Predicting the Quality of Images Compressed after Distortion in Two Steps

Xiangxu Yu, Christos G. Bampis, Praful Gupta and Alan C. Bovik

***Abstract*—In a typical communication pipeline, images undergo a series of processing steps that can cause visual distortions before being viewed. Given a high quality reference image, a reference (R) image quality assessment (IQA) algorithm can be applied after compression or transmission. However, the assumption of a high quality reference image is often not fulfilled in practice, thus contributing to less accurate quality predictions when using stand-alone R IQA models. This is particularly common on social media, where hundreds of billions of user-generated photos and videos containing diverse, mixed distortions are uploaded, compressed, and shared annually on sites like Facebook, YouTube, and Snapchat. The qualities of the pictures that are uploaded to these sites vary over a very wide range. While this is an extremely common situation, the problem of assessing the qualities of compressed images against their pre-compressed, but often severely distorted (reference) pictures has been little studied. Towards ameliorating this problem, we propose a novel two-step image quality prediction concept that combines NR with R quality measurements. Applying a first stage of NR IQA to determine the possibly degraded quality of the source image yields information that can be used to quality-modulate the R prediction to improve its accuracy. We devise a simple and efficient weighted product model of R and NR stages, which combines a pre-compression NR measurement with a post-compression R measurement. This first-of-a-kind two-step approach produces more reliable objective prediction scores. We also constructed a new, first-of-a-kind dedicated database specialized for the design and testing of two-step IQA models. Using this new resource, we show that two-step approaches yield outstanding performance when applied to compressed images whose original, pre-compression quality covers a wide range of realistic distortion types and severities. The two-step concept is versatile as it can use any desired R and NR components. We are making the source code of a particularly efficient model that we call 2stepQA publicly available at https://github.com/xiangxuyu/2stepQA . We are also providing the dedicated new two-step database free of charge at http://live.ece.utexas.edu/research/twostep/index.html .**

***Index Terms*—Image quality assessment, two-step, reference-no-reference, low quality reference image**

I. Introduction

GLOBAL mobile data traffic grew 63 percent in 2016, while mobile data traffic has grown 18-fold over the past 5 years [1]. Mobile image and video traffic comprises most of the overall mobile data that is transmitted. Online service providers like Facebook, Instagram, Netflix and YouTube generate, store, and transmit enormous quantities of visual content every day. At the same time, users increasingly expect higher quality visual data, which poses significant challenges to providers seeking to optimize the visual quality of their content under increasingly difficult bandwidth conditions.

The digital pictures captured by inexpert consumers are particularly prone to a wide variety of distortions during the capture process, before they are compressed. This makes it much more difficult to predict the perceptual quality of the pictures following compression. The innovation we make here is to devise ways of assessing the quality of the ultimately compressed pictures, while also accounting for their innate, pre-compressed state of imperfect perceptual quality.

Generally speaking, objective image quality assessment (IQA) algorithms can be classified into three broad categories, according to whether a reference image is available. Full-reference IQA algorithms require access to a complete reference image, while reduced-reference IQA algorithms require less information derived from a reference image. Since we will use them in the same way, here we will collectively refer to both of these simply as reference (R) models. If no reference image is available, then no-reference (NR) or 'blind' IQA algorithms must be used.

Given high quality reference data, R IQA models are available that yield excellent predictions of human quality judgments. Successful R models include SSIM [2], MS-SSIM [3], VIF [4], FSIM [5], VSI [6] and RRED [7]. However, high quality reference data is often not available. Indeed, a highly practical area of inquiry that has remained little studied is the design of R IQA models that account for the possibly inferior quality of a reference image to produce better quality predictions.

There are many common types of distortions that can occur before compression, such as film grain, blur, over/under-exposure and up-scaling, which can combine to degrade the quality of a captured image. These kinds of authentic, 'in-capture' artifacts are often a problem for inexpert, amateur photographers who may have unsteady hands or utilize improper lighting. These inferior quality images are then compressed, introducing further distortion. This scenario is very common, as for example on the hundreds of billions of social media images, often of imperfect quality, that are annually uploaded onto social media and subsequentially compressed (or re-compressed). These processes could greatly benefit by the introduction of perceptual compression control mechanisms that account for the intrinsic quality of each image before it is compressed.

Our problem here is different from the previously-studied multi-distortion R IQA problem, where a high quality reference image is perceptually compared against a multiply-

X. Yu, P. Gupta and A. C. Bovik are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, USA (e-mail: yuxiangxu@utexas.edu; praful_gupta@utexas.edu; bovik@ece.utexas.edu). C. G. Bampis are with Netflix Inc (e-mail: cbampis@gmail.com).

[0]IEEE Transactions on Image Processing (2019)

distorted version of it. Generally, when predicting, and often adjusting the quality of images to be compressed, an R IQA model is often applied, but it may not deliver accurate compressed quality predictions because of the imperfect reference. While an NR IQA algorithm could be applied directly on the distorted image, NR models remain limited in their efficacy, and there is still value in making the reference comparison.

Alternately, we may attempt to combine R and NR models to improve prediction performance. Our concept involves two steps. First, a NR model is applied to ascertain the innate degree of distortion of the source image. To assess or guide the overall quality following compression, an R model is then applied as a second step to measure the deviation between the source image and the compressed image, while also accounting for the NR quality measurement. In this way, the collective commingled effects of both compression and in-capture artifacts may be predicted, leading to more accurate and robust results.

Current public image quality databases, such as LIVE [8], TID2013 [9], and CSIQ [10], only contain high quality reference images. Therefore, to be able to develop and test two-step models, we have also created a new database containing source images distorted by a wide variety of mixtures of authentic distortions of diverse quality levels, along with various compressed versions of each.

The rest of the paper is organized as follows. Section II briefly discusses relevant progress on the IQA problem and work related to the two-step concept. Section III describes the new two-step IQA approach in detail. Section IV describes the new subjective image database, while Section V discusses the experiments conducted on it. Section VI concludes the paper with ideas for future work.

II. Related Work

A wide variety of generally effective R IQA models are available.

These include SSIM, VIF, MAD [11], FSIM, VSI and many others [12]–[14]. SSIM is a benchmark among modern R IQA models, and has many variations, including MS-SSIM and IW-SSIM [15]. VIF measures information extracted from both a reference image and a distorted image, a concept also used in the suite of RRED models.

MAD is based on the argument that multiple strategies should be used to assess image quality. FSIM modifies SSIM using two features, local phase congruency and gradient magnitude. The authors of [16] show that a combination of different R models can lead to improved performance, but this approach does not in any way address the problem of an imperfect reference image.

Most early NR IQA models assumed images to be distorted by a particular type of distortion. However, we are more interested in more powerful generalized models, which usually rely on natural image statistic (NSS) models [17] that are sensitive to diverse distortions, since broad application scenarios (such as social media sharing) involve pictures afflicted by diverse, complex, and commingled impairments prior to compression. General-purpose NR IQA models include DIIVINE [18], BLIINDS-II [19], BRISQUE [20], among others [21]–[24]. Among these, NIQE [25] is a 'completely blind', unsupervised IQA model that is also based on NSS but does not require any training process. In [26]–[28], the authors propose a new concept of a 'pseudo reference image (PRI)', and develop a PRI-based blind IQA framework. CORNIA [29] is a data-driven method which constructs a codebook via K-means clustering to generate features, then uses a Support Vector Regression to estimate quality. There are also a variety of NR IQA algorithms based on deep learning, such as PQR [30], DLIQA [31], RankIQA [32] and methods described in [33]–[35].

There is prior work related to, but different from the two-step concept. For example, some authors have proposed using both R and NR models within a same system, although not in direct combination. The authors of [36] apply both R and NR video quality assessment models to predict the quality of encoded videos after transmission. An R method is employed to measure the transmission loss, while an NR model is used to capture degradation from encoding at a reference node. However, the compared source video is still assumed to be of undistorted high quality.

Our proposed models are the first attempt to apply a two-step NR-then-R IQA approach to address the problem of predicting the quality compressed images when only an imperfect reference is available. This concept is of great consequence in social media (and digital camera) applications, where it is desirable to be able to accurately control the perceptual quality of the encodes that are generated before sharing. Thus far, there has been very little attention directed forwards this different problem. The two-step approach that we take here utilizes a simple product combination of R and NR IQA models. However, it delivers performance that is not exceeded by using more complicated NR-R combinations, and which significantly exceeds the performance of stand-alone R models.

