A Blind Stereoscopic Image Quality Evaluator with Segmented Stacked Autoencoders Considering The Whole Visual Perception Route

Jiachen Yang, *Member, IEEE,* Kyohoon Sim, Xinbo Gao, *Senior Member, IEEE,* Wen Lu, *Member, IEEE,* Qinggang Meng, *Senior Member, IEEE,* and Baihua Li

***Abstract*—Most of the current blind stereoscopic image quality assessment (SIQA) algorithms cannot show reliable accuracy. One reason is that they do not have the deep architectures and the other reason is that they are designed on the relatively weak biological basis, compared with findings on human visual system (HVS). In this paper, we propose a Deep Edge and COlor Signal INtegrity Evaluator (DECOSINE) based on the whole visual perception route from eyes to the frontal lobe, and especially focus on edge and color signal processing in retinal ganglion cells (RGC) and lateral geniculate nucleus (LGN). Furthermore, to model the complex and deep structure of the visual cortex, Segmented Stacked Auto-encoder (S-SAE) is used, which has not utilized for SIQA before. The utilization of the S-SAE complements weakness of deep learning-based SIQA metrics that require a very long training time. Experiments are conducted on popular SIQA databases, and the superiority of DECOSINE in terms of prediction accuracy and monotonicity is proved. The experimental results show that our model about the whole visual perception route and utilization of S-SAE are effective for SIQA.**

***Index Terms*—stereoscopic image quality assessment, retinal ganglion cell, lateral geniculate nucleus, segmented stacked auto-encoders, edge quality, color quality.**

I. Introduction

3D visual content has penetrated our lives deeply. We can easily find 3D movies, 3D TVs, 3D digital cameras and mobile phones equipped with dual cameras around us. Through them countless stereoscopic images are produced everyday. These images often suffer from perceptual quality degradation caused by distortions when they are transmitted, stored, compressed and processed. The degraded images need to be restored and SIQA indices can provide a criterion for restoration [1], [2]. Image quality assessment (IQA) models are divided into three categories according to usage of the original image: full-reference (FR) [3]–[12], reduced-reference (RR) [13]–[15] and no-reference (NR)/blind [16]–[20] IQA metrics. In the majority of cases, we cannot have direct access to the pristine images of the degraded ones [21]. Therefore, FR and RR IQA models find very limited usage. However most studies on SIQA so far have concentrated on FR and RR methods. To avoid such limitation, in this paper, we propose a SIQA method belongs to the NR metric.

The main goal of IQA is to make an accurate prediction about stereopairs' quality like a human being. It is very rational to mimic the pathway of our visual processing system. Most people see the world through two eyes, and slightly different images individually fall on the retinas [22]. The two views are compressed by retinal ganglion cells (RGC) remaining edge signals mainly. Color information is also conveyed through the RGCs. These are transmitted to visual cortex via LGNs. The right and left view signals are integrated in the primary visual cortex (V1) first [23]. Each part of the visual cortex seems to have their own particular role. However it is almost impossible to specify completely their roles because they interact with each other. Through numerous complicated steps, information about the stereopair eventually reaches the frontal lobe that determines perceived quality. Previous SIQA metrics tried to model HVS, but they were designed on the relatively weak biological basis like follows.

In the initial stage of SIQA, 2D-IQA metrics were straightforwardly applied to both left and right views, and then two obtained quality scores from each side were combined into one overall score [3]. Further, disparity information was integrated on a 2D-IQA basis [4]–[7]. But these metrics cannot deal with binocular perception such as binocular rivalry and suppression arising from V1 where information from both eyes comes together. As a result, such methods cannot predict well the quality of asymmetrically distorted stereopairs. To solve this, Wang *et al.* [8] presented a framework that integrates a spatial weighting system considering the suppression phenomenon. Ryu *et al.* [9] applied an unequal weighting system according to the respective quality of left and right images and suppression degree. Furthermore, an intermediate image called a cyclopean image was generated by Chen *et al.* [10] so as to mimic a single fused percept in V1. They produced two cyclopean images respectively from original and distorted stereopairs and then evaluated the quality using FR 2D-IQA metrics. More sophisticated algorithms have been developed since then. Lee *et al.* [11] proposed a model that divides a stereo image into binocular and monocular vision segments

This work was partially supported by National Natural Science Foundation of China (No. 61871283), Foundation of Pre-Research on Equipment of China (NO.61403120103) and Joint Foundation of pre-Research on Equipment from Education Department of China (No.6141A02022336).

J. Yang and K. Sim are with School of Electrical and Information Engineering, Tianjin University, Tianjin, China (e-mail: yangjiachen@tju.edu.cn; tlarygns0211@tju.edu.cn).

X. Gao is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an, China (e-mail: xbgao@mail.xidian.edu.cn).

W. Lu is with School of Electronic Engineering, Xidian University, Xian, China (e-mail: luwen@xidian.edu.cn).

Q. Meng and B. Li are with the Department of Computer Science, Loughborough University, UK (email: q.meng@lboro.ac.uk; B.Li@lboro.ac.uk).

and applies different visual weights to the pooling method. Zhang *et al.* [12] devised a 3D-MAD that estimates perceived quality degradation by distortion of monocular views and a cyclopean view respectively.

In addition to these FR-SIQA algorithms, some NR-SIQA metrics have been presented. Chen *et al.* [16] extracted 2D features from a cyclopean image and 3D features from a disparity map and an uncertainty map. All of the extracted features were used to train a support vector regressor (SVR). Ryu *et al.* [24] explored the relationship between the perceptual quality and visual information, and introduced a method modeling the binocular quality perception in the context of blurriness and blockiness. Su *et al.* [17] formed a convergent cyclopean image and extracted bivariate and correlation NSS features in the spatial and wavelet domains. Shao *et al.* [18] proposed a metric that learns binocular receptive field properties and quality lookups, from perspective of dictionary learning. With the development of deep learning (DL), the study on IQA was further improved. Zhang *et al.* [19] proposed a metric based on convolutional neural network (CNN) that can effectively learn the complicated mapping relations between raw images and their labels. This metric does not need handcrafted features. Shao *et al.* [20] trained two separate 2D deep belief networks (DBN) for monocular images and cyclopean images, and then combined the quality scores using weighting schemes. These two DL-based methods achieved higher consistence with subjective assessment than shallow structure metrics.

To develop biologically plausible NR-SIQA, four aspects should be addressed: 1) to establish a sufficient biological model mimicking a visual processing system, 2) to reflect binocular perception properties, 3) to have a deep structure, and 4) to deliver good prediction accuracy. But most of the previous methods did not have strong biological underpinnings, and were also based on shallow architectures. As a result, they could not achieve the satisfactory performance.

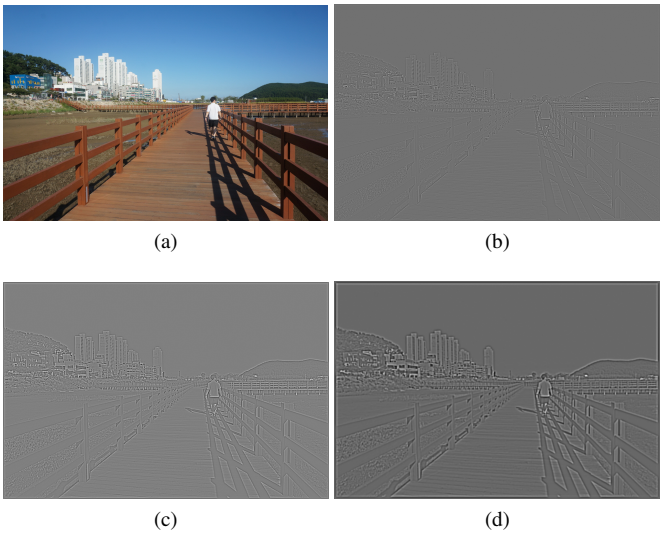
In this paper, we propose a Deep Edge and COlor Signal INtegrity Evaluator (DECOSINE) modeling the whole visual perception route that consists mainly of feature extraction about edge and color and multiple levels of abstraction. Our contributions are as follows.

1) We make a novel neuro-biological model based on the whole visual perception route from eyes to the frontal lobe unlike most SIQA metrics that imitate a part of the route. The route is organized by us into two sub-routes considering edge and color signals, and two local scores from these sub-routes are computed: edge quality score and color quality score. Concretely, contour images of left and right views are computed to model edge extraction of RGCs, and they are used to calculate intermediate maps such as sum, difference and cyclopean maps based on binocular sum and difference channel theory. Contrary to the edge information, the color information was not thoroughly applied in the field of SIQA. To deal with color information, we model the opponent coding occurred in LGNs.
2) Segmented SAE (S-SAE) is utilized to mimic deep and complex architecture of the visual cortex. It can solve a drawback of conventional SAE that requires a long training time. Using these S-SAEs, the proposed method can achieve not only low computational complexity, but also accurate prediction ability. We believe this "segmentation" idea can be also used in DBNs or CNNs-based IQA metrics [19], [20], [25]. Three separate S-SAEs for edge information and one SAE for color information are trained, and the resulting deeper features are fed into regression models, respectively.
3) According to biological discoveries of binocular vision [26], [27], two dynamic weighting systems and one static weighting system are newly designed. Especially the dynamic weighting systems consider the extent of correlation between left and right images of stereopairs. Locally measured quality scores are combined into one overall score via these weighting systems. They obviously improve prediction performance of the proposed algorithm. Comparing with previous SIQA metrics, DECOSINE provides the most precise and unbiased estimation.

The remainder of this paper is organized as follows. Section II introduces theoretical bases about the retinal ganglion cells, LGNs, V1 and S-SAE. In Section III, the proposed metric, DECOSINE, is elaborately explained. The methodology and the experimental results are presented in Section IV. Finally conclusions and future work are summarized in Section V.

II. Theoretical Background

In this section, we explain how edge and color signals are processed in the retinal ganglion cells and LGNs. Next, binocular summation and difference channels in V1 are described. The fundamentals of S-SAE are also introduced.



