ORIGINAL ARTICLE

Full-reference image quality metric for blurry images and compressed images using hybrid dictionary learning

Zihan Zhou[1] · Jing Li[2] · Yong Xu[1] · Yuhui Quan[1]

Abstract

The image quality degradation due to the loss of high-frequency components of images is often seen in real scenarios, such as artifacts caused by image compression and image blur caused by camera shake or out of focus. Quantifying such degradation is very useful for many tasks that are related to image quality. In this paper, an effective approach is proposed for the image quality assessment on images with blur as well as images with compression artifacts. Based on the relation between the dictionaries of the degraded image and the reference image, we build up a hybrid dictionary learning model to characterize the space of patches of the reference image as well as that of the degraded image. The image quality is then measured by the difference between the two resulting dictionaries. Combined with a simple sparse-coding-based metric, the proposed method shows competitive performance on five benchmark datasets, which demonstrates its effectiveness.

1 Introduction

Image degradation is a common phenomenon in the real world during the acquisition, transmission and processing of images. In many scenarios, the image degradation is caused by the loss of components of high frequency. For instance, the image blur caused by camera shake or defocus can be considered passing the image to a low-pass filter which kills and decays the Fourier coefficients in the high-frequency domain. Image compression algorithms often reduce the storage requirement or shorten the transmission time of an image by removing the image details which

correspond to global or local high-frequency components. See Fig. 1 for some examples.

An effective metric that can accurately quantify the changes on images caused by such degradations is certainly welcomed in many applications. Take image processing for example. Such a metric can be used for guiding the processing algorithms or for automatically determining the parameters of an algorithm [30]. Similarly, for image compression, the metric can be used to indicate whether the compression is overdone. Not limited to processing natural images [41], the metric also has important applications in the processing of images from different fields, such as borehole images [39], hyper-spectral images [4, 5] and SAR images [3].

Developing such a metric for computers using some computational models is one goal of objective image quality assessment (IQA). In general, the IQA metrics can be classified into full-reference (FR), reduced-reference (RR) and no-reference (NR), according to the amount of information known from the reference image. A reference image indicates the image with ultra-high quality corresponding to the degraded one, i.e., the reference image has the same content as the degraded one but with very high quality. The reference image or most of its features are fully known in the FR metrics (e.g., [8, 29, 32, 37, 43, 46, 53]) while totally inaccessible in the NR metrics

Zihan Zhou and Jing Li have contributed equally in this work.

✉ Yuhui Quan
   csyhquan@scut.edu.cn

   Zihan Zhou
   cszzh@mail.scut.edu.cn

   Jing Li
   lj225205@alibaba-inc.com

   Yong Xu
   yxu@scut.edu.cn

[1]   South China University of Technology, Guangzhou, China

[2]   Moku Lab, Alibaba Group, Beijing, China

Reference   Defocus   JPEG   JPEG2000

Fig. 1 Example of a reference image and three degraded images with blur or compression. The images are from the TID2008 dataset [23]

(e.g., [14, 17]). The RR metrics (e.g., [19, 48]) strike the balance between FR methods and RR methods, which assume only partial features of the reference image are given. This paper focuses on the FR metrics.

1.1 Motivations

In this paper, we investigate the dictionary learning with sparse coding for FR-IQA, which is motivated biologically as follows. Theoretical studies [1, 22] have shown that human primitive visual cortex uses sparse coding to represent the perceptual information from the external world such as natural scenes. Sparse codes can be considered as an analogue of neurons' responses, while the used dictionary atoms can be regarded as the corresponding active neurons in the retina. Learning an over-complete dictionary with the sparsity prior can imitate the properties, such as localization, orientation, bandpass and sparse activation, of the receptive field of simple cells in the primary visual cortex [22], and provide exact quantitative predictions that are often considered to be consistent with the measurements of the visual cortex [1]. It is also shown in [7] that the response of lateral geniculate nucleus can be accounted by the principle of sparsity and parsimony.

We are also motivated by the recent success of sparse-coding-based approaches in image representations [27, 28], as well as in IQA, e.g., FR-IQA [12, 16, 18] and NR-IQA [14, 17]. Let $\mathbf{y}_1, \ldots, \mathbf{y}_N$ denote the vectorized image patches sampled from the image, $\mathbf{d}_1, \ldots, \mathbf{d}_M$ denote the dictionary atoms, $\mathbf{c}_i$ denotes the coding vector corresponding to $\mathbf{y}_i$ for $i = 1, \ldots, N$. The existing sparse-coding-based approaches for IQA often consider the following model:

$$\mathbf{Y} \approx \mathbf{DC}, \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$, $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_M]$, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$, and $\mathbf{C}$ is assumed to be sparse. The degradation on the image will cause disturbance on both the dictionary atoms and coding vectors. Most approaches [12, 16, 18] fix the dictionary and track the changes of the coding vectors between the reference image and its distorted versions. Such approaches are generally FR, as the coding vectors contain a majority of features of the images. Some approaches [14, 17] also attempted to construct the prior knowledge for NR-IQA based on the statistical properties of the coding vectors. All these approaches only use the

sparse coding matrix $\mathbf{C}$ but without the dictionary $\mathbf{D}$ in the calculation of the final quality score. The dictionary has not been fully exploited yet.

The dictionary indeed can encode the quality degradation. An example is illustrated as follows. Without loss of generality, we consider the 1D case. Let $\mathbf{f} \in \mathbb{R}^N$ denote an image with N pixels. When sampling the patches $\mathbf{y}_1, \ldots, \mathbf{y}_N$ from $\mathbf{f}$ using a sliding window, the model in (1) can be written into the convolutional form under mild conditions [11]:

$$\mathbf{f} \approx \sum_{i=1}^{M} \mathbf{d}_i * \tilde{\mathbf{c}}_i, \tag{2}$$

where $*$ denotes convolution, and $\tilde{\mathbf{c}}_i$ denotes the transpose of the ith row of $\mathbf{C}$. Consider a degradation that removes high-frequency components, which is modeled by the convolution with a low-pass kernel $\mathbf{a}$. The resulting degraded image is given by

$$\mathbf{a} * \mathbf{f} \approx \sum_{i=1}^{M} (\mathbf{a} * \mathbf{d}_i) * \tilde{\mathbf{c}}_i. \tag{3}$$

In other words, the dictionary atoms of the degraded image are equivalent to the degraded versions of dictionary atoms of the reference image when the coefficients are fixed. Thus, the degradation caused by the loss of high-frequency information, such as image blur and image compression, can be well measured based on the difference between the dictionaries of the reference and degraded images. Based on such an observation, we proposed a hybrid dictionary learning model which sequentially learns dictionaries from the reference image and the degraded one.

1.2 Main contributions

A dictionary is essentially a set (of atoms) with permutation ambiguity, i.e., changing the order of columns (atoms) of a dictionary only generates an equivalent one. Such a property makes it difficult to compare two dictionaries. To avoid using set distances for comparing the dictionaries, we add the coding consistency during the dictionary learning on the reference image and distorted image, leading to a constrained hybrid dictionary learning approach. From the computational perspective, the dictionary atoms serve as a frame or basis of the space of patches. Comparing the dictionaries of reference and degraded image patches is to measure the distortion of the space of reference patches due to the degradation.

By using the commutativity of convolution, we can also rewrite (3) into

$$\mathbf{a} * \mathbf{f} \approx \sum_{i=1}^{M} \mathbf{d}_i * (\mathbf{a} * \tilde{\mathbf{c}}_i). \tag{4}$$

It can be seen that when fixing the dictionary, the degradation is equivalent to being applied to the sparse codes. This explains the effectiveness of the existing approaches working on sparse codes. In our method, we also use a plain sparse coding model with a fixed analytic dictionary to utilize the sparse codes for further improvement in IQA. It is worth mentioning that, from the biological perspective, the proposed dictionary-based metric can be viewed as exploiting the behaviors of active neurons when restricting same responses on them, while the sparse-coding-based metric used in the proposed approach as well as other existing approaches is based on the responses of the same neuron.

In summary, our main contribution is a proposal of a dictionary-based approach with sparse coding for the FR-IQA on blurriness and compressed artifacts. Instead of using handcrafted features such as structural features [41], edge features [47] or texture features [3, 25–27, 40], we take full advantage of dictionary learning and sparse coding which are data-driven and learning-based. The proposed approach was tested on five benchmark datasets. The experimental results have demonstrated the effectiveness of the proposed method.

2 Related work

In this section, we first review some general FR-IQA metrics which are designed for various types of distortion such as contrast reduction, noise, blur and even the mixture of multiple degradations, with particular focus on the sparse-coding-based approaches as they are closely related to ours. Then we conduct the literature review on the IQA metrics designed for blur images and compressed images.

