IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 22, 2020, accepted July 30, 2020, date of publication August 5, 2020, date of current version August 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014497

Blind Image Quality Assessment for Super Resolution via Optimal Feature Selection

JUAN BERON[1], (Student Member, IEEE), HERNAN DARIO BENITEZ-RESTREPO[1], (Member, IEEE), AND ALAN C. BOVIK[2], (Fellow, IEEE)

[1]Departamento de Electronica y Ciencias de la Computacion, Pontificia Universidad Javeriana Seccional Cali, Cali 760031, Colombia
[2]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, USA

Corresponding author: Juan Beron (juanpaberon@ieee.org)

This work was supported by the MINCIENCIAS and Pontificia Universidad Javeriana Seccional Cali with the Project Vigilancia Inteligente para la red de cámaras de la Policía Metropolitana de Cali under Grant Project 1251-745-57892.

ABSTRACT Methods for image Super Resolution (SR) have started to benefit from the development of perceptual quality predictors that are designed for super resolved images. However, extensive cross dataset validation studies have not yet been performed on Image Quality Assessment (IQA) for super resolved images. Moreover, powerful natural scene statistics-based approaches for IQA have not yet been studied for SR. To address these issues, we introduced a new dataset of super-resolved images with associated human quality scores. The dataset is based on the existing SupER dataset, which contains real low-resolution images. This new dataset also has 7 SR algorithms at three magnification scales. We selected optimal quality aware features to create two no-reference, (NR) opinion-distortion unaware (ODU) IQA models. Using the same set of selected features, we also implemented two NR-IQA opinion/distortion aware (ODA) models. The selection process identified paired-product (PP) features and those derived from discrete cosine transform coefficients (DCT) as the most relevant for the quality prediction of SR images. We conducted cross dataset validation for several state-of-the-art quality algorithms in four datasets, including our new dataset. The conducted experiments indicate that our models achieved better than state-of-the-art performance among the NR-IQA metrics. Our NR-IQA source code and the dataset are available at https://github.com/juanpaberon/IQA_SR.

INDEX TERMS No-reference image quality assessment, super resolution, image database, feature selection.

I. INTRODUCTION

Image super-resolution (SR) refers to the construction of a high-quality high-resolution (HR) image from multiple (multiple-frame SR) or a single (single-image SR) low-resolution (LR) input. In this article, we study both single-image SR (SISR) and multiple-frame SR (MFSR) models and in particular, methods for assessing the perceptual qualities of the images that they produce. SISR techniques exploit priors such as edges [1], gradients [2], [3], neighboring interpolation [4], regression [5], patches [6]–[8], and more recently, learned features extracted from deep neural networks architectures [9]. MFSR methods fuse frames with relative motion via interpolation [10], [11], iterative reconstruction [12], [13] and deep learning [14]. Both types have advanced using deep learning, which drives some of the latest and most successful SR algorithms [9], [15].

The relative performance of these algorithms have typically been evaluated using image quality assessment (IQA) models such as peak signal to noise ratio (PSNR) and the structural similarity index (SSIM) [9]. Previous works [16] have shown that PSNR and SSIM do not accurately predict perception of super-resolved image quality. Other models such as the information fidelity criterion (IFC) [17] correlate better with human perception when evaluating super resolved images. These algorithms are full reference (FR), image quality assessment and require an original pristine image, which can be impossible to obtain. By contrast, blind image quality assessment (BIQA) algorithms do not require an original image to assess quality. Previous BIQA models for super-resolved images have relied on opinion-distortion aware (ODA) image quality prediction models. Such models

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

require training on database(s) of human rated distorted images and associated human subjective opinion scores [18]. Although these prior NR models have proven effective in assessing the quality of SR images against existing IQA measures, they require a large amount of training samples, along with associated human subjective scores on a variety of distortion types. Furthermore, they often have weak generalization capability which hinders their application in practice. By comparison, opinion-distortion unaware (ODU) methods are not trained on samples of distortions, or on human subjective scores, and therefore, have greater potential for generalization. To the best of our knowledge, ODU BIQA models have not been developed to evaluate the quality of super-resolved images. This study aims to develop an ODU BIQA method, based on an optimal feature selection process, that can compete with ODA BIQA methods. In addition, we train ODA image quality prediction models using the same optimal features selected and carry out a cross dataset validation on four different datasets. These include, a new dataset of super-resolved images that we built for this work. The main contributions that we make are described as follows. First, we propose an ODU BIQA model for super-resolved IQA based on an optimal feature selection process. Second, we carry out a cross dataset generalization analysis comparing the proposed model with state-of-the-art algorithms developed for super-resolved IQA. Third, we develop a new dataset of SR images and conduct human subject studies on these images.

This article is structured as follows: Section II describes the datasets we use. Section III presents the processing and feature models we deploy to analyze the quality of super-resolved images, in addition to our method of selecting optimal perceptual quality features. Section IV develops the ODA IQA models that are derived from the selected perceptual quality features. Section V assesses the relative performances of the proposed and state-of-the-art models and discusses the results. Finally, Section VI presents concluding remarks and suggestions for further work.

II. DATASETS

Previous research into SR IQA relies on datasets that employ the popular bicubic down sampling of ground-truth high resolution (HR) images to artificially generate corresponding low resolution (LR) images [18], [19]. This strategy removes natural sensor noise and other real-world characteristics. As a result, single image SR algorithms trained on these interpolated images struggle to generalize on natural images. Here, we utilize four large benchmark IQA datasets: MY [18] and QADS [19], which are based on simulated LR images, and SupER [20], along with the new dataset that we created based on real LR images.

Table 1 summarizes the subjective datasets and SR algorithm types that we use. To limit the amount of content viewed by human subjects (and while maintaining realistic use cases and keeping the same number of scales across the datasets), we only considered the magnification scales of 2, 3 and 4.

TABLE 1. Datasets for super-resolved images with associated human scores.

Name	Num Images	Num Scenes	SR Methods	Real LR Images	SR Type
MY	1620	30	9	No	SISR
SupER	3024	14	20	Yes	SRSR MFSR
QADS	980	20	21	No	SISR
SRIJ	608	32	7	Yes	SISR

Therefore, the MY dataset is tested with 810 super-resolved images. The following section describes each of the datasets used in this study in greater detail.

A. MY

Ma *et al.* [18] proposed a database of 1620 super resolved images with associated human scores. In this article, we call this database MY after the last names of the authors. The super-resolved images derive from artificially created low-resolution images, i.e., the low-resolution images come from high-resolution images which were digitally blurred and subsampled. The 30 original high-resolution images were taken from the Berkeley segmentation dataset [21] and have a resolution of 481 $\times$ 321 pixels. By reducing the original resolution by factors of 1/2, 1/3, 1/4, 1/5, 1/6, and 1/8, they produced a total of 180 low resolution images. Then, using 8 SR methods and bicubic interpolation, they produced the 1620 super resolved images. 50 human subjects participated in their study which followed a multiple stimulus methodology. Each subject was presented with 10 images randomly selected from the dataset. The subjects then scored the images on a scale of 0 to 10.

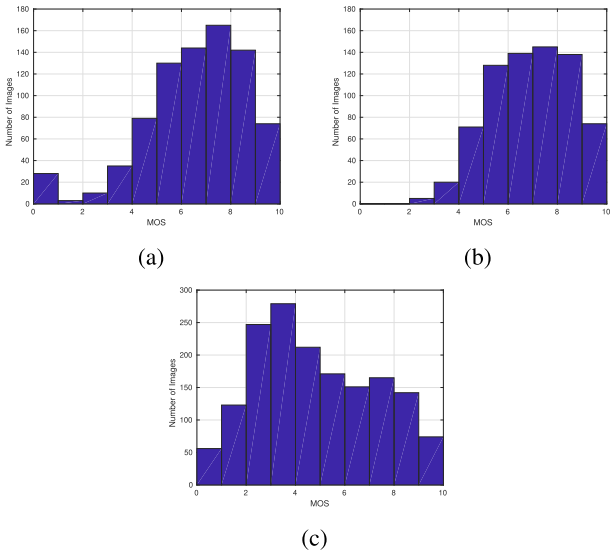


FIGURE 1. MOS distribution of the MY database. (a) MOS obtained with all scores from magnification scales 2, 3 and 4. (b) MOS obtained at magnification scales 2, 3 and 4 after removing images super resolved by Shan08. (c) MOS obtained at all magnification scales.

Our study only considered super-resolved images with magnification scales of 2, 3 and 4, since all other datasets only have images at these scales. Therefore, we only used 810 of the 1620 images. Figures 1a and 1b show the distribution of

the mean opinion scores (MOS) with and without the images super resolved by the super resolution algorithm Shan08 [2], which yielded low human scores.

Figure 2 depicts the average MOS per algorithm for the magnification scales. Shan08 had the worst performance at every scale and particularly in scale 3, which explains the outliers presented in Figure 1a. On the MY dataset, Dong11 obtained the best performance, followed by SRCNN.

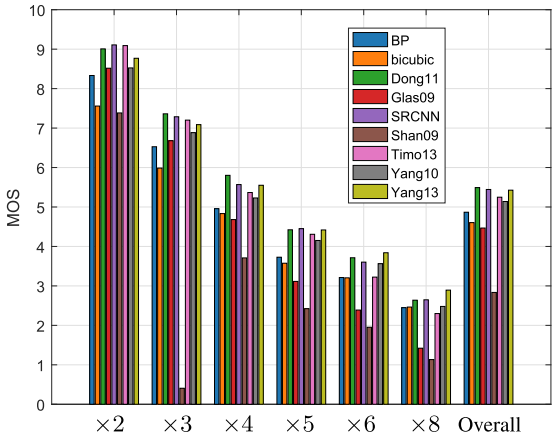


FIGURE 2. Average MOS per algorithm and per magnification scale in the MY study. For magnification scales 2, 3, 4, 5, 6 and 8, each score is the average of the MOS values of the 30 images reconstructed by each of the algorithms at that magnification scale. The group "Overall" is the average MOS value for all 180 super-resolved images produced by that particular algorithm.

B. QADS

Zhou *et al* [19] proposed a *Quality Assessment Database for SRIs* (QADS) comprised of 980 super-resolved images with associated human scores. The super resolved images came from 20 reference images thas's resolution was reduced by a factor of k ($k = 1/2,\ 1/3,\ 1/4$) by bicubic down-sampling. Thus, each SR algorithm returned the down-sampled image to its original resolution. The database used 21 methods to increase the resolution of the image: four interpolation-based methods and 17 SR algorithms. The interpolation-based methods can increase the resolutions by a factor of 2, 3 and 4. Nonetheless, not all the SR algorithms can increase the resolution by all three factors. Overall, 49 super resolved images are obtained per reference image. A total of 100 subjects participated in the study. The process used to obtain scores was based on a direct comparison of two images. Each subject was presented with two super-resolved images and the reference image of the same scene at the same time. They had to decide which super-resolved image was the best quality, or if the quality was the same (triple stimulus). Based on this information, each super-resolved image received a human score. However, the scores are only comparable between super-resolved images that show the same scene.

C. SupER

Köhler *et al* [20] developed the *Super-Resolution Erlangen* (SupER) database which provides images of 14 different scenes at four different resolutions. All of the resolutions are provided by camera hardware binning. The resolutions of the images are 2040×1080, 1020×540, 680×360 and 510×270. Images with a resolution of 2040×1080 are considered as ground truth. To test MFSR, SupER provides sequences of images under four different conditions: global motion, mixed motion, local motion and photometric variation. The authors evaluated a total of 18 SR methods: 10 MFSR and 8 SISR, as well as bicubic and nearest-neighbor interpolation. They also conducted a human study to obtain human scores for 3024 super resolved images. Their subjects were presented with two images and had to decide which of the two was better (double stimulus). They only compared images of the same scene, the same motion type and the same magnification scale. Therefore, the only difference between a pair of images was the implemented SR method. To produce scores from the comparisons, the authors utilized the Bradley Terry (B-T) model [22]. However, as a consequence of comparing only similar images (using the same scene, magnification scale and motion type), the B-T model had to be adjusted for each set of similar images. This implies that the comparison of two scores was only possible if the two scores came from the same B-T model. In other words, if the scores were associated with similar images. This is a significant difference to the MY dataset. Sets of comparable images in SupER normally include just 20 images, whereas the MY dataset can have up to 1620 images.