(a) (b) (c) (d)

Fig. 1: (a) An original image. (b), (c) and (d) LoG images modeling ON-cells with $(3, 0.5)$, $(7, 1)$ and $(13, 2)$ parameters respectively. The parameters determine the thickness of edge. Refer to Section III.B. Note that the edges are picked out, whereas areas of uniform brightness look the same mid-grey.

A. Retinal Ganglion Cells and LGNs

We have about 120 million rods and 8 million cones in each eye [22]. Photoreceptor cells comprised of rods and cones are analogous to the pixels of digital cameras. The number of these photoreceptor cells is much more than 16 million pixel of smart-phone cameras sold on the market. The more the pixels, the greater the volume of information. Our visual system has a compression function for dealing with massive amounts of visual information. It is performed by retinal ganglion cells (RGC). They have parvocellular (P), koniocellular (K) and magnocellular (M) cells [28]. The P and K pathways carry color information, whereas the M pathway carries information about movements and edges in the view. Among these, M pathway is most deeply involved in edge extraction. M cells have two types of receptive fields: ON-center and OFF-center. ON-center cells become hyperpolarized in response to light, and OFF-center cells become depolarized on exposure to light [29]. The receptive field cares about changes in a small region of the world and ignores the rest. Fig. 1 shows a scene and images after being filtered by ON-center cells. These are not the same as an ordinary grey image. Look at the shadow of banisters or mountains in Fig. 1(b)-(d), on contours including changes from bright spots to dark spots, you can see the white line first and then the black line. On the other hand, areas of uniform brightness look the same mid-grey. It is correlated with the response of the ON-center cells. In summary, only 'edges' or 'changes' in the pattern are extracted and transferred to the V1 via the left and right LGNs separately.

Fig. 2: A scene and the corresponding Lum, RG and BY aspects. The Lum, RG and BY maps are generated by opponent channel encoding.

In addition to signaling edge information, the RGCs get involved in color vision too. As mentioned above the P and K cells of RGCs respond to changes in color, and these are linked with P and K cells of LGNs. Passing the LGNs, an opponent coding is fulfilled, which encodes color activation by comparing the activities of cone types [30]. The types of cones in retinas are divided into three: L-, M- and S-cones [31]. They are sensitive to long (related to red), medium (green), and short (blue) wavelengths, respectively. There are three opponent channels encoding the red-green (RG), blue-yellow (BY) and light-dark (Lum) aspects of a scene [31]. These aspects are drawn in Fig. 2. Especially the P cells are heavily related to color vision based on RG comparison, while the color information carried through the K cells is based on the BY comparison. We thus calculate these three opponent responses for SIQA's sake, which are also conveyed to V1.

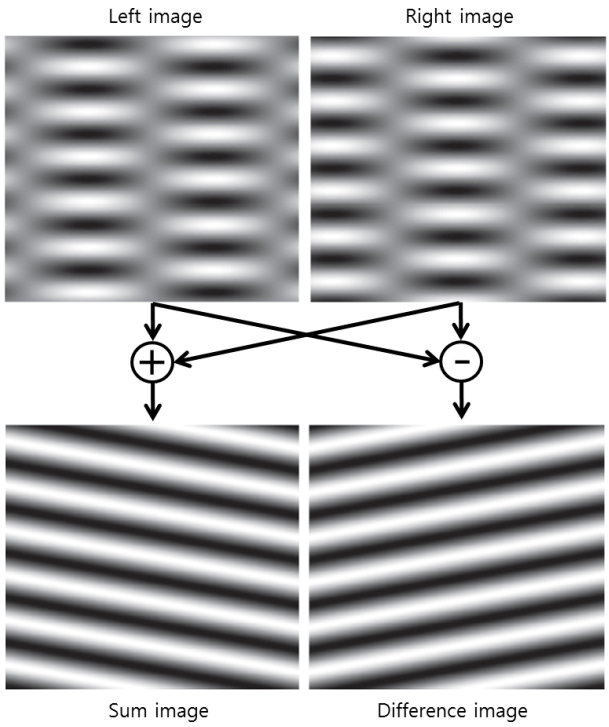


Fig. 3: Two monocular test stimuli and the corresponding binocular sum and difference images.

B. Binocular sum and difference channels in V1

When the above-mentioned signals about edge and color are transmitted into V1, the signals from the right eye are kept separate from the ones from the left eye [22]. However, about 70% of neurons in V1 can be regarded binocular and neurons in the later areas of the visual cortex are almost all binocular. The binocular neurons in V1 combine signals from the two eyes into a single fused image of the world, while the small differences between the two images are used to deduce information about depth [23].

Much previous work has shown that the fused image is dominated by the sum of left and right images [26], [32]. According to this, a sum image and a cyclopean image were generally utilized to model the fused image in SIQA. The sum image is a simple sum of the two images, whereas the cyclopean image is the weighted sum of two images considering the disparity information and the Gabor filter responses [10]. The cyclopean image reflects well binocular rivalry and suppression phenomenons. However, recent studies reported that, in addition to the binocular summation channel, there is also a binocular difference channel which subtracts one image from the other [27], [33].

In order to demonstrate the existence of difference channel, May and Zhaoping [33] designed an experiment based on

adaptation and after-effect. First, they tried to selectively adapt the binocular sum and difference channels. For adapting the difference channel selectively, they presented two images, I and $-I$, of photographic negative version each other to observer's two eyes, respectively. As a result, the difference channel, if it exists, should have a strong response because of $I-(-I)=2I$, while the sum channel, if it exists, should not respond because of $I+(-I)=0$. Conversely, to adapt the sum channel selectively, identical images were presented to the eyes, and the sum channel took $I+I=2I$, and the difference channel took $I-I=0$. As a result, only the sum channel should respond strongly. After these adaptation processes, May and Zhaoping showed two monocular test stimuli to observers' left and right eyes, respectively. The test stimuli as shown in Fig. 3 have following properties: if the brain sums the monocular images, people can see right-tilted bars, whereas if the brain subtracts them, people can see left-tilted bars. Note that there is no information about direction of tilt in these stimuli. The tilt emerges only when the monocular images are combined. After adapting the sum channel, when the test stimuli were presented, observers could see left-tilted bars that is their difference image, whereas after the difference channel was adapted, observers could see right-tilted bars as their sum image. Because it did not occur after adaptation to tilt, these observation was not a tilt after-effect. It is a compelling evidence that distinct sum and difference channels exist. For further details of this experiment, refer to [33].

Very few SIQA methods take account of the difference channel [34]. Based on that signals from the sum and difference channels are multiplexed [27], we combine edge signals of the two views into three types of intermediate images: cyclopean and sum images for modeling the sum channel, and a difference image for the difference channel.

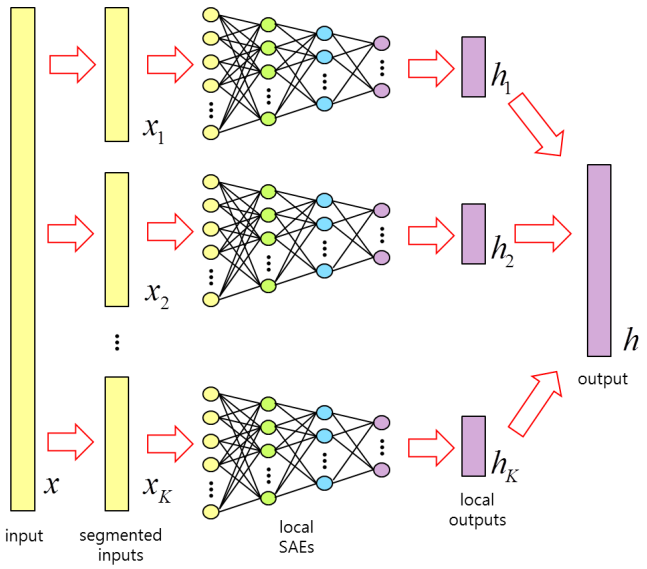


Fig. 4: A Segmented Stacked Auto-Encoder (S-SAE). An input vector is divided into smaller segments, and the segments are fed into each local SAE. The resulting local outputs are concatenated to form an output vector.

C. S-SAE

For the development of a blind IQA method, it is needed to learn a proper mapping model from quality-aware features to perceptual quality scores. Because human brain is organized in a deep architecture, e.g. from V1 to IT cortex [35], machine learning methods with the shallow structure cannot mimic it enough [36]. We thus use a SAE that is one of DL algorithms. Among many variants of the SAE, a S-SAE [37] is chosen because of its efficacy improving the characterization of features and its efficiency relieving computational complexity [37].

In a basic auto-encoder (AE), an input vector x is transformed into a reduced hidden representation h [38], and the h is mapped back into a reconstructed vector $\widetilde{x}$:

$$h = A(Wx+b), \tag{1}$$

$$\widetilde{x} = A(W'h+b'), \tag{2}$$

where A means an activation function, W, W' and b, b' are parameters for weights and biases. These parameters are optimized to minimize an average reconstruction error:

$$\underset{W,W',b,b'}{argmin}\frac{1}{n}\sum_{i=1}^{n}||x^{(i)}-\tilde{x}^{(i)}||^2, \tag{3}$$

where i is the i-th training sample and n is the total number of training samples. We can input h into a new AE and the hidden representation of h is learned. Repeating this procedure, a Stacked AE (SAE) is formed. It learns deep and abstract representation of the input vector. The learned deeper features can be used to train a regressor such as a support vector machine regressor (SVR) [39].

However the SAE has a relatively large computational complexity when it is trained, especially in an unsupervised learning phase. It thus requires a very long time for training. To alleviate it, a Segmented SAE (S-SAE) was proposed recently [37]. The S-SAE consists of several local SAEs. The input data is divided into smaller k segments according to the correlation among features. k local SAEs are applied to each segment separately, and the resulting outputs are concatenated to form an output vector, refer to Fig. 4.

