Degraded Reference Image Quality Assessment

Shahrukh Athar, *Member, IEEE*, and Zhou Wang, *Fellow, IEEE*

***Abstract*— In practical media distribution systems, visual content usually undergoes multiple stages of quality degradation along the delivery chain, but the pristine source content is rarely available at most quality monitoring points along the chain to serve as a reference for quality assessment. As a result, full-reference (FR) and reduced-reference (RR) image quality assessment (IQA) methods are generally infeasible. Although no-reference (NR) methods are readily applicable, their performance is often not reliable. On the other hand, intermediate references of degraded quality are often available, e.g., at the input of video transcoders, but how to make the best use of them in proper ways has not been deeply investigated. Here we make one of the first attempts to establish a new paradigm named degraded-reference IQA (DR IQA). Specifically, by using a two-stage distortion pipeline we lay out the architectures of DR IQA and introduce a 6-bit code to denote the choices of configurations. We construct the first large-scale databases dedicated to DR IQA and have made them publicly available. We make novel observations on distortion behavior in multi-stage distortion pipelines by comprehensively analyzing five multiple distortion combinations. Based on these observations, we develop novel DR IQA models and make extensive comparisons with a series of baseline models derived from top-performing FR and NR models. The results suggest that DR IQA may offer significant performance improvement in multiple distortion environments, thereby establishing DR IQA as a valid IQA paradigm that is worth further exploration.**

***Index Terms*— Image quality assessment, multiple distortions, degraded-reference, distortion behavior analysis, image quality databases.**

I. Introduction

THE goal of objective Image Quality Assessment (IQA) methods is to predict the quality of images as perceived by human eyes. Based on the accessibility to pristine reference content, they are traditionally classified into three paradigms, namely *full-reference* (FR), *reduced-reference* (RR) and *no-reference* (NR) or *blind* IQA (BIQA) [1], [2], as illustrated in Fig. 1. In the literature, FR, RR, and NR IQA algorithms are usually tested and at times trained on image databases where each distorted image has undergone a single (often simulated) stage of distortion. This is in clear contrast to real-world visual content distribution scenarios, as illustrated in Fig. 2, where visual content undergoes multiple stages of distortions before reaching its destination. For example, most consumer cameras and camcorders, including mobile phone cameras, store captured content using lossy compression standards such as JPEG and H.264. When these images and videos are uploaded to a social networking or video-sharing website, they usually undergo further rounds of lossy transcoding [3], [4] for onward delivery to viewers. This means two stages of compression. For another example, an image or video maybe contaminated by noise or blur during acquisition. The camera will store this content in compressed form which may be followed by further compression during its distribution. This essentially means blur or noise contamination followed by compression. Compressed medical images provide another example of content afflicted by multiple distortion stages. It is known that magnetic resonance (MR), computed tomography (CT), and ultrasound images are affected by different types of noise [5], [6], [7]. With the rapid increase in the resolution and volume of medical images and with the emergence of tele-medicine, it is now desirable to largely reduce the data rate by lossy image compression as long as it does not affect the diagnostic quality [8], [9]. This leads to a distortion combination of noise followed by compression. Compressed astronomical images provide yet another example of noise followed by compression since such images are contaminated by noise [10]. Thus, even if we start with a pristine reference image, it may be affected by multiple stages of distortions by the time it reaches the end user. The requirement for IQA methods capable of handling multiple simultaneous distortions is not new [11], but remains a major challenge [12].

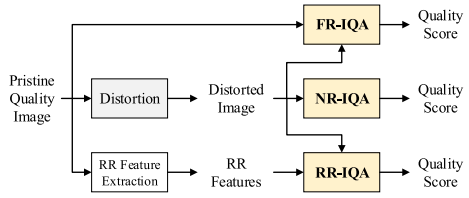


Fig. 1. General framework of FR, RR and NR IQA.

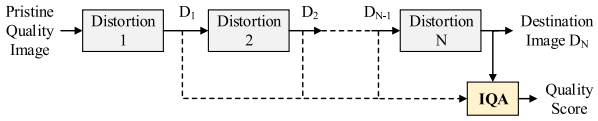


Fig. 2. The framework of practical media distribution systems.

In practical media delivery systems, access to pristine reference images in the middle of the delivery chain is either extremely rare or altogether nonexistent. This, coupled with the multiple distortion nature of such systems, makes the use of FR and RR IQA infeasible. While NR IQA methods

Manuscript received 30 October 2021; revised 19 April 2022 and 11 August 2022; accepted 27 September 2022. Date of publication 10 January 2023; date of current version 13 January 2023. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastian Bosse. *(Corresponding author: Shahrukh Athar.)*

Shahrukh Athar is with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada (e-mail: shahrukh.athar@uwaterloo.ca).

Zhou Wang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: zhou.wang@uwaterloo.ca).

Digital Object Identifier 10.1109/TIP.2023.3234498

are readily applicable, most NR methods are trained and tested on databases that have images with a single stage of distortion, and their performance lags behind that of FR IQA methods [13]. Efforts have been made recently to design NR IQA methods that handle multiply distorted images. SISBLIM [14] is a training-free metric designed for singly and multiply distorted images through the fusion of estimates of noise, blur, JPEG compression, and joint effects. BoWSF [15] selects features sensitive to different distortion types, which are encoded through a Bag-of-Words model and mapped to a quality score. LQAF [16] uses Support Vector Regression (SVR) to map features such as phase congruency, gradient magnitude, gray level gradient co-occurrence matrix and the contrast sensitivity function to quality scores. An enhanced and multi-scale version of LQAF, called MS-LQAF is proposed in [17]. The training-based GWHGLBP [18] uses the gradient-weighted histogram of the local binary pattern (LBP) generated on the gradient map of the distorted image to capture the effects of multiple distortions. Jet-LBP [19] uses color Gaussian jets to generate feature maps from a distorted image. The LBP is applied to these feature maps, followed by a weighted histogram that is mapped to quality scores through SVR. MUSIQUE [20] performs distortion identification followed by distortion parameter estimation and score generation. Nevertheless, we showed in [21] that NR IQA methods generally perform unsatisfactorily when dealing with multiply distorted images, especially when the distortion types vary and with high distortion levels at earlier stages.

In addition to the performance issues, another major limitation of NR IQA methods is that they are incapable of incorporating the mid-stage distorted images along the media delivery chain, shown as D_1 to D_{N-1} in Fig. 2, to determine the quality of the final multiply distorted image D_N. For example, at the input of a video transcoder, typically a compressed video stream of degraded quality, is often available but not used by NR IQA methods to assess the video stream at transcoder output. Similarly, FR and RR IQA are also unable to incorporate this additional information in the quality assessment task. The question is: How to best utilize the auxiliary information of the mid-stage images of degraded quality to assess the quality of the final multiply distorted images? We term this problem *degraded-reference* IQA (DR IQA) and define it as *determining the quality of a final multiply distorted image given access to its degraded reference image(s), but with no or limited access to the pristine reference image*. In Fig. 2 images D_1 to D_{N-1} are the degraded references of the final multiply distorted image D_N.

This work is by no means the first attempt to tackle the DR IQA problem. A pioneering work is the *corrupted-reference* (CR) IQA scheme laid out in the context of an image restoration problem [22], [23], [24], where the quality of a denoised image with respect to an absent pristine reference image is estimated by using a Gaussian or Poisson noise contaminated corrupted reference image. A recent interesting work targeting at DR IQA in image restoration scenarios learns a reference space for DR images by knowledge distillation from pristine images [25]. These are instantiations of Type-010010 DR IQA based on the categorization we will introduce in Section II. In our earlier work [21], we show that Type-100100 DR IQA offers the potential to substantially elevate the performance of quality prediction against two baselines: FR IQA between the degraded-reference and final distorted images, and NR IQA of the final distorted image. The two-step quality assessment (2stepQA) scheme [26], [27], [28] represents a series of Type-001100 DR IQA instantiations, where many combinations of FR methods (PSNR, MSSSIM [29], FSIM [30], VSI [31]) between the degraded-reference and final distorted images, and NR methods (NIQE [32], BRISQUE [33], CORNIA [34], PQR [35]) on the degraded reference images, have demonstrated great promises, though it does not take into account how different distortions behave in conjunction with each other. Other types of DR IQA architectures (as elaborated in Section II), to the best of our knowledge, have not been attempted in the literature.

In this paper, we make one of the first attempts to establish DR IQA as a new IQA paradigm. We lay out and discuss 53 potential architectures for DR IQA in a two-stage distortion pipeline (Section II). We construct two new large-scale synthetically annotated datasets dedicated to DR IQA (Section III-A). We study the behaviors in multiple simultaneous distortions and make some interesting observations not reported before (Section III-B). Based on these observations, we develop novel DR IQA models (Section IV) and make extensive comparisons with baseline models derived from top-performing FR and NR models (Section V). Finally, we conclude that the DR IQA paradigm offers great potentials and is worth further exploration in the future (Section VI).

II. Degraded Reference IQA Architectures

Although in practice images and videos may undergo many stages of distortions, a logical starting point to study DR IQA is to focus on the case of two stages of distortions. This allows for a thorough discussion about all potential ways to perform DR IQA without missing the main issues that may be encountered in the cases of many stages of distortions. In a two-stage distortion pipeline, a source of a *pristine reference* (PR) image undergoes stage-1 distortion, leading to a *degraded reference* (DR) image, which subsequently undergoes stage-2 distortion and results in a *final distorted* (FD) image, as shown at the top part of Fig. 3. In DR IQA, both DR and FD images are assumed available, while the PR image is generally not accessible. However, in certain circumstances, information regarding the degradation from the PR to DR images may be made available to the DR IQA module, which may help improve the performance of DR IQA algorithms.

Although traditional FR/RR/NR IQA methods alone do not directly provide adequate solutions to the DR IQA problem, they may be employed as key components in the total solution. Depending on how they are utilized, there may be many configurations of DR IQA architectures, which are summarized in Fig. 3. In particular, the FR/RR IQA computed between the PR and DR images, and between the DR and FD images, the NR IQA computed from the DR and FD images, together with the DR and FD images themselves, may all be part of the input to the DR IQA module, which

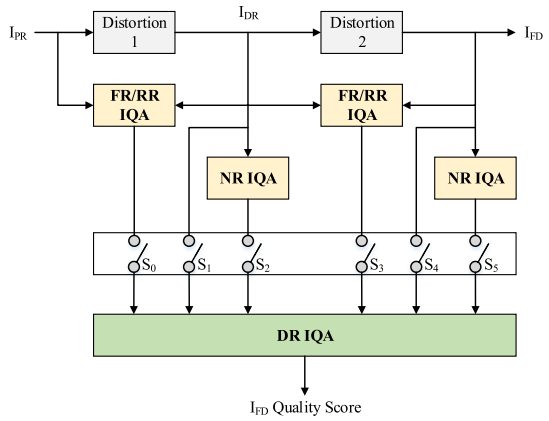


Fig. 3. General architectures of DR IQA. Configurations determined by the statuses of six switches. See Table I for more information.

