LEARNING DEEP VECTOR REGRESSION MODEL FOR NO-REFERENCE IMAGE QUALITY ASSESSMENT

Jie Gu, Gaofeng Meng, Lingfeng Wang and Chunhong Pan

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
{jie.gu, gfmeng, lfwang and chpan}@nlpr.ia.ac.cn

ABSTRACT

The goal of no-reference image quality assessment (NR-IQA) is to estimate human perceived image quality without access to either reference image or prior knowledge about distortion type. Previous approaches for this problem are typically based on a regression framework that maps the image features directly to a quality score. In contrast, psychological evidence shows that humans prefer to evaluate visual quality with qualitative descriptions, *e.g.*, using a five-grade ordinal scale: "excellent", "good", "fair", "poor" and "bad". Based on this observation, we propose a vector regression model that predicts five belief scores rather than a single quality score. The belief scores are designed to indicate the confidences of the test image being assigned with these five quality grades. In addition, with the purpose of more extensive applications, a saliency-based pooling strategy is presented to convert the predicted confidences into objective quality scores. Extensive experiments performed on two benchmark datasets demonstrate that our approach achieves state-of-the-art performance and shows great generalization ability.

***Index Terms*—** Image quality assessment, perceptual image quality, CNN, vector regression, saliency-based pooling

1. INTRODUCTION

With the increasing popularity of digital imaging and multimedia technology, image quality assessment (IQA) has become an important issue in a variety of applications. Since subjective assessment can not be routinely used in many situations for the well-known drawbacks (*e.g.*, the time-consuming and expensive data-gathering process), objective IQA methods have attracted growing attention in recent years. Generally, in addition to being used in practical applications, objective IQA is also considered as an attempt to discover how the human visual system (HVS) perceives images.

According to the accessibility of an original (reference) image, existing IQA algorithms can be roughly classified into three categories, *i.e.*, full-reference (FR) IQA, no-reference/blind (NR) IQA and reduced-reference (RR) IQA. Among them, FR-IQA metrics estimate visual quality by comparing the differences between the reference image and its distorted version. Recent FR-IQA methods, such as VIF [1] and FSIM [2], have achieved a high correlation with subjective assessment. In contrast, without access to a reference image, NR-IQA methods attempt to directly quantify image distortions by exploiting discriminative features. In this paper, we focus on the most challenging IQA problem: general-purpose NR-IQA, which predicts human perceived image quality without information from reference image and prior knowledge about distortion types.

In the past decades, many efforts have been made on designing machine learning based NR-IQA algorithms [3–11]. The main idea is to learn a regression model that maps the extracted features to the image quality score. Typically, natural scene statistics (NSS) is widely used to extract discriminative features in most traditional NR-IQA approaches [3–5]. They are based on the hypothesis that some natural scene properties will be influenced by the presence of distortion. However, NSS-based features need to be designed carefully and empirically in some particular transform domains. In order to acquire general representations from raw images, recent studies pay more attention on feature learning methods. For example, unsupervised learning has been explored in SRNSS [6] and CORNIA [7]. The former obtains the image representations in a frame of sparse coding, and the latter encodes the patches via soft-assignment and max pooling. Some other previous works try to learn informative representations in a supervised manner. Particularly, with the advances of deep learning, it has been increasingly concerned about to apply deep neural network to NR-IQA. Tang *et al.* [8] proposed a method which combines a deep belief network (DBN) and a Gaussian process. Kang *et al.* [9] described a general framework for NR-IQA based on convolutional neural network (CNN). In their later work [10], a multi-task CNN is designed for simultaneously estimating image quality and identifying distortion type.

Unlike the above regression-based approaches, Hou *et al.* [11] proposed a classification-based framework via DBN. The underlying principle of their work is that humans prefer to conduct evaluations qualitatively, typically using a five-grade ordinal scale: "excellent", "good", "fair", "poor" and "bad". However, since qualitative descriptions have been mapped to quantitative scores when constructing datasets (*e.g.*, LIVE [12]), it is still a difficult problem to learn such a five-grade classification model directly.