We compare the S-SAE with the traditional SAE in terms of computational complexity. To this end, we suppose that a SAE has N input nodes and three hidden layers with M, L and P nodes respectively. The number of connections for the SAE is as below:

$$(N\times M)+(M\times L)+(L\times P). \tag{4}$$

If a S-SAE comprising of k local SAEs is used in place of the SAE, the k-th local SAE has N_k input nodes and M_k, L_k and P_k hidden nodes, where $N=\sum_{k=1}^{K}N_k, M=\sum_{k=1}^{K}M_k, L=\sum_{k=1}^{K}L_k$ and $P=\sum_{k=1}^{K}P_k$. It has $\sum_{k=1}^{K}(N_k\times M_k+M_k\times L_k+L_k\times P_k)$ connections. If every local SAE is designed in the same form, i.e. $N_1=N_2=\cdots=N_K, M_1=M_2=\cdots=M_K, L_1=L_2=\cdots=L_K$ and $P_1=P_2=\cdots=P_K$, the complexity for the S-SAE can be expressed as:

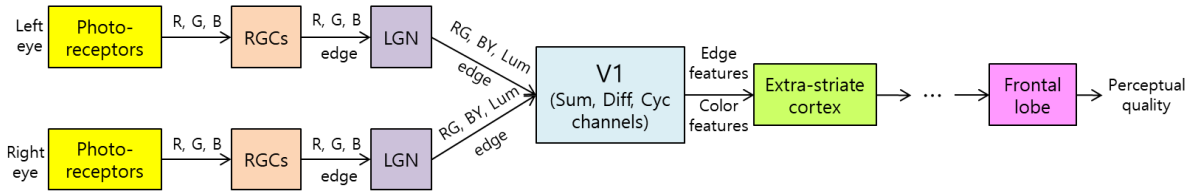


Fig. 5: The proposed biological model for SIQA.

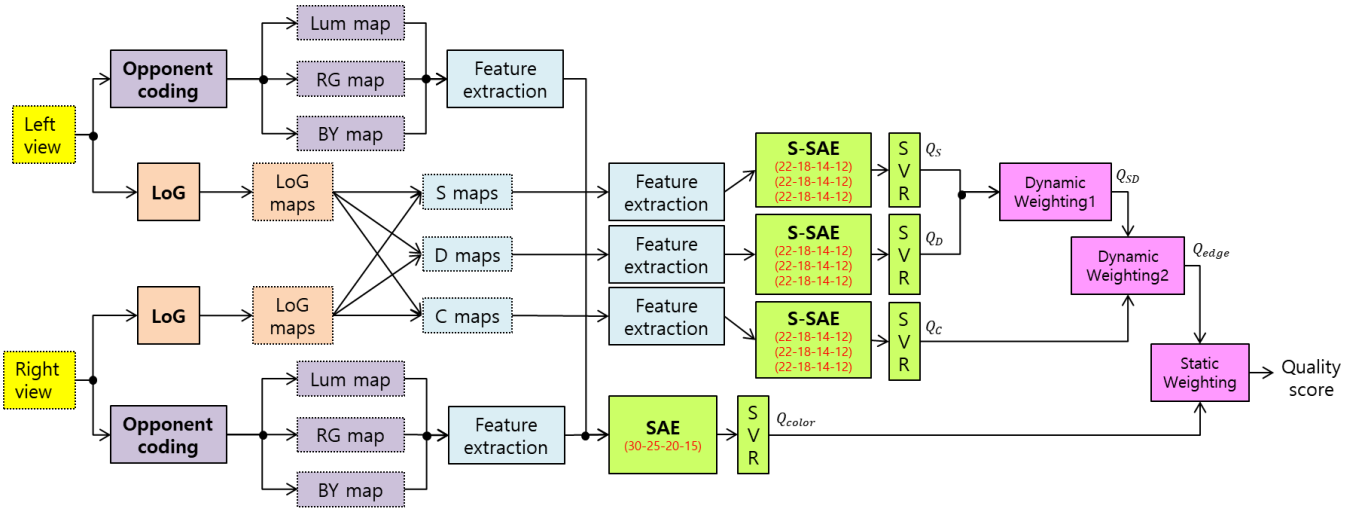


Fig. 6: The proposed algorithm, DECOSINE. It is a computational model of the biological model in Fig. 5.

$$K[(\frac{N}{K}\times\frac{M}{K})+(\frac{M}{K}\times\frac{L}{K})+(\frac{L}{K}\times\frac{P}{K})] = \frac{[(N\times M)+(M\times L)+(L\times P)]}{K} \quad (5)$$

As a result, the complexity of the SAE is reduced by K times in the S-SAE. This segmentation concept can be applied to other DL algorithms like DBNs and CNNs. In our DECOSINE, one SAE for a color signal and three S-SAEs for a edge signal are utilized to find deeper representations from quality-aware features.

III. The Proposed SIQA metric: DECOSINE

In this section, the proposed DECOSINE is explained. It is a biologically-inspired metric. Considering visual perception route of human beings, we design a model for SIQA. Based on the model, a computational algorithm is embodied. The feature extraction process and pooling systems of DECOSINE are also elaborated here.

A. The Biological Model of DECOSINE

When we look at some scenes, slightly different images are focused on the retinas. Each view is captured by L-, M- and S-cones like R, G and B components. These are conveyed to RGCs separately, and edge images of the left and right scenes are extracted from the RGCs. Next the R, G, B images and edge images are arrived at LGNs, and opponent coding occurs there. Consequentially the RG, BY, Lum signals and edge images are transmitted into the visual cortex via P, K and M streams. In V1 which is the first area of visual cortex, two eyes' signals are combined into summation (S), difference (D) and cyclopean (C) maps. S and C maps are based on the theories that the fused image is dominated by the sum of the left and right images [26], [32], while D map is inspired by the recent discovery that there is also the difference channel along with the sum channel [27], [33]. From total six types of maps, feature extraction for edge and color signals is proceeded. The extracted signals are transmitted into extra-striate cortex including more than 30 visual areas, such as V2, V3, V4 and so on [40]. There are tremendously complex connections between each area. In the extra-striate cortex, the simple features are transformed into gradually more abstract features in hierarchical ways. Synthesizing the abstract features, the frontal lobe that controls information and behavior from federal areas finally makes a decision about the perceived quality [22]. Fig. 5 describes this process organized for SIQA.

B. The Algorithm of DECOSINE

On the basis of the neuro-biological model, we design an algorithm named DECOSINE, as shown in Fig. 6. It is divided into two parts: edge quality index and color quality index. To simulate the edge extraction of RGCs, the Laplacian of a Gaussian (LoG) filter is applied on each view image:

$$\nabla^2 G(x,y) = [\frac{x^2+y^2-2\sigma^2}{\sigma^4}]e^{-\frac{x^2+y^2}{2\sigma^2}}, \quad (6)$$

where ∇^2 is the Laplacian operator, G is the 2D Gaussian function and σ is standard deviation. According to [41], we

filter the input image with a $n \times n$ Gaussian lowpass filter and compute the Laplacian of the image using the 3×3 mask in Fig. 7(a). In this paper, the parameters of the Gaussian

-1	-1	-1
-1	8	-1
-1	-1	-1

(a)

1	1	1
1	-8	1
1	1	1

(b)

Fig. 7: Simple Laplacian masks. (a) ON-cell mask. (b) OFF-cell mask.

filter are set to $(n, \sigma) \ni \{(3, 0.5), (7, 1), (13, 2)\}$ to obtain the LoG maps with different thickness of edge like Fig. 1, which models bar and edge detectors of different sizes in V1 [22]. As a result, 3 left and 3 right LoG maps are computed. The LoG maps of left and right images (L_{LoG} and R_{LoG}) are combined into three forms: S, D and C maps. The S and D maps are simply generated through summing and subtracting the LoG images as follows:

$$S(i,j) = L_{LoG}(i,j) + R_{LoG}(i,j), \tag{7}$$

$$D(i,j) = L_{LoG}(i,j) - R_{LoG}(i,j). \tag{8}$$

where i and j are spatial indices. Note that for D map we use the simple difference rather than the absolute difference because the simple difference reflects the relative difference. C map is computed as a weighted sum of the L_{LoG} and the disparity-compensated R_{LoG} as below:

$$\begin{aligned} C(i,j) =& W_L L_{LoG}(i,j) \\ &+ W_R((i + d(i,j)), j) R_{LoG}((i + d(i,j)), j), \end{aligned} \tag{9}$$

where W_L and W_R are computed from the Gabor filter responses, and d is the disparity. For further information, refer to [10].

From S, D and C maps, the quality-aware features are extracted respectively. These are further fed into the well trained S-SAEs and the obtained deep features are separately inputted into individual SVRs. The serial connection of S-SAEs and SVRs plays a role mimicking functions from the extra-striate cortex to the frontal lobe. The resulting quality scores, Q_S, Q_D and Q_C, are combined into an edge quality score Q_{edge} via pooling systems which will be explained later chapter.

For color quality index, opponent coding is firstly implemented in LGNs. We model the opponent coding by the following formulas [43]:

$$Lum = (\bar{R} + \bar{G} + \bar{B})/\sqrt{3}, \tag{10}$$

$$RG = (\bar{R} - \bar{G})/\sqrt{2}, \tag{11}$$

$$BY = (\bar{R} + \bar{G} - 2\bar{B})/\sqrt{6}, \tag{12}$$

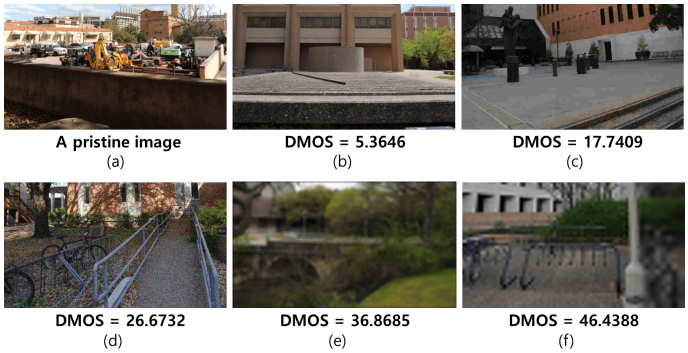


Fig. 8: Stereopairs and the corresponding differential mean opinion scores (DMOS) in LIVE 3D phase 1 database [42] for validating the usefulness of S, D and C maps. Because of symmetrically distorted pairs, left view images of stereopairs are only displayed. (a) a pristine image. (b) a JP2K compressed image. (c) a JPEG compressed image. (d) a white noise (WN)-added image. (e) a blurred image. (f) a fast-fading (FF) image. The higher DMOS, the lower quality.

where $\bar{R}, \bar{G}$ and $\bar{B}$ are mean subtracted and contrast normalized (MSCN) coefficients [44] of the $log(R)$, $log(G)$ and $log(B)$, respectively [45]. The quality-aware features are captured in the Lum, RG and BY maps. These features from maps of left and right versions are fed into one SAE network to obtain deep features about color. The deep features are inputted into a SVR, and a color quality score Q_{color} is computed. Finally, an overall quality score can be calculated via a weighed sum of Q_{edge} and Q_{color}.