In [37], the authors consider images undergoing multiple distortion stages, and point out that in such cases IQA performance on a current stage could be improved by propagating quality levels from previous stages. While they note that distorted images may be used as references, they do not propose an NR-R combination to handle them. The authors of [38] note the problems associated with a 'corrupted reference,' and take the different approach of modifying full-reference algorithms like SSIM and VIF to deal with imperfect references.

Regarding related subjective databases, the authors of [37] also introduced two new databases including a large number of images afflicted by multiple stages of distortions. However, they did not conduct a human study to obtain subjective scores, relying instead on MS-SSIM scores as proxies. The LIVE Multiply Distorted Database [39] contains images with two distortion stages and subjective scores, but the reference images are of high quality. Because of the lack of any subjective database containing low quality reference, we took the effort of developing one for public use, as described in Section IV.

III. Two-Step IQA Model

Reference IQA models assume of the availability of a reference image of high quality, and operate by predicting the quality of a distorted image by making a perceptual comparison of it with a reference image. Thus, a reference IQA model is actually a *perceptual fidelity measure* [4]. In other words, R IQA models only provide *relative* image quality scores.

Given a pristine image of high quality, such as the image in Figure 1(a), a reference IQA model (e.g., MS-SSIM) can be used to assess the quality of a JPEG compressed version of it (Figure 1(b)) by measuring perceptual fidelity deviations between the images in Figures 1(a) and 1(b). But if the quality of the reference is degraded, as in Figure 1(c), then reference IQA models become unreliable. We illustrate such a scenario in the following. The 'reference' images in Figures 1(a) and 1(c) are displayed with their associated subjective Mean Opinion Scores (MOS), which are available since these images were drawn from the LIVE In the Wild Challenge IQA database [40]. Figures 1(b) and Figure 1(d) are compressed versions of these same respective reference images with both associated MOS and Difference Mean Opinion Scores (DMOS), as well as MS-SSIM scores. These images are part of the new subjective database described in Section IV, and have both types of subjective annotations. In Figure 1, the MS-SSIM values are in monotonic agreement with DMOS (increasing MS-SSIM corresponding to decreasing DMOS), indicating that the image in Figure 1(d) is of superior quality to the one in Figure 1(b). However, the MS-SSIM score and MOS have a reverse relationship (increasing MS-SSIM corresponding to decreasing MOS). Indeed, the MOS values strongly indicate that the perceptual quality of the image in Figure 1(d) is worse than that of the image in Figure 1(b). In this case, DMOS does not accurately indicate the level of subjective quality, which is indicative of situations where reference IQA models may fail to accurately predict the quality of compressed images.

While one might consider simply using an NR IQA model to directly predict the absolute quality of the distorted-then-compressed images, this is currently not an acceptable alternative. While much progress has been made on NR IQA model design, even the best algorithms cannot yet deliver the performance needed in demanding consumer applications [34], [35], [40]. Rather than setting aside the valuable information contained in an imperfect reference image, it is a far better option to attempt to account for the *a priori* quality of the reference, and how it impacts the reference measurement. Towards this end, we introduce a combined two-step NR-then-R approach, whereby no-reference and reference quality measurements are applied in sequence, before and after compression, respectively, and are then combined in a principled way.

As is illustrated in Figure 2, given an input image $\mathbf{I}$ and its compressed version $\mathbf{I_c}$, an NR component first predicts the perceptual quality Q_{NR} of $\mathbf{I}$. Once the image is compressed, an R IQA score is generated to account for the perceptual quality difference Q_{R} between $\mathbf{I_c}$ and $\mathbf{I}$. The two-step process is then completed by combining Q_{NR} with Q_{R}. This may be viewed as a process of conditioning Q_{R} on Q_{NR}, where the predicted source image quality serves as "prior" knowledge, converting the relative quality result obtained by the reference IQA model into an absolute score, which better fits with subjective opinions.

The main advantage of the two-step model is visually illustrated in Figure 3, by considering a hypothetical image quality axis spanning the entire quality range from low quality to high quality. The true perceptual quality of an image is represented by its distance from the space of undistorted, natural images. A reference module can only measure the distance between a pristine image $\mathbf{I}$ and its compressed version $\mathbf{I_c}$: When $\mathbf{I}$ is of high quality, it will be close to the natural image space, and the reference module score may be regarded as an accurate prediction of the quality of $\mathbf{I_c}$. However, if $\mathbf{I}$ is of degraded quality, i.e. at a distance from the natural image space, then the no-reference module predicts this perceptual distance, which can then be used to augment the reference IQA result, thereby yielding a better prediction of the overall perceptual distance from $\mathbf{I_c}$ to the natural image space. While the method of combining the NR and R stages may be conceived broadly, in two-step they may also be integrated as a simple product of suitably adjusted R and NR prediction scores Q_{R} and Q_{NR}, yielding a final two-step score Q_{2step}.

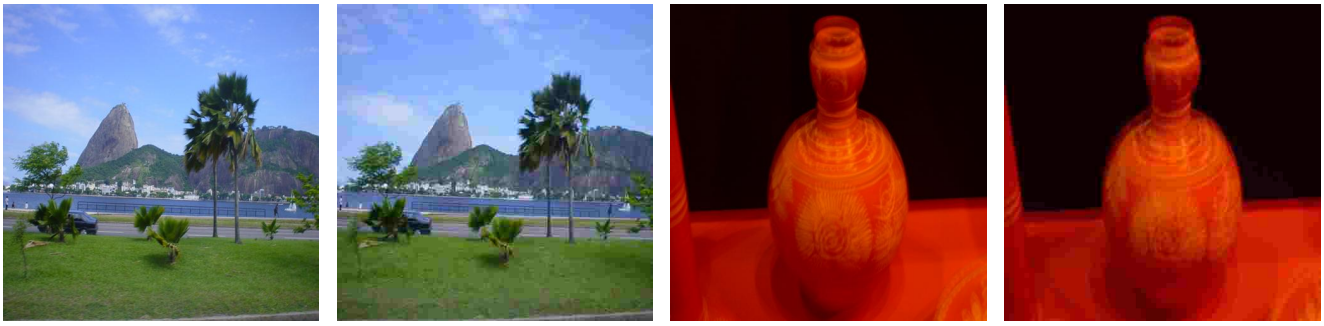
A. Reference IQA Module

A reference IQA module aims to capture perceptual quality differences between a distorted image and a reference image. Naturally, a robust, high-performance R IQA should be used in the design of a two-step model, since the system should perform well at gauging the perceptual effects of compression when the source image is not distorted. As mentioned earlier, there is now a rich variety of effective reference image quality models. From among these, we will use MS-SSIM as an exemplar R module for comparing $\mathbf{I}$ with $\mathbf{I_c}$. MS-SSIM has found considerable commercial success thanks to its simplicity and high performance. MS-SSIM compares luminance, contrast and structural quality information in a multi-scale fashion. MS-SSIM delivers quality scores that fall in the range [0, 1], where larger values correspond to better quality.

B. No-Reference IQA Module

As discussed in Section II, the majority of NR IQA models are data-driven, and depending on a process of training on one or more database(s) of distorted images labelled by human subjective opinion scores. There are also unsupervised 'opinion-unaware' NR algorithms like NIQE, and IL-NIQE [41], which are constructed using NSS.

In the two-step model, the aim of the NR module is to provide prior information about the innate perceptual quality of the source image and use it to improve the R IQA result when the source is distorted. As an effective and flexible exemplar, we will use the NIQE index, which is a completely blind IQA model, as the NR part of a simple and very effective two-step model. The empirical distributions of mean-subtracted and divisively normalized luminance coefficients of high quality images which drive NIQE are known to reliably



(a) MOS: 79

(b) MOS: 51
DMOS: 52
MS-SSIM: 0.9346

(c) MOS: 45

(d) MOS: 36
DMOS: 32
MS-SSIM: 0.9674

Fig. 1. (a) A high quality reference image. (b) JPEG compressed version of (a). (c) A low quality reference image. (d) JPEG compressed version of (c).

