Unified No-Reference Quality Assessment of Singly and Multiply Distorted Stereoscopic Images

Qiuping Jiang, Feng Shao, *Member, IEEE*, Wei Gao, *Member, IEEE*, Zhuo Chen, Gangyi Jiang, *Member, IEEE*, and Yo-Sung Ho, *Fellow, IEEE*

***Abstract*— A challenging problem in the no-reference quality assessment of multiply distorted stereoscopic images (MDSIs) is to simulate the monocular and binocular visual properties under a mixed type of distortions. Due to the joint effects of multiple distortions in MDSIs, the underlying monocular and binocular visual mechanisms have different manifestations with those of singly distorted stereoscopic images (SDSIs). This paper presents a unified no-reference quality evaluator for SDSIs and MDSIs by learning monocular and binocular local visual primitives (MB-LVPs). The main idea is to learn MB-LVPs to characterize the local receptive field properties of the visual cortex in response to SDSIs and MDSIs. Furthermore, we also consider that the learning of primitives should be performed in a task-driven manner. For this, two penalty terms including reconstruction error and quality inconsistency are jointly minimized within a supervised dictionary learning framework, generating a set of quality-oriented MB-LVPs for each single and multiple distortion modality. Given an input stereoscopic image, feature encoding is performed using the learned MB-LVPs as codebooks, resulting in the corresponding monocular and binocular responses. Finally, responses across all the modalities are fused with probabilistic weights which are determined by the modality-specific sparse reconstruction errors, yielding the final monocular and binocular features for quality regression. The superiority of our method has been verified on several SDSI and MDSI databases.**

***Index Terms*— No-reference image quality assessment, stereoscopic image, singly distorted, multiply distorted, monocular and binocular vision, receptive field, local visual primitive.**

Manuscript received February 21, 2018; revised September 12, 2018 and October 30, 2018; accepted November 7, 2018. Date of publication November 19, 2018; date of current version December 5, 2018. This work was supported in part by the Natural Science Foundation of China under Grants 61622109, 61871247, and 61801303, in part by the Zhejiang Natural Science Foundation under Grant R18F010008, in part by the Natural Science Foundation of Ningbo under Grant 2017A610112, in part by the Startup Project of Shenzhen University under Grant 2018069, and in part by the K. C. Wong Magna Fund of Ningbo University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Patrick Le Callet. *(Corresponding author: Feng Shao.)*

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

W. Gao is with the National Engineering Laboratory for Big Data System Computing Technology, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: gaowei262@126.com).

Z. Chen is with the Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore 639798 (e-mail: zchen036@e.ntu.edu.sg).

Y.-S. Ho is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea (e-mail: hoyo@gist.ac.kr).

Digital Object Identifier 10.1109/TIP.2018.2881828

I. Introduction

AUTOMATIC image quality assessment (IQA) is potentially useful for many image processing applications. QA for 2D images has been widely investigated, and many advanced 2D-IQA metrics have been developed [1]–[11]. Over the past years, owing to the emerging of stereoscopic three-dimensional (3D) contents for the use in many consumer devices such as 3D television, 3D video conference system, 3D online game and more, stereoscopic 3D content has become a hot research target of IQA.

Compared to its 2D counterpart, 3D-IQA encounters more challenges as the joint effects of image distortion, depth perception, visual discomfort, and visual presence need to be addressed simultaneously [12], [13]. However, this task is extremely challenging at the current stage given that the underlying complex interactions cannot be precisely modeled without a deep understanding of the cognitive mechanism of human brain. By this consideration, the majority of works focus on ascertaining the influence of each individual aspect on the overall 3D quality-of-experience of users [14]–[29]. As such, this paper targets to evaluate the visual quality of stereoscopic images contaminated by distortions. Similar to 2D-IQA, 3D-IQA also has three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). In view of the practicality value of assessing a stereopair without utilizing any information of its original version, we are more interested in the NR case of 3D-IQA.

A stereoscopic 3D image consists of a pair of 2D monocular images, each of which is controlled to be separately projected onto each eye of the viewer. Both the left and right images are of the same scene but captured at two slightly different perspectives. Due to the small lateral displacements between the positions of the two 2D images, our brain can have depth perception via binocular stereopsis. While most regions in one image can find their correspondence in the other, there are still some monocular regions in the left and right images since occlusion will inevitably occur [30]–[32]. For example, a small amount of background area behind the foreground object that can be seen in the left view will be occluded in the right view. Another case of monocular region is the border area. Take the toed-in camera array as an example, due to the viewing angle limit, a small amount of left (right) border area of the right (left) image only can be seen in the right (left) view. Due to the existence of monocular and binocular regions

in an input stereopair, both monocular and binocular visions are important for the perception of stereoscopic image quality. Obviously, efforts towards designing efficient visual models to resemble the monocular and binocular vision properties will be beneficial to 3D-IQA.

As an important part of the human visual system, the primary visual cortex is responsible for most of our perception of the real world's visual information [33], [34]. So, an ideal visual model for image quality evaluation should well resemble the neural response properties of the visual cortex. It has been discovered that there are two kinds of neuron cells in the visual cortex: monocular receptive field (MRF) and binocular receptive field (BRF) [35]–[37]. MRF refers to those neuron cells that only response to the stimulus presented to one particular eye while no response will be evoked if a stimulus is only presented to the other one. BRF refers to those neuron cells that have a clearly defined RF for each eye, such that an appropriate stimulus presented to either the left or the right eye will produce a response. To a simple approximation, the overall response of the binocular cells is then the sum of the responses to the left- and right-eyes' stimuli. That is, the stimuli in monocular regions will only be processed by the monocular neuron cells and the responses of the MRFs are then considered as the responses of the visual cortex towards monocular stimuli. Unlike the monocular stimuli, the stimuli in binocular regions will be processed by the binocular neuron cells and the overall responses of the BRFs in the left and right views are considered as the responses of the visual cortex towards binocular stimuli [38].

Although the above physiological mechanism seems to be natural to human visual system, formulating an efficient visual cortex-like coding model to encode monocular and binocular stimuli and adapt it to 3D-IQA is non-trivial. The critical challenge lies in simulating the MRF and BRF properties in response to stereo stimuli with different distortion types involved in 3D-IQA. It is known that, stereopairs can be either singly distorted or multiply distorted. Compared to the singly distorted case where the quality of a singly distorted stereoscopic image (SDSI) is only related to our perception of a certain distortion type, multiply distorted stereoscopic images (MDSIs) pose more challenges for quality evaluation due to the effect of interactions among different distortion types. To better cope with such challenges, how to simulate the properties of MRFs and BRFs in response to SDSIs and MDSIs needs to be addressed. Furthermore, we also consider the simulation of MRF and BRF properties for IQA should be built in a task-driven manner because quality perception is a highly subjective task. As such, the modeling of MRF and BRF properties should be well adapted to it.

Based on these considerations, this paper proposes a unified NR quality assessment method for SDSIs and MDSIs by learning task-oriented and modality-specific monocular and binocular local visual primitives (MB-LVPs) to characterize the underlying MRF and BRF properties of the visual cortex in response to stereopairs with different distortion modalities (single/multiple distortion). For this, two penalty terms including reconstruction error penalty (data-driven) and quality inconsistency penalty (task-driven) are combined and jointly minimized within a supervised dictionary learning framework to generate a set of quality-oriented MB-LVPs for each distortion modality. Traditionally, the reconstruction error is the only energy to be minimized for learning LVPs. However, the LVPs learned in such manner are not necessarily quality-aware because it fails taking the quality information into account. To obtain highly quality-aware LVPs that are suitable for the use in quality evaluation, we propose to incorporate a new quality inconsistency term into the traditional reconstruction error term to form a final objective function for optimization. Then, given an input stereoscopic image (can be either SDSI or MDSI), feature encoding is performed using the learned MB-LVPs as codebooks, resulting in the corresponding monocular and binocular responses. Finally, responses across all modalities are fused with probabilistic weights which are determined by the modality-specific reconstruction errors, yielding the final monocular and binocular features for quality regression. Overall, the contributions of this paper are three-fold:

- We propose a unified NR quality method which can be used to evaluate SDSIs and MDSIs simultaneously.
- We employ a task-driven and modality-specific dictionary learning framework to learn MB-LVPs that resemble the MB-RFs found in the visual cortex for 3D-IQA.
- We provide a cross-modality aggregation scheme based on sparse reconstruction error to characterize the masking effect of different distortion types (for MDSI) and the particularity of each individual distortion type (for SDSI).

The remainder of this paper is organized as follows. Related works are reviewed in Section II. The proposed method is described in Section III. In Section IV, experiments on both SDSI and MDSI databases are conducted. Finally, conclusions are drawn in Section V.

II. Related Work

A. No-Reference Assessment of Singly Distorted 2D Image

The problem of NR quality assessment for singly-distorted 2D images (NR-SDIQA) has long been an active research topic. Throughout the history, research efforts on NR-SDIQA have gone through two stages: distortion-specific and general-purpose. Distortion-specific approaches target at evaluating the quality of an image corrupted by one specific distortion type. Many distortion-specific approaches have been developed for evaluating sharpness [39], blocking artifacts [40], ringing artifacts [41], contrast change [42], and more. Although these distortion-specific approaches perform quite well on single distortion type, their generality across other distortion types are inadequate. Given that the distortion type is not always known in practical applications, designing effective general-purpose approaches that can handle more commonly encountered distortion types is necessary.

The past several years have witnessed tremendous progress in the development of general-purpose NR-SDIQA approaches among which natural scene statistics (NSS) features based ones dominate the landscape. The basic assumption of NSS-based general-purpose approaches is that pristine natural images inherently obey certain regular statistical rules

which will however be modified by distortions. With this, several NSS properties in spatial and transform domains have been exploited and utilized to extract quality-aware features [43]–[48]. By taking advantage of the machine learning algorithms, such as support vector regression, random forest, neural network, and more, the extracted NSS quality-aware features are mapped to quality scores in a convenient way. Another pipeline of general-purpose NR-SDIQA approaches follows a feature learning-based paradigm. In contrast to the handcrafted NSS features which rely heavily on the domain knowledge of natural scenes, feature learning-based approaches directly generate quality-aware features by feature encoding over a codebook learned from a set of raw patches or local feature descriptors [49]–[52]. The key steps are codebook construction and feature encoding. In practice, codebooks can be constructed in either unsupervised or supervised way, and feature encoding also can be performed in many different ways such as hard assignment, soft assignment, sparse coding, locality-constrained linear coding, and more. It is considered that, the feature extraction module is purely data-driven if the codebook is learned in an unsupervised way, while it is deemed to be both data-driven and task-driven if the codebook is learned in a supervised way. Our previous work has demonstrated that a certain amount of performance improvement can be achieved when adding a proper task-related constraint term to guide the codebook optimization [53], [54].

B. No-Reference Assessment of Multiply Distorted 2D Image

Although the above NR-SDIQA methods can be used to evaluate multiply-distorted images with moderate performance, there also have been some NR-IQA methods specifically designed for multiply distorted images (NR-MDIQA) to handle the newly raised challenges. Gu *et al*. [55] proposed a NR-MDIQA method containing several image processing units to simulate the quality assessment process of the human visual system. To be specific, the noise strength is first estimated, followed by blur and JPEG metrics applied on the denoised image. The final quality score is derived by incorporating a so-called free energy term to characterize the interaction among different distortion types to fuse the results of noise, blur, and JPEG metrics. Lu *et al*. [56] first performed feature selection on a set of NSS features to screen the features which are sensitive to one distortion even in the presence of another distortion. Then, the selected features are then encoded through an improved Bag-of-Word (BoW) model. Lastly, the joint effects of multiple distortions are modeled using a linear combination strategy for quality prediction. Li *et al*. [57] extracted a novel image-level structural feature representation called the gradient-weighted histogram of local binary pattern (LBP) calculated on the gradient map (GWH-GLBP) to describe the sophisticated quality degradation pattern introduced by multiple distortions. Inspired by the success of GWH-GLBP, Hadizadeh and Bajić [58] also proposed to first construct a set of feature maps based on the color Gaussian jet of an image and then apply the LBP operator on all the estimated feature maps to describe the potential quality degradation patterns caused by multiple distortions.

