Blind Stereoscopic Video Quality Assessment: From Depth Perception to Overall Experience

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

***Abstract*—Stereoscopic video quality assessment (SVQA) is a challenging problem. It has not been well investigated on how to measure depth perception quality independently under different distortion categories and degrees, especially exploit the depth perception to assist the overall quality assessment of 3D videos. In this paper, we propose a new Depth Perception Quality Metric (DPQM) and verify that it outperforms existing metrics on our published 3D-HEVC video database. Further, we validate its effectiveness by applying the crucial part of the DPQM to a novel Blind Stereoscopic Video Quality Evaluator (BSVQE) for overall 3D video quality assessment. In the DPQM, we introduce the feature of Auto-Regressive prediction based Disparity Entropy (ARDE) measurement and the feature of energy weighted video content measurement, which are inspired by the free-energy principle and the binocular vision mechanism. In the BSVQE, the binocular summation and difference operations are integrated together with the Fusion Natural Scene Statistic (FNSS) measurement and the ARDE measurement to reveal the key influence from texture and disparity. Experimental results on three stereoscopic video databases demonstrate that our method outperforms state-of-the-art SVQA algorithms for both symmetrically and asymmetrically distorted stereoscopic video pairs of various distortion types.**

***Index Terms*—Stereoscopic video quality assessment, depth perception quality, binocular summation and difference channels, natural scene statistic, autoregressive prediction.**

I. Introduction

THREE-DIMENSIONAL television (3D-TV) provides an entirely new viewing experience. However, there are still many quality issues in stereoscopic contents. Therefore, stereoscopic image/video quality assessment is an important and challenging research problem, which attracts a lot of attentions [1]. Stereoscopic image/video quality assessment contains multi-dimensional qualities. Three basic perceptual quality dimensions, namely picture quality, depth quality and visual discomfort, are identified in [2] to synthetically affect the overall quality of experience (QoE) of 3D image/video. It is essential to evaluate stereoscopic contents in all of the dimensions, not simply in picture quality. In other words, the ultimate goal of stereoscopic video quality assessment is to develop an evaluation criterion that reflects total user experience. Moreover, ocular and cognitive conflicts may cause visual fatigue and discomfort [3], which include vergence-accommodation conflict [4], cognitive integration of conflicting depth cues, and so on. In addition, visual fatigue and discomfort are also caused by display difference, viewing distance, duration of viewing, and subject variation [5], [6]. Meanwhile, several studies and proposed models on 3D visual discomfort have arisen recently. For example in [7], a study on the relationship of 3D video characteristics, eye blinking rate, and visual discomfort is conducted. In [8], a new concept named the percentage of un-linked pixels map (PUP map) is built to predict the degree of 3D visual discomfort. Basically, the experimental methods and models of visual discomfort are quite independent of that of image quality and depth quality. Consequently, when viewing stereoscopic contents, apart from visual discomfort, image quality and depth quality are two significant aspects of overall 3D QoE which this paper concentrates on.

For the overall quality assessment of stereoscopic images, existing objective models can be grouped into three categories. In the first category, some successful 2D image quality assessment (IQA) metrics, which do not explicitly utilize depth-related information, are directly applied to assess 3D image quality. For example, four kinds of 2D IQA metrics are extended to assess stereoscopic image quality [9]. The second category of methods combines depth perception information with image distortion measurement to predict ultimate 3D overall quality. Disparity information is integrated into two 2D image quality metrics (SSIM [10] and C4 [11]) to obtain the overall perceived quality of stereoscopic images [12]. Also, image quality and stereo sensation are designed as separate metrics and can be combined as an objective quality assessment model for 3D images [13]. In [14], three approaches based on 2D image quality metrics are used to integrate disparity images and original images to compute the stereoscopic image quality. In the third category, the binocular vision properties of the human vision system (HVS) are modeled into conventional 2D IQA approaches. Binocular rivalry is one of the widely used physiological models, which incorporates left and right view signals by weights based on their energies, and is utilized in several 3D IQA metrics [15]–[17].

Compared to stereoscopic image quality assessment (SIQA) metrics, the quality evaluation of 3D/stereoscopic videos is quite complex owing to temporal information. Lots of efforts have been devoted to the study of stereoscopic video quality assessment (SVQA) in the last few years. Based on conventional 2D objective quality assessment metrics, the perceptual quality metric (PQM) for overall 3D video quality perception has been proposed [18]. Moreover, the PHVS-3D is a novel SVQA method based on the 3D-DCT transform [19]. Also, the spatial frequency dominance (SFD) model considers the observed phenomenon that spatial frequency

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

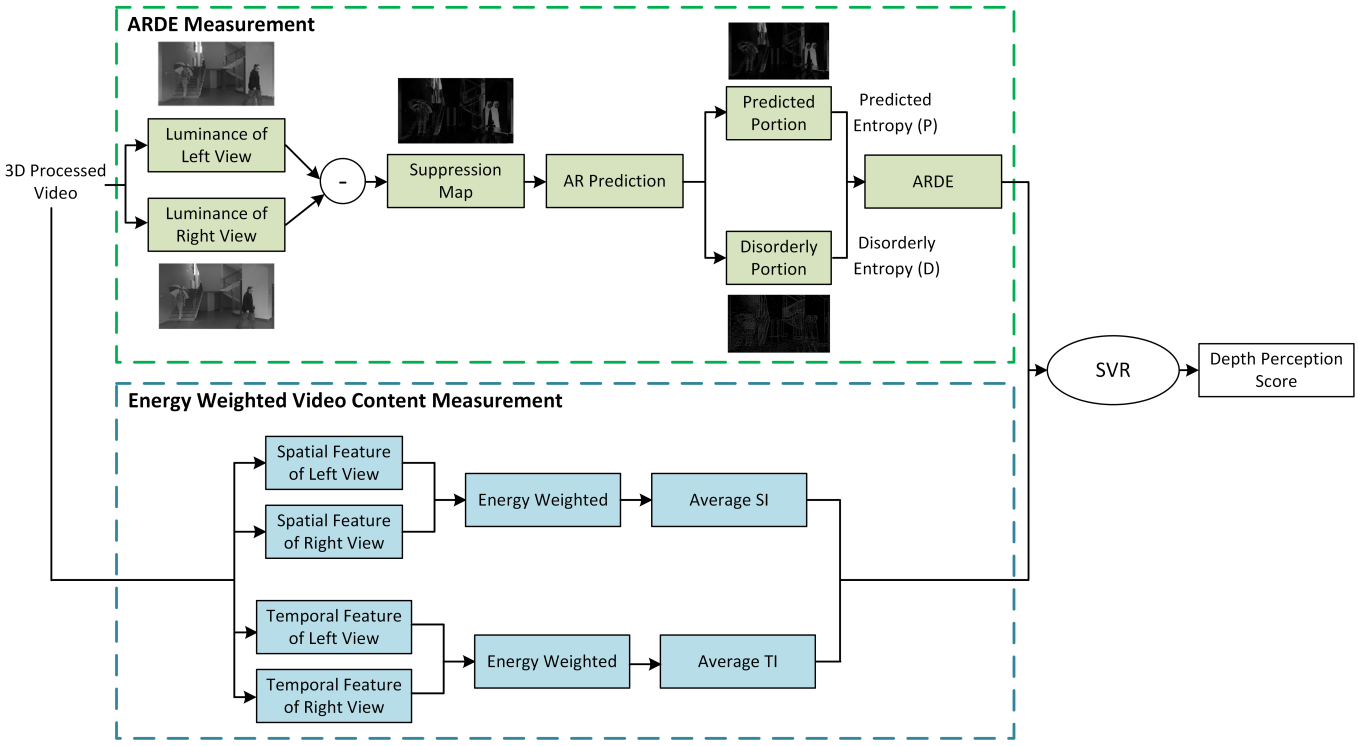

Fig. 1: Flow diagram of the DPQM scheme. The top dotted line block is the Auto-Regressive prediction based Disparity Entropy (ARDE) measurement and the bottom dotted line block is the energy weighted video content measurement.

determines view domination under the ability of the human visual system (HVS) [20]. The 3D spatial-temporal structural (3D-STS) metric has been designed to evaluate the inter-view correlation of spatial-temporal structural information extracted from adjacent frames [21]. Recently, an objective SVQA metric (i.e. the SJND-SVA) has been developed by incorporating the stereoscopic visual attention (SVA) with the stereoscopic just-noticeable difference (SJND) model [22]. However, separate quality assessment models for image quality and depth perception quality are needed to describe different aspects of overall 3D QoE for stereoscopic videos [23]. In this paper, we propose a depth perception quality metric. Then, we also apply the key part of it to the overall quality assessment for 3D/stereoscopic videos.

