Signal Processing 178 (2021) 107771

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Learning stacking regression for no-reference super-resolution image quality assessment☆

Kaibing Zhang[a], Danni Zhu[a], Jie Li[b], Xinbo Gao[c,*], Fei Gao[b], Jian Lu[a]

[a] School of Electronics and Information, Xi'an Polytechnic University, Xi'an, 710048, China
[b] School of Electronic Engineering, Xidian University, Xi'an 710071, China
[c] The Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

ARTICLE INFO

Article history:
Received 15 December 2019
Revised 18 June 2020
Accepted 21 August 2020
Available online 26 August 2020

Keywords:
No-reference (NR)
Super-resolution (SR) image quality assessment (SRIQA)
Stacking regression

ABSTRACT

No-reference super-resolution (SR) image quality assessment (NR-SRIQA) aims to evaluate the quality of SR images without relying on any reference images. Currently, most previous methods usually utilize a certain handcrafted perceptual statistical features to quantify the degradation of SR images and a simple regression model to learn the mapping relationship from the features to the perceptual quality. Although these methods achieved promising performance, they still have some limitations: 1) the handcrafted features cannot accurately quantify the degradation of SR images; 2) the complex mapping relationship between the features and the quality scores cannot be well approximated by a simple regression model. To alleviate the above problems, we propose a novel stacking regression framework for NR-SRIQA. In the proposed method, we use a pre-trained VGGNet to extract the deep features for measuring the degradation of SR images, and then develop a stacking regression framework to establish the relationship between the learned deep features and the quality scores to achieve the NR-SRIQA. The stacking regression integrates two base regressors, namely Support Vector Regression (SVR) and K-Nearest Neighbor (K-NN) regression, and a simple linear regression as a meta-regressor. Thanks to the feature representation capability of deep neural networks (DNNs) and the complementary features of the two base regressors, the experimental results indicate that the proposed stacking regression framework is capable of yielding higher consistency with human visual judgments on the quality of SR images than other state-of-the-art SRIQA methods.

1. Introduction

The objective of image super-resolution (SR) reconstruction is to generate a high-resolution (HR) image with more details by using one or several low-resolution (LR) images from the same scene [1]. The SR technology has great potential applications in many fields such as computer vision, medical image analysis, remote sensing imaging, and life entertainment. In order to assess the quality of SR images and to further optimize the performance of SR algorithms, one of the most key tasks is to evaluate the quality of resultant SR images. There is no wonder that human's opinion is the ultimate receiver of images, so subjective quality assessment is regarded as the most direct yet effective way to reflect the quality of SR images [2]. Nevertheless, the process of subjective quality assessment is time-consuming and energy-draining. As a result, this kind of methods cannot be easily integrated into an SR application system for real-world scenarios.

In contrast to subjective quality assessment on SR images, the other kind of SRIQA methods is objective quality assessment, which automatically evaluates the quality of SR images through a computational model. In general, these methods can be classified into three major categories [3], i.e., full reference image quality assessment (FRIQA) [4], reduced-reference image quality assessment (RRIQA) [5], and no reference image quality assessment (NRIQA) [6,7]. When applying the FRIQA metrics such as mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [8], and information fidelity criterion (IFC) [9] to SRIQA, the corresponding original HR image is required as a reference to calculate the quality. However, the results obtained by these FRIQA metrics sometimes are not consistent with human's perceptual quality. Moreover, the original HR images are not available in practice. Consequently, the conventional FRIQA metrics are not suitable for evaluating the quality of SR images and the capability of SR algorithms. The second kind of image quality assessment

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.
E-mail address: xbgao@mail.xidian.edu.cn (X. Gao).

https://doi.org/10.1016/j.sigpro.2020.107771
0165-1684/

methods is called RRIQA metrics which only need reduced amount of reference images, so they are better for application in practice than FRIQA. Yeganeh et al. [10] quantified artifacts of SR images by distilling the natural scene statistical features from the frequency domain to the spatial domain. He et al. [11] operated a quality aware collection of local similarity features to predict the SR images degradation. Fang et al. [12] measured the quality of the SR images by virtue of the energy change and texture variation. Although this type of evaluation methods is more flexible than the FRIQA metrics, these methods still require partial information of the original HR images. The third subclass of IQA is named as NRIQA. This kind of methods does not need any information of original images, so they can overcome the shortcomings of the two aforementioned kinds of methods, leading to much attention in the literature.

To target NRIQA, most previous methods usually elaborate on using handcrafted perceptual statistical features to quantify the degradation of SR images. For example, Moorthy et al. [13] adopted natural statistical properties of images in the wavelet domain to characterize the quality of images. Observing that human visual mechanisms are more sensitive to the structural characteristics and the contrast of images, Saad et al. [14] combined the contrast, structural and anisotropic features of DCT statistical coefficients to represent the image distortion. Ma et al. [15] utilized local frequency features, global frequency features, and spatial domain features to quantify the degradation of SR images. Although these approaches are promising for IQA, the image quality cannot be well quantified by the used handcrafted features.

In contrast to the handcrafted features, learning high-level features from deep neural networks (DNNs) has gained much attention and shows successful applications in many fields such as computer vision, pattern recognition, and image processing. The DNNs show powerful representation capability and are able to automatically learn high-level feature representations to fully capture rich intrinsic information of image quality. Recently, a number of DNNs are used to address the task of NRIQA. For instance, Sun et al. [16] exploited the AlexNet architecture to extract semantic features implied in the global image content, and utilized saliency detection and Gabor filters to capture low-level features related to local image content. The overall image quality was estimated by combining these features. Li et al. [17] proposed to employ the ResNet architecture to represent the depth features of each overlapping image block in a statistical manner, by which the image quality is evaluated by a linear regression model. Gao et al. [18] utilized the VGGNet to extract image features of each layer, and the final image quality is estimated by averaging the predicted scores of multiple layers. Kang et al. [19] developed an IQA model by extracting features from Convolution Neural Network (CNN), where the feature extraction and the regression task are integrated into an optimization process. Bosse et al. [20] proposed to employ the CNN to extract high-level features from unpreprocessed image patches for quality prediction. Bianco et al. [21] elaborated on extracting deep features by fine-tuning a pre-trained network with an IQA dataset. With the obtained deep features, a support vector regression (SVR) model is trained to evaluate the quality of images. Experimental results suggested that these applied deep perceptual features perform better on IQA than many traditional handcrafted features due to their powerful capability of quantifying image quality. DNNs models are beneficial to capture the high-level semantics of images and the learned features are highly correlated with the quality degradation, which provides a potential application for NRIQA.

Besides perceptual statistical features, the other key component for IQA is how to build an accurate computational model to map the image features into the quality scores. In the literature of IQA, many predecessors tend to learn a single model, including SVR, for the quality prediction. However, in most cases only an individual model is insufficient enough to reveal the complicated relationship between the perceptual statistical features and the quality scores. To overcome this bottleneck, an alternative way is to introduce ensemble learning for IQA, by which multiple models such as different regression methods, are strategically generated and combined to amend the possible deviation on the quality estimate.

Stacking is an effective ensemble learning technique that builds a new model by combining the predictions from multiple models (e.g., decision tree, KNN or SVM) for a particular task. In principle, the method is the process of integrating different machine learning algorithms through holdout cross-validation [22]. Later Breiman [23] further improved stacking regression by replacing holdout cross-validation with k-fold cross-validation. Unlike bagging and boosting, the stacking regression is a useful ensemble learning technique that combines multiple diverse regression models via a meta-regressor. Previous works [24,25] have proved that, as a heterogeneous ensemble approach, stacking regression can significantly boost up the prediction performance by maximizing the complementary merits of different models.

