Signal Processing: Image Communication 126 (2024) 117138

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Deep neural network based distortion parameter estimation for blind quality measurement of stereoscopic images

Yi Zhang [a,*], Damon M. Chandler [b], Xuanqin Mou [a]

[a] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710029, China
[b] College of Information Science and Engineering, Ritsumeikan University, Shiga, 525-8577, Japan

ARTICLE INFO

Keywords:
Quality measurement
Stereoscopic image
Distortion parameter estimation
Deep neural network

ABSTRACT

Stereoscopic/3D image quality measurement (SIQM) has emerged as an active and important research branch in image processing/computer vision field. Existing methods for blind/no-reference SIQM often train machine-learning models on degraded stereoscopic images for which human subjective quality ratings have been obtained, and they are thus constrained by the fact that only a limited number of 3D image quality datasets currently exist. Although methods have been proposed to overcome this restriction by predicting distortion parameters rather than quality scores, the approach is still limited to the time-consuming, hand-crafted features extracted to train the corresponding classification/regression models as well as the rather complicated binocular fusion/rivalry models used to predict the cyclopean view. In this paper, we explore the use of deep learning to predict distortion parameters, giving rise to a more efficient opinion-unaware SIQM technique. Specifically, a deep fusion-and-excitation network which takes into account the multiple-distortion interactions is proposed to perform distortion parameter estimation, thus avoiding hand-crafted features by using convolution layers while simultaneously accelerating the algorithm by using the GPU. Moreover, we measure distortion parameter values of the cyclopean view by using support vector regression models which are trained on the data obtained from a newly-designed subjective test. In this way, the potential errors in computing the disparity map and cyclopean view can be prevented, leading to a more rapid and precise 3D-vision distortion parameter estimation. Experimental results tested on various 3D image quality datasets demonstrate that our proposed method, in most cases, offers improved predictive performance over existing state-of-the-art methods.

1. Introduction

1.1. Background

In recent years, 3D imaging technology has experienced rapid development, giving rise to a stereoscopic/3D viewing experience in both consumer and industrial settings (e.g., 3D television/cinema/gaming, 3D teleoperation, 3D video meetings). However, despite the impressive level of immersion in 3D vision, visual discomfort and/or fatigue can be easily introduced if the quality is not properly maintained. For example, asymmetric image degradations, inter-view mismatches, incorrect depth-of-focus, and unnatural binocular disparity are not encountered in 2D scenarios, but can seriously impact the 3D quality. Even the retargeting of 3D images to different display devices can bring noticeable artifacts such as shape twisting and visually important content loss [1]. Thus, designing effective stereoscopic image quality measurement (SIQM) techniques continues to be an important but extremely challenging task.

In this paper, we address the SIQM scenario in which the quality is measured without using a reference image (or associated side-information). This so-called blind/no-reference (NR) SIQM task is more realistic and applicable, as in many practical applications the reference information is unavailable. A stereoscopic image normally contains two monocular views (called the "stereopair") captured by two individual cameras, and humans judge its quality based mainly on the so-called "cyclopean view", which is a merged 3D view created in the brain. Therefore, an effective SIQM method must somehow mimic this process based only on the two available views. Unfortunately, accurately modeling the binocular vision processes of the human visual system (HVS) to properly construct the cyclopean view is nontrivial. Various issues such as the presence of occlusion and border areas, and differences in the types/amounts of distortions in the stereopair can complicate the cyclopean view synthesis process. Moreover, another level of difficulty can be added when multiple distortions are introduced, in which case

* Corresponding author.
E-mail address: yi.zhang.osu@xjtu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.image.2024.117138
Received 31 August 2023; Received in revised form 28 March 2024; Accepted 21 April 2024
Available online 4 May 2024

the joint effect of different distortion types/levels on the perceived image quality also requires consideration.

Because of the aforementioned difficulties, most existing NR SIQM approaches (e.g., [2–14]) have relied on machine-learning by training models on multiply/singly-distorted 3D images with MOS/DMOS values obtained from human subjects. Unfortunately, such datasets of 3D quality ratings are limited in both quantity and diversity. In particular, because different databases may have images with different contents, distortion types/levels, and even different quality-judging standards, these "opinion-aware" NR SIQM approaches often achieve impressive results on cross-validation tests but relatively weak quality measurement (QM) performance on cross-database tests. In addition, most of these algorithms except [12] were initially developed to work with singly-distorted stereoscopic images, and not with multiply-distorted stereoscopic images.

To remove the dependence on human subjective ratings, a number of "opinion-unaware" approaches have been presented, which perform the SIQM task by using quality-aware measures/models (e.g., [15–23]). These methods often follow a similar pipeline that collapses quality-aware features to the corresponding quality score via various pooling strategies. However, without analyzing/modeling the binocular fusion/rivalry properties of the HVS in stereoscopic vision, these approaches often suffer from relatively weak QM performance when different dataset images are tested.

1.2. Motivation

In light of the aforementioned limitations, we suggest that a better approach would be to build a model that (1) can learn from distorted stereoscopic images in order to provide accurate measurements of quality on various multiply/singly distorted stereoscopic images, but (2) requires no training on human subjective ratings. In our previous work, we presented one such model [24] which operated by indirectly measuring quality via distortion parameter estimation. However, in our attempt to remove the dependence on training data, that model ultimately proved quite lacking from the machine-learning perspective. In particular, both the features and the distortion parameter estimation framework were hand-crafted, making it difficult to select the optimal feature/model combination. Furthermore, the cyclopean view used to predict the *3D-vision* distortion parameters[1] was computed based on a rather complicated binocular model, an inaccurate disparity map, and relatively weak SVM models. Thus, two natural questions arise: (1) is it possible to learn the non-handcrafted features/models that are needed to estimate the distortion parameters and thus the quality measurements of stereopairs, and (2) can the 3D-vision distortion parameters be estimated without explicitly computing the cyclopean view?

We answer the first question with the assistance of a deep convolution neural network (CNN). Specifically, we propose a four-branch fusion-and-excitation network (FFENet) which predicts four distortion parameters of the multiply and singly distorted 2D images corresponding to four common distortion types: white noise, Gaussian blur, JPEG compression, and JPEG2000 compression. As shown in Fig. 2, the fusion of different network branches represents the joint effect of different distortions, and the excitation operation applied on each branch represents the influence of multiple distortions on the distortion parameter estimation of each individual distortion type. Subsequently, the four distortion parameter values are fed into a multilayer perceptron (MLP) which classifies the distortion into one of three categories. Given the fact that different distortion types may display similar distortion artifacts, and the fact that the same distortion parameter can cause images of different sizes to have different image quality measurement (IQM) scores (as demonstrated in Fig. 3), a QM-oriented method was proposed to adjust/modify the recorded distortion parameter values such that the training labels can be more reasonable and consequently the network can be more effectively trained.

We answer the second question by conducting a subjective experiment to find a mapping function between the 2D and 3D distortion parameter values. The assumption is that in stereoscopic vision, the strength of the perceived 3D distortion depends more on the distortion types/levels of the two views and less on the image content. Specifically, in the test, subjects were presented by stereoscopic images that had been asymmetrically distorted with different distortion types and amounts. The subjects were asked to adjust a slider in the graphical user interface such that the generated 2D distorted image reflected a perceptually equivalent amount of distortion as that perceived in the 3D image (see Fig. 8). As a result, the ground-truth distortion parameters which reflect distortion levels in stereoscopic vision can be obtained, and these data can be used to train machine-learning models to predict 3D-vision distortion parameters without explicitly computing the cyclopean view.

1.3. Proposal and contributions

Based on the aforementioned points, we present in this paper an efficient SIQM method, called *CNN-based stereoscopic image quality estimator* (CNN-SIQE), to blindly measure the qualities of both singly-distorted stereoscopic images (SDSIs) and multiply-distorted stereoscopic images (MDSIs). Our method consists of the three stages as illustrated in Fig. 1. First, an FFENet followed by an MLP is employed to predict the distortion parameters as well as the distortion label of each monocular view. Each distortion parameter is mapped into a quality score, and the four scores are adaptively combined based on the distortion label to produce an overall quality measurement of each monocular view. Second, SVR models are trained to predict the equivalent distortion parameters corresponding to stereoscopic vision, based on which the cyclopean view quality is estimated. In the third stage, the quality measurements of the two monocular views are adaptively combined, followed by the incorporation of the cyclopean view quality, and this yields the final measurement of the stereoscopic image quality.

Compared with existing approaches, our method has several distinctive properties. First, compared with all other NR SIQM models, our work is the first to incorporate results from a subjective test to explicitly model the binocular fusion/rivalry properties of the HVS in stereoscopic vision. Second, compared with all other opinion-unaware SIQM works, our method does not require hand-crafted features, but rather uses a CNN to automatically learn the optimal features. Third, compared with all opinion-aware CNN-based SIQM techniques, our method predicts distortion parameters instead of MOS/DMOS scores such that consistently strong performance can be achieved on both singly- and multiply-distorted images using a consistent set of network parameters. Finally, compared with [24], our method is much faster and can still maintain excellent QM performance. We summarize the main contributions of this work as follows:

(1) We present a deep-learning-based opinion-unaware NR 2D IQM model that operates by predicting distortion parameters instead of human subjective ratings, thus allowing the QM of both multiply distorted and singly distorted images. In addition, we propose a QM-oriented method to allow more effective training.
(2) We present a method to predict the 3D-vision distortion parameters without the need to compute the cyclopean view, thus avoiding the complex computation and potential errors in binocular vision modeling.
(3) We present a new strategy which adaptively combines the quality measurements of the stereopair by considering the impact of different distortion types on the overall 3D image quality.

[1] Here, the equivalent 3D-vision distortion parameters are defined as the distortion parameters that would give rise to the same level of distortion if it were possible to directly distort the merged 3D mental view.

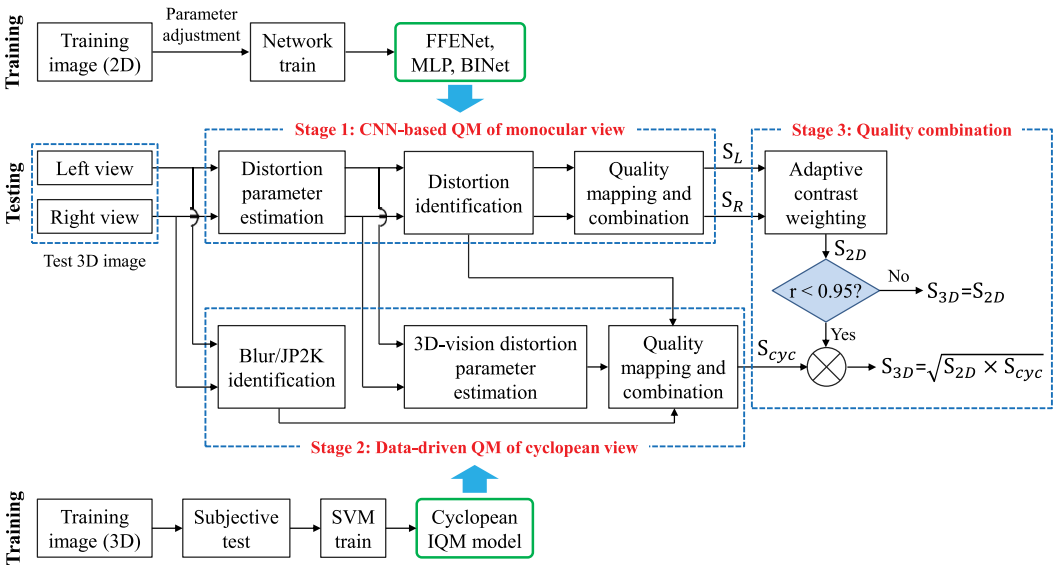

Fig. 1. A block diagram of the proposed CNN-SIQE method. CNN-SIQE first employs FFENet, MLP, and BINet to perform distortion parameter estimation and distortion identification, based on which the qualities of the two monocular views (i.e., S_L and S_R) are measured. Then, in the second stage, the estimated distortion parameters are fed into the data-driven cyclopean model built upon a subjective test for 3D-vision distortion parameter estimation, based on which the cyclopean view quality (i.e., S_{cyc}) is measured. Finally, all quality measurements are adaptively combined to yield the final 3D image quality estimate. Note that MLP is used for WN, WN+GB/JPEG/JP2K, and GB/JPEG/JP2K identification; BINet is used for GB and JP2K identification.

This paper is organized as follows. Section 2 briefly reviews the current NR SIQM approaches. Section 3 describes details of the proposed CNN-SIQE method. Section 4 analyzes the QM performance of CNN-SIQE on various stereoscopic image datasets. General conclusions are provided in Section 5.

2. Related work

Existing techniques for NR SIQM can be roughly classified into two groups, based on whether or not human subjective ratings of stereoscopic images are required. We refer to those techniques that require subjective ratings as *opinion-aware SIQM* approaches. Those techniques that do not require subjective ratings are referred to as *opinion-unaware SIQM* approaches. Here, we briefly review these related works.

2.1. Opinion-aware NR SIQM approaches

As stated in Section 1, most existing SIQM methods that trained regression models or deep neural networks on distorted 3D images with MOS/DMOS values are opinion-aware NR SIQM approaches. These methods can be further classified into two sub-types: (1) those that use hand-crafted features, and (2) those based on deep learning.

The methods based on hand-crafted features operate by mapping quality-aware features to a quality estimate using regression models such as SVR, k-nearest neighbor (KNN), and random forest regression. The features are often extracted based on modeling the natural scene statistics (NSS) or the properties of the HVS in stereoscopic vision. For example, Chen et al. [2] extracted both 2D BRISQUE [16] features and 3D NSS-based features; the 2D features were taken from a synthesized cyclopean image, whereas the 3D features were obtained from the estimated disparity and uncertainty maps. Su et al. [25] extracted wavelet-domain features from a convergent cyclopean image based on bivariate density and correlation NSS models. Shao et al. [26] formulated the stereoscopic quality prediction as a combination of a feature prior and a feature distribution. The feature prior was characterized by using SVR, and the feature distribution was implemented via sparsity regularization. Zhou et al. [4] utilized the complementary local patterns of the binocular energy response and the binocular rivalry response as the quality-aware features, which were mapped to quality scores through KNN-based machine learning. Later, Zhou et al. [6] proposed another NR SIQM method based on a binocular combination and extreme learning machine. The various quality-aware features were extracted from the two binocular combinations of stimuli based on local binary pattern operators. Fang et al. [27] extracted features to train an SVM-based quality model; the features used included statistical intensity, depth, and structure. Liu et al. [28] extracted color/luminance features to quantify the monocular quality perception, and summation/difference features to quantify the binocular quality perception. Moreover, by considering the impacts of viewport, user behavior, and stereoscopic perception on the HVS, Qi et al. [29] proposed a viewport-perception-based blind stereoscopic omnidirectional IQM method using random forest regression.