2.1 General FR-IQA metrics

Many FR-IQA metrics characterize the structural information of images, such as SSIM [46], MS-SSIM [45], IW-SSIM [43] and FSIM [53], using manually designed features. Such handcrafted features are not adaptive to data. Another line of research is learning structural features from data for FR-IQA. Along this line, the sparse-coding-based methods proposed in recent years, e.g., [12, 16, 18], have emerged as a promising direction. These methods calculate the sparse coefficients of images or image patches under some dictionaries, based on which the visual quality scores are computed.

A pioneer work of sparse-coding-based IQA is conducted in [8], which trained a feature detector using independent component analysis for extracting local image features for IQA. The feature detector can be viewed as a dictionary, and its output can be viewed as the sparse coefficients.

Guha et al. [12] proposed to learn an individual dictionary from each reference image and use it for the sparse coding of the corresponding reference image and distorted images. Such a scheme may be time-consuming since the dictionary learning process is run for every input reference image. To overcome this weakness, Li et al. [16] proposed to pre-learn a universal dictionary from a set of clear natural images rather than learning individual ones. This approach is extended in [18] to further utilize the color cues.

2.2 IQA for blurry images and compressed images

The IQA metrics on blurry images and compressed images are often referred to as sharpness metrics. Most of existing sharpness metrics are NR. Although the degradation of blurry images and compressed images are both caused by high-frequency information loss, the perceptive experiences as well as the statistic characteristics are different on these two types of degradation. Therefore, the NR sharpness metrics are developed for blurry images and compressed images, respectively; see, e.g., [17, 21, 36, 42] and [33, 34, 44].

Regarding the FR approaches for compressed images, one early work was done by Wang et al. [44]. They proposed to measure the level of block artifacts at the edges of square blocks for estimating the JPEG quality. Later, plenty of IQA methods were proposed based on the statistics of natural images. Most these methods are designed for individual compression type, such as JPEG2000 [34] and JPEG [10].

For blurry images, existing FR approaches resort to various statistics properties of clean image and blurry images for the IQA task. Narvekar et al. [21] proposed to model the probability of the blur detection results and then utilize the cumulative probability to estimate image quality. Wang et al. [42] proposed to model the properties of blurred images using the statistics of image gradient magnitudes and calculate the final score through the extreme learning machine. There are also some sparse-coding-based methods in this field. Shi et al. [36] proposed a sparse model to estimate the just noticeable defocus blur. They extract blur-related features based on the numbers of used dictionary atoms in the sparse representation and image decomposition. Li et al. [17] proposed to calculate the energy of sparse codes and the variance of image blocks for measuring the quality of a blurry image.

3 Proposed method

The proposed method is outlined in Fig. 2, which mainly contains two modules: hybrid dictionary learning and sparse coding. The former module fixes the sparse code and
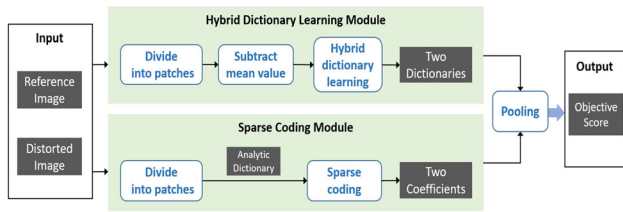
Fig. 2 Outline of the proposed method

learns the dictionary from the reference and distorted images, while the latter part fixes the dictionary and generates the sparse coefficients. The results from these two modules are pooled as the final score. The details of each module will be presented in the following subsections.

Throughout the paper, unless specified, bold upper letters denote matrices, bold lower letters represent vectors, light lower letters for scalars, bold hollow letters for sets, and calligraphic letters for operators or functions. Let $\mathbf{x}(i)$ denote the ith element of $\mathbf{x}$. For $\mathbf{x} \in \mathbb{R}^N$, its ℓ_0 pseudo-norm is denoted by $\|\mathbf{x}\|_0$ and defined as $\#\{j|\mathbf{x}(j) \neq 0\}$, and its ℓ_2 norm is denoted by $\|\mathbf{x}\|_2$. The Frobenius norm of a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ is denoted by $\|\mathbf{X}\|_F$. Let $*, \otimes$ denote the convolution and element-wise multiplication, respectively. The superscripts 'r' and 'd' on matrices or vectors are the abbreviations of *reference* and *distortion*, respectively.

3.1 Hybrid dictionary learning module

The hybrid dictionary learning module is the main contribution of this paper. Given a reference image $\mathbf{I}^r \in \mathbb{R}^{M_1 \times M_2}$, we first sample Z image patches denoted by $\{\mathbf{P}_i^r \in \mathbb{R}^{\sqrt{B} \times \sqrt{B}}\}_{i=1}^Z$ from $\mathbf{I}^r$ using a sliding window with step size S, where $Z = \lfloor \frac{M_1 - \sqrt{B} + 1}{S} \rfloor \times \lfloor \frac{M_2 - \sqrt{B} + 1}{S} \rfloor$, the patch size is $\sqrt{B} \times \sqrt{B}$ and B is a perfect square. On the distorted image $\mathbf{I}^d \in \mathbb{R}^{M_1 \times M_2}$, the same sampling process is done and we can collect Z image patches from $\mathbf{I}^d$, which are denoted by $\{\mathbf{P}_i^d \in \mathbb{R}^{\sqrt{B} \times \sqrt{B}}\}_{i=1}^Z$. The patches $\{\mathbf{P}_i^r\}_{i=1}^Z$ and $\{\mathbf{P}_i^d\}_{i=1}^Z$ are ordered, respectively, such that $\mathbf{P}_i^r$ and $\mathbf{P}_i^d$ correspond to the

same spatial locations in images for all i. Then we transform the image patches to $\mathbf{Y}^r$ and $\mathbf{Y}^d$ via vectorization, where $\mathbf{Y}^r = [\mathbf{y}_1^r, \mathbf{y}_2^r \ldots \mathbf{y}_Z^r] \in \mathbb{R}^{B \times Z}$ and $\mathbf{Y}^d = [\mathbf{y}_1^d, \mathbf{y}_2^d \ldots \mathbf{y}_Z^d] \in \mathbb{R}^{B \times Z}$.

The next procedure contains two steps. First, the dictionary learning is conducted on the reference image by solving the following minimization problem:

$$\min_{\mathbf{D} \in \mathbb{R}^{B \times K}, \mathbf{C} \in \mathbb{R}^{K \times Z}} \|\mathbf{Y}^r - \mathbf{DC}\|_F^2 \tag{5}$$
$$s.t. \quad \|\mathbf{c}_i\|_0 \leq T_1, \ \|\mathbf{d}_j\|_2 = 1, \ \forall i, j,$$

where T_1 is a predefined sparsity. The norm constraint of atoms is to remove the scaling ambiguity of solutions. This is a classic sparse dictionary learning which is solved by K-SVD [2]. The learned dictionary is denoted by $\mathbf{D}^r$, and the obtained sparse coding matrix is denoted by $\mathbf{C}^r$. Next, we conduct dictionary learning using the following model:

$$\min_{\mathbf{D} \in \mathbb{R}^{B \times K}} \|\mathbf{Y}^d - \mathbf{DC}^r\|_F^2. \tag{6}$$

The learned dictionary is denoted by $\mathbf{D}^d$. In other words, we encourage the dictionary learned on the distorted image to generate the same sparse codes, which relates the information of $\mathbf{I}^r, \mathbf{I}^d$ and implicitly aligns the atoms of the two dictionaries $\mathbf{D}^r, \mathbf{D}^d$ for comparability. Note that there are no normalization constraints on the dictionary. The problem of (6) is a least-squares problem which is solved by the conjugate gradient algorithm.