C. No-Reference Assessment of Singly Distorted 3D Image

The problem of NR-IQA for singly-distorted stereoscopic 3D images (NR-SDSIQA) is less investigated. Chen *et al*. [59] proposed to construct a cyclopean image for stereopair quality analysis by considering the disparity information and Gabor filter response. Then, 2D NSS features extracted from the cyclopean image along with the 3D NSS features extracted from the disparity map and uncertainty map constitute the final feature vector for quality regression. The Stereoscopic/3D BLind Image Naturalness Quality (S3D-BLINQ) index presented in [60] first estimated a cyclopean image using disparity map, then extracted both spatial-domain and wavelet-domain univariate and bivariate natural scene statistics to predict quality. In [61], a Bivariate Generalized Gaussian Density (BGGD) model was used to fit the joint statistics of luminance and disparity, resulting in an effective NR-SDSIQA approach dubbed Stereo Quality Evaluator (StereoQUE). Zhou and Yu [62] proposed a NR-SDSIQA method from the perspective of simulating the critical binocular combination and rivalry properties of the HVS to create binocular response maps from which the quality-aware features were extracted. Shao *et al*. [63], [64] proposed a feature-based binocular combination framework for NR-SDSIQA. It is claimed that the weights should be adaptive with respect to different distortion types in binocular combination and can be approximated by the sparse feature distribution index. Liu *et al*. [65] developed a new model for NR-SDSIQA that considered the impact of binocular fusion, rivalry, suppression, and a reverse saliency effect on the perception of distortion, resulting in a Stereo 3D INtegrated Quality (StereoINQ) Predictor. Zhang *et al*. [66] proposed to learn structures from stereopairs based on convolutional neural network (CNN) for NR-SDSIQA. Jiang *et al*. [67] designed a three-column Deep Nonnegativity Constrained Sparse Auto-Encoder (DNCSAE) with each individual DNCSAE module coping with the left image, the right image, and the cyclopean image, respectively. Oh *et al*. [68] explored a deep learning approach called DNR-S3DIQE for NR-SDSIQA based on local to global deep feature aggregation. It consists of two parts: the automatic extraction of meaningful local features, and their aggregation into global features that contain holistic information for 3D image quality. To make this available, a two-step training process is used for local and global feature extraction and bi-directional update.

D. No-Reference Assessment of Multiply Distorted 3D Image

In spite of the high possibility of stereoscopic images to be contaminated by multiple distortions, there is very limited work focusing on NR quality assessment of multiply-distorted stereoscopic images (NR-MDSIQA). In the literature, Shao *et al*. [69] made an attempt on this problem from both subjective and objective aspects. On the subjective aspect, they constructed a new MDSI database (NBU-MDSID) consisting of 270 MDSIs each of which is contaminated by all three distortion types (Gaussian blur (GB), Gaussian white noise (WN), and JPEG compression (JPEG)) and 90 SDSIs contaminated by one of the three distortion types. On the objective aspect, a new MUlti-Modal BLInd Metric (MUMBLIM) is proposed as the solution for NR-MDSIQA.

However, this objective method suffers from the following problems. First, it still follows the traditional pipeline that first evaluates the left and right images individually and then applies a binocular combination scheme to fuse the above two results into a final score. Although this pipeline has achieved a certain amount of performance improvement by enforcing proper combination weights, it is still lack of interpretability and inconsistent with the cognitive process of the human visual system when viewing a stereoscopic image (the visual information from the two images will be merged via stereo vision for subsequent differential neural coding with respect to different cortical RFs). Second, the different roles of MRFs and BRFs in creating the stereo perception are not distinctively characterized. It is known that, the corresponding and non-corresponding regions in a stereopair are processed by BRFs and MRFs, respectively. Therefore, a more reasonable way is to first model the MRF and BRF properties, respectively, then deploy such models to encode the monocular and binocular regions for quality-aware feature extraction. In this paper, we propose a unified no-reference quality assessment method for SDSIs and MDSIs by learning task-oriented MB-LVPs to better address the above problems.

III. Proposed Method

The quality-aware features in our method are obtained by feature encoding using a set of learned task-oriented and modality-specific MB-LVPs. Given an input stereoscopic image (either SDSI or MDSI), the feature encoding of its monocular (*i.e.*, non-corresponding) and binocular (*i.e.*, corresponding) regions are performed separately with respect to the learned M-LVPs and B-LVPs, resulting in the corresponding monocular and binocular responses. Finally, responses across all modalities are fused with probabilistic weights which are determined by the modality-specific sparse reconstruction errors (SREs), yielding the final monocular and binocular features for quality regression. obviously, the key to the success of our proposed method is to learn a set of MB-LVPs in a task-oriented and modality-specific manner so that the monocular/binocular quality perception issue and the multiple-distortion interaction issue can be well characterized.

A. Local Visual Primitive (LVP)

The goal of LVP learning is to simulate the properties of the cortical RFs. It has been discovered that the cortical RFs could be characterized as being spatially localized and oriented patterns [70]. Meanwhile, such properties of cortical RFs were found to be similar with the characteristics of the basis functions learned from natural images. In order to learn appropriate basis functions, a sparse coding approach with the over-complete dictionary has been presented in [70]. Performing sparse coding with an over-complete dictionary can lead to interesting interactions among the code elements, since sparsification weeds out those basis functions not needed to describe image structures. These interactions lead to deviations from a strictly linear input-output relationship, some of which have already been observed in the responses of cortical simple cells. An example of the learned basis functions (*i.e.*, LVPs) by the traditional sparse coding approach are shown in Fig. 1.

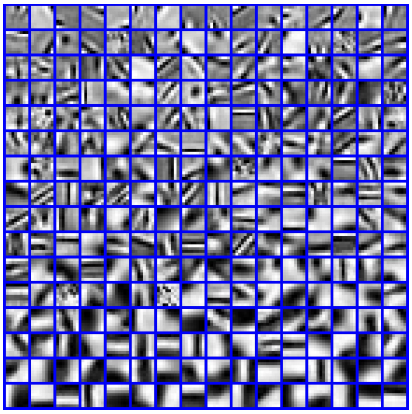


Fig. 1. The learned basis functions by the sparse coding algorithm.

B. Task-Oriented and Modality-Specific MB-LVP

1) Motivation: According to the existing studies in visual physiology [35]–[37], two types of RFs have been found in the visual cortex, *i.e.*, MRF and BRF. These two types of RFs work together in creating stereopsis when two monocular images with disparity are presented to the two eyes, respectively. In order to simulate such visual cortex-like MRFs and BRFs and adapt them to better address the NR-SDSIQA and NR-MDSIQA tasks, we are inspired to extend the above sparse coding approach based on the following principles:

- The MRF and BRF properties should be respectively simulated based on monocular and binocular stimuli.
- The RF properties in response to stimuli with different distortion modalities should be independently simulated.
- The simulation of RF properties should be adapted to the quality prediction task.

To be more specific, we in this paper propose a task-oriented (to account for the third principle) and modality-specific (to account for the second principle) dictionary learning framework to learn M-LVP and B-LVP (to account for the first principle) from monocular and binocular images, respectively.

2) Problem Formulation: We present an overview learning task-oriented and modality-specific M-LVPs and B-LVPs in Fig. 2. Without loss of generality, this figure is depicted in terms of a certain distortion modality as an example. Note that, each individual single and mixed distortion types are considered as different modalities in our method and the depicted process will be applied to all modalities. For each modality, the learning of M-LVP (B-LVP) is performed based on a set of distorted monocular patches (binocular patch pairs) along with their corresponding quality-discriminative codes. We formulate the task-oriented learning framework of M-LVP (B-LVP) associated with the k-th modality as follows:

$$\begin{aligned}&\left\langle \hat{\mathbf{D}}_k^{\mho}, \hat{\mathbf{W}}_k^{\mho}, \hat{\mathbf{A}}_k^{\mho} \right\rangle \\ &= \underset{\mathbf{D}_k^{\mho}, \mathbf{W}_k^{\mho}, \mathbf{A}_k^{\mho}}{\arg\min} \left(\left\| \mathbf{P}_k^{\mho} - \mathbf{D}_k^{\mho} \mathbf{A}_k^{\mho} \right\|_F^2 + \lambda \left\| \mathbf{S}_k^{\mho} - \mathbf{W}_k^{\mho} \mathbf{A}_k^{\mho} \right\|_F^2 \right), \\ &\quad s.t.\ \forall n, \quad \left\| \mathbf{a}_{k,n}^{\mho} \right\|_0 \leq \Psi, \end{aligned} \tag{1}$$

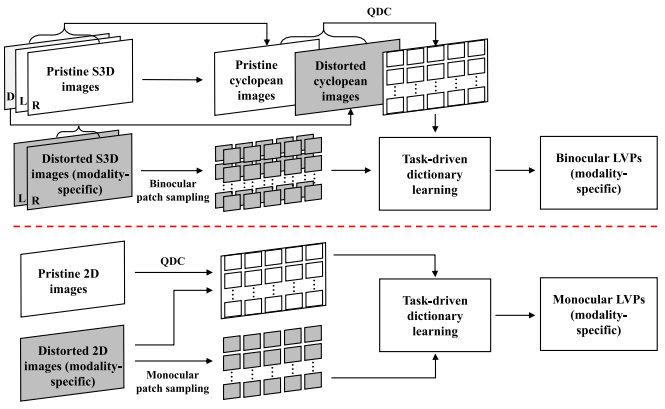


Fig. 2. Overview of learning M-LVPs and B-LVPs by using a task-oriented and modality-specific dictionary learning framework.

where the superscript symbol $\upsilon \in \{\mathcal{M}, \mathcal{B}\}$ indicates the monocular and binocular stimuli, the subscript index k indicates the k-th modality, λ is a parameter to balance the relative importance of reconstruction error (data-driven) and quality inconsistency (task-driven) penalties, and is set to 4 according to [75], Ψ is a positive constant indicating the sparsity, $\hat{\mathbf{D}}_k^{\upsilon} \in \mathbb{R}^{p_\upsilon \times d_k}$ is the learned LVPs over which the input patches $\mathbf{P}_k^{\upsilon} \in \mathbb{R}^{p_\upsilon \times n_k}$ have sparse representation codes $\hat{\mathbf{A}}_k^{\upsilon} \in \mathbb{R}^{d_k \times n_k}$, $\mathbf{S}_k^{\upsilon} \in \mathbb{R}^{d_k \times n_k}$ is the quality-discriminative code (QDC) of $\mathbf{P}_k^{\upsilon}$, $\hat{\mathbf{W}}_k^{\upsilon} \in \mathbb{R}^{d_k \times d_k}$ is a learned linear transformation matrix which encourages the original sparse codes $\hat{\mathbf{A}}_k^{\upsilon}$ to be most discriminative in terms of quality in the new space. Note that, for each input patch used for learning the LVPs, the QDC is a corresponding vector that is only determined by the quality score of this patch. We then convert such scalars (*i.e.*, quality scores) into vectors (*i.e.*, QDCs) so that the quality information can be well incorporated into LVP learning by the above task-oriented dictionary learning framework.

It is emphasized that the optimization of the above objective function will lead to a joint minimization of reconstruction error and quality inconsistency. It is expectable that the learned M-LVPs and B-LVPs in such a task-oriented and modality-specific optimization framework is able to well characterize the MRF and BRF properties of the visual cortex in responses to SDSIs and MDSIs. In the next, we first introduce how to generate monocular/binocular patches/patch pairs from 2D/3D images and their corresponding QDCs. Then, the optimization of (1) will be presented.