The deep-network-based methods often train a neural network model that directly maps the stereoscopic image to an associated quality score, in which case the quality-aware features are automatically extracted by the different network layers. Different methods usually have different network architectures/inputs. For example, Shao et al. [3] trained two separate 2D deep neural networks for the monocular and cyclopean images, respectively, and in the testing stage the quality scores predicted by the two networks were combined based on different weighting schemes. Fang et al. [11] designed a Siamese Network to extract the high-level semantic features from stereopairs to simulate the information extraction process in the brain. Then, the features of both views were combined followed by convolutional operations to imitate the information interaction process in the HVS. Chai et al. [30] proposed a monocular/binocular-interaction-oriented three-channel network to estimate the quality of stereoscopic omnidirectional images. Shen et al. [31] proposed a global feature fusion sub-network and local feature enhancement sub-network to extract features from the fused and single views to estimate the visual quality of stereoscopic images. Yang et al. [32] proposed a segmented stacked auto-encoder to model the visual perception route from the eyes to the frontal lobe. Zhou et al. [10] proposed an end-to-end dual-stream interactive network for QM of stereoscopic images. By using a pre-trained model, Sim et al. [13] extracted binocular semantic features and manually-designed binocular quality-aware features to address the problem of limited SIQM dataset size. There are also some other deep-network-based works such as [5,9,33,34].

However, these opinion-aware SIQM techniques suffer from an inevitable limitation in robustness. Due to the diversity of distortion types/levels in different datasets, training on one dataset's images cannot alway guarantee decent QM performance on other datasets. Thus, the limited number of existing 3D image quality datasets significantly restricts the wider application of these kinds of methods.

2.2. Opinion-unaware NR SIQM approaches

To overcome the potential limitation of the aforementioned opinion-aware methods, opinion-unaware methods, which require no training on human subjective ratings, have been proposed. These NR SIQM approaches usually operate by using one of three techniques: (1) collapsing feature values to quality scores based on empirical rules, experimental results, or quality lookup tables; (2) measuring the quality difference between the distorted image and the pristine images in some feature space; and (3) training regression models on a specific measurement that is representative of the image quality.

For example, in [15], the perceptual quality of a stereoscopic image was measured by using the local blurriness, blockiness, and visual saliency information. In [17], the multivariate Gaussian distribution (MGD) was employed to model the features extracted from the distorted and pristine image patches, and the Mahalanobis distance computed between the two sets of MGD parameters was adopted as the quality measurement. In [18], the phase-tuned quality lookup (PTQL) and phase-tuned visual codebook (PTVC) were constructed from the binocular energy responses, and the quality of a test image was obtained by averaging the largest quality values of all image patches, where the PTQL and PTVC were searched to determine the quality of the image patch. In [19], local receptive fields and global receptive fields learned from the reference and distorted stereoscopic images were used to construct local quality and global quality lookup tables. The quality of a test image was obtained by searching for the optimal receptive field indexes in the learned local and global lookup tables. In [20], view-specific feature and quality dictionaries were learned from a category-deviation database such that a semantic framework between the source feature domain and the target quality domain could be established. Then, the stereoscopic image quality was measured based on the classification probability given by the LC-KSVD classification framework. In [21], the modality specific dictionaries and the corresponding projection matrices were learned from a singly-distorted training dataset to predict the quality of MDSIs based on the reconstruction errors. In [22], a multimodal sparse representation framework was established to map the feature space to quality space for the phase and amplitude components. The quality of a MDSI was estimated by a multi-stage pooling scheme using multi-modal quality pooling, feature pooling, binocular pooling, and phase–amplitude quality pooling. In [24], distortion parameters corresponding to the monocular and cyclopean views were estimated by two-layer classification and regression models, from which the qualities of stereoscopic images were measured.

Admittedly, these opinion-unaware methods are effective in releasing SIQM models from the dependence on existing 3D image quality datasets. However, because these methods largely employ handcrafted features as quality indicators, there is still room for improvement (e.g., by using deep learning). In addition, these methods often employ empirical rules/equations to compute the stereoscopic image quality, but to the best of our knowledge, none of them have used psychophysics to investigate what the human visual system actually sees/perceives in binocular vision, particularly when asymmetrically-distorted stereopairs are presented. Thus, how distortion artifacts affect binocular visual quality is still an open question.

In the following section, we describe our new opinion-unaware SIQM method which differs from the existing methods in two ways: (1) instead of using handcrafted features, we take a deep-learning approach for feature extraction; consequently, our method simultaneously offers the advantage of being database-independent and benefits from the power of deep learning; (2) instead of relying on empirical rules, we use a data-driven cyclopean model created based on the results of a subjective test for statistical modeling of the HVS's binocular behavior in the QM task. Owing to the more flexible representation of image quality (i.e., distortion parameters), our model is capable of handling both MDSIs and SDSIs.

3. Algorithm

Our proposed CNN-SIQE method assumes that the QM task of a 3D scene can be achieved by estimating the quality of each monocular view followed by a strategic combination of these per-view qualities. We acknowledge that there is a wide range of distortions that can degrade image quality; however, we follow many of the previous works that only four distortion types and their combinations are considered: white noise (WN), Gaussian blur (GB), JPEG compression (JPEG), and JPEG2000 compression (JP2K), all of which are commonly encountered in daily life. For example, in image acquisition, noise can be introduced due to the different light conditions and imaging device performances; blur can be introduced by motion of the subject and camera shaking/defocus. In image transmission and storage, JPEG/JPEG2000 compression is often required due to the limited bandwidth and capacity of the devices. Moreover, these distortion types were selected, in part, due to ease of modeling: the intensity of each individual distortion can be roughly described by a single parameter. In other words, image quality often displays a high correlation with the corresponding distortion parameters (see Fig. 6 and Table 7). Thus, the aforementioned combination can be achieved in two ways: (1) directly combine quality scores computed from distortion parameter values; and (2) combine distortion parameter values first and then compute the quality score. Accordingly, these two combination strategies give rise to the three main stages of CNN-SIQE as illustrated in Fig. 1: (1) CNN-based QM of the monocular views; (2) data-driven QM of the cyclopean view; and (3) a stage to combine the quality measurements from (1) and (2). Note that the three stages have to be performed sequentially, since the latter stage requires the information derived from the former stage to operate. The details of each stage are provided in the following subsections.

3.1. CNN-based QM of monocular view

3.1.1. Network architecture

As shown in Fig. 2, the proposed FFENet consists of four parallel branches, each of which contains several convolution layers and fully-connected (FC) layers to predict distortion parameters corresponding to the four distortion types. Since we are dealing with multiply-distorted images, the feature maps in different branches are fused (averaged) to model the joint effect of different distortion types in the multiple-distortion scenario. In addition, because of this joint effect, two images with different perceived qualities can share the same distortion parameter value corresponding to certain distortion types. Thus, to model the influence of multiple distortions on distortion parameter estimation of an individual distortion type, we next employ an excitation operation which analyzes the fused feature maps by convolution layers and then concatenates them back to each corresponding branch. This fusion-and-excitation block (FEB) is repeated three times, and the output feature map channels are 128, 256, and 512, respectively. Note that each FEB contains only two groups of the fused features because only two multiple-distortion scenarios (i.e., JP2K+WN and GB+JPEG+WN) were considered.

To enable a scalar output representing different distortion parameter values, the output feature maps of the third FEB in each branch are collapsed into a single vector through average pooling. These vectors are then passed through a number of FC layers in a similar fusion-and-excitation fashion to produce the final distortion parameter estimate.

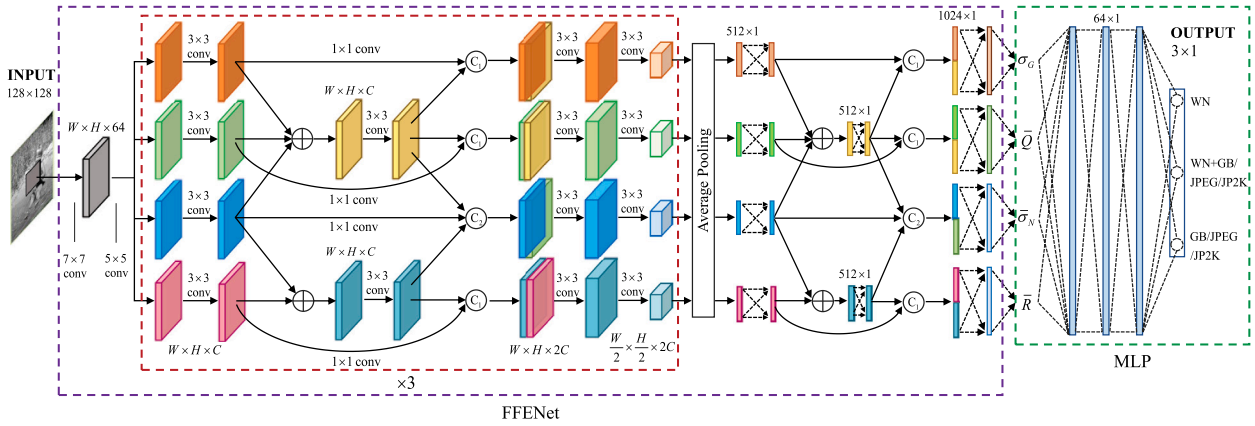

Fig. 2. Network architecture of FFENet and MLP. Note that "C_1" denotes a channel concatenation of two feature maps/vectors; "C_2" denotes concatenation of two $W \times H \times C$ fused feature/vector maps followed by a 1×1 convolution/FC layer to yield another $W \times H \times C$ feature map/vector, which is then concatenated with the third $W \times H \times C$ feature map/vector coming from the corresponding branch. Also note that FFENet predicts σ_G instead of $\bar{\sigma}_G$.

Note that one distortion type may occur in two or more multiple-distortion scenarios (e.g., white noise occurs in both GB+JPEG+WN and JP2K+WN images). For this case, in each FEB, the different fused feature maps corresponding to different multiple-distortion scenarios are first concatenated and then transferred (by a 1×1 convolution layer) to have the same number of feature channels as that in the corresponding branch before being concatenated. The same methodology is applied to the FC layer; here, the convolution operation applied on feature maps is replaced by a linear projection applied on feature vectors. Fig. 2 shows the dimensions of each layer's output assuming an input image patch of 128×128 pixels.

After predicting the four distortion parameters, we classify the image distortion into one of three categories via a MLP: (1) WN only, (2) WN + GB/JPEG/JP2K, and (3) GB/JPEG/JP2K. As shown in Fig. 2, the MLP consists of one input layer, three hidden layers, and one output layer. The input layer contains four units corresponding to the four distortion parameter values. Each hidden layer contains 64 units, each of which is connected to the previous layer through a dot product between the input vector and its weight vector, and then with the addition of a bias. The output layer computes the weighted sum (denoted by z_k for unit k) of the 64 units in the previous hidden layer, and the state of unit k (denoted by σ_k) is computed via a softmax function given by

$$\sigma_k = \frac{\exp(z_k)}{\sum_{k=1}^{m} \exp(z_k)}, k = 1, 2, \ldots, m \tag{1}$$

where $m = 3$ denotes the three categories. Consequently, the MLP has three output values, each of which represents the probability that the input belongs to each of the three classes. As we will show later, the classification label predicted by the MLP will be used for QM of both the monocular views and the cyclopean view.

3.1.2. Training data generation

To train the FFENet and MLP, training data has to be generated. To this end, the same approach as in [24] was adopted to add single/multiple distortions to 270 pristine 2D images, among which 150 images were collected from the Berkeley segmentation database [35], 45 from the Waterloo exploration database [36], 45 from the left view of the stereoscopic images adopted in [37], and 30 were taken by using our own camera. While generating the distorted versions of these images, the following four distortion parameters were recorded: (1) the standard deviation σ_G of the Gaussian blurring filter, (2) the compression quality factor Q for the JPEG compression, (3) the compression ratio R of the JPEG2000 compression, and (4) the variance σ_N of the white noise. To model these parameters more effectively, a logarithm is applied as follows:

$$\bar{\sigma}_G = \ln\left(1 + \sigma_G\right) \tag{2}$$

Table 1
Detailed information of the generated training dataset. Note that L_1 denotes the ground-truth classification labels assigned to each distorted patch for training MLP, and L_2 denotes the ground-truth classification labels for training BINet.

Distortion type	Distortion level	# Image	# Patch	L_1	L_2
GB	10	2700	30 930	2	0
JPEG	10	2700	30 930	2	0
JP2K	10	2700	30 930	2	1
WN	10	2700	30 930	0	0
JP2K+WN	30	1350	25 920	1	1
GB+JPEG+WN	294	13 230	111 720	1	0

$$\bar{Q} = \ln\left(1 + 80 \cdot (Q/80)^{1.5}\right) \tag{3}$$

$$\bar{R} = \ln\left(1 + 10^3 \cdot \left(R/10^3\right)^2\right) \tag{4}$$

$$\bar{\sigma}_N = \ln\left(1 + 10^3 \cdot \sigma_N\right). \tag{5}$$

Notice that the logarithm is used to bring the four distortion parameter values to approximately the same range, as well as to allow a more linear correlation between the distortion parameter change and the image quality variation. Also notice that FFENet predicts σ_G instead of $\bar{\sigma}_G$, the latter of which is mainly used for polynomial curve fitting as shown in Fig. 6. For multiple distortions, we followed [38] to first perform Gaussian blurring, then perform JPEG/JPEG2000 compression, and then add the white noise. After generating the multiply and singly distorted images, non-overlapping 128×128-pixel patches were extracted. Consequently, we extracted in total 261,360 image patches along with their distortion parameters (the same as those of the distorted images) and labels (denoted by L_1 and L_2) as the training data. The details of these data are summarized in Table 1, and are available at https://vinelab.jp/cnnsiqe/.

