This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3102401, IEEE Transactions on Multimedia

Learning-Based Quality Assessment for Image Super-Resolution

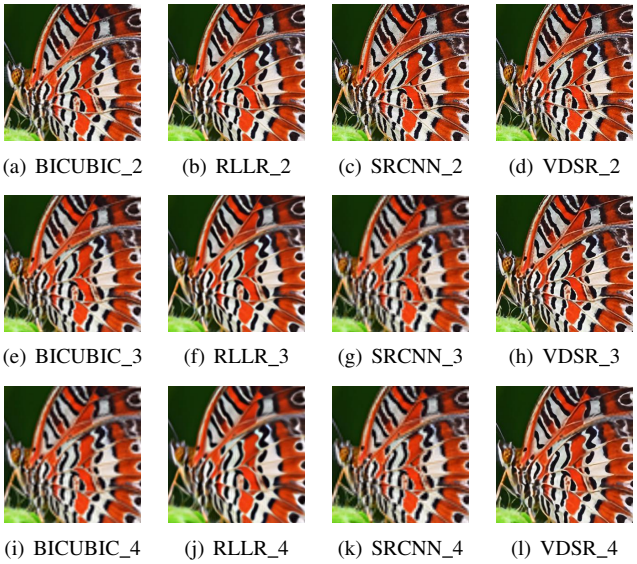
Tiesong Zhao, *Senior Member, IEEE,* Yuting Lin, Yiwen Xu, *Member, IEEE,* Weiling Chen, *Member, IEEE* and Zhou Wang, *Fellow, IEEE*

***Abstract*—Image Super-Resolution (SR) techniques improve visual quality by enhancing the spatial resolution of images. Quality evaluation metrics play a critical role in comparing and optimizing SR algorithms, but current metrics achieve only limited success, largely due to the lack of large-scale quality databases, which are essential for learning accurate and robust SR quality metrics. In this work, we first build a large-scale SR image database using a novel semi-automatic labeling approach, which allows us to label a large number of images with manageable human workload. The resulting SR Image quality database with Semi-Automatic Ratings (SISAR), so far the largest of SR-IQA database, contains 12,600 images of 100 natural scenes. We train an end-to-end Deep Image SR Quality (DISQ) model by employing two-stream Deep Neural Networks (DNNs) for feature extraction, followed by a feature fusion network for quality prediction. Experimental results demonstrate that the proposed method outperforms state-of-the-art metrics and achieves promising generalization performance in cross-database tests. The SISAR database and DISQ model will be made publicly available to facilitate reproducible research.**

***Index Terms*—Image Quality Assessment, Image Super-Resolution, Reduced-Reference.**

I. Introduction

With the rapid development of high-definition displays, the demand for high resolution image/video content has been increasing rapidly. To improve the user-end visual experience, image Super-Resolution (SR) technique is developed to interpolate High-Resolution (HR) images from their Low-Resolution (LR) references. Examples of these interpolated HR images can be seen in Fig. 1. The past two decades have witnessed a booming of image SR algorithms with widespread applications including medical image processing, video surveillance, remote sensing, and face recognition, among many others. In the SR process, an Image Quality Assessment (IQA) metric is critical as both a performance indicator and a guidance for further improvement. However, the mainstream IQA methods, such as Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) index [1], do not have high correlation with subjective opinions in image SR [2].



(a) BICUBIC_2 (b) RLLR_2 (c) SRCNN_2 (d) VDSR_2

(e) BICUBIC_3 (f) RLLR_3 (g) SRCNN_3 (h) VDSR_3

(i) BICUBIC_4 (j) RLLR_4 (k) SRCNN_4 (l) VDSR_4

Fig. 1. Examples of HR images generated by SR. (a-d), (e-h) and (i-l) are reconstructed HR images via BICUBIC [3], RLLR [4], SRCNN [5] and VDSR [6] algorithms with scaling factors of 2, 3 and 4 respectively. Quality variations are observed across both scaling factor and SR methods.

Existing IQA methods can be divided into two categories: subjective tests by human observers and objective models by automatic algorithms. Between them, the more reliable way is the subjective test, since human users are the ultimate viewers of images. Several SR methods assess their performances using small-scale subjective tests [7], [8]. However subjective test is time-consuming and hard to be embedded into practical applications. Instead, it is often utilized to construct databases, which serve as standard testing sets to compare objective models.

Numerous objective IQA models have also been proposed. According to the availability of reference, objective models can be classified into full-reference, reduced-reference and no-reference IQA models. In a typical image SR problem, the perfect quality HR image is unavailable as a reference, therefore, full-reference models are generally inapplicable. In addition, SR algorithms often introduce mixed impairments, including blurring, ringing and aliasing artifacts, which are not well measured by the existing methods [9], [10]. Therefore, conventional general-purpose reduced-reference and no-reference IQA models should be redesigned specifically for image SR problem.

This work was supported by the National Natural Science Foundation of China (61901119). (*Corresponding author: Weiling Chen.*)

T. Zhao is with the Fujian Key Lab for Intelligent Processing and Wireless Transmission of Media Information, Fuzhou University, Fuzhou 350116, China. He is also with Peng Cheng Laboratory, Shenzhen 518055, China. (e-mail: t.zhao@fzu.edu.cn).

Y. Lin, Y. Xu, W. Chen are with the Fujian Key Lab for Intelligent Processing and Wireless Transmission of Media Information, Fuzhou University, Fuzhou 350116, China (E-mail: {N181120063, xu_yiwen, weiling.chen}@fzu.edu.cn).

Z. Wang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada (E-mail: Z.Wang@ece.uwaterloo.ca).

The key challenge in SR-IQA is how to effectively learn distinguishing feature representations of various LR and HR images and then map the features to image quality prediction. Convolutional Neural Network (CNN) has shown its advantage in IQA with promising successes in recent years [11]–[13]. Compared with other data-driven models, CNN is capable of learning features and making quality predictors jointly in an end-to-end manner. Despite its superiority, CNN has not yet been well exploited in SR-IQA, for which most methods are not based on deep learning [9], [10], [14]–[16]. In this work, we propose a CNN-based SR-IQA algorithm. Considering the importance of reference information [17], we employ CNNs to extract the features from both images before and after the SR process. A two-stream CNN architecture is thereby designed to simultaneously take the test HR and the corresponding LR reference images, followed by feature fusion and quality prediction of the HR image.

The performance of deep learning based models relies heavily on the quality and quantity of training images. Although several benchmark databases of image SR qualities have been constructed [9], [15], [18]–[20], they are limited in their sample sizes. The largest database contains only 1,620 HR images. A larger image SR database is thus imperative, for which the biggest challenge is in the enormous workload of human labeling. In this work, we observe a negatively exponential decay behavior of subjective scores after iterative downsample-and-SR processing of natural images. This helps us develop a large-scale database with reduced human labeling workload. Experiments on randomly selected samples demonstrate the high accuracy of this database. Based on this database, we are able to develop and train the two-stream deep network that predicts the quality of SR images with a high correlation to human scores.

Our major contributions are as follows:

1) Developed so far the largest IQA database for image SR with a novel semi-automatic labeling method, which greatly reduced the workload of human labeling.
2) Proposed a two-stream deep network that incorporates the available LR image into the quality evaluation of HR image. It results in an end-to-end model that jointly learns perceptually consistent features from the two images and a quality predictor.
3) Designed a feature fusion method that combines the two-stream deep CNN, leading to superior performance against state-of-the-art quality prediction.

The remaining of this paper is organized as follows. Section II introduces the related work. Section III explains the methodology and process to construct the proposed large-scale database. Section IV elaborates the proposed deep network, including network architecture, feature fusion method, and model training. Section V provides the experimental results. Finally, the paper is concluded in Section VI.

II. RELATED WORK

Image SR has been an active research topic in recent years. Existing single image SR algorithms may be divided into three categories: interpolation-based [3], [4], [23], reconstruction-based [24]–[26] and learning-based [5], [6], [21], [22], [27], [28]. Interpolation-based methods are simple, efficient and of low computational cost but has limited restoration performance, because such methods often introduce severe artifacts. Built upon models of prior domain knowledge, reconstruction-based methods suppress artifacts better, but often have much higher computational complexity. Learning-based methods learn LR-to-HR image mapping from training data.