TABLE I

INFORMATION CONNECTED WITH SWITCHES (S_0 TO S_5) IN THE DR IQA ARCHITECTURE SHOWN IN FIG. 3. EACH SWITCH IS CONTROLLED BY A CODE BIT FROM THE 6-BIT ARCHITECTURE CODE

Switch	Code bit	Type of Information Connected
S_0	bit0	FR/RR IQA between PR & DR images (score or features)
S_1	bit1	DR image
S_2	bit2	NR IQA applied to DR image (score or features)
S_3	bit3	FR/RR IQA between DR & FD images (score or features)
S_4	bit4	FD image
S_5	bit5	NR IQA applied to FD image (score or features)

TABLE II

DR IQA ARCHITECTURE TYPE GROUPS AND SWITCH CODE

Architecture Type	Architecture Switch Code
Type-0 Group	000100, 000101, 000110, 000111, 001001, 001010, 001011, 001100, 001101, 001110, 001111, 010001, 010010, 010011, 010100, 010101, 010110, 010111, 011001, 011010, 011011, 011100, 011101, 011110, 011111
Type-1 Group	100001, 100010, 100011, 100100, 100101, 100110, 100111, 101001, 101010, 101011, 101100, 101101, 101110, 101111, 110001, 110010, 110011, 110100, 110101, 110110, 110111, 111001, 111010, 111011, 111100, 111101, 111110, 111111
Invalid Group	000000, 000001, 000010, 000011, 001000, 010000, 011000, 100000, 101000, 110000, 111000

produces a single quality score of the FD image. Depending upon the nature of the DR IQA module, the respective output of each FR/RR and NR IQA module, shown in Fig. 3, may be a predicted quality score, a quality map, or extracted features. The inputs to the DR IQA module may be controlled by six switches, each of which may be turned on or off independently, resulting in 64 potential configurations of DR IQA architectures, each represented by a 6-bit architecture code, where a bit of 1 or 0 denotes a switch being ON or OFF, respectively. The six switches, S_0 to S_5, are respectively controlled by bit0 to bit5 of the 6-bit architecture code and their order is given in Fig. 3. Table I provides details about the type of information connected with each switch along with its respective controlling code bit.

The 64 potential architectures may be further classified into three groups. In the case that the PR image is completely inaccessible (even in gauging the FR/RR quality degradation from the PR image to the DR image), the first bit of the architecture code equals 0. Therefore, we categorize these architectures into the Type-0 group. Correspondingly, when information regarding the PR image is accessed through the FR/RR measure between PR and DR images, we classify the architectures into the Type-1 group. There are also invalid configurations for DR IQA when either the DR or the FD image is completely inaccessible to the DR IQA module. The classifications and the corresponding architecture codes are listed in Table II, where there are 25, 28, and 11 architectures in the Type-0, Type-1 and invalid groups, respectively.

In practice, not all of the 53 (25+28) valid DR IQA architectures are equally favored, and the choice is likely dependent on the application scenarios. For instance, FR IQA methods are generally more reliable than other IQA approaches, and thus Type-1***** or Type-***1** are desirable. But in practical video delivery pipelines, obtaining both PR and DR videos or both DR and FD videos at the same monitoring points may not be easy, and even when the condition is met, the videos are often not aligned along the temporal dimension. For another example, when no existing FR/RR/NR IQA method is available, then Type-010010 is the only option, but this would require a completely new design of the DR IQA approach. More sophisticated cases may also occur with certain mixtures of distortion types in the two stages of distortions, which will be elaborated in more detail in Section III-B.

Scrutinizing observers may find some of the 53 valid DR IQA architectures to be redundant. For example, when both the DR and FD images are available, as in the case of Type-010010, then the related architectures of Type-01**1* may all be redundant, because the associated FR or NR IQA computations can be done inside the DR IQA module. However, this is based on the assumption that all the FR, NR and DR computations are performed at a centralized location. In practice, this may not always be the case, as the FR and NR IQA measures may be computed at local monitoring points (which may serve certain local decision making purposes), while DR IQA is conducted at a central location, for example, in the cloud, where the FR and NR models may not be available, or recomputing them is not energy efficient but reusing them may largely reduce the computational burden of DR IQA. Therefore, it is meaningful to identify them as distinct types.

In the rest of this paper, we focus on two scenarios corresponding to two DR IQA architectures, based on which we study the behaviors of two-stage quality variations under different distortion combinations in Section III-B and develop DR IQA models in Section IV. For this work, we limit ourselves at taking a model predicted score from each FR and NR IQA module that may be present within the two DR IQA architectures. The first scenario corresponds to Type-100100 DR IQA, and this architecture is given in Fig. 4. Since FR methods provide the most reliable quality prediction performance [13] and the PR image is assumed to have perfect quality, to simplify the analysis, we define the *absolute scores* (AS) of a test image as the FR measure between the test image and its corresponding PR image. In the case of two distortion stages, we may compute the AS scores of the DR and FD images as Eq. 1 and 2:

$$\mathrm{AS}_{\mathrm{DR}} = \mathrm{FR}(\mathrm{I}_{\mathrm{PR}}, \mathrm{I}_{\mathrm{DR}}), \tag{1}$$

$$\mathrm{AS}_{\mathrm{FD}} = \mathrm{FR}(\mathrm{I}_{\mathrm{PR}}, \mathrm{I}_{\mathrm{FD}}), \tag{2}$$

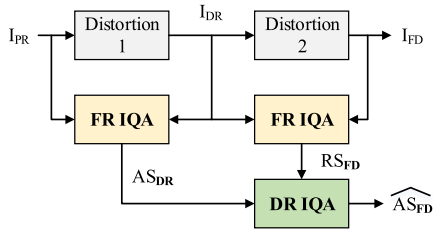


Fig. 4. Scenario 1: Type-100100 DR IQA architecture.

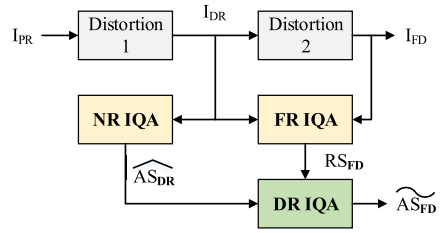


Fig. 5. Scenario 2: Type-001100 DR IQA architecture.

respectively, where I_{PR}, I_{DR}, and I_{FD} are the PR, DR and FD images, respectively. The third possible FR comparison is between the DR and FD images, which assesses the quality of the FD image relative to the DR image. Thus, we regard it as the *relative score* (RS) of the FD image:

$$RS_{FD} = FR(I_{DR}, I_{FD}). \tag{3}$$

Ideally, if FR IQA is fully trusted, then AS_{FD} yields the best quality estimate of the FD image, but cannot be computed as the PR image is not available, at the end user level. Therefore, in Scenario 1, which assumes the availability of the PR image early on in the media distribution system, the goal of the DR IQA module in Fig. 4 is to make the best prediction of AS_{FD} using AS_{DR} and RS_{FD}, i.e.,

$$\widehat{AS}_{FD} = f(AS_{DR}, RS_{FD}), \tag{4}$$

where f is the prediction function of the DR IQA module. Scenario 1 is practically applicable only when AS_{DR} is pre-computed at the first distortion stage and transmitted as side information with the DR image to the second distortion stage. Apparently, transferring AS_{DR} would require minor protocol changes of the media delivery system.

The second scenario follows Type-001100 DR IQA architecture, as shown in Fig. 5, where the PR image is completely inaccessible, and thus AS_{DR} cannot be computed. Instead, an NR IQA method is employed to produce an estimated score $\widehat{AS}_{DR}$:

$$\widehat{AS}_{DR} = NR(I_{DR}). \tag{5}$$

Correspondingly, the DR IQA module is designed to predict AS_{FD} by using NR IQA predicted $\widehat{AS}_{DR}$ and RS_{FD}:

$$\widetilde{AS}_{FD} = g(\widehat{AS}_{DR}, RS_{FD}), \tag{6}$$

where g is the prediction function of the DR IQA module and may take a similar form as f in Eq. 4. Unlike the first scenario, there is no need here to make any change to the existing media distribution system, and the DR IQA systems can be deployed as passive quality monitoring probes. This makes Scenario 2 both practically applicable and readily deploy-able.

TABLE III
COMPOSITION OF DR IQA DATABASES V1 AND V2

Reference Images in each Database (Pristine Quality)	Stage-1 Distorted Images in each Database (Singly Distorted DRs)		Stage-2 Distorted Images in each Database (Multiply Distorted FDs)	
Number of Images	Distortion	Number of Images	Distortion Combination	Number of Images
34	Blur	374	Blur-JPEG	6,358
			Blur-Noise	6,358
	JPEG	374	JPEG-JPEG	6,358
	Noise	374	Noise-JPEG	6,358
			Noise-JP2K	6,358
	Total	1,122	Total	31,790
	Overall 32,912 Distorted Images in each Database			

III. MULTIPLE DISTORTIONS: DATABASE CONSTRUCTION AND BEHAVIOR ANALYSIS

A. Database Construction

Most existing IQA datasets of multiply distorted content such as LIVE MD [36], MDIVL [37], MDID [38], MDID2013 [14], and LIVE WCmp [27], use a limited number of distortion levels (typically 3 or 4) per distortion type per stage, making it difficult to analyze how different constituent distortions behave in conjunction with each other. The Waterloo Exploration-II (Exp-II) database, that we developed in [39], has 3,570 PR, 117,810 singly distorted, and 3,337,950 multiply distorted images, offering an excellent testbed for two-stage DR IQA. Here we construct two new datasets, namely DR IQA database Version 1 (V1) and DR IQA database Version 2 (V2), by following the same procedure as [39], but without any cross-dataset content overlap. The purpose here is to use one dataset for training and the other for validation in a machine learning process, and then use the Waterloo Exp-II database for independent testing.

A total of 68 pristine quality reference images were taken from the following sources: IQA databases CSIQ [40], IVC [41], LIVE R2 [42], TID2013 [43], Toyoma [44] and some pristine images were extracted from raw videos available at CDVL [45]. These images were divided into two disjoint groups of 34 images each, with one group forming the pristine image set of DR IQA database V1 and the other forming the pristine image set of DR IQA database V2. Table III outlines the composition of DR IQA databases V1 and V2.

We include singly distorted DR images belonging to three distortion categories of Gaussian white noise, Gaussian blur, and JPEG compression. The stage-1 DR images are created in the *fair* to *excellent* perceptual quality range at distortion levels 1 to 11 based on a content adaptive distortion process [39]. This leads to 374 DR images for each of the three single distortion types in each dataset.

The stage-2 multiply distorted FD images belong to the following five distortion combinations which simulate real-world multiple distortion scenarios and are examples of practical applications of DR IQA: 1) Gaussian blur followed by JPEG compression (Blur-JPEG or B-JPG) simulates the storage of blurry images through JPEG compression; 2) Gaussian white noise followed by JPEG compression (Noise-JPEG or N-JPG) simulates the storage of noisy images through JPEG compression; 3) Gaussian white noise followed by JPEG2000 compression (Noise-JP2K or N-JP2) simulates the storage of noisy images through JPEG2000 compression;

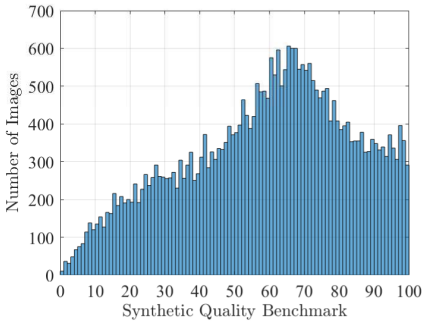


Fig. 6. SQB histogram of DR IQA database V1.

4) JPEG compression followed by JPEG compression (JPEG-JPEG or JPG-JPG) simulates multiple levels of compression (e.g., images taken by cell phone cameras are usually JPEG compressed and may undergo further compression when uploaded to a social media platform); and 5) Gaussian blur followed by Gaussian white noise (Blur-Noise or B-N) not only simulates different image capture conditions but can also simulate image acquisition followed by transmission [14]. Stage-2 FD images are created by starting with the respective DR images and applying content adaptive distortion parameters [39] at levels 1 to 17, which correspond to the entire *bad* to *excellent* perceptual quality range. This leads to 6,358 FD images for each distortion combination in each dataset. By having 11 stage-1 and 17 stage-2 distortion levels, we have ensured an adequate density of distortion levels per distortion stage. Overall, each database has 32,912 distorted images.