C. Training Data Generation

1) Training Data From Monocular Stimuli: From (1), we know that the learning of M-LVP requires both $\mathbf{P}_k^{\mathcal{M}}$ and $\mathbf{S}_k^{\mathcal{M}}$ as input. In order to generate the monocular patch set $\mathbf{P}_k^{\mathcal{M}}$, a subtractive and divisive local normalization method as in [45] is applied to each distorted 2D image. The normalized image $\hat{I}'$ is estimated by subtracting the local mean followed by dividing the local contrast of the distorted 2D image I':

$$\hat{I}'(x, y) = \frac{I'(x, y) - \mu(x, y)}{\sigma(x, y) + 1}, \tag{2}$$

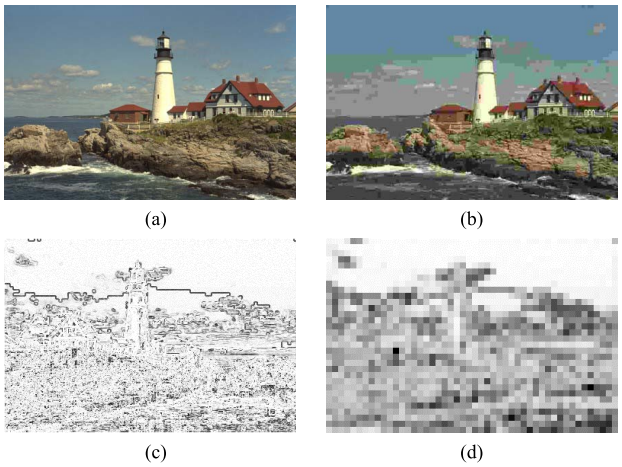


Fig. 3. Visualization of local quality estimation results on 2D images using FSIM: (a) pristine image, (b) distorted image, (c) pixel-wise FSIM map, (d) patch-level FSIM map, where the brighter areas indicate better quality.

where

$$\mu(x, y) = \sum_{h=-H}^{H} \sum_{w=-W}^{W} \beta_{\{h,w\}} I'_{\{h,w\}}(x, y), \tag{3}$$

$$\sigma(x, y) = \sqrt{\sum_{h=-H}^{H} \sum_{w=-W}^{W} \left(\beta_{\{h,w\}} (I'_{\{h,w\}}(x, y) - \mu(x, y))^2 \right)}, \tag{4}$$

are calculated to be the local mean and local contrast measures, and $\{\beta_{\{h,w\}} | h = -H, \ldots, H; w = -W, \ldots, W\}$ defines a unit-volume Gaussian window. This local normalization is found to well resemble the primate cortical visual process of the human brain [71], [72]. By local normalization, a set of modality-specific normalized images $\hat{\mathbf{I}}'_k = \{\hat{I}'_{k,1}, \hat{I}'_{k,2}, \ldots, \hat{I}'_{k,l_k}\}$ are generated, where l_k represents the total number of distorted 2D images associated with the k-th modality. Then, each normalized image is divided into non-overlapped patches of size $\sqrt{p} \times \sqrt{p}$. As a result, for the k-th modality, we can obtain an associated monocular patch set $\mathbf{P}_k^{\mathcal{M}} = [\mathbf{p}_{k,1}^{\mathcal{M}}, \mathbf{p}_{k,2}^{\mathcal{M}}, , \ldots, \mathbf{p}_{k,n_k}^{\mathcal{M}}] \in \mathbb{R}^{p_{\mathcal{M}} \times n_k}$, where $p_{\mathcal{M}} = p$, and n_k represents the total number of monocular patches extracted for the k-th modality.

To construct the QDC matrix $\mathbf{S}_k^{\mathcal{M}}$, we resort to Feature SIMilarity (FSIM) [3], a popular FR-IQA metric, which is able to provide a reasonable local quality measure [73]. By comparing a distorted 2D image I' with its pristine version I using FSIM, we can obtain a pixel-wise quality map. Then, the quality of a specific monocular patch $\mathbf{p}_{k,n}^{\mathcal{M}}$, $n = 1, 2, \ldots, n_k$ is estimated by averaging the FSIM scores of all the pixels inside this patch, resulting in patch-level quality scores $q_{k,n}^{\mathcal{M}} \in [0, 1]$. Fig. 3 (a)-(d) show examples of a pristine image, a distorted image, its pixel-wise quality map, and its patch-level quality map. Based on $q_{k,n}^{\mathcal{M}}$, the construction of QDC involves two steps: 1) quality level quantization, and 2) binary code assignment. 1) Quality level quantization: We define the quantified quality level of $q_{k,n}^{\mathcal{M}}$ as $z_{k,n}^{\mathcal{M}} \in \{0, 1, 2, \cdots, Z-1\}$ which is determined by:

$$z_{k,n}^{\mathcal{M}} = \left\lfloor Z \cdot q_{k,n}^{\mathcal{M}} \right\rfloor, \tag{5}$$

where the symbol $\lfloor \cdot \rfloor$ is the floor operation, Z is the number of quantified quality levels and is empirically set to 20. 2) Binary code assignment: Based on the quantified quality level $z_{k,n}^{\mathcal{M}}$, the QDC is described as a binary vector $\mathbf{s}_{k,n}^{\mathcal{M}} = [s_{k,n}^{\mathcal{M}}(1), s_{k,n}^{\mathcal{M}}(2), \ldots, s_{k,n}^{\mathcal{M}}(d_k)]^T \in \mathbb{R}^{d_k}$. It is obvious that each quality level $z_{k,n}^{\mathcal{M}}$ corresponds to d_k/Z elements in $\mathbf{s}_{k,n}^{\mathcal{M}}$. Therefore, for each quality level $z_{k,n}^{\mathcal{M}}$, only the corresponding d_k/Z elements in $\mathbf{s}_{k,n}^{\mathcal{M}}$ are assigned to be ones while the remaining elements are all assigned to be zeros, such that:

$$s_{k,n}^{\mathcal{M}}(i) = \begin{cases} 1, & \left\lfloor z_{k,n}^{\mathcal{M}} \cdot \frac{d_k}{Z} \right\rfloor < i \leq \left\lfloor (z_{k,n}^{\mathcal{M}} + 1) \cdot \frac{d_k}{Z} \right\rfloor \\ 0, & otherwise, \end{cases} \tag{6}$$

Finally, the QDC matrix $\mathbf{S}_k^{\mathcal{M}}$ is given as follows:

$$\begin{aligned} \mathbf{S}_k^{\mathcal{M}} &= \left[\mathbf{s}_{k,1}^{\mathcal{M}}, \mathbf{s}_{k,2}^{\mathcal{M}}, \ldots, \mathbf{s}_{k,n_k}^{\mathcal{M}}\right] \\ &= \begin{bmatrix} s_{k,1}^{\mathcal{M}}(1) & s_{k,2}^{\mathcal{M}}(1) & \cdots & s_{k,n_k}^{\mathcal{M}}(1) \\ s_{k,1}^{\mathcal{M}}(2) & s_{k,2}^{\mathcal{M}}(2) & \cdots & s_{k,n_k}^{\mathcal{M}}(2) \\ \vdots & \vdots & \cdots & \vdots \\ s_{k,1}^{\mathcal{M}}(d_k) & s_{k,2}^{\mathcal{M}}(d_k) & \cdots & s_{k,n_k}^{\mathcal{M}}(d_k) \end{bmatrix} \in \mathbb{R}^{d_k \times n_k}. \end{aligned} \tag{7}$$

The constructed QDCs provide an effective way to incorporate a task-oriented quality inconsistency penalty into the traditional LVP learning framework which accounts for only a data-driven sparse reconstruction error penalty.

2) Training Data From Binocular Stimuli: Similarly, the learning of B-LVP requires $\mathbf{P}_k^{\mathcal{B}}$ and $\mathbf{S}_k^{\mathcal{B}}$ as input. To generate the binocular patch pairs $\mathbf{P}_k^{\mathcal{B}}$, local normalization described in (2) is applied to both the left and right images of each distorted 3D image pair. Finally, for the k-th modality, we can obtain an associated binocular patch pair set $\mathbf{P}_k^{\mathcal{B}} = \left[[\mathbf{p}_{k,1}^{\mathcal{L}}, \mathbf{p}_{k,1}^{\mathcal{R}}]^T, [\mathbf{p}_{k,2}^{\mathcal{L}}, \mathbf{p}_{k,2}^{\mathcal{R}}]^T, \ldots, [\mathbf{p}_{k,n_k}^{\mathcal{L}}, \mathbf{p}_{k,n_k}^{\mathcal{R}}]^T\right] \in \mathbb{R}^{2p \times n_k}$. Note that, the two monocular patches from the left and right images are linked according to the reference disparity maps to form the binocular patch pairs.

To construct the QDC matrix $\mathbf{S}_k^{\mathcal{B}}$, a synthesized cyclopean image is first generated. From a perceptual sense, each stereo 3D image pair is merged into a single cyclopean view via binocular stereopsis. In the context of 3D-IQA, Chen *et al.* [14] have introduced a simplified model that synthesizes a cyclopean view from the left and right images of a stereopair by accounting for the critical binocular rivalry:

$$I'_{\mathcal{C}}(x, y) = \Phi_{\mathcal{L}}(x, y) \cdot I'_{\mathcal{L}}(x, y) + \Phi_{\mathcal{R}}(x+d, y) \cdot I'_{\mathcal{R}}(x+d, y), \tag{8}$$

where d is the pixel disparity between the reference left and right images $I_{\mathcal{L}}$ and $I_{\mathcal{R}}$, $\Phi_{\mathcal{L}}$ and $\Phi_{\mathcal{R}}$ are the normalized weights determined by Gabor filter response:

$$\Phi_{\mathcal{L}}(x, y) = \frac{E'_{\mathcal{L}}(x, y)}{E'_{\mathcal{L}}(x, y) + E'_{\mathcal{R}}(x+d, y)}, \tag{9}$$

$$\Phi_{\mathcal{R}}(x, y) = \frac{E'_{\mathcal{R}}(x+d, y)}{E'_{\mathcal{L}}(x, y) + E'_{\mathcal{R}}(x+d, y)}, \tag{10}$$

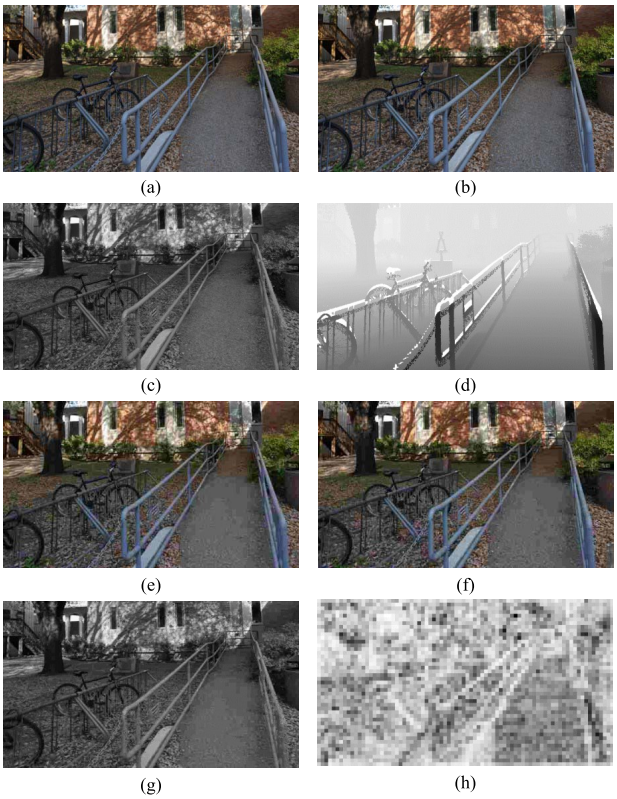


Fig. 4. Visualization of local quality estimation results on cyclopean images using the SSIM metric: (a) pristine left image, (b) pristine right image, (c) pristine cyclopean image synthesized from (a) and (b), (d) left disparity map, (e) distorted left image, (f) distorted right image, (g) distorted cyclopean image synthesized from (e) and (f), (h) patch-level SSIM map, where the brighter areas indicate better quality.

where $E'_{\mathcal{L}}$ and $E'_{\mathcal{R}}$ represent the response maps of $I'_{\mathcal{L}}$ and $I'_{\mathcal{R}}$, by deploying the Gabor filter banks described in [14]. Fig. 4 (c) and (g) show an example of the pristine and distorted cyclopean images. (c) is synthesized from the pristine left image in (a) and the pristine right image in (b), (g) is synthesized from the distorted left image in (e) and the distorted right image in (f). Note that, the reference left disparity map in (d) is utilized to support the synthesis process.

It has been experimentally demonstrated that the direct application of existing 2D FR-IQA metrics to cyclopean images can achieve a high consistency with subjective 3D quality perception and the popular structural similarity index (SSIM) metric [1] can provide reasonable performance within such a cyclopean framework [14]. Therefore, it is reasonable to estimate a SSIM-based quality map from the synthesized pristine and distorted cyclopean images for local quality measurement of binocular patch pairs. To be specific, for a certain binocular patch pair with the k-th distortion modality $\mathbf{p}_{k,n}^{\mathcal{B}} = [\mathbf{p}_{k,n}^{\mathcal{L}}, \mathbf{p}_{k,n}^{\mathcal{R}}]^T$, its quality $q_{k,n}^{\mathcal{B}}$ is computed as the average of the cyclopean image-based SSIM scores over the locations inside this patch:

$$q_{k,n}^{\mathcal{B}} = \frac{1}{\sqrt{p} \times \sqrt{p}} \sum_{(x,y) \in \mathbf{p}_{k,n}^{\mathcal{B}}} LQM_{SSIM}(x, y), \tag{11}$$

where $LQM_{SSIM}(x, y)$ represents a pixel-wise SSIM map estimated from the corresponding pristine and distorted cyclopean images. Fig. 4 (h) shows the patch-level SSIM map estimated from the pristine cyclopean image in (c) and distorted cyclopean image in (g). Based on $q_{k,n}^{\mathcal{B}}$, the final QDC matrix $\mathbf{S}_k^{\mathcal{B}} = [\mathbf{s}_{k,1}^{\mathcal{B}}, \mathbf{s}_{k,2}^{\mathcal{B}}, \ldots, \mathbf{s}_{k,n_k}^{\mathcal{B}}] \in \mathbb{R}^{d_k \times n_k}$ can be obtained in the same way according to Eqs (5)-(7).