During the past decades, numerous perceptual IQA metrics have been proposed to predict the visual quality of images with full-reference, reduced-reference and no-reference. Among them, the scope of application of full-reference metric is limited due to its requirement of unimpaired source. Recently, researchers have paid more attention to no-reference IQA. Early no-reference methods were based on Natural Scene Statistic (NSS) [29]. Common NSS features include wavelet coefficients [30], locally normalized illumination coefficients [31]. Learning-based methods have been growing steadily [32]–[34], which automatically learn the mapping between image features and the perceptual quality.

Meanwhile, Image Sharpness Assessment (ISA) technique has emerged as an effective method for IQA [35]. Hassen *et al.* identified the Local Phase Coherence (LPC) of images computed in wavelet transform domain to assess image quality [36]. Bahrami *et al.* proposed a fast no-reference ISA method based on the standard deviation of weighted Maximum Local Variation (MLV) distribution of images [37]. Blanchet *et al.* introduced an indicator of Global Phase Coherence (GPC), which decreases with blur, noise, and ringing [38]. Li *et al.* proposed a no-reference SPArse Representation-based Image SHarpness index (SPARISH), which used an overcomplete dictionary learned from natural images to measure the extent of blur [39]. Hosseini *et al.* designed two no-reference ISA metrics, called Synthetic-MaxPol [40] and HVS-MaxPol [35], which were based on MaxPol convolution kernels and MaxPol filter library, respectively.

The use of deep learning has been a strong trend in recent IQA algorithms. Kang *et al.* designed a shallow CNN to learn features from contrast normalized image patches [11]. Ma *et al.* proposed a multi-task end-to-end learning framework for no-reference IQA [41]. Yang *et al.* proposed an end-to-end SGDNet for no-reference IQA, which introduced saliency information to facilitate quality prediction [12]. Yan *et al.* integrated the NSS features prediction to the deep learning-based no-reference IQA to improve the representation and generalization ability [13].

However, most existing IQA methods are not suitable for SR images, since they are designed for images degraded by common distortions such as compression, white noise and blur. The distortions produced by SR are often mixed and more sophisticated. The study [19] has shown that the popular no-reference IQA metrics are difficult to predict the perceptual quality of SR image based on an SR Image Database (SRID).

In a lab testing environment, the pristine reference HR image may be available when evaluating SR algorithms, for which objective full-reference metrics are directly applicable. It was shown that the Mean Squared Error (MSE) is a poor

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3102401, IEEE Transactions on Multimedia

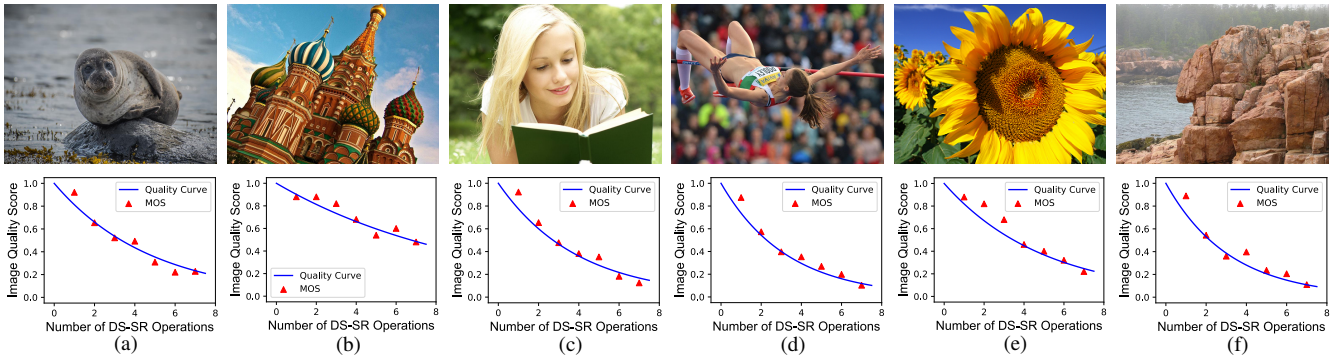


Fig. 2. Examples of exponential decay curves: (a) Animal, SRCNN [5], PLCC=0.9882; (b) Building, SAN [21], PLCC=0.9935; (c) Human, RLLR [4], PLCC=0.9923; (d) Sports, BIUCBIC [3], PLCC=0.9932; (e) Plant, RCAN [22], PLCC=0.9804; (f) Scenery, VDSR [6], PLCC=0.9804.

criterion in many cases [42]. In [43], Thapa *et al.* adopted PSNR and SSIM to compare the performance of several SR algorithms. In SR benchmark study [18], Yang *et al.* proposed a subjective IQA database of image SR (namely ECCV-14 in this paper) to analyze the effectiveness of six common full-reference IQA indicators. It was shown that these metrics fail to match the perceptual qualities of HR images. Small-scale subjective tests have also been carried out. Reibman *et al.* evaluated SR enhanced image quality through subjective tests, and pointed out that the full-reference IQA metrics cannot always capture visual quality of HR image [44].

In most practical scenarios, the reference HR images are not available, thus reduced-reference and no-reference IQA algorithms are preferred. Yeganeh *et al.* (namely Waterloo-15), based on which they proposed a weighted pooling of frequency energy falloff, dominant orientation and spatial continuity to derive an objective metric for integer-interpolated images [15]. Zhou *et al.* proposed a Quality Assessment Database for Super-resolved images (QADS) and an IQA method considering the structural and textural components of images [20]. Chen *et al.* presented a hybrid quality metric for non-integer image interpolation that combined both reduced-reference and no-reference philosophies [14]. Fang *et al.* introduced a reduced-reference quality assessment method for image SR by predicting the energy and texture similarity between LR and HR images [16]. In [10], Tang *et al.* proposed another reduced-reference IQA algorithm for SR reconstructed images with information gain and texture similarity combining saliency detection. Ma *et al.* proposed a no-reference metric by supervised learning on a database of reference-free HR images (called CVIU-17 in this paper) [9]. Furthermore, Fang *et al.* designed a Blind model based on CNN for SR-IQA (BSRIQA) [45]. In addition to the SR-IQA works, some generic IQA database also presented subject-labeled SR images, such as PieAPP [46], PIPAL [47] and BAPPS [48]. In Table I, we summarize publicly available information of existing SR-IQA databases, where PieAPP-SR represents the SR-related subset of PieAPP. These databases have greatly promoted the development of SR-IQA metrics.

Despite the significant effort, existing SR-IQA methods are limited in three aspects. First, most of them are constrained to integer scaling factors. Second, the public image SR databases are limited in size, making it difficult to train SR-IQA models without encountering serious overfitting problems. Third, there is no deep learning based model for reduced-reference SR-IQA.

TABLE I
THE EXISTING SR-IQA DATABASES

Database	Source Images	Number of SR Methods	Number of SR Images	Year
ECCV-14	10	9	540	2014
Waterloo-15	13	8	312	2015
CVIU-17	30	9	1,620	2016
SRID	20	8	480	2017
PieAPP-SR	110	5	415	2018
QADS	20	21	980	2019

Focusing on the issues above, we firstly establish a large-scale image SR database, in which the scaling factor of HR images can be arbitrary. A reduced-reference CNN-based SR-IQA method is then proposed, which combines with LR images as references and is trained on the constructed large-scale database.

III. PROPOSED LARGE-SCALE SR-IQA DATABASE

Learning-based IQA methods desire large-scale databases for training. The main challenge in building such databases is how to label a large number of images with quality ratings. Human subjective testing is desirable, but is time-consuming and expensive. Moreover, the fatigue effect when labeling large-scale datasets often affects the consistency and reliability of subjective ratings. Here we opt to a semi-automatic approach, aiming for largely reducing of workload of subjective testing.

A. The Exponential Law in Image SR

To construct a large-scale database that supports deep IQA models, we need to increase the number of SR images by an order of magnitude. This requirement inevitably leads to an explosive growth of the workload of subjective test. In this work, we address this issue by a semi-automatic rating mechanism, which is inspired by an observation of iterative DownSample-SR (DR-SR) that generates a batch of SR images of regularly distributed quality levels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3102401, IEEE Transactions on Multimedia

Due to the Nyquist-Shannon theorem, a natural image shows inferior visual quality after downsampling and its quality cannot be fully recovered by image DR. As a result, a DS-SR operation decays visual quality, and with iterative DS-SR, the quality of image would decay even further. To investigate the behavior, we perform iterative DS-SR on diverse images with different image SR algorithms, and conduct subjective tests to obtain their Mean Opinion Score (MOS) values. Interestingly, we find that the MOS value decreases with iteration, approximately following an exponentially decaying curve (with very few outliers), regardless of the image content or interpolation method, as shown in Fig. 2. Let Q and t denote the quality score and the number of SR iterations respectively, we have

$$Q(t) = e^{-bt}, t \geq 0, \tag{1}$$

where b is a positive constant depending on image content and SR method. We normalize the quality of pristine HR images to 1, thus all curves pass through a fixed point: Q (0) = 1. Given b, the quality scores of a set of HR images can be quickly acquired.

