IEEE
IEEE Signal Processing Society
IEEE Transactions on Image Processing

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

Journal:	*Transactions on Image Processing*
Manuscript ID	TIP-18507-2018
Manuscript Type:	Regular Paper
Date Submitted by the Author:	21-Feb-2018
Complete List of Authors:	Jiang, Qiuping; Ningbo University, Faculty of Information Science and Engineering; Nanyang Technological University, School of Computer Science and Engineering Shao, Feng; Ningbo University, Faculty of Information Science and Engineering Gao, Wei; City University of Hong Kong, Department of Computer Science Chen, Zhuo; Nanyang Technological University, School of Computer Science and Engineering Jiang, Gangyi; Ningbo University, Faculty of Information Science and Engineering Ho, Yo-Sung; Gwangju Institute of Science and Technology, School of Electrical Engineering and Computer Science
EDICS:	5. SMR-HPM Perception and Quality Models for Images and Video < Image & Video Sensing, Modeling, and Representation, 23. ELI-STE Stereoscopic and Multiview Processing and Display < Electronic Imaging, 18. COM-MMC Image and Video Multimedia Communications < Image & Video Communications

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

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

***Abstract*—A challenging problem in no-reference quality assessment of multiply distorted stereoscopic images (MDSIs) is to accurately simulate the monocular and binocular visual properties under the condition where multiple distortion types are involved simultaneously. Due to the joint effects of multiple distortion types in MDSIs, the underlying visual mechanisms, including monocular quality perception and binocular combination, show 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 principal idea is to learn a set of MB-LVPs to characterize the underlying local receptive field (RF) properties of the visual cortex in response to SDSIs and MDSIs. Furthermore, we also consider that the learning of LVPs should be performed in a task-driven manner. For this, two penalty entities, including reconstruction error penalty and quality inconsistency penalty, respectively defined from bottom-up and top-down aspects, are combined and jointly minimized within a dictionary learning framework to generate a set of quality-oriented MB-LVPs for each single and multiple distortion modality. Given a test stereoscopic image (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 the modalities are fused with probabilistic weights which are determined by the modality-specific sparse reconstruction errors, yielding the final monocular and binocular feature representations for quality regression. The superiority of our method has been verified on both SDSI and MDSI benchmark databases.**

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

I. INTRODUCTION

AUTOMATIC image quality assessment (IQA) is potentially useful for the optimization and monitoring of many image processing and enhancement 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 image/video has become a hot research target of IQA.

Compared to its 2D counterpart, 3D-IQA encounters more challenges as the joint effects of different perceptual dimensions, such as image distortion, depth perception, visual discomfort, visual presence and more, need to be considered 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 mainly focus on ascertaining the influence of each individual aspect on the overall 3D quality-of-experience (QoE) 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 evaluating a stereopair without utilizing any information from its original version, we are more interested in the NR case of 3D-IQA.

A stereoscopic 3D image consists of a pair of slightly offset 2D 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 and these corresponding regions are then processed by binocular vision, there are still some monocular regions in the left and right images since occlusion will inevitably occur due to the viewing angle difference [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. That is, the right image does not provide any correspondence information to those background areas which only can be perceived in the left one. 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 right border area of the left image and a small amount of left border area of the right image only can be seen in the left view and right view, respectively. When viewing a stereoscopic image, contents in monocular and binocular region will be processed by monocular and binocular vision, respectively. Obviously, efforts towards designing efficient visual models to resemble the properties of monocular and binocular vision provide unique benefits to 3D-IQA.

Q. Jiang is with the School of Information Science and Engineering, Ningbo University, Ningbo, China, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

F. Shao, and G. Jiang are with the School of Information Science and Engineering, Ningbo University, Ningbo, China.

W. Gao and Z. Chen are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

Y.-S. Ho is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Republic of Korea.

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 the vision 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.

With the above considerations, we propose 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), respectively defined from bottom-up and top-down aspects, are combined and jointly minimized within a dictionary learning framework so as to generate a set of quality-oriented M-LVPs and B-LVPs for each distortion modality. Given a test stereoscopic 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 probabilistic weights which are determined by the modality-specific reconstruction errors, yielding the final monocular and binocular feature representations for quality regression. In a nutshell, our main contributions 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. A detailed description of the proposed method is presented in Section III. In Section IV, a series of experiments on both SDSI and MDSI databases are conducted to validate the performance. Finally, conclusions are drawn in Section V.

II. Related Work

We are focusing on the NR-IQA category because it is generally more applicable than other IQA research branches. Moreover, we target to design a unified quality evaluation method for both SDSIs and MDSIs. Therefore, representative works on NR-IQA for singly/multiply distorted 2D/3D images are briefly revisited in this section.

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 all commonly encountered distortion types is necessary.

Due to the great success of natural scene statistics (NSS) features in capturing quality degradation caused by distortions, the past several years have witnessed tremendous progress in the development of general-purpose NR-SDIQA approaches. The basic assumption of NSS-based general-purpose approaches is that pristine natural images inherently obey certain regular statistical rules while the distortions will modify such regularities. With this, several NSS properties in spatial and transform domains have been exploited and utilized to extract quality-aware features, resulting in the corresponding NR-SDIQA approaches with different NSS 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 features can be 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 for the development of feature learning-based NR-IQA methods [53,54].