In this paper, inspired by the significant advantages of stacking learning, we propose a novel quality metric for NR-SRIQA. In the method, the perceptual statistical features extracted from off-the-shelf VGGNet model are used to quantify the degradation of SR images. And then an effective stacking framework, which employs SVR and KNN regression as the base regressors, is developed to learn a mapping model from the obtained deep features to the perceived quality scores. With the stacking regression model, an NR-SRIQA metric can be used to predict the quality of any an given SR image. In summary, the unique contributions of this paper are mainly two aspects:

(1) We propose to employ a pre-trained VGGNet model to extract deep visual features rather than hand-crafted statistical features, to quantify the quality of SR images. The used features are propitious to reveal the fundamental artifacts of SR images than the traditional hand-crafted features.
(2) We develop a novel stacking regression-based framework to learn a coarse-to-fine metric mapping from deep features to quality scores for NR-SRIQA. The proposed quality metric can yield more accurate quality prediction on SR images than other state-of-the-art predecessors.

The remainder of the paper is organized as follows. Section 2 provides the VGG deep features to measure the degradation of SR images and presents a two-layer stacking regression framework for NR-SRIQA. In Section 3, we evaluate the performance of the proposed method and experimentally compare it with the state-of-the-art IQA metrics. Finally, Section 4 concludes the paper and outlooks the future work.

2. The proposed method

In this section, we first elaborate on the deep feature representation of SR images based on the VGG network. Next, a two-layer stacking regression model for NR-SRIQA is detailed.

2.1. Deep feature representation

As one of the most powerful presentation learning methods, CNN has been successfully applied to various computer vision tasks. To extract high-level features to quantify the degradation of SR images, we propose to employ a traditional VGG network to represent the image quality. Considering that training a very deep neural network model is often expensive and time-consuming, we leverage a pre-trained VGGNet model for image classification to accomplish the feature extraction of SR images [26].

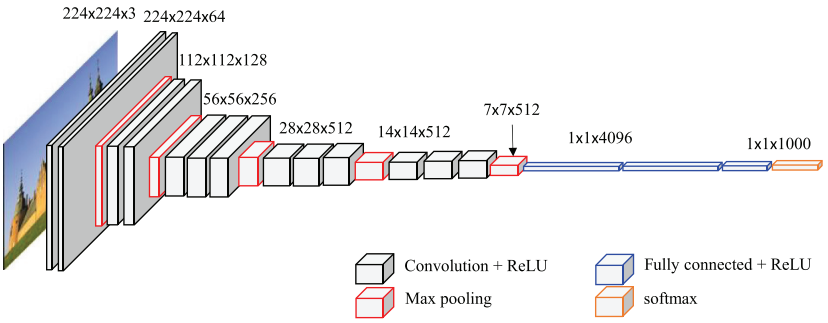

Fig. 1. Architecture of VGG16 network.

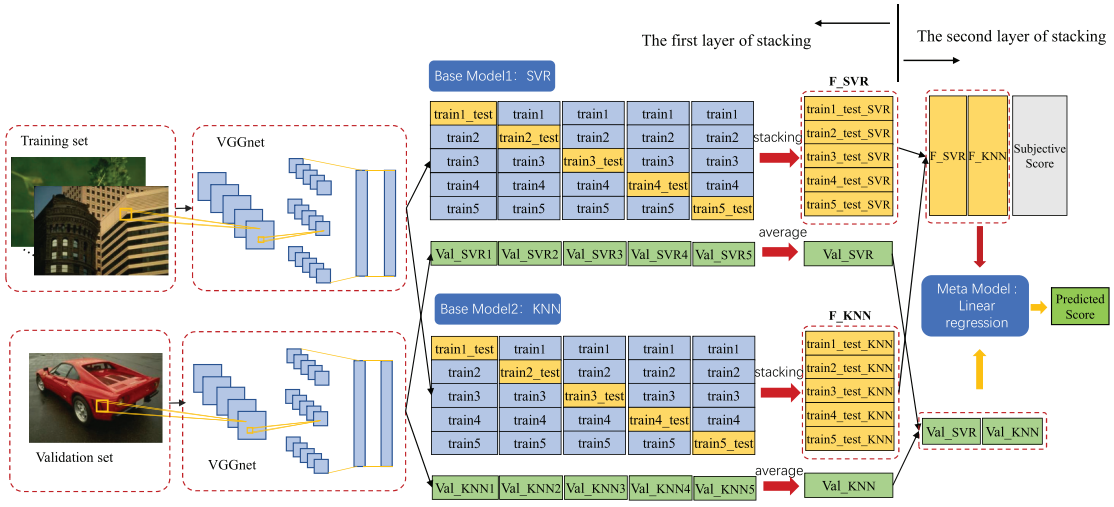

Fig. 2. The flowchart of the training-validating procedure in the proposed NR-SRIQA method.

Fig. 1 describes the architecture of a VGGNet model used in our work. As shown, the architecture of a VGGNet is mainly composed of 16 convolution (conv) layers, 3 fully connected (FC) layers, max-pooling (mpool) layer, and rectified linear units (relu). The convolution layers start from 64 to 512 feature maps and the size of each filter is 3×3. We specify the convolution stride and the spatial padding with 1-pixel size. Each max-pooling can generate a 2×2 window with a stride of 2. For these three fully-connected layers, the first two layers produce 4096 feature maps while the last one contains 1000 feature maps. In order to speed up the convergence of the training process, a rectified linear unit is followed by each of convolution layers and the first two fully-connected layers.

As the depth becomes deeper, the deep features of images tend to indicate the global features. Moreover, the output dimension of the last fully connected layer in the pre-training model is closely related to the training samples. To be more specific, for the image classification task, the output dimension of the last fully connected layer should be consistent with the number of classes of the training samples. Considering that directly applying the learned features at the last layer to IQA is obviously incompatible in that IQA is a typical regression problem, we carry out the NR-SRIQA by extracting a 4096-dimensional feature from the fully connected at the seventh layer in the pre-trained VGGNet. The obtained deep features facilitate to reveal the high-level semantic information in SR images for accurate SRIQA.

2.2. Stacking regression for NR-SRIQA

In this section, we focus on the structure of the stacking regression model for NR-SRIQA. In the model, we stack a two-layer regression model in an ensemble manner, where SVR and KNN are utilized as a list of base regressors in the first layer to obtain the meta-data by using the complete training set through five-fold cross-validation. In the second layer, a simple linear regression model is used as the meta-regressor to build the mapping relationship between the outputs of the base regressors in the ensemble and the expected outputs. Fig. 2 illustrates the flowchart of proposed NR-SRIQA framework.

From Fig. 2, we can see that the training process is divided into two parts. First, the deep feature extraction is performed on the training images through a VGGNet. Next a stacking regression framework is constructed to establish the mapping relationship between the deep features and the quality scores. In the validating stage, the feature extraction is first performed on a given SR test image and then the deep features are fed into the trained stacking regression model to obtain the predicted quality score. The pipeline of the stacking regression procedure for NR-SRIQA is implemented as follows. Given a training set D, the proposed stacking regression model first divides the training dataset into P equal subsets (e.g. P=5, $D = \{D_1, D_2, \cdots, D_5\}$). It is worth noting that the P subsets of the training dataset do not overlap each other. Afterwards, it follows a P-fold cross-validation process. Let $\hat{D}_p = D - D_p$ and D_p $(p = 1, 2, \cdots, 5)$ denote the training subset and the test subset at the p-th fold in the cross-validation, respectively, where each group $\hat{D}_p$ is utilized to learn the base regression models at the first layer by S (S=2) regression algorithms and D_p is used to generate the meta-data from the results obtained from the models at the first layer. At the end of the cross-validation process, the training set at the second layer, called the meta-data, consists of the predicted quality scores and its corresponding subjective

quality scores. Once the meta-data have been built from all training subsets in D, a linear regression algorithm is employed to generate the meta-model. Through the above process, the obtained two-layer stacking regression model can be applied to predict the quality of SR images.