C. Feature Extraction for DECOSINE

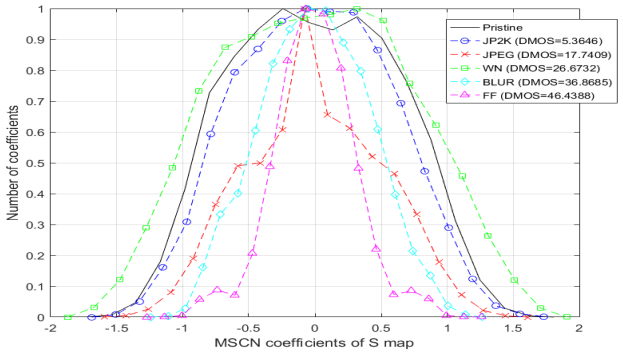


Fig. 9: A histogram of the MSCN coefficients of the S maps. Six stereopairs in Fig. 8 were used. The kurtosis and variance become smaller as the DMOSs are getting larger except for white noise.

Total 6 types of feature maps are exploited for extracting features. Because the S, D and C maps computed from left and right LoG maps have not been used in the field of SIQA, we test to see whether the maps can provide quality-aware features. To do this, we select a pristine stereopair and five distorted stereopairs (Fig. 8). From LoG maps of six pairs, the corresponding S maps are computed. In order to visualize

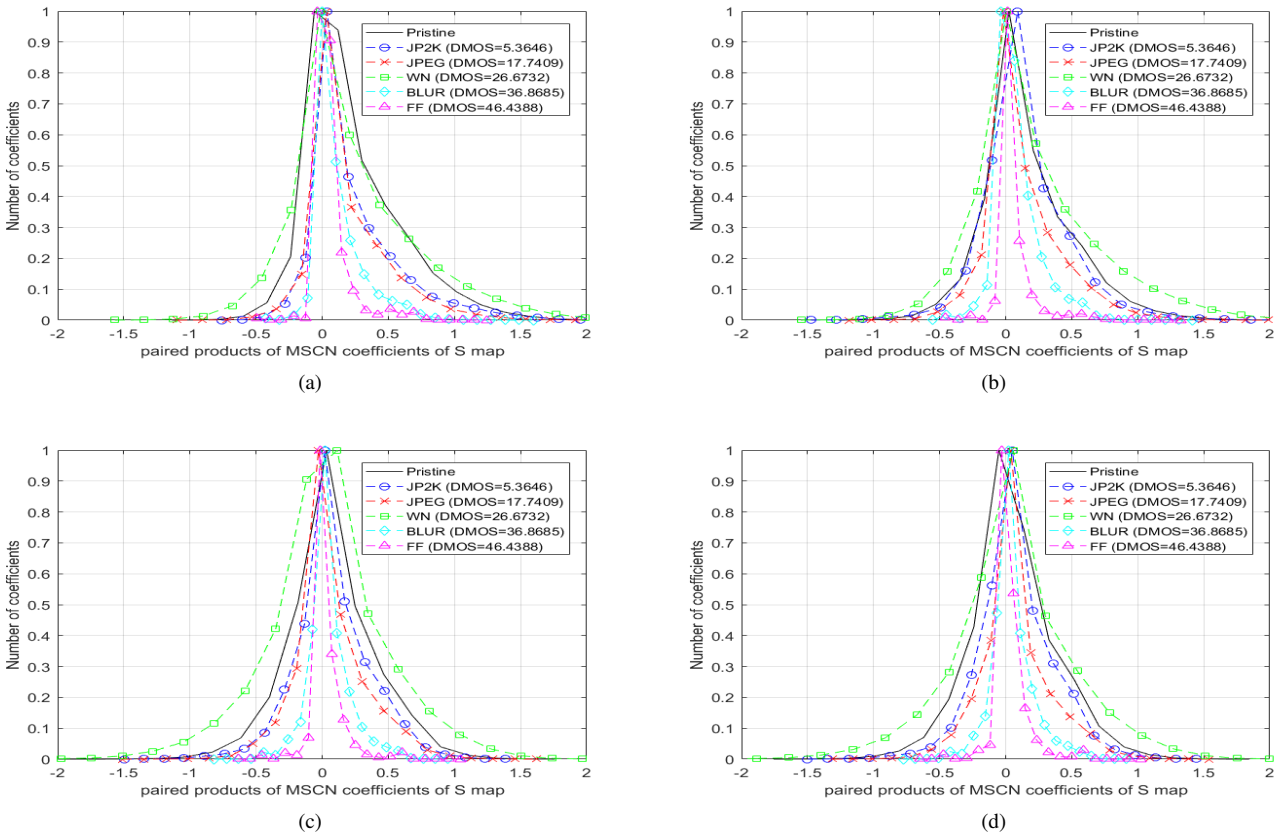


Fig. 10: A histogram of the pair-wised products of MSCN coefficients of the S maps. Six stereopairs in Fig. 8 were used. (a) Horizontal product. (b) Vertical product. (c) Main-diagonal product. (d) Secondary-diagonal product. The kurtosis and variance become smaller as the DMOSs are getting larger except for white noise.

that MSCN coefficients of the S maps vary with certain rules according to subjective quality scores, we plot histograms of the MSCN coefficients in Fig. 9. The kurtosis and variance of the histograms are clearly changed depending on DMOSs. Except for the case of the stereopair degraded with white noise (WN), the kurtosis and variance are getting smaller as DMOS values are getting higher. The pairwise products of neighboring MSCN coefficients [44] also display similar tendencies, as shown in Fig. 10. The analysis about features obtained from D and C maps are omitted due to the similarity.

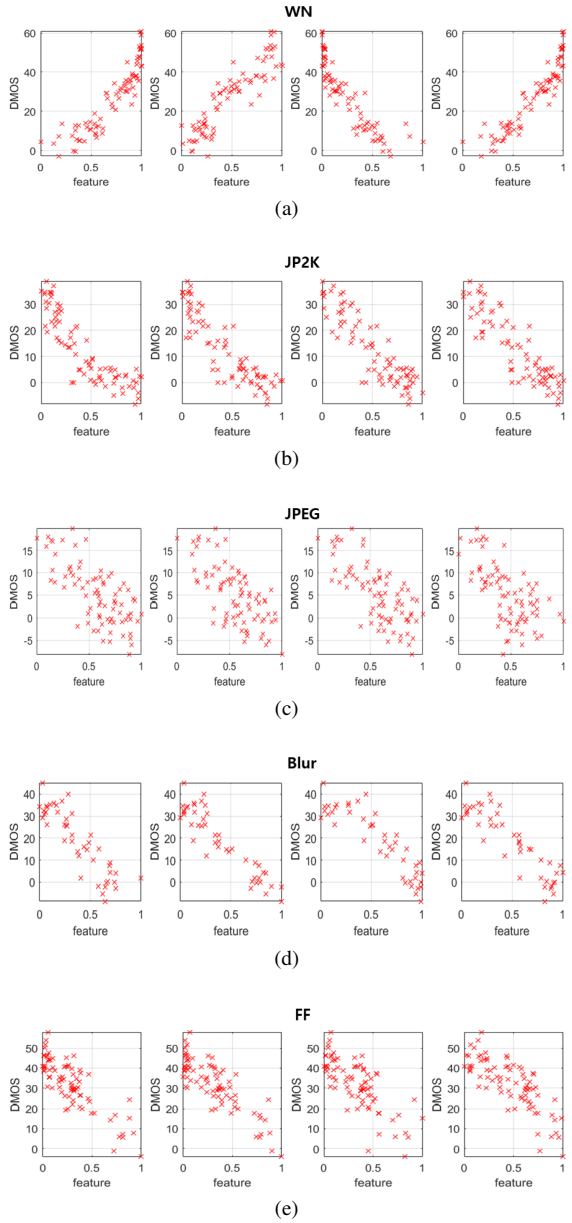
From S, D and C maps, the 2 generalized Gaussian distribution (GGD) and 16 asymmetric generalized Gaussian distribution (AGGD) fitting parameters are extracted like [44]. According to [46], magnitude, variance and entropy features are calculated. In addition, we calculate the contrast as standard deviation minus the mean value of MSCN coefficient of the S, D and C maps. The 22 features are extracted in three versions of the S, D and C maps resulting from three different left and right LoG images. Total 66 features (22 $features\ per\ version$ $\times 3$ $versions$) are thus obtained for S, D and C maps, respectively. These features are used to train three S-SAEs for edge signal processing.

For an in-depth analysis on the potential for the utilization of these new features, we representatively choose features of S maps of 365 stereopairs on LIVE-1 database [42]. We plot several features versus DMOSs by distortion type in Fig. 11. For the five distortion types in LIVE-1 database, subjective scores decrease or increase monotonically with the increase of feature values. These monotonic and linear correlations can be easily interpreted and learned by regressor models. The features from D and C maps also have the equivalent potential in learning degree of distortion.

For Lum, RG and BY maps for left and right eyes, 3 AGGD fitting parameters (shape, left variance and right variance) are extracted. Refer to [45], two sample parameters (kurtosis and skewness) are also calculated. Because the 5 features are captured in Lum, RG and BY maps for left and right views, so total 30 features (5 $features\ per\ map$ $\times$ 6 $maps$) are captured to represent color quality degradation. We use these 30 features to train a SAE for color signal processing.

