A Full-Reference Stereoscopic Image Quality Measurement via Hierarchical Deep Feature Degradation Fusion

Qiuping Jiang, Wei Zhou, Xiongli Chai, Guanghui Yue,
Feng Shao, *Member, IEEE,* and Zhibo Chen, *Senior Member, IEEE*

***Abstract*—For the problem of stereoscopic image quality measurement (SIQM), it is difficult to design an efficient yet reliable full reference (FR) SIQM method due to our limited knowledge about the properties of human binocular vision. Inspired by the fact that the input visual information is hierarchically processed in our human brain, we consider different levels of distortion in an image cause individual degradations on hierarchical features, and propose to fuse the degradations on hierarchical features to facilitate the task of FR-SIQM. As one of the most classical convolutional neural network (CNN) architectures, the VGG-16 network is first applied to each view of the stereopair to build hierarchical deep feature representations based on which monocular quality estimation (MQE) and binocular quality fusion (BQF) are then performed. Specifically, the MQE stage estimates a set of layer-wise monocular quality scores by measuring the similarity between the hierarchical feature maps of the distorted monocular view and those of the reference monocular view. The BQF stage estimates a set of layer-wise binocular quality scores via a weighted average of the corresponding layer-wise monocular quality scores. The adaptive weights are determined by a modified hierarchical feature energy-based Gain-Control model. Finally, the layer-wise binocular quality scores across all layers are fused into an overall binocular quality score via regression. Experiments on three benchmark databases validate the state-of-the-art performance of our method.**

***Index Terms*—Image quality measurement, stereoscopic image, full reference, monocular quality estimation, binocular quality fusion, convolutional neural network.**

I. Introduction

OBJECTIVE image quality measurement (IQM) is important for various image processing and computer vision applications such as diagnostic medical imaging [1], [2], object detection [3], face recognition [4], and environment monitoring [5], [6]. Quality measurement of traditional 2D images have been extensively studied, and many promising quality estimators have been proposed [7-13]. Over the past years, owing to the increase of stereoscopic three-dimensional (3D) image/video services in our daily life, stereoscopic 3D visual contents are becoming the new research target of IQM [14-23].

Compared with its 2D counterpart, stereoscopic 3D image quality measurement (SIQM) encounters more challenges as more influential factors such as image distortion, depth perception, visual discomfort, and visual presence, need to be considered simultaneously [24]. However, this task is quite challenging given that the interactions among these factors are complex and it is difficult to precisely simulate them without sufficient understanding about human binocular visual properties. Due to this, the existing works mainly focus on investigating the influence of each individual factor on 3D quality of experience (QoE) [14-23] and this paper focus on evaluating the visual quality of distorted stereoscopic images in a full-reference (FR) manner.

A stereoscopic image pair contains two slightly different monocular views (captured by two position-shifted cameras), each of which is controlled to be separately projected onto the human retina. Owing to the slight inter-view difference (i.e., disparity), depth perception emerges with binocular fusion in human brain. In practice, a stereopair can be either symmetrically or asymmetrically distorted. For example, some researchers have proposed asymmetric coding methods for stereoscopic videos to achieve maximal coding efficiency while without causing any perceptible quality degradation [25]. An effective FR-SIQM metric should be well applicable to handle both symmetrically and asymmetrically distorted stereoscopic images. In general, the FR-SIQM task can be much easier for the symmetric case because even a simple application of traditional FR 2D-IQM metrics to the two monocular views separately can achieve fairly reliable performance [26-29]. Unfortunately, when it comes to the asymmetric case, the simple averaged FR 2D-IQM estimators are found to deliver much less effectiveness. The asymmetrically distorted stereopairs can make the FR-SIQM problem much more challenging primarily due to the different distortion types and levels of the two monocular views. Previous works have observed that the overall quality of an asymmetrically distorted stereopair is a result jointly affected by monocular quality and binocular fusion, depending on the distortion types and levels [30-32].

This work was supported in part by the Natural Science Foundation of China (61901236, U1908209), in part by the Natural Science Foundation of Ningbo (2019A610097), in part by the Zhejiang Natural Science Foundation of China (R18F010008), and in part by the National Science Foundation of Shenzhen University (860-000002110122). It was also sponsored by K.C. Wong Magna Fund in Ningbo University. *(Q. Jiang and W. Zhou contributed equally to this work.) (Corresponding author: Zhibo Chen.)*

Q. Jiang, X. Chai and F. Shao are with the School of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: jiangqiuping@nbu.edu.cn).

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

G. Yue is with the School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China (e-mail: guanghuiyue.doctor@gmail.com).

To design a robust FR-SIQM metric that is well applicable to both symmetric and asymmetric distortions, an intuitive idea is to address the above two issues in a way coincident with the visual perception processes. It is known that the input visual information is hierarchically processed in human brain and the HVS understands an image according to hierarchical (i.e., low-level, mid-level, and high-level) features [33]. The human brain will try to understand the semantic information of an image and rate the quality according to hierarchical features. Scene understanding and quality evaluation are intrinsically related since both of them depend on how the HVS perceives an image and distortions can be a strong effector of scene understanding [34,35]. This inspires us to investigate the relationship between the change of hierarchical features and perceived quality degradation. The key principle is that perceptible quality degradations can lead to measurable changes of hierarchical features.

Convolutional neural network (CNN) [36], which has achieved outstanding performance in many different computer vision tasks [37-39], is capable of capturing hierarchical features from images. In CNNs, the lower layers respond to low-level primitive image elements such as edges, corners and shared common patterns, and the higher layers tend to extract high-level semantics like object parts or faces. Inspired by this, we propose to use deep CNNs to build hierarchical feature representations of stereoscopic images and propose a FR-SIQM method by measuring and fusing the degradations on the extracted hierarchical features. Although the similar idea has been utilized to address the traditional FR 2D-IQM problem [40], it remains a great challenge to adapt it to stereoscopic images where the binocular fusion is deemed as a key step. For our concerned FR-SIQM task, the extracted hierarchical features are utilized as the bases for both monocular quality estimation (MQE) and binocular quality fusion (BQF). The main contributions are summarized as follows:

- We make the first attempt to extract hierarchical features from pre-trained deep CNN to facilitate FR-SIQM and demonstrate that perceptible quality degradations of stereoscopic images can lead to measurable changes of hierarchical deep features.
- We utilize the extracted hierarchical deep features as the bases for both MQE and BQF. Especially for BQF, a hierarchical feature energy-based Gain-Control model is formulated to adaptively fuse the two monocular-view quality scores.
- We demonstrate that both MQE and BQF should be performed in a hierarchical manner in the context of FR-SIQM. This is justified by the fact that the structure of human brain is inherently hierarchical so that our designed quality metric should well resemble this property.

II. Related Works

A. Existing FR-SIQM Algorithms

In the past decades, many FR-SIQM algorithms have been proposed. Some pioneering FR-SIQM metrics [13,14,22,23] tried to assess a distorted stereopair by extending the traditional 2D FR-SIQM metrics with further considering the depth information. However, estimating the depth information from a certain stereopair is difficult and time-consuming. In addition, whether evaluating the depth information using 2D-IQM metrics is suitable for SIQM still remains an open problem. Unlike these pioneering methods, the recent studies have found that the performance of FR-SIQM models is largely beneficial from the simulation of binocular visual characteristics. Bensalma and Larabi [56] proposed a binocular energy-based quality metric (BEQM) by computing the binocular energy difference between the original and distorted stereopairs. It simulates the critical binocular fusion process that characterizes the human stereopsis perception. Lin and Wu [57] simulated the binocular frequency integration behaviors and utilized them as the bases for binocular combination when adapting the existing 2D-IQM metrics to evaluate stereoscopic images. Chen *et al.* [58] proposed to synthesize a cyclopean image for FR-SIQM from the stereopair and its corresponding disparity map to characterize the underlying binocular rivalry caused by asymmetrical distortions. Then, by applying traditional 2D FR-IQM quality metrics on this synthesized cyclopean image, a final 3D quality score was obtained. Zhang and Chandler [30] extended the traditional 2D-most apparent distortion (2D-MAD) method [10] to a 3D-MAD measure which decomposed the problem of FR-SIQM into two sub-modules: 1) distortion of the monocular views and 2) distortion of the cyclopean view. Wang *et al.* [52] devised a binocular rivalry-inspired multi-scale model for binocular combination which particularly focuses on the quality evaluation of asymmetrically distorted stereopairs. This method is developed based on the hypothesize that the strength of view dominance in binocular rivalry is related to the relative energy of the two views. Lin *et al.* [59] proposed a FR-SIQM method by estimating the cyclopean amplitude map and the cyclopean phase map based on low-level features and binocular fusion. In addition, visual saliency detection results are also used to modulate the amplitude component for subsequent cyclopean amplitude map generation. Khan and Channappayya [60] estimates the quality of stereopairs by extracting depth-salient edges to refine the local quality maps of both monocular images for spatial pooling. Ma *et al.* [61] first applied the existing FR 2D-IQM algorithms on each view of the reference and distorted stereopairs, and then combined the monocular quality scores via a linear weighting method to obtain the binocular quality score. Zhou *et al.* [62] proposed an SIQM metric based on sparse representation and binocular combination. Both sparse coefficients and gradient magnitude are used for monocular view quality measurement. Then, both first-order and second-order binocular combination strategies are adopted to fuse the monocular view quality scores.