As described, we have applied the FSIM (SSIM) metric to estimate the quality of monocular patches (binocular patch pairs). Actually, we have experimentally tested three popular FR-IQA metrics (*i.e.*, SSIM [1], GMSD [7], and FSIM [3]) and finally observed that 1) such metric combination (FSIM for monocular and SSIM for binocular) leads to best performance, and 2) the influence of different FR-IQA metrics is not obvious.

D. Optimization

For the optimization purpose, we further rewrite the objective function defined in (1) as follows:

$$\left\langle \hat{\mathbf{D}}_k^{\mho}, \hat{\mathbf{W}}_k^{\mho}, \hat{\mathbf{A}}_k^{\mho} \right\rangle = \underset{\hat{\mathbf{D}}_k^{\mho}, \hat{\mathbf{W}}_k^{\mho}, \hat{\mathbf{A}}_k^{\mho}}{\arg\min} \left\| \begin{bmatrix} \mathbf{P}_k^{\mho} \\ \sqrt{\lambda}\mathbf{S}_k^{\mho} \end{bmatrix} - \begin{bmatrix} \mathbf{D}_k^{\mho} \\ \sqrt{\lambda}\mathbf{W}_k^{\mho} \end{bmatrix} \mathbf{A}_k^{\mho} \right\|_F^2, \quad s.t.\ \forall n, \quad \left\| \mathbf{a}_{k,n}^{\mho} \right\|_0 \leq \Psi. \tag{12}$$

By defining $\mathbf{F}_k^{\mho} = [\mathbf{P}_k^{\mho}, \sqrt{\lambda}\mathbf{S}_k^{\mho}]^T$, $\mathbf{G}_k^{\mho} = [\mathbf{D}_k^{\mho}, \sqrt{\lambda}\mathbf{W}_k^{\mho}]^T$, the optimization of Eq. (12) is transformed to solve

$$\left\langle \hat{\mathbf{G}}_k^{\mho}, \hat{\mathbf{A}}_k^{\mho} \right\rangle = \underset{\mathbf{G}_k^{\mho}, \mathbf{A}_k^{\mho}}{\arg\min} \left\| \mathbf{F}_k^{\mho} - \mathbf{G}_k^{\mho}\mathbf{A}_k^{\mho} \right\|_F^2, \quad s.t.\ \forall n, \quad \left\| \mathbf{a}_{k,n}^{\mho} \right\|_0 \leq \Psi. \tag{13}$$

The above objective function can be well solved by the K-SVD algorithm [74]. Especially, K-SVD is a generalization of the k-means clustering method, and it works by iteratively alternating between sparse coding the input data based on the current dictionary, and updating the atoms in the dictionary to better fit the data samples. Before applying the K-SVD to solve this problem, both $\mathbf{D}_k^{\mho}$ and $\mathbf{W}_k^{\mho}$ need to be initialized. Towards this end, according to [75], we perform several iterations of K-SVD within each quantified quality level $z_{k,n}^{\mathcal{M}}$ and combine all the results to form the initial dictionary $\dot{\mathbf{D}}_k^{\mho}$ based on which the initial sparse codes $\dot{\mathbf{A}}_k^{\mho}$ for $\mathbf{P}_k^{\mho}$ are estimated by solving

$$\dot{\mathbf{a}}_{k,n}^{\mho} = \underset{\dot{\mathbf{a}}_{k,n}^{\mho}}{\arg\min} \left\| \mathbf{p}_{k,n}^{\mho} - \dot{\mathbf{D}}_k^{\mho}\mathbf{a}_{k,n}^{\mho} \right\|_2^2, \ s.t. \ \left\| \mathbf{a}_{k,n}^{\mho} \right\|_0 \leq \Psi, \tag{14}$$

where $\mathbf{p}_{k,n}^{\mho}$ is the n-th sample in $\mathbf{P}_k^{\mho}$ and $\dot{\mathbf{a}}_{k,n}^{\mho}$ is the n-th column of $\dot{\mathbf{A}}_k^{\mho}$. The classical orthogonal matching pursuit (OMP) algorithm [76] is utilized to get the solution of the above problem. Based on $\dot{\mathbf{A}}_k^{\mho}$, the multivariate ridge regression model with the quadratic loss and ℓ_2-norm regularization is applied to initialize $\mathbf{W}_k^{\mho}$, such that:

$$\dot{\mathbf{W}}_k^{\mho} = \underset{\dot{\mathbf{W}}_k^{\mho}}{\arg\min} \left\| \mathbf{S}_k^{\mho} - \mathbf{W}_k^{\mho}\dot{\mathbf{A}}_k^{\mho} \right\|_2^2 + \lambda_1 \left\| \mathbf{W}_k^{\mho} \right\|_F^2. \tag{15}$$

The above optimization problem actually has a closed-form solution which can be expressed as:

$$\dot{\mathbf{W}}_k^{\mho} = \mathbf{S}_k^{\mho} \left(\dot{\mathbf{A}}_k^{\mho} \right)^T \left(\dot{\mathbf{A}}_k^{\mho} \left(\dot{\mathbf{A}}_k^{\mho} \right)^T + \lambda_1 \mathbf{I} \right). \tag{16}$$

Once the initialization is completed, K-SVD is applied to get the solution of $\hat{\mathbf{G}}_k^{\mho}$ from which $\hat{\mathbf{D}}_k^{\mho} = [\hat{\mathbf{d}}_{k,1}^{\mho}, \hat{\mathbf{d}}_{k,2}^{\mho}, \cdots, \hat{\mathbf{d}}_{k,d_k}^{\mho}]$ can be obtained. However, the current $\hat{\mathbf{D}}_k^{\mho}$ still cannot be directly used for subsequent feature encoding because $\hat{\mathbf{D}}_k^{\mho}$ and $\hat{\mathbf{W}}_k^{\mho}$ are previously joint ℓ_2-normalized in $\hat{\mathbf{G}}_k^{\mho}$, *i.e.*, $\forall d, \|(\mathbf{d}_{k,d}^{\mho})^T, \sqrt{\lambda}(\mathbf{w}_{k,d}^{\mho})^T\|_2 = 1$. Finally, the desired LVP $\tilde{\mathbf{D}}_k^{\mho}$ can be calculated as:

$$\tilde{\mathbf{D}}_k^{\mho} = \left[\frac{\hat{\mathbf{d}}_{k,1}^{\mho}}{\|\hat{\mathbf{d}}_{k,1}^{\mho}\|_2}, \frac{\hat{\mathbf{d}}_{k,2}^{\mho}}{\|\hat{\mathbf{d}}_{k,2}^{\mho}\|_2}, \cdots, \frac{\hat{\mathbf{d}}_{k,d_k}^{\mho}}{\|\hat{\mathbf{d}}_{k,d_k}^{\mho}\|_2} \right], \tag{17}$$

where $k \in \{\text{GB, WN, JPEG, GB+JPEG+WN}\}$ indicates the distortion modality and $\mho \in \{\mathcal{M}, \mathcal{B}\}$ indicates the monocular and binocular stimuli. This ultimately leads to four M-LVPs (each M-LVP for GB, JPEG, WN, GB+JPEG+WN, respectively) and four B-LVPs (each for GB, JPEG, WN, GB+JPEG+WN, respectively). All these MB-LVPs will be used as the priors for feature encoding of a query stereopair to produce quality-aware features for quality regression.

E. Monocular and Binocular Feature Responses

1) Pixel Visibility Analysis: Given a query stereopair, we first classify all the pixels into the monocular and binocular ones, according to their visibility in the left and right views. For example, a pixel will be classified as monocular if it is only visible in either the left or the right view, while it will be classified as binocular if it is visible in both the left and right views. Consequently, all the pixels belonging to each class constitute the left monocular region (LMR), right monocular region (RMR), left binocular region (LBR), and right binocular region (RBR), respectively.

For the consideration of efficiency, we resort to an existing method for pixel visibility analysis of stereopairs [77]. A pixel in the left image $p_{\mathcal{L}} = (p_{\mathcal{L},x}, p_{\mathcal{L},y})$ is classified into LBR if the following two constraints are both satisfied:

$$0 \leq p_{\mathcal{L},x} + d_{\mathcal{L}}(p_{\mathcal{L}}) < R_w; \tag{18}$$

$$\forall q_{\mathcal{L}} | (q_{\mathcal{L},x} > p_{\mathcal{L},x}) \cap (q_{\mathcal{L},y} = p_{\mathcal{L},y}), \quad p_{\mathcal{L},x} + d_{\mathcal{L}}(p_{\mathcal{L}}) \neq q_{\mathcal{L},x} + d_{\mathcal{L}}(q_{\mathcal{L}}). \tag{19}$$

where R_w is the width of the image, and $q_{\mathcal{L}}$ represents a certain pixel on the right side of $p_{\mathcal{L}}$. Actually, Eq. (18) verifies that $p_{\mathcal{L}}$ stays within the image bound, and Eq. (19) ensures that $p_{\mathcal{L}}$ is not occluded in the right view. Once $p_{\mathcal{L}}$ is classified into LBR, its corresponding pixel on the right image will be classified into RBR. Both LBR and RBR constitute the overall BR. Then, the rest pixels in the left and right images are classified as LMR and RMR, respectively. An example is presented in Fig. 5 where the regions marked in blue indicate the LMR, the regions marked in green indicate the RMR, and the regions marked yellow indicate the BR.

We acknowledge that, the visibility analysis is dependent on the estimated disparity maps which can be somewhat

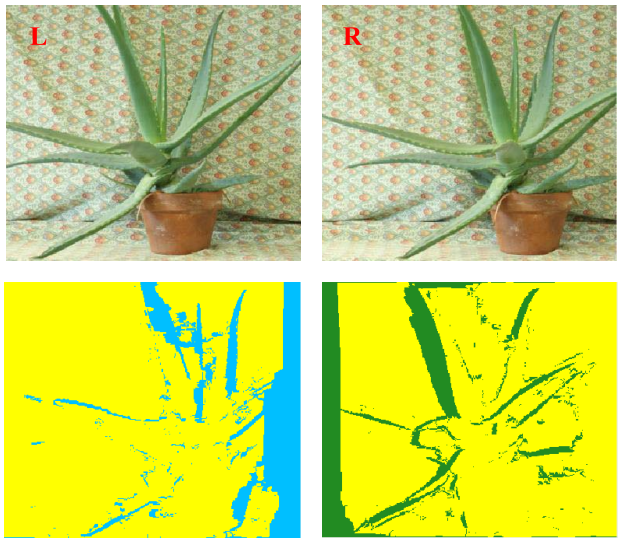


Fig. 5. An example of monocular and binocular regions. (a) left image, (b) right image. The blue regions indicate the LMR, the green regions indicate the RMR, the yellow regions indicate the BR. Best viewed in color version.

problematic especially for the severe distortion case. However, experimental results in Section-IV.E have revealed that our method can tolerate modest inaccuracy of disparity maps and still deliver better performance in comparison with those without considering the discrepancies between monocular and binocular regions in terms of the neural coding strategy.

2) Feature Encoding: As stated beforehand, the stimuli in MR and BR of a stereopair will be processed by the MRFs and BRFs in the visual cortex, respectively. For this consideration, the monocular patches centered at each pixel inside the MR (LMR and RMR) are encoded uisng the learned M-LVPs, while the binocular patch pairs centered at each pixel inside the BR (LBR and RBR) are encoded using the learned B-LVPs.

