Which Has Better Visual Quality: The Clear Blue Sky or a Blurry Animal?

Dingquan Li, *Member, IEEE*, Tingting Jiang, *Member, IEEE*, Weisi Lin, *Fellow, IEEE*, and Ming Jiang, *Senior Member, IEEE*

***Abstract*—Image content variation is a typical and challenging problem in no-reference image-quality assessment (NR-IQA). This work pays special attention to the impact of image content variation on NR-IQA methods. To better analyze this impact, we focus on blur-dominated distortions to exclude the impacts of distortion-type variations. We empirically show that current NR-IQA methods are inconsistent with human visual perception when predicting the relative quality of image pairs with different image contents. In view of deep semantic features of pretrained image classification neural networks always containing discriminative image content information, we put forward a new NR-IQA method based on semantic feature aggregation (SFA) to alleviate the impact of image content variation. Specifically, instead of resizing the image, we first crop multiple overlapping patches over the entire distorted image to avoid introducing geometric deformations. Then, according to an adaptive layer selection procedure, we extract deep semantic features by leveraging the power of a pretrained image classification model for its inherent content-aware property. After that, the local patch features are aggregated using several statistical structures. Finally, a linear regression model is trained for mapping the aggregated global features to image-quality scores. The proposed method, SFA, is compared with nine representative blur-specific NR-IQA methods, two general-purpose NR-IQA methods, and two extra full-reference IQA methods on Gaussian blur images (with and without Gaussian noise/JPEG compression) and realistic blur images from multiple databases, including LIVE, TID2008, TID2013, MLIVE1, MLIVE2, BID, and CLIVE. Experimental results show that SFA is superior to the state-of-the-art NR methods on all seven databases. It is also verified that deep semantic features play a crucial role in addressing image content variation, and this provides a new perspective for NR-IQA.**

***Index Terms*—Deep semantic features, image content variation, no-reference image quality assessment, realistic blur, statistical aggregation.**

I. Introduction

DIGITAL images can undergo various distortions during image acquisition, compression, transmission, display, *etc*. To monitor and improve the image quality, image quality assessment (IQA) has become a fundamental aspect of modern multimedia systems. Since humans are the end-users of most multimedia devices, the most accurate method to evaluate the image quality is achieved using subjective ratings. However, subjective evaluation is difficult to conduct in real-time applications for the handicaps of its inconvenience, high price, and inefficiency. This leads to the need for efficient and effective objective IQA methods that can automatically give image quality predictions. Objective IQA can be divided into full-reference IQA (FR-IQA) [1]–[4], reduced-reference IQA (RR-IQA) [5]–[7] and no-reference IQA (NR-IQA) [8]–[12]. Because reference information needed by FR/RR methods is often unavailable, FR- and RR-IQA methods are inapplicable in most practical applications. The absence of the reference information calls for NR-IQA methods, which are more applicable but also more difficult.

Humans can perceive image quality among various image contents; however, **image content variation** is a common and challenging problem in NR-IQA. For example, humans rate higher quality for the clear blur sky instead of the blurry animal, while most current objective IQA methods wrongly predict the relationship. In this work, we focus on blur-dominated distortions for analyzing the impact of image content variation on NR-IQA methods since most images captured by users in the real world suffer from out-of-focus blur or motion blur. Objective NR-IQA methods are expected to be consistent with subjective ratings. However, we find that current NR-IQA methods show many inconsistencies when predicting the relative quality of image pairs with different image contents. First, we show that objective NR-IQA methods have inconsistent cases whereby they predict quite different objective quality for image pairs with indiscriminate subjective quality ratings but different image contents. Based on the variation in objective scores, we define a criterion to quantitatively measure the impact of image content variation for quality-indiscriminate images. Sec-

Manuscript received March 26, 2018; revised July 21, 2018; accepted September 4, 2018. Date of publication October 11, 2018; date of current version April 23, 2019. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2015CB351803, in part by the National Natural Science Foundation of China under Grant 61390514, Grant 61527804, Grant 61572042, and Grant 61520106004, in part by the Tier 2 Fund of Ministry of Education, Singapore: MOE2016-T2-2-057 (S), and in part by Sino-German Center (GZ 1025). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hantao Liu. *(Corresponding author: Tingting Jiang.)*

D. Li and M. Jiang are with the Key Laboratory of Mathematics and Its Applications (LMAM), School of Mathematical Sciences, Beijing International Center for Mathematical Research, Cooperative Medianet Innovation Center, Peking University, Beijing 100871, China (e-mail: dingquanli@pku.edu.cn; ming-jiang@pku.edu.cn).

T. Jiang is with the National Engineering Laboratory for Video Technology, School of Electrical Engineering and Computer Science, Cooperative Medianet Innovation Center, Peking University, Beijing 100871, China (e-mail: ttjiang@pku.edu.cn).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2875354

ond, NR-IQA methods also have inconsistent cases whereby they give inverse relative quality predictions for image pairs with discriminable subjective quality ratings but different image contents. To explore the impact of image content variation on NR-IQA in this scenario, we construct a quality-discriminable image pair dataset, containing in total 22,792 image pairs with quite discriminable subjective qualities. According to the analysis on both quality-indiscriminate image pairs and quality-discriminable image pairs, we experimentally prove that current NR-IQA methods are indeed inconsistent with human visual perception due to image content variation.

Image-content-aware features can help to alleviate the impact of image content variation on NR-IQA models. Therefore, in this work, we address the above problem by resorting to deep semantic features extracted from an off-the-shelf deep convolutional neural network (DCNN) model, which is pre-trained for image classification tasks. However, it remains an open question of how to better adopt a pre-trained DCNN on NR-IQA. Herein, several key factors need to be taken into considerations. First, we have to choose a suitable format of image representation since a fixed size input is required by pre-trained DCNN models (e.g., ResNet-50 [13])]. To both cover the entire image and avoid introducing geometric deformation, we use the multi-patch representation for an image. Meanwhile, different from image classification, IQA focuses on human perception of image distortions. To extract powerful and effective features that are sensitive to content distortions, we should decide which pre-trained DCNN model to use and from which layer to extract features. We empirically choose the pre-trained ResNet-50 to extract the semantic features. In addition, to obtain fixed-dimension features, we use an adaptive layer selection procedure to output suitable feature maps and apply global average pooling to the maps. Last, since we extract multiple local patch features from an image, global information will be somewhat weakened or even disregarded, which is contradictory with the IQA task. To address this issue, effective aggregation mechanisms are desired. A simple strategy is to use the mean feature vector. However, this loses other important characteristics of the feature set (e.g., the standard deviation). Therefore, we adopt three statistical structures for feature aggregation: mean&std aggregation, quantile aggregation, and moment aggregation. Due to the high dimensionality of the aggregated global features, we ultimately train a linear regression model for mapping the aggregated global features to image quality scores.

We conduct experiments on not only simulated Gaussian blur images (from LIVE [14], TID2008 [15] and TID2013 [16]) but also Gaussian blur images with Gaussian noise or JPEG compression (from MLIVE1 and MLIVE2 [17]) and realistic blur images (from BID [18] and CLIVE [19]). The proposed Semantic-Feature-Aggregation-based method, SFA, is compared with nine blur-specific NR-IQA methods, two general-purpose NR-IQA methods and two extra FR-IQA methods. Our experiments show that SFA is superior to existing NR-IQA methods on two realistic image databases and five simulated image databases.

The main contributions of this work are as follows:

1) An analysis of the impact of image content variation on NR-IQA methods verifies that deep semantic features can alleviate the impact of image content variation. The deep semantic features indeed play a key role in predicting image quality among various image contents. This introduces a new viewpoint for designing NR-IQA methods in terms of the semantic aspect.
2) A novel NR-IQA framework is proposed based on semantic feature aggregation, where a pre-trained DCNN model with an adaptive layer selection procedure is used as the feature extractor, and some statistical characteristics are used for feature aggregation.
3) Experiments on seven databases (containing simulated and realistic blur images) verified the superiority and generalization capability of the proposed method.

This paper extends our conference paper [20] with the following distinctions made. 1) A major extension is that we design new experiments, construct new datasets and define new quantitative criteria for analyzing the impact of image content variation. 2) A new adaptive layer selection procedure for feature selection that can significantly improve our method's performance on simulated images is adopted. 3) More experiments and analysis are performed, for which we provide a discussion on realistic blur, report the results of statistical significance tests, present test results on the Waterloo exploration database [21], and summarize the computational efficiency.

II. Related Works

Image blur is a common distortion, and often occurs in the following situations: (1) defocus, (2) relative motion (i.e., object motion and camera shake), (3) imperfect imaging systems (such as lens aberration), (4) atmospheric turbulences, and (5) image post-processing techniques (e.g., denoising and compression) [22]–[24]. In the blur-specific NR-IQA literature, researchers have mainly considered Gaussian blur images. In this aspect, image sharpness and image quality can be discussed as synonyms. Simultaneously, image sharpness and image blurriness can be used as antonyms. Therefore, in this section, we review not only blur-specific NR-IQA methods but also methods for sharpness/blurriness estimation. However, we should note that sharpness is not always the antonym of blurriness. For example, a clear blue sky (dominated by flat regions) can neither be considered sharp nor blurred. The same example also shows that image sharpness does not equal image quality, i.e., a clear blue sky is not sharp but has good quality.

Learning-free methods: Some methods use the blur properties in the spatial domain, e.g., the edge spread [22], [23], [25] and the smoothing effect [26]–[28], while other methods further use the blur properties in the transform domains, e.g., reductions in high-frequency components [29]–[31] and the loss of phase coherence [24], [32], [33]. Since blur makes edges spread, Marziliano *et al.* [25] detect vertical edges and consider the average edge width as a quality measure. The above method is further enhanced using the concept of just noticeable blur (JNB) [22]. Noticing that blur is unlikely to be detected at an edge whose width is negligible, Narvekar and Karam [23] measure the image quality by the probability of edges whose widths are smaller than the JNB width. Based on the smoothing effect of image blur, Gu *et al.* [26] estimate the image quality using the energy

differences and contrast differences in the autoregressive parameter space, while Bahrami and Kot [27], [28] consider modeling the distribution of the total variation (TV) or the maximum local variation (MLV). In the frequency domain, blur leads to the attenuation of the high-frequency energy, which can be a cue for quality estimation [29]. Vu and Chandler [30] measure image sharpness by considering the steepness of the local magnitude spectrum and the local total variation. In addition, Li *et al.* [31] evaluate the image quality based on the energy of the non-DC Tchebichef moments [34] of gradient blocks. It is shown that step edges imply strong local phase coherence (LPC) structures across scales and space, which can be destroyed by blur [32]. This suggests that the strength of the LPC near edges and lines can be used as a sharpness measure [24]. In addition, Leclaire and Moisan [33] define the global phase coherence (GPC) of an image to relate the image quality to the loss of image regularity when its phase information is disturbed.