B. Problem Statement

Given a distorted stereoscopic image $\{I_L^D, I_R^D\}$ and its corresponding original version $\{I_L^O, I_R^O\}$ as reference, the target of FR-SIQM is to evaluate the visual quality of $\{I_L^D, I_R^D\}$ in a perceptually consistent manner based on some extracted features from the distorted and original stereoscopic images. Once the features are extracted, MQE and BQF can be

performed. We categorize the existing FR-SIQM methods into two classes according to the execution sequence of MQE and BQF:

Early combination-based: The methods belonging to this category first perform binocular combination to generate a merged view called cyclopean image from the given left and right view images and then perform MQE on this cyclopean image to obtain the final quality score [30,58,59,61,62]. This framework can be described as follows:

$$Q = \mathbb{M}(\mathbb{C}(\mathbb{F}(I_L^O), \mathbb{F}(I_R^O)), \mathbb{C}(\mathbb{F}(I_L^D), \mathbb{F}(I_R^D))), \quad (1)$$

where $\mathbb{F}$, $\mathbb{C}$, and $\mathbb{M}$ denote the feature extractor, the cyclopean image generation operator, and the MQE operator, respectively. The existing methods belonging to this category are different with each other in terms of either all or part of these three operators. Different with those later combination-based SIQM methods, these early combination-based SIQM methods usually relies on the disparity map to link the left and right view images for cyclopean image generation. However, reliable and efficient disparity estimation is still an open problem especially for the distorted stimuli, making such frameworks problematic and not applicable to real-time applications.

Later combination-based: The methods belonging to this category first perform MQE of the left and right view images separately and then perform BQF to combine the two monocular quality scores into a single one [28,52-57,60]. This framework can be described as follows:

$$Q = \mathbb{B}(\mathbb{M}(\mathbb{F}(I_L^O), \mathbb{F}(I_L^D)), \mathbb{M}(\mathbb{F}(I_R^O), \mathbb{F}(I_R^D))), \quad (2)$$

where $\mathbb{F}$, $\mathbb{M}$, and $\mathbb{B}$ denote the feature extractor, the MQE operator, and the BQF operator, respectively. The existing methods belonging to this category are different with each other in terms of either all or part of these three operators. Overall, the design of feature extractor $\mathbb{F}$ should be quality-aware and characterizable of the quality perception mechanisms of the HVS, the design of the MQE metric $\mathbb{M}$ should be able to well reflect the perceptual quality difference between the distorted and original inputs in the considered feature domain, and the design of the BQF metric $\mathbb{B}$ should well characterize the binocular visual properties of the HVS.

For the sake of efficiency, our proposed method follows the later combination-based pipeline. To be specific, we make use of the pre-trained deep CNN as the feature extractor to extract deep hierarchical features which are used as the basis for the subsequent MQE and BQF operations. The technical details will be illustrated in the following section.

III. Hierarchical Deep Feature-Based SIQM

A. Overview

The framework of our proposed method is shown in Fig. 1. Firstly, distorted and reference stereoscopic images are separately fed into a pre-trained deep CNN for hierarchical deep feature extraction. Note that, a set of filters can be used in each layer. Usually, one filter corresponds to one specific channel. Thus, a set of channel-wise feature maps are generated in each layer. For simplicity, we first aggregate these channel-wise feature maps into a single layer-wise feature map

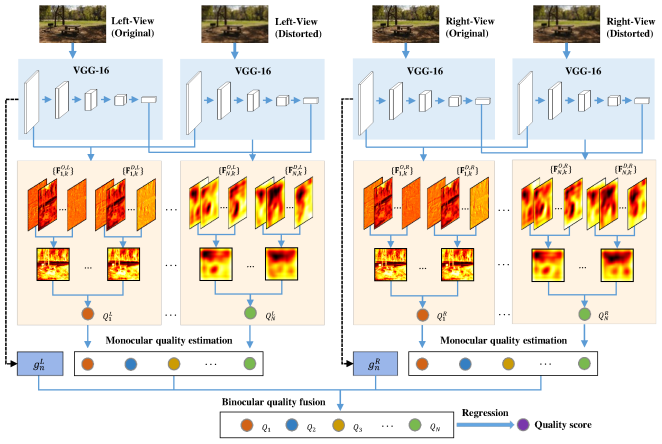


Fig. 1. The framework of the proposed hierarchical deep feature-based FR-SIQM method.

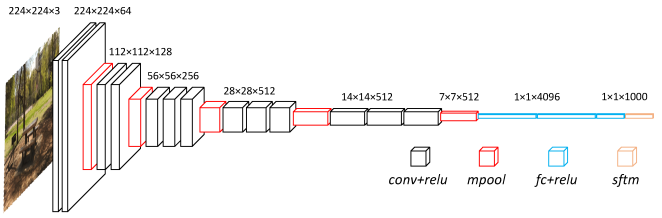


Fig. 2. The VGG-16 network architecture. It contains a total number of 16 convolutional (Conv) and fully connected (FC) layers. Some of the Conv layers are followed by max-pooling layers (Mpool). All the Conv and FC layers are equipped with the rectification (ReLU) non-linearity. For conciseness, we directly merge the ReLU layers with the Conv and FC layers together in the figure. The final layer is the soft-max layer (Sftm).

with average pooling. Then, these layer wise feature maps are used for the subsequent MQE and BQF stages. To be specific, the MQE stage estimates a set of layer-wise monocular quality scores with a similarity metric performed on the layer-wise feature maps of the distorted monocular view and those of the reference monocular view. The BQF stage estimates a set of layer-wise binocular quality scores via a weighted average of the corresponding layer-wise monocular quality scores. The weights are determined based on a hierarchical feature energy-based Gain-Control model. Finally, the layer-wise binocular quality scores across all layers are fused into an overall quality socre via regression.

B. Hierarchical Deep Feature Representation

1) Feature extractor: We use the VGG network [41] as the feature extractor. While there are now CNN models that outperform VGG, VGG remains appealing due to its relatively simple architecture and competitive performance [42]. Among several VGG network variants, the widely used VGG-16 is adopted. The architecture of VGG-16 is depicted in Fig. 2. In VGG-16, an input image is first resized into 224×224 and then passed through a stack of convolution layers (*Conv*), where the filters are with a constant size of 3×3. The number of filters in *Conv* layers starts from 64 to 512. The max-pooling layers (*Mpool*) are attached to some of the *Conv* layers. Max-pooling operation is performed over a 2×2 pixel window with a stride of 2. Finally, the output of the stacked *Conv* and *Mpool* layers

TABLE I
THE DETAILED CONFIGURATION INFORMATION REGARDING THE VGG-16 NETWORK.

Layer n	1	2	3	4	5	6	7	8
Type	Conv	ReLU	Conv	ReLU	Mpool	Conv	ReLU	Conv
Filter size	3×3	1×1	3×3	1×1	2×2	3×3	1×1	3×3
Channel number	64	64	64	64	64	128	128	128
Input size	224×224×3	224×224×64	224×224×64	224×224×64	224×224×64	112×112×64	112×112×128	112×112×128
Output size	224×224×64	224×224×64	224×224×64	224×224×64	112×112×64	112×112×128	112×112×128	112×112×128
Layer n	9	10	11	12	13	14	15	16
Type	ReLU	Mpool	Conv	ReLU	Conv	ReLU	Conv	ReLU
Filter size	1×1	2×2	3×3	1×1	3×3	1×1	3×3	1×1
Channel number	128	128	256	256	256	256	256	256
Input size	112×112×128	112×112×128	56×56×128	56×56×256	56×56×256	56×56×256	56×56×256	56×56×256
Output size	112×112×128	56×56×128	56×56×256	56×56×256	56×56×256	56×56×256	56×56×256	56×56×256
Layer n	17	18	19	20	21	22	23	24
Type	Mpool	Conv	ReLU	Conv	ReLU	Conv	ReLU	Mpool
Filter size	2×2	3×3	1×1	3×3	1×1	3×3	1×1	2×2
Channel number	256	512	512	512	512	512	512	512
Input size	56×56×256	28×28×256	28×28×512	28×28×512	28×28×512	28×28×512	28×28×512	28×28×512
Output size	28×28×256	28×28×512	28×28×512	28×28×512	28×28×512	28×28×512	28×28×512	14×14×512
Layer n	25	26	27	28	29	30	31	32
Type	Conv	ReLU	Conv	ReLU	Conv	ReLU	Mpool	FC
Filter size	3×3	1×1	3×3	1×1	3×3	1×1	2×2	1×1
Channel number	512	512	512	512	512	512	512	4096
Input size	14×14×512	14×14×512	14×14×512	14×14×512	14×14×512	14×14×512	14×14×512	7×7×512
Output size	14×14×512	14×14×512	14×14×512	14×14×512	14×14×512	14×14×512	7×7×512	1×1×4096
Layer n	33	34	35	36	37			
Type	ReLU	FC	ReLU	FC	Sftm			
Filter size	1×1	1×1	1×1	1×1	1×1			
Channel number	4096	4096	4096	1000	1000			
Input size	1×1×4096	1×1×4096	1×1×4096	1×1×4096	1×1×1000			
Output size	1×1×4096	1×1×4096	1×1×4096	1×1×1000	1×1×1000			

is followed by three fully connected layers (*FC*): the first two *FC* layers have 4096 channels each, the third one contains 1000 channels. The final layer is the soft-max layer (*Sftm*) for classification. In addition, all *Conv* and *FC* layers are equipped with the Rectified Linear Unit layers (*ReLU*). We summarize the detailed configuration information in Table I to present the input and output of each layer and other key information. In the following, we briefly introduce the formulations of these layers.

