IVC Phase II but cannot work so well on LIVE 3D dataset.

H. Statistical Significance

To verify the statistical significance of comparison results on LIVE 3D database, a two-sample t-test is adopted in our experiments to investigate whether the difference between the samples represents a true difference in the study. Further, the t-test with confidence at 90% over 100 trials for PLCC and SROCC is conducted. Due to the limitation of source code availability, we compared the StereoIF-Net with Zhou's [13] and Chen's method [58], and the statistical significance results can be seen in Table XVI. In the table, the test result is represented by "1" when StereoIF-Net is superior to compared metric. From the table, it can be found that the superiority of StereoIF-Net in statistical significance is over the two methods.

I. Cross Database Tests

In this subsection, a set of experiments are conducted to test the generalization performance of StereoIF-Net on the LIVE 3D and Waterloo IVC databases. The comparison results can be seen in Table XVII and Table XVIII, where "Phase I/Phase II" means that the trained model on Phase I is employed to the testing dataset on Phase II, and vice versa. So as on the Waterloo IVC dataset. In these tables, the top two metric values are in bold. It can be found that the proposed StereoIF-Net has a relatively better generalization ability than other methods. Specifically, the StereoIF-Net occupies the maximum number of bold values in Table XVII, in other words, three bold values (0.8595, 0.9184 and 0.9149). From the comparison on WaterlooIVC dataset in Table XVIII, it can be found that the StereoIF-Net is the second best while the model in Sim [15] is the best. But the generalization performance of Sim's method on LIVE 3D database is not so good, as shown in Table XVII. **Generally, "Phase II/Phase I" pattern always achieves**

TABLE XVII

THE GENERALIZATION PERFORMANCE OF DIFFERENT TESTS ON LIVE 3D DATABASE

Train/Test	Phase I/ Phase II		Phase II/ Phase I	
Criteria	PLCC	SROCC	PLCC	SROCC
Liu [39]	**0.8620**	0.8320	0.8880	0.8740
Shen [10]	0.8480	**0.8330**	0.9150	**0.9210**
Fang [11]	0.8110	0.7970	0.8990	0.8980
Yang [45]	0.8520	**0.8490**	0.8690	0.8600
Zhou [13]	0.7100	-	**0.9320**	-
Shi [12]	-	0.7930	-	0.9010
Messai [44]	0.8320	0.8230	0.8970	0.8870
Sim [15]	0.8041	0.7704	0.9083	0.8974
StereoIF-Net	**0.8595**	0.7972	**0.9184**	**0.9149**

TABLE XVIII

THE GENERALIZATION PERFORMANCE OF DIFFERENT TESTS ON WATERLOO-IVC SIQA DATABASE

Train/Test	WIVC-I/ WIVC-II		WIVC-II/ WIVC-I	
Criteria	PLCC	SROCC	PLCC	SROCC
Sim [15]	**0.7876**	**0.7547**	**0.9139**	**0.9032**
Liu [39]	0.6960	0.6270	0.7010	0.7080
Zhou [13]	0.7230	0.7300	0.8215	0.8000
StereoIF-Net	**0.7816**	**0.7502**	**0.8981**	**0.8712**

better results than "Phase I/Phase II" pattern, since image complexity in Phase II is higher than that in Phase I. The random separation of the training and testing dataset in different methods may lead to variations of generalization ability. For example, the datasets (e.g. training on Phase I and testing on Phase II) with similar distortion extend can derive better linear correlation and thus better cross testing results. In addition, by comparing the stereoscopic pairs in LIVE 3D and WaterlooIVC databases, it can be found that images in WaterlooIVC dataset have more apparent semantic meanings of scenes, especially in WIVC-II. Sim's method proposed a deep network including semantic and quality channels, by the hypothesis that people refer to semantic content of images when assessing perceptual quality of images. Semantic features can be regarded as indirect features on image quality. This maybe the reason why Sim's method can perform best on WIVC-II but cannot work so well on LIVE 3D dataset.

In addition, we also test StereoIF-Net's generalization ability on IVC SIQA database because it includes more complicated application scenarios than LIVE 3D database. In other words, the trained networks on LIVE 3D Phase I and Phase II, named StereoIF-Net-Phase-I and StereoIF-Net-Phase-II, are applied on IVC SIQA database respectively. As comparison, Zhou's network [13] trained on LIVE 3D Phase I and Phase II, named Zhou-Phase-I and Zhou-Phase-II, are employed to test the performance on IVC SIQA database respectively. The comparison results can be shown in Table XIX. From the table, it can be concluded that due to scenery complexity all experiments have a low performance on IVC SIQA database. Further, the trained networks of Phase I or Phase II both have a better performance than Zhou's network.