We show some examples of the learned dictionaries on the reference image and the distorted image in Fig. 3. It can be seen that the learned atoms on the distorted images have the correspondence with those on the reference image. Furthermore, we can observe that the more severe the visual degradation is, the more change the learned atoms have. Such a trend demonstrates the effectiveness of the hybrid dictionary learning for encoding the image blur and compression. The following part is a simple explanation for the observation. Recall from Eq. (3) which demonstrates the mathematical relation between the dictionary atom of

(a) Reference **(b)** Blur-L1 **(c)** Blur-L3 **(d)** Jpeg-L2 **(e)** Jpeg-L4 **(f)** Jpeg2k-L1 **(g)** Jpeg2k-L4
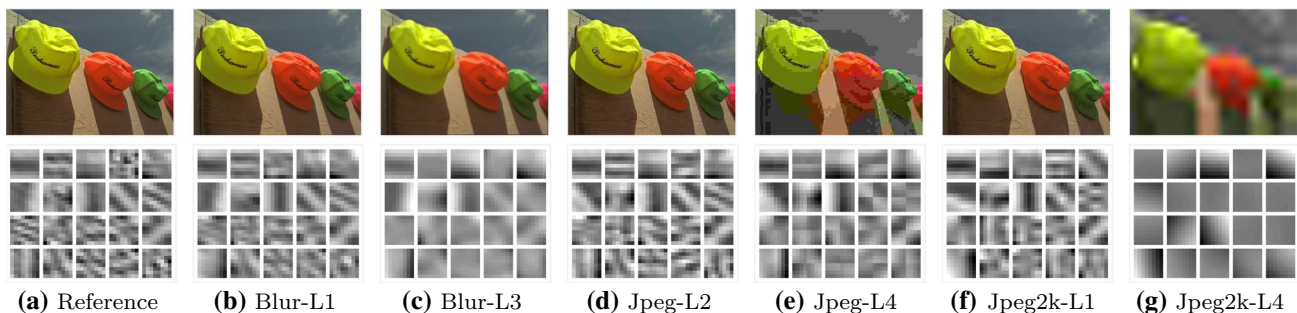
Fig. 3 Images with different degradations and the corresponding learned dictionaries by our approach. There are 4 degrees of distortion denoted by L1 to L4 (from minor to severe), which are marked in the image sub-titles

Fig. 4 Images with different blur degrees. The values indicate $\sum_i \|\mathbf{d}_i^{\mathrm{r}} - \mathbf{d}_i^{\mathrm{d}}\|_2$

the reference image and that of the distorted one: $\mathbf{d}_i^{\mathrm{d}} = \mathbf{a} * \mathbf{d}_i^{\mathrm{r}}$ where $\mathbf{a}$ is a blur kernel. Let $\mathcal{F}(\cdot)$ denote Fourier Transform. Then based on the Parseval's identity [13] and the convolution theorem [50], we have

$$\begin{aligned}
\|\mathbf{d}_i^{\mathrm{r}} - \mathbf{d}_i^{\mathrm{d}}\|_2 &= \|\mathbf{d}_i^{\mathrm{r}} - \mathbf{a} * \mathbf{d}_i^{\mathrm{r}}\|_2 = \|\mathcal{F}(\mathbf{d}_i^{\mathrm{r}}) - \mathcal{F}(\mathbf{a} * \mathbf{d}_i^{\mathrm{r}})\|_2 \\
&= \|\mathcal{F}(\mathbf{d}_i^{\mathrm{r}}) - \mathcal{F}(\mathbf{a}) \otimes \mathcal{F}(\mathbf{d}_i^{\mathrm{r}})\|_2 \\
&\geq \|\mathcal{F}(\mathbf{d}_i^{\mathrm{r}})\|_2 - \|\mathcal{F}(\mathbf{a}) \otimes \mathcal{F}(\mathbf{d}_i^{\mathrm{r}})\|_2 \\
&\geq \|\mathcal{F}(\mathbf{d}_i^{\mathrm{r}})\|_2 - \max(|\mathcal{F}(\mathbf{a})|)\|\mathcal{F}(\mathbf{d}_i^{\mathrm{r}})\|_2 \\
&= (1 - \max(|\mathcal{F}(\mathbf{a}))\|\mathcal{F}(\mathbf{d}_i^{\mathrm{r}})\|_2.
\end{aligned}$$

Note that a blur kernel can be modeled by a low-pass filter with ℓ_1-normalization for preserving image brightness. Then, it is easy to verify that $\max(|\mathcal{F}(\mathbf{a})|) \leq 1$. When the blur degree of $\mathbf{a}$ increases, $\mathcal{F}(\mathbf{a})$ will decay faster which implies the difference between $\mathbf{d}_i^{\mathrm{r}}$ and $\mathbf{d}_i^{\mathrm{d}}$ likely becomes larger as the blurriness becomes more severe. Figure 4 shows a demonstration on the relation between the blur degree and the difference of dictionary atoms. It is obvious that $\sum_i \|\mathbf{d}_i^{\mathrm{r}} - \mathbf{d}_i^{\mathrm{d}}\|_2$ increases as the image becomes more blurry.

3.2 Sparse coding module

Both dictionary and sparse coefficients can encode the degradation by blur and compression, as shown in (3) and (4). Thus, we also utilize the sparse coefficients for our IQA task, which is the main goal in the sparse coding module. Different from these recent studies [51, 52] in the field of sparse coding, we adopt a simple strategy in this module. Similar to that in the hybrid dictionary learning module, with the sampling and vectorization operations, we obtain the vectorized image patches $\mathbf{X}^{\mathrm{r}} = [\mathbf{x}_1^{\mathrm{r}}, \mathbf{x}_2^{\mathrm{r}} \ldots \mathbf{x}_{\bar{Z}}^{\mathrm{r}}] \in \mathbb{R}^{\bar{B} \times \bar{Z}}$ on the reference image $\mathbf{I}^{\mathrm{r}}$ and the vectorized image patches $\mathbf{X}^{\mathrm{d}} = [\mathbf{x}_1^{\mathrm{d}}, \mathbf{x}_2^{\mathrm{d}} \ldots \mathbf{x}_{\bar{Z}}^{\mathrm{d}}] \in \mathbb{R}^{\bar{B} \times \bar{Z}}$ on the distorted image $\mathbf{I}^{\mathrm{d}}$ with the image block size $\sqrt{\bar{B}} \times \sqrt{\bar{B}}$ and the sliding step of sampling $\bar{S}$. To capture both the low-frequency and high-frequency information, we use the Haar wavelet dictionary for sparse coding.

Let $\mathbf{W} = [\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{L-1}] \in \mathbb{R}^{\bar{B} \times L}$ denote a Haar wavelet dictionary with L atoms, where $\mathbf{w}_0$ denotes the low-frequency atom and $\mathbf{w}_l$ denotes the high-frequency

atoms for $l = 1, \ldots, L - 1$. We calculate the coefficient matrix $\mathbf{E}^{\mathrm{r}} \in \mathbb{R}^{L \times \bar{Z}}$ of $\mathbf{I}^{\mathrm{r}}$ via solving

$$\min_{\mathbf{E}} \|\mathbf{X}^{\mathrm{r}} - \mathbf{W}\mathbf{E}\|_{\mathrm{F}}^2 + \beta\|\mathbf{E}\|_0, \quad \text{s.t. } \beta > 0. \tag{7}$$

Since the wavelet dictionary is an orthogonal dictionary, the problem of (7) has the explicit solution given by

$$\mathbf{E}^{\mathrm{r}} = \mathcal{T}_\beta(\mathbf{W}^\top \mathbf{X}^{\mathrm{r}}), \tag{8}$$

where $\mathcal{T}_\beta(\cdot)$ keeps the values of elements whose absolute value is not smaller than $\sqrt{\beta}$ and sets the rest to 0. The sparse codes $\mathbf{E}^{\mathrm{d}} \in \mathbb{R}^{L \times \bar{Z}}$ of $\mathbf{I}^{\mathrm{d}}$ are calculated in the similar way, i.e., replacing $\mathbf{E}^{\mathrm{r}}$ in (7) with $\mathbf{E}^{\mathrm{d}}$, which is given by

$$\mathbf{E}^{\mathrm{d}} = \mathcal{T}_\beta(\mathbf{W}^\top \mathbf{X}^{\mathrm{d}}). \tag{9}$$