D. Pooling systems of DECOSINE

As we mentioned above, the obtained partial scores are pooled into an overall quality score. First, the quality scores that are obtained using the features of the S and D maps, Q_S and Q_D, are combined. According to [26], [27], weakly correlated two eyes' images induce similar weight to sum and difference channels. But, in general, the images are strongly correlated, and a larger weight is assigned to the difference



(a) (b) (c) (d) (e)

Fig. 11: The relation between some features from the S map and DMOSs by distortion type. (a) WN, feature #2-#5. (b) JP2K, feature #22-#25. (c) JPEG, feature #1-#4. (d) Blur, feature #23-#26. (e) FF, feature #24, #25, #30 and #55.

channel. Based on this theory, we design a dynamic weighting system:

$$W_D = C_1 - \frac{1-\rho(L,R)}{C_2}, \tag{13}$$

where

$$\rho(L,R) = \frac{E[(L-\mu_L)(R-\mu_R)]}{\sigma_L \sigma_R}. \tag{14}$$

In Eq. (14), μ_L and μ_R are expected values and σ_L and σ_R are standard deviations of L and R. Because there are disparities between left and right images, it is not elaborate to use L and R directly for calculating correlation in Eq. (14). Instead of L and R themselves, the 22 features that are identical to those for edge signals are extracted from left and right images to represent the images. These features are hardly affected by the disparities. For this work, we set $C_1 = 0.6, C_2 = 5$ to give a larger weight to the difference channel when two images are weakly correlated. As a result, $Q_{SD} = W_D Q_D + (1 - W_D)Q_S$. Second, the Q_{SD} and Q_C are pooled. The C map has strengths for treating asymmetrically distorted stereopairs because C map reflects the binocular suppression well. Thus, when the left and right images are weakly correlated, the C channel will have a greater impact. Accordingly, we design an another dynamic weighting system:

$$W_C = C_3 - \frac{1-\rho(L,R)}{C_4}. \tag{15}$$

We set $C_3 = 0.55, C_4 = 0.8$ to give a similar weight to two Q_{SD} and Q_C when the two images are strongly correlated. Likewise, 22 features from left and right images are used instead of L and R. Using this system, the quality for edge is determined by:

$$Q_{edge} = W_C Q_C + (1 - W_C)Q_{SD}. \tag{16}$$

Lastly, an overall quality score is computed. Because an edge signal is more important than color signal with regard to perceived quality [47], it is rational to give a larger weight on Q_{edge} than Q_{color}. We thus make a static weighting system:

$$Q = W_{edge} Q_{edge} + W_{color} Q_{color}, \tag{17}$$

where $W_{edge} + W_{color} = 1$. Fig. 12 shows Spearman rank-order correlation coefficients (SROCC) values according to the variation of the weights in the experiment conducted on the LIVE-1 database [42]. Note that the higher the SROCC, the better the performance. As shown in the graph, when two weights are set to $W_{edge} = 0.7$ and $W_{color} = 0.3$, the proposed DECOSINE performs best.

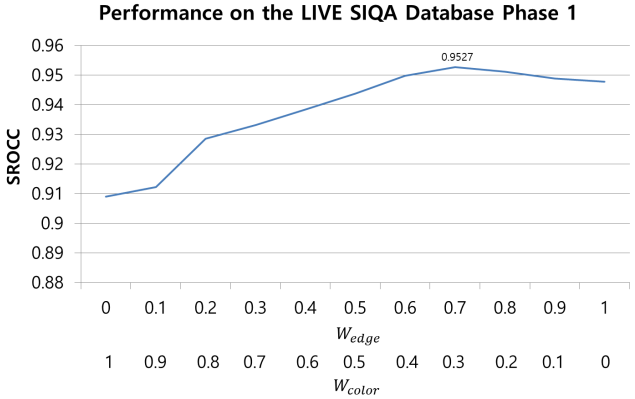


Fig. 12: The performance of DECOSINE according to the variation of W_{edge} and W_{color} in an experiment on the LIVE SIQA Database Phase 1. $W_{edge} = 0.7$ and $W_{color} = 0.3$ make DECOSINE works best.

IV. Experiments

A. S-SAEs, SAE and SVR configurations

DECOSINE needs three separate S-SAEs for edge signal and one SAE for color signal. We design each S-SAE including three local SAEs. The input vectors of these local SAEs are segments of the 66 features. Recall that there are three versions of S, D and C maps, respectively, because the three LoG maps having different thickness of the edges are computed for left and right views. We thus segment 66 features of S, D and C into three groups which are fed into local SAEs, and every local SAE has three hidden layers with $18-14-12$ nodes. The three S-SAEs for S, D and C channels are designed with the same structure. A SAE for color has three hidden layers with $25-20-15$ nodes. The reason we give three hidden layers to SAEs lies in better experimental results than SAEs having different number of hidden layers. Refer to Fig. 13.

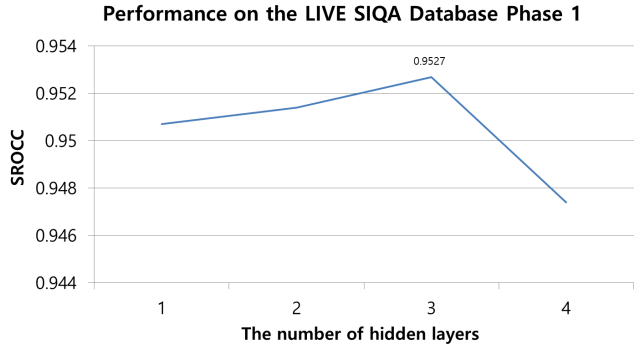


Fig. 13: The performance of DECOSINE according to the variation of the number of hidden layers in an experiment on the LIVE SIQA Database Phase 1. The three hidden layers make DECOSINE works best.

For training DECOSINE, some parameters need to be set. We obtained the optimum hyper-parameters considering experimental results of five popular SIQA databases synthetically, so DECOSINE was not overfit to a specific database. We set up with batch size as 1 due to lack of the number of samples in SIQA databases. As a result, the weights of networks are trained in the manner of stochastic gradient descent. A learning rate is set to 0.5 for all networks, and the number of epoch is 1000. To speed up the learning, we stop the process if full-batch train errors are less than 0.005 for first hidden layer and 0.001 for other layers.

In addition, the SVR also has two parameters: penalty (C) and kernel (γ) parameters. In this paper, $(C, \gamma) = (2^9, 1)$ is selected for all the SVRs. For your information, we use Deep Learning Toolbox [48] to train the S-SAEs and SAE, and LIBSVM [49] are utilized for training the SVRs.

B. SIQA Databases

To evaluate the performance of DECOSINE, five famous SIQA databases are utilized.

The LIVE SIQA Database Phase 1 (LIVE-1) [42] contains 20 reference stereopairs and 365 distorted ones. The size of left and right view images is 640×360 pixels. The left and right images of stereopairs were symmetrically degraded by WN, gaussian blur, JPEG, JP2K and fast fading. Differential mean opinion score (DMOS) values are provided as subjective quality scores on distorted ones.

The LIVE SIQA Database Phase 2 (LIVE-2) [16] consists of 8 reference stereopairs and 360 distorted ones that have size of 640×360 pixels like LIVE-1. Among them, 120 pairs were symmetrically distorted and 240 pairs were asymmetrically distorted. The distortion types are the same as LIVE-1. Due to the presence of asymmetrically distorted ones, the LIVE-2 is more challenging than LIVE-1. The DMOS values are also provided for distorted stereopairs.

The Waterloo IVC SIQA Database Phase 1 (WIVC-1) [50] has 6 reference stereopairs and 330 symmetrically or asymmetrically degraded ones. The size of left and right view images is 1390×1080. The types of distortions are WN, gaussian blur and JPEG. Each distortion type has 4 distortion levels that ensures a good perceptual separation. Mean opinion score (MOS) values are presented for subjective quality scores. Note that stereopairs in WIVC-1 has larger horizontal disparities than the other databases.

The Waterloo IVC SIQA Database Phase 2 (WIVC-2) [50] has 10 reference stereopairs and 460 symmetrically or asymmetrically degraded ones. The size of left and right view images is 1920×1080. Note that the resolution of images in WIVC-1 and WIVC-2 is better than the LIVE-1, LIVE-2 and IVC. The types of distortions are the same as them of WIVC-1. MOS values are provided for subjective quality scores.

The IVC SIQA Database (IVC) [5] contains 6 reference stereopairs, 90 symmetrically distorted ones and their associated DMOS values. Images have size of about 512×512. Four types of distortion like JPEG, JP2K, blur and down/up scaling were used to deteriorate stereopairs.

C. Algorithms and Performance Measures

To compare performance of DECOSINE, we exploit existing 9 IQA metrics: 3 FR 2D-IQA metrics (IWSSIM [47], VSI [51], VIF [52]), 3 FR SIQA metrics (Benoit *et al.* [5]'s scheme, Chen *et al.* [10]'s scheme, STRIQE [7]) and 3 NR 2D-IQA metrics (DIIVINE [53], BLIINDS-II [54], BRISQUE [44]). For convenience's sake, we call the algorithms authors' names for cases of no particular algorithm names. For quality prediction of stereopairs, FR 2D-IQA metrics are applied to left and right images, respectively. A mean value of the obtained scores is selected as an overall quality score. We name the 2D-FR extended algorithms 3D-IWSSIM, 3D-VSI and 3D-VIF, respectively. FR SIQA metrics are tested according to each researcher's instruction. For Benoit [5], we choose a ssim-d2 version among many others. One parameter α of STRIQE [7] is set to 0.8. Because previous NR SIQA metrics are not opened to the public, we model NR SIQA metrics using the NR 2D-IQA metrics. We first generate cyclopean images to deal with binocular perception and then extract their features from them. The features are used to train a SVR. For each algorithm, we obtained optimum SVR parameter sets by a grid search like: $(C, \gamma) = (2^9, 0.5), (2^{11}, 0.5)$ and $(2^{11}, 0.5)$, respectively. We name the algorithms CYC-DIIVINE, CYC-BLIINDS and CYC-BRISQUE, respectively. In addition to the

TABLE I: The performance evaluation on LIVE SIQA Database Phase I, Phase II, Waterloo IVC SIQA Database Phase I, Phase II and IVC SIQA Database. Top three performed metrics are highlighted in bold type.