Previous research efforts on general-purpose NR-SDIQA have also focused on the so-called completely blind case. The approaches belonging to this category generally have better generalization capacity because they do not require the training process to calibrate a quality prediction module. Two most representative works are: Natural Image Quality Evaluator (NIQE) [55] and Integrated Local-NIQE (IL-NIQE) [56]. NIQE first extracts a set of local features and then fits the local feature vectors to a global multivariate Gaussian (MVG) model. The MVG model fitted on a distorted image is compared to the one fitted on a set of pristine natural images to derive a distance metric as the quality score. IL-NIQE is an extended version of NIQE in that it enriches the quality-aware features for statistical model construction and integrates the local and global MVG models together.

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*., [57] 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*., [58] 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*., [59] 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 *et al*., [60] 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*., [61] 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 [62] 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 [63], 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 *et al*., [64] 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*., [65,66] 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*., [67] 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*., [68] proposed to learn structures from stereopairs based on convolutional neural network (CNN) for NR-SDSIQA. Jiang *et al.* [69] 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. The results estimated by each DNCSAE module are combined based on a Bayesian framework.

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.*, [70] 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 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 local 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. The details will be illustrated in the next section.

III. Proposed Method

The quality-aware features in our method are obtained by automatic feature encoding over a set of learned task-oriented and modality-specific MB-LVPs as neuron codebooks. Given a test 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 MB-LVPs as MRF and BRF 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 sparse reconstruction error (SREs)s, yielding the final monocular and binocular feature representation for quality regression. 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 biological behaviors of RFs found in the visual cortex. It has been widely discovered that RFs in the visual cortex could be characterized to be spatially localized and oriented patterns. Meanwhile, such properties of RFs were found to be similar with the characteristics of the basis functions learned from natural scene images. In order to learn the basis functions, an unsupervised and data-driven learning approach is adopted in [71], i.e., using a set of patches extracted from natural scenes images and seeking to maximize the sparseness of the encoded visual information. The basis functions that emerge, are an over-complete dictionary containing distinctive local visual patterns such as lines, edges, corners, and smooth patterns, as shown in Fig. 1. These learned basis functions, shown to well resemble the RFs found in the visual cortex, are called LVPs hereinafter.

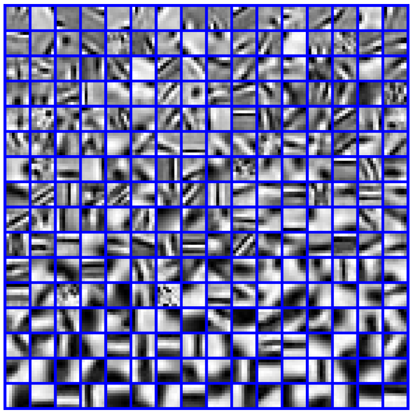


Fig. 1. Visualization of the over-complete dictionary containing a set of basis functions learned from natural scene images.

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

1) Motivations: According to the existing studies in the field of visual physiology [35-37], two types of RFs have been found existence 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 unsupervised data-driven learning method 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 a specific task, i.e., quality perception in this study.

To inherit the idea of learning an over-complete set of basis functions from natural scene images to resemble RFs, and also to account for the above principles, we propose to learn M-LVP and B-LVP (to account for the first principle) respectively from a set of monocular and binocular images based on a task-oriented (to account for the third principle) and modality-specific (to account for the second principle) dictionary learning framework.

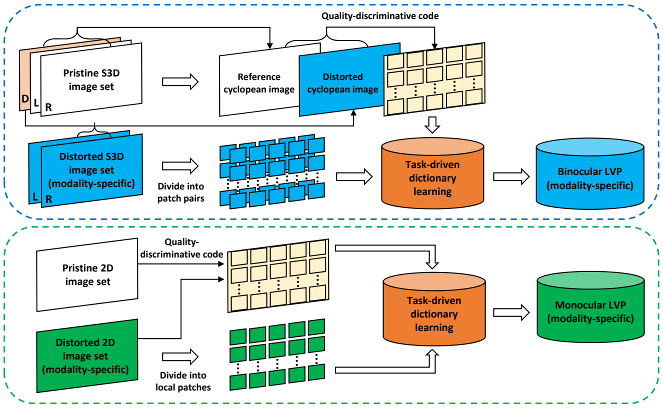

Quality-discriminative code
Pristine S3D image set
D
L
R
Reference cyclopean image
Distorted cyclopean image
Distorted S3D image set (modality-specific)
L
R
Divide into patch pairs
Task-driven dictionary learning
Binocular LVP (modality-specific)
Pristine 2D image set
Quality-discriminative code
Distorted 2D image set (modality-specific)
Divide into local patches
Task-driven dictionary learning
Monocular LVP (modality-specific)

Fig. 2. Overview of the process for learning M-LVPs and B-LVPs in a task-oriented and modality-specific manner.

2) Problem Formulation: We present an overview of the process for 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:

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

where the superscript symbol $\mho \in \{\mathcal{M}, \mathcal{B}\}$ 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, Ψ is a positive constant indicating the sparsity, $\hat{\mathbf{D}}_k^{\mho} \in \mathbb{R}^{p_{\mho} \times d_k}$ is the learned LVPs over which the input patches $\mathbf{P}_k^{\mho} \in \mathbb{R}^{p_{\mho} \times n_k}$ have sparse representation codes $\hat{\mathbf{A}}_k^{\mho} \in \mathbb{R}^{d_k \times n_k}$, $\mathbf{S}_k^{\mho} \in \mathbb{R}^{d_k \times n_k}$ is the quality-discriminative code (QDC) of $\mathbf{P}_k^{\mho}$, $\hat{\mathbf{W}}_k^{\mho} \in \mathbb{R}^{d_k \times d_k}$ is a learned linear transformation matrix which encourages the original sparse codes $\hat{\mathbf{A}}_k^{\mho}$ to be most discriminative in terms of quality in the new space.