The exponential decay relationship could greatly benefit our subjective test by reducing the workload. We are able to label a subset of images with reduced workload and infer the quality of the remaining images. To examine the feasibility of this semi-automatic rating, we carry out a dedicated experiment as follows. Firstly, we randomly select over 500 images of 30 natural scenes and perform iterative DS-SR on these images. Secondly, we utilize two approaches to obtain the subjective scores. In the semi-automatic rating, 21 subjects are asked to score a subset of HR images. Using the exponential decay law of Eq. (1), we interpolate the MOS values of the remaining images. Detailed information of this approach is illustrated in the following Section III. *C*. In the full subjective test, all images are scored by all subjects to obtain all MOS values. Thirdly, we compare the MOS values obtained by the above two approaches.

In most of IQA tasks, the MOS values tend to converge with an increased number of subjects, which is termed as data saturation [49]. Our test in Fig. 3 shows that the data saturation happens with 15-20 subjects, where the correlation between MOS values approximates 1. Therefore, we collect 21 groups of valid scores through subjective experiments to ensure the number of the subjects is sufficient.

TABLE II
PLCC EVALUATION BETWEEN SEMI-AUTOMATIC RATING AND FULL SUBJECTIVE TEST

Image Types	LR Images	HR Images	PLCC	SRCC	KRCC
Animals	7	102	0.9675	0.9558	0.8606
Buildings	7	102	0.9892	0.9482	0.8682
Humans	7	102	0.9771	0.9625	0.8777
Sports	7	102	0.9837	0.9539	0.8716
Plants	4	72	0.9716	0.9261	0.8357
Scenery	4	72	0.9717	0.9701	0.9015
Average			0.9787	0.9522	0.8623
Average Subject Performance			0.9621	0.9250	0.8374

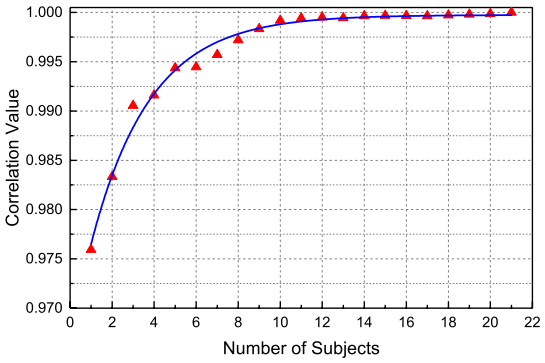


Fig. 3. Data Saturation Curve.

The comparison results are presented in Table II, where Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC) and Kendall Rank Correlation Coefficient (KRCC) are utilized as performance indicators. The table indicates an encouragingly high correlation between the two approaches. To provide a reference point, we also compute the correlations between each individual subject's labeling and MOS. The average across all subjects is a meaningful indicator of averaged subject reliability. Clearly, the semi-automatic rating achieves higher correlation than an average subject, which is another strong evidence that justifies the semi-automatic rating approach. Owning the advantages of low complexity and high accuracy, the semi-automatic rating is then utilized to generate our large-scale quality database of image SR.

B. Source Images and SR Algorithms

The first step to construct the database is to collect a batch of HR images generated by SR algorithms. In the semi-automatic rating approach, it is achieved by iterative DS-SR on source images. The source images are selected to cover diversified scenes including animals, buildings, humans, sports, plants and scenery. In total, 100 natural images of 1024×768 resolution are selected. Then, six typical SR algorithms are employed to generate HR images, including two interpolation-based methods (BICUBIC [3], RLLR [4]), two learning-based SR methods (SRCNN [5], VDSR [6]) and two GAN-based SR methods(RCAN [22], SAN [21]). In particular, SRCNN, VDSR, RCAN and SAN are constrained to image SR with integer scaling factors. We integrate these four methods and the BICUBIC algorithm to achieve non-integer image SR. In order to obtain HR images with varying qualities, different scaling factors, including 1.5, 2, 2.7, 3, 3.6 and 4, are utilized in the iterative DS-SR process. Different SR algorithms have different performance, resulting in different quality ranges under the same scaling factor and DS-SR iterations, which is unfavorable to the application of Eq. (1). To avoid this issue, we employ different scaling factors and different iteration frequencies for different SR algorithms, as shown in Table III. In particular, the scaling factor and iteration frequency are constrained to avoid fatal quality error of images, which are impractical in real-world applications.

TABLE III
COMPOSITION OF SISAR DATABASE

Source Images	Algorithms	Factors	Iteration Frequency	Total of HR Images
100	BICUBIC RLLR	1.5, 2, 2.7	8, 7, 6	4200
100	SRCNN VDSR	2	7	1400
100	SRCNN+BICUBIC VDSR+BICUBIC	1.5, 2.7	8, 6	2800
100	RCAN SAN	3, 4	7	2800
100	RCAN+BICUBIC SAN+BICUBIC	3.6	7	1400

(a) MOS: 0.7769 (b) MOS: 0.6036 (c) MOS: 0.4290

(d) MOS: 0.2831 (e) MOS: 0.1709 (f) MOS: 0.0958

Fig. 4. Sample HR Images created by SR algorithms.

From Table III, there are 12,600 HR images generated by 100 natural LR images. These LR images are processed by 10 SR algorithms or combinations of algorithms, with 6 scaling factors and a maximum iteration frequency of 8. In total, there are 12,600 HR images generated by image SR. An example of HR images of the same source is presented in Fig. 4, where the SR algorithm BICUBIC is used with a scaling factor of 2.

C. Semi-automatic Labeling

We use the semi-automatic rating approach to label the aforementioned 12,600 HR images. Each batch of images are processed by three steps: firstly, a subset of this batch is selected and subject-labeled to obtain the MOS values; secondly, the parameter b is derived by Eq. (1) and the above MOS values; thirdly, the inferred MOS (iMOS) values of the remaining images are calculated with Eq. (1) and b.

The size of subset should achieve a tradeoff between complexity and accuracy. A larger subset for subjective labeling may increase the overall accuracy but also leads to higher complexity. However, the formulation of Eq. (1) implies that the whole curve may be determined by one point, which we call the anchor point in this work. The anchor point is primarily set at the bottom half of the curve, or the second half of iteration frequency. Experimental results with 18 batches of images generated by different SR algorithms and factors show that in all cases, the correlations between semi-automatic rating and full subjective test are very high. In other words, the position of the anchor point has little impact on the interpretation performance, which may be contributed to the effectiveness of Eq. (1). As a result, we randomly select the anchor point at the bottom half of curve.

Subjective testing is conducted on the anchor point in the database, which is the most time-consuming part in our test. We recruit 23 subjects aged between 20 and 30 with regular visual acuities. The testing procedure follows the Double Stimulus Continuous Quality Scale (DSCQS) method defined in ITU R BT. 500-13 [50], where all images are randomly sorted and presented with unimpaired references. The user grading follows the rule of Absolute Category Rating (ACR)-11 scores. After statistical analysis, the scores of 2 subjects are identified to be outliers while the remaining 21 scores are averaged to obtain the MOS values. Then the MOS values are normalized to [0,1] with the maximum value 10 in ACR-11. Substituting the MOS value of anchor point into Eq. (1), we obtain the parameter b as:

$$b = -\frac{\ln(\text{MOS})}{k}, \tag{2}$$

where k is the number of iterations for the anchor point. With the value b we can derive all iMOS values by Eq. (1).

By the above process, we construct the SR Image quality database with Semi-Automatic Ratings (SISAR), which contains a total of 12,600 labeled HR images generated by SR. Among them, only 1,800 images are manually labeled with a workload of 3 hours per subject while the other images are calculated by Eq. (1). By contrast, a full subjective test of all images takes 21 hours per subject. Therefore, the semi-automatic rating significantly reduces the workload of subjective test but generates iMOS values that are highly correlated to human ratings.