Learning-based methods: These methods mainly utilize two steps: feature extraction and quality prediction. The most important aspect is to extract features that are quality relevant. Handcrafted features are generally extracted from nature scene statistics (NSS) models. They can also be obtained by some low-level image features (e.g., contrast and brightness). It is assumed that distortions alter the statistical properties of natural scenes; therefore, Wang *et al.* [35] estimate the image quality by applying extreme learning machine [36] to the statistics of a gradient distribution model. Ciancio *et al.* [18] fuse features of traditional methods and low-level features using a neural network for quality assessment. Li *et al.* [37] utilize a support vector regression (SVR) model to map the features extracted from gradient similarity, singular value similarity and DCT domain entropies. Instead of handcrafted features, powerful quality relevant features can be learnt directly by machine learning methods. Having observed that over-complete dictionaries learned from natural images can capture edge patterns, Li *et al.* [38] learn an over-complete dictionary and use it to construct a sparse coding model for the image blocks; then, they estimate image quality based on the block energy normalized by the block variance. Lu *et al.* [39] learn structure-related features from the sparse dictionary and map the learned features to quality scores using SVR. Recently, deep learning techniques have been applied for general-purpose IQA [40]–[45]. Yu *et al.* [46] attempt to apply deep learning architectures to blur image quality assessment. An image pre-processed using the local contrast normalization is passed through a convolutional layer, a down-sampling layer and a fully connected layer to extract image features; then, the features are mapped to an image quality score by regression.

Other blur-relevant research: In addition to the quality assessment of blur images, other blur-relevant research, including local blur detection [47]–[49], deblurring [50]–[53], and blur manipulation [54], has been performed. Previous studies often considered extracting features from the edge models, power spectral slopes, and image gradient statistics [47], [50], [54]. However, with the explosive development of deep learning, increasingly more blur-relevant research is being addressed by convolutional neural networks, recurrent neural networks, and generative adversarial networks [48], [49], [51]–[53]. The quality assessment of blur images can guide such related research. For example, the quality-aware loss can be integrated into deblurring algorithms. But at the same time, probing into these blur-relevant research can also provide insight for quality assessment of blur images. For example, the overall blur degree can be obtained by pooling the local blur map [48].

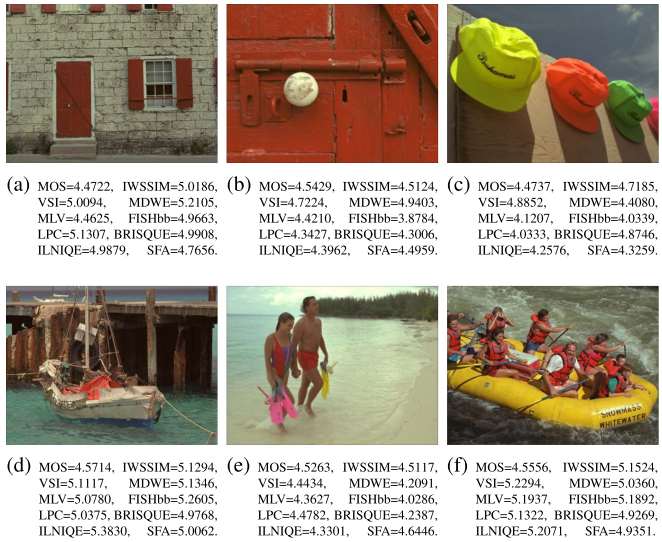


Fig. 1. A quality-indiscriminate image pair set with six images from TID2013 [16]. Their quality scores given by IQA methods are shown in sub-captions.

III. Impact of Image Content Variation

Humans can perceive image quality differences among various image contents. Therefore, it is interesting to explore the impact of image content variation on NR-IQA methods. A reasonable NR-IQA method should be consistent with subjective ratings; at the least, it should give consistent relative quality predictions of image pairs with different image contents.

A. Quality-Indiscriminate Image Pairs

A quality-indiscriminate image pair indicates an image pair with indiscriminate subjective quality but different image contents. We extend quality-indiscriminate image pairs to quality-indiscriminate image sets since a quality-indiscriminate image set containing N images can generate $\frac{N(N-1)}{2}$ quality-indiscriminate image pairs. Fig. 1 shows a quality-indiscriminate image set with six images, where the sub-captions show the mean opinion score (MOS), predicted scores given by two FR-IQA methods (IWSSIM [2], VSI [3]), four traditional blur-specific methods (MDWE [25] based on the edge spread, MLV [27] based on the smoothing effect, FISHbb [29] based on a reduction in high-frequency energy, LPC [24] based on a loss of phase coherence), and two general-purpose NR-IQA methods (BRISQUE [8], ILNIQE [9]). Objective IQA methods should give similar objective quality scores on a quality-indiscriminate image set. However, this is not the case due to the impact of image content variation. IWSSIM is a structure-based method, and VSI is a saliency-based method; they all overestimate the image quality of Fig. 1(a), (d), and

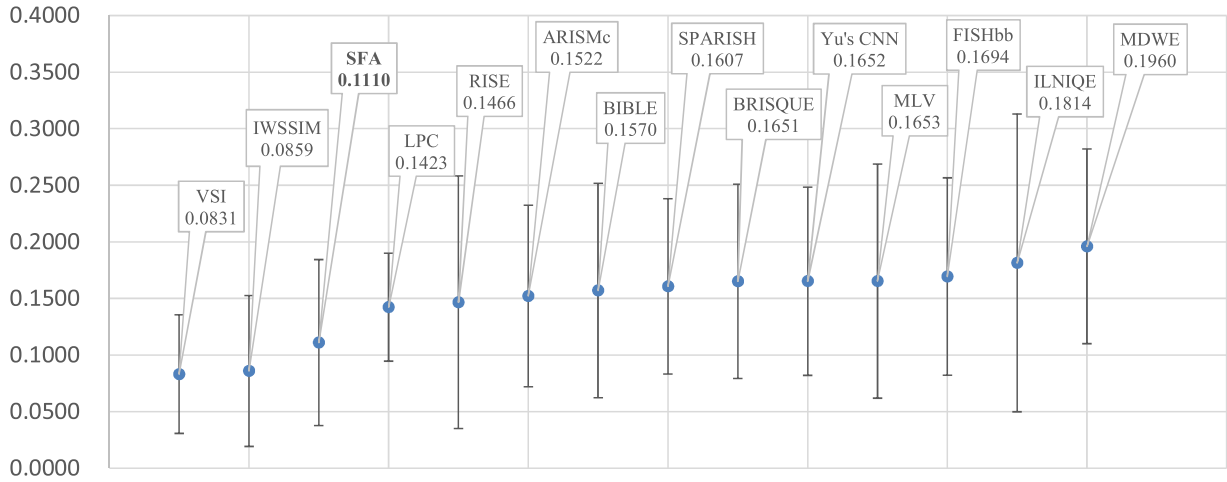
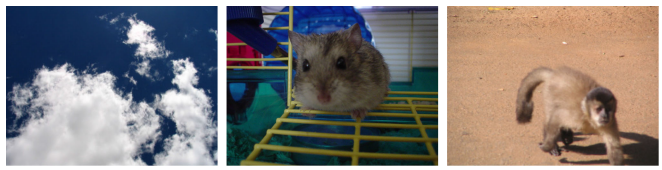


Fig. 2. The mean and standard deviation values of NSD over all datasets $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{379}\}$ constructed from TID2013. The mean NSD increases from left to right, which means that the performance drops from left to right. A smaller standard deviation of NSD indicates a more reliable reported mean NSD. Note that the learning-based methods are trained on LIVE.

(f), which contain abundant textures. MDWE overestimates the image quality of Fig. 1(a), which contains sharp edges, and it underestimates the image quality of Fig. 1(e), which contains smooth edges. The maximum local variations in Fig. 1(d) and (f) are large, which cause the over-estimation of MLV on these two images. Fig. 1(b), 1(c), and (e) contain many flat regions, mainly consisting of low-frequency components. As a result, the image quality is underestimated by FISHbb in the three images. Conversely, FISHbb overestimates the quality of Fig. 1(a) and (f) since there is a substantial amount of high-frequency energy in the image structures. Strong edges correspond to high LPC strength, while smooth edges result in low LPC strength, which explains the results that LPC overestimates the quality of Fig. 1(a) and underestimates the quality of Fig. 1(c). As the above methods, BRISQUE and ILNIQE also overestimate the quality of Fig. 1(d) and (f) due to the impact of image content variation.

To statistically analyze the impact of image content variation on IQA methods for quality-indiscriminate image sets, we first introduce some assumptions. Given a set $\mathcal{S}$ containing images with indiscriminate subjective quality, we assume the following: (i) The impact of image content variation is larger when the standard deviation (std) of the objective scores is larger. (ii) The impact of image content variation on a method is noticeable when the standard deviation of the objective scores is more than double the standard deviation of the subjective scores. (iii) To ensure the comparability among datasets with different subjective score ranges, the standard deviation should be normalized. In the extreme case that the objective scores of an infinite number of images from a quality-indiscriminate image set are uniformly distributed in the subjective score range $[0, R]$, the impact of image content variation is expected to be very strong, and the standard deviation of objective scores in this case $(R/2\sqrt{3})$ can be used as a normalization factor. Based on the above assumptions, we quantitatively measure the impact of image content variation by the normalized standard deviation (NSD):

$$\mathrm{NSD} = \frac{[\mathrm{std}_o(\mathcal{S}) - 2\mathrm{std}_s(\mathcal{S})]_+}{R/2\sqrt{3}} \tag{1}$$



(a) The clear blue sky (b) A blurry mouse (c) A blurry monkey

Fig. 3. The three images are from BID [18]. The subjective quality of (a) is better than the subjective quality of (b)/(c). The four traditional methods (MDWE [25], MLV [27], FISHbb [29], and LPC [24]) predict that (a) is worse than (b)/(c), which is inconsistent with human visual perception.

where $\mathrm{std}_o(\mathcal{S})$ and $\mathrm{std}_s(\mathcal{S})$ are the standard deviation of the mapped objective scores and the subjective scores on the dataset S. $[x]_+$ represents the positive part of x. The impact is smaller when NSD is closer to zero.

