Exploiting High-Level Semantics for No-Reference Image Quality Assessment of Realistic Blur Images

Dingquan Li[1,3], Tingting Jiang[1,2], Ming Jiang[1,3]

[1]National Engineering Lab for Video Technology, Cooperative Medianet Innovation Center; [2]School of EECS;
[3]LMAM, School of Mathematical Sciences & BICMR, Peking University, Beijing 100871, China
{dingquanli,ttjiang,ming-jiang}@pku.edu.cn

ABSTRACT

To guarantee a satisfying Quality of Experience (QoE) for consumers, it is required to measure image quality efficiently and reliably. The neglect of the high-level semantic information may result in predicting a clear blue sky as bad quality, which is inconsistent with human perception. Therefore, in this paper, we tackle this problem by exploiting the high-level semantics and propose a novel no-reference image quality assessment method for realistic blur images. Firstly, the whole image is divided into multiple overlapping patches. Secondly, each patch is represented by the high-level feature extracted from the pre-trained deep convolutional neural network model. Thirdly, three different kinds of statistical structures are adopted to aggregate the information from different patches, which mainly contain some common statistics (*i.e.*, the mean&standard deviation, quantiles and moments). Finally, the aggregated features are fed into a linear regression model to predict the image quality. Experiments show that, compared with low-level features, high-level features indeed play a more critical role in resolving the aforementioned challenging problem for quality estimation. Besides, the proposed method significantly outperforms the state-of-the-art methods on two realistic blur image databases and achieves comparable performance on two synthetic blur image databases.

KEYWORDS

Image Quality Assessment; Semantic; Realistic Blur; No-Reference Quality Metric; Statistical Aggregation Structure

1 INTRODUCTION

In the era of big data, images have become the primary carrier of information in human's daily life. Before ultimately received by a human observer, digital images may suffer from a variety of distortions. Quality of Experience (QoE), whose goal is to provide a satisfying end-user experience, has drawn increasing attention. To reach this goal, a critical precondition is to conduct image quality assessment (IQA). The most reliable way to assess image quality is subjective ratings, but it is often cumbersome, expensive and difficult to carry out in reality. Thus, objective IQA methods that can automatically predict image quality efficiently and effectively are needed. Objective IQA can be categorized into full-reference IQA (FR-IQA) [39, 44], reduced-reference IQA (RR-IQA) [27, 46] and no-reference IQA (NR-IQA) [9, 47]. Due to the unavailability of the reference image in most practical applications, NR-IQA is preferable but also more challenging.

In this paper, we focus on NR-IQA of realistic blur images. Blur is often induced by following reasons: (1) out-of-focus, (2) relative motion between the camera and the objects (object motion & camera shake), (3) non-ideal imaging systems (*e.g.*, lens aberration), (4) atmospheric turbulence, and (5) image post-processing steps (such as compression and denoising) [5, 10, 24]. Except the blur in Bokeh to strengthen the photo's expressiveness, it is a definite fact that unintentional blur impairs image quality.

(a) MOS=4.0637　　　　　(b) MOS=2.3413

Figure 1: The two images are from BID [4], and larger MOS indicates better subjective image quality. The three traditional methods (MDWE [21], FISH [38], LPC [10]) predict that (a) is worse than (b). Our method predicts that (a) is better than (b), which is consistent with subjective ratings.

Traditional NR-IQA methods of blur images are mainly based on the assumptions that blur leads to the spread of edges (*e.g.*, MDWE [21]), the reduction of high-frequency energy (*e.g.*, FISH [38]) or the loss of local phase coherence (*e.g.*, LPC [10]). However, these methods neglect the high-level semantic information, and can not distinguish neither between intrinsic flat regions and blurry regions, nor between structures with and without blurring. As a result, it is shown in Figure 1 that they predict the quality of a clear sky being worse than the quality of a blurry mouse, which is not consistent with human perception.

In this work, we tackle the problem by exploiting the high-level semantic features extracted from the pre-trained deep convolutional neural network (DCNN) models. First of all, since the pre-trained DCNN models (*e.g.*, AlexNet [15]) require a fixed input size, we need to determine **how to represent an image**. We compare four different image representations, and find that the multi-patch representation significantly better than the others. Secondly, we need to decide **which pre-trained DCNN model and which layer to extract image features.** We first explore the effectiveness of features extracted from different layers in a same pre-trained DCNN model, and find out high-level features from the top third or second layer more effective in realistic blur image assessment. Then we investigate the impact of different pre-trained models, and find out the one using residual learning (*i.e.*,ResNet-50 [11]) more suitable for NR-IQA of realistic blur images. Thirdly, as a result of the multi-patch representation, we derive a set of features for an image. So another question arises: **how to aggregate a set of extracted features?** One simple way is to use the mean feature vector to represent the feature set. However, it will lose important information (*e.g.*, the standard deviation in each dimension) of the feature set. So we propose three different statistical structures for feature aggregation, namely, mean&std aggregation, quantile aggregation and moment aggregation. As the dimension of the aggregated feature is still very high, we finally feed the aggregated feature into a linear regression model, known as partial least square regression (PLSR) [28], to predict the image quality.

Experiments are conducted on two realistic blur image databases (BID [4] and CLIVE [6]), as well as two synthetic blur image databases (TID2008 [26] and LIVE [31]). Our best proposal, named Semantic Feature Aggregation metric using PLSR (SFA-PLSR), is compared with the state-of-the-art methods. Experiments show that our method significantly outperforms the state-of-the-art on BID and CLIVE, and achieves comparable performance on TID2008 and LIVE. The good generalization ability of SFA-PLSR is validated by the cross dataset evaluation. We have also experimentally shown that high-level semantic features indeed play a more critical role than low-level features in resolving the challenging issue for NR-IQA of realistic blur images (see Figure 1). This indicates a new perspective of blur perception in terms of the semantic loss.

