Learning a no-reference quality metric for single-image super-resolution

Chao Ma [a,b,1], Chih-Yuan Yang [c], Xiaokang Yang [b], Ming-Hsuan Yang [c,*]

[a] *Australia Center for Robotic Vision, The University of Adelaide, Adelaide, Australia*
[b] *Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, PR China*
[c] *Electrical Engineering and Computer Science, University of California, Merced, CA, 95344, United States*

ARTICLE INFO

ABSTRACT

Numerous single-image super-resolution algorithms have been proposed in the literature, but few studies address the problem of performance evaluation based on visual perception. While most super-resolution images are evaluated by full-reference metrics, the effectiveness is not clear and the required ground-truth images are not always available in practice. To address these problems, we conduct human subject studies using a large set of super-resolution images and propose a no-reference metric learned from visual perceptual scores. Specifically, we design three types of low-level statistical features in both spatial and frequency domains to quantify super-resolved artifacts, and learn a two-stage regression model to predict the quality scores of super-resolution images without referring to ground-truth images. Extensive experimental results show that the proposed metric is effective and efficient to assess the quality of super-resolution images based on human perception.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Single-image super-resolution (SR) algorithms aim to construct a high-quality high-resolution (HR) image from a single low-resolution (LR) input. Numerous single-image SR algorithms have been recently proposed for generic images that exploit priors based on edges (Sun et al., 2008), gradients (Kim and Kwon, 2010; Shan et al., 2008), neighboring interpolation (Irani and Peleg, 1991; Timofte et al., 2014), regression (Dong et al., 2014), and patches (Dong et al., 2011; Farsiu et al., 2004; Glasner et al., 2009; Schulter et al., 2015; Sun et al., 2011; Timofte et al., 2013; Yang and Yang, 2013; Yang et al., 2013; 2010). Most SR methods focus on generating sharper edges with richer textures, and are usually evaluated by measuring the similarity between super-resolved HR and ground-truth images through full-reference metrics such as the mean squared error (MSE), peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index (Wang et al., 2004). In our recent SR benchmark study (Yang et al., 2014), we show that the information fidelity criterion (IFC) (Sheikh et al., 2005) performs favorably among full-reference metrics for SR performance evaluation. However, full-reference metrics are originally designed to account for image signal and noise rather than human visual perception (Girod, 1993), even for several recently proposed methods . We present 9 example SR images generated from a same LR image in Fig. 1. Table 1 shows that those full-reference metrics fail to match visual perception of human subjects well for SR performance evaluation. In addition, full-reference metrics require ground-truth images for evaluation which are often unavailable in practice. The question how we can effectively evaluate the quality of SR images based on visual perception still remains open. In this work, we propose to learn a no-reference metric for evaluating the performance of single-image SR algorithms. It is because no-reference metrics are designed to mimic visual perception (i.e., learned from large-scale perceptual scores) without requiring ground-truth images as reference. With the increase of training data, no-reference metrics have greater potential to match visual perception for SR performance evaluation.

We first conduct human subject studies using a large set of SR images to collect perceptual scores. With these scores for training, we propose a novel no-reference quality assessment algorithm that matches visual perception well. Our work, in essence, uses the same methodology as that of general image quality assessment (IQA) approaches. However, we evaluate the effectiveness of the signal reconstruction by SR algorithms rather than analyzing noise and distortions (e.g., compression and fading) as in existing IQA

---

* Corresponding author.
 E-mail addresses: c.ma@adelaide.edu.au (C. Ma), yangchihyuan@gmail.com (C.-Y. Yang), xkyang@sjtu.edu.cn (X. Yang), mhyang@ucmerced.edu (M.-H. Yang).

(a) Bicubic interpolation　　(b) Back projection (BP) [4]　　(c) Shan08 [2]

(d) Glasner09 [7]　　(e) Yang10 [8]　　(f) Dong11 [9]

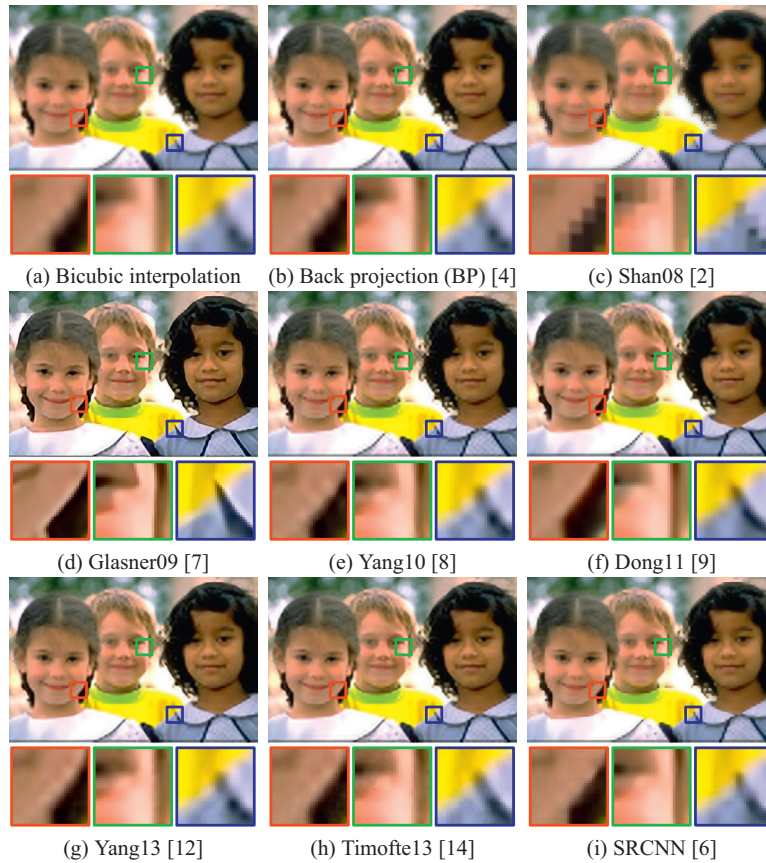(g) Yang13 [12]　　(h) Timofte13 [14]　　(i) SRCNN [6]

Fig. 1. SR images generated from the same LR image using (1) ($s = 4, \sigma = 1.2$). The quality scores of these SR images are compared in Table 1. The images are best viewed on a high-resolution displayer with an adequate zoom level, where each SR image is shown with at least 320×480 pixels (full-resolution).

Table 1

Quality scores of SR images in Fig. 1 from human subjects, the proposed metric, rescaled PSNR, SSIM and IFC (0 for worst and 10 for best). Note that human subjects favor Dong11 over Glasner09 as the SR image in Fig. 1(d) is over-sharpened (best viewed on a high-resolution displayer). However, the PSNR, SSIM and IFC metrics show opposite results as the image in Fig. 1(f) is misaligned to the reference image by 0.5 pixel. In contrast, the proposed metric matches visual perception well.

	Bicubic	BP	Shan08	Glasner09	Yang10	Dong11	Yang13	Timofte13	SRCNN
Subject	5.65	6.13	4.13	**4.78**	5.48	**6.10**	6.55	6.63	6.72
Proposed	3.79	4.20	2.93	**4.96**	5.58	**5.75**	5.80	6.05	6.36
PSNR	2.52	2.61	1.59	**2.23**	2.68	**1.16**	2.82	2.83	2.83
SSIM	7.50	7.65	6.84	**7.59**	7.69	**6.68**	7.83	7.83	7.81
IFC	4.72	4.85	4.02	**4.61**	4.97	**3.42**	5.16	5.13	5.24

methods (Moorthy and Bovik, 2010; 2011; Saad et al., 2012; Tang et al., 2011, 2014; Ye et al., 2012). We quantify SR artifacts based on their statistical properties in both spatial and frequency domains, and regress them to collected perceptual scores. Experimental results demonstrate the effectiveness of the proposed no-reference metric in assessing the quality of SR images against existing IQA measures.

The main contributions of this work are summarized as follows. First, we propose a novel no-reference IQA metric, which matches visual perception well, to evaluate the performance of SR

algorithms. Second, we develop a large-scale dataset of SR images and conduct human subject studies on these images. We make the SR dataset with collected perceptual scores publicly available at https://sites.google.com/site/chaoma99/sr-metric.

2. Related work and problem context

The problem how to evaluate the SR performance can be posed as assessing the quality of super-resolved images. Numerous metrics for general image quality assessment have been used to eval-

uate SR performance in the literature. According to whether the ground-truth HR images are referred, existing metrics fall into the following three classes.

2.1. Full-reference metrics

Full reference IQA methods such as the MSE, PSNR, and SSIM indices (Wang et al., 2004) are widely used in the SR literature (Dong et al., 2011; Kim and Kwon, 2010; Shan et al., 2008; Sun et al., 2011; Yang and Yang, 2013; Yang et al., 2013; 2010). However, these measures are developed for analyzing generic image signals and do not match human perception (e.g., MSE) (Girod, 1993). Reibman et al. (2006) conduct subject studies to examine the limitations of SR performance in terms of scaling factors using a set of three images and existing metrics. Subjects are given two SR images each time and asked to select the preferred one. The perceptual scores of the whole test SR images are analyzed with the Bradley–Terry model (Handley, 2001). The results show that while SSIM performs better than others, it is still not correlated with visual perception well. In our recent SR benchmark work (Yang et al., 2014), we conduct subject studies in a subset of generated SR images, and show that the IFC (Sheikh et al., 2005) metric performs well among full-reference measures. Since subject studies are always time-consuming and expensive, Reibman et al. use only six ground-truth images to generate test SR images while we use only 10 in Yang et al. (2014). It is therefore of great importance to conduct larger subject study to address the question how to effectively evaluate the performance of SR algorithms based on visual perception.

2.2. Semi-reference metric

In addition to the issues on matching visual perception, full-reference metrics can only be used for assessment when the ground-truth images are available. Some efforts have been made to address this problem by using the LR input images as references rather than the HR ground-truth ones, which do not always exist in real-world applications. Yeganeh et al. (2012) extract two-dimension statistical features in the spatial and frequency domains to compute assessment scores from either a test LR image or a generated SR image. However, only 8 images and 4 SR algorithms are analyzed in their work. Our experiments with a larger number of test images and SR algorithms show that this method is less effective due to the lack of holistic statistical features.

2.3. No-reference metrics

When the ground-truth images are not available, SR images can be evaluated by the no-reference IQA methods (Moorthy and Bovik, 2010; 2011; Saad et al., 2012; Tang et al., 2011) based on the hypothesis that natural images possess certain statistical properties, which are altered in the presence of distortions (e.g., noise) and this alternation can be quantified for quality assessment. In Ye et al. (2012) and Tang et al. (2014), features learned from auxiliary datasets are used to quantify the natural image degradations as alternatives of statistical properties. Existing no-reference IQA methods are all learning-based, but the training images are degraded by noise, compression or fast fading rather than super-resolution. As a result, the state-of-the-art no-reference IQA methods are less effective accounting for the artifacts such as incorrect high-frequency details introduced by SR algorithms. On the other hand, since SR images usually contain blur and ringing artifacts, the proposed algorithm bears some resemblance to existing metrics for blur and sharpness estimation (Cho et al., 2010; Ferzli and Karam, 2009; Liu et al., 2013). While the most significant difference lies in that we focus on SR images, where numerous arti-

Table 2
The scaling factors (s) in our experiments with their corresponding kernel width values (σ).