J. Computational Complexity

In this subsection, computational complexity of the StereoIF-Net and other methods is compared. These

TABLE XIX

THE GENERALIZATION PERFORMANCE OF DIFFERENT NETWORKS ON IVC SIQA DATABASE

Trained networks	PLCC	SROCC
Zhou-Phase-I	0.5601	0.5408
Zhou-Phase-II	0.5211	0.5102
StereoIF-Net-Phase-I	0.5860	0.5774
StereoIF-Net-Phase-II	**0.6115**	**0.5972**

TABLE XX

COMPUTATIONAL COMPLEXITY PERFORMANCE ON PHASE I

Methods	Computational time (s)
Khan [24]	2.905
Liu [39]	2.070
Shen [10]	7.961
Chen [58]	2.671
Zhou [13]	2.377
Oh [8]	8.308
StereoIF-Net	**1.995**

experiments are all based on a V100 Linux server equipped with 8 Nvidia GPUs of 32 GB. In Table XX, the mean computational time of each 3D image pair on LIVE 3D Phase I is displayed. From the table, it can be found that the complicated training processes of Oh's network leads to high computational time. Due to the complexity of network structure, Shen's network has a relatively high computational time. Compared with other networks, StereoIF-Net needs a lower computational time, which can prove the superiority of the proposed StereoIF-Net.

V. CONCLUSION

Considering binocular features of the HVS such as binocular interaction and fusion, in this paper, a NR stereoscopic image quality network named StereoIF-Net is proposed. The StereoIF-Net has fully considered the visual responses to stereoscopic visual signals in human cortex regions. It provides a befitting and detailed architecture to imitate the cross-over between right and left visual signals in visual cortex regions (i.e. V1-V5). The BIMs can well reflect the receptive fields and feature extraction procedures in V2-V5 areas, while the designed cross convolution can explore interaction details in each cortex region, which can enable StereoIF-Net to extract more cross characteristics. Automatic learned fusion weights of left and right views can be obtained via the proposed BFM, which can benefit the mapping from the convolution features to predicted quality scores. Compared with state-of-the-art works, the proposed StereoIF-Net achieves comparatively better prediction results on datasets including LIVE 3D, IVC and WaterlooIVC 3D SIQA datasets. Specifically, the metric values on LIVE 3D, IVC and WIVC-I datasets are the best, and are the second-best on the WIVC-II database.

REFERENCES

[1] J. Xu, W. Zhou, Z. Chen, S. Ling, and P. Le Callet, "Binocular rivalry oriented predictive autoencoding network for blind stereoscopic image quality measurement," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[2] W. Zhou, L. Yu, Y. Zhou, W. Qiu, M.-W. Wu, and T. Luo, "Local and global feature learning for blind quality evaluation of screen content and natural scene images," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2086–2095, May 2018.

[3] 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 Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[4] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, 2021.

[5] M. Yousefi and J. H. L. Hansen, "Frame-based overlapping speech detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6744–6748.

[6] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, and M.-M. Cheng, "Spatial information guided convolution for real-time RGBD semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 2313–2324, 2021.

[7] W. Ren, X. Wang, J. Tian, Y. Tang, and A. B. Chan, "Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets," *IEEE Trans. Image Process.*, vol. 30, pp. 1439–1452, 2021.

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

[9] W. Zhang, C. Qu, L. Ma, J. Guan, and R. Huang, "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network," *Pattern Recognit.*, vol. 59, pp. 176–187, Nov. 2016.

[10] 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*, vol. 424, pp. 132–142, Feb. 2021.

[11] Y. Fang, J. Yan, X. Liu, and J. Wang, "Stereoscopic image quality assessment by deep convolutional neural network," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 400–406, Jan. 2019.

[12] 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 Recognit.*, vol. 100, pp. 1–14, Apr. 2020.

[13] W. Zhou, Z. Chen, and W. Li, "Dual-stream interactive networks for no-reference stereoscopic image quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3946–3958, Aug. 2019.

[14] Y. Ding, S. Li, and Y. Chang, "Stereoscopic image quality assessment weighted guidance by disparity map using convolutional neural network," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.

[15] K. Sim, J. Yang, W. Lu, and X. Gao, "Blind stereoscopic image quality evaluator based on binocular semantic and quality channels," *IEEE Trans. Multimedia*, vol. 24, pp. 1389–1398, 2022.