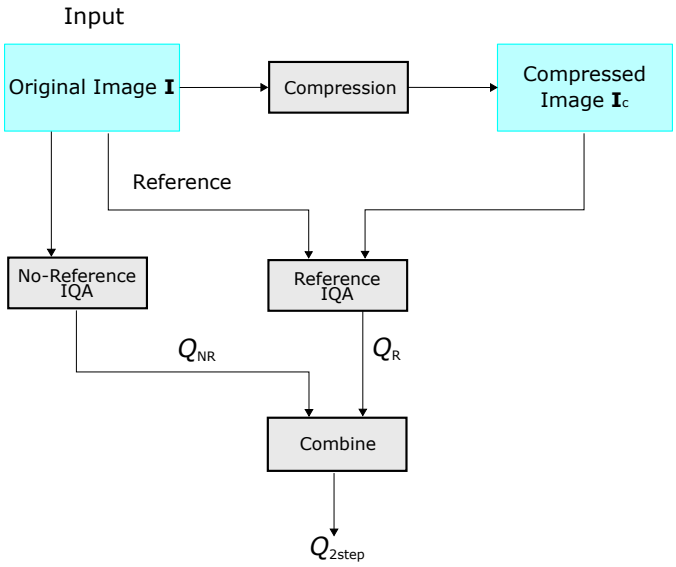


Fig. 2. Overview of two-step model. The original image $\mathbf{I}$ is compressed to obtain $\mathbf{I_c}$. A reference module is then applied to $\mathbf{I}$ and $\mathbf{I_c}$ resulting a predicted quality score Q_{R}. $\mathbf{I}$ is also processed by a no-reference module to generate a predicted score Q_{NR}. Q_{NR} and Q_{R} are then together fed into the two-step model outputting a final predicted quality score Q_{2step}.

follow a Gaussian distribution, while they tend to stray from gaussianity in the presence of distortions. NIQE measures these statistical deviations using a simple Mahalanobis-like measure of the distance between the NSS feature distribution of a test image, and of a pristine model. Unlike many trained IQA models, NIQE is very general, while delivering good prediction performance.

C. Two-Step Model

The goal of a two-step IQA model is to combine NR and R IQA modules to improve the accuracy of systems that predict the quality of compressed images that may have been degraded before compression. Generally, such a two-step model should fulfill three important properties:

1) If compression does not occur, or has an imperceptible effect on quality, then the two-step model should report the innate source (reference) image quality.
2) If the source is pristine, then the two-step model should accurately predict the effect of compression on perceived quality.
3) If the source is already distorted and then compressed with perceptual loss, then the two-step model should yield a better prediction than either the R and NR components applied on the compression image.

While there are different ways to achieve the basic two-step concept, a straightforward, simple, and effective method is to define a two-step model as a product of suitably re-mapped versions of the constituent NR and R components:

$$Q_{\mathsf{2step}} = Q_{\mathsf{R}} \cdot Q_{\mathsf{NR}}, \tag{1}$$

where Q_{R} is the reference IQA score that perceptually compares a compressed image with its reference, and Q_{NR} is NR prediction of the reference image quality. The remapping process, which will be discussed in detail, accounts for the different ranges of the NR and R outputs.

As a simple canonical example, let the NR and R components be NIQE and MS-SSIM respectively, which, following rescaling, yields a particularly simple and effective two-step model that we call 2stepQA:

$$Q_{\mathsf{2stepQA}} = \text{MS-SSIM} \cdot (1 - \frac{\text{NIQE}}{\alpha}), \tag{2}$$

where $Q_{\mathsf{R}} = \text{MS-SSIM}$ and $Q_{\mathsf{NR}} = 1 - \frac{\text{NIQE}}{\alpha}$, and α is a scaling constant. If the MS-SSIM scores fall within [0, 1], where MS-SSIM $= 1$ indicates perfect quality (the usual assumption), then the raw NIQE scores should be rescaled to the same interval prior to taking the product (1). Since NIQE scores increase with worsening picture quality on a scale of about [0,100] on known databases, we simply fix $\alpha = 100$.

Of course, for a variety of possible reasons it may be desired to use NR and/or R IQA models other than NIQE or MS-SSIM. This may arise because of known, specific distortions or a desire to use more sophisticated models. These components can also be integrated using the two-step concept to obtain better performance. However, the constituent NR and R models must be remapped before combining them via (1). Next, we describe a generalized way of remapping the R and NR elements so that they can be combined by multiplication. This lends a high degree of versatility to two-step IQA modeling, as it provides a general design framework.

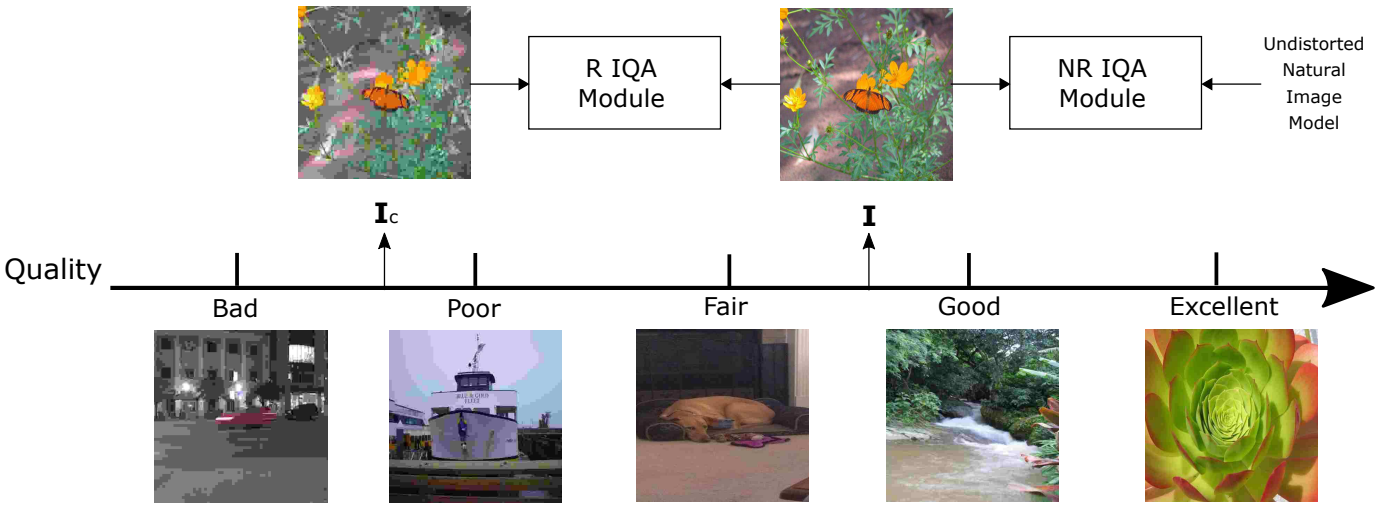


Fig. 3. Visual illustration of the two-step IQA concept. The quality axis spans low to high quality, with five exemplar distorted images shown below each Likert mark. Above the axis, image $\mathbf{I}$ is an imperfect reference image having a fair quality level, while $\mathbf{I_c}$ is its compressed version, which has a much worse quality. The reference module measures the deviation of $\mathbf{I_c}$ from $\mathbf{I}$, while the no-reference module evaluates the distance between $\mathbf{I}$ and the undistorted natural image space.

D. Logistic Remapping

To properly develop a multiplicative combination of NR and R models, it is beneficial to map them to the same range and trend. The ranges of quality scores generated by different IQA algorithms varies significantly. Many R IQA models, such as SSIM, MS-SSIM, and FSIM deliver output quality scores on [0, 1], whereas many NR IQA models, which are trained on human subjective scores are mapped to MOS/DMOS on [0, 100]. Thus, in our basic two-step model the R and NR scores to be combined are mapped to the same range to avoid influencing their relationship to perceptual quality. To preserve monotonicity, allow for generalizability, and to scale the scores to either [0, 1] or the MOS range, we deploy a simple logistic mapping of the reference and no-reference IQA scores.

Specifically, we use a four-parameter, monotonic logistic function to fit each predicted NR or R quality score Q to [0, 100]:

$$Q' = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-(Q-\beta_3/|\beta_4|)}}, \tag{3}$$

where Q' is the rescaled score after finding the least-squares best-fitting logistic function over the four parameters $\{\beta_i; i = 1, .., 4\}$.