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

facts are introduced by more than one blur kernel. In this work, we propose a novel no-reference metric for SR image quality assessment by learning from perceptual scores based on subject studies involving a large number of SR images and algorithms.

3. Human subject studies

We use the Berkeley segmentation dataset (Martin et al., 2001) to carry out the experiments as the images are diverse and widely used for SR evaluation (Glasner et al., 2009; Sun et al., 2011; Yang and Yang, 2013). For an HR source image I_h, let s be a scaling factor, and the width and height of I_h be $s \times n$ and $s \times m$. We generate a downsampled LR image I_l as follows:

$$I_l(u, v) = \sum_{x,y} k(x - su, y - sv)I_h(x, y), \qquad (1)$$

where $u \in \{1, \ldots, n\}$ and $v \in \{1, \ldots, m\}$ are indices of I_l, and k is a matrix of Gaussian kernel weight determined by a parameter σ, e.g., $k(\Delta x, \Delta y) = \frac{1}{Z}e^{-(\Delta x^2 + \Delta y^2)/2\sigma^2}$, where Z is a normalization term. Compared to our benchmark work (Yang et al., 2014), we remove the noise term from (1) to reduce uncertainty. The quality of the super-resolved images from those LR images are used to evaluate the SR performance. In this work, we select 30 ground truth images from the BSD200 dataset (Martin et al., 2001) according to the PSNR values. In order to obtain a representative image set that covers a wide range of high-frequency details, we compute the PSNR values as the quality scores of the SR images generated from the LR images using (1) with a scaling factor (s) of 2 and a Gaussian kernel width (σ) of 0.8 by the bicubic interpolation algorithm. The selected 30 images are evenly divided into three sets as shown in Fig. 2.

The LR image formation of (1) can be viewed as a combination of a downsampling and a blurring operation which is determined by the scaling factor s and kernel width σ, respectively. As subject studies are time-consuming and expensive, our current work focuses on large differences caused by scaling factors, which are critical to the quality assessment of SR images. We focus on how to effectively quantify the upper bound of SR performance based on human perception. Similar to Yang et al. (2014), we assume the kernel width is known, and compute the mean PSNR values of the SR images generated by 9 SR methods under various settings ($s \in \{2, 3, 4, 5, 6, 8\}$ and $\sigma \in \{0.4, 0.6, \ldots, 2\}$) using 30 ground truth images. Fig. 3 shows that the larger subsampling factor requires larger blur kernel width for better performance. We thus select an optimal σ for each scaling factor (s) as shown in Table 2.

In the subject studies, we use absolute rating scores rather than pairwise comparison scores as we have 1620 test images, which would require millions of pairwise comparisons (i.e., $C_2^{1620} \approx 1.3M$). Although the sampling strategy (Ye and Doermann, 2014) could alleviate this burden partially, pairwise comparison is infeasible given the number of subjects, images and time constraints. We note that subject studies in Sheikh et al. (2006) and Yang et al. (2014) are also based on absolute rating. In this work, we develop a user interface (See Fig. 4) to collect perceptual scores for these SR images. At each time, we simultaneously show 9 images generated from one LR image by different SR algorithms on a high-resolution display. These images are displayed in random order to reduce bias caused by correlation of image contents. Subjects are asked to give scores from 0 to 10 to indicate image quality
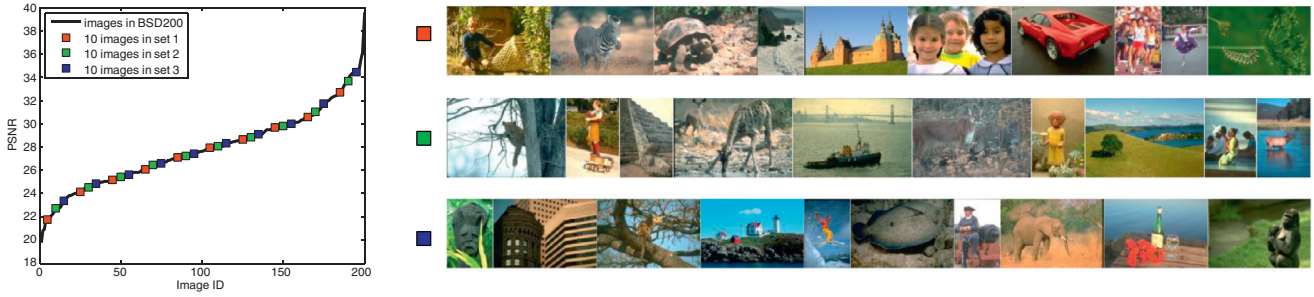
Fig. 2. Ranked PSNR values on the BSD200 dataset and the evenly selected three sets of images. The PSNR values indicate the quality scores of the SR images generated from the LR images using (1) with scaling factor (s) of 2 and Gaussian kernel width (σ) of 0.8 by the bicubic interpolation algorithm.
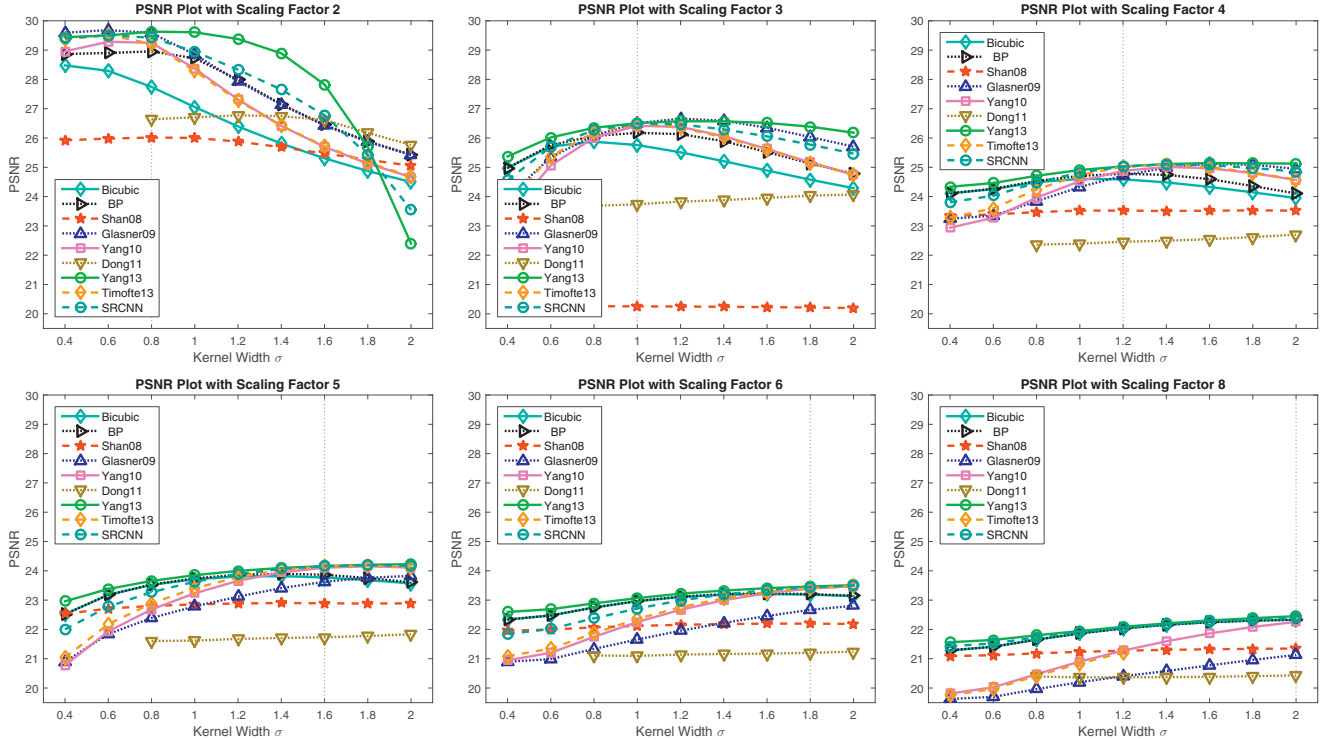

Fig. 3. Distribution of mean PSNR on the selected images. Note the increasing trend of the kernel width along the increase of the scale factor to generate the peak PSNR values. The SR algorithm Dong11 does not converge when $\sigma < 0.8$. The vertical dash line highlights the optimal kernel width for each scaling factor.

Table 3
Empirical quality scores on SR images generated by bicubic interpolation. GT indicates the ground-truth HR images.

s	GT	2	3	4	5	6	8
Score ($\approx$)	10	8–9	5–7	4–6	3–5	2–4	< 2

based on visual preference. We divide the whole test into 3 sections evenly such that subjects can take a break after each section and keep high attention span in our studies. To reduce the inconsistency among the individual quality criterion, we design a training process to conduct the test at the beginning of each section, i.e., giving subjects an overview of all the ground-truth and SR images generated by bicubic interpolation with the referred scale of quality scores as shown in Table 3.

We collect 50 scores from 50 subjects for each image, and compute the perceptual quality index as the mean of the median 40 scores to remove outliers. To the best of our knowledge, our subject study is the largest so far in terms of SR images, algorithms, and subject scores (See Table 4). In addition to using more images than Yang et al. (2014), we present subjects color SR images for

evaluation as we observe that monochrome SR images introduce larger individual bias as demonstrated in Fig. 5(a). It is reasonable that gray-scale images are rear in daily life and subjects hold different quality criterion. Fig. 5(b) shows that the mean perceptual scores are more stable after removing outliers.

Fig. 6 shows the computed mean perceptual quality indices in terms of scaling factor and kernel width. From the human subject studies, we have the following observations. First, the performance rank of 9 SR algorithms remains the same (i.e., the curves are similar) across all images in Fig. 6(a)–(f), which shows consistency of perceptual scores on evaluating SR algorithms. Second, the performance rank changes with scaling factors, e.g., Glasner09 outperforms Bicubic with higher perceptual scores in Fig. 6(a) while it is the opposite in Fig. 6(c). Since the image quality degradation caused by scaling factors is larger than that by different SR methods, the statistical properties for quantifying SR artifacts have to be discriminative to both scaling variations and SR algorithms. Third, SR results generated from LR images with more smooth contents have higher perceptual scores, e.g., the score of the image in Fig. 7(a) is higher than that of Fig. 7(b). This may be explained by the fact that visual perception is sensitive to edges and tex-
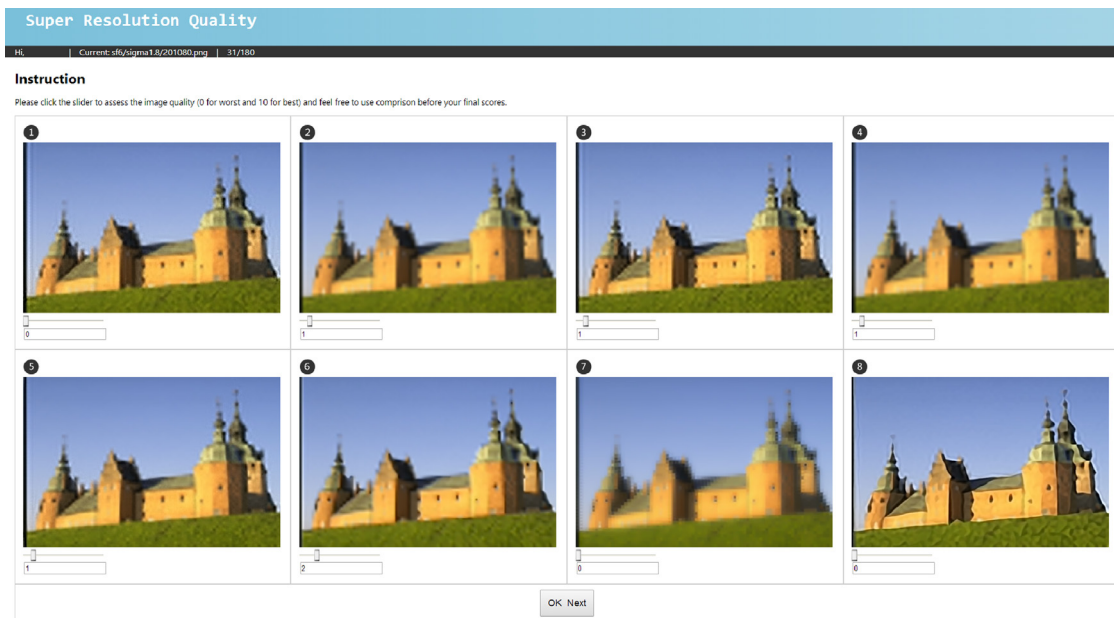
Fig. 4. One screenshot of human subject study. Subjects assign scores between 0 and 10 to displayed SR images. Test images are randomly presented in order to reduce bias caused by similarity of image contents.