We construct the quality-indiscriminate image set $\mathcal{S}$ from TID2013 [16]. The set $\mathcal{S}$ is expected to meet the following two requirements: (1) Since most of the subjective ratings for an image are in the range of [MOS-std, MOS+std], it suggests that an image pair is supposed to be reliably indiscriminate in terms of the subjective quality if their absolute MOS difference is smaller than the std of the MOS. Specifically, to guarantee the reliability of the quality-indiscriminate image sets, the extreme MOS difference in the set $\mathcal{S}$ is expected to be smaller than the average std of MOS ($\sigma = 0.1200$ in TID2013). (2) To ensure the stability, $|\mathcal{S}|$, the size of the set $\mathcal{S}$, is expected to be at least $N_0 = 4$. In the range of subjective scores, we randomly select 1000 windows with a length equal to σ. Then, we judge whether the images whose subjective scores in the selected window can meet requirement (2). Finally, we obtain 379 quality-indiscriminate image sets $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{379}\}$ that satisfy the two requirements. In Fig. 2, we report the mean and standard deviation values of NSD over all datasets. The image content variation has the smallest impact on the two FR-IQA methods, VSI and IWSSIM, in terms of the defined criterion, which is consistent with the fact that FR-IQA methods use both reference and distorted images. The NSD values of other NR-IQA methods are almost two-times larger than the NSD values of FR-IQA methods. This

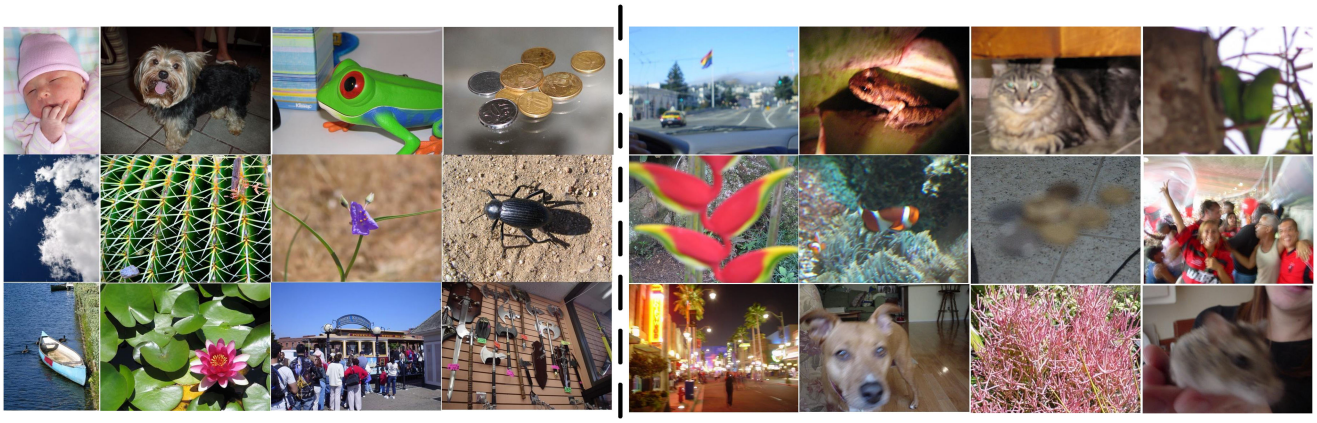


Fig. 4. The thumbnail sample images of the Content-Variation-Impact-Quality dataset (CVIQ). Most methods have many failure cases on predicting the relative quality, i.e., predicting that the quality of images on the left side worse than images on the right side. This is due to the impact of image content variation.

means that existing NR-IQA methods are strongly affected by image content variation.

B. Quality-Discriminable Image Pairs

As described in the previous subsection, the NR-IQA methods may give quite different objective scores for quality-indiscriminate image sets due to image content variation. Moreover, due to this impact, the NR-IQA methods may give inverse objective scores to quality-discriminable image pairs, which have quite discriminable subjective quality but different image contents. In Fig. 3, the image quality of the clear blue sky is worse than the image quality of the blurry monkey/mouse according to the prediction of traditional methods, which is counter to human visual perception. This can be explained by the fact that traditional methods are mainly based on low-level features, which overlook the impact of image content variation.

To quantitatively measure this impact on different NR-IQA methods, we construct a Content-Variation-Impact-Quality dataset (CVIQ) from BID. According to [55], an image pair is supposed to be quality-discriminable if their absolute MOS difference is larger than 2 std of the MOS. The average std of MOS for all images on BID is 0.8309. Therefore, choosing one image with MOS $>$ 4 and another image with MOS $<$ 2 to form an image pair is a simple procedure for ensuring that the image pair will be reliably discriminable in terms of subjective quality. Thumbnail sample images of CVIQ are shown in Fig. 4, which contains images with intrinsic flat regions, blurry regions, sharp structures and blurry structures. Overall, CVIQ consists of 148 images of high quality (MOS $>$ 4, e.g., the left side of Fig. 4) and 154 blur images of low quality (MOS $<$ 2, e.g., the right side of Fig. 4). By selecting one image with low quality and one image with high quality, we obtain altogether 22,792 (154 $\times$ 148) quality-discriminable image pairs. Table I shows the accuracy of predicting the relative quality on the 22,792 image pairs, where the highest accuracy is indicated in bold. Note that the learning-based methods are trained on CLIVE. The best NR-IQA method (excluding the method proposed in our paper), ILNIQE, only achieves an 85.01% accuracy, which means that 3,417 quality-discriminable image pairs are given incorrect relative predictions by ILNIQE. Moreover, over 11,000 quality-discriminable image pairs are improperly predicted by ARISMc. The results show that existing NR-IQA methods suffer a lot from the image content variation.

TABLE I
THE ACCURACY OF PREDICTING THE RELATIVE QUALITY OF IMAGE PAIRS IN CVIQ, WHERE THE HIGHEST ACCURACY IS IN BOLD

Category	Method	Accuracy
Learning-free	MDWE [25]	73.61%
	MLV [27]	73.21%
	ARISMc [26]	51.68%
	FISHbb [29]	84.49%
	LPC [24]	73.26%
	BIBLE [31]	73.01%
Learning-based	SPARISH [38]	76.54%
	RISE [37]	76.65%
	Yu's CNN [46]	75.95%
	BRISQUE [8]	58.17%
	ILNIQE [9]	85.01%
	SFA (Proposed)	**96.87%**

IV. A Novel NR-IQA Method Based on Semantic Feature Aggregation

Existing methods overlook the impact of image content variation, thereby causing inconsistent predictions on image pairs. Image-content-aware features can help to alleviate the impact of image content variation. Therefore, in this work, we propose a new NR-IQA method that facilitates image-content-aware features with effective aggregation mechanisms. The overall framework is shown in Fig. 5, and includes four key components: image representation, feature extraction, feature aggregation, and quality prediction.

A. Image Representation

To perform the preprocessing step and forward propagation, the pre-trained DCNN models (e.g., ResNet-50 [13]) require a fixed-size input. Therefore, images should be cropped or resized

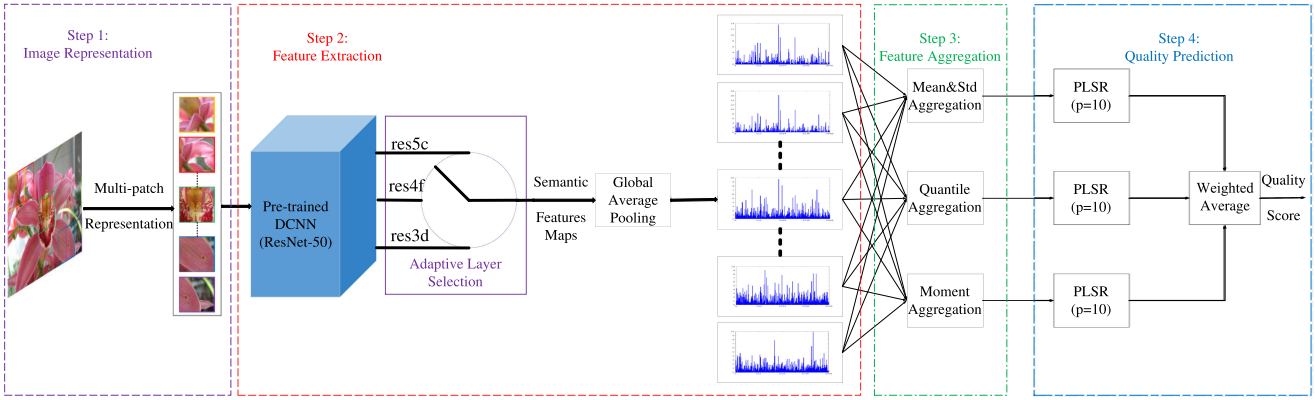


Fig. 5. [Best viewed in color.] The framework of the proposed method, which includes four main steps: multi-patch representation of an image, feature extraction by a pre-trained DCNN model with an adaptive layer selection procedure (in this paper, the 'res3d', 'res4f', and 'res5c' layers of ResNet-50 [13] are considered), feature aggregation by three different statistical structures, and quality prediction by a partial least square regression (PLSR) model with $p = 10$ components.

to a fixed size. Resizing is not a good choice since this can introduce geometric deformation, which can alter the image quality. Meanwhile, cropping the central patch alone is not sufficient to cover most pixels of a large image. Therefore, to avoid these two issues, we empirically consider using multi-patch representation, where we represent an image using multiple overlapping patches (with a stride equaling half the patch size). In this way, it not only covers the entire image but also avoids introducing the unwanted geometric deformation. We have experimentally verified in the conference paper [20] that the multi-patch representation performs well for NR-IQA.

B. Feature Extraction

We represent the given an image $\mathbf{I}$ with a set of multiple overlapping patches $\{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$, where n is the total number of image patches. Then, these patches are fed into a pre-trained DCNN model to extract the features. For each patch $\mathbf{p}_k$, a feature is extracted and denoted by

$$\mathbf{d}_k = \mathrm{GAP}(\mathrm{DCNN}(\mathbf{p}_k; L, \theta)),\ k = 1, \ldots, n. \tag{2}$$

where L indicates from which layer we extract the feature, θ is the model parameter, and GAP means a global average pooling operation that pools the feature maps.

