Signal Processing: Image Communication 85 (2020) 115839

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Subjective and objective quality assessment for image restoration: A critical survey

Bo Hu [a], Leida Li [b,*], Jinjian Wu [b], Jiansheng Qian [a,*]

[a] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China
[b] School of Artificial Intelligence, Xidian University, Xi'an 710071, China

ARTICLE INFO

Keywords:
Image restoration
Subjective quality databases
Objective quality metrics
Parameter selection
Benchmarking image restoration algorithms

ABSTRACT

Image restoration is the process of restoring latent clear images from degraded images, and it has received substantial attention in various restoration scenarios recently. Contrary to the considerable progress of image restoration algorithms, the quality evaluation for image restoration falls behind, which may hinder the further development of advanced image restoration techniques. So far, only several objective quality metrics have been proposed to evaluate the quality of restored images for specific restoration scenarios, and little work has been dedicated to the applications of image restoration quality metrics. Besides, the performance of these metrics remains an open problem, and an obvious disadvantage of these metrics is that their robustness is quite weak. To bridge this gap, we present this survey paper. First, the difference between the traditional image quality assessment and quality assessment for image restoration was analyzed profoundly. Then, for different image restoration scenarios, this paper provides (i) a comprehensive description of the existing subjective quality databases, (ii) a thorough insight into existing objective quality metrics and (iii) two applications of image restoration quality metrics, namely, parameter selection and benchmarking image restoration algorithms. After that, the experimental results and a detailed performance analysis on (i) quality assessment of restored images, (ii) benchmarking image restoration algorithms and (iii) time complexity are given based on public restored image databases. Finally, the paper outlines the challenges and future trends of subjective and objective quality assessment for image restoration in various aspects.

1. Introduction

Image restoration is the process of restoring a high quality image from one or several degraded images. Recently, research on image restoration has received substantial attention in various fields, such as, image denoising (IDN) [1–3], image deblurring [4–7], super-resolution image reconstruction (SRR) [8–10], image dehazing (IDH) [11–13], compressive sensing image recovery (CSR) [14–16], etc. For an observed image $\mathbf{I}_o$, the image restoration can be generally formulated by [17]:

$$\mathbf{I}_o = \mathbf{H}\mathbf{I}_c + n, \tag{1}$$

where $\mathbf{H}$ denotes a degradation matrix, $\mathbf{I}_c$ is the latent clear image and n is the noise. With different settings of matrix $\mathbf{H}$, this can represent specific image restoration scenario. For example, when $\mathbf{H}$ becomes an identity operator, it denotes IDN; when $\mathbf{H}$ is a blurring operator, it denotes image deblurring; when $\mathbf{H}$ is a composite operator of blurring and down-sampling, it denotes SRR; when $\mathbf{H}$ is a random projection matrix, it denotes CSR. For IDH, the model can be represented as [12]:

$$\mathbf{I}_o(x) = \mathbf{I}_c(x)t(x) + \mathbf{A}(1 - t(x)), \tag{2}$$

where x is the location of each pixel within the image, $\mathbf{A}$ is the global atmospheric light, and t is the medium transmission. Obviously, image restoration process is ill-posed as many different pairs of $\mathbf{I}_c$ and $\mathbf{H}$ render the same $\mathbf{I}_o$. To make the problem well-posed, various prior models of degradation matrix and natural images have been developed. Although a significant amount of research has been made toward image restoration algorithms in various restoration scenarios, the quality evaluation for image restoration falls behind. Naturally, perceptual visual quality of image restoration can be evaluated by either subjective quality method with appropriate standard procedures or objective quality method under refined design. Since human beings are the ultimate consumers in most restoration scenarios and thus subjective assessment is a direct way. However, subjective assessment is time consuming and not efficient, and thereby cannot be embedded into practical applications. As an alternative way, objective quality assessment is therefore highly desired.

* Corresponding authors.
E-mail addresses: ldli@xidian.edu.cn (L. Li), qianzhangiqa@163.com (J. Qian).

https://doi.org/10.1016/j.image.2020.115839
Received 6 May 2019; Received in revised form 27 September 2019; Accepted 18 March 2020
Available online 1 April 2020

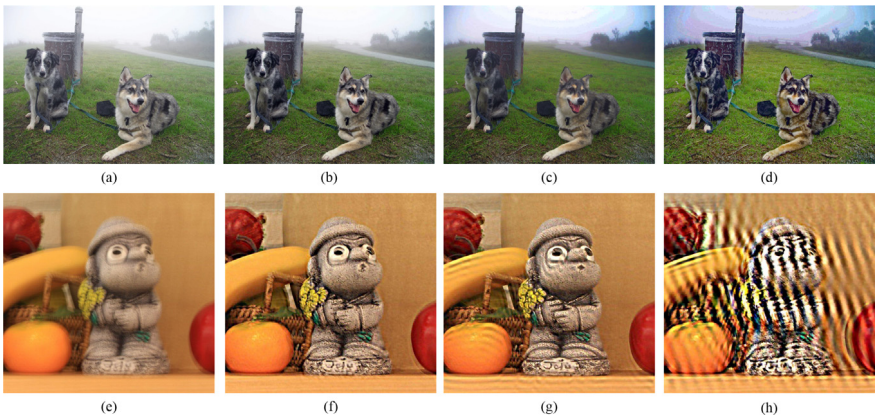

Fig. 1. Sample restored images of two different restoration scenarios. (a) is a hazy image; (b)–(d) are the dehazed images of image (a). (e) is a motion blurred image; (f)–(h) are the motion deblurred images of image (e).

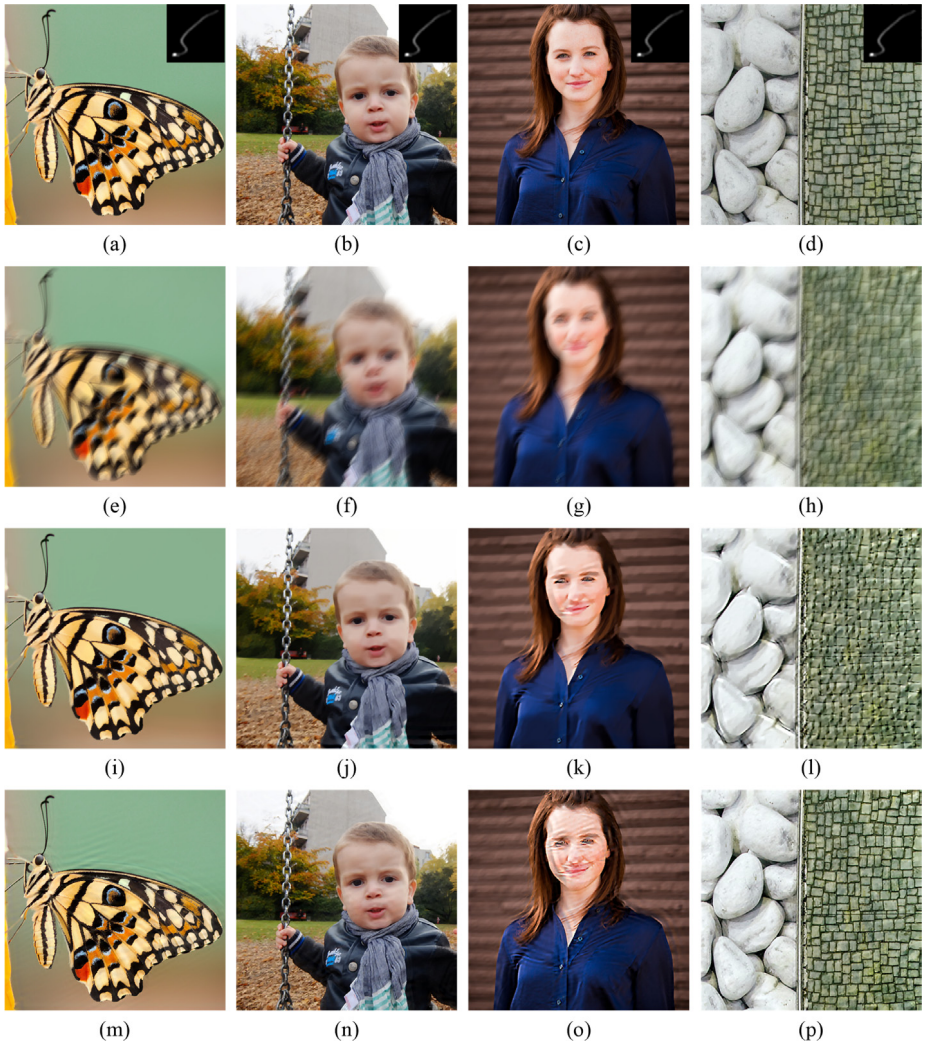

Fig. 2. Examples of motion deblurring on different image contents. (a)–(d) are original high-quality images with size 512 × 512. Blur kernel with size 25 × 25 is shown upper right corner of images (a)–(d). (e)–(h) are blurred images of (a)–(d) using the blur kernel. (i)–(l) are deblurred images of (a)–(d) using algorithm [7]. (m)–(p) are deblurred images of (a)–(d) using algorithm [5].

Objective quality evaluation for image restoration is a complex and challenging task due to the following problems. First, it is impossible to entirely remove the original distortion during the image restoration process, and thus there is residual original distortion in restored images. For intuitive observation, Fig. 1 shows an example of restored images in two specific restoration scenarios, including IDH and motion deblurring (MDB). In the figure, (a) is a hazy image; (b)–(d) are the dehazed images of image (a). (e) is a motion blurred image; (f)–(h) are the motion deblurred images of image (e). It is observed from Fig. 1 that residual haze obviously appears in dehazed images (b) and (c) and

residual motion blur still appears in motion deblurred images (f) and (g). Second, in addition to the residual original distortions, the restoration processes themselves are very likely to introduce side effects, which can be seen in Fig. 1(d) and (h). Halo artifacts and ringing effects are introduced in images (d) and (h), respectively. These artifacts do not present in degraded images (Fig. 1(a) and (e)), and they are side effects. In fact, the residual distortions and side effects usually co-exist in restored images. In other words, the restored images are usually characterized by multiple distortions. Third, although the aim of image restoration is to improve the quality of degraded images, it is possible to degrade the quality during the image restoration process. Such an example can be seen from Fig. 1(d) and (h). The quality of images (d) and (h) are obviously poor than that of images (a) and (e). Fourth, the performance of image restoration algorithm quite depends on image contents. Such an example can be seen in Fig. 2. Fig. 2 shows the examples of MDB on different image contents. (a)–(d) are original high-quality images with size 512 × 512. Blur kernel with size 25 × 25 is shown upper right corner of images (a)–(d). (e)–(h) are motion blurred images of (a)–(d) using the blur kernel. (i)–(l) are motion deblurred images of (a)–(d) using algorithm [7]. (m)–(p) are motion deblurred images of (a)–(d) using algorithm [5]. It is observed from the figure that for a MDB algorithm, with different image contents, the visual quality of these motion deblurred images are quite different. Finally, unlike commonly seen artifacts, such as Gaussian blur and noise, the distribution of distortions in resorted images is usually localized and disparate. For example, in Fig. 1(c) and (d), the halo artifacts only occur in the sky region, not in the grassland. The ringing effects usually occur around high-contrast edges, which can be seen in Fig. 2(k), (m) and (o).

Based on the aforementioned analysis, it is easy to find that the quality evaluation of image restoration is quite different from traditional image quality assessment (IQA), and traditional quality metrics are not well suitable for evaluating the quality of restored images directly. First, traditional quality metrics are mainly devoted to gauging commonly seen artifacts, such as, Gaussian blur, noise, and JPEG/JPEG2000 compression. However, the distortions in restored images, such as, haze, texture unnaturalness, halo artifacts and ringing artifacts, are quite different from these commonly seen artifacts. Second, the traditional quality metrics usually measure evenly spread distortions, while the distortions in resorted images are usually localized and disparate. Third, the restored images are usually characterized by multiple distortions. However, the current quality metrics are usually designed and tested on computed-generated images with separate distortion types. In addition, the joint effect of residual distortions and side effects is important to visual quality of restored images and needs to be considered. However, it is usually ignored in the traditional quality metrics. So, traditional quality metrics are usually limited in predicting the quality of restored images due to the difference between two types of quality evaluation. This will be shown in the experiment section.

The traditional quality metrics are also limited in conducting the quality-related applications of image restoration, such as, parameter selection and benchmarking image restoration algorithms. For parameter selection, it is easy to understand that the limited performances of traditional quality metrics will lead to a wrong selection of optimal parameter. On the other hand, due to the limited performances of traditional quality metrics, it is easy to get wrong ranking for benchmarking image restoration algorithms, where the relative quality ranking of restored images that are generated via various algorithms is the most heavily considered factor.

Up to now, several objective quality metrics have been done to the perceptual evaluation of restored images for specific restoration scenarios, and several works have been dedicated to the applications of image restoration quality metrics [18–22]. Although notable success have been achieved, there are several disadvantages for these quality metrics. First, the application scope of these metrics is quite limited since they are scenario-specific. A scenario-specific quality metric can only evaluate the quality of restored images of the corresponding scenario and cannot address the restored images of other scenarios. Second, the performance of these metrics remains an open problem. Furthermore, they are usually designed and tested on one database with limited restored images and thus the robustness, which denotes the performance stability of a quality metric on different databases, is quite weak, which will be shown in the experiment section.

In parallel with the quality assessment of image restoration, the quality assessment of image enhancement becomes a novel task and some studies have been done towards it [23–25]. In [23], the authors proposed a quality assessment framework that focuses on the relative quality ranking between enhanced images for comparing image enhancement algorithms. To achieve this goal, an image enhancement dataset that contains three topics was built. For objective metric, a combination feature of the GIST descriptor [26] and the color motion [27] and rank SVM [28] were adopted in the proposed framework. Gu et al. systematically studied this issue. Two databases dedicated to image contrast change, three quality metrics and one quality optimization based image enhancement framework have been proposed in [24,29,30]. To analyze and compare the performance of different contrast enhancement evaluation measures, Qureshi et al. [25] contributed a comprehensive survey work. In [25], a detailed analysis and discussion on the subjective and objective quality assessment for image enhancement was first provided. Then, a psychophysical-based performance comparison of different contrast enhancement evaluation measures was presented based on a newly developed contrast enhancement evaluation database. One of the most important purpose of this study is expected to contribute substantially promoting the research area of image enhancement evaluation. Such comprehensive study can also be found in many quality assessment fields, such as image and video quality assessment [31–33], image inpainting quality assessment [34], document image quality assessment [35], etc. However, to the best knowledge of the authors, there is no such work in the field of quality assessment of image restoration, to date.

Given the unique characteristics of distortions introduced by image restoration process and the inappropriateness of traditional quality metrics to evaluate the quality of restored images, here we present this survey paper, which is expected to help current and emerging researchers working in this area in developing more accurate and robust quality metrics for quality evaluation and quality-related applications of image restoration. The paper first provides a comprehensive description of existing subjective quality databases of image restoration in terms of restoration scenarios. Then, a thorough insight into existing objective quality metrics for restored images is provided. After that, we provide two applications of image restoration quality metrics, namely, parameter selection and benchmarking image restoration algorithms. An insightful analysis and discussion about the existing work is also presented. Extensive experiments on quality assessment of restored images, benchmarking image restoration algorithms and time complexity are conducted based on public restored image databases. A detailed performance analysis is performed based on the experimental results. Finally, challenges and future trends of subjective and objective quality assessment for image restoration are given.