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.

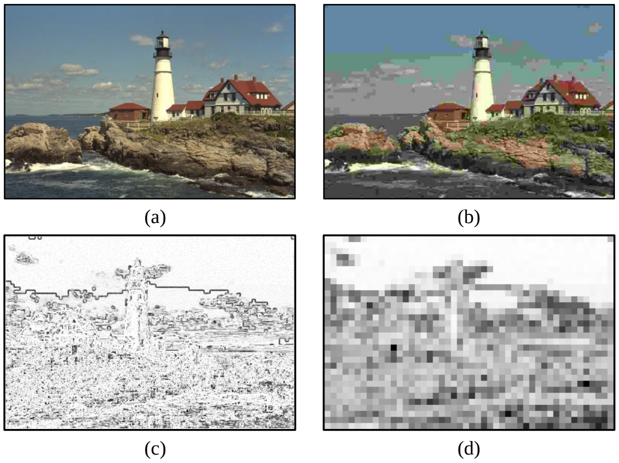


Fig. 3. Visualization of local quality estimation results on 2D images using FSIM: (a) pristine image, (b) distorted image, (c) pixel-wise FSIM map, (d) patch-level FSIM map, where the brighter areas indicate better quality.

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':

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

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 $\{\beta_{\{h,w\}} | h = -H, ..., H; w = -W, ..., W\}$ defines a unit-volume Gaussian window. This local normalization is found to well resemble the primate cortical visual process of the human brain. By local normalization, a set of modality-specific normalized images $\hat{\mathbf{I}}_k' = \{\hat{I}_{k,1}', \hat{I}_{k,2}', ..., \hat{I}_{k,l_k}'\}$ 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 $\mathbf{P}_k^{\mathcal{M}} = [\mathbf{p}_{k,1}^{\mathcal{M}}, \mathbf{p}_{k,2}^{\mathcal{M}}, , ..., \mathbf{p}_{k,n_k}^{\mathcal{M}}] \in \mathbb{R}^{p_{\mathcal{M}} \times n_k}$, 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. 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 $\mathbf{p}_{k,n}^{\mathcal{M}}$, $n = 1, 2, ..., n_k$

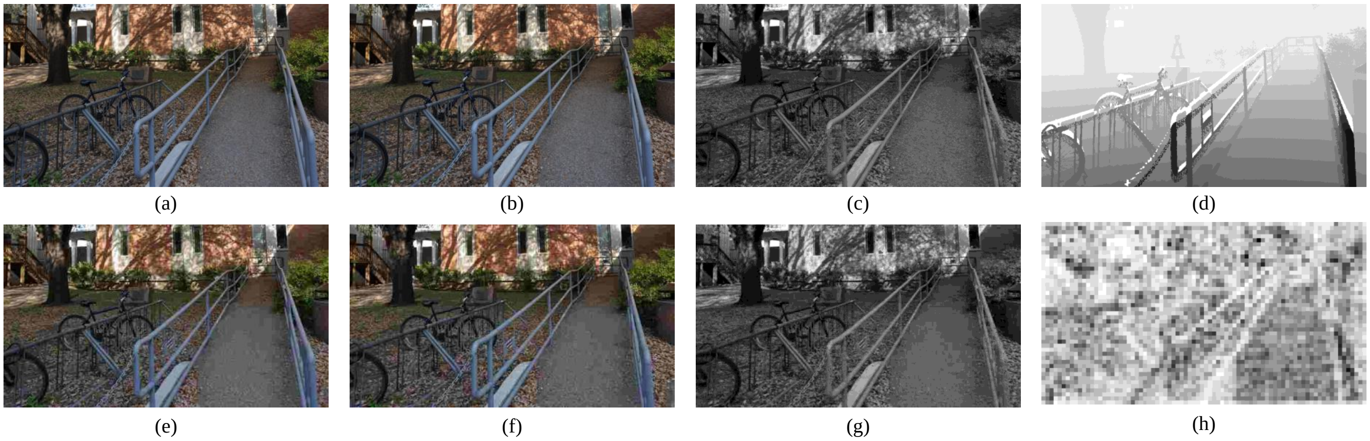


Fig. 4. Visualization of local quality estimation results on cyclopean images using the SSIM metric: (a) pristine left image, (b) pristine right image, (c) pristine cyclopean image synthesized from (a) and (b), (d) left disparity map, (e) distorted left image, (f) distorted right image, (g) distorted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

is estimated by averaging the FSIM scores of all the pixels inside this patch, resulting in patch-level quality scores $q^{\mathcal{M}}_{k,n} \in [0,1]$. 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 $q^{\mathcal{M}}_{k,n}$, we further define the QDC $\mathbf{s}^{\mathcal{M}}_{k,n} = [s^{\mathcal{M}}_{k,n}(1), s^{\mathcal{M}}_{k,n}(2), ..., s^{\mathcal{M}}_{k,n}(d_k)]^T \in \mathbb{R}^{d_k}$ as follows:

$$s^{\mathcal{M}}_{k,n}(i) = \begin{cases} 1, & \left\lfloor (z^{\mathcal{M}}_{k,n} - 1) \cdot \frac{d_k}{Z} \right\rfloor \le i < \left\lfloor z^{\mathcal{M}}_{k,n} \cdot \frac{d_k}{Z} \right\rfloor \\ 0, & otherwise \end{cases}, \quad (5)$$

where the symbol $\lfloor \cdot \rfloor$ represents the floor operation, $z^{\mathcal{M}}_{k,n} \in \{1, 2, \cdots, Z = 20\}$ is the quantified quality interval index of $q^{\mathcal{M}}_{k,n}$, and is inferred as follows:

$$Z \cdot \left\lfloor q^{\mathcal{M}}_{k,n} \right\rfloor \le z^{\mathcal{M}}_{k,n} < Z \cdot \left\lfloor q^{\mathcal{M}}_{k,n} \right\rfloor + 1. \quad (6)$$