Conducting subjective testing for such large datasets is extremely difficult. To find an alternative automatic data annotation mechanism, we conducted the largest IQA performance evaluation study to-date [13], where we find that fusing multiple FR IQA methods through a training-free reciprocal rank fusion (RRF) strategy [46] (first used in IQA in [47]) offers robust perceptual quality prediction performance. Based on RRF, we develop a synthetic quality benchmark (SQB) scheme [39] that fuses four FR methods including CIDMS [48], DSS [49], IWSSIM [50], and VIFDWT [51]. SQB outperforms all state-of-the-art FR and fused FR methods, on subject-rated datasets, that it was tested against [39]. Furthermore, by utilizing the SQB-annotated Waterloo Exp-II dataset, we trained a deep neural network (DNN) based NR IQA model called EONSS [39]. Exclusively trained on SQB-annotated Waterloo Exp-II dataset, and exclusively tested on nine subject-rated datasets, EONSS outperforms all state-of-the-art NR IQA methods that it was tested against, which in turn, justifies the reliability of the SQB annotation strategy. More information about SQB can be found in [39].

We use SQB to annotate DR IQA databases V1 and V2, and provide the SQB histogram of DR IQA database V1 in Fig. 6 which shows a wide representation of the full quality range (SQB dynamic range is between 0 and 100, where the best quality is represented by the latter), with a higher concentration on the higher quality half, which is the working range in most real-world applications. The SQB histogram of DR IQA database V2 is similar to the one shown in Fig. 6. We have made both DR IQA databases V1 and V2 publicly available on IEEE DataPort. DR IQA database V1 is available at https://dx.doi.org/10.21227/4795-vv06 and DR IQA database V2 is available at https://dx.doi.org/10.21227/8r47-gp07.

B. Multiple Distortions: Behavior Analysis

The large-scale DR IQA databases V1 and V2 provide us with a platform to investigate the behaviors of images undergoing multiple stages of distortions, which is essential in building effective DR IQA methods. We start with a visual example given in Figs. 7 and 8, where the PR *Barbara* image undergoes three types of stage-1 distortions (Gaussian noise, Gaussian blur, JPEG compression) at level 7, which then undergo stage-2 distortions (Gaussian noise, JPEG and JPEG2000 compression) at level 11 to create five distortion combinations (Blur-JPEG, Blur-Noise, JPEG-JPEG, Noise-JPEG, Noise-JP2K), resulting in 3 DR and 5 FD images, for which we compute their SSIM [52] quality maps that indicate quality variations over space (brighter suggests better quality). The various image-level AS_{DR}, AS_{FD}, and RS_{FD} quality scores for these images are also computed by SSIM[1] [52] (higher indicates better quality) and are given in Table IV.

There are several useful observations from these visual examples and SSIM results. 1) The relative quality maps of the FD images (Fig. 8(d-h)) are drastically different from their respective absolute quality maps (Fig. 8(i-m)), suggesting RS_{FD} is not a good predictor of AS_{FD} in general and FR methods should be used with caution in the absence of PR images. 2) For the cases of Blur-JPEG, Blur-Noise, and JPEG-JPEG, the relative quality maps are lighter than the absolute quality maps, showing an over-estimation of AS_{FD} by RS_{FD}. 3) However, the opposite is observed for the cases of Noise-JPEG and Noise-JP2K, where their relative quality maps are darker, and RS_{FD} under-estimates AS_{FD}. 4) For the cases of Blur-JPEG, Blur-Noise, and JPEG-JPEG, the absolute quality map of the FD image appears to be roughly an accumulative combination of the absolute quality map of the DR image and the relative quality map of the FD image. 5) However, this is obviously not the case for Noise-JPEG and Noise-JP2K, suggesting sophisticated distortion combination behaviors.

To investigate further, we use the FR IQA method FSIMc [30], which was found to be among the top performing FR methods in [13], to compute the AS_{DR}, RS_{FD}, and AS_{FD} scores (higher indicates better quality) for various DR and FD *Barbara* images created from all 11-level stage-1 distortions and 17-level stage-2 distortion combinations, and plot AS_{FD} versus RS_{FD} scores in Fig. 9. For each distortion combination, there are 11 curves, each corresponding to one of the 11 stage-1 DR images and containing 17 points that represent the corresponding stage-2 FD images. The dotted horizontal lines represent the quality level of the DR images.

[1]To enhance the visibility of the quality maps in Fig. 8, we have chosen the original version of the FR IQA method SSIM [52] that does not implement automatic downsampling. Consequently the same version of SSIM has been used to generate the scores in Table IV. While we have used SSIM due to its quality map feature to give the visual demonstration in Fig. 8, we recognize that better FR methods exist and have used FSIMc [30] in subsequent analysis.

(a) PR (b) DR (Blur) (c) DR (JPEG) (d) DR (Noise)

(e) FD (Blur-JPEG) (f) FD (Blur-Noise) (g) FD (JPEG-JPEG) (h) FD (Noise-JPEG) (i) FD (Noise-JP2K)

Fig. 7. Example *Barbara* images: (a) PR image; (b-d) DR images (Stage 1 distortion level 7); (e-i) FD images (Stage 2 distortion level 11).

(a) DR (Blur) (b) DR (JPEG) (c) DR (Noise)

(d) FD (Blur-JPEG) (e) FD (Blur-Noise) (f) FD (JPEG-JPEG) (g) FD (Noise-JPEG) (h) FD (Noise-JP2K)

(i) FD (Blur-JPEG) (j) FD (Blur-Noise) (k) FD (JPEG-JPEG) (l) FD (Noise-JPEG) (m) FD (Noise-JP2K)

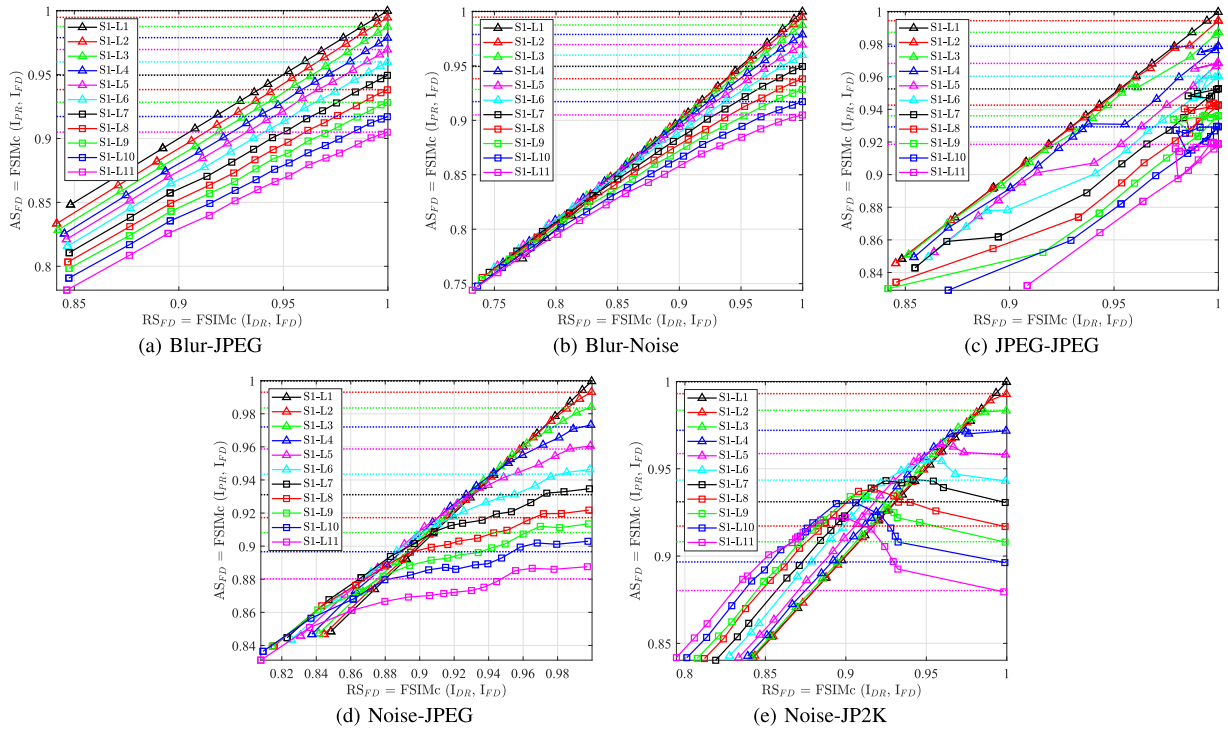
Fig. 8. SSIM [52] Quality Maps of the example *Barbara* images of Fig. 7: (a-c) Absolute Quality Maps of the DR images with respect to PR; (d-h) Relative Quality Maps of the FD images with respect to their DR images; (i-m) Absolute Quality Maps of the FD images with respect to PR.

TABLE IV

SSIM AS_{DR}, RS_{FD}, AND AS_{FD} SCORES FOR EXAMPLES IN FIG. 7

Distortion Combination	SSIM AS_{DR}		SSIM RS_{FD}		SSIM AS_{FD}	
	between	score	between	score	between	score
Blur-JPEG	(a)&(b)	0.7245	(b)&(e)	0.8727	(a)&(e)	0.6129
Blur-Noise	(a)&(b)	0.7245	(b)&(f)	0.3184	(a)&(f)	0.2860
JPEG-JPEG	(a)&(c)	0.8137	(c)&(g)	0.9353	(a)&(g)	0.7525
Noise-JPEG	(a)&(d)	0.6369	(d)&(h)	0.5066	(a)&(h)	0.7205
Noise-JP2K	(a)&(d)	0.6369	(d)&(i)	0.4841	(a)&(i)	0.7615

The key observations are summarized as follows: 1) When stage-1 distortion is minimum (level-1), the DR image is almost as good as the PR image, and not surprisingly the prediction from RS_{FD} to AS_{FD} is nearly perfect, demonstrated as the straight lines of the Stage-1 Level-1 (S1-L1) curves in all five distortion combinations in Fig. 9. However, the trend changes dramatically with the increasing stage-1 distortion levels; 2) For all five distortion combinations, with the exception of some Noise-JPEG curves, at minimum stage-2 distortion (Stage-2 Level-1), $RS_{FD} \simeq 1$ and $AS_{FD} \simeq AS_{DR}$, which is not surprising since stage-2 is not adding any further distortion to the DR image; 3) For Blur-JPEG and



(a) Blur-JPEG (b) Blur-Noise (c) JPEG-JPEG

(d) Noise-JPEG (e) Noise-JP2K

Fig. 9. AS_{FD} versus RS_{FD} plots for all stage-1 (AS_{DR}) distortion levels (S1-L1 to S1-L11) corresponding to five distortion combination types for the *Barbara* image. Dotted lines represent AS_{DR} scores. The FR IQA method FSIMc [30] was used to obtain all scores.

JPEG-JPEG[2] distortion combinations (Fig. 9(a) and (c)), the curves consistently move away nearly in parallel (especially for Blur-JPEG) from the S1-L1 curve with increasing stage-1 distortion level, suggesting additive quality degradation of the two distortion stages; 4) The Blur-Noise distortion combination (Fig. 9(b)) follows a similar behavior as the Blur-JPEG and JPEG-JPEG cases, but the curves converge with increasing stage-2 distortion, implying that as the magnitude of the stage-2 distortion, i.e., Gaussian noise, increases, it overshadows the stage-1 distortion (Gaussian blur) and becomes the dominant distortion factor; 5) The Noise-JPEG and Noise-JP2K distortion combination cases (Fig. 9(d) and (e)) exhibit more interesting nonlinear behaviors. Notably some portions of the curves go above their respective AS_{DR} baseline, i.e., overshoots take place. This behavior is most apparent in the low to mid-level stage-2 distortion levels corresponding to mid to high level stage-1 distortion levels and is much more pronounced in the Noise-JP2K case. Most interestingly, the overshoot phenomenon indicates that the corresponding FD images have better quality than their respective DR images from which they are created, suggesting that the denoising effect of compression may help improve image quality at certain noise and compression levels. A visual demonstration is given in Fig. 10, where the PR *Barbara* image is distorted at Gaussian noise level 11 to generate the DR image, which is then further distorted by JPEG2000 compression level 6 to generate the FD image. The quality maps of the DR and FD images with respect to the PR image and their SSIM scores clearly show that the quality of the FD image improves upon the DR image. 6) For Noise-JPEG and Noise-JP2K, especially the latter, we also note that unlike the other three distortion combinations, curves corresponding to higher stage-1 distortion levels are not always below those with lower stage-1 distortion. Instead as stage-2 distortion increases, a crossover takes places which is more evident for higher stage-1 distortion levels. Again this points to complex joint effects of the two distortion types involved in the combination where noisier stage-1 images are better impacted by the denoising effect of compression. 7) Although Blur-JPEG, JPEG-JPEG, and Noise-JPEG, have different stage-1 distortions but the same stage-2 distortion (JPEG compression), Figures 9(a), (c), and (d) respectively show that their quality behavior is quite different which indicates that the behavior of constituent distortions in multiply distorted content needs to considered in conjunction with each other rather than separately. Overall, the multi-stage quality variation behavior is highly dependent on the distortion combinations, and the relationship of the two distortion stages on quality ranges from nearly linear to highly nonlinear. While we have presented the above behavior analysis using one example due to space limit, similar behaviors have been observed in the whole DR IQA V1 and V2, and Waterloo Exp-II [39] databases.