D. Summary of SISAR Database

The 12,600 HR images in the SISAR database are generated by six scaling factors with ten types of SR algorithms that include six SR algorithms (BICUBIC, RLLR, SRCNN, VDSR, RCAN and SAN) and four combined SR operations (SRCNN+BICUBIC, VDSR+BICUBIC, RCAN+BICUBIC and SAN+BICUBIC). The proposed database covers diverse image contents as shown in Table IV. The histogram of the final iMOS values obtained by the proposed semi-automatic rating approach are shown in Fig. 5, where it appears that the test HR images well cover the full range of quality levels.

TABLE IV
COMPOSITION OF SISAR DATABASE

	Animals	Buildings	Humans	Sports	Plants	Scenery
Source Images	20	20	20	20	11	9
HR Images	2520	2520	2520	2520	1386	1134

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3102401, IEEE Transactions on Multimedia

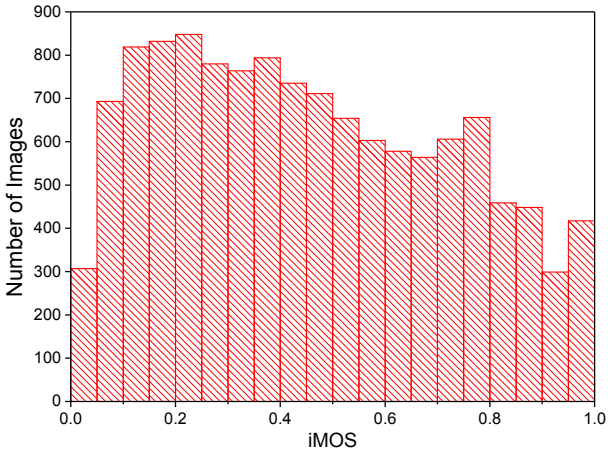


Fig. 5. Histogram of the iMOS Values in the SISAR database.

IV. Proposed Deep SR-IQA Model

We propose an end-to-end reduced-reference method for SR-IQA. The proposed model, namely Deep Image Super-resolution Quality (DISQ) index, is a two-stream CNN to evaluate the perceptual quality of HR images with LR images as references. The network architecture is illustrated in Fig. 6. Firstly, the HR and LR images are fed into different streams of CNN to extract global image features. Then, a feature fusion step is conducted in order to combine discriminative features for the quality regression. Finally, the fused features are regressed onto the quality score of HR image by fully connected layers. In the following subsections, the network structure, feature fusion and model training will be discussed in detail.

A. Network Architecture

The proposed DISQ is built on an end-to-end learning framework that estimates the quality score of an HR image with both the HR image and its corresponding LR image as inputs. All images are first split into non-overlapping patches and then fed into sub-streams of network for feature extraction and pooling to obtain global features. Considering the diversity of SR factors including non-integer ones, it is difficult to process all types of patches with a unified network. The proposed DISQ firsts utilize two convolutional modules to assemble patch-level features to obtain global features of the input LR and HR images, respectively. Then, the feature fusion module is employed to reduce all features to the same dimension through pooling and feature subtraction. Finally, the fully connected module is used to construct the mapping between the fused features and the quality of HR images.

The sub-streams of network are inspired by the architecture of VGG-16 network [51]. In the convolutional modules, all convolutions have a kernel size of 3×3 with altered kernels. Combining with max-pooling layers with a stride of 2, the network extracts local features while reducing the size of the feature maps. We employ Rectified Linear Unit (ReLU) as the nonlinear activation function. The first convolutional module, named as $\mathrm{CNN_{LR}}$, is designed to learn features of LR images. The inputs are the patches of an LR image with the size of 32×32. Another convolutional module, denoted as $\mathrm{CNN_{HR}}$, is a feature extractor of HR images. We divide each HR image into 128×128 patches as inputs to generate feature maps of a whole HR image. To facilitate feature fusion, the two CNNs contain 2 and 4 max-pooling layers correspondingly, and several convolutional layers to obtain features with similar shapes.

After feature fusion, the fused features are fed into the fully connected module to be regressed onto the perceptual scores. There are four fully connected layers with 2048, 1024, 256 and 1 neurons, respectively. We apply dropout into the first three fully connected layers with a probability of 0.5. By randomly masking out the neurons, dropout helps prevent overfitting. The last layer is a simple linear regression with a scalar output that predicts the quality score.

B. Feature Fusion

We adopt $\mathrm{CNN_{LR}}$ and $\mathrm{CNN_{HR}}$ as the feature extractors to produce the assembled patch-level feature maps F_{H} and F_{L} directly from the input HR image patches I_{p} and LR image patches I_{rp}, respectively. According to the internal structure of two convolutional modules, we can calculate the shape of output feature maps.

Feature Extraction:

$$\begin{aligned} F_{\mathrm{L}} &= \mathrm{CNN_{LR}}(I_{\mathrm{rp}};\theta_1), \mathrm{shape} = (N_{\mathrm{L}},8,8,512), \\ F_{\mathrm{H}} &= \mathrm{CNN_{HR}}(I_{\mathrm{p}};\theta_2), \mathrm{shape} = (N_{\mathrm{H}},8,8,512). \end{aligned} \quad (3)$$

Here θ_1 and θ_2 indicate the parameters of CNNs. N_{H} and N_{L} denote the numbers of HR and LR image patches. The F_{H} and F_{L} are collections of image patch features.

In order to transform the patch-level features into the image-level features, we first perform a pooling operation before image feature fusion. Specifically, the feature map is pooled into one mean tensor, max tensor and min tensor in the first dimension. Then the three tensors are concatenated into a new feature set without any further modifications.

Feature Pooling:

$$\begin{aligned} F_{\mathrm{mean}} &= \mathrm{mean}(F;\mathrm{axis}=1), \mathrm{shape} = (1,8,8,512), \\ F_{\mathrm{max}} &= \max(F;\mathrm{axis}=1), \mathrm{shape} = (1,8,8,512), \\ F_{\mathrm{min}} &= \min(F;\mathrm{axis}=1), \mathrm{shape} = (1,8,8,512), \\ F_{\mathrm{pool}} &= (F_{\mathrm{mean}},F_{\mathrm{max}},F_{\mathrm{min}}), \mathrm{shape} = (3,8,8,512). \end{aligned} \quad (4)$$

After feature pooling, there are two pooled global features F_{Hpool} and F_{Lpool} with identical structure created by F_{H} and F_{L}, respectively. We fuse the extracted image feature maps F_{Hpool} and F_{Lpool} before inputting to the regression part of the network. In image SR, the HR images are reconstructed based on the global information of LR images, which renders the difference between HR and LR image features to a meaningful representation in the feature space. In order to calculate their distance, the feature fusion step follows [17]:

$$\textbf{Feature Fusion:} \quad F_{\mathrm{fuse}} = F_{\mathrm{Hpool}} - F_{\mathrm{Lpool}}. \quad (5)$$

Idealy, F_{Lpool} in Eq. (5) should be replaced by F^0_{Hpool}, where F^0_{Hpool} denotes the pooled feature set of the perfect

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3102401, IEEE Transactions on Multimedia

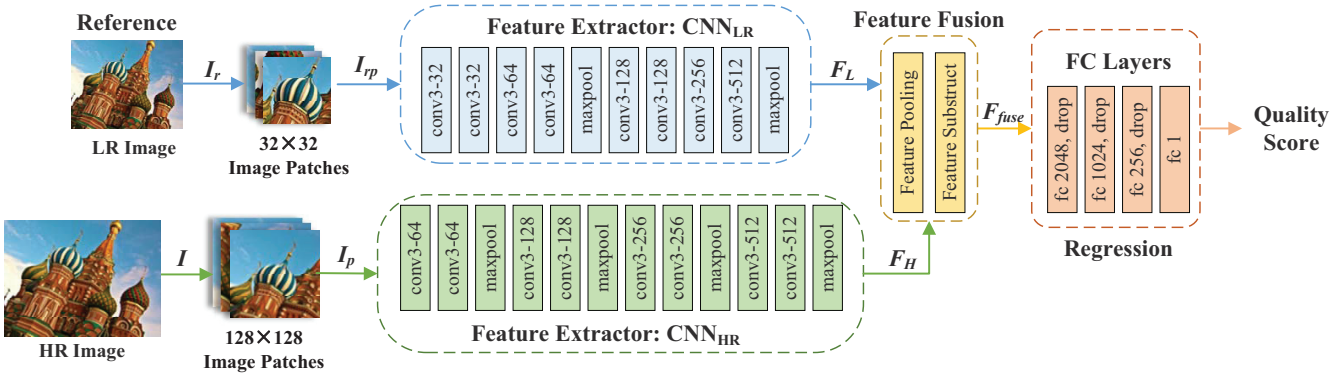


Fig. 6. Proposed DISQ Framework.