The parameters β can be effectively determined by using the subjective data from one or more IQA databases. For example, one could find the optimal βs for a number of IQA models by minimizing the squared error between the remapped objective scores and the MOS values from the LIVE IQA Database. Since a degraded image may be used as the reference image, the entire LIVE Database distorted image corpus could be used to fit the logistic function to obtain the parameters $\beta_{\mathsf{NR}i}, i = 1, .., 4$ for each NR model. Since in our design the possibly distorted image is then subjected to compression, then the JPEG subset of the LIVE IQA database could be used to determine the parameters $\beta_{\mathsf{R}1} - \beta_{\mathsf{R}4}$ for any R model.

Given a compressed image and its possibly distorted reference version, the NR module is applied on the reference image to generate an NR quality score Q_{NR}, while the R component is conducted on both the distorted and the reference images to obtain an R quality score Q_{R}. The rescaled scores Q'_{NR} and Q'_{R} can then be computed using (3) using $\{\beta_{\mathsf{NR}i}; i = 1, .., 4\}$ and $\{\beta_{\mathsf{R}i}; i = 1, .., 4\}$.

In this way, the scores predicted by the R and NR models are remapped to the same range as MOS (or by similar process, to [0, 1] if desired) without loss of information or accuracy. Of course, if a model is trained on MOS, it does not need remapping, since it already has the same score range as MOS.

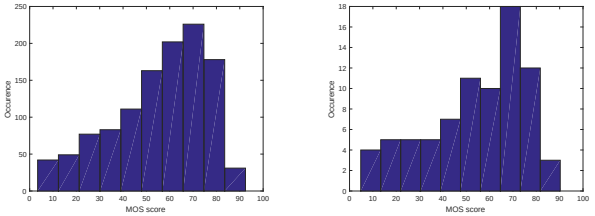
We introduce an additional fitting exponential parameter γ to control the relative weighting of the NR and R modules. Thus the remapped scores Q'_{NR} and Q'_{R}, which have the same MOS range, are combined as follows:

$$Q_{\mathsf{G}} = (Q'_{\mathsf{NR}})^{\gamma} \cdot (Q'_{\mathsf{R}})^{1-\gamma}, \tag{4}$$

where $\gamma \in [0, 1]$ adjusts the relative contributions of the R and NR components. As discussed in Section V-C, the value of γ can depend on such factors as the relative accuracy of the R or NR IQA models. We find that the performances of R and NR models can be significantly improved using this generalized model.

IV. A New Distorted-Then-Compressed Image Database

Current mainstream image quality databases, such as LIVE IQA, TID2013, and CSIQ, are widely used in IQA research. The LIVE IQA Database, which contains 29 reference images and 779 distorted images of five distortion types, was the first large public-domain IQA database. TID2013, which extends TID2008, contains 3000 images with 24 different kinds of distortions. CSIQ contains 30 original images, each distorted by one of six different types of distortions. These major databases have largely support the development of modern IQA algorithms over the past 15 years. However, since they all



(a) (b)

Fig. 4. (a) MOS distribution of the entire LIVE In the Wild Challenge IQA Database. (b) MOS distribution of the 80 selected reference images in the new LIVE Wild Compressed Database.

make use of high quality pristine images as reference images, these databases are not useful tools for studying the influence of distorted reference images on reference quality prediction performance.

A recently published database, called the LIVE In the Wild Challenge IQA Database, contains more than 1100 authentically distorted images captured by a wide variety of mobile devices. The distortions in it are representative of those encountered in practical consumer applications, where the images produced by uncertain amateur hands are often of reduced quality. Towards the development of algorithms that can assess the overall quality of these kinds of image after they are also compressed, we have created a new database which we introduce here, and call the LIVE Wild Compressed Picture Quality Database, which uses real-world, authentically degraded images as reference images. In the following, we detail the new database.

A. Content

A total of 80 images were chosen from [40] to serve as references in the new LIVE Wild Compressed Picture Quality Database. These were selected to span a wide range of content and quality levels. Figure 4 shows that the 80 selected images have a similar MOS distribution as the entire parent database [40]. The MOS of the 80 images nearly span the entire image quality range. These authentic reference images contain numerous types and complex combinations of in-capture distortions such as blur, over/under-exposure, lighting etc.

The reference images were then JPEG compressed using the Matlab JPEG tool into four different, broadly distinguishable severity levels. Following the design procedure used in the creation of other leading IQA databases [8], [9], [42], the four levels of image compression were designed to create a wide range of perceptually separable impaired pictures. Only four levels were used, since as in [8], [9], [42] this number was deemed adequate to cover the distortion space, and importantly, was necessary to limit the size of the human study. For each content, there are four compressed versions, yielding 320 compressed images. Some examples of both pristine and compressed versions of images in the database are shown in Figure 5.

B. Human study

We conducted a human study in the LIVE subjective study lab. Most of the subjects participated in the study were UT-Austin students inexperienced in understanding image quality assessment, or compression impairments. Each subject participated in two ~30 minute sessions at least 24 hours apart. The database was divided equally and randomly into two parts in each session, each containing 40 contents, including 40 pristine images and their respective four different compressed images, hence each subject viewed 200 images per session. The images were displayed in random order with each image shown only once during each session. Presentations of each unique content were separated by at least 5 images. For each subject, two sessions were generated and assigned in random order. The total number of subjects taking part in the study was 29, and all of them successfully finished both sessions. Most subjects completed each session within 20 minutes.

All of the subjects participated in a visual acuity test, and were asked whether they had any uncorrected visual deficiency. A viewing distance of 2 feet was measured and approximately maintained during testing. Before starting the experiment, each subject was required to read and sign a consent form including general information about the human study, then the procedure and requirements of the test were explained. A short training session was presented before the first test session using a different set of images than the test experiment. Given each image, the subject was asked to provide an opinion score of picture quality by dragging a slider along a continuous rating bar. As shown in Figure 6, the possible quality range was labelled from low to high with five adjectives: Bad, Poor, Fair, Good, and Excellent. The subjective scores obtained from the subjects were sampled onto numerical quality scores in [1, 100]. A screenshot of the subjective study interface is shown in Figure 7. The interface was developed on a Windows PC using PsychoPy software [43].

The subjective MOS were then computed according to the procedures described in [42]. The raw scores were first converted into Z-scores. Let s_{ijk} denote the score assigned by the i-th subject on the j-th image in session $k = \{1, 2\}$. The raw scores were converted into Z-scores per session:

$$z_{ijk} = \frac{s_{ijk} - \bar{s}_{ik}}{\sigma_{ik}}, \tag{5}$$

where $\bar{s}_{ik}$ is the mean of the raw scores over all images assessed by subject i in session k, and σ_{ik} is the standard deviation.

The subject rejection procedure described in ITU-R BT 500.13 [44] was then conducted to remove outliers. After performing the rejection procedure, 6 of the 29 subjects were rejected. The Z-scores of the remaining 23 subjects were then linearly rescaled to [0, 100]. Finally, the MOS of each image was obtained by computing the mean of the rescaled Z-scores. The overall MOS distribution of the LIVE Wild Compressed Picture Quality Database is plotted in Figure 8 for several different compression levels.

C. Analysis

To examine subject consistency, we split the subjective scores obtained on each image into two disjoint equal groups, and compared the MOS on every image, one from each group.

(a) MOS: 83 (b) MOS: 60 (c) MOS: 55 (d) MOS: 40 (e) MOS: 35

(f) MOS: 58 (g) MOS: 51 (h) MOS: 46 (i) MOS: 41 (j) MOS: 32

Fig. 5. (a) High quality reference image. (b)-(e) four JPEG compressed versions of (a) using compression parameter (distortion level) 18, 12, 6 and 3 (from left to right). (f) Low quality reference image. (g)-(j) four JPEG compressed versions of (f) using compression parameter (distortion level) 18, 12, 6 and 3 (from left to right).