[2]Fig. 9(c): Some JPEG-JPEG AS_{FD} versus RS_{FD} curves, especially corresponding to lower AS_{DR} scores, seem to form closed loops. This happens when stage-2 JPEG compression (used to create the FD image) is lower than stage-1 JPEG compression (used to create the parent DR image).

IV. DR IQA Model Design

Here we make one of the first attempts to develop DR IQA models. Rather than adopting sophisticated methods such as deep neural networks, we opt for straightforward and empirical approaches because our main goal is to establish DR IQA as a new paradigm, that offers advantages in handling multiple-distortion cases, through head-to-head comparisons with top-performing FR or NR methods (baselines in Section V). The transparency of these approaches also demonstrates

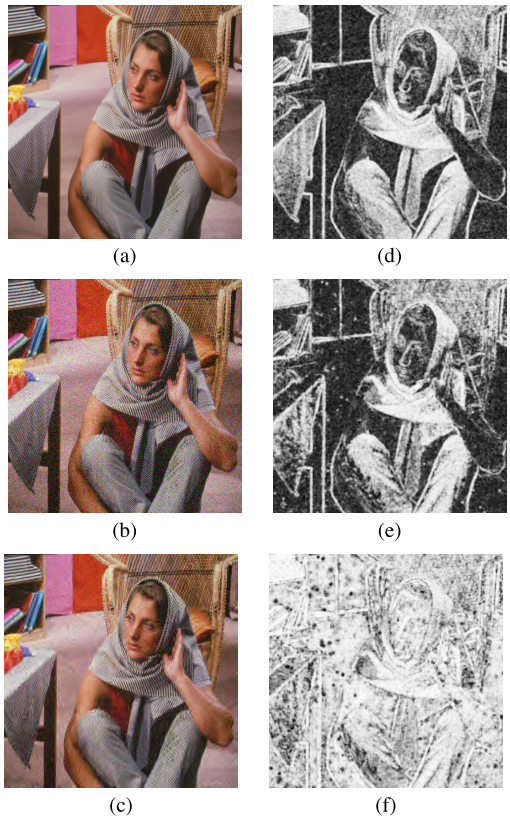


Fig. 10. Noise-JP2K distortion combination example. (a) PR *Barbara* image; (b) DR image contaminated by white Gaussian noise (level 11), SSIM AS_{DR} = 0.5053; (c) FD image obtained by compressing the DR image with JPEG2000 (level 6), SSIM RS_{FD} = 0.4825, SSIM AS_{FD} = 0.7889; SSIM quality maps of: (d) DR with respect to PR image; (e) FD with respect to DR image; (f) FD with respect to PR image.

the value of understanding the multiple-distortion behaviors (as discussed in Section III-B) in DR IQA modeling.

A. Model 1: Distortion Behavior Model

The first model is motivated by the observations on distortion behaviors from Fig. 9, where for each given stage-1 distortion level, the relationship between AS_{FD} and RS_{FD} is often well fitted by a straight line (with the exception of Noise-JP2K), anchored by the rightmost point at (RS_{FD}, AS_{FD}) = (1, AS_{DR}) for each curve, especially in Blur-JPEG, Blur-Noise, JPEG-JPEG, and Noise-JP2K cases. In Scenario 1, or Type-100100 DR IQA, shown in Fig. 4 and represented by Eq. 4, both FR computed AS_{DR} and RS_{FD} are known, and thus an estimate of AS_{FD} directly follows from the point-slope formula for each curve, given by

$$\widehat{\text{AS}_{\text{FD}}} = m \cdot (\text{RS}_{\text{FD}} - 1) + \text{AS}_{\text{DR}}. \tag{7}$$

It remains to determine the slope parameter m. For each distortion combination, we use least-square regression to obtain the best value of m for each of the 11 stage-1 distortion levels for all the DR images in DR-IQA databases V1 and V2. The plots of the best coefficient m versus AS_{DR} are shown in Fig. 11. Somewhat surprisingly, for each distortion combination type (with the exception of JPEG-JPEG at high stage-1 distortion), the behavior of coefficient m with respect to AS_{DR} is rather quite linear, suggesting that the optimal value of m may be directly predicted from AS_{DR} by

$$\widehat{m} = P_1 \cdot \text{AS}_{\text{DR}} + P_2, \tag{8}$$

where P_1 and P_2 are the slope and intercept coefficients, respectively. Replacing m with $\widehat{m}$ and plugging into Eq. 7, we obtain

$$\begin{aligned}\widehat{\text{AS}_{\text{FD}}} = {} & P_1 \cdot \text{AS}_{\text{DR}} \cdot \text{RS}_{\text{FD}} + P_2 \cdot \text{RS}_{\text{FD}} \\ & + (1 - P_1) \cdot \text{AS}_{\text{DR}} - P_2.\end{aligned} \tag{9}$$

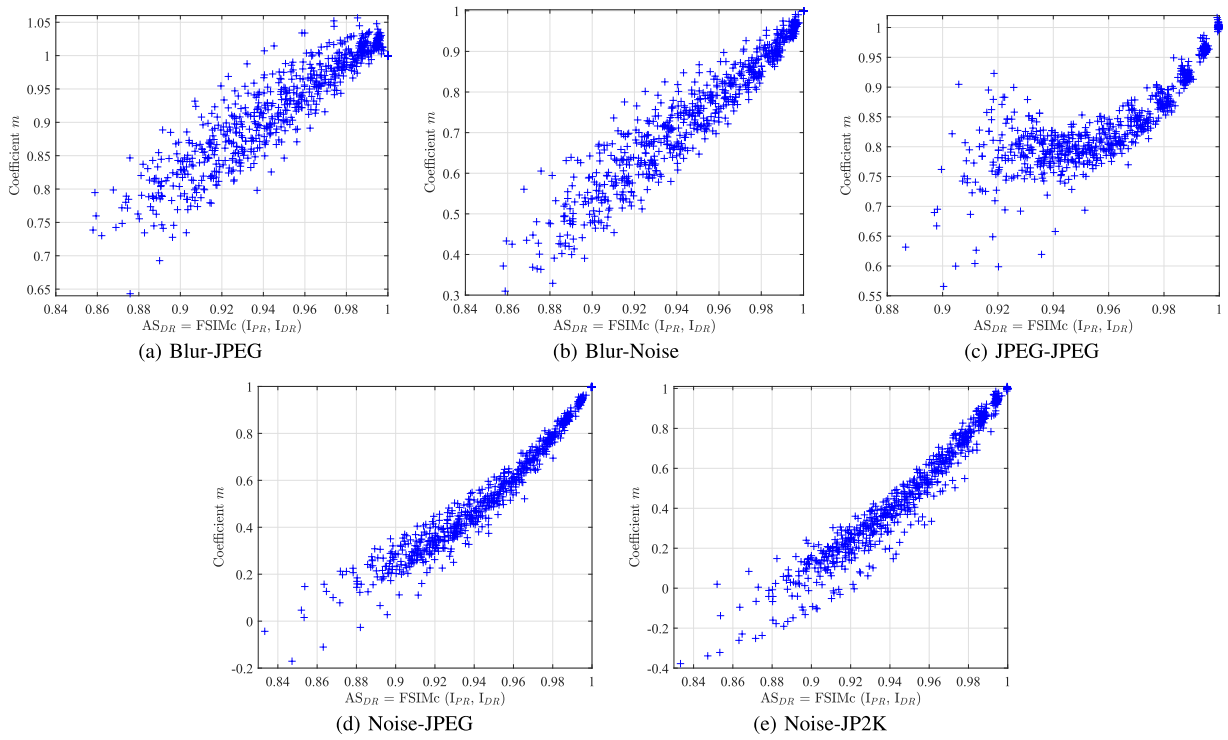
As such, by following a 2-tier modeling approach, we narrow down the number of parameters to two (P_1 and P_2) that need to be found separately for each distortion combination. Since five multiple distortion combinations are being considered in this work (Blur-JPEG, Blur-Noise, JPEG-JPEG, Noise-JPEG, and Noise-JP2K), this requires that parameters P_1 and P_2 be separately determined for all five cases. This also makes the application of the resulting Model 1 versions distortion combination specific.

In addition to developing the 2-tier Model 1 above for the five distortion combinations separately, we also construct it (i.e., determine parameters P_1 and P_2) for two general-purpose cases: 1) NBJ-JPEG (NBJ-JPG), where the distortion combinations of Noise-JPEG, Blur-JPEG, and JPEG-JPEG are considered together (so that comparisons can be made with 2stepQA [26], [27], which is designed for the case where the second distortion stage is JPEG compression); and 2) the *All Data* case, where all five distortion combinations are considered together. These two additional versions of Model 1, especially the *All Data* case, are not specific to any particular distortion combination, i.e., they are more generally applicable compared to the earlier five versions. For Scenario 1, we use the FR method FSIMc [30] to compute both AS_{DR} and RS_{FD}, and call this AS_{DR}-RS_{FD} combination FSIMc-FSIMc.

For Scenario 2 or Type-001100 DR IQA as shown in Fig. 5 and represented by Eq. 6, we use three NR IQA methods, CORNIA [34], dipIQ [53], and NIQE [32] to predict the quality of DR images, i.e., to find $\widehat{\text{AS}_{\text{DR}}}$ (Eq. 5). CORNIA and dipIQ are selected for their top performance in [13], and NIQE has been used in the 2stepQA model [26], [27]. Combining with the FR method FSIMc [30] (used to find RS_{FD}), this leads to three $\widehat{\text{AS}_{\text{DR}}}$-$\text{RS}_{\text{FD}}$ combinations: CORNIA-FSIMc, dipIQ-FSIMc, and NIQE-FSIMc. We also include the NIQE-MSSSIM combination (where the FR method MSSSIM [29] is used to compute RS_{FD}), so as to make direct comparisons with 2stepQA [26], [27], which also uses this combination. Specifically, we learn a nonlinear mapping from CORNIA, dipIQ, and NIQE to FSIMc for the CORNIA-FSIMc, dipIQ-FSIMc, and NIQE-FSIMc combinations, and from NIQE to MSSSIM for the NIQE-MSSSIM combination, with a five-parameter modified logistic function [42]:

$$F(N) = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + e^{\{\beta_2 (N - \beta_3)\}}} \right] + \beta_4 N + \beta_5, \tag{10}$$

where N denotes NR scores, F denotes mapped FR scores after the mapping step, and β_1, β_2, β_3, β_4, and β_5 are parameters tuned using DR IQA databases V1 and V2, and fixed for testing. The NR-predicted and FR-mapped $\widehat{\text{AS}_{\text{DR}}}$



(a) Blur-JPEG (b) Blur-Noise (c) JPEG-JPEG

(d) Noise-JPEG (e) Noise-JP2K

Fig. 11. Scatter plots of coefficient m versus $\mathrm{AS_{DR}}$ for the entire DR IQA databases V1 and V2 for the five distortion combinations under consideration.

score is then used in Model 1 (Eq. 9) which becomes:

$$\widetilde{\mathrm{AS_{FD}}} = P_1 \cdot \widehat{\mathrm{AS_{DR}}} \cdot \mathrm{RS_{FD}} + P_2 \cdot \mathrm{RS_{FD}} + (1 - P_1) \cdot \widehat{\mathrm{AS_{DR}}} - P_2. \quad (11)$$

Altogether, with five $\mathrm{AS_{DR}}/\widehat{\mathrm{AS_{DR}}}$ and $\mathrm{RS_{FD}}$ combinations (Scenario 1: FSIMc-FSIMc; Scenario 2: CORNIA-FSIMc; Scenario 2: dipIQ-FSIMc; Scenario 2: NIQE-FSIMc; Scenario 2: NIQE-MSSSIM) and seven multiple distortion combinations (Blur-JPEG; Blur-Noise; JPEG-JPEG; Noise-JPEG; Noise-JP2K; NBJ-JPEG; All data), we develop 35 sets of parameter settings for Model 1.