The remainder of this paper is organized as follows. Section 2 reviews the related work on NR-IQA of blur images. Section 3 introduces the benchmark databases and performance criteria. Section 4 describes our method in details. Section 5 discusses the experimental results and analysis. And conclusions are made in Section 6.

2 RELATED WORK

2.1 Learning free methods

Learning free methods use the characteristics of blur in terms of the spread of edges, the smoothing effects, the reduction of high frequency components or the loss of phase coherence.

The spread of edges can be used as a cue for blur estimation. Marziliano *et al.* [21] used the average edge spread over all detected vertical Sobel edge locations as a quality metric for blur images. It can be further improved by incorporating the concept of just noticeable blur (JNB) [5] to adapt for the perception of human visual system (HVS). Since blur is not likely to be perceived when the edge width is small enough (below the width corresponding to JNB), Narvekar and Karam [24] assigned the quality score as the percentage of edges whose blur cannot be detected.

The smoothing effects of the blur process is useful information for NR-IQA. Gu *et al.* [8] estimated image quality based on the energy-differences and contrast-differences of the locally estimated coefficients in the autoregressive parameter space. Bahrami and Kot [1] considered the content based weighting distribution of the maximum local variation, which was modeled by the generalized Gaussian distribution (GGD). The estimated standard deviation was then used as an indicator of image quality. Later they also parameterized the image total variation distribution, and predicted image quality using the standard deviation modified by the shape-parameter to account for image content variation [2].

Image blur results in the reduction of high frequency components. Vu and Chandler [38] estimated image quality using weighted average of the log-energies of the high-frequency coefficients. In [37], they generated a quality map based on a geometric mean of spectral and spatial measures. In view of the reduction of high-frequency components, the spectral measure was initially defined as the slope of the local magnitude spectrum, then rectified by a sigmoid function to account for HVS. To further consider the contrast effect, the spatial measure was calculated by the local total variation. Sang *et al.* [30] estimated image quality using the exponent of the truncated singular value curve of an image. Li *et al.* [16] considered the moment energy, which can be affected by noticeable blur.

Blur also causes the loss of phase coherence, which gives a different perspective for understanding blur perception [40]. So Hassen *et al.* [10] estimated image quality based on the strength of the local phase coherence near edges and lines.

2.2 Learning-based methods

Traditional learning free methods can not accurately express the diversity of blur process and the complexity of HVS. So recently machine learning technologies appear in IQA field. Learning-based methods mainly consist of two steps: feature extraction and quality prediction. In terms of feature extraction, these methods fall into two classes: the one using hand-crafted features and the other using learnt features.

Features can be manually designed using the natural scene statistics (NSS) of the image. NSS models of image coefficients in the spatial domain, wavelet domain and DCT domain are utilized in [22, 23, 29] to extract quality relevant features, respectively. Tang *et al.* [35] derived a set of low-level image quality features from NSS models, texture characteristics and blur/noise estimation. Ciancio *et al.* [4] used a neural network

to combine eight existing methods and low-level features for blur image quality assessment. Oh *et al.* [25] evaluated image quality of camera-shaken images through mapping the spectral direction and shape features using support vector regression (SVR). Li *et al.* [18] took gradient similarity, singular value similarity and DCT domain entropies as quality features in a multi-scale framework. Li *et al.* [19] jointly considered the structural and luminance information in predicting image quality, where the structure information was described by the local binary pattern distribution and the normalized luminance magnitudes distribution portrayed the luminance information.

Machine learning techniques can learn quality relevant features. Li *et al.* [17] and Lu *et al.* [20] extracted learnt features based on dictionary learning. Visual codebook is used to learn quality features in [41, 43]. Convolutional neural networks (CNN) have also been used to learn quality relevant features in NR-IQA [3, 13, 14, 32, 33, 45]. Kang *et al.* [13] integrated feature learning and patch quality prediction into an end-to-end network, and the image quality was estimated by the average score of all sampling patches. Following [13], the network was deeper and weights for patch scores were also integrated into the learning process [3]. In [45], CNN was used to learn features and the general regression neural network was used as the predictor. In [14], a sub-network was first trained on patches using the FR-IQA scores, and then a whole network from images to quality was trained.

The most related works to ours are [32, 33], which resize the image to meet the required input size of the pre-trained AlexNet so as to extract the image features. Our work differs from them mainly in three ways: (1) Unlike [32, 33], we use multiple overlapping image patches instead of the resized image to represent the image, which can avoid introducing deformation as well as cover the image information. Correspondingly, we propose three effective statistical structures to conduct feature aggregation. (2) The features extracted from the pre-trained DCNN model in [32, 33] are only used as the auxiliary to boost the performance of methods based on low-level features, while our aggregated semantic features are directly used as quality relevant features. (3) We focus on realistic blur, and since residual images contain important cues about image blur, the residual learning based network (ResNet-50 [11]) is selected as the feature extractor instead of the one in [32, 33] without residual learning.

3 BENCHMARK DATABASES AND PERFORMANCE CRITERIA

3.1 Benchmark Databases

In this work, we consider two realistic image databases (BID [4] and CLIVE [6]), as well as two synthetic blur image datasets from TID2008 [26] and LIVE [31].

BID includes totally 586 realistic blur images taken from real world along with a variety of scenes, light conditions, camera apertures and exposure time. Subjective quality scores are provided in the form of mean opinion score (MOS) ranging from 0 to 5.

CLIVE includes 1162 realistic distorted images captured using real-world mobile cameras, most of which suffer from motion blur or out-of-focus blur. Subjective quality scores are provided in the form of MOS ranging from 0 to 100.