Table 4
Data sets used for image quality assessment based on subject studies.

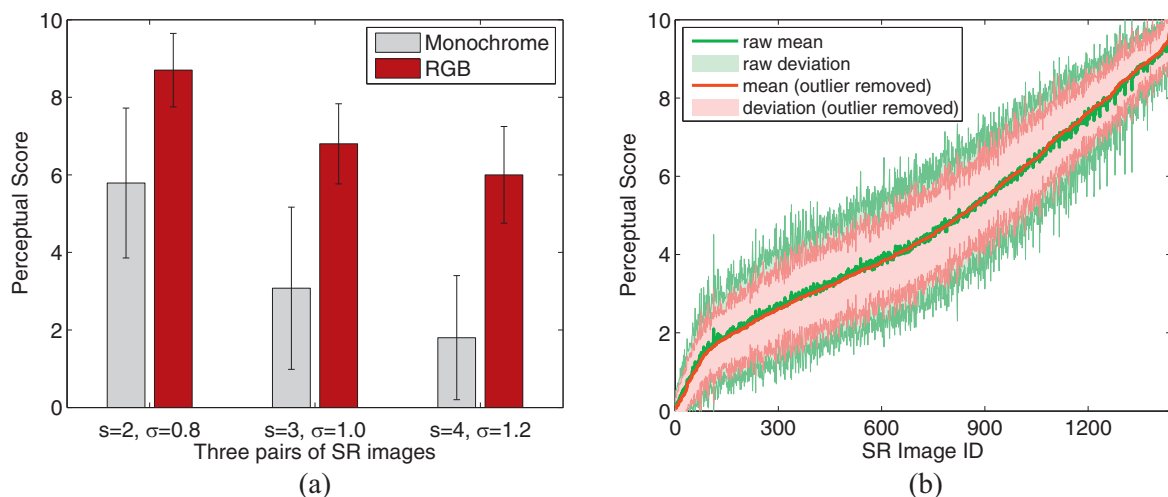| Dataset	# Reference images	# Distortions	# Subject scores
LIVE (Sheikh et al., 2006)	29	982	22,457
ASQA (Ye and Doermann, 2014)	20	120	35,700
SRAB (Yang et al., 2014)	10	540	16,200
Our study	30	1620	81,000

Fig. 5. (a) Deviation of 50 perceptual scores on three pairs of SR images generated by bicubic interpolation from the same test image in Fig. 1. (b) Sorted mean perceptual scores and deviations before and after removing outliers. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

tures and most algorithms do not perform well for images such as Fig. 7(b).

4. Proposed algorithm

We exploit three types of statistical properties as features, including local and global frequency variations and spatial discontinuity, to quantify artifacts and assess the quality of SR images. Each set of statistical features is computed on a pyramid to alleviate the scale sensitivity of SR artifacts. Fig. 8 shows the main steps of the proposed algorithm for learning no-reference quality metric. Fig. 9 shows an overview of the statistical properties of each type of features.

4.1. Local frequency features

The statistics of coefficients from the discrete cosine transform (DCT) have been shown to effectively quantify the degree and type of image degradation (Srivastava et al., 2003), and used for natural image quality assessment (Saad et al., 2012). Since SR im-
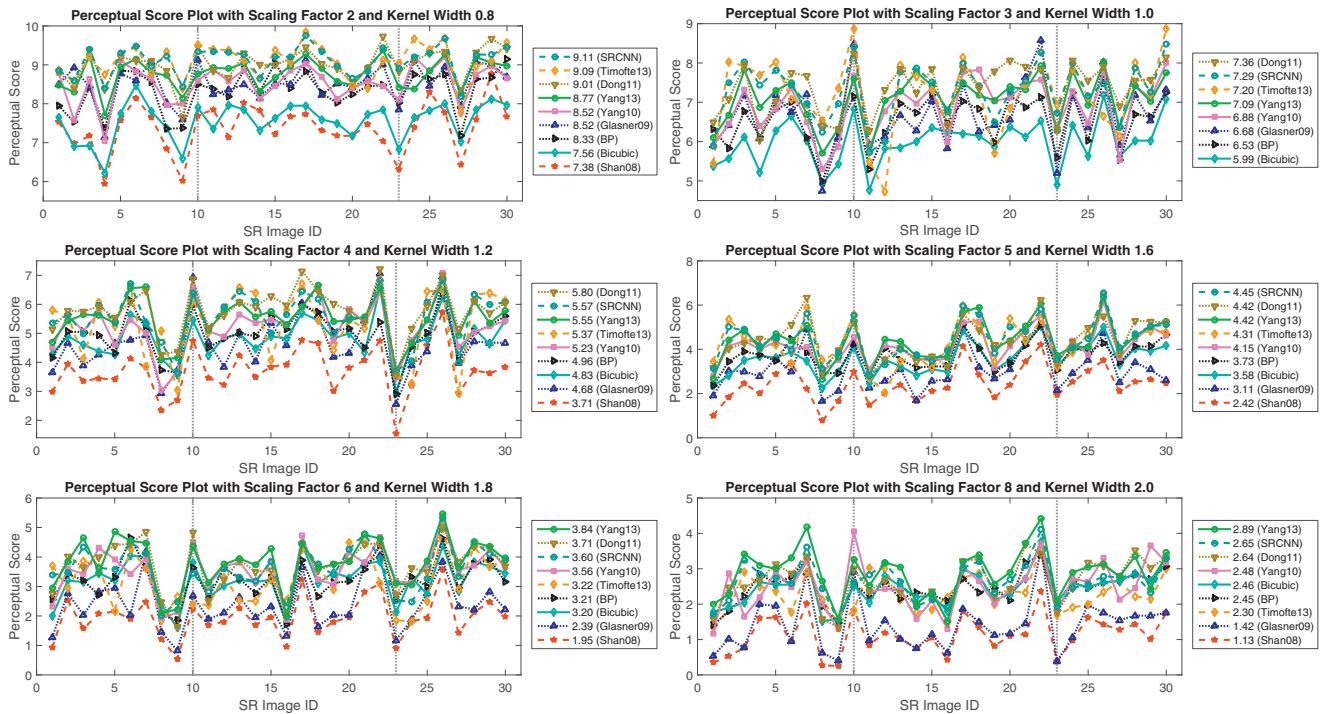
Fig. 6. Perceptual scores of SR images under 6 pairs of scaling factor (s) and kernel width (σ). The performance rank of SR algorithms remains relatively consistent, even while score values change under different scaling factors and kernel widths. The average perceptual scores of each SR algorithm are shown in the legend (Shan08 with $s = 3$, $\sigma = 1.0$ is excluded as the SR images contain severe noise and their perceptual scores are close to 0).

(a) Image ID 10　　　　　　　　　　　　　　　　(b) Image ID 23

Fig. 7. The scores in Fig. 6 indicated by vertical dash lines for the SR images generated from (a) are much higher than that of (b).
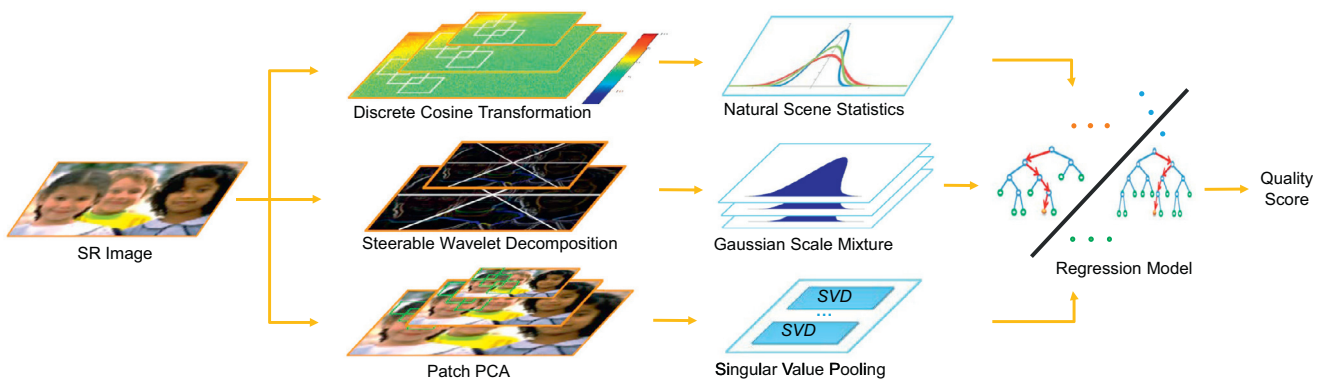