Let $\mathbf{E}^{\mathrm{r}} = [\tilde{\mathbf{e}}_0^{\mathrm{r}}, \tilde{\mathbf{e}}_1^{\mathrm{r}}, \ldots, \tilde{\mathbf{e}}_{L-1}^{\mathrm{r}}]^\top$, $\mathbf{E}^{\mathrm{d}} = [\tilde{\mathbf{e}}_0^{\mathrm{d}}, \tilde{\mathbf{e}}_1^{\mathrm{d}}, \ldots, \tilde{\mathbf{e}}_{L-1}^{\mathrm{d}}]^\top$, where $(\tilde{\mathbf{e}}_l^{\mathrm{r}})^\top, (\tilde{\mathbf{e}}_l^{\mathrm{d}})^\top$ are the lth row of $\mathbf{E}^{\mathrm{r}}, \mathbf{E}^{\mathrm{d}}$, respectively, which correspond to the lth atom $\mathbf{w}_l$, for $l = 0, \ldots, L - 1$. Note that $\tilde{\mathbf{e}}_0^{\mathrm{r}}, \tilde{\mathbf{e}}_0^{\mathrm{d}}$ correspond to $\mathbf{w}_0$ which is a low-frequency component, while $\tilde{\mathbf{e}}_l^{\mathrm{r}}, \tilde{\mathbf{e}}_l^{\mathrm{d}}, l = 1, \ldots, L - 1$ correspond to high-frequency components. We give different treatments to these two types of components. We use $\tilde{\mathbf{e}}_0^{\mathrm{r}}, \tilde{\mathbf{e}}_0^{\mathrm{d}}$ as the low-frequency features and calculate the high-frequency features $\mathbf{f}^{\mathrm{r}}, \mathbf{f}^{\mathrm{d}}$ as follows:

$$\mathbf{f}^{\mathrm{r}}(j) = \sum_{l=1}^{L-1} |\tilde{\mathbf{e}}_l^{\mathrm{r}}(j)|, \ \mathbf{f}^{\mathrm{d}}(j) = \sum_{l=1}^{L-1} |\tilde{\mathbf{e}}_l^{\mathrm{d}}(j)|, \ j = 1, \ldots, \bar{Z}.$$

3.3 Calculations of visual quality score

From the above modules, we generate the dictionaries $\mathbf{D}^{\mathrm{r}} = [\mathbf{d}_1^{\mathrm{r}}, \ldots, \mathbf{d}_K^{\mathrm{r}}], \mathbf{D}^{\mathrm{d}} = [\mathbf{d}_1^{\mathrm{d}}, \ldots, \mathbf{d}_K^{\mathrm{d}}]$ and sparse-coding-based features $\mathbf{e}_0^{\mathrm{r}}, \mathbf{e}_0^{\mathrm{d}}, \mathbf{f}^{\mathrm{r}}, \mathbf{f}^{\mathrm{d}}$ on the reference and distorted images. Then, using them as the image features, we apply different metrics to different features to generate the final score. Regarding the dictionaries, we define two metrics. The first one is defined by:

$$M_{\cos}(\mathbf{D}^{\mathrm{r}}, \mathbf{D}^{\mathrm{d}}) = \frac{1}{K} \sum_{i=1}^K \frac{|(\mathbf{d}_i^{\mathrm{r}})^\top (\mathbf{d}_i^{\mathrm{d}})|}{\|\mathbf{d}_i^{\mathrm{r}}\|_2 \|\mathbf{d}_i^{\mathrm{d}}\|_2}, \tag{10}$$

which measures cosine values of the angles between pairs of atoms. The second metric is defined by:

$$M_{\mathrm{das}}(\mathbf{D}^{\mathrm{r}}, \mathbf{D}^{\mathrm{d}}) = \frac{1}{K} \left(\sum_{i=1}^K \frac{\|\mathbf{d}_i^{\mathrm{d}}\|_2}{\|\mathbf{d}_i^{\mathrm{r}}\|_2} \right) = \frac{1}{K} \sum_{i=1}^K \|\mathbf{d}_i^{\mathrm{d}}\|_2, \tag{11}$$

which measures the energy changes on the dictionary atoms. The equality holds due to $\|\mathbf{d}_i^{\mathrm{r}}\|_2 = 1$ for all i.

Regarding the sparse codes, we separately treat the low-frequency features and the high-frequency ones, as they are quite different with each other in terms of magnitude and

distribution. The low-frequency feature is related to the luminance mean [8, 16]. We use the Pearson correlation coefficient to measure the statistical linear correlation between $\tilde{\mathbf{e}}_0^r$ and $\tilde{\mathbf{e}}_0^d$, which is defined by

$$M_{\mathrm{pcc}}(\tilde{\mathbf{e}}_0^r, \tilde{\mathbf{e}}_0^d) = \frac{1}{\bar{Z}} \frac{(\tilde{\mathbf{e}}_0^r - \mathcal{M}(\tilde{\mathbf{e}}_0^r))^\top (\tilde{\mathbf{e}}_0^d - \mathcal{M}(\tilde{\mathbf{e}}_0^d))}{\sigma(\tilde{\mathbf{e}}_0^r)\sigma(\tilde{\mathbf{e}}_0^d)}, \quad (12)$$

where $\mathcal{M}(\mathbf{x})$ outputs a vector obtained by repeating the mean value of $\mathbf{x} \in \mathbb{R}^N$ by N times, and $\sigma(\cdot)$ denotes the standard deviation. On the high-frequency features, they often correspond to the local image structures such as edges. Thus, we use the structural metric [8, 41] to measure the distortion:

$$M_{\mathrm{crs}}(\mathbf{f}^r, \mathbf{f}^d) = \frac{1}{\bar{Z}} \sum_{j=1}^{\bar{Z}} \frac{2\mathbf{f}^r(j)\mathbf{f}^d(j) + c}{(\mathbf{f}^r(j))^2 + (\mathbf{f}^d(j))^2 + c}, \quad (13)$$

where $c > 0$ is a stabilizer set to a small constant.

The final quality score of the distorted image is the combination of the four metrics as follows:

$$\begin{aligned} S_{\mathrm{JDL}}(\mathbf{I}^r, \mathbf{I}^d) = {} & \lambda_1 M_{\cos}(\mathbf{D}^r, \mathbf{D}^d) + \lambda_2 M_{\mathrm{das}}(\mathbf{D}^r, \mathbf{D}^d) \\ & + \lambda_3 M_{\mathrm{pcc}}(\tilde{\mathbf{e}}_0^r, \tilde{\mathbf{e}}_0^d) + \lambda_4 M_{\mathrm{crs}}(\mathbf{f}^r, \mathbf{f}^d) \end{aligned} \quad (14)$$

with constraints that $\lambda_j > 0, \sum \lambda_j = 1$ for $j = 1, \ldots, 4$.

To clarify the pipeline of the proposed method, we give a detailed algorithmic description in Algorithm 1.

Algorithm 1 The proposed method.

Input: The reference image $\mathbf{I}^r$; The distorted image $\mathbf{I}^d$;
Output: The quality score $S_{\mathrm{JDL}}(\mathbf{I}^r, \mathbf{I}^d)$ of $\mathbf{I}^d$;
1: Calculate the corresponding dictionaries $\mathbf{D}^r$ and $\mathbf{D}^d$ of $\mathbf{I}^r$ and $\mathbf{I}^d$ by solving (5) and (6) respectively.
2: Compute the corresponding coefficients $\mathbf{C}^r$ and $\mathbf{C}^d$ of $\mathbf{I}^r$ and $\mathbf{I}^d$ according to (8) and (9) individually;
3: Operate on $\mathbf{C}^r$ and $\mathbf{C}^d$ to get the low-frequency features $\tilde{\mathbf{e}}_0^r, \tilde{\mathbf{e}}_0^d$ and high-frequency features $\mathbf{f}^r, \mathbf{f}^d$;
4: Quantify the differences between dictionaries $\mathbf{D}^r$ and $\mathbf{D}^d$ based on (10) and (11) to get sub-scores $M_{\cos}(\mathbf{D}^r, \mathbf{D}^d)$ and $M_{\mathrm{das}}(\mathbf{D}^r, \mathbf{D}^d)$;
5: Quantify the differences between coefficients $\mathbf{C}^r$ and $\mathbf{C}^d$ based on (12) and (13) to get sub-scores $M_{\mathrm{pcc}}(\tilde{\mathbf{e}}_0^r, \tilde{\mathbf{e}}_0^d)$ and $M_{\mathrm{crs}}(\mathbf{f}^r, \mathbf{f}^d)$;
6: Combine four sub-scores obtained in the last two steps with individual weights $\lambda_1 \sim \lambda_4$, as shown in (14);
7: **return** S_{JDL};

---

4 Experiments

The proposed method is evaluated using the subsets of five public benchmark datasets including blurry images and compressed images, with five common criteria used to measure the performance.

4.1 Experimental settings and implementation details

Five datasets are used for the evaluation, including LIVE [35], CSIQ [15], TID2008 [23], TID2013 [24] and MICT

[49]. These datasets contain multiple distortion types, and we only select the images related to blur and compression for our test. The first four datasets contain Gaussian blurry images whose numbers are 145, 150, 100 and 125, respectively, and they also contain compressed images (JPEG and JPEG without error) whose numbers are 344, 300, 200 and 250, respectively. MICT contains 168 compressed images from JPEG and JPEG2000 compression. Then, there are totally 1782 images used for our evaluation.