Type	Metric	LIVE-1		LIVE-2		WIVC-1		WIVC-2		IVC	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR	3D-IWSSIM [47]	0.9420	0.9285	0.6912	0.7412	0.6795	0.5717	0.3669	0.2952	0.5339	0.6861
	3D-VSI [51]	0.8565	0.8583	0.6381	0.7473	0.7904	0.7378	0.5196	0.3992	0.5415	0.7274
	3D-VIF [52]	0.9255	0.9132	0.8537	0.8198	0.8013	0.7787	0.5722	0.4406	0.7544	0.7053
	Benoit [5]	0.7954	0.7891	0.7301	0.7018	0.4635	0.3321	0.3209	0.1651	0.5854	0.4523
	Chen [10]	0.9244	0.9113	0.8721	0.8975	0.6318	0.4577	0.4420	0.2724	0.4938	0.5852
	STRIQE [7]	0.9319	0.9164	0.8973	0.8819	0.6515	0.4681	0.5026	0.3441	0.8561	0.7978
NR	CYC-DIIVINE [53]	0.9457	0.9359	0.9235	0.9070	**0.9265**	0.9003	0.9098	0.8872	0.7820	0.7449
	CYC-BLIINDS [54]	0.9265	0.9167	0.9045	0.8992	0.8765	0.8490	0.8341	0.7965	0.8980	0.8789
	CYC-BRISQUE [44]	**0.9503**	**0.9408**	**0.9296**	**0.9210**	0.9180	**0.9067**	**0.9247**	**0.9072**	**0.9017**	**0.8834**
	DECOSINE-edge	**0.9561**	**0.9478**	**0.9415**	**0.9331**	**0.9476**	**0.9324**	**0.9261**	**0.9025**	**0.9366**	**0.9089**
	DECOSINE-color	0.9235	0.9091	0.9199	0.9096	0.8843	0.8610	0.8208	0.8042	0.8761	0.8537
	DECOSINE	**0.9615**	**0.9527**	**0.9497**	**0.9412**	**0.9439**	**0.9246**	**0.9331**	**0.9143**	**0.9469**	**0.9270**

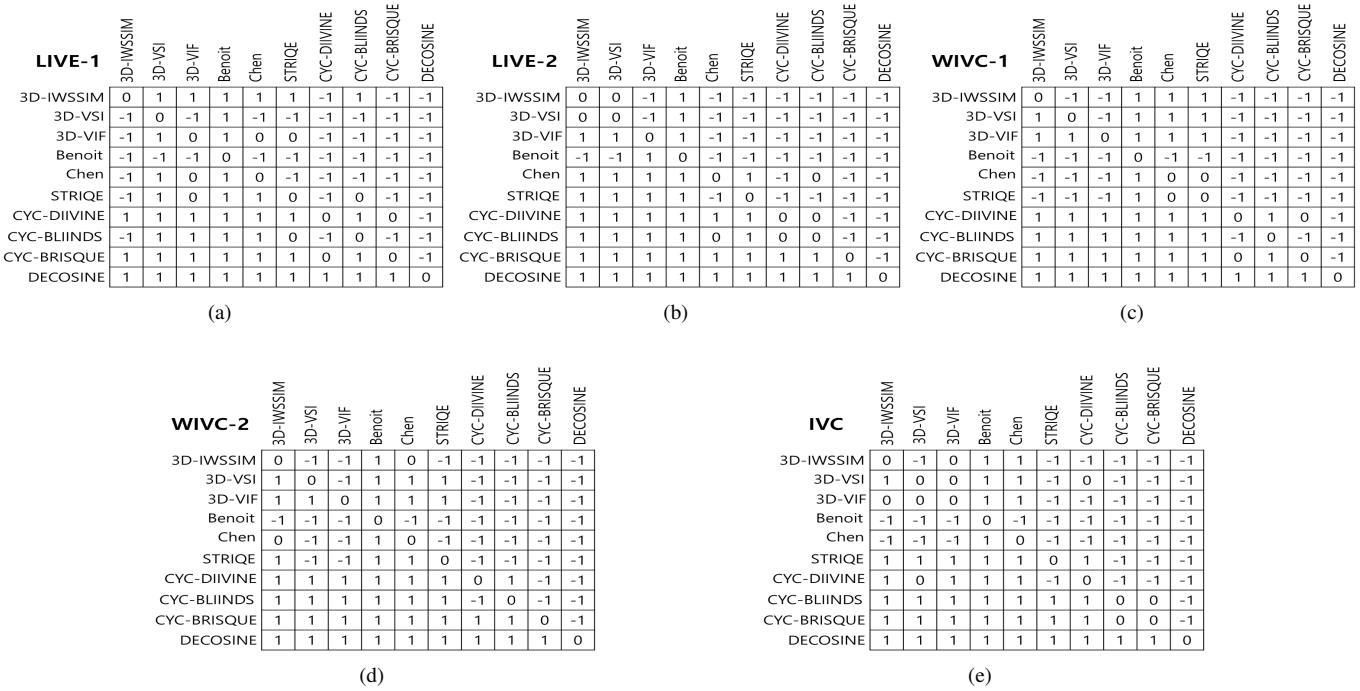


Fig. 14: Results of the t-test performed between SROCC values from the algorithms.

devised DECOSINE, its edge (DECOSINE-edge) and color (DECOSINE-color) parts are also tested separately. All parameters of the metrics are equally adapted regardless of databases. After nonlinear regression with a 5-parameter logistic function suggested from VQEG [55], we compute Pearson linear correlation coefficients ($PLCC$) between subjective scores and predicted scores. Calculating Spearman rank-order correlation coefficients ($SROCC$) does not need the nonlinear fitting process [47]. Generally, a good metric produces high $PLCC$ and $SROCC$ values. That is, $PLCC = SROCC = 1$ means a perfect match between the predicted scores and subjective scores. For evaluating NR methods including the proposed one, we randomly split the datasets into 80% training sets and 20% testing sets. Using the training sets and the corresponding DMOS or MOS values, NR metrics are trained. Then, we compute the predicted quality scores on the testing sets. For a fair comparison, the FR IQA metrics which do not require training are also tested on the 20% testing sets. It is repeated 100 times to remove the influence of the selection of training sets. Every time we repeat it, the $PLCC$ and $SROCC$ are computed for performance comparison, and each mean value is finally reported.

D. Test on Individual Databases

The test results on LIVE-1, LIVE-2, WIVC-1, WIVC-2 and IVC databases are shown in Table I. The proposed metric delivers the best performance on all the databases. Although it belongs to a NR method, the performance is better than that of FR metrics. On LIVE-1, its $SROCC$ exceeds 0.95. Although

TABLE II: The performance of the NR methods under different partition proportions on WIVC-2. SROCC values are reported for performance evaluation.

Metric	80%/20%	70%/30%	60%/40%	50%/50%
CYC-DIIVINE [53]	0.8872	0.8779	0.8679	0.8500
CYC-BLIINDS [54]	0.7965	0.7907	0.7834	0.7719
CYC-BRISQUE [44]	0.9072	0.9035	0.8976	0.8892
DECOSINE	0.9143	0.9101	0.9090	0.8996

LIVE-2 contains asymmetrically distorted stereopairs, it does not significantly affect the performance of DECOSINE, as shown in $SROCC = 0.9412$. On WIVC-1 and WIVC-2 which have high resolution stereopairs distorted asymmetrically, the proposed method yields the best prediction among the algorithms. In addition, DECOSINE shows robust ability of prediction on IVC that has only 90 stereopairs and it is demonstrated from $SROCC = 0.9270$. Although training-based NR metrics generally require a lot of training samples to achieve reliable prediction, the proposed method produces satisfactory results on IVC. The possible reasons why our algorithm performs well are due to the decent model about the whole visual perception route and utilization of deep learning.

We can observe that the predicted scores about edge quality are more consistent with subjective scores than the color quality scores. It seems that the edge part plays a more important role in DECOSINE. However, the integration of these two parts increases prediction accuracy except on WIVC-1. It verifies that the color part is also helpful for prediction of perceived quality as well as the edge part.

Many other algorithms show quite good performance on LIVE-1. Especially CYC-BRISQUE delivers $SROCC = 0.9408$ that is competitive with DECOSINE. However, the performance of many algorithms is weakened on the LIVE-2, WIVC-1 and WIVC-2 that contain asymmetrically distorted stereopairs. It means most of the methods cannot interpret binocular visual properties properly. On IVC, only CYC-BLIINDS and CYC-BRISQUE among other algorithms deliver the fine performance.

To assess the statistical significance of the performance difference between any two metrics, we further conduct Welch's t-test [56] using the 100 SROCC values. The number '1' indicates that the row metric is statistically superior to the column metric, whereas the number '-1' indicates that the row is statistically worse than the column. The number '0' indicates that the two metrics are statistically indistinguishable. The results of the t-test are shown in Fig. 14. On all the databases, the proposed method is statistically superior to the others.

We further report results under other three partition proportions on WIVC-2 which is the largest database among the databases we used: 70%, 60% and 50% samples are used for training and the remaining 30%, 40% and 50% are used for testing, respectively. As shown in Table II, the partition ratio has little effect on the performance of DECOSINE and it does not suffer from an over-fitting problem.

E. Cross-Database Test

To verify the generalization capability of our proposed algorithm, we implement cross-database tests. Among the databases, IVC is excluded because it has very few stereopairs to be used for the cross-database test. LIVE-1 and LIVE-2 present DMOS values for subjective quality scores, while WIVC-1 and WIVC-2 provide MOS values. Because DMOS and MOS values are produced by different process [24], cross-database tests between LIVE databases and WIVC databases are not proper. Thus, total four tests are implemented: 1) the algorithms are trained on LIVE-1 and tested on LIVE-2 (LIVE-1/LIVE-2), 2) LIVE-2/LIVE-1, 3) WIVC-1/WIVC-2 and 4) WIVC-2/WIVC-1. NR methods are trained on the former database and tested on the latter one. On the contrary, FR methods are tested using whole samples of the latter one without training them because they do not need a training phase.