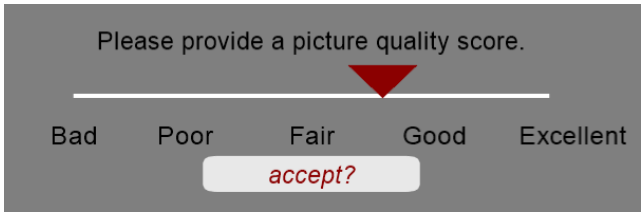
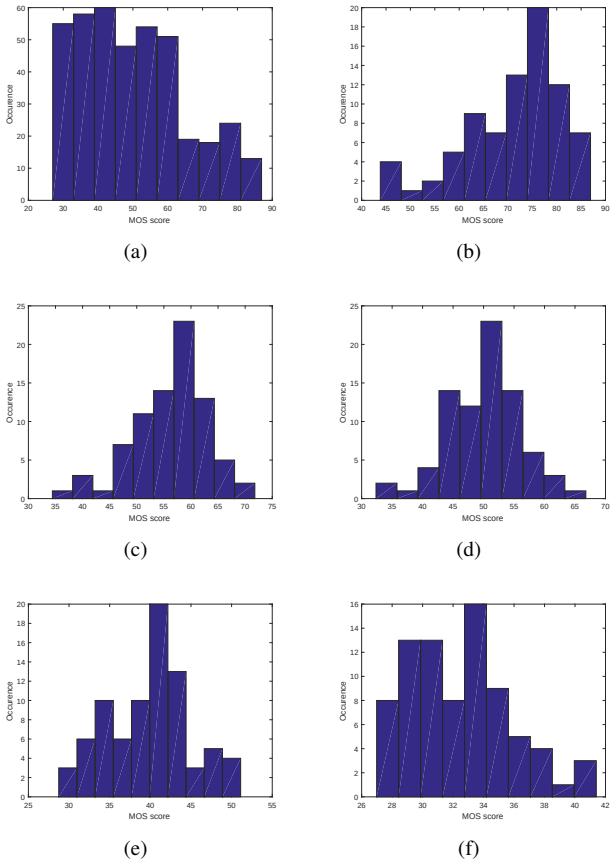


Fig. 6. The rating bar.

Fig. 7. Screenshot of the subjective study interface showing the test image shown to the subject.

The random splits were repeated 25 times and the median Spearman's rank ordered correlation coefficient (SROCC) between the two groups was found to be 0.9805.

Figure 9 shows a box plot of MOS from the LIVE Wild Compressed Picture Quality Database for different compression levels. The MOS decreases, with reducing variance, as the compression is increased. Figure 10 shows the MOS across all contents with each color coded curve at a different compression level. While the curves are nicely



(a) (b) (c) (d) (e) (f)

Fig. 8. (a) MOS distribution across the entire LIVE Wild Compressed Picture Quality Database. (b) MOS distribution of reference images. (c) MOS distribution of compressed images at distortion level 18. (d) MOS distribution of compressed images at distortion level 12. (e) MOS distribution of compressed images at distortion level 6. (f) MOS distribution of compressed images at distortion level 3.

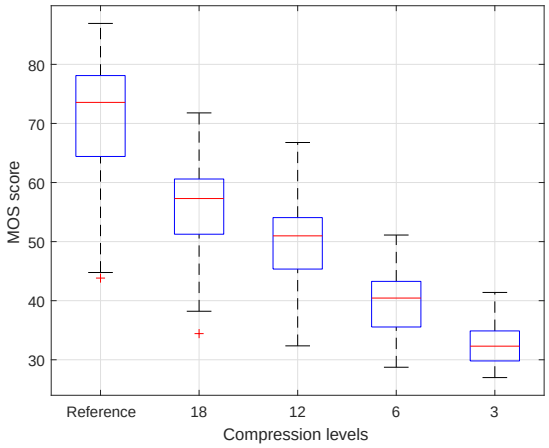


Fig. 9. Box plot of MOS of images in the LIVE Wild Compressed Picture Quality Database for different compression levels. The central red mark represents the median, while the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol.

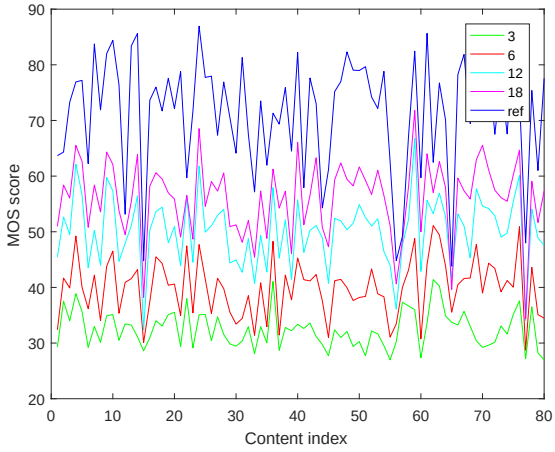


Fig. 10. MOS of all contents for five different compression (distortion) levels, coded by color.

separated by content, it is important to observe the mixing of MOS across contents, caused by the reference distortions.

V. Performance Evaluation

We used the new LIVE Wild Compressed Picture Quality Database to compare the performance of various two-step IQA models, including the 2stepQA algorithms (2) against each other and against other R and NR IQA measures. Most available R IQA databases contain pristine image as references against which to evaluate the fidelity of distorted images. In such scenarios, DMOS is typically used to reduce any biases arising from image content. However, if the reference image is affected by distortion, as in the aforementioned database, DMOS is less likely to reflect subjective opinions correctly. Thus, we only compute and correlate objective quality predictions against MOS, which represents absolute subjective quality.

We evaluated the performance between predicted quality scores and subjective MOS using SROCC and the Pearson's (linear) correlation coefficient (LCC). The predicted IQA scores were passed through a logistic non-linearity (following usual practice) before computing the LCC measure [8]. Larger values of both SROCC and LCC indicate better performance.

Although the 2stepQA model and some of the other compared IQA algorithms (both one-step and two-step) do not require training processes, we divided the database into non-overlapping 80% training sets and 20% test sets by content, to ensure fair comparisons against other learning-based IQA algorithms. Such random train-test splits were repeated for 1000 iterations to avoid unbiased results.

We utilized a number of prominent R IQA algorithms including PSNR, MS-SSIM, FSIM and VSI. Among perceptually-relevant NR IQA algorithms, we tested NIQE, BRISQUE, CORNIA, and PQR implemented using a shallow convolutional neural network (S_CNN) model. Since PQR (S_CNN) is a learned model, we pretrained it on the LIVE IQA Database, then tested the model on the LIVE Wild Compressed Picture Quality Database. These popular IQA algorithms are well established in the IQA literature and have been shown to correlate well against subjective opinions of image quality.

A. Comparisons Against Mainstream IQA Methods

We first conducted a performance comparison between the 2stepQA model (2) and several one-step R and NR IQA algorithms, and report the results in Table I. As expected PSNR, which is not a perceptually-relevant measure of image quality, performed poorly as compared with the other R and NR IQA algorithms, which all correlated at least reasonably well against subjective judgments of quality. However, the 2stepQA index (2) significantly outperformed all of the compared one-step IQA algorithms.

To determine whether the differences in correlations reported in Table I were statistically significant, we conducted a statistical significance test. We utilized the distribution of the obtained SROCC scores computed over 1000 random train-test iterations. The nonparametric Wilcoxon Rank Sum Test [45], which exclusively compares the rank of two sets of observations, was used to conduct hypothesis testing. The null hypothesis was that the median for the (row) algorithm was equal to the median of the (column) algorithm at the 95% significance level. The alternate hypothesis was that the median of the row was different from the median of the column. A value of '1' in the table represents that the row algorithm was statically superior to the column algorithm, while a value of '-1' means the counter result. A value of '0' indicates that the row and column algorithms were statistically indistinguishable (or equivalent). The statistical significance results comparing the performances of the compared IQA algorithms are tabulated in Table II.

To illustrate how the distributions of the SROCC and LCC scores varied by algorithm, Figures 11 and 12 show box-plots of the correlations computed over 1000 iterations for each of the compared algorithms. A lower standard deviation with a higher median SROCC indicates better performance. As may be inferred from the Tables and Figures, the 2stepQA

TABLE I
PERFORMANCES OF THE 2STEPQA MODEL (2) AGAINST VARIOUS ONE-STEP REFERENCE AND NO-REFERENCE IQA MODELS ON THE LIVE WILD COMPRESSED PICTURE QUALITY DATABASE. THE BEST PERFORMING ALGORITHM IS HIGHLIGHTED IN BOLD FONT. ITALICS INDICATE NO-REFERENCE ALGORITHMS.