Five criteria can be employed to measure the performance of the proposed method from different aspects, which are widely used in the evaluation of IQA methods, including:

- Spearman rank order correlation coefficient (SROCC),
- Kendall rank order correlation coefficient (KROCC),
- Pearson linear correlation coefficient (PLCC),
- Root mean square error (RMSE),
- Mean absolute error (MAE).

The criteria SROCC and KROCC are used to measure the prediction monotonicity (i.e., whether the increase or decrease in the objective prediction is associated with an increase or decrease in the subjective measurement), while the criteria PLCC, RMSE and MAE are used to evaluate the prediction accuracy (i.e., whether the objective prediction is consistent with the subjective measurement). An effective IQA method should yield high values of PLCC, SROCC and KROCC, while result in low RMSE and MAE. It is noted that due to the nonlinear relationship between objective ratings and subjective scores, all the criteria are calculated after mapping the objective score x to the subjective one by the logistic function [35]:

$$\begin{aligned} f(x) = {} & \omega_1 \left(\frac{1}{2} - \frac{1}{1 + exp(\omega_2(x - \omega_3))} \right) \\ & + \omega_4 x + \omega_5, \end{aligned} \quad (15)$$

where the parameters ω_i for $i = 1, 2, .., 5$ are all calculated by least-squares fitting tools such as MATLAB least-squares toolbox. Such a mapping scheme is widely used in existing IQA methods, which can well bridge the gap between the objective domain and subjective domain; see, e.g., [35, 48].

Implementation details of our experiments are presented as follows. In the module of hybrid dictionary learning, the size of image patches and the sliding step of sampling are set to 8×8 and 4, respectively (i.e., $S = 4, B = 64$). The input images are scaled to their half size for saving computational cost. The number of dictionary atoms K is set to 20. In the sparse coding module, we use a 4×4 wavelet Haar dictionary so that $L = 4, \bar{B} = 4$. The sampling step is set to 2, implying non-overlapping sampling. The parameter β is empirically set to 0.001. For blurriness artifacts,

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are set to be (0.03, 0.61, 0.30, 0.06) and for compression artifacts, the parameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are set to be (0.03, 0.06, 0.34, 0.57). These two sets of parameters are selected in a defined parameter grid which has four dimensions of the range from 0 to 1 with step 0.01 using the images with the corresponding distortion from the CSIQ dataset. For color images, the variation of color saturation should also be considered in our method for better perception, so we first convert them from the RGB color space to the YUV color space. Then we apply hybrid dictionary learning on the Y channel and conduct sparse coding on all there three channels.

4.2 Individual component contribution

To reveal the individual importance of the four components $M_{cos}(\mathbf{D}^r, \mathbf{D}^d)$, $M_{das}(\mathbf{D}^r, \mathbf{D}^d)$, $M_{pcc}(\tilde{\mathbf{e}}_0^r, \tilde{\mathbf{e}}_0^d)$ and $M_{crs}(\mathbf{f}^r, \mathbf{f}^d)$ of the proposed method, we evaluate their performances separately on blurry images and compressed images. The results are listed in Tables 1 and 2. From these two tables, we can see that the performance of the dictionaries' energy (i.e., M_{das}) and the high-frequency information (i.e., M_{crs}) are the two metrics most closed to the overall performance (i.e., All). According to the weighting of each component in the final prediction, M_{das} contributes the most in blurriness, while M_{crs} contributes the most in compression. M_{pcc} contributes significantly while M_{cos} contributes slightly in both distortions.

Table 1 Performance of components on blurry images

Database	Criteria	M_{cos}	M_{das}	M_{pcc}	M_{crs}	All
LIVE	PLCC	0.825	0.964	0.873	0.966	0.974
	SROCC	0.736	0.958	0.869	0.966	0.968
	KROCC	0.551	0.837	0.690	0.842	0.847
	RMSE	10.43	4.223	8.998	4.758	4.221
	MAE	8.668	3.304	7.356	3.649	3.286
CSIQ	PLCC	0.893	0.965	0.945	0.954	0.975
	SROCC	0.772	0.960	0.966	0.962	0.970
	KROCC	0.591	0.838	0.840	0.826	0.850
	RMSE	0.129	0.074	0.094	0.086	0.064
	MAE	0.103	0.059	0.078	0.061	0.049
TID2008	PLCC	0.788	0.929	0.905	0.935	0.940
	SROCC	0.733	0.941	0.912	0.938	0.951
	KROCC	0.522	0.790	0.722	0.773	0.801
	RMSE	0.723	0.414	0.500	0.442	0.400
	MAE	0.581	0.324	0.406	0.359	0.306
TID2013	PLCC	0.723	0.940	0.922	0.935	0.950
	SROCC	0.653	0.951	0.929	0.938	0.961
	KROCC	0.446	0.815	0.750	0.773	0.824
	RMSE	0.862	0.399	0.483	0.442	0.388
	MAE	0.705	0.330	0.394	0.359	0.319

Table 2 Performance of components on compressed images

Database	Criteria	M_{cos}	M_{das}	M_{pcc}	M_{crs}	All
LIVE	PLCC	0.563	0.924	0.935	0.955	0.968
	SROCC	0.498	0.949	0.946	0.952	0.969
	KROCC	0.341	0.795	0.794	0.827	0.839
	RMSE	23.97	11.12	10.26	7.661	7.263
	MAE	19.95	9.281	7.940	5.978	5.661
CSIQ	PLCC	0.707	0.964	0.967	0.958	0.972
	SROCC	0.625	0.945	0.964	0.940	0.954
	KROCC	0.447	0.789	0.827	0.775	0.810
	RMSE	0.220	0.083	0.079	0.090	0.074
	MAE	0.181	0.062	0.062	0.066	0.054
TID2008	PLCC	0.683	0.938	0.908	0.916	0.974
	SROCC	0.610	0.937	0.906	0.915	0.969
	KROCC	0.446	0.776	0.723	0.729	0.846
	RMSE	1.382	0.657	0.794	0.760	0.431
	MAE	1.101	0.498	0.630	0.594	0.327
TID2013	PLCC	0.666	0.933	0.915	0.930	0.975
	SROCC	0.564	0.922	0.900	0.914	0.970
	KROCC	0.405	0.751	0.717	0.727	0.849
	RMSE	1.223	0.592	0.661	0.605	0.418
	MAE	1.004	0.441	0.506	0.472	0.318

4.3 Performance comparison

The performance of the proposed method is evaluated on CSIQ, LIVE, TID2008 and TID2013 datasets. For comparison, five FR-IQA approaches are chosen, including the famous methods PSNR, SSIM [46], the recent novel method SPSIM [38], as well as the powerful methods MS-SSIM [45] and FSIM [53] which are regarded as the top-performing ones among the sixteen FR-IQA methods compared in [20]. All above metrics are applicable for general distortion types and most approaches above are based on capturing structural changes of images as we did. Moreover, the proposed method is compared with several metrics designed specifically for sharpness measurement, including state-of-the-art JNBM [9], CPBD [21], MLV [6] and a sparse-representation-based method SRM [17]. All the compared methods have available results for comparison or published source codes which can produce the results.

The experimental results with three criteria are summarized in Table 3. The top three good results are marked in boldface, and the best result among all the compared methods is underlined. To evaluate the overall performance over the four datasets, for each method, we also calculate the average performance weighted by the number of distorted images used in each dataset. As shown in Table 3, among all the compared distortion-specific methods (i.e., JNBM, CPBD, MLV, SRM), the proposed one