D. SRIJ

We created the new dataset that we call the Super Resolution Image quality assessment Javeriana (SRIJ) Database, by implementing 7 different SR methods: Dong11 [6], SRCNN [5], SRGAN [23], Timo13 [4], Yang10 [7], Yang13 [8]. We also included the bicubic method in order to complement the MY dataset. This is because 5 of the 6 SR methods have been tested on the MY dataset. We cropped the low- and high-resolution SupER images (captured using a hardware binning sensor) into 32 different scenes, as shown in Figure 3. The patches were captured in such a way that for the same scene the super resolved and reference images are aligned. Five of the six SR methods and bicubic interpolation yielded super-resolved images at three different magnification scales (2, 3, 4), while SRGAN only produced magnification scale 4. This yielded a total of 608 super-resolved images.

We conducted our human study following a similar procedure as described in [24], [25] and [26]. The set up was a *single stimulus categorical rating* [27]. One image was presented to each subject at a time in the middle of the screen, as depicted in Figure 4. After five seconds, the option to continue was shown, but the subjects could continue to view the image for as long as needed. The observers scored the image on a continuous sliding Likert quality bar with the labels; "Bad," "Poor," "Fair," "Good," and "Excellent," as shown in Figure 5.

A total of 43 volunteers participated in the study. Each of them had to score the quality of 640 images: 608 super

FIGURE 3. Sample images from the SRIJ Database.

FIGURE 4. Exemplar viewed by the subjects.

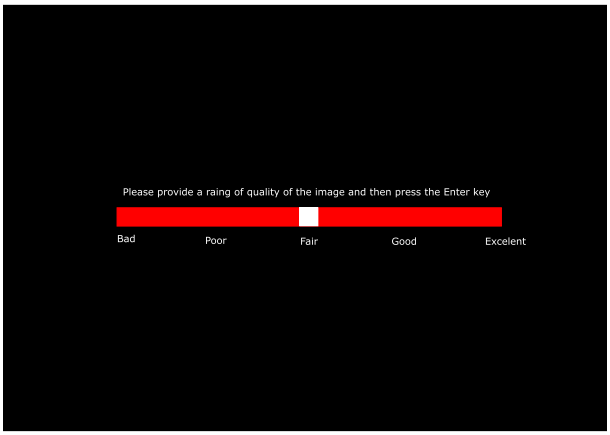


FIGURE 5. Sliding rating bar used by the subjects.

resolved and 32 ground truth images. Written informed consent was obtained from all subjects before the study. The Ethics Committee at the Faculty of Engineering and Science at Pontificia Universidad Javeriana issued an approval to carry out this study. We followed a variant of the absolute category rating with hidden references from ITU-T Rec. P.910, in which the ground truth images used as references were included in the study, but were not revealed to the subjects. To minimize subject fatigue, each subject's participation was divided into three sessions of 213, 213, and 214 images. At the beginning of each session, subjects were given a practice test of 6 images which were selected to broadly cover the quality spectrum. The images were randomly shuffled, then displayed to each subject. The illumination levels did not change significantly between sessions. In addition, the Spyder5 PRO calibrated the display to an industry color reference standard [28]. The procedure was implemented with Matlab Psychophysics Toolbox [29].

After collecting the subjective data, we computed the MOS values gathered as follows. Let d_{ij} be the score that subject i has given to the image j. To compare the subjects' scores, they are centered around zero and normalized by the standard deviation of each subject, by transforming d_{ij} into Z-scores defined as

$$z_{ij} = \frac{d_{ij} - \mu_i}{\sigma_i}$$

where

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} d_{ij} \qquad \sigma_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(d_{ij} - \mu_i)^2},$$

and n is the number of images evaluated per subject, which in this case is 640.

We also performed a subject rejection procedure as indicated in the ITU-R BT 500.11. A subject was rejected if more than 5% of his or her scores are considered abnormal. After applying this procedure, 7 of the 43 participants were rejected.

The remaining Z-scores are in the range [-3,3]. They were then linearly rescaled to [0,100] using

$$z'_{ij} = 100\frac{z_{ij} - 3}{6}.$$

The final image subjective scores were calculated as

$$MOS_j = \frac{1}{|M|}\sum_{i \in M} z'_{ij},$$

where M is the set of indices of the remaining participants.

Figures 6a and 6b depict the distributions of the subjective scores before and after rejection, showing a broad distribution skewed towards smaller values. The highest values correspond to the reference and the $\times 2$ scaled images, while the scores located around 40 correspond to the super resolved images with a magnification scales 3 and 4. This attribute can be seen more clearly in Figure 7 which shows MOS separated according to the magnification scale. While the MOS allow us to discriminate between the magnification scales well, there is a larger separation between magnification scales 2 and 3 than between 3 and 4. It is also worth noting, that even though

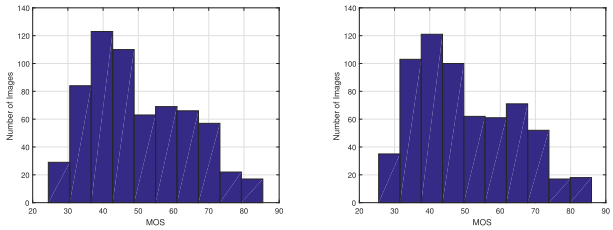


FIGURE 6. MOS distribution. (a) MOS obtained before subject rejection. (b) MOS obtained after subject rejection.

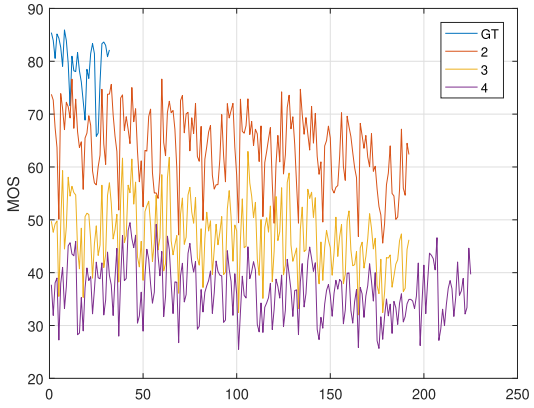


FIGURE 7. MOS obtained in the SRIJ study against the magnification scale of the images. "GT" represents ground truth images while 2, 3 and 4 are the magnification scales. The *x* axis represents the index of the image. Images with the same index were super resolved using the same method and correspond to the same reference.

some participants assigned lower scores to the reference image than to their corresponding super-resolved versions, there is generally a strong separation of the MOS between the reference and the super resolved images.

The MOS distributions presented in Figures 6a and 6b are similar to the distribution presented shown in Figure 1c which could imply that human subjects tended to prefer giving lower scores to images. The MOS distributions displayed in SRIJ has fewer images with high values, since we included the reference images in the subjective study. Nevertheless, after normalization of the SRIJ scores to the range 0-10, we found that the standard deviation of the MOS is 1.67, without including the reference images. Whereas the MOS values from MY at magnification scales 2, 3 and 4 without Shan08 have a standard deviation of 1.61.

In the same way as the analysis presented in Figure 2 for the MY study, we plotted the average MOS in SRIJ per algorithm according to the magnification scale in Figure 8. A clear difference between the MOS at different scales of magnification is also noticeable and the MOS across the evaluated SR algorithms are also similar. Nevertheless, SRCNN achieved the best MOS, on average, and at almost every magnification scale. Although SupER and SRIJ share images, it is hard to compare the human scores due to the different methodologies used to collect the human scores. SupER human scores can

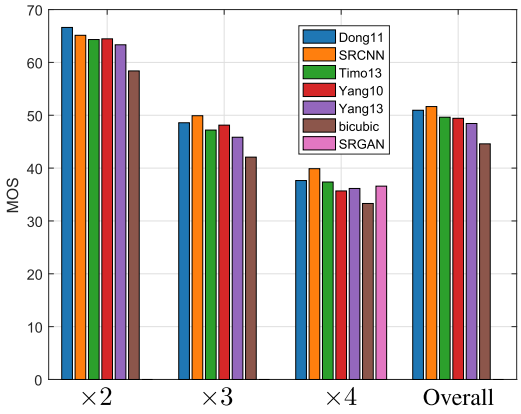


FIGURE 8. Average MOS across contents, per algorithm and magnification scale obtained in the SRIJ study. For magnification scales 2, 3 and 4, each score is the average of the MOS values for the 32 images produced by each algorithm at that magnification scale. The group "Overall" is the average MOS value of all the 96 super-resolved images produced by each algorithm.

only be compared for images of the same scene, magnification scale and movement type, while all scores are comparable for SRIJ. Hence, only the MOS from the SRIJ and MY datasets are comparable, because these studies followed the same methodology.

III. ODU NR METRIC

Existing blind image quality assessment (BIQA) models for evaluating SR algorithms are nearly all opinion-aware [18], [30], [31]. They are generally learned regression models trained on databases of images with associated human subjective scores. To achieve best performance, opinion-aware methods require large amounts of training samples with associated human subjective scores, derived from a large set of super-resolved images. The BIQA models learned by opinion-aware methods often have weak generalization capability, which limits their usability in practice. By comparison, opinion-unaware methods do not need human subjective scores for training, and thus have greater potential for good generalization capability. This research aims to develop an ODU BIQA quality prediction model based on a selection of perceptual optimal features that can compete with (and perhaps outperform) existing opinion-aware methods deployed in the quality assessment of super-resolved images.

A. PERCEPTUAL FEATURES

It has been observed that natural images, under certain transformations, such as bandpass processing, or the removal of the lowest spatial frequency [32] strongly tend towards probability distributions that can be effectively captured using several (but ultimately equivalent) parametric density models. The *Generalized Gaussian Distribution* (GGD) and the *Asymmetric Generalized Gaussian Distribution* (AGGD) are examples of such statistical models that have been widely used in previous IQA studies [24], [26], [33], [34]. The GGD

is defined as

$$f(x;\alpha,\sigma)=\frac{\alpha}{2\beta\Gamma(1/\alpha)}\exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right) \tag{1}$$

where $\Gamma(\cdot)$ is the gamma function

$$\Gamma(a)=\int_0^{\infty}t^{a-1}e^{-t}dt \quad a>0, \tag{2}$$

and

$$\beta=\sigma\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}. \tag{3}$$

This model has two parameters: α controls the "shape" and σ the standard deviation. The AGGD is defined

$$f(x;\nu,\sigma_l^2,\sigma_r^2)=\begin{cases}\frac{\nu}{(\beta_l+\beta_r)\Gamma(1/\nu)}e^{-(-x/\beta_l)^{\nu}} & x<0\\ \frac{\nu}{(\beta_l+\beta_r)\Gamma(1/\nu)}e^{-(-x/\beta_r)^{\nu}} & x\geq 0\end{cases} \tag{4}$$

where

$$\beta_l=\sigma_l\sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}},\qquad \beta_r=\sigma_r\sqrt{\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}}. \tag{5}$$

The parameters in this model are ν, σ_l, and σ_r. These parameters are frequently interpreted as distortion-sensitive, and other useful features can be derived from them, such as

$$\hat{\beta}=\frac{\beta_r+\beta_l}{2},\quad \eta=(\beta_r-\beta_l)\frac{\Gamma(2/\nu)}{\Gamma(1/\nu)}. \tag{6}$$

Moreover, the sample kurtosis and skewness can also be used to describe the distortions found in an image [35].

We fitted the distribution models (1) and (4) to a variety of bandpass coefficients obtained from the image, from which the features are derived. The bandpass processing that is applied to the image includes the following below. All of them have been successfully used in IQA models:

- Mean Subtracted Contrast Normalized (MSCN) [36]
- Paired Products [36]
- Log-Derivative [37]
- Steerable Pyramid [38], [39]
- Discrete Cosine Transform (DCT) [18], [40]
- Sigma Map [35]
- Difference of Gaussian (DoG) of Sigma Map [35]

Furthermore, we also utilize features derived from a local principal component analysis and include the principal components as features. We, then describe each of these transformations. In every case the luminance channel $I(i,j)$ is processed.

