IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 21, 2019, accepted September 3, 2019, date of publication September 24, 2019, date of current version October 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943319

A Comprehensive Performance Evaluation of Image Quality Assessment Algorithms

SHAHRUKH ATHAR, (Student Member, IEEE), AND ZHOU WANG, (Fellow, IEEE)
Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Corresponding author: Shahrukh Athar (shahrukh.athar@uwaterloo.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

ABSTRACT Image quality assessment (IQA) algorithms aim to predict perceived image quality by human observers. Over the last two decades, a large amount of work has been carried out in the field. New algorithms are being developed at a rapid rate in different areas of IQA, but are often tested and compared with limited existing models using out-of-date test data. There is a significant gap when it comes to large-scale performance evaluation studies that include a wide variety of test data and competing algorithms. In this work we aim to fill this gap by carrying out the largest performance evaluation study so far. We test the performance of 43 full-reference (FR), seven fused FR (22 versions), and 14 no-reference (NR) methods on nine subject-rated IQA datasets, of which five contain singly distorted images and four contain multiply distorted content. We use a variety of performance evaluation and statistical significance testing criteria. Our findings not only point to the top performing FR and NR IQA methods, but also highlight the performance gap between them. In addition, we have also conducted a comparative study on FR fusion methods, and an important discovery is that rank aggregation based FR fusion is able to outperform not only other FR fusion approaches but also the top performing FR methods. It may be used to annotate IQA datasets as a possible alternative to subjective ratings, especially in situations where it is not possible to obtain human opinions, such as in the case of large-scale datasets composed of thousands or even millions of images.

INDEX TERMS Image quality assessment, performance evaluation, image quality study, full-reference IQA, no-reference IQA, FR fusion, rank aggregation, image databases.

I. INTRODUCTION

Image quality assessment (IQA) can be broadly categorized into *subjective* and *objective* quality assessment (QA). In subjective QA, humans are tasked to evaluate the visual quality of content and the average of subjective ratings is termed as Mean Opinion Score (MOS). Subjective QA is usually regarded as the most reliable method of quantifying perceptual quality of content since in most cases such content is meant to be viewed by humans. However, subjective QA is time consuming, expensive, and cannot be embedded in image processing algorithms for optimization purposes. It is thus the goal of objective QA algorithms to automatically predict the quality of images as perceived by humans. Significant progress has been made in the last two decades in the design of objective QA methods and three major frameworks are now well-established in IQA research [1], [2]: 1) Full-Reference (FR) IQA, 2) Reduced-Reference (RR) IQA, and 3) No-Reference (NR) or Blind IQA. To evaluate the quality of a distorted image, FR methods require the complete availability of its pristine quality version termed as a reference image, while RR methods require access to certain features that have been extracted from the reference image. On the other hand, NR methods evaluate the quality of the distorted image in the absence of the reference image.

Since the beginning of this century, with the availability of subject-rated datasets, a large number of IQA methods belonging to all three frameworks (FR, RR, NR) have been proposed. These methods are tested on one or more subject-rated datasets and claim state-of-the-art performance. Given the large number of IQA methods that now exist, a number of challenges arise when it comes to selecting the top performing methods within and across different IQA frameworks for various purposes: 1) It can be respectively seen from Tables 3, 4, and 5 that different FR, fused FR, and NR methods are tested on different sets of subject-rated

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

datasets, and thus straightforward performance comparison becomes difficult. 2) It is also evident from Tables 3, 4, and 5 that IQA methods are usually tested (and at times trained) on singly distorted subject-rated datasets that contain different distortion types, but typically, each distorted image has been afflicted with a single stage of distortion [3]–[10]. This is in contrast to real world media distribution systems where the same visual content can undergo a number of distortions, during the processes of acquisition, transmission, and storage, before reaching the end user. While some IQA datasets with multiply distorted images are now available [11]–[14], only a limited number of IQA methods have used some of them for testing purposes. 3) General-purpose NR methods, which either rely on handcrafted features or on end-to-end learning, require training which is usually done on subject-rated IQA datasets where MOS acts as ground truth. While such training requires the availability of a large amount of data, subject-rated datasets offer only a small amount of annotated data. For example, the largest well-known subject-rated singly distorted database has a total of 3000 distorted images [5], while there are only 1600 distorted images in the largest multiply distorted database [13]. The number of images in individual distortion categories is even smaller. Such constraints make it difficult to avoid model overfitting and raises questions about the generalizability of NR methods trained on these datasets (as will become evident later in this work). To circumvent these issues, large-scale annotated datasets are required that consist of thousands of pristine reference and hundreds of thousands if not millions of distorted images. These datasets should have a wide variety of distortions and distortion combinations along with appropriately selected distortion intensity levels that cover the entire range of the quality spectrum with adequate density. However, given the limitations of subjective testing, it is not possible to obtain quality ratings from humans for such large datasets. Clearly, alternative methods for annotating large-scale IQA datasets are desired. Since the area of FR IQA has matured quite well, one possible alternative is to replace subjective ratings with scores from reliable FR methods. In fact, a number of works in IQA literature have already used either FR scores [15]–[20] or fused FR scores [21] as replacement of subjective ratings. However, their choice of FR methods seems rather ad hoc as detailed analysis about method selection has not been provided. Essentially the following questions remain unanswered while using FR scores for annotating large-scale IQA datasets as alternatives to subjective ratings: i) Which FR method or methods should be selected? ii) Can fused FR methods offer any further advantages over individual methods?

To address the above-mentioned challenges, a comprehensive survey of the performance of IQA methods, especially FR and fused FR methods, is desired that gauges their performance on a large and diverse set of subject-rated IQA datasets. A number of reviews and surveys have been conducted in the field of IQA over the past decade or so. The performance of ten FR IQA methods was evaluated on the LIVE R2 database [22] in [3]. Performance evaluation criteria included the Pearson Linear Correlation Coefficient (PLCC), Root-Mean-Squared Error (RMSE), Spearman Rank-order Correlation Coefficient (SRCC), and statistical significance testing. A description of 111 FR IQA methods is given in [23], however performance evaluation was not carried out. A comprehensive review of basic computational building blocks used in the design of perceptual IQA metrics is given in [24] along with a description of six FR IQA methods. The performance of these methods is evaluated on seven IQA databases (A57 [7], CSIQ [6], IVC [10], LIVE R2 [3], MICT [9], TID2008 [4], and WIQ [8]) in terms of PLCC and SRCC. A classification, description, and evaluation of 22 FR methods is provided in [25], where PLCC and SRCC are used for performance evaluation on six datasets which include IVC [10], TID2008 [4], and four other datasets whose description can be found in [25]. In [26], the performance of 11 FR methods was evaluated on seven IQA datasets (A57 [7], CSIQ [6], IVC [10], LIVE R2 [3], MICT [9], TID2008 [4], and WIQ [8]). PLCC, RMSE, SRCC, and Kendall Rank-order Correlation Coefficient (KRCC) were used as evaluation criteria. The computational complexity of these methods was evaluated in terms of their running speed. Various aspects of subjective and objective IQA are surveyed in [27] including: description of four subjective testing methods, description of seven FR IQA methods for standard dynamic range (SDR) images, description of two FR methods for the IQA of reference and test images with different dynamic ranges, description of six IQA datasets, and performance evaluation of seven SDR FR IQA methods on three datasets (CSIQ [6], LIVE R2 [3], TID2008 [4]) in terms of PLCC, SRCC, KRCC, RMSE, and Mean Absolute Error (MAE). In addition, the performance of an FR method for the IQA of tone mapped images (TMQI [28]) is evaluated, the computation time of different FR methods is presented, and the IQA of three-dimensional images is discussed. In [29], several objective IQA methods along with seven datasets are briefly discussed, and the performance of eight FR, three RR, and eight NR methods is evaluated on the LIVE R2 database [3] in terms of PLCC, SRCC, RMSE, and MAE. In [30], the performance of 60 FR methods was evaluated on the CIDIQ database [31] which provides subjective ratings at two viewing distances. PLCC, SRCC, and KRCC were used as performance evaluation criteria. A survey of Natural Scene Statistics (NSS) and learning based non-distortion-specific (general-purpose) NR IQA methods was performed in [32], where the design of 12 NR methods was reviewed and the performance of nine such methods was evaluated on three IQA databases (LIVE R2 [3], CSIQ [6], and TID2008 [4]). PLCC, SRCC, and statistical significance testing (only on LIVE R2 database) were used as performance evaluation criteria. For comparison, four FR methods are included in the performance evaluation. The computational complexity of six NR methods was also compared. Several distortion-specific and general-purpose NR IQA approaches were reviewed in [33], along with the performance evaluation

TABLE 1. Summary of IQA performance evaluation surveys.

Survey	Year	Number of Methods Evaluated	Number of Databases Used		Statistical Significance Testing
			SDB[a]	MDB[b]	
Ref [3]	2006	10 FR	1	0	Yes
Ref [23]	2009	Description of 111 FR Methods			No
Ref [24]	2011	6 FR	7	0	No
Ref [25]	2012	22 FR	6	0	No
Ref [26]	2012	11 FR	7	0	No
Ref [27]	2014	7 FR	3	0	No
Ref [29]	2014	19 (8 FR, 3 RR, 8 NR)	1	0	No
Ref [30]	2015	60 FR	1	0	No
Ref [32]	2015	13 (9 NR, 4 FR)	3	0	Yes
Ref [33]	2017	8 NR	3	0	No
Ref [34]	2019	36 (20 FR, 10 NR, 5 RR, 1 Fused FR)	4	0	No
This work	2019	64[c] (43 FR, 14 NR, 7[c] Fused FR)	5	4	Yes

[a]SDB: Singly Distorted Databases (Images afflicted with one distortion at a time).
[a]MDB: Multiply Distorted Databases (Images afflicted with multiple distortions at the same time).
[c]22 versions of the seven Fused FR methods were tested, which if taken into account separately means that we evaluated 79 methods.

of eight NR methods on three datasets (CSIQ [6], LIVE R2 [3], TID2013 [5]) in terms of PLCC and SRCC. The computational complexity of these methods was determined in terms of their execution time. In a recent survey [34], different areas of IQA are reviewed including two-dimensional (2D) image fidelity assessment (FR, RR, NR), three-dimensional (3D) image fidelity assessment (FR, NR), image aesthetics assessment, and 3D image visual comfort assessment. In the category of 2D image fidelity assessment, the performance of 20 FR, one fused FR, five RR, and 10 NR IQA methods is evaluated on four datasets (CSIQ [6], LIVE R2 [3], TID2008 [4], TID2013 [5]) in terms of PLCC, SRCC and RMSE. A summary of these earlier IQA reviews and surveys is given in Table 1.

Existing IQA surveys suffer from a number of shortcomings: 1) The earlier ones [3], [23]–[25] do not include state-of-the-art FR methods. 2) While conducting performance evaluation, none of these surveys utilize multiply distorted IQA datasets (in some cases this is because such datasets did not exist at the time of the survey). This puts into question the assumptions made about algorithm performance while being tested on limited data (singly distorted datasets only). 3) With the exception of [30], some recent singly distorted datasets (VCLFER [35], CIDIQ [31]) are missing in these surveys. 4) Some surveys use a single dataset [3], [29], [30], which limits content diversity and raises concerns about the generalization of their findings. 5) None of the surveys evaluates the performance of fused FR methods with the exception of [34] which evaluates only a single FR fusion method. 6) Some surveys [32], [33] are specific to the evaluation of NR methods. 7) With the exception of [3], [32], statistical significance testing is missing in these surveys. Since IQA datasets can only be regarded as small and sparse random samples from the space of all possible natural images and their distorted versions, the lack of such testing puts into question the universal nature of the findings in these surveys. 8) Although the survey in [34] is quite recent, it does not evaluate the performance of IQA methods on multiply distorted datasets, does not use the singly distorted datasets VCLFER [35] and CIDIQ [31], does not perform statistical significance testing, evaluates only a single fused FR method, and does not evaluate the performance of some state-of-the-art FR and NR IQA methods. Reference [34] uses both TID2008 [4] and TID2013 [5] datasets, where the latter contains all the reference and distorted images of the former. Given these shortcomings, it is evident that existing surveys are unable to identify the top performing FR, fused FR, and NR methods in a competitive and comparative setting. They are also unable to answer the question about the choice of FR or fused FR methods as alternatives to subjective ratings.

In this work, we attempt to address the limitations of existing IQA surveys by carrying out a comprehensive review and performance evaluation of 64 IQA methods, of which 43 are FR and seven are fused FR methods. We also include 14 NR methods in our study to provide a more thorough snapshot of the field. We tested 22 versions of the seven fused FR methods, and thus collectively a total of 79 IQA methods were evaluated. We test on nine subject-rated datasets, of which five are singly distorted and four are multiply distorted datasets. This ensures that the methods under evaluation are tested on as wide a range of reference and distorted content as possible. Apart from the usual correlation coefficient based comparison criteria, we also compare IQA methods through statistical significance testing in order to make statistically sound conclusions. To the best of our knowledge, this is the largest evaluation study carried out in IQA literature. In addition to FR and NR IQA that are surveyed and evaluated in this paper, there are other types of IQA problems such as reduced-reference (RR) IQA [1], [2], and IQA of reference/test images across different spatial resolutions [36], frame rates [37], [38], dynamic ranges [28], exposure levels [39], focus points [40], color/gray tones [41], and viewing devices [42], that are beyond the major focus of the current work.

The rest of the paper is organized as follows. A review of IQA datasets and methods included in this study is provided in Sections II and III respectively. The performance of FR and fused FR IQA methods is thoroughly evaluated in Section IV while that of NR methods is evaluated in Section V. Section VI concludes this work.

II. REVIEW OF IQA DATABASES

Over the last 15 years, a significant number of IQA databases with human rated image quality ratings have come out. Although recommendations have been made about the conduct of subjective testing and content selection [51]–[53], a *gold standard* remains elusive and the optimal method for subjective testing is still an open problem. As is evident from Table 2 and the following sections, IQA datasets use a variety of subjective testing methodologies, viewing distances, and ratings per image. Their benchmark quality ratings have different ranges and are either in the form of Difference Mean Opinion Scores (DMOS) or Mean Opinion Scores (MOS).

TABLE 2. Summary of IQA databases used in this work.

Database	Year	No. of Images Ref.	No. of Images Dist.	Distortion List (No. of Images)	Distortions per Image	Subjective Test Method	Subjective Data Type	Score Range	Ratings per Image	Viewing Distance
LIVE R2 [3], [22]	2006	29	779	1. White Gaussian Noise (145) 2. Gaussian Blur (145) 3. JPEG Compression (175) 4. JPEG2000 Compression (169) 5. Fast Fading Rayleigh Channel (145)	1	Single Stimulus	DMOS	-2.64 to 111.77	≈ 23	2 - 2.5 Screen Heights
TID2013 [5], [43]	2013	25	3000	1. Additive Gaussian Noise (125) 2. Additive Noise is more intensive in color components (125) 3. Spatially Correlated Noise (125) 4. Masked Noise (125) 5. High Frequency Noise (125) 6. Impulse Noise (125) 7. Quantization Noise (125) 8. Gaussian Blur (125) 9. Image Denoising (125) 10. JPEG Compression (125) 11. JPEG2000 Compression (125) 12. JPEG Transmission Errors (125) 13. JPEG2000 Transmission Errors (125) 14. Non Eccentricity Pattern Noise (125) 15. Local Block-wise Distortions of different intensity (125) 16. Mean Shift (Intensity Shift) (125) 17. Contrast Change (125) 18. Change of Color Saturation (125) 19. Multiplicative Gaussian Noise (125) 20. Comfort Noise (125) 21. Lossy Compression of Noisy Images (125) 22. Image Color Quantization with Dither (125) 23. Chromatic Aberrations (125) 24. Sparse Sampling and Reconstruction (125)	1 to 2	Pair-wise Comparison	MOS	0.24 to 7.21	≈ 30	Varying
CSIQ [6], [44]	2010	30	866	1. Additive White Gaussian Noise (150) 2. Gaussian Blur (150) 3. JPEG Compression (150) 4. JPEG2000 Compression (150) 5. Additive Pink Gaussian Noise (150) 6. Global Contrast Decrements (116)	1	Simultaneous Comparison	DMOS	0 to 1	≈ 6	70 cm
VCLFER [35], [45]	2012	23	552	1. Additive White Gaussian Noise (138) 2. Gaussian Blur (138) 3. JPEG Compression (138) 4. JPEG2000 Compression (138)	1	Single Stimulus	MOS	1.57 to 96.52	16 to 36	INP*
CIDIQ [31], [46]	2014	23	690	1. Poisson Noise (115) 2. Gaussian Blur (115) 3. JPEG Compression (115) 4. JPEG2000 Compression (115) 5. SGCK Gamut Mapping (115) 6. ΔE Gamut Mapping (115)	1	Double Stimulus	MOS	1.18 to 7.65 1 to 7.76	17	50 cm 100 cm
LIVE MD [11], [47]	2012	15	405	1. Gaussian Noise (45) 2. Gaussian Blur (45) 3. JPEG Compression (45) 4. Gaussian Blur followed by JPEG compression (135) 5. Gaussian Blur followed by Gaussian Noise (135)	1 1 1 2 2	Single Stimulus	DMOS	0.61 to 84.67	≈ 19	4 Screen Heights
MDID2013 [12]	2014	12	324	1. Gaussian Blur followed by JPEG compression followed by White Gaussian Noise (324)	3	Single Stimulus	DMOS	0.32 to 0.55	25	4 Image Heights
MDID [13], [48]	2017	20	1600	May include (Gaussian blur and/or contrast change) followed by (JPEG or JPEG2000 compression) followed by (Gaussian noise)	1 to 4	Pair Comparison Sorting	MOS	0.08 to 7.92	33 to 35	2 Screen Heights
MDIVL [14], [49] [50]	2017	10	750	1. Gaussian Blur followed by JPEG Compression (350) 2. Gaussian Noise followed by JPEG Compression (400)	2	Single Stimulus	MOS	1.41 to 97.97	≈ 12	INP*

*INP: Information Not Provided by authors.

Reference image content is usually selected in an ad hoc manner and different distortions are simulated by degrading the reference content at different distortion intensity levels which are themselves picked in an ad hoc manner. While the target is to have distorted images such that the quality spectrum is uniformly represented, this is often not the case (as discussed later). A majority of IQA datasets consider the simplified case of images undergoing a single distortion which is in contradiction to practical scenarios where content typically undergoes multiple distortions. Given the arbitrary nature of such benchmark data, it is unsurprising that at times the performance of IQA methods varies widely across

different datasets. Thus, it is vital to test the performance of IQA methods on as many publicly available datasets as possible [54] in order to reliably test their robustness.

To mitigate dataset specific impacts on the performance evaluation of IQA methods, in this work we choose a large number of databases to carry out such an assessment. We use four database selection criteria, specifically we use databases that contain: 1) Natural images, 2) Color images, 3) Both reference and distorted content to enable evaluation of FR IQA methods, and 4) Standard Dynamic Range (SDR) images, that is, images with a bit depth of 8 bits per pixel per color channel. Following these criteria, we have selected nine databases which simulate distortions at various intensity levels. Five of these datasets can be classified as singly distorted databases while four fall under the multiply distorted category. Table 2 presents a summary of these databases while they are briefly introduced in the next two sub-sections. This is followed by a description of some other IQA databases and the reasons for not including them in our current work. We close this section by a discussion on the range of reference and distorted content in the datasets used in this work for algorithm testing.

A. SINGLE DISTORTION DATABASES

These datasets are also referred to as *singly distorted* databases. While they contain a wide range of distortions, each distorted image is afflicted with only one kind of distortion. Until recently, a majority of IQA datasets fell under this category.

The LIVE Release 2 (LIVE R2) database [3], [22], developed by the Laboratory for Image and Video Engineering at UT Austin, is one of the most widely used IQA datasets. It consists of 29 reference and 779 distorted images. The database has five distortion types and up to five distortion intensity levels within each type. Images either have a resolution of 480 × 720 or up to 768 × 512. Subjective testing was carried out on 21″ CRT monitors and followed the single stimulus methodology [51] where reference images were also evaluated. After undergoing a short training session, subjects rated the quality of test images by moving a slider on a quality scale that was demarcated with five words: Bad, Poor, Fair, Good, and Excellent. A quality score in the range of [1, 100] was obtained from the slider location. Seven sessions of testing were done in order to minimize observer fatigue and scale realignment was carried out to match the quality scale of all sessions. The database provides subjective data in the form of DMOS after outlier removal, where better quality is represented by a lower DMOS. Further details about the database are provided in Table 2.

The Tampere Image Database 2013 (TID2013) [5], [43] builds further upon the earlier TID2008 database [4]. It consists of 25 reference images (of which 24 are natural and one is artificial) and 3000 distorted images. The database has 24 distortion types and five distortion levels per type. All images have a resolution of 512×384. A total of 971 subjects in five different countries took part in subjective testing. Experiments were carried out either in the laboratory environment or remotely via internet, and subjects were given prior instructions about the testing process. A tristimulus methodology [5] was adopted to conduct the subjective tests where subjects observe a pair of distorted images in the presence of their reference image and select the better of the two. Tests were conducted mostly on 19″ LCD or CRT monitors. Each distorted image was part of nine pair-wise comparisons. The winning image in each pair received one point and a final score for an image was obtained by summing the winning points. After outlier removal, MOS was obtained for the database, where higher MOS represents better quality. Although we are classifying TID2013 under the single distortion category, it should be noted that some of its distortion types are multiply distorted in nature (for example, lossy compression of noisy images). See Table 2 for more details.

The Computational and Subjective Image Quality (CSIQ) database [6], [44] consists of 30 reference and 866 distorted images. It has six distortion types and four to five levels of distortion per type. All images have a resolution of 512×512. Subjective tests were carried out by placing four 24″ LCD monitors side-by-side such that their viewing distance from the subject was equal. All the distorted images derived from the same reference were simultaneously displayed on the monitor array and each subject horizontally ordered images based on their perceived quality [44]. Cross-image ratings were obtained in order to carry out realignment of the quality scale between different content. After outlier removal DMOS was obtained, where a lower DMOS value represents better quality. Further details about the database are provided in Table 2.

The Video Communications Laboratory @ FER (VCLFER) database [35], [45] is composed of 23 reference and 552 distorted images. It has four distortion types and six distortion levels per type. Images in VCLFER either have a resolution of up to 771 × 512 or up to 512 × 771. Subjective testing was conducted by following the single stimulus methodology [51] and by employing a numeric scale with 100 grades. After removing outliers, the results for each subject were rescaled in the range of [0, 100], and MOS for the overall database was computed. A higher MOS value is indicative of better visual quality. See Table 2 for more details.

The Colourlab Image Database: Image Quality (CIDIQ) [31], [46] consists of 23 reference and 690 distorted images. It has six distortion types and five distortion levels per type. All images in CIDIQ have a resolution of 800 × 800. Subjective testing was carried out in accordance with the recommendations of CIE [55] and ITU [51]. A double stimulus methodology was followed where two images were displayed simultaneously, and category judgment was used to record responses from subjects. The rating scale had nine categories where the odd numbered categories from 1 to 9 were respectively labeled as Bad, Poor, Fair, Good, and Excellent quality. The actual subjective test was preceded by a training sequence. The CIDIQ database is unique in that it carried out subjective testing at two viewing distances, that of 50 cm

and 100 cm. Therefore, it provides two sets of MOS, one for each viewing distance. A higher MOS value represents better visual quality. Further details about the database are provided in Table 2.

B. MULTIPLE DISTORTION DATABASES

These datasets are also referred to as *multiply distorted* databases and contain images such that an individual distorted image may have undergone multiple (two or more) distortions, thereby better mimicking practical content distribution scenarios.

The LIVE Multiply Distorted (LIVE MD) database [11], [47] is the first IQA dataset that has been specifically designed for images with multiple simultaneous distortions. The database has 15 reference and 405 distorted images of which 135 are singly distorted while 270 are multiply distorted. LIVE MD has three distortion types (Gaussian blur, JPEG compression, and white Gaussian noise) and three distortion levels per type. Apart from containing singly distorted images belonging to each of the three distortion types, the database has two multiple distortion combinations of 1) Gaussian blur followed by JPEG compression and 2) Gaussian blur followed by white Gaussian noise contamination. All images in the database have a resolution of 1280×720. Subjective testing was conducted by following the single stimulus [51] with hidden reference methodology. After going through a training session, subjects rated the quality of test images by moving a slider on a continuous scale from 0 to 100 which was also labeled with the words, Bad, Poor, Fair, Good, and Excellent. The test was divided into two parts based on the multiple distortion combinations and each part had two sessions of 30 minutes each. The database provides subjective scores in the form of DMOS, where a lower value is indicative of better visual quality. Further details about the database are provided in Table 2.

The Multiply Distorted Image Database 2013 (MDID2013) [12] is composed of 12 reference and 324 multiply distorted images. The database uses the same distortion parameters as the LIVE MD database [11]. MDID2013 uses three distortion types (Gaussian blur, JPEG compression, and white Gaussian noise) and three distortion levels per type. It contains just one multiple distortion combination, where a reference image first undergoes Gaussian blurring which is followed by JPEG compression followed by white noise contamination. Images in MDID2013 have a resolution of up to 1280×720. The single stimulus methodology [51] was followed to conduct the subjective test and ratings were obtained on a continuous quality scale from 0 to 1. After outlier removal, DMOS for the database was computed where a lower value signifies better visual quality. See Table 2 for more details.

The Multiply Distorted Image Database (MDID) [13], [48] (different from MDID2013) contains 20 reference and 1600 multiply distorted images. The database uses five types of distortions: Gaussian noise, Gaussian blur, contrast change, JPEG, and JPEG2000 compression. Four intensity levels are set for each distortion type. Distortions are introduced in three steps in the following order: 1) To simulate image acquisition, Gaussian blur and/or contrast change are added first in either order. 2) Image transmission is simulated by compressing the image from the first step, either by using JPEG or JPEG2000 compression (one compression technique only). 3) Finally, display imperfections are simulated by adding Gaussian noise to the image from the second step. In each of these steps, distortion intensity levels, including the no-distortion case, are picked at random. However, it is ensured that the following three rules are obeyed: 1) At least one distortion is introduced, 2) Only one compression technique (JPEG or JPEG2000) is used, and 3) Repetition of distortions is avoided. Thus, each distorted image may be afflicted with one to four distortions. MDID creates 80 distorted images for each reference image and provides details about the distortion process for each image. All images in MDID have a resolution of 512×384. The pair comparison sorting methodology [13] is used to conduct subjective testing, where two images are simultaneously displayed along with their reference and subjects are required to rate the quality of one distorted image with respect to the other by using one of three possible rating options: better, worse, or equal quality. Testing was carried out on a $19''$ LCD monitor and was preceded by a training session. Following outlier removal and data normalization, MOS for the database is computed, where a higher value is indicative of better visual quality. Further details about the database are provided in Table 2.

The Multiple Distorted IVL database (MDIVL) [14], [49], [50] consists of 10 reference and 750 multiply distorted images. The database is divided into two parts based on two multiple distortion combinations: 1) Blur-JPEG, where each reference image undergoes seven levels of Gaussian blur and then each blurred image undergoes five levels of JPEG compression, and 2) Noise-JPEG, where each reference image undergoes ten levels of Gaussian noise and then each noisy image undergoes four levels of JPEG compression. All images in the database have a resolution of 886×591. Subjective testing followed the single stimulus methodology [51]. Subjects recorded their ratings on a continuous quality scale from 0 (Worst quality) to 100 (Best quality). To minimize fatigue effect, subjective testing was conducted in several sessions where each session had around 100 images and did not exceed 30 minutes. MOS was computed for the database after outlier removal, where a higher value indicates better quality. See Table 2 for more details.

C. OTHER IQA DATABASES

Apart from the nine datasets mentioned in Sections II-A and II-B, a number of other datasets have been mentioned in Tables 3, 4 and 5 that follow in the subsequent sections. Information about these and some other datasets follows.

The A57 database [7], [56] contains three reference and 54 distorted images. It consists of grayscale images with a resolution of 512×512. The dataset has six distortion types which include: 1) Gaussian white noise, 2) Gaussian blur, 3) Baseline JPEG compression, 4) Baseline JPEG2000

compression, 5) JPEG2000 compression with dynamic contrast based quantization, and 6) Flat allocation (equal distortion contrast at all scales). Each distortion was applied at three distortion intensity levels. The MICT-Toyama database [9] contains 14 reference and 168 distorted images. It consists of color images with a resolution of 768 × 512. The dataset contains two distortion types: 1) JPEG compression and 2) JPEG2000 compression, and six distortion levels per type. The single stimulus methodology was used to acquire subjective ratings, using a five category discrete quality scale and through the participation of 16 subjects, on a 17″ CRT display at a viewing distance of four times the picture height. The IVC database [10] contains ten reference and 185 distorted images. The dataset consists of color images with a resolution of 512 × 512. It has four distortion types which include: 1) JPEG compression, 2) JPEG2000 compression, 3) Local adaptive resolution (LAR) coding, and 4) Blurring. The subjective ratings for IVC were obtained by following the double stimulus methodology with five rating categories. 15 observers participated in the test and viewed the content at a distance of six times the screen height. The TID2008 database [4], [57] is an earlier version of the TID2013 database [5], and contains 25 reference and 1700 distorted images. It has color images with a resolution of 512×384. The dataset contains 17 distortion types and four distortion levels per type. For a list of distortions contained in TID2008, refer to the first 17 distortions listed in Table 2 for the TID2013 database. Subjective testing for TID2008 was carried out by using the same methodology as was later used for TID2013 (described in Section II-A). The Wireless Imaging Quality (WIQ) database [8] contains seven reference and 80 distorted images. It consists of grayscale images with a resolution of 512 × 512. The dataset simulates a wireless link distortion model by passing JPEG encoded images through an uncorrelated Rayleigh flat fading channel in the presence of additive white Gaussian noise. Two subjective tests were performed at different locations, on 17″ CRT monitors at a viewing distance of four times the picture height. The double stimulus continuous quality scale (DSCQS) [51] methodology was followed to conduct the tests. The Waterloo Exploration database [58] is a very large dataset that is composed of 4,744 reference and 94,880 distorted images. It has color images of various resolutions. The dataset contains four distortion types: 1) White Gaussian noise, 2) Gaussian blur, 3) JPEG compression, and 4) JPEG2000 compression. Each distortion is applied at five fixed intensity levels. Since the database consists of such a large number of images, subjective testing is not possible. Instead, three alternative testing criteria are proposed in [58] for the performance evaluation of objective IQA models. These include: 1) the pristine/distorted image discriminability test (D-Test), 2) the listwise ranking consistency test (L-Test), and 3) the pairwise preference consistency test (P-Test).