As a result, a binary QDC $\mathbf{s}^{\mathcal{M}}_{k,n} \in \mathbb{R}^{d_k}$ is generated for each monocular patch $\mathbf{p}^{\mathcal{M}}_{k,n} \in \mathbb{R}^{p_{\mathcal{M}}}$. The QDC matrix $\mathbf{S}^{\mathcal{M}}_{k}$ is finally described as follows:

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

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 $\mathbf{P}^{\mathcal{B}}_{k}$ and $\mathbf{S}^{\mathcal{B}}_{k}$ as input. To generate the binocular patch pairs $\mathbf{P}^{\mathcal{B}}_{k}$, 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 $\mathbf{P}^{\mathcal{B}}_{k} = \left[[\mathbf{p}^{\mathcal{L}}_{k,1}, \mathbf{p}^{\mathcal{R}}_{k,1}]^T, [\mathbf{p}^{\mathcal{L}}_{k,2}, \mathbf{p}^{\mathcal{R}}_{k,2}]^T, ..., [\mathbf{p}^{\mathcal{L}}_{k,n_k}, \mathbf{p}^{\mathcal{R}}_{k,n_k}]^T\right] \in \mathbb{R}^{2p \times n_k}$. 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 $\mathbf{S}^{\mathcal{B}}_{k}$, 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, the authors in [14] have reported a simplified model that synthesizes a cyclopean view from the left and right images of a stereopair by accounting for the critical binocular rivalry:

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

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:

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

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

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 $\mathbf{p}^{\mathcal{B}}_{k,n} = [\mathbf{p}^{\mathcal{L}}_{k,n}, \mathbf{p}^{\mathcal{R}}_{k,n}]^T$, its quality $q^{\mathcal{B}}_{k,n}$ 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 $\mathbf{S}_k^{\mathcal{B}} = [\mathbf{s}_{k,1}^{\mathcal{B}}, \mathbf{s}_{k,2}^{\mathcal{B}}, ..., \mathbf{s}_{k,n_k}^{\mathcal{B}}] \in \mathbb{R}^{d_k \times n_k}$ can be obtained in the same way according to (5)-(7).

D. Optimization

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

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

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

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

The above objective function can be well solved by the K-SVD algorithm [72]. 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 [73], we perform several iterations of K-SVD within each quality interval 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

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

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 [74] 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 $k \in \{$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.

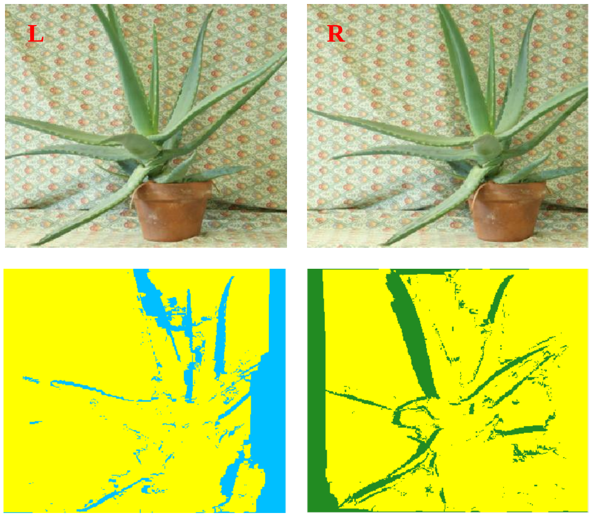


Fig. 5. An illustration of monocular and binocular regions. (a) left image, (b) right image. The blue regions indicate the LMR, the green regions indicate the RMR, the yellow regions indicate the BR.

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 a simple light-weight rule-based method for pixel visibility analysis of stereopairs [75]. 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:

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

$$\begin{aligned} &\forall q_{\mathcal{L}} | (q_{\mathcal{L},x} > p_{\mathcal{L},x}) \cap (q_{\mathcal{L},y} = p_{\mathcal{L},y}), \\ &p_{\mathcal{L},x} + d_{\mathcal{L}}(p_{\mathcal{L}}) \ne q_{\mathcal{L},x} + d_{\mathcal{L}}(q_{\mathcal{L}}). \end{aligned} \tag{19}$$

where R_w is the width of the image, and $q_{\mathcal{L}}$ represents a certain pixel on the right side of $p_{\mathcal{L}}$. 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 problematic especially for the severe distortion case. However, experimental results support 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 $p^{\mathcal{M}}_{\mathcal{L}} \in$ LMR ($p^{\mathcal{M}}_{\mathcal{R}} \in$ RMR) is denoted by $\mathbf{p}^{\mathcal{M}}_{\mathcal{L}} \in \mathbb{R}^{p\times 1}$ ($\mathbf{p}^{\mathcal{M}}_{\mathcal{R}} \in \mathbb{R}^{p\times 1}$). The neural coding process is simply approximated by sparse coding, such that:

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

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

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

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

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

where the mathematical operator $\max$ is performed on each dimension of $\hat{\mathbf{a}}^{\mathcal{M}}_{\mathcal{L},k}(i),\ i = 1, 2, \cdots, N_{\mathcal{L}}$ and $\hat{\mathbf{a}}^{\mathcal{M}}_{\mathcal{R},k}(j),\ j = 1, 2, \cdots, N_{\mathcal{R}}$, $N_{\mathcal{L}}$ 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 $p^{\mathcal{B}}_{\mathcal{L}} \in$ LBR and its corresponding patch in the right image constitute a binocular patch pair denoted by $\mathbf{p}^{\mathcal{B}} = [\mathbf{p}^{\mathcal{B}}_{\mathcal{L}}, \mathbf{p}^{\mathcal{B}}_{\mathcal{R}}] \in \mathbb{R}^{2p\times 1}$. With sparse coding, the neural coding response of $\mathbf{p}^{\mathcal{B}}$ is similarly computed as follows:

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

Finally, we can obtain a max-pooled binocular response vector denoted by $\bar{\mathbf{a}}^{\mathcal{B}}_{k}$. As a highly efficient algorithm, the batch-OMP algorithm [76] is implemented to get the solution.

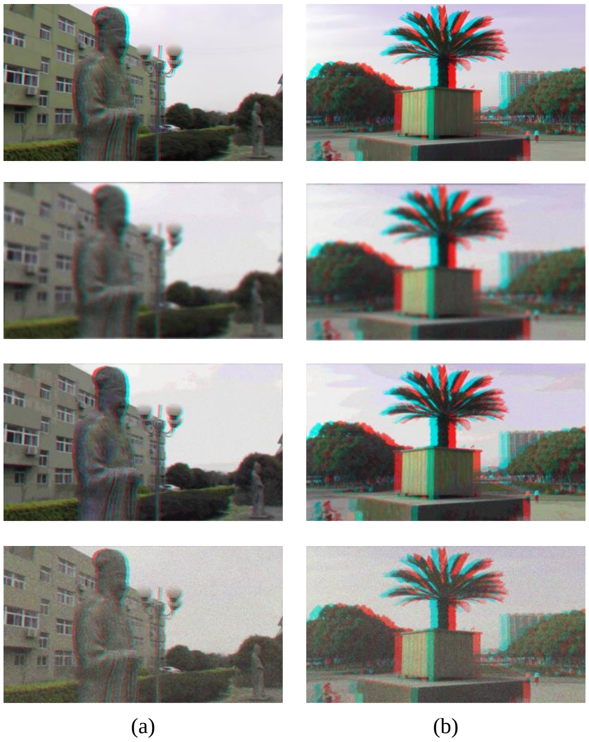


Fig. 6. An illustration of the masking effect of different distortion types in MDSIs. (a) and (b) are two examples with different scenes selected from the MDSID database. The rows from top to bottom are the pristine stereoscopic image, GB-dominant MDSI, JPEG-dominant MDSI, and WN-dominant MDSI. All the stereopairs are visualized in an anaglypic format.

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 $\mathbf{p}^{\mathcal{M}}_{\mathcal{L}}$ as an example, the corresponding SRE is computed as follows:

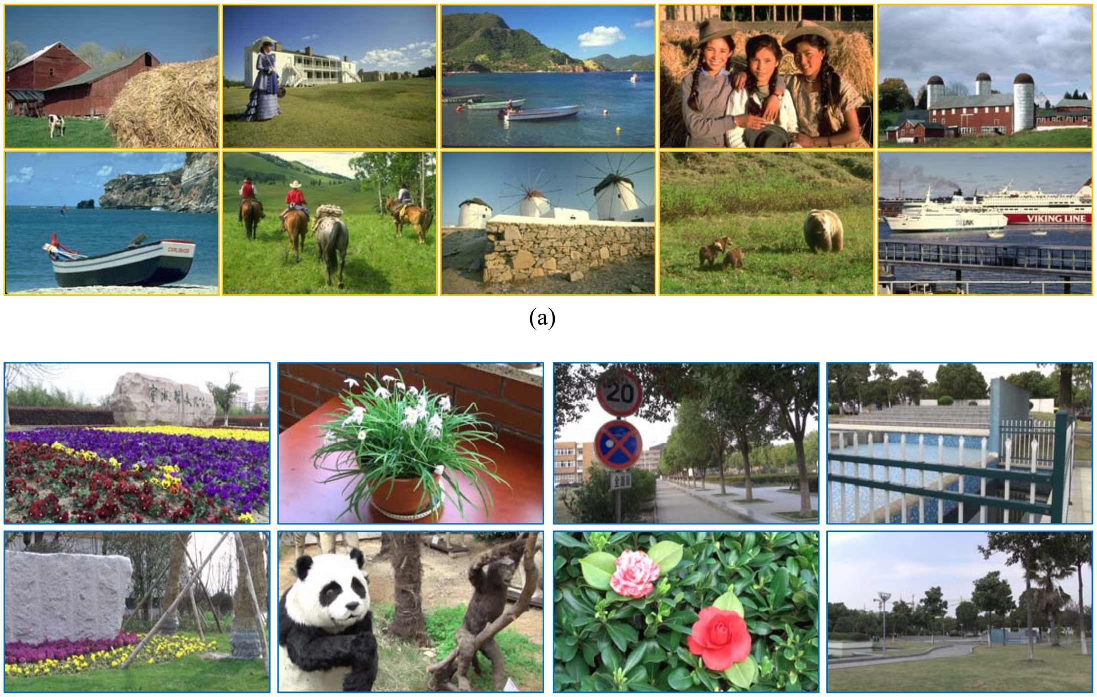
$$e^{\mathcal{M}}_{\mathcal{L},k} = \left\| \mathbf{p}^{\mathcal{M}}_{\mathcal{L}} - \tilde{\mathbf{D}}^{\mathcal{M}}_{k} \hat{\mathbf{a}}^{\mathcal{M}}_{\mathcal{L},k} \right\|_2^2, \tag{25}$$

