Blind Quality Assessment for Image Superresolution Using Deep Two-Stream Convolutional Networks

Wei Zhou, Qiuping Jiang, Yuwang Wang, Zhibo Chen, Weiping Li

Abstract

Numerous image superresolution (SR) algorithms have been proposed for reconstructing high-resolution (HR) images from input images with lower spatial resolutions. However, effectively evaluating the perceptual quality of SR images remains a challenging research problem. In this paper, we propose a no-reference/blind deep neural network-based SR image quality assessor (DeepSRQ). To learn more discriminative feature representations of various distorted SR images, the proposed DeepSRQ is a two-stream convolutional network including two subcomponents for distorted structure and texture SR images. Different from traditional image distortions, the artifacts of SR images cause both image structure and texture quality degradation. Therefore, we choose the two-stream scheme that captures different properties of SR inputs instead of directly learning features from one image stream. Considering the human visual system (HVS) characteristics, the structure stream focuses on extracting features in structural degradations, while the texture stream focuses on the change in textural distributions. In addition, to augment the training data and ensure the category balance, we propose a stride-based adaptive cropping approach for further improvement. Experimental results on three publicly available SR image quality databases demonstrate the effectiveness and generalization ability of our proposed DeepSRQ method compared with state-of-the-art image quality assessment algorithms.

Keywords: Image superresolution, blind image quality assessment, two-stream convolutional networks, stride-based adaptive cropping, human vision

1. Introduction

Image superresolution (SR) aims at constructing a high-resolution (HR) image with fine details using one or several low-resolution (LR) images as inputs [29], which is desirable in various practical scenarios, such as medical imaging, video surveillance, and high-definition television (HDTV) [23]. Through SR technologies, people can better view LR images on HR displays. During the past few decades, many generic single image superresolution algorithms have been proposed [40]. However, much less has been done to fairly evaluate the perceptual quality of superresolved images (SRIs) and the performance of SR algorithms [2].

For image superresolution quality assessment, in the literature, small-scale subjective experiments are usually used for evaluation. Specifically, subjects are asked to rate the visual quality of SR images generated by different SR algorithms. The rating provided for each SR image under examination is termed the opinion score. The mean of these ratings, i.e., the mean opinion score (MOS), is calculated as the ground-truth image quality measurement, which is a common practice in quality assessment. Thus, although some viewers may have different feelings, the MOS score is a statistical concept. Subjective tests were performed in [41] to explore the visually subjective quality of SR images. Such subjective testing is reliable for providing a fair evaluation of SR image quality but is expensive, labor intensive, and time consuming. More importantly, subjective tests cannot be integrated into the automatic design process of perception-driven SR algorithms. Therefore, it is desirable to develop objective IQA methods for automatically predicting the subjective visual quality of SR images.

When the original distortion-free image is available, full-reference image quality assessment (FR-IQA) can be carried out by comparing the distorted image with the reference image. Two classic FR-IQA metrics, namely, the peak signal-to-noise-ratio (PSNR) and the structural similarity index (SSIM) [33], are usually employed to evaluate the visual perceptual quality of reconstructed SR images. However, they do not match well with a subjective evaluation. Three publicly available SR image quality databases considering several typical SR algorithms have been built [17, 32, 48], upon which state-of-the-art IQA metrics are tested. The results demonstrate that it is difficult

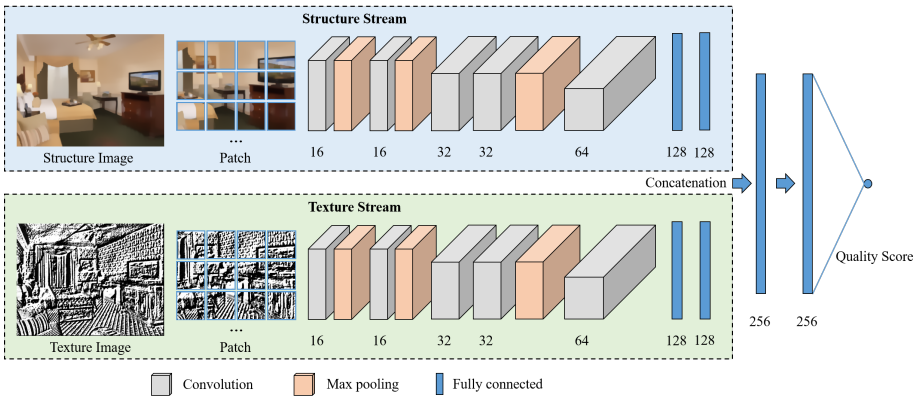


Figure 1: Overview of the network architecture of our proposed model for learning the perceptual quality of image superresolution. In DeepSRQ, representative features are extracted from distorted structure and texture SR images by two subnetworks. The extracted features are then used to estimate the quality scores by late fusion regression layers.

for the existing IQA metrics to effectively predict the perceptual quality of SR images. In addition, reference images seldom exist in most situations, and these FR-IQA metrics require the resolution of the distorted image to be the same as that of the original image. As a result, a no-reference/blind image quality assessment (NR-IQA) metric specifically designed for image superresolution, which evaluates perceptual quality with no access to reference LR or HR images, is directly applicable and highly demanded.

In recent years, deep learning-related methods have been extensively used and have achieved great achievements for a variety of image processing and computer vision problems, such as image recognition [8], video quality assessment [49], and social image understanding [15]. In this paper, we exploit deep convolutional neural networks (DCNNs) to address the no-reference image superresolution quality assessment problem. Specifically, we propose a deep neural network-based SR image quality assessor (DeepSRQ). Inspired by [48], the proposed model shown in Figure 1 contains two subcomponents accounting for structure and texture SR images. To the best of our knowledge, for objective SR IQA models, this is the first study of applying the two-stream network architecture to learn the perceptual SR image quality. The human

visual system (HVS) is not only sensitive to structural information but also considers textural details [1]. Moreover, the distortions of SR images are different from conventional artifacts, which degenerate both image structure and texture [48]. Thus, we first extract structure and texture images from distorted SR images. Second, since existing image superresolution quality databases (i.e., the superresolution (SR) quality database [17], superresolution reconstructed image database (SRID) [32]) and quality assessment database for SRIs (QADS)) are relatively small-scale, and we opt to utilize a stride-based adaptive cropping approach to augment the training data and ensure the category balance. This approach also considers the local visual information of the whole distorted structure and texture SR images. Third, we take the distorted structure and texture SR patches as inputs and train the two-stream convolutional network to extract the discriminative feature representations. Fourth, different from the original classification-based DCNN [13], we aim to map the feature representations to estimated scores, which adopts a fully connected layer instead of a softmax layer. Specifically, two fully connected layers are followed by each substream. Then, we use concatenation to further obtain one quality score for each SR image patch. Finally, we average these estimated scores as the perceptual quality for the entire SR image.