We design our model inspired by the principles of hierarchical human visual cortex responses to 3D visual signals. Specifically, when the human brain is processing stereoscopic visual signals, the response of binocular disparity is initially formed in the primary visual cortex (V1) area. Further, the depth perception is enhanced through disparity-selective neurons in the secondary cortical area V2. The output of V2 is then used for the processing of dorsal and ventral pathways. It is generally assumed that the dorsal pathway manages the coarse stereopsis, while the ventral pathway focuses on the fine stereopsis [24]. Also, an fMRI study [25] showed that 3D vision stimuli led to V3a activations in the visual cortex. Moreover, V4 visual area plays a crucial role in the aspects of fine depth perception and 3D imaging [26]. Therefore, the neuronal responses to binocular disparity and depth perception exist in both low-level and high-level visual areas.

Besides, the free-energy principle and the binocular vision mechanism have been widely utilized in image/video quality assessment [27], [28]. Inspired by these theories, we build our stereoscopic perception model containing the Auto-Regressive prediction based Disparity Entropy (ARDE) measurement and the energy weighted video content measurement. Firstly, in the ARDE measurement, we apply the auto-regressive approach to decompose inter-ocular difference images into the predicted and the disorderly portions, which is inspired by the free-energy principle. When perceiving and understanding an input visual scene, the free-energy principle indicates that the human brain works as an internal inference process for minimizing the free-energy and always attempts to reduce uncertainty through the internal generative model [29]. Specifically, in addition to the forward prediction from lower cortical areas to higher cortical areas, the feedback from higher-level areas to lower-level areas should also be used to influence the inference, which is known as a circulation process [30], [31]. Secondly, we propose the energy weighted video content measurement inspired by the binocular vision mechanism. According to the psychophysical studies about stereoscopic vision, if similar monocular contents fall on corresponding retinal regions in left and right eyes, binocular fusion occurs and can integrate two retinal regions into a single and stable binocular perception [32]. The fusional region is known as the Panum's area. When perceived contents presented to left and right eyes are obviously different, perception alternates between left and right views, which is called binocular rivalry. Moreover, the HVS cannot tolerate the binocular rivalry for a long time, which results in binocular suppression, and then the entire content

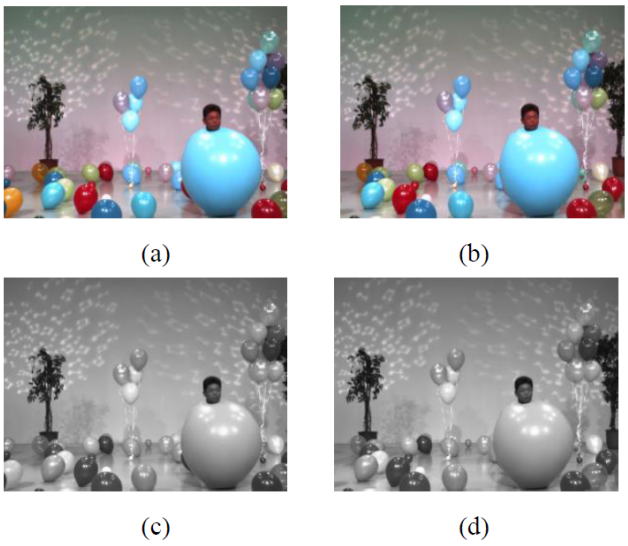

Fig. 2: Luminance extraction of distorted left and right view YUV format videos in 3D-HEVC video database [42]. (a) The last frame of left view video; (b) The last frame of right view video; (c) Gray-scale distorted map of left view; (d) Gray-scale distorted map of right view.

from one of the retina may be suppressed [33]. Based on the binocular rivalry mechanism, high energy region is more likely to contain more important and useful visual information. Therefore, left and right view signals should be integrated by assigning different weights according to their binocular energies [34]–[37]. Also, the binocular rivalry involves neural competition in low-level and high-level cortical areas [38]–[40] as well as the increase of neuron activity in V1, V2, V3a and V4v areas [41]. They are consistent with the responses to binocular disparity and depth perception in the visual cortical areas.

In this paper, motivated by the above observations, we apply the AR model to decompose inter-ocular difference images, and then utilize entropy to reveal binocular disparity variation as well as measure depth perception quality. Meanwhile, we use binocular weights on the spatial and temporal features of 3D videos to reflect video content difference and further influence the ultima depth perception quality. Note that, the depth perception quality is related to disparity and video content according to the subjective experiment in our previous work [42]. Therefore, we synthesize the two measurements to develop a Depth Perception Quality Metric (DPQM). In addition, we propose a Fusion Natural Scene Statistic (FNSS) measurement to represent the binocular fusion peculiarity and complement the ARDE measurement. Also, the FNSS measurement and the ARDE measurement can be integrated to form a Blind Stereoscopic Video Quality Evaluator (B-SVQE) in the binocular summation and difference channels. Furthermore, experimental results show the effectiveness of the proposed stereoscopic perception model.

Since depth perception is important in the overall 3D perceptual quality assessment for stereoscopic videos, we first develop a Depth Perception Quality Metric (DPQM) considering two primary affecting factors (i.e. disparity and video content) for depth perception quality as follows: i) we propose the new AR-based Disparity Entropy (ARDE) feature to measure disparity variation; ii) we propose the Energy Weighted Spatial Information (EWSI) and Temporal Information (EWTI) features to reflect video content difference. These three different types of features are combined by a support vector regression (SVR) model to predict depth perception scores. For the first aspect, suppression maps are generated by subtracting left and right view videos in the luminance plane. The free-energy principle based AR prediction is then conducted on the suppression maps to decompose them into the predicted and the disorderly portions. Then, the statistical entropy feature of these two portions is applied to represent disparity quality. Also, we verify the effectiveness of the depth perception quality assessment model on the latest 3D-HEVC video database.

Based on the DPQM, we then propose a Blind Stereoscopic Video Quality Evaluator (BSVQE) containing three key aspects: i) we apply the binocular summation and difference operations [43] to obtain fusion maps and suppression maps from the prescribed stereoscopic video; ii) we propose some Fusion Natural Scene Statistic (FNSS) features after the zero-phase component analysis (ZCA) whitening filter in the fusion maps; iii) we also utilize the ARDE feature in our depth perception quality model for the suppression maps. Our experimental results show that the performance of our BSVQE correlates well with human visual perception and is validated to be effective and robust on three stereoscopic video databases compared with other SVQA metrics. Our 3D-HEVC stereo video database and a software release of the BSVQE are available online: http://staff.ustc.edu.cn/~chenzhibo/resources.html for public research usage.

The remainder of this paper is organized as follows. Section II introduces the proposed Depth Perception Quality Metric (DPQM) and the experiments on 3D-HEVC video database containing subjective depth perception quality scores. In Section III, we propose the Blind Stereoscopic Video Quality Evaluator (BSVQE), which integrates the image quality from the fusion map and the depth quality from the suppression map. We present experimental results and analysis in Section IV, and then conclude in Section V.

II. Proposed Depth Perception Quality Metric

As depth perception is a fundamental aspect of human quality of experience (QoE) when viewing stereoscopic videos, the evaluation of depth perception quality is important. Therefore, we propose a Depth Perception Quality Metric (DPQM), as depicted in Fig. 1. According to the subjective experiment in our previous work [42], disparity and video content are two dominating factors related to depth perception quality. Firstly, inspired by the free-energy principle, the entropy feature of the suppression map after autoregressive (AR) prediction is extracted to reflect disparity variation. We name it AR-based Disparity Entropy (ARDE) feature. Secondly, according to the binocular vision mechanism, the 3D Energy Weighted Spatial-temporal Information of left and right views, i.e. the

(a) (b) (c)

Fig. 3: Image decomposition with AR prediction model. (a) Suppression map by subtracting the left and right gray-scale distorted maps in Fig. 2; (b) Predicted portion; (c) Disorderly portion, i.e. predicted residual.

EWSI and EWTI features, are then used to demonstrate diverse video contents. Finally, these three different types of features are taken as the input to a support vector regression (SVR) model for training quality prediction to obtain depth perception scores.

A. AR-based Disparity Entropy Measurement (ARDE)

The human vision system (HVS) is more sensitive to luminance. Hence, the gray-scale distorted map, i.e. the Y component of the input YUV format video, is computed first as shown in Fig. 2. The suppression map reflecting disparity information can be obtained using the following operation by subtracting left and right stereo-halves [44], [45]:

$$S^{-} = D_L - D_R \tag{1}$$

where D_L and D_R are the distorted images for left and right views in the luminance channel.