In the verification stage, the deep perceptual features of an given SR image is first extracted by the pre-trained VGGNet model. Next P quality scores are separately predicted by each $\hat{D}_p(p = 1, \cdots, P)$. Subsequently, the average of P predicted results obtained from P base regressors is calculated as the inputs at the second layer. Finally, the quality of the test SR image is estimated via the obtained meta-model.

2.3. Base regressors in ensemble

In this subsection, we introduce two base regressors used in the proposed stacking NR-SRIQA framework.

2.3.1. Support vector regression

SVR is one of the most popular regression models due to its powerful capability of non-linear mapping. In the stacking regression at the first layer, we utilize SVR as one of base regressors to roughly estimate the perceptual score of an SR image. The relationship between the feature map x_i and the subjective score q_i can be formulated as below:

$$q_i = \langle w, \varphi(x_i) \rangle + b, \tag{1}$$

where w and b respectively denote the weight of feature and the bias, and $\varphi(.)$ represents a kernel function for mapping the raw data to a high-dimensional space. The Radial Basis Function (RBF) is widely used as a kernel function, which is described as

$$K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{2}$$

where σ is the standard deviance of the RBF kernel; x_i and x_j are deep features of the i-th and j-th SR images, respectively.

By applying SVR in a stacking structure, the m training images can be used to predict m quality scores through a five-fold cross-validation experiment. These predicted results Q are stacked together as the meta-data for the inputs at the second-layer regression.

2.3.2. K-Nearest neighbor regression

In the field of machine learning, one of the most effective non-parametric models is KNN regression [27]. Principally, the KNN regression uses a simple majority voting scheme to obtain a rough prediction by averaging the outputs of k nearest neighbors. Assume that $X = \{x_1, x_2, \cdots, x_m\} \in \mathbb{R}^{n \times m}$ denotes the features of the training images corresponding to m SR images and $\{q_i\}_{i=1}^m \in \mathbb{R}^m$ represents the subjective scores. Given that y_j is the feature vector of the j-th test image, we first employ the distance metric to measure the similarity between all feature vectors in the training set and the feature vector of the j-th test image. The most common distance metric of similarity is the Euclidean distance as below:

$$L(x_i, y_j) = \left(\sum_{i=1}^{m}\sum_{l=1}^{n} \|x_i^{(l)} - y_j^{(l)}\|^2\right)^{\frac{1}{2}}, \tag{3}$$

where n is the dimension of feature vector. By averaging the subjective scores $\{q_k\}_{k=1}^K$ corresponding to the K nearest neighbors, the predicted score of the j-th test image is computed by

$$\hat{Q} = \frac{1}{K}\sum_{k=1}^{K} q_k. \tag{4}$$

In view of the philosophy of stacking learning, the meta-regressor used in the stacking regression can further improve the accuracy of the quality prediction. To avoid overfitting, we select a simple linear regression as the meta-regressor at the second layer. Algorithm 1 summarizes the major procedures of the proposed NR-SRIQA method.

Algorithm 1: Learning a stacking regression model for NR-SRIQA.

Input: The training set $D = \{x_i, q_i\}_{i=1}^m$ with m SR images,where $x_i \in \mathbb{R}^n$ represents the ith features and $q_i \in \mathbb{R}$ denotes the ith subjective score.
Output: A stacking ensemble regression model H.

- **Step1:** Learn the base regression models at the first level.

for $s = 1$ *to* S **do**
  for $p = 1$ *to* P **do**
    Learn the base regressor h_{sp}.

- **Step2:** Construct the meta-data from D.

for $s = 1$ *to* S **do**
  for $i = 1$ *to* m **do**
    Construct meta-data set composed of $D_s' = \{x_i', q_i\}$, where $x_i' = \{h_{s1}(x_i), h_{s2}(x_i), \cdots, h_{sP}(x_i)\}$.

- **Step3:** Learn the meta regression model at the second level.

Learn a new regression model h' from the newly generated meta-data set $\{D_s'\}_{s=1}^S$.
return $H(X) = h'(h_{sp}(X)), s = 1, \cdots, S, p = 1, \cdots, P$.

3. Experimental results and analysis

In this section, we first introduce the SRIQA database used in the experiments. And then we carry out a set of validation experiments to validate the effectiveness of feature selection, the base regressor selection, and the heterogeneous ensemble regression. Next we probe how the the scale of training set affects the predicted performance. Finally, we further verify the superiority of the proposed method by comparing the performance of the existing state-of-the-art IQA methods.

3.1. Benchmark database

In the experiments, we adopt the SRIQA dataset prepared in [15] as benchmark to evaluate the performance on NR-SRIQA. The database includes 1620 super-resolved images of 180 LR images obtained by nine different SR algorithms. The MOS value of each SR image is derived from the median of 40 subjective perception scores from 40 observers. Quantitatively, four widely used indicators including Root Mean Square Error (RMSE), Pearson Linear Correlation Coefficient (PLCC) [28], Spearman Rank Order Correlation Coefficient (SROCC) [29], and Kendall Rank Order Correlation Coefficient (KROCC) [30], are employed to evaluate the performance of different methods. RMSE indictor reflects the accuracy of the evaluation algorithm. The smaller the value means the higher accuracy of the evaluated algorithm. PLCC indictor indicates the accuracy and the correlation of the evaluated algorithm. Under the normal circumstances, the better correlation with human perception means a closer value to 1. SROCC and KROCC are used to evaluate the prediction monotonicity between the objective and subjective evaluation scores. The larger the two indicators are, the better

Table 1
The performance indicators predicted by different deep features in a validation set.

Performance	Deep Features		
	AlexNet [31]	ResNet50 [32]	VGGNet
RMSE	1.0687	0.7121	**0.6961**
PLCC	0.9011	0.9563	**0.9574**
SROCC	0.8922	0.9506	**0.9528**
KROCC	0.7245	0.8112	**0.8173**

Table 2
Performance comparison of different features for describing SR images. The bold in the table indicates the best results.

Performance	Features				
	FC7	FC8	FC7(PCA)	FC8(PCA)	Ma et al. [15]
RMSE	**0.5802**	0.6091	0.6319	0.7075	0.8234
PLCC	**0.9689**	0.9657	0.9629	0.9533	0.9363
SROCC	**0.9622**	0.9583	0.9523	0.9436	0.9277
KROCC	**0.8399**	0.8361	0.8211	0.8061	0.7832

correlation with the human perceptual quality the algorithm means.

3.2. Effectiveness of deep features

For NR-SRIQA, how to select effective perceptual features to quantify the degradation of SR images plays a paramount role in the design of computational quality metric. In our experiments, we use a VGGNet to extract the deep perceived features for measuring the quality of SR images. Then the deep features at the seventh layer of the fully connected layer, which is denoted as FC7, are employed to train a stacking regression model for NR-SRIQA.

Table 1 tubulates the averaged values of the performance indicators obtained from 100 repeated experiments. The experimental results indicate that the VGGNet-based deep features can yield more competitive results than the deep features obtained from AlexNet [31] and ResNet [32]. Furthermore, in order to further prove the representation capability of the applied features, we also extract the features at the eighth layer of the fully connected layer, which is denoted as FC8, to compare the predicted performance. Meanwhile, Principle Component Analysis (PCA) algorithm is applied to reduce the dimensional size of features FC7 and FC8, where 99 percent of energy is maintained for efficient learning.