The above-mentioned datasets have not been used in the current work for the following reasons. The A57 and WIQ datasets are composed of grayscale images which does not fulfill one of our database selection conditions, that a dataset should be composed of color images. This condition is required to provide a uniform comparison basis, as some of the objective IQA methods that we test are designed to take the color aspect into account. Besides, these datasets are composed of only a small amount of source and distorted content. The MICT-Toyama dataset has not been selected as 11 of its 14 reference images are found in LIVE R2 dataset while the cropped versions of all its reference images are found in the TID2013 reference image set. Both LIVE R2 and TID2013 datasets contain the two distortion types found in MICT-Toyama. Since we are including LIVE R2 and TID2013 in our analysis, we believe that including MICT-Toyama would be redundant. The TID2008 dataset has not been included since all of its reference content, distortion types and levels are found in its enhanced version TID2013. The IVC dataset contains a small number of test images per distortion type and three of its four distortion types (Blur, JPEG and JPEG2000 compression) are effectively covered in the five single distortion databases that we have selected for testing. Although the Waterloo Exploration database is one of the largest available IQA datasets, we have not used it because of the unavailability of subjective ratings.

The databases discussed above, and in Sections II-A and II-B, belong to the category of *simulated distortion* databases, where a number of pristine reference images are first obtained and then artificially degraded with different types and levels of distortions in a controlled manner. By contrast, *authentic distortion* databases constitute another category of IQA datasets, where distortions are captured directly in real-world environments. It is difficult to categorize images into different distortion types and intensity levels in such datasets. The following four databases fall in the authentic distortion category. The Blurred Image Database (BID) [59] consists of 585 images, taken by human users, that represent realistic blur distortions. Images are classified into five blur classes which include unblurred images, out-of-focus blur, simple motion blur, complex motion blur, and other kinds of blur. Subjective testing was carried out by using a single stimulus methodology on a continuous quality scale marked with labels (Excellent, Good, Fair, Poor, and Bad). The Camera Image Database 2013 (CID2013) [60] consists of 480 images captured by 79 different cameras of varying quality. Different types of cameras were used to capture images, including mobile phone cameras, compact cameras and SLR cameras. The database is divided into six smaller datasets each of which is composed of six different scenes that have been captured by 12-14 different cameras. A dynamic reference method [60] was proposed and used to conduct the subjective test. The subjects first saw a slideshow of the test images to get an overall idea of quality variation, and then saw each image in a single stimulus manner where they could give quality ratings on a continuous scale. Besides MOS, subjective evaluations for the attributes of sharpness, graininess, brightness, and color saturation are also provided. The LIVE in the Wild Image Quality Challenge (LIVE WC)

database [61] is composed of 1162 images taken by a diverse set of mobile device cameras. The images in this dataset depict a wide variety of real-world scenes. The subjective study was performed online by using the Amazon Mechanical Turk [62], which is a crowdsourcing platform. The single stimulus methodology was employed where subjects recorded their quality ratings on a continuous scale that was divided into five parts with appropriate labels (Excellent, Good, Fair, Poor, and Bad). Besides the subjective test to provide MOS, a separate experiment was conducted to obtain subjective opinion about the distortion category that a test image may belong to. Distortion categories included blurry, grainy, overexposed, underexposed, and no apparent distortion. A majority voting policy was adopted to arrive at a distortion category for an image. A recent database called KonIQ-10K [63] consists of 10,073 images and is by far the largest among the authentic distortion databases. The source of the KonIQ-10K images is the very large-scale YFCC100M multimedia database [64] which has 100 million Flickr based media objects (images and videos). Initially 10 million images were randomly picked from the YFCC100M database from which 10,073 authentically distorted images were sampled through the use of content and quality based indicators. The subjective study was carried out online through a crowdsourcing platform [65]. A five-point absolute category rating (ACR) scale was used to obtain subject ratings, where a rating of 1 indicated bad while that of 5 indicated excellent quality. The database provides subjective ratings in terms of MOS. In this work, we have not used authentic distortion datasets because they lack the presence of reference images, which renders them unusable for the evaluation of FR IQA methods. Nevertheless, these datasets are a valuable resource and should be used in studies that are exclusive to NR IQA methods.

A number of datasets composed of content other than natural images have been constructed. The Screen Image Quality Assessment Database (SIQAD) [66] consists of 20 reference and 980 distorted screen content images. It follows the single stimulus methodology to obtain subjective scores on an 11 point numerical scale. The Document Image Quality dataset [67] selected 25 documents from publicly available document datasets and used a smart phone camera at varying distances to capture 175 document images. The dataset provides Optical Character Recognition (OCR) accuracy as a measure of quality that has to be predicted by objective methods. The Newspaper dataset [68] is composed of 521 grayscale text zone images derived from a collection of newspaper images. As ground truth, the dataset provides OCR accuracy results. Since our focus is on natural images, we have not utilized these datasets in this work.

A valuable compilation of various image and video quality databases can be found at [69].

D. CONTENT ANALYSIS

The space of all possible natural images is enormous. Ideally, an IQA database should properly reflect the statistical distribution of natural image content, or contain diverse content type for a wide coverage. In practical IQA databases, however, the large natural image space is often represented by just a few source or reference images. From Table 2, it can be seen that subject-rated datasets usually have 10 to 30 reference images. Limitations on the amount of source content are encountered due to the constraints of subjective testing. For example, even with just 25 reference images, the TID2013 database [5] has 3000 distorted images, which leads to significant challenges in obtaining human ratings. The limited source content that a dataset has, should thus be as diverse as possible in order to sample different parts of the space of all possible natural images. This is also an important reason for selecting as many subject-rated IQA databases as possible while testing a new algorithm, so that its performance can be gauged on as wide a set of source content as possible.

Usually the variety in reference content is described in subjective terms, such as the presence of people, human faces, landscapes, animals, closeup or wide-angle shots, buildings, indoor or outdoor shots, and so on. However, a few quantitative descriptors have also been used to describe such content. In [70], image spatial information (SI) and colorfulness (CF) have been used to represent the dimensions of space and color respectively, and the SI versus CF space has been proposed as a two-dimensional space to represent the diversity of source content. In this work, we use the SI versus CF space to examine the range of source content in the nine IQA datasets under consideration.

SI is used to determine edge energy in an image [70]. Different SI measures have been found to have high correlation with compression based image complexity measures [71]. A standard deviation based SI measure (SI_{std}) was recommended in [52] while a root mean square based measure (SI_{rms}) was used in [70]. However in [71], SI_{std}, SI_{rms}, and a mean based SI measure (SI_{mean}) were compared and it was found that SI_{mean} has the highest correlation with compression based image complexity measures. Therefore, we will use SI_{mean} for further analysis in this work. To obtain SI_{mean}, a color image is first converted to grayscale and then filtered with horizontal $\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$ and vertical $\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$ Sobel filters, leading to images s_h and s_v respectively. The pixel-level edge magnitude is then defined as [70], [71]:

$$s_{mag} = \sqrt{s_h^2 + s_v^2} \tag{1}$$

And SI_{mean} is obtained as [71]:

$$\mathrm{SI}_{mean} = \frac{1}{N} \sum s_{mag} \tag{2}$$

where N is the number of pixels in the image.

CF is an indicator of the variety and intensity of colors in an image [70]. A computationally efficient CF measure was proposed in [72] which correlates well with subjective measurements of colorfulness. Assuming an image in the

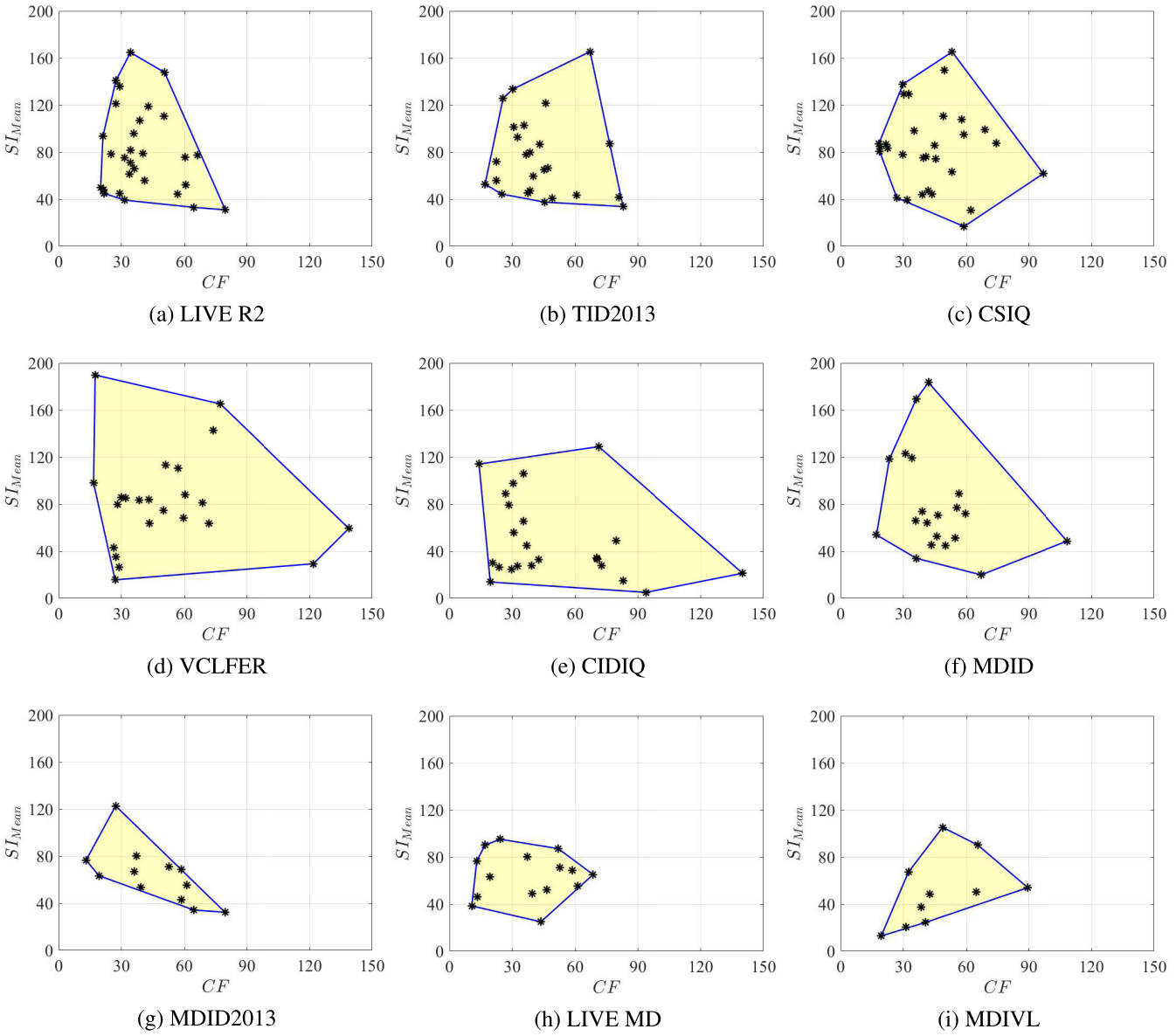

FIGURE 1. Spatial information (SI_{Mean}) versus colorfulness (CF) plots of the reference images belonging to the nine databases being used for method performance evaluation in this work. The blue lines represent the convex hull in each case.

sRGB color space, it is first transformed to an opponent color space as follows [72]:

$$rg = R - G \tag{3}$$

$$yb = \frac{1}{2}(R + G) - B \tag{4}$$

Then CF is defined as:

$$\mathrm{CF} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3 \cdot \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{5}$$

where σ_{rg} and σ_{yb} are the standard deviations, while μ_{rg} and μ_{yb} are the mean values, in the rg and yb directions respectively.

We computed the SI and CF values of all reference images in the nine IQA databases by using the definitions given in (2) and (5) respectively. The SI versus CF plots for these databases are given in Fig. 1 where the blue outer boundary marks the convex hull in each case and the area inside is marked yellow. For convenience, we have used the same scale for each axis in all the plots of Fig. 1. It is evident that the source content in these datasets occupies different regions in the SI versus CF space. While the VCLFER [35] and CIDIQ [31] datasets seem to cover the most area in this space, a majority of their images are clustered in smaller regions. On the other hand, the content in the LIVE R2 [3] and CSIQ [6] datasets is more uniformly distributed inside their respective convex hulls. Among the multiply distorted datasets, MDID [13] appears to have a wider coverage region while the other datasets in this category seem to have a limited range of source content. Apart from such subjective analysis of the SI versus CF coverage of datasets, efforts have been

made to quantify this coverage as well. A two-dimensional criteria called the *relative total coverage* (RTC) was defined in [70] as the square root of the area of the convex hull of all points in the normalized SI versus CF space. One drawback of using RTC as a coverage metric is that it does not take into account empty spaces within the convex hull. Thus, a single image that is located further away from the rest of the content in the SI versus CF space can lead to elevated RTC values giving a false sense of better coverage. To address this issue, another metric called *total effective coverage* (TEF) was proposed in [73] which builds upon the RTC concept. TEF introduces a fill rate factor to weigh the RTC value obtained for a dataset. A circle of certain radius r is considered around each image point in the SI versus CF space, within which a *presence* parameter p is considered as 1. The fill rate factor is then determined as a ratio of the area inside the convex hull where $p = 1$ to the area of the entire convex hull. By using a hypothetical database, it is demonstrated in [73] that TEF is a more effective coverage metric than RTC. Apart from the MDID2013 dataset, the RTC and TEF analysis for the eight other datasets can be found in [73] (it should be noted that the root mean square definition of SI is used in [73]).

E. DISTORTION ANALYSIS

In addition to wide content coverage, another important property of an ideal IQA database is diversity in terms of distortion types and levels. For a complete list of the types of distortions included in the nine IQA databases under consideration refer to Table 2, where this information is provided along with the number of images in each distortion type. While creating distorted content, the goal should be to simulate varying degrees of distortions such that the perceptual quality scale is uniformly sampled. This will ensure that objective IQA methods are tested across the quality spectrum. To accomplish this, IQA databases include different intensity levels for each distortion type. This information is provided in Sections II-A and II-B for the datasets under consideration. To ascertain the range of distortions in each database, the histograms of their subjective ratings (MOS or DMOS) are plotted in Fig. 2. A higher MOS value represents better visual quality while the opposite is true for DMOS where lower values signify better visual quality. The distribution of distorted content across the quality spectrum can be regarded as relatively uniform in MDID database [13] and mildly uniform in LIVE R2 [3], VCLFER [35], and CIDIQ (at viewing distance of 50 cm) [31] databases. On the other hand, TID2013 [5] and CSIQ [6] databases contain a relatively larger amount of better quality content while LIVE MD [11] and MDIVL [14] databases contain relatively more low quality content. It has been shown that objective IQA methods find it more difficult to evaluate better quality images as compared to low quality ones [5]. Thus, a dataset with a higher proportion of low quality content may not be as challenging as one with more better quality content. The impact of viewing distance on perceptual quality can be observed while comparing the MOS histogram of the CIDIQ database at a viewing distance of 50 cm (Fig. 2e) with the one obtained at 100 cm (Fig. 2f). While the distorted content remains the same in both cases, the presence of more higher quality ratings in the latter case demonstrates the challenge that objective IQA methods need to overcome

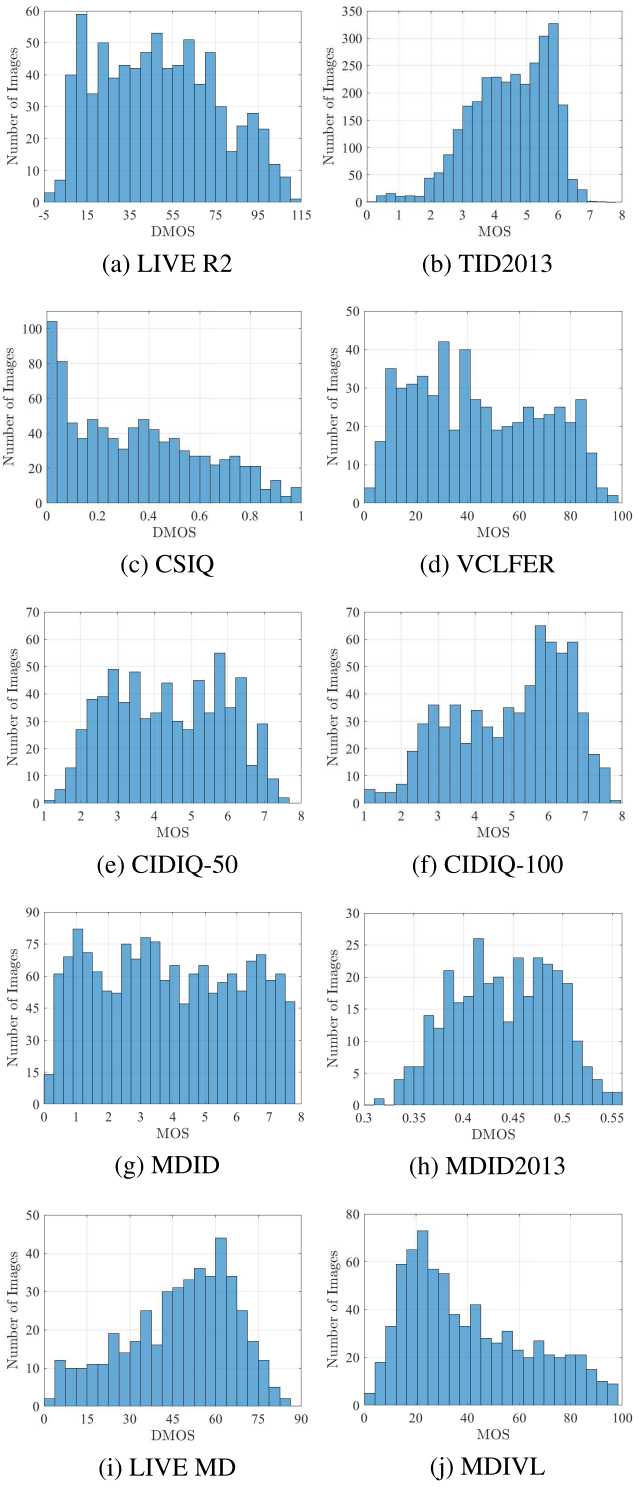

FIGURE 2. Histograms of MOS/DMOS of the nine IQA databases being used for method performance evaluation in this work. Note: The MOS of CIDIQ database has been obtained at two viewing distances of 50 cm and 100 cm [31].

and also highlights the importance of IQA databases which provide ratings at different viewing distances.

The non-uniform distribution of distorted content in most databases can be attributed to the way in which distortions are simulated. In all datasets being considered here, fixed parameters for each distortion type are used to simulate different levels of distortions across the dataset. While convenient, such an approach does not take into account the nature of source content and the masking effect that it can have upon different distortions. For example, the same compression ratio may lead to very different results when applied to images with different spatial information levels and the same amount of noise may appear quite different when applied to images that differ in texture characteristics. Thus, a reasonable alternative method is to simulate distortions in a content adaptive manner, that is, content specific distortion parameters should be found for each constituent image that roughly correspond to predefined perceptual quality levels. Nevertheless, in the current context, the variation of distorted content across the quality spectrum for different datasets provides one more reason to use as many databases as possible in the performance analysis of objective IQA methods.

While the histograms in Fig. 2 allow for observing the distribution of distorted content within each database, it is difficult to compare one dataset with another because they use different quality scales and subjective testing methods. To provide a unified, albeit weak [70], basis for comparing different datasets with each other, we compute the peak signal-to-noise ratio (PSNR) of all distorted images in each dataset and provide the corresponding boxplots in Fig. 3 where the range of distortions in different datasets can be compared. It can be observed that single distortion databases offer a wider range of distortion intensities while this range is quite limited in multiple distortion datasets. However, this comparison is weak because: 1) PSNR is not a perceptual metric [74], and 2) Even if the individual distortion intensities are wide-ranging in multiple distortion datasets, the interaction of one distortion with another may diminish the effect of the overall distortion, for example, JPEG compression of noisy images may have a denoising effect. The opposite is also true, and thus more research is needed to understand how multiple distortions interact with each other and with image content.

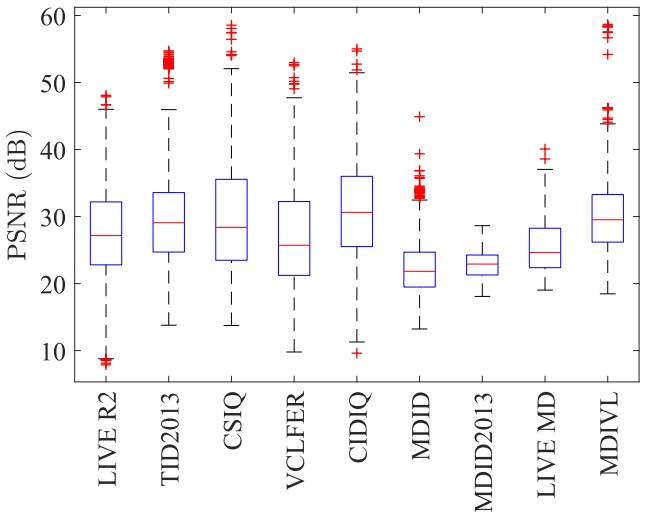

FIGURE 3. PSNR box plots for all databases. The top and bottom edges of the blue boxes represent the 75th and 25th percentiles, respectively, while the red line represents the median (50th percentile). The top and bottom black lines (whiskers) represent the extreme data points while the outliers are represented by red + symbols.

III. REVIEW OF IQA ALGORITHMS

Our focus in this work is to evaluate representative FR and NR IQA methods, designed for 2D natural images. We will also evaluate fusion based methods where the aim is to achieve better performance by combining results from multiple FR methods. We will provide a brief description of the design philosophies of the methods under consideration. As mentioned earlier, we have not evaluated the performance of RR and other types of IQA methods [75] in this work.

A. FULL-REFERENCE IMAGE QUALITY ASSESSMENT

Full-Reference (FR) IQA methods evaluate the quality of a distorted image with respect to the corresponding original (reference) image that is assumed to be distortion-free and of pristine quality [1]. In this work we evaluate the performance of 43 FR IQA methods which are listed in Table 3 along with information about whether a method operates on color or grayscale images, year of publication, and the number and names of the IQA databases that it was tested on. Although this list is not exhaustive, it is representative of different IQA design philosophies. The FR IQA methods being considered are reviewed next and are classified based on their design philosophies.

1) ERROR BASED METHODS

Historically, the *mean squared error* (MSE) and the related *peak signal-to-noise ratio* (PSNR) have been used as the "standard" quality measures [1]. Let $\mathbf{X} = \{x_i | i = 1, 2, ..., N\}$ and $\mathbf{Y} = \{y_i | i = 1, 2, ..., N\}$ represent the reference and distorted images respectively, where x_i and y_i represent the intensities of the i-th samples in the images $\mathbf{X}$ and $\mathbf{Y}$, respectively, and N is the number of image samples (pixels). MSE and PSNR are defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{6}$$

$$\text{PSNR} = 10\, log_{10}\frac{L^2}{\text{MSE}} \tag{7}$$

where L is the dynamic range of image pixel intensities. For gray-scale images with a bit depth of 8 bits/pixel, $L = 2^8 - 1 = 255$. The PSNR is similar to the *signal-to-noise ratio* (SNR) which is defined as:

$$\text{SNR} = 10\, log_{10}\frac{\frac{1}{N}\sum_{i=1}^{N} x_i^2}{\text{MSE}} \tag{8}$$

TABLE 3. Information about 43 FR IQA methods under performance evaluation.

FR Method	Color/Gray	Year	No. of Test Databases	Single Distortion Test Databases Used	Multiple Distortion Test Databases Used
AD_DWT [76]	Gray	2013	3	IVC LIVE R2 TID2008	None
ADM [77]	Gray	2011	5	CSIQ IVC LIVE R2 MICT TID2008	None
CID_MS [78] CID_SS [78]	Color	2013	2	TID2008 Images from six Gamut Mapping Datasets (See references [11], [41]-[44] of the CID Paper [78])	None
DSS [79]	Gray	2015	3	CSIQ LIVE R2 TID2008	None
DVICOM [80] DVICOM_F [80]	Gray	2018	3	CSIQ LIVE R2 TID2008	None
DWT_VIF [81]	Gray	2010	1	LIVE R2	None
ESSIM [82]	Gray	2013	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
FSIM [83]	Gray	2011	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
FSIMc [83]	Color	2011	5	CSIQ IVC LIVE R2 MICT TID2008	None
GMSD [84]	Gray	2014	3	CSIQ LIVE R2 TID2008	None
GSIM [85]	Gray	2012	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
IFC [86]	Gray	2005	1	LIVE R2	None
IW_PSNR [87]	Gray	2011	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
IWSSIM [87]	Gray	2011	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
MAD [6]	Gray	2010	4	CSIQ LIVE R2 MICT TID2008	None
MCSD [88]	Gray	2016	6	CSIQ IVC LIVE R2 MICT TID2008 TID2013	None
MSSSIM [89]	Gray	2003	1	Earlier Version of LIVE R2	None
NQM [90]	Gray	2000	—	Barbara, Boats, Lena, Mandrill, Peppers images and their distorted versions	None
PSNR	Gray	—	—	Legacy Method	
PSNR_DWT [76]	Gray	2013	3	IVC LIVE R2 TID2008	None
PSNR_HAc [91]	Color	2011	1	TID2008	None
PSNR_HA [91]	Gray	2011	1	TID2008	None
PNSR_HMAc [91]	Color	2011	1	TID2008	None
PSNR_HMA [91]	Gray	2011	1	TID2008	None
PSNR_HVS [92]	Gray	2006	—	Barbara, Lena images and their distorted versions	None
PSNR_HVSM [93]	Gray	2007	—	Test set composed of 19 images	None
QASD [94]	Color	2016	5	CSIQ IVC LIVE R2 TID2008 TID2013	None
RFSIM [95]	Gray	2010	1	TID2008	None
SFF [96]	Color	2013	5	CSIQ IVC LIVE R2 MICT TID2008	None
SNR	Gray	—	—	Legacy Method	
SRSIM [97]	Gray	2012	3	CSIQ LIVE R2 TID2008	None
SSIM [98]	Gray	2004	1	Earlier Version of LIVE R2	None
SSIM_DWT [76]	Gray	2013	3	IVC LIVE R2 TID2008	None
UQI [99]	Gray	2002	—	Lena image and its distorted versions	None
VIF [100]	Gray	2006	1	LIVE R2	None
VIF_DWT [76]	Gray	2013	3	IVC LIVE R2 TID2008	None
VIF_P [100], [101]	Gray	2005	—	Faster version of VIF, not tested in original paper [100]	None
VSI [102]	Color	2014	4	CSIQ LIVE R2 TID2008 TID2013	None
VSNR [56]	Gray	2007	1	LIVE R2	None
WSNR [90]	Gray	2000	—	Barbara, Boats, Lena, Mandrill, Peppers images and their distorted versions	None
WSSI [103]	Gray	2009	1	LIVE R2	None

The MSE has certain advantages [74] such as ease of use, clear physical meaning since it is the energy of the error signal and thus satisfying the Parseval's theorem, and ability to be used for algorithm optimization leading to closed-form solutions, etc. However, it has been repeatedly shown that MSE and PSNR have poor correlation with perceptual image quality, i.e., relative to subjective quality assessment by humans. This is because MSE-type of measures make the following underlying assumptions about perceptual image (and video) quality [74]: 1) It is independent of any spatial and temporal relationships between samples, 2) It is independent of the relationships between the image (and video) signals and error signals, 3) It is determined by the magnitude of the error signal only but ignoring the signs of errors, and 4) All signal samples are of equal importance. Unfortunately, not even one of these assumptions hold in the context of perceptual image (and video) quality assessment [1], [74]. It was also shown in [74] that images along the equal-MSE hypersphere have drastically different perceptual quality. Thus, the advantages of using signal-to-noise ratio based methods are negated by their shortcomings in the context of perceptual quality assessment.

To address the shortcomings of PSNR and SNR, several efforts have been made to modify these methods in order to make them perceptually better suited for IQA. In [90] the Contrast Sensitivity Function (CSF), which is used to approximate the behavior of the Human Visual System (HVS), was used to weigh the signal and noise powers, leading to a linear quality measure called Weighted SNR (WSNR). The Noise Quality Measure (NQM) was also presented in [90] and uses a nonlinear quasi-local processing model of the HVS to accomplish quality assessment. An HVS based version of PSNR, called PSNR-HVS, was proposed in [92] which uses the CSF. PSNR-HVS was modified by incorporating a model that takes into account the between-coefficient contrast masking of Discrete Cosine Transform (DCT) basis functions leading to a new method called PSNR-HVSM [93]. PSNR-HVS and PSNR-HVSM were further modified by incorporating human perception of contrast and mean brightness distortions, leading to modified methods called

PSNR-HA and PSNR-HMA respectively [91]. To deal with color images, PSNR-HA and PSNR-HMA were applied separately to each component of YCbCr transformed images and the results were combined into a quality score, leading to PSNR-HAc and PSNR-HMAc respectively [91]. The Visual Signal-to-Noise Ratio (VSNR) [56] is another HVS based method, which uses wavelet based models of visual masking and visual summation to first ascertain if the distortions are beyond contrast thresholds of detection, in which case they are deemed visible. For suprathreshold distortions, low-level and mid-level visual properties of perceived contrast and global precedence respectively, are modeled as Euclidean distances in the distortion-contrast space of a multiscale wavelet decomposition. VSNR is then calculated as the ratio of the RMS contrast of the pristine reference image to the weighted sum of the two Euclidean distances. An information content weighted version of PSNR, called IW-PSNR is proposed in [87], where the underlying premise is that some regions of visual content are perceptually more important than others, either due to the visual attention property of the HVS or due to the influence of distortions [104], [105]. IW-PSNR uses information theoretic principles to compute information content weights which are used in the pooling stage of quality score generation. In [76] a Haar wavelet based Discrete Wavelet Transform (DWT) framework is developed to compute image quality methods in the DWT domain. Image quality methods are separately applied to the approximation subbands and edge-maps obtained from detail subbands, leading to approximation and edge quality scores which are linearly combined to yield the final quality scores. Of the four developed methods in [76], two are error-based methods and include PSNR-DWT and absolute difference based AD-DWT.