The rest of the paper is structured as follows: Section 2 briefly reviews traditional IQA. Section 3 provides the subjective quality databases for image restoration. Objective quality assessment for image restoration is provided in Section 4. Extensive experimental results are discussed in Section 5. Section 6 presents the challenges and future trends. Finally, the paper is concluded in Section 7.

2. Traditional image quality assessment

In the past few years, it has witnessed an explosive growth of objective IQA metrics. According to the availability of the reference image, objective IQA metrics can be classified into three categories, i.e., full-reference (FR) IQA, reduced-reference (RR) IQA and no-reference

(NR)/blind IQA. Since this paper aims to provide a critical review of subjective and objective quality assessment for image restoration instead of the traditional IQA themselves, here we will only provide a brief discussion of their advantages, limitations and disadvantages.

FR quality metrics require both reference and distorted images, and it can be further classified into several categories. There are a long array of classifications of FR quality metrics in the literature [36]. A brief summary and survey of more than 100 FR quality metrics was given by Pedersen and Hardeberg [37]. Afterwards, FR quality metrics were divided into four groups, namely, mathematically based metrics, low-level metrics, high-level metrics and a group of other metrics [38]. In [36], based on the available information during the design process, the authors proposed a novel knowledge-based taxonomic scheme for FR quality metrics. The hierarchical classification framework contains six layers and each layer represents a single type of knowledge about the available information. It is worth noting that motivated by the recent success of convolutional neural networks (CNN) for classification and detection tasks, the CNN-based FR metrics were rapidly developed [39–41] for superior performance. For more detailed discussions about FR quality metrics, the reader is referred to Refs. [36–38].

As a compromise between the FR IQA and NR IQA methods, the RR IQA method aims to predict image quality accurately with a limited amount of reference information. It can be further classified into statistics-based RR IQA methods and non-statistical RR IQA methods. Based on the assumption that natural images present statistical properties, the statistics-based RR IQA methods were built by extracting the statistical features in different domains [42–48]. However, these metrics do not perform consistently well across different distortion types, and the robustness of those statistics is questionable [47]. The second type of RR metrics does not use statistical features [49–52]. For these metrics, the performance is closely related to the amount of reference data [47].

For NR quality metrics, the reference images are unavailable and only the distorted images are used. It can be further classified into two categories, namely, distortion-specific metrics and general-purpose metrics. In distortion-specific metrics, only one type of distortion is evaluated, such as blocking artifacts [53,54], blur/sharpness [55], contrast distortion [56] and ringing effects [57], etc. One obvious weakness of these metrics is that they fail to evaluate other distortions. The general-purpose metrics are designed for evaluating image quality under various types of distortions [58–69]. Although the existing general-purpose NR quality metrics have achieved notable success in measuring distortions of common types, they are limited in evaluating the quality of authentically distorted images. The possible reasons are as follows: (1) the limited representation capability of handcrafted features towards complex mixtures of multiple distortions; (2) the limitation of traditional machine learning tools. In addition, the generalization ability of there metrics is quite limited. Relatively recently, CNN-based NR metrics have been rapidly developed and proposed for higher prediction accuracy [70–74]. It has been demonstrated the advantages of the CNN-based NR metrics compared with handcrafted features based metrics. However, the performance of CNN-based metrics heavily depends on the number of training data, and an overfitting problem is easy to occur with a small training data set.

3. Subjective quality assessment for image restoration

With the tremendous increase of research on image restoration algorithms, it is crucial to develop comprehensive subjective databases for quality evaluation of restored images and applications of image restoration quality metrics, such as, parameter selection and benchmarking image restoration algorithms. In this section, we describe the existing subjective quality databases for image restoration in terms of restoration scenarios. These restoration scenarios are IDN, MDB, defocus deblurring (DFD), SRR, IDH, image deblocking (IDK) and CSR. In order to provide valuable insights about the target topic history and serve as a "critical survey" to the readers, here we propose a clear logic framework to survey each restoration scenario. The logic framework, which is shown in Fig. 3, contains four logically related layers and each layer contains one issue, including what has been done related to the topic, what are their advantages and disadvantages, what does the recent work do and what problem has been solved by the recent work.

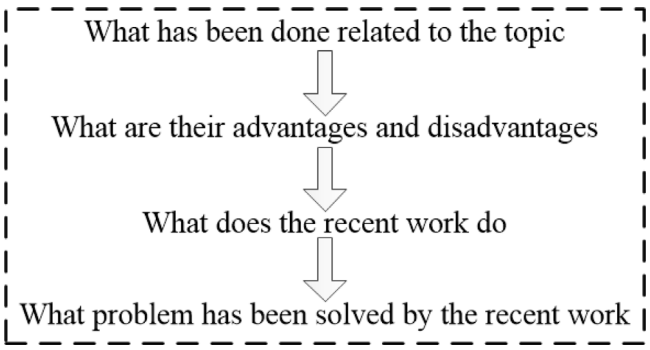

Fig. 3. The logic framework of surveying each restoration scenario.

IDN: In [75], a Tampere image database 2013 (TID2013) was built. IDN is included in this database. First, 25 reference images were distorted by noise at 5 distortion levels. Then, one denoising method was used to generate denoised images. In subjective experiment, the reference image and a pair of distorted images were simultaneously presented, which is a multi-stimulus (MS) method. Finally, each image was associated with a mean opinion score (MOS) in the range from 0 to 9. The TID2013 has been widely used to benchmark objective quality metrics. However, it cannot be used to benchmark image restoration algorithms, since only one denoising method was included in this database. To overcome this, Zeng et al. [76] built an image denoising quality database (IDQD) that contains several algorithms. First, white Gaussian noise was added to ten original high-quality natural images at three different standard deviations, producing thirty noisy images. Then, these images were processed by eight denoising algorithms. A total of 240 denoised images were generated after image denoising. In subjective experiment, all 8 denoised images from one noisy image were shown to the subject in random order. The subject was asked to rank the perceptual quality of these images from the best to the worst with 8 levels. After analysis of subjective data, namely, removing the invalid data of outlier subjects, the average ranking was computed as quality score. Although the database has been built under elaborate design, there are several limitations and disadvantages. First, showing all denoised images from one noisy image is impractical when there are a lot of algorithms. It is difficult to rank the perceptual quality of these images once and show so many images with their original size on a monitor. Second, the noise is synthetic and simple, which is far from the real noise type. More recently, the real noise type is considered in [77]. In [77], the authors subjectively evaluated the performance of four speckle reduction algorithms for phase shifting holograms using five different parametrizations for each case. For this purpose, a perceptual evaluation of speckle noise reduction database (PESNR) was first built. Five holographic images and four speckle reduction algorithms were selected for denoising. A total of 80 holographic images were generated from a set of 5 different filter parametrization levels. Then, the subjective test was carried out under the recommendation environment issued by ITU-R. The pair comparison method was used to collect the forced choice data. After that, the Thurstone Case V model [78] was used to estimate the quality score of each denoised images. Finally, the optimal parameter index of each denoising algorithm and the performance of these algorithms at the respective optimal parameter value were evaluated by making full use of the quality scores. Although the real noise was investigated, it is far from the comprehensive subjective database.

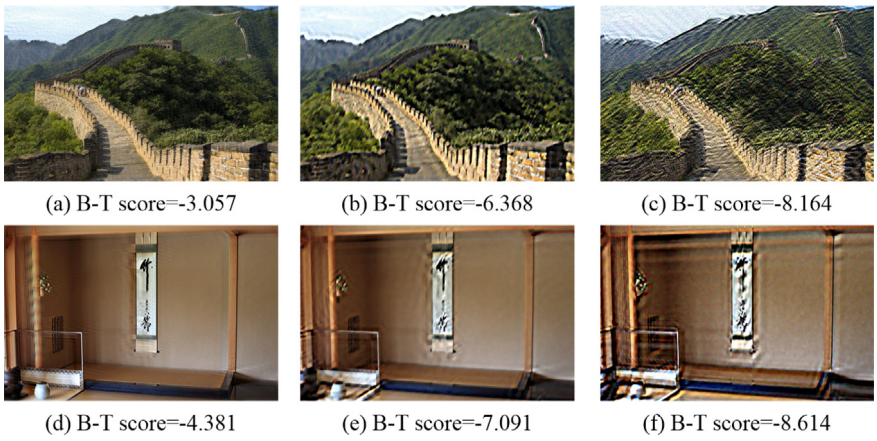

Fig. 4. Illustration of some examples in MDD13 database.

Fig. 5. Illustration of some examples in DDID database.

First, only speckle noise of phase shifting holograms was considered, other noise types, such as quantization noise, high frequency noise and non-eccentricity pattern noise, are neglected. Second, only four speckle reduction algorithms were included, and they are difficult to represent state-of-the-art denoising algorithms. In addition, too few images to build a comprehensive database.

MDB: In [18], the authors built a motion deblurring database 2013 (MDD13). First, forty high-quality natural images were selected, and then two different point spread functions (23×23 and 27×27) were used to blur these images. To simulate more realistic motion blur, each blurred image was added Gaussian noise of three different levels. A total of 240 blurred images were generated. Then, five motion deblurring algorithms were used to generate 1199 deblurred results. In user study, the Amazon Mechanical Turk (MTurk) was employed. A pair of deblurred images was shown side-by-side, and each user was asked to compare and rank the quality of deblurred results. Finally, the Bradley–Terry (B-T) model [79] was used to generate a global quality score for each image from the pairwise comparison results. Illustration of some examples are shown in Fig. 4. Although some efforts have been made, the following aspects could be done better. First, the point spread functions with larger size should be used and it can simulate more realistic motion blur. Second, the motion blur is synthetic and thus it is difficult to simulate complex real scenario. To solve these problems, a large scale motion deblurring quality database was built in [80]. It can be divided into three image subsets, namely, motion deblurring for real image database (MDRID), motion deblurring for non-uniform blurred image database (MDNID) and motion deblurring for uniform blurred image database (MDUID). For MDRID, 100 real blurred images were selected from multiple sources, such as images from the Internet and pictures captured by authors, etc. For MDNID, 25 high-quality natural images from the Internet were selected as ground truth. Then, 100 non-uniform blurred images were obtained by applying four camera trajectories and adding 1% Gaussian noise to simulate camera noise. For MDUID, 100 uniform blurred images were obtained by a convolution operation between the high-quality natural images and four uniform blur kernels with size ranging from 51×51 and 101×101. Then, 13 representative state-of-the-art algorithms were used to generate deblurred images. After image deblurring, a total of 1291 real deblurred images, 1296 non-uniform deblurred images and 1291 uniform deblurred images were generated in MDRID, MDNID and MDUID databases. In subjective study, the pair comparison approach was used. Finally, the B-T model was used to obtain a global quality score for each deblurred image. Compared with the work [18], it is more comprehensive and well designed. However, the pair comparison

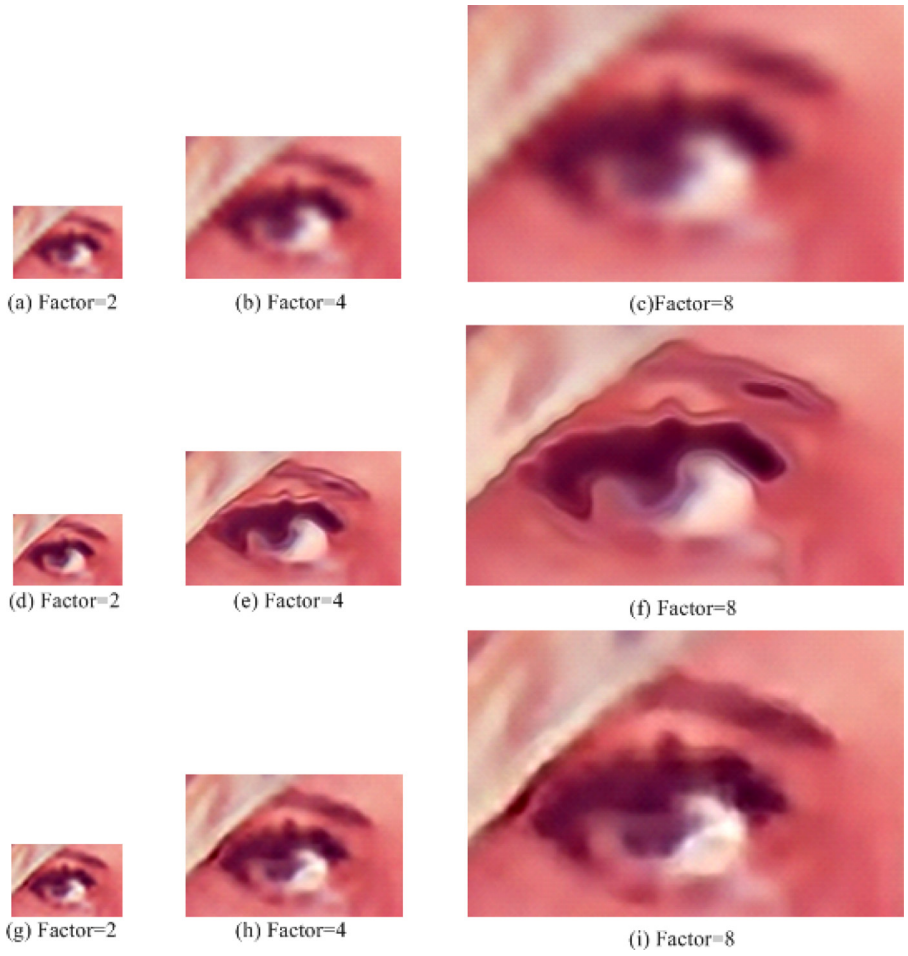

Fig. 6. An example of SR reconstructed images in SRID database.

approach is time consuming and not efficient when lots of images need to be compared. In addition, it is sometimes difficult to differentiate between two stimuli. Therefore, more effective and efficient compare-based methods are highly desirable.

DFD: A defocus deblurred image database (DDID) was built in [81]. To simulate different distortion levels, namely, slight, medium and heavy, ten original images were first processed using the Gaussian low-pass filter with different parameters. A total of 30 defocus blurred images were generated. Then, each blurred image was processed by eight state-of-the-art image defocus deblurring algorithms. Next, the subjective test was conducted using single-stimulus (SS) method. Finally, MOS was provided after necessary postprocessing of subjective scores. Illustration of some examples are shown in Fig. 5. Although the SS method was widely used in literature, it is more sensitive to the context and may be biased by the visual content [22].

