Blind Image Quality Assessment with Hierarchy: Degradation From Local Structure to Deep Semantics

Jinjian Wu[a,*], Jichen Zeng[a], Weisheng Dong[a], Guangming Shi[a], Weisi Lin[b]

[a] *School of Artificial Intelligence, Xidian University, Xi'an, 710071, China*
[b] *School of Computer Engineering, Nanyang Technological University, 639798, Singapore*

Abstract

Though blind image quality assessment (BIQA) is highly desired in perceptual-oriented image processing systems, it is extremely difficult to design a reliable BIQA method. With the help of the prior knowledge, the human visual system (HVS) hierarchically perceives the quality degradation during the visual recognition. Inspired by this, we suggest different levels of distortion generate individual degradations on hierarchical features, and propose to consider the degradations on both low and high level features for quality prediction. By mimicking the orientation selectivity (OS) mechanism in the primary visual cortex, an OS based local structure is designed for low-level visual information representation. At the meantime, the deep residual network, which possesses multiple levels for feature integration, is employed to extract the deep semantics for high-level visual content representation. By fusing the local structure and the deep semantics, a hierarchical feature set is acquired. Next, the correlations between the degradations of image qualities and their corresponding hierarchical feature sets are analyzed, and a novel hierarchical feature degradation (HFD) based BIQA (HFD-BIQA) method is built. Experimental results on the legacy and wild image quality assessment databases demonstrate the prediction accuracy of the proposed HFD-BIQA method, and verify that the HFD-BIQA performs highly consistent with the subjective perception.

Keywords: Blind Image Quality Assessment, Hierarchical Feature Degradation, Local Structure, Deep Semantics

☆This work was supported by the National Natural Science Foundation of China (No. 61772388).

*Corresponding author
Email address: jinjian.wu@mail.xidian.edu.cn (Jinjian Wu)

1. Introduction

With the tremendous increase of digital photographs in our daily life, it is highly desired to faithfully evaluate the visual qualities in many signal processing systems, e.g., digital signal acquisition, compression, transmission, and so on [1]. Though the subjective image quality assessment (IQA) by human returns credible evaluation result, it is cumbersome, laborious, and cannot be embed into the real-time signal processing system [2]. Thus, how to design a reliable objective IQA method, which performs consistently with the subjective perception, has became one of the most challenging issues in image processing and computation vision societies.

A large amount of IQA methods have been introduced in the last decade. The largest number of these IQA methods are full-reference (FR, e.g., the peak signal-to-noise ratio and structure similarity [3]) and reduced-reference (RR, e.g., reduced-reference entropic differencing [4] and reduced-reference IQA with visual information fidelity [5]), for which the whole reference image or part of the reference information are required. However, the reference information is unavailable for most situations, and thus the application scopes for FR and RR IQAs are severely limited. No-reference (NR) IQA, which requires no more reference information during quality evaluation [6], has attracted increasing interest in recent years. And this work focuses on developing a novel NR IQA method.

Without the help and guidance from the reference information, it becomes extremely difficult for NR IQA to accurately evaluate the quality of images [7]. Early NR IQA methods commonly use the prior knowledge of the distortion type for quality prediction, which are called distortion-specific NR IQA [2, 8]. For such type of methods, the distortion-specific features are extracted for quality prediction. e.g., sharpness for blur [9], blockiness for JPEG [10], ringing for JPEG2000 [11], and so on. These distortion-specific NR IQAs have a limited application scope, which only work for a certain type of distortion.

Recently, the non-distortion-specific NR IQA methods have been emphatically studied [12–14], for which the prior knowledge of distortion is unavailable and is called blind IQA (BIQA). In general, some kind of statistical characteristic on low-level features are analyzed on a vast number of images, and a common prior knowledge is learned to guide the BIQA. In [15], Moorthy et al.

learned the natural scene statistical (NSS) with the generalized Gaussian distribution (GGD) in the wavelet domain, and measured the quality with the changes on the GGD coefficients (called DIIVINE). Following DIIVINE, Saad et al. [16] extended the NSS characteristic to the DCT domain, and proposed a BLIINDS method for BIQA. Moreover, Mittal et al. [17] directly calculated the NSS feature in the spatial domain with both GGD and asymmetric GGD, and introduced the BRISQUE for quality estimation. In the recent, Zhang et al. [18] integrated a large set of NSS features in several domains, and proposed the IL-NIQE for BIQA. Besides these NSS based methods, Ye and Doermann [19] trained a codebook directly from image block to guide BIQA. Liu et al. [20] analyzed the spatial and spectral entropies for quality assessment. And Zhang et al. [21] learned a local quantized pattern based visual codebook for distortion estimation. Though these low-level feature based methods have greatly improved the BIQA performance, there still exist a large gap between the objective method and the human subjective perception.

In order to design a more reliable objective BIQA method, we turn to investigate the characteristic of the human visual system (HVS) during visual signal processing. It is well known that the visual perception in the HVS is classically modeled as a hierarchy with increasingly sophisticated representations, i.e., from simple low-level structure (e.g., edge and line) to complicated high-level semantics (e.g., object and categories) [22, 23]. Thus, besides the degradation on the low-level structure, we also need consider the degradation on the high-level semantics for quality prediction.

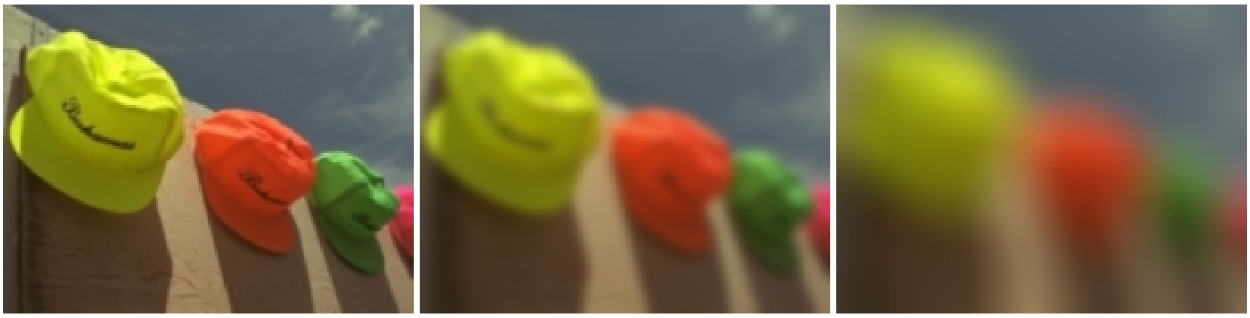
By hierarchically learning high-level representation with multiple hidden layers, the deep neural network (especially the convolutional neural network (CNN)) has been adopted for BIQA. In [24], the CNN was adopted to automatically extract image features (without hand-crafted features) for BIQA. Moreover, the predicted qualities from CNN for patches of an image were weighted pooled according to their magnitudes in [25]. However, these CNN based BIQA methods mainly predict the quality with the degradation on the high-level semantics (i.e., the last layer of the CNN), and have not fully considered the degradation on the low-level structure (the first few layers which represent the low-level features are difficult to be used, because the number of them is huge and these optimized filters can not directly represent local structures). Meanwhile, with limited size of the IQA database (the largest IQA database, TID2013 [26], contains only 25 reference images and 3000 corresponding distorted images), it is hard to optimize the huge number of coefficients in the CNN. As a result, the performance of these CNN based BIQA methods are always unstable on the public available IQA databases.