B. Model 2: Distortion Behavior Model

Motivated by the simplicity of the distortion behavior analysis based 2-tier Model 1 with only two parameters, and also to better account for the non-linear behavior of certain distortion combinations such as Noise-JPEG and Noise-JP2K, we adopt a direct six-parameter polynomial model with quadratic terms as Model 2 for Scenario 1 DR IQA:

$$\widehat{\mathrm{AS_{FD}}} = a \cdot {\mathrm{AS_{DR}}}^2 + b \cdot {\mathrm{RS_{FD}}}^2 + c \cdot \mathrm{AS_{DR}} + d \cdot \mathrm{RS_{FD}} + e \cdot \mathrm{AS_{DR}} \cdot \mathrm{RS_{FD}} + f, \quad (12)$$

where a, b, c, d, e and f are model coefficients, which are estimated directly using DR IQA databases V1 and V2. Similar to Model 1, the six model coefficients of Model 2 are separately determined for the five multiple distortion combinations as well as for the more generally applicable NBJ-JPEG and *All Data* cases. Model 2 reduces to Model 1 when: $a = 0$, $b = 0$, $c = (1 - P_1)$, $d = P_2$, $e = P_1$ and $f = -P_2$. Analogous to Model 1, for the case of Scenario 2 DR IQA, Model 2 becomes:

$$\widehat{\mathrm{AS_{FD}}} = a \cdot \widehat{\mathrm{AS_{DR}}}^2 + b \cdot {\mathrm{RS_{FD}}}^2 + c \cdot \widehat{\mathrm{AS_{DR}}} + d \cdot \mathrm{RS_{FD}} + e \cdot \widehat{\mathrm{AS_{DR}}} \cdot \mathrm{RS_{FD}} + f. \quad (13)$$

Specifically, the NR (CORNIA, dipIQ, NIQE) predicted DR image quality scores are mapped to respective FR (FSIMc or MSSSIM) scores using the nonlinear mapping function of Eq. 10. Similarly, with five $\mathrm{AS_{DR}}/\widehat{\mathrm{AS_{DR}}}$ and $\mathrm{RS_{FD}}$ combinations, and seven multiple distortion combinations, 35 sets of parameter settings are developed for Model 2.

C. Model 3: Support Vector Regression Model

To better understand how well the distortion behavior based Models 1 and 2 (which use very few parameters) capture the nature of the DR IQA problem, we opt to use support vector regression (SVR) [54], [55] to construct Model 3, which serves as an additional reference point, and also act as DR IQA models in their own right.

Specifically, we develop Model 3 by using nu-SVR that employs the radial basis function (RBF) kernel [55], [56] and four control parameters. For each of the 35 settings, the predictors are the FR FSIMc/MSSSIM $\mathrm{RS_{FD}}$ scores and either the FR FSIMc $\mathrm{AS_{DR}}$ scores or the NR CORNIA/dipIQ/NIQE $\widehat{\mathrm{AS_{DR}}}$ scores. The training targets are the $\mathrm{AS_{FD}}$ scores given by the SQB of the FD images. We use DR IQA database V1 for model training and DR IQA database V2 for model validation. The finalized models are later tested on a separate set of datasets (Section V). Before training, we ensure that the data has been scaled properly as recommended in [56]. During training, we determine the best possible SVR control parameters for a particular model through an extensive grid search by training the model on DR IQA database V1 hundreds

TABLE V

ABSOLUTE PERFORMANCE OF FR METHODS WHEN DETERMINING THE QUALITY OF FD IMAGES WITH RESPECT TO PR IMAGES IN TERMS OF PLCC. THE BEST SCORES FOR EACH DATABASE AND DISTORTION COMBINATION ARE HIGHLIGHTED IN BOLD

Database	FR Method	Distortion Combination						
		B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Waterloo Exp-II[a]	FSIMc	0.9153	0.8990	0.9157	0.8932	**0.9077**	0.9110	**0.9094**
	MSSSIM	**0.9363**	**0.9804**	**0.9470**	**0.8980**	0.8989	**0.9178**	0.9043
LIVE MD[b]	FSIMc	**0.7563**	**0.7884**	–	–	–	–	**0.7690**
	MSSSIM	0.7074	0.7738	–	–	–	–	0.6990
MDIVL[b]	FSIMc	**0.8909**	–	–	**0.9193**	–	–	**0.8874**
	MSSSIM	0.8370	–	–	0.8996	–	–	0.8645

[a]PLCC is computed with respect to SQB. [b]PLCC is computed with respect to MOS/DMOS.
Note: SRCC test results were found to be similar to PLCC, and hence they are not shown due to space limit.

to thousands of times using different combinations of control parameters, and then selecting the parameters that lead to the best model performance in terms of both the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-order Correlation Coefficient (SRCC) on the validation data (DR IQA database V2). Since model training by using a large grid is quite time consuming, we use a two-tier grid search. First a coarse-level grid search is performed that identifies the region of the grid that should be focused on. This is followed by a fine-level grid search to finalize the SVR parameters which are used to train the final model on DR IQA database V1.

V. PERFORMANCE EVALUATION

A. Databases and Evaluation Criteria

To the best of our knowledge, there are only four datasets [27], [36], [37], [39] that provide both singly distorted DR and their respective multiply distorted FD images, together with quality labels. Although two other datasets, MDID [38] and MDID2013 [14], contain multiply distorted images, they do not provide DR images. Therefore, we use these four datasets, as discussed below, for performance evaluation. These datasets do not have any content overlap with DR IQA databases V1 and V2 used in the development of our DR IQA models.

The Waterloo Exploration-II (Waterloo Exp-II) database [39] has 3,570 PR images, 39,270 singly distorted images each for Blur, JPEG compression, and Noise, and 667,590 multiply distorted images each for the distortion combinations of Blur-JPEG, Blur-Noise, JPEG-JPEG, Noise-JPEG, and Noise-JP2K. The singly and multiply distorted images are the DR and FD images, respectively, in a 2-stage distortion process. All distorted images are annotated with synthetic quality benchmark (SQB) labels that have been generated by fusing the results from four state-of-the-art FR methods [39].

The LIVE Multiply Distorted (LIVE MD) database [36] consists of 15 PR images, 45 singly distorted images each for Blur, JPEG compression and Noise, and 135 multiply distorted images each for the distortion combinations of Blur-JPEG and Blur-Noise. We consider the singly distorted Blur images as DR and the multiply distorted images as the FD images. Subjective ratings are available in the form of difference mean opinion scores (DMOS).

The Multiply Distorted IVL (MDIVL) database [37], [57] consists of 10 PR and 750 multiply distorted images of which 350 belong to the Blur-JPEG combination while 400 belong to the Noise-JPEG combination. Although the database does not explicitly contain singly distorted images, in both Blur-JPEG and Noise-JPEG combinations, the least compression distortion level utilizes MATLAB quality factor of 100, which produces nearly perceptually lossless compression. Thus, we regard 70 out of 350 Blur-JPEG and 100 out of 400 Noise-JPEG images as singly distorted Blur and Noise images, respectively, thereby providing us with DR and FD images. MDIVL provides subjective ratings in the form of mean opinion scores (MOS).

The LIVE Wild Compressed (LIVE WCmp) database [26], [27] is composed of 400 images. It starts with 80 authentically distorted images taken from the LIVE Wild Challenge database [58] which can be regarded as DR images. Each of these 80 images are further JPEG compressed at four fixed compression levels regardless of content, leading to 320 FD images. LIVE WCmp does not have PR images and provides subjective ratings in the form of MOS.

Among the above-mentioned datasets, Waterloo Exp-II [39] is synthetically annotated while the other three (LIVE MD [36], MDIVL [37], and LIVE WCmp [27]) contain subjective ratings. On the other hand, three datasets (Waterloo Exp-II [39], LIVE MD [36], and MDIVL [37]) are simulated distortion datasets, whereas LIVE WCmp [27] has authentically distorted DR images. It is pertinent to mention that while the development of DR IQA models in this work (Section IV) was done by using simulated distortion datasets (DR IQA databases V1 and V2), testing these models on the LIVE WCmp [27] database allows us to evaluate their performance for the real world scenario of the storage of authentically distorted user generated content through compression.

We use PLCC and SRCC as measures of a model's prediction accuracy and prediction monotonicity, respectively [59]. PLCC is computed after a nonlinear mapping step between model predictions and target scores, whereas SRCC is computed directly [13]. Due to space limit, only PLCC results are reported here but SRCC results are found to be similar in all test cases.

B. Absolute FR Performance and Baseline IQA Models

With access to the pristine reference images, FR IQA methods offer the best quality prediction performance [13], and thus serve as an approximate upper bound for the baseline and DR IQA models that we will test later. Here we select FSIMc [30] and MSSSIM [29] as the reference FR models. The former outperforms most other FR methods [13], and the

TABLE VI

PERFORMANCE OF BASELINE METHODS IN TERMS OF PLCC. FOR BASELINES 1 AND 2 THE BEST SCORES FOR EACH DATABASE AND DISTORTION COMBINATION ARE HIGHLIGHTED IN BOLD