Table 3 Performance comparison on benchmark datasets

Dataset	Criterion	OURS	PSNR	SSIM	MS-SSIM	FSIM	SPSIM	JNBM	CPBD	MLV	SRM
(a) Blurry images											
CSIQ	PLCC	**0.9746**	0.9252	0.9496	**0.9660**	**0.9645**	0.9559	0.8711	0.9155	0.9489	0.9391
	SROCC	**0.9700**	0.9291	0.9608	**0.9711**	0.9713	0.9654	0.8338	0.8845	0.9245	0.9139
	RMSE	**0.0642**	0.1087	0.0898	**0.0741**	0.0757	0.0842	0.1407	0.1153	0.0904	0.0984
LIVE	PLCC	**0.9735**	0.7841	0.9483	0.9565	0.9737	0.9602	0.7744	0.8638	0.9057	0.9596
	SROCC	**0.9679**	0.7823	0.9517	0.9542	0.9706	0.9607	0.7872	0.9182	0.9312	0.9593
	RMSE	**4.2208**	11.464	5.8612	5.3889	4.2107	5.1613	11.686	9.3057	7.8283	5.1958
TID2008	PLCC	0.9400	0.8736	0.9508	0.9503	0.9454	0.9395	0.6932	0.8231	0.8593	0.8896
	SROCC	**0.9513**	0.8684	0.9544	0.9563	0.9472	0.9450	0.6667	0.8412	0.8548	0.8869
	RMSE	0.4004	0.5712	0.3634	0.3655	0.3823	0.4020	0.8458	0.6663	0.6001	0.5361
TID2013	PLCC	0.9504	0.9148	0.9648	0.9634	0.9549	0.9488	0.7115	0.8555	0.8830	0.9014
	SROCC	**0.9607**	0.9147	0.9668	0.9673	0.9550	0.9531	0.6902	0.8531	0.8785	0.8925
	RMSE	0.3881	0.5041	0.3283	0.3347	0.3704	0.3941	0.8769	0.6461	0.5858	0.5404
Weighted average	PLCC	**0.9640**	0.8734	0.9531	**0.9597**	0.9611	0.9533	0.7716	0.8689	0.9038	0.9262
	SROCC	0.9347	0.8730	0.9585	**0.9626**	0.9628	0.9585	0.7542	0.8780	0.9019	0.9162
	RMSE	**1.3476**	3.4591	1.8091	**1.6748**	1.3585	2.0284	3.6726	2.9116	2.4652	1.7102
(b) Compressed images											
CSIQ	PLCC	0.9715	0.9182	0.9711	**0.9778**	0.9794	0.9755	0.5468	0.5651	0.5469	0.8608
	SROCC	0.9539	0.9159	0.9541	**0.9639**	0.9640	0.9610	0.5286	0.4421	0.3837	0.8086
	RMSE	**0.0739**	0.1234	0.0743	**0.0653**	0.0629	0.0685	0.2608	0.257	0.2608	0.1586
LIVE	PLCC	0.9681	0.8878	0.9645	**0.9713**	0.9778	0.9779	0.3501	0.4766	0.2939	0.8869
	SROCC	0.9689	0.8911	0.9641	**0.9706**	0.9758	0.9759	0.3384	0.3756	0.2764	0.8779
	RMSE	7.2632	13.3452	7.6572	**6.9033**	6.0758	6.0638	27.164	25.494	27.719	13.397
TID2008	PLCC	**0.9748**	0.8699	0.9615	0.9694	**0.9746**	0.9734	0.6753	0.6937	0.6765	0.9076
	SROCC	**0.9696**	0.8664	0.9584	**0.9645**	0.9694	0.9694	0.6467	0.5701	0.5843	0.8842
	RMSE	**0.4305**	0.9329	0.5196	0.4644	0.4240	0.4336	1.3951	1.3625	1.3931	0.794
TID2013	PLCC	**0.9753**	0.9167	0.9604	0.9656	**0.9709**	0.9754	0.6636	0.6891	0.6504	0.9073
	SROCC	**0.9705**	0.9170	0.9453	0.9495	0.9581	0.9613	0.6383	0.6004	0.5324	0.8522
	RMSE	**0.4180**	0.6554	0.4569	0.4264	0.3927	0.3615	1.2272	1.1888	1.2461	0.6899
Weighted average	PLCC	**0.9721**	0.9001	0.9648	0.9714	0.9760	0.9757	0.5447	0.5949	0.5262	0.8883
	SROCC	**0.9651**	0.8997	0.9556	0.9624	0.9669	0.9668	0.5246	0.4862	0.4281	0.8529
	RMSE	2.2299	4.0938	2.3664	**2.1356**	1.8883	1.8809	8.2133	7.7309	8.3721	4.1001

consistently performs the best across all the datasets and all criteria. Moreover, for blurry images (Table 3 (a)), the proposed method performs comparably or even better than the other top performed metrics in CSIQ and LIVE datasets. Overall speaking, in terms of PLCC and RMSE (the weighted average value across all datasets), our method is the best. For compressed images (Table 3 (b)), the proposed method performs the best in TID2008 and TID2013 datasets, and the overall performance across all datasets is comparable with the other top metrics such as FSIM and SPSIM. To further demonstrate the advantage of our method, we evaluate the performance on the summations of final scores produced by ours and other methods directly. The results are listed in Table 4. It implies that the

proposed method utilizing dictionary learning and sparse coding may extract complementary features with other competitive methods, which means that our method can be utilized to ensemble with some existing methods.

To draw a statistical conclusion for the performance, we conduct a two-tailed F-test to the prediction residuals after nonlinear regression (i.e., RMSE) at the 5% significance level [48] for every pair of the compared methods on the dataset CSIQ and the dataset LIVE with specific types, where the model residuals are assumed coming from the normal distributions. The results of statistical significance tests are displayed in Fig. 5. It is observed that all compared IQA methods are not statistically better than the proposed approach marked as 'OURS' in Fig. 5 on the

Table 4 Performance of combined features on blurry images in the TID2013 dataset

Database	Criteria	MS-SSIM	OURS+MS-SSIM	SPSIM	OURS+SPSIM	FSIM	OURS+FSIM
TID2013	PLCC	0.9634	0.9657	0.9488	0.9657	0.9549	0.9649
	SROCC	0.9673	0.9708	0.9531	0.9708	0.9550	0.9697
	KROCC	0.8429	0.8556	0.8132	0.8556	0.8158	0.8509
	RMSE	0.3347	0.3241	0.3941	0.3241	0.3704	0.3277
	MAE	0.2557	0.2487	0.2998	0.2487	0.2912	0.2519

(a) Blurry images   **(b)** Compressed images   **(c)** Blurry images   **(d)** Compressed images
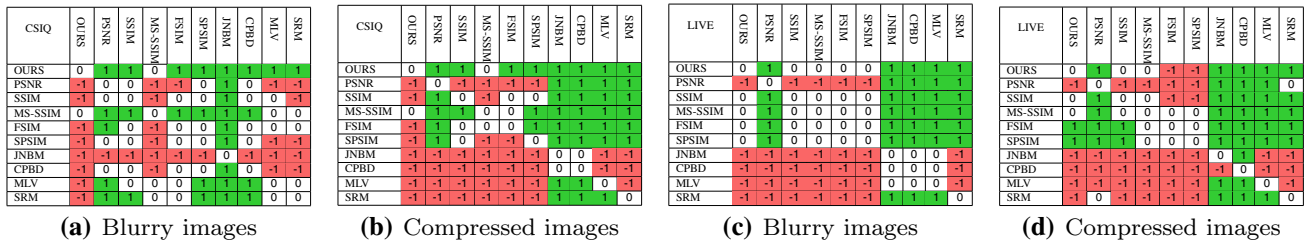
Fig. 5 The results of statistical significance tests of several IQA metrics on the datasets CSIQ and LIVE of specific types. The value of '1' (highlighted in green) implies that the metric in the row is significantly better than the metric in the column, while the value of '−1' (highlighted in red) indicates the inverse condition. '0' implies there is no significant difference
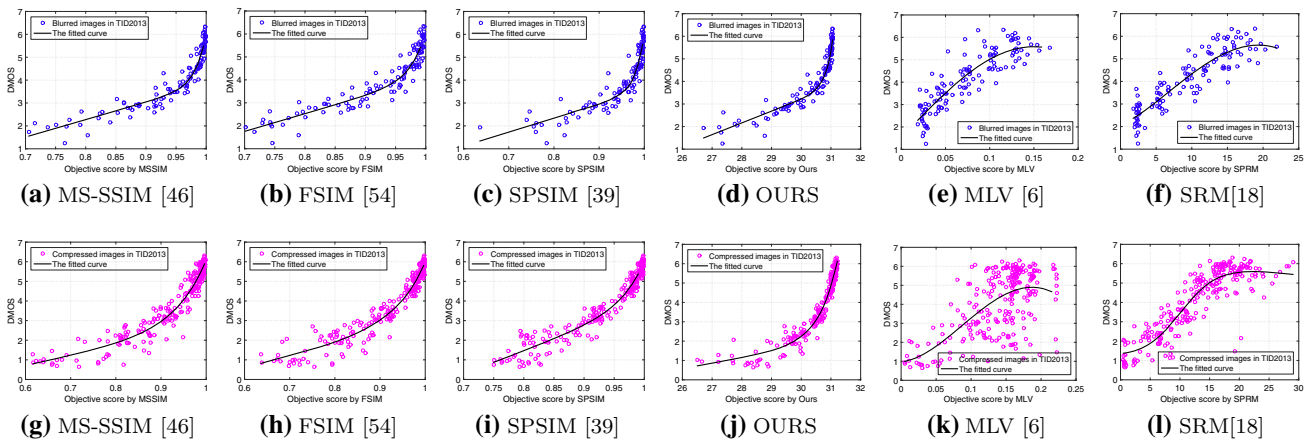
CSIQ sub-dataset. For LIVE sub-dataset, our method is superior than or equivalent to almost all of existing new and classical IQA metrics tested in this paper, expect for the FR methods FSIM and SPSIM for compressed images.