In this work, we introduce a novel BIQA method based on hierarchical feature degradation (HFD). The primary visual cortex presents obvious orientation selectivity (OS) mechanism for low-level feature extraction [27, 28]. Inspired by this mechanism, an OS based local structure has been designed for low-level feature extraction. Meanwhile, with multiple processing layers to learn hierarchical representations of data, the later layers of the deep neural network can efficiently represent the high-level feature of visual contents [29]. As one of the most powerful deep learning architectures, the residual network (ResNet) [30] is adopted for deep semantics extraction. Next, the local structure and deep semantics are fused for HFD analysis. By analyzing the correlation between the perceptive quality and the degradation on the hierarchical features with support vector regression (SVR), a novel HFD based BIQA (HFD-BIQA) method is proposed. Experimental results demonstrate that the proposed HFD-BIQA has a remarkable improvement against these existing methods.

The main contributions of our model are as follows

- Firstly, we thoroughly analyze the hierarchical degradation from different distortion levels, and suggest to consider the degradations on both low and high level features for quality prediction.
- Secondly, an orientation selectivity based local structure is designed to extract the low-level feature; combing with the high-level feature obtained from deep learning network, a hierarchical feature set is built.
- Finally, by analyzing the degradation on the hierarchical feature set, a novel HFD-BIQA method is proposed. The HFD-BIQA presents promising performance.

The rest of this paper is organized as following. In Section 2, the hierarchical visual quality degradation is firstly analyzed. And then, the hierarchical feature set is built for HFD-BIQA method modeling in Section 3. In Section 4, comparative studies of the HFD-BIQA with the



(a) PSNR=36.71dB (b) PSNR=26.37dB (c) PSNR=20.93dB

Figure 1: Hierarchical visual quality degradation under different noise levels.

existing IQA methods on both legacy and wild IQA databases are demonstrated. Finally, some conclusions are drawn in Section 5.

2. Hierarchical Feature Degradation

Researches on cognitive neuroscience indicate that the HVS is a hierarchy of cortical areas, in which the input visual signal is hierarchically processed with increasingly sophisticated representation (from local features to global abstract/semantics) [22, 23, 31]. For an input visual signal, the primary visual areas (V1 and V2) are highly adapted to extract simple features (e.g., local edge and orientation). By integrating these simple features from the primary visual areas, the successive areas (V3, V4, and medial-temporal area) generalize more complicated and regional representations (e.g., contour and shape). Then, the contour/shape information is further integrated at the high-level visual areas (inferotemporal and prefrontal areas), and finally generate the global semantics (e.g., abstract and categories) for visual recognition and scene understanding.

Inspired by the hierarchical feature extraction and visual recognition in the HVS, we suggest distortions will generate individual degradations on the hierarchical features. Moreover, different levels of distortions cause different destructive effects on these hierarchical features. As shown in Fig. 1, the original *Hats* scene is distorted by three different levels of Gaussian blur noise (WBN), which cause different quality degradations (Fig. 1 (a) has a much better quality than Fig. 1 (b), while Fig. 1 (c) has the worst quality). With further analysis on noise level, we can see that a weak noise level (PSNR=36.71dB) in Fig. 1 (a) has slightly blurred the local edge, while has little effect on the shape of the hats. In other words, the weakly WBN only degrades the low-level feature, while has no influence on the high-level semantics on Fig. 1 (a). With the increasing of noise level, the local edge in Fig. 1 (b) (with PSNR=$26.37dB$) is severely distorted. Though the shape of the hat and the characters are obviously destroyed, the main concept can still be extracted (i.e., understanding the general hats in this image). With further increasing of the noise level (PSNR=$20.93dB$), the local edge and the regional shape in Fig. 1 (c) are seriously distorted, which made it impossible to extract the accurate concept (hats or air balloon or something else) for recognition.

Therefore, different noise levels usually cause different degradations on these hierarchical features. Weak noise mainly effects the low-level features, and has limited effect on the high-level features. And thus, the perceptual quality of an image is usually good under weak noise. Strong noise not only severely distorts the low-level structure, but also directly destroys the high-level semantics, which results in obvious quality degradation. In order to perform more consistent with the subjective perception, we need consider the degradations on multi-levels of features (e.g., low and high level features) for BIQA modeling.

3. Blind Quality Measurement

In this section, the low-level feature extraction with the OS based local structure is firstly introduced. Then, the high-level feature from the latest layer is extracted for deep semantics representation. Finally, the degradation on both low and high features are analyzed for BIQA modeling. The architecture of the proposed BIQA model is shown in Fig. 2.

3.1. Local Visual Structure Extraction

The HVS is highly sensitive to changes on image structure, and thus the structural degradation is widely used for quality assessment [5, 32]. Neuroscience researches have demonstrated that neurons on the primary visual cortex present substantial OS mechanism for low-level structure extraction [27, 28]. Moreover, the OS arises from the intracortical responses (i.e., excitatory and inhibitory interactions) among cortical cells in a local receptive field [33]. Inspired by the OS

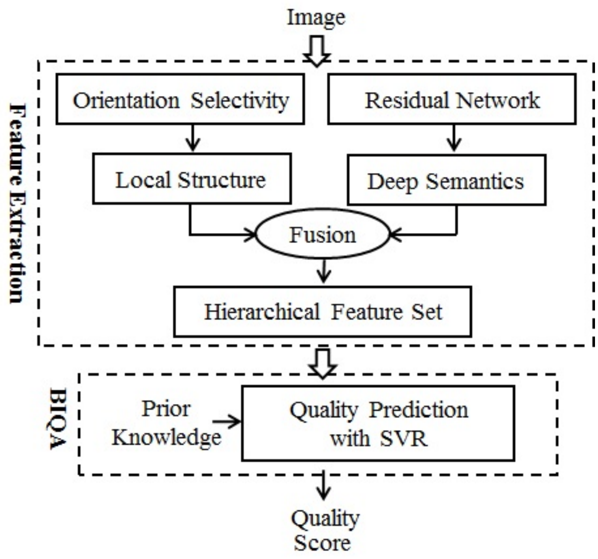


Figure 2: The architecture of the proposed BIQA model.

mechanism, we try to describe the local structure ($\mathcal{S}_l$) with response intensity ($\mathcal{I}_r$) and response pattern ($\mathcal{P}_r$) in a local neighborhood.