In this paper, we introduce Deep Vector Regression Model (DVRM), an effective NR-IQA method to address this problem. Rather than training a five-category classifier, we prefer to learn a regression network with five output variables. The network aims to predict five belief scores, which are designed to indicate the confidences of an input patch being assigned with the five corresponding grades. At test stage, we adapt the network into fully convolutional network and then obtain five belief score maps. The output score maps can be directly applied to the qualitative evaluation on test image. Moreover, since quantitative estimation is more widely used in practical applications, a saliency-based pooling strategy is presented to compute the objective quality scores.

Our method benefits from two main contributions. The first is a vector regression framework for NR-IQA. Different from the existing regression methods, our framework yields a vector of five belief scores to measure the corresponding five qualitative grades. Exper-

This work was supported in part by the National Natural Science Foundation of China (Grants No. 61370039, 91646207, 61620106003, and 61403376), and the Beijing Nature Science Foundation (Grant No. 4162064).

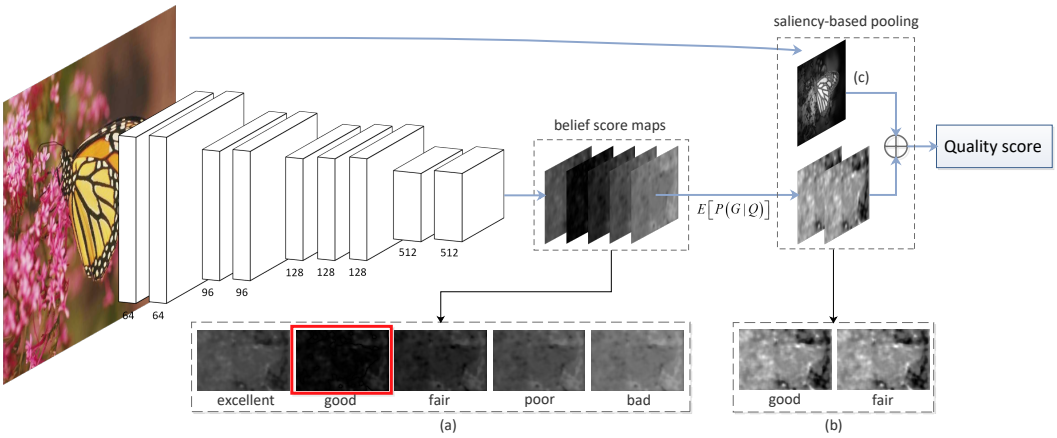

Fig. 1. Pipeline of our approach. (a) Defined belief score maps and their corresponding quality grades. Since the second map has the smallest mean absolute value, the quality of the test image can be described as "good" (shown in a red box); (b) Local score maps in Eq. (3). In this case, "good" and "fair" are selected for global score pooling; (c) Saliency map produced by SDSP [13], which is used as a weighting function to characterize the visual importance of local regions.

imental results show that this framework works consistently better than the state-of-the-art NR-IQA methods. The second contribution is a pooling strategy that employs a saliency map as weighting function to compute the global objective quality score. The saliency map incorporates the semantic information of image contents into quality assessment, thus further improving the evaluation accuracy.

2. THE PROPOSED APPROACH

Our approach consists of two processes, *i.e.*, belief score estimation based on vector regression and global score pooling based on saliency. The pipeline of our approach is shown in Fig. 1.

2.1. Vector Regression

Psychological evidence has shown that humans prefer to evaluate visual quality qualitatively, typically using a five-grade ordinal scale: "excellent", "good", "fair", "poor" and "bad" [11]. However, due to the individual diversity in perception, different people may classify the same image into different grades. To characterize this diversity, we introduce a quality distribution for each grade. The quality distribution of k-th grade is denoted by $P(G = k|Q)$, where $k = 1, 2, \ldots, 5$, corresponding to the five grades, Q and G are random variables that represent intrinsic score and grade respectively.

The main idea of previous works is to learn a single regression model that maps the features directly to a quality score. In our study, notice that qualitative descriptions are preferred by humans, we thereby predict the confidences of a given image being labeled as the five explicit grades (*i.e.*, "excellent", "good", "fair", "poor" and "bad"). To this end, a vector regression model is proposed that estimates a vector of five belief scores instead of a single quality score. For an input image $\mathbf{x}$, the belief scores are defined as the distances from the ground truth to the expectations of quality distributions

$$s_k = y - \mu_k \quad (k = 1, 2, \ldots, 5) \,, \tag{1}$$

where μ_k denotes the expectation of $P(G = k|Q)$, y is the ground truth score. An illustration for the belief score is provided in Fig. 2. Physically, the distance $|s_k|$ provides a natural way to measure the confidence that the image $\mathbf{x}$ belongs to the k-th grade.