Generally, when perceiving and understanding visual information outside, the human brain always works under the instruction of the free-energy principle and the Bayesian theory [29]–[31]. Here, we utilize an autoregressive (AR) prediction model [46], [47] for image content active inference. Specifically, in order to predict an input image $I(x, U)$, a probabilistic model is adopted by minimizing the prediction error, which is equivalent to maximizing the posterior probability as follows:

$$\max p\left(x/U\right) \quad s.t. \quad U = \{x_1, x_2, \ldots, x_N\} \tag{2}$$

where U represents the 21×21 pixels surrounding the central pixel x in the input image, which is local compared with the relatively larger image size in the experiment. Additionally, it can be seen that those x_i values have the strong correlation with point x and play dominant roles for the maximization goal [48]. Therefore, the mutual information $I\left(x; x_i\right)$ is set as the autoregressive coefficient, and the AR model used to predict the value of central pixel x is given by [49]:

$$x' = \sum_{x_i \epsilon U} a_i x_i + \varepsilon \tag{3}$$

where x_i are all of the adjacent pixels to central point x in a surrounding region, ε is the white noise added to the prediction process, and also the coefficients are computed as follows:

$$a_i = \frac{I(x; x_i)}{\sum\limits_{x_j \in U} I(x; x_j)} \tag{4}$$

With the AR prediction model, i.e. equations (2), (3) and (4), we can obtain the predicted image, then the disorderly image is obtained by directly using the original suppression map to subtract the predicted image. In other words, an input suppression map is decomposed into two portions, namely, the predicted image and the disorderly image which are shown in Fig. 3.

The 'surprise' determined by the entropy of a given image is then computed for the predicted image I_p and the disorderly image I_d respectively:

$$P_H = -\sum P\left(I_p\right) log P\left(I_p\right) \tag{5}$$

$$D_H = -\sum P\left(I_d\right) log P\left(I_d\right) \tag{6}$$

After getting the two entropy features, we combine them by the product operation as:

$$Q_{entropy} = P_H D_H \tag{7}$$

where P_H and D_H are the entropy values for the predicted and the disorderly portions. $Q_{entropy}$ is the disparity quality, which represents the AR-based Disparity Entropy (ARDE) measurement and exposes the disparity difference among various stereoscopic videos. In addition, the relationship between depth score and entropy on 3D-HEVC video database is shown in Fig. 4. In general, from figures (a-b), we can see that the entropy of the predicted portion influences the variation of depth perception quality positively, so does the entropy of the disorderly portion. Also, in figure (c), higher entropy value of the suppression map results in higher depth score. Moreover, the product operation of the predicted and the disorderly parts makes the points in figure (c) cluster into three groups generally, which is more obvious than the distribution of the points in figure (a) and figure (b). And when there is no disparity (i.e. $Q_{entropy} = 0$), the MOS for depth are all

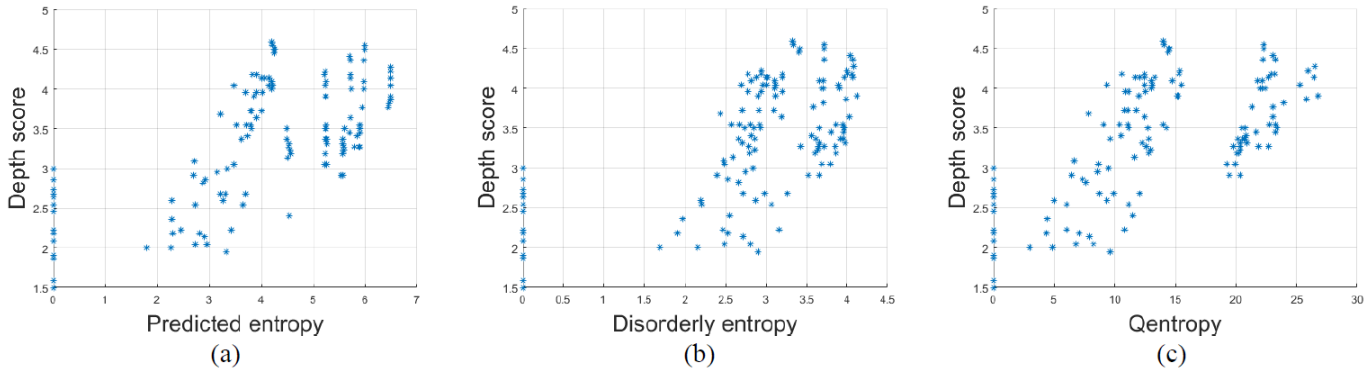

Fig. 4: Demonstration of the relationship between depth score and entropy on 3D-HEVC video database [42]. (a) The relationship between depth score and the entropy of the predicted portion; (b) The relationship between depth score and the entropy of the disorderly portion; (c) The relationship between depth score and Qentropy.

below 3. Note that, the 3D-HEVC video database contains three levels of camera baselines which correspondingly reflect three different perceived depth. Therefore, in this way, the entropy feature after the product operation (i.e. $Q_{entropy}$) can predict the MOS for depth perception effectively.

For example, as presented in Fig. 5, we dispose of the predicted and the disorderly parts separately. In Fig. 5, (a) has a better perceptual depth quality than (d), i.e. the MOS value for depth of (a) is higher than that of (d), though the image qualities of the two stereoscopic videos are the same. Correspondingly, figures (b) and (c) are the predicted and the disorderly portions for figure (a). Also, figures (e) and (f) are the predicted and the disorderly portions for figure (d). The entropy values of (b) and (c) are 5.7132 and 4.0585 (both relatively larger), while the entropy values of (e) and (f) are 5.2400 and 3.9596 (both relatively smaller). Hence, the statistical characteristic of entropy can reflect the variation of depth perception quality.

Specifically, when we have a larger amount of disparity, the suppression map between left and right views has fewer pixels equivalent to 0. Also, most of the pixels in the suppression map is 0. Therefore, the entropy value of the suppression map is higher. In other words, more entropy in the suppression map can reveal larger disparity between the left and right views of stereoscopic videos, which ensures a better perception of disparity and possibly makes the perceived depth quality higher.

The pseudocode of the algorithm for the new ARDE measurement is shown in Algorithm 1. Since different camera baselines represent various depth perception levels, given a stereoscopic video, the AR-based Disparity Entropy (ARDE) values are almost the same for each frame. For convenience, we can utilize the last frame to compute the ARDE feature. Therefore, the feature input into the SVR model is the product of the entropy values which are extracted from the predicted and the disorderly portions of the suppression map in the luminance plane.

Algorithm 1 Disparity model based on AR prediction

Input: Luminance maps of distorted left and right views, i.e. D_L, D_R

Output: Entropy feature after AR prediction of suppression maps $Q_{entropy}$

1: **for** each stereo-halves D_L and D_R **do**
2: $S^- = D_L - D_R \leftarrow suppression\ map$
3: Decompose the suppression map into predicted portion by Eq. (2,3,4) $\leftarrow AR\ prediction$, and disorderly portion (the suppression map subtracts the predicted image)
4: Generate $P_H = -\sum P(I_p) logP(I_p)$
5: Generate $D_H = -\sum P(I_d) logP(I_d)$
6: $Q_{entropy} = P_H D_H$ (entropy feature for predicted and disorderly portions)
7: **end for**
8: **return** $Q_{entropy}$

B. Energy Weighted Video Content Measurement

In order to investigate the impact of different video contents on depth perception quality, we adopt the spatial information (SI) and temporal information (TI) [50] to reflect the spatiotemporal features for left and right view videos. Since the neuronal responses in the visual cortex are almost separate in the space-time domain [51], we use the binocular energies as weights for left and right views on both the spatial and temporal features. This is likely to reveal high-level cortical processing because the features are extracted from global video frames. The SI based on the Sobel filter is computed as follows:

$$SI = max_T \{std_S [Sobel(F_n)]\} \tag{8}$$

where F_n is each frame in the luminance plane at time $n = 1, 2, ..., N$, std_S is the standard deviation over the pixels in the image space, and max_T is the maximum value in a time series T of spatial information for the video. Moreover, the TI is based upon the motion difference feature as below:

$$TI = max_T \{std_S [M_n(i,j)]\} \tag{9}$$

where $M_n(i,j)$ is the difference between the pixel values at the same location in the image space but at successive times or frames of the luminance plane as:

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \tag{10}$$

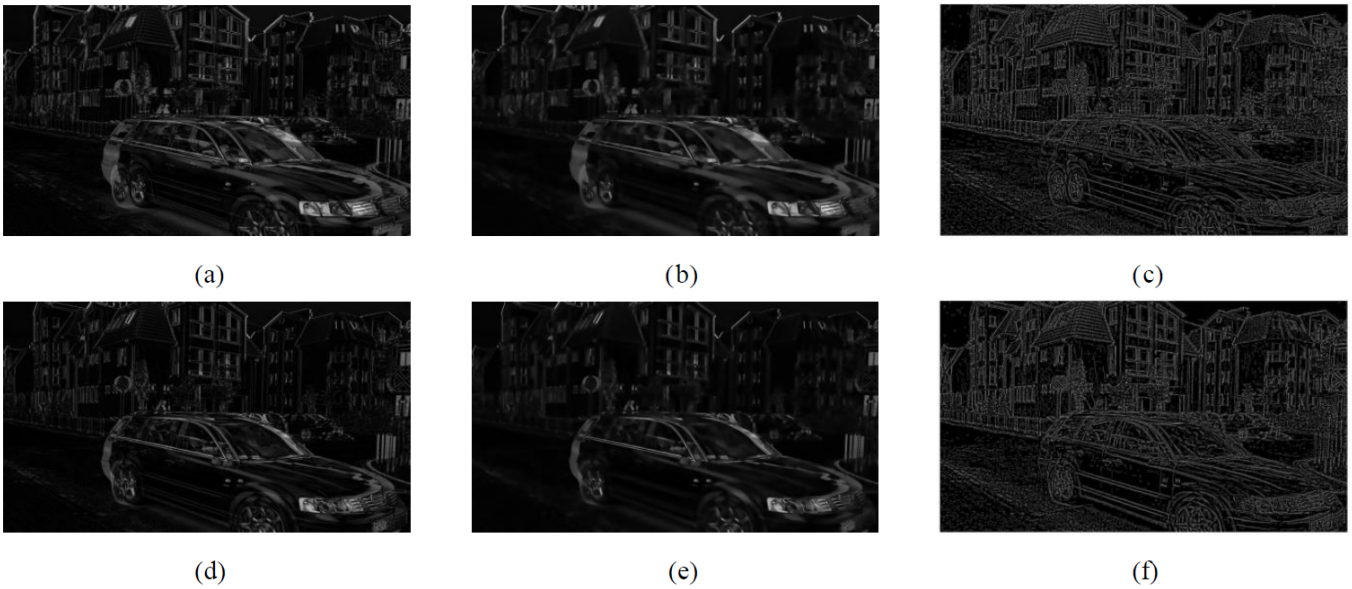

Fig. 5: Demonstration of the effectiveness of the disparity feature ARDE. The first column contains the suppression maps, while the middle and last columns contain the corresponding predicted and disorderly portions, respectively. (a) MOS for depth is 4.1818; (b) Decomposed predicted portion of (a), entropy=5.7132; (c) Decomposed disorderly portion of (a), entropy=4.0585; (d) MOS for depth is 3.3636; (e) Decomposed predicted portion of (c), entropy=5.2400; (f) Decomposed disorderly portion of (c), entropy=3.9596.

TABLE I: 2D FR METRICS ON VIDEO SEQUENCES AND DEPTH SEQUENCES

	Video Sequences		**Depth Sequences**	
Metrics	SROCC	LCC	SROCC	LCC
PSNR	0.0488	0.0459	-0.2430	-0.2382
SSIM	0.1521	0.1403	-0.1635	-0.2059
FSIM	0.2171	0.2215	-0.1098	-0.1987
MS-SSIM	0.2213	0.2272	-0.1475	-0.1749

TABLE II: COMPARISON WITH 2D FR METRICS ON SUPPRESSION MAPS

Metrics	**SROCC**	**LCC**
PSNR	0.6818	0.8215
SSIM	0.4874	0.5981
FSIM	0.7108	0.8246
MS-SSIM	0.4894	0.6062
Proposed DPQM	**0.8654**	**0.9187**

where $F_n(i,j)$ is the pixel at position coordinate (i,j) of the nth frame in time. Therefore, more motion in adjacent frames results in higher values of TI.

In addition, the binocular rivalry mechanism indicates that high energy region is more likely to contain more important and useful visual information. In other words, left and right views should be assigned different weights according to their binocular energies [34]–[37]. Hence, we utilize an energy weighted pooling method [52] given as follows:

$$G_l = \frac{\sum E_{dl} R_l}{\sum E_{dl}} \quad and \quad G_r = \frac{\sum E_{dr} R_r}{\sum E_{dr}} \tag{11}$$

where the summations are performed on full energy and ratio maps. G_l and G_r represent the dominant levels of left and right views respectively. Also, E_{ol}, E_{or}, E_{dl}, and E_{dr} are the

TABLE III: COMPARISON WITH METRICS USING DIFFERENT DISPARITY FEATURES ON SUPPRESSION MAPS

Metrics	**SROCC**	**LCC**
Inter-ocular Difference	0.8188	0.8803
Separate Entropy	0.8501	0.9015
Weighted Sum	0.8297	0.8947
Product (Proposed DPQM)	**0.8654**	**0.9187**

energy maps of original and distorted videos by computing the local variances at each spatial location [53]. The local energy ratio maps in both views are computed as follows:

$$R_l = \frac{E_{dl}}{E_{ol}} \quad and \quad R_r = \frac{E_{dr}}{E_{or}} \tag{12}$$

where l and r denote left and right views respectively. Also, d represents the distorted video, while o is the original video. Given the values of G_l and G_r in (11), we compute the weights assigned to left and right views by:

$$w_l = \frac{G_l^2}{G_l^2 + G_r^2} \quad and \quad w_r = \frac{G_r^2}{G_l^2 + G_r^2} \tag{13}$$

Then, we can obtain the Energy Weighted average Spatial Information (EWSI) as well as the Energy Weighted average Temporal Information (EWTI) for the input stereoscopic left and right view videos V_l and V_r as follows:

$$EWSI_{avg} = w_l SI_l + w_r SI_r \tag{14}$$

$$EWTI_{avg} = w_l TI_l + w_r TI_r \tag{15}$$

respectively. We apply the energy weighted method to the spatial and temporal features that can reveal the significance of the spatiotemporal features for left and right views and

Algorithm 2 Energy weighted video content measurement

Input: Left and right view videos V_l, V_r
Output: Average spatial and temporal features $EWSI_{avg}$, $EWTI_{avg}$

for each luminance stereopairs **do**
  Initialize $EWSI_{avg} = 0$, $EWTI_{avg} = 0$
  for $F_n, n = 1 \rightarrow N$ **do**
    $Sobel\,(F_n) \leftarrow Sobel\ filter$
    $std_S[Sobel(F_n)]$ (standard deviation in the spatial domain)
    $SI = max_T\{std_S[Sobel\,(F_n)]\}$ (maximum value in the temporal domain)
  end for
  for pixel at position (i, j) **do**
    Generate $M_n\,(i,j) = F_n\,(i,j) - F_{n-1}\,(i,j)$
    $std_S[M_n(i,j)]$
    $TI = max_T\{std_S[M_n\,(i,j)]\}$
  end for
end for
Generate local energy maps E_{ol}, E_{or}, E_{dl}, E_{dr} (local variances at each spatial location)
$R_l = E_{dl}/E_{ol}, R_r = E_{dr}/E_{or} \leftarrow energy\ radios$
$G_l = (\sum E_{dl}R_l)/\sum E_{dl}, G_r = (\sum E_{dr}R_r)/\sum E_{dr} \leftarrow pooling\ method$
$w_l = G_l^2/(G_l^2 + G_r^2), w_r = G_r^2/(G_l^2 + G_r^2) \leftarrow weights\ assignment$
$EWSI_{avg} = w_l SI_l + w_r SI_r$ and $EWSI_{avg} = w_l SI_l + w_r SI_r$
return $EWSI_{avg}$, $EWTI_{avg}$

ulteriorly reflect the impact of different video contents on the depth perception quality. Also, the pseudocode of the energy weighted video content measurement is presented in Algorithm 2.

C. Depth Perception Quality Evaluation

After extracting the disparity feature and the energy weighted spatial-temporal features, we adopt the SVR to train a regression model that maps these three different types of features into predicted depth perception scores. To our best knowledge, only the 3D-HEVC video database created in our previous work [42] correspondingly provides the subjective depth perception quality score for each stereoscopic video. Also, it contains three levels of camera baselines which represent different perceived depth levels. Therefore, we use this database to validate the effectiveness of our proposed depth perception quality metric.

Due to lack of 3D depth perception quality assessment metrics, we compute the performance of some state-of-the-art 2D FR metrics on the 3D-HEVC database for both video and depth sequences, as shown in Table I. We present the average Spearman rank-order correlation coefficient (SROCC) as well as the average linear correlation coefficient (LCC) performance values of left and right view videos. From Table I, we can find that the qualities of texture image and depth image are not the same as the depth perception. Based on