In addition, the training process of DeepSRQ produces both low-level visual information and high-level semantic features. Our experimental results show that the synthetically learned features are more effective than both the handcrafted low-level features and high-level semantic features extracted from pretrained DCNN models. Each substream is verified in our experiments. We also conducted an ablation study to demonstrate that the proposed stride-based adaptive cropping approach indeed plays a critical role in DeepSRQ.

The contributions of this work are summarized as follows:

- Since the distortions of SR images cause both image structure and texture quality degradation, we propose a two-stream deep convolutional network for the blind quality estimation of superresolution images. The proposed network extracts the discriminative features from various distorted structure and texture SR images, where each subnetwork adapts and differs from the classification-based

architecture.

- To ensure the category balance, we propose a stride-based adaptive cropping approach for augmenting the training data. We show the effectiveness of the proposed adaptive cropping method for further improving the performance of the whole framework.

- We conduct extensive experiments on various databases demonstrating the effectiveness of the proposed network and the corresponding adopted techniques. In addition, the synthetic features learned from the proposed two-stream network are more effective than traditional features.

The remainder of this paper is organized as follows. In Section II, we introduce the proposed deep neural network-based SR image quality assessor (DeepSRQ) for no-reference/blind superresolution image quality prediction in detail. Section III presents the experimental results and analysis. In Section IV, we conclude the paper and discuss future research directions.

2. Proposed DeepSRQ

The network architecture of the proposed blind quality assessor for learning the perceptual quality of image superresolution is shown in Figure 1. We design our proposed DeepSRQ network with a two-stream architecture, which takes distorted structure and texture SR images as inputs through two subnetworks, i.e., structure stream and texture stream. Each subnetwork adapts and differs from AlexNet [13]. It replaces the last softmax layer with a fully connected layer for the regression task. We thus adopt MSE loss rather than cross-entropy loss. The whole network design, such as kernel number and network parameters, is also different. Since existing image superresolution quality databases are relatively small-scale, we opt to augment the training set by cropping the distorted structure and texture SR images into multiple patches. In addition, to ensure the category balance, we propose a stride-based adaptive cropping operation in the preprocessing stage for further boosting the performance of our DeepSRQ method. Next,

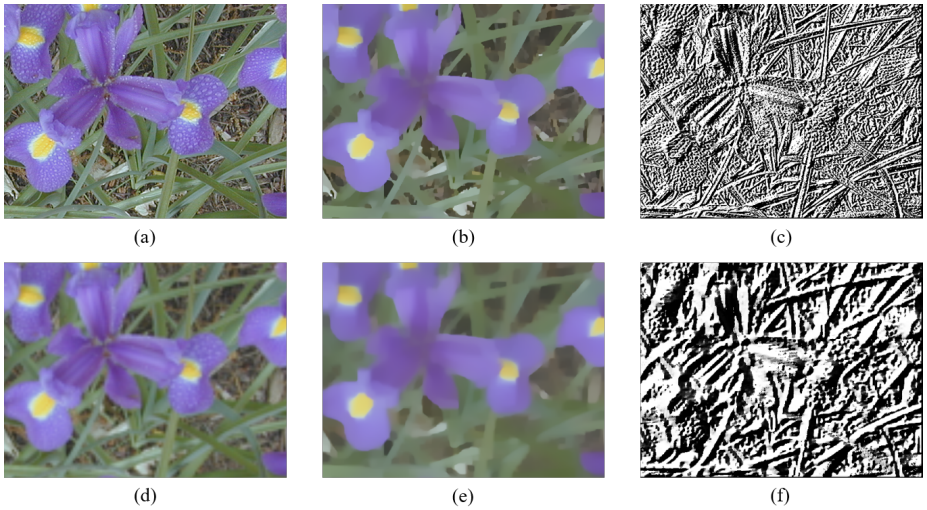


Figure 2: Examples of natural images and the corresponding structure as well as texture images in the QADS database [48]. (a) Original HR image with higher visual quality. (b) Extracted structure image of (a). (c) Extracted texture image of (a). (d) Distorted SR image with lower visual quality. (e) Extracted structure image of (d). (f) Extracted texture image of (d).

we present the details of our proposed DeepSRQ. The training and quality prediction steps are also presented.

2.1. Image Representation

As demonstrated in previous studies, the distortions of SR images cause both image structure and texture degradation [48]; we first extract structure and texture images from distorted SR images. Specifically, we adopt the relative total variation-based structure extraction method [38] with default parameters to extract structure images from distorted SR images. From the perspective of perception, the texture map obtained directly by subtracting the structure map from the original image is not necessarily the texture perceived by human eyes. The texture map obtained by the texture descriptors is more consistent with human perception. For texture description, the local binary pattern (LBP) is an effective texture description operator and has significant advantages of rotation invariance and gray invariance [22]. Moreover, the LBP is widely used in image quality evaluation tasks [50]. Therefore, we utilize it to extract texture images

from distorted SR images. Some examples of natural images and the corresponding structure as well as texture images in the QADS database [48] can be seen in Figure 2. We observe that the structure and texture images can discriminate images with different visual qualities, which demonstrates the effectiveness of the extracted structure and texture images.

To train the two-stream deep neural network, a large quantity of training data is needed. Moreover, we need to use the same fixed sizes during training and testing because fully connected layers exist in the proposed network, and their number of parameters is not flexible. Since the existing publicly available image superresolution quality databases are small-scale, cropping SR images is an effective method for increasing the quantity of training data. Moreover, compared with resizing, cropping ensures that the perceptual image quality is unchanged [12]. Hence, we choose to crop multiple patches from different spatial locations to cover the local visual information of the whole SR image without introducing any geometric deformation. The resolution of SR images in the SR quality database [17] and QADS database [48] is fixed for different scaling factors and the corresponding Gaussian kernel widths. Therefore, the total number of nonoverlapping cropped patches for each SR image is given by:

$$Num_p = \left\lfloor \frac{M}{m} \right\rfloor \times \left\lfloor \frac{N}{n} \right\rfloor, \tag{1}$$

where M and m are the SR image width and cropped patch width, respectively, while N and n are the SR image height and cropped patch height, respectively. In other words, $M \times N$ is the SR image resolution, and $m \times n$ denotes the cropped patch resolution. Note that $M > m$, $N > n$, and $m = n$ in our experiments.