1) MSCN

Following the works presented [36] and [33], we define the local weighted luminance mean and spread as:

$$\mu(i,j)=\sum_{k=-K}^{K}\sum_{l=-K}^{K}\omega_{k,l}I(i-k,j-l), \tag{7}$$

$$\sigma(i,j)=\sqrt{\sum_{k=-K}^{K}\sum_{l=-K}^{K}\omega_{k,l}\big(I(i-k,j-l)-\mu(i,j)\big)^2}, \tag{8}$$

where ω is a 2D circular-symmetrical Gaussian filter sampled out to 3 standard deviations. In our implementation, we used $K=3$. The MSCN coefficients are then defined as

$$\hat{I}(i,j)=\frac{I(i,j)-\mu(i,j)}{\sigma(i,j)+C}, \tag{9}$$

where $C=1$ is a constant that prevents instabilities. An AGGD model is fitted to the MSCN coefficients to calculate the first two features $(\nu,\hat{\beta})$ which will be called f_1 and f_2, respectively.

As a way of capturing asymmetries in shape, we also fit the density (1) to the negative and positive MSCN coefficients. Let (α_l,ζ_l) and (α_r,ζ_r) be the parameters yielded after the fitting for the negative and positive MSCN, respectively, where α_l, α_r are the shape parameters, and ζ_l, ζ_r are the standard deviations. Then, we define the asymmetry features

$$f_3=\alpha_r-\alpha_l;\quad f_4=\zeta_r-\zeta_l. \tag{10}$$

Finally, two additional features f_5 and f_6 are defined as the sample kurtosis and the skewness of the MSCN coefficients. Thereafter, we will refer to features derived from the MSCN coefficients as $f_i,\ i=1,\ldots,4$.

2) PAIRED PRODUCTS

Paired products of MSCN coefficients are a way of capturing the correlations between them [34], [36]. If $\hat{I}(i,j)$ is the matrix of MSCN coefficients, then the paired products' coefficients are defined as

$$\begin{aligned}H(i,j)&=\hat{I}(i,j)\hat{I}(i,j+1)\\ V(i,j)&=\hat{I}(i,j)\hat{I}(i+1,j)\\ D1(i,j)&=\hat{I}(i,j)\hat{I}(i+1,j+1)\\ D2(i,j)&=\hat{I}(i,j)\hat{I}(i+1,j-1)\end{aligned} \tag{11}$$

corresponding to the horizontal (H), vertical (V), and two diagonal ($D1$ and $D2$) directions. Each of these four sets of coefficients is fitted with an AGGD. Let $(\nu^R,\eta^R,\sigma_l^R,\sigma_r^R)$ be the parameter $(\nu,\eta,\sigma_l,\sigma_r)$ estimate related to the set of paired product coefficients $R\in\{H,V,D1,D2\}$. Thus 16 more quality-aware features are obtained, which we group into four sets:

$$\begin{aligned}P_1&=\{\nu^H,\nu^V,\nu^{D1},\nu^{D2}\},\\ P_2&=\{\eta^H,\eta^V,\eta^{D1},\eta^{D2}\},\\ P_3&=\{\sigma_l^H,\sigma_l^V,\sigma_l^{D1},\sigma_l^{D2}\},\\ P_4&=\{\sigma_r^H,\sigma_r^V,\sigma_r^{D1},\sigma_r^{D2}\}.\end{aligned}$$

From here on, we will refer to the features in the sets P_i as PP (Paired Product) features.

3) LOG-DERIVATIVE COEFFICIENTS

Other useful quality sensitive features may be derived from the MSCN coefficients $\hat{I}(i,j)$ [24], [26], [37]. Let

$$J(i,j) = \log(|\hat{I}(i,j)| + G), \tag{12}$$

where $G = 0.1$. Therefore, the directional Log-Derivative coefficients are defined as:

$$\begin{aligned}
PD_1 &= J(i,j+1) - J(i,j)\\
PD_2 &= J(i+1,j) - J(i,j)\\
PD_3 &= J(i+1,j+1) - J(i,j)\\
PD_4 &= J(i+1,j-1) - J(i,j)\\
PD_5 &= J(i-1,j) + J(i+1,j)\\
&\quad - J(i,j-1) - J(i,j+1)\\
PD_6 &= J(i,j) + J(i+1,j+1)\\
&\quad - J(i,j+1) - J(i+1,j)\\
PD_7 &= J(i-1,j-1) + J(i+1,j+1)\\
&\quad - J(i-1,j+1) - J(i+1,j-1).
\end{aligned} \tag{13}$$

An AGGD is fitted to each PD_i. Let $(\nu^R, \hat{\beta}^R)$ be the parameters $(\nu, \hat{\beta})$ estimated from each set of Log-Derivative coefficients $R \in \{PD_1, \ldots, PD_7\}$. Hence, 14 features are estimated, which we arranged in two feature sets:

$$\begin{aligned}
PLD_1 &= \{\nu^{PD_1}, \ldots, \nu^{PD_7}\}\\
PLD_2 &= \{\hat{\beta}^{PD_1}, \ldots, \hat{\beta}^{PD_7}\}
\end{aligned}$$

We will collectively refer to the sets PLD_i as PLD.

4) DISCRETE COSINE TRANSFORM COEFFICIENTS

The discrete cosine transform (DCT) was implemented following the work of Ma *et al* [18], which was applied to patches of 7 by 7 pixels. From each transformed patch $\mathcal{P}$ three features are calculated. A GGD curve is fitted to the coefficient of $\mathcal{P}$ to obtain the parameter $\alpha^{\mathcal{P}}$ which is the shape parameter of (1). The standard deviation $\sigma^{\mathcal{P}}$ and the mean value $\mu^{\mathcal{P}}$ of $\mathcal{P}$ are used to obtain the normalized deviation

$$\hat{\sigma}^{\mathcal{P}} = \frac{\sigma^{\mathcal{P}}}{\mu^{\mathcal{P}}}.$$

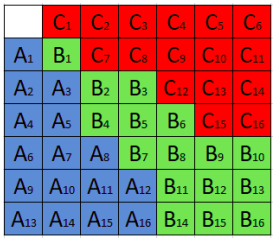


FIGURE 9. Division of the patch in three sections as indicated in Ma *et al* [18].

The patch is then divided into three sections as indicated in Figure 9, and from each section the normalized deviation $\hat{\sigma}_i$, $i = 1, 2, 3$ is calculated to obtain a final patch feature $\Sigma^{\mathcal{P}}$ as the variance of the set $\{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$. Thus, from patch $\mathcal{P}$ three values are obtained: $(\alpha^{\mathcal{P}}, \hat{\sigma}^{\mathcal{P}}, \Sigma^{\mathcal{P}})$. Consider the set $\{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$ of all the possible patches that can be obtained for the images. Therefore, three lists can be created by feature type

$$\begin{aligned}
l_1 &= [\alpha^{\mathcal{P}_1}, \ldots, \alpha^{\mathcal{P}_n}],\\
l_2 &= [\hat{\sigma}^{\mathcal{P}_1}, \ldots, \hat{\sigma}^{\mathcal{P}_n}],\\
l_3 &= [\Sigma^{\mathcal{P}_1}, \ldots, \Sigma^{\mathcal{P}_n}],
\end{aligned}$$

which are then sorted. Let dc_1, dc_2 and dc_3 be the mean values of the highest tenth percentile of l_1, l_2, and l_3; dc_4, dc_5 and dc_6 be the mean values of the lowest tenth percentile; and dc_7, dc_8 and dc_9 be their mean values, respectively. We refer to the features dc_i as DCT features.

5) SIGMA MAP

The sigma map derives from the coefficients obtained in (8). In [35] the sample kurtosis, skewness and mean of this map were used as features to train a regressor to predict picture quality scores. We denote these features as sm_1, sm_2 and sm_3, and collectively refer to them as SM hereafter.

6) DoG OF SIGMA MAP

The sigma map is further transformed by applying a Difference of Gaussians (DoG) filter, to produce another map denoted as $\text{DoG}_{\text{sigma}}$. The DoG filter is defined as

$$\text{DoG} = \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sigma_1}e^{\frac{-(x^2+y^2)}{2\sigma_1^2}} - \frac{1}{\sigma_2}e^{\frac{-(x^2+y^2)}{2\sigma_2^2}}\right),$$

where $\sigma_2 = 1.5\sigma_1$, $\sigma_1 = 1.16$ and where we use a window size of 7 by 7 pixels as defined in [35]. The MSCN coefficients are extracted from $\text{DoG}_{\text{sigma}}$ to obtain four features $g_1, \ldots, g_4$. These are the shape parameter of GGD (1) fitted to the histogram of $\text{DoG}_{\text{sigma}}$, as well as the standard deviation, kurtosis and skewness of $\text{DoG}_{\text{sigma}}$, respectively.

The MSCN coefficients of $\text{DoG}_{\text{sigma}}$ are then also calculated, to obtain two additional features g_5 and g_6 that are kurtosis and the skewness respectively. The $g_1, \ldots, g_6$ are referred to as DoG features.

7) STEERABLE PYRAMID COEFFICIENTS

Steerable Pyramid decompositions [41] have previously been implemented in IQA models [18], [24], [26], [42], and resemble the decomposition that occurs in area V1 of the visual cortex [43].

In [26] and [24] the authors suggested using six orientations of θ ($\theta \in \{0°, 30°, 60°, 90°, 120°, 150°\}$), yielding six bands D_i $i = 1, \ldots, 6$. Let $(\nu^{D_i}, \hat{\beta}^{D_i})$ be the parameters $(\nu, \hat{\beta})$ obtained after fitting the AGGD model (4) to the histogram of D_i. Hence, 12 features are obtained, which are grouped into two sets:

$$\begin{aligned}
SP_1 &= \{\nu^{D_1}, \ldots, \nu^{D_6}\}\\
SP_2 &= \{\hat{\beta}^{D_1}, \ldots, \hat{\beta}^{D_6}\}.
\end{aligned}$$

TABLE 2. Feature summary of MSCN, Paired Products (PP), Log-Derivatives (PLD), Steerable Pyramid (SP and W), Discrete Cosine Transform (DCT), Principal Component Analysis (PCA), Sigma Map (SM), Difference of Gaussians of Sigma Map (DoG). The variables in the Feature Sets column that are in lower case are taken as unitary sets while the variables in capital letters are sets with more than one feature. The second and fourth columns refer to the number of features and sets calculated at the original resolution, whereas, the last two columns refer to the total number of sets and features in all the scales.

Feature label	Number Features	Features Sets	Number of sets	Total Sets	Total Features
MSCN	6	$f_1 - f_6$	6	18	18
PP	16	$P_1 - P_4$	4	12	48
PLD	14	$PLD_1 - PLD_2$	2	6	42
SP	12	$SP_1 - SP_2$	2	6	36
DCT	9	$dc_1 - dc_9$	9	27	27
W	45	$W_1 - W_4$	4	4	45
PCA	25	PC	1	3	75
SM	3	$sm_1 - sm_3$	3	9	9
DoG	6	$g_1 - g_6$	6	6	6
TOTAL				91	306

The features in the sets SP_i will be refered to as SP (Steerable Pyramid) features.

Additionally, Ma *et al* [18] suggested the use of other features that result from a steerable pyramid decomposition. They suggested using six orientations of θ ($\theta \in \{0°, 30°, 60°, 90°, 120°, 150°\}$), but computed at two scales (original size and half resolution), thereby obtaining 12 sub bands. On the histogram of each sub band i, a GGD model is fitted to obtain 12 features that will be denoted by as_i and correspond to the parameter shape, $i = 1, \ldots, 12$. The sub bands that correspond to the same orientation are concatenated, in order to compute their histograms. A GGD model is fitted to the histograms obtaining six new shape parameters (one per orientation), yielding six new features which will be denoted by ac_i.

The windowed structural correlation [42], [44] defined as

$$p = \frac{2\sigma_{xy} + C}{\sigma_x^2 + \sigma_y^2 + C}$$

is also calculated between the high pass band and each of the sub bands as well as between sub bands. In our implementation, we used a 15×15 gaussian window with $\sigma = 1.5$ as in [18]. σ_{xy} is the cross-covariance between windowed regions, σ_x σ_y are the windowed standard deviations, and C is a constant for stabilization. Since there are 12 sub bands, p is calculated between the high pass band and each of the sub bands obtaining 12 features Phb_i, $i = 1, \ldots, 12$. Furthermore, since there are six sub bands, we can obtain 15 values of p corresponding to each possible correlation between sub bands, which we refer to as Psb_i, $i = 1, \ldots, 15$.