It is important to note that two problems occur when determining the ground-truth distortion parameters of an image. First, different from [24], the blur parameter for a JPEG2000-compressed image cannot be zero, and likewise the JPEG2000 parameter for a Gaussian blur image cannot be one. The reason is that both Gaussian blur and JPEG2000-compressed images will display blurring artifacts. In [24], each regression model was trained to predict only one distortion parameter on the corresponding distorted images. However, in this work, the FFENet is trained to predict four distortion parameters at the same time on all distorted patches. Since it is the manifested/displayed distortion artifact that determines the network output, a network can easily get confused if it were forced to output different values for images with the same/similar amount of distortion artifact. In other words, it is inappropriate to say that a JPEG2000-compressed image is not blurred. Thus, equivalent parameter values have to be calculated for

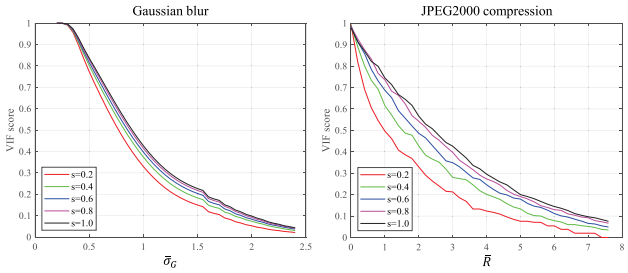

Fig. 3. VIF scores computed for Gaussian blur and JPEG2000-compressed images generated from the different scales of the pristine image shown in Fig. 5(a). The horizontal and vertical axes denote, respectively, the distortion parameter values, and the corresponding VIF scores. "$s = x$" denotes that the rescaled image is x times of its original size; "$s = 1.0$" denotes the original-size image.

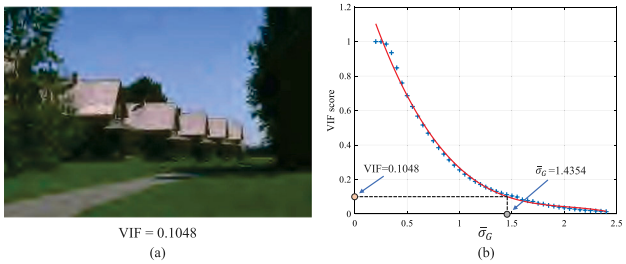

Fig. 4. An illustration of computing the equivalent blur parameter for a JPEG2000-compressed image.

the two distortion types such that the network can be trained more effectively. Second, the same distortion parameter may give rise to different quality estimates for an image resized to different scaling factors (as demonstrated in Fig. 3), and thus it may be difficult to draw an accurate monotonic mapping function which maps distortion parameter values to quality scores. Thus, the training data have to be made that the same distortion parameter value produces images with approximately the same objective IQM scores.

To solve these two problems, in this paper we propose a QM-oriented method which computes equivalent distortion parameters based on objective IQM methods. Specifically, for the first problem, an equivalent distortion parameter implies that its corresponding distorted image shares the same IQM score as the image under question. For example, to compute the equivalent blur parameter of a JPEG2000-compressed image in Fig. 4(a), we first compute the VIF [39] scores for the same-content Gaussian blurred images generated with various blur parameters. As shown in Fig. 4(b), these data can be fitted by a polynomial curve from which the polynomial coefficients can be calculated. Then, we compute the VIF score of the JPEG2000-compressed image, and find a point on the curve that has the same/closest VIF score value. The corresponding x-axis value of that point is the estimated equivalent blur parameter.

For solving the second problem, a similar QM-oriented method is adopted. For example, as shown in Fig. 5, given a JPEG2000-compressed image I_1 generated from the pristine image I with $\bar{R} = 6.266$, the equivalent JPEG2000 parameter can be calculated as follows. First, we rescale I_1 to I_2 [Fig. 5(e)] and compute its VIF score by referring to the similarly resized reference image I_0 [Fig. 5(c)] obtained from I. In this paper, bicubic interpolation is used to rescale an image until the minimum of the height and width is 512 pixels. Next, a number of JPEG2000-compressed images are generated from I_0 with different JPEG2000 parameters $\bar{R}_i$ and meanwhile the corresponding VIF scores are saved. These data can be fitted by a polynomial curve [Fig. 5(d)] from which the polynomial coefficients can be calculated. Finally, the point on the curve with the same/closest VIF score is found, and its corresponding x-axis value is the computed equivalent

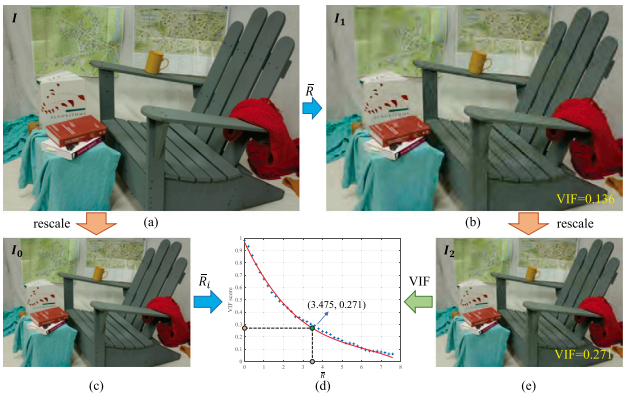

Fig. 5. An illustration of computing the equivalent JPEG2000 parameter for a large-sized JPEG2000-compressed image.

JPEG2000 parameter for I_2. Note that image patches are also extracted from I_2 for training, not from I_1.

3.1.3. Quality mapping and combination

Armed with the trained FFENet and MLP, we can now possibly predict the distortion parameters and use those parameters to measure the quality of each monocular view. For distortion parameter estimation, we first extract the Y-channel (luminance) of the test RGB image and divide it into 128 × 128-pixel patches with 64-pixel overlap. Then, these patches are fed into the FFENet to obtain their four distortion parameters. Since our goal is to predict distortion parameters of the whole image, a pooling operation is applied to collapse all parameter values into a scalar. Motivated by [40], we use image patches with sharp edges/textures for pooling, and the local standard deviation (LSD) is employed to measure the local sharpness which is given by

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K}\sum_{l=-L}^{L} \omega_{k,l}(I_{k,l}(i,j) - \mu(i,j))^2}, \tag{6}$$

where $I(i,j)$ is the (grayscale) image; $\omega_{k,l}$ is a circularly-symmetric 2D Gaussian weighting function with a standard deviation of 1.5 and rescaled to unit volume; $I_{k,l}(i,j) = I(i+k, j+l)$ denotes the local image pixel value; $\mu(i,j)$ denotes the weighted sum of $I_{k,l}(i,j)$ computed by

$$\mu(i,j) = \sum_{k=-K}^{K}\sum_{l=-L}^{L} \omega_{k,l} I_{k,l}(i,j), \tag{7}$$

where we set $K = L = 5$ by following [41].

In this paper, image patches with the top 25% largest LSD values were selected. Let ξ_m^p and ζ_n^p ($m = 1,2,3,4$; $n = 0,1,2$) denote, respectively, the four distortion parameters predicted by the FFENet and the three probabilities predicted by MLP for each patch p. The distortion parameters and the classification label of the overall image are computed by

$$\xi_m = \frac{1}{N_p}\sum_{p=1}^{N_p} \xi_m^p \tag{8}$$

$$l_1 = \arg\max_n \left(\frac{1}{N_p}\sum_{p=1}^{N_p} \zeta_n^p \right), \tag{9}$$

where N_p denotes the total number of selected patches.

For quality evaluation, the VIF [39] algorithm is first employed to measure the qualities of the generated distorted training images. Then for each distortion type, we convert distortion parameters into quality scores based on modeling the shape of scatter plots of the distortion parameter values versus VIF quality scores (see Fig. 6). As shown in

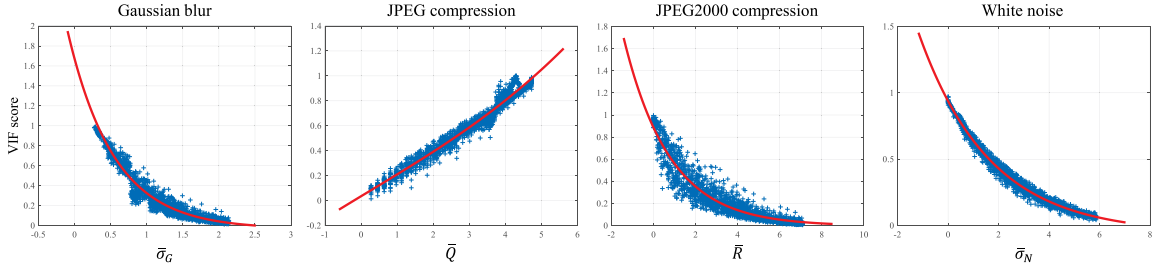

Fig. 6. Scatter plots showing how VIF quality scores change for different distortion parameter values. Observe that different trends emerge for different distortion types. For each scatter plot, the horizontal and vertical axes represent, respectively, the distortion parameter value, and the corresponding VIF quality score computed from the distorted image regenerated by using that parameter value.

Table 2
Values of polynomial and Laplacian distribution parameters for fitting the four curves in Fig. 6.

	λ_1	λ_2	λ_3	λ_4
GB	1.0117	3.3485×10^{-1}	6.4174×10^{-1}	-3.4143×10^{-2}
JPEG	1.3007×10^{-3}	-8.5939×10^{-4}	1.7473×10^{-1}	3.5380×10^{-2}
JP2K	1.2269	-7.0408×10^{-1}	2.1899	-3.7466×10^{-3}
WN	1.0006	-1.1771×10^{-2}	2.8117	-5.8794×10^{-2}

Table 3
Values of polynomial parameters for fitting the four curves in Fig. 7.

	λ_1	λ_2	λ_3	λ_4
GB	1.1920	−1.2030	8.6122×10^{-1}	-4.6376×10^{-2}
JPEG	1.9831	−2.6949	1.7567	-1.8692×10^{-1}
JP2K	1.4828	−1.5964	1.0707	-7.5016×10^{-2}
WN	9.6335×10^{-1}	-8.4899×10^{-1}	7.9886×10^{-1}	-3.0893×10^{-2}

Fig. 6, the scatter plot shape of the JPEG compression can be modeled by a polynomial curve defined as

$$y = \lambda_1 \cdot x^3 + \lambda_2 \cdot x^2 + \lambda_3 \cdot x + \lambda_4. \tag{10}$$

For the other three distortion types, their scatter plot shapes are modeled by a single-sided Laplacian distribution curve defined as

$$y = \lambda_1 \cdot e^{-\frac{(x-\lambda_2)}{\lambda_3}} + \lambda_4. \tag{11}$$

For both equations, λ_i $(i = 1, 2, 3, 4)$ denotes the curve parameters, and their fitted values for each distortion type are provided in Table 2.

As mentioned previously, training images with large sizes have to be rescaled in order that the same distortion parameter value will produce distorted images with approximately the same VIF quality score. Likewise, the training patches are extracted from the rescaled image with the equivalent distortion parameters recorded for supervised learning. Accordingly, a large-sized test image also needs to be rescaled in the testing stage. As shown in Fig. 5(b)(e), when an image is rescaled, the VIF score also changes. In other words, the VIF scores mapped by Eqs. (10) and (11) represent the qualities of the rescaled image, if the image is of a large size. Thus, for a large-sized input, the mapped VIF scores have to be modified before the quality-combination stage.

Admittedly, the VIF score relationship between a resized distorted image and its original version largely depends on the scaling factor. To simplify the problem, we only consider a limited number of the scaling factors in this work. Specifically, for images with large sizes in the training dataset, the VIF quality scores are computed for both the original and rescaled versions. Accordingly, the scatter plots of VIF scores corresponding to each distortion type are shown in Fig. 7, wherein the x and y axes denote the VIF scores calculated for the rescaled and the original-sized images, respectively. Again, each scatter plot can be modeled by the polynomial curve defined in Eq. (10), and the fitted parameter values are provided in Table 3.

After mapping distortion parameters to corresponding VIF scores, we combine these scores to generate an overall quality measurement. For this combination, we employ the most-apparent-distortion strategy [24]. Specifically, let VIF_G, VIF_Q, VIF_R, and VIF_N denote the mapped qualities for Gaussian blur, JPEG compression, JPEG2000 compression, and white noise, respectively. Then, the quality degradation (denoted by D_G, D_Q, D_R, and D_N, respectively) for each of the four distortion types is computed via

$$\mathrm{D}_G = 1 - \mathrm{VIF}_G \tag{12}$$

$$\mathrm{D}_Q = 1 - \mathrm{VIF}_Q \tag{13}$$

$$\mathrm{D}_R = 1 - \mathrm{VIF}_R \tag{14}$$

$$\mathrm{D}_N = 1 - \mathrm{VIF}_N. \tag{15}$$

Notice that the estimated distortion parameter values are clipped to certain ranges.[2] before applying Eqs. (10) and (11) to ensure that the mapped VIF scores are reasonable. The final quality measurement of each monocular view (denoted by S_L and S_R for the left and right views, respectively) is then given by

$$\mathrm{S}_{L(R)} = \begin{cases} \mathrm{D}_N, & l_1 = 0 \\ \mathrm{D}_1^{\gamma} \times \mathrm{D}_2^{1-\gamma}, & l_1 = 1 \\ \max\left(\mathrm{D}_{GR}, \mathrm{D}_Q\right) \times \rho^{\min\left(\mathrm{D}_{GR}, \mathrm{D}_Q\right)}, & l_1 = 2 \end{cases} \tag{16}$$

where l_1 denotes the classification labels predicted by the MLP; D_{GR} denotes the average value of D_G and D_R; and D_1 and D_2 represent the two quality measurements computed under different noise conditions. Specifically, D_1 is computed assuming that images are corrupted by a small amount of noise, and D_2 is computed assuming that the noise corruption is severe in which case the blurring/compression artifacts are to some extent visually masked. Accordingly, we compute the two quality measurements by:

$$\mathrm{D}_1 = d_1 \times \rho^{d_2} \tag{17}$$

$$\mathrm{D}_2 = \max\left(\mathrm{D}_{GRQ} - \beta, \mathrm{D}_N\right) \times \rho^{\min\left(\mathrm{D}_{GRQ} - \beta, \mathrm{D}_N\right)} \tag{18}$$

where d_1 and d_2 denote the first and second largest values of D_{GR}, D_Q, and D_N; and D_{GRQ} denotes the larger value of D_{GR} and D_Q. The parameter β attempts to account for the visual masking caused by the noise; the parameter ρ attempts to take into account the impact of the second-most-apparent distortion; and the parameter γ, which controls the influence of D_1 vs. D_2, is determined by $\bar{\sigma}_N$ as follows:

$$\gamma = A/[1 + e^{t_1(\bar{\sigma}_N - t_2)}] + B. \tag{19}$$

We set the following parameter values: $\rho = 1.2$, $\beta = 0.3$, $A = 1.5$, $B = 0$, $t_1 = 1.5$, and $t_2 = 0.5$ such that the best performance can be achieved across different 3D image datasets.