2.2. Stride-Based Adaptive Cropping Method

It should be noted that there exist three amplification factors (i.e., 2, 4, and 8) in the SRID [32]. A larger amplification factor leads to a higher resolution of SR images. Therefore, we propose a stride-based adaptive cropping approach to ensure the category balance for different amplification factors. Specifically, the stride of the maximum amplification factor (i.e., $f_{\max}$) is the cropped patch size denoted by m.

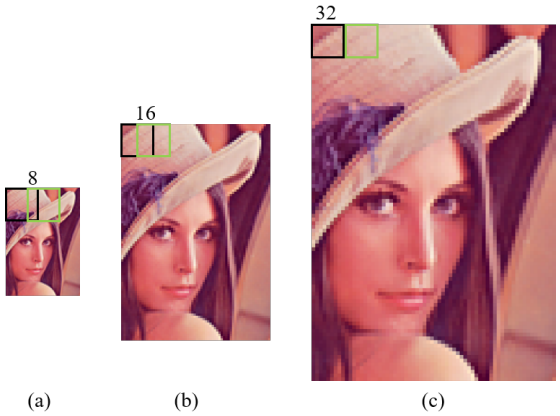


Figure 3: Illustration of the proposed stride-based adaptive cropping method with several SR image examples, where 8, 16, and 32 denote the stride sizes for SR images with amplification factors 2, 4, and 8, respectively. The cropping patch size is 32×32. (a) SR image with amplification factors equal to 2. (b) SR image with amplification factors equal to 4. (c) SR image with amplification factors equal to 8.

Then, the strides for the other amplification factors (i.e., f) are computed as:

$$Stride_f = \frac{f}{f_{\max}} \times m. \tag{2}$$

Specifically, as shown in Figure 3, we illustrate several SR image examples with three amplification factors. As the cropping patch size is 32×32, the stride sizes are 8, 16, and 32 for SR images with amplification factors equal to 2, 4, and 8, respectively. By using the proposed stride-based adaptive cropping approach, the quantity of training data for each amplification factor generally remains balanced. Furthermore, our ablation experiments show that this adaptive cropping approach brings about improved performance.

2.3. Network Architecture

Given a distorted SR image I, we first extract the structure and texture images from it. Then, we can represent the structure and texture images by a set of cropped patches. Let $\{s_1, s_2, ..., s_{Num_p}\}$ and $\{t_1, t_2, ..., t_{Num_p}\}$ be the cropped structure and texture patches used to train the DCNN. As shown in Figure 1, we first take distorted structure and texture SR image patches as inputs. Following local normalization [42], we rescale

Table 1: Detailed subnetwork configurations of the proposed method. CONV: convolution, ELU: activation, POOL: max pooling, DENSE: fully connected, DROP: dropout.

Layer	Output Shape	Parameter Number
INPUT	(32, 32, 3)	0
CONV1+ELU1	(32, 32, 16)	448
POOL1	(16, 16, 16)	0
CONV2+ELU2	(16, 16, 16)	2320
POOL2	(8, 8, 16)	0
CONV3+ELU3	(8, 8, 32)	4640
CONV4+ELU4	(8, 8, 32)	9248
CONV5+ELU5	(8, 8, 64)	18496
POOL5	(4, 4, 64)	0
DENSE1+ELU6+DROP1	128	131200
DENSE2+ELU7+DROP2	128	16512
DENSE3	1	129

each patch to the range [0, 1] by dividing all channels by 255 before feeding it into the two-stream network. The last fully connected layer of each subnetwork is concatenated to obtain a $256 - dim$ feature vector, and then two fully connected layers with sizes equal to 256 and 1 are employed to regress the inputs onto one single visual quality score.

Moreover, we show the ablation substream of our proposed two-stream network. The detailed subnetwork configurations of the proposed method can be found in Table 1. Inspired by the work in [13], the designed framework of the subnetwork consists of 12 layers, which include one input layer, five convolutional layers, three max pooling layers, and three fully connected layers. First, for the convolutional layers Conv1 and Conv2, the kernel number is 16. The kernel number is 32 for the convolutional layers Conv3 and Conv4. The kernel number of the last convolutional layer (i.e., Conv5) is 64. Note that each kernel size for the convolutional layers is set to 3×3. We also adopt padding to ensure the unchanged patch size during the process of convolution operations.

In addition, we apply the exponential linear unit (ELU) instead of traditional activation functions such as the sigmoid, tanh, and rectified linear unit (ReLU) after the

convolution layer. Formally, the output of the nonlinear activation function ELU can be represented by:

$$u(x) = \begin{cases} x, x \geq 0 \\ \alpha(e^x - 1), x < 0 \end{cases}, \tag{3}$$

where α denotes the parameter to control negative factors and can output information even if the input is negative. Moreover, the mean of the overall output is approximately 0, which is more robust than other traditional activation functions. After the convolution and ELU layers, we exploit max pooling with a window size of 2×2. Therefore, for the feature extractor, the output shape of each feature map is 1/8 of the original input patch size.

We then flatten these feature maps to obtain $128 - dim$ feature vectors that can represent the input distorted structure or texture SR image patches. To conduct the perceptual quality regression, we use three fully connected layers with sizes equal to 128, 128, and 1. The ELU layers are also employed after the first two fully connected layers. Meanwhile, dropout is utilized to avoid overfitting. Specifically, the outputs of neurons are set to zero randomly with a particular probability. In our experiments, the probability is either 0.35 or 0.5. The last fully connected layer (i.e., the output layer) has one dimension that represents the predicted quality score.

Finally, to estimate the entire SR image quality, we assume that the visual distortions in reconstructed SR images are roughly homogeneous, which is appropriate for most practical situations. In this case, we thus take estimated scores of image patches as inputs and outputs the average mean of these scores, which is computed as the perceptual quality for the whole SR image.