Fig. 8. Main steps of the proposed no-reference metric. For each input SR image, statistics computed from the spatial and frequency domains are used as features to represent SR images. Each set of extracted features are trained in separate ensemble regression trees, and a linear regression model is used to predict a quality score by learning from a large number of visual perceptual scores.
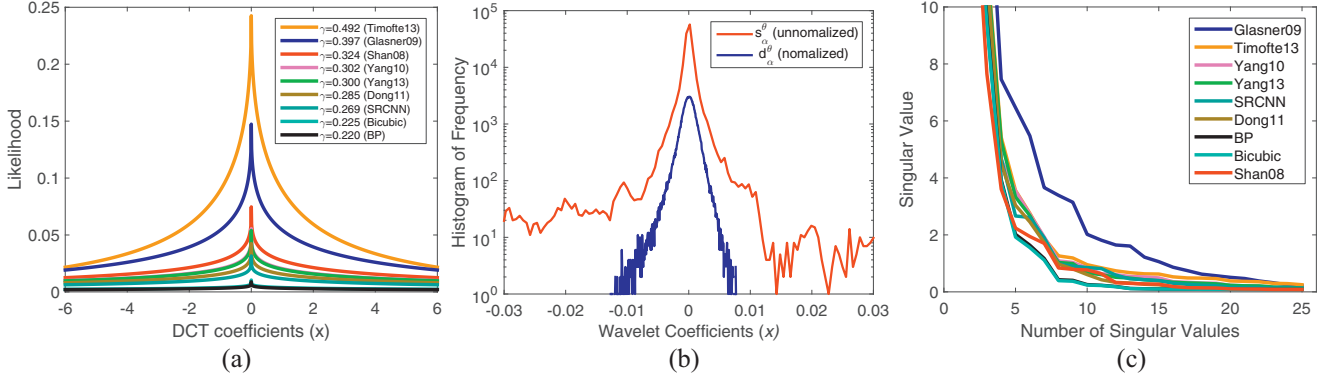
Fig. 9. (a) Estimated GGD distribution of the normalized DCT coefficient in the first block of the images from Fig. 1. Note that the shape parameter γ effectively characterizes the distribution difference between SR algorithms (μ is disregarded). (b) Wavelet coefficient distribution in one subband. The GSM makes the distribution of subband more Gaussian-like (blue). (c) Distribution of patch singular values of SR images in Fig. 1. For SR images generated by Bicubic and BP containing more edge blur (smoothness), their singular values fall off more rapidly. In contrast, Glasner09 strengthens the sharpness and the singular values of its generated SR image decrease more slowly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
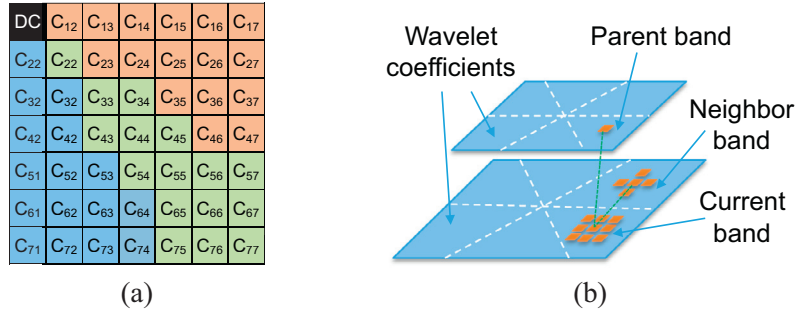

Fig. 10. (a) Three groups of DCT coefficients for one block are shown by different colors. The DC coefficients are excluded. (b) $N = 15$ neighboring filters. 3×3 adjacent positions in the current band, 5 locations in the neighboring band and 1 from the parent band. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ages are generated from LR inputs, the task can be considered as a restoration of high-frequency components on LR images. To quantify the high-frequency artifacts introduced by SR restoration, we propose to transform SR images into the DCT domain and fit the DCT coefficients by the generalized Gaussian distribution (GGD) as in Saad et al. (2012).

$$f(x|\mu, \gamma) = \frac{1}{2\Gamma(1 + \gamma^{-1})} e^{-(|x-\mu|^\gamma)}, \tag{2}$$

where μ is the mean of the random variable x, γ is the shape parameter and $\Gamma(\cdot)$ is the gamma function, e.g., $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$. We observe that the shape factor γ is more discriminative than the mean value μ to characterize the distribution of DCT coefficients (See Fig. 9(a)). We thus select the value of γ as one statistical feature to describe SR images. Let σ be the standard deviation of a DCT block, we use $\bar{\sigma} = \frac{\sigma}{\mu}$ to describe the perturbation within one block. We further group DCT coefficients of each block into three sets (See Fig. 10(a)) and compute the normalized deviation $\bar{\sigma}_i$ ($i = 1, 2, 3$) of each set and their variation Σ of $\{\bar{\sigma}_i\}$ as features. As all the statistics are computed on individual blocks, large bias is likely to be introduced if these measures are simply concatenated. We thus pool those block statistics and use the mean values to represent each SR image. To increase their discriminative strength, we add the first and last 10% pooled variations as features.

4.2. Global frequency features

The global distribution of the wavelet coefficients of one SR image might not be fitted well by a specific distribution (e.g., GGD). We sort to the Gaussian scale mixture (GSM) model, which shows effective in describing the marginal and joint statistics of natural images (Moorthy and Bovik, 2011; Wainwright and Simoncelli, 2000) using a set of neighboring wavelet bands. An N-dimensional random vector Y belongs to a GSM if $Y \equiv z \cdot U$, where $\equiv$ denotes equality in probability distribution, and U is a zero-mean Gaussian random vector with covariance Q. The variable z is a non-negative mixing multiplier. The density of Y is given by an integral as

$$p_Y(y) = \int \frac{1}{(2\pi)^{N/2}|z^2 Q|^{1/2}} e^{\left(-\frac{Y^T Q^{-1} Y}{2z^2}\right)} p_z(z) dz, \tag{3}$$

where $p_z(\cdot)$ is the probability of the mixing variable z. We first apply the steerable pyramid decomposition (Simoncelli et al., 1992) on an SR image to generate neighboring wavelet coefficients. Compared to Wainwright and Simoncelli (2000) and Moorthy and Bovik (2011), we apply the decomposition in both the real and imaginary domains rather than only in the real domain. We observe that the wavelet coefficients in the complex domain have better discriminative strength. As shown in Fig. 10(b), we assume that N (e.g., $N = 15$) filters in neighborhoods that share a mixer estimated by $\hat{z} = \sqrt{Y^T Q^{-1} Y/N}$. Such estimation is identical to divisive normalization (Moorthy and Bovik, 2011; Wang et al., 2004) and makes the probability distribution of wavelet band more Gaussian-like (See Fig. 9(b)). Let d_α^θ be the normalized wavelet subband with scale α and orientation θ. We estimate the shape parameter γ using (2) on d_α^θ and concatenated bands d^θ across scales. In addition, we compute the structural correlation (Moorthy and Bovik, 2011; Wang et al., 2004) between high-pass response and their band-pass counterparts to measure the global SR artifacts. Specifically, the band-pass and high-pass responses are filtered across-scale by a 15×15 Gaussian window with kernel width $\sigma = 1.5$. The struc-
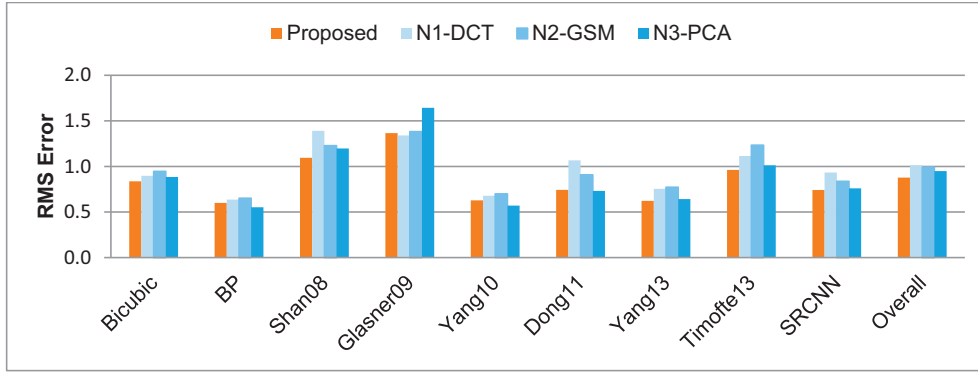
(a) 5-fold cross validation
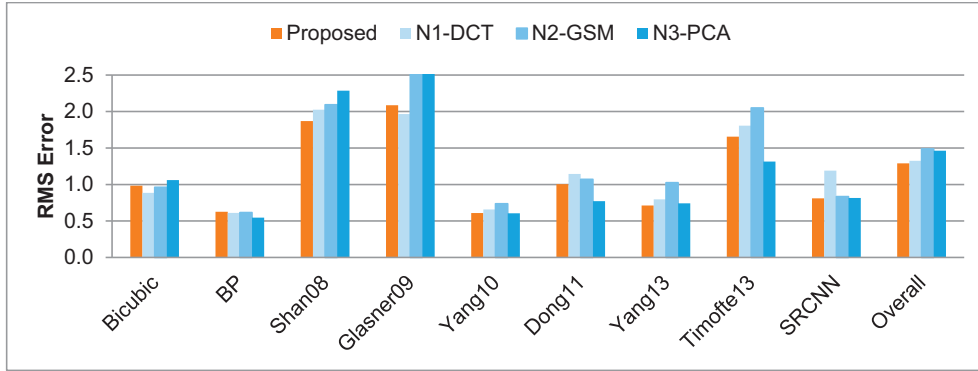
(b) Leaving 6 reference images out
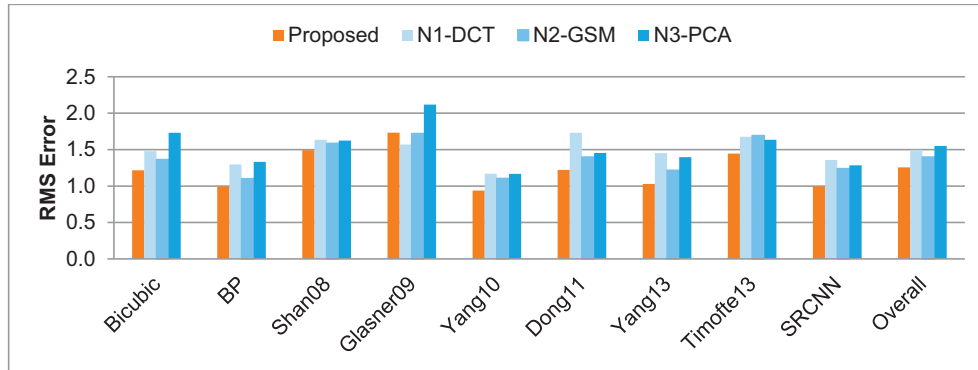
(c) Leaving 2 SR methods out

Fig. 11. Root-mean-square error between the estimated score and the subjective score (measures with smaller values are closer to human visual perception) using 3 validation schemes. Note that the proposed two-stage regression model (orange bar) on three types of low-level features (blue bar) reduces the error between perceptual scores significantly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tural correlation is computed by $\rho = \frac{2\sigma_{xy} + c_0}{\sigma_x^2 + \sigma_y^2 + c_0}$, where σ_{xy} is the cross-covariance between the windowed regions; σ_x as well as σ_y are their windowed variances; and c_0 is a constant for stabilization.

4.3. Spatial features

Since the spatial discontinuity of pixel intensity is closely related to perceptual scores for SR images in subject studies (See Fig. 6), we model this property in a way similar to Yeganeh et al. (2012). We extract features from patches rather than pixels to increase discriminative strength. We apply principal component analysis (PCA) on patches and use the corresponding singular values to describe the spatial discontinuity.

Singular values of images with smooth contents are squeezed to zero more rapidly than for those with sharp contents (as they correspond to less significant eigenvectors). Fig. 9(c) shows the singular values of SR images generated from Bicubic and BP fall off more rapidly as the generated contents tend to be smooth.