[2] In this paper, we set $\bar{\sigma}_G = \max\left(0.4, \ln\left[1 + \max\left(0, \sigma_G\right)\right]\right)$, $\bar{Q} = \min\left(4.5, \bar{Q}\right)$, $\bar{R} = \max\left(0.001, \bar{R}\right)$, and $\bar{\sigma}_N = \max\left(0, \bar{\sigma}_N\right)$ according to Fig. 6.

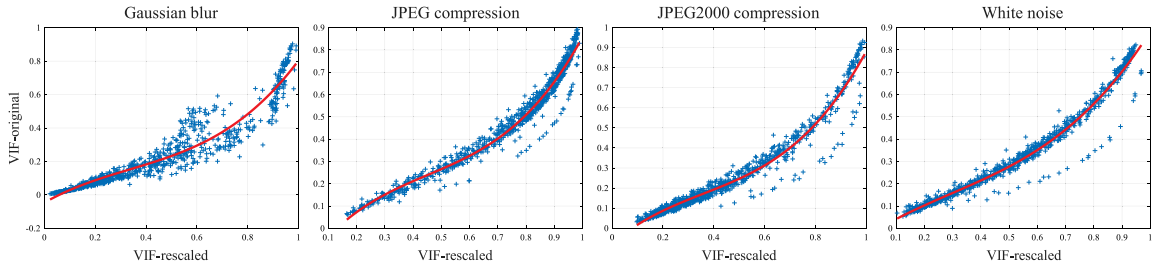

Fig. 7. Scatter plots of VIF score values computed for the original versus the rescaled distorted images. For each scatter plot, the horizontal and vertical axes represent VIF scores computed for the rescaled and the original-sized distorted images, respectively.

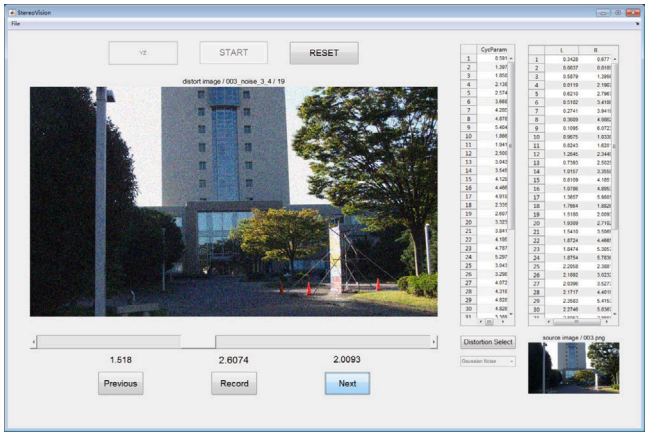

Fig. 8. A GUI for conducting the human subjective experiment. The left table on the interface shows the selected 3D-vision distortion parameters, and the right table shows the ground-truth distortion parameters of the left and right views.

3.2. Data-driven QM of cyclopean view

As mentioned previously, humans judge the quality of a stereoscopic image based, in part, on the merged 3D mental view formed after the binocular fusion and rivalry. In our previous work [24], we attempted to model this binocular visual behavior via a complicated multi-pathway contrast gain-control model used to compute the cyclopean view, whose distortion parameters were then estimated based on training SVM models using only symmetrically-distorted 3D images. Due to the potential errors introduced in the estimated disparity map, the inaccuracy of the MCM, and the difference between the training and testing data, the cyclopean model proposed in [24] can perform poorly at times. Hence in this work, we propose a new data-driven method which directly predicts 3D-vision distortion parameters without explicitly computing the cyclopean view.

3.2.1. 3D-vision distortion parameter estimation

Our method to predict the 3D-vision distortion parameters is based on the assumption that in stereoscopic vision, the strength of the perceived 3D distortion depends more on the distortion types and the distortion levels of the two monocular views than it does on the image contents, especially when images are asymmetrically distorted. Asymmetrically distorted images give rise to discomfort due to binocular rivalry, and this discomfort plays a much bigger role than image-content-based effects (e.g., masking) in determining the perceived quality. Thus, a regression model can be trained for the specific distortion type which maps the two distortion levels of the two views to a single parameter representing the 3D-vision distortion level. To obtain the training data, we designed a psychophysical experiment, in which subjects were presented with asymmetrically-distorted stereoscopic images and were asked to choose a value by which a 2D image generated to have the same image content was perceived to contain a visually equivalent amount of distortion as that being perceived in the 3D view. For each distortion type, we selected eight distortion levels, and assumed that the right view had an equal/better quality than the left view. Consequently, for each distortion type of each pristine 3D image, a total of 45 asymmetrically-distorted versions were generated.

The GUI for the experiment is shown in Fig. 8. Moving the slider will generate images with different distortion levels. The "Previous" and "Next" buttons allow the subject to double-check the results. When the "Record" button is clicked, the results are saved. Note that the ground-truth distortion parameters of the left and right views are always presented to the subject during the test. This information is necessary because, when images are less distorted, it becomes difficult for subjects to perceive the minor distortion level change, thus giving rise to inconsistent scores.

The experiment was conducted on distorted versions of nine pristine 3D images captured by a FUJIFILM 3D digital camera. To simplify the experiment, we used only asymmetrically-distorted images, and by default the left view was always distorted to a greater extent than the right view. Ten different distortion levels were randomly applied to the two monocular views, which gave rise to 45 distorted versions for each pristine 3D image. The distortion intensity ranged from just noticeable to severely visibly degraded (in which case the image's structures/contents were severely destroyed) in an effort to make the resulting model robust to a wide range of distortion severities. Consequently, 405 ground-truth 3D-vision parameters were generated for each of the three distortions: Gaussian blur, JPEG2000 compression, and white noise. Note that the experiment was not conducted for JPEG-compressed images because, as we have found, blocking artifacts always dominate the overall 3D quality; thus, for an asymmetrically JPEG-compressed image, the 3D-vision JPEG parameter is simply computed by

$$\bar{Q}_{cyc} = \min\left(\bar{Q}_L, \bar{Q}_R\right) + \ln\left(1 + \left|\bar{Q}_L - \bar{Q}_R\right|\right), \tag{20}$$

where $\bar{Q}_L$ and $\bar{Q}_R$ denote the estimated JPEG parameters for the left and right views, respectively.

During the test, both 2D and 3D images were displayed on a single 24-inch 3D monitor (ASUS VG248QE), and NVIDIA 3D vision-2 wireless glasses were used for stereoscopic viewing. The subject was allowed to pause/stop/resume the test as needed without losing the results, in an effort to reduce potential fatigue. The obtained 405 training data[3] were then used to train an SVR model based on the ϵ-SVR [42] approach.

Specifically, given the training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$ where $\mathbf{x}_l \in \mathbb{R}^2$ and $y \in \mathbb{R}$, ϵ-SVR aims to find a function $f(\mathbf{x})$ that has at most ϵ deviation from the desired target y. Usually $f(\mathbf{x})$ takes the following form

$$f(\mathbf{x}) = \langle w, \phi(\mathbf{x}) \rangle + b, \tag{21}$$

[3] Available at https://vinelab.jp/cnnsiqe/. Each training data sample actually contains three values corresponding to the distortion parameters of the left, right, and cyclopean views, respectively. The former two values are the features, and the third value is the label.

where $\phi(\mathbf{x})$ is a non-linear projection which maps the data to the high-dimensional feature space; w is the weight vector, and b is the bias. Computing $f(\mathbf{x})$ is equivalent to solving the following optimization problem:

$$\min_{w,b,\xi,\xi^*} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right)$$
$$\text{subject to} \quad \begin{cases} y_i - \langle w, \phi(\mathbf{x}_i)\rangle - b & \leq \epsilon + \xi_i \\ \langle w, \phi(\mathbf{x}_i)\rangle + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \tag{22}$$

where ξ_i and ξ_i^* are the slack variables; $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ϵ are tolerated. By introducing a dual set of variables to construct the Lagrange function from both the objective function and the corresponding constraints, the solution of Eq. (22) can be finally obtained:

$$f(\mathbf{x}) = \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)\cdot K\left(\mathbf{x}_i, \mathbf{x}\right) + b, \tag{23}$$

where $K(\mathbf{x_i}, \mathbf{x}) = \phi(\mathbf{x_i})^T\phi(\mathbf{x})$ is a kernel function; α_i and α_i^* are the Lagrange multipliers; and b is computed by

$$\begin{cases} b = y_j - \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)\cdot K\left(\mathbf{x}_i, \mathbf{x}_j\right) - \epsilon & for\ \alpha_j \in (0, C) \\ b = y_j - \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)\cdot K\left(\mathbf{x}_i, \mathbf{x}_j\right) + \epsilon & for\ \alpha_j^* \in (0, C) \end{cases} \tag{24}$$

Consequently, for each of the three distortion types, we obtain an SVR model, which will be later used to predict the corresponding 3D-vision distortion parameter which is given by

$$u_m = \mathrm{SVR}_m\left[\mathbf{v}_m\right] = \sum_{i=1}^{l_m}\left(\alpha_{m_i} - \alpha_{m_i}^*\right)\cdot K_m\left(\mathbf{x_{m_i}}, \mathbf{v}_m\right) + b_m, \tag{25}$$

where $\mathbf{v}_m \in \mathbb{R}^2$ denotes the parameter values estimated for the two monocular views for each distortion type; and $m = 1, 2, 3$ denotes the three SVR models.

3.2.2. Quality measurement of cyclopean view

With the predicted 3D-vision distortion parameters, it is now possible to measure the quality of the cyclopean view. Although the same most-apparent-distortion strategy can be applied, an important issue raised by 3D vision has to be addressed. In Section 3.1.3, D_{GR} is set as the average of D_G and D_R, which means that the blurring artifact introduced by either Gaussian blur distortion or JPEG2000 compression will indiscriminately cause quality degradation. However, this assumption does not hold for the 3D-vision case, because the two distortion types, if applied only to a single monocular view, will have different impacts on the overall 3D image quality. For example, consider two asymmetrically-distorted 3D images both of which have the same high-quality left views, but the right view of one image is corrupted by Gaussian blur and the other by JPEG2000 compression. According to Section 3.1.2, if the two right views have the same or similar blur and JPEG2000 parameters, their 3D-vision parameters should also be the same or similar. However, in reality, the perceived quality of the JPEG2000-compressed image is much lower than that of the Gaussian blur image, because the blurring can be suppressed by the high-quality view while ringing artifacts cannot due to the additive information introduced to the image content.

Based on the above observations, it seems to be necessary to first identify the distortion type leading to the perception of blur in the cyclopean view such that a more appropriate strategy of modeling the 3D-vision quality-judging behavior of the HVS can be adopted. To this end, we propose a blur identification network (BINet), whose architecture is shown in Table 4. The network generally follows a similar pipeline as is used in VGG [43] in which an image is first passed through a stack of convolution (Conv) layers followed by a max-pooling (MaxPool) layer for feature extraction. Then, the feature maps are flattened after an adaptive average-pooling (AdaptAvgPool) operation, and the obtained feature vectors are processed by several fully-connected (FC) layers and a soft-max layer for classification. All convolution layers use batch normalization (BN) and a rectified linear unit (ReLU) for the nonlinearity. Apart from the first and second convolution layers that contain filters of size 7×7 and 5×5 pixels for large receptive fields, all other convolution layers use 3×3-pixel size filters. For all convolution operations, we use one pixel stride, and the padding is selected such that the spatial resolution is preserved after convolution. The max-pooling operation is performed over 2×2-pixel size windows with a stride of 2, and the adaptive average-pooling layer pools any feature map to 4×4 pixels. Given an image patch of 128×128-pixel size, the dimensions of the output feature map of each layer are shown in Table 4. The network is trained on the same 261,360 image patches as in Section 3.1.2, and their corresponding ground-truth labels are shown in Table 1. As we will demonstrate in Section 4.6, the use of BINet can significantly improve the QM performance of the algorithm especially when asymmetrically-distorted images are presented.

Table 4
An architecture of the proposed blur identification network.