2.4. Learning Quality

We need to obtain a large quantity of training data to train the proposed DeepSRQ network. Meanwhile, the input sizes should be fixed. Hence, we train our network on 32×32 cropped patches extracted from relatively larger distorted structure and texture SR images with the stride-based adaptive cropping approach. The subjective quality scores are taken as ground-truth labels. During the training process, the learning objective function of the network with weights $\mathbf{w}$ and updated weights $\widehat{\mathbf{w}}$ is defined to

minimize the mean squared error (MSE) as follows:

$$\mathcal{L}_{MSE} = ||c_{\mathbf{w}}(s_i, t_i) - y_i||_2^2$$

$$\widehat{\mathbf{w}} = \min_{\mathbf{w}} L, \tag{4}$$

where y_i represents the label of the distorted SR image patch, which is the human rating score. $c_w(p_i, t_i)$ denotes the quality score computed by the proposed DeepSRQ method. In addition, $s_i \in \{s_1, s_2, ..., s_{Num_p}\}$, $t_i \in \{t_1, t_2, ..., t_{Num_p}\}$, and $y_i \in \{y_1, y_2, ..., y_{Num_p}\}$. Num_p is the total number of input image patches.

To estimate the whole SR image quality by the predicted quality scores of input image patches, we adopt the average quality pooling as:

$$Q = \frac{1}{Num_p} \sum_{i=1}^{Num_p} c_{\mathbf{w}}(s_i, t_i), \tag{5}$$

where Q is the final perceptual quality prediction for the SR image. Note that using average pooling for quality evaluation tasks is a common practice because the spatial distortion in SR images is generally homogeneous. Indeed, we also test different saliency models, but the performance does not improve. One possible explanation is that state-of-the-art saliency detection algorithms focus on objects rather than distortion, which is more important for image quality assessment.

3. Experimental Results and Analysis

In this section, we report the results of several experiments to test the performance of DeepSRQ on three existing publicly available image superresolution quality databases, i.e., the SR quality database [17], SRID [32], and QADS [48]. We also pretrain DeepSRQ on the SR quality database, then perform cross validation on SRID, and vice versa. Moreover, we examine the effects of several parameter settings, including patch size and kernel size, and visualize the feature map to discover what has been learned from our proposed two-stream deep learning architecture. Furthermore, we carry out ablation experiments to test and quantify the performance gain of each key technique for learning the perceptual quality of image superresolution.

Table 2: The values of scaling factors (s) and the corresponding kernel width values (σ) in SR quality database [17].

s	2	3	4	5	6	8
σ	0.8	1.0	1.2	1.6	1.8	2.0

3.1. Protocol

We briefly introduce three publicly available image superresolution quality databases and the three commonly used criteria employed in the experiments.

The **SR quality database** includes a total of 1,620 SR images that are generated from LR images by nine SR algorithms. These SR algorithms are applied with a variety of scaling factors and kernel widths, denoted by s and σ, respectively. The numbers of ground-truth HR images and LR images are 30 and 60, respectively. Note that the larger subsampling factor requires a larger blur kernel width for better performance. Therefore, the optimal kernel width is applied for each scaling factor. The parameter selection details of this database can be found in Table 2. A subjective experiment is conducted to collect the subjective quality scores from 50 subjects. The mean of the median 40 subject scores is computed as the ground truth in the form of the mean opinion score (MOS), ranging from 0 to 10. Here, higher MOS means better perceptual quality.

SRID consists of 480 distorted SR images that are directly generated by LR images using two interpolation methods and six SR enhancement algorithms with three amplification factors of 2, 4, and 8. Nondistorted HR images are unavailable in this database. Subjective quality scores are provided in the form of MOS ranging from 0 to 10; the higher the value, the better the perceptual quality.

QADS contains 20 source images and 980 SRIs. The source images are selected from the MDID database [28] and Set14 database [43]. Three magnification scales are introduced to obtain the 980 SRIs, including 2 times, 3 times, and 4 times. Twenty-one image superresolution algorithms are applied to obtain the distorted SR images. Subjective quality ratings are given in MOS ranging from 0 to 1. A higher value indicates better perceptual quality.

We adopt three commonly used criteria to evaluate the performance of DeepSRQ:

Spearman rank-order correlation coefficient (SROCC), Pearson linear correlation coefficient (PLCC), and root mean square error (RMSE). Here, SROCC is used to evaluate prediction monotonicity, while PLCC and RMSE are used to evaluate prediction accuracy. Higher correlation coefficients and lower error indicate better agreement with human quality ratings. Moreover, before calculating the PLCC and RMSE performance of objective quality assessment methods, a nonlinear logistic fitting is applied to map the predicted scores to the same scales of subjective quality scores. Following [30], we adopt a four-parameter logistic function as follows:

$$g(x) = \frac{\tau_1 - \tau_2}{1 + e^{\frac{x-\tau_3}{\tau_4}}} + \tau_2, \tag{6}$$

where τ_1 to τ_4 are four free parameters to be determined in the curve fitting process. x denotes the raw objective score, and $g(x)$ represents the mapped score after the fitting.

3.2. Training Details

In the experiments, we use the stochastic gradient descent (SGD) as the optimization algorithm with 0.9 momentum. The learning rate is initially set to 10^{-2} with 10^{-6} decay. We update the network weights through backpropagation. The batch size is 128 in our experiments. For each image superresolution quality database, we randomly select 80% image data as the training set and the remaining 20% for testing. There is no overlap of source image content between the training and testing sets. The performance of the proposed DeepSRQ method is reported after 1,000 epochs.

3.3. Baselines

The proposed DeepSRQ method not only combines feature extraction with quality regression as well as pooling evaluation in a joint learning process but also creates synthetically learned features (i.e., both low-level visual information and high-level semantic features). To verify the effectiveness of our proposed method, it is compared with state-of-the-art IQA metrics using the handcrafted low-level features and the high-level semantic features extracted from pretrained DCNN models.

Due to the existence of original distortion-free HR images in the SR quality database [17], we compare our method with both the FR-IQA (i.e., PSNR, SSIM [33], multi-scale SSIM denoted by MS-SSIM [35]) and NR-IQA algorithms, which include the

blind/referenceless image spatial quality evaluator (BRISQUE) [18], natural image quality evaluator (NIQE) [19], local natural image quality evaluator (ILNIQE) [44], convolutional neural networks for no-reference image quality assessment (CNN-IQA) [11], and the shallow convolutional neural network for SR IQA (CNNSR) [4]. Moreover, we employ the pretrained DCNN (i.e., the well-known ResNet50 [8]) to extract high-level semantic features using the Caffe framework [10], and then combine the high-level semantic features (i.e., the 2,048 dimensions output of the pool5 layer) with other handcrafted features such as the NSS features. We then input these features into the support vector regression (SVR) model to predict the quality score.