For the monocular case, a monocular patch of size $\sqrt{p}\times\sqrt{p}$ centered at pixel $p_{\mathcal{L}}^{\mathcal{M}} \in$ LMR ($p_{\mathcal{R}}^{\mathcal{M}} \in$ RMR) is denoted by $\mathbf{p}_{\mathcal{L}}^{\mathcal{M}} \in \mathbb{R}^{p\times 1}$ ($\mathbf{p}_{\mathcal{R}}^{\mathcal{M}} \in \mathbb{R}^{p\times 1}$). The neural coding process is simply approximated by sparse coding, such that:

$$\hat{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}} = \arg\min_{\hat{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}}} \left\|\mathbf{p}_{\mathcal{L}}^{\mathcal{M}} - \tilde{\mathbf{D}}_k^{\mathcal{M}}\mathbf{a}_{\mathcal{L},k}^{\mathcal{M}}\right\|_2^2, \quad s.t. \ \left\|\mathbf{a}_{\mathcal{L},k}^{\mathcal{M}}\right\|_0 \le \Psi, \tag{20}$$

$$\hat{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}} = \arg\min_{\hat{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}}} \left\|\mathbf{p}_{\mathcal{R}}^{\mathcal{M}} - \tilde{\mathbf{D}}_k^{\mathcal{M}}\mathbf{a}_{\mathcal{R},k}^{\mathcal{M}}\right\|_2^2, \quad s.t. \ \left\|\mathbf{a}_{\mathcal{R},k}^{\mathcal{M}}\right\|_0 \le \Psi, \tag{21}$$

where $\hat{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}}$ ($\hat{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}}$) represents the monocular response of $\mathbf{p}_{\mathcal{L}}^{\mathcal{M}}$ ($\mathbf{p}_{\mathcal{R}}^{\mathcal{M}}$) with respect to the k-th M-LVP $\tilde{\mathbf{D}}_k^{\mathcal{M}}$. Then, max-pooling is applied to obtain $\bar{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}}$ and $\bar{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}}$:

$$\bar{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}} = \max\left[\hat{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}}(1), \hat{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}}(2), \cdots, \hat{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}}(N_{\mathcal{L}})\right], \tag{22}$$

$$\bar{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}} = \max\left[\hat{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}}(1), \hat{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}}(2), \cdots, \hat{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}}(N_{\mathcal{R}})\right], \tag{23}$$

where the mathematical operator max is performed on each dimension of $\hat{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}}(i)$, $i = 1, 2, \cdots, N_{\mathcal{L}}$ and $\hat{\mathbf{a}}_{\mathcal{R},k}^{\mathcal{M}}(j)$, $j = 1, 2, \cdots, N_{\mathcal{R}}$, $N_{\mathcal{L}}$ and $N_{\mathcal{R}}$ represent the total number of pixels contained in LMR and RMR, respectively.

For the binocular case, a monocular patch of size $\sqrt{p}\times\sqrt{p}$ centered at pixel $p_{\mathcal{L}}^{\mathcal{B}} \in$ LBR and its corresponding patch in the right image constitute a binocular patch pair denoted by $\mathbf{p}^{\mathcal{B}} = [\mathbf{p}_{\mathcal{L}}^{\mathcal{B}}, \mathbf{p}_{\mathcal{R}}^{\mathcal{B}}] \in \mathbb{R}^{2p\times 1}$. With sparse coding, the neural coding response of $\mathbf{p}^{\mathcal{B}}$ is similarly computed as follows:

$$\hat{\mathbf{a}}_k^{\mathcal{B}} = \arg\min_{\hat{\mathbf{a}}_k^{\mathcal{B}}} \left\|\mathbf{p}^{\mathcal{B}} - \tilde{\mathbf{D}}_k^{\mathcal{B}}\mathbf{a}_k^{\mathcal{B}}\right\|_2^2, \quad s.t. \ \left\|\mathbf{a}_k^{\mathcal{B}}\right\|_0 \le \Psi, \tag{24}$$

Finally, we can obtain a max-pooled binocular response vector denoted by $\bar{\mathbf{a}}_k^{\mathcal{B}}$. As a highly efficient algorithm, the batch-OMP algorithm [78] is implemented to get the solution.

F. Cross-Modality Feature Response Aggregation

Besides the characterization of local MRF and BRF properties, another challenge in NR-MDSIQA is to model the effect of interactions among different distortion types. We propose to address this problem based on a simple yet effective linear combination framework where the weights are determined by the estimated modality-specific SRE. Take $\mathbf{p}_{\mathcal{L}}^{\mathcal{M}}$ as an example, the corresponding SRE is computed as follows:

$$e_{\mathcal{L},k}^{\mathcal{M}} = \left\|\mathbf{p}_{\mathcal{L}}^{\mathcal{M}} - \tilde{\mathbf{D}}_k^{\mathcal{M}}\hat{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}}\right\|_2^2, \tag{25}$$

Then, the SRE-based weights can be derived and the finally aggregated left monocular response vector is also computed by:

$$\bar{\mathbf{a}}_{\mathcal{L}}^{\mathcal{M}} = \sum_k \bar{\mathbf{a}}_{\mathcal{L},k}^{\mathcal{M}} \cdot \exp\left(-\frac{\sum_{n=1}^{N_{\mathcal{L}}} e_{\mathcal{L},k}^{\mathcal{M}}(n)}{N_{\mathcal{L}}}\right), \tag{26}$$

The finally aggregated right monocular response vector $\bar{\mathbf{a}}_{\mathcal{R}}^{\mathcal{M}}$ and binocular response vector $\bar{\mathbf{a}}^{\mathcal{B}}$ can be computed in a similar manner. The aggregated left and right monocular response vectors are further combined to form a final monocular response vector $\bar{\mathbf{a}}^{\mathcal{M}} = [\bar{\mathbf{a}}_{\mathcal{L}}^{\mathcal{M}}, \bar{\mathbf{a}}_{\mathcal{R}}^{\mathcal{M}}]$. As observed from Eq. (26), the weights decrease with increasing SREs. The rationale is that, when encoding a patch using all the learned modality-specific LVPs, a larger modality-specific SRE implies a weaker capacity of this specific modality in representing the patch, thus a smaller weight is assigned to this modality accordingly. In order to better demonstrate the effectiveness of the proposed SRE-based weighting scheme, we compute the modality-specific SREs as features based on which a distortion type classifier is built by using support vector machine (SVM) [79]. Here, we investigate the median accuracies of this distortion type classification across 100 train-test trials on the LIVE 3D Phase-I database (only the GB, JPEG, and WN distortions are considered). Table I lists the results of the classification accuracy for each distortion type. It is obvious that the classification accuracies are much better than random guess, which actually validate the rationality of using SREs as the basis for determining the weights in cross-modality aggregation.

The proposed cross-modality aggregation scheme actually provides a unified and effective way to characterize 1) the masking effect of different distortion types (for multiple distortion), and 2) the particularity of each individual distortion type (for single distortion). For the multiple distortion case,

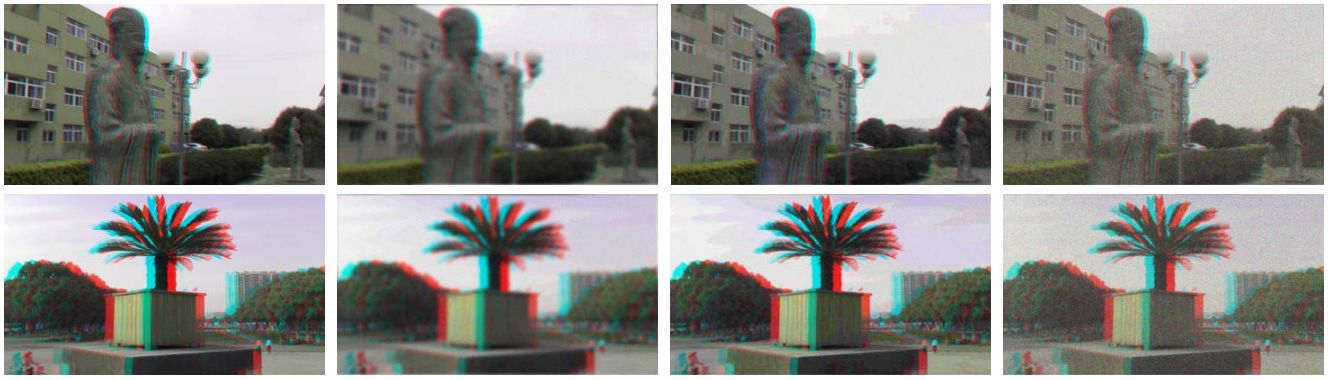


Fig. 6. An illustration of the masking effect of different distortion types in MDSIs. (a) pristine MDSIs, (b) GB-dominant MDSIs, (c) JPEG-dominant MDSIs, and (d) WN-dominant MDSIs. The stereopairs are visualized in an anaglypic format.

TABLE I
MEDIAN CLASSIFICATION ACCURACY(%) ACROSS 100 TRAIN-TEST TRIALS ON THE LIVE 3D PHASE-I DATABASE

Distortion Type	GB	JPEG	WN
Median Accuracy	84.44	82.50	86.25
Standard Derivation	8.80	7.50	6.25

a specific distortion type may play a dominant role and the other types are somewhat masked. Therefore, larger SREs are produced for those masked distortion modalities. To facilitate understanding, examples are presented in Fig. 6 where the two samples in columns (a) and (b) are selected from the MDSID database. The rows from top to bottom correspond to the pristine stereoscopic image, GB-dominant MDSI, JPEG-dominant MDSI, and WN-dominant MDSI, respectively. For the single distortion case, it is obvious that each individual distortion type shows an appearance with strong particularity and therefore the modality associated with smallest SRE is considered to have the largest weight in this regard.

G. Quality Inference

After achieving the finally aggregated monocular and binocular response vectors $\bar{\mathbf{a}}^{\mathcal{M}}$ and $\bar{\mathbf{a}}^{\mathcal{R}}$, a quality predictor is built via support vector regression (SVR) as in the relevant NR-IQA works [43]–[47], [49], [51], [54], [57]–[62]. Specifically, a SVR model is learned based on a set of distorted stereo images along with their corresponding subjective rating scores. The learned SVR model is used to evaluate the quality of any testing samples. We use the LIBSVM package [79] to implementSVR.

IV. EXPERIMENTAL RESULTS

In this section, we analyze the proposed method's capability to predict stereo image quality by testing several SDSI and MDSI databases. First, we present the details of training data collection, and introduce the benchmark databases as well as the evaluation protocols. We also compare the performance of the proposed method against other relevant NR-IQA algorithms. Finally, we evaluate the effectiveness of some key components in the proposed method.

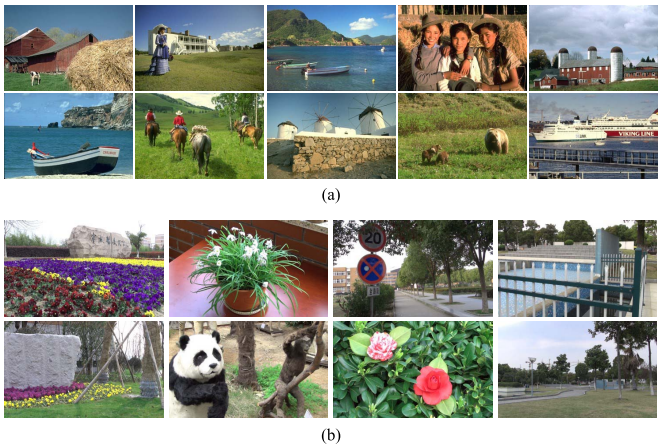


Fig. 7. The pristine 2D and 3D images. (a) 2D images selected from the Berkeley Segmentation Data Set (BSDS500) [80], (b) 3D images (only the left images are presented) captured by ourselves.

A. Training Data Collection

As mentioned in Section III-C, the learning of MB-LVPs requires monocular patches and binocular patch pairs along with their corresponding QDCs as input. For the training data collection from monocular stimuli, three types of distortions (*i.e.*, GB, JPEG, WN) are added either singly or multiply to ten 2D natural images (shown in Fig. 7(a)) selected from the Berkeley Segmentation Data Set (BSDS500) [80] at four distortion levels, which finally leads to 120 singly-distorted (*i.e.*, GB, JPEG, WN) and 640 multiply-distorted (*i.e.*, GB+JPEG+WN) 2D images. For the training data collection from binocular stimuli, three types of distortions (*i.e.*, GB, JPEG, WN) are added either singly or multiply to eight 3D natural images (shown in Fig. 7(b)) captured by ourselves at four distortion levels, which finally leads to 96 singly-distorted (*i.e.*, GB, JPEG, WN) and 512 multiply-distorted (*i.e.*, GB+JPEG+WN) 3D images. To be specific, for the simulation of GB, 2D Gaussian kernels with standard deviation σ_G were used for blurring with a square kernel window of side $3 \times \sigma_G$ using the Matlab *fspecial* and *imfilter* functions. For the simulation of JPEG compression, the Matlab function *imwrite* was used to produce JPEG compressed

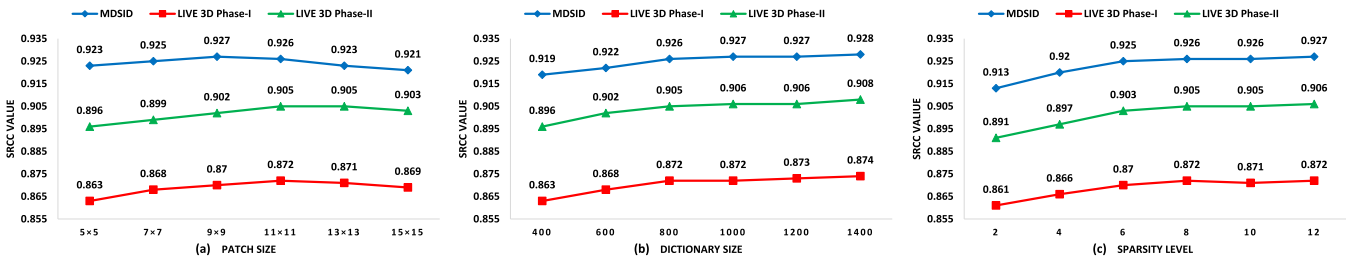


Fig. 8. The SRCC results obtained by different parameter settings. (a) SRCC values of different patch sizes, (b) SRCC values of different dictionary sizes, and (c) SRCC values of different sparsity levels.