TID2008 contains 1700 distorted images, in which we only consider the 100 Gaussian blur images. There are only 25 reference images, and 4 blur kernels for each reference image. Subjective quality scores are provided in the form of MOS ranging from 0 to 9.

LIVE contains 779 distorted images, in which we only consider the 145 Gaussian blur images. There are only 29 reference images, and 5 blur kernels for each reference image. Subjective quality scores are provided in the form of Difference of MOS (DMOS) ranging from 0 to 100.

3.2 Performance Evaluation Criteria

Three evaluation criteria are chosen to evaluate the performance of NR-IQA methods: Spearman's rank-order correlation coefficient (SROCC), Pearson's linear correlation coefficient (PLCC) and root mean square error (RMSE). PLCC and RMSE are used for measuring prediction accuracy, while SROCC is used for measuring prediction monotonicity. For these three criteria, larger PLCC/SROCC and smaller RMSE indicate better performance. Before calculating PLCC and RMSE values of the learning free methods, a nonlinear fitting is needed to map the objective scores to the same scales of the subjective scores. In this paper, we adopt the the following four-parameter logistic function recommended in [36].

$$f(x) = \frac{\tau_1 - \tau_2}{1 + e^{\frac{x - \tau_3}{\tau_4}}} + \tau_2 \tag{1}$$

where τ_1 to τ_4 are free parameters to be determined during the curve fitting process.

Monte-Carlo cross validation is used for learning-based methods. For each database, 80% data are for training and 20% data are for testing. There is no same "original images" between training data and testing data. This procedure is repeated 1000 times and the median or mean values are reported. It should be noted that the learning free methods are tested on the same data as learning-based methods. Besides, we should specifically point out that the training data on BID are used in Section 4 for the comparative study.

4 THE PROPOSED METHOD

The framework of the proposed method is shown in Figure 2, including four steps: image representation, feature extraction, feature aggregation, and quality prediction. In this section, we will conduct an in-depth comparative study to determine the best choice for each step.
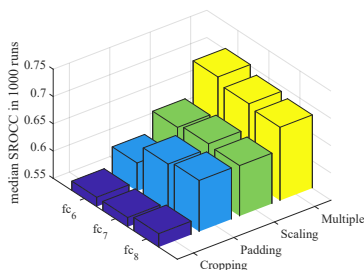
4.1 Image Representation

The pre-trained DCNN models (*e.g.*, AlexNet) require a fixed input size. To meet this requirement, images can be cropped, or resized to the fixed size. Since the resizing operation can introduce geometric deformation, which may change the image quality, it is not a good way. In the mean time, cropping
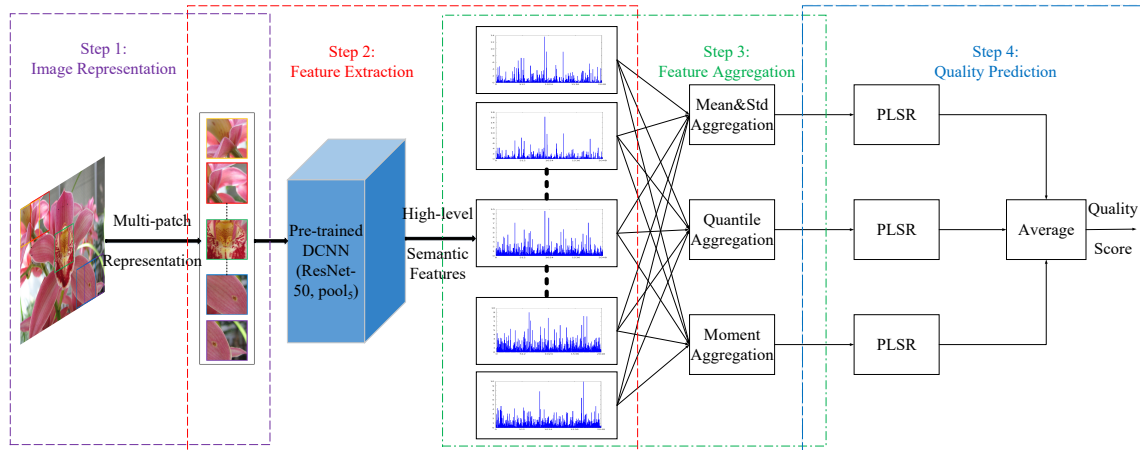
Figure 2: The overall framework of the proposed method, mainly includes four steps: image representation, feature extraction, feature aggregation, and quality prediction.

only the central patch is not enough to cover the information of a large image. Because of these two issues, we consider using multiple overlapping patches to represent the image, which not only covers information of the whole image but also avoids introducing geometric deformation.

We compare the impact of four different image representations, including the cropping, scaling, padding and multi-patch representation. Cropping representation uses the central patch to represent the image. Padding representation preserves the aspect ratio by resizing the larger dimension to the required length and then padding zeros to the smaller dimension. Scaling representation directly resizes the image without keeping the original aspect ratio. Multi-patch representation generates multiple overlapping patches that are uniformly sampled over the whole image with a sampling stride[1].

Figure 3: Comparison among different image representations. No matter which layer is used to extract features, the multi-patch representation achieves the best performance.

To perform the comparison, we need a baseline for feature extraction, feature aggregation and quality prediction. Before the comparative study on the following steps, we choose the classical pre-trained DCNN model AlexNet and extract features from the frequently-used fully connected layers (*i.e.*, fc_6, fc_7 and fc_8). For feature aggregation, we choose the mean feature vector for simplicity. PLSR is used for quality prediction. The comparative study is conducted on the training data of BID, where 20% of the training data are used as validation data and the performance on the validation set is used for comparison (the same below). It can be seen from Figure 3 that (1) cropping representation obtains the worst performance, (2) since resizing operation keeps most of the image information, padding and scaling representation achieve better performance than cropping representation, (3) the use of multi-patch representation significantly outperforms the other three. So we decide to use the multi-patch representation in our framework.