From the results present in Table III, we can observe four points. 1) As shown in weighted average PLCC and SROCC values across the four tests, the proposed DECOSINE computes a reliable prediction about the quality of stereopairs despite cross-database tests. Some algorithms deliver decent performance in one or two cross-database tests, but DECOSINE is the only one which computes good performance on all the tests. 2) The performance of the other NR algorithms in the cross-database tests is not as good as that in the individual database tests. CYC-BRISQUE that shows impressive performance on the individual database tests can hardly predict perceptual quality well on the cross-database tests. 3) The performance of FR algorithms remains no matter what kind of cross-database tests. Especially, 3D-VIF ranks second following DECOSINE. 4) When quality scores about edge and color are combined, the predictive performance is obviously improved. The performance improvement degree in cross-database tests is larger than that of individual database tests. It means the integration of edge and color quality has potential in real-life applications of SIQA methods.

F. Performance on Individual Distortion Types

Further experiments have been conducted to demonstrate the performance of DECOSINE on individual distortion types. For training-based DECOSINE, we select the LIVE-1 and LIVE-2 for our experiments as these two datasets contain the same types of distortion. After DECOSINE is trained by a subset of the LIVE-2 among five subsets which are partitioned by distortion type, we test it on a subset of the LIVE-1 which of stereopairs are degraded by the same distortion type. For comparing the performance, the other NR metrics are also examined. The test results are presented in Table IV which proves that DECOSINE predicts perceptual quality well regardless of types of distortion, allowing for the cross-database test. Comparing with the NR metrics, our proposed metric delivers the most stable performance. On JPEG subsets, the prediction accuracy of DECOSINE is not so good, but it is better than that of the other algorithms. Note that the performance of metrics on whole LIVE-2 is better than that on other subsets. It can be explained by that the nonlinear

TABLE III: Results of cross-database tests. If metrics were trained on LIVE-1 and tested on LIVE-2, we presented as (training database)/(testing database). Top three performed metrics are highlighted in bold types.

Metric	LIVE-1 / LIVE-2		LIVE-2 / LIVE-1		WIVC-1 / WIVC-1		WIVC-2 / WIVC-1		Weighted average	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
3D-IWSSIM [47]	0.6357	0.7491	**0.9370**	**0.9327**	0.3771	0.3025	0.7023	0.5900	0.6443	0.6231
3D-VSI [51]	0.7656	0.7480	0.8627	0.8649	0.5344	0.4089	0.7892	0.7368	0.7239	0.6708
3D-VIF [52]	0.8366	0.8186	**0.9243**	**0.9195**	0.5486	0.4449	**0.8075**	**0.7835**	**0.7639**	0.7218
Benoit [5]	0.7346	0.7226	0.8057	0.7936	0.3157	0.1578	0.2186	0.3389	0.5259	0.4846
Chen [10]	**0.9073**	**0.9013**	0.9200	0.9150	0.4818	0.2671	0.6320	0.4724	0.7212	0.6186
STRIQE [7]	**0.8886**	**0.8826**	**0.9260**	**0.9211**	0.4560	0.3561	0.7252	0.4772	0.7307	0.6437
CYC-DIIVINE [53]	0.5528	0.5230	0.5220	0.4839	0.4483	0.4201	0.2337	0.2207	0.4441	0.4165
CYC-BLIINDS [54]	0.7470	0.7430	0.8070	0.8062	0.0893	0.1102	0.5940	0.5011	0.5284	0.5134
CYC-BRISQUE [44]	0.4961	0.4920	0.8428	0.8325	0.6740	0.6474	0.4131	0.4566	0.6156	0.6135
DECOSINE-edge	0.8110	0.7870	0.8911	0.8871	**0.7968**	**0.7940**	**0.8489**	**0.8077**	**0.8342**	**0.8178**
DECOSINE-color	0.7896	0.7526	0.7883	0.7898	**0.7203**	**0.6719**	0.7405	0.7385	0.7576	**0.7340**
DECOSINE	**0.8456**	**0.8231**	0.9161	0.9149	**0.8421**	**0.8313**	**0.8739**	**0.8687**	**0.8677**	**0.8676**

TABLE IV: The performance of DECOSINE on different types of distortion. DECOSINE is trained on each distortion subset of the LIVE-2 and tested on the same distortion subset of LIVE-1. The top performing metric is highlighted in bold face.

Metric	WN		JP2K		JPEG		Blur		FF		All	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
CYC-DIIVINE [53]	0.9002	0.8488	0.8829	0.8027	0.3683	0.2737	0.9114	0.8704	0.7079	0.6253	0.5220	0.4839
CYC-BLIINDS [54]	**0.9283**	0.9323	0.6530	0.6136	0.6507	0.5868	0.9212	0.8901	0.6765	0.6172	0.8070	0.8062
CYC-BRISQUE [44]	0.8922	0.8923	0.7713	0.7308	0.1825	0.1085	**0.9317**	0.8794	0.6974	0.6879	0.8428	0.8325
DECOSINE	0.9020	**0.9443**	**0.9098**	**0.8493**	**0.6750**	**0.5879**	0.9251	**0.9004**	**0.8124**	**0.7575**	**0.9161**	**0.9149**

regression using the 5-parameter logistic function is largely influenced by the number of samples. As described in this experimental results, DECOSINE can be used for general purpose IQA tasks [57], [58] unlike distortion-specific IQA metrics [59], [60].

G. Time Efficiency of S-SAE

When implementing deep learning algorithms, the biggest challenge is the need of a very long time for a training phase. In our proposed approach, this problem is alleviated by using segmentation concept (Section II-B). To compare training times of a S-SAE and a traditional SAE, we select a S map. In our DECOSINE, the 66 features of the S map are inputted into a S-SAE that consists of three local SAEs having an input layer and three hidden layers of $22-18-14-12$ nodes, respectively. For comparison, we also implement a conventional SAE for the S map instead of the S-SAE. The SAE is set to an input layer and three hidden layers with $66-54-42-36$ nodes. In Table V, we list the time taken on a PC with Intel Core i7 CPU at 2.80 GHz, 8.00 GB RAM, Windows 10 64-bit, and MATLAB R2017a. The segmentation does not show much performance improvement on IVC containing only 90 stereopairs, whereas the computation times on comparatively larger databases are substantially decreased. If the size of the test database is not too small, using the S-SAE reduces the time complexity in comparison to using the SAE. An average reduction rate, 76.00%, reflects the decent time efficiency of using S-SAEs. At the same time, the accuracy and monotonicity of prediction are maintained. On LIVE-1, we compute one of local quality scores, Q_S, using SAE + SVR and S-SAE + SVR, respectively. The result shows similar or better mean $PLCC$ and $SROCC$ values as shown in $0.9491, 0.9380$ when SAE is used and $0.9531, 0.9398$ when S-SAE is used. Through it we can know using S-SAE is better than using SAE in terms of time complexity.

TABLE V: The computation time in a SAE and a S-SAE.

Dataset	SAE (sec)	S-SAE (sec)	Reduction
LIVE-1	35.30	21.22	60.11%
LIVE-2	30.26	19.74	65.23%
WIVC-1	16.58	12.80	77.20%
WIVC-2	18.25	14.01	76.77%
IVC	20.20	20.35	100.7%
Average	24.12	17.62	76.00%

TABLE VI: The computation time of NR algorithms.

Metrics	CYC-DIIV.	CYC-BLII.	CYC-BRIS.	DECOSINE
Time (sec)	20.82	30.31	9.10	28.26

In addition, we compare the time complexity between DECOSINE and other NR methods. The run time of metrics for predicting quality of a stereopair from LIVE-1 is calculated. The algorithms are first trained on LIVE-2, and tested on the stereopair. Table VI shows DECOSINE has a moderate time complexity.

V. Conclusions

We propose a NR SIQA algorithm named DECOSINE based on the whole visual perception route from eyes to the frontal lobe. Especially functions of retinal ganglion cells (RGC) and lateral geniculate nucleus (LGN) about edge and color signal processing are importantly considered, and segmented stacked autoencoders (S-SAE) is utilized to model deep and complex structure of the visual cortex. Our DECOSINE computes two locally estimated scores: edge quality and color quality scores. Inspired by that binocular integration occurred in V1 after edge extraction of retinal ganglion cells, sum, difference and cyclopean maps are computed from LoG filtered left and right images. The opponent coding theory is utilized for modeling color information processing occurred in LGNs. The quality-aware features are mapped into local quality scores via combination of S-SAEs/SAE and SVRs. These scores are combined into an overall score through two dynamic and one static weighting systems. Experiments have been conducted on popular five SIQA databases and the results verify a good and reliable performance of DECOSINE in comparison with previous IQA metrics.

Although the proposed metric shows good performance, there is still a room for improvement. We did not deal with visual comfort aspect directly. This aspect is closely connected with the development of 3D images and movies because most viewers take count of it [61], [62]. We will make an effort to add it to DECOSINE in future work.

For the development of SIQA field, larger and realistic databases are urgently required. Previous databases contain stereopairs corrupted by only one of a few synthetically introduced distortions. In addition, the number of samples are not sufficient. In LIVE-1, LIVE-2, WIVC-1, WIVC-2 and IVC SIQA databases, there are 365, 360, 330, 460 and 90 stereopairs, respectively. We hope new databases consisting of more samples and corresponding DMOS or MOS values will be constructed.

References

[1] H. Zhang, J. Yang, Y. Zhang, and T. S. Huang, "Image and video restorations via nonlocal kernel regression," *Cybernetics IEEE Transactions on*, vol. 43, no. 3, pp. 1035–1046, 2012.

[2] L. Shao, R. Yan, X. Li, and Y. Liu, "From heuristic optimization to dictionary learning: a review and comprehensive comparison of image denoising algorithms." *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1001–1013, 2014.

[3] P. Campisi, P. L. Callet, and E. Marini, "Stereoscopic images quality assessment," in *Signal Processing Conference, 2007 European*, 2007, pp. 2110–2114.

[4] J. Yang, C. Hou, Y. Zhou, and Z. Zhang, "Objective quality assessment method of stereo images," in *3dtv Conference: the True Vision - Capture, Transmission and Display of 3d Video*, 2009, pp. 1–4.

[5] A. Benoit, P. L. Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *Eurasip Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–13, 2009.