It is well known that the HVS is extremely sensitive to luminance changes, and the response intensity is directly related to the luminance change. Thus, for a given image I, the local structure intensity of each pixel can be demanded as its luminance change, and is calculated as,

$$\mathcal{I}_r(x) = \sqrt{(G_h(x))^2 + (G_v(x))^2}, \tag{1}$$

where $G_h(x)$ and $G_v(x)$ are the changes alone horizontal and vertical directions.

Visual pattern, which represents the repeated local content in an image, has been widely used in visual recognition works [34]. The response pattern $\mathcal{P}_r$ that a local receptive field represents is determined by the arrangement of intracortical responses (i.e., excitatory and inhibitory interactions). Moreover, neighbor neurons with similar preferred orientations always present excitatory interactions, and these dissimilar ones present inhibitory interactions [35]. Inspired by this, we try to describe the pattern $\mathcal{P}_r(x)$ of a pixel as the arrangement of interactions between the central pixel x and its local neighbors ($\mathcal{R}(x)$=$\{x_1, x_2, \cdots, x_n\}$),

$$\mathcal{P}_r(x) = \mathcal{A}(\mathcal{I}(x|x_1), \mathcal{I}(x|x_2), \cdots, \mathcal{I}(x|x_n)), \tag{2}$$

where $\mathcal{A}$ represents the spatial arrangement, and $\mathcal{I}(x|x_i)$ is the interaction type between two pixels,

$$\mathcal{I}(x|x_i) = \begin{cases} 1 & \text{if } |\theta(x) - \theta(x_i)| < \mathcal{T} \\ 0 & \text{else} \end{cases}, \tag{3}$$

$$\theta(x) = \arctan \frac{G_v(x)}{G_h(x)}, \tag{4}$$

where '1' ('0') represents excitation (inhibition) interaction. The parameter $\mathcal{T}$ judges the interaction type, and in this work we set $\mathcal{T}$=6° according to the visual masking threshold [36].

With the arrangement of binary interaction type ('0' or '1'), the number of pattern generated with Eq. (2) is growing exponentially with the pixel number in $\mathcal{R}(x)$ (i.e., 2^n different types). As a result, a 5×5 local region (i.e., n=24) will present more than 10 million (i.e., 2^{24}) different pattern forms, which is too huge for structure representation. With further analysis, we have found that not all of these patterns appeared equally (some types of patterns are more frequently appeared, e.g., patterns which represent smooth and edge regions). Moreover, some patterns have similar format and represent homogeneous visual contents. Therefore, we can select these representative patterns for visual structure representation.

In order to select these representative patterns, the often used saliency objective detection database (has no overlap/correlation with all of these IQA databases) [37], which contains 1000 different scenes, is chosen. Firstly, 200 images are randomly chosen from the database. Then, the pattern form for each pixel is calculated with Eq. (2). With all of these patterns from these 200 images, the K-Means clustering algorithm is employed for representative pattern selection,

$$\{\hat{\mathcal{P}}_r^k, k = 1, 2, \cdots, K\} = \arg\min \sum_{k=1}^{K} \sum_{m=1}^{M} ||w_m \cdot (\mathcal{P}_r^m - \hat{\mathcal{P}}_r^k)||^2, \tag{5}$$

where K is the number of representative patterns, $\hat{\mathcal{P}}_r^k$ represents the k-*th* clustering centroid, and we set K=1000 in this work (to make sure that the numbers of the low and high level features are the same). w_m is the weight factor and is computed as the appearance probability of $\mathcal{P}_r^m$ (i.e., the proportion of pattern $\mathcal{P}_r^m$ among all patterns that appears in the 200 chosen images).

Furthermore, in order to verify the robustness of the representative selection result, we have repeated this procedure (randomly choosing 200 images for clustering) for many times, and we

(a) I1: Lady-Face (b) I2: Green-House (c) Local structure based histogram

Figure 3: An example of Local structure based low-level visual content extraction, where different images present individual histograms

have found that the returned representative pattern sets are extremely similar, which confirms that the proposed procedure can efficiently select these fundamental representative patterns from nature scenes for low-level visual content extraction.

With Eqs. (1) and (5), the response intensity ($\mathcal{I}_r$) and response pattern ($\hat{\mathcal{P}}_r$) for each pixel are calculated for its local structure representation. And the low-level visual content ($\mathcal{F}_l$) of an image can be mapped into a structure based histogram,

$$\mathcal{F}_l(k) = \sum_{x=1}^{N} \mathcal{I}_r(x) \cdot \delta(\hat{\mathcal{P}}_r^x, \hat{\mathcal{P}}_r^k) \tag{6}$$

$$\delta(\hat{\mathcal{P}}_r^x, \hat{\mathcal{P}}_r^k) = \begin{cases} 1 & \text{if } \hat{\mathcal{P}}_r^x = \hat{\mathcal{P}}_r^k \\ 0 & \text{else} \end{cases}, \tag{7}$$

where N is the number of pixels in an image, and $\hat{\mathcal{P}}_r^x$ represents the pattern form that pixel x belongs to. An intuitive example of low-level visual content representation with Eq. (6) is shown in Fig. 3. We can see that different images with individual visual contents represent different histogram forms.