SRR: Yang et al. [82] built a SR reconstructed image database, which we call SRIDY. To generate low-resolution images, ten source images were processed by nine different combinations of downsampling and blurring. Then, each low-resolution image was processed by six SR image reconstruction algorithms, producing 540 SR reconstructed images. Next, human subject study was conducted to assess the quality of these images based on their visual perception. Finally, MOS was provided. In [83], the authors expanded the SRIDY database, which we call SRIDM. First, thirty source images were selected, and each image was processed by six different combinations of downsampling and blurring to generate 180 low-resolution images. Then, nine SR image reconstruction algorithms were used to generate SR reconstructed images. In the subject study, absolute rating scores were used to collect subjective scores by simultaneously showing 9 SR reconstructed images. Finally, MOS was utilized as the subjective scores. The main disadvantage of the SRIDM database is that the low-resolution images were generated by a composite operator of blurring and down-sampling based on high-resolution original images, which is not natural. To simulate the real application scenario, another SR reconstructed image database (SRID) was built in [84]. First, twenty low-resolution natural images with diversified contents were selected. Then, two interpolation algorithms and six SR image reconstruction algorithms were used to generate high-resolution images. To obtain different distortion levels, each low-resolution image was processed with three different amplification factors, namely, 2, 4 and 8. An example of SR reconstructed images is shown in Fig. 6. A total of 480 SR reconstructed images were obtained. The SS method was used in subjective experiment. Fig. 7 shows the graphical user interface. After removing the maximum and minimum scores for each image, MOS was obtained by averaging the remaining subjective scores. One of the main disadvantages is that these algorithms are all traditional and deep learning based algorithms are excluded. Therefore, it is not considered as a comprehensive database. To solve this, both traditional and deep learning-based algorithms are used to build a large-scale database in [85]. In [85], a quality assessment database for SR reconstructed images (QADS) was built. Twenty reference images with size 504×384 were selected, and then low-resolution images were obtained using bicubic down-sampling by a factor of k ($k = 2, 3, 4$). Next, 21 SR image reconstruction algorithms were used to super-resolve these low-resolution images back to their original sizes. In subjective evaluation, the MS method was used to collect subjective scores. After data analysis and processing, MOS was provided. Although notable efforts have been dedicated to develop the comprehensive database for SR reconstructed images, there is an

obvious limitation. The process of generating low-resolution images is manned and not natural.

IDH: In [23], the authors built an enhanced image quality database, which contains three topics, i.e., image dehazing, low light scene enhancement and underwater image enhancement. Here we only discuss the scenario of image dehazing and the corresponding database is called IDHD for short. First, 100 hazy images from image search engines and related literature were collected, and then 5 enhancement algorithms were used to process these images. After that, subjective test was carried out under pair comparison method. The interface of labeling was shown in Fig. 8. Finally, the quality rank was given. Although the database is well designed, there are several limitations. One limitation is that the number of enhancement algorithms is too small. Another limitation is that the quality rank cannot accurately reflect the relative quality difference of the images. In [86], the authors built a dehazed image quality database (DHID). 25 hazy images were processed by eight dehazing algorithms, producing 200 dehazed images. Then, subjective user study was conducted using MS method by showing 9 images (one hazy image and eight dehazed images) simultaneously. Then, MOS was obtained by averaging the subjective results of all subjects. In [87], a synthetic haze removing quality (SHRQ) database was built, and it can be divided into regular and aerial image subsets, which we call SHRQR and SHRQA, respectively. For SHRQR, 45 high quality haze-free regular images were selected, and then the atmospheric scattering model was used to generate synthesized hazy images. Then, eight state-of-the-art image dehazing algorithms were used to process these images. For SHRQA, 30 high quality aerial images were selected and the synthesized hazy images were obtained via atmospheric scattering model. The same 8 image dehazing algorithms were used to generate 240 aerial dehazed images. In subjective testing, the MS method was used by showing the reference haze-free image, the hazy image and the dehazed image simultaneously. After data processing and analysis, the MOSs were used to represent image quality. One advantage of the work [87] is that it created an aerial image subset, since aerial imaging is an important application area of dehazing. However, these images are synthetic hazy images and thus the distortion characteristics of these dehazed images may be quite different from the real dehazed images. To solve this, Min et al. [88] further built a dehazing quality database (DHQD) using real hazy images. First, 250 hazy images that have various haze densities were selected, and then seven representative image dehazing algorithms were used to generate 1750 dehazed images. After that, subjective quality evaluation study were conducted, and DS strategy was adopted. Finally, MOSs are taken as the ground-truth quality of the dehazing after subjective data processing. This database is the largest of its kind so far. However, CNN-based image dehazing algorithms are not included.

IDK: In [89], a deblocked image database (DBID) was built. Twenty natural images with size 512 × 512 were selected, and each image was subject to JPEG compression with three different distortion levels to simulate the blocking artifacts. Then, six state-or-the-art image deblocking algorithms were used to process these JPEG images, producing 360 deblocked images. In subjective experiment, DS method was used to obtain more accurate subjective scores by showing the reference images. Finally, MOS was obtained after postprocessing and analysis of subjective scores. Limited number of image deblocking algorithms and natural images is one of the main limitations of this database.

CSR: In [90], the authors built a CS recovered image database (CSRID). Ten original high-quality natural images were selected, and ten representative CS recovery algorithms were used to generate the recovered images with three different sensing rates, producing 300 recovered images. Then, subjective experiment was conducted to collect image quality score by SS method. After analysis of subjective scores, MOS was obtained to represent image quality. Fig. 9 shows an example of CS recovered images in CSRID database.

For the convenience of the reader, the information of these databases is summarized in Table 1, where "Nof" denotes the number of reference images, "Nol" denotes the number of distortion levels, "Nod" denotes the number of degraded images, "Noa" denotes the number of algorithms, "Yes/No" denotes if the CNN-based algorithms are included in the database, "Nor" denotes the number of restored images, "Real/Synthetic" denotes the type of distortion in degraded images. From the table, we conclude that: (1) The research of subjective quality assessment for DFD, IDK and CSR is relatively less than that of MDB, SRR and IDH; (2) There are limited restored images in the existing databases due to the limited reference images (the maximum number of reference images is only 45), the limited levels of distortion and the limited algorithms (there are no more than 10 algorithms in most databases); (3) For most databases, the selected algorithms are traditional and the deep learning based algorithms are not included; (4) Most of the distortions in degraded images are synthetic, and little work has been done to the degraded images with real distortions.

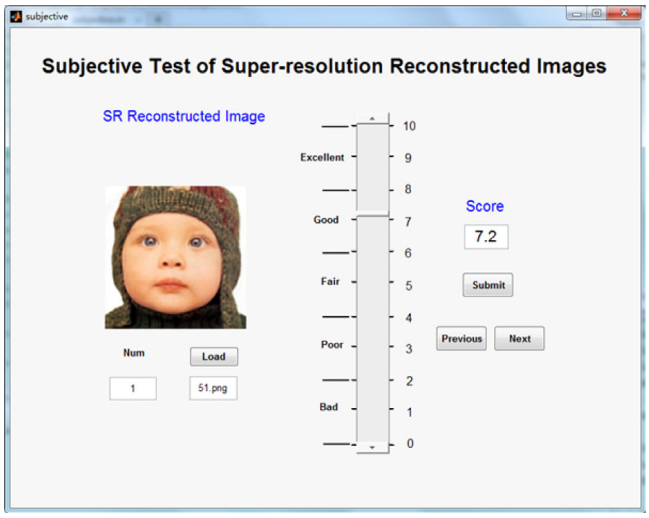

Fig. 7. The graphical user interface used in [84].

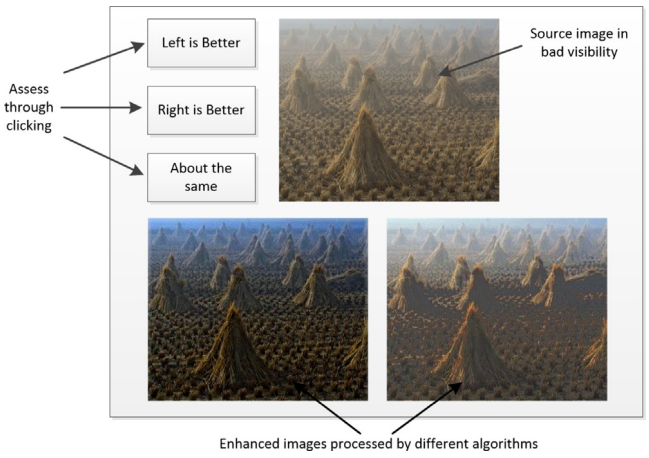

Fig. 8. The interface of labeling used in [23].

4. Objective quality assessment for image restoration

Guided by the proposed logic framework, we give the up-to-date overviews on objective quality assessment of restored images and two applications of image restoration quality metrics in this section.

4.1. Quality assessment of restored images

Objective quality assessment of restored images can be grouped into three categories, namely, FR, RR and NR, which is shown in Fig. 10.

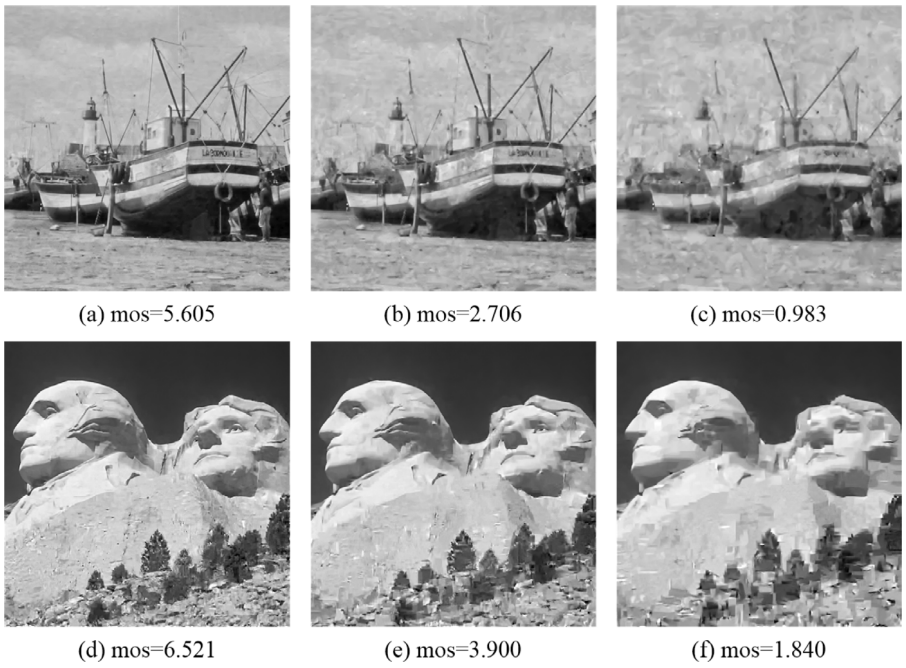

Fig. 9. An example of CS recovered images in CSRID database [90].

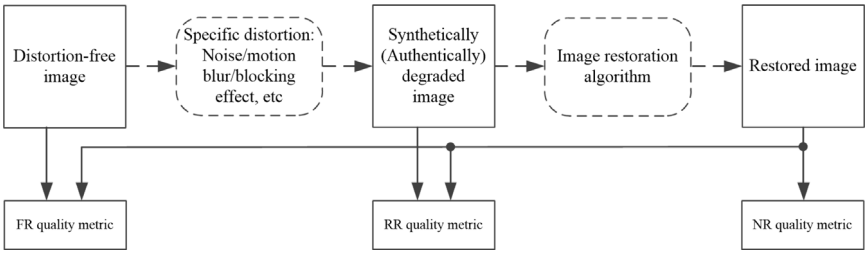

Fig. 10. Three categories of objective quality assessment of restored images.

Table 1
A summary of subjective quality assessment databases for image restoration. "Nof" denotes the number of reference images, "Nol" denotes the number of distortion levels, "Nod" denotes the number of degraded images, "Noa" denotes the number of algorithms, "Yes/No" denotes if the CNN-based algorithms are included in the database, "Nor" denotes the number of restored images, "Real/Synthetic" denotes the type of distortion in degraded images.

Scenario	Database	Nof	Nol	Nod	Noa	Yes/No	Nor	Stimulus	Sub. score	Real/Synthetic	Year	Ref.
IDN	TID13	25	5	125	1	No	125	MS	MOS	Synthetic	2013	Ref. [75]
	IDQD	10	3	30	8	No	240	MS	MOS	Synthetic	2013	Ref. [76]
	PESNR	–	5	4	4	No	80	DS	MOS	Real	2019	Ref. [77]
MDB	9mmMDD13	40	6	240	5	No	1199	DS	B-T	Synthetic	2013	Ref. [18]
	9mmMDRID	–	–	100	13	No	1291	DS	B-T	Real	2016	Ref. [80]
	9mmMDNID	25	4	100	13	No	1296	DS	B-T	Synthetic	2016	Ref. [80]
	9mmMDUID	25	4	100	13	No	1291	DS	B-T	Synthetic	2016	Ref. [80]
DFD	DDID	10	3	30	8	No	240	SS	MOS	Synthetic	2016	Ref. [81]
SRR	SRIDY	10	9	90	6	No	540	–	MOS	Synthetic	2014	Ref. [82]
	SRIDM	30	6	180	9	Yes	1620	MS	MOS	Synthetic	2017	Ref. [83]
	SRID	20	3	60	8	No	480	SS	MOS	Real	2018	Ref. [84]
	QADS	20	3	60	21	Yes	980	MS	MOS	Synthetic	2019	Ref. [85]
IDH	IDHD	–	–	100	5	No	500	MS	Rank	Real	2014	Ref. [23]
	DHID	–	–	25	8	No	200	MS	MOS	Both	2015	Ref. [86]
	SHRQR	45	1	45	8	Yes	360	MS	MOS	Synthetic	2019	Ref. [87]
	SHRQA	30	1	30	8	Yes	240	MS	MOS	Synthetic	2019	Ref. [87]
	DHQD	–	–	250	7	No	1750	DS	MOS	Real	2019	Ref. [88]
IDK	DBID	20	3	60	6	No	360	DS	MOS	Synthetic	2016	Ref. [89]
CSR	CSRID	10	3	30	10	No	300	SS	MOS	Synthetic	2016	Ref. [90]

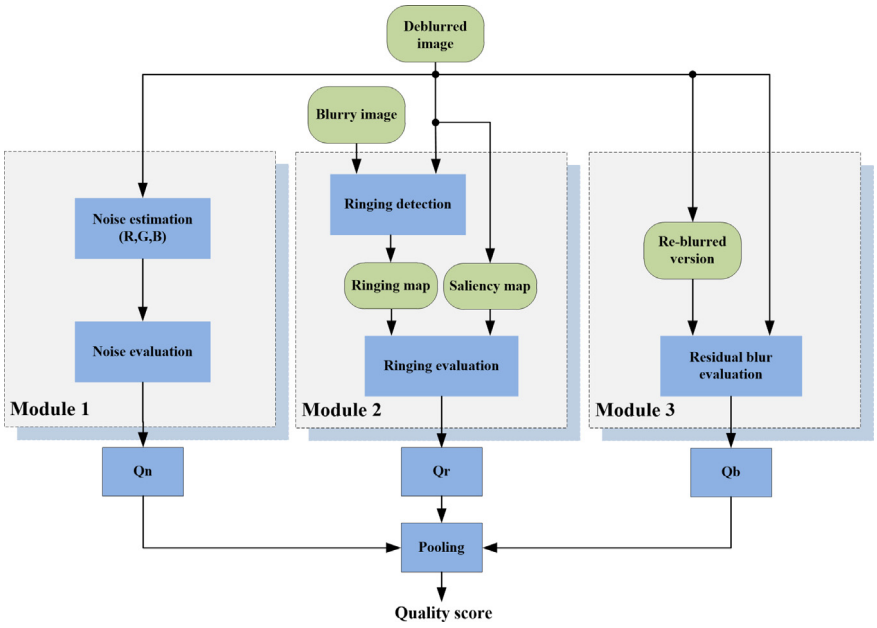

Fig. 11. The framework of the NRRB [91].