These 45 features are clustered into four sets:

$$W_1 = \{as_1, \ldots, as_{12}\}$$
$$W_2 = \{ac_1, \ldots, ac_6\}$$
$$W_3 = \{Phb_1, \ldots, Phb_{12}\}$$
$$W_4 = \{Psb_1, \ldots, Psb_{15}\}.$$

We will collectively refer to these features as W.

8) PRINCIPAL COMPONENT ANALYSIS COEFFICIENTS

Ma *et al* [18] and Yeganeh *et al* [45] proposed the application of Principal Component Analysis (PCA) for super resolved image quality assessment. We followed the procedure in [18] by dividing the image into 5×5 patches, and creating a 25 elements column vector per patch. These vector compose a matrix that's singular values are extracted, yielding 25 features. We will refer to this set of 25 features PC as PCA.

9) SUMMARY OF FEATURES

All features except DoG and W were calculated at three different scales (original, half resolution and quarter resolution), yielding a total of 306 features. These scales are different than the magnification scales for the SR algorithms. During subsequent analysis, we grouped the features into sets according to the type of features calculated, yielding 91 sets of features. Table 2 summarizes the features and sets of features per scale and overall.

B. OPINION-DISTORTION-UNAWARE IMAGE QUALITY ANALYZER

Natural images obey certain invariant statistics that are modified by distortions [46]. The natural images become "less natural" when they become distorted, which is reflected in the behavior of their statistics. This phenomenon is used by BIQA metrics such as NIQE and IL-NIQE, which extract sets of local features from an image, then fit the feature vectors to a multivariate Gaussian (MVG) model. The quality of a test image is then predicted as the distance between its MVG model and the MVG model learned from a corpus of pristine naturalistic images. Following the work in [47], the pristine set that we used was made up of 170 images from BSD200 [21], 29 images from LIVE IQA [48], and 90 images from the pristine set of IL-NIQE [34].

The image features that we used are the ones presented in section III-A. The quality of a test image is predicted by the standardized Euclidean distance between its features and the MVG model learned from the corpus of pristine naturalistic

images:

$$Q(x) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{c_i^2}}, \tag{14}$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ and $(c_1, c_2, \ldots, c_n)$ are the mean and variance of each feature of the same type of the pristine set, n is the number of features, which in this case is $n = 306$, and $x = (x_1, x_2, \ldots, x_n)$ are the features extracted from the image that's perceptual quality is to be predicted.

C. PERCEPTUAL FEATURE SELECTION ALGORITHM

The collected features were selected because they have previously been successfully used for IQA. However, not all the 306 features are necessarily sensitive to distortions present in super resolved images. Additionally, it is uncertain which features are more predictive of the quality of super resolved images. Therefore, we deployed a process of feature selection.

Consider any dataset of n images with associated human scores. Let (X_i, y_i) be a tuple where X_i is the vector of extracted features and y_i the associated human score for the image i. To select the optimal perceptual features, we defined families of features that's elements are not separated during the selection process. These families are the feature sets $f_1, \ldots f_6, P_1, \ldots, P_4, \ldots$ and so on which are defined in section III-A and presented in Table 2. Let Θ be the set of families of features. For $\Phi \subseteq \Theta$, define Q_Φ as in (14), with the condition that X_i is reduced to only features that are in the family of features within Φ. Therefore, by evaluating the image i in Q_Φ, the predicted scores $\hat{y}_i$ are obtained. Define the function

$$c(\Phi) = c_p(\hat{Y}_\Phi, Y), \tag{15}$$

where $\hat{Y}_\Phi$ and Y are the lists of predicted scores and human scores of the n images in the dataset, respectively, and c_p is the function that returns the Pearson correlation coefficient (PCC) obtained from two ordered lists. Let P_Θ be the set of all subsets of Θ without the empty set. Thus our optimization process is defined as:

$$\max_{\Phi \in P_\Theta} c(\Phi). \tag{16}$$

Algorithm 1 presents the optimal feature selection procedure used to determine the most representative perceptual quality features. It is based on the sequential forward floating search (SFFS) from Pudil *et al* [49] which returns the optimal subset of a set of features Γ at each possible size for the subset. This is, if $|\Gamma| = n$, therefore, SFFS returns a sequence of sets $(\gamma_i)_{i=1}^{n-1}$, where $\gamma_i \subset \Gamma$ and $|\gamma_i| = i$. The optimization is defined by a cost function c that depends on a combination of features from Θ. We have used SFFS previously in [47], with the difference that in this work our set of features $\Gamma = \Theta$ is made of families of features. We used families of features, since some features should not be separated in the selection procedure. If they are kept together, the generalization capacity of the new model will be increased.

Algorithm 1 Perceptual Feature Selection Algorithm

Require: *hScores*, *iFeats*.
  (*featSets*, *correlations*) ← sffs(*iFeats*, *hScores*)
  (*bestCorrelation*, *bIndex*) ← max(*correlations*)
  dCorrelations ← **dev**(*correlations*)
  sIndex ← {*bIndex*}
  if $|dCorrelations(bIndex)| < 10^{-4}$ **then**
    index ← *bIndex*
    while *index* + 1 ≤ 91 **do**
      index ← *index* + 1
      if $|dCorrelations(index)| < 10^{-4}$ **then**
        add *index* to *sIndex*
      else
        break
      end if
    end while
    index ← *bIndex*
    while *index* − 1 ≥ 1 **do**
      index ← *index* − 1
      if $|dCorrelations(index)| < 10^{-4}$ **then**
        add *index* to *sIndex*
      else
        break
      end if
    end while
  end if
  {*j* = 2, 3, 4}
  hScoresSR ← GetScoresSR(*hScores*)
  $hScoresSR_j$ ← *hScoresSR* from the magnification scale *j*
  perfm ← Vector of length(*sIndex*) elements
  for *i* = 1 **to** *i* = length(*sIndex*) **do**
    mScores ← EvalMetric(*iFeats*, *featSets*(*sIndex*(*i*))
    mScoresSR ← GetScoresSR(*mScores*)
    $mScoresSR_j$ ← *mScoresSR* from the magnification scale *j*
    p_j ← corr($mScoresSR_j$,$hScoresSR_j$)
    perfm(*i*) ← joinCorr(p_2,p_3,p_4)
  end for
  (*bPerfm*, *index*) ← max(*perfm*)
  return *featSets*(*sIndex*(*index*))

The families of features selected by SFFS at each possible feature vector size are represented by the variable *featSets*. Additionally the PCC for each combination is returned in the variable *correlations*.

By using MY and SRIJ, we obtained different *featSets* which we call $Q1$ and $Q2$, respectively. Figure 10 shows the *correlations* obtained in each case, where we observed an interval in which PCC has a small variation. In both cases, the elements of *featSets* that gave the best performance are located in stable correlation region, hence the selection of the element from *featSets* that provides the maximum correlation is ambiguous.

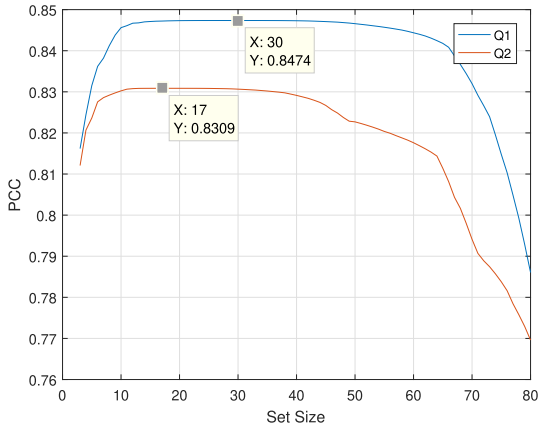


FIGURE 10. Pearson correlation coefficient for the best performing group of sets of features per feature set size. *Q*1 was evaluated on the MY dataset, while *Q*2 was evaluated on the SRIJ dataset. The values indicated on each curve correspond to the maximum correlation.

To obtain the final feature selection, we defined a no variation interval on the curves in Figure 10. This is the set of points near the maximum where the numerical derivative is less than 10^{-4}. The derivative was estimated as

$$d[n] = \frac{p[n+1] - p[n-1]}{2},$$

where n is the size of the group and $p[n]$ is the PCC obtained for the best performing group of size n. The no variation interval is represented in Algorithm 1 by *sIndex*, which corresponds to the indices of the combination of feature sets whose performance belong to the no variation interval. The function dev returns the numerical derivative of the correlations and the function max returns the greatest value, and the index of the greatest value.

For $Q1$, combinations of 15 to 47 elements were located in the no variation interval. For $Q2$, combinations of 12 to 33 elements were considered in the no variation interval. Each of these groups, with almost the same performance, were considered as a SR IQA model M. The selection process within the no variation interval was based on the performance of a model applied to images having the same magnification scale. However, MY and SRIJ datasets were not designed for accurate comparisons between images of similar quality. Hence, instead of predicting the quality of images, we predicted the performance of SR algorithms. We averaged all the scores given to the super resolved images built with the same SR algorithm and magnification scale (as explained below in detail).

Let I and J be the sets of indices corresponding to an SR algorithm and magnification scale respectively. Set Ω_{ij} with $i \in I$ and $j \in J$ corresponds to the set of super resolved images that come from the SR algorithm i and their magnification scale is j. Regarding the MY dataset, if only magnification scales 2, 3 and 4 are used, then there will be 27 sets Ω_{ij}, while in SRIJ there will be 19. Let the set of human scores and model scores on the images of the set Ω_{ij} be denoted Ω_{ij}^h and Ω_{ij}^m respectively. Thus, the human score H_{ij} and model score M_{ij} associated with the set Ω_{ij} are defined as

$$H_{ij} = \frac{1}{|\Omega_{ij}|}\left(\sum_{x\in\Omega_{ij}^h} x\right), \quad M_{ij} = \frac{1}{|\Omega_{ij}|}\left(\sum_{x\in\Omega_{ij}^m} x\right).$$

Here Ω_{ij}^h and Ω_{ij}^m correspond to *hScores* and *mScores* in Algorithm 1. The process of obtaining H_{ij} or M_{ij} is accomplished in Algorithm 1 with the function GetScoresSR. H_{ij} and M_{ij} are represented by the variables *hScoresSR* and *mScoresSR*, respectively.

The PCC at magnification scale j can be calculated between H_{ij} and M_{ij}, with j fixed. Since there are three elements in J that correspond to the three magnification scales (2, 3, and 4), then three PCC values are calculated on each model. To obtain a single correlation coefficient from a list of correlation coefficients $\{p_k : k = 1, \ldots, \bar{K} \in \mathbb{N}\}$, it was suggested in [50] to apply the following transformation to each p_k:

$$\hat{p}_k = 0.5 \ln\left(\frac{1+p_k}{1-p_k}\right), \tag{17}$$

then obtain the mean value $\bar{p}$ of $\{\hat{p}_k : k = 1, \ldots, \bar{K} \in \mathbb{N}\}$. Therefore, the representative correlation coefficient will be given by the inverse of (17) on $\bar{p}$:

$$\bar{P} = \frac{e^{2\bar{p}} - 1}{e^{2\bar{p}} + 1}. \tag{18}$$

The correlations at each scale j are represented by p_j in Algorithm 1, and the function joinCorr yields $\bar{P}$, which is saved in the vector *perfm*. This vector logs the performances of the combinations on the no variation interval. Finally, from *perfm* we identify the combination attaining the highest combined correlation value. For $Q1$, a combination of 17 elements was selected, corresponding to 45 features, while for $Q2$ a combination of 33 elements was selected, corresponding to 73 features.

The simplest approach to feature selection is testing the performance of every subset of features, which is computationally impractical. As the features are previously calculated and our predictor does not require training, then approaches such as sequential forward search and sequential backward search are possible. However, these approaches suffer from the "nesting effect" which is solved by the SFFS proposed by Pudil *et al* [49].