3.2. Deep Visual Semantics Extraction

The high-level visual feature plays a key role in visual perception. As the highest visual area of the HVS, the inferotemporal cortex (IT) integrates the former outputs and generates the high-level

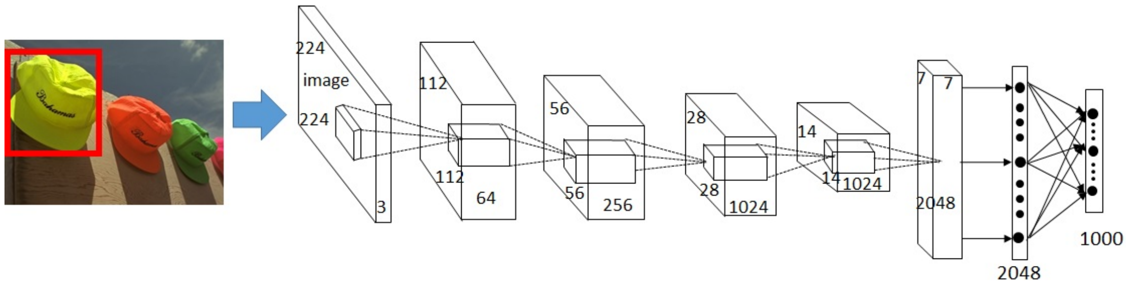


Figure 4: Architecture of the 50-layer ResNet for deep semantics extraction (the latest layer with 1×1000 features).

visual feature (e.g., abstract) for objective recognition [38]. Thus, distortions on the high-level feature directly disturb the understanding of the visual content, which result in severe quality degradation.

Deep learning network can efficiently extract high-level feature for visual recognition. With the inspiration of the hierarchy in the HVS for visual perception, deep learning network uses multiple processing layers to learn and integrate representations, and assemble high-level feature (i.e., deep semantics) in the later layers [29, 39]. Moreover, with the increase of stacked layer number (i.e., the depth of the network), more complex and enrich semantics information can be acquired in the later layers. Therefore, the deep learning network has been directly used for BIQA [24, 25]. However, with the size limitation of the existing IQA database (the largest one contains only 3000 distorted images, and all of them are generated by 25 original scenes/reference images), it is hard to optimize the huge (tens of thousands) coefficients in the network.

Different from these existing deep learning based BIQA, we only need to extract the high-level features from images for HFD based PKB creation. Thus, these existing trained deep learning networks, which are succeed in objective detection or recognition, can be directly adopted for high-level feature extraction. As a powerful and deeper neural network, the trained ResNet [30] is adopted for deep semantics extraction in this work. Considering the efficiency and computational complexity, the standard 50-layer ResNet (with batch normalization and average pool for regularization) is chosen, whose architecture is shown in Fig. 4. And the output of the latest layer (with 1×1000 features) is used as the deep semantics information (i.e., $\mathcal{F}_h \in R^{1\times1000}$). Since no retraining procedure is required (the ResNet was trained by stochastic gradient descent with backpropaga-

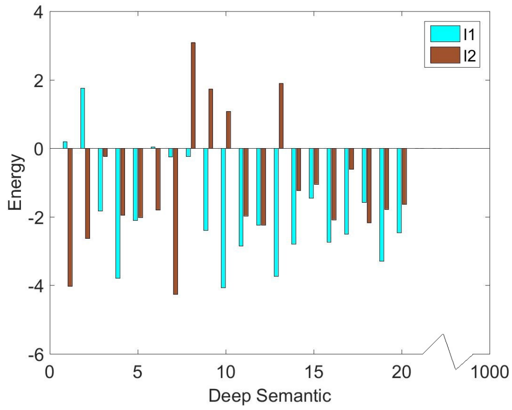


Figure 5: An example of deep semantics based high-level visual content extraction, and its corresponding original scenes are shown in Fig. 3

.

tion on the ImageNet dataset), this step can overcome the size limitation of the IQA database. An intuitive example of high-level visual content representation with deep semantics is shown in Fig. 5 (the corresponding original scenes are shown in Fig. 3). It is obvious that the two different original scenes (i.e., the Lady-Face and Green-House) possess different high level features, which confirms the efficiency of the deep semantics extraction procedure.

3.3. Blind Quality Assessment

As analyzed in Section 2, different distortion type/level generate different changes on hierarchical features. Thus, we try to measure the hierarchical degradation for quality prediction. Firstly, the low/high level features are normalized for fusion,

$$\hat{\mathcal{F}}_i(j) = \frac{\mathcal{F}_i(j)}{\sqrt{\sum_n (\mathcal{F}_i(n))^2}}, \tag{8}$$

where $\mathcal{F}_i$ represents the local structure ($\mathcal{F}_l$) or the deep semantics ($\mathcal{F}_h$), and $\hat{\mathcal{F}}_i(n)$ is the n-th normalized feature.

Next, the two types of features are combined and the hierarchical feature set (i.e., $\mathcal{F}=\{\hat{\mathcal{F}}_l, \hat{\mathcal{F}}_h\}$) is acquire for quality degradation analysis. The correlations between the hierarchical feature sets ($\mathcal{F}$) and the subjective quality scores (Q, i.e., MOS or DMOS) of distorted images are analyzed. As an efficient regression procedure from a high dimension to a lower one, the classical support vector regression (SVR) is adopted to learn the mapping relationship between $\mathcal{F}$ and Q. In this work, the LIBSVM [40] with the radial basis function kernel is used,

$$\mathcal{M}_d = \mathrm{SVR}_{\mathrm{learn}}(\mathcal{F}, Q). \tag{9}$$

Finally, with the guidance of the prior degradation knowledge ($\mathcal{M}_d$), the quality of an image I can be predicted as,

$$\hat{Q}(I) = \mathrm{SVR}_{\mathrm{predict}}(\mathcal{F}(I), \mathcal{M}_d), \tag{10}$$

where $\mathcal{F}(I)$ is the hierarchical feature set of the input image I, and $\hat{Q}$ is the predicted quality score

4. Experimental Result Analysis

In this section, the databases and protocols that used in the experiments are firstly given. Then, the efficiency of the HFD is illustrated. Next, the prediction accuracy of the HFD-BIQA method is demonstrated by comparing with the existing state-of-the-art BIQA methods on the public available databases. Finally, the robustness of the HFD-BIQA method is testify through cross-validation experiments on different databases.

4.1. Database and Protocol

Four large-scale IQA databases are chosen for experimental result analysis, including three legacy databases and one wide database. The three legacy databases, i.e., CSIQ [41], LIVE [42], and TID2013 [26], are composed by several types of distortions under different noise levels. The CSIQ database contains 866 images (30 original scenes degraded by 6 types of distortions under 5 noise levels). The LIVE database contains 779 images (29 original scenes degraded by 5 types of distortions under 7 noise levels). And the TID2013 contains 3000 images (25 original scenes degraded by 24 types of distortions under 5 noise levels). While the wild database, i.e., the LIVE In the Wild Image Quality Challenge Database (Wild-LIVE for short) [43], contains 1163 different original scenes and each is distorted by a wide variety of randomly occurring and unknow mixture distortion types.