TABLE II
LIST OF THE PARAMETER SETTINGS IN DISTORTION SIMULATION

Distortion Type	Parameter	Level-1	Level-2	Level-3	Level-4
GB	σ_G	3.2	3.8	4.4	5.0
JPEG	Q	27	22	17	12
WN	σ_N^2	0.002	0.008	0.032	0.128

images by varying the quality parameter Q as specified in the JPEG standard. For the simulation of WN, the Matlab function *imnoise* was used to produce noise images by adding the noise generated from a standard normal distribution of variance σ_N^2. All these parameter settings have been provided in Table II. They were selected in a way that the resulting distorted images were perceptually separable from each other and from the reference ones in the sense of perceived quality. In addition, following the previous relevant works [55], [81], for multiple distortion simulation, the GB is simulated first, followed by JPEG compression, and finally the WN injection.

With the generated distorted 2D and 3D images at hand, and following the processes described in Section III-C, a monocular patch set (binocular patch pair set) and the corresponding QDC set are generated for each modality (*i.e.*, GB, JPEG, WN, GB+JPEG+WN) and served as the input data for task-driven and modality-specific M-LVP (B-LVP) learning. In the implementation, the monocular patches (binocular patch pairs) are selected within each distortion modality to guarantee the involved distortion levels span the whole quality scale ranging from the worst to the best. The patch size is set to be 11×11 (p=121), the number of the learned modality-specific LVP is set to be 800 (d_k=800), the sparsity is set to be 8 (Ψ=8), and the balance parameter in Eq. (1) is set to be 4 (λ=4) according to [75]. Actually, all these parameters are involved only in the LVP learning stage. Once the MB-LVPs have been learned, the testing stage (*i.e.*, feature encoding and quality inference) is free of these parameters except for the sparsity level Ψ. Even so, we still test the results obtained by different parameter settings, as shown in Fig. 8. By observing the performance variations to different parameter values, we find that the chosen parameters generally work well in most cases.

B. Database and Protocol

For performance evaluation, three databases including LIVE 3D Phase-I [82], LIVE 3D Phase-II [59], and MDSID [69], are used as benchmarks. LIVE 3D Phase-I and Phase-II contain only SDSIs and the subjective scores of each SDSI in the form of DMOS. The difference between these two is that, SDSIs in LIVE 3D Phase-I are corrupted by symmetric single distortion, while SDSIs in LIVE 3D Phase-II are corrupted by either symmetric or asymmetric single distortion. The MDSID database contains both SDSIs and MDSIs along with the DMOS values. Note that, MDSIs in MDSID database are corrupted by symmetric multiple distortions (GB+JPEG+WN). Overall, in view of the scene (different reference images) and distortion (symmetric/asymmetric, single/multiple) diversities of these three databases, the performance evaluation on them is considered to be comprehensive.

For performance evaluation on each database, 100 trails of train-test process are conducted and the median results over 100 trails are reported to best avoid the performance bias. Each trail involves randomly splitting the database into two non-overlap subsets: 80% samples out of the entire database for training and the remaining 20% for testing. In this paper, the used performance criteria include: Pearson's linear correlation coefficient (PLCC), Spearman's rank-order correlation coefficient (SRCC), and Root mean square error (RMSE). A better model should deliver higher PLCC and SRCC values but lower RMSE value. Before computing the performance criteria, a logistic function is applied first to bring the prediction values to the same scale of the DMOS values [83],

$$Q' = \alpha_1\left(\frac{1}{2} - \frac{1}{1+\exp\left(\alpha_2(Q-\alpha_3)\right)}\right) + \alpha_4 Q + \alpha_5, \quad (27)$$

where Q is the predict score by the algorithm, and α_1, α_2, α_3, α_4, and α_5 are the parameters to be fitted. Note that, this regression step will only affect the PLCC and RMSE results.

C. Evaluation on MDSID

In the experiments, we consider all the MDSIs in the entire MDSID database for training and testing. We compare the proposed method with two NR-SDIQA methods (BRISQUE [45], GM-LOG [46]), two NR-MDIQA methods (SISBLIM [55], Color-JET [58]), three NR-SDSIQA methods (3D-DQE [64], StereoINQ [65], 3D-DNCSAE [67]), and one NR-MDSIQA method (MUMBLIM) [69]. We adapt the compared NR-SDIQA and NR-MDIQA methods to 3D case as follows: for SISBLIM which is actually training-free, the left and right views of a MDSI are first evaluated separately, resulting in two individual quality scores whose mean value is computed as the final predict score; for BRISQUE, GM-LOG,

TABLE III
PERFORMANCE RESULTS ON MDSID. FOR EACH CRITERION, THE BEST VALUE IS MARKED IN BOLDFACE

Method	PLCC	SRCC	RMSE
BRISQUE	0.925	0.913	3.551
GM-LOG	0.934	0.919	3.345
SISBLIM	0.835	0.828	5.265
Color-JET	0.932	0.916	3.439
3D-DQE	0.934	0.920	3.348
StereoINQ	0.907	0.905	3.943
3D-DNCSAE	0.931	0.914	3.447
MUMBLIM	0.878	0.882	4.570
Proposed	**0.938**	**0.926**	**3.062**

and Color-JET, the features extracted from the left and right views are combined into an overall feature vector for training and testing. The performance results of all the competing methods on MDSID are tabulated in Table III where the best values are highlighted in boldface. In addition, to understand whether the advantages of the proposed method over the competitors are statistically significant, the one-sided t-test was conducted as in [46], [64], and [65]. In practice, the one-sided t-test was conducted to test the equivalence of the mean values of two samples drawn from independent populations of a normal distribution. It was performed at a significance level of 0.01 using the 100 SRCC values of all pairs of compared models. The null hypothesis is that the SRCC values of the pair of models are drawn from populations with equal mean. The alternative hypothesis is that the mean of one model is greater than the other. The results are presented in Table IV where "1" and "−1" indicate the row model is statistically better and worse than the column model, respectively, while "0" indicates the row and column models are statistically equivalent.

It can be observed from the tables that our proposed method performs the best in terms of each criterion among all the competing methods. In addition, more observations can be illustrated as follows. First, the two popular 2D NR-IQA methods, *i.e.*, BRISQUE and GM-LOG, with a simple extension, can achieve rather competitive performance in evaluating MDSIs, although they are not designed for handling neither the multiple distortion nor stereo 3D case. The SISBLIM method, a representative method for 2D NR-MDIQA, performs the worst among all the methods. It is expectable because SISBLIM is a training-free metric. The Color-JET method, although specifically designed for 2D multiple distortion, performs slightly worse than GM-LOG in the 3D multiple distortion case. These results support our statement that the quality issues caused by multiple distortions in 2D and 3D images are different. The three compared NR-SDSIQA methods, *i.e.*, 3D-DQE, StereoINQ, and 3D-DNCSAE, have shown different abilities for evaluating MDSIs: *i.e.*, 3D-DQE and 3D-DNCSAE perform much better than StereoINQ. The reason may be that both 3D-DQE and 3D-DNCSAE take the advantages of deep learning techniques in different ways. However, they still take hybrid NSS features as input and the limitations of using existing NSS features to quantify the mixed distortion type are not well addressed. This also indicates that the investigations on effective NSS features for the measurement of multiple distortion suffered by stereopairs need further research effort. Towards the circumvention of exploring potential NSS for evaluating MDSIs, the MUMBLIM method tries to construct an implicit mapping function for space transfer in a multi-modal sparse representation framework. It is claimed that the interactions between different distortion types can be characterized by exploiting a joint sparse representation of each modality and the modality-specific space transfer also can be differentially treated via the joint optimization. Since MUMBLIM also does not require subjective ratings for training, it only delivers moderate performance. Moreover, the different roles of MRFs and BRFs in stereo perception are not differentially characterized in MUMBLIM.

With the similar consideration, our proposed method also avoids extracting any assumed NSS features from the input stereopairs. Instead, the used quality-aware features are obtained via automatic feature encoding with respect to the pre-learned MB-LVPs. Due to the task-oriented and modality-specific MB-LVP learning, the underlying monocular and binocular primitive representations in response to different distortion modalities suitable for feature encoding in NR-SIQA tasks can be well built and the interactions between different distortion types can be reasonably approximated by the proposed SRE-based weighting scheme as well.

D. Evaluation on LIVE 3D Phase-I and Phase-II

As mentioned, the LIVE 3D Phase-I and Phase-II databases contain only SDSIs. Therefore, experiments on these two databases are conducted to ascertain the ability of a specific quality model to evaluate SDSIs. In the experiments, we only consider the stereopairs corrupted by one of the three distortion types (*i.e.*, GB, JPEG, WN) for training and testing. We compare the proposed method with seven state-of-the-art NR-SIQA algorithms which are all designed for evaluating the visual quality of SDSIs. The compared seven NR-SIQA algorithms are Chen's method (Chen-TIP) [59], Su's method (S3D-BLINQ) [60], Apinna's method (StereoQUE) [61], Zhou's method (Zhou-TMM) [62], Liu's method (StereoINQ) [65], Jiang's method (3D-DNCSAE) [67], and Oh's method (DNR-S3DIQE) [68]. The individual distortion type performance results as well as the averaged ones in terms of PLCC, SRCC, and RMSE on the two databases are summarized in Table V and Table VI, respectively. To facilitate presentation, the top three values are marked in boldface.

It can be seen that the proposed method performs quite stably on both databases as it always ranks top three for all the cases except the SRCC value for the GB subset in LIVE 3D Phase-I and the PLCC value for the JPEG subset in LIVE 3D Phase-II. When comparing the averaged results, our method performs the best on LIVE 3D Phase-I in terms of PLCC and SRCC while takes the third place on RMSE (the best two RMSE values are obtained by StereoINQ and S3D-BLINQ). However, it needs to be noticed that StereoINQ is outside the top three in terms of SRCC and S3D-BLINQ is outside the top three in terms of PLCC, which inversely neutralize their slight advantages in RMSE. Another point needs to be emphasized is that, StereoINQ does not provide

TABLE IV

ONE-SIDED T-TEST RESULTS ON MDSID. IN THE TABLE, "1" AND "−1" INDICATE THE ROW MODEL IS STATISTICALLY BETTER AND WORSE THAN THE COLUMN MODEL, RESPECTIVELY, WHILE "0" INDICATES THE ROW AND COLUMN MODELS ARE STATISTICALLY EQUIVALENT

Method	BRISQUE	GM-LOG	SISBLIM	Color-JET	3D-DQE	StereoINQ	3D-DNCSAE	MUMBLIM	Proposed
BRISQUE	0	-1	1	-1	-1	1	0	1	-1
GM-LOG	1	0	1	1	0	1	1	1	-1
SISBLIM	-1	-1	0	-1	-1	-1	-1	-1	-1
Color-JET	1	-1	1	0	-1	1	1	1	-1
3D-DQE	1	0	1	1	0	1	1	1	-1
StereoINQ	-1	-1	1	-1	-1	0	-1	1	-1
3D-DNCSAE	0	-1	1	-1	-1	1	0	1	-1
MUMBLIM	-1	-1	1	-1	-1	-1	-1	0	-1
Proposed	1	1	1	1	1	1	1	1	0