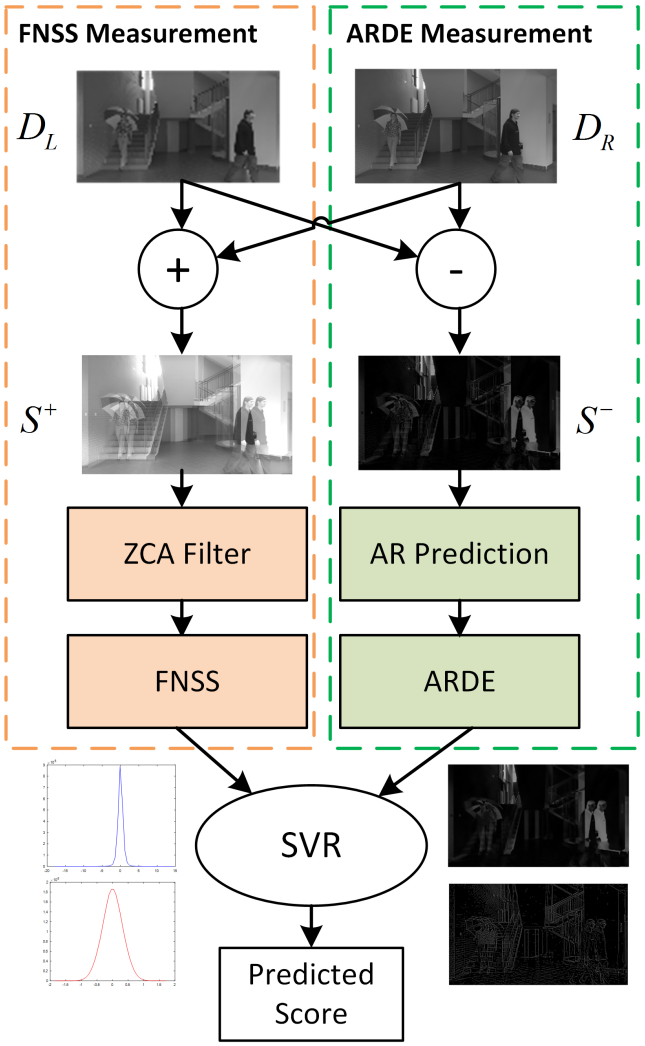

Fig. 6: Flow diagram of the BSVQE method. The left dotted line block is the Fusion Natural Scene Statistic (FNSS) measurement and the right dotted line block represents the Auto-Regressive prediction based Disparity Entropy (ARDE) measurement.

the above analysis, we compare our metric with classical 2D FR metrics on the suppression maps in equation (1). Table II gives the comparison results showing that our depth perception quality evaluation metric outperforms the others.

In addition, we also try to apply different disparity features including the entropy of the inter-ocular difference channel, the separate entropy features of the two portions, and a weighted sum of the entropy values for the predicted and the disorderly portions. The results are shown in Table III. For the weighted sum metric, the weights for each entropy of the predicted and the disorderly portions is set to 0.5 as an example. As can be seen in Table III, the product operation performs better than other metrics. Therefore, it can be demonstrated that the combination of the entropy product operation and the energy weighted spatial-temporal features is effective to develop the depth perception quality metric (DPQM). One possible explanation is that, since the

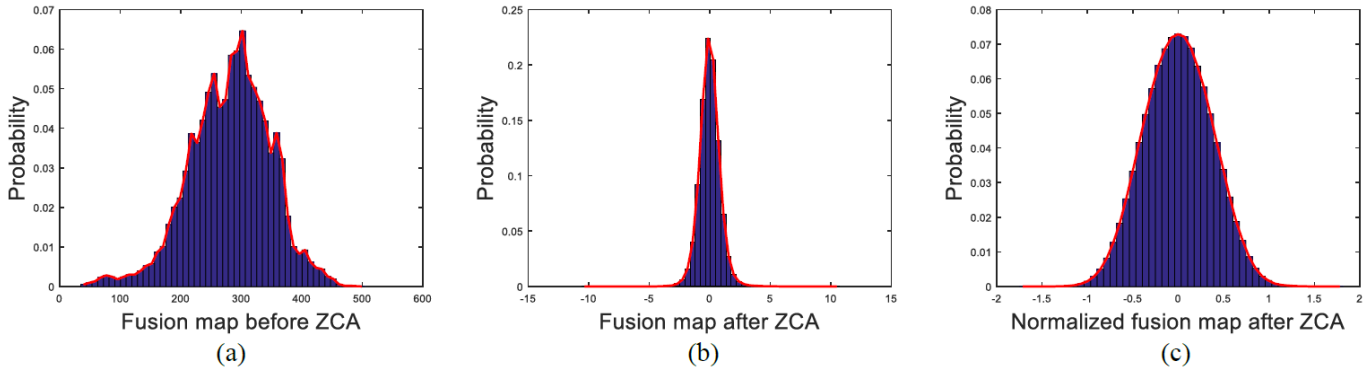

Fig. 7: Demonstration of the effect of the ZCA whitening filter and the divisive normalization on the statistical distribution of the fusion map. (a) Statistical distribution before ZCA; (b) Statistical distribution before normalization of the fusion map after ZCA; (c) Statistical distribution after normalization.

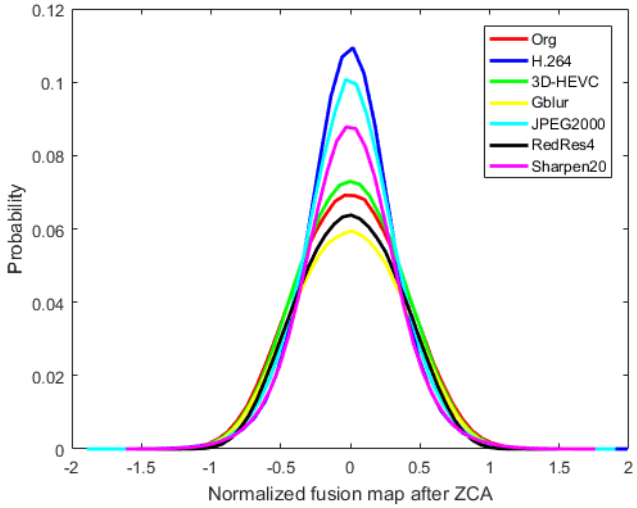

Fig. 8: Statistical distribution of normalized fusion map after ZCA for different distortion types.

entropy values are either 0 or greater than 1 which can be seen in Fig. 4 (a-b), the product of entropy values for the predicted and the disorderly portions is enough to enlarge the discrepancies among various perceived depth levels, compared with other methods. Furthermore, the effectiveness of the proposed DPQM scheme is also validated by incorporating the key part of it into the proposed overall stereoscopic video quality assessment metric, which outperforms existing SVQA metrics on three stereoscopic video databases.

III. Proposed BSVQE Method

For the reason of multi-dimensional quality assessment characteristic for 3D videos, the efficient AR-based Disparity Entropy (ARDE) measurement described in section II-A is adopted by combining with the Fusion Natural Scene Statistic (FNSS) features after the zero-phase component analysis (ZCA) whitening filter. They are applied to develop a more general SVQA method. The block diagram of the proposed Blind Stereoscopic Video Quality Evaluator (BSVQE) is shown in Fig. 6. Firstly, we utilize the binocular summation and difference operations to obtain the fusion map as well as the suppression map of the distorted left and right views in the luminance plane. Secondly, the ZCA whitening filter is applied to the fusion map, and then the FNSS features are extracted from the filtered image. Thirdly, the AR prediction based depth perception is applied to the subtracted suppression map to extract the ARDE feature. Finally, the support vector regression (SVR) is adopted to predict the overall quality scores for stereoscopic videos.

A. Binocular Summation and Difference Channels

Depending on the scenes viewed by left and right eyes, binocular vision operates in several kinds of 'modes' [54]. If left and right images are completely incompatible, then the binocular rivalry occurs and our visual perception alternates between two views. Otherwise, our eyes fuse left and right views into a single percept which is usually close to the summation of left and right images as:

$$S^{+} = D_L + D_R \tag{16}$$

where D_L and D_R are left and right view images respectively.

The summation and difference of a stereo-pair is shown in Fig. 6 as an example. We can see that the images from the two channels are quite different. Specifically, the summation image reflects the fusion ability of the stereo-halves, while the difference image reveals the disparity information between left and right views. Then, the signals from the binocular summation and difference channels are multiplexed so that each primary visual cortex (V1) neuron receives a weighted sum of the visual signals from these two channels [55]. Therefore, we adopt the binocular summation and difference operations of the distorted left and right view images in the luminance plane, as denoted in equations (1) and (16). This way, we obtain the fusion map and the suppression map simultaneously.