In order to verify the performance of IQA methods on these databases, three classical criteria are adopted in this experiment, which are the Spearman rank order correlation coefficient (SRCC), the Pearson linear correlation coefficient (PLCC), and the root mean squared error (RMSE). The correlation between the predicted qualities (i.e., the quality scores from the BIQA model) and the ground truth scores (i.e., MOS/DMOS) are analyzed with these criteria. The SRCC represents the prediction monotonicity, and a better IQA method returns a larger SRCC value. The PLCC measures the prediction accuracy (the higher PLCC the better performance), and the RMSE represents the prediction deviation (the smaller RMSE the better performance). More details about the three criteria can be found in [44].

When using SVR for quality prediction, a training procedure is required in the regression module. Similar to the training procedure in these existing BIQA methods (e.g., in [21, 45]), we randomly divide the images that a database contained into two subsets (training and testing subsets). To make sure that there is no overlap between the two subsets, 80% original scenes are randomly selected, and their corresponding distorted images are used for training; the left 20% distorted images are used for testing. Moreover, in order to eliminate the performance bias (not governed by a specific training result), the 80% training - 20% testing procedure is repeated for 1000 times, and the median performance across the 1000 times is calculated as the final result.

4.2. Analysis on Hierarchical Degradation

The HVS hierarchically processes the input visual content, and different levels of distortion generate different degradation on the hierarchical visual features. An example is shown in Fig. 6, in which two different scenes (i.e., Lady-Face and Green-House from TID2013 [26]) are distorted by JPEG noise under different levels, and the corresponding index values are listed in Tab. 1.

Weak noise mainly degrade the local structure, and has limited influence on the deep semantics. As shown in Fig. 6 (a) and (c), the two images are distorted by weak JPEG noise (with PSNR 28.23 dB and 28.68 dB, resepctively). As can be seen, though there are obvious degradations on the local structures (e.g., the facial contour in Fig. 6 (a) and the edge of barriers in Fig. 6 (c)), we can still easily extract the primary visual contents of the two images for understanding (i.e., can still understand that Fig. 6 (a) contains a lady face, and Fig. 6 (b) is a green house). Meanwhile, the measurement with local structure can accurately represent the perceptual qualities of the two images. As listed in Tab. 1, Fig.6 (a) (with MOS=3.26) has worse subjective perceptual qual-

(a) Lady-Face with weak noise (PSNR=28.23dB)

(b) Lady-Face with strong noise (PSNR=22.88dB)

(c) Green-House with weak noise (PSNR=28.68dB)

(d) Green-House with strong noise (PSNR=21.61dB)

Figure 6: An example of hierarchical degradation on two different scenes distorted by JPEG noise under two different levels.

Table 1: An example of hierarchical degradation on two different scenes

Feat. \ Image	Fig.6 (a)	Fig.6 (c)	Fig.6 (b)	Fig.6 (d)
MOS	3.26	4.86	2.19	1.66
PSNR	28.23	28.68	22.88	21.61
Local Structure	3.27	4.64	2.41	2.53
Deep Semantics	3.61	3.20	2.11	1.92
HFD-BIQA	3.61	3.63	2.35	1.87

ity (smaller MOS value) than that of Fig.6 (c) (with MOS=4.64). And the measurement results from the local structure is 3.27 and 4.64 for them, which are consistent with the subjective perception (MOS). However, the deep semantics returns an opposite result for the two images (3.61 and 3.20 for them, which means Fig.6 (a) has better quality than Fig.6 (b)).

Strong noise severely degrades the local structure, and directly destroys the deep semantics. As a result, the quality mainly relates to the degradation on the deep semantics, and has little relationship with the degradation on the local structure. As shown in Fig. 6 (b) and (d), the two images are distorted by strong JPEG noise (with PSNR 22.88 dB and 21.61 dB, respectively). As a result, we can hardly extract complete information from the two images, e.g., the nose in Fig. 6 (b) or the roof in Fig. 6 (c). Since the local structure is severely distorted, its distortion degree cannot represent the perceptual quality anymore. As shown in Tab. 1, the measurement from the local structure returns an opposite result (Fig. 6 (b) has worse quality (2.41) than Fig. 6 (d) (2.53)) against the subjective perception (the MOS for Fig. 6 (b) and (d) are 2.19 and 1.66, respectively). The quality predictions on the two images with the deep semantics show that Fig. 6 (b) (with 2.11) has better quality than that of Fig. 6 (d) (with 1.92), which is consistent with the subjective perception.

Table 2: Comprehensive analysis of hierarchical degradation on the CSIQ Database

Feat. \ Crit.	PLCC	SRCC	RMSE
Local Structure	0.847	0.790	0.136
Deep Semantics	0.832	0.762	0.147
HFD-BIQA	0.890	0.842	0.120

The proposed HFD-BIQA can accurately represent the quality degradations on the four images in Fig. 6. By fusing both the low and high features for quality prediction, the proposed HFD-BIQA contains a hierarchical degradation measurement, which can efficiently measure the quality degradation by weak or strong noise. As shown in Tab. 1, the predicted qualities for Fig. 6 (a)-(d) are 3.61, 3.63, 2.35, and 1.87, respectively. The prediction results show that Fig. 6 (c) has the best quality, Fig. 6 (a) is the second best, and Fig. 6 (d) is the worst one, which is consistent with the subjective perception.

In order to give a comprehensive analysis on HFD, the performances of the local structure, the deep semantics, and the proposed HFD-BIQA on the whole CSIQ database [41] are compared, and the comparison results are listed in Tab. 2. By fusing the local structure and the deep semantics, the proposed HFD-BIQA has the highest PLCC and SRCC values, and the lowest RMSE value, which demonstrates that the measurement on the HFD is more consistent with the subjective perception than that on only one type of feature (i.e., the local structure or the deep semantics).

4.3. IQA Performance Comparison

4.3.1. Performance on The Legacy Databases

In order to demonstrate the performance, the proposed HFD-BIQA is firstly compared with 7 state-of-the-art BIQA methods (i.e., IMNSS [21], DL-IQA [46], IL-NIQE [18], NIQE [47], BRISQUE [17], CBIQ [19], and DIIVINE [15]) on the three legacy IQA databases.