Layer	Kernel size	Stride	Padding	Output size
Conv+BN+ReLU	7×7	1	3	$128 \times 128 \times 64$
Conv+BN+ReLU	5×5	1	2	$128 \times 128 \times 64$
MaxPool	–	–	–	$64 \times 64 \times 64$
Conv+BN+ReLU	3×3	1	1	$64 \times 64 \times 64$
Conv+BN+ReLU	3×3	1	1	$64 \times 64 \times 64$
MaxPool	–	–	–	$32 \times 32 \times 64$
Conv+BN+ReLU	3×3	1	1	$32 \times 32 \times 64$
Conv+BN+ReLU	3×3	1	1	$32 \times 32 \times 64$
MaxPool	–	–	–	$16 \times 16 \times 64$
AdaptAvgPool+Flatten	–	–	–	1024×1
FC+ReLU	–	–	–	64×1
FC+ReLU	–	–	–	64×1
FC	–	–	–	2×1
Soft-max	–	–	–	2×1

Using the BINet to predict the label of an overall image follows the same procedure as that required when using the MLP. Accordingly, we use the same approach as in Section 3.1.3 to evaluate the overall quality of the cyclopean view (denoted by S_{cyc}), in which case all variables in Eq. (16) are computed based upon 3D-vision distortion parameters. The only difference is the computation of D_{GR}, which is given by

$$\mathrm{D}_{GR} = \begin{cases} \mathrm{D}_G, & l_2 = 0 \\ \mathrm{D}_R, & l_2 = 1 \end{cases} \tag{26}$$

where l_2 denotes the classification label predicted by the BINet. Since there are two l_2 labels corresponding to the two monocular views, the label of the lower-quality view is used as the label of the cyclopean view. The same principle is also applied to l_1 in Eq. (16). We set $\beta = 0.15$ in Eq. (18) to account for the decreased masking effect in 3D vision. Also, we set $\mathrm{D_{GR}} = \mathrm{D_{GR}} - 0.1$ in Eq. (16) such that the inaccuracy of the VIF algorithm is taken into account (see [44] for more details). For other parameters, the same values as stated in Section 3.1.3 are adopted.

3.3. CNN-SIQE quality measurement

Given the three quality measurements S_L, S_R, and S_{cyc} for the left, right, and cyclopean views respectively, the final stage of CNN-SIQE combines them into an overall quality measurement of the stereoscopic image. First, we combine S_L and S_R by using an adaptive contrast-weighting strategy. Specifically, we use a similar approach in [24] to

compute image contrast which is given by

$$\bar{C}_{L/R} = \left[\frac{1}{N_b}\sum_{b=1}^{N_b} C_{L/R}(b)\right]^{\frac{3}{2}} \cdot \left[\frac{1}{N_p}\sum_{p=1}^{N_p} \mathrm{F}_{L/R}(p)\right], \tag{27}$$

where $C_{L/R}(b)$ denotes the RMS contrast computed for block b (the total number of blocks is N_b) in the left/right view by using the approach in [45]; $\mathrm{F}_{L/R}(p)$ denotes the sharpness value computed by applying the FISH_{bb} algorithm to image patch p in the left/right view (N_p is the total number of patches). Note that when sharpness is computed, the same image patches as described in Section 3.1.3 are selected.

Although contrast is an effective measurement to indicate the weight of the dominant view when images are corrupted by noise/blur, it does not work for JPEG/JPEG2000 compression. The reason is that, given a 3D image with one view undistorted and the other view JPEG/JPEG2000-compressed, the compressed view always dominates the perception of the overall quality of the image (due to the uncompensated blocking/ringing artifact), but its contrast is always similar to that of the undistorted view. To emphasize this importance of the JPEG/JPEG2000-compressed view in the QM process, we suppress the weight of the other view via a sigmoid function previously defined in Eq. (19).

Specifically, let $w_{L/R}$ denote the weight of the left/right view; let $l_1^{L/R}$ and $l_2^{L/R}$ denote the classification labels predicted by the MLP and the BINet for the left/right views; and let $\mathrm{D}_{jpg}^{L/R}$ and $\mathrm{D}_{j2k}^{L/R}$ denote the quality degradations computed from the JPEG/JPEG2000 parameters via Eqs. (13)–(14) for the left/right views. If $S_L > S_R$, then $w_L = \bar{C}_L$, and the weight of the right view is computed by

$$w_R = \bar{C}_R \cdot C_1^R \cdot C_2^R, \tag{28}$$

where C_1^R and C_2^R denote two regularization factors designed for JPEG/JPEG2000 compression and are given by

$$\mathrm{C}_1^R = \begin{cases} A_1/[1+e^{t_1\left(\mathrm{D}_{jpg}^L - t\right)}] + (1-A_1), & l_1^L = 2, l_2^L = 0 \\ 1, & \text{otherwise} \end{cases} \tag{29}$$

$$\mathrm{C}_2^R = \begin{cases} A_2/[1+e^{t_2\left(\mathrm{D}_{j2k}^L - t\right)}] + (1-A_2), & l_2^L = 1 \\ 1, & \text{otherwise.} \end{cases} \tag{30}$$

Similarly, if $S_L < S_R$, then $w_R = \bar{C}_R$, and the weight of the left view is computed by

$$w_L = \bar{C}_L \cdot C_1^L \cdot C_2^L, \tag{31}$$

where

$$\mathrm{C}_1^L = \begin{cases} A_1/[1+e^{t_1\left(\mathrm{D}_{jpg}^R - t\right)}] + (1-A_1), & l_1^R = 2, l_2^R = 0 \\ 1, & \text{otherwise} \end{cases} \tag{32}$$

$$\mathrm{C}_2^L = \begin{cases} A_2/[1+e^{t_2\left(\mathrm{D}_{j2k}^R - t\right)}] + (1-A_2), & l_2^R = 1 \\ 1, & \text{otherwise.} \end{cases} \tag{33}$$

Note that the joint condition of $l_1 = 2$ and $l_2 = 0$ indicates that images are corrupted by blur and/or JPEG-compression distortion; $l_2 = 1$ indicates that images are JPEG2000-compressed. Let S_{2D} denote the combined quality measurement obtained from the two monocular views; we compute S_{2D} via

$$\mathrm{S}_{2D} = \frac{w_L \cdot S_L + w_R \cdot S_R}{w_L + w_R}, \tag{34}$$

In this paper, we empirically set $A_1 = A_2 = 1, t_1 = -30, t_2 = -15, t = 0.3$ to help achieve the best performance.

The next step is to combine S_{2D} with S_{cyc}. Note that the subjective experiment described in Section 3.2.1 was conducted to investigate the binocular rivalry properties of the HVS in perceiving mainly asymmetrically-distorted stereoscopic images. Thus, we combine S_{2D} with S_{cyc} (which is computed based upon the results of the subjective experiment) only when the distortion levels of the two views are different. In this paper, the distortion similarity between the two views is roughly estimated from their quality similarity which is given by

$$r = \left(\frac{2\mathrm{S}_L\mathrm{S}_R}{\mathrm{S}_L^2 + \mathrm{S}_R^2}\right)^2. \tag{35}$$

Consequently, the overall quality measurement of the stereoscopic image, which we denote by S_{3D}, is computed by

$$\mathrm{S}_{3D} = \begin{cases} S_{2D}, & r > T \\ \sqrt{\mathrm{S}_{2D} \times \mathrm{S}_{cyc}}, & r \le T \end{cases} \tag{36}$$

where $T = 0.95$ is a threshold to indicate symmetric or asymmetric distortion. Smaller values of S_{3D} indicate greater stereoscopic image quality.

4. Results

In this section, we examine and discuss the ability of CNN-SIQE to predict the quality ratings from various 3D image datasets. We also compare the performance of CNN-SIQE with existing SIQM approaches.

4.1. Training

Our CNN-SIQE model contains three neural networks and three SVR models that need training. The three neural networks are FFENet, MLP, and BINet; and the three SVR models are used to predict the distortion parameters of the Gaussian blur, JPEG2000 compression, and white noise, respectively. As mentioned in Sections 3.1.2 and 3.2.2, all three networks were trained on the same 261,360 image patches corrupted by four distortion types and their two combinations as shown in Table 1. To train the three SVR models, 405 training data were collected for each of the three distortion types in the subjective experiment as mentioned in Section 3.2.1. Training of the three neural networks was conducted on a remote server with four NVIDIA GeForce RTX 3090 GPUs; all other experiments, including the SVR model training and database testing, were conducted on a local workstation with an i9-9900K CPU (8-core, 3.6 GHz) and a GeForce RTX 2080 SUPER GPU.

Specifically, the three networks were trained by using the Adam [46] optimizer (0.9 and 0.999 were used for the decay rates for the first/second moment estimates, respectively; the default settings were used for all other hyperparameters). The network parameters were initialized with values sampled from a normal distribution $N(1, 0.02)$. The initial learning rate was set to 2×10^{-4}, and was scaled down by a factor of 0.9 after each epoch until 10^{-6}. The batch size was set to 100 for training FFENet, 128 for training MLP, and 16 for training BINet. Note that MLP takes as input the output of FFENet; thus, we first trained FFENet, and then MLP by freezing the FFENet parameters. Also note that when training BINet, image patches were randomly and equally selected from the four "$L_2 = 0$" sets (shown in Table 1), and the patch numbers of the two classes were also roughly the same to prevent model bias caused by imbalanced classes. Thus, the training data corresponding to the "$L_2 = 0$" label were changed dynamically for every epoch. The L1 loss was used to train FFENet, and the cross-entropy loss was used to train MLP and BINet. Consequently, it took about five days to train FFENet for 90 epochs, 11 h to train BINet for 140 epochs, and 12 h to train MLP for 18 epochs before convergence.

For training the three SVR models, the Python version of LIBSVM [47] was employed, and the radial basis function kernel was used as the kernel function. More details about training SVR models can be found in [47,48].

Table 5
Overall performances of CNN-SIQE and other IQM methods on the NBU-MDSID, LIVE 3D, and WaterlooIVC 3D datasets.

		Chen	*Lin*	*Shao*	*SOIQE*	SBM	SINQ	DCNN	StereoQA	PADNet	BSIQE	QAC	NIQE	ILNIQE	SISBLIM	BPRI	BMPRI	M-3D	CNN-SIQE
PLCC	NBUMD-I	*0.915*	***0.931***	*0.827*	*0.807*	0.578	0.828	0.755	0.834	0.718	0.640	0.845	0.926	0.708	0.599	0.843	0.900	**0.948**	0.939
	NBUMD-II	*0.835*	*0.832*	*0.732*	***0.858***	0.425	0.685	0.635	0.675	0.547	0.490	0.724	0.849	0.716	0.479	0.664	0.801	0.900	**0.913**
	LIVE-I	*0.930*	*0.872*	***0.933***	*0.799*	0.742	0.901	0.866	0.908	0.867	–	0.893	0.858	0.897	0.820	0.894	0.918	0.924	**0.926**
	LIVE-II	***0.911***	*0.655*	*0.819*	*0.824*	–	–	–	–	–	0.723	0.846	0.832	0.784	0.641	0.870	0.848	0.911	**0.913**
	WIVC-I	*0.711*	*0.688*	***0.833***	*0.830*	0.157	0.419	0.667	0.702	0.492	0.178	0.741	0.776	0.731	0.894	0.783	0.830	0.931	**0.949**
	WIVC-II	*0.573*	*0.578*	*0.737*	***0.751***	0.035	0.356	0.640	0.643	0.474	0.148	0.702	0.668	0.653	0.802	0.698	0.777	0.908	**0.931**
	Average	*0.806*	*0.754*	***0.810***	***0.810***	0.314	0.526	0.592	0.624	0.513	0.359	0.789	0.814	0.745	0.705	0.788	0.843	0.919	**0.928**
SROCC	NBUMD-I	*0.899*	***0.917***	*0.807*	*0.766*	0.671	0.822	0.737	0.835	0.698	0.648	0.855	0.901	0.592	0.706	0.837	0.853	**0.934**	0.920
	NBUMD-II	***0.822***	*0.797*	*0.721*	***0.822***	0.501	0.676	0.619	0.671	0.522	0.513	0.740	0.840	0.682	0.587	0.675	0.780	0.884	**0.889**
	LIVE-I	*0.903*	*0.830*	***0.904***	*0.870*	0.738	0.849	0.885	0.873	0.828	–	0.891	0.825	0.870	0.786	0.888	**0.892**	0.890	0.890
	LIVE-II	***0.904***	*0.639*	*0.797*	*0.815*	–	–	–	–	–	0.661	0.828	0.820	0.764	0.528	0.859	0.845	0.903	**0.911**
	WIVC-I	*0.626*	*0.611*	***0.817***	*0.783*	0.069	0.259	0.472	0.667	0.415	0.148	0.547	0.601	0.696	0.872	0.718	0.808	0.918	**0.937**
	WIVC-II	*0.489*	*0.479*	***0.721***	*0.717*	0.183	0.150	0.530	0.628	0.387	0.139	0.541	0.515	0.618	0.778	0.660	0.763	0.885	**0.929**
	Average	*0.767*	*0.705*	*0.791*	***0.794***	0.357	0.451	0.539	0.609	0.471	0.347	0.730	0.745	0.702	0.708	0.769	0.821	0.901	**0.913**
KROCC	NBUMD-I	*0.716*	***0.738***	*0.608*	*0.570*	0.515	0.607	0.519	0.620	0.489	0.472	0.651	0.718	0.427	0.631	0.631	0.648	**0.770**	0.745
	NBUMD-II	*0.634*	*0.611*	*0.534*	***0.637***	0.360	0.482	0.440	0.481	0.354	0.357	0.548	0.649	0.488	0.530	0.489	0.586	0.704	**0.708**
	LIVE-I	*0.726*	*0.637*	***0.734***	*0.681*	0.503	0.657	0.685	0.684	0.623	–	0.704	0.625	0.676	0.589	0.701	**0.708**	0.707	**0.708**
	LIVE-II	***0.731***	*0.480*	*0.605*	*0.651*	–	–	–	–	–	0.472	0.630	0.612	0.566	0.395	0.661	0.644	0.720	**0.732**
	WIVC-I	*0.469*	*0.458*	***0.636***	*0.596*	0.062	0.172	0.319	0.483	0.305	0.111	0.393	0.456	0.523	0.706	0.542	0.638	0.764	**0.791**
	WIVC-II	*0.346*	*0.344*	***0.535***	*0.527*	0.017	0.110	0.369	0.440	0.276	0.099	0.385	0.366	0.451	0.596	0.489	0.592	0.707	**0.770**
	Average	*0.597*	*0.539*	*0.605*	***0.608***	0.237	0.332	0.388	0.449	0.338	0.248	0.548	0.566	0.520	0.572	0.582	0.634	0.727	**0.742**
RMSE	NBUMD-I	*3.860*	***3.484***	*5.377*	*5.658*	9.573	5.363	6.282	5.281	6.659	7.356	5.121	3.608	6.765	7.663	5.154	4.174	**3.047**	3.295
	NBUMD-II	*6.610*	*6.669*	*12.020*	***6.172***	12.020	8.760	9.281	8.872	10.063	10.479	8.294	6.349	8.391	10.552	8.984	7.191	5.251	**4.915**
	LIVE-I	*5.723*	*7.620*	***5.576***	*15.532*	10.424	6.746	7.774	6.501	7.740	–	6.988	7.979	6.858	8.895	6.974	6.157	5.949	**5.884**
	LIVE-II	***4.619***	*8.452*	*6.409*	*6.338*	–	–	–	–	–	7.724	5.969	6.200	6.942	8.583	5.515	5.923	4.613	**4.562**
	WIVC-I	*11.936*	*12.319*	***9.397***	*9.457*	16.950	15.405	12.641	12.085	14.777	16.696	11.395	10.704	11.583	7.606	16.968	9.469	6.174	**5.352**
	WIVC-II	*16.627*	*16.556*	*13.705*	***13.403***	20.286	18.962	15.593	15.544	17.865	20.070	14.459	15.101	15.377	12.116	20.294	12.784	8.488	**7.399**

4.2. Testing

Three datasets which include the NBU-MDSID [21,22], LIVE 3D [49], and WaterlooIVC 3D [50] datasets were used for QM performance evaluation. In addition, we generated the forth dataset from the selected 224 pristine images in the Flickr1024 dataset [51] to test the performance of the three network models (see Section 4.5 for more details). Each SIQM dataset consists of two phases; we denote them by NBUMD-I/II, LIVE-I/II, and WIVC-I/II, respectively. Among the six sub-datasets, NBUMD-I and LIVE3D-I contain only symmetrically-distorted 3D images, while the others contain both symmetrically and asymmetrically distorted images. Multiple distortions (i.e., GB+WN+JPEG) are found in NBU-MDSID, whereas images in LIVE 3D and WaterlooIVC 3D contain only single distortions. All distorted images in NBU-MDSID were used for testing, while images with the same four distortion types in LIVE 3D and WaterlooIVC 3D were tested. Consequently, the testing data used for QM include 270 images from NBUMD-I, 300 images from NBUMD-II, 285 images from LIVE-I, 288 images from LIVE-II, 258 images from WIVC-I, and 340 images from WIVC-II.