`Conv layer`: A *Conv* layer is defined on a translation invariance basis and shared weights across different spatial locations. The input and output of each *Conv* layer are three-dimensional tensors, called feature maps. If we denote the input feature maps of a *Conv* layer n as $\mathbf{F}_{n-1}$, where $n \in \{1, 2, \cdots, N\}$, the output feature maps of this *Conv* layer can be computed by:

$$\mathbf{F}_n^{Conv} = \mathbf{F}_{n-1} * \mathbf{W}_n^{Conv} + \mathbf{b}_n^{Conv}, \tag{3}$$

where the symbol $*$ denotes the 2-D convolution operation, $\mathbf{W}_n^{Conv}$ is the set of filter weights in this *Conv* layer, and $\mathbf{b}_n^{Conv}$ is the set of bias values added to the corresponding convolutional responses.

`Mpool layer`: Next, a *Mpool* layer is followed to aggregate the features over local non-overlapping windows at each location per feature map. If we denote the local pooling window at location p as Ω_p and the input feature maps of a *Mpool* layer n as $\mathbf{F}_{n-1}$, where $n \in \{1, 2, \cdots, N\}$, the output feature maps of the *Mpool* layer n can be computed by:

$$\mathbf{F}_{n,p}^{MP} = \max_{q \in \Omega_p}(\mathbf{F}_{n-1,q}). \tag{4}$$

`FC layer`: When several *Conv* layers and *Mpool* layers are stacked alternately in depth, hierarchical features with increasing receptive fields can be obtained. Finally, the extracted feature maps are aggregated into a 1-D global feature vector via *FC* layer. If we denote the input features (or feature maps) of an FC layer n as $\mathbf{F}_{n-1}$, where $n \in \{1, 2, \cdots, N\}$, the output feature vector of the FC layer n can be computed by:

$$\mathbf{F}_n^{FC} = \mathbf{F}_{n-1} \cdot \mathbf{W}_n^{FC} + \mathbf{b}_n^{FC}. \tag{5}$$

`ReLU layer`: The ReLU layer is defined to restrict all the negative values of the input to zero while preserving the positive values. If we denote the input feature maps of a ReLU layer n as $\mathbf{F}_{n-1}$, where $n \in \{1, 2, \cdots, N\}$, the output feature maps of this *ReLU* layer can be computed by:

$$\mathbf{F}_n^{ReLU} = \max(0, \mathbf{F}_{n-1}). \tag{6}$$

2) Hierarchical feature representation: Feeding an image into the pre-trained VGG network on ImageNet, hierarchical feature maps with increasing receptive fields are generated. For the task of SIQM, each view of the distorted and original

TABLE II
IMPORTANT NOTATIONS AND DEFINITIONS.

Notations	Definitions
N	Total number of layers in the VGG network
$I^{D,L}$	The left view image of a stereoscopic image pair
$I^{D,R}$	The right view image of a stereoscopic image pair
$\mathbf{F}_{n,k}^{O,L}$	The k-th feature map in the n-th layer of the original left image
$\mathbf{F}_{n,k}^{D,L}$	The k-th feature map in the n-th layer of the distorted left image
$\mathbf{F}_{n,k}^{O,R}$	The k-th feature map in the n-th layer of the original right image
$\mathbf{F}_{n,k}^{D,R}$	The k-th feature map in the n-th layer of the distorted right image
$\mathbf{S}_n^L$	The similarity map in the n-th layer of the left-view image
$\mathbf{S}_n^R$	The similarity map in the n-th layer of the right-view image
Q_n^L	The quality score in the n-th layer of the left-view image
Q_n^R	The quality score in the n-th layer of the right-view image
g_n^L	The left-view weights in the n-th layer
g_n^R	The right-view weights in the n-th layer
Q_n	The fused binocular quality score in the n-th layer

stereopairs is first resized to 224×224×3 followed by mean value subtraction in each channel, and then fed into VGG separately to extract the corresponding hierarchical feature maps. The number of feature maps in each layer equals to the relevant number of filters/channels defined in the network. In our method, we only extract deep feature maps from the former 31 layers and discard the last six *FC+ReLU* and *Sftm* layers which actually produce global feature vectors instead of local feature maps. The default input image size in the standard VGG-16 network is 224×224×3 and VGG-16 can only receive such an input image size due to the existence of the FC layers. Although we have removed the FC layers from the standard VGG-16 network architecture for local feature extraction, it is more reasonable to still use the default input image size because the pre-trained VGG-16 on ImageNet is directly used without any re-training process in our implementation. That is, the parameters are optimized for the standard VGG-16 architecture and all these parameters are optimal only when using the default input image size.

Before describing the technical details, we first summarize some important notations and their definitions in Table II. Given a distorted stereoscopic image $\{I^{D,L}, I^{D,R}\}$ and its relevant original version $\{I^{O,L}, I^{O,R}\}$, we denote the k-th feature map in the n-th layer of the distorted left-view image $I^{D,L}$ as $\mathbf{F}_{n,k}^{D,L}$ with $k = \{1, 2, \cdots, K_n\}$ and $n = \{1, 2, \cdots, N\}$, respectively. Similarly, we have $\mathbf{F}_{n,k}^{D,R}$ for the distorted right-view image $I^{D,R}$, $\mathbf{F}_{n,k}^{O,L}$ and $\mathbf{F}_{n,k}^{O,R}$ for the relevant original left-view image $I^{O,L}$ and original right-view image $I^{O,R}$, respectively. To facilitate analysis, the feature maps in different layers are shown in Fig. 3. One can see that the feature maps in lower layer (*e.g.*, (c) and (f)) highlight the low-level visual features and image micro-structures while the feature map in higher layer (*e.g.*, (e) and (h)) tend to reflect the high-level semantic features and image macro-structures. Since distortion will have impact on both micro- and macro-structures of an image, using the extracted hierarchical deep feature representations for quality degradation evaluation of stereoscopic images is deemed reasonable.

C. Monocular Quality Estimation

Based on the built hierarchical deep feature representations, monocular quality estimation can be performed. Taking the left-view as an example, monocular quality estimation is performed by measuring the similarity between $\mathbf{F}_n^{D,L}$ and $\mathbf{F}_n^{O,L}$ which represent the averaged feature map of the distorted and original left view images in each layer, respectively. In the proposed method, instead of directly comparing the feature maps, we perform the similarity measurement in the corresponding gradient domain. The rationale is that the HVS is highly adapted for extracting structural information from scenes and therefore the similarity metric should be capable of reflecting the structural degradation between the distorted and original feature maps. Gradient magnitude, defined as the root mean square of image directional gradients along the horizontal and vertical directions, is known as a simple yet effective local structure descriptor of images. The similarity metric performed on the gradient magnitude maps of the

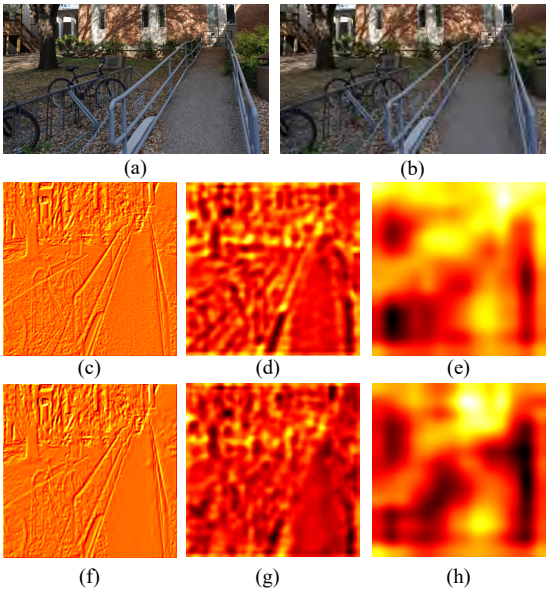


Fig. 3. The feature maps in different layers. (a) Original left-view image, (b) Distorted left-view image with JPEG2000 compression, (c)-(e) are respectively the 1-st channel feature maps in the 1-st, 15-th, and 30-th layers of (a), (f)-(h) are respectively the 1-st channel feature maps in the 1-st, 15-th, and 30-th layers of (b). Note that, all the feature maps are resized to 224×224 for visualization.

hierarchical feature maps is considered to be more in line with human perception. In addition, previous works on 2D-IQM have also verified the efficiency and effectiveness of gradient magnitude.