2) STRUCTURAL SIMILARITY BASED METHODS

It can be seen from the previous section that HVS characteristics have been used to modify error based methods such as the MSE and PSNR. This is essentially a *bottom-up* approach to IQA design since the functionality of different HVS components is being simulated. By contrast, the *top-down* approach to IQA design does not try to model the functionality of individual HVS components. Instead, it tries to mimic the functionality of HVS as a whole [1]. The last two decades have seen the advent of a number of successful IQA methods that follow the *top-down* approach, some of which will be briefly explained in this and subsequent sub-sections.

One of the most well-known FR methods following the *top-down* approach is the Structural Similarity (SSIM) index [98], which is a modified version of the Universal image Quality Index (UQI) [99], and is based on the assumption that the HVS is adapted for extracting structural information from visual content. SSIM operates in the spatial domain and performs three types of comparisons between the reference and distorted images: luminance, contrast and structure. Luminance comparison is a function of mean intensity of the images being compared, while contrast comparison is a function of standard deviations. Structural comparison is done through correlation between the image patches being compared after mean subtraction and variance normalization. All comparisons are done locally by a sliding window and the three SSIM components are combined, leading to local quality scores, which together lead to a quality map. The overall quality score for the entire distorted image with respect to the reference image is obtained by taking the mean of all the local quality scores. The SSIM index is a single-scale approach, that is, it can take into account only one set of viewing conditions. To account for the variations in viewing conditions, a multi-scale version of SSIM called MSSSIM was developed in [89] and uses 5 scales. Images at different scales are obtained by downsampling the images at the previous scale by a factor of 2. The contrast and structural comparisons are performed at all scales, while the luminance comparison takes place only at the final scale. The quality scores obtained at each scale are combined through a weighted product, where the weights assigned to different scales are obtained through an image synthesis calibration experiment that involved subjective testing. To generate final quality scores, both SSIM and MSSSIM use mean pooling, which assigns equal importance to all areas of visual content. As discussed earlier, some regions of visual content are perceptually more important, either because of the visual attention property of the HVS or due to the influence of distortions [104], [105]. In [87], a modified version of MSSSIM, called IWSSIM was presented. IWSSIM operates at 5 scales and uses information theoretic principles to compute information content weights that are used in the pooling stage. A wavelet domain implementation of SSIM, called Wavelet Structural Similarity Index (WSSI) was proposed in [103] which uses the Haar wavelet for image decomposition. In WSSI, edge-maps are obtained from detail subbands followed by the generation of approximation and edge structural similarity maps. A contrast map is used to pool together the different wavelet domain structural similarity maps leading to approximation and edge similarity scores which are then linearly combined into the final WSSI quality score. Another SSIM based wavelet domain method called the SSIM-DWT was developed in [76] and uses the same design philosophy as WSSI.

Besides SSIM and methods that are directly based on it, several other FR IQA methods have been proposed that utilize the SSIM design philosophy. The Riesz-transform based Feature SIMilarity metric (RFSIM) was proposed in [95]. RFSIM uses first and second order Riesz Transform coefficients as features and compares them only at key locations identified by an edge-based feature mask which is obtained by using the Canny edge detection operator without thinning. The final RFSIM quality score is obtained as a product of similarity scores of individual feature maps. The Feature Similarity index (FSIM) was proposed in [83] and uses phase congruency as the primary feature to evaluate image similarity. Since phase congruency is contrast invariant, the gradient magnitude is used as a secondary feature in FSIM to capture contrast information. The phase congruency and gradient magnitude maps of the reference and distorted images are

compared leading to phase congruency and gradient magnitude quality maps which are then combined into a single quality map for the luminance channel of the images through a weighted product. The final FSIM quality score is obtained by pooling this quality map by using a weighting function that is derived from the phase congruency maps of the images being compared. A color version of FSIM, called FSIMc, has also been proposed in [83]. RGB color versions of the images being compared are first converted to the YIQ color space [106]. Phase congruence and gradient magnitude based comparisons are performed on the luminance channel Y, as in FSIM, leading to the luminance similarity map. Additionally, the I and Q chromatic channels are compared leading to I and Q similarity maps whose product leads to a chrominance similarity map. The luminance and chrominance similarity maps are pooled into the final FSIMc score by using the phase congruence based weighting function. The Spectral Residual based Similarity index (SRSIM) is proposed in [97] and uses a spectral residual based visual saliency model (SRVS) [107] to perform two functions: 1) SRVS maps act as features to ascertain local quality and 2) A weighting function is derived from the SRVS map to highlight the importance of visual regions when pooling to obtain the final quality score. To account for the lack of contrast sensitivity of SRVS, SRSIM uses gradient magnitude maps of the images being compared as supplementary features. Following a similar design approach as SRSIM, the Visual Saliency-based Index (VSI) is proposed in [102] which is able to handle color images. VSI uses the visual saliency model called Saliency Detection by combining Simple Priors (SDSP) [108] to generate visual saliency maps, which are used as features in local quality estimation and also act as a weighting function during the pooling stage for final quality score generation. It was shown in [102] that visual saliency maps are insensitive to change of contrast and color saturation, which thus requires VSI to include additional features. This is accomplished by first transforming the RGB color images into an opponent color space. Next, gradient magnitude is used as a feature to generate gradient similarity maps in order to make VSI contrast sensitive, while chrominance similarity maps are generated through the two chromatic channels to make VSI color saturation sensitive. A Gradient Similarity based IQA method (GSIM) is proposed in [85], where changes in contrast and structure in images being compared are measured through gradient comparison. It also takes into account masking effects, visibility threshold and luminance distortions. GSIM combines the measurement of luminance distortion and contrast-structure distortion in an adaptive manner to give a final quality score, where more weight is given to the latter. An Edge Strength Similarity based IQA method (ESSIM) is proposed in [82], where it is assumed that the edge-strength of each pixel fully represents the semantic information of images. Based on the characteristics of the edge in images, ESSIM defines edge-strength to take both anisotropic regularity and irregularity into account. Another FR IQA method based on gradient similarity called Gradient Magnitude Similarity Deviation (GMSD) is proposed in [84]. While GMSD compares the gradient magnitude maps of the reference and distorted images to compute a local quality map, it uses standard deviation as the pooling strategy to generate the final quality score from the local quality map. The underlying premise is that the global variation of local image quality is an indicator of overall image quality. Following the design philosophy of GMSD, an FR IQA method called Multiscale Contrast Similarity Deviation (MCSD) is proposed in [88]. First, the pristine reference and distorted images are downsampled by a factor of 2 and a contrast similarity map for the images being compared is computed by using their respective contrast maps. Next, standard deviation is used as a pooling strategy to generate a contrast similarity deviation (CSD) quality score from the contrast similarity map. To incorporate the effect of viewing distance, this process is repeated at two further scales by downsampling by a factor of 2 each time and computing the CSD at each scale. The product of the three CSD scores gives the final MCSD quality score. A Discrete Cosine Transform (DCT) domain FR IQA method called the DCT Subbands Similarity (DSS) is proposed in [79]. DSS measures the amount of local change of respective subband coefficients by comparing the local variance and generates a quality score for each subband. A final DSS quality score is obtained by combining the individual subband scores such that more weight is given to subbands corresponding to lower spatial frequencies in accordance with the characteristics of the HVS. The Color-Image-Difference measure (CID) [78] is a FR IQA method for color images. CID uses an image-appearance model to normalize the images being compared and transforms them to a working color space. It then extracts features from both the reference and distorted images, which are compared for similarity. Feature comparisons include lightness, chroma, hue, contrast and structure comparisons. A multiscale approach similar to MSSSIM [89] is used for contrast and structure comparisons. Lightness comparison is made on the smallest scale. A factorial combination model is finally used to combine the scores from different feature comparisons into a single CID quality score.

3) NATURAL SCENE STATISTICS BASED METHODS

IQA methods belonging to this paradigm regard natural images as entities with certain statistical properties which can be defined in terms of representative models and that are effected due to distortions [2]. Statistical models of the reference and distorted images are compared using principles of information theory, thereby providing an opportunity for quality assessment. Early works apply the idea to reduced-reference (RR) IQA [109], [110], where the original reference image is not fully available, but certain statistical features (in this case natural scene statistics features) are extracted and compared with those extracted from the test image to yield a quality evaluation. The idea was later extended for FR IQA.

A well-known FR IQA method following this design approach is the Information Fidelity Criterion (IFC) that was proposed in [86]. IFC treats IQA as an information fidelity problem where the reference image from the natural image source is being communicated to a receiver who is a human observer, through a channel which is the distortion process. Here, the reference and distorted images are the input and output of the channel respectively. IFC uses a Natural Scene Statistics (NSS) [111] based Gaussian Scale Mixtures (GSM) model [112] in the wavelet domain to represent the source where the steerable pyramid decomposition [113] with six orientations is used. The distortion model is obtained by attenuating the source model and adding Gaussian noise to it. The task of image fidelity measurement is then accomplished by determining the mutual information between respective wavelet subbands of the reference and distorted images represented through the source and distortion models respectively. The final IFC fidelity or quality score is obtained by summing the mutual information for all subbands. Using the IFC as a base, the FR IQA method called Visual Information Fidelity (VIF) was proposed in [100]. Like IFC, the VIF uses a NSS [111] based GSM model [112] in the wavelet domain to model the source and uses the same steerable pyramid decomposition [113]. VIF also uses a similar distortion model as the IFC. However, the VIF introduces a HVS model in the wavelet domain to incorporate the uncertainty that is introduced by the HVS channel as it processes the visual signal. VIF models the HVS channel through a stationary, zero mean, additive white Gaussian noise model. VIF then defines two types of information: 1) The reference image information represents the information in the reference image and is defined as the mutual information between the input and output of the HVS channel without the distortion channel. 2) The test image information is the information in the distorted image and is defined as the mutual information between the input of the distortion channel and the output of the HVS channel, where these two channels are in series (distortion channel followed by the HVS channel). VIF is then defined as the ratio of the test image information to the reference image information (for all subbands). The designers of VIF [100] provide a pixel domain version of VIF, called VIFP which is computationally simpler. Although the implementation details of VIFP have not been provided in [100], some information and its implementation code can be found at [101]. While VIF [100] uses a vector GSM implementation, VIFP [101] uses a scalar GSM implementation and is multi-scale in nature. A low-complexity wavelet-domain version of VIF, called the DWT-VIF has been proposed in [81]. To reduce the computational complexity, DWT-VIF adopts a one-level decomposition using the Haar wavelet instead of the over-complete steerable pyramid decomposition as in VIF. This allowed the use of a scalar GSM model in DWT-VIF instead of the vector GSM model that was required in VIF. DWT-VIF computes quality scores separately between approximation subbands and edge maps extracted from the detail subbands of the reference and distorted images being compared. A linear combination of the approximation and edge similarity scores gives the final DWT-VIF quality score. The designers of DWT-VIF [81] provide a similar method called VIF-DWT in [76].

Since natural images are known to possess sparse structures, sparsity based approaches to IQA can also be placed under the NSS category. A sparse coding based FR IQA method for color images called Sparse Feature Fidelity (SFF) is proposed in [96]. SFF computes the fidelity of the distorted image with respect to the reference image by using two sub-tasks, feature similarity and luminance correlation. A universal feature detector is trained once on a set of natural images using Independent Component Analysis (ICA) and then used to transform a given image into a sparse coefficient vector. The reference and distorted images are first split into corresponding patches and only those patches are selected for further processing which display suprathreshold distortions. Next, the feature detector is applied to the selected reference and distorted image patches to extract sparse feature vectors. The feature vectors of the reference image are used to determine a visual threshold to identify visually important patches. This process of patch and feature vector selection is done to incorporate the HVS properties of visual attention and visual thresholding [104], [105]. Once features have been selected from the reference and distorted images, similarity between them is determined. Separately, correlation between the mean values of selected image patches from the reference and distorted images is used to represent luminance correlation. Finally, the feature similarity and luminance correlation values are linearly combined to yield the final SFF quality score. Another sparsity based FR IQA method for color images called sparse representation based image Quality index with Adaptive Sub-Dictionaries (QASD) has recently been proposed in [94]. QASD utilizes a universal overcomplete dictionary, which is trained by using natural images, to extract sparse features which are the primary features being used for quality assessment. First, QASD utilizes the universal overcomplete dictionary to extract sparse coefficients from blocks of the reference image. Next, it adaptively forms sub-dictionaries for respective image blocks by using only the basis vectors obtained in the sparse representation of the reference image. The sparse representation of the distorted image blocks is then obtained only by using the respective sub-dictionaries. This ensures that the same set of basis vectors are used in the sparse representation of both the reference and distorted images, therefore ensuring meaningful comparison for IQA. Using the sparse representations, feature maps are generated for the reference and distorted images which are then compared for similarity. It is mentioned in [94] that weak distortions have limited influence on sparse representations, therefore, supplementary features are employed to capture the effect of such distortions. Three supplementary features are used which include image gradient, color and luminance. The RGB color image is first converted to the YCbCr color space, which is followed by image gradient similarity computation in the Y channel and color similarity computation in the

chroma channels. Luminance similarity is determined as in SFF [96]. The sparse feature maps are used to generate a weighting map, which is used in the weighted pooling of the sparse feature similarity map, gradient similarity map, and chroma similarity map. The final QASD quality score is obtained as a weighted product of the various similarity maps.

4) MIXED STRATEGY BASED METHODS

Some other design philosophies have also been used for the task of FR IQA, which use an overlap of different strategies. The Most Apparent Distortion (MAD) [6] method, assumes that the HVS adopts two different strategies to determine image quality: 1) For high quality images with only near-threshold distortions, MAD uses a detection based strategy. A spatial domain local visual mask, based on the CSF, luminance and contrast masking, is used to find regions in which the near-threshold distortions are considered as visible. Image quality of the distorted image with respect to the reference is then estimated in the identified regions through the mean squared error. 2) For low quality images with clearly suprathreshold distortions, MAD uses an appearance based strategy. A log-Gabor filter bank is used to decompose the reference and distorted images into coefficients, with greater weight given to coarser scales. Image quality is determined as the absolute difference between low level statistics including the mean, variance, skewness and kurtosis, of the weighted coefficients. Based on the amount of distortion, the detection and appearance quality scores are then combined through a weighted geometric mean to give the final MAD score.

A FR IQA method (ADM) was proposed in [77] which uses a wavelet domain decoupling algorithm for impairment separation and then evaluates detail losses and additive impairments. It simulates the HVS by incorporating the CSF and contrast masking characteristics of the HVS. Detail loss, defined as the loss of useful visual information, is computed after the decoupling process as the ratio of the Minkowski sum of the restored image to that of the original image. Additive impairment, defined as redundant visual information due to the influence of distortions, is computed as the Minkowski sum of the additive impairment image obtained after the decoupling process. The detail loss and additive impairment quality scores are then adaptively combined such that more weight is given to the detail loss based score for low quality images.

The Detail Virtual Cognitive Model (DVICOM) [80] combines two separate metrics that measure the perceptual impact of detail losses and spurious details. Using the images being compared and Least Squares decomposition, DVICOM breaks down the gradient field of the distorted image into two components, a prediction of the gradient field of the original image and an unpredictable gradient residual. Detail loss is then determined by the attenuation of the predicted gradient, measured through the loss of positional information. The gradient residual is used to measure the spurious detail component as the ratio of the original gradient energy and the residual gradient energy. These two components are considered as coordinates of a two-dimensional space and mapping is done to a DMOS estimate by using a parametric function that has been trained on experimental data. In addition to the standard version of DVICOM, a computationally faster version has also been provided by its inventors, which we refer to as DVICOM_F.

B. FR FUSION BASED IMAGE QUALITY ASSESSMENT

It is evident from Sections IV and V that state-of-the-art FR IQA methods achieve good correlation with human perception of quality (where the weighted average SRCC of top performing FR methods is around 0.86 on nine subject-rated databases), while there is significant room for improvement in the performance of general-purpose NR IQA methods (where the top performing NR method has a weighted average SRCC of around 0.61 on the same set of data). However, it has been observed in the past [54] and we shall demonstrate later in this paper as well that the performance of state-of-the-art FR methods fluctuates across different IQA databases that have different sets of distortions. The question is: How to achieve objective IQA that has stable, robust, and perceptually well-correlated performance across different distortion types? Researchers have tried to answer this question by combining or fusing the results from different FR IQA methods together, in the hope that the deficiencies of one method will be covered by another method in the combination set. Such FR fusion methods can be classified into three categories: 1) Empirical fusion methods, 2) Learning based fusion methods, and 3) Rank aggregation based fusion methods. In this work we evaluate the performance of seven FR fusion based methods which are listed in Table 4 along with information about whether they operate on grayscale or color images, year of publication, and number and names of the IQA databases that they were tested on. A brief description of these methods and their categories follows.

1) EMPIRICAL FUSION

In this rather simple approach, the results from two or more FR IQA methods are combined through a weighted product procedure. The weights assigned to different FR methods are obtained by optimizing on some subject-rated database.

The Hybrid Feature Similarity (HFSIMc) index [117] combines results from two feature similarity based FR methods, FSIMc [83] and RFSIM [95], in the following manner:

$$\text{HFSIMc} = (\text{RFSIM})^a \cdot (\text{FSIMc})^b \tag{9}$$

where the exponent values of $a = 0.4$ and $b = 3.5$ have been optimized on the TID2008 database [4].

The Combined Image Similarity Index (CISI) [114] combines results from three FR methods, FSIMc [83], MSSSIM [89] and VIF [100], as follows:

$$\text{CISI} = (\text{MSSSIM})^a \cdot (\text{VIF})^b \cdot (\text{FSIMc})^c \tag{10}$$

where the exponent values of $a = 0.5$, $b = 0.3$, and $c = 5$, have been optimized on the TID2008 database [4].

TABLE 4. Information about seven FR fusion based methods under performance evaluation.

FR Fusion Method	Color/Gray	Year	No. of Test Databases	Single Distortion Test Databases Used							Multiple Distortion Test Databases Used
CISI [114]	Color	2012	7	A57	CSIQ	IVC	LIVE R2	MICT	TID2008	WIQ	None
CM3 [115]	Gray	2014	1	None							LIVE MD
CM4 [115]	Gray	2014	1	None							LIVE MD
CNNM [116]	Color	2015	1	TID2013							None
HFSIMc [117]	Color	2012	7	A57	CSIQ	IVC	LIVE R2	MICT	TID2008	WIQ	None
MMF [118]	Gray/Color	2013	6	A57	CSIQ	IVC	LIVE R2	MICT	TID2008		None
RAS [21]	Gray/Color	2014	3	CSIQ	LIVE R2	TID2008					None

Two combined metrics designed for multiply distorted images are proposed in [115]. They are called CM3 and CM4, and are respectively defined as:

$$\mathrm{CM3} = (\mathrm{IFC})^{0.34} \cdot (\mathrm{NQM})^{2.4} \cdot (\mathrm{VSNR})^{-0.3} \quad (11)$$

$$\mathrm{CM4} = (\mathrm{IFC})^{0.2} \cdot (\mathrm{NQM})^{2.9} \cdot (\mathrm{VSNR})^{-0.54} \cdot (\mathrm{VIF})^{0.5} \quad (12)$$

where IFC [86], NQM [90], VSNR [56], and VIF [100] are FR methods, and the exponent values have been optimized on the LIVE MD database [11].

Although some other FR fusion based methods that follow the weighted product approach have been proposed, such as the CQM [119] and the EHIS [120], we will use the above-mentioned four methods as representatives of this category.

2) LEARNING BASED FUSION

A general-purpose learning based FR fusion approach called Multi-Method Fusion (MMF) was first proposed in [121] and then further refined in [118]. Given an annotated training dataset, MMF selects a subset of FR IQA methods from a larger pool, and then uses support vector regression (SVR) to learn a model that is a non-linear combination of the methods being fused. Defining similar distortion types as a *context*, two kinds of fusion methods are constructed: 1) Context-Free MMF (CF-MMF) is independent of distortion type where regression is done at the level of the entire training set. 2) Context-Dependent MMF (CD-MMF) takes distortion type into account and performs regression within each group of similar distortions. In the published version of MMF [118], the pool of FR IQA methods is composed of 10 methods which include: MSSSIM [89], SSIM [98], VIF [100], VSNR [56], NQM [90], PSNR-HVS [92], IFC [86], PSNR, FSIM [83], and MAD [6]. However, it is noted that any other FR method pool can be used for MMF construction. To ensure a level playing field, scores from different FR methods are linearly rescaled to the range of [0, 1], as per the recommendations in [122], before learning a combination model through SVR. For CD-MMF, SVM is used to learn a classification algorithm to automatically determine the context of a given image. To accomplish this, the distortions in known IQA databases are divided into five groups based on similarity among distortions and five spatial domain features are used to learn the classification algorithm. With the context determined, FR method fusion is carried out through a SVR based model which may involve a different set of FR methods for each context. To determine the best possible set of FR methods to be fused, for both CF-MMF and CD-MMF exhaustive search becomes infeasible if the FR method pool is large. Two algorithms are proposed in [118] for FR method selection: 1) Sequential Forward Method Selection (SFMS) uses PLCC as the objective function and starts with a single FR method that has the highest PLCC with respect to the training subjective data. It then combines this method with every other FR method in the pool one at a time and trains the MMF model, where the method that gives the highest PLCC is selected as the second FR method. This process is repeated sequentially until all the FR methods in the pool have been exhausted. The number of FR methods being combined is then selected based on computational complexity requirements. 2) Biggest Index Ranking Difference (BIRD) selects FR methods that are most dissimilar to each other in order to have a FR set that works well for a wide variety of distortions. The number of FR methods to be fused for a particular training dataset is determined based on a formula that balances performance and complexity. For example, the fusion count is estimated to be six for the TID2008 database [4] while using the SFMS algorithm, and the following methods are selected: FSIM [83], VIF [100], IFC [86], MAD [6], PSNR-HVS [92], and MSSSIM [89]. This combination will be used later in this work while evaluating the performance of MMF where we have restricted ourselves to CF-MMF. We will also use three other pools for FR method selection, details of which are provided in Section IV-C.

A neural networks based general-purpose supervised FR fusion based approach called Combined Neural Network Metric (CNNM) was proposed in [116]. As input, CNNM takes the scores from six FR IQA methods without any pre-processing and gives a combined quality score at its output. In order to select FR methods for fusion, 27 different methods were analyzed on the 24 different types of distortions in the TID2013 database [5]. Based on results from this analysis and the evaluation done in [123], six FR methods were chosen such that they reliably cover the distortions in TID2013 between them. The selected FR methods include VIF [100], PSNR-HVS [92], PSNR-HMAc [91], FSIMc [83], SFF [96], and SRSIM [97]. A 4-layer cascade-forward backprop neural network with 10, 10, and 20, neurons in hidden layers was used with training being done on the TID2013 database [5] using MATLAB. The TID2013 database has 3000 images, of which 1500 were used for training while the remainder were used for later analysis.

During training itself, MATLAB used 500 of the 1500 images for training, while 500 were used for validation and 500 for testing.

3) RANK AGGREGATION BASED FUSION

The FR fusion methods discussed above require training with respect to subject-rated databases. The empirical fusion approaches need such databases to optimize exponent values while the learning based fusion approaches need them to learn the combination model. These approaches often suffer from overfitting problems, as will be demonstrated in Section IV. On the other hand, a training-free fusion approach could potentially alleviate these issues.

A recently proposed framework called Blind Learning of Image quality using Synthetic Scores (BLISS) [21] replaces human opinion scores with synthetic quality scores that act as ground truth data. Such synthetic quality scores are generated by using a training-free FR fusion method which involves two steps: 1) Generation of consensus ranking through unsupervised rank aggregation, and 2) Score adjustment of a base FR method based on the consensus ranking. Since different FR measures have different score ranges, their outcomes cannot be combined by averaging their values. Instead rank aggregation is used as an alternative. Given a set of test images and their associated scores assigned by a number of FR methods, a consensus ranking is first obtained by using the unsupervised rank aggregation method called Reciprocal Rank Fusion (RRF) [124], which was first developed for combining document rankings from multiple information retrieval systems. The RRF score of an image I_i is defined as [21], [124]:

$$RRF_{score}(I_i) = \sum_{j=1}^{J} \frac{1}{k + r_j(i)} \tag{13}$$

where J is the number of FR methods being combined, $r_j(i)$ is the rank given by the j-th FR method to the image I_i, and $k = 60$ is a constant that counters the impact of high rankings by outliers. The value of the constant k was determined through a pilot investigation in [124].

It is mentioned in [21] that RRF values cannot be directly used as quality scores since they indicate the quality of an image relative to other images in the dataset. Instead, a quality measure is obtained by adjusting the scores of a base FR method with respect to the consensus ranking obtained through RRF. While generating the final synthetic quality scores, the mean squared error between the combined scores and the base FR scores is minimized and a penalty is applied when there is an inconsistency with respect to the consensus ranking. The entire process of FR method fusion and synthetic score generation is training-free. In this work we will call the FR fusion approach proposed in [21] as RRF based Adjusted Scores (RAS). In [21], five FR methods are used in fusion which include GMSD [84], VIF [100], FSIM [83], FSIMc [83], and IWSSIM [87]. Two combinations are adopted, where the first fuses all five FR methods while the second one excludes VIF. We shall respectively call them as RAS_B1 and RAS_B2 in this work and will evaluate their performance in addition to several other RAS fusion combinations in Section IV.

C. NO-REFERENCE IMAGE QUALITY ASSESSMENT

No-Reference (NR) IQA methods evaluate the quality of a distorted image in the absence of any reference information [1], and thus they are also referred to as *blind* IQA methods. By its very nature, blind IQA is a difficult task and early efforts were made towards the design of NR IQA methods for specific distortions, such as for blur [125], JPEG compression [126], JPEG2000 compression [127]. However, with advances in domain-knowledge, technology and with the availability of subject-rated IQA databases, several general-purpose NR methods have been designed in the last decade that work with a number of distortions. Contemporary NR IQA methods are usually classified into two categories [128]: 1) Opinion-Aware (OA) methods which are trained on distorted images whose quality has been rated by human subjects, and 2) Opinion-Unaware (OU) methods (also referred to as Opinion-Free) which do not train on human-rated distorted images. In this work we evaluate the performance of 14 NR IQA methods (8 OA and 6 OU) which are listed in Table 5 along with information about whether they operate on grayscale or color images, year of publication, and number and names of the IQA databases that they were tested on. Although this is not an exhaustive list, we selected NR methods for a good representation of various blind IQA design philosophies in addition to computational time constraints. A brief description of the NR IQA methods being evaluated in this work is given next.

1) OPINION-AWARE NR METHODS

OA NR methods can be further classified into two categories based on whether handcrafted or learned features are used.

In the handcrafted features based approach, features that correlate well with image quality, such as NSS based statistical parameters representing the empirical distributions of image coefficients in either the spatial or some transform domain, are extracted from the distorted images. Next, these feature vectors and associated image subjective ratings are used to train a model by using machine learning techniques such as support vector regression (SVR) [139]. In the testing phase, the OA NR method extracts features from the test image and uses the learned quality model to map them to a quality score. The Blind Image Quality Index (BIQI) [129] is a pioneering general-purpose NR IQA method based on the premise that different distortions affect the natural scene statistics (NSS) of images in a specific manner. BIQI uses the Daubechies 9/7 wavelet basis [140] to decompose an image into three-scales and three-orientations. The Generalized Gaussian Distribution (GGD) is then used to represent the coefficients of each subband. GGD parameters are estimated using the approach proposed in [141] and form a feature vector to represent the image. BIQI then follows a

TABLE 5. Information about 14 NR IQA methods under performance evaluation.