We compared CNN-SIQE with various IQM models. The four full-reference (FR) IQM models include: the cyclopean MS-SSIM (Cyc-MS) [52], frequency-integrated PSNR (FI-PSNR) [53], BJND [54], and SOIQE [55]. The 13 NR IQM models include: the saliency-guided binocular model (SBM) [56], SINQ [57], deep convolution neural network (DCNN) [11], StereoQA [10], PADNet [58], BSIQE [13], QAC [59], NIQE [60], IL-NIQE [61], SISBLIM [62], BPRI [63], BMPRI [64], and MUSIQUE-3D (denoted by "M-3D" in Tables 5 and 6) [24]. The former six opinion-aware SIQM methods (i.e., SBM, SINQ, DCNN, StereoQA, PADNet, and BSIQE) were trained on either Phase I or Phase II of the LIVE 3D dataset. For the extra opinion-unaware 2D-IQM methods (i.e., QAC, NIQE, ILNIQE, SISBLIM, BPRI, and BMPRI), we took as the 3D quality a weighted combination of the quality measurements computed for the two views, and the weighting/combination strategy followed the same technique as in [45]. Four criteria including the Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC), Kendall Rank-Order Correlation Coefficient (KROCC), and Root Mean Square Error (RMSE) were used to quantify the performance of each IQM model. Note that a logistic transform was applied prior to computing these measures, yet only PLCC and RMSE are affected.

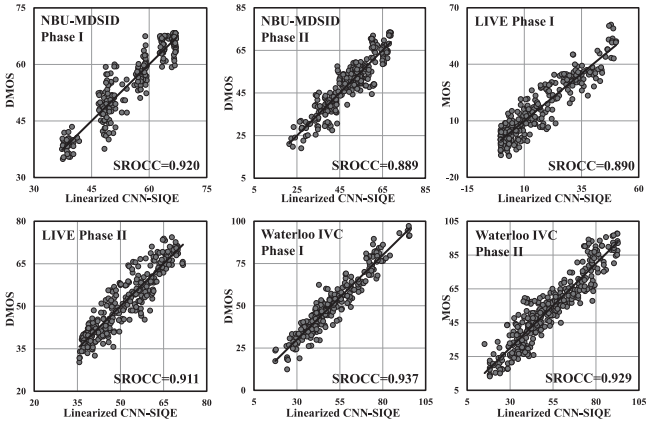

Fig. 9. Scatter plots of predicted (after logistic transform) vs. subjective quality ratings for CNN-SIQE on different 3D image quality datasets.

4.3. Overall performance

The overall performance of CNN-SIQE and other FR/NR IQM methods are presented in Table 5, in which the average values of PLCC/SROCC/KROCC were weighted by the numbers of distorted images tested from each sub-dataset. Due to the different MOS/DMOS ranges for different datasets, the average RMSE values were not computed. Moreover, because SBM, SINQ, DCNN, StereoQA, and PADNet were trained on LIVE-II, the testing results of these methods on that dataset are not presented. The same restriction was used for BSIQE, which was trained on LIVE-I. The best results achieved by the NR IQM methods are shown in bold. The best results achieved by the FR IQM methods are shown in italics and bold.

As shown in Table 5, CNN-SIQE performs quite well in quality measurement of both MDSIs and SDSIs. The six opinion-aware NR SIQM methods were trained on singly-distorted images (i.e., LIVE-I/II), and thus their weak performances on NBUMD-I/II are as expected. Yet, these methods also perform rather poorly on the WaterlooIVC 3D dataset which contains only singly-distorted stereoscopic images. This fact indicates that training on one dataset does not guarantee good performance on other datasets because different datasets vary in terms of image content, distortion types and levels, and quality-judging

Table 6
Performances of CNN-SIQE and other IQM methods tested on different distortion types in the LIVE 3D and WaterlooIVC 3D datasets.

			Chen	*Lin*	*Shao*	*SOIQE*	SBM	SINQ	DCNN	StereoQA	PADNet	BSIQE	QAC	NIQE	ILNIQE	SISBLIM	BPRI	BMPRI	M-3D	CNN-SIQE
PLCC	LIVE-I	WN	***0.934***	*0.907*	*0.932*	*0.883*	0.541	0.895	0.757	0.895	0.774	–	0.824	0.918	0.918	0.932	0.903	**0.941**	0.927	0.922
		JP2K	*0.908*	*0.838*	***0.919***	*0.884*	0.864	0.885	**0.937**	0.931	0.933	–	0.919	0.751	0.854	0.845	0.904	0.917	0.891	0.924
		JPEG	*0.591*	*0.208*	***0.638***	*0.593*	0.369	0.618	**0.748**	0.605	0.461	–	0.743	0.613	0.580	0.706	0.685	0.588	0.708	0.663
		GB	*0.929*	***0.946***	*0.940*	*0.864*	0.697	0.881	0.877	0.904	0.866	–	0.926	0.915	0.919	**0.940**	0.790	0.864	0.924	0.903
	LIVE-II	WN	***0.956***	*0.893*	*0.799*	*0.950*	–	–	–	–	–	0.579	0.706	0.893	**0.940**	0.788	0.894	0.903	0.927	0.912
		JP2K	*0.838*	*0.742*	*0.793*	***0.919***	–	–	–	–	–	0.759	0.788	0.637	0.780	0.548	0.831	0.807	0.857	**0.907**
		JPEG	*0.833*	*0.597*	*0.780*	***0.866***	–	–	–	–	–	0.682	0.844	0.672	0.719	0.812	0.861	0.624	**0.862**	0.841
		GB	***0.962***	*0.584*	*0.956*	*0.777*	–	–	–	–	–	0.909	0.953	0.946	0.931	0.955	0.893	0.926	0.973	**0.974**
	WIVC-I	WN	*0.775*	*0.768*	*0.815*	***0.863***	0.454	0.457	0.459	0.675	0.775	0.789	0.633	0.664	0.829	0.840	**0.925**	0.892	0.910	**0.925**
		JPEG	*0.941*	***0.970***	*0.833*	*0.930*	0.623	0.341	0.535	0.781	0.198	0.437	0.596	0.809	0.729	0.879	0.807	0.782	0.883	**0.927**
		GB	*0.595*	*0.668*	***0.907***	*0.886*	0.331	0.535	0.810	0.817	0.850	0.109	0.775	0.894	0.924	0.929	0.903	0.858	**0.966**	0.960
	WIVC-II	WN	*0.571*	*0.568*	*0.757*	***0.782***	0.471	0.337	0.539	0.632	0.601	0.636	0.515	0.404	0.800	0.839	0.749	0.695	0.842	**0.900**
		JPEG	*0.819*	***0.923***	*0.774*	*0.908*	0.227	0.122	0.520	0.664	0.040	0.473	0.583	0.742	0.695	0.777	0.763	0.706	0.814	**0.944**
		GB	*0.292*	*0.459*	***0.886***	*0.777*	0.300	0.439	0.820	0.835	0.836	0.009	0.793	0.791	0.834	0.875	0.944	0.878	0.950	**0.953**
SROCC	LIVE-I	WN	***0.948***	*0.928*	*0.930*	*0.925*	0.033	0.910	0.619	0.921	0.698	–	0.855	0.914	0.926	0.931	0.928	**0.937**	0.935	0.920
		JP2K	*0.896*	*0.839*	*0.883*	***0.901***	0.874	0.819	**0.919**	0.882	0.865	–	0.917	0.744	0.845	0.853	0.888	0.886	0.869	0.899
		JPEG	*0.558*	*0.207*	*0.611*	***0.618***	0.434	0.561	**0.730**	0.560	0.370	–	0.701	0.597	0.587	0.678	0.639	0.506	0.673	0.609
		GB	*0.926*	***0.935***	*0.910*	*0.908*	0.792	0.882	0.884	**0.901**	0.897	–	0.894	0.881	0.885	**0.901**	0.688	0.764	0.879	0.899
	LIVE-II	WN	*0.955*	*0.907*	*0.807*	***0.963***	–	–	–	–	–	0.399	0.686	0.895	**0.948**	0.863	0.904	0.898	0.917	0.915
		JP2K	*0.833*	*0.719*	*0.788*	***0.915***	–	–	–	–	–	0.828	0.782	0.632	0.768	0.634	0.843	0.814	0.860	**0.924**
		JPEG	***0.840***	*0.613*	*0.745*	*0.833*	–	–	–	–	–	0.721	0.809	0.637	0.663	**0.871**	0.851	0.679	0.832	0.837
		GB	*0.910*	*0.711*	***0.939***	*0.820*	–	–	–	–	–	0.825	0.846	0.864	0.889	0.900	0.777	0.792	0.877	**0.914**
	WIVC-I	WN	*0.837*	*0.784*	*0.829*	***0.912***	0.489	0.768	0.655	0.867	0.757	0.807	0.873	0.889	0.807	0.805	0.923	0.902	0.898	**0.924**
		JPEG	*0.935*	***0.960***	*0.828*	*0.919*	0.749	0.290	0.585	0.819	0.052	0.554	0.797	0.861	0.726	0.876	0.839	0.808	0.862	**0.911**
		GB	*0.600*	*0.758*	*0.924*	***0.932***	0.037	0.268	0.562	0.805	0.851	0.540	0.451	0.935	0.945	0.932	0.869	0.790	0.959	**0.960**
	WIVC-II	WN	*0.783*	*0.797*	*0.799*	***0.893***	0.459	0.721	0.700	0.811	0.619	0.660	0.903	0.834	0.797	0.839	0.812	0.804	0.852	**0.934**
		JPEG	*0.859*	***0.903***	*0.727*	*0.875*	0.231	0.111	0.649	0.664	0.020	0.529	0.596	0.680	0.651	0.733	0.753	0.701	0.822	**0.928**
		GB	*0.312*	*0.669*	***0.912***	*0.823*	0.279	0.137	0.591	0.810	0.838	0.277	0.717	0.936	0.883	0.880	0.942	0.891	0.937	**0.943**
KROCC	LIVE-I	WN	***0.803***	*0.761*	*0.775*	*0.765*	0.031	0.735	0.398	0.747	0.473	–	0.646	0.735	0.759	0.767	0.759	**0.780**	0.778	0.756
		JP2K	*0.714*	*0.639*	*0.696*	***0.727***	0.726	0.610	**0.755**	0.685	0.675	–	0.747	0.523	0.635	0.665	0.696	0.696	0.686	0.723
		JPEG	*0.372*	*0.132*	***0.445***	*0.423*	0.336	0.388	**0.526**	0.380	0.249	–	0.499	0.414	0.402	0.480	0.447	0.344	0.476	0.414
		GB	*0.772*	***0.786***	*0.739*	*0.746*	0.648	0.697	0.719	**0.727**	0.721	–	0.713	0.687	0.705	0.717	0.491	0.578	0.687	0.701
	LIVE-II	WN	*0.815*	*0.743*	*0.635*	***0.833***	–	–	–	–	–	0.284	0.500	0.714	**0.800**	0.682	0.728	0.725	0.761	0.743
		JP2K	*0.642*	*0.541*	*0.601*	***0.752***	–	–	–	–	–	0.634	0.601	0.437	0.573	0.473	0.664	0.628	0.676	**0.758**
		JPEG	*0.636*	*0.428*	*0.546*	***0.638***	–	–	–	–	–	0.525	0.620	0.457	0.486	**0.690**	0.656	0.483	0.643	0.640
		GB	***0.744***	*0.534*	*0.789*	*0.620*	–	–	–	–	–	0.615	0.626	0.661	0.714	0.716	0.552	0.559	0.689	**0.754**
	WIVC-I	WN	*0.674*	*0.602*	*0.633*	***0.751***	0.343	0.573	0.470	0.692	0.571	0.588	0.701	0.717	0.636	0.640	**0.775**	0.738	0.723	**0.775**
		JPEG	*0.788*	***0.841***	*0.659*	*0.771*	0.573	0.198	0.433	0.651	0.041	0.418	0.656	0.720	0.532	0.732	0.694	0.668	0.689	**0.762**
		GB	*0.495*	*0.594*	*0.760*	***0.783***	0.011	0.140	0.418	0.623	0.661	0.396	0.306	0.787	0.798	0.782	0.695	0.607	**0.826**	0.821
	WIVC-II	WN	*0.609*	*0.616*	*0.636*	***0.726***	0.376	0.526	0.503	0.634	0.439	0.473	0.736	0.656	0.610	0.667	0.679	0.665	0.677	**0.791**
		JPEG	*0.662*	***0.728***	*0.555*	*0.705*	0.190	0.088	0.494	0.500	0.007	0.382	0.433	0.504	0.479	0.570	0.589	0.539	0.633	**0.773**
		GB	*0.226*	*0.490*	***0.741***	*0.635*	0.225	0.048	0.429	0.635	0.643	0.175	0.519	0.783	0.708	0.702	0.792	0.718	0.780	**0.801**

standards. In comparison, the opinion-unaware methods (e.g., NIQE, BPRI, and BMPRI) seem to be more effective, suggesting that a model not trained on human subjective ratings might be the best option to solve the SIQM problem, given the limited quantity of the existing 3D image databases. Compared with MUSIQUE-3D, we observe that the advantage of CNN-SIQE is not quite impressive when tested on symmetrically-distorted images (e.g., NBUMD-I and LIVE-I). The reason is that the proposed human subjective experiment was mainly designed for modeling the properties of the HVS when asymmetrically-distorted images are viewed, and thus the trained SVR models do not actually help when symmetric distortions are present according to Eq. (36). Despite that, the result is still very competitive and promising, and we believe that the slight performance drop is justified given the markedly increased algorithm speed as demonstrated in Table 10.