Table 3: Performances comparison on individual distortion type of LIVE database, and the best performed BIQA method is emphasized with bold

Distortion	Crit.	HFD-BIQA	IMNSS	DL-IQA	IL-NIQE	NIQE	BRISQUE	CBIQ	DIIVINE
J2K	PLCC	**0.957**	0.950	0.947	0.918	0.927	0.923	0.913	0.922
	SRCC	**0.943**	0.934	0.928	0.905	0.914	0.914	0.903	0.937
	RMSE	**7.236**	7.580	–	9.846	9.394	9.945	9.938	9.013
JPG	PLCC	**0.971**	0.951	0.940	0.970	0.956	0.956	0.942	0.921
	SRCC	0.951	0.933	0.912	0.950	0.937	**0.956**	0.942	0.910
	RMSE	**7.614**	7.877	–	7.840	8.906	8.282	9.302	12.77
WGN	PLCC	0.979	0.982	0.955	**0.988**	0.976	0.985	0.958	0.987
	SRCC	0.972	**0.986**	0.968	0.980	0.967	0.979	0.932	0.984
	RMSE	5.761	4.419	–	4.380	5.440	**3.767**	6.31	5.047
GBN	PLCC	0.942	0.948	0.944	0.943	0.948	**0.949**	0.929	0.923
	SRCC	0.919	0.949	0.946	0.923	0.931	**0.951**	0.935	0.921
	RMSE	6.304	6.943	–	6.280	5.490	**4.656**	8.634	7.788
FFN	PLCC	**0.931**	0.922	0.890	0.879	0.888	0.903	0.904	0.888
	SRCC	**0.905**	0.895	0.861	0.851	0.861	0.877	0.856	0.863
	RMSE	**10.37**	10.56	–	13.11	12.76	13.22	13.68	11.84

Firstly, the performances of these IQA methods on the individual distortion type of LIVE database are compared. There are five different distortion types in LIVE database, namely, JPEG compression noise (JPG), JPEG2000 compression noise (J2K), white Gaussian noise (WGN), Gaussian blur noise (GBN), and fastfading noise (FFN).

The performances of these IQA methods on each distortion type of LIVE database are listed in Tab. 3. It is apparent that the HFD-BIQA performs highly consistent with the subjective perception (the PLCC and the SRCC values are larger than 0.9 in all of these distortion types). More concretely, the HFD-BIQA performs the best on three types of distortion (i.e., J2K, JPG, and FFN) among these BIQA methods, and performs a slightly worse than the best one on the other two types. In summary, the HFD-BIQA gains 8 of 15 (3 criteria $\times$ 5 distortion type) best performance among these BIQA methods.

Besides on individual distortion type, the overall performance on the whole database is further

Table 4: Performance Comparison on the whole database (LIVE, CSIQ and TID2013), and the best performed BIQA method is emphasized with bold

DB	Crit.	HFD-BIQA	IMNSS	DL-IQA	IL-NIQE	NIQE	BRISQUE	CORNIA	DIIVINE
LIVE	PLCC	**0.951**	0.943	0.930	0.905	0.908	0.929	0.937	0.892
	SRCC	**0.948**	0.944	0.927	0.902	0.908	0.920	0.938	0.882
	RMSE	**8.437**	8.705	–	11.622	11.423	10.421	9.645	12.33
CSIQ	PLCC	**0.890**	0.835	–	0.863	0.726	0.812	0.750	0.804
	SRCC	**0.842**	0.789	–	0.822	0.629	0.748	0.676	0.776
	RMSE	**0.120**	0.142	–	0.130	0.179	0.154	0.172	0.154
TID2013	PLCC	**0.764**	0.598	–	0.641	0.421	0.626	0.552	0.643
	SRCC	**0.681**	0.522	–	0.518	0.330	0.571	0.434	0.567
	RMSE	**0.797**	0.997	–	0.955	1.130	0.931	1.035	0.952
Mean	PLCC	**0.868**	0.792	–	0.803	0.685	0.789	0.746	0.780
	SRCC	**0.824**	0.752	–	0.747	0.622	0.746	0.683	0.742

analyzed. The performance results of these IQA methods on the three legacy databases (LIVE, CSIQ, and TID2013) are listed in Tab. 4. By comparing with these BIQA methods,we can see that the prediction accuracy of the HFD-BIQA is completely higher than the others (with larger SRCC and PLCC values, and smaller RMSE values on all of the three databases). Especially for the TID2013 (the largest database, on which the existing IQA methods usually perform no good enough), the HFD-BIQA achieves a remarkable improvement against these existing BIQA (the PLCC of the HFD-BIQA VS. the second best on TID2013 is 0.764:0.643, and the SRCC is 0.681:0.571). Furthermore, the weighted mean (weighting the the size of the database) performance of these methods on the three databases are calculated, which is tabulated at the bottom of Tab. 4. The HFD-BIQA has much larger SRCC (with 0.868) and PLCC (with 0.824) values than the other BIQA methods, which further verify the advantage of the proposed method.

Besides direct comparisons, the statistical significances of the HFD-BIQA against the other BIQA methods are calculated to further demonstrate whether the HFD-BIQA performs significantly better than others. In this work, the f-test metric [48], which counts the residuals between the quality scores for IQA methods and the subjective qualities (MOS/DMOS), is employed for

Table 5: Statistical significance comparison between the HFD-BIQA and the other BIQA methods on LIVE, CSIQ, and TID2013 Database

Algo. / DB	IMNSS	IL-NIQE	NIQE	BRISQUE	CORNIA	DIIVINE
LIVE	0	1	1	1	1	1
CSIQ	1	1	1	1	1	1
TID2013	1	1	1	1	1	1

Table 6: Performance Comparison on the Wild-LIVE Database, and the best performed BIQA method is emphasized with bold

Crit.	HFD-BIQA	IMNSS	IL-NIQE	NIQE	BRISQUE	DIIVINE	FRIQUEE
PLCC	**0.776**	0.53	0.5	0.48	0.61	0.59	0.72
SRCC	**0.760**	0.52	0.44	0.42	0.58	0.56	0.72

statistical significance measurement. And the confidence level is set as 95% in this experiment.

The comparison results from f-test about the HFD-BIQA against the other BIQA methods are listed in Tab. 5, in which a value of '1' ('-1') represents the HFD-BIQA is statistically superior (worse) than the compared method, and '0' indicates that their performances are statistically indistinguishable. As can be seen, almost all of the values in Tab. 5 are '1' (only one with '0' value), which confirms that the HFD-BIQA performs statistically better than the other BIQA methods on the three legacy IQA databases (except for LIVE database, on which the HFD-BIQA performs equivalently with IMNSS).