For the QADS database [48], we compare the proposed DeepSRQ with 15 types of both traditional FR-IQA and NR-IQA metrics, including PSNR, SSIM [33], MS-SSIM [35], information fidelity criterion (IFC) [26], visual information fidelity (VIF) [25], most apparent distortion (MAD) [14], information content weighted SSIM (IW-SSIM) [34], feature similarity (FSIM) index [45], gradient similarity (GSIM) index [16], internal generative mechanism (IGM) [36], gradient magnitude similarity deviation (GMS-D) [39], directional anisotropy structure measurement (DASM) [3], superpixel-based similarity (SPSIM) index [27], structure-texture decomposition-based IQA approach called SIS [48], and local pattern statistics index (LPSI) [37]. It should be noted that the SIS [48] is an FR method that considers structure and texture information based on traditional handcrafted features. However, our proposed DeepSRQ is a DCNN-based NR algorithm. Therefore, we also compare the proposed DeepSRQ with two deep learning-based image quality evaluation methods, namely, CNN-IQA [11] and deep bilinear CNN (DBCNN) [46].

Since the SRID database [32] applies three different amplification factors to generate the SR images, the resolutions of these SR images vary. Moreover, the input size of ResNet50 is fixed to 224×224. Therefore, only the performance result of this method on the SR quality database [17] is shown. In addition, note that the SRID database has no-reference HR images; thus, more NR-IQA metrics are used for performance comparison except for those used on the SR quality database, which include no-reference free energy-based robust metric (NFERM) [6], blind image quality index (BIQI) [20], blind image integrity notator using DCT statistics (BLIINDS-II) [24], codebook rep-

resentation for no-reference image assessment (CORNIA) [42], derivative statistics-based image quality evaluator (DESIQUE) [47], distortion identification-based image verity and integrity evaluation index (DIIVINE) [21], and six-step blind metric (SISBLIM) [5].

In addition, we compare the proposed DeepSRQ with several state-of-the-art image sharpness assessment methods, including spectral and spatial sharpness measure (S3) [31], local phase coherence-based sharpness index (LPC-SI) [7], HVS-MaxPol-1 [9] using the best single kernel, and HVS-MaxPol-2 [9] adopting the combination of the best two kernels.

3.4. Performance Comparison

In this part, three publicly available SR image quality databases [17, 32] are used for performance comparison. For the SR quality database [17], we compare the performance of our proposed DeepSRQ method with three classical FR-IQA metrics, namely PSNR, SSIM [33], and MS-SSIM [35]. Additionally, several state-of-the-art NR-IQA metrics, including BRISQUE [18], NIQE [19], ILNIQE [44], CNN-IQA [11] and CNNSR [4], are taken for performance comparison. Among these four NR-IQA metrics, the CNNSR [4] is a shallow convolutional neural network specifically designed for evaluating the quality of SR images. As shown in Table 3, our proposed DeepSRQ outperforms both the FR and NR algorithms. The reason why the RMSE values of BRISQUE and ILNIQUE are much larger might be that they are not designed, optimized and tested for image superresolution applications.

Furthermore, since DeepSRQ considers both low-level visual information and high-level semantic features, we experimentally show that the synthetically learned features are more effective than both the handcrafted low-level features and the high-level semantic features extracted from pretrained DCNN models. Specifically, we employ the remarkable residual learning-based network (i.e., ResNet50) to extract high-level semantic features from its pool5 layer. In addition, before extracting the features, we crop the SR images into patches with 224×224 pixels due to the fixed input size of ResNet50. For each cropped image patch, we then obtain a $2,048 - dim$ feature. Since each patch in a particular SR image is equally important to the contribution of

Table 3: Performance comparison on SR quality database [17]. FR: full-reference, NR: no-reference.

Type	Method	SROCC	PLCC	RMSE
FR	PSNR	0.3110	0.3335	2.9383
	SSIM [33]	0.5562	0.5726	1.7980
	MS-SSIM [35]	0.6452	0.6218	1.0272
NR	BRISQUE [18]	0.5721	0.6176	10.0747
	NIQE [19]	0.6254	0.6364	1.5582
	ILNIQE [44]	0.6282	0.6198	18.3748
	LPSI [37]	0.4896	0.5276	2.0422
	S3 [31]	0.5066	0.5494	2.0087
	LPC-SI [7]	0.5441	0.5665	1.9812
	HVS-MaxPol-1 [9]	0.6423	0.6706	1.7834
	HVS-MaxPol-2 [9]	0.6314	0.6417	1.8438
	CNN-IQA [11]	0.7983	0.8398	1.312
	CNNSR [4]	0.8394	0.9156	1.2527
	ResNet50-pool5+NSS+SVR	0.8734	0.8873	1.1060
	Proposed DeepSRQ	**0.9206**	**0.9273**	**0.9042**

final perceptual quality, we input the average values of these high-level semantic features and the handcrafted low-level features (i.e., NSS features) into the SVR model for predicting the quality scores of SR images. The database is also divided randomly into 80% for training and 20% for testing. Finally, the procedure is repeated 1,000 times, and the median values are taken as the experimental results reported in Table 3. We find that our proposed DeepSRQ outperforms ResNet50-pool5+NSS+SVR, which further demonstrates the effectiveness of synthetically learned features in a deep neural network.

For the QADS [48], we compare the proposed DeepSRQ with state-of-the-art FR-IQA and NR-IQA metrics. The performance comparison results are shown in Table 4, which demonstrate that our proposed DeepSRQ method outperforms the other FR-IQA approaches. The compared FR-IQA algorithms include PSNR, SSIM [33], MS-SSIM [35], IFC [26], VIF [25], MAD [14], IW-SSIM [34], FSIM [45], GSIM [16], IGM [36], GMSD [39], DASM [3], SPSIM [27], SIS [48], and LPSI [37]. Note that the SIS adopts a structure-texture decomposition method and then calculates similarities from textural, structural and high-frequency aspects to form a parametric model. Our pro-

Table 4: Performance comparison on QADS database [48]. FR: full-reference, NR: no-reference.