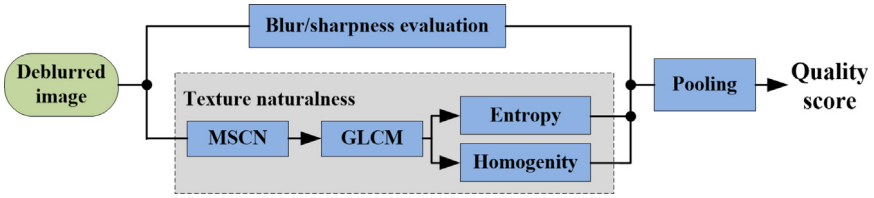

Fig. 12. The framework of the proposed method [81].

The only difference from the traditional IQA is that the degraded image rather than the reference image is used in RR quality metric for image restoration. For better shape of presentation, here we discuss each quality metric in terms of restoration scenarios.

IDN: After IDQD database construction, several meaningful experiments were carried out by the authors [76]. First, the agreement among different subjects on ranking the quality of denoised images was evaluated by computing the consistency between individual subject and average subject rankings. Then, a relative performance comparison of the image denoising algorithms was provided. Finally, 12 FR, 2 RR, and 4 NR objective IQA metrics were tested on the IDQD database. The experimental results show that there is significant space for future improvement of objective IQA metrics. Another conclusion is that more accurate objective IQA metrics for denoised images can be built by combining structural fidelity and naturalness measures. As discussed above, this study only focuses on simple Gaussian noise, and thus it is limited in real noise scene. In [92], a database of color images corrupted by natural noise due to low-light conditions was built and several issues about noise reduction have been studied based on the established database. First, the authors collected low-light uncompressed natural images of 120 scenes using two digital cameras and a mobile phone camera. Then, the authors evaluated six popular image denoising algorithms using three FR quality metrics, namely, peak signal to noise ratio (PSNR), structural similarity (SSIM) [93] index and visual signal to noise ratio (VSNR) [94]. In addition, the accuracy of the proposed noise estimation method was evaluated and the Poisson–Gaussian noise model were investigated based on the acquired real low-light noisy images. From these reviews, more efforts should be put into developing effective and robust quality metrics for real denoised images.

MDB: In [18], a RR quality metric for evaluating the quality of motion deblurring (MMD) was proposed by measuring naturalness and sharpness. Specifically, 11 different features were extracted to portray the noise, ringing, saturation and sharpness in deblurred images. Then, a feature selection strategy was proposed to select the optimal subset of collected features. Finally, eight features were selected and logistic regression model was used to produce the final quality score. One of the main limitations of this algorithm is that it performs not well on motion deblurred images with large-scale ringing effects. The reason is that when measuring large-scale ringing effects, the motion blur image is not suitable as a reference image. It makes imprecise ringing detection. Inspired by [18], Hu et al. [91] proposed a new quality metric for motion deblurred images by measuring noise, ringing and residual blur (NRRB). The diagram of the NRRB is shown in Fig. 11. First, noise estimation and evaluation were conducted in red, green and blue color channels. Then, with the guidance of motion blurred images, ringing effects were detected and then evaluated by fusing the characteristic of visual saliency of HVS. A re-blurred method was proposed to measure residual blur. Finally, the overall quality score of a motion deblurred image was obtained by pooling these three scores. Notable advances have been made in evaluating the quality of deblurred images. However, the ringing detection method are largely depends on the distortion intensity of motion blurred images; hence, they usually do not perform well on the deblurred images that recovered from the images with severe motion blur.

DFD: In [81], a quality enhancement module was proposed based on gray level co-occurrence matrix (GLCM) for the existing blur/sharpness evaluation metrics. The framework of the metric is shown in Fig. 12. Specifically, entropy and homogeneity were used to measure the loss of texture naturalness caused by deblurring. The final quality score of a defocus deblurred image was obtained by a combination of blur score, entropy score and homogeneity score. It proves the effectiveness of additional measurement of local texture

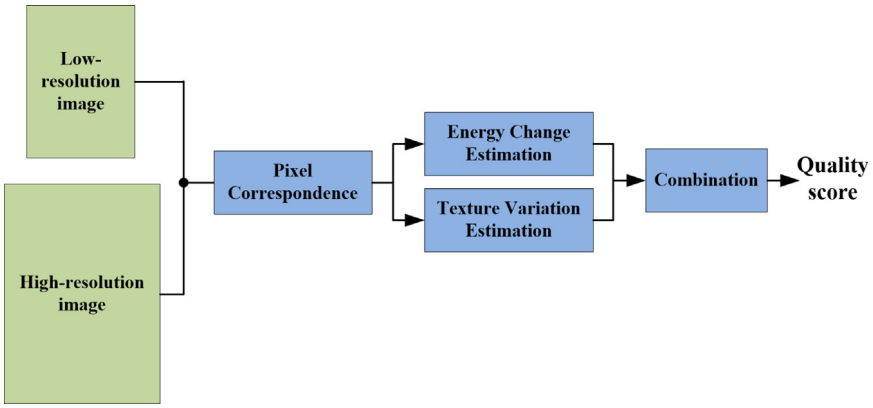

Fig. 13. The framework of the proposed method in [95].

naturalness. Inspired by this, Li et al. [96] further proposed a natural scene statistics (NSS) based no-reference quality metric for defocus deblurred images. The spatial and frequency domain features were extracted to account for both the global and local aspects of distortions in deblurred images. Specifically, in spatial domain, the naturalness factor was used to characterize the global naturalness. In frequency domain, the distribution of Log-Gabor coefficients was employed to portray the local structural distortions. After feature extraction, SVR was used to learn a quality model. Experimental results have demonstrated the effectiveness of the NSS based method. Although massive efforts have been made, there are several limitations and disadvantages. First, with so many handcrafted features, some of the NSS features may be correlated with others. Therefore, it is necessary to reduce its dimension, which can both make the quality prediction more efficient and reduce the computational cost. Second, the generalization ability, which is an important issue for learning-based quality metrics, is not investigated due to the lack of other databases.

SRR: In [83], a no-reference quality metric for single-image SR (SISR) was proposed based on low-level statistical features. Three types of statistical features were extracted to quantify artifacts in SR reconstructed images. Specifically, the statistics of coefficients from the discrete cosine transform (DCT) were extracted as local frequency features. The statistics of wavelet coefficients were used to quantify global super-resolved artifacts. Singular values based features were extracted as spatial domain features. Then, a two-stage regression model was used to estimate the quality score. Although the SISR performs favorably against state-of-the-art quality metrics, it runs slowly due to complex feature extraction process. In addition, the generalization ability needs further investigation. Fang et al. [19] proposed a deep learning based metric to predict the visual quality of image SR. A six-layer convolutional neural network was designed to extract the high-level intrinsic features. Finally, one regression layer was used to produce quality score. The small patches (160×160) taken from SR reconstructed images to form the training set, and the same score was assigned to the corresponding patches. The proposed metric has a simple network structure and can be easy to follow. In [97], the authors proposed a deep CNN model for NR quality assessment of SR metric. The network architecture consists of six convolutional layers, two max pooling layer, three skip connections, and two fully connected layers. The final quality score of a SR reconstructed image was computed by averaging the predicted score of each patches (32×32). As everyone knows, lack of training data is one of the main obstacles for CNN-based quality metrics. To alleviate the problem, the input image is divided into multiple patches [19,97]. However, the patch-based method is problematic due to imprecise local scores. The local score varies dramatically depending on the local image structure and distortions. More effective strategies for assigning local scores need to be investigated and proposed for more accurate quality prediction. The aforementioned methods are NR quality metrics. However, the low-resolution images can provide plenty of useful information. Inspired by this, Fang et al. [95] further proposed a RR quality assessment metric for image SR (RRIQA-SR) based on the low-resolution image information. The framework of the RRIQA-SR is shown in Fig. 13. First, pixel correspondence between low-resolution and high-resolution images was conducted using scale-invariant feature transform (SIFT) descriptors. Then, the similarities were computed to measure energy change and texture variation between low-resolution and high-resolution images in DCT domain. Finally, the visual quality of SR reconstructed image was obtained by combining these two components. More recently, a FR quality assessment metric for image SR (FRIQA-SR) was proposed by separately considering the structural and textural components of images [85]. Dissimilar texture or checkerboard artifacts were measured by computing the changes of textural distributions. Blurring and jaggies artifacts were also measured to portray the structural degenerations. Finally, to obtain the overall quality score of a SR reconstructed image, a new pooling mechanism was proposed to fuse the different similarities together. However, this metric is a FR one, so a high-quality pristine image is needed and thus the application scope is quite limited.

IDH: In [98], the authors proposed three metrics for dehazed images by computing the changes of the visible edges between the image before and after contrast restoration. Specifically, they are the percentage of pixels that become pure black or white, the rate of appearance of new edges and the mean ratio of the gradients at the visible edges. They are widely used in the performance comparison of image dehazing algorithms [11]. However, measuring these simple aspects alone is difficult to accurately predict the quality of dehazed images. This will be shown in the experiment section. In [87], two FR quality assessment metrics for image dehazing were proposed. For regular image dehazing, image structure recovering, color rendition and over-enhancement of low-contrast areas were measured, and the overall score of a dehazed image was obtained by pooling these aspects. For aerial image dehazing, by incorporating the specific characteristics of aerial images, an improved method was built. Experimental results validated the effectiveness of these two metrics. However, they are FR quality metrics that use both distortion-free images and dehazed images, and thus they are very limited in real image dehazing. In addition, the metric that is designed based on synthetic hazy images may not be reliable due to the differences between synthetic haze and real haze. To solve these problems, a RR dehazing quality index (DHQI) was proposed by extracting and integrating three groups of features based on real dehazed images [88]. First, with the guidance of hazy images, three groups of features, namely, haze-removing features, structure-preserving features and over-enhancement features were extracted to account for the most key aspects of dehazing. Then, all features were combined to train a SVR model for quality prediction of dehazed images. However, the hazy image, as an approximate reference, was directly used to measure structural distortion in this metric, which is questionable. Another

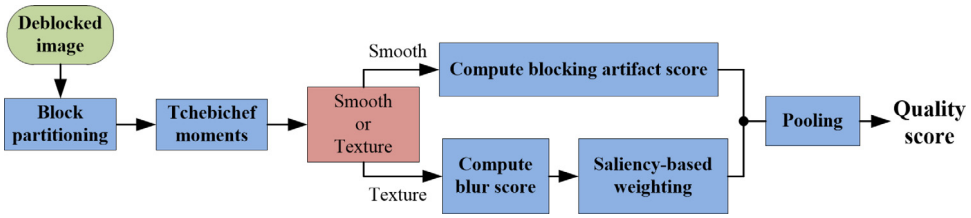

Fig. 14. The framework of the DBIQ metric [89].

limitation is that the internal relations of these three groups of features are not investigated. These can be the future work.

IDK: Yim et al. [99] proposed a FR PSNR including blocking effects (PSNR-B) metric, which modified the classical PSNR by including a blocking effect factor. Although it performed better than the classical PSNR, the application scope is still limited because of the need for distortion-free reference image. In [89], by simultaneously evaluating blur in textured regions and blocking artifacts in smooth regions, a no-reference deblocked image quality (DBIQ) metric was proposed. The framework of the DBIQ metric is shown in Fig. 14. First, a deblocked image was divided into smooth regions and textured regions based on the block energy. Then, a discrete moment-based blocking artifact metric [53] was used to measure the strength of blockiness in smooth regions. In textured regions, a further-blurred method was used for blur evaluation. Finally, an overall score of a deblocked image was generated by pooling blocking artifact score and blur score. This is a pioneering work that has been done towards the perceptual evaluation of image deblocking. The DBIQ is designed based on the discrete orthogonal moments of gray-scale images. However, color information also affects the visual quality of deblocked images. There, further improvement of the performance could be investigated by incorporating color information. A straightforward way is to use quaternion-type moments [100]. In [101], a no-reference quality assessment technique for deblocked images using continuous wavelet transform (DBCWT) was proposed. A given deblocked image was first divided into blocks of size 32×32. Then, these squared blocks were further divided into smooth and textured blocks based on the average value of CWT coefficients. After block classification, estimation of blocking artifacts in smooth region and estimation of blur in textured region were conducted to obtain blocking artifacts score and blur score. Finally, the overall quality score was computed by combining these two aspects. Although the above quality metrics have achieved notable success towards the quality evaluation of deblocked images, the joint effect of blur and blocking artifacts is ignored. To solve this problem, Hu et al. [102] proposed an internal generative mechanism driven blind quality metric for deblocked images (IGMDB). A deblocked image was first decomposed into two the predicted and disorderly portions. Then, blocking artifacts evaluation, blur evaluation and joint effect evaluation were conducted in the predicted portion. Finally, all extracted features were fed into random forest (RF) to learn a quality prediction model. For further improvements, an extension method of IGMDB was proposed in [103], which we call IGMDB-II. In [103], both the predicted and disorderly portions were evaluated. For the predicted portion, blocking artifacts were evaluated based on the ratio of regular corner and the block boundary strength, blur was evaluated based on the strength of gradient, and joint effect was evaluated in Curvelet domain. For disorderly portion, uncertain information was evaluated based on Renyi entropy feature. After feature extraction, RF was used to map all features into an overall quality score. Although it outperforms the relevant quality metrics, it takes a long time to run. In addition, generalization ability remains an open problem.

CSR: In [104], Hu et al. proposed a no-reference CS recovered image quality (CSRIQ) metric based on the measurement of both local and global distortions in CS recovered images. The framework of the CSRIQ metric is shown in Fig. 15. The local features consist of a gray level co-occurrence matrix based texture measure and a local phase coherence based edge sharpness measure. The global features were computed based on the statistics of the coefficients of singular value decomposition. Also, statistics of MSCN coefficients were extracted as a group of global features. After feature extraction, SVR was used to learn a quality prediction model for CS recovered images. This is one of the first studies on perceptual evaluation of CS recovered images. Further improvement could be investigated by emphasizing high distortion regions. The reason is that salient local artifacts greatly affect the perception of HVS and the CS recovered images are usually degraded by localized and disparate distortions.

Although notable efforts have been dedicated to the quality evaluation of restored images, there are still some limitations and disadvantages. From the survey, we conclude that: (1) There is no general image restoration quality metric, and the existing image restoration quality metrics are usually scenario-specific. One common weakness of these metrics is that their application scope is quite limited, since they can only address restored images of the corresponding scenario and cannot evaluate the quality of restored images in other scenarios; (2) In most quality metrics, only distortion-specific features were extracted and used to characterize the individual distortions. The joint effect of these individual distortions is usually neglected; (3) The features used in the existing image restoration quality metrics are almost handcrafted and low-level, and thus it may not be sufficient for quality evaluation of restored images.

4.2. Application of image restoration quality metrics

In this subsection, we give the overviews on two applications of image restoration quality metrics, namely, parameter selection and benchmarking image restoration algorithms. Most image restoration algorithms contain user-defined parameters, and thus parameter selection is of importance to these image restoration algorithms for faster convergence rate or better visual quality of restored image. On the other hand, with so many image restoration algorithms at hand, how to selecting the best algorithm is of importance to obtain a better restored image; hence, benchmarking image restoration algorithms is another important application of image restoration quality metrics.