Mathematically, the gradient magnitudes of $\mathbf{F}_n^{D,L}$ and $\mathbf{F}_n^{O,L}$ at location i, denoted by $\mathbf{GM}_n^{D,L}(i)$ and $\mathbf{GM}_n^{O,L}(i)$, are calculated as follows:

$$\mathbf{GM}_n^{D,L}(i) = \sqrt{(\mathbf{F}_n^{D,L} * \mathbf{G}_h)^2(i) + (\mathbf{F}_n^{D,L} * \mathbf{G}_v)^2(i)}, \quad (7)$$

$$\mathbf{GM}_n^{O,L}(i) = \sqrt{(\mathbf{F}_n^{O,L} * \mathbf{G}_h)^2(i) + (\mathbf{F}_n^{O,L} * \mathbf{G}_v)^2(i)}, \quad (8)$$

where the symbol $*$ denotes the 2-D convolution operation, $\mathbf{G}_h$ and $\mathbf{G}_v$ denotes the Prewitt filters along horizontal and vertical directions, respectively, and are defined as follows:

$$\mathbf{G}_h = \begin{bmatrix} 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \end{bmatrix}, \quad (9)$$

$$\mathbf{G}_v = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 \\ -1/3 & -1/3 & -1/3 \end{bmatrix}. \quad (10)$$

Then, we measure the similarity between $\mathbf{GM}_n^{D,L}$ and $\mathbf{GM}_n^{O,L}$ at location i as follows:

$$\mathbf{S}_n^L(i) = \frac{2\mathbf{GM}_n^{D,L}(i)\mathbf{GM}_n^{O,L}(i) + c}{[\mathbf{GM}_n^{D,L}(i)]^2 + [\mathbf{GM}_n^{O,L}(i)]^2 + c}, \quad (11)$$

where c is a small positive constant that prevents the occurrence of numerical instability. In the proposed method, we set c=0.01 empirically.

With the obtained gradient magnitude similarity map $\mathbf{S}_n^L$ in hand and with the inspiration that the global variation of image

local quality degradation can reflect the overall quality, the final monocular quality score for the left-view is computed via a simple standard deviation-based pooling scheme as follows:

$$Q_n^L = \sqrt{\frac{1}{H \times W} \sum_{i=1}^{H \times W} (\mathbf{S}_n^L(i) - \overline{S_n^L})^2}, \tag{12}$$

where H and W are the height and width of the image, $\overline{S_n^L}$ denotes the mean value of $\mathbf{S}_n^L$. Similarly, we can compute the final monocular quality score for the right-view, denoted by Q_n^R. Since the above described monocular quality estimation is carried out in a layer-wise manner, the estimated Q_n^L and Q_n^R are called layer-wise monocular quality scores in this paper.

D. Binocular Quality Fusion

Based on the obtained layer-wise monocular quality scores $Q_{n,k}^L$ and $Q_{n,k}^R$, binocular quality fusion targets at combining the corresponding layer-wise monocular quality scores into layer-wise binocular quality scores in a manner that is consistent with the binocular combination mechanisms of the HVS. In the literature, there have been several biological models, such as eye-weighting model [43], quadratic summation model [44], vector summation model [45], neural network model [46], and gain-control theory model [47], proposed to simulate the critical binocular combination behavior. Among them, the latest gain-control theory model simulates an early stage of binocular combination and explains both Fechner's paradox [48] and cyclopean perception well [47]. In general, the gain-control theory model is expressed as follows:

$$f_B(I^L, I^R) = \left(\frac{1+E^L}{1+E^L+E^R}\right) I^L + \left(\frac{1+E^R}{1+E^L+E^R}\right) I^R, \tag{13}$$

where E^L and E^R denote the cumulative sums of energy over layers of the left-view and right-view images, respectively.

As described, the distorted left-view image $I^{D,L}$ and right-view image $I^{D,R}$ can respectively be represented by the VGG model as

$$\mathcal{F}(I^{D,L}) = \{\mathbf{F}_n^{D,L}\} = \{\mathbf{F}_1^{D,L}, \mathbf{F}_2^{D,L}, \cdots, \mathbf{F}_N^{D,L}\}, \tag{14}$$

$$\mathcal{F}(I^{D,R}) = \{\mathbf{F}_n^{D,R}\} = \{\mathbf{F}_1^{D,R}, \mathbf{F}_2^{D,R}, \cdots, \mathbf{F}_N^{D,R}\}, \tag{15}$$

where n is the index of layer.

Based on the above hierarchical feature representations, we further compute the energies of a certain layer as follows:

$$E(\mathbf{F}_n^{D,L}) = \sum_p [\mathbf{F}_n^{D,L}(p)]^2, \tag{16}$$

$$E(\mathbf{F}_n^{D,R}) = \sum_p [\mathbf{F}_n^{D,R}(p)]^2, \tag{17}$$

where p denotes the pixel intensity value. Besides, the total energies over all layers are computed as follows:

$$E(\mathbf{F}^{D,L}) = \sum_{n=1}^{N} E(\mathbf{F}_n^{D,L}), \tag{18}$$

$$E(\mathbf{F}^{D,R}) = \sum_{n=1}^{N} E(\mathbf{F}_n^{D,R}), \tag{19}$$

where N is the number of layers.

Finally, we derive the gain g_n for each layer based on the Gain-Control theory model expressed in Eq. (13):

$$g_n^L = \frac{1+E(\mathbf{F}_n^{D,L})}{1+E(\mathbf{F}^{D,L})+E(\mathbf{F}^{D,R})}, \tag{20}$$

$$g_n^R = \frac{1+E(\mathbf{F}_n^{D,R})}{1+E(\mathbf{F}^{D,L})+E(\mathbf{F}^{D,R})}. \tag{21}$$

In the above equations, the rationality of using total energies over all layers (instead of a certain layer) in the denominator is illustrated as follows. In the original gain control model, the gain value of the left view is defined as the ratio between $(1+E^L)$ and $(1+E^L+E^R)$ where E^L and E^R are the sums of energy over all spatial frequency band components (can be obtained by Difference-of-Gaussian decomposition) of the left and the right views, respectively. If we consider each specific layer in our method as multiple spatial frequency bands, an intuitive consideration is that the gain value of a certain layer should not only accounts for the dominance of each view relative to the other one in this layer but also accounts for the importance of this layer among all layers.

Finally, according to the Gain-Control theory model given in Eq. (13), the layer-wise binocular quality score can be derived as follows:

$$Q_n = g_n^L Q_n^L + g_n^R Q_n^R, \tag{22}$$

where Q_n^L and Q_n^R denote the layer-wise monocular quality scores of the left and right views, respectively.

E. Hierarchical Layer Fusion

With the layer-wise binocular quality scores in hand, it is necessary to fuse them into a single quality scalar. In our proposed method, we resort to support vector regression (SVR) [49] to learn the mapping from layer-wise binocular quality scores to final binocular quality scores. Considering a set of training data $\{(\mathbf{Q}_1, S_1), \cdots, (\mathbf{Q}_l, S_l)\}$, where $\mathbf{Q}_i \in \mathbb{R}^N$ is the estimated layer-wise binocular quality scores and S_i is the corresponding subjective opinion quality score (*e.g.*, DMOS). Given parameters $C > 0$ and $\epsilon > 0$, the standard form of SVR is expressed as [49]:

$$\min_{\boldsymbol{\omega}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \ \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + C\left\{\sum_{i=1}^{l}\xi_i + \sum_{i=1}^{l}{\xi_i}^*\right\} \tag{23}$$

$$\text{subject to } \ \boldsymbol{\omega}^T\phi(\mathbf{Q}_i) + b - S_i \le \epsilon + \xi_i \tag{24}$$

$$S_i - \boldsymbol{\omega}^T\phi(\mathbf{Q}_i) - b \le \epsilon + {\xi_i}^* \tag{25}$$

$$\xi_i, {\xi_i}^* \ge 0, \ i = 1, 2, \cdots, l \tag{26}$$

where $\mathcal{K}(\mathbf{Q}_i, \mathbf{Q}_j) = \phi(\mathbf{Q}_i)^T\phi(\mathbf{Q}_j)$ is the kernel function. We resort to the LIBSVM package [50] to implement SVR with the radial basis function kernel. Once the SVR model SVR_TRAIN is built, it can be used for quality prediction of an arbitrary distorted stereoscopic image with its corresponding layer-wise binocular quality scores as input: $Q = \text{SVR\_TRAIN}(\mathbf{Q}_i)$.

TABLE III
KEY INFORMATION OF THE USED BENCHMARK DATABASES.

Database	LIVE-I [28]	LIVE-II [51]	WIVC-II [52]
# of Reference Image	20	8	10
# of Distortion Type	5	5	3
# of Distortion Level	4	9	4
# of Distorted Image	365	360	460
Distortion Pattern	Sym.	Sym./Asym.	Sym./Asym.
Subjective Score Type	DMOS	DMOS	DMOS

IV. EXPERIMENTAL RESULTS

A. Experimental Setups

1) Benchmark database: Three benchmark databases including Laboratory of Image and Video Evaluation (LIVE) 3D-IQA Phase-I database [28], LIVE 3D-IQA Phase-II database [51], and Waterloo IVC 3D-IQM Phase-II database [52] are used in the experiments.