In addition, the performance of the traditional handcrafted perceptual statistical features proposed by Ma et al. [15] is also applied to validate the effectiveness of the VGGNet features. All the compared features are fed into the proposed stacking regression framework to establish SRIQA models. To ensure the reliability of the experiments, we select 80% of the samples as the training set and the remaining as the validating set. This experiment is repeated 100 times under the same experimental configuration. Table 2 reports the averaged values of the performance indices. As shown from the table, we can find that the four performance indicators obtained by 4096-dimensional FC7 features constantly exceed those by the FC8 features. It is worthy of noticing that the performance indicators achieved by the PCA-reduced FC7 features are slightly degraded in comparison to the original ones. In contrast to FC7 features, the four indicators of the PCA-reduced FC8 features are seriously degraded and are not comparable to these of the original FC8 features. Based on the above analysis, one can find that the extracted FC7 features are conducive to reliably revealing the quality of SR images than the FC8 features. In addition, we also find that the VGGNet-based features indicate obvious superiority over the handcrafted perceptual statistical features used by Ma et al. [15].

3.3. Selection of base regressors

Stacking regression is an ensemble learning technique that combines multiple base regression models through a meta-regressor. The individual regression models are trained based on the complete training set; then the meta-regressor is fitted based on the output meta-features of the individual regression models in the ensemble model [21]. Therefore, stacking regression is capable of improving the prediction accuracy by a linear combination of different predictors.

In general, two fundamental principles should be followed when selecting base regressors. The first one is that the correlation between the individual regressors should be as small as possible so as to exploit the complementary characteristics between different base regressors. The second is that the performance gap between different base regressors cannot be too large, otherwise the base regressor with poor performance will inevitably influence

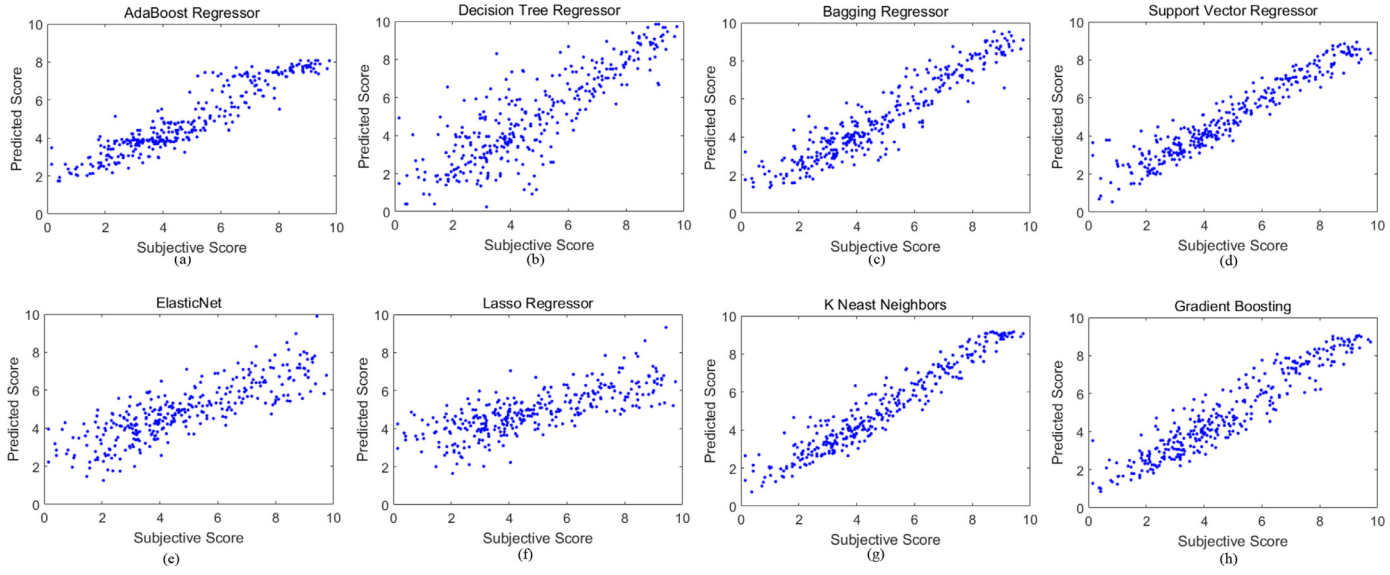

Fig. 3. The prediction results of a validation set obtained from eight regression models.

Table 3
The performance indicators predicted by eight different regressors in the validation set.

Performance	Regression Models							
	AdaBoost	Decision Tree	Bagging	SVR	Elastic Net	Lasso	KNN	Gradient Boosting
RMSE	0.9711	1.2564	0.8888	0.7225	1.3902	1.6534	0.6901	0.7765
PLCC	0.9274	0.8594	0.9292	0.9572	0.8270	0.7500	0.9618	0.9452
SROCC	0.9180	0.8356	0.9171	0.9515	0.8418	0.7721	0.9510	0.9356
KROCC	0.7593	0.6593	0.7581	0.8149	0.6448	0.5711	0.8202	0.7836

Table 4
Comparison of performance obtained from SVR & AdaBoost, SVR & Gradient Boosting, SVR & Bagging, KNN & AdaBoost, KNN & Gradient Boosting, and KNN & Bagging as base regressor, respectively.

Performance	Regression Models						
	SVR & AdaBoost	SVR & Gradient Boosting	SVR & Bagging	KNN & AdaBoost	KNN & Gradient Boosting	KNN & Bagging	SVR & KNN
RMSE	0.7198	0.6956	0.7303	0.7061	0.6741	0.6896	**0.5802**
PLCC	0.9536	0.9576	0.9533	0.9555	0.9596	0.9572	**0.9689**
SROCC	0.9486	0.9505	0.9474	0.9489	0.9514	0.9472	**0.9622**
KROCC	0.8070	0.8144	0.8093	0.8122	0.8161	0.8105	**0.8399**

the final performance of the stacking model. To validate the reasonability of our selection, eight different regressions including AdaBoost, Decision Tree, Bagging, SVR, Elastic Net, Lasso, KNN, and Gradient Boosting, are separately used to predict the quality of super-resolved images. Like the aforementioned experiment, we still choose 80 percent of the samples to train each model and the remaining 20 percent of the samples as the validation set. Fig. 3 demonstrates the predicted results obtained from eight different regressors. Based on the figure, we can clearly notice that either SVR or KNN can yields more consistent predicted scores with subjective scores than other base regressors.

Table 3 further demonstrates the performance indicators predicted by eight different regressors under the same validation set. We observe that when AdaBoost, Gradient Boosting, Bagging, SVR, and KNN are used as the base regressors in the stacking regression, all the RMSE values are less than 1. Particularly, noticing that both SVR and KNN achieve better predicted accuracy than other base regressors, we recommend choosing SVR and KNN as the candidate base regressors in our stacking regression model. Moreover, the other three different regressors, i.e., AdaBoost, Gradient Boosting, and Bagging, also yield smaller performance differences and higher accuracy, so one of them can be integrated with SVR or KNN for heterogeneous ensemble regression. In order to seek an optimal combination of two base regressors, Table 4 reports the training-validating experimental results achieved by different combinations. The higher accuracy implies the better combination of base regressors. Based on the contrastive results in Table 4, the heterogeneous ensemble of the two base regressors SVR and KNN produces the best results among all the possible candidates.

We further plot the predicted results obtained by SVR and KNN with a scatter plot and a fitting curve as shown in Fig. 4. In order to clearly show the complementary features of two base regressors, Fig. 4 demonstrates the local green and yellow boxes to display two intersected curves in red color and blue color. Based on the results in the figure we can see that the two base regressions of SVR and KNN have smaller correlation and are complementary each other at a certain degree. According to the two principles of ensemble learning, it is reasonable to select SVR and KNN as the base regressors to build the stacking regression model for NR-SRIQA.