4.2 Feature Extraction

Given an image $\mathbf{I}$, we represent it with a set of multiple overlapping patches $\{\mathbf{p}_1, \cdots, \mathbf{p}_n\}$, and then feed these patches into an off-the-shelf DCNN model to extract features. For each patch $\mathbf{p}_j$, the extracted feature is denoted by

$$\mathbf{d}_j = DCNN(\mathbf{p}_j, L; \theta), \ j = 1, \cdots, n. \tag{2}$$

where L indicates which layer (*e.g.*, fc_8 layer in AlexNet) to extract features and θ is the trained network parameter.

The role of high-level semantics: Pre-trained DCNN models for image classification or scene recognition have encoded semantics in high-level features. Here, we conduct a comparative study to investigate the role of high-level semantics in NR-IQA of realistic blur images. We take the AlexNet as the pre-trained model, and extract features of multiple patches from its different layers[2] ($conv_1$ to $conv_5$

---

[1]There is no significant performance variation among different sampling strides when it is subjected to cover the whole information, so the sampling stride is simply fixed to be half of the patch size.

[2]Since the response of the convolutional layer is a set of feature maps, we derive features by global average pooling.

and fc_6 to fc_8). PLSR maps the mean feature vector to the quality score. From the plot in Figure 4, we have the following observations. First, high-level features are better than the low-level features, which indicates that high-level semantic features play an important role in NR-IQA of realistic blur images. However, the feature extracted from the top layer (fc_8) is slightly worse than the second and third top layers (fc_7, fc_6). This is because the top layer is directly linked to the classifier and the extracted feature is task-specific, which may contain only the classification information. The third top layer (fc_6), close to the last convolutional layer, achieves the best performance in terms of SROCC. Therefore, in our framework, we consider the second or third top layer close to the last convolutional layer to extract features.
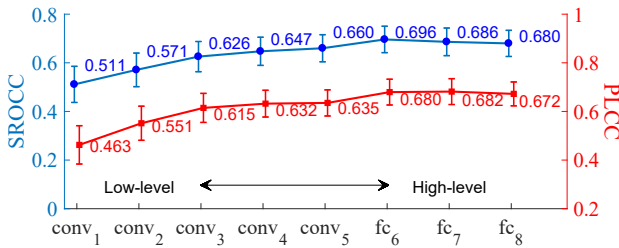
Figure 4: [Best viewed in color.] Mean and standard deviation of SROCC and PLCC. x-axis indicates from which layer (in AlexNet) we extract features. The curve indicates the mean values and the error bars indicate the standard deviations.

Impact of different pre-trained DCNN models: We also compare different pre-trained DCNN models in the proposed framework, including AlexNet [15], GoogleNet [34] and ResNet-50 [11], where the features are extracted from the fc_6, $pool_5/7 \times 7\_s1$ and $pool_5$ layer, respectively. The quality prediction step is still based on PLSR. Figure 5 shows the performance values, from which we can observe that ResNet-50 achieves the best performance. It is shown that the residual image contains important information in capturing quality relevant features [42]. Besides, image blur can be more easily captured in residual images. So the significant gain in ResNet-50 may due to the residual learning, and we choose ResNet-50 as the feature extractor.

4.3 Feature Aggregation

With the extracted features, we need to aggregate them into a single one. One straightforward way is to concatenate all these n features into a long feature vector, *i.e.*,

$$\mathbf{f}_{concat} = \mathbf{d}_1 \oplus \cdots \mathbf{d}_j \cdots \oplus \mathbf{d}_n \qquad (3)$$

where $\oplus$ is the concatenation operator.

However, it will result in a very high dimension of the feature space. Besides, the dimension of the concatenated feature vector will depend on the number of patches, which is not the same among the images with different resolutions.
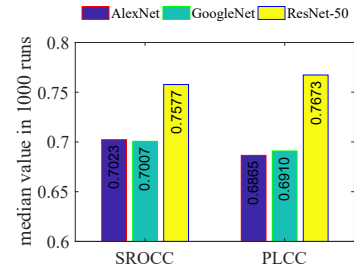
Figure 5: Comparison among different pre-trained DCNN models.

To avoid this, we can take the mean value of all features in each dimension, that is,

$$\mathbf{f}_{mean} = (m_1, \cdots, m_i, \cdots, m_l)^T, \qquad (4)$$

$$m_i = \frac{\sum_{j=1}^{n} d_{ji}}{n}, i = 1, \cdots, l.$$

where d_{ji} is the i-th element of $\mathbf{d}_j$ and l is the dimension of $\mathbf{d}_j$.

The mean aggregation structure loses important information (*e.g.*, the standard deviation in each dimension) of the feature set. So we propose three different statistical structures for feature aggregation, namely, mean&std aggregation, quantile aggregation and moment aggregation.

Mean&std aggregation: The standard deviation in each dimension is further considered, and the first aggregated feature $\mathbf{f}_1$ is obtained by:

$$\mathbf{f}_1 = \mathbf{f}_{mean} \oplus \mathbf{f}_{std} \qquad (5)$$

$$\mathbf{f}_{std} = \left(\sqrt{\frac{\sum_{j=1}^{n}(d_{j1}-m_1)^2}{n-1}}, \cdots, \sqrt{\frac{\sum_{j=1}^{n}(d_{jl}-m_l)^2}{n-1}} \right)^T$$

where m_i is the i-th element of $\mathbf{f}_{mean}, i = 1, \cdots, l$.