4.2.1. Parameter selection

In [105], an image content metric Q (MetricQ) was proposed based upon singular value decomposition of local image gradient matrix for automatic parameter selection for denoising algorithms. Two indexes, i.e, image content index and coherence index, was defined to reflect the quality of a local patch. The overall score of an image was computed as a combination of two indexes. Finally, the proposed MetricQ was used to optimize the parameters of two denoising algorithms. It provides good visual performance in balancing between denoising and detail preservation and does not require intensive computation. However, it does not work as a general NR quality metric, since only structured regions is measured. From this perspective, it cannot handle situations where noise only appears in smooth regions. In [106], the authors proposed an approach for parameter trimming. This approach consists of three elements, namely, IQA, image reconstruction and parameter trimming. It used MetricQ [105] for IQA after each iteration and then

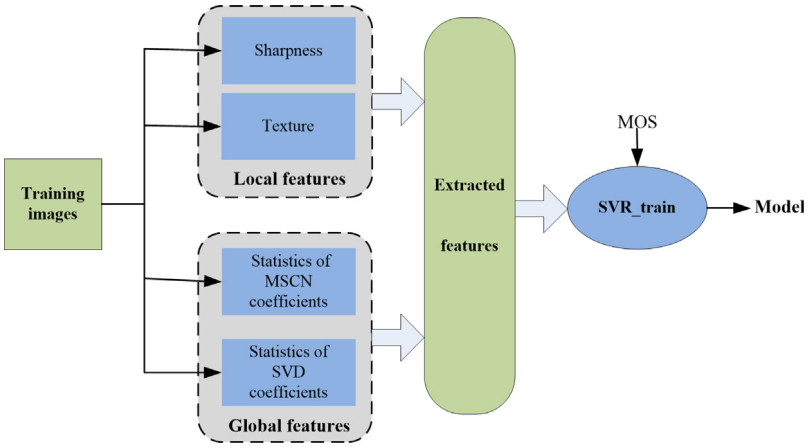

Fig. 15. The framework of the CSRIQ metric [104].

predicted the convergence trend corresponding to each value of the parameter. Finally, a parameter trimming criteria was proposed for reduction in total number of iterations while still selecting the best parameter. In this paper, a single parameter is measured. However, for most image restoration algorithms, there are multiple parameters. So, more comprehensive parameter trimming methods should be investigated in future. In addition, the MetricQ [105] estimates image quality based on a single denoised image, ignoring the information provided by the degraded images. To make full use of the available information, Liang et al. [20] further proposed a comparison-based IQA (C-IQA) framework for selecting image restoration parameters. The C-IQA method consists of two basic modules, i.e, content detection and contribution and one optional module, i.e, distortion sensitivity weighting. Then, the proposed C-IQA was used for parameter selection of image reconstruction and image denoising algorithms. The parameter trimming framework combined with C-IQA saves plenty of computation time during iterative image reconstruction process. One of the main advantages is that the proposed framework can be used for different application scenarios by integrating other image quality metrics. In [21], two NR quality metrics were proposed for image auto-denoising. The first metric, which named NR structure similarity measure (NRSS), evaluates the denoised images with two measurements, i.e, the noise reduction and the structure preservation. Based on the fact that more noisy measurements will be required when the noise level increases, the second metric, which named patch matching measure (NRPM), was proposed for high-level noise. It works to iteratively find and add noisy measurements for averaging until the difference between two successively averaged images is no visible. Then, both the proposed metrics were used for image auto-denoising. Although very promising results have been achieved, there are limitations of these two metrics. For NRSS, it is only robust to low-level noise and cannot directly give an estimate of the noise level. For NRPM, it is invalid for random textures (e.g., tree and grass) and spatially varying noise. In [107], a whiteness-based criteria was proposed to select the regularization parameter and to stop iterative blind and non-blind image deconvolution algorithms. The rationale is that if the deblurred image is well estimated, the residual image is relatively white; contrarily, a poorly deblurred image typically exhibits structured artifacts, such as, ringing and over-smoothness, yielding residual signals that are not relatively white.

4.2.2. Benchmarking image restoration algorithms

In [108], a denoising algorithm selection framework that chooses among different denoising algorithms using comparison-based image quality assessment was proposed. The proposed metric consists of distortion detection, contribution and texture compensation, and outputs a scalar number indicating the relative quality of the first image based on the second. For denoising method selection, the bubble sort algorithm was adopted to extend the proposed metric to compare multiple images, and then ranked them. In [22], a pairwise-comparison-based rank learning framework was proposed for benchmarking image restoration algorithms. The proposed framework was inspired by the fact that the ranking of restored images that are generated via various algorithms is the most heavily considered factor. First, a rank learning model was trained using the differential feature vectors (DFVs) of a set of restored image pairs and the corresponding preference labels. Then, the trained model was used to predict the preference labels of test image pairs. Finally, the ranking of these images was obtained via a simple statistical strategy. If the mean rank value of the restored images that were generated by this algorithm is high, then it is considered to have good restoration performance. This framework have achieved notable success in benchmarking image restoration algorithms with different restoration scenarios. Meantime, it was designed based on binary classification and cannot rank multiple images directly. A straightforward way to solve this problem is to use list-wise learning-to-rank algorithm by concatenating different image pairs.

Although great efforts have been done to the two applications of image restoration quality metrics, there are still some limitations and disadvantages. From the survey, we conclude that: (1) For parameter selection, most efforts have focused on image denoising and little work has been done to other restoration scenarios. In addition, these parameter selection frameworks were usually tested on several commonly seen denoising algorithms, and the performance of these frameworks on other denoising algorithms remains an open problem; (2) For benchmarking image restoration algorithms, the existing metrics, which are designed based on a pairwise comparison, cannot rank multiple images directly. Furthermore, the performance of these metrics on benchmarking deep learning based restoration algorithms is quite questionable, since they are designed based on restored images generated by traditional algorithms.

5. Experimental results and discussions

In order to investigate the performances of the traditional quality metrics and quality metrics for image restoration, extensive experiments are conducted on nine public databases of three commonly seen restoration scenarios in terms of quality assessment of restored images, benchmarking image restoration algorithms and time complexity in this section.

5.1. Experimental settings

Nine public databases of three commonly seen restoration scenarios are used in experiments. They are: (1) three motion deblurring

databases, namely, MDD13 [18], MDUID [80] and MDRID [80]; (2) three image dehazing databases, namely, DHID [86], SHRQR [87] and DHQD [88]; (3) three SR image reconstruction databases, namely, SRIDM [83], SRID [84] and QADS [85]. It is noteworthy that image denoising, as one of most common image restoration tasks, is not included due to the lack of suitable database. Specifically, IDQD [76] is not public and thus unavailable. In TID13 [75], only one denoising algorithm was used to generate denoised images, and therefore it cannot be used to conduct the experiment of benchmarking image denoising algorithms.

Four criteria are used to quantitatively measure the performance, including Spearman rank order correlation coefficient (SRCC), Kendalls rank order correlation coefficient (KRCC), Pearsons linear correlation coefficient (PLCC) and root mean square error (RMSE). SRCC and KRCC are used to measure the prediction monotonicity, while the prediction accuracy is evaluated by PLCC and RMSE. To compute PLCC and RMSE, a non-linear fitting function with five parameters is employed to map the predicted scores to subjective scores. Specifically, it is defined as [109]:

$$f(x) = \tau_1\left(\frac{1}{2} - \frac{1}{1 + e^{\tau_2(x-\tau_3)}}\right) + \tau_4 x + \tau_5, \tag{3}$$

where τ_i, $i = 1, 2, \ldots, 5$, are the fitting parameters. Generally, a better metric produces higher SRCC, KRCC and PLCC values, as well as a lower RMSE value.

5.2. Performance evaluation on quality assessment of restored images

In this subsection, the performances of the traditional quality metrics and quality metrics for restored images are thoroughly investigated and compared in terms of the prediction monotonicity and prediction accuracy on nine public databases of three restoration scenarios. It is worth noting that the FR and RR metrics are not included and tested due to the unavailability of reference images in most restoration scenarios. These metrics are:

(1) Eleven general-purpose NR-IQA methods: Natural Image Quality Evaluator (NIQE) [59], Integrated Local NIQE (ILNIQE) [60], Quality-Aware Clustering (QAC) [58], Blind Image Quality Index (BIQI) [61], BLind Image Integrity Notator using DCT Statistic (BLIINDS2) [64], Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [65], COdebook Representation for No-reference Image quality Assessment (CORNIA) [68], Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) Index [63], Gradient Magnitude and Laplacian of Gaussian responses (GMLOG) [69], NR free energy based robust metric (NFERM) [67] and dubbed spatial–spectral entropy-based quality (SSEQ) [66].

(2) Three state-of-the-art multiply-distorted image quality metrics, including FIve-Step BLInd Metric (FISBLIM) [110], SIx-Step BLInd Metric (SISBLIM) [111] and gradient-weighted histogram of local binary pattern calculated on the gradient map (GLBP) [112].

(3) Seven quality metrics for specific restoration applications: two metrics for MDB, namely, MMD [18] and NRRB [91]; four metrics for IDH, namely, e [98], Σ [98], $\overline{\gamma}$ [98] and DHQI [88]; one metric for SRR, namely, SISR [83].

For fair comparison, all aforementioned training-based metrics are re-trained based on the corresponding database, and the parameters of the corresponding metrics are maintained the same without elaborate selection for different databases. Specifically, we randomly separate the images into two no overlap subsets: 80% for model training and the rest 20% for testing. Moreover, this separation guarantees that the restored images generated from the same degraded image are either in the training set or in the test set. To avoid bias, we repeat the training–testing process 1000 times. Then, the median values of performances are reported. The MATLAB source codes of all the competing quality metrics are obtained from original authors or websites. The experimental results are listed in Table 2. In the table, the type "GNR" represents the general-purpose quality metric, "MIM" represents multiply-distorted image quality metric, "MMDB" represents the metric for MDB, "MIDH" represents the metric for IDH, and "MSRR" represents the metric for SRR.

From these results, we conclude that: (1) The traditional quality metrics are limited in predicting the quality of restored images, which is mainly due to the following reasons. First, the traditional quality metrics are usually designed and tested on the common distortion categories, such as JPEG2000/JPEG compression, additive white Gaussian noise, Gaussian blur, and fast fading Rayleigh, while the types of distortions of restored images are quite different from these common distortion categories. Second, the distortions of restored images are more complex and real, since they are usually localized and disparate and characterized by multiple distortions, i.e., a mixture of residual distortions and side effects. Third, the interaction of the residual distortions and side effects is vital to the quality evaluation of restored images, and it is naturally not considered and measured in the traditional quality metrics; (2) The metrics designed for specific restoration applications usually outperform the traditional quality metrics, mainly due to their specific design; (3) The robustness of these metrics for restored images remains an open problem. For example, NRRB [91] achieves the best performances on MDD13 database, whose SRCC and PLCC are 0.870 and 0.882, respectively. However, it does not perform well on MDUID and MDRID databases. Similarly, DHQI [88] does not perform well on DHID database and it is even inferior to some traditional quality metrics, while the performance of DHQI is much better than that of all compared metrics on DHQD database. The reasons are as follows. First, these metrics are usually designed on an individual database, and thus the features may not be robust across different databases. Second, even in the same restoration scenario, there are differences between synthetic and real distortions. Generally speaking, the database with real distortion is more challenging and to be conquered; (4) There is no a single quality metric that performs well in all restoration scenarios.

5.3. Performance evaluation on benchmarking image restoration algorithms

In this subsection, the performances of the traditional quality metrics and quality metrics for restored images on benchmarking image restoration algorithms are investigated and compared on the aforementioned databases. For this task, two criteria, SRCC and KRCC, are used and computed based on a group of images, which are from the same degraded image but processed via different restoration algorithms. Then, the average results of these SRCC/KRCC values are computed and reported as the final performance. The SRCC and KRCC values range from −1 to 1, where "1" indicates that the predicted scores are fully consistent with the subjective ratings and "−1" the opposite. In addition to the above methods, the recently proposed pairwise-comparison-based metric, Ref. [22], is also included for comparison. The experimental results are listed in Tables 3–5, where "GIR" represents the general image restoration quality metric.

From these results, we conclude that: (1) Both the traditional quality metrics and quality metrics for restored images are quite limited for benchmarking image restoration algorithms. The reasons are as follows. First, it is easy to get wrong ranking of a set of restored images, since they are quite limited in evaluating the quality of restored images. Second, these metrics are usually trained/operated with individual images to assign an absolute score. Thus, the relative quality ranking, which is the most heavily considered factor for this task [22], is typically ignored in these metrics. As a result, they are weak at ordering a set of images, result in limited performances; (2) The Ref. [22], which focuses on the relative quality ranking of a set of images, performs consistently well on all databases. Further investigation demonstrates that although the performance of Ref. [22] is much better than that of the other metrics, there is still much room for improvement. For example, the highest SRCC value produced by the Ref. [22] is only 0.682 for IDH, which is far from the ideal. The main reason may

Table 2
Performances of the traditional quality metrics and quality metrics for restored images on nine public databases of three restoration scenarios.