According to comparison results in our conference paper [20], we choose ResNet-50 as the pre-trained DCNN model. As indicated in the conference paper, when there is only a small set of images with different image contents, the role of deeper features is weakened, and the impact of shallower features is enhanced. Therefore, we adaptively extract suitable semantic feature maps from its 'res3d', 'res4f' or 'res5c' (lower to higher) layer based on the five-fold cross validation on training data.

C. Feature Aggregation

Since we extract features from local patches, one of the challenges in applying pre-train DCNN to NR-IQA is that global information will be somewhat weakened or even disregarded. We need effective mechanisms to aggregate the extracted local patch features into a single global feature. A straightforward strategy to do this is to concatenate all n features, i.e.,

$$\mathbf{f}_{\mathrm{concat}} = \mathbf{d}_1 \oplus \cdots \oplus \mathbf{d}_k \oplus \cdots \oplus \mathbf{d}_n \tag{3}$$

where $\oplus$ indicates a concatenation operator.

However, the above concatenation will result in a very high dimensionality of the feature space. Moreover, the dimensions of this concatenated feature vector depend on the number of image patches, which may be differ among the images with various resolutions. To avoid this, we can take the mean feature vector, which is

$$\mathbf{f}_{\mathrm{mean}} = (m_1, \ldots, m_i, \ldots, m_l)^T,$$
$$m_i = \frac{\sum_{k=1}^{n} d_{ki}}{n}, i = 1, \ldots, l. \tag{4}$$

where l is the dimension of $\mathbf{d}_k$, d_{ki} is the i-th element of $\mathbf{d}_k$, and T indicates a transposed operator.

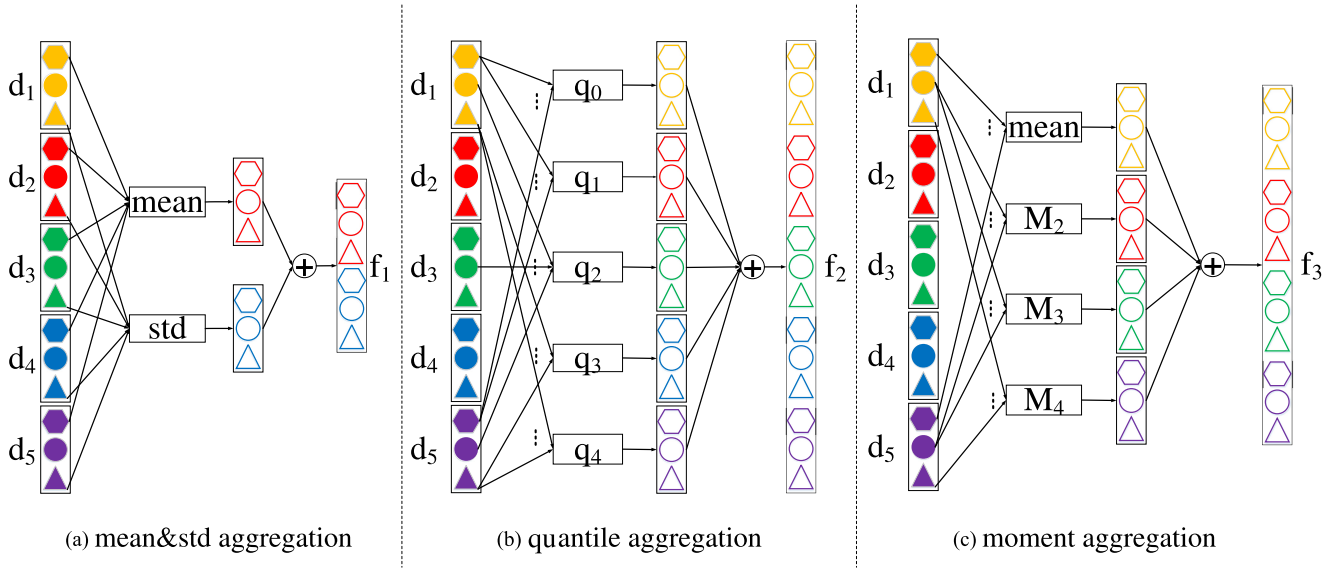
The mean aggregation structure may lose important characteristics (such as the standard deviation in each dimension) of the local patch features, which can harm the final prediction. Therefore, to better deliver the information, we adopt three different statistical structures for feature aggregation: mean&std aggregation, quantile aggregation, and moment aggregation.

Mean&std aggregation: Noticing that the mean cannot reflect the variations, mean and standard deviation are jointly considered in different tasks [56]–[58]. Here, we also consider these two statistics for the same concern, and the 1st aggregated feature $\mathbf{f}_1$ can be calculated as

$$\mathbf{f}_1 = \mathbf{f}_{\mathrm{mean}} \oplus \mathbf{f}_{\mathrm{std}}$$
$$\mathbf{f}_{\mathrm{std}} = \left(\sqrt{\frac{\sum_{k=1}^{n}(d_{k1} - m_1)^2}{n-1}}, \ldots, \sqrt{\frac{\sum_{k=1}^{n}(d_{kl} - m_l)^2}{n-1}} \right)^T \tag{5}$$

where $m_i (i = 1, \ldots, l)$ indicates the i-th element of $\mathbf{f}_{\mathrm{mean}}$.

Quantile aggregation: Lu *et al.* [56] proposed a sorting layer to aggregate the features of several random patches. However, the sorting layer cannot handle images with different numbers of patches. To address this situation, we propose quantile



(a) mean&std aggregation (b) quantile aggregation (c) moment aggregation

Fig. 6. [Best viewed in color.] An illustration of the three statistical structures used for feature aggregation. The inputs are $n = 5$ features $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5\}$, and the feature dimension is $l = 3$. $(\mathbf{q}_0, \mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4)$ indicates the 5 quartiles, and $\mathbf{M}_r^r$ equals the central moment of order r ($r = 2, 3, 4$). For clarity, some links between patch features and statistical functions are omitted.

aggregation since quantiles are important order statistics for describing a distribution. In this work, we choose the widely used quartiles.[1] The zeroth to fourth quartiles of $(d_{1i}, \ldots, d_{ni})$ are denoted as $d_i^{\{0\}}, d_i^{\{1\}}, d_i^{\{2\}}, d_i^{\{3\}}, d_i^{\{4\}}, i = 1, \ldots, l$. Thus, the 2nd aggregated feature $\mathbf{f}_2$ can be obtained by

$$\mathbf{f}_2 = \mathbf{q}_0 \oplus \mathbf{q}_1 \oplus \mathbf{q}_2 \oplus \mathbf{q}_3 \oplus \mathbf{q}_4$$
$$\mathbf{q}_j = \left(d_1^{\{j\}}, \ldots, d_l^{\{j\}}\right)^T, j = 0,1, 2, 3, 4. \quad (6)$$

Moment aggregation: Moments also play a key role in characterizing a distribution.[2] We know that the mean and standard variation can represent Gaussian distributions, but real distributions need more statistics (e.g., the generalized Gaussian distribution (GGD) needs the mean, standard variation and one additional shape parameter). Thus, we also consider the higher order moments in our work. To achieve a balance between the need for more information and the dimensionality reduction of the feature space, we further consider the r-th root of the central moment of order r ($r = 3, 4$) and obtain the 3rd aggregated feature $\mathbf{f}_3$ by

$$\mathbf{f}_3 = \mathbf{f}_{\text{mean}} \oplus \mathbf{M}_2 \oplus \mathbf{M}_3 \oplus \mathbf{M}_4$$
$$\mathbf{M}_r = \left(\sqrt[r]{\frac{\sum_{k=1}^{n}(d_{k1} - m_1)^r}{n}}, \ldots, \sqrt[r]{\frac{\sum_{k=1}^{n}(d_{kl} - m_l)^r}{n}}\right)^T \quad (7)$$

The three aforementioned statistical aggregation structures can result in $2l$-, $5l$-, and $4l$-dimensional feature vectors. An illustration of the three aggregation structures is shown in Fig. 6, where $n = 5, l = 3$. The effectiveness of the three statistical aggregation structures has been verified in our conference paper [20].

D. Quality Prediction

Using statistical structures for feature aggregation, we decrease the dimensionality of the feature space ($nl \rightarrow 2l, 5l, 4l$). In addition, we make the dimensions independent of the number of patches. However, l is usually very large in the pre-trained DCNN model. The feature dimensions remain substantially larger than the size of the database. An effective and efficient regression model for quality prediction is desired. Specifically, in this work, we adopt a simple linear regression model, partial least square regression (PLSR) [59], because of its low complexity and remarkable capability to deal with high-dimensional data. PLSR first reduces the aggregated features to p uncorrelated latent components, where p is the only hyper-parameter in PLSR. Then, the least square regression is performed on these components. p is set globally to 10 for simplicity. To take advantage of ensemble learning, the quality prediction is given by the weighted average of the scores predicted by the three PLSR models.

E. Deep Semantic Features for NR-IQA: Aware of Image Content

The deep semantic features are extracted from the image classification DCNN models pre-trained on ImageNet. Therefore, the deep semantic features contain crucial image content information, i.e., they are aware of image content and expected to help to overcome the issue of image content variation. We also test the proposed NR-IQA method, SFA, in the two scenarios as described in Section III, to see if deep semantic features indeed alleviate the impact of image content variation.

[1]The min, median and max are the zeroth, second, and fourth quartiles, respectively.

[2]The mean is actually the first-order origin moment, and the standard variation is actually the square root of the second-order central moment.

In Fig. 1, compared with existing NR-IQA methods, our proposed method, SFA, shows the smallest difference (5.0062-4.3259 = 0.6803) in the predictions among the six quality-indiscriminate images. In addition, the image content variation has the smallest impact on our method for quality-indiscriminate images among the NR-IQA methods (see Fig. 2).

In Fig. 3, our SFA method predicts that (b)/(c) is worse than (a), which is in accordance with subjective ratings. In Table I, SFA achieves the best accuracy (96.87%, with 10+% gains over the second and third best methods, ILNIQE and FISHbb), which means that our semantic-feature-aggregation-based method can reduce the impact of image content variation on relative quality predictions.

These results verify the effectiveness of deep semantic features for addressing image content variation in NR-IQA, and they provide a new perspective for NR-IQA in terms of the semantic aspect.