HR image. However, this perfect image is unfortunately unavailable. To investigate the impact of replacing F^0_{Hpool} with F_{Lpool}, we carry out two empirical studies. First, we perform test on a reference image and several corresponding LR images to calculate the difference of F^0_{Hpool} and F_{Lpool}. The results are given in Table V, which shows that the MSE of the two features are negligible relative to $F_{\mathrm{Hpool}} - F_{\mathrm{Lpool}}$. Second, we compute the correlation between the norm of $F_{\mathrm{Hpool}} - F_{\mathrm{Lpool}}$ and subjective scores of images. The results shown in Fig. 7 suggest that the correlation is strong. These empirical studies lead us to adopt Eq. (5). Furthermore, the effectiveness of this fusion method is also validated by experiments in the following Section V.*C*.

TABLE V
MSE BETWEEN F^0_{Hpool} AND F_{Lpool}, WHICH IS RELATIVELY SMALL COMPARE WITH THE MSE BETWEEN F_{Hpool} AND F_{Lpool} (10^{-2})

	Downsample by 2	Downsample by 2.5	Downsample by 2.7	Downsample by 3
MSE	3.357E-03	3.363E-03	3.401E-03	3.380E-03

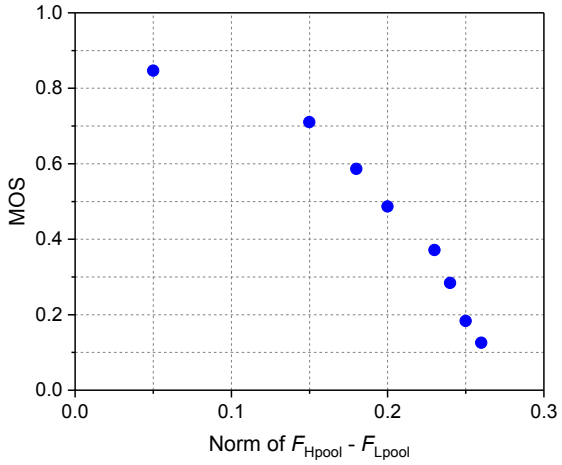


Fig. 7. Correlation between MOS and the Norm of $F_{\mathrm{Hpool}} - F_{\mathrm{Lpool}}$.

C. Model Learning

For an input HR image I and a reference LR image I_{r}, the proposed SR-IQA network M is used to predict the perceptual quality of HR image Q_{pred}:

$$Q_{\mathrm{pred}} = M((I_{\mathrm{r}}, I); \theta), \tag{6}$$

where θ indicates all parameters of this network.

Denote the ground truth quality of the input as Q_{gt}. The training goal of network M is to find the optimal parameter setting, so as to minimize the overall quality prediction loss between Q_{pred} and Q_{gt} of all HR images in the training dataset. We apply the MSE as loss function in the training process, which is widely used in various regression tasks. Moreover, in order to avoid overfitting, $l2$-norm is added to the loss as a penalty term. Thus, the loss function is:

$$\mathrm{Loss} = \frac{1}{N}\sum_{i=0}^{N} \left\| Q_{\mathrm{pred,i}} - Q_{\mathrm{gt,i}} \right\|^2 + \lambda \left\| \theta \right\|^2, \tag{7}$$

where the subscript i of $Q_{\mathrm{pred,i}}$ and $Q_{\mathrm{gt,i}}$ represent the predicted quality and group truth of the i-th image, respectively. λ is the regularization coefficient set to 0.0005 in our work.

The Adam optimizer [52] is adopted to minimize the loss function. The main advantage of Adam is that after bias correction, the learning rate of each epoch has a definite range, which makes the parameters more stable. The learning rate is η=0.0001, and other parameters of Adam optimizer are of default settings. During model training, each batch only contains one HR image and the corresponding LR image, and the batch size is equal to 1 strictly. The number of the training epochs is set to 9.

V. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed DISQ model and compare it with 16 other IQA metrics on four datasets. The cross-database test is also performed to assess the generalizability of the algorithms.

A. Experimental Setups

Databases: We train the DISQ model on the proposed SISAR database. To verify the generalizability of the proposed method, we also employ four publicly available databases,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3102401, IEEE Transactions on Multimedia

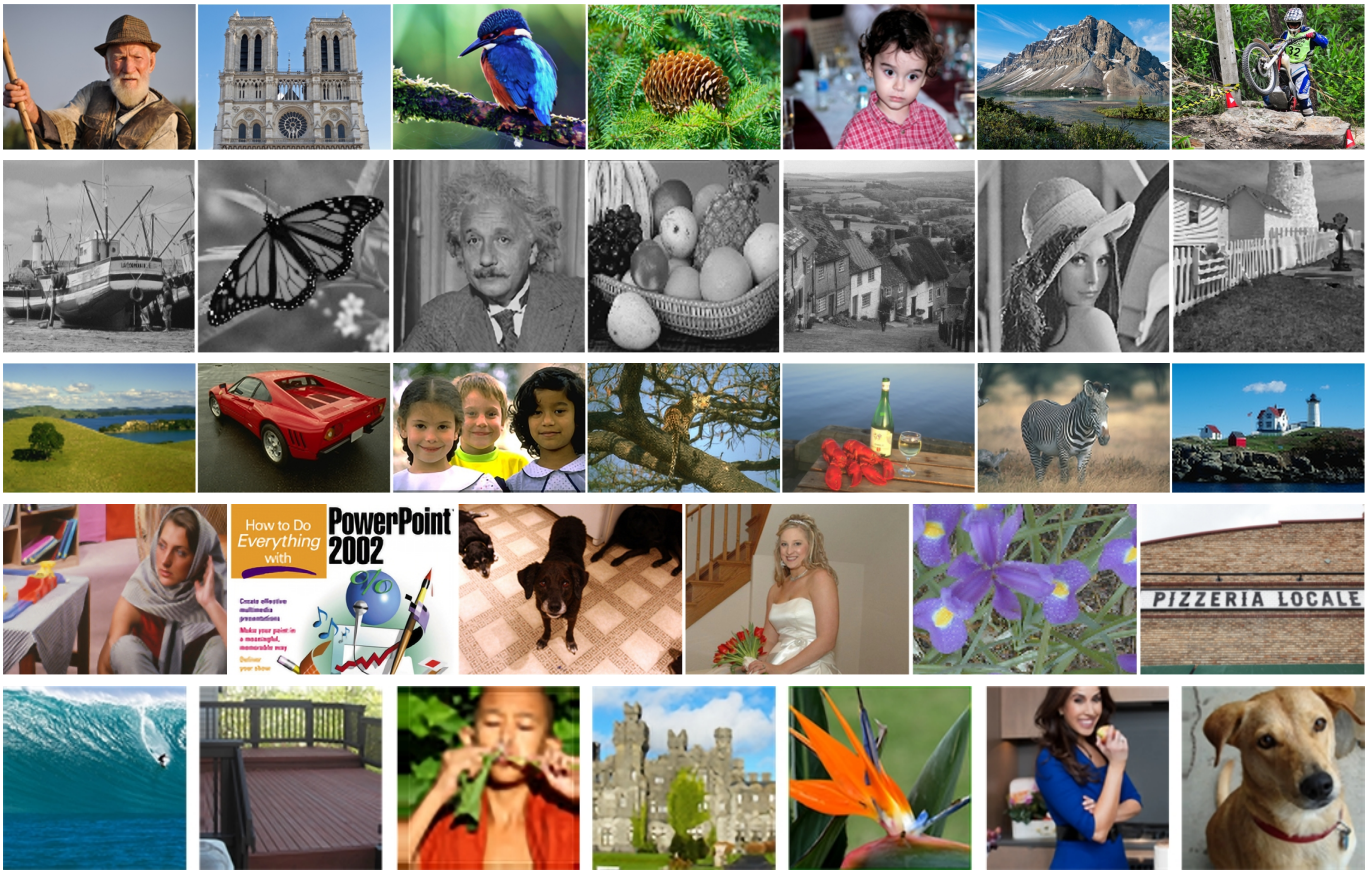


Fig. 8. Typical HR images of SISAR (line 1), Waterloo-15 (line 2), CVIU-17 (line 3), QADS (line 4) and PieAPP-SR (line 5).