4.4. Two-stage regression model

We model the features of local frequency, global frequency and spatial discontinuity with three independent regression forests (Breiman, 2001; Criminisi et al., 2011). Their outputs are linearly regressed on perceptual scores to predict the quality of evaluated SR images. Let x_n ($n = 1, 2, 3$) denote one type of low-level features, and y be the perceptual scores of SR images. The jth node of the tth decision tree ($t = 1, 2, \ldots, T$) in the forest is
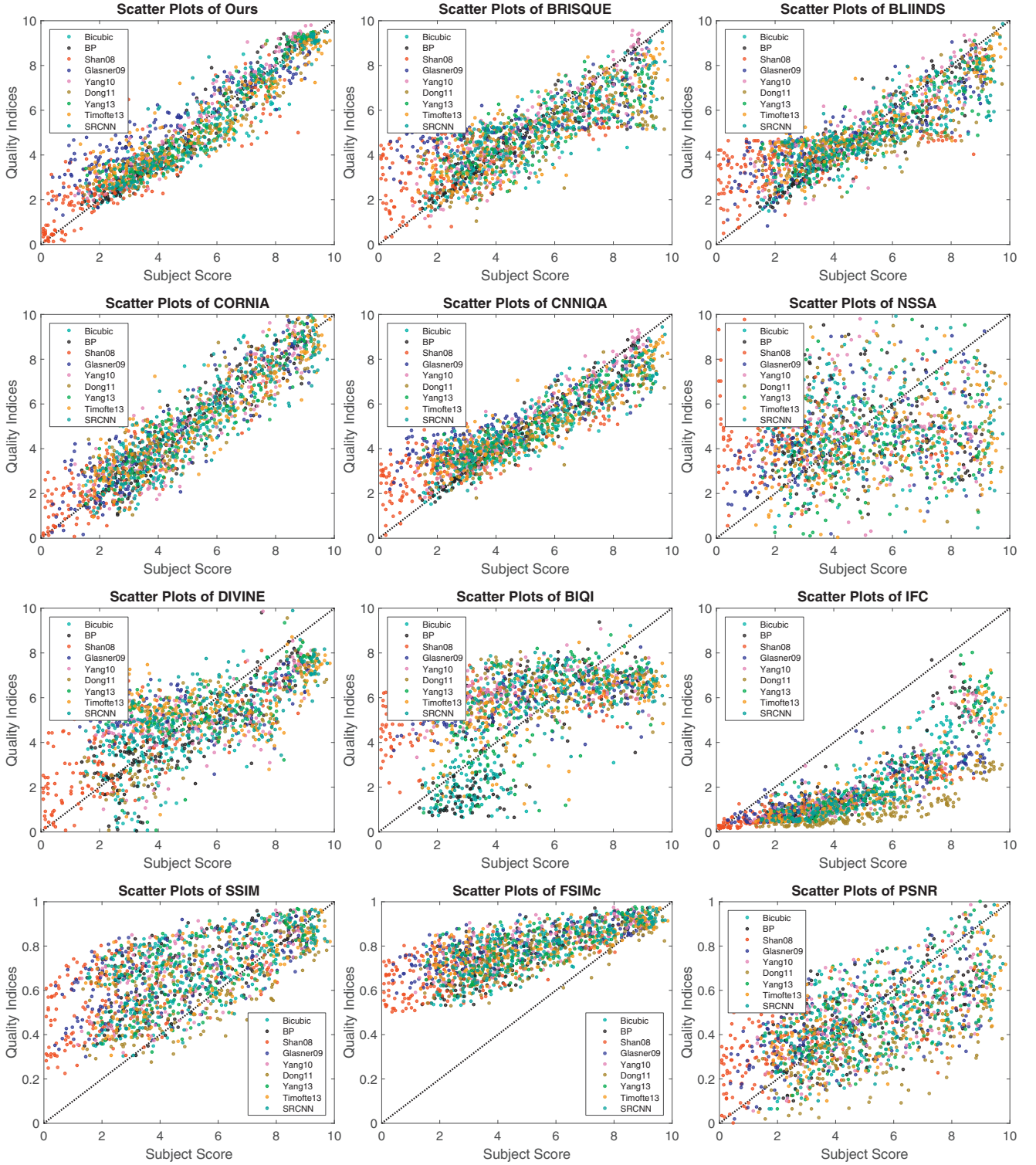
Fig. 12. Quality indices generated by different methods to perceptual scores. The proposed metric and other no-reference baseline methods (except DIVINE and BIQI) are leaned under 5-fold cross validation. A metric matches visual perception well if the distribution is compact and spreads out along the diagonal.

learned as:

$$\theta_j^{n*} = \arg\max_{\theta_j^n \in \mathcal{T}_j} I_j^n, \tag{4}$$

where $\mathcal{T}_j$ controls the size of a random subset of training data to train node j. The objective function I_j^n is defined as:

$$I_j^n = \sum_{x_n \in \mathcal{S}_j} \log(|\Lambda_y(x_n)|) - \sum_{i \in \{L,R\}} \Big(\sum_{x_n \in \mathcal{S}_j^i} \log(|\Lambda_y(x_n)|) \Big) \tag{5}$$

with Λ_y the conditional covariance matrix computed from probabilistic linear fitting, where $\mathcal{S}_j$ denotes the set of training data arriving at node j, and $\mathcal{S}_j^L$, $\mathcal{S}_j^R$ the left and right split sets. We refer readers to Criminisi et al. (2011) for more details about regression forest. The predicted score $\hat{y}_n$ is thus computed by averaging the outputs of T regression trees as:

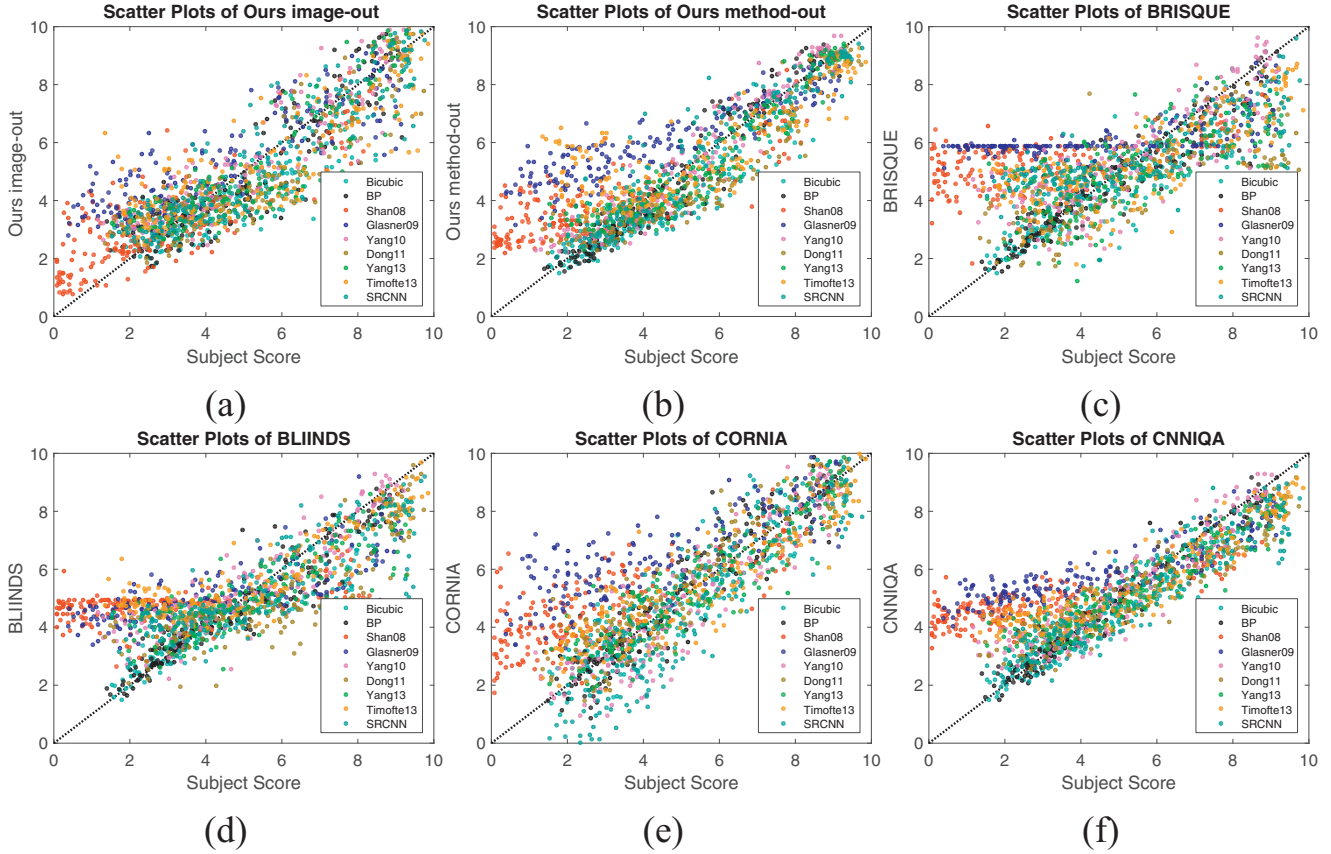$$\hat{y}_n = \frac{1}{T} \sum_t^T p_t(x_n | \Theta). \tag{6}$$

Fig. 13. Quality indices generated by different methods to perceptual scores. (a) and (b): Proposed metric under *leave-image-out* and *leave-method-out* validation schemes. (c)–(f): Original baseline no-reference algorithms without retraining on our SR dataset. The proposed metric under two challenging validations still performs well against state-of-the-art metrics. A metric matches visual perception well if the distribution is compact and spread out along the diagonal.

Consequently, we linearly regress the outputs from all three types of features to perceptual scores, and estimate the final quality score as $\hat{y} = \sum_n \lambda_n \cdot \hat{y}_n$, where the weight λ is learned by minimizing

$$\lambda^* = \operatorname{argmin}_\lambda (\sum_n \lambda_n \cdot \hat{y}_n - y)^2. \tag{7}$$

5. Experimental validation

In the human subject studies, we generate 1620 SR images from 180 LR inputs using 9 different SR algorithms, and collect their perceptual scores from 50 subjects. The mean of the median 40 subject scores is used as perceptual score. We randomly split the dataset into 5 sets, and recursively select one set for test and the remaining for training. After this loop, we obtain the quality scores estimated by the proposed metric for all SR images. We then compare the Spearman rank correlation coefficients between the predicted quality scores and perceptual scores. In addition to the 5-fold cross validation, we split the training and test sets according to the reference images and SR methods to verify the generality of the proposed metric. Given that there are 30 reference images and 9 SR methods, we leave 6 reference images or 2 methods out in each experiment. Several state-of-the-art no-reference IQA methods and 4 most widely used full-reference metrics for SR images are included for experimental validation. More results and the source code of the proposed metric can be found at https://sites.google.com/site/chaoma99/sr-metric.

Table 5
List of features used in this work.