Fig. 9 shows scatter plots of predicted (after logistic transform) vs. subjective quality ratings for the proposed CNN-SIQE algorithm on all six datasets (more visual result comparisons can be viewed in the online supplement at https://vinelab.jp/cnnsiqe/). In each plot, the horizontal axis denotes the logistic-transformed quality prediction; the vertical axis denotes the subjective ratings (DMOS values for NBU-MDSID and LIVE 3D; MOS values for WaterlooIVC). The scatter plots generally exhibit homoscedasticity, indicating that the algorithm's performance is consistent across different quality levels. In summary, by investigating the testing results on all dataset images, CNN-SIQE appears to offer better QM performance over other NR IQM methods.

4.4. Performance on individual distortion type

To evaluate the variation in performance as a function of distortion type, we also report the PLCC, SROCC, and KROCC values of each IQM method tested on the four individual distortion types in Table 6. Again, the results of SBM, SINQ, DCNN, StereoQA, and PADNet on LIVE-II, and the results of BSIQE on LIVE-I are not presented, because the corresponding dataset images were used for training. Entries shown in italics denote FR IQM methods. Italicized entries shown in bold denote the best-performing FR IQM method. Entries shown in bold denote the best-performing NR IQM method.

Observe from Table 6 that CNN-SIQE provides better or competitive predictions as compared with other IQM methods on most distortion types. Specifically, compared with FR SIQM methods, observe that some methods perform quite well on individual distortion types (e.g., SOIQE on LIVE-II and WIVC-I). However, their weak overall performances in Table 5 suggest that these methods apparently struggle to bring all quality measures to the same scale for the whole dataset. Compared with NR SIQM methods, especially the opinion-aware approaches, we observe a significant performance improvement on WIVC-I/II, suggesting that our method has better generalization and robustness. In particular, compared with MUSIQUE-3D, an equally-good performance is observed on LIVE-I which contains only symmetrically-distorted images, and much better results are observed on LIVE-II, WIVC-I, and WIVC-II, all of which contain both symmetrically and asymmetrically distorted images. These observations can be attributed to the same symmetric vs. asymmetric argument described in Section 4.3.

To further explore the effectiveness of using distortion parameter values as indicators of image quality, we report the PLCC, SROCC, and KROCC values computed between the estimated distortion parameters and the MOS/DMOS values in Table 7. Note that in this test, only singly-distorted images were considered such that the correlation between a specific distortion parameter and the image quality

Table 7
Correlation between human subjective ratings and distortion parameters estimated for the different distortion type images in the LIVE 3D and WaterlooIVC 3D datasets.

		2D			3D		
		PLCC	SROCC	KROCC	PLCC	SROCC	KROCC
LIVE-I	WN	0.925	0.921	0.756	0.921	0.920	0.755
	JP2K	0.919	0.891	0.709	0.920	0.892	0.709
	JPEG	0.632	0.603	0.412	0.620	0.586	0.401
	GB	0.944	0.915	0.747	0.929	0.904	0.719
LIVE-II	WN	0.904	0.911	0.737	0.922	0.922	0.757
	JP2K	0.896	0.897	0.719	0.925	0.923	0.756
	JPEG	0.828	0.849	0.647	0.704	0.745	0.563
	GB	0.937	0.879	0.696	0.925	0.924	0.764
WIVC-I	WN	0.943	0.939	0.790	0.894	0.875	0.708
	JPEG	0.910	0.887	0.710	0.881	0.870	0.683
	GB	0.875	0.947	0.805	0.846	0.943	0.800
WIVC-II	WN	0.943	0.932	0.792	0.871	0.868	0.691
	JPEG	0.911	0.887	0.707	0.881	0.872	0.686
	GB	0.858	0.937	0.789	0.814	0.912	0.751

can be easily modeled. In Table 7, entries of "2D" denote the correlation between the MOS/DMOS values and the linearly-combined 2D-distortion parameters, where the weights were computed based on Eqs. (28) and (31); entries of "3D" denote the correlation between the MOS/DMOS values and the 3D-vision distortion parameters predicted by the SVR models. As can be observed, the estimated distortion parameters correlate well with human subjective ratings for most distortion types, following a similar trend as Table 6. These results demonstrate the rationality and feasibility of measuring the image quality in terms of distortion parameters. In summary, by investigating the QM performance on individual distortion types, CNN-SIQE still performs competitively well.

4.5. Performance of individual network model

In CNN-SIQE, three network models are employed, among which MLP and BINet are the classification models, and FFENet is the regression model. As the ground-truth distortion parameters of the images in the three IQM datasets (i.e., NBU-MDSID [21,22], LIVE 3D [49], and WaterlooIVC 3D [50]) are publicly unavailable, we generated our own dataset to evaluate the performance of the three trained networks. Specifically, for each of the 224 pristine images selected from the test and validation sets of the Flickr1024 dataset [51], random distortion levels corresponding to the six distortion types in Table 1 were applied. As CNN-SIQE rescales large-size images into a fixed size to reduce the computational complexity, these 224 pristine images were also rescaled (if needed) before being distorted in order to avoid the scaling problem mentioned in Section 3.1.2, and consequently the ground-truth distortion parameters of each image could easily be accessed. Note that in our experiment the two views of each stereoscopic image share the same distortion type but different distortion parameter values. Accordingly, we generated in total 1,344 distorted stereopairs, each of which is associated with eight distortion parameters and two classification labels (i.e., L_1 and L_2 in Table 1).

Table 8 shows the performance of the three network models tested on the 1,344 distorted stereopairs. Entries of MLP and BINet denote the percentages of images with the distortion type in the row that were classified to have labels in the column; and the sum of each row for MLP/BINet equals one. Entries of FFENet denote the SROCC values computed between the estimated distortion parameters and the ground-truth; each value indicates how accurate the parameter of distortion in the column can be predicted for the singly/multiply-distorted images in the row. As can be observed, all network models perform quite well when images are singly distorted, but less effective when images are multiply distorted. The results are as expected, because the artifacts caused by the previously added distortions can be easily destroyed by the subsequently added distortions. For example, the finally-introduced noise distortion will make it difficult for the algorithm to detect the blurring/blocking artifacts, let alone predict their intensities. Yet our method still seems to be reasonable, as it mimics the properties of the HVS by analyzing the most apparent distortions, whose classification labels and intensity parameters can be well estimated even in multiple-distortion scenarios.

Table 8
Performances of MLP, BINet, and FFENet tested on the 1344 distorted stereopairs generated from the 224 pristine images in Flickr1024 [51].

Distortion type	MLP			BINet		FFENet			
	$L_1=0$	$L_1=1$	$L_1=2$	$L_2=0$	$L_2=1$	GB	JPEG	JP2K	WN
GB	0.000	0.011	0.989	0.998	0.002	0.995	–	–	–
JPEG	0.016	0.000	0.984	1.000	0.000	–	0.981	–	–
JP2K	0.000	0.000	1.000	0.002	0.998	–	–	0.972	–
WN	0.989	0.002	0.009	1.000	0.000	–	–	–	0.990
JP2K+WN	0.487	0.513	0.000	0.449	0.551	–	–	0.408	0.932
GB+JPEG+WN	0.225	0.770	0.004	1.000	0.000	0.662	0.575	0.000	0.935

Table 9
Performances of applying different quality-weighting strategies and different stages of CNN-SIQE to the NBU-MDSID, LIVE 3D, and WaterlooIVC 3D datasets.

		$S_{2D(avg)}$	$S_{2D(cst)}$	S_{2D}	$S_{3D(we)}$	$S_{3D(wb)}$	S_{cyc}	S_{3D}
PLCC	NBUMD-I	0.940	0.940	0.940	0.915	0.939	0.905	0.939
	NBUMD-II	0.931	0.919	0.911	0.913	0.913	0.904	0.913
	LIVE-I	0.923	0.923	0.922	0.925	0.926	0.907	0.926
	LIVE-II	0.817	0.860	0.881	0.920	0.907	0.914	0.913
	WIVC-I	0.812	0.890	0.912	0.949	0.942	0.864	0.949
	WIVC-II	0.752	0.871	0.920	0.931	0.922	0.824	0.931
SROCC	NBUMD-I	0.920	0.920	0.920	0.896	0.920	0.887	0.920
	NBUMD-II	0.889	0.888	0.886	0.887	0.891	0.875	0.889
	LIVE-I	0.889	0.889	0.889	0.886	0.889	0.874	0.890
	LIVE-II	0.794	0.854	0.876	0.917	0.903	0.908	0.911
	WIVC-I	0.807	0.861	0.899	0.930	0.930	0.840	0.937
	WIVC-II	0.760	0.854	0.919	0.930	0.921	0.803	0.929
KROCC	NBUMD-I	0.745	0.745	0.745	0.710	0.745	0.697	0.745
	NBUMD-II	0.713	0.708	0.704	0.704	0.713	0.691	0.708
	LIVE-I	0.706	0.706	0.706	0.703	0.708	0.685	0.708
	LIVE-II	0.601	0.656	0.681	0.736	0.719	0.731	0.732
	WIVC-I	0.625	0.689	0.723	0.783	0.779	0.664	0.791
	WIVC-II	0.572	0.681	0.752	0.774	0.758	0.616	0.770
RMSE	NBUMD-I	3.271	3.271	3.271	3.868	3.292	4.078	3.295
	NBUMD-II	4.375	4.733	4.963	4.899	4.896	5.133	4.916
	LIVE-I	5.992	5.966	6.012	5.914	5.884	6.557	5.884
	LIVE-II	6.446	5.713	5.299	4.378	4.710	4.539	4.562
	WIVC-I	9.914	7.727	6.950	5.351	5.679	8.535	5.352
	WIVC-II	13.376	9.964	7.974	7.414	7.851	11.500	7.399

4.6. Ablation study

We performed an ablation study to analyze the contributions of different stages in CNN-SIQE, as well as to investigate the effectiveness of the proposed quality-weighting strategy for the SIQM task. Accordingly, two extra weighting strategies (i.e., average weighting and contrast weighting using only Eq. (27) to compute the weight) were investigated; the QM performances corresponding to the first stage of CNN-SIQE using the three different weighting strategies are reported (denoted by $S_{2D(avg)}$, $S_{2D(cst)}$, and S_{2D}, respectively). We also report the QM performance of CNN-SIQE without using BINet [in which case D_{GR} in Eq. (26) was computed as the average of D_G and D_R; denoted by $S_{3D(wb)}$], and without using the equivalent blur/JP2K parameters for training FFENet and MLP [in which case D_{GR} in Eqs. (16) and (26) were computed as the maximum value of D_G and D_R; denoted by $S_{3D(we)}$]. All variant models were tested on the same three datasets, and with the same parameter settings. Results in terms of PLCC, SROCC, and KROCC are shown in Table 9, where S_{cyc} and S_{3D} denote, respectively, the second and final stages of CNN-SIQE for reference.

As can be observed, the second stage of CNN-SIQE can significantly improve the overall SIQM performance when testing on asymmetrically-distorted images, demonstrating that both stages are crucial. However,

Table 10
Runtimes in seconds/image for eight NR SIQM algorithms as a function of image size.

	180 × 320	360 × 640	540 × 960	1080 × 1920
SBM [56]	0.401	1.028	2.040	7.835
SINQ [57]	0.464	1.955	4.196	18.203
DCNN [11]	0.007	0.012	0.024	0.104
StereoQA [10]	0.009	0.038	0.085	0.336
PADNet [58]	–	0.091	0.090	0.095
BSIQE [13]	0.110	0.180	0.300	1.081
MUSIQUE-3D [24]	0.351	1.171	2.504	10.686
CNN-SIQE	0.056	0.257	0.577	0.577

we do observe that when testing on NBUMD-I and LIVE-I, which contain only symmetrically-distorted images, the second stage of CNN-SIQE is of little help and may even harm the performance. We suspect that this finding is due to the fact that the subjective experiment described in Section 3.2.1 used only asymmetrically-distorted image data to train the SVM models. Thus, a threshold in Eq. (36) is required to distinguish between symmetrically and asymmetrically-distorted images in order to achieve decent performance on both types of distortion. Comparing with $S_{3D(wb)}$, we conclude that using BINet can help improve the performance of CNN-SIQE especially when asymmetrically-distorted images are presented. Meanwhile, the performance drop on NBUMD for $S_{3D(we)}$ seems to indicate that an equivalent blur/JP2K parameter setting for JP2K/blur images can benefit the training process especially when multiple distortions are observed. As for the different quality-weighting strategies, we observe that the proposed adaptive contrast-weighting strategy which takes into account both image contrast and the characteristics of different distortion types is the best choice to help achieve equally-good performance on most datasets.