The LIVE 3D-IQA Phase-I database consists of 20 reference and 365 distorted stereoscopic images with human subjective rating scores in the form of difference mean opinion score (DMOS). Five types of distortions, including additive Gaussian White Noise (WN), Gaussian Blur (GBLUR), JPEG Compression (JPEG), JPEG2000 Compression (JP2K), and Fast Fading (FF) Channel Distortion, are simulated. Meanwhile, all distortions are symmetrically applied. The LIVE 3D-IQA Phase-II database consists of 8 reference and 360 distorted stereoscopic images with their DMOSs. The same five distortion types as in LIVE Phase-I database are considered. For each distortion type, each reference stereoscopic image was degraded to generate 3 symmetrically distorted and 6 asymmetrically distorted stereoscopic images. The Waterloo IVC 3D-IQA Phase-II database consists of 10 reference and 460 distorted stereoscopic images with DMOSs. Each single-view reference image was either symmetrically or asymmetrically processed by three types of distortions, i.e., additive white Gaussian noise contamination (WN), Gaussian blur (GBLUR), and JPEG compression (JPEG). Each distortion type has four distortion levels. Some key information regarding these databases are summarized in Table II.

2) Performance criteria: Four performance criteria including the Spearman rank correlation coefficient (SRCC), the Pearson linear correlation coefficient (PLCC), the Kendall rank correlation coefficient (KRCC), and the Root Mean Square Error (RMSE) between the predicted scores and DMOSs, are used. A better SIQM approach should have higher SRCC, PLCC, and KRCC values (with a maximum of 1) while lower RMSE values (with a minimum of 0). As a common strategy, before calculating the PLCC and RMSE, the following logistic regression function suggested in [53] is implemented:

$$Q_p = \lambda_1 \left[\frac{1}{2} - \frac{1}{1 + \exp(\lambda_2 (Q - \lambda_3))} \right] + \lambda_4 Q + \lambda_5, \quad (27)$$

where Q is the model predicted score as input and Q_p is the output, and λ_1, λ_2, λ_3, λ_4, and λ_5 are the parameters to be fitted. A set of model predicted scores and the corresponding DMOSs are used to determine these parameters by using the *nlinfit* function in MATLAB.

TABLE IV
OVERALL PERFORMANCE COMPARISON ON LIVE 3D PHASE-I AND PHASE-II DATABASES. RESULTS OF THE BEST-PERFORMING SIQM METHOD ARE BOLDED.

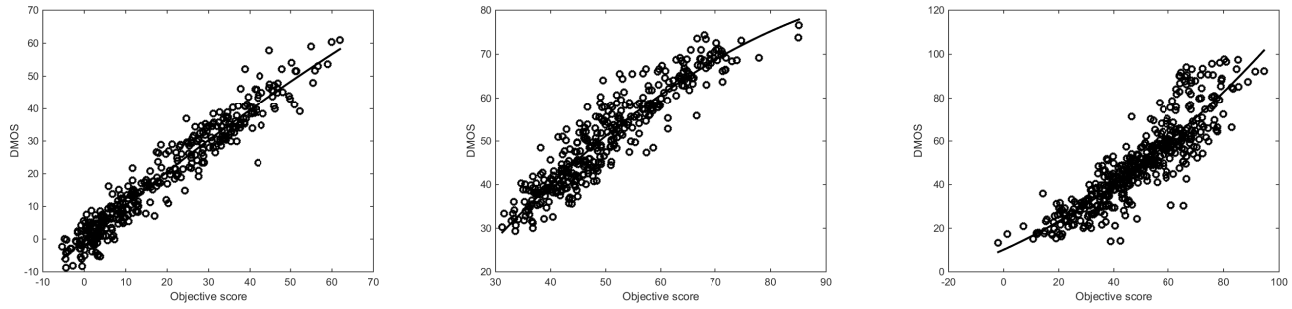
Methods	LIVE 3D Phase-I			LIVE 3D Phase-II		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
SSIM [7]	0.872	0.876	8.033	0.801	0.792	6.757
MS-SSIM [8]	0.926	0.922	6.193	0.778	0.772	7.096
VIF [9]	0.925	0.920	6.228	0.840	0.817	6.132
MAD [10]	0.942	0.939	5.498	0.854	0.842	5.869
DeepSIM [40]	0.945	0.940	5.476	0.859	0.848	5.694
Benoit's [14]	0.915	0.911	6.633	0.812	0.806	6.582
You's [15]	0.895	0.896	7.312	0.729	0.681	7.727
Wang's [54]	0.888	0.890	7.536	0.817	0.805	6.502
Ko's [55]	0.910	0.907	6.804	0.760	0.756	7.341
Bensalma's [56]	0.887	0.875	7.559	0.770	0.751	7.204
Lin&Wu's [57]	0.864	0.856	8.242	0.658	0.638	8.496
Chen's [58]	0.917	0.916	6.550	0.906	0.901	4.767
Shao's [29]	0.932	0.927	5.941	0.836	0.819	6.196
Zhang's [30]	0.951	0.944	5.052	0.927	0.924	4.220
Lin's [59]	0.937	0.931	5.744	0.911	0.894	4.648
Khan's [60]	0.927	0.916	-	**0.932**	0.922	-
Ma's [61]	0.951	0.949	5.074	0.911	0.905	4.652
Zhou's [62]	0.939	0.926	5.535	0.912	0.896	4.667
Proposed	**0.960**	**0.953**	**4.455**	**0.932**	**0.927**	**4.041**

3) Evaluation protocol: Since the proposed method requires learning a regression model using SVR to estimate the final quality score, it is required to construct a training set for SVR model training. For performance evaluation on each database, the entire database is divided into two non-overlapping subsets, i.e., training subset and testing subset, according to the content of original stereoscopic images. To be specific, for each individual database, distorted stereoscopic images associated with 80% original contents constitute the training subset while the remaining distorted stereoscopic images associated with 20% original contents are considered as the testing subset. To ensure the results are not biased to specific train-test splits, such random split process is randomly repeated 100 times to calculate the median prediction results as the final performance.

B. Performance Evaluation

1) Evaluation on LIVE 3D Phase-I and Phase-II databases: The performance evaluation is first conducted on LIVE 3D Phase-I and Phase-II databases. We compare the performance of the proposed method against that of several mainstream FR 2D-IQM and FR-SIQM methods. The competing FR 2D-IQM methods include SSIM [7], MS-SSIM [8], VIF [9], MAD [10], and DeepSIM [40]. For these FR 2D-IQM approaches, the predicted quality of a certain stereopair is directly taken to be the averaged quality estimated from the both monocular views without considering any binocular fusion mechanism. The competing FR-SIQM methods include Benoit's method [14], You's method [15], Wang's method [54], Ko's method [55], Bensalma's method [56], Lin & Wu's method [57], Chen's method [58], Shao's method [29], Zhang's method [30], Lin's method [59], Khan's method [60], Ma's method [61], and Zhou's method [62]. We present the SRCC, PLCC and RMSE results of these methods on the entire LIVE 3D Phase-I and Phase-II databases in Table IV where the indicators providing the best performances are bolded.

From this table, we have the following observations. First, the proposed method performs better than all the competing



(a) LIVE 3D Phase-I (b) LIVE 3D Phase-II (c) WIVC 3D Phase-II

Fig. 4. The scatter plots of the proposed method on LIVE 3D Phase-I, LIVE 3D Phase-II, and WIVC 3D Phase-II databases, respectively.

TABLE V
PLCC VALUES ON DIFFERENT TYPES OF DISTORTIONS ON THE LIVE 3D PHASE-I AND PHASE-II DATABASES. RESULTS OF THE BEST-PERFORMING SIQM METHODS ON EACH DISTORTION TYPE ARE BOLDED.

Methods	LIVE 3D Phase-I					LIVE 3D Phase-II				
	JP2K	JPEG	WN	GBLUR	FF	JP2K	JPEG	WN	GBLUR	FF
SSIM [7]	0.859	0.480	0.893	0.911	0.689	0.697	0.645	0.925	0.828	0.856
MS-SSIM [8]	0.916	0.622	0.929	0.936	0.803	0.833	0.814	0.941	0.688	0.841
VIF [9]	0.936	0.633	0.930	0.941	0.858	0.808	0.782	0.821	0.983	0.940
MAD [10]	0.951	0.762	0.935	0.961	0.842	0.873	0.858	0.887	0.939	0.920
DeepSIM [40]	0.950	0.774	0.935	0.952	0.860	0.881	0.863	0.892	0.938	0.937
Benoit's [14]	0.914	0.607	0.907	0.924	0.782	0.853	0.821	0.933	0.747	0.867
You's [15]	0.921	0.602	0.908	0.899	0.730	0.776	0.776	0.877	0.576	0.788
Wang's [54]	0.870	0.482	0.875	0.934	0.764	0.731	0.662	0.938	0.878	0.879
Ko's [55]	0.910	0.529	0.885	0.953	0.828	0.904	0.707	0.899	0.714	0.824
Bensalma's [56]	0.839	0.380	0.915	0.937	0.734	0.667	0.858	0.944	0.908	0.910
Lin&Wu's [57]	0.865	0.477	0.937	0.918	0.720	0.722	0.642	0.927	0.842	0.856
Chen's [58]	0.912	0.590	0.926	0.935	0.747	0.839	0.831	0.957	0.962	0.907
Shao's [29]	0.921	0.629	0.929	0.942	0.843	0.793	0.759	0.811	0.957	0.936
Zhang's [30]	**0.955**	0.732	0.952	0.959	0.881	0.886	0.889	0.957	**0.984**	0.932
Lin's [59]	0.952	0.755	0.927	0.958	0.862	-	-	-	-	-
Khan's [60]	0.951	0.711	0.947	0.959	0.858	0.927	0.893	0.970	0.978	0.899
Ma's [61]	0.950	0.768	0.952	**0.963**	0.872	0.887	0.899	0.957	0.978	0.901
Proposed	0.953	**0.799**	**0.963**	0.952	**0.887**	**0.936**	**0.900**	**0.971**	0.983	**0.950**

TABLE VI
SRCC VALUES ON DIFFERENT TYPES OF DISTORTIONS ON THE LIVE 3D PHASE-I AND PHASE-II DATABASES. RESULTS OF THE BEST-PERFORMING SIQM METHODS ON EACH DISTORTION TYPE ARE BOLDED.