NR Method	Color/Gray	Year	No. of Test Databases	Single Distortion Test Databases Used	Multiple Distortion Test Databases Used
BIQI [129]	Gray	2010	1	LIVE R2	None
BRISQUE [130]	Gray	2012	2	LIVE R2 TID2008	None
CORNIA [131]	Gray	2012	2	LIVE R2 TID2008	None
dipIQ [16]	Gray	2017	4[b]	CSIQ LIVE R2 TID2013	None
GWHGLBP [132]	Gray	2016	2	None	LIVE MD MDID2013
HOSA [133]	Gray	2016	10[a]	CSIQ LIVE R2 MICT TID2013	LIVE MD
ILNIQE [134]	Color	2015	4	CSIQ LIVE R2 TID2013	LIVE MD
LPSI [135]	Gray	2015	2	LIVE R2 TID2008	None
MEON [136]	Color	2018	4[b]	CSIQ LIVE R2 TID2013	None
NIQE [128]	Gray	2013	1	LIVE R2	None
NRSL [137]	Gray	2016	7[c]	CSIQ LIVE R2 TID2013	LIVE MD
QAC [15]	Gray	2013	3	CSIQ LIVE R2 TID2008	None
SISBLIM [12]	Color	2014	7	CSIQ IVC LIVE R2 MICT TID2008	LIVE MD MDID2013
WaDIQaM-NR [138]	Color	2018	4[d]	CSIQ LIVE R2 TID2013	None

[a] HOSA was also tested on two authentic distortion databases: CID2013 [60], LIVE WC [61]; one database of screen content images: SIQAD [66]; and on two document image databases: Newspaper database [68], Document Image Quality database [67].
[b] dipIQ and MEON are also tested on the Waterloo Exploration database [58] which is a single distortion database that does not have subject rated quality scores.
[c] NRSL was also tested on three authentic distortion databases: BID [59], CID2013 [60], LIVE WC [61].
[d] WaDIQaM-NR was also tested on one authentic distortion database: LIVE WC [61].

two-step process to determine image quality. First, the feature vector is used to determine the presence of various distortions. Since BIQI is trained on the LIVE R2 database [3], [22], its published version uses a distortion set of JPEG compression, JPEG2000 compression, white noise, Gaussian blur and fast fading, since these distortions are present in LIVE R2. In the training phase, BIQI uses support vector machine (SVM) [139] to learn the classification model which assigns probability scores to various distortions based on their perceived magnitude. In the second step, the same feature vector is used for quality score assignment for each distortion category. In the training phase SVR [139] is used to learn a regression based model. The final BIQI quality score is then determined as a probability-weighted sum of the quality scores for various distortions. The Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [130] is a NSS based NR method that operates in the spatial domain. BRISQUE operates on locally normalized luminance values which are termed as Mean Subtracted Contrast Normalized (MSCN) coefficients. A benefit of this normalization process is that it leads to relatively decorrelated neighboring coefficients as compared to non-normalized pixel values. NSS features are extracted from the models of the MSCN coefficients and their pairwise products. The GGD is used to fit the empirical MSCN distributions, where the procedure proposed in [141] is used to estimate GGD parameters which form one set of features. The relationships between neighboring pixels are modeled through the pairwise products of neighboring MSCN coefficients along four orientations. The Asymmetric Generalized Gaussian Distribution (AGGD) [142] is used to fit the empirical distributions of these pairwise products and the estimated fitting parameters lead to another set of features. To incorporate multiscale operation, BRISQUE extracts features at two scales. It is shown in [130] that distortions affect these NSS features such that they occupy different regions in the GGD and AGGD parameter spaces, thereby providing an opportunity to learn quality models. BRISQUE uses SVR [143] to learn a model to map features to a quality score and uses the LIVE R2 database [3], [22] for training. The degradation of structural features has been used in the design of OA NR methods, such as the GWHGLBP [132] which has been designed for multiply distorted images. First, the gradient map of a distorted image is obtained through the Prewitt filter. Structural information is extracted from the gradient map by applying the Local Binary Pattern (LBP) operator [144] leading to GLBP codes. It is claimed that these codes are affected in unique ways by different distortions, making them effective features for IQA. Contrast information is incorporated with structural information by accumulating the gradient magnitude of pixels that have the same GLBP pattern, thereby leading to a histogram of gradient-weighted GLBP codes which forms the feature space. Feature extraction is done at two scales and SVR is used to learn a mapping from this feature space to quality scores. In this work, we have used the version of GWHGLBP that has been trained on the LIVE MD database [11]. NRSL [137] is an OA NR method that uses both structural and luminance based features. NRSL begins by performing local contrast normalization as a means to reduce redundancy in a manner similar to [130]. The LBP operator [144] is locally applied to the contrast normalized image to obtain the LBP code of each pixel. These codes are then used to build a structural histogram. Separately a luminance histogram is built from the absolute magnitudes of the contrast normalized image. The structural and luminance histograms represent the feature space of NRSL, and feature extraction is done at three scales. SVR is then used to learn a mapping from the feature space to quality scores. In this work, we have used the version of NRSL that has been trained on the LIVE R2 database [3], [22].

The handcrafted features based approach is designed around features that have been selected based on domain knowledge. An alternative approach is to automatically learn features which are then used in the training process along with subjective ratings to design OA NR models. A pioneering

method following this approach is called COdebook Representation for No-reference Image quality Assessment (CORNIA) [131], which uses unsupervised feature learning. CORNIA extracts a number of local descriptors by randomly sampling patches from an image, which are normalized and whitened before being used as local features. K-means clustering is performed on local features belonging to unlabeled training images to construct a visual codebook which is also normalized. Soft-assignment coding is performed on local descriptors by using the visual codebook which leads to a coefficient matrix that is converted to a fixed-length feature vector through maxpooling. In the publicly released version of CORNIA, the CSIQ database [6] is used for codebook construction and SVR with a linear kernel is used to learn a mapping from the feature vector to quality scores, where the LIVE R2 database [3], [22] has been used for model training. Compared to CORNIA, which uses low order statistics and a large codebook composed of 10000 codewords, a recent OA NR method called High Order Statistics Aggregation (HOSA) [133] also utilizes higher order statistics and a much smaller codebook composed of only 100 codewords. HOSA extracts local features in a manner similar to CORNIA [131] and also uses K-means clustering for codebook construction. However, in addition to calculating the mean of each cluster, higher order statistics including covariance and coskewness of each cluster are also calculated. A quality aware representation of an image is obtained through soft weighted differences of image statistics, including high order statistics. SVR with a linear kernel is used to learn a mapping from the feature space to quality scores. In the publicly available version of HOSA, the codebook is constructed by using the CSIQ database [6], while the LIVE R2 database [3], [22] is used for model training.

Recently deep neural networks (DNN) based approaches, mostly using convolutional neural networks (CNN), have been used to learn features and quality models. An end-to-end optimized DNN based approach is proposed in [138] that is capable of performing both FR and NR quality assessment, and built upon an earlier version [145]. The CNN used in [138] is based on the VGG network [146] and has ten convolutional layers, five pooling layers for feature extraction, and two fully connected layers for regression. Since CNNs require large training data and quality annotated IQA datasets are quite small, the size of the training set is augmented by randomly sampling multiple patches from each training image, which are assigned the same quality label as the parent image. The network takes image patches of size 32×32 pixels as input. Rectified Linear Unit (ReLU) [147] is used as the activation function. To perform IQA, an image is divided into 32×32 sized patches and local quality scores are pooled into a global image quality score either by simple or weighted average. The latter functionality aims to pool local quality scores based on the principles of visual saliency and is incorporated by adding a second branch that runs parallel to the quality regression branch of the network. This additional branch gives patchwise weights that are then used in pooling. For our tests we have selected the weighted average version of the NR approach proposed in [138] which is called Weighted Average Deep Image QuAlity Measure for NR IQA (WaDIQaM-NR) that is trained on the LIVE R2 database [3], [22]. The Multi-task End-to-end Optimized deep neural Network (MEON) [136] is another recent DNN based approach. MEON breaks the IQA task into two subtasks that are performed by respective sub-networks: 1) Distortion identification, and 2) Quality score prediction. Instead of using ReLU [147], MEON uses the bio-inspired generalized divisive normalization (GDN) transform [110] as the activation function which allows for a reduction of model parameters. The two sub-networks in MEON share the early layers, specifically four stages are shared where each stage consists of a convolutional, GDN, and maxpooling layers. Thereafter, sub-network 1 which is responsible for distortion identification and has two fully connected layers with a GDN layer in between, produces a probability vector to identify the likelihood of each distortion. Sub-network 2 which itself has two dedicated fully connected layers with a GDN layer in between, is responsible for quality prediction and produces a score vector containing quality scores corresponding to each distortion. The probability vector from sub-network 1 is fed into sub-network 2 where it is combined with the score vector to give a final quality score in terms of a scalar value, thereby giving the network a causal structure. Due to its multi-task nature, MEON is able to break the training phase into two steps. The loss function of subtask 1 is minimized in the initial pre-training step. Since training for distortion type identification does not require subject-rated data, MEON is able to train the shared layers and sub-network 1 on a large amount of data in the pre-training step. In the second training step, the entire network is joint optimized in an end-to-end manner by using a subject-rated database. In its publicly available version, MEON used the LIVE R2 database [3], [22] for performing joint optimization. Although a number of other deep learning based approaches have been proposed recently [18], [19], [148]–[158], in this work we have evaluated the performance of WaDIQaM-NR [138], [145] and MEON [136] as only their author-trained models are publicly available.

2) OPINION-UNAWARE NR METHODS

OU NR methods may be training-free or they may require some form of training that does not involve subject-rated images.

The Natural Image Quality Evaluator (NIQE) [128] is a pioneering general-purpose OU NR method. Like BRISQUE [130] (discussed in the previous sub-section), NIQE operates at two scales in the spatial domain by converting an image into MSCN coefficients, uses the GGD to fit the empirical distribution of these coefficients, uses the AGGD to fit the empirical distribution of pairwise coefficient products, and uses the estimated GGD and AGGD parameters as NSS features. However, unlike BRISQUE, NIQE does not use these features in conjunction with subject-rated distorted images to train a quality model. Instead, NIQE uses the features obtained from

a distorted image to fit a multivariate Gaussian (MVG) model whose distance from a universally learned MVG model of pristine natural images is regarded as a measure of quality. Although some training is required to obtain the MVG model representing pristine natural images, no training is necessary with respect to quality annotated distorted images which is what makes NIQE an OU NR method. The Integrated Local NIQE (ILNIQE) index [134] further builds upon the approach taken in NIQE. In addition to the two NSS features employed in NIQE (statistics of MSCN coefficients and their pairwise products), three additional NSS features are included. Information about structural degradations is incorporated by including image gradient features that include image gradient components through empirical fitting parameters of a GGD and gradient magnitude through the empirical fitting parameters of a Weibull distribution. To capture the selective response of neurons in the visual cortex to stimulus orientation and frequency, multi-scale multi-orientation filter responses are obtained through log-Gabor filters. NSS features are then extracted from response maps through GGD fitting and another round of gradient statistics extraction. ILNIQE also includes color based NSS features which are obtained by first taking the RGB color image to the logarithmic scale and then converting it to an opponent color space. A Gaussian model is then used to empirically fit the coefficient distributions in the opponent color space, thereby providing another set of NSS features. Like NIQE, ILNIQE determines the quality of a distorted image by measuring the distance between the MVG fit of its NSS features and the universal MVG model of pristine natural images. However, instead of using a single MVG model for the distorted image, image quality is determined at the patch level and then pooling is done to obtain a final quality score. ILNIQE also uses principal component analysis (PCA) to reduce correlation between features and for dimensional reduction.

The Quality Aware Clustering (QAC) method [15] takes an alternative approach to OU NR design. QAC partitions an image into a set of overlapping patches which are first divided into groups based on similar quality and then patches with similar local structures are clustered together. The local features are extracted through the application of a high pass filter. A set of centroids are learned for each quality group and form a codebook which is used to determine the quality of each patch. QAC has the capability to give a local quality map as well as an overall quality score. Although during its development QAC needs to divide image patches into groups based on quality, it does not use subject-rated databases to accomplish this. Instead it builds a new database starting from 10 source images from the Berkeley Segmentation database [159], and uses the FR IQA method FSIM [83] to annotate patch quality which is normalized through a percentile pooling procedure. Although QAC training does involve working with distorted images, it is still an OU NR method since it does not train against subject-rated distorted images. A recent OU NR method called DIP inferred quality (dipIQ) index [16] uses quality-discriminable image pairs (DIPs) for training. First a new dataset is constructed that has 840 source and 16,800 distorted images (which include Gaussian noise, Gaussian blur, JPEG and JPEG2000 compression). A DIP generation engine is constructed which uses three FR IQA methods, GMSD [84], MSSSIM [89], and VIF [100], to annotate distorted image quality. Each candidate image pair is assigned with a non-negative T value equivalent to the smallest score difference of the FR models. A raised-cosine function is used to quantify the quality discriminability uncertainty level based on T values. 80 million DIPs are produced using this DIP generation engine. Using these DIPs with their associated uncertainty levels and CORNIA features [131] as base features for image representation, RankNet [160] which is a neural network based pairwise learning-to-rank algorithm, is employed to learn an OU NR model.

Some OU NR methods take a training-free approach. The SIx-Step BLInd Metric (SISBLIM) [12], which is itself an improved version of FISBLIM [161], has been developed for singly and multiply distorted images and operates by determining the individual and joint impact of different distortions. It first uses the approach in [162] to estimate the amount of noise in a distorted image and then denoises the image by using the BM3D method [163]. The estimates of blur and JPEG quality are determined from the denoised image by using the methods proposed in [125] and [126] respectively. To take into account the interaction between different distortions and the masking effect due to image content, a model based on the free energy theory [164] is used to quantify the joint effects. Finally, the SISBLIM score is obtained as a linear combination of weighted quality estimates of noise, blur, JPEG compression and joint effects. The Local Pattern Statistics Index (LPSI) [135] is another recent training-free OU NR method that utilizes the LBP operator [144]. To reduce computational complexity, LPSI uses only four neighbors of each image pixel to compute LBP codes, which leads to six distinct binary patterns. Based on analysis, LPSI picks the locally weighted statistic associated with one of these six binary patterns as a quality measure since it offers the best discriminant ability to distinguish most distortions from pristine natural images.

IV. PERFORMANCE ANALYSIS OF FR AND FUSED FR METHODS

A. EVALUATION CRITERIA

1) PREDICTION ACCURACY

The Pearson Linear Correlation Coefficient (PLCC) is used as a measure of a method's prediction accuracy [53]. Since the scores produced by objective IQA methods are usually not linear with respect to subjective ratings, a nonlinear regression step is necessary before the computation of PLCC. We do this by adopting the five-parameter modified logistic function used in [3]:

$$P(Q) = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + e^{\{\beta_2 (Q - \beta_3)\}}} \right] + \beta_4 Q + \beta_5 \quad (14)$$

TABLE 6. Test results of 43 FR methods on nine subject-rated IQA databases in terms of PLCC. All distortions in each dataset were considered.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
AD_DWT	0.9384	0.3624	0.8163	0.8692	0.4100	0.5379	0.6010	0.7159	0.8501	0.8506
ADM	0.9360	0.8355	0.9285	0.9182	0.7791	0.8196	0.8349	0.6428	0.9062	0.9060
CID_MS	0.9159	0.8362	0.8732	0.9375	0.8364	0.8171	0.8414	0.6183	0.8917	0.8961
CID_SS	0.9279	0.8038	0.9079	0.9357	0.8534	0.7806	0.8617	0.6258	0.8822	0.8750
DSS	0.9618	0.8530	0.9612	0.9259	0.7715	0.8267	0.8733	0.8168	0.9023	0.8973
DVICOM	0.9734	0.8194	0.9179	0.9144	0.8035	0.8018	0.8919	0.8161	0.8873	0.8773
DVICOM_F	0.9735	0.8194	0.9191	0.9170	0.8037	0.8001	0.8916	0.8097	0.8858	0.8797
DWT_VIF	0.9658	0.7406	0.9009	0.8901	0.6952	0.5516	0.8941	0.7212	0.8716	0.8393
ESSIM	0.9566	0.8645	0.9224	0.9094	0.7953	0.8255	0.8451	0.6953	0.8861	0.9081
FSIM	0.9597	0.8589	0.9120	0.9185	0.7410	0.8265	0.8969	0.6474	0.8933	0.9037
FSIMc	0.9613	0.8769	0.9191	0.9329	0.7583	0.8410	0.8998	0.6412	0.8965	0.9039
GMSD	0.9603	0.8590	0.9541	0.9176	0.7387	0.7585	0.8776	0.8309	0.8808	0.8685
GSIM	0.9512	0.8464	0.8964	0.9155	0.7700	0.8342	0.8352	0.6647	0.8808	0.9072
IFC	0.9268	0.1737	0.8366	0.8614	0.5479	0.1724	0.9162	0.6279	0.9058	0.7990
IW_PSNR	0.9329	0.5984	0.8024	0.9212	0.6273	0.7200	0.6951	0.7649	0.8284	0.8771
IWSSIM	0.9522	0.8319	0.9144	0.9191	0.8476	0.8698	0.8983	0.8513	0.9109	0.9056
MAD	0.9675	0.8464	0.9500	0.9053	0.7809	0.8411	0.7552	0.7471	0.8948	0.8985
MCSD	0.9675	0.8648	0.9560	0.9217	0.7532	0.7727	0.8637	0.8275	0.8847	0.8787
MSSSIM	0.9489	0.8329	0.8991	0.9232	0.8180	0.8039	0.8419	0.7273	0.8747	0.8805
NQM	0.9129	0.6794	0.7200	0.9429	0.4879	0.6712	0.6170	0.3946	0.9086	0.7931
PSNR	0.8723	0.6775	0.7512	0.8321	0.6232	0.6814	0.6091	0.5564	0.7398	0.6806
PSNR_DWT	0.9301	0.6921	0.7631	0.8902	0.5792	0.6722	0.6393	0.5725	0.8630	0.8186
PSNR_HAc	0.9164	0.8418	0.9017	0.8759	0.7408	0.7624	0.7436	0.6768	0.7851	0.7322
PSNR_HA	0.9130	0.8511	0.8592	0.8697	0.6913	0.7292	0.7269	0.6825	0.8004	0.8093
PSNR_HMAc	0.9295	0.8329	0.8672	0.8977	0.7314	0.7896	0.7655	0.7255	0.8090	0.7560
PSNR_HMA	0.9249	0.8275	0.8342	0.8951	0.6831	0.7459	0.7437	0.7296	0.8192	0.8512
PSNR_HVS	0.9134	0.7031	0.7808	0.8843	0.6346	0.7073	0.6764	0.6813	0.7996	0.8085
PSNR_HVSM	0.9251	0.6709	0.7725	0.8841	0.6303	0.7042	0.6814	0.7281	0.8182	0.8506
QASD	0.9574	0.8897	0.9481	0.9253	0.7257	0.8116	0.8063	0.6312	0.8966	0.8827
RFSIM	0.9386	0.8329	0.9164	0.8904	0.6943	0.7621	0.7084	0.4738	0.8713	0.8200
SFF	0.9632	0.8706	0.9643	0.7761	0.7834	0.7721	0.8590	0.7952	0.8893	0.8904
SNR	0.8616	0.6498	0.7414	0.8228	0.6374	0.6888	0.6474	0.4264	0.7283	0.6414
SRSIM	0.9555	0.8664	0.9244	0.9022	0.7066	0.8147	0.8685	0.6401	0.8883	0.8928
SSIM	0.9449	0.7895	0.8612	0.9144	0.7674	0.8230	0.8457	0.5249	0.8915	0.8623
SSIM_DWT	0.9559	0.7799	0.9050	0.8955	0.8405	0.7821	0.8810	0.7624	0.8913	0.8594
UQI	0.8984	0.6427	0.8294	0.7981	0.6078	0.4980	0.8277	0.5318	0.8540	0.7723
VIF	0.9604	0.7720	0.9278	0.8938	0.7267	0.6415	0.9367	0.8376	0.9030	0.8736
VIF_DWT	0.9657	0.7657	0.9123	0.8969	0.7259	0.5845	0.9031	0.7531	0.8839	0.8653
VIF_P	0.9596	0.7529	0.9044	0.8921	0.7073	0.5629	0.8827	0.7589	0.8712	0.8126
VSI	0.9482	0.9000	0.9279	0.9320	0.7226	0.8240	0.8703	0.5512	0.8789	0.8749
VSNR	0.9236	0.7138	0.7355	0.8794	0.6261	0.7424	0.6805	0.3775	0.8309	0.8037
WSNR	0.9144	0.6031	0.7337	0.8468	0.5752	0.6766	0.5889	0.6853	0.8185	0.7942
WSSI	0.9549	0.7698	0.9001	0.9072	0.8406	0.7691	0.8785	0.7543	0.8843	0.8551

where Q denotes the objective quality scores directly from an IQA method, P denotes the IQA scores after the regression step, and β_1, β_2, β_3, β_4, and β_5 are model parameters that are found numerically in MATLAB to maximize the correlation between subjective and objective scores. Given a database with its subjective scores denoted by S, the PLCC value of an IQA method is then calculated as:

$$PLCC(P,S)=\frac{\sum_{i=1}^{N}(P_i-\bar{P})\cdot(S_i-\bar{S})}{\sqrt{\sum_{i=1}^{N}(P_i-\bar{P})^2\cdot\sum_{i=1}^{N}(S_i-\bar{S})^2}} \tag{15}$$

where P_i and S_i are respectively the values in the vectors P and S for the image i, $\bar{P}$ and $\bar{S}$ are respectively the mean values of vectors P and S, while N is the number of images in the database.

2) PREDICTION MONOTONICITY

The Spearman Rank-order Correlation Coefficient (SRCC) is used as a measure of a method's prediction monotonicity [53]. SRCC is a non-parametric rank-order based correlation metric and does not require the preceding nonlinear mapping step. The SRCC value of an IQA method on a database with N images is calculated as:

$$SRCC(Q,S)=1-\left[\frac{6\sum_{i=1}^{N}d_i^2}{N(N^2-1)}\right] \tag{16}$$

where d_i is the difference between the i-th image's ranks in the objective (Q) and subjective (S) scores. Other rank-order based methods such as Kendall's Rank-order Correlation Coefficient (KRCC) are found to be highly consistent with the SRCC measure and provide minimal additional information, and thus are not included in the current report.

3) STATISTICAL SIGNIFICANCE TESTING

Conclusions drawn about the performance of IQA methods based on PLCC and SRCC values can only be considered *universal* if testing is done on the entire population of concerned data, which in this case is the space of all possible natural images and their distorted versions. Since this is not possible and subject-rated IQA databases can only be regarded as sparse random samples from this enormous population, hypothesis testing is performed to ascertain whether the drawn inferences on a given sample size are statistically significant at a particular confidence level. The term

TABLE 7. Test results of 43 FR methods on nine subject-rated IQA databases in terms of SRCC. All distortions in each dataset were considered.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
AD_DWT	0.9412	0.5967	0.8029	0.8628	0.5522	0.6244	0.6027	0.7750	0.8040	0.7810
ADM	0.9460	0.7874	0.9333	0.9138	0.7794	0.8185	0.8186	0.6248	0.8815	0.8490
CID_MS	0.9103	0.8314	0.8789	0.9366	0.8350	0.8062	0.8330	0.6168	0.8608	0.8778
CID_SS	0.9270	0.7879	0.9116	0.9304	0.8528	0.7789	0.8535	0.6236	0.8408	0.8208
DSS	0.9616	0.7921	0.9555	0.9272	0.7755	0.8246	0.8658	0.8078	0.8714	0.8759
DVICOM	0.9750	0.7598	0.9181	0.9155	0.8034	0.7903	0.8840	0.8168	0.8672	0.8374
DVICOM_F	0.9748	0.7606	0.9226	0.9181	0.8028	0.7909	0.8837	0.8104	0.8642	0.8411
DWT_VIF	0.9671	0.6093	0.8909	0.8833	0.6909	0.5434	0.8836	0.7229	0.8269	0.7921
ESSIM	0.9597	0.8035	0.9325	0.9075	0.7968	0.8253	0.8250	0.6966	0.8517	0.8682
FSIM	0.9634	0.8015	0.9242	0.9178	0.7438	0.8149	0.8872	0.5817	0.8635	0.8585
FSIMc	0.9645	0.8510	0.9309	0.9323	0.7608	0.8285	0.8904	0.5806	0.8666	0.8613
GMSD	0.9603	0.8044	0.9570	0.9177	0.7427	0.7675	0.8613	0.8283	0.8448	0.8210
GSIM	0.9561	0.7946	0.9107	0.9121	0.7709	0.8299	0.8137	0.6637	0.8454	0.8485
IFC	0.9259	0.5389	0.7671	0.8570	0.4929	0.3427	0.9119	0.6861	0.8839	0.7807
IW_PSNR	0.9328	0.6913	0.8310	0.9166	0.6013	0.7137	0.6719	0.7816	0.7572	0.8178
IWSSIM	0.9567	0.7779	0.9212	0.9163	0.8484	0.8564	0.8911	0.8551	0.8836	0.8588
MAD	0.9669	0.7807	0.9466	0.9061	0.7815	0.8391	0.7249	0.7507	0.8646	0.8643
MCSD	0.9668	0.8089	0.9592	0.9224	0.7562	0.7808	0.8451	0.8269	0.8517	0.8370
MSSSIM	0.9513	0.7859	0.9132	0.9227	0.8196	0.7988	0.8296	0.7238	0.8363	0.8274
NQM	0.9093	0.6465	0.7411	0.9436	0.4694	0.6323	0.5827	0.4016	0.8999	0.7460
PSNR	0.8756	0.6394	0.8057	0.8246	0.6254	0.6701	0.5784	0.5604	0.6771	0.6136
PSNR_DWT	0.9325	0.6426	0.8052	0.8819	0.5401	0.6419	0.6070	0.5797	0.8206	0.7385
PSNR_HAc	0.9216	0.8187	0.9261	0.8702	0.7430	0.7684	0.7240	0.6724	0.7112	0.6789
PSNR_HA	0.9192	0.7792	0.9147	0.8610	0.6875	0.7295	0.7055	0.6785	0.7146	0.7284
PSNR_HMAc	0.9338	0.8128	0.9121	0.8907	0.7278	0.7877	0.7461	0.7249	0.7403	0.7114
PSNR_HMA	0.9298	0.7568	0.8997	0.8847	0.6634	0.7388	0.7239	0.7281	0.7423	0.7625
PSNR_HVS	0.9186	0.6533	0.8294	0.8781	0.6313	0.7011	0.6490	0.6779	0.7126	0.7278
PSNR_HVSM	0.9295	0.6246	0.8221	0.8756	0.6122	0.6969	0.6559	0.7273	0.7410	0.7619
QASD	0.9629	0.8674	0.9530	0.9231	0.7307	0.8079	0.7778	0.6687	0.8766	0.8315
RFSIM	0.9434	0.7743	0.9291	0.8871	0.6795	0.7450	0.6766	0.4151	0.8330	0.7756
SFF	0.9649	0.8513	0.9627	0.7738	0.7834	0.7689	0.8396	0.8005	0.8700	0.8535
SNR	0.8650	0.6127	0.7994	0.8101	0.6358	0.6709	0.6278	0.4383	0.6135	0.5767
SRSIM	0.9620	0.8076	0.9317	0.9021	0.7087	0.7966	0.8521	0.6238	0.8666	0.8350
SSIM	0.9479	0.7417	0.8755	0.9112	0.7697	0.8094	0.8328	0.4873	0.8604	0.7966
SSIM_DWT	0.9603	0.7093	0.9111	0.8877	0.8410	0.7815	0.8690	0.7650	0.8587	0.7929
UQI	0.8941	0.5507	0.8098	0.7984	0.5937	0.4743	0.8183	0.5334	0.8149	0.7311
VIF	0.9636	0.6769	0.9194	0.8866	0.7203	0.6257	0.9306	0.8444	0.8823	0.8381
VIF_DWT	0.9681	0.6439	0.9020	0.8930	0.7224	0.5826	0.8943	0.7553	0.8479	0.8243
VIF_P	0.9618	0.6101	0.8807	0.8919	0.7029	0.5471	0.8770	0.7594	0.8367	0.7711
VSI	0.9524	0.8965	0.9422	0.9317	0.7213	0.8106	0.8569	0.5700	0.8414	0.8269
VSNR	0.9279	0.6817	0.8108	0.8741	0.6145	0.7200	0.6594	0.3923	0.7719	0.7420
WSNR	0.9158	0.5782	0.7729	0.8381	0.5600	0.6542	0.5428	0.6998	0.7611	0.7243
WSSI	0.9586	0.6937	0.9075	0.9004	0.8411	0.7705	0.8690	0.7479	0.8494	0.7866

statistical significance signifies whether the difference in the performance of one IQA method with respect to another, on a set of sample points, is purely due to chance or due to some genuine underlying effect [165]. Generalizations about the difference in method performance can only be made in the latter case at the stated confidence level.

In the field of IQA, statistical significance testing is usually carried out on model prediction residuals. Given the objective scores of different IQA methods to be compared, they are converted to prediction residuals by first mapping them to the MOS/DMOS range of the database being used for testing by using the nonlinear mapping procedure explained for PLCC calculation in Section IV-A.1, and then subtracting the actual subjective scores from these *predicted* subjective scores. In this work we use the one-sided (left-tailed) two-sample *F*-test [165] to statistically compare the performance of any two given IQA methods at the 5% significance level (95% confidence). The *null* hypothesis is that the data in the two residual vectors comes from normal distributions with the same variance, making them statistically indistinguishable. The *alternative* hypothesis is that the data in the residual vectors comes from normal distributions with different variances, making them statistically distinguishable. The test statistic is the ratio of the variances of the two residual vectors. Given the number of residuals and the confidence level, a critical threshold is determined. If the value of the test statistic is smaller than the critical threshold, then this indicates a failure to reject the null hypothesis. By performing the one-sided test twice with the order of the methods swapped, we were able to determine if their performance is statistically indistinguishable or whether one method performed better than the other. In the statistical significance testing tables that follow, a "1", "_", or "0" mean that the method in the row is statistically (with 95% confidence) better, indistinguishable, or worse than the method in the column respectively. Since the tests assume the Gaussianity of prediction residuals, we use a simple kurtosis based check for Gaussianity as in [3]. If the kurtosis of prediction residuals of an IQA method is between 2 and 4, then they are accepted for the Gaussianity assumption.

B. PERFORMANCE OF FR METHODS ON INDIVIDUAL DATABASES

We tested the 43 FR methods discussed in Section III-A and given in Table 3, on each of the nine subject-rated IQA databases mentioned in Table 2, of which five are single

TABLE 8. RAS exhaustive search set composition.

S. No.	Fast Set	Medium Set	Full Set
1	ADM	ADM	ADM
2	DSS	CID_MS	CID_MS
3	ESSIM	DSS	DSS
4	FSIM	DVICOM_F	DVICOM
5	FSIMc	ESSIM	ESSIM
6	GMSD	FSIMc	FSIMc
7	GSIM	GMSD	GMSD
8	IWSSIM	GSIM	IWSSIM
9	MCSD	IWSSIM	MAD
10	MSSSIM	MCSD	MCSD
11	SFF	MSSSIM	QASD
12	SRSIM	SFF	SFF
13	SSIM_DWT	SRSIM	SRSIM
14	VIF_DWT	VIF_DWT	VIF
15	VSI	VSI	VSI

distortion datasets discussed in Section II-A and four are multiple distortion datasets discussed in Section II-B. Since the single distortion database CIDIQ [31] contains subjective scores at two viewing distances, testing was done separately for each case and results are mentioned under the headings of CIDIQ50 and CIDIQ100 for the viewing distances of 50 cm and 100 cm, respectively. For each database, testing was done on the entire dataset, that is, all distortions were considered. The test results are given in Table 6 in terms of PLCC and in Table 7 in terms of SRCC.

C. SELECTION OF FR METHODS FOR FUSION IN FUSED FR METHODS

We described seven fusion based FR methods in Section III-B and listed them in Table 4. The four methods belonging to the *empirical fusion* category (HFSIMc [117], CISI [114], CM3 [115], and CM4 [115]) combine specific FR methods and hence do not need to select methods from a large pool. The authors of the *learning based fusion* method CNNM [116] provide a pre-trained model that combines six FR methods, and we use the same selection and order for CNNM.