For visualizing the performance of compared metrics, we show the scatter plots of subjective scores against the predicted objective scores for four distortion-specific metrics (MLV, SRM) and three distortion-general metrics (MS-SSIM, FSIM, SPSIM) with better performance in Fig. 6 on the TID2013 sub-datasets. The black curve is obtained by fitting the data points with the logistic function in (15). From the visual results, it can be seen that the data points of our method are more tight to the fitting curve than

others, implying that the proposed method is highly consistent with human visual systems.

4.4 Hybrid dictionary versus single dictionary

To further explore the effectiveness of using hybrid dictionary, we compare it with a sparse-coding-based approach using traditional single dictionary learning on two commonly compression types, e.g., JPEG and JPEG2K. MICT and TID2008 datasets are used in this experiment. The single-dictionary-based method used for comparison is SPARQ [12] which is a classic sparse-representation-based FR-IQA method. The results are listed in

(a) MS-SSIM [46]   **(b)** FSIM [54]   **(c)** SPSIM [39]   **(d)** OURS   **(e)** MLV [6]   **(f)** SRM[18]

(g) MS-SSIM [46]   **(h)** FSIM [54]   **(i)** SPSIM [39]   **(j)** OURS   **(k)** MLV [6]   **(l)** SRM[18]

Fig. 6 Scatter plots of metric values against subjective scores by several IQA metrics. The top row and the bottom row are for blurry images and compressed images on the TID2013 database, respectively

**Table 5** The performance on two datasets for different compression types

Dataset	Type	Metric	PLCC	SROCC	RMSE
MICT	JPG2K	SPARQ	0.927	0.928	**0.462**
		OURS	**0.932**	**0.935**	0.487
	JPEG	SPARQ	0.864	0.859	0.622
		OURS	**0.895**	**0.877**	**0.541**
TID2008	JPG2K	SPARQ	**0.973**	**0.966**	**0.447**
		OURS	0.969	0.963	0.485
	JPEG	SPARQ	0.943	0.919	0.565
		OURS	**0.965**	**0.923**	**0.445**

Table 5 where the better one between the two approaches is marked in boldface. It is observed that the proposed method outperforms the SPARQ method in most cases, which demonstrates that the learned dictionaries and features obtained by hybrid dictionary learning are effective in capturing the intrinsic structure changes for JPEG and JPG2K compression types.

4.5 Impact of parameters

The number of dictionary atoms K is an important parameter in hybrid dictionary learning. In general, the dictionary with more multifarious atoms has stronger representation ability when the dimension of data is fixed. It may span a larger subspace that the training image patches lie on. To see the influence of this parameter, we vary it from 10 to 60 with step 10 and then test the performance on blurry images and compressed images in CSIQ and TID2008 datasets. The results are drawn in Fig. 7, and the patch size is set to 8. According to the results, the K is set to 20 in our experiments since it gets the best performance.

The patch size $\sqrt{B}$ is another important parameter. Its square, i.e., B, is the dimension of training data which also determines the dimension of the learned subspace. The nonzero mean image patches with larger scale possess more structural information within a reasonable range. We vary $\sqrt{B}$ from 4 to 10, respectively, and then evaluate the performance to see the influence when K is fixed to 20. The

(a) Blur   **(b)** Compression
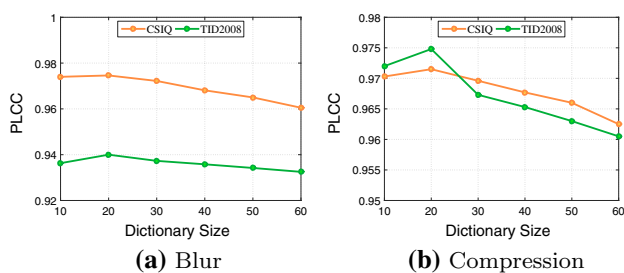
Fig. 7 Impact of dictionary size on performance

(a) Blur   **(b)** Compression
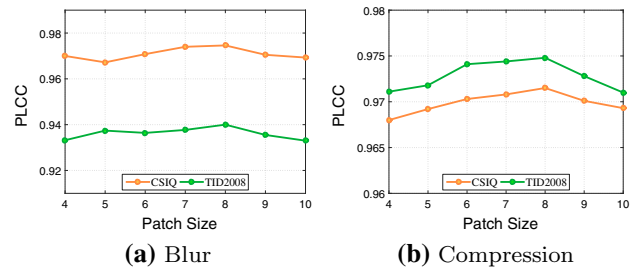
Fig. 8 Impact of image patch size on performance

results are drawn in Fig. 8. According to the results, the $\sqrt{B}$ is set to 8 in our experiments as it shows the best performance in both distortions.

4.6 Computational cost

We compare the average running time of the proposed method with other methods for their complete evaluation processes. The experiment is conducted on a Windows-10 PC with Intel Core-I7 7700K CPU (4.20 GHz) and 32 GB RAM. The proposed method is implemented using MATLAB R2018b. We chose 10 test images randomly from the CSIQ dataset, whose size is 512×512. The settings are same for all compared methods including 10 testing images. All the Matlab source codes of the compared methods were obtained from the original authors. The results are displayed in Table 6. It is observed that compared with another sparse-representation-based method SRM, ours is much faster. In comparison with other compared methods, the proposed method is slower. It should be noted that the time cost of our proposed method is still in an applicable range.

To further explore the computational cost of different modules of our method, we list the time cost of them for ten test images in Table 7. From the table, it can be seen that the most of running time is spent by the procedure of hybrid dictionary learning, which may be accelerated by other fast-run dictionary learning methods (e.g., [26]) or more powerful computational resources.

4.7 Limitation and failure case

In the experiments, we also evaluate the performance for different reference images and their corresponding degraded versions with blur of different levels in the CSIQ and LIVE databases. In detail, for a reference image and its blurry versions, we calculate their quality scores and compare the generated scores with subjective scores (SS) such as DMOS or MOS to obtain the performance measurements such as PLCC. From these measurements, we can observe that the performance varies for different image contents. An example is shown in Fig. 9, where the

Table 6 Average time cost of the compared methods

Model Type	OURS FR	PSNR FR	SSIM FR	MS-SSIM FR	FSIM FR	SPSIM FR	JNBM NR	CPBD NR	MLV NR	SRM NR
Time (seconds)	0.192	0.001	0.001	0.032	0.117	0.098	0.140	0.074	0.039	1.421

Table 7 Time cost (seconds) of two modules of our method

Image Index	1	2	3	4	5	6	7	8	9	10	Mean
Dictionary learning	0.141	0.119	0.149	0.222	0.072	0.137	0.117	0.215	0.071	0.049	0.129
Sparse coding	0.073	0.074	0.070	0.068	0.068	0.069	0.070	0.073	0.078	0.072	0.071

	Reference	Blur-L2	Blur-L3	Blur-L4
SS	1.000	0.897	0.731	0.496
Ours	1.000	0.776	0.242	0.007
MS-SSIM	1.000	0.947	0.739	0.424
SPSIM	1.000	0.824	0.447	0.247
FSIM	1.000	0.863	0.497	0.202

Fig. 9 Images with dense and weak details and corresponding quality scores evaluated by four methods with normalization. SS indicates the normalized subjective score

proposed method gets the worst PLCC compared to three FR-IQA metrics. The images in the CSIQ database shown (from left to right) in the figure are ordered in the degradation level from light to severe (L2 to L4) monotonically. Additionally, the listed scores are normalized into the same scale for comparison. From Fig. 9, it is observed that the score generated by ours decreases suddenly with increasing distortion level (e.g., L2(0.776) to L3(0.242)) and the scores are not compatible with subjective scores (e.g., L2(0.897) to L3(0.496)). A probable reason is that the reference image shown in Fig. 9 contains dense and weak details, which may be represented with small sparse codes in K-SVD. Recall the procedure of learning the degraded dictionary we described in Sect. 3. As the sparse codes are fixed in the procedure, the atoms related to codes with small values are sensitive to the changes of the corresponding patches. As a result, the proposed method is sensitive to images with dense and weak details and may produce incompatible results in such cases.

5 Conclusions

In this paper, an effective metric is proposed for the full-reference image quality assessment on blurry images and compressed images. The proposed metric is built upon a hybrid dictionary learning and sparse coding scheme. The experimental results show that the proposed metric performs competitively with the existing top ones. In the future, we would like to extend our method to handle more image degradation types.