In addition, in the experiments we utilize the strategy of cross-validation and grid optimization to choose the appropriate parameters for SVR and KNN. The penalty factor of SVR is set to 10 and the number of k nearest neighbors used in KNN is set to 3 throughout the experiments.

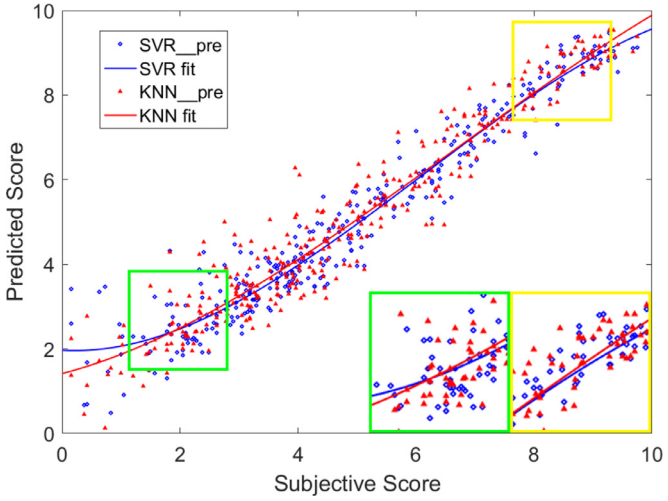

Fig. 4. The predicted results of a validation set obtained by SVR and KNN regression models.

Table 5
Comparison of averaged performance obtained by SVR & SVR, KNN & KNN, and SVR & KNN as base regressors. The bold in the table indicates the best results.

Performance	Base Regressor		
	SVR &SVR	KNN &KNN	SVR & KNN
RMSE	0.6440	0.6652	**0.5802**
PLCC	0.9616	0.9589	**0.9689**
SROCC	0.9554	0.9414	**0.9622**
KROCC	0.8264	0.8043	**0.8399**

3.4. Effectiveness of heterogeneous ensemble regression

As we all know, the ensemble learning can be divided into two fundamental categories: homogeneous ensemble learning and heterogeneous ensemble learning. Although the heterogeneous ensemble learning is one representative scheme, the homogeneous ensemble scheme is adopted in many cases. In this section, we use three cross combinations towards two base regressors, which are denoted as SVR & SVR, KNN & KNN, and SVR & KNN, respectively, to perform three training-test trials. Among them, the parameter settings of all the SVRs are the same configuration. Accordingly, all the KNNs are also set the same experimental settings in the ensemble. Table 5 reports the experimental results obtained by

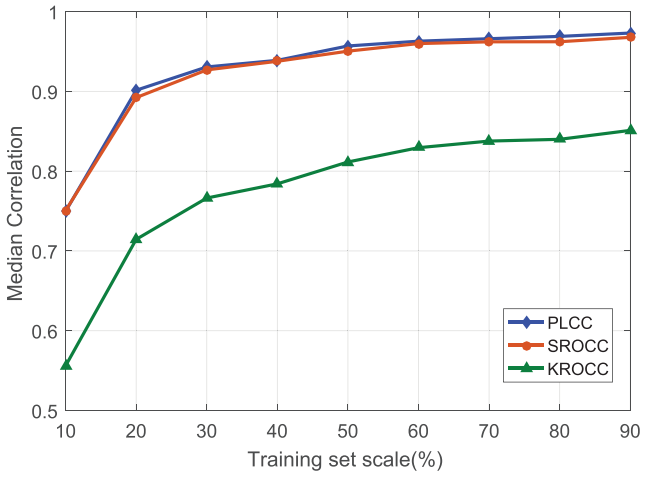

Fig. 5. Performance influence on different scales of training set.

different ensembles. As shown, although the overall performance of the homogeneous ensemble regressions is improved in comparison to single regression, their performance are still inferior to the results obtained by the heterogeneous ensemble regression. Based on the above analysis, we suggest selecting heterogeneous ensemble regression of SVR and KNN for NR-SRIQA.

3.5. Performance influence on the scale of training dataset

In this section, we will experiment on how the scale of training set impacts on the accuracy of predicted quality scores. To this end, the scales of training set are changed from 10% to 90% with a step of 10% and the remaining of the database is used for validating. To obtain a fair evaluation, all these experiments are repeated 100 times and the averaged performance indicators on different validation sets are demonstrated in Fig. 5. In terms of the results demonstrated in the figure, similar to other machine learning methods, the performance of the proposed stacking regression model is also improved as the increase of the scale of the used training set. In particular, when the scale of the training set reaches 30%, both PLCC and SROCC exceed 0.9 or even approach 0.95. The empirical study indicates that the proposed stacking regression-based NR-SRIQA can achieve a stable and effective prediction model with a small-scale training dataset.

3.6. Contrastive results

In this section, we will compare the proposed method with six different NRIQA methods, namely BLIINDS [14], BRISQUE [33], SSEQ [34], ILNIQE [35], Kang et al. [19] method, and Ma et al. [15] method. In these experiments, we randomly select 80 percent of samples in the SRIQA dataset to train the NR-SRIQA model and the remaining as a validation set for the performance evaluation. Fig. 6 shows the scatter plots from the collection of the compared experimental results, where the abscissa of the scatter plot is the subjective quality evaluation scores and the ordinate of the scatter plot is the predicted scores obtained from each algorithm. Each point in the scatter plot represents a predicted image, and the red line is obtained by a logic function. When the scatter distribution is closer to the fitted line, the evaluated results imply better consistency with the subjective perceptual scores. It can be seen from all the scatter plots that the compared NRIQA methods can yield considerable promising experimental results than ILNIQE method. Compared with other regression-based methods, the points near the curve fitted by the ILNIQE method are more scattered. ILNIQE method learns a multivariate Gaussian (MVG) model of image patches from a collection of pristine natural images. Using a learned MVG model, a Bhattacharyya-like distance is adopted to measure the quality of each image patch, and then an overall quality score is calculated by average pooling. However, in most cases only a single distance metric is insufficient enough to reveal the complicated relationship between the perceptual features and the quality scores. Conversely, Kang et al. method [19], Ma et al.