V. Experimental Results and Analysis

We first describe the experimental settings. Then, the performance comparison of our method SFA with 9 representative blur-specific NR-IQA methods, 2 general-purpose NR-IQA methods and 2 extra FR-IQA methods is conducted on Gaussian blur images (without and with Gaussian noise/JPEG compression) and realistic blur images from multiple databases. Moreover, we conduct a statistical significance test to determine whether the comparison results are significant. Then, we verify the performance on the Waterloo exploration database [21] and generalization capability of SFA. After that, the contribution of the adaptive layer selection procedure is shown. Finally, the computational efficiency of each method is reported.

A. Experimental Settings

Compared methods: We choose 9 representative blur-specific NR-IQA methods, 2 general-purpose NR-IQA methods and 2 extra FR-IQA methods for comparison. The nine NR methods are MDWE [25], MLV [27], ARISMc [26], FISHbb [29], LPC [24], BIBLE [31], SPARISH [38], RISE [37], and Yu's CNN [46]. The two general-purpose NR-IQA methods are BRISQUE [8] and ILNIQE [9]. The two FR methods are IWSSIM [2] and VSI [3]. Note that the codes of these compared methods are obtained from the original authors.

Basic evaluation criteria: For methods without training in the quality prediction step, we refer to the suggestion of the Video Quality Experts Group (VQEG) [60] and adopt a four-parameter logistic function for mapping the objective score o to the subjective score s:

$$f(o) = \frac{\tau_1 - \tau_2}{1 + e^{-\frac{o-\tau_3}{\tau_4}}} + \tau_2 \tag{8}$$

where τ_1 to τ_4 are fitting parameters initialized with $\tau_1 = \max(s), \tau_2 = \min(s), \tau_3 = \text{mean}(o), \tau_4 = \text{std}(o)/4$.

Five basic criteria are chosen for the performance comparison:

1) Spearman's Rank-order Correlation Coefficient (SROCC) computes the prediction monotonicity and indicates how well the relationship between subjective and objective quality can be depicted by a monotonic function:

$$\text{SROCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{9}$$

where N represents the size of the testing dataset and d_i is the rank difference of the i-th image's subjective and objective scores.

2) Kendall's rank-order correlation coefficient (KROCC) is another prediction monotonicity criterion and indicates the ordinal association between the subjective and objective quality:

$$\text{KROCC} = \frac{2(F_c - F_d)}{N(N-1)} \tag{10}$$

where F_c and F_d are the numbers of concordant and discordant pairs on the testing database.

3) Pearson's linear correlation coefficient (PLCC) is a measure of the linear correlation between the subjective scores and the mapped scores, which means the prediction accuracy:

$$\text{PLCC} = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2 \sum_{i=1}^{N}(f_i - \bar{f})^2}} \tag{11}$$

where s_i and $\bar{s}$ are the i-th subjective score and the mean of all s_i, respectively, and f_i and $\bar{f}$ are the i-th mapped objective score after the non-linear mapping and the mean of all f_i.

4) Root mean square error (RMSE) is another prediction accuracy criterion that represents the distance between the subjective scores and the mapped scores:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - f_i)^2} \tag{12}$$

5) Outlier's ratio (OR) is a prediction consistency measure that gives the percentage of the mapped scores deviating from the subjective scores in a "2 standard deviation" sense:

$$\text{OR} = \frac{\sum_{i=1}^{N}(|s_i - f_i| > 2\sigma_i)}{N} \times 100\% \tag{13}$$

where σ_i is the i-th standard deviation of the raw subjective scores.

Image databases: We consider blur images from seven databases: LIVE [14], TID2008 [15], TID2013 [16], MLIVE1 [17], MLIVE2 [17], BID [18] and CLIVE [19]. Table II provides general information, covering the number of reference images (# Reference image), number of blur images (# Blur image), blur type, score type and score range.

1) Gaussian blur images from LIVE, TID2008, and TID2013 are obtained using Gaussian filters. There are 29, 25, and 25 reference images, respectively, and 5, 4, and 5 blur kernels, respectively, for each reference image. These kernels generate a total of 145, 100, and 125 blur images, respectively.
2) Noisy blurred images in MLIVE1 simulate image acquisition where images are out of focus and corrupted by

TABLE II
GENERAL INFORMATION OF THE SEVEN DATABASES USED FOR THE COMPARATIVE EXPERIMENTS. "# REFERENCE IMAGE" MEANS THE NUMBER OF REFERENCE IMAGES, AND "# BLUR IMAGE" MEANS THE NUMBER OF BLUR IMAGES

Database	# Reference image	# Blur image	Blur type	Score type*	Score range
LIVE [14]	29	145	Gaussian blur	DMOS	[0 100]
TID2008 [15]	25	100	Gaussian blur	MOS	[0 9]
TID2013 [16]	25	125	Gaussian blur	MOS	[0 9]
MLIVE1 [17]	15	225	Gaussian blur with white Gaussian noise	DMOS	[0 100]
MLIVE2 [17]	15	225	Gaussian blur with JPEG compression	DMOS	[0 100]
BID [18]	-	586	Realistic blur (out-of-focus, motion, *etc.*)	MOS	[0 5]
CLIVE [19]	-	1162	Realistic blur	MOS	[0 100]

*DMOS indicates the difference of mean opinion scores (MOS) between the test image and its reference image.

sensor noise. There are 15 reference images with 4 levels of Gaussian blur and 4 levels of additive white Gaussian noise, resulting in 225 distorted images (the 15 reference images are excluded).

3) Compressed blurred images in MLIVE2 simulate image storage where images are out of focus and compressed by the JPEG encoder. There are 15 reference images with 4 levels of Gaussian blur and 4 compression levels, which results in 225 distorted images (The 15 reference images are excluded).
4) Realistic blur images from BID and CLIVE are taken from the real world and include a variety of scenes, camera apertures and exposure times. There are 586 images in BID and 1162 images in CLIVE in total.

B. Performance Comparison

In the intra-database experiments, 80% of the data are randomly selected as training data, and the other 20% is used for testing on each database. Training and testing data do not share the same reference image. To avoid bias, this procedure is run 1000 times. Table III summarizes the median performance values on the seven databases (LIVE, TID2008, TID2013, MLIVE1, MLIVE2, BID and CLIVE). The proposed SFA method achieves high performances on all seven databases.

For Gaussian blur images, our SFA method is comparable with BIBLE and SPARISH on LIVE, while it outperforms the other NR methods on TID2008 and TID2013. The four best NR methods on LIVE, i.e., BIBLE, SPARISH, SFA and ARISMc achieve better performances than the FR method, VSI.

For Gaussian blur images with Gaussian noise/JPEG compression, the proposed method significantly outperforms both the NR and FR methods. Noise increases the high-frequency components of the image, while blur decreases them; therefore, most of the other methods suffer poor performance on MLIVE1 due to the presence of noise (SROCC values are less than 0.5, some of which even being negative). On MLIVE2, most methods have an SROCC value of greater than 0.8 since JPEG compression (similar to blur) also causes a reduction in the high-frequency components. Noted that in addition to SFA, ILNIQE is also a good NR-IQA method for Gaussian blur images with Gaussian noise/JPEG compression.

For realistic blur images, SFA achieves the best performance on CLIVE and BID, and it achieves a significant performance gain over the other methods in both prediction monotonicity (i.e., SROCC and KROCC), accuracy (i.e., PLCC and RMSE) and consistency (i.e., OR). The two general-purpose methods suffer poor performances on BID and CLIVE. The first seven blur-specific NR-IQA methods do not work well on realistic blur image databases (PLCC < 0.55, KROCC and SROCC < 0.5) due to their neglect of deep semantic features. RISE and Yu's CNN are slightly better than the first nine NR methods. This can be explained by the fact that RISE considers multi-scale and multi-resolution features, while Yu's CNN attempts to learn quality-relevant features.

Discussion on realistic blur: It is difficult to model all the influencing factors in the real world. In addition to the Gaussian and out-of-focus blur, there are other crucial factors to be considered, e.g., the motion blur in Fig. 7(a), the ghosting in Fig. 7(b), the macrophotography in Fig. 7(c) and the image content variation in Fig. 7(d). A substantial portion of Section III has discussed the impact of image content variation. Here, we give a few comments on the other factors.

1) Motion blur: there are few NR-IQA methods for estimating motion blur, although its related problem "motion deblurring", has become a hot topic [50], [53]. Motion blur has directionality, whereas Gaussian blur is isotropic. In terms of the specific characteristic of motion blur, one may further consider the directionality and the directional features for quality estimation on motion blur images. We believe the availability of a large realistic motion blur image database with subjective ratings will certainly facilitate such work.
2) Ghosting: ghosting effects arise when the motion degree is very high, in contrast to ordinary motion blur. Some related articles (e.g., [61]) have considered ghosting effects in designing NR-IQA methods.
3) Macrophotography: the blur in Bokeh is used to strengthen a photo's expressiveness. In light of this, to assess the perceptual quality of macrophotography images, aesthetic factors may need to be considered.