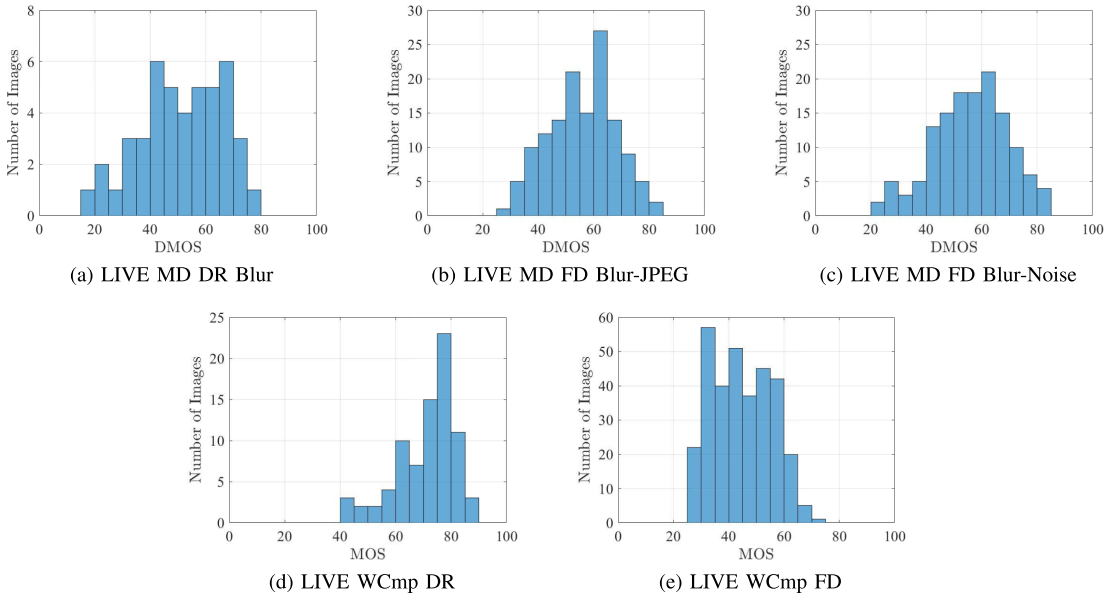
Baseline Type	Database	Method	Distortion Combination						
			B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Baseline 1 FR Methods	Waterloo Exp-II[a]	FSIMc	0.8436	0.8826	0.8276	0.8280	**0.8223**	**0.7980**	**0.7926**
		MSSSIM	**0.8567**	**0.9473**	**0.8340**	**0.8809**	0.8039	0.7498	0.7425
	LIVE MD[b]	FSIMc	0.2256	0.3882	–	–	–	–	**0.3045**
		MSSSIM	**0.2366**	**0.4270**	–	–	–	–	0.2254
	MDIVL[b]	FSIMc	**0.5207**	–	–	0.8111	–	–	**0.6238**
		MSSSIM	0.4984	–	–	**0.8770**	–	–	0.5985
	LIVE WCmp[b,c]	FSIMc	–	–	–	–	–	–	**0.9030**
		MSSSIM	–	–	–	–	–	–	0.8498
Baseline 2 NR Methods	Waterloo Exp-II[a]	CORNIA	**0.8918**	0.6205	0.7512	0.7832	0.6943	0.8172	0.7553
		dipIQ	0.8522	**0.9414**	**0.8790**	**0.8462**	**0.8380**	**0.8422**	**0.8532**
		NIQE	0.7741	0.8941	0.7084	0.6368	0.6913	0.7030	0.7137
	LIVE MD[b]	CORNIA	0.7141	**0.8144**	–	–	–	–	**0.7360**
		dipIQ	0.5238	0.6603	–	–	–	–	0.5531
		NIQE	**0.7677**	0.6670	–	–	–	–	0.5802
	MDIVL[b]	CORNIA	**0.9331**	–	–	0.7748	–	–	**0.7963**
		dipIQ	0.8298	–	–	**0.8074**	–	–	0.7514
		NIQE	0.7910	–	–	0.5357	–	–	0.5731
	LIVE WCmp[b,c]	CORNIA	–	–	–	–	–	–	**0.8424**
		dipIQ	–	–	–	–	–	–	0.7978
		NIQE	–	–	–	–	–	–	0.8314
Baseline 3 2stepQA	Waterloo Exp-II[a]	2stepQA	0.9340	0.9696	0.8951	0.8420	0.7213	0.7709	0.7140
	LIVE MD[b]		0.7746	0.6730	–	–	–	–	0.6500
	MDIVL[b]		0.8697	–	–	0.7964	–	–	0.8149
	LIVE WCmp[b,c]		–	–	–	–	–	–	0.9229

[a]PLCC is computed with respect to SQB. [b]PLCC is computed with respect to MOS/DMOS.
[c]LIVE WCmp has authentic distortions followed by JPEG compression. It cannot be placed in a specific distortion combination.
Note: SRCC test results were found to be similar to PLCC, and hence they are not shown due to space limit.

latter is the FR component in 2stepQA [27] besides being a competitive method [13]. Table V shows the performance of the FR methods in terms of PLCC, where the results are termed as absolute performance since testing is done against PR images. The LIVE WCmp database [27] is not present because it does not have PR images.

Given the two-stage DR IQA framework, three types of approaches may be applied using existing IQA models in the literature, and are included as baseline models. 1) Baseline-1: Use FR methods to assess the relative quality of FD images with respect to DR images, i.e., RS_{FD}, and use it to predict the absolute quality of FD images. Specifically, we use FSIMc [30] and MSSSIM [29] as the representative FR measures; 2) Baseline-2: Use NR methods to assess the FD images directly, without referencing to the DR images. Based on the performance analysis of 14 NR methods on nine datasets in [13] and another six recent NR methods [60], [61], [62], [63], [64], [65] on the LIVE MD and/or MDIVL datasets, we select CORNIA [34] and dipIQ [53] as representative opinion-aware and opinion-unaware baseline NR methods, respectively, due to their good performance. We also select NIQE [32] as it is the NR component in 2stepQA [27]; 3) Baseline-3: 2stepQA [26], [27] is the only IQA model that utilizes the quality information about the DR image while determining the quality of a multiply distorted FD image. Specifically, NIQE [32] is used for NR assessment of the DR image, MSSSIM [29] is used to compare the DR and FD images, and a product of the scores produces a final quality assessment of the FD image.

Table VI provides the quality prediction performance of the three baseline approaches in terms of PLCC. There are several important observations. First, close comparison with Table V shows that there is generally a large gap in performance between the absolute FR quality scores and the Baseline-1 relative FR quality scores, even though the same top-performing FR models are employed in both cases, suggesting that FR IQA models are reliable only when the pristine quality reference is accessible. Table VI also shows that Baseline-1 performance varies drastically between databases, where it is quite acceptable in case of LIVE WCmp but poor for LIVE MD. This is explained by the amount of distortion(s) present in the DR images and if there is perceptual separation between the quality ranges of the DR and FD images in a dataset. To elaborate on this further, we provide the histograms of the subjective ratings for the DR and FD images in LIVE MD and LIVE WCmp databases in Fig. 12. It should be noted that LIVE MD [36] provides subjective ratings in the form of DMOS for which a lower value indicates better visual quality, whereas LIVE WCmp [27] provides these ratings as MOS for which a higher value indicates better visual quality. The histogram of the DR images in LIVE MD, shown in Fig. 12(a), shows that most of its DR images are quite distorted and when compared with the histograms of the FD images in LIVE MD, shown in Fig. 12(b,c), it becomes clear that not only do the DR and FD images share a similar perceptual quality range, but they are also similarly distributed in terms of quality. Since FR IQA methods are essentially image fidelity measures and require the reference image to be of pristine quality, the nature of DR and FD images in LIVE MD render them completely ineffective as Baseline-1 models. On the other hand, Fig. 12(d) shows that most of the DR images in the LIVE WCmp database have relatively better quality and when compared with the histogram of the FD images in this dataset, shown in Fig. 12(e), it is evident that the DR and FD images in LIVE WCmp are differently distributed in terms of quality. This shows that as the quality of DR images improves, the performance of FR IQA models, used as Baseline-1, improves as well. Overall, this discussion



(a) LIVE MD DR Blur (b) LIVE MD FD Blur-JPEG (c) LIVE MD FD Blur-Noise

(d) LIVE WCmp DR (e) LIVE WCmp FD

Fig. 12. Histograms of MOS/DMOS for the DR and FD images in LIVE MD [36] (a-c) and LIVE WCmp [27] (d, e) databases.

demonstrates the ineffectiveness of FR IQA when dealing with degraded references on their own.

Second, the Baseline-2 NR models perform highly inconsistently and largely depend on the test dataset, the distortion types, and the NR model being used. Since most NR models are developed, trained and/or validated with singly distorted images, there is generally a large performance drop when they evaluate multiply distorted images. Furthermore, as shown in [13], NR methods SISBLIM [14] and GWHGLBP [18], that are designed for multiply distorted content, are unable to outperform CORNIA [34] even on multiply distorted datasets, and their performance drops further when testing also incorporates singly distorted content. Such inconsistent performance of the NR IQA paradigm, owing to the difficult nature of the problem and its inability to use auxiliary information about a distorted image even if it is available, is a strong motivation for the development of the new DR IQA paradigm that is able to utilize additional information provided by DR images.

Third, the Baseline-3 2stepQA model may sometimes significantly improve upon the first two Baseline models, but the performance gain varies drastically, and is mostly limited to certain types of distortion combinations such as B-JPG and JPG-JPG. Overall, all three baseline models exhibit significant performance gaps against the absolute FR IQA scores shown in Table V, suggesting the potential space for improvement by deeper investigation on DR IQA.

C. Performance of DR IQA Models

Table VII provides the performance of DR IQA Models 1, 2, and 3 in terms of PLCC, where the test datasets have no content overlap with DR IQA databases V1 and V2 used for model development.

1) Comparison With Baseline Models: A comparison of Table VII with Table VI, shows that DR IQA Models 1, 2, and 3 outperform the FR based Baseline-1 approach significantly and nearly comprehensively. Their superior performance relative to Baseline-1 demonstrates the shortcomings of the FR paradigm in the absence of PR images at the final destination and establishes the value of the DR IQA framework.

Comparing Table VII with Table V, we find that the DR IQA models perform better than or at par in most cases against FR computed AS_{FD} scores. This is no small achievement given that FR performance is usually considered as an upper bound in IQA when the PR images are accessible. There are a few exceptions on the N-JP2, NBJ-JPG and All data cases. This highlights the difficult nature of the N-JP2 case, as can be seen in the distortion behavior plot of Fig. 9(e). It also highlights the difficult nature of the NBJ-JPG and All data cases, where multiple distortion combinations are considered together.

Comparing Table VII with Table VI shows that DR IQA Models nearly comprehensively outperform the NR based Baseline-2 approach on all test datasets. A few NR models perform exceptionally well on certain test cases (e.g., CORNIA [34] on B-N and B-JPG cases of LIVE MD and MDIVL, respectively, and dipIQ [53] on Waterloo Exp-II), but their performance drops drastically on other cases. These results suggest the benefit of incorporating the additional information in the DR images, and again demonstrate the value of the DR IQA paradigm. The superior performance of Scenario 2 or Type-001100 DR IQA, compared to Baseline-2, also shows that in the absence of PR images, NR methods can be effectively used to compute $\widehat{AS}_{DR}$ scores for DR images, which together with the FR computed RS_{FD} scores between the DR and FD images, can lead to effective DR IQA models, again highlighting the value of using additional information provided by DR images even if it is through their NR predicted quality.

The 2stepQA-based Baseline-3 approach is most relevant as an early Type-001100 DR IQA instantiation. Since 2stepQA [27] combines NIQE and MSSSIM, a direct comparison can be made with the NIQE-MSSSIM DR IQA models by comparing Tables VI and VII, where it can be

TABLE VII

PERFORMANCE OF DR IQA MODELS IN TERMS OF PLCC. FOR MODELS 1, 2, AND 3 THE BEST SCORES FOR EACH DATABASE AND DISTORTION COMBINATION ARE HIGHLIGHTED IN BOLD