	PSNR	MS-SSIM	FSIM	VSI	*NIQE*	*BRISQUE*	*CORNIA*	*PQR (S_CNN)*	2stepQA
SROCC	0.4227	0.8930	0.9101	0.7953	0.8457	*0.9091*	*0.9005*	*0.8944*	**0.9311**
LCC	0.4299	0.8923	0.9134	0.8153	0.8407	*0.8966*	*0.8955*	*0.8939*	**0.9305**

TABLE II
RESULTS OF ONE-SIDED WILCOXON RANK SUM TEST PERFORMED BETWEEN SROCC VALUES OF THE IQA ALGORITHMS COMPARED IN TABLE I. A VALUE OF "1" INDICATES THAT THE ROW ALGORITHM WAS STATISTICALLY SUPERIOR TO THE COLUMN ALGORITHM; " − 1" INDICATES THAT THE ROW WAS WORSE THAN THE COLUMN; A VALUE OF "0" INDICATES THAT THE TWO ALGORITHMS WERE STATISTICALLY INDISTINGUISHABLE. ITALICS INDICATE NO-REFERENCE ALGORITHMS.

	PSNR	MS-SSIM	FSIM	VSI	*NIQE*	*BRISQUE*	*CORNIA*	*PQR (S_CNN)*	2stepQA
PSNR	0	-1	-1	-1	-1	-1	-1	-1	-1
MS-SSIM	1	0	-1	1	1	-1	-1	0	-1
FSIM	1	1	0	1	1	1	1	1	-1
VSI	1	-1	-1	0	-1	-1	-1	-1	-1
NIQE	1	-1	-1	1	0	-1	-1	-1	-1
BRISQUE	1	1	-1	1	1	0	1	1	-1
CORNIA	1	1	-1	1	1	-1	0	1	-1
PQR (S_CNN)	1	0	-1	1	1	-1	-1	0	-1
2stepQA	1	1	1	1	1	1	1	1	0

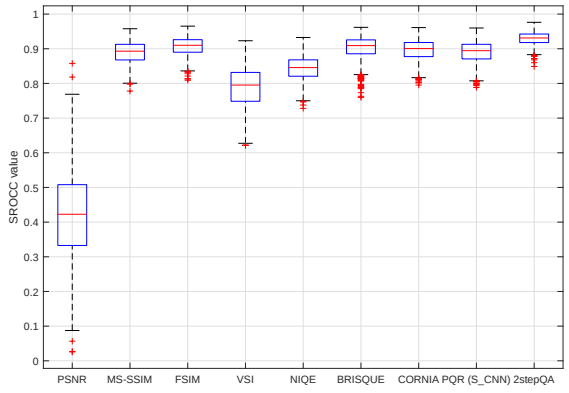


Fig. 11. Box plot of SROCC distributions of the compared algorithms in Table I over 1000 trials on the LIVE Wild Compressed Picture Quality Database.

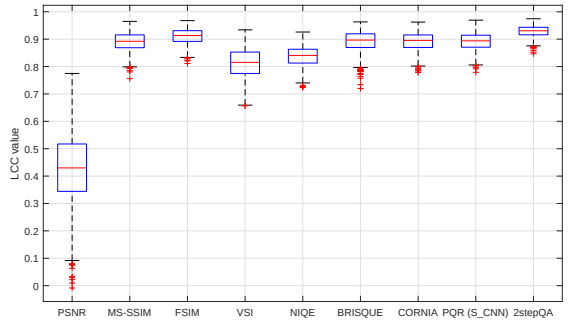


Fig. 12. Box plot of LCC distributions of the compared algorithms in Table I over 1000 trials on the LIVE Wild Compressed Picture Quality Database.

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED 2STEPQA AGAINST MS-SSIM ON TWO EQUALLY DIVIDED SUBSETS OF THE LIVE WILD COMPRESSED PICTURE QUALITY DATABASE. SUBSET 1 CONTAINED COMPRESSED IMAGES HAVING BETTER QUALITY REFERENCE IMAGES (OR LOWER NIQE SCORES), WHILE SUBSET 2 CONTAINED COMPRESSED IMAGES HAVING WORSE QUALITY REFERENCE IMAGES.

	Subset 1		Subset 2	
	MS-SSIM	2stepQA	MS-SSIM	2stepQA
SROCC	0.9395	0.9434	0.8546	0.8991
LCC	0.9419	0.9458	0.8551	0.8980

model (2) exhibited significantly higher and more reliable correlations against subjective quality than all of the compared one-step R and NR IQA algorithms.

Since the design of 2stepQA involves MS-SSIM as its integral component, it is of interest to explore why 2stepQA is able to improve on MS-SSIM on the LIVE Wild Compressed Picture Quality Database. To do this, we divided the database into two equal-sized subsets based on the quality of the reference images. The no-reference NIQE engine was used to evaluate the quality of the reference images and to divide the references into two quality classes. The first class comprised of 160 compressed images derived from 40 high quality references, while the second class used the remaining 160 images with lower quality references. As may be seen in Table III, both MS-SSIM and 2stepQA correlated similarly with subjectivity on the subset of high quality reference images (Subset 1). However, 2stepQA significantly outperformed MS-SSIM on the subset of poor quality reference images (Subset 2) because of the contribution of the NR component, indicating that 2stepQA can significantly improve on the performance of stand-alone R IQA models operating on low-quality reference images.

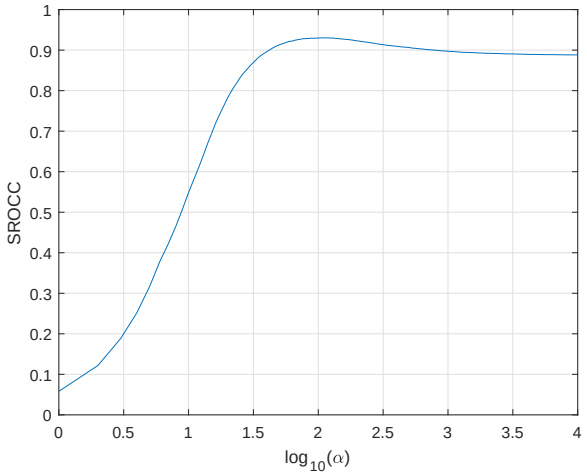


Fig. 13. Performance of the 2stepQA index (2) as the parameter α is varied, showing a peak at $\alpha = 100$.

B. Selection of 2stepQA Parameter (α)

The 2stepQA index involves a free parameter α which affects the mapping to MOS. Figure 13 shows the SROCC values of the 2stepQA model (2) for a wide range of values of α. The model attains its best performance at $\alpha \approx 100$. However, the performance of 2stepQA is robust over a wide range of the values of $\alpha \in [50, 150]$.

C. More General Two-Step Models

The general two-step model outlined earlier lends a more flexible approach towards combining different R and NR components. Table IV and V plot the performance of general two-step models incorporate several combinations of R and NR IQA components.

We considered four R IQA algorithms: PSNR, MS-SSIM, FSIM and VSI, and four NR IQA algorithms: NIQE, BRISQUE, CORNIA and PQR (S_CNN). The logistic function parameters (βs) used in (3) were optimized on the LIVE Image Quality Database. Learning-based NR IQA models, such as BRISQUE, CORNIA and PQR (S_CNN), were then trained on the subset of LIVE Wild Challenge Database that excludes the 80 reference images of the LIVE Wild Compressed Picture Quality Database.

To highlight the importance of using accurate NR algorithms, we also included experimental results by replacing the NR scores with the actual MOS of the reference images. This serves as an idealized basis of comparison of NR algorithms evaluated on the same reference images. The median SROCC and LCC of the various two-step models over 1000 iterations of randomly chosen disjoint 80% training and 20% test subsets are reported in Tables IV and V, respectively. The optimal exponents γ in (4) are reported along with the correlation scores.

As one would expect, a high-performing R algorithm is assigned larger weights $(1-\gamma)$ in the general two-step model (4) when the NR component is fixed, as reflected by the γ values reported in Tables IV and V. When tested on the LIVE Wild Compressed Picture Quality Database, FSIM and MS-SSIM outperform other one-step algorithms, including VSI and PSNR (Table I), and for these models, the optimal γ values were smaller. When the low-performing PSNR was combined with any of the four NR models, γ took much larger values (≥ 0.5) than for other R models, implying that NR models dominate the two-step product when combined with low-performing R models like PSNR. However, when a high-performing R model is used, such as FSIM or MS-SSIM, the corresponding optimal values of γ are smaller (< 0.5), emphasizing the value of having a high-performing R model in the product.