B. Fusion Natural Scene Statistic Measurement (FNSS)

In order to develop a No Reference (NR) stereoscopic video quality assessment metric, we adopt the NSS features

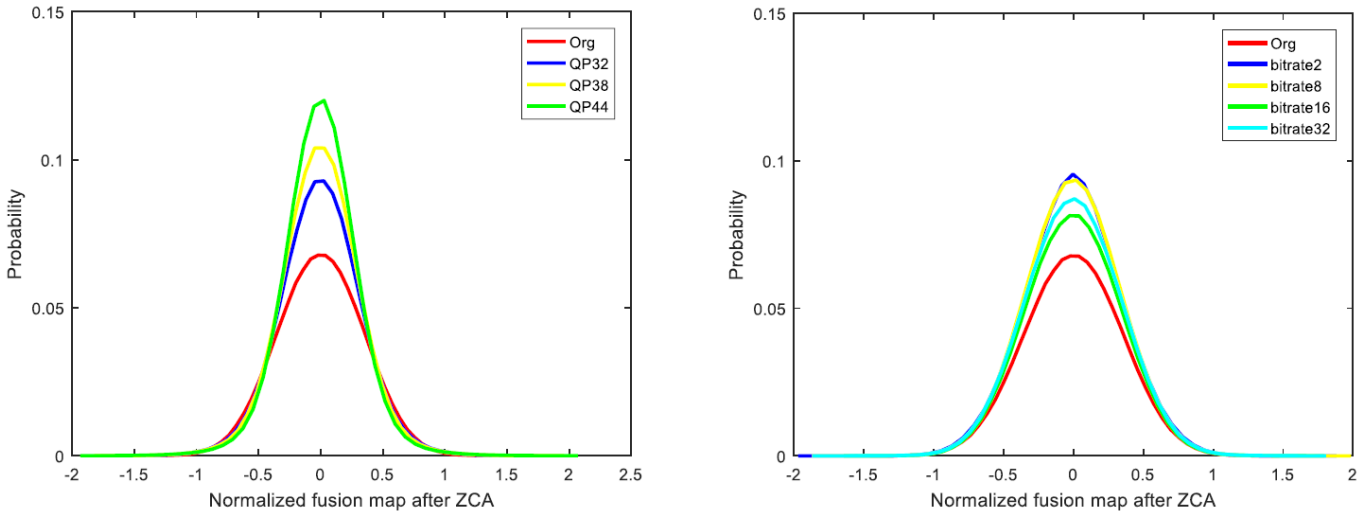

Fig. 9: Illustration of how the statistical distribution of normalized fusion map after ZCA vary with different distortion levels from NAMA3DS1-COSPAD1 database [56]. (a) H.264 video coding distortion; (b) JPEG2000 still image compression distortion.

of Fusion map (FNSS) from the distorted videos. Specifically, before extracting the FNSS features of the distorted left and right views, the zero-phase component analysis (ZCA) whitening filter is applied, in order to reduce the correlation among adjacent pixels, i.e. the spatial redundancy as:

$$Z^{+} = ZCA\left(S^{+}\right) = ZCA\left(D_L + D_R\right) \tag{17}$$

Then, we adopt an NSS model in the spatial domain to extract the features revealing the perceived quality of fusion image. They also complement the disparity feature in the difference channel. We implement the decorrelating effect on the fusion map after ZCA by divisive normalization transform and local mean subtract [54] as below:

$$\widehat{Z^{+}}\left(x,y\right) = \frac{Z^{+}\left(x,y\right) - \mu\left(x,y\right)}{\sigma\left(x,y\right) + C} \tag{18}$$

where C is a small constant to avoid the instability of the denominator, $\mu\left(x,y\right)$ and $\sigma\left(x,y\right)$ are the mean and standard deviation of the input fusion image $Z^{+}\left(x,y\right)$ respectively as:

$$\mu\left(x,y\right) = \sum_{i=-I}^{I}\sum_{j=-J}^{J} w_{i,j} Z_{i,j}^{+}\left(x,y\right) \tag{19}$$

$$\sigma\left(x,y\right) = \sqrt{\sum_{i=-I}^{I}\sum_{j=-J}^{J} w_{i,j}\left(Z_{i,j}^{+}\left(x,y\right) - \mu\left(x,y\right)\right)^{2}} \tag{20}$$

where $w = \{w_{i,j} | i = -I, \ldots, I, j = -J, \ldots, J\}$ is a 2D circularly-symmetric Gaussian weighted function.

Given a fusion image from left and right views in the luminance plane, Fig. 7 shows the statistical distribution before and after the ZCA whitening filter and the decorrelating process. From figures (a-c), we can find that the ZCA filter and the divisive normalization both make the probability distribution of the fusion map more Gaussian-like and statistically significant.

Algorithm 3 Fusion NSS model

Input: Luminance maps of distorted left and right views, i.e. D_L, D_R

Output: FNSS features extracted from the fusion maps

1: **for** each stereo-halves D_L and D_R **do**
2: $S^{+} = D_L + D_R \leftarrow fusion\ map$
3: Generate $Z^{+} = ZCA\left(S^{+}\right) \leftarrow ZCA\ filter$
4: Generate $\widehat{Z^{+}} \leftarrow Z^{+}$ by Eq. (18,19,20)
5: Quantify the statistical distribution using AGGD model $f(x;\lambda,\sigma_l^2,\sigma_r^2)$ by Eq. (21,22,23,24)
6: Utilize original image and reduced resolution scales to perform as the FNSS features
7: **end for**
8: **return** $FNSS\ features$

In addition, the probability distribution of various distortion types for the normalized fusion map after ZCA compared with the pristine stereoscopic video is shown in Fig. 8. The original video and 3D-HEVC type distortion are from 3D-HEVC database [42], while the Gaussian blur distortion comes from SVQA database [22]. Also, the H.264, JPEG2000, reduction of resolution for 4 downsampling and image sharpening (edge enhancement) are from NAMA3DS1-COSPAD1 database [56]. Fig. 9 shows that the probability distribution of the normalized fusion images after ZCA is affected by different distortion levels of 3D videos. We use the H.264 video coding and JPEG2000 still image compression distortion from the NAMA3DS1-COSPAD1 database, as illustrated in Fig. 9.

Then, we quantify the statistical distribution using the asymmetric generalized Gaussian distribution (AGGD) [57]. The AGGD with zero mean value mode to fit the distribution

TABLE IV: SROCC COMPARISON ON 3D-HEVC DATABASE

Metrics	2D Direct Average	3D Weighted Average [61]	2 Pooling (BEST) [62]
PSNR	0.3804	0.3851	0.3429
SSIM	0.3811	0.3629	0.3777
FSIM	0.6993	0.6930	0.6865
MS-SSIM	0.6149	0.6027	0.6189
VQM	0.6800	0.6602	0.6766
Proposed BSVQE	**0.8970**		

TABLE V: COMPARISON WITH 3D METRICS ON BOTH SVQA AND NAMA3DS1-COSPAD1 DATABASES

	SVQA database		**NAMA3DS1-COSPAD1 database**	
Metrics	SROCC	LCC	SROCC	LCC
PQM in [18]	0.8165	0.7852	0.6006	0.6340
PHVS-3D in [19]	0.7195	0.7082	0.5146	0.5480
SFD in [20]	0.6633	0.6483	0.5896	0.5965
3D-STS in [21]	0.8338	0.8311	0.6214	0.6417
SJND-SVA in [22]	0.8379	0.8415	0.6229	0.6503
Proposed BSVQE	**0.9387**	**0.9394**	**0.9086**	**0.9239**

is given by:

$$f(x;\lambda,\sigma_l^2,\sigma_r^2)=\begin{cases}\frac{\lambda}{(\rho_l+\rho_r)\Gamma(1/\lambda)}e^{-(\frac{-x}{\rho_l})^{\lambda}} & x<0\\ \frac{\lambda}{(\rho_l+\rho_r)\Gamma(1/\lambda)}e^{-(\frac{x}{\rho_r})^{\lambda}} & x\geq 0\end{cases} \tag{21}$$

where

$$\rho_l=\sigma_l\sqrt{\frac{\Gamma\left(\frac{1}{\lambda}\right)}{\Gamma\left(\frac{3}{\lambda}\right)}} \tag{22}$$

$$\rho_r=\sigma_r\sqrt{\frac{\Gamma\left(\frac{1}{\lambda}\right)}{\Gamma\left(\frac{3}{\lambda}\right)}} \tag{23}$$

and λ is the shape parameter controlling the shape of the statistic distribution, while σ_l^2, σ_r^2 are the scale parameters of the left and right sides respectively. Further, the AGGD becomes the generalized Gaussian distribution (GGD) when $\rho_l=\rho_r$. For each fusion map of stereoscopic videos, the parameters $\left(\lambda,\sigma_l^2,\sigma_r^2\right)$ are estimated using the moment-matching based approach [58]. Also, the parameters $\left(\eta,\lambda,\sigma_l^2,\sigma_r^2\right)$ of the best AGGD fit are computed where η is given by:

$$\eta=(\rho_r-\rho_l)\frac{\Gamma\left(\frac{2}{\lambda}\right)}{\Gamma\left(\frac{1}{\lambda}\right)} \tag{24}$$

Thus the two scales, i.e. the original image scale and a reduced resolution scale (low pass filtered and downsampled by a factor of 2) proposed in [59] are used to perform as the FNSS features extracted from the fusion map. The pseudocode of the FNSS measurement is shown in Algorithm 3.

C. Overall 3D QoE with AR Based Depth Perception

As can be seen in section II, the AR prediction is adopted to gray-scale suppression maps. Then, the entropy feature along with the spatial-temporal features are used to assess the depth perception quality. Also, we validate the effectiveness of the depth perception model on 3D-HEVC video database. Furthermore, the 3D-HEVC video sequences have been the only available stereoscopic videos of multi-view plus depth (MVD) format to date. Hence, this disparity feature (i.e. the ARDE feature) saves the computation complexity with no need for applying different algorithms to estimate depth maps. Then according to the relationship between depth and disparity to compute the disparity as follows [60]:

$$d=\frac{fB}{z_p} \tag{25}$$

where f is the focal length, B is the baseline of the camera, and z_p is the depth value.