In practice, we split the benchmark image dataset (*e.g.*, LIVE) into five groups according to the provided mean opinion scores. The principle of the partition is that the number of samples should be uniform across the five groups. Then, μ_k can be estimated as the mean score of k-th group.

In our work, the vector regression model is implemented by a deep convolutional network. A saliency-based pooling strategy is further used to convert the predicted belief scores into objective quality score. The details will be discussed later in Subsection 2.2 and Subsection 2.3.

2.2. Patch-wise Belief Score Mapping using CNN

We use a CNN to implement the belief score regression. To train the deep network with limited labeled data, we divide image into patches and feed them to the network. Each patch is assigned a 5-D vector of belief score, which can be computed from the ground truth score of its source image. The details can be found below.

Local Normalization. The preprocessing step employs a local contrast normalization method as in [5, 9]. Unlike common global normalization, it applies a local operation to compute the mean and variance on each pixel individually. Note that the normalized image exhibits a largely homogeneous appearance [5], which will benefit the subsequent training process.

Training Data. Labeling subjective scores for a large scale image dataset is generally impractical. To obtain sufficient samples for training a deep model, the following data augmentation method is used. We collect samples by cropping non-overlapping 32×32 patches and assign each patch a quality score as the ground truth of its source image.

Network Configurations. The deep network is built by stacking 3×3 convolutional layers as shown in Fig. 1. Max-pooling is applied to some of the convolutional layers over a 3×3 pixel window with stride 2. During training, the last pooling layer is followed by three fully connected layers: the first two have 512 nodes each, the third performs five-way regression and thus contains five nodes. In addition, Rectified Linear Unit (ReLU) non-linearity [14] is equipped after all hidden layers, and Dropout [15] is used after the second fully connected layer to avoid overfitting.

In our framework, the loss function can be expressed as

$$L(\mathbf{W}, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^{N} \| f_{\mathbf{W},\mathbf{b}}(\mathbf{x}_i) - \mathbf{s}_i \|_2^2 + \lambda \|\mathbf{W}\|_F^2 \,, \tag{2}$$

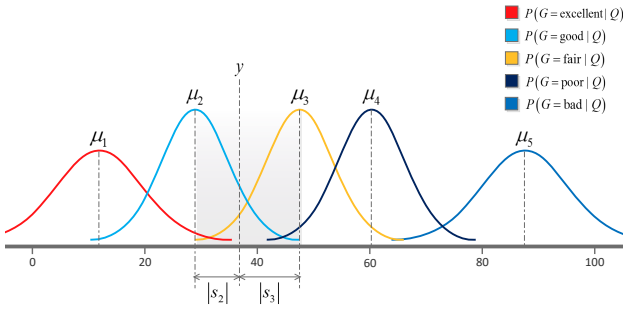

Fig. 2. Illustration of the belief scores and the five quality distributions. μ_k, $(k = 1, 2, \ldots, 5)$ denotes the expectation that can be estimated from benchmark dataset. s_k, $(k = 1, 2, \ldots, 5)$ denotes the defined belief score. y is the ground truth score of test image. The smaller the $|s_k|$ is, the more likely the quality of test image can be described as the k-th grade. In this case, "good" and "fair" are the most relative grades to the ground truth score.

where $\mathbf{x}_i$ and $\mathbf{s}_i$ denote the sampled patch and the vector of belief score, $f_{\mathbf{W},\mathbf{b}}(\mathbf{x}_i)$ is the output of the deep model, $\mathbf{W}$ and $\mathbf{b}$ represent the weights and bias of the network respectively, and λ is the hyper-parameter. We employ stochastic gradient descent (SGD) to optimize the loss function, with the momentum set to 0.9 and the batch size set to 256. The learning rate is initially set to 0.01 and then decreased by a factor of 0.1 at regular intervals. In total, the learning is stopped after about 100 epoches, and the learning rate is decreased 3 times. Moreover, the weight decay is set to 0.0005 and the dropout ratio is set to 0.5.