Quantile aggregation: Quantiles are important order statistics. We consider the widely used quartiles. The min, the median and the max are the zeroth, second, and fourth quartile, respectively. We denote the zeroth to fourth quartile of $(d_{1i}, \cdots, d_{ni})$ as $d_i^{(0)}, d_i^{(1)}, d_i^{(2)}, d_i^{(3)}, d_i^{(4)}, i = 1, \cdots, l$, respectively. So the second aggregated feature $\mathbf{f}_2$ based on quartiles can be defined as:

$$\mathbf{f}_2 = \mathbf{q}_0 \oplus \mathbf{q}_1 \oplus \mathbf{q}_2 \oplus \mathbf{q}_3 \oplus \mathbf{q}_4 \qquad (6)$$

$$\mathbf{q}_t = \left(d_1^{(t)}, \cdots, d_l^{(t)} \right)^T, t = 0, 1, 2, 3, 4.$$

Moment aggregation: Moments also play an important role in describing the statistics of a distribution. Mean is actually the origin moment of first-order. In order to balance between the need of more information and the dimension reduction of the feature space, we further consider the k-th root of the central moment of order k $(k = 2, 3, 4)$[3], and

---

[3]Note that the first central moment is zero, and here the second central moment is the variance computed using a divisor of n rather than $n - 1$.
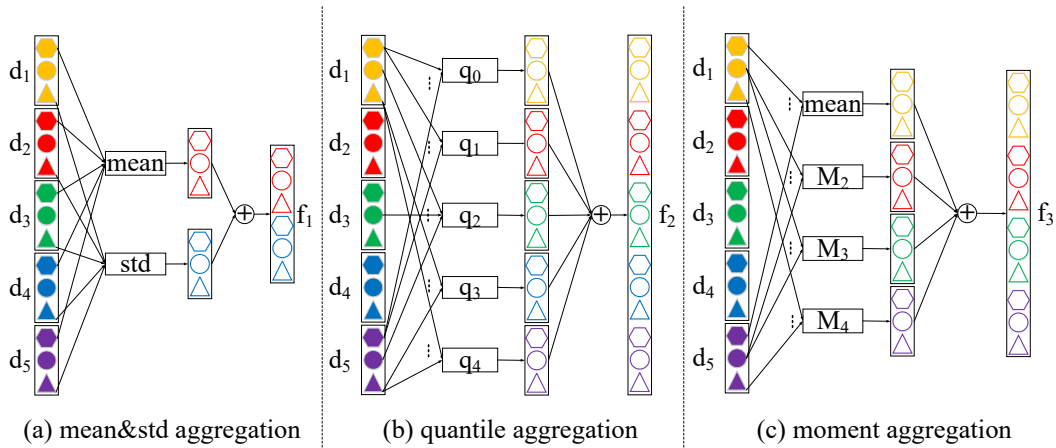
Figure 6: [Best viewed in color.] An example of the three statistical structures for feature aggregation. The input is $n = 5$ features $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5\}$, where the feature dimension is $l = 3$. $(\mathbf{q}_0, \mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4)$ indicate the five quartiles, and $\mathbf{M}_k$ represents the k-th root of the central moment of order k ($k = 2, 3, 4$). Not all the connections are shown between input and statistical functions for clarity.

obtain the third aggregated feature $\mathbf{f}_3$:

$$\mathbf{f}_3 = \mathbf{f}_{mean} \oplus \mathbf{M}_2 \oplus \mathbf{M}_3 \oplus \mathbf{M}_4 \tag{7}$$

$$\mathbf{M}_k = \left(\sqrt[k]{\frac{\sum_{j=1}^{n}(d_{j1} - m_1)^k}{n}}, \cdots, \sqrt[k]{\frac{\sum_{j=1}^{n}(d_{jl} - m_l)^k}{n}} \right)^T$$

where $k = 2, 3, 4$. m_i is the i-th element of $\mathbf{f}_{mean}, i = 1, \cdots, l$.

The aforementioned three statistical structures for feature aggregation result in a $2l$, $5l$ and $4l$-dimensional feature vector, respectively. An example of these aggregation structures when $n = 5, l = 3$ is shown in Figure 6.

4.3.1 Contribution of Different Statistical Aggregation Structures. We compare the mean aggregation (baseline) with the three proposed statistical aggregation structures. The ResNet-50 is used as the feature extractor of multiple patches and PLSR is used as the regression model. Table 1 summarizes the median values of SROCC, PLCC and RMSE. The best result comes from the ensemble of the three statistical structures and has been marked in boldface. We can see that the three proposed statistical structures have significant gain over the baseline, from which we verify the effectiveness of the proposed aggregation structures on capturing the information of the feature set.

4.4 Quality Prediction

With the help of statistical structures for feature aggregation, we reduce the dimension of feature space ($nl \rightarrow 2l, 4l, 5l$) and make the dimension independent of the number of patches. However, in the pre-trained DCNN, l is also a large number ($l = 2048$ in ResNet-50's $pool_5$ layer). Since the dimension of the feature space is much larger than the number of our training samples, we consider the linear regression model. Specifically, partial least square regression (PLSR) [28] is

Table 1: Comparison among different aggregation structures. "average-quality" means averaging scores of the three proposed structures.

Aggregated Feature	SROCC	PLCC	RMSE
mean ($\mathbf{f}_{mean}$)	0.7577	0.7673	0.8283
mean&std ($\mathbf{f}_1$)	0.8022	0.8174	0.7333
quantile ($\mathbf{f}_2$)	0.8109	0.8254	0.7135
moment ($\mathbf{f}_3$)	0.8100	0.8254	0.7171
$\mathbf{f}_1 \oplus \mathbf{f}_2$	0.8123	0.8269	0.7116
$\mathbf{f}_3 \oplus \mathbf{f}_2$	0.8127	0.8270	0.7121
average-quality ($\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$)	**0.8154**	**0.8305**	**0.7055**

adopted in our work because of its low-complexity and remarkable capability to handle high-dimensional data. PLSR reduces the input high-dimensional features to several uncorrelated latent components and then performs least squares regression on these components. There is only one parameter p (the number of components) in PLSR, which can be determined by cross validation.