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

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

The finally aggregated right monocular response vector $\bar{\mathbf{a}}^{\mathcal{M}}_{\mathcal{R}}$ 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 $\bar{\mathbf{a}}^{\mathcal{M}} = [\bar{\mathbf{a}}^{\mathcal{M}}_{\mathcal{L}}, \bar{\mathbf{a}}^{\mathcal{M}}_{\mathcal{R}}]$. As observed from (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.

The proposed cross-modality aggregation scheme actually provides a unified and effective way to characterize 1) the



(a)

(b)

Fig. 7. The pristine 2D and 3D natural images. (a) The ten 2D natural images selected from the Berkeley Segmentation Data Set (BSDS500) [78], (b) The eight 3D natural images (only the left images are presented) captured by ourselves using a stereo digital camera.

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, 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). 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 [77] to implement SVR with the radial basis function as the kernel.

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.

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 QCDs as inputs. 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) [78] at four distortion levels (the distortion control parameters are decided to ensure a good perceptual separation), 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.

Following the previous relevant works, for multiple distortion simulation, the GB is simulated first, followed by JPEG compression, and finally the WN injection. With 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), and the sparsity is set to be 8 (Ψ=8). All these parameter settings are kept fixed in our experiments.

B. Database and Protocol

For performance evaluation, three stereo 3D image quality databases are used: (1) LIVE 3D Phase-I database [79], (2) LIVE 3D Phase-II database [14,61], and (3) MDSID database [70]. Among these three databases, LIVE 3D Phase-I and Phase-II contain only SDSIs and the subjective rating scores of each SDSI in the form of DMOS (the difference between these two is that, SDSIs contained in LIVE 3D Phase-I are corrupted by symmetric single distortion, while SDSIs contained in LIVE 3D Phase-II are corrupted by either symmetric or asymmetric single distortion), while MDSID contains both SDSIs and MDSIs as well as their corresponding DMOS values. Note that, MDSIs contained in MDSID database are corrupted by symmetric multiple distortion (GB+JPEG+WN). Detailed information regarding the databases can be found in the authors' original papers. Overall, in viewing of the scene (different reference images) and distortion (symmetric and asymmetric, single and multiple) diversities of these three databases, the performance evaluation on them is considered to be comprehensive.

For performance evaluation on each database, 100 repetitions of train-test process are conducted and the median results for each criteria over 100 repetitions are reported to best avoid the performance bias. Each repetition involves a random database split into two non-overlap subsets: 80% samples out of the entire database for model training and the remaining 20% for model testing. Such protocol has been widely adopted for performance evaluation of learning-based NR-IQA methods. 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 performance should deliver higher PLCC and SRCC values but lower RMSE value. Before computing the performance criteria, a five-parameter logistic function is applied to bring the prediction values to the same scale of the DMOS values [80],

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

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 determined in the fitting process. Note that, this logistic regression step will affect only PLCC and RMSE results.

C. Evaluation on MDSID

To the best of our knowledge, MDSID is the only database suitable for assessing the ability of different SIQA methods in evaluating MDSIs. In the experiments, we consider all the MDSIs in the entire MDSID database for model training and testing. We compare the proposed method with two NR-SDIQA methods (BRISQUE [45], GM-LOG [46]), two NR-MDIQA methods (SISBLIM [57], Color-JET [60]), three NR-SDSIQA methods (3D-DQE [66], StereoINQ [67], 3D-DNCSAE [69]), and one NR-MDSIQA method (MUMBLIM) [70]. All the compared methods are popular and representative in the NR-IQA research. 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, 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 I where the best value for each criterion has been highlighted in boldface. In addition, we also conduct the statistical significance test as in [66,67] to further prove the superiority of our method over other competitors. The results are presented in Table II where the value "1" indicates the row model is statistically better than the column model, the value "-1" indicates the row model is statistically worse than the column model, and the value "0" indicates the row and column models are statistically equivalent.

TABLE I
PERFORMANCE RESULTS ON THE MDSID DATABASE. FOR EACH CRITERION, THE BEST VALUE IS HIGHLIGHTED 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**

It can be observed from the table 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 either the multiple distortion or 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, i.e., without utilizing the subjective rating scores to support learning a quality prediction model. The recently proposed Color-JET method, although specifically designed for multiple distortion evaluation of 2D images, performs slightly worse than GM-LOG in the 3D multiple distortion case. These results actually support our statement that the quality issues caused by multiple distortions in 2D and 3D images are different. The three compared NR-SDSIQA

TABLE II
STATISTICAL SIGNIFICANCE TEST RESULTS ON THE MDSID DATABASE. IN THE TABLE, "1" INDICATES THE ROW MODEL IS STATISTICALLY BETTER THAN THE COLUMN MODEL; "-1" INDICATES THE ROW MODEL IS STATISTICALLY WORSE THAN THE COLUMN MODEL; "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

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 efforts. Towards the circumvention of exploring potential NSS for evaluating MDSIs, the MUMBLIM method tries to construct an implicit mapping function for space transfer, i.e., from the original raw patch space to target quality space, 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 to-be-assessed stereopairs. Instead, the used quality-aware features in our method are obtained in a more flexible data-driven manner 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.