Although the authors of the *rank aggregation based fusion* method RAS [21] (discussed in Section III-B.3) provide a selection of methods to be fused, we believe that a more extensive search needs to be done to select FR methods for fusion, especially if the resulting scores are to be used as alternative ground truth for annotating large datasets. We begin by identifying a pool of FR methods to be combined in RAS [21]. Since RAS [21] not only combines FR methods but then adjusts the score of a base FR method with respect to the consensus ranking, an exhaustive search would require testing all possible combinations for each FR method being used as the base method. Given that we are considering 43 FR methods (Table 3), this would require testing more than 189 trillion combinations, which is computationally infeasible. To reduce the computational load, we make three sets of 15 FR methods each based on time constraints and thus test 245,760 combinations in each case for a total of 737,280 tests. The sets were formed subject to the following three conditions for a color test image of size 1024×1024: 1) The first set, called the Fast Set, only contains top performing FR methods that require less than 1.5 seconds to determine the quality of the test image. 2) The second set, called the Medium Set, contains top performing FR methods that take less than 2.7 seconds to determine the quality of the test image. 3) The final set, called the Full Set, has no time constraints. The FR methods in each of the three sets for the RAS exhaustive search are given in Table 8. Based on weighted average SRCC, top performing FR method combinations were selected in each set.

Instead of just computing the weighted average SRCC across all nine subject-rated databases, we compute weighted average SRCC for three categories: 1) Across all databases, 2) Across only the five single distortion databases, and 3) Across only the four multiple distortion databases. This was done to more thoroughly analyze how the performance of RAS varies for these different conditions. Within each category, all distorted images of the constituent databases were considered. These three categories were considered for each of the three sets of FR methods (Table 8), leading to a total of nine possibilities. The top performing combinations obtained in the exhaustive search for all these possibilities are given in Table 9, where each distinct combination is assigned a unique name (RAS1 to RAS7). The following observations can be made: 1) Although combinations of up to 15 FR methods were tested, the top performing combinations only include two to four FR methods in the fusion process. Thus, the notion of *the more the better* is not valid when it comes to fusion based FR methods. 2) The methods in each combination usually follow different design philosophies. While RAS4 and RAS5 differ in the base FR method, they combine the same three FR methods that include CID_MS [78] which follows a multiscale similarity based approach with emphasis on color features, SFF [96] which follows a sparsity based approach, and VSI [102] which follows a similarity based approach that incorporates visual saliency based weighted pooling. RAS2 combines a similarity based approach (VSI) with a sparsity based approach (SFF). RAS3 combines two similarity based approaches, DSS [79] (similarity in the DCT domain) and IWSSIM [87] (multiscale similarity measure that employs information content weighting in the pooling stage), with VIF_DWT [76] which follows a NSS based approach to IQA. RAS6 builds on RAS3 by adding CID_MS [78] to the combination which emphasizes on color based similarity. It is thus evident that RAS prefers combining different IQA design philosophies, such that they complement each other. The deficiencies in one design philosophy with regard to a particular distortion may be addressed by the strengths of another design approach. 3) RAS favors color based FR methods. All FR methods combined in RAS1, RAS2, RAS4, and RAS5, are color based, while RAS6 and RAS7 combine both color and grayscale based methods. Only RAS3 combines exclusively grayscale based methods.

As stated earlier, for the *learning based fusion* method MMF [118] (discussed in Section III-B.2) we test its context free version called CF-MMF. Since MMF follows a supervised learning based approach using SVR, different sets of FR methods can be combined. For this work, we select

TABLE 9. RAS exhaustive search outcome for each search set and database category.

Search Set	Database Category	Individual FR Methods included in Fusion				Base FR Method	Name Given to Fused FR Method
		Method 1	Method 2	Method 3	Method 4		
Fast	All Databases	FSIMc	SFF	VSI	–	SFF	RAS1
	Single Distortion Databases	SFF	VSI	–	–	VSI	RAS2
	Multiple Distortion Databases	DSS	IWSSIM	VIF_DWT	–	DSS	RAS3
Medium	All Databases[a]	CID_MS	SFF	VSI	–	CID_MS	RAS4
	Single Distortion Databases[b]	CID_MS	SFF	VSI	–	VSI	RAS5
	Multiple Distortion Databases	CID_MS	DSS	IWSSIM	VIF_DWT	DSS	RAS6
Full	All Databases[a]	CID_MS	SFF	VSI	–	CID_MS	RAS4
	Single Distortion Databases[b]	CID_MS	SFF	VSI	–	VSI	RAS5
	Multiple Distortion Databases	CID_MS	DSS	VIF	–	VIF	RAS7

[a]Exhaustive Search for the *All Databases* category leads to the same outcome for the Medium and Full Sets (RAS4).
[b]Exhaustive Search for the *Single Distortion Databases* category leads to the same outcome for the Medium and Full Sets (RAS5).

TABLE 10. Fused FR methods information table.

Fused FR Method	No. of Methods Fused	Individual FR Methods included in Fusion						Notes	
		Method 1	Method 2	Method 3	Method 4	Method 5	Method 6		
CISI	3	MSSSIM	VIF	FSIMc	–	–	–	–	
CM3	3	IFC	NQM	VSNR	–	–	–	–	
CM4	4	IFC	NQM	VSNR	VIF	–	–	–	
CNNM	6	FSIMc	PSNR_HMAc	PSNR_HVS	SFF	SRSIM	VIF	–	
HFSIMc	2	RFSIM	FSIMc	–	–	–	–	–	
MMF1	6	FSIM	IFC	MAD	MSSSIM	PSNR_HVS	VIF	Selection Method: SFMS	
MMF2	6	VSI	ADM	VIF_DWT	MCSD	IWSSIM	SFF		
MMF3	6	VSI	ADM	VIF_DWT	CID_MS	GMSD	SRSIM		
MMF4	6	VSI	ADM	CID_MS	MCSD	GMSD	IWSSIM		
RAS_B1	5	FSIM	FSIMc	GMSD	IWSSIM	VIF	–		GMSD
RAS_B2	4	FSIM	FSIMc	GMSD	IWSSIM	–	–		GMSD
RAS_MMF1	6	FSIM	IFC	MAD	MSSSIM	PSNR_HVS	VIF		FSIM
RAS_MMF2	6	VSI	ADM	VIF_DWT	MCSD	IWSSIM	SFF		VSI
RAS_MMF3	6	VSI	ADM	VIF_DWT	CID_MS	GMSD	SRSIM		VSI
RAS_MMF4	6	VSI	ADM	CID_MS	MCSD	GMSD	IWSSIM		VSI
RAS1	3	FSIMc	SFF	VSI	–	–	–	Base FR	SFF
RAS2	2	SFF	VSI	–	–	–	–		VSI
RAS3	3	DSS	IWSSIM	VIF_DWT	–	–	–		DSS
RAS4	3	SFF	CID_MS	VSI	–	–	–		CID_MS
RAS5	3	SFF	CID_MS	VSI	–	–	–		VSI
RAS6	4	DSS	IWSSIM	CID_MS	VIF_DWT	–	–		DSS
RAS7	3	CID_MS	DSS	VIF	–	–	–		VIF

TABLE 11. Test results of 22 fused FR methods on nine subject-rated IQA databases in terms of PLCC. All distortions in each dataset were considered.

Fused FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
CISI	0.9625	0.8575	0.9364	0.9289	0.8239	0.8193	0.9220	0.7095	0.9032	0.9007
CM3	0.8337	0.6058	0.6870	0.9435	0.6383	0.7238	0.6362	0.4604	0.8718	0.8062
CM4	0.8072	0.5891	0.6597	0.9432	0.6238	0.7086	0.6128	0.5622	0.8422	0.8219
CNNM	0.8892	0.9338	0.9007	0.8741	0.6439	0.6548	0.8328	0.6378	0.8249	0.7665
HFSIMc	0.9579	0.8635	0.9304	0.9211	0.7365	0.8120	0.8357	0.5242	0.8918	0.8893
MMF1	0.8561	0.9504	0.9202	0.8624	0.7326	0.7572	0.8185	0.6736	0.8523	0.8075
MMF2	0.8887	0.9512	0.9120	0.8608	0.6437	0.5736	0.8416	0.6056	0.8678	0.7964
MMF3	0.8831	0.9516	0.9274	0.8463	0.6186	0.6962	0.8692	0.7012	0.8241	0.8047
MMF4	0.8818	0.9532	0.9394	0.8681	0.6392	0.7131	0.8751	0.5785	0.7911	0.8391
RAS_B1	0.9683	0.8582	0.9539	0.9241	0.8255	0.8299	0.9177	0.7991	0.8980	0.9030
RAS_B2	0.9647	0.8701	0.9408	0.9255	0.7905	0.8350	0.9030	0.7730	0.8958	0.9041
RAS_MMF1	0.9696	0.8273	0.9622	0.9284	0.8521	0.8097	0.9059	0.7850	0.9034	0.9108
RAS_MMF2	0.9662	0.8596	0.9604	0.9200	0.8502	0.8436	0.9059	0.7643	0.8990	0.9084
RAS_MMF3	0.9638	0.8616	0.9568	0.9384	0.8569	0.8646	0.9104	0.7080	0.9015	0.9121
RAS_MMF4	0.9620	0.8815	0.9420	0.9409	0.8412	0.8719	0.9039	0.7620	0.9023	0.9187
RAS1	0.9659	0.8958	0.9567	0.9008	0.7995	0.8427	0.8958	0.7160	0.8945	0.9006
RAS2	0.9617	0.9003	0.9514	0.8930	0.7983	0.8387	0.8873	0.7100	0.8896	0.8897
RAS3	0.9701	0.8423	0.9660	0.9266	0.8583	0.8313	0.9262	0.8424	0.9111	0.9111
RAS4	0.9590	0.8912	0.9348	0.9383	0.8555	0.8739	0.8986	0.7173	0.9108	0.9167
RAS5	0.9588	0.8980	0.9406	0.9313	0.8462	0.8713	0.8999	0.7119	0.9048	0.9124
RAS6	0.9682	0.8488	0.9640	0.9408	0.8832	0.8585	0.9294	0.8181	0.9150	0.9202
RAS7	0.9687	0.8320	0.9670	0.9393	0.8788	0.8428	0.9434	0.8189	0.9144	0.9167

the version of CF-MMF recommended for the TID2008 database [4] through the sequential forward method selection (SFMS) strategy in [118], where it is computed that for this dataset, six FR methods should be combined. For the said version of CF-MMF [118], the six FR methods that are part of the fusion process are: FSIM [83], VIF [100], IFC [86], MAD [6], PSNR_HVS [92], and MSSSIM [89]. Since a pre-trained version of this model is not available, we follow the approach in [118] and train the model ourselves through SVR with a radial basis function (RBF). Instead

TABLE 12. Test results of 22 fused FR methods on nine subject-rated IQA databases in terms of SRCC. All distortions in each dataset were considered.

Fused FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
CISI	0.9680	0.8150	0.9425	0.9270	0.8231	0.8063	0.9135	0.6920	0.8740	0.8612
CM3	0.9207	0.7136	0.8073	0.9450	0.6452	0.7659	0.7114	0.5055	0.9206	0.7733
CM4	0.9316	0.7195	0.8247	0.9441	0.6417	0.7686	0.7661	0.6209	0.9224	0.7891
CNNM	0.8928	0.9201	0.8850	0.8763	0.6270	0.6451	0.8218	0.6720	0.8048	0.7260
HFSIMc	0.9610	0.8228	0.9423	0.9205	0.7315	0.7982	0.8202	0.5075	0.8624	0.8453
MMF1	0.8741	0.9409	0.9043	0.8594	0.7241	0.7379	0.8084	0.6799	0.8085	0.7703
MMF2	0.8907	0.9436	0.8910	0.8448	0.5720	0.5318	0.8196	0.6111	0.8533	0.7785
MMF3	0.8947	0.9455	0.9303	0.8345	0.6286	0.6517	0.8580	0.6606	0.7253	0.7265
MMF4	0.8852	0.9452	0.9438	0.8685	0.5990	0.6422	0.8596	0.6004	0.7533	0.8123
RAS_B1	0.9690	0.8034	0.9563	0.9223	0.8252	0.8312	0.9089	0.8013	0.8688	0.8595
RAS_B2	0.9653	0.8116	0.9464	0.9235	0.7917	0.8319	0.8932	0.7741	0.8654	0.8580
RAS_MMF1	0.9717	0.7355	0.9607	0.9268	0.8536	0.8133	0.8984	0.7923	0.8756	0.8720
RAS_MMF2	0.9689	0.8158	0.9642	0.9185	0.8490	0.8392	0.8952	0.7686	0.8714	0.8635
RAS_MMF3	0.9663	0.8195	0.9610	0.9383	0.8573	0.8583	0.9010	0.7132	0.8733	0.8699
RAS_MMF4	0.9642	0.8350	0.9493	0.9404	0.8430	0.8662	0.8953	0.7698	0.8730	0.8763
RAS1	0.9672	0.8756	0.9602	0.8958	0.7986	0.8375	0.8857	0.7205	0.8675	0.8593
RAS2	0.9637	0.8876	0.9591	0.8902	0.7975	0.8313	0.8759	0.7164	0.8589	0.8473
RAS3	0.9712	0.7794	0.9625	0.9261	0.8575	0.8271	0.9204	0.8455	0.8842	0.8796
RAS4	0.9590	0.8819	0.9422	0.9395	0.8562	0.8638	0.8913	0.7230	0.8836	0.8914
RAS5	0.9599	0.8864	0.9471	0.9313	0.8471	0.8632	0.8920	0.7168	0.8770	0.8822
RAS6	0.9680	0.7930	0.9603	0.9405	0.8840	0.8532	0.9250	0.8214	0.8867	0.8954
RAS7	0.9687	0.7724	0.9606	0.9388	0.8788	0.8363	0.9397	0.8250	0.8886	0.8986

of the TID2008 database [4], we use its enhanced version TID2013 [5], and the above-mentioned six FR methods to learn a fusion model. Half of the TID2013 dataset is used for training, half for validation, and grid search is employed to ascertain optimal SVR parameters. We refer the corresponding model as MMF1.

To provide a more thorough comparison of MMF [118] with RAS [21], we train three other CF-MMF models, one each for the three FR method pools identified for RAS in Table 8. As computed in [118] six FR methods should be combined for the TID2008 database [4], and we follow this recommendation for TID2013 [5] as well. We use the SFMS strategy [118] to identify the methods to be fused for each of the three FR method pools (Table 8) and built the following three CF-MMF models: 1) MMF2 developed on the Fast Set combines VSI [102], ADM [77], VIF_DWT [76], MCSD [88], IWSSIM [87], and SFF [96]. 2) MMF3 developed on the Medium Set combines VSI [102], ADM [77], VIF_DWT [76], CID_MS [78], GMSD [84], and SRSIM [97]. 3) MMF4 developed on the Full Set combines VSI [102], ADM [77], CID_MS [78], MCSD [88], GMSD [84], and IWSSIM [87]. Training for each of these MMF models was done in a manner similar to MMF1 and data scaling was applied where necessary as recommended in [118]. In each case, eight FR methods were combined but performance gain beyond the combination of six methods was negligible, and hence we combine six methods in the final models. Since the combinations in the four CF-MMF methods are not being used in the seven RAS methods discussed earlier, to provide another comparison point between RAS and MMF, we construct four additional RAS models that use the same FR method combinations as in each of the CF-MMF models. Specifically, RAS_MMF1, RAS_MMF2, RAS_MMF3, and RAS_MMF4, use the RAS technique [21] to fuse the FR methods that are selected for combination in MMF1, MMF2, MMF3, and MMF4, respectively. For these additional RAS methods, the base FR method was selected as the first one identified by the SFMS strategy. The details of all the fusion based methods whose performance is being evaluated in this work, including the various versions of RAS and MMF, are given in Table 10. RAS_B1 and RAS_B2 are the versions of RAS discussed in [21]. Although overall, we are evaluating the performance of seven different fused FR techniques, it can be noted from Table 10 that we are considering four different versions of MMF and 13 different versions of RAS. Thus, in total, 22 fused FR methods are being evaluated in this work.

D. PERFORMANCE OF FUSED FR METHODS ON INDIVIDUAL DATABASES

We tested the performance of the 22 fused FR methods mentioned in Table 10 on each of the nine subject-rated databases mentioned in Table 2 (CIDIQ database [31] at two viewing distances). The test results are given in Table 11 in terms of PLCC and in Table 12 in terms of SRCC. Testing was done on all distortion types included in each dataset.

E. OVERALL PERFORMANCE

Since we are evaluating the performance of IQA methods on nine different databases, a measure of overall performance is necessary. We provide this measure by computing the weighted average PLCC and SRCC values for each IQA method over different databases (as in [87]). The weight assigned to a database depends on its size in terms of the number of distorted images. The weighted average PLCC and SRCC for an IQA method over different databases are computed as:

$$PLCC_{WA} = \frac{\sum_{i=1}^{D} n_i \cdot PLCC_i}{\sum_{i=1}^{D} n_i} \tag{17}$$

$$SRCC_{WA} = \frac{\sum_{i=1}^{D} n_i \cdot SRCC_i}{\sum_{i=1}^{D} n_i} \tag{18}$$

TABLE 13. Weighted average PLCC and SRCC values of individual and fused FR methods for the three cases of: 1) All databases, 2) Single distortion databases, and 3) Multiple distortion databases. Methods in each case are sorted in descending order with respect to PLCC/SRCC values. Fused FR Methods are highlighted in bold.

Part 1: All Databases				Part 2: Single Distortion Databases				Part 3: Multiple Distortion Databases			
FR Method	PLCC	FR Method	SRCC	FR Method	PLCC	FR Method	SRCC	FR Method	PLCC	FR Method	SRCC
RAS5	0.8985	**RAS4**	0.8907	**RAS5**	0.9054	**RAS5**	0.9003	**RAS7**	0.9199	**RAS7**	0.9106
RAS6	0.8979	**RAS5**	0.8903	**RAS4**	0.9034	**RAS4**	0.8992	**RAS6**	0.9136	**RAS6**	0.9016
RAS4	0.8977	**RAS1**	0.8783	**RAS_MMF4**	0.8988	**RAS2**	0.8909	**RAS3**	0.9117	**RAS3**	0.8976
RAS_MMF4	0.8967	**RAS2**	0.8777	**RAS1**	0.8969	**RAS1**	0.8872	VIF	0.9064	VIF	0.8925
RAS7	0.8935	**RAS_MMF4**	0.8771	**RAS2**	0.8965	VSI	0.8847	**RAS_B1**	0.8990	**RAS_B1**	0.8801
RAS_MMF3	0.8912	**RAS6**	0.8761	**RAS_MMF3**	0.8925	**RAS_MMF4**	0.8783	IWSSIM	0.8970	IWSSIM	0.8785
RAS3	0.8911	**RAS_MMF3**	0.8724	**RAS6**	0.8905	**MMF1**	0.8773	**RAS_MMF1**	0.8942	**RAS_MMF1**	0.8778
RAS1	0.8908	**RAS7**	0.8710	**RAS_MMF2**	0.8879	QASD	0.8741	**RAS_MMF4**	0.8925	**RAS_MMF4**	0.8745
RAS_MMF2	0.8888	**RAS_MMF2**	0.8690	VSI	0.8855	**RAS_MMF3**	0.8735	**CISI**	0.8921	**RAS4**	0.8728
RAS_B1	0.8881	**RAS3**	0.8665	**MMF1**	0.8848	FSIMc	0.8700	**RAS_MMF2**	0.8908	**CISI**	0.8723
RAS2	0.8879	**RAS_B1**	0.8653	**RAS_B2**	0.8832	**RAS_MMF2**	0.8680	**RAS_B2**	0.8887	**RAS_MMF2**	0.8710
RAS_B2	0.8850	**CISI**	0.8634	QASD	0.8830	**MMF3**	0.8641	**RAS_MMF3**	0.8885	**RAS_MMF3**	0.8701
CISI	0.8831	VSI	0.8631	**RAS_B1**	0.8830	**RAS6**	0.8640	**RAS4**	0.8859	**RAS5**	0.8693
IWSSIM	0.8787	FSIMc	0.8628	**RAS3**	0.8813	**MMF4**	0.8634	**RAS5**	0.8841	**RAS_B2**	0.8684
RAS_MMF1	0.8786	**RAS_B2**	0.8607	**RAS7**	0.8810	**CISI**	0.8592	DVICOM	0.8799	DVICOM	0.8634
FSIMc	0.8785	IWSSIM	0.8559	FSIMc	0.8809	**RAS_B1**	0.8583	DVICOM_F	0.8794	DVICOM_F	0.8631
DSS	0.8757	SFF	0.8527	**CISI**	0.8788	SFF	0.8572	**RAS1**	0.8781	DSS	0.8630
VSI	0.8707	DSS	0.8520	**MMF4**	0.8777	**RAS_B2**	0.8570	DSS	0.8774	**RAS1**	0.8596
MCSD	0.8705	QASD	0.8482	ESSIM	0.8754	CID_MS	0.8536	VIF_DWT	0.8757	VIF_DWT	0.8564
FSIM	0.8687	**MMF1**	0.8479	DSS	0.8749	**RAS7**	0.8523	FSIMc	0.8735	IFC	0.8530
ESSIM	0.8674	MCSD	0.8464	MCSD	0.8724	**RAS3**	0.8517	FSIM	0.8721	**RAS2**	0.8500
GMSD	0.8671	**RAS_MMF1**	0.8452	MAD	0.8719	**HFSIMc**	0.8509	GMSD	0.8710	FSIMc	0.8479
SFF	0.8658	**MMF4**	0.8449	**RAS_MMF1**	0.8712	ESSIM	0.8493	**RAS2**	0.8698	GMSD	0.8458
DVICOM	0.8631	CID_MS	0.8445	IWSSIM	0.8700	**CNNM**	0.8490	MCSD	0.8666	FSIM	0.8452
DVICOM_F	0.8631	GMSD	0.8433	**MMF3**	0.8697	MCSD	0.8484	SSIM_DWT	0.8650	SFF	0.8433
QASD	0.8625	FSIM	0.8430	**HFSIMc**	0.8696	DSS	0.8467	SFF	0.8643	MCSD	0.8422
SRSIM	0.8616	ESSIM	0.8418	FSIM	0.8671	IWSSIM	0.8452	WSSI	0.8608	SSIM_DWT	0.8385
MMF4	0.8602	DVICOM_F	0.8394	SFF	0.8665	GMSD	0.8421	DWT_VIF	0.8598	DWT_VIF	0.8368
MMF1	0.8593	**MMF3**	0.8392	SRSIM	0.8654	FSIM	0.8419	IFC	0.8567	WSSI	0.8338
MMF3	0.8569	DVICOM	0.8387	GMSD	0.8653	MAD	0.8413	SRSIM	0.8535	VIF_P	0.8336
GSIM	0.8553	SRSIM	0.8347	GSIM	0.8619	GSIM	0.8401	VIF_P	0.8514	SRSIM	0.8264
HFSIMc	0.8550	**HFSIMc**	0.8345	**CNNM**	0.8595	**MMF2**	0.8399	ESSIM	0.8506	ESSIM	0.8259
MSSSIM	0.8537	CID_SS	0.8325	**MMF2**	0.8592	MSSSIM	0.8386	MSSSIM	0.8440	CID_MS	0.8253
ADM	0.8536	MSSSIM	0.8323	ADM	0.8590	SRSIM	0.8386	CID_SS	0.8434	CID_SS	0.8200
MAD	0.8516	ADM	0.8308	MSSSIM	0.8583	CID_SS	0.8385	ADM	0.8423	MSSSIM	0.8191
CID_MS	0.8511	GSIM	0.8307	CID_MS	0.8570	ADM	0.8384	GSIM	0.8414	VSI	0.8177
CID_SS	0.8452	**CNNM**	0.8270	DVICOM_F	0.8553	PSNR_HAc	0.8361	VSI	0.8395	ADM	0.8149
SSIM_DWT	0.8436	**MMF2**	0.8248	DVICOM	0.8551	PSNR_HMAc	0.8352	CID_MS	0.8386	GSIM	0.8111
MMF2	0.8434	MAD	0.8220	CID_SS	0.8460	**RAS_MMF1**	0.8297	**MMF3**	0.8298	**MMF4**	0.8060
CNNM	0.8389	SSIM_DWT	0.8137	PSNR_HAc	0.8425	DVICOM_F	0.8281	**HFSIMc**	0.8243	**HFSIMc**	0.7999
VIF	0.8388	WSSI	0.8069	RFSIM	0.8393	DVICOM	0.8270	**MMF4**	0.8236	QASD	0.7936
WSSI	0.8384	SSIM	0.8029	PSNR_HMAc	0.8391	RFSIM	0.8112	SSIM	0.8230	**MMF2**	0.7930
SSIM	0.8271	PSNR_HMAc	0.8028	SSIM_DWT	0.8335	SSIM	0.8080	QASD	0.8195	SSIM	0.7923
VIF_DWT	0.8220	VIF	0.8024	PSNR_HA	0.8315	PSNR_HA	0.8057	**MMF2**	0.8100	**MMF3**	0.7868
PSNR_HMAc	0.8153	PSNR_HAc	0.7942	SSIM	0.8290	SSIM_DWT	0.8020	MAD	0.8089	**MMF1**	0.7859
PSNR_HAc	0.8094	VIF_DWT	0.7768	WSSI	0.8278	PSNR_HMA	0.7952	**MMF1**	0.8057	MAD	0.7812
PSNR_HMA	0.8080	PSNR_HMA	0.7762	PSNR_HMA	0.8219	WSSI	0.7941	**CNNM**	0.7955	**CNNM**	0.7808
PSNR_HA	0.8061	**CM4**	0.7758	VIF	0.8066	**CM4**	0.7743	UQI	0.7875	**CM4**	0.7791
VIF_P	0.8059	PSNR_HA	0.7747	VIF_DWT	0.7965	**CM3**	0.7682	PSNR_HMA	0.7789	UQI	0.7673
RFSIM	0.8055	RFSIM	0.7740	VIF_P	0.7843	VIF	0.7596	PSNR_HMAc	0.7653	PSNR_HMA	0.7363
DWT_VIF	0.8032	**CM3**	0.7575	DWT_VIF	0.7763	IW_PSNR	0.7501	IW_PSNR	0.7652	**CM3**	0.7350
PSNR_HVS	0.7402	DWT_VIF	0.7531	VSNR	0.7492	VSNR	0.7410	PSNR_HA	0.7527	PSNR_HMAc	0.7347
PSNR_HVSM	0.7364	VIF_P	0.7526	PSNR_HVS	0.7467	VIF_DWT	0.7390	PSNR_HVSM	0.7466	IW_PSNR	0.7306
VSNR	0.7335	IW_PSNR	0.7438	PSNR_DWT	0.7323	PSNR_HVS	0.7295	PSNR_HAc	0.7399	PSNR_HA	0.7095
IW_PSNR	0.7263	VSNR	0.7174	PSNR_HVSM	0.7315	VIF_P	0.7142	RFSIM	0.7343	PSNR_HAc	0.7060
PSNR_DWT	0.7244	PSNR_HVS	0.7136	PSNR	0.7180	PSNR_HVSM	0.7141	PSNR_HVS	0.7264	PSNR_HVSM	0.7010
UQI	0.7226	PSNR_HVSM	0.7099	NQM	0.7136	DWT_VIF	0.7134	AD_DWT	0.7087	RFSIM	0.6958
NQM	0.7022	PSNR_DWT	0.6944	IW_PSNR	0.7078	PSNR_DWT	0.7076	PSNR_DWT	0.7076	AD_DWT	0.6924
PSNR	0.6927	IFC	0.6924	SNR	0.7043	PSNR	0.7066	VSNR	0.7003	PSNR_HVS	0.6801
CM3	0.6893	AD_DWT	0.6875	UQI	0.6918	NQM	0.6949	**CM3**	0.6927	VSNR	0.6677
WSNR	0.6820	UQI	0.6829	**CM3**	0.6876	SNR	0.6923	**CM4**	0.6908	PSNR_DWT	0.6665
SNR	0.6819	NQM	0.6801	WSNR	0.6824	AD_DWT	0.6852	WSNR	0.6813	NQM	0.6488
CM4	0.6768	PSNR	0.6720	**CM4**	0.6701	WSNR	0.6717	NQM	0.6782	WSNR	0.6341
AD_DWT	0.6054	SNR	0.6606	AD_DWT	0.5563	UQI	0.6428	PSNR	0.6396	PSNR	0.5992
IFC	0.5789	WSNR	0.6596	IFC	0.4470	IFC	0.6161	SNR	0.6347	SNR	0.5938

where $PLCC_i$ and $SRCC_i$ are respectively the PLCC and SRCC values of the IQA method for database i, n_i is the number of images in database i, and D is the number of databases being considered. Although we are using nine IQA databases in this work (five singly distorted and four multiply distorted), since the singly distorted database CIDIQ [31] provides MOS at two viewing distances, it will be regarded as two datasets. We compute weighted average PLCC and SRCC for three cases: 1) All databases ($D = 10$), 2) Only single distortion databases ($D = 6$), and 3) Only multiple distortion databases ($D = 4$). Information about the number of distorted images in each dataset is provided in Table 2.