4.3.2. Performance on The Wild Database

Different from the legacy IQA databases (which are well-modeled by the synthetic distortions) the wild-LIVE database is composed by a large set of widely diverse authentic distorted images [43]. Therefore, it is a great challenge for NR IQA methods to accurately predict the image quality on this database. Here, the HFD-BIQA is compared with these state-of-the-art BIQA methods and a latest BIQA method (i.e., FRIQUEE [49], which achieves the best performance on the wild-LIVE until now) on the wild-LIVE database. The outputs from different BIQA methods

Table 7: Performance Comparison on TID2013 and CSIQ when Trained on LIVE

DB / Algo.	CSIQ			TID2013		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
HFD-BIQA	0.900	0.843	0.125	**0.921**	**0.899**	**0.545**
IL-NIQE	**0.906**	**0.880**	**0.119**	0.873	0.877	0.683
NIQE	0.890	0.866	0.128	0.822	0.814	0.795
BRISQUE	0.840	0.826	0.153	0.721	0.726	0.969
CBIQ	0.835	0.842	0.155	0.811	0.817	0.819
DIIVINE	0.875	0.854	0.137	0.859	0.849	0.714

are listed in Tab. 6. The HFD-BIQA has much larger PLCC and SRCC values than the five state-of-the-art BIQA methods, which means the HFD-BIQA performs obviously better than these BIQA methods. Meanwhile, the HFD-BIQA also has larger PLCC and SRCC values than that from the latest FRIQUEE method, which further confirms the superiority of the proposed method.

4.4. Cross Validation

The efficiency of the HFD-BIQA on each individual database has been demonstrated in the former subsection, here we try to prove that the HFD-BIQA is not limited by the database that it be trained. Therefore, the cross validation among the three legacy databases (i.e., LIVE, CSIQ, and TID2013) is used to demonstrate the robustness of the HFD-BIQA. Though the number and types of distortion for the three databases are different, they contain four common distortion types, i.e., WGN, GBN, JPG, and J2K. Thus, images with the four common distortion types are firstly extracted. Then, all of the images from one database is used for training, and the left images from the other two databases are used for testing.

Tab. 7 lists the performances on CSIQ and TID2013 databases when training on LIVE database.

Table 8: Performance Comparison on TID2013 and LIVE when Trained on CSIQ

DB / Algo.	LIVE			TID2013		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
HFD-BIQA	0.910	**0.918**	11.18	**0.917**	**0.888**	**0.559**
IL-NIQE	0.913	0.915	10.99	0.873	0.877	0.685
NIQE	**0.917**	**0.918**	**10.72**	0.822	0.814	0.795
BRISQUE	0.643	0.632	12.25	0.583	0.570	1.135
CBIQ	0.828	0.811	11.97	0.851	0.803	0.733
DIIVINE	0.522	0.520	13.65	0.812	0.764	0.814

As can be seen, the HFD-BIQA performs much better than other BIQA methods on TID2013 database (has the largest PLCC and SRCC values against the other BIQA methods, and the smallest RMSE value), and performs almost the same with the best one on CSIQ database (has similar PLCC, SRCC, and RMSE values with IL-NIQE).

Moreover, Tab. 8 lists the results that training on CSIQ database and testing on LIVE and TID2013 databases. Tab. 9 shows the results that training on TID2013 database and testing on LIVE and CSIQ database. It is apparent that the HFD-BIQA performs highly coincidently to the HVS (with large PLCC and SRCC values). More concretely, the HFD-BIQA always performs the best or a slightly worse than the best one as shown in these tables.

With these cross-validation results among the three legacy databases, we can conclude that the HFD based PKB can efficiently represent the generalized quality degradation, and the HFD-BIQA has achieved a remarkable and robust quality prediction accuracy under the guidance of the HFD based PKB.

Table 9: Performance Comparison on CSIQ and LIVE when Trained on TID2013

Algo. \ DB	LIVE			CSIQ		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
HFD-BIQA	0.874	0.890	13.09	0.877	0.821	0.142
IL-NIQE	0.913	0.915	10.99	**0.906**	**0.880**	**0.119**
NIQE	**0.917**	**0.918**	**10.72**	0.890	0.866	0.128
BRISQUE	0.789	0.795	11.82	0.839	0.808	0.153
CBIQ	0.663	0.617	11.98	0.824	0.794	0.159
DIIVINE	0.627	0.621	12.46	0.658	0.641	0.212

5. Conclusion

In this paper, we have introduced a novel HFD-BIQA method. Since the HVS presents a hierarchical procedure for visual signal processing, we have suggested that different levels of distortion generate individual degradations on hierarchical features. For example, weak distortion mainly degrades the low-level feature (local structure), and strong distortion directly destroys the high-level feature (deep semantics). And thus, we have proposed to consider the degradations on hierarchical features for quality assessment.

By mimicking the OS mechanism in the primary visual cortex, an OS based local structure has been designed for low-level visual content extraction. Meanwhile, the deeper residual network has been employed to extract the deep semantics for high-level visual content representation. Next, the local structure and the deep semantics have been fused to generate the hierarchical feature set. By measuring the degradations on the hierarchical feature set, the novel HFD-BIQA method has been introduced. Experimental results on the three legacy IQA databases (i.e., CSIQ, LIVE, and TID2013) have demonstrated the prediction accuracy of the HFD-BIQA, and the performance on the wild IQA database (i.e., Wild-LIVE) has further verified that the HFD-BIQA performs highly consistent with the subjective perception.

6. Reference

[1] W. Lin, C. J. Kuo, Perceptual visual quality metrics: A survey, J. Visual Communication and Image Representation 22 (4) (2011) 297–312.

[2] R. A. Manap, L. Shao, Non-distortion-specific no-reference image quality assessment: A survey, Information Sciences 301 (2015) 141–160.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.

[4] R. Soundararajan, A. Bovik, RRED indices: Reduced reference entropic differencing for image quality assessment, IEEE Transactions on Image Processing 21 (2) (2012) 517 –526.

[5] J. Wu, W. Lin, G. Shi, A. Liu, Reduced-reference image quality assessment with visual information fidelity, IEEE Transactions on Multimedia 15 (7) (2013) 1700–1705.

[6] Q. Wu, H. Li, F. Meng, K. N. Ngan, S. Zhu, No reference image quality assessment metric via multi-domain structural information and piecewise regression, Journal of Visual Communication and Image Representation 32 (2015) 205–216.

[7] Z. Ni, L. Ma, H. Zeng, C. Cai, K.-K. Ma, Gradient direction for screen content image quality assessment, IEEE Signal Processing Letters 23 (10) (2016) 1394–1398.

[8] L. Li, W. Xia, Y. Fang, K. Gu, J. Wu, W. Lin, J. Qian, Color image quality assessment based on sparse representation and reconstruction residual, Journal of Visual Communication and Image Representation 38 (2016) 550–560.