Other works have also explored feature selection for BIQA [51], [52]. They analyzed each feature individually by training a model and determining its performance with different metrics such as Spearman rank correlation coefficient (SRCC), PCC, root mean square error and Kendall correlation coefficient. Even though this approach is successful at detecting the best features for image quality prediction, it cannot detect features that complement each other. Additionally, since our ODU model does not require training, it is possible that the contribution to the final score of the selected features is not appropriately distributed. Therefore it is necessary to

assess the performance of the sets of features instead of only individual features.

D. SELECTED PERCEPTUAL FEATURES

Tables 3 and 4 tabulate the features selected for $Q1$ and $Q2$. These features (along with the function (14), and the pristine set defined in section III-B) define two ODU NR-IQA models which we will call $Q1_{\text{ODU}}$ and $Q2_{\text{ODU}}$, respectively. We also implemented the overall model Q_{ODU} which uses all 306 features and the same function and pristine set.

TABLE 3. Selected sets of features used in $Q1$.

	Resolution Scale		
F. label	Original	Half	Quarter
MSCN	f_1, f_3, f_6	f_4	f_4
PP	P_1, P_3	P_2, P_3	P_2, P_3
PLD	-	-	-
SP	SP_1	-	-
DCT	dc_9	dc_5	dc_5
W	W_2	-	-
PCA	-	-	-
SM	-	-	-
DoG	g_3	-	-

TABLE 4. Selected sets of features used in $Q2$.

	Resolution Scale		
F. label	Original	Half	Quarter
MSCN	f_3	f_3	f_1, f_2, f_3, f_5
PP	-	P_2	P_2
PLD	-	-	-
SP	SP_1, SP_2	-	-
DCT	dc_1, dc_3, dc_5, dc_8	dc_1, dc_5, dc_7	dc_1, dc_3, dc_7,
W	-	-	-
PCA	-	PC	-
SM	$sm_1 - sm_3$	$sm_1 - sm_3$	$sm_1 - sm_3$
DoG	g_4, g_5, g_6	-	-

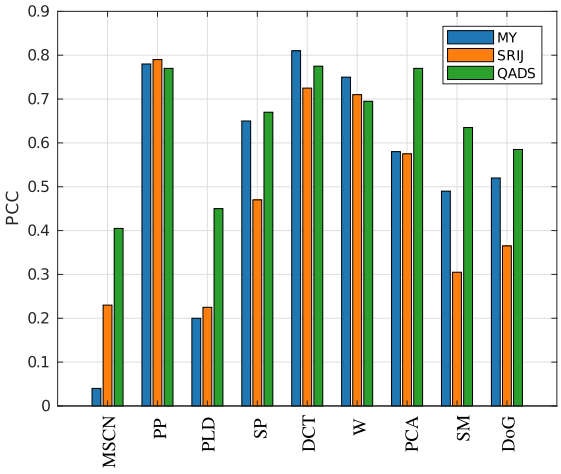


FIGURE 11. Highest PCC achieved with SFFS using only one type of feature set.

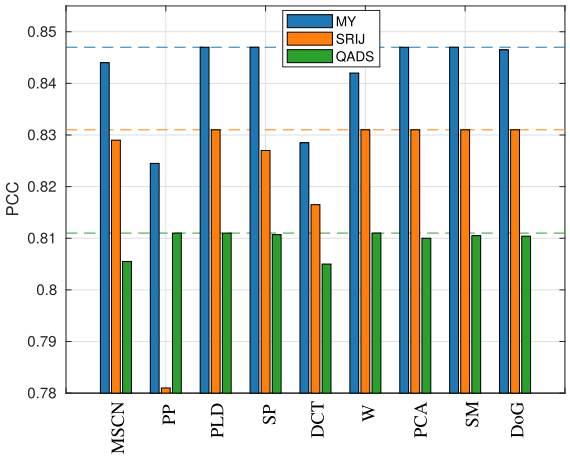


FIGURE 12. Highest PCC achieved while dropping one type of feature set. The dashed lines indicate the reference performance achieved by considering all types of feature sets.

Although Q_1 and Q_2 were optimally selected to predict the quality of super-resolved images under the definition given in III-B, it is hard to determine which of the feature sets within Q_1 and Q_2 are the most relevant to the SR image quality prediction task. In addition, it is possible that some features sets were not selected because they resemble another combination of feature sets.

To identify the feature types that achieve the best performances we applied the SFFS procedure to each type of feature and recorded the highest attained PCC as shown in Figure 11. These results indicate that PP and DCT provide the highest performance in image quality assessment. However, the information contained in a particular type of feature can also be obtained with the correct combination of other types of features. In order to find the features that's distortion sensitivity is irreplaceable, we again applied the SFFS procedure and recorded the highest PCC. However, this time removing one feature type from the selection options. The results seen in Figure 12 show that removing the *PP* features to measure the distortions found in the MY and SRIJ databases produced the largest performance loss. While, in the case of the QADS database, the *DCT* features were the ones that produced the largest performance loss.

The images created by SR algorithms may be impaired by one or several artifacts that depend on the content of the image and the type of SR algorithm. Therefore, the search for a particular type of feature able to represent the level of distortion is a challenging task. We have found that the best approach is to have a selected group of different features. However, our results show that DCT is able to provide relevant information to predict the quality of a super resolved image. DCT preserves high frequency details and it is based on a local image analysis [18]. Furthermore, it presents excellent sparsity properties in images such as energy packing efficiency, rate distortion, residual correlations, and variance distribution [53]. As it provides such compact space representations, distortions from up-scaling/SR are more likely to stand out. In addition, PP features model unnatural spatial dependencies which are common artifacts produced by SR algorithms.

An effective SR algorithm should be capable of recreating the natural correlations that exist between neighboring pixels and PP feature capture this local correlation [35], [54]. We believe that local analysis and the ability to model high frequency components and unnatural spatial dependencies provide rich SR distortion information that is not given by less relevant features, rendering the DCT and PP features irreplaceable.

IV. ODA NR METRIC DESIGN

A common practice in IQA algorithm design is to learn a regression model on a group of perceptual quality features [18], [35], [36]. Two common regression techniques used in BIQA are random forest regression (RFR) and support vector regression (SVR). Ma *et al* [18] used RFR and a two stage regression model to learn their super resolved image qualifier. We applied the same RFR procedure to produce an ODA quality analyzer from our selected features. From the features selected in $Q1$ and $Q2$, we produced two models called $Q1_{\mathrm{ODA}}$ and $Q2_{\mathrm{ODA}}$, respectively. By using all 306 features we obtained a third model, Q_{ODA}.

V. PERFORMANCE COMPARISON

To assess the importance of our feature selection procedure we compared the performance of Q_{ODU}, $Q1_{\mathrm{ODU}}$ and $Q2_{\mathrm{ODU}}$, with models using the same definition (14), but with features selected using methods from [51]. We implemented three methods, the main difference of which is the correlation coefficient used to determine the quality of the features selected. These selection procedures evaluate the performance of each feature. This is achieved by defining the model using just one feature, then, predicting and calculating a correlation coefficient between the predictions and the human scores. Hence, the feature that's correlation coefficient is above average is selected.

The correlation coefficients for the selection methods are PCC and SRCC. The metrics Q_p, Q_s and Q_{ps} rely on features selected using PCC and SRCC, and a combination of PCC and SRCC respectively. Algorithms 1, 2 and 5 from [51] were implemented for this work. Tables 5 and 6 depict the correlation coefficients obtained across the datasets. These results show that our feature selection method provides a better combination of features.

TABLE 5. Cross-Dataset evaluation with SRCC.

Metric	MY	SRIJ	QADS	SupER
Q_{ODU}	0.635	0.598	0.723	0.294
$Q1_{\mathrm{ODU}}$	**0.842**	0.793	**0.757**	0.275
$Q2_{\mathrm{ODU}}$	0.789	**0.833**	0.747	**0.338**
Q_p (MY)	0.634	0.598	0.724	0.294
Q_s (MY)	0.635	0.597	0.724	0.294
Q_{ps} (MY)	0.635	0.598	0.724	0.294
Q_p (SRIJ)	0.638	0.640	0.720	0.292
Q_s (SRIJ)	0.638	0.640	0.721	0.292
Q_{ps} (SRIJ)	0.638	0.640	0.721	0.292

While human scores given to any image in the MY and SRIJ datasets can be used to compare with any other image, the procedure is different for QADS and SupER, because the human score of an image can not be compared with any other image. The human scores associated with images in QADS are only comparable if the images are extracted from the same scene. On the other hand, scores in SupER are only comparable if the images come from the same scene, have the same magnification scale and motion type. This implies that the correlation coefficients can only be calculated from groups of scores that are comparable. QADS has 20 groups of 49 images as there are 20 different scenes. However, one of the scenes is not a natural image, therefore it was not included in the analysis. SupER has 126 groups of 20 images with comparable scores, since there are 3 magnification scales, 14 images and 3 motion types were used in this study. We then calculated 19 and 126 correlation coefficients from QADS and SupER respectively and obtained a single correlation coefficient using (17) and (18).

TABLE 6. Cross-Dataset evaluation with PCC.

Metric	MY	SRIJ	QADS	SupER
Q_{ODU}	0.742	0.580	0.725	0.507
$Q1_{\mathrm{ODU}}$	**0.847**	0.803	**0.744**	0.457
$Q2_{\mathrm{ODU}}$	0.824	**0.830**	0.742	**0.556**
Q_p (MY)	0.742	0.580	0.725	0.507
Q_s (MY)	0.742	0.580	0.725	0.506
Q_{ps} (MY)	0.742	0.580	0.725	0.507
Q_p (SRIJ)	0.742	0.637	0.725	0.491
Q_s (SRIJ)	0.742	0.637	0.725	0.491
Q_{ps} (SRIJ)	0.742	0.637	0.725	0.491

The performance obtained by the newly designed metrics were compared with the performances of state-of-the-art quality prediction models. We tested NIQE [33] and IL-NIQE [34] (which are NR ODU metrics), MY [18], PI [55] and FRIQUEE [35] (which are NR ODA) and also MS-SSIM [56], FSIM [57], SSIM, VIF [38], IFC [17], and STD [19] (which are FR metrics). We included FR models because some of them (IFC and VIF) have provided good performances on image quality prediction in previous super-resolved IQA problems [16], [20]. MY, PI and STD were specifically designed to predict human scores of super resolved images. In all the tests with NIQE and IL-NIQE, we used the pristine set defined in section III-B. We carried out some initial tests following the same procedure as in Ma *et al* [18]. 80 % of the dataset was utilized for training and the remaining 20% was used for testing. This cross validation was applied to the MY dataset and SRIJ, in which 648 and 486 randomly selected images were used for training and the remaining 162 and 122 images were designated for testing, respectively. In all cases, we kept the image content in the training and testing sets separate. We conducted 250 of these random iterations and in each iteration, SRCC and PCC were calculated against the human scores. To compare the performance of all the ODA metrics, we also calculated the correlation coefficients between the scores predicted by the models, that do not require training, and human scores. Table 7 tabulates the median values of the obtained correlations for all iterations.

TABLE 7. Correlation coefficients obtained when training with a random 80% of the dataset and testing with the remaining 20%. The process was repeated 250 times and the median results are tabulated below. The values in bold and underlined, identify the algorithm that achieved the best and second-best scores respectively.

	MY		SRIJ	
Metric	**SRCC**	**PCC**	**SRCC**	**PCC**
Q_{ODU}	0.666	0.761	0.700	0.675
$Q1_{\mathrm{ODU}}$	0.847	0.853	<u>0.846</u>	0.844
$Q2_{\mathrm{ODU}}$	0.803	0.838	**0.863**	**0.858**
NIQE	0.645	0.584	0.584	0.598
IL-NIQE	0.781	0.758	0.747	0.702
Q_{ODA}	<u>0.863</u>	**0.896**	0.828	<u>0.846</u>
$Q1_{\mathrm{ODA}}$	0.852	0.880	0.775	0.795
$Q2_{\mathrm{ODA}}$	0.845	0.883	0.767	0.754
MY	0.854	0.889	0.752	0.743
PI	0.796	0.816	0.769	0.767
FRIQUEE	0.842	0.850	-	-
MS-SSIM	0.715	0.730	0.554	0.480
FSIM	0.728	0.763	0.580	0.542
SSIM	0.594	0.635	0.577	0.521
VIF	0.761	0.749	0.630	0.672
IFC	0.766	0.724	0.573	0.615
STD	**0.877**	<u>0.893</u>	0.778	0.751

TABLE 8. Cross-Dataset evaluation with Spearman rank correlation coefficients. The evaluation dataset is indicated on the top of the columns. The dataset used for training is indicated in parenthesis. The values in bold and underlined indicate the models obtaining the best and second-best scores, respectively.