TABLE V

PERFORMANCE RESULTS ON LIVE 3D PHASE-I. FOR EACH CRITERION, THE TOP THREE VALUES ARE MARKED IN BOLDFACE

Criteria	Method	GB	JPEG	WN	Average
PLCC	Chen-TIP	0.917	0.695	0.917	0.843
	S3D-BLINQ	0.953	0.746	**0.961**	**0.887**
	StereoQUE	0.881	**0.806**	0.919	0.869
	Zhou-TMM	**0.973**	0.695	0.945	0.871
	StereoINQ	**0.967**	0.734	**0.970**	**0.891**
	3D-DNCSAE	0.956	0.739	0.946	0.880
	DNR-S3DIQE	0.950	**0.767**	0.910	0.876
	Proposed	**0.968**	**0.795**	**0.948**	**0.904**
SRCC	Chen-TIP	0.878	0.617	0.919	0.805
	S3D-BLINQ	0.791	0.603	0.906	0.767
	StereoQUE	0.865	**0.782**	0.910	**0.852**
	Zhou-TMM	**0.916**	0.614	0.915	0.815
	StereoINQ	0.883	0.656	**0.954**	0.831
	3D-DNCSAE	**0.938**	0.662	**0.932**	0.844
	DNR-S3DIQE	**0.930**	**0.765**	0.921	**0.872**
	Proposed	0.915	**0.774**	**0.927**	**0.872**
RMSE	Chen-TIP	5.898	4.523	6.433	5.618
	S3D-BLINQ	4.326	**3.959**	**3.931**	**4.072**
	StereoQUE	6.938	4.391	6.664	5.998
	Zhou-TMM	**3.127**	4.286	5.086	4.166
	StereoINQ	**3.554**	4.049	**3.834**	**3.812**
	3D-DNCSAE	4.206	**4.005**	5.083	4.431
	DNR-S3DIQE	-	-	-	-
	Proposed	**3.548**	**3.740**	**5.079**	**4.122**

TABLE VI

PERFORMANCE RESULTS ON LIVE 3D PHASE-II. FOR EACH CRITERION, THE TOP THREE VALUES ARE MARKED IN BOLDFACE

Criteria	Method	GB	JPEG	WN	Average
PLCC	Chen-TIP	0.941	**0.901**	0.947	0.930
	S3D-BLINQ	0.968	**0.888**	0.953	**0.936**
	StereoQUE	0.878	0.829	0.920	0.876
	Zhou-TMM	**0.983**	0.757	0.936	0.892
	StereoINQ	**0.984**	0.871	**0.970**	**0.942**
	3D-DNCSAE	0.963	**0.874**	**0.966**	0.934
	DNR-S3DIQE	0.934	0.821	0.836	0.864
	Proposed	**0.972**	0.873	**0.968**	**0.938**
SRCC	Chen-TIP	0.900	**0.867**	0.950	0.906
	S3D-BLINQ	0.903	0.818	0.946	0.889
	StereoQUE	0.846	0.839	0.932	0.872
	Zhou-TMM	0.903	0.593	0.891	0.796
	StereoINQ	**0.909**	0.839	**0.957**	**0.902**
	3D-DNCSAE	**0.918**	**0.851**	**0.956**	**0.908**
	DNR-S3DIQE	0.889	0.822	0.833	0.848
	Proposed	**0.915**	**0.842**	**0.959**	**0.905**
RMSE	Chen-TIP	4.725	**3.342**	3.513	3.860
	S3D-BLINQ	4.453	4.169	3.547	4.056
	StereoQUE	6.662	4.756	4.325	5.248
	Zhou-TMM	**2.455**	4.502	3.575	**3.511**
	StereoINQ	**2.481**	3.476	**2.519**	**2.825**
	3D-DNCSAE	4.512	**3.359**	**2.861**	3.577
	DNR-S3DIQE	-	-	-	-
	Proposed	**3.550**	**3.361**	**2.692**	**3.201**

satisfactory performance on the MDSID database. Although our method indeed does not provide the best performance on LIVE 3D Phase-II, the PLCC, SRCC, and RMSE values all take the second place. Given that our proposed method is designed to be a unified method for both NR-SDSIQA and NR-MDSIQA applications, we believe such performance results on singly-distorted stereo image quality databases are still competitive as a reasonable choice in the cases where the distortion profile (single or multiple) of stereopairs is unkown.

E. Model Ablation Analysis

Compared to the previous works, our proposed method has three new components (modules) which make the method particularly suitable for evaluating both SDSIs and MDSIs in a unified manner. The three components include 1) learning M-LVPs and B-LVPs from monocular and binocular stimuli, respectively; 2) learning M-LVPs and B-LVPs in a task-oriented and modality-specific manner; 3) computing modality-specific SREs as the combination weights for cross-modality feature response aggregation.

We are interested to understand the contribution of each component as indicated above. Towards this end, we implement three ablation models to: 1) verify the contribution of B-LVP learning, *i.e.*, with/without B-LVPs for feature encoding, 2) verify the contribution of Task-Oriented Penalty (TOP), *i.e.*, with/without task-oriented penalty in Eq. (1), and 3) verify the contribution of modality-specific SRE-based weighting scheme, *i.e.*, use average pooling (AVE), max pooling (MAX), and SRE-based pooling (SRE) in Eq. (26). The comparison results are shown in Fig. 9. It can be observed that, as compared to our proposed one which takes all these components into account, without each of the above components leads to performance deterioration at varying degrees. All these results support our contributions and all these components together make our method an effective solution for unified NR quality evaluation of both SDSIs and MDSIs.

F. Validation of M-LVPs on 2D Image Databases

Although the above model ablation analyses have demonstrated that combining M-LVPs and B-LVPs for respective

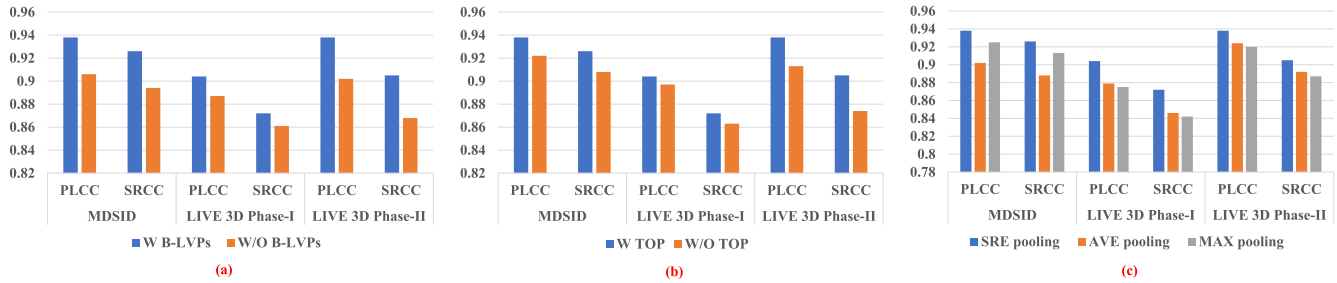


Fig. 9. Performance results of different model ablations. (a) with/without using B-LVPs for feature encoding, (b) with/without combining task-oriented penalty (TOP) for LVP learning, (c) use average pooling, max pooling, and SRE-based pooling for cross-modality response aggregation.

TABLE VII
PERFORMANCE RESULTS ON 2D IMAGE DATABASES

Database	Distortion	PLCC	SRCC	RMSE
CSIQ	GB	0.9146	0.9015	0.1227
	JPEG	0.9172	0.8957	0.1123
	WN	0.9114	0.8908	0.1130
	All	0.9133	0.8925	0.1128
LIVE	GB	0.9572	0.9396	7.8986
	JPEG	0.9224	0.9057	10.5940
	WN	0.9501	0.9358	8.6358
	All	0.9190	0.9023	10.6971
LIVE-MD	GB+JPEG	0.8965	0.8772	8.2528
	GB+WN	0.9116	0.8918	7.8538
	All	0.8938	0.8745	8.3841

feature encoding of monocular and binocular regions can lead to performance improvement, the effectiveness of M-LVPs for NR-IQA deserves further investigations by conducting more experiments on 2D image databases. Through such experiments, we may have a sense about how well the M-LVPs can contribute. Here, three most widely used 2D image databases including CSIQ [2], LIVE [83], and LIVE-MD [81] are selected as the benchmarks. To be more specific, we only use the learned M-LVPs to encode 2D images for feature extraction and also use SVR for quality regression. Similarly, 100 trails of train-test process are conducted on each database and the median results over 100 trails are reported. The results are shown in Table VII. From the Table, we can observe that the model only using M-LVPs for feature encoding can also make a reasonable quality prediction of both singly and multiply distorted 2D images. This further demonstrate the effectiveness of the learned M-LVPs.

G. Evaluation on Unknown Distortions

To validate the performance on other unknown distortion types, we also evaluate the proposed method on stereoscopic images suffered from fast fading and JPEG2000 compression. This experiment was conducted on LIVE 3D Phase-I and Phase-II databases where all the stereoscopic images suffered from fast fading and JPEG2000 compression were considered as the test set. The corresponding experimental results are shown in Table VIII. From the results, we can find that, although these two distortion types were not considered in LVP learning, our proposed method still delivered moderate performance on handling these two distortions. The reason

TABLE VIII
PERFORMANCE RESULTS ON UNKNOWN DISTORTIONS

Database	Distortion	PLCC	SRCC	RMSE
LIVE 3D Phase-I	JPEG2000	0.929	0.918	6.103
	Fast Fading	0.811	0.788	9.835
LIVE 3D Phase-II	JPEG2000	0.873	0.859	5.698
	Fast Fading	0.915	0.906	4.566

may be that these two distortions may have some potential commonalities with the three distortions we have considered, while our learned LVPs are able to capture such commonalities to a certain extent. However, for those distortions that are dramatically different from our considered distortions, such as contrast change, color distortion, and more, our proposed method may lose the power. Actually, the generalization capability is still the most challenging problem in NR S3D-IQA thus far.

H. Discussion

Although our proposed method has outstanding ability in NR quality evaluation of both SDSIs and MDSIs, the following issues deserve further discussions:

1) As the MV-LVP learning can be performed off-line, its computational time will not be considered for calculating the overall computational time in testing. Actually, the testing stage involves pixel visibility analysis, feature encoding, feature aggregation, and quality regression. Among these steps, the most time consuming step is pixel visibility analysis which takes about 3 seconds for a 640×480 stereopair when testing on a PC with Intel Core i5-6200 CPU @ 2.4 GHz and an 8 GB RAM. The software platform is MATLAB R2014b. As a result, the overall computational time is about 4.2 seconds for a 640×480 stereopair. In addition, the computational time can be further reduced by taking the advantage of parallel computation because the monocular and binocular regions can be processed in a parallel manner.

2) The B-LVPs are learned from a set of binocular patch pairs suffered from only symmetric distortion profile so that the binocular quality perception under the asymmetric distortion condition may not be fully exploited. In the future, it is interesting to generate more binocular patch pairs with more comprehensive distortion profiles, *e.g.*, both symmetric and asymmetric, both single and multiple, for MB-LVP learning.

Based on such data, the cortical RF properties in response to various MDSIs can be better simulated.

3) According to the evidences in visual physiology [70], our method resort to sparse coding as an approximation to the complex neuron encoding mechanism (we deem the learned MB-LVPs as local RFs found in the visual cortex). However, whether the sophisticated neuron encoding mechanism can be well addressed in such a simple way remains an open problem which requires further investigations.

4) The proposed method still follows the general learning-based NR-IQA framework which requires subjective rating scores as labels to calibrate a quality prediction model. However, obtaining subjective rating scores in terms of the perceived quality is always expensive and labor-consuming. Therefore, how to develop effective opinion-unaware solutions is the future research direction.

V. Conclusion

We have presented a unified NR quality evaluation method for both SDSIs and MDSIs by learning a set of MB-LVPs based on a novel task-oriented and modality-specific dictionary learning framework. The learned MB-LVPs can well characterize the underlying MRF and BRF properties of the visual cortex in response to stereopairs with different distortion modalities (single/multiple distortion). Two penalty terms, including reconstruction error penalty (data-driven) and quality inconsistency penalty (task-driven), are jointly minimized so as to generate a set of quality-oriented M-LVPs and B-LVPs for each distortion modality. Given a query stereo image (can be either SDSI or MDSI), feature encoding is performed using the learned MB-LVPs as MRF and BRF codebooks, resulting in the corresponding monocular and binocular responses. Finally, responses across all modalities are fused with the modality-specific SRE-based weights, yielding the final monocular and binocular feature representations for quality prediction using SVR. Our method, whose superiority has been well demonstrated by the experimental results on both SDSI and MDSI benchmark databases, achieves better consistency with subjective perception.

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[2] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, 2010, Art. no. 011006.

[3] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[4] S. Li, F. Zhang, M. Lin, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[5] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

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

[7] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[8] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE Trans. Broadcast.*, vol. 61, no. 3, pp. 520–531, Sep. 2015.