Methods	LIVE 3D Phase-I					LIVE 3D Phase-II				
	JP2K	JPEG	WN	GBLUR	FF	JP2K	JPEG	WN	GBLUR	FF
SSIM [7]	0.857	0.436	0.938	0.879	0.586	0.703	0.679	0.920	0.836	0.835
MS-SSIM [8]	0.898	0.599	0.942	0.928	0.735	0.817	0.827	0.947	0.801	0.830
VIF [9]	0.902	0.582	0.932	0.931	0.804	0.826	0.778	0.820	0.950	0.934
MAD [10]	0.925	0.736	0.950	0.954	0.772	0.869	0.839	0.885	0.924	0.918
DeepSIM [40]	0.923	0.747	0.948	0.945	0.785	0.873	0.841	0.891	0.922	0.920
Benoit's [14]	0.887	0.565	0.939	0.911	0.683	0.842	0.839	0.926	0.766	0.862
You's [15]	0.884	0.547	0.929	0.910	0.629	0.834	0.755	0.878	0.275	0.740
Wang's [54]	0.870	0.445	0.939	0.918	0.654	0.727	0.694	0.934	0.882	0.865
Ko's [55]	0.891	0.527	0.933	0.941	0.756	0.902	0.728	0.900	0.836	0.811
Bensalma's [56]	0.817	0.328	0.906	0.916	0.650	0.804	0.846	0.939	0.884	0.874
Lin&Wu's [57]	0.839	0.207	0.928	0.935	0.658	0.719	0.613	0.907	0.711	0.701
Chen's [58]	0.896	0.558	0.948	0.926	0.688	0.833	0.840	0.955	0.910	0.889
Shao's [29]	0.883	0.599	0.930	0.910	0.793	0.788	0.745	0.807	0.939	0.935
Zhang's [30]	0.916	0.700	0.950	0.942	0.833	0.895	0.866	0.952	0.942	0.922
Lin's [59]	0.913	0.716	0.929	0.933	0.829	-	-	-	-	-
Khan's [60]	0.907	0.606	0.938	0.930	0.809	0.913	0.867	**0.958**	0.885	0.865
Ma's [61]	**0.924**	**0.736**	**0.952**	0.945	0.826	0.878	**0.879**	0.949	0.906	0.893
Zhou's [62]	0.906	0.693	0.941	0.917	0.744	0.872	0.849	0.946	0.911	0.905
Proposed	0.895	0.702	0.939	**0.948**	**0.839**	**0.939**	0.865	0.943	**0.946**	**0.936**

metrics on both databases by a large margin except for the PLCC value of Khans method on LIVE 3D Phase-II is on a par with our method. Second, although the SIQM methods more or less take into account either the disparity information or binocular visual properties, they are not always better than the 2D-extended IQM methods. The reasons are explained

as follows. For Benoit's [14] and You's [15] methods, their performances heavily depend on the accuracy of the used stereo matching algorithms for disparity estimation. Meanwhile, directly using 2D-IQM metrics to evaluate disparity maps may not be necessarily perceptually interpretable. Although Shao's [29], Wang's [54], Ko's [55], and Lin & Wu's [57] methods are designed by taking into account binocular combination, their methods are still based on linear weighting two quality scores from simple FR 2D-IQM methods on both monocular views. Bensalma's [56] method performs the worst among all the FR-SIQM methods on LIVE Phase-I and also not so well on LIVE Phase-I. This indicates that the accurate calculation of binocular energy still has a long way to go for achieving a satisfactory performance in the application of FR-SIQM. Zhang's [30], Chen's [58], Lin's [59], Ma's [61], Zhou's [62] methods measure the quality of stereoscopic images based on the synthesized cyclopean image by modeling the binocular rivalry mechanism. Therefore, these methods are more consistent with stereopsis and perform much better on LIVE 3D Phase-II which contains both symmetric and asymmetric distortions. This is expectable because stereopairs with asymmetrical distortions are more likely to cause binocular rivalry which is well characterized by the synthesized cyclopean image. Khan and Channappayya's method [60] estimates the quality of stereopairs by extracting depth-salient edges to refine the quality maps associated with the gradient features of both monocular images. This method achieves fairly good performance due to the consideration of salient edges (middle-level features) that contribute to depth perception. However, only low-level and middle-level features, i.e., inter-gradient and saliency maps, are taken into account for monocular quality estimation, ignoring the contribution of high-level semantic features. Moreover, a simple multiplication operator used for binocular fusion cannot well characterize the complex binocular interaction mechanism.

We further draw the scatter plots of the proposed method on each database by using the standard leave-one-out evaluation strategy, as shown in Fig. 4(a), (b). The vertical axis denotes the DMOSs and the horizontal axis denotes the predicted scores. A better convergence of the points to the fitted curve in the scatter plots means a better consistency with the DMOSs. As one can see, our method can achieve high consistency with human subjective perception.

In spite of the prominent performance on the entire database, it is also necessary to know the capacity of the proposed method for evaluating each individual distortion type. Therefore, we also report the performance results on subsets of the LIVE 3D Phase-I and Phase-II databases corresponding to each individual distortion type. Test results are shown in Table V and Table VI in terms of PLCC and SRCC, respectively. Results of the best-performing SIQM methods on each distortion type are bolded in the table. Since RMSE generally have the opposite tendency with PLCC and SRCC, we do not report them here for brevity. Note that the PLCC values of Zhou's method [62] on each individual distortion type (Table V) are not reported because these results are not reported in the corresponding paper. As shown in Table V and Table VI, we observe that our proposed method delivers the best results on the majority of all distortion types among all competitors, *i.e.*, it delivers the best results for 12 times, followed by Ma's method [61] for only 5 times. In particular, our method can evaluate the GBLUR and FF distortion type quite well (refer to the SRCC values shown in Table VI) because these two types of distortions usually destroy the semantic information of image contents. Fortunately, our proposed method can well characterize both low-level visual feature and high-level semantic information degradations by applying a hierarchical deep feature degradation fusion strategy for SIQM.

2) Evaluation on WIVC Phase-II databases: We also evaluate the proposed method on the WIVC 3D Phase-II database. The competing methods include two classical FR 2D-IQM metrics, *i.e.*, PSNR and SSIM [7], and several mainstream FR-SIQM metrics, *i.e.*, Benoit's method [14], You's method [15], Yang's method [26], Chen's method [58], and Wang's method [52]. Similar to the evaluation protocol on LIVE, for the FR 2D-IQM approaches, the predicted quality of a certain stereopair is directly taken to be the average value of the quality scores estimated from the two views. The PLCC, SRCC, and KRCC results of the proposed method and that of other competing methods on the entire WIVC 3D Phase-II database are presented in Table VII where the indicators providing the best performances are highlighted with boldface. It can be seen that a simple average of FR 2D-IQM scores of both monocular views cannot accurately predict the quality of stereoscopic images contained in WIVC 3D Phase-II which is constructed mainly for SIQM with asymmetric distortions (although both symmetric and asymmetric distortions are involved, asymmetric distortions occupies the great majority). While those competing FR-SIQM methods are somewhat better than the extended FR 2D-IQM methods, most of them are still far from satisfactory. Among the competing FR-SIQM methods, the latest method proposed by Wang *et al*. [52] delivers the best performance. However, our proposed method is still superior to this state-of-the-art method in terms of all the performance criteria. Finally, we show the scatter plot of our proposed method on the entire WIVC 3D Phase-II database in Fig. 4(c). All these results, together with the results shown in Table IV, Table V, and Table VI, verify the outstanding performance of our method for evaluating both symmetrically and asymmetrically distorted stereopairs.

C. Model Ablation Study

Since our proposed method addresses the SIQM problem by fusing the degradations on hierarchical deep features, it is necessary to validate the reasonability of integrating the quality estimators across different layers in deep neural network. Meanwhile, in order to fuse the layer-wise monocular quality scores into a single final binocular quality score, BQF is first performed to obtain layer-wise binocular quality scores which are then regressed into a single binocular quality score as the final prediction in our method. In this section, we conduct a model ablation study on the LIVE 3D Phase-I and Phase-II databases to investigate the contribution of BQF and hierarchical layer fusion, respectively. To demonstrate the importance of BQF, we implement a model without the

TABLE VII
PERFORMANCE COMPARISON ON THE WATERLOO IVC 3D PHASE-II DATABASE. RESULTS OF THE BEST-PERFORMING METHOD ARE BOLDED.