Additionally, the depth map, as computed from numerical images, is a concept that we use in order to represent depth information. However perceptually, current studies suggest that the depth map is not available in the human vision system (HVS). If there exists, it is likely that such a depth map is the output of a high-level visual cortical area, which is not always correlating well with the depth perception quality. Therefore, here we utilize the disparity feature by computing the entropy after AR prediction in the difference channel, instead of the depth map, as a vital input feature to get the 3D overall quality for stereoscopic videos.

TABLE VI: PERFORMANCE OF DIFFERENT DISTORTION TYPES ON SVQA DATABASE

Distortion type	SROCC	LCC
H.264	0.9379	0.9371
Gaussian blur	0.9505	0.9568

IV. Experimental Results and Analysis

We conduct experiments on three 3D/stereoscopic video databases, namely, two widely-used and publicly available databases (i.e. SVQA database [22] and NAMA3DS1-COSPAD1 database [56]) as well as our established 3D-HEVC video database [42], to examine the validity of our proposed overall SVQA model BSVQE. The 3D-HEVC video database contains 138 stereoscopic distorted video sequences obtained from 6 source 3D videos. They cover various spatial and temporal complexities of texture video and of depth video. In the 3D-HEVC database, the distorted videos are under different artifact levels of symmetric and asymmetric 3D-HEVC compression, and have diverse depth levels. Moreover, when watching stereoscopic videos, viewers need to rate three types of scores, i.e. image quality, depth quality, and 3D

TABLE VII: SROCC OF DIFFERENT DISTORTION TYPES ON NAMA3DS1-COSPAD1 DATABASE

Metrics	H.264	JPEG2000	Downsampling and Sharpening
SJND-SVA in [22]	0.6810	0.6901	0.5071
Proposed BSVQE	**0.8857**	**0.8383**	**0.8000**

TABLE VIII: LCC OF DIFFERENT DISTORTION TYPES ON NAMA3DS1-COSPAD1 DATABASE

Metrics	H.264	JPEG2000	Downsampling and Sharpening
SJND-SVA in [22]	0.5834	0.8062	0.6153
Proposed BSVQE	**0.9168**	**0.8953**	**0.9750**

overall quality, while all of the videos are in the zone of visual comfort. The image quality is the perceived quality of the pictures. Also, the depth quality refers to the ability of the video to deliver an enhanced sensation of depth. Moreover, the 3D overall quality is given by comprehensively considering both the image quality and the depth quality.

The SVQA database contains 450 symmetric and asymmetric stereoscopic video clips that can be classified into two distortion types. To be specific, half of them are under H.264 video coding compression distortion and the remaining are under Gaussian blur artifacts.

The NAMA3DS1-COSPAD1 database takes 10 source sequences (SRCs) from NAMA3DS1 to be impaired by various spatial or coding degradations, which has 100 symmetric distorted stereoscopic videos. The coding impairments are introduced through the H.264/AVC video coder and JPEG 2000 still image coder. Also, losses in resolution have been considered. Specifically, two hypothetical reference conditions (HRCs) sequences are either downsampled by a factor of 4 or sharpened by image edge enhancement.

In all of the above three stereoscopic video databases, the Absolute Category Rating with Hidden Reference (ACR-HR) on 5 discrete scales has been performed. Also, the associated mean opinion score (MOS) value is provided for each stereoscopic video. Additionally, each database is divided randomly into 80% for training and 20% for testing. We perform 1000 iterations of cross validation on each database, and provide the median Spearman rank-order correlation coefficient (SROCC) and linear correlation coefficient (LCC) performance as the final measurement.

In 3D-HEVC database, for the evaluation of our metric performance, SROCC and LCC are used. Meanwhile, higher correlation coefficient means better correlation with human perceived quality judgment. We compare with three different kinds of 2D and 3D algorithms by SROCC, which are 2D direct average, 3D weighted average, and two times pooling strategy on the texture and depth sequences of the videos. Furthermore, in SVQA and NAMA3DS1-COSPAD1 databases, SROCC and LCC are also adopted to compare the performance with other state-of-the-art stereoscopic video quality assessment metrics to verify the effectiveness of our proposed BSVQE method.

A. Correlation with MOS on 3D-HEVC Database

As can be seen in Table IV, we compare the SROCC performance of our proposed BSVQE method with other state-of-the-art 2D and 3D Full Reference (FR) objective stereoscopic video quality metrics on 3D-HEVC video database, including 2D direct average, 3D weighted average [61], and two times pooling approach [62]. The 2D direct average is computed by averaging the SROCCs performance of the classical 2D metrics for left and right view videos to derive the SROCC of a 3D video. In the 3D weighted average metric, a 2D-to-3D weighted scheme is added to 2D FR metrics, which accounts for the effective binocular vision perception mechanism of the HVS [61]. Here, we apply the weighted scheme to the widely used 2D FR metrics. The twice pooling method aims to evaluate the 3D/stereoscopic video coding quality with 2D objective metrics by using both the texture and depth sequences for two times pooling [62]. Three different types of pooling functions are used and we take the best results presented in Table IV. From Table IV, we can observe that our proposed BSVQE method achieves superior SROCC performance compared with those three state-of-the-art algorithms. Meanwhile, we also compute the LCC achieving 0.9273 for our metric.

B. Comparison with Other 3D SVQA Metrics

In order to demonstrate the robustness and effectiveness of our proposed BSVQE method on more stereoscopic video databases, we conduct more experiments on SVQA and NAMA3DS1-COSPAD1 databases. Table V shows the comparison of SROCC and LCC with several state-of-the-art 3D objective quality assessment metrics, such as PQM [18], PHVS-3D [19], SFD [20], 3D-STS [21] and SJND-SVA [22], for stereoscopic videos in SVQA database and NAMA3DS1-COSPAD1 database. The results show that our method outperforms the others notably. Furthermore, in order to discover how the percentage number of training affects the overall performance of our BSVQE algorithm, we vary the percentage of training and testing sets to plot the median performance for 3D-HEVC video database [42], SVQA database [22], and NAMA3DS1-COSPAD1 database [56]. Fig. 10 shows the change of SROCC and LCC performance with respect to the training percentage. We can observe that a large number of training data bring about the increase of performance on all of the three stereoscopic video databases.

C. Performance on Individual Distortion Types

As SVQA and NAMA3DS1-COSPAD1 databases both consist of different distortion types, it is interesting to know the performance on each individual distortion type. In this experiment, we examine the SROCC and LCC performance of our proposed BSVQE method for each separate distortion type on two stereoscopic video databases as shown in Table VI, VII,

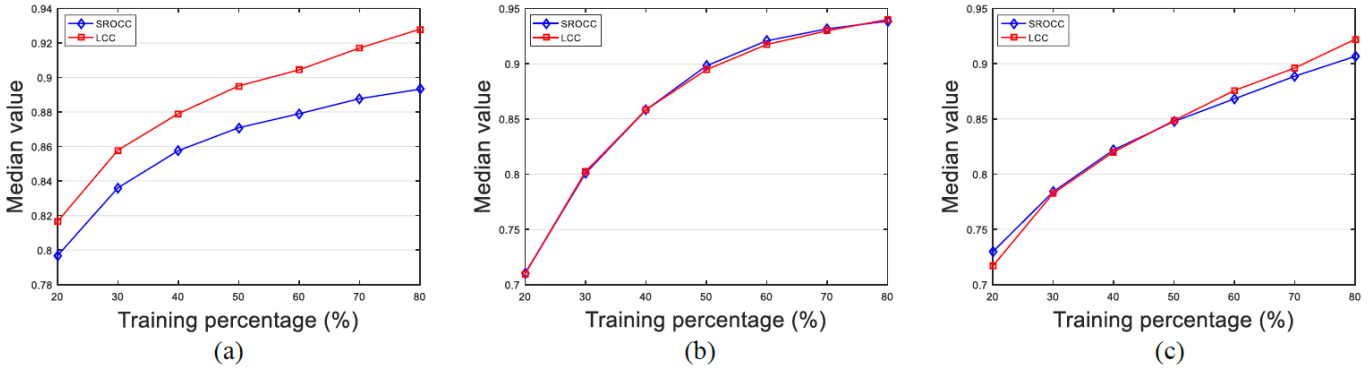

Fig. 10: Mean SROCC and LCC performance with respect to the training percentage over 1000 iterations. (a) Run on 3D-HEVC video database [42]; (b) Run on SVQA database [22]; (c) Run on NAMA3DS1-COSPAD1 database [56].

and VIII. In Table VII and VIII, we also compare with the state-of-the-art 3D video quality assessment algorithm SJND-SVA [22] on NAMA3DS1-COSPAD1 database. From the results presented in the tables, we can find that our proposed BSVQE method is suited to various distortion types for both symmetric and asymmetric distorted stereoscopic videos.