After the above investigations, we obtain our best proposal, dubbed as Semantic Feature Aggregation metric using PLSR (SFA-PLSR). It uses multiple overlapping patches to represent images, and extracts features from the $pool_5$ layer of the pre-trained ResNet-50 model, as well as averages the scores of the mean&std aggregation, quantile aggregation and moment aggregation.

5 EXPERIMENTS

In the following parts, we compare the performance of the proposed SFA-PLSR method with the state-of-the-art NR-IQA methods in both intra-database and inter-database scenarios. As for the software platform to implement our proposed

Table 2: Performance comparison on four databases. In each column, the best performance value is marked in boldface and the second best performance value is underlined. The last column indicates the weighted-average of SROCC over all four databases, where the weights are proportional to the database-sizes.

Category	Method	BID [4] SROCC	BID [4] PLCC	BID [4] RMSE	CLIVE [6] SROCC	CLIVE [6] PLCC	CLIVE [6] RMSE	TID2008 [26] SROCC	LIVE [31] SROCC	Overall SROCC
NR-IQA of blur images	MDWE [21]	0.3067	0.3538	1.1639	0.4313	0.4988	17.5025	0.8556	0.9188	0.4514
	CPBD [24]	0.0202	0.2181	1.2166	0.3027	0.4026	18.4602	0.8723	0.9390	0.2945
	FISH [38]	0.4736	0.4853	1.0894	0.4865	0.5380	17.0310	0.8737	0.9008	0.5323
	S3 [37]	0.4109	0.4471	1.1177	0.4034	0.4864	17.6224	0.8650	0.9515	0.4686
	LPC [10]	0.3150	0.4053	1.1408	0.1483	0.3490	18.9205	0.8805	0.9469	0.2922
	MLV [1]	0.3169	0.3750	1.1561	0.3412	0.4076	18.4350	0.8977	0.9431	0.4058
	ARISM [8]	0.0151	0.1929	1.2245	0.2427	0.3554	18.8947	0.8851	0.9585	0.2601
	BIBLE [16]	0.3609	0.3923	1.1469	0.4260	0.5178	17.3007	0.9114	**0.9638**	0.4703
	SPARISH [17]	0.3071	0.3555	1.1659	0.4015	0.4843	17.6702	0.9126	**0.9638**	0.4403
	RISE [18]	0.5839	0.6017	0.9936	-	-	-	<u>0.9218</u>	0.9493	0.6833
Proposed	**SFA-PLSR**	**0.8269**	**0.8401**	**0.6854**	**0.8130**	**0.8313**	**11.3905**	0.9098	0.9523	**0.8321**
General purpose NR-IQA	BRISQUE [22]	0.5795	0.5754	1.0624	0.5950	0.6195	16.0273	0.8737	0.8892	0.6258
	Kang's CNN [13]	0.4818	0.4977	1.1030	0.4964	0.5218	17.8567	0.9000	0.9429	0.5448
	FRIQUEE [7]	<u>0.7359</u>	<u>0.7477</u>	<u>0.8433</u>	0.6916	0.7069	14.4244	**0.9261**	0.9515	<u>0.7353</u>
	NRSL [19]	0.638	0.663	0.931	0.631	0.654	15.317	-	<u>0.959</u>	0.658
	S-HOSA [32]	0.6869	0.6913	0.9112	<u>0.7051</u>	<u>0.7241</u>	<u>14.0237</u>	0.8729	0.9469	0.7258

The results of RISE and NRSL are from their **original papers**. The code of Kang's CNN and S-HOSA are written by ourselves following the detail of their papers, and the codes of other compared methods are from **original authors**.

method, we use the Caffe [12] framework to extract the features from the pre-trained DCNN model. PLSR is performed by the MATLAB function *plsregress*, and its parameter p is globally set to 10 based on the 5-fold cross-validation using the training data of a single run (on BID), where p is selected from the set $\{5, 10, 15, 20, 25, 30\}$ for simplicity.

5.1 Performance Comparison

In this part, we compare the performance of SFA-PLSR with ten existing (from classical to the most up to date) NR-IQA methods of blur images, which are MDWE [21], CPBD [24],FISH [38], S3 [37], LPC [10], MLV [1], ARISM [8], BIBLE [16], SPARISH [17] and RISE [18]. Five remarkable general-purpose NR-IQA methods, including BRISQUE [22], Kang's CNN [13], FRIQUEE [7], NRSL [19], and S-HOSA [32], are also taken for comparison.

Table 2 reports the median SROCC, PLCC and RMSE in 1000 runs on the four databases. We also report the weighted-average SROCC over all four databases as the overall performance, where the weights are proportional to the database-sizes (see the last column of Table 2). Among the ten NR-IQA methods of blur images, the first nine methods fail on the two realistic databases (SROCC< 0.5 on BID and CLIVE) due to their neglect of global semantic information, and RISE achieves the best performance on BID. The proposed method SFA-PLSR significantly outperforms others on BID and CLIVE in both prediction accuracy (PLCC, RMSE) and monotonicity (SROCC). As for the general purpose NR-IQA methods, FRIQUEE and S-HOSA achieve better performance on the realistic databases than the others. Kang's CNN [13] does not perform well because it assumes that patch quality equals to image quality, which is not true

for these two realistic image datasets. On TID2008 and LIVE, there are less than 30 images with different contents, which is much smaller than BID (586) and CLIVE (1162), so the role of semantic information is weakened and the impact of low-level features is enhanced. Nevertheless, our method SFA-PLSR still achieves comparable performance on the two synthetic databases. In general, our method also achieves the best overall performance.