Methods	Symmetric Distortion			Asymmetric Distortion			All		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
PSNR	0.688	0.535	0.391	0.627	0.485	0.326	0.639	0.496	0.352
SSIM [7]	0.736	0.562	0.398	0.547	0.452	0.318	0.550	0.468	0.332
Benoit's [14]	0.755	0.571	0.401	0.555	0.454	0.317	0.551	0.460	0.321
You's [15]	0.763	0.560	0.401	0.686	0.600	0.423	0.682	0.587	0.418
Yang's [26]	0.792	0.663	0.485	0.641	0.595	0.415	0.639	0.588	0.414
Chen's [58]	0.837	0.758	0.564	0.633	0.563	0.406	0.613	0.578	0.417
Wang's [52]	0.938	0.905	0.732	0.880	0.848	0.665	0.892	0.869	0.690
Proposed	**0.944**	**0.906**	**0.752**	**0.884**	**0.853**	**0.677**	**0.911**	**0.905**	**0.741**

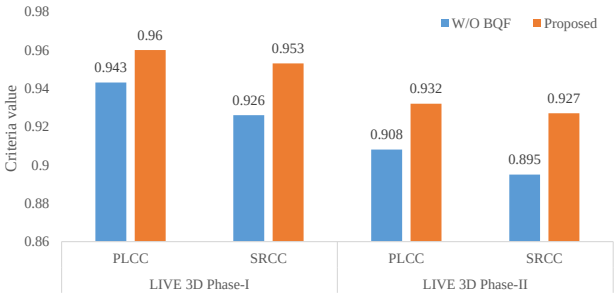


Fig. 5. Performance comparison of the models with/without BQF module.

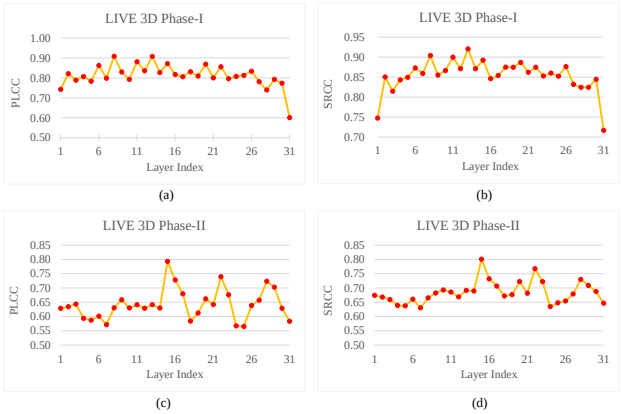


Fig. 6. Performance results of the binocular quality score in each single layer. (a) PLCC on LIVE 3D Phase-I, (b) SRCC on LIVE 3D Phase-I, (c) PLCC on LIVE 3D Phase-II, (d) SRCC on LIVE 3D Phase-II.

BQF operation for comparison. To be specific, the layer-wise monocular quality scores are directly averaged to obtain the layer-wise binocular quality scores which are also regressed using SVR. We denote this comparison model as W/O BQF hereinafter. The performance comparison between W/O BQF and the proposed method is shown in Fig. 5. It is obvious that the proposed method outperforms the W/O BQF model by a large margin especially on the LIVE 3D Phase-II database which contains both symmetrically and asymmetrically distorted stereopairs. Such observation is consistent with a previous conclusion that an adaptive binocular combination scheme is particularly important for evaluating asymmetrically distorted stereopairs. To verify the importance of hierarchical layer fusion, we compute the performance results of the binocular quality score in each single layer. The results are depicted in Fig. 6. One can see that the performance results associated with each individual layer-wise binocular quality scores are all worse than the proposed method. This demonstrates the reasonability and effectiveness of fusing the quality degradations on hierarchical deep features for solving the problem of SIQM. Through such model ablation studies, an important conclusion can be drawn that both MQE and BQF should be performed in a hierarchical manner in FR-SIQM. This can be justified by the fact that the structure of human brain is inherently hierarchical so that our designed SIQM quality metric should well resemble this property.

Additionally, we also compared VGG-16 [41] with other three networks including AlexNet [36], GoogleNet [63], and ResNet-50 [64] to demonstrate the effectiveness of VGG-16 as the initial feature extractor. Similarly, the compared three networks are also pre-trained on the ImageNet dataset. The MQE and BQF strategies remain unchanged. Fig. 7 shows the performance values of different models. We can observe that the proposed method using VGG-16, although almost the same with the one using ResNet-50, performances better than the ones using AlexNet and GooleNet. Considering the number of feature maps in ResNet-50 is much more than that in VGG-16, we finally adopt the VGG-16 network for feature extraction to achieve a better tradeoff between efficacy and efficiency.

Finally, in order to demonstrate the effectiveness of the proposed hierarchical deep feature-based gain control model for BQF, we also compared with other two binocular combination schemes including the quadratic summation (QS) model [44] and vector summation (VS) model [45]. For both the standard QS model and VS model, the suggested binocular combination schemes are based on the brightness information. Denote the monocular brightness flux signals of the left-eye and the right-eye by ϵ^L and ϵ^R, respectively. The binocular combination scheme suggested by the standard QS model is described as follows:

$$f_B(I^L, I^R) = \sqrt{(\epsilon^L)^2 + (\epsilon^R)^2}, \tag{28}$$

The binocular combination scheme suggested by the standard VS model is described as follows:

$$f_B(I^L, I^R) = \frac{(\epsilon^L)^2 + (\epsilon^R)^2}{\epsilon^L + \epsilon^R}, \tag{29}$$

To adapt these models to BQF, we directly fed the layer-wise monocular quality scores into the above equations to compute the corresponding layer-wise binocular quality scores. The comparison results by using different binocular combination

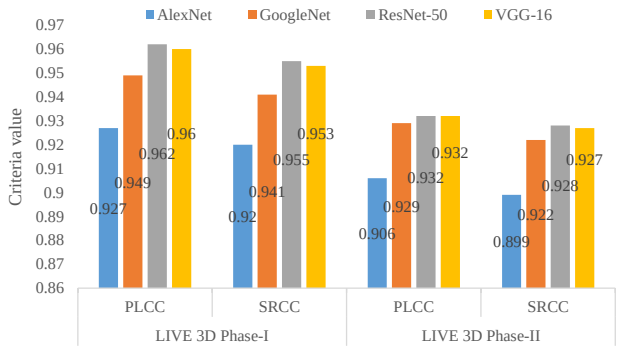


Fig. 7. Performance comparison by using different networks for initial feature extraction. The compared networks include AlexNet [36], GoogleNet [65], and ResNet-50 [66].

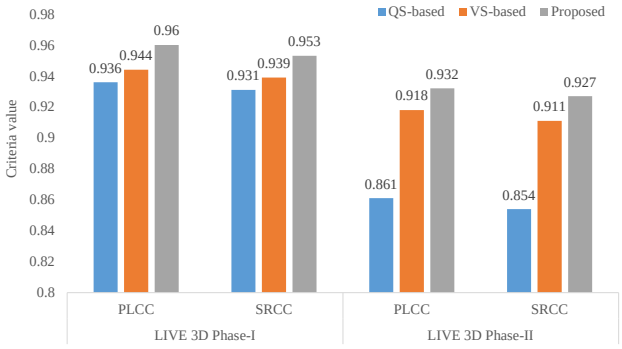


Fig. 8. Performance comparison by using different binocular combination schemes including the quadratic summation (QS) model [44] and vector summation (VS) model [45].

schemes are presented in Fig. 8. We can observe that the proposed method and the VS-based method perform much better than the QS-based method. The reason is that the QS model can only characterize the binocular combination behaviors for the case that the input brightness is symmetric for both eyes, while failing to characterize the well-known the cyclopean perception [47] and the Fechner's paradox [48]. Although the VS-based method is able to somehow characterize the cyclopean perception in human brain, it still fails to resemble the fact that human brain is inherently hierarchical. Therefore, the VS-based method only performs moderately. Overall, the proposed method performs the best as it not only takes the Fechner's paradox and the cyclopean perception into account, but also well resemble the hierarchical visual information processing mechanism of the human brain.

D. Running Time

A good SIQA algorithm is expected to have high predictive accuracy, while being computationally efficient. Computational speed is another important factor for evaluating a SIQA metric, as the quality of an input stereoscopic image needs to be judged online in many practical applications. Therefore, we also test and report the running time of the proposed method. It takes 2.905 seconds to obtain the quality score of a 640×480 stereoscopic image on a personal computer with an Intel(R) Core(TM) i5-4200M CPU Processor at 2.5 GHz, 8 GB RAM, Windows 7 Pro 64-bit. Note that, the running time will be further reduced based on a parallel computing paradigm where the layer-wise monocular quality scores and layer-wise binocular quality scores can be estimated in a parallel manner. Overall, considering the outstanding predictive performance, we believe that such running time is promising.

E. Limitations

There are two major limitations of the proposed method. First, we use a machine learning-based regression model to fuse the layer-wise binocular quality scores into a single quality score. Thus, the proposed method may not be traceable because the machine learning model works as a black-box module in our system. In the future work, we plan to find a more explicit way to fuse the layer-wise binocular quality scores so that the role of each layer can be better understood. Second, the used datasets involve five distortion types which are commonly encountered in stereoscopic image processing systems. The five distortion types include WN, GBLUR, JPEG, JP2K, and FF. However, the much more complex real camera distortions are not specifically treated in our SIQM system, which also remains a future work.