Metric	MY	SRIJ	QADS	SupER
Q_{ODU}	0.635	0.598	0.723	0.294
$Q1_{\mathrm{ODU}}$	<u>0.842</u>	0.793	0.757	0.275
$Q2_{\mathrm{ODU}}$	0.789	**0.833**	0.747	0.338
NIQE	0.641	0.491	0.634	0.344
IL-NIQE	0.767	0.649	0.859	0.200
Q_{ODA} (MY)	-	0.762	0.752	0.292
$Q1_{\mathrm{ODA}}$ (MY)	-	<u>0.809</u>	0.859	0.349
$Q2_{\mathrm{ODA}}$ (MY)	-	0.790	0.797	0.316
Q_{ODA} (SRIJ)	0.812	-	0.803	0.350
$Q1_{\mathrm{ODA}}$ (SRIJ)	0.822	-	0.801	0.342
$Q2_{\mathrm{ODA}}$ (SRIJ)	0.817	-	0.828	0.332
MY (MY)	-	0.676	0.727	0.300
MY (SRIJ)	0.786	-	0.810	0.351
PI (MY)	-	0.625	0.772	0.399
PI (SRIJ)	0.799	-	0.839	0.352
FRIQUEE(MY)	-	-	0.798	-
WaDIQaM (MY)	-	0.754	0.402	0.256
WaDIQaM (SRIJ)	0.393	-	0.481	0.158
MS-SSIM	0.694	0.557	0.918	0.568
FSIM	0.686	0.552	0.930	0.518
SSIM	0.561	0.579	0.925	0.496
VIF	0.741	0.613	<u>0.932</u>	**0.753**
IFC	0.748	0.560	0.906	0.620
STD	**0.867**	0.757	**0.952**	<u>0.740</u>

The results indicate that Q_{ODA} achieved the best performance. Q_{ODA} is a NR ODA model that utilizes all features. Tables 8 and 9 tabulate the SRCC and PCC results of the cross-dataset validation, in which we trained each model on either the MY or SRIJ dataset, then tested on the other datasets. FRIQUEE was only evaluated on the QUADS dataset because it relies on color features, while the SRIJ and SUPER datasets contain only grey level images. Furthermore, since we trained FRIQUEE on the MY dataset, we excluded it from the test datasets. We observe that even

TABLE 9. Cross-Dataset evaluation with Pearson correlation coefficients. The evaluation dataset is indicated on the top of the columns The dataset used for training is indicated in parenthesis. The values in bold and underlined indicate the best and second best scores, respectively.

Metric	MY	SRIJ	QADS	SupER
Q_{ODU}	0.742	0.580	0.725	0.507
$Q1_{\mathrm{ODU}}$	<u>0.847</u>	0.803	0.744	0.457
$Q2_{\mathrm{ODU}}$	0.824	<u>0.830</u>	0.742	0.556
NIQE	0.581	0.503	0.571	0.412
IL-NIQE	0.743	0.616	0.783	0.384
Q_{ODA} (MY)	-	0.783	0.747	0.307
$Q1_{\mathrm{ODA}}$ (MY)	-	**0.834**	0.851	0.519
$Q2_{\mathrm{ODA}}$ (MY)	-	0.803	0.796	0.471
Q_{ODA} (SRIJ)	0.743	-	0.797	0.440
$Q1_{\mathrm{ODA}}$ (SRIJ)	0.753	-	0.799	0.453
$Q2_{\mathrm{ODA}}$ (SRIJ)	0.765	-	0.824	0.481
MY (MY)	-	0.687	0.719	0.422
MY (SRIJ)	0.745	-	0.804	0.476
PI (MY)	-	0.643	0.721	0.506
PI (SRIJ)	0.759	-	0.828	0.524
FRIQUEE(MY)	-	-	0.799	-
WaDIQaM (MY)	-	0.581	0.252	0.255
WaDIQaM (SRIJ)	0.340	-	0.476	0.204
MS-SSIM	0.721	0.461	0.895	0.491
FSIM	0.731	0.500	0.917	0.446
SSIM	0.610	0.491	<u>0.922</u>	0.453
VIF	0.732	0.643	0.911	<u>0.815</u>
IFC	0.710	0.599	0.873	0.663
STD	**0.881**	0.728	**0.938**	**0.835**

though FRIQUEE relies on 560 features, the proposed models ($Q1_{\mathrm{ODA}}$ trained on the MY dataset and $Q1_{\mathrm{ODA}}$ and $Q2_{\mathrm{ODA}}$ both trained on SRIJ) outperform FRIQUEE. The state-of-the-art and deep learning-based picture quality prediction model WaDIQaM-NR, developed by Bosse *et al* [58], was included in the cross-dataset validation. Since WaDIQaM-NR requires long training times, this quality predictor is only studied in the cross-dataset evaluation. For this implementation, we fine-tuned the last two layers of the architecture based on the scores of either the MY or SRIJ datasets. We utilized a dropout rate 0.5 as used in [58] and we augmented the size of the datasets by flipping the images horizontally. Despite the fine-tuning process and the outstanding results of this quality predictor on the major synthetic-distortion picture-quality databases, the prediction accuracies obtained were not competitive against traditional methods. This was partly because there was not enough data to train it adequately [59], [60].

$Q1_{ODU}$ obtained the second-best SRCC and PCC when tested on MY which could be misleading, because the features that it uses were selected using the human scores from the MY dataset. Nevertheless, the third best SRCC and PCC were achieved by $Q2_{ODA}$ and $Q2_{ODU}$, respectively. Similarly, $Q2_{ODU}$ obtained almost the best SRCC and PCC when tested on SRIJ, while $Q1_{ODA}$ was the second best. On the other hand, when tested on the QADS and SupER datasets, the best results were obtained by the FR-IQA models, while $Q1_{ODA}$ had the best performance among the NR-IQA models when tested in QADS. SupER is a particularly hard dataset, especially for NR-IQA models which achieved very low SRCC and PCC scores on it.

Apart from STD, the FR-IQA models achieved poor performance when tested on MY and SRIJ. However, on QADS and SupER they had the most competitive scores. One main difference among the datasets is that the human scores in QADS and SupER can only be used to compare images of the same scene. Therefore, we calculated the SRCC and PCC between predicted scores and human scores of images from the same scene on MY and SRIJ datasets. This followed the same procedure as QADS and SupER for testing the relevance of image content on the prediction of image quality. Since MY has 9 SR algorithms (using the magnification scales 2, 3 and 4), per scene there are 27 human scores and predicted scores to obtain the correlation coefficients. SRIJ has 7 SR algorithms, 6 of which were used to produce images at magnification scales 2, 3 and 4. For SRGAN, only the magnification scale 4 was applied. Thus, there are 19 human and predicted scores to calculate correlation coefficients. To determine whether changes in the values of SRCC and PCC are affected by the number of collected scores, we randomly selected 27 images from the MY dataset, and maintained the proportions of the magnification scales by randomly selecting 9 images per scale. This process was conducted 1000 times. The median correlation coefficients were calculated for each iteration. A similar procedure was performed on the SRIJ database. Tables 10 and 11 show an increase in performance of almost all the models in both datasets. Nonetheless, the increase was greater for the FR-IQA models on the MY dataset, allowing them to achieve competitive scores, unlike the low performance shown in Tables 8 and 9.

TABLE 10. SRCC results obtained when selecting images randomly or only within the same scene.

	MY		SRIJ	
Metric	**Random**	**Scene**	**Random**	**Scene**
Q_{ODU}	0.649	0.858	0.621	0.796
$Q1_{\text{ODU}}$	0.847	0.898	0.800	0.858
$Q2_{\text{ODU}}$	0.798	0.886	**0.840**	**0.906**
NIQE	0.652	0.710	0.506	0.570
IL-NIQE	0.779	0.854	0.666	0.848
Q_{ODA}	0.822	0.899	0.776	0.895
$Q1_{\text{ODA}}$	0.830	0.891	0.817	0.904
$Q2_{\text{ODA}}$	0.826	0.904	0.802	0.892
MY	0.794	0.866	0.691	0.825
PI	0.807	0.875	0.642	0.753
WaDIQaM	0.404	0.592	0.771	0.852
MS-SSIM	0.702	0.848	0.573	0.570
FSIM	0.697	0.896	0.566	0.560
SSIM	0.573	0.867	0.595	0.588
VIF	0.749	0.821	0.627	0.655
IFC	0.755	0.811	0.572	0.619
STD	**0.871**	**0.907**	0.771	0.858

Köhler *et al* [20] noted that the performances of models tested on SupER at different magnification scales differed, hence we evaluated the models across scales and found a similar trend as [20]. NR-IQA models had a stronger performance at magnification scales 2 and 3 while the FR-IQA models obtained the best performance at magnification scale

TABLE 11. PCC outcomes calculated when selecting images randomly or only within the same scene.

	MY		SRIJ	
Metric	**Random**	**Scene**	**Random**	**Scene**
Q_{ODU}	0.761	0.861	0.611	0.804
$Q1_{\text{ODU}}$	**0.858**	0.903	0.819	0.881
$Q2_{\text{ODU}}$	0.839	0.909	**0.850**	0.908
NIQE	0.593	0.632	0.530	0.618
IL-NIQE	0.762	0.833	0.645	0.809
Q_{ODA}	0.754	0.827	0.803	0.926
$Q1_{\text{ODA}}$	0.764	0.818	**0.849**	0.932
$Q2_{\text{ODA}}$	0.776	0.847	0.822	**0.933**
MY	0.757	0.815	0.713	0.870
PI	0.769	0.823	0.668	0.787
WaDIQaM	0.361	0.558	0.609	0.674
MS-SSIM	0.750	0.813	0.495	0.614
FSIM	0.749	**0.922**	0.528	0.622
SSIM	0.628	0.903	0.528	0.658
VIF	0.742	0.810	0.667	0.779
IFC	0.720	0.776	0.621	0.755
STD	0.891	0.924	0.757	0.856

TABLE 12. Cross-Dataset evaluation using the Spearman Rank Correlation Coefficient on SupER. The magnification scale is indicated on the top and the column "Overall" contains the scores obtained using the whole dataset.

Metric	×2	×3	×4	Overall
Q_{ODU}	-0.0801	-0.5210	-0.2467	-0.2944
$Q1_{\text{ODU}}$	-0.2241	-0.3598	-0.2386	-0.2753
$Q2_{\text{ODU}}$	-0.3427	-0.3967	-0.2727	-0.3383
NIQE	-0.1977	-0.4877	-0.3297	-0.3440
IL-NIQE	0.2427	-0.2248	-0.5558	-0.1999
Q_{ODA} (MY)	0.4029	0.3098	0.1524	0.2916
$Q1_{\text{ODA}}$ (MY)	0.3017	0.4531	0.2860	0.3494
$Q2_{\text{ODA}}$ (MY)	0.3735	0.3430	0.2289	0.3164
Q_{ODA} (SRIJ)	0.4562	0.3538	0.2314	0.3504
$Q1_{\text{ODA}}$ (SRIJ)	0.4523	0.3610	0.1996	0.3417
$Q2_{\text{ODA}}$ (SRIJ)	0.3797	0.3621	0.2520	0.3324
MY (MY)	0.3657	0.3133	0.2168	0.2998
MY (SRIJ)	0.3558	0.4152	0.2790	0.3512
PI (MY)	-0.3238	-0.5486	-0.3048	-0.3989
PI (SRIJ)	-0.3055	-0.4561	-0.2868	-0.3520
WaDIQaM (MY)	0.088	0.262	0.403	0.226
WaDIQaM (SRIJ)	0.173	0.219	0.082	0.158
MS-SSIM	0.3306	0.5065	0.7748	0.5680
FSIM	-0.0453	0.4835	0.8454	0.5182
SSIM	0.0263	0.4435	0.8110	0.4962
VIF	0.5500	0.7889	0.8485	0.7527
IFC	0.3688	0.6389	0.7739	0.6196
STD	0.5140	0.7288	0.8753	0.7397

4. We decided to keep the sign on the correlation coefficients from Tables 12 and 13 to show that it was common to see that for some sets of images, the sign of the correlation coefficient was the opposite of that expected. For example, FSIM and SSIM are supposed to provide positive correlation coefficients, but on images with magnification scale 2, the average value was negative. We found that the only models that had a strong performance at every magnification scale were VIF and STD, while the third best model, IFC, encountered problems at the magnification scale 2.