Type	Method	SROCC	PLCC	RMSE
FR	PSNR	0.3544	0.3897	0.2530
	SSIM [33]	0.5290	0.5327	0.2325
	MS-SSIM [35]	0.7172	0.7240	0.1895
	IFC [26]	0.8609	0.8657	0.1375
	VIF [25]	0.8152	0.8210	0.1568
	MAD [14]	0.7234	0.7311	0.1874
	IW-SSIM [34]	0.8195	0.8234	0.1559
	FSIM [45]	0.6885	0.6902	0.1988
	GSIM [16]	0.5538	0.5684	0.2260
	IGM [36]	0.7145	0.7192	0.1907
	GMSD [39]	0.7650	0.7749	0.1736
	DASM [3]	0.7512	0.7585	0.1790
	SPSIM [27]	0.5751	0.5822	0.2233
	SIS [48]	0.9232	0.9230	0.1057
NR	LPSI [37]	0.4051	0.4207	0.2492
	S3 [31]	0.4636	0.4671	0.2429
	LPC-SI [7]	0.4902	0.4846	0.2403
	HVS-MaxPol-1 [9]	0.6160	0.6169	0.2162
	HVS-MaxPol-2 [9]	0.5739	0.5817	0.2234
	CNN-IQA [11]	0.8665	0.8709	0.1280
	DBCNN [46]	0.8707	0.8589	0.1508
	Proposed DeepSRQ	**0.9528**	**0.9557**	**0.0767**

posed DeepSRQ performs better than this SIS method due to the powerful learned discriminative features from the two-stream network. Moreover, the proposed DeepSRQ outperforms state-of-the-art NR-IQA methods, such as CNN-IQA [11] and DBCNN [46]. This is mainly because the characteristics of SR images are not well considered in these algorithms.

For the SRID [32], since the originally nondistorted HR images are unavailable in this database, we compare our method with more state-of-the-art NR-IQA metrics, which include BRISQUE [18], NIQE [19], ILNIQE [44], NFERM [6], BIQI [20], BLIINDS-II [24], CORNIA [42], DESIQUE [47], DIIVINE [21], and SISBLIM [5]. The performance comparison values are provided in Table 5. It can be seen that the

Table 5: Performance comparison on the SRID database [32].

Method	SROCC	PLCC	RMSE
BRISQUE[18]	0.6666	0.6738	1.1953
NIQE [19]	0.4759	0.5247	1.3769
ILNIQE [44]	0.4233	0.4136	1.4729
NFERM [6]	0.6177	0.6011	1.2927
BIQI [20]	0.4336	0.4253	1.2682
BLIINDS-II [24]	0.3687	0.3783	1.4973
CORNIA [42]	0.5985	0.6767	1.1909
DESIQUE [47]	0.5453	0.5253	1.3763
DIIVINE [21]	0.4826	0.4286	1.4614
SISBLIM [5]	0.5965	0.6223	1.2661
LPSI [37]	0.7454	0.7457	1.0777
S3 [31]	0.1797	0.1800	1.5910
LPC-SI [7]	0.0234	0.1978	1.6613
HVS-MaxPol-1 [9]	0.3736	0.3307	1.5264
HVS-MaxPol-2 [9]	0.4561	0.4237	1.4651
CNN-IQA [11]	0.8541	0.8783	0.7753
DBCNN [46]	0.6439	0.7422	4.5729
Proposed DeepSRQ	**0.9138**	**0.9309**	**0.5922**

proposed DeepSRQ method outperforms the other NR-IQA metrics.

In all three adopted SR image quality databases, several state-of-the-art image sharpness assessment methods are compared with our proposed DeepSRQ. As shown in Tables 3, 4 and 5, the proposed DeepSRQ can achieve better performance on all SR image quality databases.

3.5. Cross Dataset Validation

In addition, we test the generalization ability of our proposed DeepSRQ method through cross dataset validation. Since a similar data distribution is assumed between the training and testing images, we report the performance on the three image super-resolution quality databases [17, 32].

As shown in Table 6, DeepSRQ has a promising generalization ability for different databases. In other words, our proposed DeepSRQ is independent and robust for the

Table 6: Performance values in cross dataset evaluation.

Train → Test	SROCC	PLCC	RMSE
SR quality database → SRID	0.7225	0.7486	1.0749
SRID → SR quality database	0.8431	0.8415	1.3055

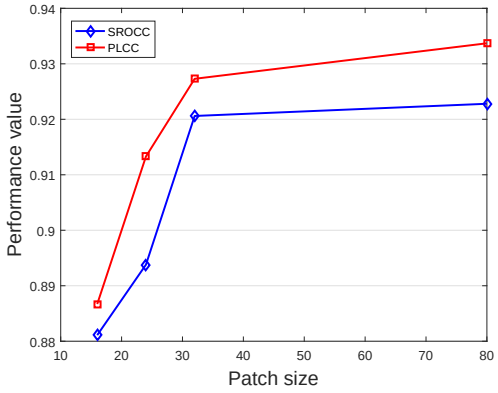


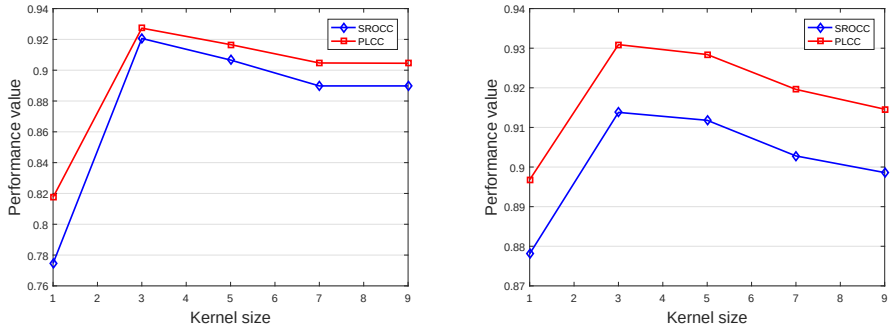
Figure 4: SROCC and PLCC performance with respect to the patch sizes including 16×16, 24×24, 32×32, and 80×80 on SR quality database [17].

used image superresolution quality databases. Note that the reason for the performance difference is due to different distributions in the three SR image quality databases.

3.6. Effects of Parameters

Several network parameters are involved in the proposed DeepSRQ design. To understand how these network parameters affect the performance of our proposed DeepSRQ method, we carry out experiments to test the DeepSRQ with different parameter settings.

Since the cropping approach is applied to increase our training data and the local visual information of the whole SR images is considered in our experiments, we examine how the patch size affects the performance of the proposed DeepSRQ method. Note that a fixed sampling stride (i.e., 32) is used for the SR quality database [17] to ensure that the amount of training data remains unchanged. Then, we vary the patch size while fixing the rest of the network architecture to plot the performance for the SR quality database [17].