V. Conclusions

In this paper, we have presented a depth perception quality metric and extended it to a No Reference stereoscopic video quality assessment (SVQA) method for stereoscopic videos. The main contributions of this work are: 1) according to the subjective experiment in our previous work, we derive a depth perception quality prediction model based on the free-energy principle as well as the binocular vision mechanism and verify the effectiveness of this depth perception quality assessment model on 3D-HEVC video database; 2) we propose a Blind Stereoscopic Video Quality Evaluator (BSVQE) for assessing the 3D overall quality of distorted stereoscopic videos, which is different from other SVQA metrics, in the sense that we consider the depth feature ARDE along with the texture feature FNSS into the SVQA problem; 3) in the BSVQE method, we combine the binocular summation and difference channels with NSS and entropy; 4) a comparison of our method with some 2D/3D state-of-the-art SVQA metrics is conducted on three databases. Our results show that our introduced metric is promising at handling the stereoscopic video quality assessment problem of both symmetrically and asymmetrically distorted 3D videos, as well as for different distortion types.

In addition, we would like to point out that the stereoscopic video databases used in our current study include mostly artifacts due to various codecs. Also, how to apply different color spaces and extend our database as well as quality assessment model to investigate the influence of synthesis distortions [63] should be considered in future research. Meanwhile, future work could also involve modeling direct effects of binocular rivalry on the difference in perceived depth quality, such as local depth discrepancies, conflicting depth cues, and depth flickering, etc. Besides, we presume that it may be worthwhile to investigate the use of alternative statistical features and introduce more cortical functions to conduct further psychophysical studies on visual cortex in the stereoscopic perception model. Furthermore, apart from depth perception, it is important to understand human opinions on visual discomfort, aiming to develop a more complete objective quality assessment model for 3D QoE.

References

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

[2] I. Union, "Subjective methods for the assessment of stereoscopic 3dtv systems," *Recommendation ITU-R BT*, vol. 2021, 2015.

[3] M. Urvoy, M. Barkowsky, and P. Le Callet, "How visual fatigue and discomfort impact 3D-TV quality of experience: a comprehensive review of technological, psychophysical, and psychological factors," *annals of telecommunications-annales des télécommunications*, vol. 68, no. 11-12, pp. 641–655, 2013.

[4] R. Patterson, "Human factors of 3-D displays," *Journal of the Society for Information Display*, vol. 15, no. 11, pp. 861–871, 2007.

[5] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, "Visual discomfort and visual fatigue of stereoscopic displays: a review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 30 201–1, 2009.

[6] J. Park, S. Lee, and A. C. Bovik, "3D visual discomfort prediction: vergence, foveation, and the physiological optics of accommodation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 415–427, 2014.

[7] J. Li, M. Barkowsky, and P. Le Callet, "Visual discomfort is not always proportional to eye blinking rate: exploring some effects of planar and in-depth motion on 3DTV QoE," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM 2013*, 2013, pp. pp–1.

[8] J. Chen, J. Zhou, J. Sun, and A. C. Bovik, "Visual discomfort prediction on stereoscopic 3D images without explicit disparities," *Signal Processing: Image Communication*, vol. 51, pp. 50–60, 2017.

[9] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic images quality assessment," in *Signal Processing Conference, 2007 15th European*. IEEE, 2007, pp. 2110–2114.

[10] 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.

[11] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3. IEEE, 2003, pp. III–185.

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

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

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

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

[16] 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.

[17] 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.

[18] P. Joveluro, H. Malekmohamadi, W. C. Fernando, and A. Kondoz, "Perceptual video quality metric for 3D video quality assessment," in *2010 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video*. IEEE, 2010, pp. 1–4.

[19] L. Jin, A. Boev, A. Gotchev, and K. Egiazarian, "3D-DCT based perceptual quality assessment of stereo video," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 2521–2524.

[20] F. Lu, H. Wang, X. Ji, and G. Er, "Quality assessment of 3D asymmetric view coding using spatial frequency dominance model," in *2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*. IEEE, 2009, pp. 1–4.

[21] J. Han, T. Jiang, and S. Ma, "Stereoscopic video quality assessment model based on spatial-temporal structural information," in *Visual Communications and Image Processing (VCIP), 2012 IEEE*. IEEE, 2012, pp. 1–6.

[22] F. Qi, D. Zhao, X. Fan, and T. Jiang, "Stereoscopic video quality assessment based on visual attention and just-noticeable difference models," *Signal, Image and Video Processing*, vol. 10, no. 4, pp. 737–744, 2016.

[23] M.-J. Chen, D.-K. Kwon, and A. C. Bovik, "Study of subject agreement on stereoscopic video quality," in *Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on*. IEEE, 2012, pp. 173–176.

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

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

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

[27] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2015.

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

[29] D. C. Knill and A. Pouget, "The bayesian brain: the role of uncertainty in neural coding and computation," *TRENDS in Neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.

[30] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology-Paris*, vol. 100, no. 1, pp. 70–87, 2006.

[31] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[32] I. P. Howard and B. J. Rogers, *Binocular vision and stereopsis*. Oxford University Press, USA, 1995.

[33] S. Steinman, B. Steinman, and R. Garzia, *Foundations of binocular vision: a clinical perspective*. McGraw Hill Professional, 2000.

[34] W. J. Levelt, "The alternation process in binocular rivalry," *British Journal of Psychology*, vol. 57, no. 3-4, pp. 225–238, 1966.

[35] R. Blake, "Threshold conditions for binocular rivalry." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3, no. 2, p. 251, 1977.

[36] M. Fahle, "Binocular rivalry: Suppression depends on orientation and spatial frequency," *Vision research*, vol. 22, no. 7, pp. 787–800, 1982.

[37] 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.

[38] H. R. Wilson, "Computational evidence for a rivalry hierarchy in vision," *Proceedings of the National Academy of Sciences*, vol. 100, no. 24, pp. 14 499–14 503, 2003.

[39] A. W. Freeman, "Multistage model for binocular rivalry," *Journal of Neurophysiology*, vol. 94, no. 6, pp. 4412–4420, 2005.

[40] F. Tong, M. Meng, and R. Blake, "Neural bases of binocular rivalry," *Trends in cognitive sciences*, vol. 10, no. 11, pp. 502–511, 2006.

[41] A. Polonsky, R. Blake, J. Braun, and D. J. Heeger, "Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry," *Nature neuroscience*, vol. 3, no. 11, pp. 1153–1159, 2000.

[42] W. Zhou, N. Liao, Z. Chen, and W. Li, "3D-HEVC visual quality assessment: Database and bitstream model," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, pp. 1–6.

[43] Z. Li and J. J. Atick, "Efficient stereo coding in the multiscale representation*," *Network: computation in neural systems*, vol. 5, no. 2, pp. 157–174, 1994.

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

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

[46] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, 2009.

[47] X. Zhang and X. Wu, "Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 887–896, 2008.

[48] M. Vasconcelos and N. Vasconcelos, "Natural image statistics and low-complexity feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 228–244, 2009.

[49] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 43–54, 2013.

[50] I. Rec, "P. 910: Subjective video quality assessment methods for multimedia applications," *Int. Telecomm. Union, Geneva*, 2008.

[51] L. Wang, Y. Kaneoke, and R. Kakigi, "Spatiotemporal separability in the human cortical response to visual motion speed: a magnetoencephalography study," *Neuroscience research*, vol. 47, no. 1, pp. 109–116, 2003.

[52] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *2006 International Conference on Image Processing*. IEEE, 2006, pp. 2945–2948.

[53] 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.

[54] S. Lyu, "Dependency reduction with divisive normalization: justification and effectiveness," *Neural computation*, vol. 23, no. 11, pp. 2942–2973, 2011.

[55] K. A. May and L. Zhaoping, "Efficient coding theory predicts a tilt aftereffect from viewing untilted patterns," *Current Biology*, 2016.

[56] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutiérrez, and N. Garcia, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 109–114.

[57] N.-E. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 2281–2284.

[58] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995.

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

[60] S. A. Mahmood and R. F. Ghani, "Objective quality assessment of 3D stereoscopic video based on motion vectors and depth map features," in *Computer Science and Electronic Engineering Conference (CEEC), 2015 7th*. IEEE, 2015, pp. 179–183.

[61] J. Wang, S. Wang, and Z. Wang, "Quality prediction of asymmetrically compressed stereoscopic videos," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3427–3431.

[62] K. Wang, K. Brunnström, M. Barkowsky, M. Urvoy, M. Sjöström, P. Le Callet, S. Tourancheau, and B. Andrén, "Stereoscopic 3D video coding quality evaluation with 2D objective metrics," in *IS&T/SPIE*

Electronic Imaging. International Society for Optics and Photonics, 2013, pp. 86 481L–86 481L.

[63] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Processing: Image Communication*, vol. 30, pp. 78–88, 2015.