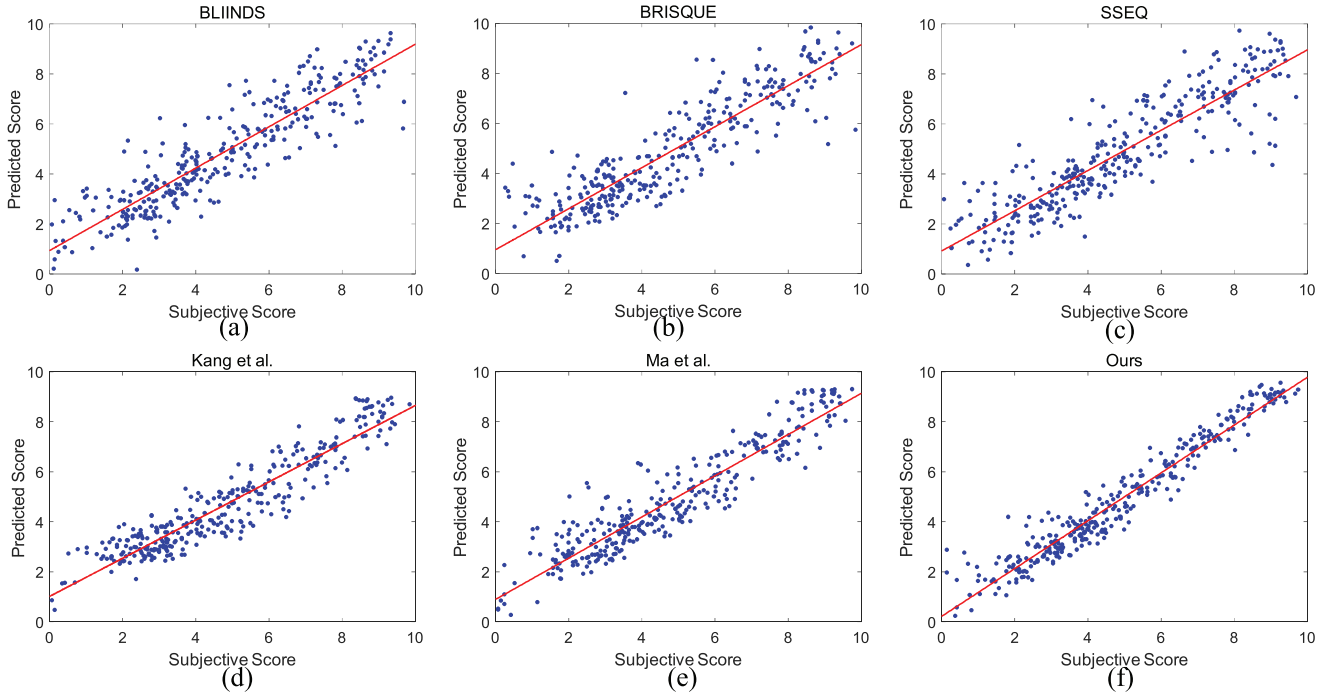

Fig. 6. Scatter plots of different NRIQA methods.

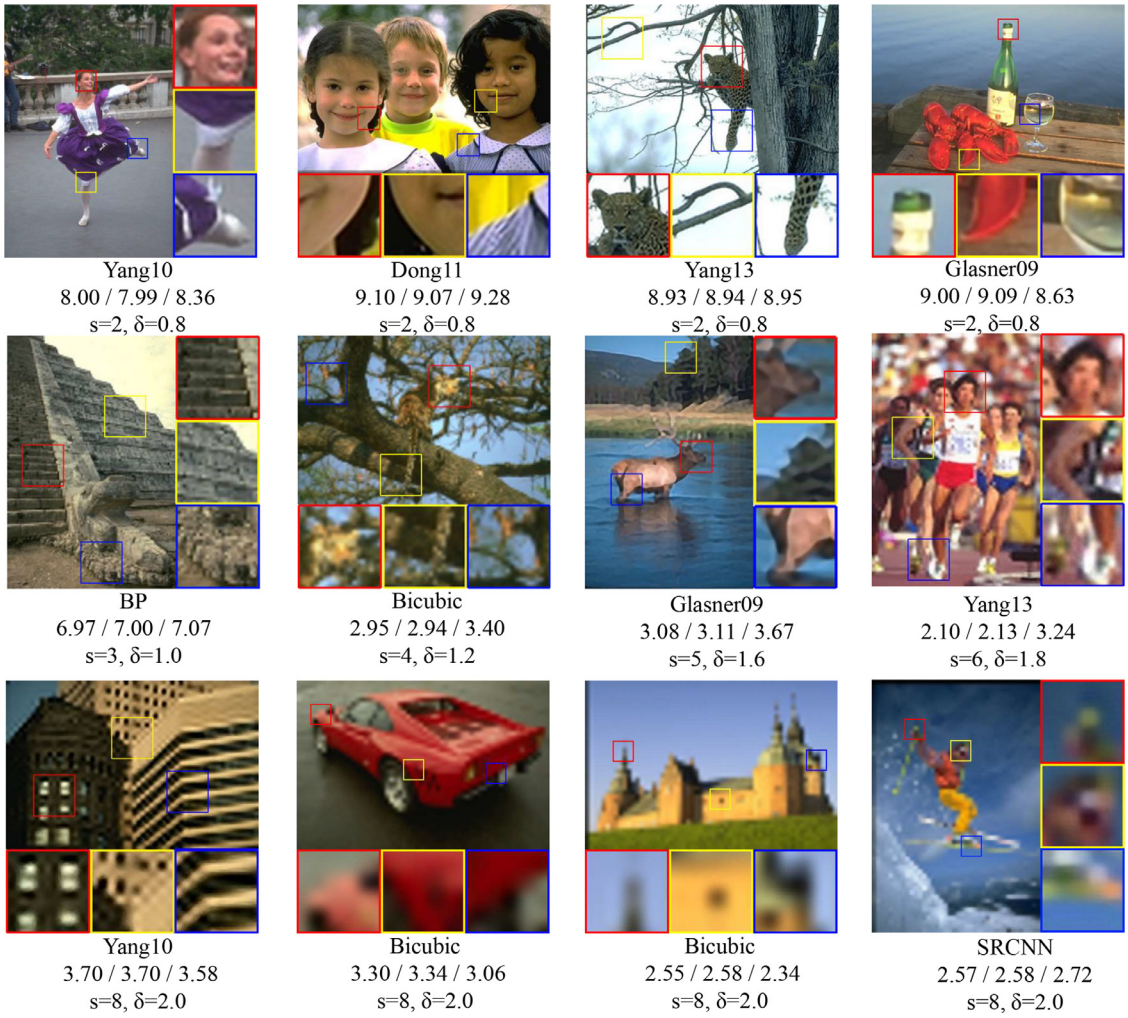

Fig. 7. Twelve best cases stem from the SRIQA dataset by the proposed evaluation method in a training-validating experiment. The left / middle / right values denote the perceptual scores, the predicted scores using the proposed method and the predicted scores using Ma's method. Note that s is the upscaling factor and δ is the width of the kernel function.

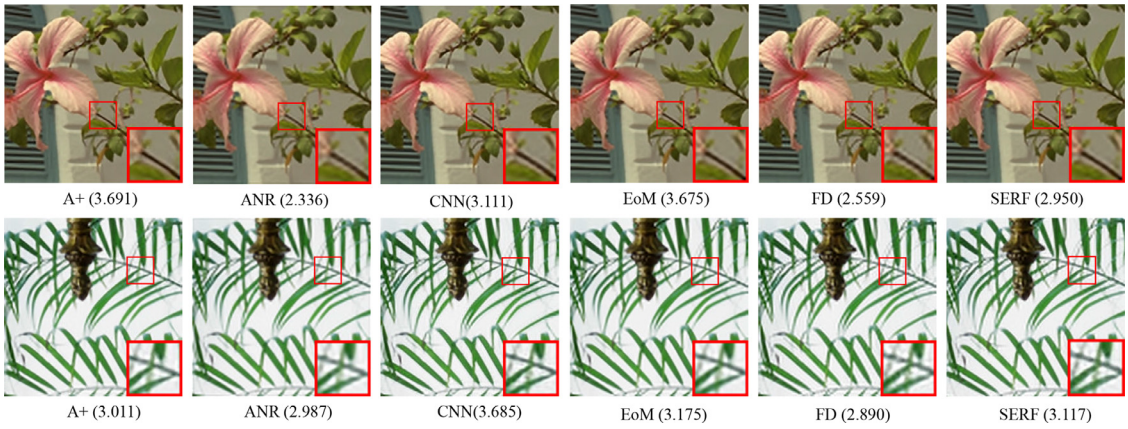

Fig. 8. The quality scores of different SR images (flowers and leaves) with the up-sampling factors of 3 predicted by the proposed method.

method [15], and the proposed method contribute to more consistency than the remaining competitors. In particular, in terms of the contrastive results demonstrated in Fig. 6(d), (e), and (f), we can confirm that the proposed method is capable of obtaining a better consistency than other two approaches.