Feature domain	Feature description	#
Local frequency	γ (mean, first 10% percentile)	6
	$\bar{\sigma}$ (mean, last 10% percentile)	6
	Σ (mean, last 10% percentile)	6
Global frequency	γ for each band d_α^θ and d^θ	18
	Across-scale correlation	12
	Across-band correlation	15
Spatial discontinuity	Singular values of patches	75
Total		138

5.1. Parameter settings

We use a three-level pyramid on 7×7 blocks of DCT coefficients to compute local frequency features. For steerable pyramid wavelet decomposition, we set α and θ to be 2 and 6, respectively. The resulting 12 subbands are denoted by s_α^θ, where $\alpha \in \{1, 2\}$ and $\theta \in \{0°, 30°, 60°, 90°, 120°, 150°\}$. We set the number N of neighboring filters to 15, i.e., 3×3 adjacent positions in the current band, 5 adjacent locations in the neighboring band and 1 from the parent band share a mixer (See Fig. 10(b)). For spatial discontinuity, we compute singular values on 5×5 patches on a three-level pyramid. We list the detailed feature information in Table 5. We vary the parameter T of regression trees from 100 to 5000 with a step of 50 and find the proposed algorithm performs best when T is set to 2000.
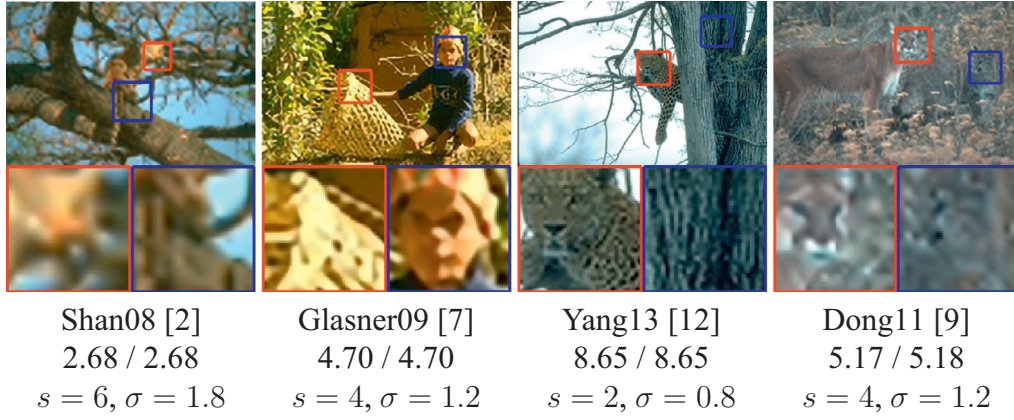
Fig. 14. Four best cases using the proposed metric to evaluate the quality of SR images under the 5-fold cross validation. The left / right values under each image are the predicted score and the perceptual score respectively.

Fig. 15. Four worst cases using the proposed metric to evaluate the quality of SR images under the 5-fold cross validation. The left / right values under each image are the predicted score and the perceptual score respectively.
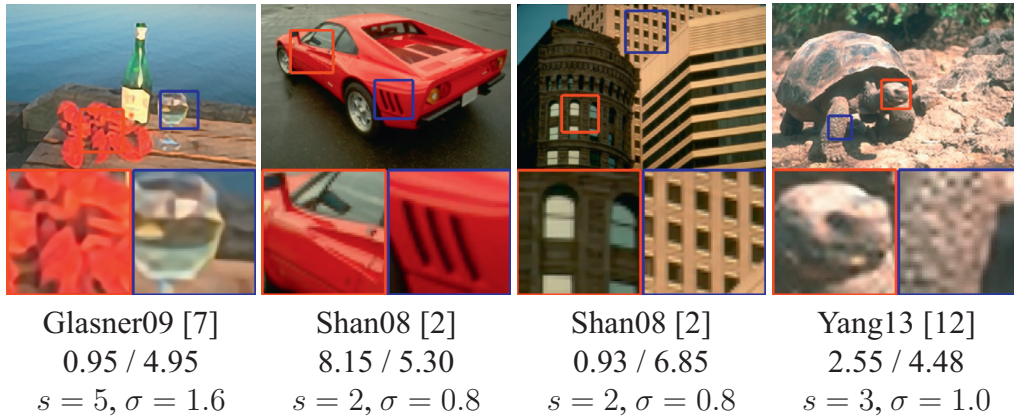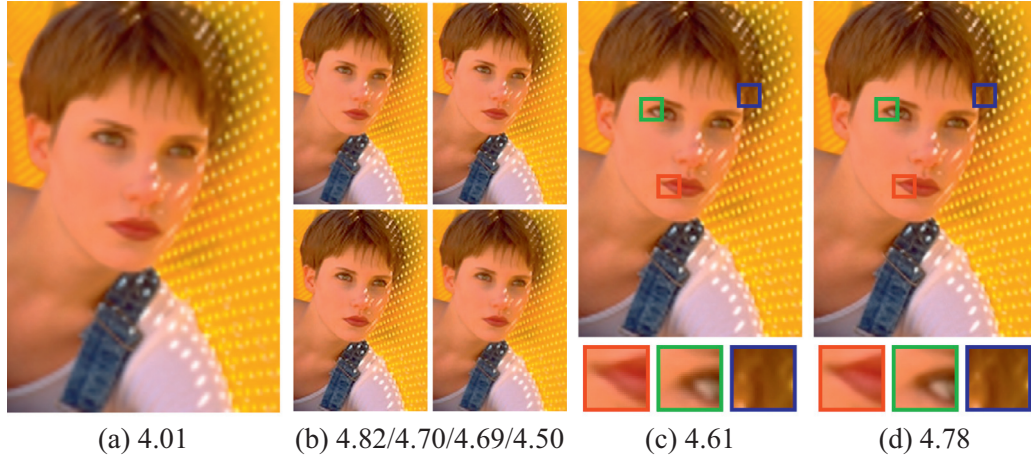

Fig. 16. Perception guided SR results (best viewed on a high-resolution displayer) with quality scores predicted by the proposed metric. (a) Input LR image ($s = 4, \sigma = 1.8$). (b) Selected best SR images with the Dong11, Yang13, Timofte13 and Yang10 methods using the proposed metric. (c) 3×3 grid integration. (d) Pixel-level integration.

5.2. Quantitative validations

We run the proposed measure 100 times in each validation and choose the mean values as the estimated quality scores. We compare the contribution of each feature type using root-mean-square errors (RMSEs) in Fig. 11.

The small overall error values, 0.87 in (a) and less than 1.4 in (b) and (c) compared to the score range (0–10), indicate the ef-

fectiveness of the proposed method by linearly combining three types of statistical features. In addition, we carry out an ablation study replacing the random forest regression (RFR) by the support vector regression (SVR) on each type of features. The SVR model is widely used in existing no-reference image quality metrics (Mittal et al., 2012; Moorthy and Bovik, 2010; 2011; Saad et al., 2012; Ye et al., 2012). Table 6 shows that RFR is more robust to the outliers than SVR on each type of features or a simple concatenation of
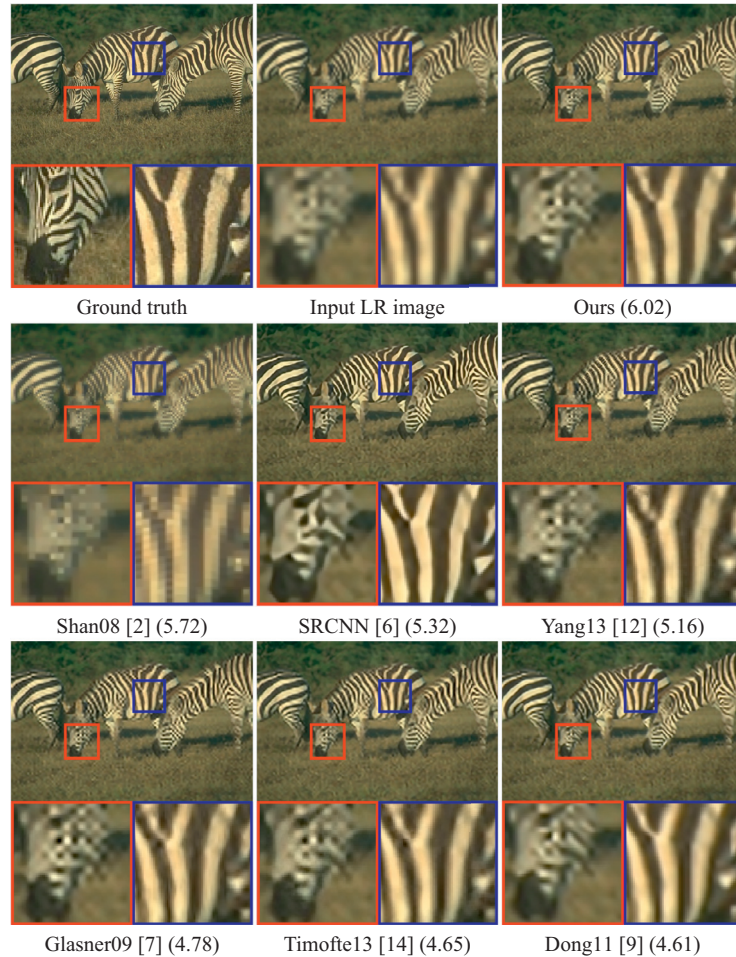
Fig. 17. Visual comparison of SR results. The input low resolution images are generated using (1) with $s = 4$ and $\sigma = 1.2$. We show the best 6 results based on their quality scores in parentheses predicted by the proposed metric, and select the best 4 algorithms to integrate our SR results.

Table 6
Spearman rank correlation coefficients (Hogg et al., 2005) (metric with higher coefficient matches perceptual score better). The random forest regression (RFR) uniformly performs better than the support vector regression (SVR) for each type of features or the concatenation (-con) of three type of features. The proposed two-stage regression approach (-all) combining three types of features improves the accuracy for both RFR and SVR. Bold: best; underline: second best.

	Ours	RFR-con	RFR-DCT	RFR-GSM	RFR-PCA	SVR-all	SVR-con	SRV-DCT	SVR-GSM	SVR-PCA
Bicubic	**0.933**	0.922	0.910	0.898	<u>0.923</u>	0.851	0.772	0.630	0.713	0.862
BP	**0.966**	<u>0.962</u>	0.956	0.952	**0.966**	0.881	0.876	0.776	0.838	0.889
Shan08	**0.891**	<u>0.887</u>	0.830	0.870	0.874	0.504	0.373	0.499	0.522	0.044
Glasner09	**0.931**	<u>0.926</u>	0.911	0.897	0.878	0.841	0.717	0.766	0.685	0.599
Yang10	<u>0.968</u>	0.961	0.954	0.948	**0.969**	0.929	0.905	0.874	0.834	0.877
Dong11	<u>0.954</u>	0.946	0.922	0.929	**0.960**	0.885	0.892	0.792	0.883	0.874
Yang13	**0.958**	<u>0.955</u>	0.937	0.932	**0.958**	0.898	0.855	0.801	0.770	0.874
Timofte13	**0.930**	<u>0.928</u>	0.911	0.880	0.927	0.883	0.814	0.859	0.628	0.839
SRCNN	**0.949**	0.938	0.917	0.936	<u>0.945</u>	0.866	0.853	0.778	0.816	0.843
Overall	**0.931**	<u>0.921</u>	0.909	0.913	<u>0.921</u>	0.752	0.696	0.711	0.616	0.663

three types of features. The proposed two stage-regression model effectively exploits three types of features and performs best.