Similarly, an effective NR module is essential to achieving better performance of two-step models. The contributions of different NR algorithms in the general two-step model are shown in Table VI. Although the performance of the two-step model (4) is not influenced as much by the choice of NR algorithm as by R algorithm, it is clear that higher-performing NR algorithms result in better overall performance.

Tables VII and VIII plot the performances of general two-step models but fixing $\gamma = 0.5$. As compared with the results in Tables IV and V, where optimal γ values were used, the generalized models still achieved nearly optimal performance when $\gamma = 0.5$ for most combinations of R and NR IQA modules.

In the general two-step model, the parameter γ reflects the weight or importance of the NR component relative to the R component. If in a given compression application it is determined that the pre-compressed reference images are of high-quality, then the relative contribution of the NR component may be reduced or even eliminated (γ=0), while the importance of the R component is increased. If the reference images are known to present with a wide range of perceptual qualities, then γ may be increased to better reflect the importance of the NR component in the final quality evaluation of the compressed image. Different training sets may result in different values of γ. For example, since the reference images in the LIVE Challenge Compressed Database take on a wide range of perceptual qualities, the NR component may be assigned a larger weight (larger γ). Conversely, when training models on the LIVE IQA Database, where the reference images are of exceptionally high-quality, then the NR component becomes much less important, and the value of γ may be reduced. Overall, γ depends on the application scenario.

D. Weighted 2stepQA

The general two-step model allows the choice of parameter γ to adjust the relative weights assigned to the NR and R IQA components. Small values of γ correspond to less emphasis on the NR score, and conversely, larger values of γ increase the importance of the NR contribution. Figure 14 plots the performance of the general two-step model (4) against γ using NIQE and MS-SSIM as the combination components, i.e., "weighted 2stepQA." The best SROCC result was attained for $\gamma = 0.47$, as shown in Figure IV. The performance of the generalized model was robust over the range $\gamma \in [0.4, 0.55]$, indicating that the NR and R components of 2stepQA are of roughly equal importance.

TABLE IV
SROCC PERFORMANCES OF TWO-STEP COMBINATIONS OF REFERENCE AND NO-REFERENCE IQA MODELS ON THE LIVE WILD COMPRESSED PICTURE QUALITY DATABASE. THE PARAMETER γ IS GIVEN IN PARENTHESES. RESULTS USING MOS AS THE 'IDEAL' NR MODULE SCORE IS SHOWN FOR COMPARISON.

	PSNR	MS-SSIM	FSIM	VSI
NIQE	0.6609(0.63)	0.9283(0.47)	0.9263(0.37)	0.8775(0.54)
BRISQUE	0.6833(0.62)	0.9333(0.46)	0.9357(0.41)	0.8980(0.53)
CORNIA	0.6807(0.51)	0.9356(0.39)	0.9375(0.36)	0.8992(0.46)
PQR (S_CNN)	0.6769(0.60)	0.9382(0.41)	0.9367(0.36)	0.8970(0.49)
MOS	0.6156(0.71)	0.9401(0.61)	0.9474(0.56)	0.8946(0.64)

TABLE V
LCC PERFORMANCES OF TWO-STEP COMBINATIONS OF REFERENCE AND NO-REFERENCE IQA MODELS ON THE LIVE WILD COMPRESSED PICTURE QUALITY DATABASE. THE PARAMETER γ IS GIVEN IN PARENTHESES. RESULTS USING MOS AS THE 'IDEAL' NR MODULE SCORE IS SHOWN FOR COMPARISON.

	PSNR	MS-SSIM	FSIM	VSI
NIQE	0.6830(0.63)	0.9268(0.47)	0.9278(0.37)	0.8839(0.54)
BRISQUE	0.6743(0.62)	0.9309(0.46)	0.9355(0.41)	0.8988(0.53)
CORNIA	0.6747(0.51)	0.9353(0.39)	0.9394(0.36)	0.9026(0.46)
PQR (S_CNN)	0.6662(0.60)	0.9378(0.41)	0.9392(0.36)	0.8976(0.49)
MOS	0.6064(0.71)	0.9403(0.61)	0.9499(0.56)	0.8981(0.64)

TABLE VI
PERFORMANCES OF NO-REFERENCE IQA MODULES ON 80 REFERENCE IMAGES ON THE LIVE WILD COMPRESSED PICTURE QUALITY DATABASE.

	SROCC	LCC
NIQE	0.5350	0.6742
BRISQUE	0.7217	0.7282
CORNIA	0.6772	0.7523
PQR (S_CNN)	0.7451	0.7175

TABLE VII
SROCC OF GENERAL TWO-STEP MODELS USING DIFFERENT COMBINATIONS OF REFERENCE AND NO-REFERENCE IQA MODELS ON THE LIVE WILD COMPRESSED PICTURE QUALITY DATABASE, FOR $\gamma = 0.5$. MOS AS AN IDEAL NR ALGORITHM IS INCLUDED FOR COMPARISON

	PSNR	MS-SSIM	FSIM	VSI
NIQE	0.6423	0.9312	0.9254	0.8805
BRISQUE	0.6711	0.9339	0.9327	0.8982
CORNIA	0.6863	0.9331	0.9287	0.9022
PQR (S_CNN)	0.6745	0.9367	0.9325	0.8993
MOS	0.5692	0.9384	0.9473	0.8832

TABLE VIII
LCC OF GENERAL TWO-STEP MODELS USING DIFFERENT COMBINATIONS OF REFERENCE AND NO-REFERENCE IQA MODELS ON THE LIVE WILD COMPRESSED PICTURE QUALITY DATABASE, FOR $\gamma = 0.5$. MOS AS AN IDEAL NR ALGORITHM IS INCLUDED FOR COMPARISON.

	PSNR	MS-SSIM	FSIM	VSI
NIQE	0.6578	0.9302	0.9269	0.8871
BRISQUE	0.6626	0.9312	0.9310	0.9004
CORNIA	0.6799	0.9314	0.9286	0.9036
PQR (S_CNN)	0.6663	0.9377	0.9330	0.9005
MOS	0.5572	0.9381	0.9489	0.8858

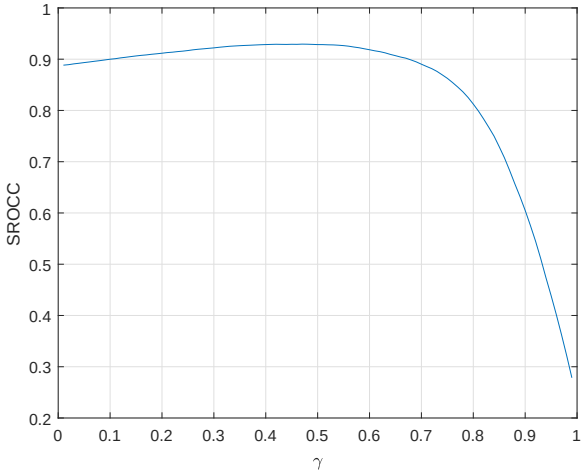


Fig. 14. Performance of the 2stepQA (MS-SSIM + NIQE) model with exponent γ allowed to vary.

TABLE IX
PERFORMANCE OF 2STEPQA ON THE JPEG SUBSET OF THE LIVE IMAGE QUALITY DATABASE AS COMPARED WITH MS-SSIM AND NIQE.

	MS-SSIM	NIQE	2stepQA
SROCC	0.9787	0.9355	0.9632
LCC	0.9819	0.9483	0.9744

E. Performance When the Reference Images are of High Quality

Table IX illustrates the performance of the 2stepQA model on the JPEG subset of LIVE Image Database, where the reference images are of extremely good quality. In this case, as would be expected, 2stepQA does not outperform MS-SSIM, but neither does it significantly underperform MS-SSIM, since the NR component does not contribute much

TABLE X
PERFORMANCE OF DIFFERENT COMBINATION METHODS OTHER THAN MULTIPLICATION ON THE LIVE WILD COMPRESSED PICTURE QUALITY DATABASE

	SROCC
Linear Regression	0.9289
Polynomial Regression (degree 2)	0.9154
Polynomial Regression (degree 3)	0.9253
Polynomial Regression (degree 4)	0.9195
2stepQA	0.9311

to the overall product (see also Table III. However, 2stepQA is statistically superior to MS-SSIM and the other compared R IQA models, when applied to imperfect reference settings, which is a very large and important application space (e.g., social media pictures).

