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 been well investigated on how to measure depth perception quality independently under different distortion categories and degrees, especially exploiting 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 prove 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...
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...
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 (BSVQE) 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 (ETWI) 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
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]:

$$\sum_{x\in U} D_-(x) = D_L - D_R$$

(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:

$$\max_{I_p, I_d} p(X/Y) \quad s.t. \quad U = \{x_1, x_2, \ldots, x_N\}$$

(2)

where $U$ represents the $21 \times 21$ 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 $I(x; x_i)$ is set as the autoregressive coefficient, and the AR model used to predict the value of central pixel $x$ is given by [49]:

$$x' = \sum_{x_i \in U} a_i x_i + \varepsilon$$

(3)

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

$$a_i = \frac{I(x; x_i)}{\sum_{x_j \in U} I(x; x_j)}$$

(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:

$$P_H = -\sum_{x \in P} P(I_p(x)) \log P(I_p(x))$$

(5)

$$D_H = -\sum_{x \in D} P(I_d(x)) \log P(I_d(x))$$

(6)

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

$$Q_{\text{entropy}} = P_H D_H$$

(7)

where $P_H$ and $D_H$ are the entropy values for the predicted and the disorderly portions. $Q_{\text{entropy}}$ 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. $Q_{\text{entropy}} = 0$), the MOS for depth are all

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.
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. $Q_{\text{entropy}}$) 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.

### 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:

$$SI = \max_T \left\{ \text{std}_S [\text{Sobel} (F_n)] \right\}$$

(8)

where $F_n$ is each frame in the luminance plane at time $n = 1, 2, ..., N$, $\text{std}_S$ 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:

$$TI = \max_T \left\{ \text{std}_S [M_n (i,j)] \right\}$$

(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:

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

(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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Video Sequences</th>
<th>Depth Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.0488</td>
<td>-0.2430</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.1521</td>
<td>-0.1635</td>
</tr>
<tr>
<td>FSIM</td>
<td>0.2171</td>
<td>-0.1098</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.2213</td>
<td>-0.1475</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SROCC</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.6818</td>
<td>0.8215</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.4874</td>
<td>0.5981</td>
</tr>
<tr>
<td>FSIM</td>
<td>0.7108</td>
<td>0.8246</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.4894</td>
<td>0.6062</td>
</tr>
<tr>
<td>Proposed DPQM</td>
<td>0.8654</td>
<td>0.9187</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SROCC</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-ocular Difference</td>
<td>0.8188</td>
<td>0.8803</td>
</tr>
<tr>
<td>Separate Entropy</td>
<td>0.8501</td>
<td>0.9015</td>
</tr>
<tr>
<td>Weighted Sum</td>
<td>0.8297</td>
<td>0.8947</td>
</tr>
<tr>
<td>Product (Proposed DPQM)</td>
<td>0.8654</td>
<td>0.9187</td>
</tr>
</tbody>
</table>

where $F_n(i,j)$ is the pixel at position coordinate $(i,j)$ of the $n$th 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:

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

(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 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:

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

(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:

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

(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:

$$EWSI_{avg} = w_l SI_l + w_r SI_r$$

(14)

$$EWTI_{avg} = w_l TI_l + w_r TI_r$$

(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
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.

Algorithm 2 Energy weighted video content measurement

**Input:** Left and right view videos $V_l, V_r$

**Output:** Average spatial and temporal features $EWSI_{avg}$, $EWTI_{avg}$

1: for each luminance stereopairs do
2: Initialize $EWSI_{avg} = 0$, $EWTI_{avg} = 0$
3: for $F_n, n = 1 \rightarrow N$ do
4: $Sobel(F_n) \leftrightarrow Sobel$ filter
5: $std_S[Sobel(F_n)]$ (standard deviation in the spatial domain)
6: $SI = \max_T \{ std_S \{Sobel(F_n)\}\}$ (maximum value in the temporal domain)
7: end for
8: for pixel at position $(i, j)$ do
9: $M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$
10: $std_S[M_n(i, j)]$
11: $TI = \max_T \{ std_S \{M_n(i, j)\}\}$
12: end for
13: end for
14: Generate local energy maps $E_{dl}, E_{or}, E_{dt}, E_{dr}$ (local variances at each spatial location)
15: $R_l = E_{dl}/E_{dt}, R_r = E_{dr}/E_{or}$ $\leftrightarrow$ energy ratios
16: $G_l = (\sum E_{dl}R_l)/\sum E_{dl}, G_r = (\sum E_{dr}R_r)/\sum E_{dr}$ $\leftrightarrow$ pooling method
17: $w_l = G_l^2/(G_l^2 + G_r^2), w_r = G_r^2/(G_l^2 + G_r^2)$ $\leftrightarrow$ weights assignment
18: $EWSI_{avg} = w_lSI_l + w_rSI_r$ and $EWSI_{avg} = w_lSI_l + w_rSI_r$
19: return $EWSI_{avg}$, $EWTI_{avg}$

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 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. 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.
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:

$$S^+ = D_L + D_R$$ (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
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:

$$Z^+ = ZCA \left( S^+ \right) = ZCA \left( D_L + D_R \right)$$ (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:

$$\hat{Z}^+ (x, y) = \frac{Z^+ (x, y) - \mu (x, y)}{\sigma (x, y)} + C$$ (18)

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

$$\mu (x, y) = \sum_{i=-I}^{I} \sum_{j=-J}^{J} w_{i,j} Z^+_{i,j} (x, y)$$ (19)

$$\sigma (x, y) = \sqrt{\sum_{i=-I}^{I} \sum_{j=-J}^{J} w_{i,j} (Z^+_{i,j} (x, y) - \mu (x, y))^2}$$ (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.