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

[17] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

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

[19] R. B. Tootell *et al.*, "Functional analysis of V3A and related areas in human visual cortex," *J. Neurosci., Off. J. Soc. Neurosci.*, vol. 17, no. 18, pp. 7060–7078, 1997.

[20] A. W. Roe *et al.*, "Toward a unified theory of visual area V4," *Neuron*, vol. 74, no. 1, pp. 12–29, 2012.

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

[22] S. L. P. Yasakethu, C. T. E. R. Hewage, W. A. C. Fernando, and A. M. Kondoz, "Quality analysis for 3D video using 2D video quality models," *IEEE Trans. Consum. Electron.*, vol. 54, no. 4, pp. 1969–1976, Nov. 2008.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[24] S. Khan and S. S. Channappayya, "Estimating depth-salient edges and its application to stereoscopic image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5892–5903, Dec. 2018.

[25] Y. Liu, F. Kong, and Z. Zhen, "Toward a quality predictor for stereoscopic images via analysis of human binocular visual perception," *IEEE Access*, vol. 7, pp. 69283–69291, 2019.

[26] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.

[27] Y. Fan, M.-C. Larabi, F. A. Cheikh, and C. Fernandez-Maloigne, "Stereoscopic image quality assessment based on the binocular properties of the human visual system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2037–2041.

[28] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[29] X. K. Yang, W. S. Ling, Z. K. Lu, E. P. Ong, and S. S. Yao, "Just noticeable distortion model and its applications in video coding," *Signal Process., Image Commun.*, vol. 20, no. 7, pp. 662–680, Aug. 2005.

[30] R. Bensalma and M. C. Larabi, "A perceptual metric for stereoscopic image quality assessment based on the binocular energy," *Multidimensional Syst. Signal Process.*, vol. 24, no. 2, pp. 281–316, 2013.

[31] Y. Zhu, G. Zhai, K. Gu, Z. Che, and D. Li, "Stereoscopic image quality assessment with the dual-weight model," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2016, pp. 1–6.

[32] Y. Zhao, Z. Chen, C. Zhu, Y.-P. Tan, and L. Yu, "Binocular just-noticeable-difference model for stereoscopic images," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 19–22, Jan. 2011.

[33] S. A. Fezza, M. Larabi, and K. M. Faraoun, "Stereoscopic image quality metric based on local entropy and binocular just noticeable difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 2002–2006.

[34] J. Si, H. Yang, B. Huang, Z. Pan, and H. Su, "A full-reference stereoscopic image quality assessment index based on stable aggregation of monocular and binocular visual features," *IET Image Process.*, vol. 15, no. 8, pp. 1629–1643, Jun. 2021.

[35] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2165–2176, Aug. 2004.

[36] T. Abeywickrama, M. A. Cheema, and D. Taniar, "K-nearest neighbors on road networks: A journey in experimentation and in-memory implementation," *Proc. VLDB Endowment*, vol. 9, no. 6, p. 492–503, Jan. 2016.

[37] 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 Recognit.*, vol. 71, pp. 207–217, Nov. 2017.

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

[39] Y. Liu, C. Tang, Z. Zheng, and L. Lin, "No-reference stereoscopic image quality evaluator with segmented monocular features and perceptual binocular features," *Neurocomputing*, vol. 405, pp. 126–137, Sep. 2020.

[40] Y. Liu, W. Yan, Z. Zheng, B. Huang, and H. Yu, "Blind stereoscopic image quality assessment accounting for human monocular visual properties and binocular interactions," *IEEE Access*, vol. 8, pp. 33666–33678, 2020.

[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 Trans. Image Process.*, vol. 24, no. 5, pp. 1685–1699, May 2015.

[42] Y. Fang, J. Yan, J. Wang, X. Liu, G. Zhai, and P. Le Callet, "Learning a no-reference quality predictor of stereoscopic images by visual binocular properties," *IEEE Access*, vol. 7, pp. 132649–132661, 2019.

[43] W. Zhou and L. Yu, "Binocular responses for no-reference 3D image quality assessment," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1077–1084, Jun. 2016.

[44] O. Messai, F. Hachouf, and Z. A. Seghir, "AdaBoost neural network and cyclopean view for no-reference stereoscopic image quality assessment," *Signal Process., Image Commun.*, vol. 82, Mar. 2020, Art. no. 115772.

[45] 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," *Inf. Sci.*, vol. 474, pp. 1–17, Feb. 2019.

[46] 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 Trans. Image Process.*, vol. 25, no. 5, pp. 2059–2074, Mar. 2016.