DR IQA Model	Database	Predictors		Distortion Combination and Model Type						
		$AS_{DR}/\widehat{AS_{DR}}$	RS_{FD}	B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Model 1 (Distortion Behavior Based)	Waterloo Exp-II[a]	FSIMc	FSIMc	0.9126	0.8991	0.9201	0.8568	0.8147	0.8538	0.8217
		CORNIA	FSIMc	0.9085	0.9119	0.9088	0.8559	**0.8291**	0.8540	0.8264
		dipIQ	FSIMc	0.9079	0.9114	**0.9219**	**0.8922**	0.7998	**0.8653**	**0.8335**
		NIQE	FSIMc	0.8917	0.9041	0.8860	0.8626	0.8041	0.8367	0.8142
		NIQE	MSSSIM	**0.9341**	**0.9682**	0.9165	0.7815	0.5839	0.7754	0.7628
	LIVE MD[b]	FSIMc	FSIMc	0.7576	0.7911	–	–	–	–	0.7662
		CORNIA	FSIMc	0.7786	0.7616	–	–	–	–	0.7635
		dipIQ	FSIMc	0.7797	**0.7932**	–	–	–	–	**0.7797**
		NIQE	FSIMc	**0.7827**	0.7677	–	–	–	–	0.7637
		NIQE	MSSSIM	0.7744	0.6818	–	–	–	–	0.6705
	MDIVL[b,c]	FSIMc	FSIMc	0.8964	–	–	**0.9179**	–	**0.9001**	**0.8997**
		CORNIA	FSIMc	**0.9192**	–	–	0.8551	–	0.8858	0.8859
		dipIQ	FSIMc	0.9167	–	–	0.8928	–	0.8921	0.8921
		NIQE	FSIMc	0.8605	–	–	0.7856	–	0.8234	0.8223
		NIQE	MSSSIM	0.8607	–	–	0.7430	–	0.8091	0.8075
	LIVE WCmp[b,d]	CORNIA	FSIMc	0.9100	0.9096	0.9084	0.9055	0.8971	0.9095	0.9094
		dipIQ	FSIMc	0.9081	0.9080	0.9080	0.9076	0.9071	0.9080	0.9079
		NIQE	FSIMc	**0.9280**	**0.9271**	**0.9271**	**0.9218**	**0.9139**	**0.9271**	**0.9258**
		NIQE	MSSSIM	0.9264	0.9261	0.9259	0.9175	0.9134	0.9156	0.9133
Model 2 (Distortion Behavior Based)	Waterloo Exp-II[a]	FSIMc	FSIMc	0.9135	0.9003	0.9206	0.8751	**0.8857**	0.8567	0.8296
		CORNIA	FSIMc	0.9090	0.9117	0.9085	0.8654	0.8432	0.8550	0.8288
		dipIQ	FSIMc	0.9078	0.9116	**0.9228**	**0.9075**	0.8751	**0.8685**	**0.8416**
		NIQE	FSIMc	0.8911	0.9042	0.8854	0.8723	0.8606	0.8414	0.8248
		NIQE	MSSSIM	**0.9336**	**0.9686**	0.9182	0.8726	0.8490	0.8132	0.7980
	LIVE MD[b]	FSIMc	FSIMc	0.7575	0.7911	–	–	–	–	0.7628
		CORNIA	FSIMc	0.7745	0.7679	–	–	–	–	**0.7901**
		dipIQ	FSIMc	0.7797	**0.7963**	–	–	–	–	0.7772
		NIQE	FSIMc	**0.7825**	0.7736	–	–	–	–	0.7615
		NIQE	MSSSIM	0.7071	0.7089	–	–	–	–	0.6347
	MDIVL[b,c]	FSIMc	FSIMc	0.8970	–	–	**0.9221**	–	**0.9008**	**0.8990**
		CORNIA	FSIMc	0.9147	–	–	0.8793	–	0.8953	0.8958
		dipIQ	FSIMc	**0.9164**	–	–	0.9076	–	0.8906	0.8863
		NIQE	FSIMc	0.8586	–	–	0.7988	–	0.8202	0.8167
		NIQE	MSSSIM	0.8572	–	–	0.7633	–	0.7874	0.7550
	LIVE WCmp[b,d]	CORNIA	FSIMc	0.9084	0.9109	0.9093	0.9122	0.9035	0.9141	0.9140
		dipIQ	FSIMc	0.9081	0.9079	0.9081	0.9058	0.9058	0.9079	0.9077
		NIQE	FSIMc	**0.9277**	**0.9271**	**0.9265**	**0.9201**	**0.9100**	**0.9255**	**0.9233**
		NIQE	MSSSIM	0.9255	0.9258	0.9242	0.9067	0.8886	0.9187	0.9134
Model 3 (SVR Based)	Waterloo Exp-II[a]	FSIMc	FSIMc	0.9287	0.9104	0.9195	0.8877	**0.9074**	0.8629	0.8416
		CORNIA	FSIMc	0.9215	0.9180	0.9062	0.8547	0.8449	0.8643	0.8389
		dipIQ	FSIMc	0.9228	0.9195	**0.9224**	0.9112	0.8825	**0.8690**	**0.8448**
		NIQE	FSIMc	0.9017	0.9089	0.8853	0.8809	0.8660	0.8422	0.8317
		NIQE	MSSSIM	**0.9383**	**0.9671**	0.9159	**0.9327**	0.8746	0.8172	0.7952
	LIVE MD[b]	FSIM	FSIMc	0.7539	0.8082	–	–	–	–	0.7329
		CORNIA	FSIMc	**0.7769**	**0.8175**	–	–	–	–	0.7032
		dipIQ	FSIMc	0.7371	0.7791	–	–	–	–	**0.7839**
		NIQE	FSIMc	0.7712	0.7641	–	–	–	–	0.7602
		NIQE	MSSSIM	0.7349	0.7270	–	–	–	–	0.6468
	MDIVL[b,c]	FSIMc	FSIMc	0.8975	–	–	**0.9227**	–	0.9063	**0.9048**
		CORNIA	FSIMc	**0.9397**	–	–	0.8767	–	**0.9085**	0.9001
		dipIQ	FSIMc	0.9203	–	–	0.9052	–	0.8999	0.8950
		NIQE	FSIMc	0.8578	–	–	0.8182	–	0.8296	0.8046
		NIQE	MSSSIM	0.8671	–	–	0.8161	–	0.7737	0.7799
	LIVE WCmp[b,d]	CORNIA	FSIMc	0.9117	0.9156	0.9166	0.9087	0.9023	0.9166	0.9133
		dipIQ	FSIMc	0.9086	0.9083	0.9065	0.9010	0.9010	0.9067	0.9069
		NIQE	FSIMc	0.9239	**0.9237**	**0.9207**	**0.9169**	**0.9056**	**0.9176**	**0.9202**
		NIQE	MSSSIM	**0.9247**	0.9217	0.9188	0.8609	0.8514	0.9152	0.9131

[a] PLCC is computed with respect to SQB. [b] PLCC is computed with respect to MOS/DMOS.
[c] The NBJ-JPG and All Data model versions are applied to the entire MDIVL database.
[d] LIVE WCmp has authentic distortions followed by JPEG compression. It cannot be placed in a specific distortion combination. DR IQA models, developed for the seven distortion combinations, are applied to the entire dataset and results have been reported in respective columns.
Note: SRCC test results were found to be similar to PLCC, and hence they are not shown due to space limit.

seen that 2stepQA performs better than the NIQE-MSSSIM DR IQA models in more than half of the test cases. However, when other base FR/NR-FR combinations (FSIMc-FSIMc, CORNIA-FSIMc, dipIQ-FSIMc, and NIQE-FSIMc), are adopted, DR IQA Models 1, 2, and 3, outperform 2stepQA nearly comprehensively (with the exception of B-JPG and B-N), and the gaps are large in the difficult N-JPG (except Model 1 on Waterloo Exp-II), N-JP2 (except Model 1 on Waterloo Exp-II), NBJ-JPG, and the most difficult All data cases of Waterloo Exp-II, LIVE MD, and MDIVL databases. For the B-N combination, DR IQA models perform better than 2stepQA on LIVE MD while the reverse is true for the Waterloo Exp-II database. For the B-JPG combination, 2stepQA mostly performs better than the DR IQA models. It can also be seen that both the 2stepQA and DR IQA models offer similar performance on the LIVE WCmp database, where for the majority of test cases 2stepQA is slightly better while DR IQA models are slightly better for a few cases. However, almost all DR IQA models have a PLCC above 0.9 suggesting competitive performance (see Section V-C.3

for more discussion). Overall, these results suggest that the selection of base NR and FR models, and the method of combination (i.e., considering the behavior of multiple simultaneous distortions instead of simple product) are both important in yielding superior performance.

It is also worth noting that the DR IQA Models 1, 2, and 3, are developed by using DR IQA databases V1 and V2, where all training images are annotated by SQB scores [39] rather than subjective ratings. Thus, the superior performance of these models on subject-rated test datasets (LIVE MD [36], MDIVL [37], and LIVE WCmp [27]) has in turn provided another strong demonstration of the value of using SQB [39] as an alternative IQA data annotation mechanism.

2) Inter-Model Comparisons: We perform three kinds of inter-model comparisons. First, the complexity and number of parameters increase from the proposed Models 1, 2 to 3, thus presumably, one would expect their performance to improve correspondingly. Somewhat surprisingly, this is not necessarily always the case. Close observation of Table VII concludes that in a majority of cases, these models offer similar performance. In particular, Model 1, which is constructed empirically from distortion behavior analysis and uses only 2 parameters, often produces similar performance when compared with the 6-parameter Model 2 and the much more sophisticated SVR-based Model 3. This reveals the value of distortion behavior analysis as discussed in Sections III-B and IV. Also note that the performance of Model 1 is not as competitive as the other two models in the N-JPG and N-JP2 combinations. This is explained by the construction of Eq. 7 which makes it difficult to adequately capture the complex distortion behavior for these cases as depicted in Figs. 9(d) and (e).

Second, we compare across DR IQA architectures. Of the five $AS_{DR}/\widehat{AS_{DR}}$ and RS_{FD} combinations, the first (FSIMc-FSIMc) belongs to Scenario 1 or Type-100100 DR IQA (Fig. 4), while the rest (CORNIA-FSIMc, dipIQ-FSIMc, NIQE-FSIMc, and NIQE-MSSSIM) belong to the more practical Scenario 2 or Type-001100 DR IQA (Fig. 5). While FR methods are used to compute RS_{FD} in both Scenarios 1 and 2, the major difference is that Scenario 1 considers the PR image to be available and uses FR methods to compute AS_{DR} whereas Scenario 2 considers such images to be unavailable and uses NR methods to compute $\widehat{AS_{DR}}$. Since FR methods generally outperform NR ones, it is natural to expect that the Scenario 1 method should outperform Scenario 2 methods. Interestingly, this is not always the case in Table VII. To further investigate this rather counter-intuitive observation, in Table VIII we evaluate the performance of the FR and NR methods used to predict the quality of DR images, i.e., compute AS_{DR} and $\widehat{AS_{DR}}$ scores, respectively, in terms of PLCC and SRCC on subject-rated IQA databases that provide PR images (i.e., for LIVE MD Blur DR images, and MDIVL Blur and Noise DR images). A comparison of Tables VII and VIII shows: 1) For LIVE MD Blur images, the FR FSIMc is outperformed by the NR methods (with the exception of NIQE in terms of PLCC) and correspondingly Scenario 1 is outperformed by Scenario 2 on both the Blur-JPEG and Blur-Noise combinations of LIVE MD. 2) Similarly for MDIVL Blur images, the FR FSIMc is outperformed by the NR methods (with the exception of NIQE) and correspondingly

TABLE VIII

IQA METHOD PERFORMANCE TO PREDICT THE QUALITY OF DR IMAGES IN THE LIVE MD (BLUR) AND MDIVL (BLUR & NOISE) DATABASES

Method	Type	LIVE MD Blur		MDIVL Blur		MDIVL Noise	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
FSIMc	FR	0.8108	0.7595	0.9502	0.9363	0.9358	0.9317
CORNIA	NR	0.8452	0.8110	0.9586	0.9550	0.8128	0.8095
dipIQ	NR	0.8350	0.8002	0.9660	0.9384	0.9317	0.9267
NIQE	NR	0.8027	0.7885	0.8465	0.8193	0.7283	0.7200

Scenario 1 is outperformed by CORNIA-FSIMc and dipIQ-FSIMc based Scenario 2 on the Blur-JPEG combination of MDIVL. 3) For MDIVL Noise images, the FR FSIMc outperforms the NR methods and correspondingly Scenario 1 outperforms Scenario 2 on the Noise-JPEG combination of MDIVL. This analysis shows that the performance of a DR IQA model is directly impacted by the performance of the IQA method evaluating the quality of the DR images. It also suggests that the lack of access to the PR image may not always be critical to DR IQA, as long as appropriate NR and FR methods are used to compute $\widehat{AS_{DR}}$ and RS_{FD}, respectively.