Achieving good performance in SupER remains a difficult challenge for SR IQA. One reason for the low performances on this database is that the main factor related to the strength

TABLE 13. Cross-Dataset evaluation using the Pearson Correlation Coefficient on SupER. The magnification scale is indicated on the top and the column "Overall" contains the scores obtained using the whole dataset.

Metric	×2	×3	×4	Overall
Q_{ODU}	-0.5092	-0.6574	-0.3165	-0.5074
$Q1_{\text{ODU}}$	-0.5856	-0.4947	-0.2628	-0.4575
$Q2_{\text{ODU}}$	-0.6281	-0.5964	-0.4254	-0.5558
NIQE	-0.1548	-0.5972	-0.4377	-0.4121
IL-NIQE	0.3415	-0.5320	-0.7512	-0.3836
Q_{ODA} (MY)	0.6402	0.2958	-0.1119	0.3068
$Q1_{\text{ODA}}$ (MY)	0.6195	0.6068	0.2879	0.5189
$Q2_{\text{ODA}}$ (MY)	0.5989	0.5203	0.2593	0.4709
Q_{ODA} (SRIJ)	0.7052	0.4413	0.0641	0.4397
$Q1_{\text{ODA}}$ (SRIJ)	0.7072	0.4736	0.0690	0.4530
$Q2_{\text{ODA}}$ (SRIJ)	0.6691	0.5210	0.1841	0.4811
MY (MY)	0.6185	0.4130	0.1858	0.4218
MY (SRIJ)	0.6240	0.5264	0.2312	0.4757
PI (MY)	-0.4683	-0.6412	-0.3820	-0.5056
PI (SRIJ)	-0.6366	-0.5952	-0.2974	-0.5238
WaDIQaM (MY)	0.060	0.225	0.456	0.255
WaDIQaM (SRIJ)	0.358	0.265	-0.026	0.204
MS-SSIM	0.0140	0.5671	0.7423	0.4912
FSIM	-0.1221	0.4922	0.7705	0.4457
SSIM	-0.1381	0.4665	0.8007	0.4535
VIF	0.6199	0.8590	0.8886	0.8155
IFC	0.2774	0.7125	0.8389	0.6630
STD	0.6915	0.8668	0.8944	0.8351

of the distortion noted by human subjects is the magnification scale. This can be observed in Figures 2 and 8. The different type of SR algorithm is a secondary factor that also affects the quality. This issue is important since in the case of the SupER database, the only difference between comparable scores is the type of SR algorithm. This implies that the quality predictor must have high sensitivity to detect small changes in quality.

To determine the statistical significance of the results, we began by randomly selecting 80% of the predicted scores from each dataset, then calculating the correlation coefficients against the human scores. From MY and SRIJ, 648 and 487 images were randomly selected, respectively. From QADS, 15 of the 19 groups were randomly selected, and from SupER, 101 of the 126 groups were randomly selected. For models that required training, the scores came from a different dataset. For example, if we selected predicted scores from MY, then the models were trained on SRIJ. We performed this process 10000 times, and obtained 10000 PCC. Using these values, we conducted a Kruskal-Wallis test, in which the null hypothesis was: the median correlation for the (row) algorithm is equal to the median correlation for the (column) algorithm with a confidence of 99%. Tables 14, 15, 16 and 17 show the results of the Kruskall-Wallis tests conducted on MY, SRIJ, QADS and SupER, respectively.

As with the previous results, STD achieved the best performance, but among the NR models the performance of the proposed methods in this study achieved the best results. Table 14 indicates that $Q1_{\text{ODU}}$ attained the best performance, while $Q2_{\text{ODU}}$ was in second place. Table 15 shows that the best results were obtained by $Q2_{\text{ODU}}$, while $Q1_{\text{ODU}}$ produced the second-best results. $Q1_{\text{ODA}}$ and $Q2_{\text{ODU}}$ were statistically superior when tested on QADS and SupER, respectively.

TABLE 14. Statistical significance test results on MY dataset using the Pearson Correlation Coefficient.

	Q_{ODU}	$Q1_{\text{ODU}}$	$Q2_{\text{ODU}}$	NIQE	IL-NIQE	Q_{ODA}	$Q1_{\text{ODA}}$	$Q2_{\text{ODA}}$	MY	PI	WaDIQaM	MS-SSIM	FSIM	SSIM	VIF	IFC	STD
Q_{ODU}	-	0	0	1	0	0	0	0	0	0	1	1	1	1	1	1	0
$Q1_{\text{ODU}}$	1	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
$Q2_{\text{ODU}}$	1	0	-	1	1	1	1	1	1	1	1	1	1	1	1	1	0
NIQE	0	0	0	-	0	0	0	0	0	0	1	0	0	0	0	0	0
IL	1	0	0	1	-	-	0	0	0	0	1	1	1	1	1	1	0
Q_{ODA}	1	0	0	1	-	-	0	0	0	0	1	1	1	1	1	1	0
$Q1_{\text{ODA}}$	1	0	0	1	1	1	-	0	1	0	1	1	1	1	1	1	0
$Q2_{\text{ODA}}$	1	0	0	1	1	1	1	-	1	1	1	1	1	1	1	1	0
MY	1	0	0	1	1	1	0	0	-	0	1	1	1	1	1	1	0
PI	1	0	0	1	1	1	1	0	1	-	1	1	1	1	1	1	0
WaDIQaM	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0
MS-SSIM	0	0	0	1	0	0	0	0	0	0	1	-	0	1	0	1	0
FSIM	0	0	0	1	0	0	0	0	0	0	1	1	-	1	0	1	0
SSIM	0	0	0	1	0	0	0	0	0	0	1	0	0	-	0	0	0
VIF	0	0	0	1	0	0	0	0	0	0	1	1	1	1	-	1	0
IFC	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	-	0
STD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-

TABLE 15. Statistical significance test results on SRIJ using the Pearson Correlation Coefficient.

	Q_{ODU}	$Q1_{\text{ODU}}$	$Q2_{\text{ODU}}$	NIQE	IL-NIQE	Q_{ODA}	$Q1_{\text{ODA}}$	$Q2_{\text{ODA}}$	MY	PI	WaDIQaM	MS-SSIM	FSIM	SSIM	VIF	IFC	STD
Q_{ODU}	-	0	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0
$Q1_{\text{ODU}}$	1	-	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1
$Q2_{\text{ODU}}$	1	1	-	1	1	1	0	1	1	1	1	1	1	1	1	1	1
NIQE	0	0	0	-	0	0	0	0	0	0	0	1	1	1	0	0	0
IL	1	0	0	1	-	0	0	0	0	0	1	1	1	1	0	1	0
Q_{ODA}	1	0	0	1	1	-	0	0	1	1	1	1	1	1	1	1	1
$Q1_{\text{ODA}}$	1	1	1	1	1	1	-	1	1	1	1	1	1	1	1	1	1
$Q2_{\text{ODA}}$	1	1	0	1	1	1	0	-	1	1	1	1	1	1	1	1	1
MY	1	0	0	1	1	0	0	0	-	1	1	1	1	1	1	1	0
PI	1	0	0	1	1	0	0	0	0	-	1	1	1	1	-	1	0
WaDIQaM	1	0	0	1	0	0	0	0	0	0	-	1	1	1	0	0	0
MS-SSIM	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0
FSIM	0	0	0	0	0	0	0	0	0	0	0	1	-	1	0	0	0
SSIM	0	0	0	0	0	0	0	0	0	0	0	1	0	-	0	0	0
VIF	1	0	0	1	1	0	0	0	0	-	1	1	1	1	-	1	0
IFC	1	0	0	1	0	0	0	0	0	0	1	1	1	1	0	-	0
STD	1	0	0	1	1	0	0	0	1	1	1	1	1	1	1	1	-

TABLE 16. Statistical significance test results on QADS using the Pearson Correlation Coefficient.

	Q_{ODU}	$Q1_{\mathrm{ODU}}$	$Q2_{\mathrm{ODU}}$	NIQE	IL-NIQE	Q_{ODA} (MY)	$Q1_{\mathrm{ODA}}$ (MY)	$Q2_{\mathrm{ODA}}$ (MY)	Q_{ODA} (SRIJ)	$Q1_{\mathrm{ODA}}$ (SRIJ)	$Q2_{\mathrm{ODA}}$ (SRIJ)	MY (MY)	MY (SRIJ)	PI (MY)	PI (SRIJ)	FRIQUEE (MY)	WaDIQaM (MY)	WaDIQaM (SRIJ)	MS-SSIM	FSIM	SSIM	VIF	IFC	STD
Q_{ODU}	-	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0
$Q1_{\mathrm{ODU}}$	1	-	1	1	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0
$Q2_{\mathrm{ODU}}$	1	0	-	1	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0
NIQE	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
IL-NIQE	1	1	1	1	-	1	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0
Q_{ODA} (MY)	1	1	1	1	0	-	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0
$Q1_{\mathrm{ODA}}$ (MY)	1	1	1	1	1	1	-	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
$Q2_{\mathrm{ODA}}$ (MY)	1	1	1	1	1	1	0	-	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0
Q_{ODA} (SRIJ)	1	1	1	1	1	1	0	1	-	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0
$Q1_{\mathrm{ODA}}$ (SRIJ)	1	1	1	1	1	1	0	1	1	-	0	1	0	1	0	1	1	1	0	0	0	0	0	0
$Q2_{\mathrm{ODA}}$ (SRIJ)	1	1	1	1	1	1	0	1	1	1	-	1	1	1	0	1	1	1	0	0	0	0	0	0
MY (MY)	0	0	0	1	0	0	0	0	0	0	0	-	0	0	0	1	1	0	0	0	0	0	0	0
MY (SRIJ)	1	1	1	1	1	1	0	1	1	1	1	1	-	1	0	1	1	1	0	0	0	0	0	0
PI (MY)	0	0	0	1	0	0	0	0	0	0	0	1	0	-	0	0	1	1	0	0	0	0	0	0
PI (SRIJ)	1	1	1	1	1	1	0	1	1	1	1	1	1	1	-	1	1	1	0	0	0	0	0	0
FRIQUEE (MY)	1	1	1	1	1	1	0	0	0	0	0	1	0	1	0	-	1	1	0	0	0	0	0	0
WaDIQaM (MY)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0
WaDIQaM (SRIJ)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-	0	0	0	0	0	0
MS-SSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	0	0	0	1	0
FSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	0	1	1	0
SSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	1	1	0
VIF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	-	1	0
IFC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	-	0
STD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-

TABLE 17. Statistical significance test results on SupER using the Pearson Correlation Coefficient.