(a) (b)

Figure 5: SROCC and PLCC performance with respect to the kernel sizes ranging from 1×1 to 9×9. (a) Run on SR quality database [17]. (b) Run on SRID [32].

Figure 4 shows the change in performance with respect to the patch sizes including 16×16, 24×24, 32×32, and 80×80. From Figure 4, we can see that a larger patch size results in better performance of the trained network. Moreover, the performance increases slightly as the patch size increases from 32×32 to 80×80. However, a larger patch size not only reduces the spatial quality resolution but also causes more processing time for training. Therefore, we prefer a relatively small patch size that can also yield promising performance.

In addition, different kernel sizes in the convolutional layers may lead to various performances because of the receptive field. Therefore, to discover how the kernel size affects the performance of our DeepSRQ algorithm, we change the kernel size while fixing the rest of the network architecture to plot the performance for the SR quality database [17] and SRID [32].

Figure 5 shows the change in SROCC and PLCC performance with respect to the kernel sizes ranging from 1×1 to 9×9. Except for the kernel size of 1×1, we can observe that a small kernel size creates an increase in both SROCC and PLCC performance. One possible explanation is that the relatively small receptive field (i.e., 3×3) is important for the image SR problem, which can effectively capture the notion of five orientations: up, down, left, right, and center.

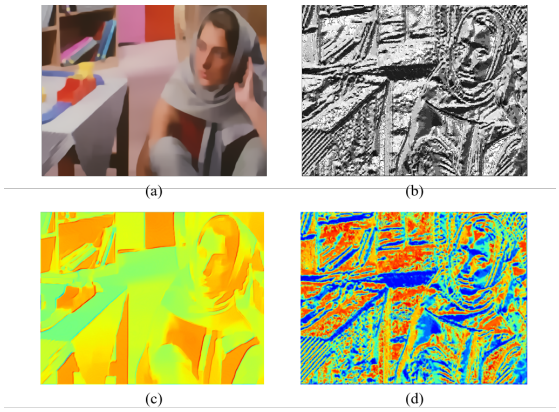


Figure 6: Examples of distorted structure and texture SR images as well as the corresponding feature maps in the first convolutional layer. (a) Distorted structure SR image. (b) Distorted texture SR image. (c) Feature map of (a). (d) Feature map of (b).

3.7. Visualize Learned Feature Map

To discover what has been learned from the proposed two-stream deep learning scheme, we visualize the feature maps in the first convolutional layer. Figure 6 depicts one of the feature maps at the first convolutional layer for both distorted structure and texture SR images. We can observe that the structural information and textural details can be separately learned from the structure and texture substreams, respectively, which further verifies the effectiveness of our DeepSRQ method.

3.8. Ablation Experiments

To validate the necessity of each subnetwork and how they contribute to the whole two-stream framework, we use each substream to perform the perceptual quality prediction of SR images. The results of this ablation study are provided in Table 7. We find that the combination of structure and texture streams helps improve the final performance. Moreover, since the LBP texture descriptor involves the radius parameter, which is denoted by r, we also vary different values of r to discover how it influences the performance of our algorithm. Table 8 shows the change in SROCC, PLCC and RMSE performance with respect to the radius r on the QADS database [48]. Examples of distorted SR images and the corresponding texture images with different LBP

Table 7: Ablation study of each substream on SR quality database [17], SRID database [32], and QADS database [48].

SR quality database [17]	SROCC	PLCC	RMSE
Structure	0.8242	0.8213	1.3787
Texture	0.9049	0.9153	0.9733
Proposed DeepSRQ	**0.9206**	**0.9273**	**0.9042**
SRID database [32]	SROCC	PLCC	RMSE
Structure	0.8619	0.8797	0.7709
Texture	0.8840	0.9094	0.6742
Proposed DeepSRQ	**0.9138**	**0.9309**	**0.5922**
QADS database [48]	SROCC	PLCC	RMSE
Structure	0.9137	0.9214	0.1012
Texture	0.9138	0.9242	0.0995
Proposed DeepSRQ	**0.9528**	**0.9557**	**0.0767**

Table 8: Parameter experiment results on QADS database [48].

Parameter Settings	SROCC	PLCC	RMSE
$r = 1$	**0.9138**	**0.9242**	**0.0995**
$r = 2$	0.9116	0.9197	0.1023
$r = 3$	0.8988	0.9028	0.1121
$r = 4$	0.8809	0.8839	0.1218
$r = 5$	0.8807	0.8808	0.1233

radii r ranging from 1 to 5 in the QADS database [48] are illustrated in Figure 7. We can observe that a smaller radius brings about an increase in performance due to more reserved texture details.

Additionally, to demonstrate that the adopted techniques are critical for the performance of DeepSRQ for perception-driven image superresolution, we further conduct several ablation experiments. Specifically, we remove the adaptive cropping approach and then test the performance of the remaining framework. As shown in Table 9, the proposed stride-based adaptive cropping approach is validated to further improve the performance of our proposed DeepSRQ.

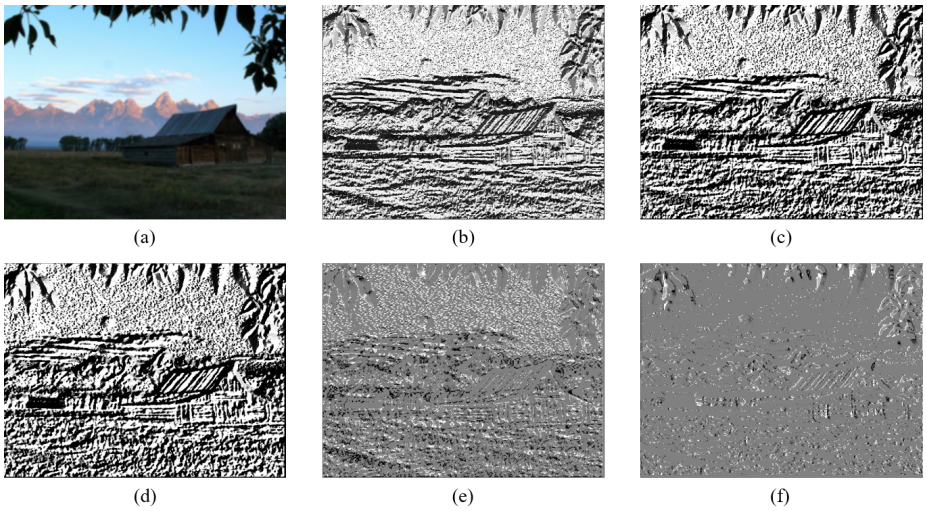


Figure 7: Examples of distorted SR images and the corresponding texture images with different LBP radii r in the QADS database [48]. (a) Distorted SR image. (b) Extracted texture image with LBP radius $r = 1$. (c) Extracted texture image with LBP radius $r = 2$. (d) Extracted texture image with LBP radius $r = 3$. (e) Extracted texture image with LBP radius $r = 4$. (f) Extracted texture image with LBP radius $r = 5$.