For fair comparisons, we generate the IQA indices from 11 state-of-the-art methods including: (1) six no-reference metrics: BRISQUE (Mittal et al., 2012), BLIINDS (Saad et al., 2012), DIVINE (Moorthy and Bovik, 2011), BIQI (Moorthy and Bovik, 2010), CORNIA (Ye et al., 2012), and CNNIQA (Kang et al., 2014); (2) one semi-reference metric: NSSA (Yeganeh et al., 2012); and (3) four full-reference metrics: IFC (Sheikh et al., 2005), SSIM (Wang et al., 2004), FSIM (Zhang et al., 2011), and PSNR. As the no-reference metrics are originally designed to measure image degradations, e.g., noise, compression and fading, rather than for SR evalua-

tion, we retrain them on our SR dataset using the same validation schemes. Note that both the DIVINE and BIQI metrics apply intermediate steps to estimate specific types of image degradations (Sheikh et al., 2006) for image quality assessment. However, SR degradation is not considered in any type of degradations in Sheikh et al. (2006). We directly regress the features generated by DIVINE and BIQI methods to the perceptual scores but this approach is not effective as the quality scores for different SR images are almost the same. We thus report the original results using the DIVINE and BIQI indices without retraining on our dataset. We empirically tune the parameters to obtain best performance during retraining. The NSSA metric is designed for evaluating SR images.
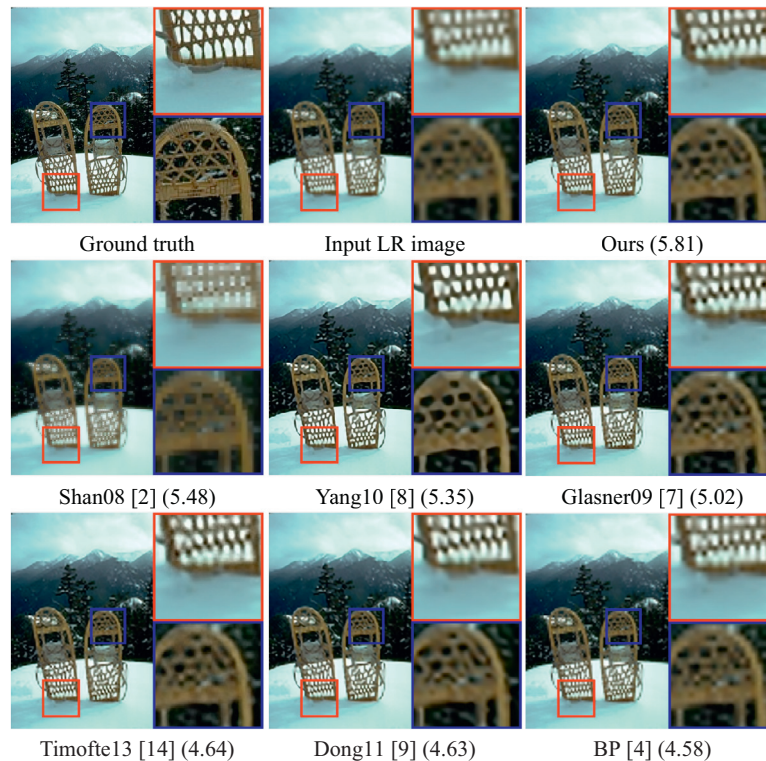
Fig. 18. Visual comparison of SR results. The input low resolution images are generated using (1) with $s = 4$ and $\sigma = 1.2$. We show the best 6 results based on their quality scores in innermost parentheses predicted by the proposed metric, and select the best 4 algorithms to integrate our SR results.

The other four full-reference metrics are widely used in SR evaluation although they are not designed for SR. Fig. 12 shows the correlation between subjective scores and IQA indices. Tables 7–9 quantitatively compares the Spearman rank correlation coefficients. In addition, we compare the original results of BRISQUE, BLIINDS, CORNIA, and CNNIQA in Table 10 and Fig. 13. Without retraining on our SR dataset, these metrics generally perform worse. This shows the contributions of this work by developing a large-scale SR image dataset and carrying out large-scale subject studies on these SR images. Note that we do not present the results of NSSA in Table 10 and Fig. 13 as the learned data file of the NSSA metric is not publicly available.

5.3. Discussion

As shown in Tables 7–9 and Fig. 12, the proposed method performs favorably against the state-of-the-art IQA methods, e.g., the overall quantitative correlation with perceptual scores is 0.931 under 5-fold cross validation. The leave-image-out and leave-method-out validations are more challenging since they take into account the independence of image contents and SR algorithms. In the leave-image-out setting, the training and test sets do not contain SR images generated from the same reference image. In the leave-method-out setting, the SR images in training and test sets are generated by different SR algorithms. Tables 8 and 9 show that the proposed metric performs well against existing IQA methods in these two challenging validations. Note that the proposed metric performs best in the 5-fold cross validation as it learns from perceptual scores and favors prior information from image contents and SR algorithms for training.

The six evaluated no-reference IQA metrics, BRISQUE, BLIINDS, DIVINE, BIQI, CORNIA, and CNNIQA, are not originally designed for SR. We retrain them (except DIVINE and BIQI) on our own SR dataset. For DIVINE and BIQI, we present the reported results as

the performance of these methods by retraining on our dataset is significantly worse. The reason is that these two metrics apply intermediate steps to quantify specific image distortions in Sheikh et al. (2006) rather than SR. Table 7 shows that for most SR algorithms, the DIVINE or BIQI metrics do not match human perception well. The retrained BRISQUE and BLIINDS metrics perform well against DIVINE and BIQI. We note that some of the features used by the BRISQUE and BLIINDS metrics are similar to the proposed DCT and GSM features. However, both BRISQUE and BLIINDS metrics are learned from one support vector regression (SVR) model (https://www.csie.ntu.edu.tw/~cjlin/libsvm/), which are less robust to the outliers of perceptual scores than the random forest regression (RFR) model. Fig. 12 shows that their quality scores scatter more than close to the diagonal. The CORNIA method learns a codebook from an auxiliary dataset (Larson and Chandler, 2010) containing various image distortions. The coefficients of densely sampled patches from a test image are computed based on the codebook as features. Table 7 shows that the CORNIA metric achieves second best results among all the baseline algorithms. The proposed metric performs favorably against CORNIA due to the effective two-stage regression model based on RFRs. While CORNIA only relies on one single SVR. The CNNIQA metric uses convolutional neural network to assess the image quality, however, it does not perform as well as the proposed method. The can be explained by insufficient amount of training examples. Overall, the proposed method exploits both global and local statistical features specifically designed to account for SR artifacts. Equipped with a novel two-stage regression model, i.e., three independent random forests are regressed on extracted three types of features and their outputs are linearly regressed with perceptual scores, our metric is more robust to outliers than the compared IQA methods, which are based on one single regression model (e.g., SVR or CNN).

Although the semi-reference NSSA method is designed for evaluating SR images and extracts both frequency and spatial features,

Table 7
Spearman rank correlation coefficients (Hogg et al., 2005) (metric with higher coefficient matches perceptual score better). The compared no-reference metrics are re-trained on our SR dataset using the 5-fold cross validation. The proposed metric performs favorably against state-of-the-art methods. Bold: best; underline: second best.

	Ours	BRISQUE (Mittal et al., 2012)	BLIINDS (Saad et al., 2012)	CORNIA (Ye et al., 2012)	CNNIQA (Kang et al., 2014)	NSSA (Yeganeh et al., 2012)	DIVINE (Moorthy and Bovik, 2011)	BIQI (Moorthy and Bovik, 2010)	IFC (Sheikh et al., 2005)	SSIM (Wang et al., 2004)	FSIM (Zhang et al., 2011)	PSNR
Bicubic	**0.933**	0.850	0.886	0.889	0.926	−0.007	0.784	0.770	0.884	0.588	0.706	0.572
BP	**0.966**	0.917	0.931	0.932	0.956	0.022	0.842	0.740	0.880	0.657	0.770	0.620
Shan08	0.891	0.667	0.664	**0.907**	0.832	−0.128	0.653	0.254	0.934	0.560	0.648	0.564
Glasner09	**0.931**	0.738	0.862	0.918	0.914	0.325	0.426	0.523	0.890	0.648	0.778	0.605
Yang10	**0.968**	0.886	0.901	0.908	0.943	0.036	0.525	0.556	0.866	0.649	0.757	0.625
Dong11	**0.954**	0.783	0.811	0.912	0.921	0.027	0.763	0.236	0.865	0.649	0.765	0.634
Yang13	**0.958**	0.784	0.864	0.923	0.927	0.168	0.537	0.646	0.870	0.652	0.768	0.631
Timofte13	**0.930**	0.843	0.903	0.911	0.924	0.320	0.122	0.563	0.881	0.656	0.756	0.620
SRCNN	**0.949**	0.812	0.843	0.898	0.908	0.165	0.625	0.617	0.885	0.660	0.780	0.645
Overall	**0.931**	0.802	0.853	0.919	0.904	0.076	0.589	0.482	0.810	0.635	0.747	0.604

it does not perform well as shown in Fig. 12 and Tables 7–9. This is because the features used in the NSSA method are two-dimension coefficients and their regressor is based on a simple linear model. The quality indices computed by weight-averaging two coefficients are less effective for evaluating the quality of SR images generated by the state-of-the-art SR methods.

For the cases when ground truth HR images are available, the proposed method performs favorably against four widely used full-reference quality metrics including PSNR, SSIM (Wang et al., 2004), IFC (Sheikh et al., 2005), and FSIM (Zhang et al., 2011). The PSNR metric performs poorly since the pixel-wise difference measurement does not effectively account for the difference in visual perception (See Table 7 and Fig. 12). For example, an SR image with slight misalignment from the ground truth data appears similarly in terms of visual perception, but the PSNR value decreases significantly. The SSIM method performs better than PSNR as it aims to mimic human vision and computes perceptual similarity between SR and ground truth images by using patches instead of pixels. However, the SSIM metric favors the false sharpness on the SR images generated by Shan08 and Glasner09 and overestimates the corresponding quality scores as shown in Fig. 12. The FSIM metric is less effective in evaluating the SR performance either. The IFC method is also designed to match visual perception and generally performs well for SR images (Yang et al., 2014). Nonetheless, its indices are less accurate for some SR images (Fig. 12). This can be explained by the fact that the IFC metric is limited by local frequency features. In other words, the IFC metric does not take global frequency and spatial properties into account, and fails to distinguish them. Thus it may underestimate the quality of SR images (See the dots cluster below the diagonal in the last sub figure of Fig. 12).

We present four best and worse cases using our metric with 5-fold cross validation to predict the quality of SR images in Fig. 14 and Fig. 15. The reasons that cause the worst cases in Fig. 15 can be explained by several factors. For the first, third and fourth SR images, the proposed metric gives low quality scores due to the fact that human subjects do not always favor oversharp SR images (see also the discussion in Table 1 in the manuscript). For the second image, the richer high-frequency contents affect the proposed metric to compute the high score.

Overall, the proposed metric performs favorably against the state-of-the-art methods, which can be attributed to two reasons. First, the proposed metric uses three sets of discriminative low-level features from the spatial and frequency domains to describe SR images. Second, an effective two-stage regression model is more robust to outliers for learning from perceptual scores collected in our large-scale subject studies. In contrast, existing methods neither learn from perceptual scores nor design effective features with focus on representing SR images. The proposed metric is implemented in Matlab on a machine with an Intel i5-4590 3.30 GHz CPU and 32GB RAM. We report the average run time (in seconds) as follows, ours: 13.31, BRISQUE: 0.14, BLIINDS: 23.57, DIVINE: 9.51, BIQI: 1.21, CORNIA: 3.02, CNNIQA: 12.68, NSSA: 0.28, IFC: 0.61, SSIM: 0.13, FSIM: 0.18, and PSNR: 0.02.

6. Perception guided super-resolution

Given an LR input image, we can apply different SR algorithms to reconstruct HR images and use the proposed metric to automatically select the best result. Fig. 1 shows such an example where the SR image generated by the Timofte13 method has the highest quality score using the proposed metric (See Fig. 1(i)) and is thus selected as the HR restoration output. Equipped with the proposed metric, we can also select the best local regions from multiple SR images and integrate them into a new SR image. Given a test LR image, we apply aforementioned 9 SR algorithms to generate 9 SR images. We first divide each of them into a 3 × 3 grid of regions.