Compliance with ethical standards

Conflict of interest We declare that no conflict of interest exists.

References

1. A Olshausen B, Field D (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607–9
2. Aharon M, Elad M, Bruckstein A (2006) K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 54(11):4311–4322
3. Akbarizadeh G (2012) A new recognition approach based on genetic algorithm for classifying textures in satellite sar images. Int J Remote Sens Appl 2(4):7–19
4. Alasvand Z, Naderan M, Akbarizadeh G (2017) Superpixel-based feature learning for joint sparse representation of hyperspectral images. In: Proceedings of international conference on pattern recognition and image Analysis, IEEE, pp 156–159
5. Andekah ZA, Naderan M, Akbarizadeh G (2017) Semi-supervised hyperspectral image classification using spatial-spectral features and superpixel-based sparse codes. In: Iranian Conference on Electrical Engineering, IEEE, pp 2229–2234
6. Bahrami K, Kot AC (2014) A fast approach for no-reference image sharpness assessment based on maximum local variation. IEEE Signal Process Lett 21(6):751–755
7. Barlow H (2001) Redundancy reduction revisited. Network (Bristol, England) 12:241–53

8. Chang HW, Yang H, Gan Y, Wang MH (2013) Sparse feature fidelity for perceptual image quality assessment. IEEE Trans Image Process 22(10):4007–18

9. Ferzli R, Karam LJ (2009) A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). IEEE Trans Image Process 18(4):717–728

10. Gore A, Gupta S (2015) Full reference image quality metrics for jpeg compressed images. AEU-Int J Electron Commun 69(2):604–608

11. Gu S, Zuo W, Xie Q, Meng D, Feng X, Zhang L (2015) Convolutional sparse coding for image super-resolution. In: Proceedings of IEEE conference on computer vision, pp 1823–1831

12. Guha T, Nezhadarya E, Ward RK (2014) Sparse representation-based image quality assessment. Signal Process Image Commun 29(10):1138–1148

13. Gunawan H (2002) A generalization of bessel" s inequality and parseval" s identity. Periodica Math Hungarica 44(2):177–181

14. He L, Tao D, Li X, Gao X (2012) Sparse representation for blind image quality assessment. In: Proceedings of IEEE conference on computer vision and pattern recognition, IEEE, pp 1146–1153

15. Larson EC, Chandler DM (2010) Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electron Imaging 19(1):011006

16. Li L, Cai H, Zhang Y, Lin W, Kot AC, Sun X (2016) Sparse representation-based image quality index with adaptive sub-dictionaries. IEEE Trans Image Process 25(8):3775–3786

17. Li L, Wu D, Wu J, Li H, Lin W, Kot AC (2016) Image sharpness assessment by sparse representation. IEEE Trans Multimedia 18(6):1085–1097

18. Li L, Xia W, Fang Y, Gu K, Wu J, Lin W, Qian J (2016) Color image quality assessment based on sparse representation and reconstruction residual. J Visual Commun Image Represent 38:550–560

19. Liu Y, Zhai G, Gu K, Liu X, Zhao D, Gao W (2018) Reduced-reference image quality assessment in free-energy principle and sparse representation. IEEE Trans Multimedia 20(2):379–391

20. Ma K, Wu Q, Wang Z, Duanmu Z, Yong H, Li H, Zhang L (2016) Group mad competition-a new methodology to compare objective image quality models. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1664–1673

21. Narvekar ND, Karam LJ (2011) A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). IEEE Trans Image Process 20(9):2678–2683

22. Olshausen BA, Field DJ (1997) Sparse coding with an over-complete basis set: a strategy employed by v1? Vision Res 37(23):3311–3325

23. Ponomarenko N, Lukin V, Zelensky A, Egiazarian K, Carli M, Battisti F (2004) Tid 2008—a database for evaluation of full-reference visual quality assessment metrics. Adv Modern Radioelectron 10:30–45

24. Ponomarenko N, Ieremeiev O, Lukin V, Egiazarian K, Jin L, Astola J, Vozel B, Chehdi K, Carli M, Battisti F (2013) Color image database tid2013: peculiarities and preliminary results. In: Proceedings of European workshop on visual information processing, pp 106–111

25. Quan Y, Xu Y, Sun Y, Luo Y (2014) Lacunarity analysis on image patterns for texture classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 160–167

26. Quan Y, Huang Y, Ji H (2015) Dynamic texture recognition via orthogonal tensor dictionary learning. In: Proceedings of IEEE conference on computer vision

27. Quan Y, Bao C, Ji H (2016) Equiangular kernel dictionary learning with applications to dynamic texture analysis. In: Proceedings of IEEE conference on computer vision pattern recognition

28. Quan Y, Xu Y, Sun Y, Huang Y (2016) Supervised dictionary learning with multiple classifier integration. Pattern Recognit 55:247–260

29. Saha A, Wu QJ (2013) Perceptual image quality assessment using phase deviation sensitive energy features. Signal Process 93(11):3182–3191

30. Shan Q, Jia J, Agarwala A (2008) High-quality motion deblurring from a single image. ACM Trans Graph 27:73

31. Sharifzadeh F, Akbarizadeh G, Kavian YS (2019) Ship classification in sar images using a new hybrid CNN-MLP classifier. J Indian Soc Remote Sens 47(4):551–562

32. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–44

33. Sheikh HR, Wang Z, Cormack L, Bovik AC (2002) Blind quality assessment for jpeg2000 compressed images. In: Proceedings of conference on signals, systems and computers, IEEE, vol 2, pp 1735–1739

34. Sheikh HR, Bovik AC, Cormack L (2005) No-reference quality assessment using natural scene statistics: Jpeg 2000. IEEE Trans Image Process 14(11):1918–1927

35. Sheikh HR, Sabir MF, Bovik AC (2006) A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process 15(11):3440–51

36. Shi J, Xu L, Jia J (2015) Just noticeable defocus blur detection and estimation. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 657–665

37. Shi Z, Zhang J, Cao Q, Pang K, Luo T (2018) Full-reference image quality assessment based on image segmentation with edge feature. Signal Process 145:99–105

38. Sun W, Liao Q, Xue JH, Zhou F (2018) SPSIM: a superpixel-based similarity index for full-reference image quality assessment. IEEE Trans Image Process 27(9):4232–4244

39. Taibi F, Akbarizadeh G, Farshidi E (2019) Robust reservoir rock fracture recognition based on a new sparse feature learning and data training method. Multidimens Syst Signal Process 30:1–34

40. Tirandaz Z, Akbarizadeh G (2015) A two-phase algorithm based on kurtosis curvelet energy and unsupervised spectral regression for segmentation of sar images. IEEE J Sel Topics Appl Earth Observ Remote Sens 9(3):1244–1264

41. Wang S, Rehman A, Wang Z, Ma S, Gao W (2012) SSIM-motivated rate-distortion optimization for video coding. IEEE Trans Circuits Syst Video Technol 22(4):516–529

42. Wang S, Deng C, Zhao B, Huang GB, Wang B (2016) Gradient-based no-reference image blur assessment using extreme learning machine. Neurocomputing 174:310–321

43. Wang Z, Li Q (2010) Information content weighting for perceptual image quality assessment. IEEE Trans Image Process 20(5):1185–1198

44. Wang Z, Bovik AC, Evan BL (2000) Blind measurement of blocking artifacts in images. In: Proceedings of international conference on image processing, vol 3, pp 981–984

45. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: Proceedings of conference signals, systems and computers, IEEE, vol 2, pp 1398–1402

46. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–12

47. Xu Y, Quan Y, Zhang Z, Ji H, Fermüller C, Nishigaki M, Dementhon D (2012) Contour-based recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition, IEEE, pp 3402–3409

48. Xu Y, Liu D, Quan Y, Le Callet P (2015) Fractal analysis for reduced reference image quality assessment. IEEE Trans Image Process 24(7):2098–2109

49. Parvez Sazzad ZM, Kawayoke Y, Horita Y (2008) MICT image quality evaluation database. http://mict.eng.u-toyama.ac.jp/mict/index2.html

50. Zayed AI (1998) A convolution and product theorem for the fractional fourier transform. IEEE Signal Process Lett 5(4):101–103

51. Zhang H, Wang S, Xu X, Chow TW, Wu QJ (2018) Tree2vector: learning a vectorial representation for tree-structured data. IEEE Trans Neural Netw Learn Syst 99:1–15

52. Zhang H, Wang S, Zhao M, Xu X, Ye Y (2018) Locality reconstruction models for book representation. IEEE Trans Knowl Data Eng 30(10):1873–1886

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