5.2 Cross Dataset Evaluation

In this subsection, we test the generalization capability of learning-based methods through cross dataset evaluation. Since learning-based methods assume testing images and training images have similar distribution, we conduct cross dataset evaluation on realistic databases (BID and CLIVE) and synthetic databases (TID2008 and LIVE), respectively. It should be noted that CLIVE contains 383 images resized from BID images, we exclude the 383 images from CLIVE in cross dataset experiments.

We compare our method with RISE (the compared NR-IQA method of blur images with the best overall performance), FRIQUEE and S-HOSA (the best two general purpose NR-IQA methods). The SROCC values are provided in Table 3. It can be seen that our method performs better than RISE, FRIQUEE and S-HOSA, which has demonstrated the database independency and robustness of the proposed SFA-PLSR method.

5.3 Impact of Training Ratio

In order to have an intuitive understanding of how the training ratio affects the performance of our methods, we also

Table 3: SROCC values in cross dataset evaluation.

train → test	RISE	FRIQUEE	S-HOSA	Ours
BID → CLIVE	-	0.3571	0.4767	**0.5729**
CLIVE → BID	-	0.3886	0.3433	**0.6838**
TID2008 → LIVE	0.8638	0.8690	0.8950	**0.9166**
LIVE → TID2008	0.9138	0.8727	0.8612	**0.9243**

The results of cross dataset evaluation on the two realistic blur datasets were not reported in the original paper of RISE.

conduct an experiment to test SFA-PLSR with different training ratios (from 10% to 90% with an increment step of 10%). It is clearly shown from Figure 7 that with the increase of training ratio, the performance values boost quickly when the training ratio is smaller than 30%. We can see that even if only 40% of images are used for training, the PLCC values are still close to 0.8. This is helpful in real-world applications, where relatively small amount of images are labeled.
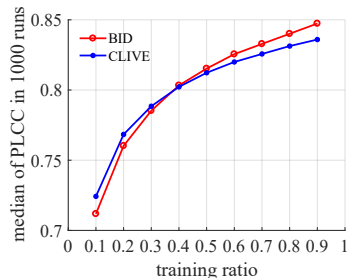

Figure 7: The PLCC of SFA-PLSR with different training ratios.

5.4 2σ-Confidence Band and Failure Case

In this part, we further consider the prediction consistency of the proposed method and FRIQUEE (the method among the compared methods with the best overall performance). The green regions shown in Figure 8(a), (b) are the 2σ-confidence bands on BID. The scatter points outside the band are regarded as outliers. It can be seen that FRIQUEE has more outliers than our method SFA-PLSR. The median values of outlier's ratio (OR) in 1000 runs are 5.98%, 11.11% for SFA-PLSR, FRIQUEE, respectively, which indicates that our method is more consistent with human perception. The outliers correspond the failure cases, and the worst case of our method is shown in Figure 8(c). The picture suffered from so complex distortions. To overcome this type of failure cases, more clues should be considered, such as saturation and ghosting.

6 CONCLUSION

In this paper, we propose a novel NR-IQA method for realistic blur images, which is based on statistically aggregating the high-level semantic features extracted from pre-trained deep
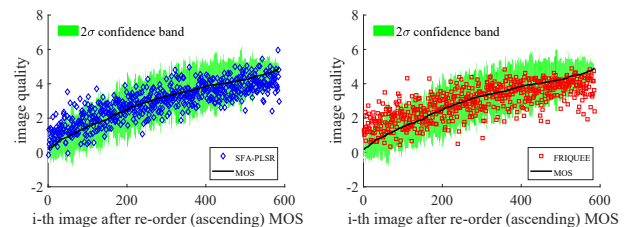
(a) SFA-PLSR   **(b) FRIQUEE**

(c) Failure case

Figure 8: (a) SFA-PLSR scores and 2σ-confidence band on BID, (b) FRIQUEE scores and 2σ-confidence band on BID, and (c) a failure case.

convolutional neural networks. The top performance and strong generalization capability of our method are validated by comparing with several state-of-the-art methods on two realistic image databases (BID, CLIVE) and two synthetic image databases (TID2008, LIVE). Experiments also show that high-level semantics indeed play a more critical role than low-level features in NR-IQA of realistic blur images. In the future study, we will consider our methods in a coarse to fine multi-scale framework, since object scale also plays a role in human blur perception.

REFERENCES
[1] Khosro Bahrami and Alex C. Kot. 2014. A fast approach for no-reference image sharpness assessment based on maximum local variation. *IEEE Signal Processing Letters* 21, 6 (2014), 751–755.
[2] Khosro Bahrami and Alex C. Kot. 2016. Efficient image sharpness assessment based on content aware total variation. *IEEE Transactions on Multimedia* 18, 8 (2016), 1568–1578.
[3] Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek. 2016. A deep neural network for image quality assessment. In *IEEE International Conference on Image Processing*. 3773–3777.
[4] Alexandre Ciancio, André Luiz N. Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador.

2011. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing* 20, 1 (2011), 64–75.

[5] Rony Ferzli and Lina J. Karam. 2009. A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). *IEEE Transactions on Image Processing* 18, 4 (2009), 717–728.

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

[7] Deepti Ghadiyaram and Alan C. Bovik. 2017. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision* 17, 32 (2017).

[8] Ke Gu, Guangtao Zhai, Weisi Lin, Xiaokang Yang, and Wenjun Zhang. 2015. No-reference image sharpness assessment in autoregressive parameter space. *IEEE Transactions on Image Processing* 24, 10 (2015), 3218–3231.