Metric	Type	Motion deblurring											
		MDD13 [18]				MDUID [80]				MDRID [80]			
		SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
NIQE [59]	GNR	0.361	0.236	0.377	2.771	0.200	0.132	0.050	4.968	0.027	0.018	0.036	8.146
ILNIQE [60]	GNR	0.650	0.456	0.647	2.282	0.294	0.200	0.150	4.918	0.109	0.072	0.053	8.139
QAC [58]	GNR	0.094	0.064	0.132	2.966	0.097	0.064	0.099	4.950	0.057	0.037	0.086	8.120
BIQI [61]	GNR	0.760	0.569	0.781	1.858	0.434	0.294	0.386	2.288	0.137	0.091	0.080	8.911
BLIINDS2 [64]	GNR	0.793	0.608	0.826	1.672	0.373	0.256	0.312	2.505	0.203	0.136	0.159	7.824
BRISQUE [65]	GNR	0.802	0.622	0.837	1.631	0.322	0.220	0.243	2.607	0.139	0.094	0.106	6.511
CORNIA [68]	GNR	0.821	0.627	0.855	1.549	0.414	0.279	0.347	2.540	0.093	0.061	0.178	7.665
DIIVINE [63]	GNR	0.725	0.541	0.757	1.930	0.283	0.197	0.262	2.587	0.074	0.049	0.113	8.198
GMLOG [69]	GNR	0.735	0.552	0.767	1.920	0.464	0.316	0.297	2.592	0.185	0.127	0.192	6.507
NFERM [67]	GNR	0.801	0.614	0.836	1.630	0.539	0.378	0.456	2.362	0.226	0.155	0.155	8.812
SSEQ [66]	GNR	0.614	0.444	0.728	2.021	0.397	0.269	0.311	2.447	0.131	0.086	0.130	8.895
FISBLIM [110]	MIM	0.608	0.425	0.635	2.313	0.039	0.025	0.033	4.972	0.040	0.027	0.040	8.144
SISBLIM [111]	MIM	0.588	0.407	0.606	2.379	0.008	0.005	0.047	4.969	0.024	0.015	0.018	8.149
GLBP [112]	MIM	0.651	0.471	0.714	2.059	0.371	0.257	0.244	2.858	0.143	0.094	0.192	6.482
MMD [18]	MMDB	0.805	0.608	0.854	1.558	0.434	0.292	0.336	2.685	0.084	0.056	0.202	6.340
NRRB [91]	MMDB	0.870	0.682	0.882	1.408	0.053	0.043	0.106	4.812	0.048	0.040	0.146	8.853
Metric	Type	Image Dehazing											
		DHID [86]				SHRQR [87]				DHQD [88]			
		SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
NIQE [59]	GNR	0.255	0.170	0.269	1.425	0.340	0.238	0.545	11.61	0.431	0.302	0.509	11.31
ILNIQE [60]	GNR	0.249	0.174	0.311	1.406	0.312	0.217	0.556	11.51	0.597	0.427	0.655	9.929
QAC [58]	GNR	0.136	0.093	0.228	1.440	0.405	0.285	0.654	10.48	0.451	0.321	0.546	11.02
BIQI [61]	GNR	0.264	0.175	0.335	1.386	0.365	0.258	0.519	11.54	0.580	0.421	0.619	10.24
BLIINDS2 [64]	GNR	0.324	0.219	0.354	1.304	0.349	0.245	0.567	11.00	0.649	0.475	0.691	9.295
BRISQUE [65]	GNR	0.373	0.258	0.407	1.296	0.312	0.214	0.532	11.43	0.668	0.492	0.699	9.405
CORNIA [68]	GNR	0.415	0.298	0.481	1.229	0.368	0.261	0.573	11.28	0.693	0.506	0.721	8.921
DIIVINE [63]	GNR	0.349	0.241	0.409	1.309	0.311	0.216	0.485	11.77	0.614	0.442	0.641	9.964
GMLOG [69]	GNR	0.297	0.207	0.356	1.304	0.335	0.232	0.544	11.49	0.694	0.515	0.729	8.999
NFERM [67]	GNR	0.339	0.237	0.379	1.309	0.398	0.279	0.619	10.76	0.699	0.519	0.731	8.887
SSEQ [66]	GNR	0.383	0.269	0.468	1.276	0.394	0.279	0.643	10.38	0.740	0.556	0.768	8.361
FISBLIM [110]	MIM	0.194	0.127	0.215	1.444	0.191	0.132	0.349	12.98	0.201	0.138	0.094	13.08
SISBLIM [111]	MIM	0.095	0.066	0.108	1.470	0.144	0.097	0.292	13.25	0.158	0.102	0.247	12.73
GLBP [112]	MIM	0.305	0.207	0.340	1.362	0.254	0.177	0.379	12.76	0.607	0.433	0.631	9.951
e [98]	MIDH	0.287	0.196	0.328	1.397	0.371	0.254	0.236	13.46	0.450	0.313	0.423	11.90
Σ [98]	MIDH	0.233	0.162	0.239	1.436	0.082	0.048	0.410	12.64	0.146	0.094	0.526	11.17
$\overline{\gamma}$ [98]	MIDH	0.210	0.142	0.263	1.427	0.107	0.072	0.125	13.74	0.414	0.269	0.507	11.32
DHQI [88]	MIDH	0.351	0.246	0.429	1.312	0.558	0.401	0.719	9.554	0.858	0.676	0.869	6.502
Metric	Type	SR image reconstruction											
		SRIDM [83]				SRID [84]				QADS [85]			
		SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
NIQE [59]	GNR	0.639	0.467	0.656	1.815	0.461	0.330	0.464	1.433	0.402	0.282	0.407	0.251
ILNIQE [60]	GNR	0.655	0.476	0.604	1.916	0.426	0.282	0.418	1.469	0.716	0.531	0.721	0.190
QAC [58]	GNR	0.410	0.279	0.416	2.187	0.369	0.249	0.366	1.505	0.348	0.243	0.364	0.256
BIQI [61]	GNR	0.706	0.519	0.702	1.696	0.609	0.454	0.675	1.174	0.688	0.501	0.678	0.202
BLIINDS2 [64]	GNR	0.801	0.608	0.827	1.355	0.682	0.505	0.723	1.098	0.642	0.471	0.647	0.209
BRISQUE [65]	GNR	0.884	0.702	0.892	1.085	0.807	0.634	0.838	0.889	0.840	0.656	0.839	0.149
CORNIA [68]	GNR	0.915	0.747	0.926	0.884	0.862	0.685	0.881	0.765	0.949	0.804	0.951	0.085
DIIVINE [63]	GNR	0.770	0.579	0.789	1.478	0.725	0.553	0.773	1.027	0.804	0.610	0.807	0.162
GMLOG [69]	GNR	0.791	0.597	0.806	1.411	0.760	0.573	0.791	0.990	0.821	0.627	0.825	0.155
NFERM [67]	GNR	0.872	0.685	0.883	1.107	0.812	0.627	0.835	0.887	0.890	0.714	0.893	0.124
SSEQ [66]	GNR	0.815	0.614	0.827	1.348	0.706	0.535	0.746	1.082	0.550	0.395	0.566	0.227
FISBLIM [110]	MIM	0.756	0.568	0.784	1.491	–	–	–	–	0.757	0.559	0.742	0.184
SISBLIM [111]	MIM	0.755	0.568	0.741	1.614	–	–	–	–	0.737	0.531	0.703	0.195
GLBP [112]	MIM	0.808	0.608	0.809	1.421	0.705	0.534	0.766	1.013	0.670	0.485	0.673	0.203
SISR [83]	MSRR	0.924	0.759	0.933	0.873	0.889	0.720	0.912	0.663	0.886	0.706	0.887	0.127

be that the features used in the Ref. [22], i.e. spatial-domain LBP features and frequency-domain log-Gabor features, are not sensitive to the distortions of dehazed images, such as, residual haze, halo effects and over-enhancement, etc.

5.4. Time complexity

Time complexity is another important aspect for evaluating the performance of a quality metric, and low-complexity quality metrics are quite desired in real-time applications. To this end, in addition to comparing the performances evaluation on quality assessment for image restoration and benchmarking image restoration algorithms, we also compute the computational time of each quality metric. Specifically, for each restoration scenario, one database is used to conduct the experiment, and the average execution time (second) for an image is reported. Experiments are performed on a computer with Intel Core i7-7700K CPU@4.20 GHz and 32.0G RAM. The software platform is MATLAB R2016b under Windows 10. The experimental results are listed in Table 6.

Table 3
Performances of the traditional quality metrics and quality metrics for MDB on benchmarking motion deblurring algorithms.

Metric	Type	Motion deblurring					
		MDD13 [18]		MDUID [80]		MDRID [80]	
		SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
NIQE [59]	GNR	−0.152	−0.110	−0.209	−0.167	−0.162	−0.120
ILNIQE [60]	GNR	−0.308	−0.256	−0.372	−0.282	−0.293	−0.219
QAC [58]	GNR	−0.126	−0.081	−0.158	−0.124	0.005	−0.009
BIQI [61]	GNR	0.671	0.579	0.479	0.364	0.284	0.218
BLIINDS2 [64]	GNR	0.625	0.537	0.486	0.368	0.318	0.245
BRISQUE [65]	GNR	0.660	0.561	0.447	0.346	0.251	0.192
CORNIA [68]	GNR	0.733	0.637	0.479	0.352	0.159	0.116
DIIVINE [63]	GNR	0.579	0.491	0.417	0.309	0.234	0.176
GMLOG [69]	GNR	0.587	0.501	0.541	0.423	0.293	0.219
NFERM [67]	GNR	0.666	0.575	0.585	0.459	0.366	0.281
SSEQ [66]	GNR	0.546	0.460	0.500	0.376	0.244	0.187
FISBLIM [110]	MIM	−0.256	−0.212	−0.164	−0.118	−0.132	−0.101
SISBLIM [111]	MIM	−0.248	−0.209	−0.227	−0.172	−0.072	−0.060
GLBP [112]	MIM	0.689	0.593	0.450	0.344	0.262	0.197
MMD [18]	MMDB	0.796	0.705	0.599	0.479	0.212	0.156
NRRB [91]	MMDB	0.685	0.585	0.274	0.223	0.083	0.096
Ref. [22]	GIR	0.810	0.723	0.759	0.614	0.522	0.416

Table 4
Performances of the traditional quality metrics and quality metrics for IDH on benchmarking image dehazing algorithms.

Metric	Type	Image dehazing					
		DHID [86]		SHRQR [87]		DHQD [88]	
		SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
NIQE [59]	GNR	−0.146	−0.092	−0.401	−0.319	−0.278	−0.211
ILNIQE [60]	GNR	−0.034	−0.037	−0.443	−0.351	−0.277	−0.208
QAC [58]	GNR	−0.100	−0.074	0.240	0.173	0.385	0.292
BIQI [61]	GNR	0.038	0.046	0.335	0.262	0.249	0.195
BLIINDS2 [64]	GNR	0.068	0.049	0.292	0.216	0.382	0.296
BRISQUE [65]	GNR	0.092	0.076	0.297	0.225	0.551	0.451
CORNIA [68]	GNR	0.239	0.182	0.289	0.222	0.492	0.387
DIIVINE [63]	GNR	0.070	0.059	0.357	0.283	0.436	0.349
GMLOG [69]	GNR	0.046	0.041	0.372	0.285	0.529	0.427
NFERM [67]	GNR	0.182	0.160	0.490	0.394	0.492	0.395
SSEQ [66]	GNR	0.338	0.269	0.430	0.348	0.441	0.349
FISBLIM [110]	MIM	0.142	0.123	0.155	0.114	0.120	0.095
SISBLIM [111]	MIM	−0.065	−0.032	0.025	0.002	0.085	0.064
GLBP [112]	MIM	0.057	0.047	0.221	0.174	0.381	0.298
e [98]	MIDH	−0.209	−0.176	0.466	0.363	−0.063	−0.061
Σ [98]	MIDH	0.334	0.267	–	–	–	–
$\overline{\gamma}$ [98]	MIDH	0.039	0.025	0.098	0.076	−0.389	−0.279
DHQI [88]	MIDH	0.252	0.207	0.563	0.462	0.742	0.631
Ref. [22]	GIR	0.411	0.335	0.617	0.520	0.682	0.581

Table 5
Performances of the traditional quality metrics and quality metrics for SRR on benchmarking SR reconstruction algorithms.

Metric	Type	SR image reconstruction					
		SRIDM [83]		SRID [84]		QADS [85]	
		SRCC	KRCC	SRCC	KRCC	SRCC	KRCC
NIQE [59]	GNR	−0.334	−0.258	−0.156	−0.119	−0.091	−0.079
ILNIQE [60]	GNR	−0.087	−0.111	0.226	0.160	−0.620	−0.492
QAC [58]	GNR	0.323	0.235	−0.206	−0.169	0.179	0.137
BIQI [61]	GNR	0.321	0.245	0.539	0.423	0.630	0.474
BLIINDS2 [64]	GNR	0.423	0.331	0.448	0.357	0.493	0.394
BRISQUE [65]	GNR	0.583	0.462	0.703	0.588	0.531	0.415
CORNIA [68]	GNR	0.593	0.472	0.641	0.516	0.839	0.686
DIIVINE [63]	GNR	0.398	0.309	0.647	0.523	0.657	0.534
GMLOG [69]	GNR	0.443	0.343	0.564	0.462	0.734	0.577
NFERM [67]	GNR	0.531	0.419	0.605	0.494	0.613	0.468
SSEQ [66]	GNR	0.422	0.325	0.469	0.370	0.407	0.304
FISBLIM [110]	MIM	−0.254	−0.208	–	–	−0.574	−0.473
SISBLIM [111]	MIM	−0.261	−0.212	–	–	−0.578	−0.478
GLBP [112]	MIM	0.384	0.293	0.632	0.522	0.430	0.345
SISR [83]	MSRR	0.480	0.382	0.730	0.605	0.778	0.658
Ref. [22]	GIR	0.762	0.649	0.823	0.714	0.829	0.716

Table 6
The average execution time (second) for an image of objective quality metrics.

Metric	Type	MDRID [80]	DHQD [88]	SRIDM [83]
NIQE [59]	GNR	0.275	0.123	0.065
ILNIQE [60]	GNR	3.618	3.424	3.475
QAC [58]	GNR	0.349	0.166	0.077
BIQI [61]	GNR	0.778	0.817	0.714
BLIINDS2 [64]	GNR	46.13	20.81	11.65
BRISQUE [65]	GNR	0.253	0.182	0.151
CORNIA [68]	GNR	1.731	1.499	1.432
DIIVINE [63]	GNR	26.61	9.742	4.912
GMLOG [69]	GNR	0.109	0.048	0.025
NFERM [67]	GNR	74.15	33.83	17.86
SSEQ [66]	GNR	1.566	0.741	0.360
FISBLIM [110]	MIM	2.508	1.726	0.661
SISBLIM [111]	MIM	2.652	1.765	0.715
GLBP [112]	MIM	0.188	0.082	0.043
MMD [18]	MMDB	89.78	–	–
NRRB [91]	MMDB	8.212	–	–
e [98]	MIDH	–	6.132	–
DHQI [88]	MIDH	–	0.173	–
SISR [83]	MSRR	–	–	11.22

It is easily observed from Table 6 that GMLOG [69] runs the least computational time among the tested metrics in all databases, and QAC [58] and GLBP [112] run much faster than other metrics. NFERM [67] requires the largest computational time, except for MDRID [80] database, and BLIINDS2 [64] and DIIVINE [63] are much slower than the others. In MDRID [80] database, MMD [18] requires the largest computational time. It takes 89.78 s to predict a quality score of a motion deblurred image, which does not meet the time requirement in most image processing applications.

Joint analysis of time complexity and prediction consistency is very meaningful, and it would be useful to know which quality metric could achieve the two criteria in an optimal way. Also, it provides a reference for the design of the new quality metrics. To this end, a detailed joint analysis is provided here. For MDRID [80] database, although some quality metrics run fast, there is no a quality metric that performs well on motion deblurred images. For DHQD [88] database, DHQI [88] achieves the best performance of joint analysis. Specifically, it produces the highest prediction consistency and runs for a lower time. SSEQ [66] and GMLOG [69] can be considered as medium quality metrics due to their ordinary performance and moderate time. e [98] performs not well, because it takes a long time to run and only produces a lower prediction consistency. For SRIDM [83], although the consistency of SISR [83] is slightly higher than that of CORNIA [68], it runs about eight times longer than CORNIA [68]. In this regard, CORNIA [68] can be considered as the best metric. The consistency of CORNIA ranks second and it takes 1.432 s to predict a quality score of a SR reconstructed image. Except for CORNIA [68] and SISR [83], BRISQUE [65] performs better than other metrics. From these results, we know that both time complexity and prediction consistency are challenging problems for these quality metrics, and there is much room for improvement.

6. Challenges and future trends

In this section, we discuss several factors, challenges and future trends of subjective and objective quality assessment for image restoration.

6.1. Subjective quality assessment for image restoration

With the tremendous increase of research in image restoration algorithms, it was crucial to develop comprehensive databases for designing quality metrics for image restoration. Based on previous observations,

we would suggest considering *Data preparation*, *Subjective study* and *New scenario* for constructing comprehensive databases in future.

Data preparation. (1) Most existing image restoration databases collect limited original images, which makes it difficult to cover a wide variety of visual contents. This can cause serious trouble to quality models trained on these databases and also lead to unfair performance evaluation due to the dependence of the restoration algorithm on the image contents; hence, creating databases to collect images with diverse scenes is crucial and should be conquered. (2) The existing databases usually set limited distortion levels to original images, while discarding images that contain heavy or slight distortion. However, real-world images usually face with much more complicated situations and thus it is difficult to cover a wide range of distortion space. More distortion levels should be included for constructing a comprehensive database. (3) Most existing databases are almost about traditional image restoration algorithms excluding deep learning based algorithms. This will lead to inevitable bias to the images that generated by deep learning based algorithms. So, to establish a comprehensive database, both traditional algorithms and deep learning based methods should be introduced. (4) Most existing databases are built based on the synthetic degraded images that generated by injecting noise, motion blur or haze, etc, into the original images. However, the characteristics distortions between synthetic degraded images and real degraded images are quite different. Besides, the reference images are usually not available in most image restoration scenarios. Thus, it is more challenging and meaningful to deal with distortions in the real world, and real-distortion image restoration databases are highly needed.