TABLE III
PERFORMANCE RESULTS ON THE LIVE 3D PHASE-I DATABASE. FOR EACH CRITERION, THE TOP THREE VALUES ARE MARKED IN RED, GREEN, AND BLUE COLOR, RESPECTIVELY.

Criteria	Method	GB	JPEG	WN	Average
PLCC	Akhter-SPIE	0.617	0.729	0.904	0.750
	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
	Proposed	0.968	0.795	0.948	0.904
SRCC	Akhter-SPIE	0.555	0.675	0.914	0.715
	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
	Proposed	0.915	0.774	0.927	0.872
RMSE	Akhter-SPIE	11.387	4.273	7.092	7.584
	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
	Proposed	3.548	3.740	5.079	4.122

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 eight state-of-the-art NR-SIQA algorithms which are all designed for evaluating the visual quality of SDSIs. The compared eight NR-SIQA algorithms are Akhter's method (Akhter-SPIE) [81], Chen's method (Chen-TIP) [61], Su's method (S3D-BLINQ) [62], Apinna's method (StereoQUE) [63], Zhou's method (Zhou-TMM) [64], Liu's method (StereoINQ) [67], and Jiang's method (3D-DNCSAE) [69]. 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 III and Table IV, respectively. To facilitate viewing, the top three values for each criterion are marked in red, green, and blue, respectively.

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 PLCC value when evaluating the JPEG subset contained in the LIVE 3D Phase-II database.

TABLE IV
PERFORMANCE RESULTS ON THE LIVE 3D PHASE-II DATABASE. FOR EACH CRITERION, THE TOP THREE VALUES ARE MARKED IN RED, GREEN, AND BLUE COLOR, RESPECTIVELY.

Criteria	Method	GB	JPEG	WN	Average
PLCC	Akhter-SPIE	0.795	0.786	0.722	0.768
	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
	Proposed	0.972	0.873	0.968	0.938
SRCC	Akhter-SPIE	0.682	0.649	0.714	0.682
	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
	Proposed	0.915	0.842	0.959	0.905
RMSE	Akhter-SPIE	8.450	4.535	7.416	6.800
	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
	Proposed	3.550	3.361	2.692	3.201

When comparing the average performance 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 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. Validation of Individual Component

Compared to the previous works, our proposed method has three unique 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 ascertain the contribution of each component. Towards this end, we have designed three experiments to: 1) investigate the contribution of B-LVP learning, i.e., with/without B-LVPs for feature encoding, 2) investigate the contribution of Task-Oriented Penalty (TOP), i.e., with/without task-oriented penalty in Eq. (1), and 3) investigate 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 depicted in Fig. 8. It can be observed that, as compared to our proposed one which takes all these components into account, without each of the above components can lead 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.

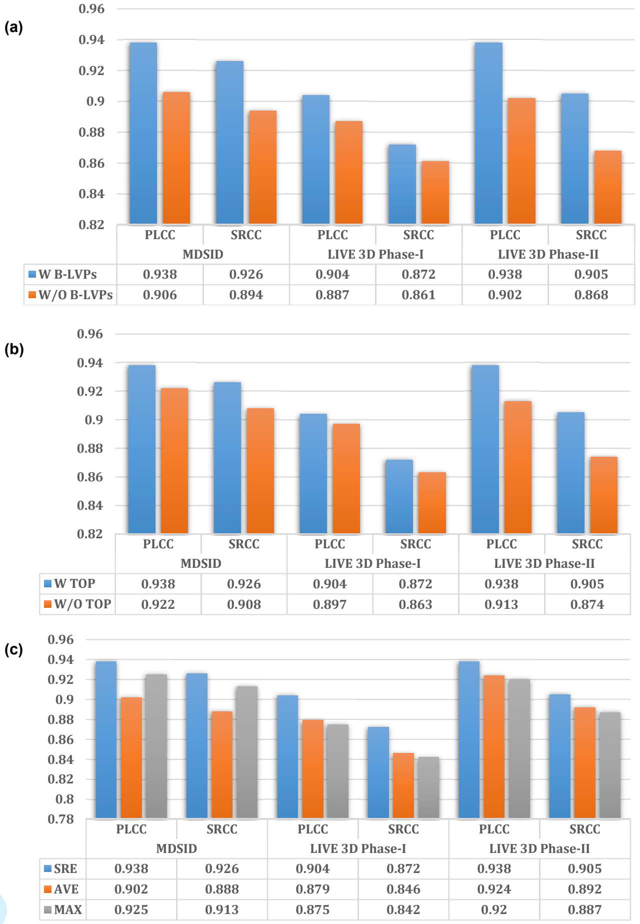


Fig. 8. Performance comparison results of individual component validation. (a) with/without B-LVPs for feature encoding, (b) with/without task-oriented penalty, (c) use average pooling, max pooling, and SRE-based pooling for cross-modality response aggregation.