[9] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. 2015. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia* 17, 1 (2015), 50–63.

[10] Rania Hassen, Zhou Wang, and Magdy M. A. Salama. 2013. Image sharpness assessment based on local phase coherence. *IEEE Transactions on Image Processing* 22, 7 (2013), 2798–2810.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*. 675–678.

[13] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *CVPR*. 1733–1740.

[14] Jongyoo Kim and Sanghoon Lee. 2017. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing* 11, 1 (January 2017), 206–220.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.

[16] Leida Li, Weisi Lin, Xuesong Wang, Gaobo Yang, Khosro Bahrami, and Alex C Kot. 2016. No-reference image blur assessment based on discrete orthogonal moments. *IEEE Transactions on Cybernetics* 46, 1 (2016), 39–50.

[17] Leida Li, Dong Wu, Jinjian Wu, Haoliang Li, Weisi Lin, and Alex C Kot. 2016. Image sharpness assessment by sparse representation. *IEEE Transactions on Multimedia* 18, 6 (2016), 1085–1097.

[18] Leida Li, Wenhan Xia, Weisi Lin, Yuming Fang, and Shiqi Wang. 2017. No-reference and robust image sharpness evaluation based on multi-scale spatial and spectral features. *IEEE Transactions on Multimedia* (2017).

[19] Qiaohong Li, Weisi Lin, Jingtao Xu, and Yuming Fang. 2016. Blind image quality assessment using statistical structural and luminance features. *IEEE Transactions on Multimedia* 18, 12 (2016), 2457–2469.

[20] Qingbo Lu, Wengang Zhou, and Houqiang Li. 2016. A no-reference image sharpness metric based on structural information using sparse representation. *Information Sciences* 369 (2016), 334–346.

[21] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. 2002. A no-reference perceptual blur metric. In *IEEE International Conference on Image Processing*, Vol. 3. 57–60.

[22] Anish Mittal, Anush Krishna Moorthy, and Alan C. Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708.

[23] Anush Krishna Moorthy and Alan C. Bovik. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing* 20, 12 (2011), 3350–3364.

[24] Niranjan D. Narvekar and Lina J. Karam. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Transactions on Image Processing* 20, 9 (2011), 2678–2683.

[25] Taegeun Oh, Jincheol Park, Kalpana Seshadrinathan, Sanghoon Lee, and Alan C. Bovik. 2014. No-reference sharpness assessment of camera-shaken images by analysis of spectral structure. *IEEE Transactions on Image Processing* 23, 12 (2014), 5428–5439.

[26] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. 2009. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics* 10, 4 (2009), 30–45.

[27] Feng Qi, Debin Zhao, and Wen Gao. 2015. Reduced reference stereoscopic image quality assessment based on binocular perceptual information. *IEEE Transactions on Multimedia* 17, 12 (2015), 2338–2344.

[28] Roman Rosipal and Nicole Krämer. 2006. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*. Springer, 34–51.

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

[30] Qingbing Sang, Huixin Qi, Xiaojun Wu, Chaofeng Li, and Alan C. Bovik. 2014. No-reference image blur index based on singular value curve. *Journal of Visual Communication and Image Representation* 25, 7 (2014), 1625 – 1630.

[31] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing* 15, 11 (2006), 3440–3451.

[32] Ernestasia Siahaan, Alan Hanjalic, and Judith A. Redi. 2016. Augmenting blind image quality assessment using image semantics. In *IEEE International Symposium on Multimedia*. 307–312.

[33] Cuirong Sun, Houqiang Li, and Weiping Li. 2016. No-reference image quality assessment based on global and local content perception. In *Visual Communications and Image Processing*. 1–4.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*. 1–9.

[35] Huixuan Tang, Neel Joshi, and Ashish Kapoor. 2011. Learning a blind measure of perceptual image quality. In *CVPR*. 305–312.

[36] VQEG. 2000. Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment. (2000).

[37] Cuong T. Vu, Thien D. Phan, and Damon M. Chandler. 2012. S_3: A spectral and spatial measure of local perceived sharpness in natural images. *IEEE Transactions on Image Processing* 21, 3 (2012), 934–945.

[38] Phong V. Vu and Damon M. Chandler. 2012. A fast wavelet-based algorithm for global and local image sharpness estimation. *IEEE Signal Processing Letters* 19, 7 (2012), 423–426.

[39] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[40] Zhou Wang and Eero P Simoncelli. 2003. Local phase coherence and the perception of blur. In *NIPS*. 1435–1442.

[41] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. 2016. Blind Image Quality Assessment Based on High Order Statistics Aggregation. *IEEE Transactions on Image Processing* 25, 9 (2016), 4444–4457.

[42] Jia Yan, Weixia Zhang, and Tianpeng Feng. 2016. Blind image quality assessment based on natural redundancy statistics. In *ACCV*. 3–18.

[43] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *CVPR*. 1098–1105.

[44] Junyong You, Touradj Ebrahimi, and Andrew Perkis. 2011. Modeling motion visual perception for video quality assessment. In *ACM MM*. 1293–1296.

[45] Shaode Yu, Shibin Wu, Fan Jiang, Leida Li, Yaoqin Xie, and Lei Wang. 2017. A shallow convolutional neural network for blind image sharpness assessment. *PloS one* (2017).

[46] Guangtao Zhai, Xiaolin Wu, Xiaokang Yang, Weisi Lin, and Wenjun Zhang. 2012. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing* 21, 1 (2012), 41–52.

[47] Jiangyang Zhang and C-C Jay Kuo. 2014. An objective quality of experience (QoE) assessment index for retargeted images. In *ACM MM*. 257–266.