V. Conclusion

This paper has presented a new FR-SIQM method by measuring and fusing the degradations on hierarchical features extracted from pre-trained VGG-16 model. The theoretical development of the proposed model to the community is that we demonstrate that both the two stages in FR-SIQM, i.e., MQE and BQF, should be performed in a hierarchical manner accounting for hierarchical features from low-level to high-level. To be specific, the role of hierarchical features in our method is two-fold: hierarchical feature maps for MQE which estimates a set of layer-wise monocular quality scores, as well as a weighting basis for BQF which estimates a set of layer-wise binocular quality scores. The layer-wise binocular quality scores over layers are fused into a final binocular quality score using SVR. The innovation of this work is that we make the first attempt to make use of hierarchical features extracted from pre-trained deep neural networks to facilitate SIQM and demonstrate its effectiveness. The proposed method is validated by experiments on several public available stereoscopic image quality databases and the experimental results confirm the state-of-the-art performance as well as the efficient computational speed of our proposed method. Future work will focus on extending the framework to address stereoscopic video quality measurement by using deep neural networks for hierarchical spatio-temporal feature extraction.

References

[1] G. Andria *et al.*, "Dosimetric characterization and image quality assessment in breast tomosynthesis," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 10, pp. 2535-2544, Oct. 2017.

[2] F. Attivissimo, G. Cavone, A. M. L. Lanzolla, and M. Spadavecchia, "A technique to improve the image quality in computer tomography," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 5, pp. 1251-1257, May 2010.

[3] B. Li, G. Thomas, and D. Williams, "Detection of ice on power cables based on image texture features," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 3, pp. 497-504, Mar. 2018.

[4] H. Sellahewa and S. A. Jassim, "Image-quality-based adaptive face recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 805-813, Apr. 2010.

[5] G. Yue, K. Gu, and J. Qiao, "Effective and efficient photo-based PM2.5 concentration estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 10, pp. 3962-3971, Oct. 2019.

[6] K. Gu, Z. Xia, and J. Qiao, "Stacked selective ensemble for PM2.5 forecast," *IEEE Transactions on Instrumentation and Measurement*, to be published, 2019.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *in Proc. of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, pp. 1398-1402, 2003.

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

[10] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, Article no. 011006, 2010.

[11] G. Yue, C. Hou, T. Zhou, and X. Zhang, "Effective and efficient blind quality evaluator for contrast distorted images," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 8, pp. 2733-2741, Aug. 2019.

[12] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, and H. Sun,"Optimizing multistage discriminative dictionaries for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2035-2048, Aug. 2018.

[13] Q. Jiang, W. Gao, S. Wang, G. Yue, F. Shao, Y.-S. Ho, S. Kwong, "Blind image quality measurement by exploiting high orderstatistics with deep dictionary encoding network," *IEEE Transactions on Instrumentation and Measurement*, doi: 10.1109/TIM.2020.2984928, 2020.

[14] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Using disparity for quality assessment of stereoscopic images," *in Proc. of the 15th IEEE International Conference on Image Processing*, pp. 389-392, Oct. 2008.

[15] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," *in Proc. of the International Workshop Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, USA, 2010, pp. 1-6.

[16] F. Shao, Z. Zhang, Q. Jiang, W. Lin, and G. Jiang, "Toward domain transfer for no-reference quality prediction of asymmetrically distorted stereoscopic images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 573-585, Mar. 2018.

[17] Y. Jung, H. Sohn, S. Lee, H. Park, and Y. Ro, "Predicting visual discomfort of stereoscopic images using human attention model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 12, pp. 2077-2082, Dec. 2013.

[18] Q. Jiang, F. Shao, W. Gao, H. Li, Y.-S. Ho, "A risk-aware pairwise rank learning approach for visual discomfort prediction of stereoscopic 3D," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1588-1592, Nov. 2019.

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

[20] Q. Jiang, F. Shao, W. Lin, G. Jiang, "Learning a referenceless stereopair quality engine with deep non-negativity constrained sparse auto-encoder," *Pattern Recognition*, vol. 76, pp. 242-255, Apr. 2018.

[21] Q. Jiang, F. Shao, W. Gao, Z. Chen, G. Jiang, and Y.-S. Ho,"Unified no-reference quality assessment of singly and multiply distorted stereoscopic images," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1866-1881, Apr. 2019.

[22] J. Wang, S. Wang, K. Ma, and Z. Wang, "Perceptual depth quality in distorted stereoscopic images," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1202-1215, Mar. 2017.

[23] Q. Jiang, F. Shao, G. Jiang, M. Yu, Z. Peng,"Leveraging visual attention and neural activity for stereoscopic 3Dvisual comfort assessment," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 9405-9425, Apr. 2017.

[24] P. Benzie, et al., "A survey of 3DTV displays: Techniques and technologies," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1647-1658, Nov. 2007.

[25] F. Shao, G. Jiang, M. Yu, K. Chen, and Y. S. Ho, "Asymmetric coding of multi-view video plus depth based 3-D video for view rendering," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 157-167, Jan. 2011.

[26] J. Yang, C. Hou, Y. Zhou, Z. Zhang, and J. Guo, "Objective quality assessment method of stereo images," *in Proc. of IEEE 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON 2009)*, Germany, May 2009.

[27] J. Yang, C. Hou, R. Xu, and J. Lei, "New metric for stereo image quality assessment based on HVS," *International Journal of Imaging Systems and Technology*, vol. 20, no. 4, pp. 301-307, Dec. 2010.

[28] 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, Sep. 2013.

[29] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual fullreference quality assessment of stereoscopic images by considering binocular visual characteristics," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1940-1953, May 2013.

[30] 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*, vol. 24, no. 11, pp. 3810-3825, Nov. 2015.

[31] P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Transactions on Applied Perception*, vol. 3, no. 2, pp. 95-109, Apr. 2006.

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

[33] D. C. Van Essen, and J. H. Maunsell, "Hierarchical organization and functional streams in the visual cortex," *Trends in Neurosciences*, vol. 6, pp. 370-375, 1983.

[34] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," *in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 2394-2402, 2015.

[35] E. Siahaan, A. Hanjalic, and J. A. Redi, "Augmenting blind image quality assessment using image semantics," *in Proc. of the IEEE International Symposium on Multimedia*, pp. 307-312, Dec. 2016.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Proc. of Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[37] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," *arXiv preprint arXiv:1202.2745*, 2012.

[38] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," *in Proc. of the European Conference on Computer Vision*, pp. 354-370, Oct. 2016.

[39] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *in Proc. of the British Machine Vision Conference*, 2015.

[40] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104-114, 2017.

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

[42] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F.-F. L, "Imagenet: A large-scale hierarchical image database," *in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, Jun. 2009.

[43] W. Levelt, "Binocular brightness averaging and contour information," *British Journal of Psychology*, vol. 56, pp. 1-13, 1965.

[44] G. R. Engel, "The autocorrelation function and binocular brightness mixing," *Vision Research*, vol. 9, pp. 1111-1130, 1969.

[45] D. MacLeod, "The Schrodinger equation in binocular brightness combination," *Perception*, vol. 1, pp. 321-324, 1972.

[46] N. Sugie, "Neural models of brightness perception and retinal rivalry in binocular vision," *Biological Cybernetics*, vol. 43, pp. 13-21, 1982.

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

[48] S. Grossberg and F. Kelly, "Neural dynamics of binocular brightness perception," *Vision Research*, vol. 39, no. 22, pp. 3796-3816, 1999.

[49] B. Schlkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[50] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, Art. no. 27, Apr. 2011.

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

[52] 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, Nov. 2015.

[53] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, Nov. 2006.

[54] X. Wang, S. Kwong, and Y. Zhang, "Considering binocular spatial sensitivity in stereoscopic image quality assessment," *in Proc. of the IEEE Conference on Visual Communication and Image Processing*, Nov. 2011, pp. 1-4.

[55] H. Ko, C.-S. Kim, S. Y. Choi, and C.-C. J. Kuo, "3D image quality index using SDP-based binocular perception model," *in Proc. of the 11th IEEE Workshop 3D Image/Video Technol. Appl.*, Jun. 2013.

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

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

[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 Processing: Image Communication*, vol. 28, no. 9, pp. 1143-1155, 2013.

[59] Y. Lin, J. Yang, W. Lu, Q. Meng, Z. Lv, and H. Song, "Quality index for stereoscopic images by jointly evaluating cyclopean amplitude and cyclopean phase," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 89-101, Feb. 2017.

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

[61] J. Ma, P. An, L. Shen, and K. Li, "Joint binocular energy-contrast perception for quality assessment of stereoscopic images," *Signal Processing: Image Processing*, vol. 65, pp. 33-45, Jul. 2018.

[62] W. Zhou, Y. Zhou, W. Qiu, T. Luo, and Z. Zhai, "Perceived quality measurement of stereoscopic 3D images based on sparse representation and binocular combination," *Digital Signal Processing*, vol. 93, pp. 128-137, Oct. 2019.

[63] C. Szegedy, et al., "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, Sep. 2014.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, Jun. 2016.