[47] P. Zhao, S. Li, and Y. Chang, "No-reference stereoscopic image quality assessment based on dilation convolution," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.

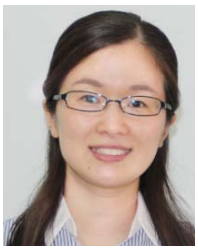
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

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

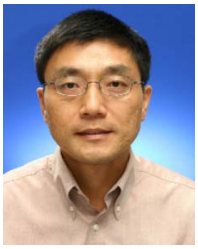
[50] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-reference quality assessment of natural stereopairs," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3379–3391, Sep. 2013.
[51] A. Benoit, P. L. Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *Eurasip J. Image Video Process.*, vol. 2008, no. 1, pp. 1–13, Jan. 2008.
[52] J. Wang, A. Rehman, K. Zeng, S. Wang, and Z. Wang, "Quality prediction of asymmetrically distorted stereoscopic 3D images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3400–3414, Nov. 2015.
[53] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines Vinod Nair," in *Proc. ICML*, vol. 27, Jun. 2010, pp. 807–814.
[54] X. Wang, L. Ma, S. Kwong, and Y. Zhou, "Quaternion representation based visual saliency for stereoscopic image quality assessment," *Signal Process.*, vol. 145, pp. 202–213, Apr. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168417304152
[55] J. Zhang, T. M. Le, S. H. Ong, and T. Q. Nguyen, "No-reference image quality assessment using structural activity," *Signal Process.*, vol. 91, no. 11, pp. 2575–2588, Nov. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168411001654
[56] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 18, 2020, doi: 10.1109/TPAMI.2020.3045810.
[57] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. Int. Conf. Learn. Represent.*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980
[58] 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 Process., Image Commun.*, vol. 28, no. 9, pp. 1143–1155, 2013.
[59] L. Chen and J. Zhao, "Quality assessment of stereoscopic 3D images based on local and global visual characteristics," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 61–66.
[60] J. Yang, Y. Zhao, B. Jiang, W. Lu, and X. Gao, "No-reference quality evaluation of stereoscopic video based on spatio-temporal texture," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2635–2644, Oct. 2020.

Jianwei Si received the B.S. degree in safety science and engineering from Liaoning Technical University, China, in 2018, the M.S. degree in computer science and technology from Qingdao University, China, in 2021. He is currently pursuing the Ph.D. degree with the Faculty of Information Science and Engineering, Ocean University of China. His research interests include deep learning, image processing, and meteorology.

Baoxiang Huang (Member, IEEE) received the B.S. degree in traffic engineering from the Shandong University of Technology, China, in 2002, and the M.S. degree in mechatronic engineering from Shandong University, China, in 2005, and the Ph.D. degree in computer engineering from the Ocean University of China, China. Currently, she is an Associate Professor with the College of Computer Science and Technology, Qingdao University, China, and also an Academic Visitor at Nottingham University. Her research interests include remote sensing image processing and analysis, environment modeling, and quality assessment.



Huan Yang (Member, IEEE) received the B.S. degree in computer science from the Heilongjiang Institute of Technology, China, in 2007, the M.S. degree in computer science from Shandong University, China, in 2010, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2015. Currently, she is an Associate Professor with the College of Computer Science and Technology, Qingdao University, China. Her research interests include intelligent image processing, perception-based modeling, quality assessment, and computer vision.



Weisi Lin (Fellow, IEEE) is a Professor and the Associate Chair (Research) with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include intelligent image processing, perceptual signal modeling, video compression, and multimedia communication. He is a Chartered Engineer and a fellow of the IET. He was the Technical Program Chair of PCM 2012, the IEEE International Conference on Multimedia and Expo (ICME) 2013, QoMEX 2014, and the IEEE Visual Communications and Image Processing (VCIP) 2017. He was a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2016 to 2017 and the AsiaPacific Signal and Information Processing Association (APSIPA) from 2012 to 2013. He is a Keynote/Invited/Panelist/Tutorial Speaker at over 40 international conferences. He is a Highly Cited Researcher 2019, 2020, and 2021 (awarded by Clarivate Analytics). He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE SIGNAL PROCESSING LETTERS.

Zhenkuan Pan (Senior Member, IEEE) received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 1987, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 1992. Currently, he is a Professor with the College of Computer Science and Technology, Qingdao University, Qingdao, China. He has authored and coauthored more than 300 academic papers in the areas of computer vision, dynamics, and control. His research interests include variational models of image and geometry processing and multibody system dynamics.