TABLE 14. Kurtosis based check for Gaussianity of prediction residuals of individual and fused FR methods. A "1" means that the kurtosis of the residuals is between 2 and 4, and they can be assumed to be Gaussian distributed. A "0" means that the kurtosis of residuals is not between 2 and 4, and they are assumed to be non-Gaussian. Fused FR Methods are highlighted in bold.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
RAS5	1	1	0	1	1	1	1	1	1	1
RAS6	1	1	1	1	1	1	1	1	1	1
RAS4	1	1	1	1	1	1	1	1	1	1
RAS_MMF4	0	0	0	1	1	1	1	1	1	1
RAS7	1	1	1	1	1	1	0	1	1	1
RAS_MMF3	1	0	0	1	1	1	0	1	1	1
RAS3	1	1	1	1	1	1	1	1	1	1
RAS1	0	1	0	1	1	1	1	1	1	1
RAS_MMF2	1	0	0	1	1	1	1	1	1	1
RAS_B1	0	1	1	1	1	1	1	1	1	1
RAS2	1	1	0	1	1	1	1	1	1	1
RAS_B2	0	0	1	1	1	1	1	1	1	1
CISI	1	1	0	1	1	1	1	1	1	1
IWSSIM	0	0	1	1	1	1	1	1	1	1
RAS_MMF1	1	0	1	1	0	1	1	1	1	1
FSIMc	0	0	0	1	1	1	1	1	1	1
DSS	0	0	1	0	1	0	1	1	1	0
VSI	1	0	0	1	1	1	1	1	1	1
MCSD	0	0	0	1	1	0	1	1	1	0
FSIM	0	0	1	1	1	1	1	1	1	1
ESSIM	0	0	0	1	1	1	1	1	1	1
GMSD	0	0	0	1	1	0	1	1	1	0
SFF	1	1	0	1	1	1	0	1	1	1
DVICOM	1	0	1	1	0	1	1	1	1	1
DVICOM_F	1	0	1	1	0	1	1	1	1	1
QASD	1	0	0	1	1	1	1	1	1	1
SRSIM	1	0	1	1	1	1	1	1	1	1
MMF4	1	0	1	1	1	1	0	1	1	1
MMF1	0	0	0	1	1	1	1	1	1	0
MMF3	0	0	1	1	1	1	1	1	1	1
GSIM	1	0	0	1	1	1	1	1	1	1
HFSIMc	1	0	0	1	1	1	1	1	1	1
MSSSIM	1	0	1	1	1	1	1	1	1	1
ADM	0	0	1	1	1	1	1	1	1	1
MAD	0	0	1	1	1	1	1	1	1	1
CID_MS	1	1	1	1	1	1	1	1	1	1
CID_SS	1	1	1	1	1	1	1	1	1	1
SSIM_DWT	1	1	1	1	0	1	0	1	1	1
MMF2	0	0	0	1	1	1	1	1	1	1
CNNM	0	0	0	1	1	0	0	1	1	1
VIF	1	1	1	1	1	1	1	1	1	1
WSSI	1	1	1	1	0	1	0	1	1	1
SSIM	1	1	0	1	1	1	1	1	1	1
VIF_DWT	1	1	1	1	1	1	1	1	1	1
PSNR_HMAc	1	0	1	1	1	1	1	1	1	1
PSNR_HAc	1	0	1	1	1	1	1	1	1	1
PSNR_HMA	0	0	1	0	1	1	1	1	1	1
PSNR_HA	1	0	1	1	1	1	1	1	1	1
VIF_P	1	1	1	1	1	1	0	1	1	1
RFSIM	1	1	0	1	1	1	1	1	1	1
DWT_VIF	1	1	1	1	1	1	0	1	1	1
PSNR_HVS	1	0	1	0	1	1	1	1	1	1
PSNR_HVSM	0	0	1	1	1	1	1	1	1	1
VSNR	1	0	0	1	1	1	1	1	1	1
IW_PSNR	0	0	1	1	1	1	1	1	1	1
PSNR_DWT	0	0	1	1	1	1	1	1	1	1
UQI	1	1	1	1	1	1	1	1	1	1
NQM	0	0	1	1	1	1	1	1	1	0
PSNR	1	0	1	1	1	1	1	1	1	1
CM3	1	0	1	1	1	1	1	1	0	1
WSNR	0	0	1	1	1	1	1	1	1	1
SNR	1	0	1	1	1	1	1	1	1	1
CM4	1	0	1	1	1	1	1	1	1	1
AD_DWT	0	1	0	1	1	1	1	1	1	1
IFC	1	1	1	1	1	1	1	1	1	1

All distorted images in each database, regardless of distortion type, have been used for the computation of PLCC and SRCC values.

While determining the overall performance, we consider the 43 individual FR methods (Table 3) and the 22 fused FR methods (Table 10) together, in order to observe if fused FR methods offer any benefits over individual methods, and if so, then by how much. Table 13 depicts the overall performance of the 65 methods in terms of weighted average PLCC and SRCC, where parts 1, 2, and 3 of the table correspond to the cases of all databases, single distortion databases, and multiple distortion databases, respectively. Within each case, the methods have been sorted in the descending order with respect to the weighted average PLCC and SRCC values. Therefore, the best performing methods for each case are towards the top of the table, while methods at the bottom of the table have the worst performance for that case. The names of the fused FR methods are mentioned in bold, in order to distinguish them from the individual FR methods.

F. STATISTICAL SIGNIFICANCE TESTING

We carried out statistical significance testing in accordance with the description given in Section IV-A.3. First, a Kurtosis based check for Gaussianity was performed on the prediction residuals of all 65 individual and fused FR methods on all the datasets. The outcome of this test is presented in Table 14, where a "1" means that the kurtosis of the residuals is between 2 and 4, while a "0" means that it is outside of this range. The prediction residuals are assumed to be Gaussian in the former case, while they are not in the latter. While doing this test, all distorted images within each dataset were considered. It can be seen from Table 14 that the kurtosis based assumption of Gaussianity of prediction residuals holds in most cases (around 82% cases). Next, the prediction residuals of all methods were compared by making all possible pairs of individual and fused FR methods, and carrying out hypothesis testing through the one-sided (left-tailed) two-sample F-test at 95% confidence (see Section IV-A.3).

Table 15 provides the outcome of statistical significance testing for 16 of the 22 fused FR methods. These methods include all four methods belonging to the *empirical fusion* category (HFSIMc [117], CISI [114], CM3 [115], and CM4 [115]). We include both methods of the *learning based fusion* category (CNNM [116] and MMF [118], [121]). As discussed in Section IV-C, we tested four versions of MMF. Here, we include the top three MMF versions that have the highest weighted average PLCC for the *All Databases* case (see Table 13). These versions are MMF1, MMF3, and MMF4. Of the 13 versions of RAS [21], which belongs to the *rank aggregation based fusion* category, we selected the following eight versions: Among the seven RAS versions found through the exhaustive search procedure in Section IV-C and listed in Table 9, the top four RAS versions that have the highest weighted average PLCC for the *All Databases* case (see Table 13) were selected. These versions include RAS4, RAS5, RAS6, and RAS7. The three RAS versions corresponding to the MMF versions included above were also selected (RAS_MMF1, RAS_MMF3, and RAS_MMF4).

TABLE 15. Statistical significance testing results of fused FR methods based on prediction residuals. Each entry is a codeword composed of ten symbols, where each symbol represents the test outcome for one IQA database. The symbol location within a codeword represents IQA databases in the following order: [LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ50, CIDIQ100, MDID, MDID2013, LIVE MD, MDIVL]. A "1" means that the method in the row, for a particular database, is statistically better than the method in the column, a "0" means that it is statistically worse, while a "_" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence).

	RAS5	RAS6	RAS4	RAS_MMF4	RAS7	RAS_MMF3	RAS_B1	CISI	RAS_MMF1	MMF4	MMF1	MMF3	HFSIMc	CNNM	CM3	CM4
RAS5	_____	01000_00__	_1_____	_1_0____	010_0100__	010___0___	010__100__	_1___10___	010__1_0__	10_1111111	1011111_11	1011111_11	_11_1111_1	1011111111	1110111111	1110111111
RAS6	10111_11__	_____	101_1_11__	101_1_11__	_1____0___	101_1_11__	__11111_11	1_111111_1	_1_1111___	1011111111	1011111111	1011111111	1011111111	1011111111	111_111111	111_111111
RAS4	_0_____	010_0_00__	_____	_10_____	010_0100__	010___0___	01011100_1	_1__110__1	0101_1_0__	10_1111111	1011111_11	10_1111_11	_1_1111111	1011111111	111_111111	111_111111
RAS_MMF4	_0_1_____	010_0_00__	_01_____	_____	010_0100__	_10_____	0101_10__1	_1_1_10__1	0101_1____	10_1111111	1011111111	1011111111	_1111111_1	1011111111	111_111111	111_111111
RAS7	101_1011__	_0____1___	101_1011__	101_1011__	_____	101_1011__	_0111_1_11	10111111_1	__11111___	1011111111	1011111111	1011111111	1011111111	1011111111	111_111111	111_111111
RAS_MMF3	101___1___	010_0_00__	101___1___	_01_____	010_0100__	_____	0__111_0__	__1_110___	0101_1_0__	1011111111	1011111_11	1011111_11	1_111111_1	1011111_11	111_111111	111_111111
RAS_B1	101__011__	__00000_00	10100011_0	1010_01__0	_1000_0_00	1__000_1__	_____	1_1_____1__	_10_0_1___	1011111111	1011111111	1011111111	1_1_1_11_1	1011111111	1110111111	1110111111
CISI	_0___01___	0_000000_0	_0__001__0	_0_0_01__0	01000000_0	__0_001___	0_0____0__	_____	010_0_10__	10_1111111	1011111_11	1011111_11	____1_11__	1011111_11	1110111111	1110111111
RAS_MMF1	101__0_1__	_0_0000___	1010_0_1__	1010_0____	__00000___	1010_0_1__	_01_1_0__	101_1_01__	_____	1011111111	1011111111	1011111111	101_1_11_1	1011111111	1110111111	1110111111
MMF4	01_0000000	0100000000	01_0000000	01_0000000	0100000000	0100000000	0100000000	01_0000000	0100000000	_____	1_1_001001	__11___0_1	0110001_00	_11__11_01	1110__1_11	1110__1_0_
MMF1	0100000_00	0100000000	0100000_00	0100000000	0100000000	0100000_00	0100000000	0100000_00	0100000000	0_0_110110	_____	0___110_1_	0100_00100	011_11__11	11101_111_	11101111__
MMF3	0100000_00	0100000000	01_0000_00	0100000000	0100000000	0100000_00	0100000000	0100000_00	0100000000	__00___1_0	1___001_0_	_____	01_0001100	_110__1__1	1110__111_	1110__11__
HFSIMc	_00_0000_0	0100000000	_0_0000_0	_0000000_0	0100000000	0_000000_0	0_0_0_00_0	____0_00__	010_0_00_0	1001110_11	1011_11011	10_1110011	_____	101111_011	1110111_11	1110111_11
CNNM	0100000000	0100000000	0100000000	0100000000	0100000000	0100000_00	0100000000	0100000_00	0100000000	_00__00_10	100_00__00	_001__0__0	010000_100	_____	1110_01110	1110_01__0
CM3	0001000000	000_000000	000_000000	000_000000	000_000000	000_000000	0001000000	0001000000	0001000000	0001__0_00	00010_000_	0001___000_	0001000_00	0001_10001	_____	1_____0_
CM4	0001000000	000_000000	000_000000	000_000000	000_000000	000_000000	0001000000	0001000000	0001000000	0001__0_1_	00010000__	0001___00__	0001000_00	0001_10__1	0_____1_	_____

TABLE 16. Statistical significance testing results of individual FR methods based on prediction residuals. Each entry is a codeword composed of ten symbols, where each symbol represents the test outcome for one IQA database. The symbol location within a codeword represents IQA databases in the following order: [LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ50, CIDIQ100, MDID, MDID2013, LIVE MD, MDIVL]. A "1" means that the method in the row, for a particular database, is statistically better than the method in the column, a "0" means that it is statistically worse, while a "_" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence). Fused FR Methods RAS6 and MMF1 are included for comparison and are highlighted in bold.

	RAS6	IWSSIM	FSIMc	DSS	VSI	ESSIM	GMSD	SFF	DVICOM_F	QASD	**MMF1**	ADM	MAD	CID_MS	VIF	PSNR
RAS6	_____	11111_10_1	101_1_1111	1__1111__1	101_111111	1011111111	1011111_11	10_1111_11	0111111_11	1011111111	1011111111	11111111_1	__111_1111	111_111111	1111110__1	1111111111
IWSSIM	00000_01_0	_____	00_011_1__	000_1111__	_000111111	_____	000_111_11	00011111_1	01__11_111	_00_1111_1	10_1111111	1_0_1111__	000111111_	1_10_1111_	0101110__1	1111111111
FSIMc	010_0_0000	11_100_0__	_____	_10___10__	10____1__1	_1_10_1___	_101_110_1	__01_110_1	01_101_0_1	_00__11__1	10_1_11_11	1101__1___	0101__10__	111_011___	_1_1_100_1	1111111_11
DSS	0__0000__0	111_0000__	_01____01__	_____	101_1__111	1011__11__	__1__1__11	_0_1_11___	011_010__1	_01_1_11_1	1011111111	111___11__	0_11__11__	11100_11__	_111110__1	1111111111
VSI	010_000000	_111000000	01_____0__0	010_0__000	_____	01_10_10_0	0101_1_0__	010101_0_0	01_10_00__	010___1___	10_1_11011	11_10_1_00	01010_10_0	111_0_1__0	01_1_1000_	1111111_11
ESSIM	0100000000	_1__00000_	_0_01_0___	0100__00__	10_01_01_1	_____	__0_1100_1	0_01_100_1	01____00_1	_0001__1__1	10_1111_11	11_____0_	010___1___	11100_____	_1_11100_1	1111111111
GMSD	0100000_00	111_000_00	_010_001_0	__0__0__00	1010_0_1__	__1_0011_0	_____	_0010_1__0	011_000___	_01__011__	1011__1111	111_001100	01__0011_0	11100011_0	_111_10_0_	1111111111
SFF	01_0000_00	1110000000	__10_001_0	_1_0_00___	101010_1_1	1_10_011_0	_1101_0__1	_____	0110__0___	10101011__	10101_1111	1110_01100	0110_011__	11100011__	_1101100_1	1110111111
DVICOM_F	1000000_00	10__00_000	10_010_1_0	100_101__0	10_01_11__	10____11_0	100_111___	1001__1___	_____	100_1_11__	10_1111111	100___1100	100__011_0	10100_11_0	11_1110___	1111111111
QASD	0100000000	_11_0000_0	_11__00__0	_10_0_00_0	101___0___	_1110_0__0	_10__100__	01010100__	011_0_00__	_____	1011_1__11	01_10010_0	01_10010_0	11100_0___	_111_100__	1111111_11
MMF1	0100000000	01_0000000	01_0_00_00	0100000000	01_0_00100	01_0000_00	0100__0000	01010_0000	01_0000000	0100_0__00	_____	01_0000_00	0100001000	0110000_00	01_0_10000	_111111111
ADM	00000000_0	0_1_0000__	0010__0___	000___00__	00_01_0_11	00_____1_	000_110011	0001_10011	011___0011	000_1_1__1	10_1111_11	_____	000___10__	1_100_____	01_11100_1	1111111_11
MAD	__000_0000	111000000_	1010__01__	1_00__00__	10101_01_1	101___0___	10__1100_1	1001_100__	011__100_1	10_01101_1	1011110111	111___01__	_____	1_100101__	111_1100_1	1111111111
CID_MS	000_000000	0_01_0000_	000_100___	00011_00__	000_1_0__1	00011_____	00011100_1	00011100__	01011_00_1	00011_1___	1001111_11	0_011_____	0_011010__	_____	01011100_1	1111111_11
VIF	0000001__0	1010001__0	_0_0_011_0	_000001__0	10_0_0111_	_0_00011_0	_000_01_1_	_0010011_0	00_0001___	_000_011__	10_1_01111	10_00011_0	000_0011_0	10100011_0	_____	11111_1111
PSNR	0000000000	0000000000	0000000_00	0000000000	0000000_00	0000000000	0000000000	0001000000	0000000000	0000000_00	_000000000	0000000_00	0000000000	0000000_00	00000_0000	_____

Finally, RAS_B1, which is one of the original RAS versions in [21] is included as well.

Table 16 provides the outcome of statistical significance testing for 14 of the 43 individual FR methods. These methods were selected by analyzing the weighted average PLCC of the *All Databases* case in Table 13 and picking the top performing methods such that: A) The overall top four methods are selected which include IWSSIM [87], FSIMc [83], DSS [79], and VSI [102], all of which are *structural similarity based* approaches. B) There is representation from each of the four categories of individual FR methods discussed in Section III-A. PSNR is selected from the *error based methods* category. In addition to the four top performing methods (IWSSIM, FSIMc, DSS, and VSI), three additional methods, CID_MS [78], ESSIM [82], and GMSD [84] are selected from the *structural similarity based methods* category. VIF [100], SFF [96], and QASD [94] represent the *NSS based methods* category. Finally, ADM [77], MAD [6], and DVICOM_F [80], represent the *mixed strategy based methods* category. To help statistically compare individual FR methods with fused ones, two fused FR methods are also included in Table 16. These include MMF1 and RAS6, as representatives of learning based and rank aggregation based fusion, respectively.

Each entry in Tables 15 and 16 is a codeword composed of ten symbols. Each symbol represents the outcome of statistical significance testing for one IQA database. The location of the symbol in the codeword represents specific IQA databases in the following order, from left to right: LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ at viewing distance of 50 cm (CIDIQ50), CIDIQ at viewing distance of 100 cm (CIDIQ100), MDID, MDID2013, LIVE MD, and MDIVL. Each symbol can take one of three possibilities, a "1", "_", or "0" meaning that the method in the row is statistically (with 95% confidence) better, indistinguishable, or worse than the method in the column respectively, for a particular database. The order of methods in Tables 15 and 16 is based on their order in the weighted average PLCC portion of the *All Databases* case in Table 13.

G. COMPUTATIONAL COMPLEXITY

The computational complexity of all 43 FR IQA methods under test was evaluated in terms of their execution time to determine the quality of a 1024 × 1024 color image on a Lenovo laptop computer with a 2.4GHz Intel Core i7-4700MQ processor, 12GB of RAM, Samsung 850 EVO Solid State Drive, and Windows 10 Home operating system. The execution times of all FR methods are given in Table 17, where methods have been sorted in ascending order with respect to execution time. Since PSNR is the fastest method, we also provide the execution time relative to PSNR for convenience in comparison.

TABLE 17. Execution time of individual FR methods for a test image. Methods are sorted in ascending order with respect to the execution time.

FR Method	Execution Time (Seconds)	Execution Time (Relative to PSNR)
PSNR	0.0044 s	1.00
SNR	0.0107 s	2.43
GMSD	0.0293 s	6.66
SRSIM	0.0309 s	7.02
SSIM	0.0462 s	10.50
MCSD	0.0475 s	10.80
GSIM	0.0510 s	11.59
RFSIM	0.0578 s	13.14
ESSIM	0.1230 s	27.95
UQI	0.1652 s	37.55
MSSSIM	0.1958 s	44.50
WSNR	0.2002 s	45.50
SFF	0.2173 s	49.39
DSS	0.2196 s	49.91
WSSI	0.2402 s	54.59
VIF_DWT	0.2527 s	57.43
VIF_P	0.2546 s	57.86
VSI	0.2727 s	61.98
DWT_VIF	0.2851 s	64.80
FSIM	0.3210 s	72.95
FSIMc	0.3210 s	72.95
VSNR	0.3861 s	87.75
SSIM_DWT	0.4473 s	101.66
ADM	0.5897 s	134.02
PSNR_DWT	0.6578 s	149.50
AD_DWT	0.7790 s	177.05
NQM	0.9878 s	224.50
IW_PSNR	1.4777 s	335.84
IWSSIM	1.5670 s	356.14
PSNR_HVS	2.2685 s	515.57
PSNR_HVSM	2.2685 s	515.57
DVICOM_F	2.3753 s	539.84
CID_SS	2.7699 s	629.52
QASD	2.9208 s	663.82
CID_MS	3.1687 s	720.16
PSNR_HA	3.2619 s	741.34
PSNR_HMA	3.2619 s	741.34
VIF	4.5277 s	1029.02
IFC	4.5797 s	1040.84
MAD	5.4482 s	1238.23
DVICOM	6.9084 s	1570.09
PSNR_HAc	9.7606 s	2218.32
PSNR_HMAc	9.7606 s	2218.32

H. ANALYSIS AND DISCUSSION

Based on the results obtained in the previous sub-sections and in particular on Table 13 (Overall performance) and Tables 15 and 16 (Statistical Significance Testing), the following observations can be made.

1) INDIVIDUAL FR METHODS

Considering the top ten methods in each category and for each evaluation criterion in Table 13, it can be seen that most top performing methods, especially for the *all databases* and *single distortion databases* categories, belong to the structural similarity based class of FR methods. These methods include IWSSIM [87], FSIMc/FSIM [83], DSS [79], VSI [102], GMSD [84], MCSD [88], ESSIM [82], and CID_MS [78]. For these categories, the sparsity based NSS methods QASD [94] and SFF [96], and the mixed strategy based methods DVICOM [80] and MAD [6] also do well. For the *multiple distortion databases* category, the NSS methods VIF [100] and VIF_DWT [76], and the mixed strategy based method DVICOM/DIVICOM_F [80], do well in addition to the structural similarity based approaches. It can be observed from Table 13 that error based FR methods do not offer competitive performance against other IQA design philosophies. From Table 13 it can be seen that overall: 1) For the *all*

databases case, IWSSIM [87] is the top performing method in terms of weighted average PLCC, while VSI [102] is the top performer in terms of weighted average SRCC, 2) For the *single distortion databases* case, VSI [102] is the top performing method both in terms of weighted average PLCC and SRCC, and 3) For the *multiple distortion databases* case, VIF [100] is the top performer both in terms of weighted average PLCC and SRCC.

While using weighted average PLCC and SRCC is one way to determine overall performance, it has the drawback of favoring larger databases. Thus, in our case, the TID2013 database [5] is given the largest weight since it has the most images, while the MDID2013 database [12] is given the smallest weight. This is done even though both these databases contain entirely different distortion processes, where TID2013 contains images afflicted with a single distortion, while MDID2013 has images that have undergone three kinds of distortions. It is thus unfair to develop an opinion solely on the basis of weighted average PLCC and SRCC. Another way to compare methods and to determine which one is performing better than others, is to observe the statistical significance testing tables. From Table 16, we can observe that IWSSIM [87] is statistically better than most other methods on most of the databases. This shows that IWSSIM is a robust method that does well across different kinds of distortion types. The FR methods DSS [79] and FSIMc [83] follow IWSSIM in performance and do quite well when statistically compared to other FR methods.

2) FUSED FR METHODS

For all three cases of *all databases*, *single distortion databases*, and *multiple distortion databases*, it is clear from Table 13 that the rank aggregation based FR fusion technique RAS [21] significantly outperforms all other FR fusion techniques, in terms of both weighted average PLCC and SRCC. The same conclusion can be comprehensively drawn from Table 15, where it can be seen that RAS based methods are statistically superior than all other fusion based methods for the vast majority of datasets. Among the 13 RAS methods, it can be observed from Table 15 that statistically, RAS6 is overall the top performer, followed closely by RAS7. It can also be observed that RAS methods selected through the exhaustive search procedure described in Section IV-C, especially those belonging to the *medium* and *full* sets (Table 8) such as RAS6 and RAS7 respectively, perform better than the FR methods combination described in the original RAS work [21], thereby highlighting the importance of finding the set of FR methods to be fused through a more structured approach.

The two learning based fusion approaches, MMF [118] and CNNM [116] do not appear to be competitive when compared to the rank aggregation based approach, as can be seen from Table 13. It can be observed from Table 15 that the different MMF approaches (MMF1, MMF3, and MMF4) and CNNM perform better than the different RAS methods only on the TID2013 database [5]. However, as described in Section III-B.2, the MMF methods and CNNM, are all trained on this very database, and hence comparing these methods with other approaches on TID2013 is unreliable and unfair. On all other datasets, the MMF methods and CNNM are statistically outperformed by the RAS methods, which shows that learning based fusion approaches suffer from model overfitting issues. Since the four RAS methods RAS_MMF1, RAS_MMF2, RAS_MMF3, and RAS_MMF4 combine the same set of individual FR methods as the four MMF methods MMF1, MMF2, MMF3, and MMF4, respectively (see Table 10), the two FR fusion approaches can be directly compared. Since the TID2013 database was used to train the four MMF methods and it contributes the largest weight to the weighted average PLCC and SRCC computation, we avoid using these evaluation criteria. Instead we statistically compare these methods by using Table 15, where it can be seen that the MMF based methods are outperformed by their RAS counterparts on all datasets except TID2013. This again highlights the superiority of the rank aggregation based fusion, which does not involve any training, and hence does not suffer from model overfitting issues. By contrast, the learning based fusion approaches, even when they use one of the largest subject-rated dataset for training, suffer from overfitting issues because the number of distorted images per distortion type are quite small even in the TID2013 database [5] (only 125 images per distortion type).

The empirical fusion based methods CM3 and CM4 [115], described in Section III-B.1 and given in Equations 11 and 12 respectively, perform inadequately, even for multiply distorted content for which they are designed, as is evident from Part 3 of Table 13. This is because of the choice of FR methods that are being fused in CM3 and CM4, especially IFC [86], NQM [90], and VSNR [56], and the way in which exponent values are obtained on a single database (LIVE MD [11]). It can be observed from Tables 6 and 7 that while IFC, NQM, and VSNR, perform quite well on the LIVE MD database, their performance is lacking on other IQA datasets. This is further substantiated from Table 13 in terms of weighted average PLCC and SRCC. However, since the exponent values in Equations 11 and 12 are only optimized on LIVE MD database, CM3 and CM4 are highly database dependent. Thus, they perform well only on a few datasets (VCLFER [35] and LIVE MD [11]), while their performance on other datasets is inferior as can be observed in Tables 11 and 12. This highlights the pitfalls of: 1) the empirical fusion based approach which is rather ad hoc, 2) the selection of FR methods to be fused on the basis of a single dataset, and 3) the use of a single dataset for parameter tuning. It can be observed from Tables 13 and 15 that the empirical fusion based methods HFSIMc [117] and CISI [114], described in Section III-B.1 and given in Equations 9 and 10 respectively, perform better than CM3 and CM4. CISI also performs better than HFSIMc. Both these methods, especially CISI, perform statistically better than the learning based fusion methods (MMF and CNNM) as can be observed from Table 15. This performance gain is because HFSIMc and CISI, especially the latter, fuse FR methods that perform well across most

databases individually as well. However, even these empirical fusion methods cannot outperform rank aggregation based fusion methods.

3) INDIVIDUAL AND FUSED FR METHODS

When individual and fused FR methods are considered together, the following observations can be made: 1) The rank aggregation based fusion methods (RAS) [21] outperform the best individual FR methods, as can be seen from Table 13. This is also evident from Table 16 where it is clear that RAS6 performs statistically better than the top performing FR methods on a majority of databases. Although statistical significance testing results for other RAS methods in comparison with individual FR methods have not been provided due to space constraints, they are also found to be statistically superior. 2) The learning based fusion methods, MMF [118] and CNNM [116], are outperformed by the best individual FR methods on datasets that are not involved in training these fusion methods. This can be seen from Table 13 in terms of weighted average PLCC and SRCC, and also from Table 16 for MMF1 in terms of statistical significance testing (statistical analysis for MMF2, MMF3, MMF4, and CNNM yielded similar conclusions). 3). Of the four empirical fusion methods, CM3 [115], CM4 [115], and HFSIMc [117], are outperformed by the best individual FR methods as can be observed from Table 13 in terms of weighted average PLCC and SRCC. The only exception is the empirical fusion method CISI [114], which performs at par with or better than top performing individual FR methods.

It can therefore be concluded that learning based fusion (MMF and CNNM) and empirical fusion techniques (CM3, CM4, HFSIMc), do not generalize very well when tested across a wide variety of IQA datasets, thereby revealing that they suffer from model overfitting and training database dependency issues. Such drawbacks make them less robust to handle unseen data, where they are outperformed by the best individual FR methods. On the other hand, the rank aggregation based fusion methods (RAS), perform better than other fusion techniques, but more importantly, they outperform the best individual FR methods across the entire range of IQA datasets used. Since these methods are completely training-free, they do not suffer from model overfitting and database dependence issues, making them truly robust. While it can be seen from Tables 6 and 7 that the performance of even the top performing FR methods varies, sometimes widely, across different IQA datasets, Tables 11 and 12 show that such performance variation across different datasets is less pronounced for RAS based methods. It can be concluded that by aggregating the ranks generated from various top performing FR IQA methods, the deficiencies of some methods in the combination are compensated by the strengths of other constituents. These characteristics of rank aggregation based fusion methods make them ideal candidates to annotate large-scale IQA datasets in place of subject ratings. While opinions provided by humans will continue to be the ultimate benchmark when it comes to annotating IQA databases, as we discussed earlier, it is quite impossible to obtain human opinions in adequate numbers for very large-scale datasets. Here, rank aggregation based fusion methods can be used to annotate such large datasets in place of human opinion scores instead of choosing one or the other individual FR method.

V. PERFORMANCE ANALYSIS OF NR METHODS

To analyze the performance of NR IQA methods, we use the same evaluation criteria as described in Section IV-A, and compute the evaluation metrics for two types of data. First, like the performance analysis of FR and fused FR methods in Section IV, all images within a database are considered, that is, all distortion types are taken into account while calculating PLCC, SRCC, and performing statistical significance testing. This will be referred to as the *all distortions* category. Second, evaluation metrics are calculated for a subset of distortion types in each dataset, which we shall refer to as the *subset distortions* category. For single distortion databases (LIVE R2 [3], TID2013 [5], CSIQ [6], VCLFER [35], and CIDIQ [31]), we constitute a subset of images belonging to four common distortion types: 1) Noise, 2) Gaussian Blur, 3) JPEG compression, and 4) JPEG2000 compression. It should be noted that the noisy images in the CIDIQ database [31] are afflicted with Poisson noise, while they are afflicted with additive white Gaussian noise in the other four single distortion datasets. However, for the purposes of the subset performance analysis, we do not make a distinction between the two. For multiply distorted databases, we constitute subsets of images by separately calculating evaluation metrics for individual distortion combinations (where possible). This means that we separately consider the Blur-JPEG and Blur-Noise combinations in the LIVE MD database [11], and the Blur-JPEG and Noise-JPEG combinations in the MDIVL database [14]. Since the MDID2013 database [12] contains only one distortion combination, while the MDID database [13] has many possible distortion combinations due to the random choice of distortions at different stages, the images in these two datasets cannot be split into subsets, and hence the entire datasets will be considered for the subset case as well. The rationale for conducting performance analysis for a subset of distortion types, especially for single distortion databases, stems from the fact that most training-based opinion-aware NR models are trained for the above-mentioned common distortion types that are found in almost all single distortion datasets. Thus, these subsets of distortions provide a more fair ground for comparison of these methods. However, we also consider the case of all distortions in each database and do not retrain these NR models on individual databases but use the original versions from the authors, in order to more rigorously test NR methods, as the ultimate goal of NR or *blind* IQA methods is to be robust to *unseen* data. The gap in performance for these two cases should highlight future research directions as well.