Third, we compare across distortion combinations. Table VII reports the performance of DR IQA models for seven multiple distortion combinations (Blur-JPEG, Blur-Noise, JPEG-JPEG, Noise-JPEG, Noise-JP2K, NBJ-JPEG, and All data). All models perform quite consistently across the LIVE MD [36], MDIVL [37], and LIVE WCmp [27] databases, but they do not have all seven combinations. Thus, we focus on the Waterloo Exp-II database [39]. It appears that the DR IQA models perform quite well for Blur-JPEG, Blur-Noise, and JPEG-JPEG cases, and reasonably well (with few exceptions) for Noise-JPEG, Noise-JP2K, NBJ-JPEG, and All data cases when considered independently, but not at the same level of the first three cases. This is understandable given the variations in distortion behaviors demonstrated in Fig. 9. This makes DR IQA an interesting and challenging problem that demands future investigations, especially for the challenging distortion combinations and the All data case.

3) Performance on Authentically Distorted Content: To the best of our knowledge, LIVE WCmp [27] is the only subject-rated IQA dataset that uses authentically distorted DR images to create FD images and provides both. A comparison of Tables VI and VII shows that on LIVE WCmp, the DR IQA models outperform both the FR-based Baseline 1 and NR-based Baseline 2, while their performance is similar to the 2stepQA-based Baseline 3. This indicates that even when developed by using simulated distorted content, DR IQA has the potential to perform adequately well when using authentically distorted degraded references. However, it should be pointed out that LIVE WCmp [27] is a small dataset with only 80 authentically distorted DR and 320 FD images. Furthermore, to create FD images, LIVE WCmp [27] uses fixed JPEG compression parameters of 18, 12, 6 and 3, where the latter two lead to excessive compression which is uncommon in practice. The MOS histogram of LIVE WCmp FD images, shown in Fig. 12(e) indicates that it has relatively more low quality images than high quality ones. In the future, it is desirable to develop new subject-rated IQA datasets that contain a more diverse set of authentically distorted DR/FD images with a wider coverage of the quality spectrum. Such

datasets will be helpful in developing new DR IQA models targeted towards authentically distorted content.

VI. Conclusion

We make one of the first attempts to establish a DR IQA paradigm, targeting at the problem when only a reference image of degraded quality is available when assessing the quality of a multiply distorted image, a problem that is of practical importance in many real-world applications such as image/video distributions. We lay out possible architectures of DR IQA in two-stage distortion pipelines, introduce a 6-bit code to denote various configurations, and focus on two specific architectures or scenarios. We establish first-of-their-kind large-scale synthetically annotated databases dedicated to DR IQA, and conduct a novel multiple distortion behavior analysis for two-stage distortion pipelines. We also develop novel DR IQA models and make extensive comparisons with different types of baseline models. The results suggest that DR IQA may offer significant performance gain in multiple distortion environments against existing FR and NR IQA paradigms, thereby establishing DR IQA as a novel IQA paradigm in its own right. We hope the current work can inspire significant future research that explores different DR IQA architectures, distortion combinations, and design philosophies. Extension of the general DR IQA architectures framework introduced in this work (Fig. 3) to more than two distortion stages should be straightforward but the development and verification of associated DR IQA models is grounds for significant future work. As with other IQA paradigms, DR IQA can be extended to undertake DR video quality assessment (VQA). In fact, pioneering work in this direction has recently emerged [66], [67], [68], which opens up a new venue for future research.

References

[1] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synth. Lect. Image, Video Multimedia Process.*, vol. 2, no. 1, pp. 1–156, Dec. 2006.

[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] *Youtube Help: Recommended Upload Encoding Settings*. Accessed: Jan. 7, 2023. [Online]. Available: https://support.google.com/youtube/answer/1722171?hl=en

[4] *Vimeo Help Center: Video and Audio Compression Guidelines*. Accessed: Jan. 7, 2023. [Online]. Available: https://help.vimeo.com/hc/en-us/articles/360056550451

[5] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data," *Mag. Reson. Med.*, vol. 34, no. 6, pp. 910–914, 1995.

[6] H. Lu, X. Li, I.-T. Hsiao, and Z. Liang, "Analytical noise treatment for low-dose CT projection data by penalized weighted least-square smoothing in the K-L domain," in *Proc. SPIE*, vol. 4682, San Diego, CA, USA, May 2002, pp. 146–152.

[7] P. Coupé, P. Hellier, C. Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," *IEEE Trans. Image Process.*, vol. 18, no. 10, pp. 2221–2229, Oct. 2009.

[8] D. A. Koff and H. Shulman, "An overview of digital compression of medical images: Can we use lossy image compression in radiology?" *Can. Assoc. Radiol. J.*, vol. 57, no. 4, pp. 211–217, Oct. 2006.

[9] D. Koff et al., "Pan-Canadian evaluation of irreversible compression ratios ('lossy' compression) for development of national guidelines," *J. Digit. Imag.*, vol. 22, no. 6, pp. 569–578, Dec. 2009.

[10] R. L. White, "High-performance compression of astronomical images," in *Proc. NASA. Goddard Space Flight Center, Space Earth Sci. Data Compress. Workshop*, Jan. 1993, pp. 117–123.

[11] N. W. Lewis and J. W. Allnatt, "Subjective quality of television pictures with multiple impairments," *IET Electron. Lett.*, vol. 1, no. 7, pp. 187–188, Sep. 1965.

[12] D. M. Chandler, "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Process.*, vol. 2013, Jan. 2013, Art. no. 905685.

[13] S. Athar and Z. Wang, "A comprehensive performance evaluation of image quality assessment algorithms," *IEEE Access*, vol. 7, pp. 140030–140070, 2019.

[14] 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.

[15] Y. Lu, F. Xie, T. Liu, Z. Jiang, and D. Tao, "No reference quality assessment for multiply-distorted images based on an improved bag-of-words model," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1811–1815, Oct. 2015.

[16] C. Li, Y. Zhang, X. Wu, W. Fang, and L. Mao, "Blind multiply distorted image quality assessment using relevant perceptual features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 4883–4886.

[17] 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.

[18] 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.

[19] H. Hadizadeh and I. V. Bajic, "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.

[20] 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.

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

[22] W. Cheng and K. Hirakawa, "Corrupted reference image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Orlando, FL, USA, Sep. 2012, pp. 1485–1488.

[23] C. Zhang and K. Hirakawa, "Blind full reference quality assessment of Poisson image denoising," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 2719–2723.

[24] C. Zhang, W. Cheng, and K. Hirakawa, "Corrupted reference image quality assessment of denoised images," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1732–1747, Apr. 2019.

[25] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, and J. Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 10222–10231.

[26] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, "Predicting the quality of images compressed after distortion in two steps," in *Proc. SPIE*, vol. 10752, San Diego, CA, USA, Sep. 2018, Art. no. 107520K.

[27] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, "Predicting the quality of images compressed after distortion in two steps," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5757–5770, Dec. 2019.

[28] A. Bovik, "Assessing quality of images or videos using a two-stage quality assessment," U.S. Patent 10 529 066, Jan. 7, 2020.

[29] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. ACSSC*, Pacific Grove, CA, USA, Mar. 2003, pp. 1398–1402.

[30] 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.

[31] 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, Aug. 2014.

[32] 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.

[33] 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.

[34] 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.

[35] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *Proc. IEEE Int. Conf. Image Process.*, Athens, Greece, Oct. 2018, pp. 609–613.
[36] 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.
[37] 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.
[38] 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.
[39] S. Athar, Z. Wang, and Z. Wang, "Deep neural networks for blind image quality assessment: Addressing the data challenge," 2021, *arXiv:2109.12161*.
[40] 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, Mar. 2010, Art. no. 011006.
[41] P. Le Callet and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database," V1, 2005. Accessed: Jan. 7, 2023. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00580755 and http://www.irccyn.ec-nantes.fr/ivcdb/
[42] 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.
[43] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.
[44] Y. Horita, K. Shibata, and K. Yoshikazu. (2011). *MICT Image Quality Evaluation Database*. [Online]. Available: https://qualinet.github.io/databases/image/mict_image_quality_database/
[45] *The Consumer Digital Video Library (CDVL)*. Accessed: Jan. 7, 2023. [Online]. Available: http://www.cdvl.org/
[46] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Boston, MA, USA, 2009, pp. 758–759.
[47] 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.
[48] 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.
[49] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, "Image quality assessment based on DCT subband similarity," in *Proc. Int. Conf. Quality Multimedia Exp.*, Quebec City, QC, Canada, Sep. 2015, pp. 2105–2109.
[50] 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.
[51] 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.
[52] 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.
[53] 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.
[54] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4757-3264-1
[55] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, Apr. 2011.
[56] C. W. Hsu, C. C. Chang, and C. J. Lin. (May 2016). *A Practical Guide to Support Vector Classification*. [Online]. Available: https://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf
[57] S. Corchs, F. Gasparini, and R. Schettini, "Noisy images-JPEG compressed: Subjective and objective image quality evaluation," in *Proc. SPIE*, vol. 9016, San Francisco, CA, USA, Feb. 2014, Art. no. 90160V.
[58] 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.
[59] *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II*, document T01-SG09-C-0060, Video Quality Experts Group and others, 2003.
[60] Y. Liu et al., "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 929–943, Apr. 2020.
[61] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Trans. Image Process.*, vol. 29, pp. 7414–7426, 2020.
[62] Y. Liu, K. Gu, X. Li, and Y. Zhang, "Blind Image quality assessment by natural scene statistics and perceptual characteristics," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 3, p. 91, Aug. 2020.
[63] D. Li, T. Jiang, and M. Jiang, "Norm-in-Norm loss with faster convergence and better performance for image quality assessment," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 789–797.
[64] X. Yang, F. Li, and H. Liu, "TTL-IQA: Transitive transfer learning based no-reference image quality assessment," *IEEE Trans. Multimedia*, vol. 23, pp. 4326–4340, 2021.
[65] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE Trans. Image Process.*, vol. 31, pp. 4149–4161, 2022.
[66] Y. Li et al., "User-generated video quality assessment: A subjective and objective study," *IEEE Trans. Multimedia*, early access, Oct.29, 2021, doi: 10.1109/TMM.2021.3122347.
[67] X. Yu, N. Birkbeck, Y. Wang, C. G. Bampis, B. Adsumilli, and A. C. Bovik, "Predicting the quality of compressed videos with pre-existing distortions," *IEEE Trans. Image Process.*, vol. 30, pp. 7511–7526, 2021.
[68] Z. Wang et al., "Unified end-to-end quality and latency measurement, optimization and management in multimedia communications," WO Patent PCT/IB2020 051 973, Oct. 1, 2020.

Shahrukh Athar (Member, IEEE) received the B.E. degree in electrical engineering from the National University of Sciences and Technology, Pakistan, in 2007, the M.S. degree in computer engineering from the Lahore University of Management Sciences, Pakistan, in 2009, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada, in 2020. He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, from 2020 to 2022. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, McMaster University, Canada. His research interests include image processing, perceptual image quality assessment, and large-scale dataset design.

Zhou Wang (Fellow, IEEE) received the Ph.D. degree from The University of Texas at Austin in 2001. He is currently the Canada Research Chair and a 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 80,000 citations (Google Scholar). He served as a member of IEEE Image, Video and Multidimensional Signal Processing Technical Committee (2020–2022) and IEEE Multimedia Signal Processing Technical Committee (2013–2015). He was elected as a fellow of Royal Society of Canada, Academy of Science, in 2018, and a fellow of Canadian Academy of Engineering in 2016. He was a recipient of 2021 Technology Emmy Award, 2016 IEEE Signal Processing Society Sustained Impact Paper Award, 2015 Primetime Engineering Emmy Award, 2014 NSERC E. W. R. Steacie Memorial Fellowship Award, 2013 IEEE Signal Processing Magazine Best Paper Award, and 2009 IEEE Signal Processing Society Best Paper Award. Since 2022, he has been serving as a Senior Editor for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He served as a Senior Area Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING (2015–2019) and an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2016–2018), IEEE TRANSACTIONS ON IMAGE PROCESSING (2009–2014), and IEEE SIGNAL PROCESSING LETTERS (2006–2010), among other journals.