[9] J. Guan, W. Zhang, J. Gu, H. Ren, No-reference blur assessment based on edge modeling, Journal of Visual Communication and Image Representation 29 (2015) 1–7.

[10] F. Pan, X. Lin, S. Rahardja, W. Lin, E. Ong, S. Yao, Z. Lu, X. Yang, A locally adaptive algorithm for measuring blocking artifacts in images and videos, Signal Processing: Image Communication 19 (6) (2004) 499–506.

[11] H. Liu, N. Klomp, I. Heynderickx, A no-reference metric for perceived ringing artifacts in images, IEEE Transactions on Circuits and Systems for Video Technology 20 (4) (2010) 529–539.

[12] F. Gao, D. Tao, X. Gao, X. Li, Learning to rank for blind image quality assessment, IEEE Transactions on Neural Networks and Learning Systems 26 (10) (2015) 2275–2290.

[13] K. Ma, W. Liu, T. Liu, Z. Wang, D. Tao, dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs, IEEE Transactions on Image Processing 26 (8) (2017) 3951–3964.

[14] S. Wang, K. Gu, K. Zeng, Z. Wang, W. Lin, Objective quality assessment and perceptual compression of screen content images, IEEE computer graphics and applications 38 (1) (2018) 47–58.

[15] A. K. Moorthy, A. C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Transactions on Image Processing 20 (12) (2011) 3350–3364.

[16] M. Saad, A. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Transactions on Image Processing 21 (8) (2012) 3339 –3352.

[17] A. Mittal, A. Moorthy, A. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on Image Processing 21 (12) (2012) 4695–4708.

[18] L. Zhang, L. Zhang, A. C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Transactions on Image Processing 24 (8) (2015) 2579–2591.

[19] P. Ye, D. Doermann, No-reference image quality assessment using visual codebooks, IEEE Transactions on Image Processing 21 (7) (2012) 3129 –3138.

[20] L. Liu, B. Liu, H. Huang, A. C. Bovik, No-reference image quality assessment based on spatial and spectral entropies, Signal Processing: Image Communication 29 (8) (2014) 856–863.

[21] X. Xie, Y. Zhang, J. Wu, G. Shi, W. Dong, Bag-of-words feature representation for blind image quality assessment with local quantized pattern, Neurocomputing 226 (2017) 176–187.

[22] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nature Neuroscience 2 (1999) 1019–1025.

[23] S. Hochstein, M. Ahissar, View from the top: Hierarchies and reverse hierarchies in the visual system, Neuron 36 (5) (2002) 791–804.

[24] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1733–1740.

[25] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, G. Xie, No-reference image quality assessment using prewitt magnitude based on convolutional neural networks, Signal, Image and Video Processing 10 (4) (2016) 609–616.

[26] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. Kuo, Color image database TID2013: Peculiarities and preliminary results, in: 2013 4th European Workshop on Visual Information Processing (EUVIP), 2013, pp. 106–111.

[27] D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, The Journal of Physiology 160 (1) (1962) 106–154.

[28] R. Ben Yishai, R. L. Bar-Or, H. Sompolinsky, Theory of orientation tuning in visual cortex, Proceedings of the National Academy of Sciences of the United States of America 92 (9) (1995) 3844–3848.

[29] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[31] D. J. Felleman, D. C. Van Essen, Distributed hierarchical processing in the primate cerebral cortex, Cerebral cortex (New York, N.Y.: 1991) 1 (1) (1991) 1–47.

[32] Y. Fang, K. Zeng, Z. Wang, W. Lin, Z. Fang, C. W. Lin, Objective quality assessment for image retargeting based on structural similarity, IEEE Journal on Emerging and Selected Topics in Circuits and Systems 4 (1) (2014) 95–105.

[33] T. W. Troyer, A. E. Krukowski, N. J. Priebe, K. D. Miller, Contrast-invariant orientation tuning in cat visual cortex: Thalamocortical input tuning and correlation-based intracortical connectivity, The Journal of Neuroscience 18 (15) (1998) 5908–5927.

[34] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987.

[35] J. A. Cardin, L. A. Palmer, D. Contreras, Stimulus feature selectivity in excitatory and inhibitory neurons in primary visual cortex, The Journal of neuroscience : the official journal of the Society for Neuroscience 27 (39) (2007) 333–344.

[36] J. Wu, W. Lin, G. Shi, Y. Zhang, W. Dong, Z. Chen, Visual orientation selectivity based structure description, IEEE Transactions on Image Processing 24 (11) (2015) 4602–4613.

[37] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, IEEE CVPR 2009, 2009, pp. 1597–1604.

[38] L. G. Ungerleider, J. V. Haxby, 'what' and 'where' in the human brain, Current Opinion in Neurobiology 4 (2) (1994) 157–165.

[39] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, Q. Dai, Supervised hash coding with deep neural network for environment perception of intelligent vehicles, IEEE Transactions on Intelligent Transportation Systems 19 (1) (2018) 284–295.

[40] C. C. Chang, C. J. Lin, Libsvm: a library for support vector machines (2001).
URL `http://www.csie.ntu.edu.tw/ cjlin/libsvm/.`

[41] E. C. Larson, D. M. Chandler, Categorical image quality (csiq) database (2004).

[42] H. R. Sheikh, K. Seshadrinathan, A. K. Moorthy, Z. Wang, A. C. Bovik, L. K. Cormack, Image and video quality assessment research at live (2006).

[43] D. Ghadiyaram, A. C. Bovik, Massive online crowdsourced study of subjective and objective picture quality, IEEE Transactions on Image Processing 25 (1) (2016) 372–387.

[44] VQEG, Final report from the video quality experts group on the validation of objective models of video quality assessment ii, video Quality Expert Group (VQEG) (2003).
URL `http://www.vqeg.org/`

[45] K. Gu, G. Zhai, X. Yang, W. Zhang, Using free energy principle for blind image quality assessment, IEEE Transactions on Multimedia 17 (1) (2015) 50–63.

[46] W. Hou, X. Gao, D. Tao, X. Li, Blind image quality assessment via deep learning, IEEE Transactions on Neural Networks and Learning Systems 26 (6) (2015) 1275–1286.

[47] A. Mittal, R. Soundararajan, A. Bovik, Making a completely blind image quality analyzer, IEEE Signal Processing Letters 20 (3) (2013) 209–212.

[48] D. Sheskin, Handbook of parametric and nonparametric statistical procedures, Chapman & Hall/CRC, Boca Raton, 2011.

[49] D. Ghadiyaram, A. C. Bovik, Perceptual quality prediction on authentically distorted images using a bag of features approach, Journal of Vision 17 (1) (2017) 32.