A. PERFORMANCE OF NR METHODS

We tested the 14 NR methods discussed in Section III-C and given in Table 5, on each of the nine subject-rated IQA

TABLE 18. PLCC of 14 NR methods on nine subject-rated IQA databases. All distortions in each dataset were considered.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
BIQI	0.9224	0.4489	0.6796	0.6106	0.3542	0.2563	0.6372	0.0169	0.7389	0.6215
BRISQUE	0.9671	0.4747	0.7006	0.8209	0.2924	0.3257	0.4558	0.1403	0.6045	0.6516
CORNIA	0.9665	0.5715	0.7325	0.8366	0.4496	0.1991	0.7907	0.6935	0.8679	0.8277
dipIQ	0.9348	0.4774	0.7787	0.8942	0.5208	0.2498	0.6738	0.4355	0.7669	0.7627
GWHGLBP	0.8079	0.4973	0.7002	0.6427	0.3653	0.2978	0.7035	0.7430	0.9663	0.5737
HOSA	0.9991	0.5481	0.7240	0.8496	0.4969	0.3761	0.6521	0.2513	0.6768	0.7167
ILNIQE	0.7061	0.5090	0.8024	0.7289	0.2768	0.3003	0.7053	0.5146	0.8923	0.6303
LPSI	0.8280	0.4892	0.7216	0.6020	0.4037	0.3981	0.4336	0.0999	0.5464	0.5715
MEON	0.9389	0.4946	0.7804	0.9221	0.4306	0.3854	0.5168	0.2430	0.2339	0.5722
NIQE	0.9052	0.4001	0.7170	0.8040	0.3703	0.2708	0.6728	0.5571	0.8387	0.5688
NRSL	0.9815	0.5345	0.7413	0.8905	0.4672	0.3069	0.6502	0.3088	0.4829	0.6794
QAC	0.8625	0.4371	0.7067	0.7615	0.3573	0.2856	0.6043	0.4240	0.4145	0.5713
SISBLIM	0.8077	0.3961	0.6945	0.7574	0.4782	0.4532	0.6700	0.8123	0.8948	0.5724
WaDIQaM-NR	0.9341	0.4707	0.7372	0.7862	0.4133	0.3481	0.4215	0.1371	0.2897	0.5213

TABLE 19. SRCC of 14 NR methods on nine subject-rated IQA databases. All distortions in each dataset were considered.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
BIQI	0.9198	0.3935	0.6186	0.6170	0.3433	0.2353	0.6276	0.0077	0.5556	0.5711
BRISQUE	0.9654	0.3672	0.5563	0.8130	0.3640	0.2496	0.4035	0.2209	0.5018	0.6647
CORNIA	0.9681	0.4288	0.6534	0.8354	0.3727	0.2071	0.7918	0.7055	0.8340	0.8336
dipIQ	0.9378	0.4377	0.5266	0.8957	0.4135	0.2100	0.6612	0.4153	0.6678	0.7131
GWHGLBP	0.7410	0.3844	0.5773	0.6243	0.3337	0.2412	0.7032	0.7555	0.9698	0.5841
HOSA	0.9990	0.4705	0.5925	0.8574	0.4494	0.3248	0.6412	0.2993	0.6393	0.7399
ILNIQE	0.8975	0.4939	0.8144	0.7391	0.2997	0.3127	0.6900	0.5148	0.8778	0.6238
LPSI	0.8181	0.3949	0.5303	0.5865	0.2060	0.1411	0.0306	0.0168	0.2717	0.5736
MEON	0.9409	0.3750	0.7248	0.9215	0.4101	0.2497	0.4861	0.2980	0.1917	0.5466
NIQE	0.9073	0.3132	0.6271	0.8126	0.3458	0.2212	0.6523	0.5451	0.7738	0.5713
NRSL	0.9796	0.4277	0.6750	0.8930	0.4249	0.2894	0.6458	0.4088	0.4145	0.6047
QAC	0.8683	0.3722	0.4900	0.7686	0.3196	0.1944	0.3239	0.2272	0.3579	0.5524
SISBLIM	0.7741	0.3177	0.6603	0.7622	0.4435	0.4098	0.6554	0.8089	0.8770	0.5375
WaDIQaM-NR	0.9417	0.4393	0.6388	0.7524	0.3588	0.2235	0.4040	0.1316	0.2379	0.5614

TABLE 20. PLCC of 14 NR methods on nine subject-rated IQA databases. A subset of distortions in each dataset were considered.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD		MDIVL	
									BJPG	BN	BJPG	NJPG
BIQI	0.9534	0.7565	0.7968	0.6106	0.4797	0.4900	0.6372	0.0169	0.7743	0.6634	0.7398	0.6035
BRISQUE	0.9760	0.8399	0.9196	0.8209	0.5257	0.3906	0.4558	0.1403	0.8663	0.4596	0.8249	0.6511
CORNIA	0.9715	0.8824	0.9135	0.8366	0.5900	0.5477	0.7907	0.6935	0.8774	0.8723	0.9419	0.7900
dipIQ	0.9559	0.8879	0.9479	0.8942	0.7433	0.6472	0.6738	0.4355	0.8235	0.7897	0.8311	0.7882
GWHGLBP	0.8088	0.7675	0.7839	0.6427	0.5196	0.5345	0.7035	0.7430	0.9677	0.9684	0.7745	0.4943
HOSA	0.9992	0.8858	0.9360	0.8496	0.6504	0.6283	0.6521	0.2513	0.8968	0.6728	0.9005	0.7022
ILNIQE	0.7031	0.8491	0.8143	0.7289	0.3127	0.3892	0.7053	0.5146	0.9048	0.8968	0.8293	0.5759
LPSI	0.8440	0.8114	0.8657	0.6020	0.5509	0.6289	0.4336	0.0999	0.8820	0.1182	0.7959	0.5075
MEON	0.9907	0.8940	0.9334	0.9221	0.6495	0.6379	0.5168	0.2430	0.0995	0.3881	0.3875	0.7405
NIQE	0.9162	0.8091	0.8876	0.8040	0.4694	0.4338	0.6728	0.5571	0.9099	0.8481	0.7996	0.4507
NRSL	0.9887	0.9108	0.9058	0.8905	0.4216	0.4500	0.6502	0.3088	0.3283	0.6263	0.6418	0.7334
QAC	0.8777	0.8051	0.8736	0.7615	0.4512	0.5068	0.6043	0.4240	0.5378	0.6722	0.6765	0.6090
SISBLIM	0.8220	0.7309	0.7967	0.7574	0.5792	0.6741	0.6700	0.8123	0.9030	0.8913	0.8056	0.4871
WaDIQaM-NR	0.9302	0.8983	0.8577	0.7862	0.4600	0.5530	0.4215	0.1371	0.6842	0.4379	0.6415	0.5231

databases mentioned in Table 2. Testing was done separately for the two viewing distances in the CIDIQ database [31], where labels of CIDIQ50 and CIDIQ100 correspond to the viewing distances of 50 cm and 100 cm, respectively. For all databases, the test results for the *all distortions* case are given in Table 18 in terms of PLCC and in Table 19 in terms of SRCC. The test results for the *subset distortions* case are given in Tables 20 and 21 in terms of PLCC and SRCC respectively. While considering Tables 18, 19, 20, and 21, it should be noted that the OA NR methods BIQI [129], BRISQUE [130], NRSL [137], CORNIA [131], HOSA [133], WaDIQaM-NR [138], and MEON [136] are trained on the LIVE R2 database [3], [22], and GWHGLBP [132] is trained on the LIVE MD database [11]. Thus, comparing these OA NR methods with other approaches on these respective databases is unreliable and unfair.

The overall performance of the 14 NR methods was determined by using the same approach as in Section IV-E. The weighted average PLCC and SRCC were computed for three cases: 1) All databases, 2) Only single distortion databases, and 3) Only multiple distortion databases. Table 22 depicts the overall performance of the 14 NR methods for *all distortions* in terms of weighted average PLCC and SRCC, where parts 1, 2, and 3 of the table correspond to the cases of all databases, single distortion databases, and multiple distortion databases, respectively. Within each case, the methods have been sorted in the descending order with respect to the weighted average PLCC and SRCC values, where the best performing methods can be found towards the top of the table. Table 23 provides the results for *subset distortions*. In both Tables 22 and 23 we are including results for the FR methods IWSSIM [87] and PSNR for quick comparison. For a thorough comparison of the overall performance of NR methods with that of individual and fused FR methods, these tables should be compared with Table 13.

TABLE 21. SRCC of 14 NR methods on nine subject-rated IQA databases. A subset of distortions in each dataset were considered.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD		MDIVL	
									BJPG	BN	BJPG	NJPG
BIQI	0.9528	0.7763	0.7972	0.6170	0.4976	0.4849	0.6276	0.0077	0.6542	0.4902	0.6591	0.5302
BRISQUE	0.9757	0.8401	0.8992	0.8130	0.4727	0.4771	0.4035	0.2209	0.7923	0.2991	0.7385	0.6612
CORNIA	0.9732	0.8727	0.8987	0.8354	0.5740	0.5053	0.7918	0.7055	0.8278	0.8523	0.9254	0.8027
dipIQ	0.9574	0.8720	0.9290	0.8957	0.7460	0.6433	0.6612	0.4153	0.6979	0.7391	0.6512	0.7730
GWHGLBP	0.7447	0.6538	0.6728	0.6243	0.4768	0.4454	0.7032	0.7555	0.9640	0.9751	0.7584	0.4502
HOSA	0.9991	0.8681	0.9111	0.8574	0.6677	0.6236	0.6412	0.2993	0.8437	0.5357	0.8789	0.7150
ILNIQE	0.9153	0.8417	0.8802	0.7391	0.3669	0.4248	0.6900	0.5148	0.8915	0.8821	0.7915	0.5797
LPSI	0.8333	0.7046	0.7711	0.5865	0.3382	0.3949	0.0306	0.0168	0.8387	0.0012	0.7348	0.4692
MEON	0.9906	0.9012	0.9300	0.9215	0.6421	0.5830	0.4861	0.2980	0.0476	0.3257	0.3255	0.7397
NIQE	0.9168	0.7972	0.8710	0.8126	0.4703	0.4180	0.6523	0.5451	0.8713	0.7938	0.7625	0.4510
NRSL	0.9880	0.8965	0.8874	0.8930	0.5732	0.5564	0.6458	0.4088	0.2634	0.5991	0.4684	0.7125
QAC	0.8857	0.8055	0.8415	0.7686	0.4450	0.4566	0.3239	0.2272	0.3959	0.4707	0.5537	0.5282
SISBLIM	0.7835	0.7703	0.8059	0.7622	0.5565	0.6314	0.6554	0.8089	0.8746	0.8782	0.7584	0.3320
WaDIQaM-NR	0.9399	0.8646	0.8636	0.7524	0.4777	0.4691	0.4040	0.1316	0.5012	0.2502	0.6121	0.4830

TABLE 22. Weighted Average PLCC and SRCC values of NR methods for *all distortions* and for the three cases of: 1) All Databases, 2) Single Distortion Databases, and 3) Multiple Distortion Databases. Methods in each case are sorted in descending order with respect to PLCC/SRCC values. FR Methods IWSSIM and PSNR are included for comparison and are highlighted in bold.

Part 1: All Databases				Part 2: Single Distortion Databases				Part 3: Multiple Distortion Databases			
NR Method	PLCC	NR Method	SRCC	NR Method	PLCC	NR Method	SRCC	NR Method	PLCC	NR Method	SRCC
IWSSIM*	0.8787	**IWSSIM***	0.8559	**IWSSIM***	0.8700	**IWSSIM***	0.8452	**IWSSIM***	0.8970	**IWSSIM***	0.8785
PSNR*	0.6927	**PSNR***	0.6720	**PSNR***	0.7180	**PSNR***	0.7066	CORNIA	0.8006	CORNIA	0.7990
CORNIA	0.6713	CORNIA	0.6147	HOSA	0.6266	ILNIQE	0.5651	GWHGLBP	0.7143	GWHGLBP	0.7184
HOSA	0.6275	ILNIQE	0.6031	NRSL	0.6136	HOSA	0.5641	ILNIQE	0.6945	ILNIQE	0.6830
dipIQ	0.6181	HOSA	0.5851	CORNIA	0.6099	NRSL	0.5499	SISBLIM	0.6937	SISBLIM	0.6749
NRSL	0.6085	dipIQ	0.5620	MEON	0.6026	CORNIA	0.5272	dipIQ	0.6838	dipIQ	0.6491
GWHGLBP	0.5949	NRSL	0.5589	dipIQ	0.5869	MEON	0.5245	NIQE	0.6597	NIQE	0.6392
ILNIQE	0.5919	SISBLIM	0.5408	WaDIQaM-NR	0.5683	dipIQ	0.5207	**PSNR***	0.6396	HOSA	0.6292
SISBLIM	0.5821	GWHGLBP	0.5377	BRISQUE	0.5571	WaDIQaM-NR	0.5203	HOSA	0.6296	**PSNR***	0.5992
NIQE	0.5642	NIQE	0.5181	LPSI	0.5509	BRISQUE	0.4877	NRSL	0.5977	NRSL	0.5780
MEON	0.5570	BIQI	0.5007	ILNIQE	0.5432	BIQI	0.4824	BIQI	0.5837	BIQI	0.5394
BIQI	0.5397	MEON	0.4969	GWHGLBP	0.5382	SISBLIM	0.4770	QAC	0.5503	BRISQUE	0.4614
BRISQUE	0.5360	BRISQUE	0.4792	SISBLIM	0.5291	NIQE	0.4606	BRISQUE	0.4915	MEON	0.4387
QAC	0.5338	WaDIQaM-NR	0.4782	QAC	0.5259	QAC	0.4556	MEON	0.4610	WaDIQaM-NR	0.3896
LPSI	0.5179	QAC	0.4292	NIQE	0.5189	GWHGLBP	0.4518	LPSI	0.4483	QAC	0.3736
WaDIQaM-NR	0.5131	LPSI	0.3558	BIQI	0.5188	LPSI	0.4325	WaDIQaM-NR	0.3970	LPSI	0.1943

*FR Methods included for comparison.

TABLE 23. Weighted Average PLCC and SRCC values of NR methods for *subset distortions* and for the three cases of: 1) All Databases, 2) Single Distortion Databases, and 3) Multiple Distortion Databases. Methods in each case are sorted in descending order with respect to PLCC/SRCC values. FR Methods IWSSIM and PSNR are included for comparison and are highlighted in bold.

Part 1: All Databases				Part 2: Single Distortion Databases				Part 3: Multiple Distortion Databases			
NR Method	PLCC	NR Method	SRCC	NR Method	PLCC	NR Method	SRCC	NR Method	PLCC	NR Method	SRCC
IWSSIM*	0.9116	**IWSSIM***	0.9002	**IWSSIM***	0.9226	**IWSSIM***	0.9179	**IWSSIM***	0.9004	**IWSSIM***	0.8820
CORNIA	0.8088	CORNIA	0.8007	dipIQ	0.8584	dipIQ	0.8527	CORNIA	0.8096	CORNIA	0.8062
dipIQ	0.7805	dipIQ	0.7562	MEON	0.8535	MEON	0.8449	GWHGLBP	0.7269	GWHGLBP	0.7208
HOSA	0.7534	HOSA	0.7438	HOSA	0.8407	HOSA	0.8364	ILNIQE	0.7110	ILNIQE	0.6974
SISBLIM	0.7227	ILNIQE	0.7078	CORNIA	0.8080	NRSL	0.8171	SISBLIM	0.7093	SISBLIM	0.6733
PSNR*	0.7148	**PSNR***	0.7048	NRSL	0.7855	**PSNR***	0.8054	dipIQ	0.7005	dipIQ	0.6571
NIQE	0.7093	SISBLIM	0.7008	**PSNR***	0.7836	CORNIA	0.7954	NIQE	0.6763	NIQE	0.6537
GWHGLBP	0.7073	NRSL	0.6996	BRISQUE	0.7689	BRISQUE	0.7685	HOSA	0.6639	HOSA	0.6488
NRSL	0.6937	NIQE	0.6954	WaDIQaM-NR	0.7653	WaDIQaM-NR	0.7477	**PSNR***	0.6441	**PSNR***	0.6015
ILNIQE	0.6801	GWHGLBP	0.6672	NIQE	0.7415	NIQE	0.7360	NRSL	0.5996	NRSL	0.5790
QAC	0.6637	MEON	0.6441	SISBLIM	0.7359	SISBLIM	0.7276	QAC	0.5944	BIQI	0.5464
MEON	0.6609	BIQI	0.6272	QAC	0.7312	QAC	0.7200	BIQI	0.5918	BRISQUE	0.4756
BIQI	0.6466	BRISQUE	0.6239	LPSI	0.7284	ILNIQE	0.7179	BRISQUE	0.5193	MEON	0.4379
BRISQUE	0.6457	WaDIQaM-NR	0.5786	BIQI	0.6999	BIQI	0.7059	MEON	0.4632	WaDIQaM-NR	0.4051
WaDIQaM-NR	0.6096	QAC	0.5529	GWHGLBP	0.6882	LPSI	0.6252	LPSI	0.4586	QAC	0.3815
LPSI	0.5953	LPSI	0.4254	ILNIQE	0.6501	GWHGLBP	0.6150	WaDIQaM-NR	0.4498	LPSI	0.2203

*FR Methods included for comparison.

Statistical significance testing was conducted in the same manner as described in Sections IV-A.3 and IV-F. The outcome of the kurtosis based check for Gaussianity of prediction residuals is presented in Table 24 where a "1" means that the residuals can be assumed to be Gaussian while a "0" means that such an assumption cannot be made. Each entry in the table may be composed of more than one symbol, and depicts the outcome of the check for either the *all* or *subset* distortions cases, as explained in the table caption. It can be observed from Table 24 that the kurtosis based assumption of Gaussianity holds in around 85% of cases. The prediction residuals of all NR methods were compared by carrying out hypothesis testing through the one-sided (left-tailed) two-sample F-test at 95% confidence (as in Section IV-F). Tables 25 and 26 provide the outcome of the statistical significance testing for the *all distortions* and *subset distortions* cases, respectively. For details of how to interpret the tables, refer to Section IV-F, and to the captions of Tables 25 and 26.

As in Section IV-G, the computational complexity of all 14 NR IQA methods under test was evaluated in terms of their execution time to determine the quality of a 1024

TABLE 24. Kurtosis based check for Gaussianity of prediction residuals of NR Methods, for *all* and *subset* (SS) distortions. The order of symbols within each entry is as follows: LIVE R2 (All, SS), TID2013 (All, SS), CSIQ (All, SS), VCLFER (All), CIDIQ50 (All, SS), CIDIQ100 (All, SS), MDID (All), MDID2013 (All), LIVE_MD (All, Blur-JPEG, Blur-Noise), MDIVL (All, Blur-JPEG, Noise-JPEG). A "1" means that the kurtosis of the residuals is between 2 and 4, and they can be assumed to be Gaussian distributed. A "0" means that the kurtosis of residuals is not between 2 and 4, and they are assumed to be non-Gaussian. FR Methods IWSSIM and PSNR are highlighted in bold.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
IWSSIM*	01	01	11	1	10	11	1	1	111	111
PSNR*	11	00	11	1	11	11	1	1	111	111
CORNIA	01	11	11	1	11	11	1	1	111	101
HOSA	01	10	11	0	11	11	1	1	111	111
dipIQ	01	11	11	1	11	11	1	1	111	111
NRSL	00	11	10	1	11	11	1	1	111	110
GWHGLBP	11	11	11	1	11	11	1	1	110	111
ILNIQE	11	00	00	1	11	11	1	1	111	101
SISBLIM	00	11	11	1	11	11	1	1	111	101
NIQE	11	11	11	0	11	11	1	1	111	101
MEON	00	11	10	1	11	11	1	1	111	111
BIQI	00	10	11	1	11	11	1	1	111	101
BRISQUE	01	10	10	0	11	11	1	1	111	101
QAC	01	10	10	1	11	11	1	1	111	111
LPSI	11	11	11	1	11	11	1	1	111	111
WaDIQaM-NR	01	11	10	0	11	11	1	1	111	111

*FR Methods included for comparison.

× 1024 color image on a Lenovo laptop computer with a 2.4GHz Intel Core i7-4700MQ processor, 12GB of RAM, Samsung 850 EVO Solid State Drive, and Windows 10 Home operating system. The execution times of all NR methods are given in Table 27, where methods have been sorted in ascending order with respect to execution time. As before, we provide the execution time of NR methods relative to the FR method PSNR for convenience in comparison with Table 17. Apart from the 14 NR methods being evaluated in this work, we have included the execution times of seven other well-known NR IQA methods in Table 27, which include: BLIINDS2 [166], DIIVINE [167], FRIQUEE [168], [169], Jet-LBP [170], MS-LQAF [171], NFERM [172], and TCLT [173]. We have not evaluated the performance of these methods because they take an excessive amount of time to estimate the quality of an image, and are infeasible for large-scale or real-time use. It should also be noted that while WaDIQaM-NR [138] takes a lot of time to determine the quality of the test image on the CPU (10.1277 seconds), it runs considerably faster when executed on the GPU. For reference, on another machine, WaDIQaM-NR ran around 40 times faster on the GPU as compared to the CPU.

B. ANALYSIS AND DISCUSSION

It can be observed from Tables 22 and 23 that in terms of weighted average PLCC and SRCC, the NR method CORNIA [131] outperforms other NR methods, sometimes by a clear margin, for the cases of *all databases* and *multiple distortion databases* in both the *all distortions* and *subset distortions* categories. In case of *single distortion databases*, HOSA [133] does well for the *all distortions* category, while dipIQ [16] and MEON [136] do well in the *subset distortions* category. Since the OA NR methods are trained on databases that are constituents in the weighted average PLCC and SRCC computation, as described in Section III-C, these results should be considered in conjunction with the statistical significance testing outcome. From Tables 25 and 26, it can be respectively observed that for both the categories of *all distortions* and *subset distortions*, the NR methods CORNIA [131], HOSA [133], and dipIQ [16], perform better than most other methods on most databases. CORNIA and HOSA are OA NR methods that first learn image features and then a quality model, while dipIQ is an OU NR method that utilizes millions of DIPs and a learning-to-rank algorithm to learn the quality model. However, HOSA itself can be regarded as a modified version of CORNIA, while dipIQ uses CORNIA features at its base. This shows that CORNIA features [131] are quite effective when it comes to blind IQA.

The following observations can be made about various NR design philosophies: 1) The OA NR methods that use handcrafted features (BIQI [129], BRISQUE [130], GWHGLBP [132], and NRSL [137]), do not show robust cross-dataset performance. While they may perform better on one class of data, such as single distortion or multiple distortion datasets, their performance degrades considerably on another class of data. This shows that such models suffer from model overfitting and database dependency issues, and also that truly general-purpose handcrafted features for perceptual IQA remain lacking. 2) OA NR techniques that utilize unsupervised feature learning, such as CORNIA [131] and HOSA [133], demonstrate relatively robust performance. For example, even though these methods are trained on singly distorted content, they perform relatively well on multiply distorted databases, which is somewhat surprising. 3) Among OA NR methods that employ deep learning, MEON [136] performs better than WaDIQaM-NR [138]. This may be because MEON uses two sub-tasks to perform IQA, where a large amount of data is used to pre-train the distortion identification aspect of the network. However, unlike CORNIA and HOSA, these methods do not perform adequately on multiply distorted content, even though they are trained on individual distortion types that make up the multiple distortion combinations. This further highlights the difficulties encountered while doing IQA for multiply distorted content and while training deep learning models on small-scale datasets. 4) dipIQ [16] performs better than most NR methods. In addition to using CORNIA features, dipIQ utilizes a novel training process which does not use human annotated data for training. Instead, it alleviates the issue of small-scale subject-rated datasets by using millions of DIPs, generated by using FR IQA methods, to train the model. This approach highlights the advantages of utilizing techniques that use large-scale datasets which employ alternative annotation techniques. 5) The performance of OU NR methods (NIQE [128], ILNIQE [134], QAC [15], SISBLIM [12], and LPSI [135]) shows considerable room for improvement, which also highlights the difficult nature of the OU NR IQA problem. 6) While many training-based NR methods are usually trained and tested on each database separately, which

TABLE 25. Statistical significance testing results of NR methods based on prediction residuals for the *all distortions* case. Each entry is a codeword composed of ten symbols, where each symbol represents the test outcome for one IQA database. The symbol location within a codeword represents IQA databases in the following order: [LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ50, CIDIQ100, MDID, MDID2013, LIVE MD, MDIVL]. A "1" means that the method in the row, for a particular database, is statistically better than the method in the column, a "0" means that it is statistically worse, while a "_" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence). FR Methods IWSSIM and PSNR are included for comparison and are highlighted in bold.

	IWSSIM	**PSNR**	CORNIA	HOSA	dipIQ	NRSL	GWHGLBP	ILNIQE	SISBLIM	NIQE	MEON	BIQI	BRISQUE	QAC	LPSI	WaDIQaM-NR
IWSSIM	_____	1111111111	0111111111	0111111111	1111111111	0111111111	1111111101	1111111111	1111111111	1111111111	111_111111	1111111111	0111111111	1111111111	1111111111	1111111111
PSNR	0000000000	_____	01__110000	01__11011_	01_0110__0	01_011011_	1111110001	1101110_0_	1111110001	01__110_01	01_0111111	011111_1_1	011_11111_	_11111__11	11_1111111	01_1111111
CORNIA	1000000000	10__001111	_____	0_____1111	1100__1111	0__0__1111	11_1__1_01	11011_1101	11_1_01001	11_1__1111	1100__1111	1111__1111	_1__1_1111	11_1__1111	11_1_01111	11_1__1111
HOSA	1000000000	10__00100_	1_____0000	_____	1100____00	1__0____1_	11_11_0001	1_011_0001	11_1___001	11_11__001	1100__1_11	11111___01	11_11_1_11	11_11_1_11	11_1__1_11	11_1__1_11
dipIQ	0000000000	10_1001__1	0011__0000	0011____11	_____	001_____11	1_111__001	1__11_0_01	1111_0_001	11111___01	___0__1_11	1_111_11_1	0_111_1111	1_111_1_11	1_111_1111	__111_1111
NRSL	1000000000	10_100100_	1__1__0000	0__1____0_	110_____00	_____	1_11__0001	1_011_000_	1111_0_001	11_1___001	1_00__1_11	1111____01	11111_1_0_	11_1__1__1	11_1__1__1	11_1__1_11
GWHGLBP	0000000010	0000001110	00_0__0_10	00_00_1110	0_000__110	0_00__1110	_____	1_00___11_	_1_0_0101_	01_0___11_	0_00___111_	0_____111_	0__0__1110	01_0___111_	_____111_	0__0__111_
ILNIQE	0000000000	0010001_1_	00100_0010	0_100_1110	0__00_1_10	0_100_111_	0_11___00_	_____	011_0010__	0110__1_1_	0__0___111_	0111___111_	0_10___111_	011___1_1_	0_11___111_	0_10__1111
SISBLIM	0000000000	0000001110	00_0_10110	00_0____110	0000_1_110	0000_1_110	_0_1_1010_	100_1101__	_____	0__0_1_11_	0000___111_	0__111_11_	00_01_1110	0____1111_	_0_1___111_	000___111_
NIQE	0000000000	10__001_10	00_0__0000	00_00__110	00000___10	00_0___110	10_1___00_	1001__0_0_	1__1_0_00_	_____	0000___111_	0__1____11_	00_____1110	1__1__1_1_	10_1___111_	00____111_
MEON	000_000000	10_1000000	0011___0000	0011__0_00	___1__0_00	0_11__0_00	1_11___000_	1__1___000_	1111__000_	1111___000_	_____	1_11__0_0_	0_11_____00	1111__0___	1_11___1_0_	__11__1___
BIQI	0000000000	100000_0_0	0000__0000	00000____10	0_000_00_0	0000____10	1_____000_	1000___000_	1__000_00_	1__0____00_	0_00__1_1_	_____	0__0__1_1_	1__0___01_	1_0___1_1_	0_00__1_11
BRISQUE	1000000000	100_00000_	_0__0_0000	00_00_0_00	1_000_0000	00000_0_1_	1__1___0001	1_01___000_	11_10_0001	11_____0001	1_00____11	1__1__0_0_	_____	1__1__0_11	1__1_____1	1__1_____11
QAC	0000000000	_00000__00	00_0__0000	00_00_0_00	0_000_0_00	00_0__0__0	10_1___000_	100____0_0_	1_____0000_	0__0__0_0_	0000__1___	0__1___10_	0__0__1_00	_____	10_1__110_	0_____1___
LPSI	0000000000	00_0000000	00_0_10000	00_0__0_00	0_000_0000	00_0__0__0	_____000_	1_00___000_	_1_0___000_	01_0___000_	0_00__0_1_	0_1____0_0_	0__0_____0	01_0__001_	_____	0__0____1_
WaDIQaM-NR	0000000000	10_0000000	00_0__0000	00_0__0_00	__000_0000	00_0__0_00	1__1___000_	1_01__0000	111____000_	11_____000_	__00__0___	1_11__0_00	0__0____00	1_____0___	1__1____0_	_____

TABLE 26. Statistical significance testing results of NR methods based on prediction residuals for the *subset distortions* case. Each entry is a codeword composed of ten symbols, where each symbol represents the test outcome for one IQA database. The symbol location within a codeword represents IQA databases in the following order: [LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ50, CIDIQ100, LIVE MD BJPG, LIVE MD BN, MDIVL BJPG, MDIVL NJPG]. A "1" means that the method in the row, for a particular database, is statistically better than the method in the column, a "0" means that it is statistically worse, while a "_" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence). FR Methods IWSSIM and PSNR are included for comparison and are highlighted in bold.