In order to reduce the possible performance bias of a single experiment, we repeat the training-validating experiment 100 times and calculate the averaged performance indicators achieved by each evaluation algorithm. The compared results are listed in Table 6. The results listed in the table reflect not only the accuracy of the NRIQA methods but also the monotony of the NRIQA methods. Note that the larger RMSE value means the poorer performance of quality assessment. As the results presented in the table, the proposed method achieves the smallest RMSE and there-

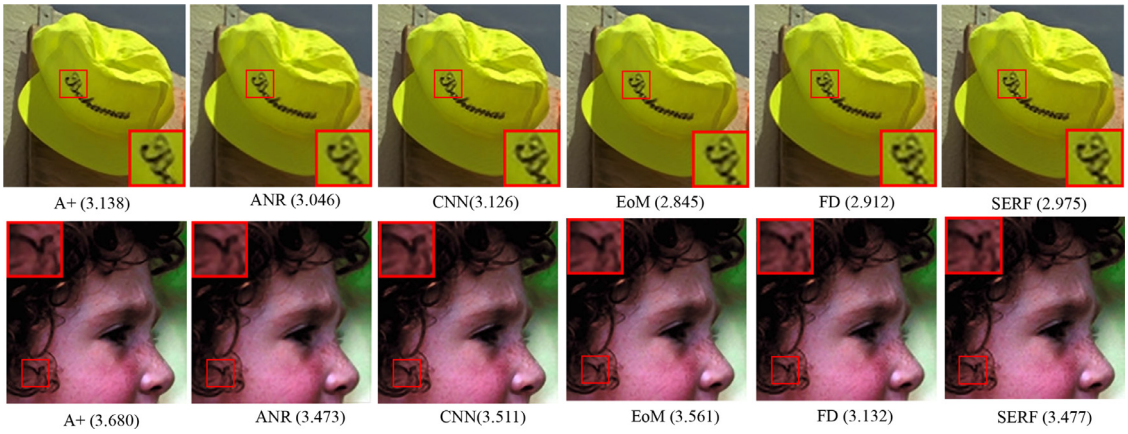

Fig. 9. The quality scores of different SR images (hats and girl) with the up-sampling factors of 3 predicted by the proposed method.

Table 6
Comparison of median indicators obtained from different NRIQA methods across 100 training-validating random splits. The bold in the table indicates the best results.

Performance	NRIQA Methods						
	BLIINDS [14]	BRISQUE [33]	SSEQ [34]	ILNIQE [35]	Ma et al. [15]	Kang et al. [19]	Ours
RMSE	1.0851	1.2134	1.1836	4.3407	0.8870	0.8836	**0.5802**
PLCC	0.8920	0.8631	0.8705	−0.2747	0.9295	0.9351	**0.9689**
SROCC	0.8960	0.8542	0.8751	−0.3187	0.9150	0.9230	**0.9622**
KROCC	0.7238	0.6698	0.6958	−0.2162	0.7554	0.7543	**0.8399**

fore exhibits the best prediction capability. Similarly, both SROCC and KROCC indicators are within the range from 0 to 1. The higher the values are and the better performance of evaluating algorithms is. As listed in the table, the two indictors obtained by the proposed algorithm show the best monotonicity and the SROCC index of our method exceeds 0.95. In addition, the PLCC indicator of proposed method also demonstrates the best correlation, which is far superior over the performance indictors of other NRIQA algorithms. Based on the above experimental analysis, one can claim that the proposed method can yield more compelling results and shows better consistency with human perceptual quality than other competitors.

In order to further substantiate the effectiveness of the proposed method, Fig. 7 presents the twelve best cases when the proposed method is applied to evaluate the quality of SR images used in our experiment. In the figure, we provide the human perceptual scores and the resultant scores of Ma et al. method and the proposed methods as reference. In general, the SR images under a lower upscaling factor contain more visual texture details and sharper edges, exhibiting better perceived quality. Conversely, the SR images under a higher upscaling factor tend to be fewer texture details and more blurring edges, leading to worse perceived quality. For the SR images with richer details at the first rows, although the Ma et al. method has excellent prediction results, the proposed method performs better than the unique competitor. For the SR images with fewer details at the last two rows, the proposed method is still superior to Ma et al. method [15]. Especially for the SR images at the last row, the proposed method can yield more consistent predicted scores with the subjective scores when the SR image quality is poor, showing obvious advantages in comparing with the Ma et al. method [15]. This is because that deep learning-based features are propitious to accurately quantify the degradation of SR images. Moreover, the proposed heterogeneous stacking regression model provides an effective way to establish the mapping relationship between the deep features and the perceptual scores.

To validate the rationality of the proposed NR-SRIQA metric, we select 24 SR images obtained by six different SR algorithms, i.e., A+ [36], ANR [37], CNN [38], EoM [39], FD [40], and SERF [41], to predict their perceptual quality. Note that these test images are the SR images by a up-sampling factor of 3 and not contained in the benchmark database. Figs. 8 and 9 show the SR results along with the predicted scores corresponding to the resulting SR images. It is found that the perceptual quality scores obtained by A+ [36] and CNN [38] are much better than those by ANR-, SERF-, EoM-, and FD-based SR approaches. The quality scores predicted by the proposed NR-SRIQA model are well consistent with the subjective perception and can effectively indicate the SR performance of different SR algorithms.

4. Conclusion

We have proposed a remarkably effective NR-SRIQA metric based on stacking regression for SR image quality evaluation. The proposed method first uses the VGGNet model to extract the deep perceptual features to quantify the quality of SR images. Next, a stacking regression model is framed to predict the quality of SR images. In the stacking model, SVR and KNN regression are used as the two base regressors at the first layer and a linear regression as the meta-regressor at the second layer. Thorough experimental results verify that the proposed framework shows considerable advantages over the state-of-the-art NRIQA algorithms and can gain compelling consistency with subjective perception. In the future work, we will continue to investigate more effective deep networks such as saliency-guided deep neural networks [42] for more accurately quantifying the quality of SR images. Besides, deeper stacking regression model may be formulated to yield more accurate quality prediction on SR images.

Declaration of Competing Interest

None.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61971339, Grant 61971172, and Grant 61471161, in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by the Key Project of the Natural Science Foundation of Shaanxi Province under Grant 2018JZ6002, and in part by the Graduate Innovation Foundation of Xi'an Polytechnic University under Grant chx2019028.