Waterloo-15 [15], CVIU-17 [9], QADS [20] and PieAPP-SR [46], for cross-database validations. Typical images of these databases are listed in Fig. 8. In particular, the SRID database in Table I is unavailable for test and ECCV-14 is a subset of CVIU-17. As a result, they are not tested in our work.

SISAR contains 12,600 HR images. We divide it into training and test sets with a 80/20 split and no content overlapped, and utilize the 5-fold cross-validation to evaluate the performance of DISQ. The performance presented in the Table VI is the average performance of all test sets.

Waterloo-15 is an interpolation image database containing 312 interpolated HR images and the corresponding LR images from 13 source images [15]. The HR images were created by eight interpolation algorithms combined with scaling factors of 2, 4 and 8.

CVIU-17 is a collection of 180 LR images and 1,620 HR images, which were generated by nine SR methods and six integer scaling factors. The database is an extension of ECCV-14 [18], which contains 540 HR images.

QADS is a super-resolved image database, which created 980 HR images using 21 image SR methods from 20 reference images.

PieAPP-SR is a collection of all SR images in the PieAPP [46], which contains 415 SR images generated by Aplus [53], SRCNN [5] algorithms and so on.

The score ranges and types are not unified in these database, we choose the settings of SISAR as our standard in these experiments. Subjective scores on the other three databases are linearly scaled to the range of [0,1].

Evaluation: We use two common measurements to evaluate the performance of our algorithm by calculating the correlation between the subjective and objective quality scores: PLCC and SRCC. PLCC is used to measure the accuracy of IQA algorithms. SRCC is used to evaluate the monotonicity of quality predictions. For these metrics, a higher value up to 1 indicates better performance of a specific IQA method. The results in this work are the average values obtained by calculating the correlation coefficients based on different image contents.

B. Performance Comparison

The performance evaluation results of the proposed DISQ model are listed in Table VI and compared with other IQA methods, where the best results are shown in **bold** and the second-best results are in ***bold italics***, respectively. The compared algorithms include four no-reference IQA methods (DIIVIVE [30], BRISQUE [31], HOSA [32], CNN-IQA [11]) designed for general distorted images, six no-reference ISA indicators (LPC-SI [36], MLV [37], GPC [38], SPARISH [39], Synthetic-MaxPol [40], HVS-MaxPol [35]), two focus quality assessment algorithms (FQPath [54], FocusLiteNN [55]), and four related SR-IQA works (NSS-SR [15], HYQM

TABLE VI
PERFORMANCE COMPARISON OF IQA METHODS

Methods	Waterloo-15		CVIU-17		QADS		PieAPP-SR		SISAR	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
DIIVIVE	0.6202	0.5827	0.4602	0.4580	0.4404	0.4654	0.5828	0.6731	0.5797	0.5685
BRISQUE	0.7527	0.7626	0.5025	0.4644	0.5242	0.5461	—	—	0.5882	0.6016
CNN-IQA	0.5282	0.6339	0.3112	0.3067	0.4251	0.4959	0.5370	0.4707	0.8523	0.8568
HOSA	0.8287	***0.8236***	***0.5927***	0.5450	0.6343	0.6409	0.8009	0.7931	0.5372	0.5197
LPC-SI	0.8017	0.4689	0.4547	0.4164	0.5027	0.4902	0.6420	0.5939	0.7854	0.6563
MLV	0.7313	0.3508	0.5655	0.4628	0.2471	0.2456	0.3084	0.3244	0.7659	0.5561
GPC	***0.8310***	0.5058	0.5878	***0.5576***	0.4002	0.4470	0.7222	0.7476	0.5545	0.7394
SPARISH	0.6988	0.6584	0.4711	0.4390	0.5849	0.6530	0.7956	0.7858	0.6444	0.6910
Synthetic-MaxPol	0.5909	0.2938	0.4920	0.4537	0.4057	0.3845	0.6376	0.6301	0.5838	0.4948
HVS-MaxPol	0.7905	0.7272	0.5448	0.5557	0.5918	0.5876	0.7410	0.7197	0.5886	0.3920
FQPath	0.8038	0.7544	0.5122	0.5111	0.4879	0.4750	0.7201	0.5506	0.4445	0.4091
FocusLiteNN	0.6887	0.6979	0.4843	0.5205	0.5667	0.5643	0.6322	0.5680	0.8739	0.8760
NSS-SR	0.7733	0.6000	0.5504	0.4993	0.3670	0.2160	0.4383	0.1143	0.4507	0.3824
HYQM	0.4108	0.2096	0.2225	0.1125	0.4506	0.4443	0.7469	0.4480	0.5613	0.5046
LNQM	0.7488	0.7022	—	—	***0.7220***	***0.7274***	0.8105	0.7679	0.6337	0.5806
BSRIQA	0.7282	0.6767	0.4938	0.5841	0.6515	0.6732	***0.8155***	**0.8141**	***0.8863***	***0.8832***
DISQ	**0.8577**	**0.8373**	**0.6690**	**0.5774**	**0.7754**	**0.7716**	**0.8427**	***0.8073***	**0.9032**	**0.9051**

[14], LNQM [9], BSRIQA [45]). The source codes of these metrics are obtained from the authors' public websites. Among them, the CNN-IQA, FocusLiteNN and BSRIQA are data-driven algorithms, and are retrained on the SISAR database following the available codes to realize model convergence. For the machine-learning based LNQM method, its training code is publicly unavailable, thus its results are provided for all databases except for its training set CVIU-17. The BRISQUE method encounters errors when processing PieAPP-SR thus the corresponding results are not presented. From Table VI, we have the following observations:

First, DISQ obtains the highest PLCC and SRCC on the Waterloo-15 database, and the GPC and HOSA achieves the second best performance on PLCC and SRCC, respectively. In this database, all images were interpolated with integers. The results show that the proposed DISQ algorithm can well generalize the interpolated images with integer scaling factors.

Second, DISQ achieves a promising performance on the CVIU-17 database, while other metrics present moderate performances. CVIU-17 was created by several popular image SR algorithms, in which HR images are not limited to the common distortions. Therefore, the features modeled in common IQA methods cannot cover diverse visual contents. Although NSS-SR and NYQM are designed for image SR, they do not work well on this database, because the hand-crafted features of these methods are built specifically for image interpolation.

Third, most non-SR-IQA methods are less effective on the QADS database. DISQ and LNQM, show excellent performance in comparison with other methods. Between them, DISQ utilizes the large-scale database and deep learning to effectively extract intrinsic features, and shows much better performance. The moderate performance of the retrained CNN-based CNN-IQA, FocusLiteNN and BSRIQA may be attributed to the relatively shallow network depth.

Fourth, the proposed DISQ presents the highest PLCC on PieAPP-SR, and the BSRIQA achieves the best SRCC performance. Both of them are CNN-based SR-IQA methods, which can predict the quality of SR images more accurately than other methods. Moreover, DISQ ranks the second in terms of SRCC.

Fifth, it can be clearly observed that DISQ outperforms other metrics on the SISAR database. The retrained CNN-IQA, FocusLiteNN and BSRIQA achieve good performance, and most no-reference ISA algorithms show a moderate correlation with SISAR database. However, the performance of the retrained models drops significantly on other databases.

To fairly investigate model generalization ability, Table VII reports the average performances of the date-driven methods trained on SISAR under comparison on other benchmark databases, and the best results are highlighted in **bold**. As can be observed, DISQ achieves the best average performance in cross-database experiments. The results demonstrate the high generalization ability of the proposed DISQ model.

TABLE VII
PERFORMANCE COMPARISON OF THE CNN-BASED METHODS TRAINED ON SISAR

CNN-based Methods	PLCC	SRCC
CNN-IQA	0.4504	0.4768
FocusLiteNN	0.5930	0.5877
BSRIQA	0.6723	0.6870
DISQ	**0.7862**	**0.7487**

C. Ablation Study

To evaluate the contribution of each component in the proposed DISQ method, we conduct a series of ablation experiments. The ablation results are presented in Table VIII to XI. The best results are highlighted in **bold**.