Table 8

Spearman rank correlation coefficients (Hogg et al., 2005) (metric with higher coefficient matches perceptual score better). The compared metrics are retrained on our SR dataset under the leave-image-out validation. Bold: best; underline: second best.

	Ours	BRISQUE (Mittal et al., 2012)	BLIINDS (Saad et al., 2012)	CORNIA (Ye et al., 2012)	CNNIQA (Kang et al., 2014)	NSSA (Yeganeh et al., 2012)
Bicubic	**0.805**	0.423	0.522	<u>0.761</u>	0.736	0.093
BP	**0.893**	0.539	0.476	<u>0.873</u>	0.853	−0.046
Shan08	<u>0.800</u>	0.442	0.474	**0.832**	0.742	0.048
Glasner09	**0.867**	0.277	0.399	<u>0.859</u>	0.803	0.023
Yang10	**0.904**	0.625	0.442	0.843	<u>0.867</u>	0.012
Dong11	**0.875**	0.527	0.411	0.819	<u>0.849</u>	−0.101
Yang13	**0.885**	0.575	0.290	<u>0.843</u>	0.841	0.108
Timofte13	<u>0.815</u>	0.500	0.406	**0.828**	0.740	−0.035
SRCNN	**0.904**	0.563	0.383	0.827	<u>0.850</u>	0.042
Overall	**0.852**	0.505	0.432	<u>0.843</u>	0.799	0.017

Table 9

Spearman rank correlation coefficients (Hogg et al., 2005) (metric with higher coefficient matches perceptual score better). The compared metrics are retrained on our SR dataset under the leave-method-out validation. Bold: best; underline: second best.

	Ours	BRISQUE (Mittal et al., 2012)	BLIINDS (Saad et al., 2012)	CORNIA (Ye et al., 2012)	CNNIQA (Kang et al., 2014)	NSSA (Yeganeh et al., 2012)
Bicubic	<u>0.932</u>	0.850	0.929	0.893	**0.941**	0.036
BP	<u>0.967</u>	0.934	0.953	0.938	**0.971**	0.021
Shan08	**0.803**	0.534	0.471	<u>0.799</u>	0.767	-0.087
Glasner09	**0.913**	0.677	0.805	0.817	<u>0.883</u>	0.393
Yang10	**0.965**	0.834	0.895	0.914	<u>0.930</u>	-0.054
Dong11	**0.932**	0.774	0.780	0.917	<u>0.920</u>	-0.062
Yang13	**0.944**	0.716	0.845	<u>0.911</u>	0.906	0.147
Timofte13	0.774	0.760	<u>0.849</u>	**0.898**	0.845	0.382
SRCNN	**0.933**	0.771	0.806	<u>0.908</u>	0.890	0.149
Overall	**0.848**	0.644	0.763	<u>0.809</u>	0.797	0.053

Table 10

Spearman rank correlation coefficients (Hogg et al., 2005) (metric with higher coefficient matches perceptual score better). The compared no-reference metrics are not retrained on our SR dataset. Bold: best; underline: second best.

	Ours 5-fold CV	Ours image-out	Ours method-out	BRISQUE Mittal et al. (2012)	BLIINDS (Saad et al., 2012)	CORNIA (Ye et al., 2012)	CNNIQA (Kang et al., 2014)
Bicubic	**0.933**	0.805	<u>0.932</u>	0.850	0.929	0.893	0.927
BP	<u>0.966</u>	0.893	**0.967**	0.934	0.953	0.938	0.931
Shan08	**0.891**	0.800	<u>0.803</u>	0.534	0.471	0.799	0.842
Glasner09	**0.931**	0.867	0.913	0.677	0.805	0.817	<u>0.896</u>
Yang10	**0.968**	0.904	<u>0.965</u>	0.834	0.895	0.914	0.938
Dong11	**0.954**	0.875	0.932	0.774	0.780	0.917	<u>0.936</u>
Yang13	**0.958**	0.885	<u>0.944</u>	0.716	0.845	0.911	0.934
Timofte13	**0.930**	0.815	0.774	0.760	0.849	0.898	<u>0.906</u>
SRCNN	**0.949**	0.904	0.933	0.771	0.806	<u>0.908</u>	0.924
Overall	**0.931**	<u>0.852</u>	0.848	0.644	0.763	0.809	0.833

We compute their quality scores based on the proposed metric and stitch the best regions to generate a new SR image (See Fig. 16(c)). For better integration, we densely sample overlapping patches of 11×11 pixels. We then apply the proposed metric on each patch and compute an evaluation score of each pixel of that SR image. For each patch, we select the one from all results with highest quality scores and stitch all the selected patches together using the graph cut and Poisson blending (Pérez et al., 2003) method (See Fig. 16(d)). It is worth noting that the proposed metric can be used to select SR regions with high perceptual scores from which a high-quality HR image is formed. Fig. 17 and Fig. 18 show two more pixel-level integrated SR results, which retain most edges and render smooth contents as well. The integrated SR results effectively exploit the merits of state-of-the-art SR algorithms, and show better visual quality.

7. Conclusion

In this paper, we propose a novel no-reference IQA algorithm to assess the visual quality of SR images by learning perceptual scores collected from large-scale subject studies. The proposed metric regress three types of low-level statistical features extracted from SR images to perceptual scores. Experimental results demonstrate that the proposed metric performs favorably against state-of-the-art quality assessment methods for SR performance evaluation.

References

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Cho, T.S., Joshi, N., Zitnick, C.L., Kang, S.B., Szeliski, R., Freeman, W.T., 2010. A content-aware image prior. CVPR.
Criminisi, A., Shotton, J., Konukoglu, E., 2011. Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Microsoft Research Technical Report MSR-TR-2011-114. (28).
Dong, C., Loy, C.C., He, K., Tang, X., 2014. Learning a deep convolutional network for image super-resolution. In: ECCV, pp. 184–199.

Dong, W., Zhang, L., Shi, G., Wu, X., 2011. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. TIP 20 (7), 1838–1857.

Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P., 2004. Advances and challenges in super-resolution. Int. J. Imag. Syst. Technol. 14 (2), 47–57.

Ferzli, R., Karam, L.J., 2009. A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). TIP 18 (4), 717–728.

Girod, B., 1993. What's wrong with mean-squared error? In: Digital Images and Human Vision. MIT Press, pp. 207–220.

Glasner, D., Bagon, S., Irani, M., 2009. Super-resolution from a single image. ICCV.

Handley, J.C., 2001. Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment. IS&T PICS.

Hogg, R., McKean, J., Craig, A., 2005. Introduction to Mathematical Statistics. Pearson Education.

Irani, M., Peleg, S., 1991. Improving resolution by image registration. CVGIP 53 (3), 231–239.

Kang, L., Ye, P., Li, Y., Doermann, D.S., 2014. Convolutional neural networks for no-reference image quality assessment. In: CVPR, pp. 1733–1740.

Kim, K.I., Kwon, Y., 2010. Single-image super-resolution using sparse regression and natural image prior. TPAMI 32 (6), 1127–1133.

Larson, E.C., Chandler, D.M., 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. J. Electron. Imag. 19 (1), 011006.

Liu, Y., Wang, J., Cho, S., Finkelstein, A., Rusinkiewicz, S., 2013. A no-reference metric for evaluating the quality of motion deblurring. ACM Trans. Graph. 32 (6), 175.

Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. ICCV.

Mittal, A., Moorthy, A.K., Bovik, A.C., 2012. No-reference image quality assessment in the spatial domain. TIP 21 (12), 4695–4708.

Moorthy, A.K., Bovik, A.C., 2010. A two-step framework for constructing blind image quality indices. SPL 17 (5), 513–516.

Moorthy, A.K., Bovik, A.C., 2011. Blind image quality assessment: from natural scene statistics to perceptual quality. TIP 20 (12), 3350–3364.

Pérez, P., Gangnet, M., Blake, A., 2003. Poisson image editing. ACM Trans. Graph. 22 (3), 313–318.

Reibman, A.R., Bell, R.M., Gray, S., 2006. Quality assessment for super-resolution image enhancement. ICIP.

Saad, M.A., Bovik, A.C., Charrier, C., 2012. Blind image quality assessment: a natural scene statistics approach in the dct domain. TIP 21 (8), 3339–3352.

Schulter, S., Leistner, C., Bischof, H., 2015. Fast and accurate image upscaling with super-resolution forests. In: CVPR, pp. 3791–3799.

Shan, Q., Li, Z., Jia, J., Tang, C., 2008. Fast image/video upsampling. ACM Trans. Graph. 27 (5), 153.

Sheikh, H.R., Bovik, A.C., de Veciana, G., 2005. An information fidelity criterion for image quality assessment using natural scene statistics. TIP 14 (12), 2117–2128.

Sheikh, H.R., Sabir, M.F., Bovik, A.C., 2006. A statistical evaluation of recent full reference image quality assessment algorithms. TIP 15 (11), 3440–3451.

Simoncelli, E.P., Freeman, W.T., Adelson, E.H., Heeger, D.J., 1992. Shiftable multiscale transforms. TIT 38 (2), 587–607.

Srivastava, A., Lee, A.B., Simoncelli, E.P., Zhu, S.-C., 2003. On advances in statistical modeling of natural images. J. Math. Imaging Vis. 30 (6–7), 17–33.

Sun, J., Sun, J., Xu, Z., Shum, H.-Y., 2011. Gradient profile prior and its applications in image super-resolution and enhancement. TIP 20 (6), 1529–1542.

Sun, J., Xu, Z., Shum, H., 2008. Image super-resolution using gradient profile prior. CVPR.

Tang, H., Joshi, N., Kapoor, A., 2011. Learning a blind measure of perceptual image quality. CVPR.

Tang, H., Joshi, N., Kapoor, A., 2014. Blind image quality assessment using semi-supervised rectifier networks. CVPR.

Timofte, R., Smet, V.D., Gool, L.J.V., 2013. Anchored neighborhood regression for fast example-based super-resolution. ICCV.

Timofte, R., Smet, V.D., Gool, L.J.V., 2014. A+: adjusted anchored neighborhood regression for fast super-resolution. In: ACCV, pp. 111–126.

Wainwright, M.J., Simoncelli, E.P., 2000. Scale mixtures of Gaussians and the statistics of natural images. NIPS.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. TIP 13 (4), 600–612.

Yang, C.-Y., Ma, C., Yang, M.-H., 2014. Single image super-resolution: a benchmark. ECCV.

Yang, C.-Y., Yang, M.-H., 2013. Fast direct super-resolution by simple functions. ICCV.

Yang, J., Lin, Z., Cohen, S., 2013. Fast image super-resolution based on in-place example regression. CVPR.

Yang, J., Wright, J., Huang, T.S., Ma, Y., 2010. Image super-resolution via sparse representation. TIP 19 (11), 2861–2873.

Ye, P., Doermann, D., 2014. Active sampling for subjective image quality assessment. CVPR.

Ye, P., Kumar, J., Kang, L., Doermann, D.S., 2012. Unsupervised feature learning framework for no-reference image quality assessment. CVPR.

Yeganeh, H., Rostami, M., Wang, Z., 2012. Objective quality assessment for image super-resolution: a natural scene statistics approach. ICIP.

Zhang, L., Zhang, L., Mou, X., Zhang, D., 2011. FSIM: a feature similarity index for image quality assessment. TIP 20 (8), 2378–2386.