4.7. Computational complexity

In this section, we compare the running time of different NR SIQM methods on different-sized images to analyze the computational complexity of the proposed method. Specifically, four different image sizes including 180×320, 360 × 640, 540×960, and 1080 × 1920 pixels were used, and all times were measured by executing unoptimized MATLAB code or Python code on the same workstation as described in Section 4.1. The average runtime of each algorithm tested for 30 trials are shown in Table 10. Note that PADNet was applied to image patches of 256 × 256-pixel size, and thus its runtime on 180 × 320-pixel images is not available.

As shown in Table 10, all CNN-based SIQM models run much faster than the traditional SVM-based methods, which is as expected due to the GPU acceleration. Our method runs a bit slower than most of the other CNN-based methods because of the time needed to compute the RMS contrast of the two views (whereas in other approaches only quality scores are predicted by the corresponding network). Note that the runtime of CNN-SIQE does not change when images become larger because large-sized images are reshaped to a fixed size before being processed. Given the increased QM performance of CNN-SIQE vs. other CNN-based opinion-aware methods, as well as the much increased running efficiency as compared with MUSIQUE-3D [24], we believe that the time cost can be justified.

5. Conclusion

We presented an NR approach for quality measurement of multiply and singly distorted stereoscopic images. Our method, called CNN-SIQE, employs (1) an FFENet for distortion parameter estimation of the two monocular views, (2) SVR models for distortion parameter estimation of the cyclopean view, and (3) an MLP, BINet, and an adaptive contrast-weighting strategy for quality combination and estimation. All of the three strategies give rise to a much faster and effective NR SIQM technique, which demonstrated better and/or highly competitive performance in comparison with other methods on various stereoscopic image quality datasets.

Yet, despite the effectiveness of CNN-SIQE, there is still room to further improve the robustness of our method because its performance did indeed decrease for some distortion types in some databases. In fact, the issue of robustness commonly exists in the NR SIQM field, since different 3D image quality datasets vary significantly in image content, in distortion types/levels, and even in the subjective quality-rating standards, making it difficult for one algorithm to outperform all others on all images/distortion types on all datasets. Also, our current model considers only four distortion types and their combinations; whether or not it can be generalized to work on other distortion types requires further investigation. Thus, future work might focus on the development of alternative training strategies and/or adaptive network architectures that could possibly help push SIQM toward achieving equally competitive performance on all different dataset images. Future work could also involve building a more generalized NR SIQM technique to handle more distortion types (e.g., contrast change and other real-world distortions), thus increasing its applicability to a wider range of usage scenarios.

CRediT authorship contribution statement

Yi Zhang: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Damon M. Chandler:** Writing – review & editing, Visualization, Software, Formal analysis, Data curation. **Xuanqin Mou:** Writing – review & editing, Validation, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61901355, 62071375, 62271384) and the China Postdoctoral Science Foundation (Grant No. 2018M640991).

References

[1] X. Wang, F. Shao, Q. Jiang, Z. Fu, X. Meng, K. Gu, Y.-S. Ho, Combining retargeting quality and depth perception measures for quality evaluation of retargeted stereopairs, IEEE Trans. Multimed. 24 (2021) 2422–2434.
[2] M.-J. Chen, L.K. Cormack, A.C. Bovik, No-reference quality assessment of natural stereopairs, IEEE Trans. Image Process. 22 (9) (2013) 3379–3391.
[3] F. Shao, W. Tian, W.-S. Lin, G.-Y. Jiang, Q.-H. Dai, Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions, IEEE Trans. Image Process. 25 (5) (2016) 2059–2074.
[4] W. Zhou, L. Yu, Binocular responses for no-reference 3D image quality assessment, IEEE Trans. Multimed. 18 (6) (2016) 1077–1084.
[5] H. Oh, S. Ahn, J. Kim, S. Lee, Blind deep S3D image quality evaluation via local to global feature aggregation, IEEE Trans. Image Process. 26 (10) (2017) 4923–4936.
[6] W. Zhou, L. Yu, Y. Zhou, W. Qiu, M.-W. Wu, T. Luo, Blind quality estimator for 3D images based on binocular combination and extreme learning machine, Pattern Recognit. 71 (2017) 207–217.
[7] W. Zhou, W. Qiu, M.-W. Wu, Utilizing dictionary learning and machine learning for blind quality assessment of 3-D images, IEEE Trans. Broadcast. 63 (2) (2017) 404–415.
[8] Q. Jiang, F. Shao, W. Lin, G. Jiang, Learning a referenceless stereopair quality engine with deep nonnegativity constrained sparse autoencoder, Pattern Recognit. 76 (2018) 242–255.

[9] J. Yang, Y. Zhao, Y. Zhu, H. Xu, W. Lu, Q. Meng, Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network, Inform. Sci. 474 (2019) 1–17.

[10] W. Zhou, Z. Chen, W. Li, Dual-stream interactive networks for no-reference stereoscopic image quality assessment, IEEE Trans. Image Process. 28 (8) (2019) 3946–3958.

[11] Y. Fang, J. Yan, X. Liu, J. Wang, Stereoscopic image quality assessment by deep convolutional neural network, J. Vis. Commun. Image Represent. 58 (2019) 400–406.

[12] Q. Jiang, F. Shao, W. Gao, Z. Chen, G. Jiang, Y.-S. Ho, Unified no-reference quality assessment of singly and multiply distorted stereoscopic images, IEEE Trans. Image Process. 28 (4) (2019) 1866–1881.

[13] K. Sim, J. Yang, W. Lu, X. Gao, Blind stereoscopic image quality evaluator based on binocular semantic and quality channels, IEEE Trans. Multimed. 24 (2021) 1389–1398.

[14] Y. Chang, S. Li, A. Liu, W. Zhang, J. Jin, W. Xiang, Bidirectional feature aggregation network for stereo image quality assessment considering parallax attention-based binocular fusion, IEEE Trans. Broadcast. (2023).

[15] S. Ryu, K. Sohn, No-reference quality assessment for stereoscopic images based on binocular quality perception, IEEE Trans. Circuits Syst. Video Technol. 24 (4) (2013) 591–602.

[16] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.

[17] W.-J. Zhou, L. Yu, W.-W. Qiu, T. Luo, Z.-P. Wang, M.-W. Wu, Utilizing binocular vision to facilitate completely blind 3D image quality measurement, Signal Process. (2016).

[18] F. Shao, W. Lin, S. Wang, G. Jiang, M. Yu, Blind image quality assessment for stereoscopic images using binocular guided quality lookup and visual codebook, IEEE Trans. Broadcast. 61 (2) (2015) 154–165.

[19] F. Shao, W. Lin, S. Wang, G. Jiang, M. Yu, Q. Dai, Learning receptive fields and quality lookups for blind quality assessment of stereoscopic images, IEEE Trans. Cybern. 46 (3) (2016) 730–743.

[20] F. Shao, Z. Zhang, Q. Jiang, W. Lin, G. Jiang, Toward domain transfer for no-reference quality prediction of asymmetrically distorted stereoscopic images, IEEE Trans. Circuits Syst. Video Technol. 28 (3) (2018) 573–585.

[21] F. Shao, W. Tian, W. Lin, G. Jiang, Q. Dai, Learning sparse representation for blind quality assessment of multiply distorted stereoscopic images, IEEE Trans. Multimed. 19 (9) (2017) 1821–1836.

[22] F. Shao, Y. Gao, Q. Jiang, G. Jiang, Y. Ho, Multistage pooling for blind quality prediction of asymmetric multiply-distorted stereoscopic images, IEEE Trans. Multimed. 20 (10) (2018) 2605–2619.

[23] Q. Jiang, Z. Peng, F. Shao, K. Gu, Y. Zhang, W. Zhang, W. Lin, StereoARS: Quality evaluation for stereoscopic image retargeting with binocular inconsistency detection, IEEE Trans. Broadcast. 68 (1) (2021) 43–57.

[24] Y. Zhang, D.M. Chandler, X. Mou, Quality assessment of multiply and singly distorted stereoscopic images via adaptive construction of cyclopean views, Signal Process., Image Commun. 94 (2021) 116175.

[25] C.-C. Su, L.K. Cormack, A.C. Bovik, Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation, IEEE Trans. Image Process. 24 (5) (2015) 1685–1699.

[26] F. Shao, K. Li, W. Lin, G. Jiang, Q. Dai, Learning blind quality evaluator for stereoscopic images using joint sparse representation, IEEE Trans. Multimed. 18 (10) (2016) 2104–2114.

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

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

[29] Y. Qi, G. Jiang, M. Yu, Y. Zhang, Y.-S. Ho, Viewport perception based blind stereoscopic omnidirectional image quality assessment, IEEE Trans. Circuits Syst. Video Technol. 31 (10) (2020) 3926–3941.

[30] X. Chai, F. Shao, Q. Jiang, X. Meng, Y.-S. Ho, Monocular and binocular interactions oriented deformable convolutional networks for blind quality assessment of stereoscopic omnidirectional images, IEEE Trans. Circuits Syst. Video Technol. 32 (6) (2021) 3407–3421.

[31] L. Shen, X. Chen, Z. Pan, K. Fan, F. Li, J. Lei, No-reference stereoscopic image quality assessment based on global and local content characteristics, Neurocomputing 424 (2021) 132–142.

[32] J. Yang, K. Sim, X. Gao, W. Lu, Q. Meng, B. Li, A blind stereoscopic image quality evaluator with segmented stacked autoencoders considering the whole visual perception route, IEEE Trans. Image Process. 28 (3) (2018) 1314–1328.

[33] W. Zhang, C. Qu, L. Ma, J. Guan, R. Huang, Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network, Pattern Recognit. 59 (2016) 176–187.

[34] J. Si, B. Huang, H. Yang, W. Lin, Z. Pan, A no-reference stereoscopic image quality assessment network based on binocular interaction and fusion mechanisms, IEEE Trans. Image Process. 31 (2022) 3066–3080.

[35] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: IEEE International Conference on Computer Vision, ICCV, vol. 2, 2001, pp. 416–423.

[36] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, L. Zhang, Waterloo exploration database: New challenges for image quality assessment models, IEEE Trans. Image Process. 26 (2) (2016) 1004–1016.

[37] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, P. Westling, High-resolution stereo datasets with subpixel-accurate ground truth, in: German Conference on Pattern Recognition, Springer, 2014, pp. 31–42.

[38] W. Sun, F. Zhou, Q. Liao, MDID: A multiply distorted image database for image quality assessment, Pattern Recognit. 61 (2017) 153–168.

[39] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.

[40] Y. Zhang, D.M. Chandler, X. Mou, Multi-domain residual encoder–decoder networks for generalized compression artifact reduction, J. Vis. Commun. Image Represent. 83 (2022) 103425.

[41] Z. Wang, A.C. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[42] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

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

[44] Y. Zhang, D.M. Chandler, Opinion-unaware blind quality assessment of multiply and singly distorted images via distortion parameter estimation, IEEE Trans. Image Process. 27 (11) (2018) 5433–5448.

[45] Y. Zhang, D.M. Chandler, 3D-MAD: A full reference stereoscopic image quality estimator based on binocular lightness and contrast perception, IEEE Trans. Image Process. 24 (11) (2015) 3810–3825, http://dx.doi.org/10.1109/tip.2015.2456414.

[46] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[47] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.

[48] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.

[49] A.K. Moorthy, C.-C. Su, A. Mittal, A.C. Bovik, Subjective evaluation of stereoscopic image quality, Signal Process., Image Commun. 28 (2012) 870–873.

[50] J. Wang, A. Rehman, K. Zeng, S. Wang, Z. Wang, Quality prediction of asymmetrically distorted stereoscopic 3D images, IEEE Trans. Image Process. 24 (11) (2015) 3400–3414.

[51] Y. Wang, L. Wang, J. Yang, W. An, Y. Guo, Flickr1024: A large-scale dataset for stereo image super-resolution, in: International Conference on Computer Vision Workshops, 2019, pp. 3852–3857.

[52] M.-J. Chen, C.-C. Su, D.-K. Kwon, L.K. Cormack, A.C. Bovik, Full-reference quality assessment of stereopairs accounting for rivalry, Signal Process., Image Commun. 28 (9) (2013) 1143–1155.

[53] Y.-H. Lin, J.-L. Wu, Quality assessment of stereoscopic 3D image compression by binocular integration behaviors, IEEE Trans. Image Process. 23 (4) (2014) 1527–1542.

[54] F. Shao, W. Lin, S. Gu, G. Jiang, T. Srikanthan, Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics, IEEE Trans. Image Process. 22 (5) (2013) 1940–1953.

[55] Z. Chen, J. Xu, C. Lin, W. Zhou, Stereoscopic omnidirectional image quality assessment based on predictive coding theory, IEEE J. Sel. Top. Sign. Proces. 14 (1) (2020) 103–117.

[56] X. Xu, Y. Zhao, Y. Ding, No-reference stereoscopic image quality assessment based on saliency-guided binocular feature consolidation, Electron. Lett. 53 (2017) 1468–1470.

[57] L. Liu, B. Liu, C.-C. Su, H. Huang, A.C. Bovik, Binocular spatial activity and reverse saliency driven no-reference stereopair quality assessment, Signal Process., Image Commun. 58 (2017) 287–299.

[58] J. Xu, W. Zhou, Z. Chen, S. Ling, P.L. Callet, Binocular rivalry oriented predictive autoencoding network for blind stereoscopic image quality measurement, IEEE Trans. Instrum. Meas. 70 (2020) 1–13.

[59] W. Xue, L. Zhang, X. Mou, Learning without human scores for blind image quality assessment, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2013, pp. 995–1002.

[60] A. Mittal, R. Soundararajan, A.C. Bovik, Making a completely blind image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.

[61] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (2015) 2579–2591.

[62] K. Gu, G. Zhai, X. Yang, W. Zhang, Hybrid no-reference quality metric for singly and multiply distorted images, IEEE Trans. Broadcast. 60 (3) (2014) 555–567.

[63] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, C. Chen, Blind quality assessment based on pseudo-reference image, IEEE Trans. Multimed. 20 (8) (2017) 2049–2062.

[64] X. Min, G. Zhai, K. Gu, Y. Liu, X. Yang, Blind image quality estimation via distortion aggravation, IEEE Trans. Broadcast. 64 (2) (2018) 508–517.