1) Influence of the references LR images: In this experiment, we exclude the reference LR image in our model, resulting in a network with $\mathrm{CNN_{HR}}$ and fully connected modules only. This no-reference model is trained on SISAR under the same parameter settings and training steps as the proposed

TABLE VIII
DISQ PERFORMANCE COMPARISON WITH OR WITHOUT LR IMAGE AS REFERENCE

Models	SISAR		Waterloo-15		CVIU-17		QADS		PieAPP-SR	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
w/o	0.8805	0.8867	0.7473	0.6522	0.5502	0.4543	0.5740	0.5691	0.5857	0.5886
w/	**0.9032**	**0.9051**	**0.8577**	**0.8373**	**0.6690**	**0.5774**	**0.7754**	**0.7716**	**0.8427**	**0.8073**

TABLE IX
DISQ PERFORMANCE COMPARISON FOR DIFFERENT SETTINGS OF PATCH SIZES

Patch Sizes	SISAR		Waterloo-15		CVIU-17		QADS		PieAPP-SR		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
LR-32, HR-128	0.9032	0.9051	0.8577	0.8373	**0.6690**	0.5774	**0.7754**	**0.7716**	**0.8427**	**0.8073**	**0.8096**	**0.7799**
LR-32, HR-256	**0.9252**	**0.9240**	0.7570	0.7719	0.5185	0.5247	0.6455	0.6488	0.6185	0.6103	0.6929	0.6959
LR-64, HR-128	0.9016	0.9119	0.8356	0.8060	0.4745	0.6002	0.6596	0.7075	0.8123	0.7473	0.7367	0.7546
LR-64, HR-256	0.9025	0.9012	**0.8675**	**0.8469**	0.5412	0.5419	0.6836	0.6755	0.4600	0.5226	0.6910	0.6976
LR-128, HR-256	0.9096	0.9019	0.7645	0.8140	0.6137	**0.6255**	0.6508	0.5871	0.5782	0.5415	0.7034	0.6940

TABLE X
DISQ PERFORMANCE COMPARISON OF DIFFERENT FEATURE POOLING METHODS

Models	SISAR		Waterloo-15		CVIU-17		QADS		PieAPP-SR	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Max Pooling	**0.9039**	**0.9119**	0.8161	0.7854	0.5311	0.4788	0.5751	0.5634	0.7485	0.7107
Mean Pooling	0.8868	0.8944	0.8451	0.8148	0.6462	**0.6058**	0.6212	0.5722	0.7051	0.7312
Min Pooling	0.8492	0.8572	0.6876	0.6646	0.4310	0.4632	0.4343	0.4230	0.5171	0.4682
Joint Pooling	0.9032	0.9051	**0.8577**	**0.8373**	**0.6690**	0.5774	**0.7754**	**0.7716**	**0.8427**	**0.8073**

TABLE XI
DISQ PERFORMANCE COMPARISON USING DIFFERENT FEATURE FUSION METHODS

Fusion Methods & Shape of F_{fuse}	SISAR		Waterloo-15		CVIU-17		QADS		PieAPP-SR	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Method 1: (3, 8, 8, 512)	**0.9032**	**0.9051**	**0.8577**	0.8373	**0.6690**	**0.5774**	**0.7754**	**0.7716**	**0.8427**	**0.8073**
Method 2: (6, 8, 8, 512)	0.8768	0.8707	0.8348	0.8029	0.5447	0.5933	0.6983	0.6740	0.7334	0.6832
Method 3: (9, 8, 8, 512)	0.8944	0.9001	0.8559	**0.8435**	0.5521	0.4885	0.6474	0.6359	0.6819	0.7292

reduced-reference DISQ model. Its testing performance on different databases is listed in Table VIII.

Compared with two-stream DISQ model, the no-reference model achieves inferior performance, which justifies the effectiveness of introducing LR image as reference in our model.

2) Selection of image patch sizes: The DISQ model utilizes a two-steam CNN network to process LR and HR images, where these images are split into patches. Table IX lists the performance of the models with different patch sizes on several databases under the same training settings. The experimental results show that the DISQ model presents the best average performance on five databases with the patch sizes of 32×32 and 128×128.

3) Selection of feature map pooling method: In the proposed network, the feature map is pooled into a concatenated tensor, which is described in Section IV.*B*. Previous studies [11], [45] have proved that the concatenated pooling feature has certain advantages compared with common mean pooling or max pooling. To intuitively illustrate the effectiveness of the joint features, we report the performance comparison of different pooling methods under identical training epochs in Table X. The results show that the joint pooling method significantly improves the accuracy of quality prediction, and the performance of the min pooling is the lowest. The merged feature map possesses rich and robust image features, which contributes to the mapping from image features to quality.

4) Effectiveness of feature fusion methods: We incorporate the LR and HR image features in two other methods, which were also discussed in [17]. In summary, the global image features F_{Hpool} and F_{Lpool} are merged in the following three methods

Method 1: $F_{\text{fuse1}} = F_{\text{Hpool}} - F_{\text{Lpool}}$,

Method 2: $F_{\text{fuse2}} = (F_{\text{Hpool}}, F_{\text{Lpool}})$,

Method 3: $F_{\text{fuse3}} = (F_{\text{Hpool}}, F_{\text{Lpool}}, F_{\text{Hpool}} - F_{\text{Lpool}})$.

Under the same training settings, the performance of our model combined with different feature fusion methods are provided in Table XI. Method 1 exhibits the optimal performance with the fewest parameters on most databases. The network that merges F_{Hpool} and F_{Lpool} in Method 2 is theoretically

capable of learning $F_{\mathrm{Hpool}} - F_{\mathrm{Lpool}}$ in the regression part. However, the performance is worse than Method 1. Method 3 unites the first two but fails to further improve the quality prediction accuracy on most benchmarks even with an increasing number of parameters. With better or similar performance, Method 1 uses only 1/2 and 1/3 of the numbers of parameters when compared with Methods 2 and 3, respectively.

VI. CONCLUSIONS

In this work, we exploit the quality distribution of iterative DS-SR operations and propose a semi-automatic rating approach that greatly reduces the labeling workload whilst keeping high labeling accuracy. With this approach, we build SISAR, the largest-of-its-kind database for SR-IQA. Then, we propose an end-to-end DISQ model for SR-IQA, which uses a two-stream DNN for feature extraction, followed by a feature fusion network for quality prediction. By training on the SISAR database, the DISQ model achieves superior performance than state-of-the-art SR-IQA algorithms. Cross-database validation also reveals the generalization ability of our DISQ model. The proposed database and quality model will be made publicly available to facilitate reproducible research.

REFERENCES

[1] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[2] Z. Wang, J. Chen and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. (Early Access)

[3] H. Hou and H. Andrews, "Cubic splines for image interpolation and digital filtering," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 6, pp. 508–517, 1978.

[4] X. Liu, D. Zhao, R. Xiong, S. Ma and W. Gao, "Image interpolation via regularized local linear regression," pp. 118–121, 2010.

[5] C. Dong, C. C. Loy, K. He and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[6] J. Kim, J. K. Lee and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[7] M. S. M. Sajjadi, B. Schölkopf and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *2017 IEEE International Conference on Computer Vision*, 2017, pp. 4501–4510.

[8] X. Wang, K. Yu, C. Dong and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 606–615.

[9] C. Ma, C. Y. Yang, X. Yang and M. H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision & Image Understanding*, vol. 158, pp. 1–16, 2017.

[10] L. Tang, K. Sun, L. Liu, G. Wang and Y. Liu, "A reduced-reference quality assessment metric for super-resolution reconstructed images with information gain and texture similarity," *Signal Processing: Image Communication*, vol. 79, pp. 32–39, 2019.

[11] L. Kang, P. Ye, Y. Li and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[12] S. Yang, Q. Jiang, W. Lin and Y. Wang, "Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1383–1391.

[13] B. Yan, B. Bare and W. Tan, "Naturalness-aware deep no-reference image quality assessment," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2603–2615, 2019.

[14] J. Chen, Y. Xu, K. Ma, H. Huang and T. Zhao, "A hybrid quality metric for non-integer image interpolation," in *2018 Tenth International Conference on Quality of Multimedia Experience*, 2018, pp. 1–3.

[15] H. Yeganeh, M. Rostami and Z. Wang, "Objective quality assessment of interpolated natural images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4651–4663, 2015.