References

[1] K. Zhang, J. Li, H. Wang, X. Liu, X. Gao, Learning local dictionaries and similarity structures for single image super-resolution, Signal Process. 142 (2018) 231–243.
[2] G. Wang, L. Li, Q. Li, K. Gu, Z. Lu, J. Qian, Perceptual evaluation of single-image super-resolution reconstruction, in: IEEE International Conference on Image Processing (ICIP), 2017, pp. 3145–3149.
[3] L. He, F. Gao, W. Hou, L. Hao, Objective image quality assessment: a survey, Int. J. Comput. Math. 91 (11) (2014) 2374–2388.
[4] T. Ahmad, S.S. Quershi, The full reference quality assessment metrics for super resolution of an image: shedding light or casting shadows? in: International Conference on Electronics and Information Engineering, vol. 2, 2010, pp. V2–224.
[5] S. Golestaneh, L.J. Karam, Reduced-reference quality assessment based on the entropy of DWT coefficients of locally weighted gradient magnitudes, IEEE Trans. Image Process. 25 (11) (2016) 5293–5303.
[6] V. Kamble, K. Bhurchandi, No-reference image quality assessment algorithms: a survey, Optik 126 (11-12) (2015) 1090–1097.
[7] S. Xu, S. Jiang, W. Min, No-reference/blind image quality assessment: a survey, IETE Tech. Rev. 34 (3) (2017) 223–245.
[8] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, et al., Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[9] H.R. Sheikh, A.C. Bovik, G. De Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128.
[10] H. Yeganeh, M. Rostami, Z. Wang, Objective quality assessment for image super-resolution: a natural scene statistics approach, in: 19th IEEE International Conference on Image Processing, 2012, pp. 1481–1484.
[11] H. Yuqing, C. Shuan, W. Jianguo, Assessment method of image super resolution reconstruction based on local similarity, in: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, 2013, pp. 154–157.
[12] Y. Fang, J. Liu, Y. Zhang, W. Lin, Z. Guo, Quality assessment for image super-resolution based on energy change and texture variation, in: IEEE International Conference on Image Processing (ICIP), 2016, pp. 2057–2061.
[13] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: from natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (12) (2011) 3350–3364.
[14] M.A. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: a natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (8) (2012) 3339–3352.
[15] C. Ma, C.-Y. Yang, X. Yang, M.-H. Yang, Learning a no-reference quality metric for single-image super-resolution, Comput. Vis. Image Underst. 158 (2017) 1–16.
[16] C. Sun, H. Li, W. Li, No-reference image quality assessment based on global and local content perception, Vis. Commun. Image Process. (VCIP) (2016) 1–4.
[17] D. Li, T. Jiang, M. Jiang, Exploiting high-level semantics for no-reference image quality assessment of realistic blur images, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 378–386.
[18] F. Gao, J. Yu, S. Zhu, Q. Huang, Q. Tian, Blind image quality prediction by exploiting multi-level deep representations, Pattern Recognit. 81 (2018) 432–442.
[19] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1733–1740.
[20] S. Bosse, D. Maniry, T. Wiegand, W. Samek, A deep neural network for image quality assessment, in: IEEE International Conference on Image Processing (ICIP), 2016, pp. 3773–3777.
[21] S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, Signal Image Video Process. 12 (2) (2018) 355–362.
[22] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.
[23] L. Breiman, Stacked regressions, Mach. Learn. 24 (1) (1996) 49–64.
[24] M.J. Van der Laan, E.C. Polley, A.E. Hubbard, Super learner, Stat. Appl. Genet. Mol.Biol. 6 (1) (2007).
[25] M.P. Sesmero, A.I. Ledezma, A. Sanchis, Generating ensembles of heterogeneous classifiers using stacked generalization, WIREs Data Min. Knowl. Discov. 5 (1) (2015) 21–34.
[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556(2014).
[27] Y. Song, J. Liang, J. Lu, X. Zhao, An efficient instance selection algorithm for k nearest neighbor regression, Neurocomputing 251 (2017) 26–34.
[28] P. Sedgwick, Pearson's correlation coefficient, BMJ 345 (2012) e4483.
[29] D.J. Sheskin, Spearman's rank-order correlation coefficient, in: Handbook of Parametric and Nonparametric Statistical Procedures, 2007, pp. 1353–1370.
[30] H. Abdi, The Kendall rank correlation coefficient, Encycl. Meas. Stat. (2007) 508–510.
[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: NIPS, 2012, pp. 1097–1105.
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[33] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.
[34] L. Liu, B. Liu, H. Huang, A.C. Bovik, No-reference image quality assessment based on spatial and spectral entropies, Signal Process. Image Commun. 29 (8) (2014) 856–863.
[35] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (2015) 2579–2591.
[36] R. Timofte, V. De Smet, L. Van Gool, A+: Adjusted anchored neighborhood regression for fast super-resolution, in: Asian Conference on Computer Vision, 2014, pp. 111–126.
[37] R. Timofte, V. De Smet, L. Van Gool, Anchored neighborhood regression for fast example-based super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1920–1927.
[38] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: European Conference on Computer Vision, 2014, pp. 184–199.
[39] K. Zhang, B. Wang, W. Zuo, H. Zhang, L. Zhang, Joint learning of multiple regressors for single image super-resolution, IEEE Signal Process. Lett. 23 (1) (2015) 102–106.
[40] C.-Y. Yang, M.-H. Yang, Fast direct super-resolution by simple functions, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 561–568.
[41] Y. Hu, N. Wang, D. Tao, X. Gao, X. Li, SERF: A simple, effective, robust, and fast image super-resolver from cascaded linear regression, IEEE Trans. Image Process. 25 (9) (2016) 4091–4102.
[42] S. Yang, G. Lin, Q. Jiang, W. Lin, A dilated inception network for visual saliency prediction, IEEE Trans. Multimed. (2019), doi:10.1109/TMM.2019.2947352.

Kaibing Zhang received the M.Sc. degree in Computer Software and Theory from Xihua University, Chengdu, China, in 2005 and the Ph.D. degree in Pattern Recognition and Intelligent System from Xidian University, Xi'an, China, in 2012, respectively. He is currently a Professor at the College of Electrics and Information, Xi'an Polytechnic University, Xi'an, China. His main research interests include pattern recognition, computer vision, and image super-resolution reconstruction. In these areas, he has published around 20 technical articles in refereed journals and proceedings including IEEE TIP, TNNLS, Signal Processing (Elsevier), Neurocmputing, CVPR, ICIP, etc.

Dan'ni Zhu received the B.S. degree in Measurement Technology and Instrument from Xi'an Jiaotong University City College, Xi'an, China, in 2015. She is currently pursuing her M.Sc. degree in the School of Electronics and Information, Xi'an Polytechnic University, Xi'an, China. Her research interests include machine learning, deep learning, and super-resolution image quality assessment.

Jie Li received the B.Sc. degree in electronic engineering, the M.Sc. degree in signal and information processing, and the Ph.D. degree in circuit and systems, from Xidian University, Xi'an, China, in 1995, 1998, and 2004, respectively. She is currently a Professor in the School of Electronic Engineering, Xidian University, China. Her research interests include image processing and machine learning. In these areas, she has published around 50 technical articles in refereed journals and proceedings including IEEE T-IP, T-CSVT, Information Sciences etc.

Xinbo Gao received the B.Eng., M.Sc. and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a post-doctoral research fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education of P. R. China, a Professor of Pattern Recognition and Intelligent System of Xidian University and a Professor of Computer Science and Technology of Chongqing University of Posts and Telecommunications. His current research interests include Image processing, computer vision, multimedia analysis, machine learning and pattern recognition. He has published six books and around 300 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology and a Fellow of the Chinese Institute of Electronics.

Fei Gao is currently with the School of Electronic Engineering, Xidian University; and the School of Computer Science and Technology, Hangzhou Dianzi University (HDU). He received his Bachelor Degree in Electronic Engineering and Ph.D. Degree in Information and Communication Engineering from Xidian University (Xi'an, China) in 2009 and 2015, respectively. From Oct. 2012 to Sep. 2013, he was a Visiting Ph.D. Candidate in University of Technology, Sydney (UTS) in Australia. He mainly applies machine learning techniques to computer vision problems. His research interests include visual quality assessment and enhancement, intelligent visual arts generation, biomedical image analysis, etc. His research results have expounded in 20 publications at prestigious journals and conferences. He served for a number of journals and conferences.

Jian Lu received the M.Sc. degree in Control Science and Engineering from the Xi'an Jiaotong University, Xi'an, China, in 2007, and the Ph.D. degree in Weapon Science and Technology from Northwestern Polytechnical University, Xi'an, in 2015. Since 2001, he has been with the College of Electrics and Information, Xi'an Polytechnic University, Xi'an. His main research interests include underwater robot location, cooperative localization, person re-identification, and small target detection.