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

![Figure 9](image-url) -- 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.

**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 $Z^+ = ZCA(S^+) \leftarrow ZCA$ filter
4. Generate $\hat{Z}^+ \leftarrow Z^+$ by Eq. (18,19,20)
5. Quantify the statistical distribution using AGGD model $f(x; \lambda, \sigma^2_l, \sigma^2_r)$ 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.
TABLE IV: SROCC COMPARISON ON 3D-HEVC DATABASE

<table>
<thead>
<tr>
<th>Metrics</th>
<th>2D Direct Average</th>
<th>3D Weighted Average [61]</th>
<th>2 Pooling (BEST) [62]</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.3804</td>
<td>0.3851</td>
<td>0.3429</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.3811</td>
<td>0.3629</td>
<td>0.3777</td>
</tr>
<tr>
<td>FSIM</td>
<td>0.6993</td>
<td>0.6930</td>
<td>0.6865</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.6149</td>
<td>0.6027</td>
<td>0.6189</td>
</tr>
<tr>
<td>VQM</td>
<td>0.6800</td>
<td>0.6602</td>
<td>0.6766</td>
</tr>
<tr>
<td>Proposed BSVQE</td>
<td></td>
<td></td>
<td>0.8970</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SVQA database</th>
<th>NAMA3DS1-COSPAD1 database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SROCC</td>
<td>LCC</td>
</tr>
<tr>
<td>PQM in [18]</td>
<td>0.8165</td>
<td>0.7852</td>
</tr>
<tr>
<td>PHVS-3D in [19]</td>
<td>0.7195</td>
<td>0.7082</td>
</tr>
<tr>
<td>SFD in [20]</td>
<td>0.6633</td>
<td>0.6483</td>
</tr>
<tr>
<td>3D-STS in [21]</td>
<td>0.8338</td>
<td>0.8311</td>
</tr>
<tr>
<td>SIND-SVA in [22]</td>
<td>0.8379</td>
<td>0.8415</td>
</tr>
<tr>
<td>Proposed BSVQE</td>
<td>0.9387</td>
<td>0.9394</td>
</tr>
</tbody>
</table>

The disparity is given by:

\[
f(x; \lambda, \sigma_l^2, \sigma_r^2) = \begin{cases} 
\lambda (\rho_l + \rho_r)/2 & x < 0 \\
\lambda (\rho_l + \rho_r)/2 e^{-(x/\rho_r)^2} & x \geq 0
\end{cases}
\]

where

\[
\rho_l = \sigma_l \sqrt{\Gamma \left(\frac{1}{2}\right) / \Gamma \left(\frac{3}{2}\right)}
\]

\[
\rho_r = \sigma_r \sqrt{\Gamma \left(\frac{1}{2}\right) / \Gamma \left(\frac{3}{2}\right)}
\]

and \( \lambda \) is the shape parameter controlling the shape of the statistic distribution, while \( \sigma_l^2, \sigma_r^2 \) are the scale parameters of the left and right sides respectively. Further, the AGGD becomes the generalized Gaussian distribution (GGD) when \( \rho_l = \rho_r \).

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.

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
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. 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 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 number of training from 10% to 90% and consider the 3D/stereoscopic video code distortion types. To be specific, half of them are under H.264 video coding compression distortion and the remaining are under Gaussian blur artifacts.

### C. Performance on Individual Distortion Types

As SVQA and PHVS-3D/SFD 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.

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>H.264</th>
<th>JPEG2000</th>
<th>Downsampling and Sharpening</th>
</tr>
</thead>
<tbody>
<tr>
<td>SJND-SVA in [22]</td>
<td>0.6810</td>
<td>0.6901</td>
<td>0.5071</td>
</tr>
<tr>
<td>Proposed BSVQE</td>
<td>0.8857</td>
<td>0.8383</td>
<td>0.8000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>H.264</th>
<th>JPEG2000</th>
<th>Downsampling and Sharpening</th>
</tr>
</thead>
<tbody>
<tr>
<td>SJND-SVA in [22]</td>
<td>0.3834</td>
<td>0.8062</td>
<td>0.0133</td>
</tr>
<tr>
<td>Proposed BSVQE</td>
<td>0.9168</td>
<td>0.8953</td>
<td>0.9750</td>
</tr>
</tbody>
</table>
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