	IWSSIM	**PSNR**	CORNIA	HOSA	dipIQ	NRSL	GWHGLBP	ILNIQE	SISBLIM	NIQE	MEON	BIQI	BRISQUE	QAC	LPSI	WaDIQaM-NR
IWSSIM	_____	1111111111	0111111101	011111_111	_111111111	0111111111	1111110011	111111__11	111111__11	111111_111	011_111111	_111111111	0111111111	1111111111	1111111111	1111111111
PSNR	0000000000	_____	0_0___0000	0_0___010_	0_00000_00	00001_1110	11111_0001	1111110001	1111_00001	010_110001	0_00__1110	01111__1__	010_11010_	_1011_11__	11_11_0101	0__11__111
CORNIA	1000000010	1_1___1111	_____	0_0____111	1_0000111_	00_01_1111	1111__0011	1111110_11	1111_00_11	11111_0_11	0_00_01111	11111_1111	01___1_111	11111_1111	1111___111	1_111_1111
HOSA	100000_000	1_1___101_	1_1____000	_____	1_000_1010	1010111_1_	1111110011	111111_011	1111___011	111111_011	1__0___111_	1111111_11	111111111_	11111_1111	11111__111	1_111_1111
dipIQ	_000000000	1_11111_11	0_1111000_	0_111_0101	_____	001_111111	1111110011	11111100_1	11111_00_1	11111100_1	0_101_1111	_11111_111	01111101_1	1111111111	11111_01_1	1_11111111
NRSL	1000000000	11110_0001	11_10_0000	0101000_0_	110_000000	_____	1111__0001	1111__0001	1111000001	11111_0001	010000_11_	1111__0_01	1101__0101	1111__0__1	1111000101	1_11__01_1
GWHGLBP	0000001100	00000_1110	0000__1100	0000001100	0000001100	0000__1110	_____	100011110_	___0_011__	00001_11__	00000001110	0_____11_0	0000_11100	0000__1110	000__011__	0000__111_
ILNIQE	000000__00	0000001110	0000001_00	000000_100	00000011_0	0000__1110	011100001_	_____	01__00____	01001__1_1	0000001110	01_10_111_	0_000_11__	010___111_	010100_1__	0000_0111_
SISBLIM	000000__00	0000_11110	0000_11_00	0000____100	00000_11_0	0000111110	___1_100__	10__11____	_____	000011_1__	0000__1110	0__1_11110	0000_111_0	000_111110	_001___1__	000_11111_
NIQE	000000_000	101_001110	00000_1_00	000000_100	00000011_0	00000_1110	11110_00__	10110__0_0	111100_0__	_____	0000001110	01110_1110	000_0_11_0	1__10_1110	1_110011__	001_00111_
MEON	100_000000	1_11__0001	1_11_10000	0__1__000_	1_010_0000	101111_00_	1111110001	1111110001	1111__0001	1111110001	_____	1111110001	1111110_01	1111110001	11111_0_01	1_11110_01
BIQI	_000000000	10000__0__	00000_0000	0000000_00	_00000_000	0000__1_10	1_____00_1	10_01_000_	1__0_00001	10001_0001	0000001110	_____	0000__010_	1000__1_1_	100__0010_	1000___111_
BRISQUE	1000000000	101_00101_	10___0_000	000000000_	10000010_0	0010__1010	1111_00011	1_111_00__	1111_000_1	111_1_00_1	0000001_10	1111__101_	_____	1111__101_	1111_0_1_1	1011_01_11
QAC	0000000000	_0100_00__	00000_0000	0000000_00	0000000000	0000__1__0	1111__0001	101___000_	111_000001	0__01_0001	0000001110	0111__0_0_	0000__010_	_____	1__1_00101	00____01__
LPSI	0000000000	00_00_1010	0000___000	00000___000	00000_10_0	0000111010	111__100__	101011_0__	_110___0__	0_001100__	00000_1_10	011__1101_	0000_1_0_0	0__0_11010	_____	00_0__1_1_
WaDIQaM-NR	0000000000	1__00__000	0_000_0000	0_000_0000	0_00000000	0_00__10_0	1111__000_	1111_1000_	111_00000_	110_11000_	0_00001_10	0111__000_	0100_10_00	11____10__	11_1__0_0_	_____

TABLE 27. Execution Time of NR methods on a test image. Methods are sorted in ascending order with respect to the execution time. FR method PSNR is included for comparison and is highlighted in bold.

NR Method	Execution Time (Seconds)	Execution Time (Relative to PSNR)
PSNR*	0.0044 s	1.00
LPSI	0.0827 s	18.80
MEON	0.2348 s	53.36
QAC	0.2811 s	63.89
HOSA	0.3312 s	75.27
NRSL[a]	0.3895 s	88.52
GWHGLBP[a]	0.3945 s	89.66
NIQE	0.4558 s	103.59
BRISQUE	0.4641 s	105.48
BIQI	1.2045 s	273.75
dipIQ	2.8367 s	644.70
Jet-LBP[a,b,c]	3.1004 s	704.64
CORNIA	3.6154 s	821.68
ILNIQE	4.0060 s	910.45
SISBLIM	5.3890 s	1224.77
TCLT[c]	7.8548 s	1785.18
WaDIQaM-NR	10.1277 s	2301.75
MS-LQAF[a,c]	36.9052 s	8387.55
DIIVINE[c]	38.2215 s	8686.70
BLIINDS2[c]	94.6167 s	21503.80
FRIQUEE[c]	109.1559 s	24808.16
NFERM[c]	128.8809 s	29291.11

*FR Method included for comparison.
[a]Feature extraction time only.
[b]The performance of Jet-LBP was not evaluated as SVR model parameters are not available.
[c]The performance of these methods was not evaluated due to their large computation times.

often leads to high PLCC and SRCC numbers, we believe that cross-dataset testing is crucial to the performance analysis of NR methods. 7) While NR methods such as CORNIA and dipIQ may be relatively better in quality prediction performance compared to other methods, they have a large execution time, as can be seen from Table 27. This implies that such methods are infeasible for real time usage. 8) We have included the FR IQA methods IWSSIM [87] and PSNR for comparison in the tables of this section, and it can be seen that the performance of all NR IQA methods is still a considerable distance away from top performing FR methods such as the IWSSIM, a disparity which is even more pronounced in the *all distortions case*. Even the perceptually inaccurate PSNR outperforms many NR methods, especially for the *all distortions* case. The above-mentioned observations highlight the significant room for improvement that exists in the area of NR IQA, both in terms of quality prediction performance and execution time.

VI. CONCLUSION AND FUTURE WORK

In this work, we carried out an extensive review and performance evaluation study of the field of IQA. In all, we evaluated the performance of 43 FR, seven fused FR and 14 NR IQA methods. If the 22 different versions of the seven fused FR methods are considered separately, then this means that we evaluated 79 IQA methods. In order to ensure the diversity of test data, we used nine subject-rated IQA datasets, five of which are composed of singly distorted content, while four contain multiply distorted content. To the best of our knowledge, this is so far the largest study of its kind, and hopefully will plug the gap that previously existed with regard to the lack of such surveys.

In summary, this work has the following findings: 1) Among the individual FR methods, structural similarity based methods IWSSIM [87], FSIMc [83], DSS [79], and VSI [102], are the top performers. 2) The empirical (HFSIMc [117], CM3 [115], CM4 [115]) and learning based (MMF [118], CNNM [116]) fusion approaches are not only outperformed by rank aggregation based fusion approach RAS [21], but also by top performing individual FR methods, thereby implying that existing empirical and learning based fusion methods do not offer clear advantages over individual FR methods. 3) However, the rank aggregation based fusion approach RAS [21] not only comprehensively outperforms other fusion approaches but also top performing individual FR methods. Its training-free nature and robust cross-dataset performance make it highly promising as a means to annotate very large-scale IQA datasets in the future. 4) Among NR methods, we have found CORNIA [131], HOSA [133], and dipIQ [16], to perform better than other methods. 5) While the perceptual quality prediction performance of FR methods has matured quite well, the performance of NR methods, both in terms of perceptual quality prediction accuracy and computational complexity is still a long distance away from top performing FR methods.

This work not only highlights the current state-of-the-art in the field of IQA of 2D natural images, but also the challenges that IQA researchers need to address, especially in the area of blind IQA. As discussed in Section V-B, the top performing NR models CORNIA [131], HOSA [133], and dipIQ [16] utilize CORNIA features which are learned automatically in an unsupervised manner, thereby highlighting the strength of learned against handcrafted features. DNN based models have enjoyed a lot of success in other areas of computer vision and image processing [174], which is largely due to the availability of very large-scale annotated datasets such as ImageNet [175]. On the other hand, DNN based blind IQA models, such as the ones evaluated in this work (WaDIQaM-NR [138] and MEON [136]) show a lot of room for improvement. These and other DNN based IQA models identified in Section III-C.1 train on the available small-scale IQA datasets (with hundreds or a few thousands of images) and may try to increase training data size by data augmentation, but achieved only limited success. The design of very large-scale annotated IQA datasets is an open problem [176]. The real challenge is that it is impossible to perform subjective tests to annotate such very large-scale datasets, thus the use of alternative data annotation techniques is highly desirable. One important discovery of this work is that rank aggregation based training-free FR fusion methods offer good promise of robust perceptual quality prediction performance when tested across a wide range of available subject-rated datasets. In the future, very large-scale simulated distortion datasets, with millions of images, may be developed where distortions are added in a content adaptive manner. Such datasets can then be synthetically annotated by using rank aggregation based FR fusion methods. DNN models should then be trained by

utilizing such new datasets. This research direction deserves deeper investigation in the future.

ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Z. Wang and A. C. Bovik, "Modern image quality assessment," in *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1. San Rafael, CA, USA: Morgan & Claypool, Jan. 2006, pp. 1–156.

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

[3] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[4] N. N. Ponomarenko, V. V. Lukin, A. A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[5] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. K. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[6] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006:1–011006:21, Jan. 2010.

[7] *A57 Image Quality Database*. Accessed: Aug. 20, 2019. [Online]. Available: http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=26

[8] U. Engelke, H.-J. Zepernick, and T. M. Kusuma, "Subjective quality assessment for wireless image communication: The wireless imaging quality database," in *Proc. 5th Int. Workshop Video Process., Qual. Metrics Consum. Electron. (VPQM)*, Scottsdale, AZ, USA, Jan. 2010, pp. 1–5.

[9] Y. Horita, K. Shibata, and K. Yoshikazu, *MICT Image Quality Evaluation Database*.

[10] P. L. Callet and F. Autrusseau. (2005). *Subjective Quality Assessment IRCCyN/IVC Database*. Accessed: Aug. 20, 2019. [Online]. Available: http://www2.irccyn.ec-nantes.fr/ivcdb/

[11] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2012, pp. 1693–1697.

[12] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Trans. Broadcast.*, vol. 60, no. 3, pp. 555–567, Sep. 2014.

[13] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, Jan. 2017.

[14] S. Corchs and F. Gasparini, "A multidistortion database for image quality," in *Proc. Int. Workshop Comput. Color Imag. (CCIW)*, Milan, Italy, Mar. 2017, pp. 95–104.

[15] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, Jun. 2013, pp. 995–1002.

[16] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[17] Y. Zhang and D. M. Chandler, "Opinion-unaware blind quality assessment of multiply and singly distorted images via distortion parameter estimation," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5433–5448, Nov. 2018.

[18] J. Kim and S. Lee, "Deep blind image quality assessment by employing FR-IQA," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3180–3184.

[19] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[20] S. Athar, A. Rehman, and Z. Wang, "Quality assessment of images undergoing multiple distortion stages," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3175–3179.

[21] P. Ye, J. Kumar, and D. Doermann, "Beyond human opinion scores: Blind image quality assessment based on synthetic scores," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 4241–4248.

[22] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. *LIVE Image Quality Assessment Database Release 2*. Accessed: Aug. 20, 2019. [Online]. Available: http://live.ece.utexas.edu/research/Quality/subjective.htm

[23] M. Pedersen and J. Y. Hardeberg, "Survey of full-reference image quality metrics," Høgskolen i Gjøviks Rapportserie 5, Norwegian Color Res. Lab., Gjøvik Univ. College Norway, Gjøvik, Norway, Tech. Rep., Jun. 2009.

[24] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," *J. Visual Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011.

[25] M. Pedersen and J. Y. Hardeberg, "Full-reference image quality metrics: Classification and evaluation," *Found. Trends Comput. Graph. Vis.*, vol. 7, no. 1, pp. 1–80, Mar. 2012.

[26] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Orlando, FL, USA, Oct. 2012, pp. 1477–1480.

[27] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," Jun. 2014, *arXiv:1406.7799*. [Online]. Available: https://arxiv.org/abs/1406.7799

[28] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb. 2013.

[29] L. He, F. Gao, W. Hou, and L. Hao, "Objective image quality assessment: A survey," *Int. J. Comput. Math.*, vol. 91, no. 11, pp. 2374–2388, 2014.

[30] M. Pedersen, "Evaluation of 60 full-reference image quality metrics on the CID:IQ," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 1588–1592.

[31] X. Liu, M. Pedersen, and J. Y. Hardeberg, "CID:IQ—A new image quality database," in *Proc. Int. Conf. Image, Signal Process. (ICISP)*, Cherbourg, France, Jul. 2014, pp. 193–202.

[32] R. A. Manap and L. Shao, "Non-distortion-specific no-reference image quality assessment: A survey," *Inf. Sci.*, vol. 301, pp. 141–160, Apr. 2015.

[33] S. Xu, S. Jiang, and W. Min, "No-reference/blind image quality assessment: A survey," *IETE Tech. Rev.*, vol. 34, no. 3, pp. 223–245, 2017.

[34] Y. Niu, Y. Zhong, W. Guo, Y. Shi, and P. Chen, "2D and 3D image quality assessment: A survey of metrics and challenges," *IEEE Access*, vol. 7, pp. 782–801, 2019.

[35] A. Zarić, N. Tatalović, N. Brajković, H. Hlevnjak, M. Lončarić, E. Dumić, and S. Grgić, "VCLFER image quality assessment database," *Proc. AUTOMATIKA*, vol. 53, no. 4, pp. 344–354, 2012.

[36] H. Yeganeh, M. Rostami, and Z. Wang, "Objective quality assessment of interpolated natural images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4651–4663, Nov. 2015.

[37] R. M. Nasiri and Z. Wang, "Perceptual aliasing factors and the impact of frame rate on video quality," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3475–3479.

[38] R. M. Nasiri, Z. Duanmu, and Z. Wang, "Temporal motion smoothness and the impact of frame rate variation on video quality," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1418–1422.

[39] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[40] R. Hassen, Z. Wang, and M. M. A. Salama, "Objective quality assessment for multiexposure multifocus image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2712–2724, Sep. 2015.

[41] K. Ma, T. Zhao, K. Zeng, and Z. Wang, "Objective quality assessment for color-to-gray image conversion," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4673–4685, Dec. 2015.

[42] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," in *Proc. SPIE Electron. Imag.*, San Francisco, CA, USA, vol. 9394, Mar. 2015, pp. 939406-1–939406-11.

[43] *Tampere Image Database 2013 (TID2013) Version 1.0*. Accessed: Aug. 20, 2019. [Online]. Available: http://www.ponomarenko.info/tid2013.htm

[44] E. C. Larson and D. M. Chandler. *Computational and Subjective Image Quality (CSIQ) Database*. Accessed: Aug. 20, 2019. [Online]. Available: http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=23

[45] *VCLFER Image Quality Assessment Database*. Accessed: Aug. 20, 2019. [Online]. Available: http://www.vcl.fer.hr/quality/vclfer.html

[46] *Colourlab Image Database: Image Quality (CID:IQ)*. Accessed: Aug. 20, 2019. [Online]. Available: https://www.ntnu.edu/web/colourlab/software

[47] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik. *LIVE Multiply Distorted Image Quality Database*. Accessed: Aug. 20, 2019. [Online]. Available: http://live.ece.utexas.edu/research/Quality/live_multidistortedimage.html

[48] *Multiply Distorted Image Database (MDID)*. Accessed: Aug. 20, 2019. [Online]. Available: http://www.sz.tsinghua.edu.cn/labs/vipl/mdid.html

[49] S. Corchs, F. Gasparini, and R. Schettini, "Noisy images-JPEG compressed: Subjective and objective image quality evaluation," in *Proc. SPIE Electron. Imag.*, San Francisco, CA, USA, vol. 9016, Feb. 2014, pp. 90160V-1–90160V-9.
[50] *Multiply Distorted Database MD-IVL*. Accessed: Aug. 20, 2019. [Online]. Available: http://www.mmsp.unimib.it/image-quality/
[51] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT.500-13, Jan. 2012.
[52] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T P.910, Apr. 2008.
[53] *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II*, document, Video Quality Experts Group, 2003. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-ii/frtv-phase-ii.aspx
[54] S. Tourancheau, F. Autrusseau, Z. M. P. Sazzad, and Y. Horita, "Impact of subjective dataset on the performance of image quality metrics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, San Diego, CA, USA, Oct. 2008, pp. 365–368.
[55] *Guidelines for the Evaluation of Gamut Mapping Algorithms*, Int. Commission Illumination, Peter Blattner, Switzerland, 2004.
[56] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.
[57] *Tampere Image Database 2008 (TID2008) Version 1.0*. Accessed: Aug. 20, 2019. [Online]. Available: http://www.ponomarenko.info/tid2008.htm
[58] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.
[59] A. Ciancio, A. L. N. T. da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, Jan. 2011.
[60] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 390–402, Jan. 2015.
[61] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.
[62] *Amazon Mechanical Turk*. Accessed: Aug. 20, 2019. [Online]. Available: https://www.mturk.com/
[63] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10K: Towards an ecologically valid and large-scale IQA database," 2018, *arXiv:1803.08489*. [Online]. Available: https://arxiv.org/abs/1803.08489
[64] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.
[65] *Figure Eight*. Accessed: Aug. 20, 2019. [Online]. Available: https://www.figure-eight.com/
[66] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4408–4421, Nov. 2015.
[67] J. Kumar, P. Ye, and D. Doermann, "A dataset for quality assessment of camera captured document images," in *Proc. Int. Workshop Camera-Based Document Anal. Recognit. (CBDAR)*, Washington, DC, USA, Aug. 2013, pp. 113–125.
[68] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, Jun. 2013, pp. 987–994.
[69] S. Winkler. *Image and Video Quality Resources*. Accessed: Aug. 20, 2019. [Online]. Available: https://stefan.winkler.site/resources.html
[70] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.
[71] H. Yu and S. Winkler, "Image complexity and spatial information," in *Proc. IEEE Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Klagenfurt, Austria, Jul. 2013, pp. 12–17.
[72] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Proc. SPIE Electron. Imag.*, Santa Clara, CA, USA, vol. 5007, Jun. 2003, pp. 87–95.
[73] M. Buczkowski and R. Stasiński, "Effective coverage as a new metric for image quality assessment databases comparison," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Poznan, Poland, May 2017, pp. 1–5.
[74] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.
[75] Z. Wang, "Objective image quality assessment: Facing the real-world challenges," in *Proc. IS&T Int. Symp. Electron. Imag.*, vol. 2016, no. 13, San Francisco, CA, USA, Feb. 2016, pp. 1–6.
[76] S. Rezazadeh and S. Coulombe, "A novel discrete wavelet transform framework for full reference image quality assessment," *Signal, Image, Video Process.*, vol. 7, no. 3, pp. 559–573, May 2013.
[77] S. Li, F. Zhang, M. Lin, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.
[78] I. Lissner, J. Preiss, P. Urban, M. S. Lichtenauer, and P. Zolliker, "Image-difference prediction: From grayscale to color," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 435–446, Feb. 2013.
[79] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, "Image quality assessment based on DCT subband similarity," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 2105–2109.
[80] E. D. Di Claudio and G. Jacovitti, "A detail-based method for linear full reference image quality prediction," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 179–193, Jan. 2018.
[81] S. Rezazadeh and S. Coulombe, "Low-complexity computation of visual information fidelity in the discrete wavelet domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 2438–2441.
[82] X. Zhang, X. Feng, W. Wang, and W. Xue, "Edge strength similarity for image quality assessment," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 319–322, Apr. 2013.
[83] 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.
[84] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.
[85] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.
[86] 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.
[87] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.
[88] T. Wang, L. Zhang, H. Jia, B. Li, and H. Shu, "Multiscale contrast similarity deviation: An effective and efficient index for perceptual image quality assessment," *Signal Process., Image Commun.*, vol. 45, pp. 1–9, Jul. 2016.
[89] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2003, pp. 1398–1402.
[90] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.
[91] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, "Modified image visual quality metrics for contrast change and mean shift accounting," in *Proc. 11th Int. Conf. Exper. Designing Appl. CAD Syst. Microelectron. (CADSM)*, Svalyava, Ukraine, Feb. 2011, pp. 305–311.
[92] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *Proc. 2nd Int. Workshop Video Process., Qual. Metrics Consum. Electron. (VPQM)*, Scottsdale, AZ, USA, vol. 4, Jan. 2006, pp. 1–4.
[93] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Proc. 3rd Int. Workshop Video Process., Qual. Metrics Consum. Electron. (VPQM)*, Scottsdale, AZ, USA, vol. 4, Jan. 2007.
[94] L. Li, H. Cai, Y. Zhang, W. Lin, A. C. Kot, and X. Sun, "Sparse representation-based image quality index with adaptive sub-dictionaries," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3775–3786, Aug. 2016.
[95] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Hong Kong, Sep. 2010, pp. 321–324.
[96] H.-W. Chang, H. Yang, Y. Gan, and M.-H. Wang, "Sparse feature fidelity for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4007–4018, Oct. 2013.

[97] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Orlando, FL, USA, Sep. 2012, pp. 1473–1476.
[98] 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.
[99] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[100] 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.
[101] H. R. Sheikh and A. C. Bovik. (2005). *Pixel Domain Version of VIF*. [Online]. Available: http://live.ece.utexas.edu/research/Quality/VIF.htm
[102] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.
[103] S. Rezazadeh and S. Coulombe, "A novel approach for computing and pooling Structural SIMilarity index in the discrete wavelet domain," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Cairo, Egypt, Nov. 2009, pp. 2209–2212.
[104] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, Nov. 2011.
[105] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.
[106] C. C. Yang and S. H. Kwok, "Efficient gamut clipping for color image processing using LHS and YIQ," *Opt. Eng.*, vol. 42, no. 3, pp. 701–711, Mar. 2003.
[107] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
[108] L. Zhang, Z. Gu, and H. Li, "SDSP: A novel saliency detection method by combining simple priors," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Melbourne, VIC, Australia, Sep. 2013, pp. 171–175.
[109] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proc. SPIE Electron. Imag.*, San Jose, CA, USA, vol. 5666, Mar. 2005, pp. 149–159.
[110] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.
[111] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, Mar. 2001.
[112] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Appl. Comput. Harmon. Anal.*, vol. 11, no. 1, pp. 89–123, Jul. 2001.
[113] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Washington, DC, USA, vol. 3, Oct. 1995, pp. 444–447.
[114] K. Okarma, "Combined image similarity index," *Opt. Rev.*, vol. 19, no. 5, pp. 349–354, Sep. 2012.
[115] K. Okarma, "Quality assessment of images with multiple distortions using combined metrics," *Elektronika Ir Elektrotechnika*, vol. 20, no. 6, pp. 128–131, 2014.
[116] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," in *Proc. SPIE Electron. Imag.*, San Francisco, CA, USA, vol. 9394, Mar. 2015, pp. 93940K-1–93940K-12.
[117] K. Okarma, "Hybrid feature similarity approach to full-reference image quality assessment," in *Proc. Int. Conf. Comput. Vis. Graph. (ICCVG)*, Warsaw, Poland, Sep. 2012, pp. 212–219.
[118] T. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1793–1807, May 2013.
[119] K. Okarma, "Combined full-reference image quality metric linearly correlated with subjective assessment," in *Proc. Int. Conf. AI Soft Comput. (ICAISC)*, Zakopane, Poland, Jun. 2010, pp. 539–546.
[120] K. Okarma, "Extended hybrid image similarity—Combined full-reference image quality metric linearly correlated with subjective scores," *Elektronika ir Elektrotechnika*, vol. 19, no. 10, pp. 129–132, 2013.
[121] T. Liu, W. Lin, and C.-C. J. Kuo, "A multi-metric fusion approach to visual quality assessment," in *Proc. Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Mechelen, Belgium, Sep. 2011, pp. 72–77.
[122] C. W. Hsu, C. C. Chang, and C. J. Lin. (2003). *A Practical Guide to Support Vector Classification*. Last updated May 2016. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[123] N. Ponomarenko, O. Ieremeiev, V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, M. Battisti, and C.-C. J. Kuo, "A new color image database TID2013: Innovations and results," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst. (ACVIS)*, Poznan, Poland, Oct. 2013, pp. 402–413.
[124] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Boston, MA, USA, Jul. 2009, pp. 758–759.
[125] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Rochester, NY, USA, vol. 3, Sep. 2002, pp. III:57–III:60.
[126] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Rochester, NY, USA, vol. 1, Sep. 2002, pp. I:477–I:480.
[127] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.
In journal information it should be Ṣno. 11Ť, not Ṣno. 1Ť as it is now.
[128] 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.
[129] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.
[130] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[131] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2012, pp. 1098–1105.
[132] Q. Li, W. Lin, and Y. Fang, "No-reference quality assessment for multiply-distorted images in gradient domain," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 541–545, Apr. 2016.
[133] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.
[134] 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.
[135] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 339–343.
[136] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.
[137] Q. Li, W. Lin, J. Xu, and Y. Fang, "Blind image quality assessment using statistical structural and luminance features," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2457–2469, Dec. 2016.
[138] 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.
[139] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000.
[140] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61. Philadelphia, PA, USA: SIAM, 1992.
[141] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, Feb. 1995.
[142] N. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Cairo, Egypt, Nov. 2009, pp. 2281–2284.
[143] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, Apr. 2011.
[144] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[145] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 3773–3777.

[146] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14.
[147] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, Jun. 2010, pp. 807–814.
[148] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1733–1740.
[149] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015.
[150] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 2791–2795.
[151] J. Fu, H. Wang, and L. Zuo, "Blind image quality assessment for multiply distorted images via convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 1075–1079.
[152] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, and G. Xie, "No-reference image quality assessment using Prewitt magnitude based on convolutional neural networks," *Signal, Image Video Process.*, vol. 10, no. 4, pp. 609–616, Apr. 2016.
[153] J. Li, J. Yan, D. Deng, W. Shi, and S. Deng, "No-reference image quality assessment based on hybrid model," *Signal, Image, Video Process.*, vol. 11, no. 6, pp. 985–992, Sep. 2017.
[154] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognit.*, vol. 81, pp. 432–442, Sep. 2018.
[155] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image, Video Process.*, vol. 12, no. 2, pp. 355–362, Feb. 2018.
[156] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.
[157] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 609–613.
[158] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.
[159] 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. IEEE Int. Conf. Comput. Vis. (ICCV)*, Vancouver, BC, Canada, vol. 2, Jul. 2001, pp. 416–423.
[160] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Bonn, Germany, Aug. 2005, pp. 89–96.
[161] K. Gu, G. Zhai, M. Liu, X. Yang, W. Zhang, X. Sun, W. Chen, and Y. Zuo, "FISBLIM: A five-step blind metric for quality assessment of multiply distorted images," in *Proc. IEEE Workshop Signal Process. Syst. (SiPS)*, Taipei, Taiwan, Oct. 2013, pp. 241–246.
[162] D. Zoran and Y. Weiss, "Scale invariance and noise in natural images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Kyoto, Japan, Sep. 2009, pp. 2209–2216.
[163] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.
[164] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41–52, Jan. 2012.
[165] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL, USA: CRC Press, 2011.
[166] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.
[167] 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.
[168] D. Ghadiyaram and A. C. Bovik, "Feature maps driven no-reference image quality prediction of authentically distorted images," in *Proc. SPIE Electron. Imag.*, San Francisco, CA, USA, vol. 9394, Mar. 2015, pp. 93940J-1–93940J-14.
[169] 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, pp. 32-1–32-25, Jan. 2017.
[170] H. Hadizadeh and I. V. Bajić, "Color Gaussian jet features for no-reference quality assessment of multiply-distorted images," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1717–1721, Dec. 2016.
[171] C. Li, Y. Zhang, X. Wu, and Y. Zheng, "A multi-scale learning local phase and amplitude blind image quality assessment for multiply distorted images," *IEEE Access*, vol. 6, pp. 64577–64586, 2018.
[172] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.
[173] Q. Wu, H. Li, F. Meng, K. N. Ngan, B. Luo, C. Huang, and B. Zeng, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 425–440, Mar. 2016.
[174] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[175] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255.
[176] 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.

SHAHRUKH ATHAR (S'14) received the B.E. degree in electrical engineering from the National University of Sciences and Technology, Rawalpindi, Pakistan, in 2007, and the M.S. degree in computer engineering from the Lahore University of Management Sciences, Lahore, Pakistan, in 2009. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Waterloo, Waterloo, ON, Canada. His research interests include perceptual image quality assessment and large-scale dataset design.

ZHOU WANG (S'99–M'02–SM'12–F'14) received the Ph.D. degree from The University of Texas at Austin, in 2001. He is currently a Canada Research Chair and Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include image and video processing and coding; visual quality assessment and optimization; computational vision and pattern analysis; multimedia communications; and biomedical signal processing. He has more than 200 publications in these fields with over 50 000 citations (Google Scholar).

Dr. Wang was elected a Fellow of the Royal Society of Canada, Academy of Sciences, in 2018, and the Canadian Academy of Engineering, in 2016. He was a recipient of the 2016 IEEE Signal Processing Society Sustained Impact Paper Award, the 2015 Primetime Engineering Emmy Award, the 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, the 2013 IEEE Signal Processing Magazine Best Paper Award, and the 2009 IEEE Signal Processing Society Best Paper Award. He has been serving as a Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, since 2015. Previously, he served as a member of the IEEE Multimedia Signal Processing Technical Committee, from 2013 to 2015, an Associate Editor or a Guest Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, from 2016 to 2018, the IEEE TRANSACTIONS ON IMAGE PROCESSING, from 2009 to 2014, the IEEE SIGNAL PROCESSING LETTERS, from 2006 to 2010, and the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, from 2013 to 2014 and from 2007 to 2009, among other journals.