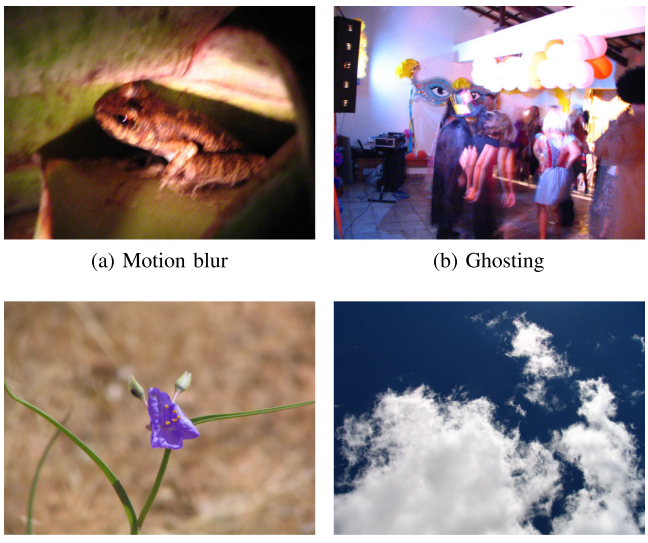
TABLE III
INTRA-DATABASE PERFORMANCE COMPARISON. IN EACH COLUMN, THE BEST PERFORMANCE VALUES OF NR METHODS ARE MARKED IN BOLDFACE

Method	LIVE [14]					TID2008 [15]					TID2013 [16]				
	SROCC↑	KROCC↑	PLCC↑	RMSE↓	OR↓	SROCC	KROCC	PLCC	RMSE	OR	SROCC	KROCC	PLCC	RMSE	OR
IWSSIM [2]	0.9723	0.8733	0.9698	4.4734	40.00%	0.9680	0.8707	0.9533	0.3459	45.00%	0.9723	0.8787	0.9526	0.3753	56.80%
VSI [3]	0.9538	0.8300	0.9535	5.4508	48.00%	0.9592	0.8496	0.9551	0.3397	50.00%	0.9669	0.8581	0.9571	0.3593	56.00%
MDWE [25]	0.9188	0.7800	0.9377	6.4427	52.00%	0.8556	0.6579	0.8660	0.5697	70.00%	0.8466	0.6467	0.8698	0.6039	72.00%
MLV [27]	0.9431	0.8133	0.9578	5.2170	48.00%	0.8977	0.7158	0.9075	0.4837	65.00%	0.9142	0.7446	0.9226	0.4762	64.00%
ARISMc [26]	0.9585	0.8467	0.9684	4.6117	**40.00%**	0.8851	0.7124	0.8872	0.5266	65.00%	0.9108	0.7513	0.9149	0.4938	64.00%
FISHbb [29]	0.9469	0.8267	0.9570	5.2410	48.00%	0.8737	0.6807	0.8916	0.5160	65.00%	0.8900	0.7067	0.9087	0.5100	68.00%
LPC [24]	0.9469	0.8133	0.9326	6.6480	56.00%	0.8805	0.6860	0.8858	0.5334	65.00%	0.9049	0.7267	0.9086	0.5132	64.00%
BIBLE [31]	**0.9638**	0.8533	0.9711	**4.3871**	**40.00%**	0.9114	0.7441	0.9178	0.4575	**60.00%**	0.9131	0.7446	0.9264	0.4615	64.00%
SPARISH [38]	**0.9638**	**0.8600**	0.9693	4.4870	**40.00%**	0.9126	0.7474	0.9164	0.4628	**60.00%**	0.9102	0.7400	0.9228	0.4716	64.00%
RISE [37]	0.9492	0.8267	0.9594	5.6563	48.00%	0.9203	0.7757	0.9235	0.4891	**60.00%**	0.9300	0.7800	0.9342	0.4971	68.00%
Yu's CNN [46]	0.9469	0.8200	0.9486	6.5674	48.00%	0.8752	0.6737	0.8784	0.6426	70.00%	0.8929	0.7067	0.9020	0.6195	76.00%
BRISQUE [8]	-	-	-	-	-	0.8782	0.6947	0.8865	0.5330	65.00%	0.8878	0.7067	0.8963	0.5536	68.00%
ILNIQE [9]	0.9308	0.7933	0.9444	6.1241	56.00%	0.8451	0.6491	0.8617	0.5782	70.00%	0.8466	0.6533	0.8675	0.6134	76.00%
SFA (Proposed)	0.9631	**0.8600**	**0.9722**	4.7469	**40.00%**	**0.9368**	**0.8000**	**0.9455**	**0.4193**	**60.00%**	**0.9477**	**0.8180**	**0.9542**	**0.4281**	**60.00%**

BRISQUE is trained on the full LIVE IQA database.

Method	MLIVE1 [17]					MLIVE2 [17]				
	SROCC	KROCC	PLCC	RMSE	OR	SROCC	KROCC	PLCC	RMSE	OR
IWSSIM [2]	0.9198	0.7624	0.9340	6.4245	0.00%	0.9103	0.7495	0.9386	6.4895	0.00%
VSI [3]	0.8882	0.7179	0.9104	7.5412	0.00%	0.8797	0.7067	0.9131	7.6284	0.00%
MDWE [25]	0.0869	0.0607	0.2447	17.9239	6.67%	0.5632	0.4107	0.6465	14.4053	2.22%
MLV [27]	0.4687	0.3175	0.6422	13.9836	4.44%	0.8256	0.6202	0.8827	8.9481	**0.00%**
ARISMc [26]	-0.2926	-0.2116	0.3960	17.0197	8.89%	0.8763	0.7125	0.9214	7.3130	**0.00%**
FISHbb [29]	0.3087	0.2114	0.2996	16.7142	6.67%	0.7598	0.5642	0.8560	9.7748	**0.00%**
LPC [24]	0.4401	0.3074	0.6585	13.7785	4.44%	0.7018	0.5023	0.8441	10.1885	**0.00%**
BIBLE [31]	0.1563	0.0971	0.3147	17.4678	8.89%	0.8337	0.6384	0.8953	8.2416	**0.00%**
SPARISH [38]	-0.0532	-0.0313	0.3370	17.3901	6.67%	0.9132	0.7556	0.9413	6.4184	**0.00%**
RISE [37]	0.8613	0.6761	0.8877	10.4500	**0.00%**	0.8846	0.7152	0.9240	8.6906	**0.00%**
Yu's CNN [46]	0.8828	0.7125	0.8959	10.4125	**0.00%**	0.8759	0.7040	0.9140	9.0764	**0.00%**
BRISQUE [8]	0.3055	0.2239	0.4071	16.8893	8.89%	0.8200	0.6458	0.9006	8.1076	**0.00%**
ILNIQE [9]	0.9219	0.7652	0.9290	**6.7615**	**0.00%**	0.9104	0.7495	0.9278	**7.1369**	**0.00%**
SFA (Proposed)	**0.9373**	**0.7899**	**0.9419**	7.5586	**0.00%**	**0.9404**	**0.8000**	**0.9468**	7.4790	**0.00%**

Method	BID [18]					CLIVE [19]				
	SROCC	KROCC	PLCC	RMSE	OR	SROCC	KROCC	PLCC	RMSE	OR
MDWE [25]	0.3067	0.2123	0.3538	1.1639	23.08%	0.4313	0.2956	0.4988	17.5025	6.90%
MLV [27]	0.3169	0.2199	0.3750	1.1561	22.22%	0.3412	0.2318	0.4076	18.4350	7.76%
ARISMc [26]	-0.0151	-0.0105	0.1929	1.2245	26.50%	0.2427	0.1631	0.3554	18.8947	8.19%
FISHbb [29]	0.4736	0.3254	0.4853	1.0894	18.80%	0.4865	0.3320	0.5380	17.0310	6.47%
LPC [24]	0.3150	0.2159	0.4053	1.1408	22.22%	0.1483	0.0968	0.3490	18.9205	7.76%
BIBLE [31]	0.3609	0.2449	0.3923	1.1469	22.22%	0.4260	0.2931	0.5178	17.3007	6.90%
SPARISH [38]	0.3074	0.2088	0.3555	1.1659	23.08%	0.4015	0.2750	0.4843	17.6702	7.33%
RISE [37]	0.5632	0.3978	0.5681	1.0543	17.09%	0.5152	0.3586	0.5550	17.1360	6.03%
Yu's CNN [46]	0.5572	0.3902	0.5600	1.0649	20.51%	0.5017	0.3491	0.5010	18.3058	8.19%
BRISQUE [8]	0.1051	0.0678	0.2246	1.2166	26.50%	0.3153	0.2136	0.3758	18.7053	8.62%
ILNIQE [9]	0.4963	0.3439	0.5192	1.0649	17.95%	0.4401	0.3013	0.5102	17.3930	6.47%
SFA (Proposed)	**0.8263**	**0.6334**	**0.8399**	**0.6859**	**5.98%**	**0.8119**	**0.6195**	**0.8331**	**11.3525**	**0.86%**



(a) Motion blur (b) Ghosting

(c) Macrophotography (d) Content variation

Fig. 7. Crucial factors (in addition to the Gaussian and out-of-focus blur) that influence image quality of realistic blur images.

TABLE IV
THE T-TEST RESULTS ON LIVE [14], TID2008 [15] AND TID2013 [16], MLIVE1 [17], MLIVE2 [17], BID [18] AND CLIVE [19]

Method	LIVE	TID2008	TID2013	MLIVE1	MLIVE2	BID	CLIVE
IWSSIM [2]	-1	-1	-1	1	1	-	-
VSI [3]	1	-1	-1	1	1	-	-
MDWE [25]	1	1	1	1	1	1	1
MLV [27]	1	1	1	1	1	1	1
ARISMc [26]	0	1	1	1	1	1	1
FISHbb [29]	1	1	1	1	1	1	1
LPC [24]	1	1	1	1	1	1	1
BIBLE [31]	0	1	1	1	1	1	1
SPARISH [38]	0	1	1	1	1	1	1
RISE [37]	1	1	1	1	1	1	1
Yu's CNN [46]	1	1	1	1	1	1	1
BRISQUE [8]	-	1	1	1	1	1	1
ILNIQE [9]	1	1	1	1	1	1	1

1 (-1) indicates SFA statistically outperforms (underperforms) the compared method, and 0 indicates SFA is statistically on par with the compared method. We use '-' since FR methods are inapplicable on realistic blur image databases and results of BRISQUE on LIVE are not available.

C. Statistical Significance

We conduct statistical significance tests to determine whether the comparison results in the previous subsection are significant.