Subjective study. (1) In most existing image restoration databases, absolute rating method was used to evaluate images in subjective experiment and then subjective scores were obtained after data processing. However, preference judgments rather than an absolute scale measurements is more in keeping with the visual characteristic of Humans when several images are presented at the same time [113,114]. Also, the pairwise comparison has a higher discriminatory power, especially when perceptual visual differences between images are relatively small [115]. So, pairwise comparison is more reliable and should be preferred used for constructing image restoration databases in future. (2) With the tremendous amount of images in subjective experiment, it is essential to propose new strategies that can collect the label efficiently to mitigate the typical problems of subjective experiment, such as, labor intensive and time-consuming. However, little work has been done to address this factor, which may hinder further development of the comprehensive databases.

New scenario. Although extensive image restoration algorithms have been proposed in various fields, the existing quality assessment databases only focus on several scenarios, as shown in Table 1. More restoration scenarios are largely under-explored and should be conquered. For example, image denoising, as one of the most common image restoration tasks, has been widely studied. However, there is no public comprehensive databases for performance evaluation. Another scenario is image deraining [116], which is important to outdoor vision systems, such as autonomous navigation systems and surveillance cameras. Although many image deraining methods have been proposed in the past decades, little work has been done to the perceptual evaluation of image deraining algorithms and derained images.

6.2. Objective quality assessment for image restoration

Although some objective quality metrics have been done to the perceptual evaluation of image restoration algorithms and restored images, there is still considerable room for improvement. Here, we would suggest considering *Feature extraction*, *Model design*, *Deep learning*, *Generalization ability* and *Model application* for developing more reliable and robust metrics in future.

Feature extraction. (1) In most existing metrics, the distortion-specific features are extracted and used to portray the individual distortion of restored images. Then, all features are combined to train a regression model for quality prediction. However, the restored images are usually characterized by multiple distortions, as shown in Introduction section. Therefore, measuring specific distortions is not sufficient for evaluating the quality of restored images, and the interaction of individual distortions, which plays an important role in visual perception but is significantly under-explored, is needed when designing a high-performance algorithm. (2) It has been demonstrated that the HVS exhibits the hierarchical characteristic, in which the input visual signal is hierarchically processed from low to high level features. However, the features used in the existing metrics are almost low-level, and thus they work effectively for weak distortion but are blind to the semantics information. The high-level features, which play an important role in visual perception and recognition, may be effective but are somewhat under-explored. Therefore, the quality metrics that take into account both low-level features and high-level features should be modestly investigated.

Model design. (1) Although NR IQA metrics have achieved notable progress, much room for improvement still remains due to limited information from restored images. To this end, we can resort to RR IQA metrics, which can obtain more information from both restored and degraded images than a single restored image. For example, by making full use of the available degraded images as references, side effects introduced by restoration process can be detected relatively easily. (2) The current metrics are usually designed for specific restoration scenarios. One obvious disadvantage is that the scope of application is largely limited. So, general image restoration quality metrics are highly desired, and it can be a future trend.

Deep learning. Deep neural network has recently achieved great successes on various computer vision tasks, and many deep neural network based metrics have been proposed for estimating the quality of images with common distortions. However, little work has been done towards image restoration. The existing well-designed network architectures provide insights for developing novel and effective quality metrics of image restoration in future. However, the limited availability of human labeled training data is a distinct challenge. As shown in Section 3, the existing databases are insufficient to train a deep neural network. Obviously, common data augmentation techniques, such as rotation, cropping and horizontal reflection, cannot be used in this task, due to these techniques would change perceptual quality of images significantly. Thus, how to solve the problem of training data is important for developing deep neural network based metrics.

One possible solution proposes new data augmented approaches to alleviate the problem. Another possible solution is that we can resort to the transfer learning, which extracts useful information from a large-scale dataset in the related domains and then adapts them for being used in target tasks [117]. Last but not least, few-shot learning [118] is a feasible way to address limited annotated data in quality assessment for image restoration.

Generalization ability. Generalization ability is an important issue for learning-based quality metrics. However, little effort has been spent to address it so far. Most existing learning-based quality metrics have limited generalization ability. The main reasons are as follows. First, the extracted features are designed and tested on one database, and they may not be robust to the complex distortions in different databases. Second, the regression tools are not robust and have limited performances. Up to now, to design more advanced image restoration quality metrics that have good generalization ability is still an open problem. More efforts should be put into this trend in future.

Model application. (1) Although some efforts have been done to parameter selection, as discuss in Section 4, almost all the works focus on image denoising. Parameter selection for other restoration scenarios is less investigated and significantly under-explored. (2) The existing metrics are quite limited for benchmarking image restoration algorithms, and it is less investigated so far. Based on previous discussions, the learning to rank technology [119,120], which focus on the orders of lists of items, applies to this task. (3) Perceptual image

restoration, which is hopefully perceptually close to the latent clear images during the model/network training, is another important but under-explored application. Traditionally, MSE based loss function was usually adopt to train the deep network for image restoration [121]. However, the loss function encourages finding pixel-wise averages with smooth results and does not focus on perceptual image quality. Thus, even with a very low loss value, the restored results may do not correlate well with human perception. Some recent efforts have been done to address the perceptual image restoration [122,123], and more efforts are encouraged to further boost this trend.

7. Conclusions

Although notable success has been made in developing robust image restoration algorithms recently, little efforts have been dedicated to the quality assessment for image restoration algorithms and restored images. It is substantially different from traditional IQA in nature due to the not commonly observed types of distortions in other applications. To help researchers in developing more effective quality metrics for image restoration, we present this review. In this paper, we have provided a comprehensive description of existing subjective quality assessment databases of image restoration, the current research efforts carried in the development of objective quality metrics and two applications of image restoration quality metrics, namely, parameter selection and benchmarking image restoration algorithms. Then, three performance comparisons of traditional quality metrics and metrics for image restoration in terms of quality assessment of restored images, benchmarking image restoration algorithms and time complexity were provided. The experimental results have demonstrated that there is still considerable room for developing more robust metrics. Finally, based on the review, we have highlighted the challenges and future trends.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Outstanding Innovation Scholarship for Doctoral Candidate of “Double First Rate” Construction Disciplines of CUMT.

References

[1] J. Liu, R.J. Liu, Y.H. Wang, J.L. Chen, Y.J. Yang, D.L. Ma, Image denoising searching similar blocks along edge directions, Signal Process., Image Commun. 57 (2017) 33–45.
[2] N. Islam, Z. Shahid, W. Puech, Denoising and error correction in noisy AES-encrypted images using statistical measures, Signal Process., Image Commun. 41 (2016) 15–27.
[3] S.H. Gu, L. Zhang, W.M. Zuo, X.C. Feng, Weighted nuclear norm minimization with application to image denoising, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit., CVPR, Columbus, OH, USA, 2014, pp. 2862–2869.
[4] S. Tang, X.Z. Xie, M. Xia, L. Luo, P.S. Liu, Z.X. Li, Spatial-scale-regularized blur kernel estimation for blind image deblurring, Signal Process., Image Commun. 68 (2018) 138–154.
[5] J.S. Pan, Z.C. Lin, Z.X. Su, M.-H. Yang, Robust kernel estimation with outliers handling for image deblurring, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit., CVPR, Las Vegas, NV, USA, 2016, pp. 2800–2808.
[6] J.X. Dong, J.S. Pan, Z.X. Su, M.-H. Yang, Blind image deblurring with outlier handling, in: Proc. Conf. on Computer Vision, ICCV, Venice, Italy, 2017, pp. 2497–2505.
[7] J.S. Pan, D.Q. Sun, H. Pfister, M.-H. Yang, Blind image deblurring using dark channel prior, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit., CVPR, Las Vegas, NV, USA, 2016, pp. 1628–1636.
[8] A. Laghrib, A. Ben-loghfyry, A. Hadri, A. Hakim, A nonconvex fractional order variational model for multi-frame image super-resolution, Signal Process., Image Commun. 67 (2018) 1–11.
[9] C.Z. Zou, Y.S. Xia, Bayesian Dictionary learning for hyperspectral image super resolution in mixed Poisson-Gaussian noise, Signal Process., Image Commun. 60 (2018) 29–41.
[10] A. Laghrib, A. Hakim, S. Raghay, An iterative image super-resolution approach based on Bregman distance, Signal Process., Image Commun. 58 (2017) 24–34.
[11] A.N. Wang, W.H. Wang, J.L. Liu, N.H. Gu, Aipnet: image-to-image single image dehazing with atmospheric illumination prior, IEEE Trans. Image Process. 28 (1) (2019) 381–393.
[12] Q. Liu, X.B. Gao, L.H. He, W. Lu, Single image dehazing with depth-aware non-local total variation regularization, IEEE Trans. Image Process. 27 (10) (2018) 5178–5191.
[13] N. Baig, M.M. Riaz, A. Ghafoor, A.M. Siddiqui, Image dehazing using quadtree decomposition and entropy-based contextual regularization, IEEE Signal Process. Lett. 23 (6) (2016) 853–857.
[14] W.S. Dong, G.M. Shi, X. Li, Y. Ma, F. Huang, Compressive sensing via nonlocal low-rank regularization, IEEE Trans. Image Process. 23 (8) (2014) 3618–3632.
[15] X.L. Wu, W.S. Dong, X.J. Zhang, G.M. Shi, Model-assisted adaptive recovery of compressed sensing with imaging applications, IEEE Trans. Image Process. 21 (2) (2012) 451–458.
[16] W.S. Dong, G.M. Shi, X.L. Wu, L. Zhang, A learning-based method for compressive image recovery, J. Vis. Commun. Image Represent. 24 (7) (2013) 1055–1063.
[17] W.S. Dong, L. Zhang, G.M. Shi, X. Li, Nonlocally centralized sparse representation for image restoration, IEEE Trans. Image Process. 22 (4) (2013) 1620–1630.
[18] Y.M. Liu, J. Wang, S. Cho, A no-reference metric for evaluating the quality of motion deblurring, ACM Trans. Graph. 32 (6) (2013) 1–12.
[19] Y.M. Fang, C. Zhang, W.H. Yang, I.Y. Liu, Z.M. Guo, Blind visual quality assessment for image super-resolution by convolutional neural network, Multimed. Tools Appl. 77 (10) (2018) 1–18.
[20] H.L. Liang, D.S. Weller, Comparison-based image quality assessment for selecting image restoration parameters, IEEE Trans. Image Process. 25 (11) (2016) 5118–5130.
[21] X.F. Kong, Q.X. Yang, No-reference image quality assessment for image auto-denoising, Int. J. Comput. Vis. 126 (5) (2018) 537–549.
[22] B. Hu, L.D. Li, H.T. Liu, W.S. Lin, J.S. Qian, Pairwise-comparison-based rank learning for benchmarking image restoration algorithms, IEEE Trans. Multimedia 21 (8) (2019) 2042–2056.
[23] Z.Y. Chen, T.T. Jiang, Y.H. Tian, Quality assessment for comparing image enhancement algorithms, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit., CVPR, Columbus, OH, USA, 2014, pp. 3003–3010.
[24] K. Gu, D.C. Tao, J.-F. Qiao, W.S. Lin, Learning a no-reference quality assessment model of enhanced images with big data, IEEE Trans. Neural Netw. Learn. Syst. 29 (4) (2018) 1301–1313.
[25] M.A. Qureshi, A. Beghdadi, M. Deriche, Towards the design of a consistent image contrast enhancement evaluation measure, Signal Process., Image Commun. 58 (2017) 212–227.
[26] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.
[27] M.A. Stricker, M. Orengo, IST/SPIE Symposium on Electronic Imaging: Science and Technology, International Society for Optics and Photonics, 1995, pp. 381–392,
[28] T. Joachims, Training linear svms in linear time, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, pp. 217–226.
[29] K. Gu, G.T. Zhai, X.J. Yang, W. Zhang, M. Liu, Subjective and objective quality assessment for images with contrast change, in: Proc. 20th IEEE Int. Conf. Image Process., ICIP, 2013, pp. 383–387.
[30] K. Gu, G.T. Zhai, W.S. Lin, M. Liu, The analysis of image contrast: From quality assessment to automatic enhancement, IEEE Trans. Cybern. 46 (1) (2016) 284–297.
[31] W.S. Lin, C.-C.J. Kuo, Perceptual visual quality metrics: A survey, J. Vis. Commun. Image Represent. 22 (4) (2011) 297–312.
[32] S. Winkler, Analysis of public image and video databases for quality assessment, IEEE J. Sel. Top. Signal Process. 6 (6) (2012) 616–625.
[33] A. Beghdadi, M.-C. Larabi, A. Bouzerdoum, K.M. Iftekharuddin, A survey of perceptual image processing methods, Signal Process., Image Commun. 28 (8) (2013) 811–831.
[34] M.A. Qureshi, M. Deriche, A. Beghdadi, A. Amin, A critical survey of state-of-the-art image inpainting quality assessment metrics, J. Vis. Commun. Image Represent. 49 (2017) 177–191.
[35] P. Ye, D. Doermann, Document image quality assessment: A brief survey, in: 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 2013, pp. 723–727.
[36] A. Lahoulou, M.C. Larabi, A. Beghdadi, E. Viennet, A. Bouridane, Knowledge-based taxonomic scheme for full-reference objective image quality measurement models, J. Imag. Sci. Technol. 60 (6) (2016) 60406-1-60406-15.
[37] M. Pedersen, J.Y. Hardeberg, Survey of full-reference image quality metrics, in: HøSkolen I Gjølks Rapportserie, Vol. 5, The Norwegian Color Research Laboratory (Gjøik University College, 2009, ISSN: 1890-520X.

[38] M. Pedersen, J.Y. Hardeberg, Full-reference image quality metrics: classification and evaluation, Found. Trends Comput. Graph. Vis. 7 (2012) 1–80.

[39] J. Kim, S. Lee, Deep learning of human visual sensitivity in image quality assessment framework, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit., CVPR, Honolulu, HI, USA, 2017, pp. 1969–1977.

[40] J. Kim, S. Lee, Fully deep blind image quality predictor, IEEE J. Sel. Top. Sign. Proces. 11 (1) (2017) 206–220.

[41] S. Bosse, D. Maniry, K.-R. Mller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Trans. Image Process. 27 (1) (2018) 206–219.

[42] Z. Wang, E.P. Simoncelli, Reduced-reference image quality assessment using a wavelet-domain natural image statistic model, in: Proceedings of the SPIE Conference, 2005, pp. 149–159.

[43] L. Ma, S. Li, F. Zhang, K.N. Ngan, Reduced-reference image quality assessment using reorganized DCT-based image representation, IEEE Trans. Multimedia 13 (4) (2011) 824–829.

[44] A. Rehman, Z. Wang, Reduced-reference image quality assessment by structural similarity estimation, IEEE Trans. Image Process. 21 (8) (2012) 3378–3389.