F. Simplicity of the 2StepQA Model

The general two-step concept, and in particular 2stepQA, are simple and very easy to implement, yet are able to significantly outperform other mainstream, stand-alone IQA algorithms. Since in the two-step product concept both the R and NR components are scaled to the same range (e.g., [0, 1]), where 1 = best quality, then the score will be lowered if either the reference image is distorted, or if the compression distorts, or both. The output quality prediction will only be high (approach 1) if the reference is of high quality, *and* the process of compression does not lower the quality.

It is important to mention that we also devised and tested a variety of other ways to combine the R and NR components. Linear regression only obtained comparable performance as the product model, as shown in Table X. Furthermore, polynomial regression of degrees 2, 3 and 4 also did not improve performance over the simple product model.

Of course, this does not mean that the two-step concept cannot be improved on. For example, given that the problem may be viewed as predicting R quality after compression, given an NR quality measurement before compression, we have been working on a conditional (Bayesian) framework, but this will require vastly more data collection (a much larger crowdsourced database than currently exists) to be able to learn accurate predictive models. The 2stepQA model is an intuitive and successful choice that delivers statistically superior performance as compared against state-of-art NR and R algorithms.

VI. CONCLUSION

We described a new two-step framework for the design of algorithms that can predict the quality of distorted pictures (e.g., during capture) after they are subjected to additional compression. The general approach is to combine NR (before compression) with R (after compression) algorithms in a simple exponentially weighted product formulation. In order to facilitate the development, testing, and benchmarking of two-step models for this application, we constructed a new subjective quality resource called the LIVE Wild Compressed Picture Quality Database. This new dedicated resource contains compressed versions of real-world reference images that have already been subjected to complex mixtures of authentic distortions (typically occurring during capture), spanning wide ranges of original quality levels. The two-step framework is general enough to encompass the design of any potential combination of suitable R and NR algorithms. We also highlight a simple exemplar two-step model called 2stepQA, which combines two highly efficient commercial algorithms (MS-SSIM and NIQE), achieving standout efficiency without any need for training. We show that the 2stepQA model outperforms other leading R and NR IQA models applied in isolation (one-step). Adding a training process produces even better results, but at the possible loss of generality, and increased effort and complexity. The standout performance is particularly significant for low quality reference images. The proposed two-step IQA concept is a simple yet efficient way to address the low quality reference IQA problem.

ACKNOWLEDGMENT

The authors thank Meixu Chen of LIVE for her help polishing the paper.

REFERENCES

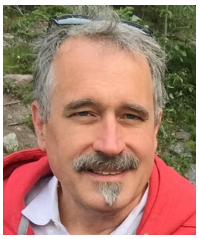
[1] Cisco Visual Networking Index. (2017) Global mobile data traffic forecast update, 2016-2021. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf
[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Asilomar Conf. Signals Syst. Comput.*, vol. 2, pp. 1398–1402, Nov. 2003.
[4] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.
[5] 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, 2011.
[6] 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, 2014.
[7] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, 2012.
[8] 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, 2006.
[9] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, and F. Battisti, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, 2015.
[10] E. C. Larson and D. Chandler, "Categorical image quality (CSIQ) database," 2010.
[11] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011 006:1–011 006:21, Mar. 2010.
[12] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, 2014.
[13] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," *IEEE Int'l. Conf. Image Process.*, pp. 1473–1476, 2012.
[14] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar wavelet-based perceptual similarity index for image quality assessment," *Signal Process., Image Commun.*, vol. 61, pp. 33–43, 2018.
[15] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, 2011.

[16] O. I. Ieremeiev, V. V. Lukin, N. N. Ponomarenko, K. O. Egiazarian, and J. Astola, "Combined full-reference image visual quality metrics," *Electronic Imaging*, vol. 2016, no. 15, pp. 1–10, 2016.
[17] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, 2010.
[18] ——, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, 2011.
[19] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, 2012.
[20] 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, 2012.
[21] 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, 2016.
[22] ——, "BSD: Blind image quality assessment based on structural degradation," *Neurocomputing*, vol. 236, pp. 93–103, 2017.
[23] P. G. Freitas, W. Y. Akamine, and M. C. Farias, "No-reference image quality assessment based on statistics of local ternary pattern," *Int. Conf. Quality Multimedia Exp.*, pp. 1–6, 2016.
[24] H. Wang, J. Fu, W. Lin, S. Hu, C.-C. J. Kuo, and L. Zuo, "Image quality assessment based on local linear information and distortion-specific compensation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 915–926, 2016.
[25] 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, 2013.
[26] X. Min, G. Zhai, K. Gu, Y. Fang, X. Yang, X. Wu, J. Zhou, and X. Liu, "Blind quality assessment of compressed images via pseudo structural similarity," *Proc. IEEE Int. Conf. Multimedia Expo*, pp. 1–6, 2016.
[27] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Trans. on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018.
[28] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.
[29] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," *IEEE Conf. Comp. Vis. Pattern Recog.*, pp. 1098–1105, 2012.
[30] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," *IEEE Int'l Conf. on Image Process.*, pp. 609–613, 2018.
[31] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, 2015.
[32] X. Liu, J. van de Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *CVPR*, 2017, pp. 1040–1049.
[33] J. Yang, B. Jiang, Y. Zhu, C. Ji, and W. Lu, "An image quality evaluation method based on joint deep learning," in *Neural Inf. Process.* Springer, 2017, pp. 658–665.
[34] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, 2018.
[35] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, 2017.
[36] J. Joskowicz, R. Sotelo, J. P. Garella, P. Zinemanas, and M. Simón, "Combining full reference and no reference models for broadcast digital tv quality monitoring in real time," *IEEE Trans. Broadcasting*, vol. 62, no. 4, pp. 770–784, 2016.
[37] S. Athar, A. Rehman, and Z. Wang, "Quality assessment of images undergoing multiple distortion stages," *IEEE Int'l Conf. Image Process.*, pp. 3175–3179, 2017.
[38] W. Cheng and K. Hirakawa, "Corrupted reference image quality assessment," *IEEE Int'l Conf. Image Process.*, pp. 1485–1488, 2012.
[39] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," *Asilomar Conf. Signals Syst. Comput.*, pp. 1693–1697, 2012.
[40] 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, 2016.
[41] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, 2015.
[42] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, 2010.
[43] J. W. Peirce, "Generating stimuli for neuroscience using psychopy," *Front. Neuroinform.*, vol. 2, p. 10, 2009.
[44] I. T. Union, "Methodology for the subjective assessment of the quality of television pictures ITU-R recommendation BT.500-13," *Tech. Rep.*, 2012.
[45] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

Xiangxu Yu received the B.Eng in Electronic and Information Engineering from The Hong Kong Polytechnic University, Hongkong, China, and the M.S. degree in Electrical and Computer Engineering from The University of Texas at Austin, Austin, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Laboratory for Image and Video Engineering, The University of Texas at Austin. His research interests focus on image and video processing, and machine learning.

Christos G. Bampis is with the Video Algorithms group at Netflix. He works on perceptual video quality and quality of experience prediction systems for adaptive video streaming. Before that, he completed his PhD studies at the University of Texas at Austin.

Praful Gupta received the B. Tech. degree in Electrical engineering from the Indian Institute of Technology Roorkee, Roorkee, India, and the M.S. degree in Electrical and Computer Engineering from The University of Texas at Austin, Austin, in 2015 and 2017, respectively. He is currently pursuing Ph.D. degree from The University of Texas at Austin. His research interests include image and video processing, machine learning, and computer vision.



Alan Conrad Bovik (F '95) is the Cockrell Family Regents Endowed Chair Professor at The University of Texas at Austin. His research interests include image processing, digital television, digital streaming video, and visual perception. For his work in these areas he has been the recipient of the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from the Optical Society of America, a 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, and the Norbert Wiener Society Award and the Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. He has also received about 10 'best journal paper' awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. A Fellow of the IEEE, his recent books include The Essential Guides to Image and Video Processing. He co-founded and was longest-serving Editor-in-Chief of the IEEE Transactions on Image Processing, and also created/Chaired the IEEE International Conference on Image Processing which was first held in Austin, Texas, 1994.