F. Discussion

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

1) The B-LVPs in our method are learned from a set of binocular patch pairs suffered from only symmetric single/multiple distortion profile so that the binocular quality perception under the asymmetric single/multiple distortion condition may not be fully exploited. In the near 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 local RF properties in response to various MDSIs can be better simulated.

2) According to the existing evidences found in visual physiology [71], our method resort to the traditional sparse coding approach 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.

3) The proposed method still follows the general learning-based NR-IQA framework which requires subjective ratings as ground truth 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 (V1) 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 Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

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

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

[4] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Broadcasting*, vol. 61, no. 3, pp. 520-531, Jul. 2015.

[9] S. H. Bae and M. Kim , "A novel image quality assessment with globally and locally consilient visual quality perception," *IEEE Transactions on Image Processing*, 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 Transactions on Image Processing*, vol. 26, no. 4, pp. 1799-1809, Apr. 2017.

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

[12] J. Schild, J. LaViola, and M. Masuch, M, "Understanding user experience in stereoscopic 3D games," *In Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 89-98, May 2012.

[13] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "Quality of experience model for 3DTV," *In Proc. of Stereoscopic Displays and Applications XXIII*, International Society for Optics and Photonics, 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 Processing: Image Communication*, 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 710-720, Jun. 2012.

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

[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 Transactions on Circuits and Systems for Video Technology*, 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 Transactions on Broadcasting*, vol. 59, no. 1, pp. 28-37, Jan. 2013.

[24] Q. Jiang, F. Shao, G. Jiang, M. Yu, and Z. Peng, "Three-dimensional visual comfort assessment via preference learning," *Journal of Electronic Imaging*, vol. 24, no. 4, 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 Processing Letters*, 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 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Pattern Analysis and Machine Intelligence*, 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. of the IEEE Computer Vision and Pattern Recognition*, 2001.

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

[33] V. A. Lamme, H. Super, R. Landman, P. R. Roelfsema, and H. Spekreijse, "The role of primary visual cortex (V1) in visual awareness," *Vision Research*, vol. 40, no.10-12, pp. 1507-1521, 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 Neuroscience*, vol. 3, no. 11, 1153, 2000.

[35] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, 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," *The Journal of Physiology*, 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 Computation*, 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," *Journal of Neurophysiology*, 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 Processing Letters*, vol. 21, no. 6, pp. 751-755, Jun. 2014.

[40] Z. Wang, A. Bovik, and B. Evans, "Blind measurement of blocking artifacts in images," in *Proc. of the IEEE International Conference on Image Processing*, 2000, pp. 981-984.

[41] H. Liu, N. Klomp, and I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 529-539, Apr. 2010.

[42] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. Chen, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 4559-4565, Jan. 2017.

[43] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Image Processing*, 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 Transactions on Multimedia*, vol. 17, no. 1, pp. 50-63, Jan. 2015.

[48] Q. Wu, H. Li, F. Meng, K. N. Ngan, B. Luo, C. Huang, and B. Zeng, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 425-440, Mar. 2016.

[49] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, 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. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[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 Transactions on Image Processing*, 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," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 123-133, Nov. 2015.

[54] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, and H. Sun, "Optimizing Multi-Stage Discriminative Dictionaries for Blind Image Quality Assessment," *IEEE Transactions on Multimedia*, 2017.

[55] A. Mittal, R. Soundararajan, and A.C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209-212, Mar. 2013.

[56] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579-2591, Aug. 2015.

[57] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Transactions on Broadcasting*, vol. 60, no. 3, pp. 555-567, Jul. 2014.

[58] 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 Processing Letters*, vol. 22, no. 10, pp. 1811-1815, Oct. 2015.

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

[60] H. Hadizadeh and I. Bajic, "Color Gaussian Jet features for no-reference quality assessment of multiply-distorted images," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1717-1721, Dec. 2016.

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

[62] 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 Transactions on Image Processing*, vol. 24, no. 5, pp. 1685-1699, May 2015.

[63] B. Appina, S. Khan, and S. S. Channappayya, "No-reference stereoscopic image quality assessment using natural scene statistics," *Signal Processing: Image Communication*, vol. 43, pp. 1-14, 2016.

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

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

[66] 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 Transactions on Image Processing*, vol. 25, no. 5, pp. 2059-2074, May 2016.

[67] 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 Processing: Image Communication*, vol. 58, pp. 287-299, 2017.

[68] 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 Recognition*, vol. 59, pp. 176-187, 2016.

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

[70] 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 Transactions on Multimedia*, vol. 19, no. 8, pp. 1821-1836, Aug. 2017.

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

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

[73] Z. Jiang, Z. Lin, L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, vol. 35, no. 11, pp. 2651-2664, Nov. 2013.

[74] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. of the IEEE Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pp. 40-44, Nov. 1993.

[75] 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 Transactions on Visualization and Computer Graphics*, vol. 21, no. 1, pp. 56-67, Jan. 2015.

[76] 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.

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

[78] 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. of the IEEE International Conference on Computer Vision*, pp. 416-423, 2001.

[79] A. K. Moorthy, C. C. Su, A. Mittal, and A. C. Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 870-883, 2013.

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

[81] R. Akhter, Z. P. Sazzad, Y. Horita, and J. Baltes, "No-reference stereoscopic image quality assessment," in *Proc. of the Stereoscopic Displays and Applications XXI, International Society for Optics and Photonics*, vol. 7524, p. 75240T, Feb. 2010.