[45] Z. Wang, G.X. Wu, H.R. Sheikh, E.P. Simoncelli, E.-H. Yang, A.C. Bovik, Quality-aware images, IEEE Trans. Image Process. 15 (6) (2006) 1680–1689.

[46] K. Ramesh, B. Das, Image quality assessment based on multiscale geometric analysis, in: International Conference on Electronics Computer Technology, Kanyakumari, India, 2011, pp. 57–60.

[47] J.J. Wu, W.S. Lin, G.M. Shi, L.D. Li, Y.M. Fang, Orientation selectivity based visual pattern for reduced-reference image quality assessment, Inform. Sci. 351 (2016) 18–29.

[48] J.J. Wu, W.S. Lin, Y.M. Fang, L.D. Li, G.M. Shi, Visual structural degradation based reduced-reference image quality assessment, Signal Process., Image Commun. 47 (2016) 16–27.

[49] R. Soundararajan, A. Bovik, RRED Indices: Reduced reference entropic differencing for image quality assessment, IEEE Trans. Image Process. 21 (2) (2012) 517–526.

[50] J.J. Wu, W.S. Lin, G.M. Shi, A. Liu, Reduced-reference image quality assessment with visual information fidelity, IEEE Trans. Multimed. 15 (7) (2013) 1700–1705.

[51] K. Gu, G.T. Zhai, X.K. Yang, W.J. Zhang, A new reduced-reference image quality assessment using structural degradation model, in: International Symposium on Circuits and Systems, Beijing, China, 2013, pp. 1095–1098.

[52] J. Redi, P. Gastaldo, I. Heynderickx, R. Zunino, Color distribution information for the reduced-reference assessment of perceived image quality, IEEE Trans. Circuits Syst. Video Technol. 20 (12) (2010) 1757–1769.

[53] L.D. Li, H.C. Zhu, G.B. Yang, J.S. Qian, Referenceless measure of blocking artifacts by tchebichef kernel analysis, IEEE Signal Process. Lett. 21 (1) (2014) 122–125.

[54] L.D. Li, W.S. Lin, H.C. Zhu, Learning structural regularity for evaluating blocking artifacts in JPEG images, IEEE Signal Process. Lett. 21 (8) (2014) 918–922.

[55] L.D. Li, W.S. Lin, X.S. Wang, G.B. Yang, K. Bahrami, A.C. Kot, No-reference image blur assessment based on discrete orthogonal moments, IEEE Trans. Cybern. 46 (1) (2016) 39–50.

[56] Y.M. Fang, K.D. Ma, Z. Wang, W.S. Lin, Z.J. Fang, G.T. Zhai, No-reference quality assessment of contrast-distorted images based on natural scene statistics, IEEE Signal Process. Lett. 22 (7) (2015) 838–842.

[57] H.T. Liu, N. Klomp, I. Heynderickx, A no-reference metric for perceived ringing artifacts in images, IEEE Trans. Circuits Syst. Video Technol. 20 (4) (2010) 529–539.

[58] W.F. Xue, L. Zhang, X.Q. Mou, Learning without human scores for blind image quality assessment, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit., CVPR, Portland, OR, USA, 2013, pp. 995–1002.

[59] A. Mittal, R. Soundararajan, A.C. Bovik, Making a completely blind image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.

[60] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (2015) 2579–2591.

[61] A.K. Moorthy, A.C. Bovik, A two-step framework for constructing blind image quality indices, IEEE Signal Process. Lett. 17 (5) (2010) 513–516.

[62] Y. Zhang, D.M. Chandler, No-reference image quality assessment based on log derivative statistics of natural scenes, J. Electr. Imag. 22 (4) (2013) 1–11.

[63] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: from natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (12) (2011) 3350–3364.

[64] M.A. Saad, A.C. Bovik, Blind image quality assessment: a natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (8) (2012) 3339–3352.

[65] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.

[66] L.X. Liu, B. Liu, H. Huang, A.C. Bovik, No-reference image quality assessment based on spatial and spectral entropies, Sig. Process.: Image Commun. 29 (8) 856–863.

[67] K. Gu, G.T. Zhai, X.K. Yang, W.J. Zhang, Using free energy principle for blind image quality assessment, IEEE Trans. Multimedia 17 (1) (2015) 50–63.

[68] P. Ye, J. Kumar, L. Kang, D. Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit., CVPR, Providence, RI, USA, 2012, pp. 1098–1105.

[69] W.F. Xue, X.Q. Mou, L. Zhang, A.C. Bovik, X.C. Feng, Blind image quality assessment using joint statistics of gradient magnitude and laplacian features, IEEE Trans. Image Process. 23 (11) (2014) 4850–4862.

[70] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Columbus, OH, USA, 2014, pp. 1733–1740.

[71] S. Bosse, D. Maniry, T. Wiegand, W. Samek, A deep neural network for image quality assessment, in: Proc. IEEE Int. Conf. Image Process. Phoenix, AZ, USA, 2016, pp. 3773–3777.

[72] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-end blind image quality assessment using deep neural networks, IEEE Trans. Image Process. 27 (3) (2018) 1202–1213.

[73] D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, Y. Zhang, Blind predicting similar quality map for image quality assessment, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Salt Lake City, UT, USA, 2018, pp. 6373–6382.

[74] K.-Y. Lin, G. Wang, Hallucinated-IQA: No-reference image quality assessment via adversarial learning, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Salt Lake City, UT, USA, 2018, pp. 732–741.

[75] N. Ponomarenko, O. Ieremeiev, V. Lukin, et al., Color image database TID2013: Peculiarities and preliminary results, in: Proc. Eur. Workshop Vis. Inf. Process. Paris, France, 2013, pp. 106–111.

[76] K. Zeng, Z. Wang, Perceptual evaluation of image denoising algorithms, in: Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 2013, pp. 1351–1355.

[77] E. Fonseca, P. Fiadeiro, V. Hajihashemi, M. Bernardo, A. Pinheiro, M. Pereira, Perceptual evaluation of speckle noise reduction techniques for phase shifting holograms, in: Eleventh International Conference on Quality of Multimedia Experience (QoMEX), Berlin, Germany, 2019.

[78] L.L. Thurstone, A law of comparative judgment, Psychol. Rev. 34 (4) (1927) 273–286.

[79] R.A. Bradley, M.E. Terry, Rank analysis of incomplete block designs: I. the method of paired comparisons, Biometrika 39 (3/4) (1952) 324–345.

[80] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, M.-H. Yang, A comparative study for single image blind deblurring, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit. Las Vegas, NV, USA, 2016, pp. 1701–1709.

[81] L.D. Li, Y. Yan, Y.M. Fang, S.Q. Wang, L. Tang, J.S. Qian, Perceptual quality evaluation for image defocus deblurring, Signal Process., Image Commun. 48 (2016) 81–91.

[82] C.-Y. Yang, C. Ma, M.-H. Yang, Single image super-resolution: a benchmark, in: European Conference on Computer Vision, 2014.

[83] C. Ma, C.-Y. Yang, X.K. Yang, M.-H. Yang, Learning a no-reference quality metric for single-image super-resolution, Comput. Vis. Image Understand. 158 (2017) 1–16.

[84] G.C. Wang, L.D. Li, Q.H. Li, K. Gu, Z.L. Lu, J.S. Qian, Perceptual evaluation of single-image super-resolution reconstruction, in: IEEE International Conference on Image Processing, Beijing, China, 2017, pp. 3145–3149.

[85] F. Zhou, R.G. Yao, B.Z. Liu, G.P. Qiu, Visual quality assessment for super-resolved resolved images: Database and method, IEEE Trans. Image Process. 28 (7) (2019) 3528–3541.

[86] K.D. Ma, W.T. Liu, Z. Wang, Perceptual evaluation of single image dehazing algorithms, in: IEEE International Conference on Image Processing, Quebec City, QC, Canada, 2015, pp. 3600–3604.

[87] X.K. Min, G.T. Zhai, K. Gu, Y.C. Zhu, J.T. Zhou, G.D. Guo, X.K. Yang, X.P. Guan, W.J. Zhang, Quality evaluation of image dehazing methods using synthetic hazy images, IEEE Trans. Multimedia 21 (9) (2019) 2319–2333.

[88] X.K. Min, G.T. Zhai, K. Gu, X.K. Yang, X.P. Guan, Objective quality evaluation of dehazed images, IEEE Trans. Intell. Transp. Syst. 20 (8) (2019) 2879–2892.

[89] L.D. Li, Y. Zhou, W.S. Lin, J.J. Wu, X.F. Zhang, B.J. Chen, No-reference quality assessment of deblocked images, Neurocomputing 177 (2016) 572–584.

[90] B. Hu, L.D. Li, J.S. Qian, Y.M. Fang, Perceptual evaluation of compressive sensing image recovery, in: Eighth International Conference on Quality of Multimedia Experience, Lisbon, Portugal, 2016, pp. 1–6.

[91] B. Hu, L.D. Li, J.S. Qian, Perceptual quality evaluation for motion deblurring, IET Comput. Vis. 12 (6) (2018) 796–805.

[92] J. Anaya, A. Barbu, RENOIR-A dataset for real low-light image noise reduction, J. Vis. Commun. Image Represent. 51 (2018) 144–154.

[93] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[94] D.M. Chandler, S.S. Hemami, VSNR: A wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (2007) 2284–2298.

[95] Y.M. Fang, J.Y. Liu, Y.B. Zhang, W.S. Lin, Z.M. Guo, Reduced-reference quality assessment of image super-resolution by energy change and texture variation, J. Vis. Commun. Image Represent. 60 (2019) 140–148.

[96] L.D. Li, Y. Yan, Z.L. Lu, J.J. Wu, K. Gu, S.Q. Wang, No-reference quality assessment of deblurred images based on natural scene statistics, IEEE Access 5 (2017) 2163–2171.

[97] B. Bare, K. Li, B. Yan, B.L. Feng, C.F. Yao, A deep learning based no-reference image quality assessment model for single image super-resolution, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Calgary, AB, Canada, 2018, pp. 15–20.

[98] N. Hautiere, J.-P. Tarel, D. Aubert, E. Dumont, Blind contrast enhancement assessment by gradient ratioing at visible edges, J. Image Anal. Stereol. 27 (2) (2008) 87–95.

[99] C. Yim, A.C. Bovik, Quality assessment of deblocked images, IEEE Trans. Image Process. 20 (1) (2011) 88–98.

[100] W. Zhang, B. Hu, Z. Xu, L.D. Li, Color image quality assessment with quaternion moments, International Conference on Cloud Computing and Security, ICCCS, Nanjing, China, 2016, pp. 301–312.

[101] P. Joshi, S. Prakash, S. Rawat, Continuous wavelet transform-based no-reference quality assessment of deblocked images, Vis. Comput. 34 (12) (2018) 1739–1748.

[102] B. Hu, L.D. Li, J.S. Qian, Internal generative mechanism driven blind quality index for deblocked images, in: IEEE International Conference on Image Processing, Athens, Greece, 2018, pp. 2476–2480.

[103] B. Hu, L.D. Li, J.S. Qian, Internal generative mechanism driven blind quality index for deblocked images, Multimed Tools Appl. 78 (9) (2019) 12583–12605.

[104] B. Hu, L.D. Li, J.J. Wu, S.Q. Wang, L. Tang, J.S. Qian, No-reference quality assessment of compressive sensing image recovery, Signal Process., Image Commun. 58 (2017) 165–174.

[105] X. Zhu, P. Milanfar, Automatic parameter selection for denoising algorithms using a no-reference measure of image content, IEEE Trans. Image Process. 19 (12) (2010) 3116–3132.

[106] H. Liang, D.S. Weller, Regularization parameter trimming for iterative image reconstruction, in: Proc. IEEE Asilomar Conf. Signals, Syst. Comput. Pacific Grove, CA, USA, 2015, pp. 755–759.

[107] M.S.C. Almeida, M.A.T. Figueiredo, Parameter estimation for blind and non-blind deblurring using residual whiteness measures, IEEE Trans. Image Process. 22 (7) (2013) 2751–2763.

[108] H.Y. Liang, D.S. Weller, Denoising method selection by comparison-based image quality assessment, in: IEEE International Conference on Image Processing, ICIP, Phoenix, AZ, USA, 2016, pp. 3106–3110.

[109] L.D. Li, W.H. Xia, W.S. Lin, Y.M. Fang, S.Q. Wang, Robust image sharpness evaluation based on multi-scale spatial and spectral features, IEEE Trans. Multimedia 19 (5) (2017) 1030–1040.

[110] K. Gu, G.T. Zhai, M. Liu, X.K. Yang, W.J. Zhang, X.H. Sun, W.H. Chen, Y. Zuo, FISBLIM: A five-step blind metric for quality assessment of multiply distorted images, in: Proc. of IEEE Workshop on Signal Processing Systems, SiPS 2013, Taipei City, Taiwan, 2013, pp. 241–246.

[111] K. Gu, G.T. Zhai, X.K. Yang, W.J. Zhang, Hybrid no-reference quality metric for singly and multiply distorted images, IEEE Trans. Broadcast. 60 (3) (2014) 555–567.

[112] Q.H. Li, W.S. Lin, Y.M. Fang, No-reference quality assessment for multiply-distorted images in gradient domain, IEEE Signal Process. Lett. 23 (4) (2016) 541–545.

[113] L. Ma, L. Xu, Y.C. Zhang, Y.H. Yan, K.N. Ngan, No-reference retargeted image quality assessment based on pairwise rank learning, IEEE Trans. Multimedia 18 (11) (2016) 31–48.

[114] L. Xu, J. Li, W.S. Lin, Y.B. Zhang, L. Ma, Y.M. Fang, Y.H. Yan, Multi-task rank learning for image quality assessment, IEEE Trans. Circuits Syst. Video Technol. 27 (9) (2017) 1833–1843.

[115] P. Hanhart, L. Krasula, P. Le Callet, T. Ebrahimi, How to benchmark objective quality metrics from paired comparison data? in: IEEE International Conference on Quality of Multimedia Experience, Lisbon, Portugal, 2016, pp. 1–6.

[116] W.H. Yang, J.Y. Liu, S. Yang, Z.M. Guo, Scale-free aingle image deraining via visibility enhanced recurrent wavelet learning, IEEE Trans. Image Process. 28 (6) (2019) 2948–2961.

[117] L. Shao, F. Zhu, X.L. Li, Transfer learning for visual Categorization: A survey, IEEE Trans. Neural Netw. Learn. Syst. 26 (5) (2015) 1019–1034.

[118] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H.S. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, Salt Lake City, UT, USA, 2018, pp. 1199–1208.

[119] W. Chen, T.-Y. Liu, Y. Lan, Z.-M. Ma, H. Li, Ranking measures and loss functions in learning to rank, in: Advances in Neural Information Processing Systems, 2009, pp. 315–323.

[120] T.-Y. Liu, Learning to rank for information retrieval, Found. Trends Inf. Retr. 3 (3) (2009) 225–331.

[121] W.S. Dong, P.Y. Wang, W.T. Yin, G.M. Shi, Denoising prior driven deep neural network for image restoration, IEEE Trans. Pattern Anal. Mach. Intell. 41 (10) (2019) 2305–2318.

[122] J. Johnson, A. Alahi, F.-F. Li, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, 2016, pp. 694–711.

[123] C. Ledig, L. Theis, F. Huszar, J. Caballero, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proc. Int. Conf. Comput. Vis. Pattern Recognit. Honolulu, HI, USA, 2017, pp. 105–114.