[6] 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 *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010.

[7] S. K. Md, B. Appina, and S. S. Channappayya, "Full-reference stereo image quality assessment using natural stereo scene statistics," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1985–1989, 2015.

[8] X. Wang, S. Kwong, Y. Zhang, and Y. Zhang, "Considering binocular spatial sensitivity in stereoscopic image quality assessment," in *Visual Communications and Image Processing (VCIP), 2011 IEEE*, 2011, pp. 1–4.

[9] S. Ryu, H. K. Dong, and K. Sohn, "Stereoscopic image quality metric based on binocular perception model," in *IEEE International Conference on Image Processing*, 2012, pp. 609–612.

[10] 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.

[11] K. Lee and S. Lee, "3d perception based quality pooling: Stereopsis, binocular rivalry, and binocular suppression," *Selected Topics in Signal Processing IEEE Journal of*, vol. 9, no. 3, pp. 533–545, 2015.

[12] Y. Zhang and D. M. Chandler, "3d-mad: A full reference stereoscopic image quality estimator based on binocular lightness and contrast perception." *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 24, no. 11, pp. 3810–3825, 2015.

[13] C. T. E. R. Hewage and M. G. Martini, "Reduced-reference quality metric for 3d depth map transmission," in *3dtv-Conference: the True Vision - Capture, Transmission and Display of 3d Video*, 2010, pp. 1–4.

[14] Q. Xu, G. Zhai, M. Liu, and K. Gu, "Using structural degradation and parallax for reduced-reference quality assessment of 3d images," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2014, pp. 1–6.

[15] W. Zhou, G. Jiang, M. Yu, F. Shao, and Z. Peng, "Reduced-reference stereoscopic image quality assessment based on view and disparity zero-watermarks," *Signal Processing Image Communication*, vol. 29, no. 1, pp. 167–176, 2014.

[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, p. 3379, 2013.

[17] C. C. Su, L. K. Cormack, and A. C. Bovik, "Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation." *Image Processing IEEE Transactions on*, vol. 24, no. 5, p. 1685, 2015.

[18] 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*, p. 1, 2015.

[19] 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 Recognition*, vol. 59, no. C, pp. 176–187, 2016.

[20] F. Shao, W. Tian, W. Lin, and G. Jiang, "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. 1–1, 2016.

[21] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[22] R. J. Snowden, P. Thompson, and T. Troscianko, *Basic vision : an introduction to visual perception*. Oxford University Press, 2006.

[23] J. D. Pettigrew, "The neurophysiology of binocular vision." *Scientific American*, vol. 227, no. 2, pp. 84–95, 1972.

[24] S. Ryu and K. Sohn, "No-reference quality assessment for stereoscopic images based on binocular quality perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 591–602, 2014.

[25] 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.

[26] K. A. May, L. Zhaoping, and P. B. Hibbard, "Perceived direction of motion determined by adaptation to static binocular images." *Current Biology Cb*, vol. 22, no. 1, pp. 28–32, 2012.

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

[28] S. H. Hendry and R. C. Reid, "The koniocellular pathway in primate vision." *Annual Review of Neuroscience*, vol. 23, no. 1, p. 127, 2000.

[29] S. PH, S. JH, and M. JH, "Functions of the on and off channels of the visual system," *Nature*, vol. 322, no. 6082, pp. 824–825, 1986.

[30] P. G. Lovell, D. J. Tolhurst, C. A. Parraga, R. Baddeley, U. Leonards, J. Troscianko, and T. Troscianko, "Stability of the color-opponent signals under changes of illuminant in natural scenes." *Journal of the Optical Society of America A Optics Image Science and Vision*, vol. 22, no. 10, pp. 2060–2071, 2005.

[31] R. L. De Valois and K. K. De Valois, "A multi-stage color model." *Vision Research*, vol. 33, no. 8, pp. 1053–1065, 1993.

[32] J. Ding and G. Sperling, "A gain-control theory of binocular combination," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 4, pp. 1141–1146, 2006.

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

[34] J. Yang, Y. Lin, Z. Gao, Z. Lv, W. Wei, and H. Song, "Quality index for stereoscopic images by separately evaluating adding and subtracting," *Plos One*, vol. 10, no. 12, p. e0145800, 2015.

[35] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio, "A quantitative theory of immediate visual recognition," *Progress in Brain Research*, vol. 165, no. 6, pp. 33–56, 2007.

[36] M. Bianchini and F. Scarselli, "On the complexity of neural network classifiers: a comparison between shallow and deep architectures." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1553–1565, 2014.

[37] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du, and S. Marshall, "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, 2016.

[38] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–55, 2009.

[39] E. Ergul, N. Arica, N. Ahuja, and S. Erturk, "Clustering through hybrid network architecture with support vectors," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2016.

[40] D. B. Chklovskii and A. A. Koulakov, "Maps in the brain: What can we learn from them?" *Annual Review of Neuroscience*, vol. 27, no. 1, pp. 369–392, 2004.

[41] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006.

[42] 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.

[43] D. L. Ruderman, T. W. Cronin, and C. C. Chiao, "Statistics of cone responses to natural images: implications for visual coding," *Journal of the Optical Society of America A*, vol. 15, no. 15, pp. 2036–2045, 1998.

[44] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain." *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 21, no. 12, pp. 4695–4708, 2012.

[45] D. Ghadiyaram and A. C. Bovik, "Feature maps driven no-reference image quality prediction of authentically distorted images," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9394, pp. 93 940J–93 940J–14, 2015.

[46] D. Tao, "Sparse representation for blind image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1146–1153.

[47] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, p. 1185, 2011.

[48] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," *Technical University of Denmark*, 2012.

[49] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[50] J. Wang, A. Rehman, K. Zeng, S. Wang, and Z. Wang, "Quality prediction of asymmetrically distorted stereoscopic 3d images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3400–3414, 2015.

[51] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 23, no. 10, p. 4270, 2014.

[52] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[53] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 20, no. 12, p. 3350, 2011.

[54] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: a natural scene statistics approach in the dct domain." *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 21, no. 8, p. 3339, 2012.

[55] V. Q. E. Group *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii (fr_tv2)," *ftp://ftp. its. bldrdoc. gov/dist/ituvidq/Boulder_VQEG_jan_04/VQEG_PhaseII_FRTV_Final_Report_SG9060 doc, 2003*, 2003.

[56] G. D. Ruxton, "The unequal variance t-test is an underused alternative to student's t-test and the mann?whitney u test," *Behavioral Ecology*, vol. 17, no. 4, pp. 688–690, 2006.

[57] X. Gao, F. Gao, D. Tao, and X. Li, "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 12, pp. 2013–2026, 2013.

[58] T. J. Liu, K. H. Liu, J. Y. Lin, W. Lin, and C. J. Kuo, "A paraboost method to image quality assessment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 1, pp. 107–121, 2015.

[59] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot, "No-reference image blur assessment based on discrete orthogonal moments." *IEEE Transactions on Cybernetics*, vol. 46, no. 1, p. 39, 2016.

[60] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: From quality assessment to automatic enhancement." *IEEE Transactions on Cybernetics*, vol. 46, no. 1, p. 284, 2016.

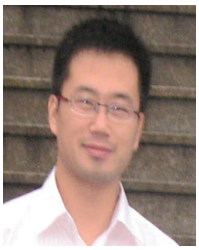
[61] F. L. Kooi and A. Toet, "Visual comfort of binocular and 3d displays," *Displays*, vol. 25, no. 2?3, pp. 99–108, 2004.

[62] L. M. J. Meesters, W. A. Ijsselsteijn, and P. J. H. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional tv," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 381–391, 2004.

Jiachen Yang (M'13) received the M.S. and Ph.D. degrees in communication and information engineering from Tianjin University, Tianjin, China, in 2005 and 2009, respectively. He is currently a professor at Tianjin University. He is also a visiting scholar with the Department of Computer Science, School of Science, Loughborough University, U.K. His research interests include stereo camera, stereo vision research, pattern recognition, stereo image displaying, and quality evaluation.

Kyohoon Sim received the B.S. degrees from the School of Electronic Engineering, Dongguk University, Seoul, Korea, in 2015. He is currently pursuing the M.S. degree at the School of Electrical Automation and Information Engineering, Tianjin University, Tianjin, China. His research interests include stereoscopic image quality assessment, 3D visual perception, computer vision and deep learning.

Xinbo Gao (M'02, SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xian, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, a Professor of Pattern Recognition and Intelligent System, and the Director of the State Key Laboratory of Integrated Services Networks, Xian. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He has authored five books and around 200 technical articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, International Journal of Computer Vision, and Pattern Recognition. Prof. Gao is currently a fellow of the Institution of Engineering and Technology. He is on the Editorial Boards of several journals, including Signal Processing (Elsevier), and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences.



Wen Lu (M'13) received the M.S. and Ph.S. degrees in electrical engineering from Xidian University, China, in 2006 and 2009, respectively. He is currently an Associate Professor at Xidian University. His research interests include image and video understanding, visual quality assessment, and computational vision.

Qinggang Meng (M'06, SM'18) received the B.S. and M.S. degrees from the School of Electronic Information Engineering, Tianjin University, China, and the Ph.D. degree in computer science from Aberystwyth University, U.K. He is a Senior Lecturer with the Department of Computer Science, Loughborough University, U.K. His research interests include biologically and psychologically inspired learning algorithms and developmental robotics, service robotics, robot learning and adaptation, multi-UAV cooperation, drivers distraction detection, human motion analysis and activity recognition, activity pattern detection, pattern recognition, artificial intelligence, and computer vision. He is a Fellow of the Higher Education Academy, U.K.

Baihua Li received her BSc and MSc degrees in Electronic Engineering from Tianjin University and her PhD degree in Computer Science from Aberystwyth University. She has worked at Tianjin University and Manchester Metropolitan University before she joined the Department of Computer Science at Loughborough University. Her research emphasizes innovations and novel applications of internet of things, computer vision and pattern recognition techniques in various fields. More than 50 papers have been published in high impact journals and conferences of international standard, such as Pattern Recognition, IEEE Trans Syst Man Cybern and IEEE Trans Biomed Eng.