2.3. Saliency-based Image Quality Pooling

At testing stage, we transform the fully connected layers into convolutional layers [16]. The adapted fully convolutional network can take the entire locally normalized image as input and outputs five belief score maps, denoted by $\mathbf{S}_k$, $k = 1, 2, \ldots, 5$.

Following the analysis above, the belief score maps can be directly used to evaluate image quality qualitatively. In specific, the mean absolute value of $\mathbf{S}_k$ points out the confidence of the given image being labeled as the k-th grade. Therefore, the image quality can be simply described as the grade corresponding to the map with the smallest mean absolute value, as shown in Fig. 1(a).

Though ordinal-scale description is considered as a natural way of evaluating image quality, quantitative estimation is more widely used in real-world applications. In our framework, a saliency-based pooling strategy is designed to obtain the objective quality score s_{img}. First, two belief score maps with the smallest mean absolute values are picked out to compute the local score maps

$$\mathbf{L}_k = \mathbf{S}_k \oplus \mu_k \quad (k = p, p+1)\,, \tag{3}$$

where $\oplus$ represents element-wise addition, p and $p+1$ are the indices of the selected belief score maps. As illustrated in Fig. 2, the selected two grades are the most relative ones to the intrinsic score of test image. After that, given that humans always fixate some particular areas of an image [17], saliency is employed as a weighting function to integrate the local quality scores

$$\begin{aligned} s_{img} &= \frac{1}{2} \cdot \sum_{i,j} (\mathbf{L}_p + \mathbf{L}_{p+1}) \odot \mathbf{F} \\ \mathbf{F} &= \frac{\mathbf{M}}{\sum_{i,j} \mathbf{M}(i,j)} \end{aligned}\,, \tag{4}$$

where $\odot$ represents element-wise product, i and j indicate the location of the maps, $\mathbf{M}$ is the resized saliency map. In our experiment, a simple yet effective saliency method SDSP [13] is employed in Eq.(4). An example for the local score map and the saliency map can be found in Fig. 1(b) and Fig. 1(c), respectively.

3. EXPERIMENTS

3.1. Experimental Protocol

Dataset description. We test the performance of our approach on two large-scale benchmark datasets: LIVE [12] and TID2008 [18].

1. LIVE: It contains 29 reference images associated with their distorted versions with five types of distortions-JPEG2000 compression (JP2K), JPEG compression (JPEG), white noise (WN), Gaussian blur (BLUR) and fast fading (FF), at 7-8 degradation levels. Each image is given an associated Differential Mean Opinion Score (DMOS), generally in the range $[0, 100]$.

2. TID2008: A total of 1700 degraded images are included in TID2008 (25 reference images, 17 types of distortions for each reference image, 4 degradation levels for each distortion type). A Mean Opinion Score (MOS) is provided for each image ranging from 0 to 9. We conduct the experiments only on 13 distortions: Additive Gaussian noise (WN), Different additive noise in color components (WNC), Spatially correlated noise (SCN), Masked noise (MN), High frequency noise (HFN), Impulse noise (IN), Quantization noise (QN), Gaussian blur (BLUR), Image denoising (IDN), JPEG compression (JPEG), JPEG2000 compression (JP2K), JPEG transmission errors (JPEGTE) and JPEG2000 transmission errors (JP2KTE). The other 4 distortions are not included as they are either strongly heterogeneous or highly subjective for NR-IQA.

Evaluation criteria. We use two common criteria for the evaluations: Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC). They are adopted here to measure the linear dependence and monotonicity between the subjective scores and predicted scores, respectively.

3.2. Evaluation on LIVE

In this section, the performance of our approach on the LIVE dataset is provided. In order to compare with other methods properly, the same experimental strategy in [11] is adopted. Specifically, we group the images in LIVE according to their reference images, and randomly select 23 groups for training, retaining the remaining 6 groups for testing. To avoid the effects of random selection, all the results are reported as the medians of 20 train-test iterations. Note that only degraded images are included in the testing set.

Table 1 shows the experimental results, where the best one is highlighted in boldface. We compare the proposed method with three benchmark FR-IQA methods, PSNR, SSIM [19] and VIF [1], and six representative NR-IQA methods, BLIINDS-II [4], DIIVINE [3], SRNSS [6], BRISQUE [5], CORNIA [7] and DLIQA [11]. The results of all these methods are taken from [11]. Different pooling strategies are also compared: DVRM_A denotes the baseline method implemented by averaging the local score maps in Eq.(3); DVRM_S denotes the proposed saliency-based method.