Table 9: Ablation experiment about adaptive cropping approach on SRID database [32]. DeepSRQ\ADA denotes the removed adaptive cropping approach. The best results are in bold.

Ablation	SROCC	PLCC	RMSE
DeepSRQ\ADA	0.8988	0.9061	0.6736
Proposed DeepSRQ	**0.9138**	**0.9309**	**0.5922**

4. Conclusions

In this paper, we propose a two-stream network to predict the perceptual quality of SR images in a no-reference manner, which is demonstrated to be more consistent with human perception. We consider both the structural and textural characteristics of the distortions in SR images. The top performance and promising generalization capacity of our proposed DeepSRQ method are validated by comparison with state-of-the-art IQA algorithms on three publicly available SR image quality databases. Experimental results also show that the synthetically learned features in a deep neural network are more effective than both the handcrafted low-level visual features and the high-level semantic features. Moreover, we validate that the two-stream scheme performs bet-

ter than each substream through an ablation study. Extensive parameter experiments also show various aspects of our proposed DeepSRQ. In addition, the proposed stride-based adaptive cropping approach is verified to further improve the performance of the proposed DeepSRQ method.

In future studies, we will apply the DeepSRQ metric to automatically optimize image SR frameworks, including both learning-free and learning-based SR algorithms. Furthermore, it is worth designing more effective and robust deep neural networks by considering more relevant characteristics of image superresolution.

Acknowledgment

This work was supported in part by NSFC under Grant 61571413, 61632001.

References

[1] Jean-François Aujol, Guy Gilboa, Tony Chan, and Stanley Osher. Structure-texture image decomposition‖modeling, algorithms, and parameter selection. *International journal of computer vision*, 67(1):111–136, 2006.

[2] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

[3] Li Ding, Hua Huang, and Yu Zang. Image quality assessment using directional anisotropy structure measurement. *IEEE Transactions on Image Processing*, 26(4):1799–1809, 2017.

[4] Yuming Fang, Chi Zhang, Wenhan Yang, Jiaying Liu, and Zongming Guo. Blind visual quality assessment for image super-resolution by convolutional neural network. *Multimedia Tools and Applications*, pages 1–18, 2018.

[5] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Transactions on Broadcasting*, 60(3):555–567, 2014.

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

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

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Mahdi S Hosseini, Yueyang Zhang, and Konstantinos N Plataniotis. Encoding visual sensitivity by MaxPol convolution filters for image sharpness assessment. *IEEE Transactions on Image Processing*, 2019.

[10] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[11] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014.

[12] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, 2017.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

[15] Zechao Li, Jinhui Tang, and Tao Mei. Deep collaborative embedding for social image understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[16] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2012.

[17] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

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

[19] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[20] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters*, 17(5):513–516, 2010.

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

[22] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):971–987, 2002.

[23] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 20(3):21–36, 2003.

[24] 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*, 21(8):3339–3352, 2012.

[25] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages iii–709. IEEE, 2004.

[26] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005.

[27] Wen Sun, Qingmin Liao, Jing-Hao Xue, and Fei Zhou. SPSIM: A superpixel-based similarity index for full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(9):4232–4244, 2018.

[28] Wen Sun, Fei Zhou, and Qingmin Liao. MDID: A multiply distorted image database for image quality assessment. *Pattern Recognition*, 61:153–168, 2017.

[29] RY Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984.

[30] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II. *VQEG, 2003*.

[31] Cuong T Vu, Thien D Phan, and Damon M Chandler. S3: A spectral and spatial measure of local perceived sharpness in natural images. *IEEE transactions on image processing*, 21(3):934–945, 2011.

[32] Guangcheng Wang, Leida Li, Qiaohong Li, Ke Gu, Zhaolin Lu, and Jiansheng Qian. Perceptual evaluation of single-image super-resolution reconstruction. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 3145–3149. IEEE, 2017.

[33] 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*, 13(4):600–612, 2004.

[34] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, 2011.

[35] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. Ieee, 2003.

[36] Jinjian Wu, Weisi Lin, Guangming Shi, and Anmin Liu. Perceptual quality metric with internal generative mechanism. *IEEE Transactions on Image Processing*, 22(1):43–54, 2013.

[37] Qingbo Wu, Zhou Wang, and Hongliang Li. A highly efficient method for blind image quality assessment. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 339–343. IEEE, 2015.

[38] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia. Structure extraction from texture via relative total variation. *ACM Transactions on Graphics (TOG)*, 31(6):139, 2012.

[39] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014.

[40] Bo Yan, Bahetiyaer Bare, Chenxi Ma, Ke Li, and Weimin Tan. Deep objective quality assessment driven single image super-resolution. *IEEE Transactions on Multimedia*, 2019.

[41] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386. Springer, 2014.

[42] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1098–1105. IEEE, 2012.

[43] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[44] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

[45] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

[46] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[47] Yi Zhang and Damon M Chandler. No-reference image quality assessment based on log-derivative statistics of natural scenes. *Journal of Electronic Imaging*, 22(4):043025, 2013.

[48] Fei Zhou, Rongguo Yao, Bozhi Liu, and Guoping Qiu. Visual quality assessment for super-resolved images: Database and method. *IEEE Transactions on Image Processing*, 28(7):3528–3541, 2019.

[49] Wei Zhou, Zhibo Chen, and Weiping Li. Stereoscopic video quality prediction based on end-to-end dual stream deep neural networks. In *Pacific Rim Conference on Multimedia*, pages 482–492. Springer, 2018.

[50] Wujie Zhou, Lu Yu, Weiwei Qiu, Yang Zhou, and Mingwei Wu. Local gradient patterns (L-GP): An effective local-statistical-feature extraction scheme for no-reference image quality assessment. *Information Sciences*, 397:1–14, 2017.