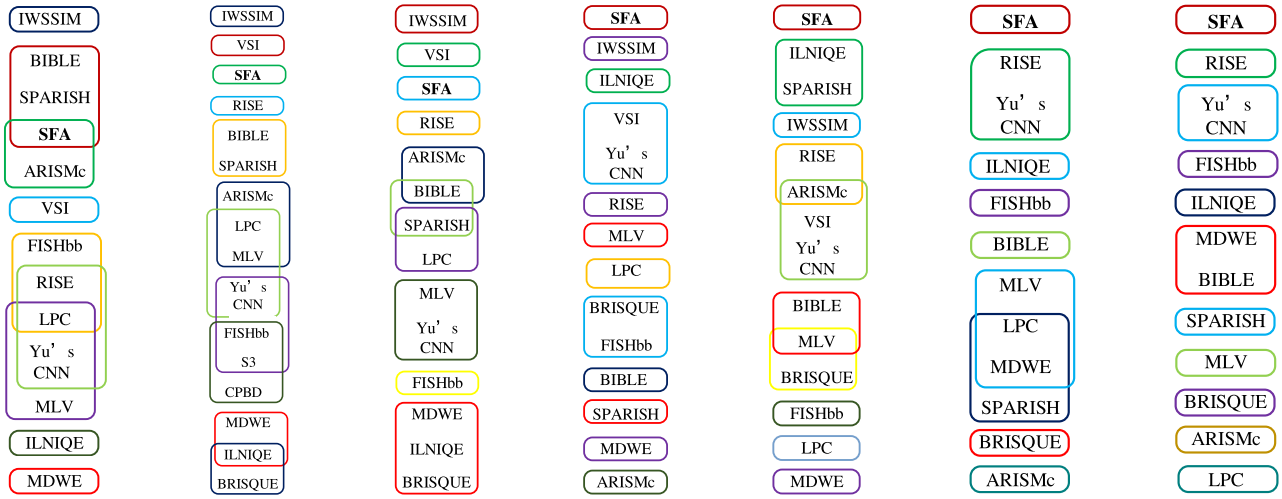


Fig. 8. [Best viewed in color.] Global ranking and grouping of methods by their statistical significance results. The methods on the upper positions achieve a better performance, and the methods within the same rectangle are statistically indistinguishable, i.e., their performances are similar.

On each database, a two sample t-test is conducted at 1‰significance level using the SROCC value pairs of 1000 runs.

Table IV lists the statistical significance test results, where 1 (-1) indicates that SFA statistically outperforms (underperforms) the compared method, and 0 indicates that SFA is on par with the compared method. On LIVE, SFA is significantly outperformed by the FR method, IWSSIM; is on par with BIBLE, SPARISH and ARISMc; and significantly outperforms the others (including the FR method, VSI). On TID2008 and TID2013, SFA method statistically outperforms the other NR methods, while it statistically underperforms the two FR methods, IWSSIM and VSI. On MLIVE, SFA is the best method based on the significance test, being even better than the two FR methods IWSSIM and VSI. On BID and CLIVE, SFA is statistically superior to all the other NR methods. Generally, SFA statistically outperforms other methods. Moreover, we conduct further statistical significance tests between each pair of methods. In addition, based on the statistical significance results, we give a global ranking and grouping of the evaluated methods on each database in Fig. 8. SFA is always in the top rectangle, which demonstrates the superior performance of SFA.

D. Performance on Waterloo Exploration Database

Waterloo exploration database [21] is a large IQA database containing various image contents. We present the test results of the proposed method on this database. RISE is considered for comparison because of its top performance on the seven databases. We first generate the blurred images by applying 2D circularly symmetric Gaussian blur kernels with standard deviations (std) of [0.5, 1.2, 2.5, 6.5, 15.2] to the source images. Similar to [21], we generate 138,274,885 discriminable image pairs (DIPs). Then we conduct a D-test, L-test and P-test [21], whose results are shown in Table V. The D-test (D) measures the ability of an IQA method to separate the pristine and distorted images. The L-test (L_S, L_K) evaluates the consistency of IQA methods when doing monotonic predictions, i.e., predicting image quality for images with different distortion levels but the same content and the same distortion type. The P-test (P, M_i) tests the ability of IQA methods to predict the relative quality predictions on a number of DIPs. The D-test results of both SFA and RISE are smaller than 0.9. This is because it is difficult for both SFA and RISE to distinguish the slightly blurred images from the pristine ones, since there exists a slight blur (i.e., std $= 0.5$). In addition, SFA is slightly inferior to RISE in the D-test because the high-level features used in SFA are less sensitive to slight blur than the low-level features used in RISE. In the L-test, SFA can achieve more consistent monotonic predictions than RISE. In the P-test, compared to RISE, SFA greatly decreases the number of incorrect preference predictions (M_i) from around 80000 to around 50000.

TABLE V
THE D-TEST, L-TEST, AND P-TEST RESULTS ON THE WATERLOO EXPLORATION DATABASE [21]

Method	$D\uparrow$	$L_S\uparrow$	$L_K\uparrow$	$P\uparrow$	$M_i\downarrow$
SFA (Proposed)	0.8714	0.9979	0.9951	0.9996	49958
RISE	0.8994	0.9763	0.9450	0.9994	79688

The definitions of D, L_S, L_K, P, M_i are referred to [21], [44], [55]

E. Generalization Capability

Generalization capability is an important issue for learning-based methods. In this subsection, we verify the generalization capability of SFA using cross-database evaluation and compare it with RISE, since RISE is better than Yu's CNN in terms of the average performance on the seven databases. Fig. 9 shows the SROCC of SFA and RISE. In most of the cross-database experiments, the SROCC value of the SFA is greater than the SROCC value of RISE, which means that our method has better generalization capability than RISE.

For Gaussian blur image databases (LIVE, TID2008 and TID2013), when the SFA model is trained on one database

Train \ Test	LIVE	TID2008	TID2013	MLIVE1	MLIVE2	BID	CLIVE
LIVE	0.9631/0.9492	**0.9313**/0.9138	**0.9460**/0.9339	**0.3732**/0.1823	**0.7168**/0.6192	**0.5267**/0.0080	**0.4972**/0.2857
TID2008	**0.9429**/0.8638	0.9368/0.9203	**0.9815**/0.8696	**0.3597**/0.0483	**0.6834**/0.6029	**0.3667**/0.1506	**0.4664**/0.0638
TID2013	**0.9165**/0.8497	**0.9839**/0.8913	0.9477/0.9300	**0.2801**/0.3383	**0.6191**/0.4543	**0.2769**/0.0900	**0.4832**/0.2317
MLIVE1	0.8534/**0.8603**	**0.8161**/0.7775	**0.7922**/0.7157	0.9373/0.8613	**0.9025**/0.6868	**0.4474**/0.3896	0.2036/**0.2334**
MLIVE2	**0.9007**/0.7926	**0.8570**/0.8056	**0.8394**/0.6544	**0.7917**/0.4859	0.9404/0.8846	**0.4609**/0.2261	**0.3682**/0834
BID	0.7945/**0.8760**	0.7600/**0.8017**	**0.7602**/0.7106	**0.7570**/0.5504	**0.8129**/7607	0.8263/0.5632	**0.6362**/0.1931
CLIVE	**0.8897**/0.8156	**0.8603**/0.7791	**0.8796**/0.7255	**0.5643**/0.0672	**0.7995**/4754	**0.7380**/0.3613	0.8119/0.5152

Fig. 9. [Viewed in color.] The SROCC values in the form of SFA/RISE in the cross-database evaluation. In each entry, the better value is indicated in bold. Note that the intra-database experimental results are also shown (in gray) as a reference. The numerical values in red mean that the corresponding SROCC values are negative. The blue blocks emphasize the results whereby both training and testing data are simulated/realistic blur.

TABLE VI
SROCC AND PLCC OF THE PROPOSED METHOD WITH AND WITHOUT THE ADAPTIVE LAYER SELECTION PROCEDURE

Adaptive Layer Selection	SROCC			PLCC		
	No	Yes	Gain	No	Yes	Gain
LIVE [14]	0.9538	0.9631	+0.0093	0.9644	0.9722	+0.0078
TID2008 [15]	0.9038	0.9368	+0.0330	0.9063	0.9455	+0.0392
TID2013 [16]	0.9077	0.9477	+0.0400	0.9133	0.9542	+0.0409
MLIVE1 [17]	0.8828	0.9373	+0.0545	0.8934	0.9419	+0.0485
MLIVE2 [17]	0.9343	0.9404	+0.0061	0.9449	0.9468	+0.0019
BID [18]	0.8269	0.8263	-0.0006	0.8401	0.8399	-0.0002
CLIVE [19]	0.8130	0.8119	-0.0011	0.8313	0.8399	+0.0086

and tested on the other two databases, the SROCC values are greater than 0.9165, which means a good generalization capability of SFA on Gaussian blur images. When SFA is trained on noisy/compressed blurred images or realistic blur images, the testing performance on Gaussian blur images is also encouraging (SROCC $>$ 0.76). Since realistic blur is substantially more complex than simulated blur, SFA trained on simulated blur images cannot be generalized to realistic situations. Therefore, the SFA model trained on simulated blur does not perform well on realistic blur. In realistic blur situations (BID and CLIVE), SFA is trained on one of the realistic databases and tested on the other database. It can be observed that the SROCC values of SFA in cross-database experiments are greater than 0.63, being far greater than the reported SROCC values of RISE in the intra-database experiments.

F. Contribution of Adaptive Layer Selection

The original layer chosen in the conference version was "res5c" [20]. Compared to the conference version, the proposed framework is modified with an adaptive layer selection procedure. Specifically, we adaptively determine the layer from which to extract features using five-fold cross validation on training data. Table VI lists the results of the proposed framework with/without the adaptive layer selection procedure, where the gain of using the adaptive layer selection procedure is also reported. We can see that the adaptive layer selection procedure improves the performance of SFA on the simulated images.

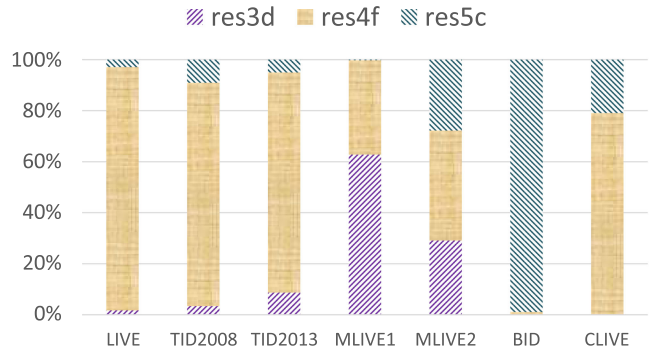


Fig. 10. [Best viewed in color.] The ratio of the selected 'res3d', 'res4f' and 'res5c' layer in the adaptive layer selection procedure. The original layer chosen in the conference version is 'pool5', which is equivalent to applying the global average pooling to the 'res5c' layer.