[9] S.-H. Bae and M. Kim, "A novel image quality assessment with globally and locally consilient visual quality perception," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2392–2406, May 2016.

[10] L. Ding, H. Huang, and Y. Zang, "Image quality assessment using directional anisotropy structure measurement," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1799–1809, Apr. 2017.

[11] E. D. Di Claudio and G. Jacovitti, "A detail-based method for linear full reference image quality prediction," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 179–193, Jan. 2018.

[12] J. Schild, J. LaViola, and M. Masuch, "Understanding user experience in stereoscopic 3D games," in *Proc. ACM SIGCHI Conf. Hum. Factors Comput. Syst.*, May 2012, pp. 89–98.

[13] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "Quality of experience model for 3DTV," *Proc. SPIE*, vol. 8288, p. 82881P, Feb. 2012.

[14] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Process., Image Commun.*, vol. 28, no. 9, pp. 1143–1155, 2013.

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

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

[17] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 2971–2983, Oct. 2015.

[18] J. Wang, A. Rehman, K. Zeng, S. Wang, and Z. Wang, "Quality prediction of asymmetrically distorted stereoscopic 3D images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3400–3414, Nov. 2015.

[19] Y. Zhang and D. M. Chandler, "3D-MAD: A full reference stereoscopic image quality estimator based on binocular lightness and contrast perception," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3810–3825, Nov. 2015.

[20] P. Lebreton, A. Raake, M. Barkowsky, and P. Le Callet, "Evaluating depth perception of 3D stereoscopic videos," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 710–720, Oct. 2012.

[21] J. Wang, S. Wang, K. Ma, and Z. Wang, "Perceptual depth quality in distorted stereoscopic images," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1202–1215, Mar. 2017.

[22] Y. J. Jung, H. Sohn, S.-I. Lee, H. W. Park, and Y. M. Ro, "Predicting visual discomfort of stereoscopic images using human attention model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2077–2082, Dec. 2013.

[23] H. Sohn, Y. J. Jung, S.-I. Lee, and Y. M. Ro, "Predicting visual discomfort using object size and disparity information in stereoscopic images," *IEEE Trans. Broadcast.*, vol. 59, no. 1, pp. 28–37, Mar. 2013.

[24] Q. Jiang, F. Shao, G. Jiang, M. Yu, and Z. Peng, "Three-dimensional visual comfort assessment via preference learning," *J. Electron. Imag.*, vol. 24, no. 4, p. 043002, Jul. 2015.

[25] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "On predicting visual comfort of stereoscopic images: A learning to rank based approach," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 302–306, Feb. 2016.

[26] Q. Jiang, F. Shao, G. Jiang, M. Yu, and Z. Peng, "Visual comfort assessment for stereoscopic images based on sparse coding with multi-scale dictionaries," *Neurocomputing*, vol. 252, pp. 77–86, Aug. 2017.

[27] J. Park, H. Oh, S. Lee, and A. C. Bovik, "3D visual discomfort predictor: Analysis of disparity and neural activity statistics," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1101–1114, Mar. 2015.

[28] H. Oh and S. Lee, "Visual presence: Viewing geometry visual information of UHD S3D entertainment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3358–3371, Jul. 2016.

[29] H. Oh, J. Kim, J. Kim, T. Kim, S. Lee, and A. C. Bovik, "Enhancement of visual comfort and sense of presence on stereoscopic 3D images," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3789–3801, Aug. 2017.

[30] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 675–684, Jul. 2000.

[31] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Dec. 2001, p. 1.

[32] L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *IEEE Trans. Broadcast.*, vol. 51, no. 2, pp. 191–199, Jun. 2005.

[33] V. A. F. Lamme, H. Supèr, R. Landman, P. R. Roelfsema, and H. Spekreijse, "The role of primary visual cortex (V1) in visual awareness," *Vis. Res.*, vol. 40, nos. 10–12, pp. 1507–1521, Jun. 2000.

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

[35] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.

[36] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *J. Physiol.*, vol. 195, no. 1, pp. 215–243, 1968.

[37] Y. D. Zhu and N. Qian, "Binocular receptive field models, disparity tuning, and characteristic disparity," *Neural Comput.*, vol. 8, no. 8, pp. 1611–1641, 1996.

[38] I. Ohzawa and R. D. Freeman, "The binocular organization of complex cells in the cat's visual cortex," *J. Neurophysiol.*, vol. 56, no. 1, pp. 243–259, 1986.

[39] K. Bahrami and A. C. Kot, "A fast approach for no-reference image sharpness assessment based on maximum local variation," *IEEE Signal Process. Lett.*, vol. 21, no. 6, pp. 751–755, Jun. 2014.

[40] Z. Wang, A. C.Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2000, pp. 981–984.

[41] H. Tao, N. Klomp, and I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 529–539, Apr. 2010.

[42] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4559–4565, Dec. 2017.

[43] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[44] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

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

[46] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

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

[48] Q. Wu *et al.*, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 425–440, Mar. 2016.

[49] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3129–3138, Jul. 2012.

[50] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

[51] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.

[52] L. Zhang, Z. Gu, X. Liu, H. Li, and J. Lu, "Training quality-aware filters for no-reference image quality assessment," *IEEE MultiMedia*, vol. 21, no. 4, pp. 67–75, Oct./Dec. 2014.

[53] Q. Jiang, F. Shao, G. Jiang, M. Yu, and Z. Peng, "Supervised dictionary learning for blind image quality assessment using quality-constraint sparse coding," *J. Vis. Commun. Image Represent.*, vol. 33, pp. 123–133, Nov. 2015.

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

[55] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Trans. Broadcast.*, vol. 60, no. 3, pp. 555–567, Sep. 2014.

[56] Y. Lu, F. Xie, T. Liu, Z. Jiang, and D. Tao, "No reference quality assessment for multiply-distorted images based on an improved bag-of-words model," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1811–1815, Oct. 2015.

[57] Q. Li, W. Lin, and Y. Fang, "No-reference quality assessment for multiply-distorted images in gradient domain," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 541–545, Apr. 2016.

[58] H. Hadizadeh and I. Bajić, "Color Gaussian jet features for no-reference quality assessment of multiply-distorted images," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1717–1721, Dec. 2016.

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

[60] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Oriented correlation models of distorted natural images with application to natural Stereopair quality evaluation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1685–1699, May 2015.

[61] B. Appina, S. Khan, and S. S. Channappayya, "No-reference stereoscopic image quality assessment using natural scene statistics," *Signal Process., Image Commun.*, vol. 43, pp. 1–14, Apr. 2016.

[62] W. Zhou and L. Yu, "Binocular responses for no-reference 3D image quality assessment," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1077–1084, Jun. 2016.

[63] F. Shao, K. Li, W. Lin, G. Jiang, and Q. Dai, "Learning blind quality evaluator for stereoscopic images using joint sparse representation," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 2104–2114, Oct. 2016.

[64] F. Shao, W. Tian, W. Lin, G. Jiang, and Q. Dai, "Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2059–2074, Mar. 2016.

[65] L. Liu, B. Liu, C.-C. Su, H. Huang, and A. C. Bovik, "Binocular spatial activity and reverse saliency driven no-reference stereopair quality assessment," *Signal Process., Image Commun.*, vol. 58, pp. 287–299, Aug. 2017.

[66] W. Zhang, C. Qu, L. Ma, J. Guan, and R. Huang, "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network," *Pattern Recognit.*, vol. 59, pp. 176–187, Nov. 2016.

[67] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "Learning a referenceless stereopair quality engine with deep nonnegativity constrained sparse autoencoder," *Pattern Recognit.*, vol. 76, pp. 242–255, Apr. 2018.

[68] H. Oh, S. Ahn, J. Kim, and S. Lee, "Blind deep S3D image quality evaluation via local to global feature aggregation," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4923–4935, Oct. 2017.

[69] F. Shao, W. Tian, W. Lin, G. Jiang, and Q. Dai, "Learning sparse representation for no-reference quality assessment of multiply distorted stereoscopic images," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1821–1836, Aug. 2017.

[70] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[71] M. Carandini, D. J. Heeger, and J. A. Movshon, "Linearity and normalization in simple cells of the macaque primary visual cortex," *J. Neurosci.*, vol. 17, no. 21, pp. 8621–8644, 1997.

[72] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons," in *Statistical Theories of the Brain*. Cambridge, MA, USA: MIT Press, 2002, pp. 203–222.

[73] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[74] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

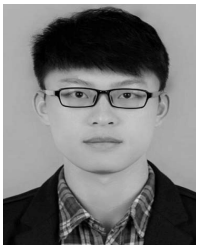
[75] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.

[76] Y. C. Pati, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 1993, pp. 40–44.

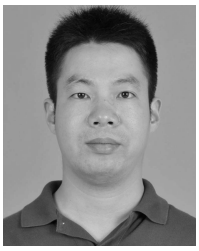
[77] S.-J. Luo, Y.-T. Sun, I.-C. Shen, B.-Y. Chen, and Y.-Y. Chuang, "Geometrically consistent stereoscopic image editing using patch-based synthesis," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 1, pp. 56–67, Jan. 2015.

[78] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," *CS Technion*, vol. 40, no. 8, pp. 1–15, 2008.

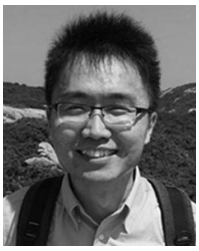
[79] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, Art. no. 27.
[80] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2001, pp. 416–423.
[81] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2012, pp. 1693–1697.
[82] A. K. Moorthy, C.-C. Su, A. Mittal, and A. C. Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Process., Image Commun.*, vol. 28, no. 8, pp. 870–883, Dec. 2013.
[83] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.



Qiuping Jiang received the Ph.D. degree from the School of Information Science and Engineering, Ningbo University, Ningbo, China, in 2018. From 2017 to 2018, he was a Visiting Student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Information Science and Engineering, Ningbo University. His research interests include image processing, visual perception modeling, and computer vision. He was a recipient of the JVCI 2017 Best Paper Award Honorable Mention. He is a reviewer for several prestigious journals and conferences, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS.



Feng Shao received the B.S. and Ph.D. degrees in electronic science and technology from Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively. In 2012, he was a Visiting Scholar with the School of Computer Engineering, Nanyang Technological University, Singapore. He is currently a Professor with the Faculty of Information Science and Engineering, Ningbo University, China. His research interests include 3D video coding, 3D image/video quality assessment, and 3D image/video enhancement. He received the Excellent Young Scholar Award by the National Natural Science Foundation of China in 2016.



Wei Gao received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2017. From 2012 to 2013, he was a Camera ISP Engineer with OmniVision Technologies, Shanghai, China. In 2016, he was a Visiting Scholar with the Electrical Engineering Department, University of California, Los Angeles, CA, USA. Since 2017, he has been a Post-Doctoral Fellow with the Department of Computer Science, City University of Hong Kong, and a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include image and video coding, perceptual image processing, multimedia communication, machine learning and optimization, and integrated circuits design.

Zhuo Chen received the B.S. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University. He is currently pursuing the Ph.D. degree with the Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore. He was a Deep Learning Engineer with LLVISION Pte. Ltd. In 2014, he joined the Chinese Academy of Sciences as a Research Assistant. His research interests include image quality assessment and deep learning.

Gangyi Jiang received the M.S. degree from Hangzhou University, Hangzhou, China, in 1992, and the Ph.D. degree from Ajou University, South Korea, in 2000. He is currently a Professor with the Faculty of Information Science and Engineering, Ningbo University, Ningbo, China. He has published over 100 referred papers in international journals and conferences. His research interests include video compression, multi-view video coding, and visual perception.

Yo-Sung Ho (SM'06–F'16) received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, South Korea, in 1981 and 1983, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1990. In 1983, he joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. From 1990 to 1993, he was with the North America Philips Laboratories, Briarcliff Manor, NY, USA, where he was involved in the development of the advanced digital high-definition television system. In 1993, he rejoined ETRI as a Technical Staff and was involved in the development of the Korean DBS digital television and high-definition television systems. Since 1995, he has been with the Gwangju Institute of Science and Technology (GIST), South Korea. Since 2003, he has been the Director of the Realistic Broadcasting Research Center, GIST, where he is currently a Professor with the School of Electrical Engineering and Computer Science. His research interests include digital image and video coding, image analysis and image restoration, 3D image modeling and representation, advanced source coding techniques, augmented reality and virtual reality, 3D television, and realistic broadcasting technologies. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the IEEE TRANSACTIONS ON MULTIMEDIA.