[16] Y. Fang, J. Liu, Y. Zhang, W. Lin and Z. Guo, "Quality assessment for image super-resolution based on energy change and texture variation," in *2016 IEEE International Conference on Image Processing*, 2016, pp. 2057–2061.

[17] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[18] C. Y. Yang, C. Ma and M. H. Yang, "Single-image super-resolution: A benchmark," 2014.

[19] G. Wang, L. Li, Q. Li, K. Gu, Z. Lu and J. Qian, "Perceptual evaluation of single-image super-resolution reconstruction," in *2017 IEEE International Conference on Image Processing*, 2017, pp. 3145–3149.

[20] F. Zhou, R. Yao, B. Liu and G. Qiu, "Visual quality assessment for super-resolved images: Database and method," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3528–3541, 2019.

[21] T. Dai, J. Cai, Y. Zhang, S. T. Xia and L. Zhang, "Second-order attention network for single image super-resolution," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 057–11 066.

[22] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong and Y. Fu, "Image super-resolution using very deep residual channel attention network," in *European Conference on Computer Vision*, 2018, pp. 294–310.

[23] Y. Romano, M. Protter and M. Elad, "Single image interpolation via adaptive nonlocal sparsity-based modeling," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3085–3098, 2014.

[24] K. Zhang, X. Gao, D. Tao and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4544–4556, 2012.

[25] V. Papyan and M. Elad, "Multi-scale patch-based image restoration," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 249–261, 2016.

[26] C. Ren, X. He and T. Q. Nguyen, "Single image super-resolution via adaptive high-dimensional non-local total variation and adaptive geometric feature," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 90–106, 2017.

[27] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei and C.-W. Lin, "Coarse-to-fine cnn for image super-resolution," *IEEE Transactions on Multimedia*, vol. 23, pp. 1489–1502, 2021.

[28] Y. Zhang, P. Wang, F. Bao, X. Yao, C. Zhang and H. Lin, "A single-image super-resolution method based on progressive-iterative approximation," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1407–1422, 2020.

[29] Z. Wang and A. C. Bovik, "Reduced- and no-reference image quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, 2011.

[30] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[31] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[32] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[33] F. Shao, W. Lin, S. Wang, G. Jiang, M. Yu and Q. Dai, "Learning receptive fields and quality lookups for blind quality assessment of stereoscopic images," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 730–743, 2016.

[34] K. Gu, D. Tao, J.-F. Qiao and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 1301–1313, 2018.

[35] M. S. Hosseini, Y. Zhang and K. N. Plataniotis, "Encoding visual sensitivity by maxpol convolution filters for image sharpness assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4510–4525, 2019.

[36] R. Hassen, Z. Wang and M. M. A. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2798–2810, 2013.

[37] K. Bahrami and A. C. Kot, "A fast approach for no-reference image sharpness assessment based on maximum local variation," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 751–755, 2014.

[38] A. Moisan, "No-reference image quality assessment and blind deblurring with sharpness metrics exploiting fourier phase information," *Journal of Mathematical Imaging and Vision*, vol. 52, no. 1, pp. 145–172, 2015.

[39] L. Li, D. Wu, J. Wu, H. Li, W. Lin and A. C. Kot, "Image sharpness assessment by sparse representation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1085–1097, 2016.

[40] M. S. Hosseini and K. N. Plataniotis, "Image sharpness metric based on maxpol convolution kernels," in *2018 25th IEEE International Conference on Image Processing*, 2018, pp. 296–300.

[41] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2018.

[42] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[43] D. Thapa, K. Raahemifar, W. R. Bobier and V. Lakshminarayanan, "A performance comparison among different super-resolution techniques," *Computers & Electrical Engineering*, 2016.

[44] A. R. Reibman, R. M. Bell and S. Gray, "Quality assessment for super-resolution image enhancement," in *2006 International Conference on Image Processing*, 2006, pp. 2017–2020.

[45] Y. Fang, Z. Chi, W. Yang, J. Liu and Z. Guo, "Blind visual quality assessment for image super-resolution by convolutional neural network," *Multimedia Tools & Applications*, vol. 77, no. 10, pp. 1–18, 2018.

[46] E. Prashnani, H. Cai, Y. Mostofi and P. Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1808–1817.

[47] H. Gu, Jinjin andCai, H. Chen, X. Ye, J. S. Ren and C. Dong, "Pipal: A large-scale image quality assessment dataset for perceptual image restoration," in *European Conference on Computer Vision*, 2020, pp. 633–651.

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[49] W. Zhang and H. Liu, "Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2424–2437, 2017.

[50] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," *Jan*, 2012.

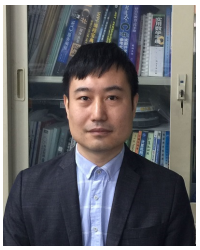
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of 3rd International Conference on Learning Representations*, 2015.

[52] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of 3rd International Conference on Learning Representations*, 2015.

[53] R. Timofte, V. De Smet and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 111–126.

[54] M. S. Hosseini, J. A. Z. Brawley-Hayes, Y. Zhang, L. Chan, K. N. Plataniotis and S. Damaskinos, "Focus quality assessment of high-throughput whole slide imaging in digital pathology," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 62–74, 2020.

[55] Z. Wang, M. Hosseini, A. Miles, K. Plataniotis and Z. Wang, "Focuslitenn: High efficiency focus quality assessment for digital pathology," in *Medical Image Computing and Computer-Assisted Intervention*. Springer International Publishing, 2020.



Tiesong Zhao (Senior member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2006, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2011. He served as a Research Associate with the Department of Computer Science, City University of Hong Kong (2011-2012), a Post-doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo (2012-2013) and a Research Scientist with the Ubiquitous Multimedia Laboratory, The State University of New York at Buffalo (2014-2015).

He is currently a Minjiang Distinguished Professor in the College of Physics and Information Engineering, Fuzhou University, China. His research interests include multimedia signal processing, coding, quality assessment and transmission. Due to his contributions in video coding and transmission, he received the Fujian Science and Technology Award for Young Scholars in 2017. He has also been serving as an Associate Editor of IET Electronics Letters since 2019.

Yuting Lin received the B.S. degree in communication engineering from Southwest University, Chongqing, China, in 2018, and the M.S. degree in communication and information system from Fuzhou University, Fuzhou, China, in 2021. Her research interests include image processing and quality perception.

Yiwen Xu (Member, IEEE) received the Ph.D degree in the department of electronic engineering from Xiamen University, Xiamen, China, in 2012. He has been an Associate Professor with the College of Physics and Information Engineering, Fuzhou University, Fujian, China, since 2013. His research interests lie in multimedia information processing, video codec and transmission, and video quality assessment.

Weiling Chen (Member, IEEE) received the B.S. and Ph.D. degrees in communication engineering from Xiamen University, Xiamen, China, in 2013 and 2018, respectively. She is currently a Lecturer with the College of Physics and Information Engineering, Fuzhou University, China. From Sep. 2016 to Dec. 2016, she was visiting at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Her current research interests include image quality perception, computer vision and underwater acoustic transmission.

Zhou Wang (S'99–M'02–SM'12–F'14) received the Ph.D. degree from The University of Texas at Austin in 2001. He is currently a Canada Research Chair and Professor in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include image and video processing and coding; visual quality assessment and optimization; computational vision and pattern analysis; multimedia communications; and biomedical signal processing. He has more than 200 publications in these fields with over 70,000 citations (Google Scholar). Dr. Wang serves as a member of IEEE Image, Video and Multidimensional Signal Processing Technical Committee (2020-2022) and IEEE Multimedia Signal Processing Technical Committee (2013-2015), a Senior Area Editor of IEEE Transactions on Image Processing (2015-2019), an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology (2016-2018), IEEE Transactions on Image Processing (2009-2014), IEEE Signal Processing Letters (2006-2010), and a Guest Editor of IEEE Journal of Selected Topics in Signal Processing (2013-2014 and 2007-2009), among other journals. He was elected a Fellow of Royal Society of Canada: Academy of Science in 2018, and a Fellow of Canadian Academy of Engineering in 2016. He is a recipient of 2021 Technology Emmy Award, 2016 IEEE Signal Processing Society Sustained Impact Paper Award, 2015 Primetime Engineering Emmy Award, 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, 2013 IEEE Signal Processing Magazine Best Paper Award, and 2009 IEEE Signal Processing Society Best Paper Award.