It can be seen that our approach is well applicable for objective evaluation. As shown in the last column of Table 1, both DVRM_A and DVRM_S outperform all the other state-of-the-art NR-IQA methods on the entire database, which indicates that our approach is robust to the influence of distortion types. Furthermore, one can see that the proposed method works consistently well on all the individual distortion types, especially on WN, BLUR and FF.

Table 1. Performance comparison on the specific distortion types of LIVE and the entire LIVE dataset. DVRM_A: baseline method, implemented by directly averageing the local score maps in Eq. (3); DVRM_S: saliency-based method.

SROCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.890	0.841	0.985	0.782	0.890	0.820
SSIM	0.932	0.903	0.963	0.894	0.941	0.851
VIF	0.953	0.913	0.986	0.973	0.965	0.953
BLIINDS-II	**0.951**	**0.942**	0.978	0.944	0.927	0.920
DIIVINE	0.913	0.910	0.984	0.921	0.863	0.916
SRNSS	0.928	0.931	0.938	0.933	0.941	0.930
BRISQUE	0.910	0.919	0.955	0.941	0.874	0.920
CORNIA	0.903	0.889	0.958	0.946	0.915	0.906
DLIQA	0.933	0.914	0.968	0.947	0.857	0.929
DVRM_A	0.937	0.915	0.988	0.950	0.955	0.932
DVRM_S	0.945	0.909	**0.988**	**0.968**	**0.956**	**0.937**

LCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.896	0.860	0.986	0.783	0.890	0.824
SSIM	0.937	0.928	0.970	0.874	0.943	0.863
VIF	0.962	0.943	0.984	0.974	0.962	0.950
BLIINDS-II	**0.963**	**0.979**	0.985	0.948	0.944	0.923
DIIVINE	0.922	0.921	**0.988**	0.923	0.888	0.917
SRNSS	0.936	0.939	0.940	0.936	0.947	0.932
BRISQUE	0.936	0.937	0.958	0.935	0.898	0.917
CORNIA	0.915	0.902	0.952	0.940	0.913	0.903
DLIQA	0.953	0.948	0.961	**0.950**	0.892	0.934
DVRM_A	0.923	0.956	0.984	0.942	0.970	0.935
DVRM_S	0.935	0.954	0.985	0.949	**0.974**	**0.942**

Table 2. Performance comparison on the entire TID2008 dataset. Refer to Table 1 for notations.

	SSIM	CORNIA	CNN++	DVRM_A	DVRM_S
SROCC	0.878	0.813	0.870	0.903	**0.916**
LCC	0.857	0.837	0.880	0.890	**0.904**

3.3. Evaluation on TID2008

To further evaluate the proposed method, we conduct experiments on TID2008 in this section, which is notoriously more difficult than LIVE. Similarly, all the degraded images in TID2008 are classified into 25 groups according to the reference images. The results are computed as the medians across 20 train-test iterations, where 20 groups are randomly chosen as the training set and the remaining 5 groups as the testing set. It is noteworthy that few NR-IQA methods report available results on this challenging dataset.

In Table 2, two state-of-the-art NR-IQA methods and one benchmark FR-IQA method are listed for comparison. The results, including SSIM, CORNIA and CNN++, are all taken from [10]. It can be seen that both DVRM_A and DVRM_S perform greatly better than the other two competitors on this large database. Particularly, even without reference image, our approach still outperforms the FR-IQA method SSIM.

3.4. Cross Dataset Evaluation

In this section, cross dateset evaluation is designed to verify the generalization ability of our approach. To this end, the model is trained on LIVE and then tested on TID2008. We compare our approach against two NR-IQA methods (*i.e.*, CORNIA and BRISQUE), since their codes are available online. Moreover, as the range of DMOS in LIVE is different from that of MOS in TID2008, a non-linear mapping needs to be performed before computing the LCC metric, as suggested by [12]. In summary, we apply the non-linear mapping on the predicted scores and compute the evaluation criteria as the way of evaluating FR measures.