TABLE VII
THE AVERAGE COMPUTATION TIME (SECONDS/IMAGE) ON TID2013 (512 $\times$ 384) AND MLIVE1 (1280 $\times$ 720)

Time (sec)	IWSSIM	VSI	MDWE	MLV	ARISMc	FISHbb	LPC
TID2013	0.1060	0.0606	0.1594	0.0176	8.2553	0.1436	0.1772
MLIVE	0.4097	0.1117	0.8398	0.0847	39.7846	0.6656	0.8944
Time (sec)	**BIBLE**	**SPARISH**	**RISE**	**Yu's CNN**	**BRISQUE**	**ILNIQE**	**SFA**
TID2013	4.2321	4.9975	0.3860	3.1171	0.0635	3.4382	4.9311
MLIVE	19.3742	22.6687	1.6065	14.8752	0.1945	3.5103	25.2485

In Fig. 10, we take a closer look at the ratio of the selected 'res3d', 'res4f' and 'res5c' layers in the adaptive layer selection procedure. It can be observed that the ratios of the selected lower level ('res3d' and 'res4f') layers on the simulated blur databases are greater than the ratios on realistic blur databases. These results experimentally verify that, the deeper features are less effective in the quality assessment of the simulated blur images compared to the realistic blur images. This can be explained by the fact that there are less than 30 reference images in the simulated blur database, which means less image content; thus, the role of deeper semantic features is weakened. With the adaptive layer selection procedure, the model chooses more suitable deep features; therefore, the performance on simulated blur images is significantly improved, while the performance on realistic blur is barely changed.

G. Computational Efficiency

To compare the computational efficiency of different methods, all tests are performed on a desktop computer with an Intel Core i7-6700K CPU at 4.00 GHz, 64 GB of RAM, Ubuntu 14.04 and MATLAB 2016b (Yu's CNN is implemented using Python 2.7.6 on the same computer). We use the default settings of the original codes and do not optimize them. The average computation time (seconds/image) on TID2013 (512 $\times$ 384) and MLIVE1 (1280 $\times$ 720) for each method is shown in Table VII, which suggests that the computational cost of our method is of the same order as the computational cost of certain complex methods. However, our method can be more than 6x faster when using a TITAN Xp GPU (0.8130 seconds/image on TID2013 and 2.4164 seconds/image on MLIVE1). In addition, we can enlarge the patch stride in the multi-patch representation step to obtain a more efficient model.

VI. Conclusion

In this work, we have shown that existing NR-IQA methods contradict human visual perception on account of image content variation. To alleviate the impact of image content variation, we have proposed a novel NR-IQA method based on semantic feature aggregation (SFA), where semantic features are extracted from the pre-trained ResNet-50 with an adaptive layer (feature) selection procedure, and the aggregation is achieved by merging statistical characteristics. For quantitatively measuring the impact of image content variation, we have designed new experiments, constructed new datasets and defined new quantitative criteria in two scenarios. The experimental results have verified the key role of deep semantic features in addressing image content variation and have demonstrated the superiority and generalization capability of SFA on both realistic and synthetic images.

Acknowledgment

The authors would like to thank Diqi Chen and Daochang Liu for their helpful discussions, and the anonymous reviewers for constructive comments.

References

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

[2] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[3] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[4] H. Hadizadeh and I. V. Baji, "Full-reference objective quality assessment of tone-mapped images," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 392–404, Feb. 2018.

[5] Y. Xu, D. Liu, Y. Quan, and P. Le Callet, "Fractal analysis for reduced reference image quality assessment," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2098–2109, Jul. 2015.

[6] S. Golestaneh and L. J. Karam, "Reduced-reference quality assessment based on the entropy of DWT coefficients of locally weighted gradient magnitudes," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5293–5303, Nov. 2016.

[7] Y. Liu *et al.*, "Reduced-reference image quality assessment in free-energy principle and sparse representation," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 379–391, Feb. 2018.

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

[9] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[10] J. Xu *et al.*, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.

[11] Q. Li, W. Lin, J. Xu, and Y. Fang, "Blind image quality assessment using statistical structural and luminance features," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2457–2469, Dec. 2016.

[12] Q. Wu *et al.*, "Blind image quality assessment based on rank-order regularized regression," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2490–2504, Nov. 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

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

[15] N. Ponomarenko *et al.*, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[16] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process.: Image Commun.*, vol. 30, pp. 57–77, 2015.

[17] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Asilomar Conf. Signals Syst. Comput.*, 2012, pp. 1693–1697.

[18] A. Ciancio *et al.*, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, Jan. 2011.

[19] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[20] D. Li, T. Jiang, and M. Jiang, "Exploiting high-level semantics for no-reference image quality assessment of realistic blur images," in *Proc. ACM Multimedia Conf.*, 2017, pp. 378–386.

[21] K. Ma *et al.*, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.

[22] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 717–728, Apr. 2009.

[23] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Sep. 2011.

[24] R. Hassen, Z. Wang, and M. M. A. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2798–2810, Jul. 2013.

[25] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," *Signal Process.: Image Commun.*, vol. 19, no. 2, pp. 163–172, 2004.

[26] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "No-reference image sharpness assessment in autoregressive parameter space," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3218–3231, Oct. 2015.

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

[28] K. Bahrami and A. C. Kot, "Efficient image sharpness assessment based on content aware total variation," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1568–1578, Aug. 2016.

[29] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Process. Lett.*, vol. 19, no. 7, pp. 423–426, Jul. 2012.

[30] C. T. Vu, T. D. Phan, and D. M. Chandler, "S_3: A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 934–945, Mar. 2012.

[31] L. Li *et al.*, "No-reference image blur assessment based on discrete orthogonal moments," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 39–50, Jan. 2016.

[32] Z. Wang and E. P. Simoncelli, "Local phase coherence and the perception of blur," in *Proc. Proc. Neural Inf. Process. Syst. Conf.*, 2003, pp. 1435–1442.

[33] A. Leclaire and L. Moisan, "No-reference image quality assessment and blind deblurring with sharpness metrics exploiting Fourier phase information," *J. Math. Imag. Vis.*, vol. 52, no. 1, pp. 145–172, 2015.

[34] R. Mukundan, S. Ong, and P. A. Lee, "Image analysis by Tchebichef moments," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1357–1364, Sep. 2001.

[35] S. Wang, C. Deng, B. Zhao, G.-B. Huang, and B. Wang, "Gradient-based no-reference image blur assessment using extreme learning machine," *Neurocomputing*, vol. 174, pp. 310–321, 2016.

[36] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[37] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, "No-reference and robust image sharpness evaluation based on multi-scale spatial and spectral features," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1030–1040, May 2017.

[38] L. Li *et al.*, "Image sharpness assessment by sparse representation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1085–1097, Jun. 2016.

[39] Q. Lu, W. Zhou, and H. Li, "A no-reference image sharpness metric based on structural information using sparse representation," *Inf. Sci.*, vol. 369, pp. 334–346, 2016.

[40] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1733–1740.

[41] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2877–2884.

[42] J. Gu, G. Meng, J. Redi, S. Xiang, and C. Pan, "Blind image quality assessment via vector regression and object oriented pooling," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1140–1153, May 2018.

[43] J. Guan, S. Yi, X. Zeng, W. K. Cham, and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2505–2520, Nov. 2017.

[44] K. Ma *et al.*, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[45] S. Bosse, D. Maniry, K. R. Mller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[46] S. Yu *et al.*, "A shallow convolutional neural network for blind image sharpness assessment," *PLOS ONE*, vol. 12, no. 5, 2017, Art. no. e0176632.

[47] J. Shi, L. Xu, and J. Jia, "Discriminative blur detection features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2965–2972.

[48] K. Ma, H. Fu, T. Liu, Z. Wang, and D. Tao, "Deep blur mapping: Exploiting high-level semantics by deep neural networks," *IEEE Trans. Image Process.*, vol. 27, 10, pp. 5155–5166, Oct. 2018.

[49] W. Zhao, F. Zhao, D. Wang, and H. Lu, "Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3080–3088.

[50] A. Levin, "Blind motion deblurring using image statistics," in *Proc. Proc. Neural Inf. Process. Syst. Conf.*, 2007, pp. 841–848.

[51] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8174–8182.

[52] L. Li *et al.*, "Learning a discriminative prior for blind image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6616–6625.

[53] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8183–8192.

[54] S. Bae and F. Durand, "Defocus magnification," in *Proc. Eurographics*, 2007, pp. 571–579.

[55] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[56] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 990–998.

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

[58] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.

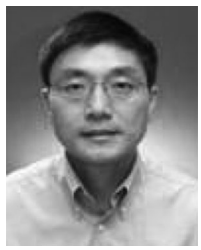
[59] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection*. Berlin, germany: Springer, 2006, pp. 34–51.

[60] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," 2000.

[61] H. Liu, J. Koonen, M. Fuderer, and I. Heynderickx, "The relative impact of ghosting and noise on the perceived quality of MR images," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3087–3098, Jul. 2016.

Dingquan Li received dual B.S. degrees in electronic science and technology and applied mathematics from Nankai University, Tianjin, China, in 2015. He is currently working toward the Ph.D. degree in applied mathematics with Peking University, Beijing, China. His research interests include image/video quality assessment, perceptual optimization, and machine learning. Dr. Li is a member of the National Engineering Lab for Video Technology.

Tingting Jiang received the B.S. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2001, and the Ph.D. degree in computer science from Duke University, Durham, NC, USA, in 2007. She is currently an Associate Professor of computer science with Peking University, Beijing, China. Her research interests include computer vision and image/video quality assessment.



Weisi Lin (M'92–SM'98–F'16) received the B.Sc. degree in electronics and the M.Sc. degree in digital signal processing from Sun Yat-Sen University, Guangzhou, China, in 1982 and 1985, respectively, and the Ph.D. degree in computer vision from King's College, London University, London, U.K., in 1993. He was with Sun Yat-Sen University, Shantou University, Shantou, China; Bath University, Bath, U.K.; the National University of Singapore, the Institute of Microelectronics, Singapore; and the Institute for Infocomm Research, Singapore. He has been the Project Leader of more than ten major successful projects in digital multimedia technology development. He was the Laboratory Head of Visual Processing and the Acting Department Manager of Media Processing with the Institute for Infocomm Research. He is currently a Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include image processing, perceptual signal modeling, video compression, multimedia communication, and computer vision.

Ming Jiang received the B.Sc. and Ph.D. degrees in mathematics from Peking University, Beijing, China, in 1984 and 1989, respectively. He has been a Professor with the Department of Information Science, School of Mathematical Science, Peking University, since 2002. He has been the Managing Director of the Microsoft Statistics and Information Technology Laboratory, Peking University, since 2005. His research interests include mathematical and technical innovations in biomedical imaging and image processing, with X-ray computed tomography, optical molecular tomography, and multimodality biomedical imaging as the main focus.