	Q_{ODU}	$Q1_{\mathrm{ODU}}$	$Q2_{\mathrm{ODU}}$	NIQE	IL-NIQE	Q_{ODA} (MY)	$Q1_{\mathrm{ODA}}$ (MY)	$Q2_{\mathrm{ODA}}$ (MY)	Q_{ODA} (SRIJ)	$Q1_{\mathrm{ODA}}$ (SRIJ)	$Q2_{\mathrm{ODA}}$ (SRIJ)	MY (MY)	MY (SRIJ)	PI (MY)	PI (SRIJ)	WaDIQaM (MY)	WaDIQaM (SRIJ)	MS-SSIM	FSIM	SSIM	VIF	IFC	STD
Q_{ODU}	-	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0
$Q1_{\mathrm{ODU}}$	0	-	0	1	1	1	0	0	1	1	0	1	0	0	0	1	1	0	1	1	0	0	0
$Q2_{\mathrm{ODU}}$	1	1	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
NIQE	0	0	0	-	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
IL-NIQE	0	0	0	0	-	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Q_{ODA} (MY)	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
$Q1_{\mathrm{ODA}}$ (MY)	1	1	0	1	1	1	-	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0
$Q2_{\mathrm{ODA}}$ (MY)	0	1	0	1	1	1	0	-	1	1	0	1	0	0	0	1	1	0	1	1	0	0	0
Q_{ODA} (SRIJ)	0	0	0	1	1	1	0	0	-	0	0	1	0	0	0	1	1	0	0	0	0	0	0
$Q1_{\mathrm{ODA}}$ (SRIJ)	0	0	0	1	1	1	0	0	1	-	0	1	0	0	0	1	1	0	1	-	0	0	0
$Q2_{\mathrm{ODA}}$ (SRIJ)	0	1	0	1	1	1	0	1	1	1	-	1	1	0	0	1	1	0	1	1	0	0	0
MY (MY)	0	0	0	1	1	1	0	0	0	0	0	-	0	0	0	1	1	0	0	0	0	0	0
MY (SRIJ)	0	1	0	1	1	1	0	1	1	1	0	1	-	0	0	1	1	0	1	1	0	0	0
PI (MY)	0	1	0	1	1	1	0	1	1	1	1	1	1	-	0	1	1	1	1	1	0	0	0
PI (SRIJ)	1	1	0	1	1	1	1	1	1	1	1	1	1	1	-	1	1	1	1	1	0	0	0
WaDIQaM (MY)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	1	0	0	0	0	0	0
WaDIQaM (SRIJ)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0
MS-SSIM	0	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	-	1	1	0	0	0
FSIM	0	0	0	1	1	1	0	0	1	0	0	1	0	0	0	1	1	0	-	0	0	0	0
SSIM	0	0	0	1	1	1	0	0	1	-	0	1	0	0	0	1	1	0	1	-	0	0	0
VIF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	1	0
IFC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	0
STD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-

The improvement in performance of $Q1_{\mathrm{ODU}}$ and $Q2_{\mathrm{ODU}}$ in comparison to NIQE and IL-NIQE shows that using SR-specific features is important. However, blindly adding many features to try to remedy performance is not correct either. This is shown by the performance of Q_{ODU}, which uses all 306 features. The results presented in Table 7 show that Q_{ODA} had the highest performance, which disagrees with the results depicted in Tables 8 and 9, implying a lack of generalization by the model. Under cross-dataset validation, we found that our metrics $Q1_{\mathrm{ODA}}$ and $Q2_{\mathrm{ODA}}$ obtained the best results among the ODA NR models. MY, PI and FRIQUEE share features with $Q1_{\mathrm{ODA}}$ and $Q2_{\mathrm{ODA}}$, since MY and PI use *DCT* features and FRIQUEE uses *PP* features. The selection procedure in this study grouped the best features derived from *DCT* and *PP* (in addition to some complementary features) to create models with better than state of the art performance under a cross-dataset validation set up.

VI. CONCLUSION

Two sets of 45 and 73 perceptual quality-aware features were selected from a group of 306 features to create two NR-IQA metrics based on the working principle of NIQE [33] and IL-NIQE [34]. IL-NIQE and NIQE are models built on feature sets that have been selected for distortions which are different to the impairments possibly presented in the outcomes of SR algorithms. In addition, any perceptual feature set is not a guarantee of good SR quality prediction as shown by the performances of IL-NIQE and NIQE. They deploy feature sets that provide lower performance than the outcomes of our proposed models. This makes it necessary to select a particular set of features specific to SR distortions. Nonetheless, the approach of simply adding new features to satisfy a possible lack of perceptual quality prediction performance is not efficacious. This is shown by the lower performance of Q_{ODU}, which uses all 306 features. In conclusion, our contribution is a formal and optimal feature selection procedure that selects the best feature set for predicting the perceptual quality of SR images. We showed the capability of image quality prediction using these selected features, by creating two NR-IQA models based on a two-stage regression model proposed by Ma *et al* [18]. These proposed new IQA models achieved state-of-the-art performance as assessed by a cross dataset validation on four different datasets: MY, QADS, SupER, and our new dataset, SRIJ. We found that the performance of the IQA models could be greatly improved when tested on the MY dataset, by only comparing images from the same scene, as was achieved in SupER and QADS. In this case, all the different types of models increased their performance. Nevertheless, the NR-IQA models and most of the FR-IQA models performed poorly on SupER. This was possibly the result of the small differences in distortion between the images. Achieving competitive performance on SupER is a challenging goal for future research.

ACKNOWLEDGMENT

The authors would also like to thank the NVIDIA Corporation for the donation of a TITAN XP GPU used in these experiments. We would also like to acknowledge the grant provided by Comision Fulbright Colombia to fund the Visiting Scholar Scholarship granted to H.D.B.-R.

REFERENCES

[1] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[2] L. Shan, F. Liu, L. Wang, and Y. Ji, "Predictive group handover scheme with channel borrowing for mobile relay systems," in *Proc. Int. Wireless Commun. Mobile Comput. Conf.*, Aug. 2008, pp. 153–158.

[3] K. In Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[4] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 184–199.

[6] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.

[7] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[8] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 561–568.

[9] R. Timofte, S. Gu, J. Wu, L. Van Gool, L. Zhang, and M. Yang, "NTIRE 2018 challenge on single image super-resolution: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 965–996.

[10] M. Batz, A. Eichenseer, and A. Kaup, "Multi-image super-resolution using a dual weighting scheme based on Voronoi tessellation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2822–2826.

[11] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 349–366, Feb. 2007.

[12] T. Kohler, X. Huang, F. Schebesch, A. Aichert, A. Maier, and J. Hornegger, "Robust multiframe super-resolution employing iteratively re-weighted minimization," *IEEE Trans. Comput. Imag.*, vol. 2, no. 1, pp. 42–58, Mar. 2016.

[13] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[14] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016.

[15] S. Nah, R. Timofte, S. Gu, S. Baik, S.-I. Hong, G. Moon, S. Son, and K. M. Lee, "NTIRE 2019 challenge on video super-resolution: Methods and results," in *Proc. CVPR Workshops*, 2019.

[16] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 372–386.

[17] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[18] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understand.*, vol. 158, pp. 1–16, May 2017.

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

[20] T. Köhler, M. Batz, F. Naderi, A. Kaup, A. Maier, and C. Riess, "Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 16, 2019, doi: 10.1109/TPAMI.2019.2917037.

[21] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 416–423.

[22] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang, "A comparative study for single image blind deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1701–1709.

[23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[24] D. E. Moreno-Villamarin, H. D. Benitez-Restrepo, and A. C. Bovik, "Predicting the quality of fused long wave infrared and visible light images," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3479–3491, Jul. 2017.

[25] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[26] T. R. Goodall, A. C. Bovik, and N. G. Paulter, "Tasking on natural statistics of infrared images," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 65–79, Jan. 2016.

[27] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, Dec. 2012.

[28] Datacolor. *Datacolor Spyder5 Family*. Accessed: Nov. 8, 2019. [Online]. Available: http://www.datacolor.com/photography-design/product-overview/spyder5-family/#spyder5pro

[29] D. H. Brainard, "The psychophysics toolbox," *Spatial Vis.*, vol. 10, no. 4, pp. 433–436, 1997.

[30] Y. Fang, C. Zhang, W. Yang, J. Liu, and Z. Guo, "Blind visual quality assessment for image super-resolution by convolutional neural network," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29829–29846, Nov. 2018.

[31] B. Bare, K. Li, B. Yan, B. Feng, and C. Yao, "A deep learning based no-reference image quality assessment model for single-image super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1223–1227.

[32] D. L. Ruderman, "The statistics of natural images," *Network-Comp Neural*, vol. 5, no. 4, pp. 517–548, 1994.

[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[34] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[35] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, Jan. 2017.

[36] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/Referenceless image spatial quality evaluator," in *Proc. Conf. Rec. 45th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2011, pp. 723–727.

[37] Y. Zhang and D. M. Chandler, "An algorithm for no-reference image quality assessment based on log-derivative statistics of natural scenes," in *Image Quality and System Performance X*, vol. 8653, P. D. Burns and S. Triantaphillidou, Eds. Bellingham, WA, USA: SPIE, Feb. 2013, pp. 156–165, doi: 10.1117/12.2001342.

[38] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[39] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[40] M. A. Saad, A. C. Bovik, and C. Charrier, "DCT statistics model-based blind image quality assessment," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 3093–3096.

[41] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

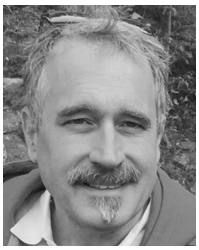
[42] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[43] B. A. Olshausen and D. J. Field, "How close are we to understanding V1?" *Neural Comput.*, vol. 17, no. 8, pp. 1665–1699, Aug. 2005.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[45] H. Yeganeh, M. Rostami, and Z. Wang, "Objective quality assessment for image super-resolution: A natural scene statistics approach," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1481–1484.
[46] A. K. Moorthy and A. C. Bovik, "Statistics of natural image distortions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 962–965.
[47] J. Beron, H. D. B. Restrepo, and A. C. Bovik, "Optimal feature selection for blind super-resolution image quality evaluation," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1842–1846.
[48] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. *Live Image Quality Assessment Database Release 2*. Accessed: Oct. 10, 2018. [Online]. Available: http://live.ece.utexas.edu/research/quality
[49] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, Nov. 1994.
[50] D. M. Corey, W. P. Dunlap, and M. J. Burke, "Averaging correlations: Expected values and bias in combined Pearson rs and fisher's z transformations," *J. Gen. Psychol.*, vol. 125, no. 3, pp. 245–261, 1998, doi: 10.1080/00221309809595548.
[51] I. F. Nizami, M. Majid, and K. Khurshid, "New feature selection algorithms for no-reference image quality assessment," *Int. J. Speech Technol.*, vol. 48, no. 10, pp. 3482–3501, Oct. 2018.
[52] I. F. Nizami, M. Majid, W. Manzoor, K. Khurshid, and B. Jeon, "Distortion-specific feature selection algorithm for universal blind image quality assessment," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, p. 19, Jan. 2019.
[53] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Boston, MA, USA: Academic, 1990.
[54] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.
[55] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "2018 PIRM challenge on perceptual image super-resolution," in *Proc. ECCV Workshops*, 2018, pp. 334–355.
[56] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.
[57] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[58] 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 Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
[59] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.
[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 586–595, vol. 34, no. 6, Jun. 2018.

JUAN BERON (Student Member, IEEE) received the B.S. degree (Hons.) in electronics engineering for his undergraduate project and academic performance from Pontificia Universidad Javeriana, Cali, Colombia, in 2019. His main research interests include computer vision, image and video processing, and image quality assessment.

HERNAN DARIO BENITEZ-RESTREPO (Member, IEEE) received the B.S. degree in electronics engineering from Pontificia Universidad Javeriana, Cali, Colombia, in 2002, and the Ph.D. degree in electrical engineering from the Universidad del Valle, Cali, in 2008. Since 2008, he has been with the Department of Electronics and Computing, Pontificia Universidad Javeriana Seccional Cali. Since 2010, he has been an Adjunct Professor with the Laboratory of Computer Vision and Systems, Université Laval, Quebec City, Canada. His main research interests encompass image and video quality assessment, infrared vision, and digital signal processing. He has been a member of the Scientific Editorial Board of the *Quantitative Infrared Thermography Journal*, since 2014. He is a member of SPIE. In 2011, he received the Merit Scholarship for short-term research from the Ministére de l'Education, du Québec to pursue research on infrared vision with the Laboratory of Computer Vision and Systems, Université Laval. He was a recipient of a Fulbright Visiting Researcher Scholarship to carry out research on video quality assessment with the Laboratory of Image and Video Engineering (LIVE), The University of Texas at Austin, in 2019. He was a Chair of the Colombia's IEEE Signal Processing, from 2012 to 2017.



ALAN C. BOVIK (Fellow, IEEE) is currently a Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. His research interests include image processing, digital television, digital streaming video, and visual perception. He was a recipient of the 2019 Progress Medal from the Royal Photographic Society, the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from the Optical Society of America, the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, and the Norbert Wiener Society Award and the Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. A perennial Web of Science Group Highly-Cited Researcher, he has also received about ten best journal paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. His recent books include *The Essential Guides to Image and Video Processing*. He co-founded and was the longest-serving Editor-in-Chief of the IEEE Transactions on Image Processing and created/chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, in 1994.