Table 3. SROCC and LCC obtained by training on LIVE and testing on TID2008. Only four types of distortions shared by LIVE and TID2008 are included. Refer to Table 1 for notations.

SROCC	JP2K	JPEG	BLUR	WN	ALL
BRISQUE	0.902	0.875	0.857	0.821	0.865
CORNIA	0.920	0.899	**0.901**	0.647	0.886
DVRM_A	0.922	0.926	0.764	**0.920**	0.885
DVRM_S	**0.943**	**0.930**	0.773	0.909	**0.894**

LCC	JP2K	JPEG	BLUR	WN	ALL
BRISQUE	0.908	0.909	0.855	0.812	0.873
CORNIA	0.904	0.928	**0.887**	0.642	0.878
DVRM_A	0.893	0.940	0.818	**0.925**	0.896
DVRM_S	**0.945**	**0.942**	0.801	0.919	**0.911**

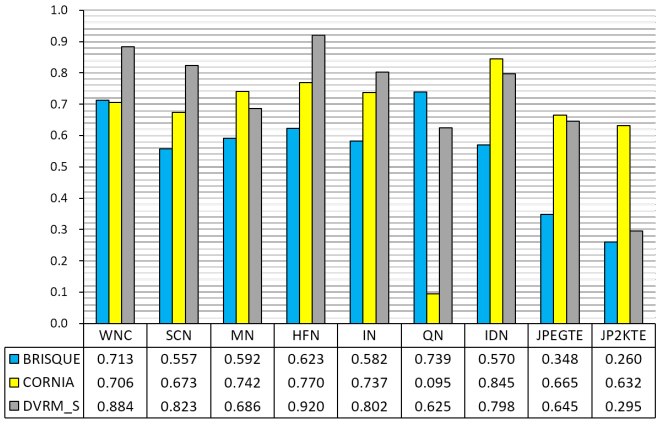

Fig. 3. SROCC obtained by training on LIVE and testing on TID2008. The horizontal axis represents the distortion types and the vertical axis represents the SROCC metric. Notice that these types of distortions are not included in the training dataset.

For better illustration, the distortion types in TID2008 are divided into two categories, according to whether they are shared by LIVE and TID2008. Table 3 shows the experimental results on the four common distortions, namely JP2K, JPEG, BLUR and WN. It can be observed that the proposed method outperforms the other two competitors in both distortion-specific and overall experiments. In addition, Fig. 3 shows the SROCC metric on the other 9 distortions. One can see that even without training on these types of distortions, our approach still achieves convincing results in some cases, such as WNC, SCN, HFN, IN and IDN.

4. CONCLUSION

In this paper, we have developed a vector regression model for NR-IQA problem by combing two processes: belief score estimation and saliency-based quality pooling. Our algorithm is based on the observation that humans prefer to conduct evaluations with qualitative descriptions rather than exact scores. The proposed method has shown great performance and generalization ability in comparison with current state-of-the-art NR-IQA approaches.

5. REFERENCES

[1] Hamid R. Sheikh and Alan C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[2] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[3] Anush K. Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[4] Michele A. Saad, Alan C. Bovik, and Christophe Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[5] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[6] Lihuo He, Dacheng Tao, Xuelong Li, and Xinbo Gao, "Sparse representation for blind image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1146–1153.

[7] Peng Ye, Jayant Kumar, Le Kang, and David S. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[8] Huixuan Tang, Neel Joshi, and Ashish Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2877–2884.

[9] Le Kang, Peng Ye, Yi Li, and David S. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[10] Le Kang, Peng Ye, Yi Li, and David S. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *IEEE International Conference on Image Processing*, 2015, pp. 2791–2795.

[11] Weilong Hou, Xinbo Gao, Dacheng Tao, and Xuelong Li, "Blind image quality assessment via deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1275–1286, 2015.

[12] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[13] Lin Zhang, Zhongyi Gu, and Hongyu Li, "SDSP: A novel saliency detection method by combining simple priors," in *IEEE International Conference on Image Processing*, 2013, pp. 171–175.

[14] Vinod Nair and Geoffrey E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[15] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[17] Ulrich Engelke, Hagen Kaprykowsky, Hans-Jürgen Zepernick, and Patrick Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 50–59, 2011.

[18] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Jaakko Astola, Marco Carli, and Federica Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[19] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.