Predicting the Quality of Compressed Videos With Pre-Existing Distortions

Xiangxu Yu, Neil Birkbeck, Yilin Wang, *Member, IEEE*, Christos G. Bampis, Balu Adsumilli, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*— Because of the increasing ease of video capture, many millions of consumers create and upload large volumes of User-Generated-Content (UGC) videos to social and streaming media sites over the Internet. UGC videos are commonly captured by naive users having limited skills and imperfect techniques, and tend to be afflicted by mixtures of highly diverse in-capture distortions. These UGC videos are then often uploaded for sharing onto cloud servers, where they are further compressed for storage and transmission. Our paper tackles the highly practical problem of predicting the quality of compressed videos (perhaps during the process of compression, to help guide it), with only (possibly severely) distorted UGC videos as references. To address this problem, we have developed a novel Video Quality Assessment (VQA) framework that we call 1stepVQA (to distinguish it from two-step methods that we discuss). 1stepVQA overcomes limitations of Full-Reference, Reduced-Reference and No-Reference VQA models by exploiting the statistical regularities of both natural videos and distorted videos. We also describe a new dedicated video database, which was created by applying a realistic VMAF-Guided perceptual rate distortion optimization (RDO) criterion to create realistically compressed versions of UGC source videos, which typically have pre-existing distortions. We show that 1stepVQA is able to more accurately predict the quality of compressed videos, given imperfect reference videos, and outperforms other VQA models in this scenario.**

***Index Terms*— Video quality assessment, natural scene statistics, video compression, user-generated-content.**

I. Introduction

In RECENT years, digital images and videos have become remarkably ubiquitous and now constitute the majority of Internet traffic. According to [1], video streaming continues to occupy a growing share of Internet bandwidth, and by 2022, it is expected that 82% of global "moving bits" will be picture and video content. A substantial portion of this data is generated by streaming providers like Netflix, Hulu, and Amazon Prime Video. The content they provide, with some exceptions, has been created by expert photographers using professional capture devices. These pristine videos are generally (but not always) of high quality, and can usually be used as references in subsequent Video Quality Assessment (VQA)/compression processes. Because of the availability of high quality reference videos, perceptual rate distortion optimization (RDO) is often used to optimize the compression process, which is widely adopted by industry, such as Netflix [2]. However, another category of videos are also uploaded and downloaded in gigantic volumes by casual users, called User-Generated-Content (UGC) videos. These imperfect videos are very often uploaded onto social platforms like YouTube, Snapchat, Facebook, Instagram, and TikTok. UGC is commonly captured by a wide variety of consumers, ranging from skilled professionals to inexpert users having uncertain techniques and often unsteady hands, resulting in numerous, often commingled impairments of perceived quality. These UGC videos have often undergone a series of processing steps, such as editing, aesthetic modification, and compression, before the user uploads them to an online server, where they inevitably undergo *another* round of compression. The highly diverse mixtures and severities of distortion that these UGC videos contain are very difficult to model. Because of the absence of high quality reference videos, traditional RDO protocols that use reference VQA models as perceptual metrics are less reliable. Since high quality pristine reference videos cannot be counted on to guide compression, it is highly desirable to find ways of assisting compression decisions by accounting for the qualities of the source videos. This challenging problem presents unique difficulties in perceptual distortion modeling.

We remind the reader that VQA models can be conveniently placed into two categories. Those that require a reference signal, which includes Full-Reference (FR) and Reduced-Reference (RR) models, and those that do not: No-Reference (NR) models. Current mainstream reference VQA models include the FR VMAF [3], VQM [4], SSIM [5] and VIF [6], and the RR model ST-RRED [7]. The second main category is that of NR models, which aim to accurately predict picture quality without the aid of any reference videos. In applications involving UGC videos, reference models (FR or RR) are problematic, since comparing a possibly distorted test video

Manuscript received July 30, 2020; revised March 26, 2021 and July 9, 2021; accepted August 2, 2021. Date of publication August 30, 2021; date of current version September 3, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo. *(Corresponding author: Xiangxu Yu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB), The University of Texas at Austin, TX, USA under Protocol No. 2007-11-0066.

Xiangxu Yu and Alan C. Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: yuxiangxu@utexas.edu; bovik@ece.utexas.edu).

Neil Birkbeck, Yilin Wang, and Balu Adsumilli are with YouTube Media Algorithms Team, Google Inc., Mountain View, CA 94043 USA (e-mail: birkbeck@google.com; yilin@google.com; badsumilli@google.com).

Christos G. Bampis is with Netflix Inc., Los Gatos, CA 95032 USA (e-mail: cbampis@gmail.com).

Digital Object Identifier 10.1109/TIP.2021.3107213

against a reference that is also distorted must lead to errors, and possibly severe losses of quality prediction accuracy. While NR models are intended to be able to predict the quality of distorted-then-compressed videos, NR VQA remains a challenging research topic and the most competitive models still have difficulty in predicting the quality of complex distorted UGC videos [8].

Here we seek to advance progress on solving the problem of predicting the quality of distorted-then-compressed videos. For example, in a cloud server, many distorted and already-compressed videos will be uploaded by users, only to be further compressed without the benefit of a guiding pristine reference. Instead, the source video may already contain any of many kinds of possibly mixed distortions, including almost-inevitable prior compression. Thus, standard reference VQA models cannot be expected to produce optimal quality predictions. To address this issue, we propose a method called 1stepVQA, which learns to predict compressed UGC video quality, including relationships between native distorted UGC video quality and that of further video compression. 1stepVQA accomplishes this by monitoring losses of expected statistical regularity in videos being analyzed. Since distortions can cause the statistical properties of videos to deviate from well-modeled natural regularities, it is possible to learn how distortions affect "natural video statistics" and to use what is learned to predict perceptual quality. By a 'natural video,' we mean any high-quality video captured by an optical camera. 'Natural video statistics' refers to a class of statistical models that have been shown to accurately describe natural pictures and videos that have been subjected to perceptually relevant bandpass and nonlinear normalization processes [7], [9]. We use them to measure deviations between statistical models of UGC videos that have been subjected to compression, and to predict their perceptual quality. In the approach taken here, we seek to avoid the drawbacks of reference VQA models by also making measurements of quality-aware features on the typically flawed reference videos.

The essential tools needed to solve VQA problems are databases of appropriate diverse video contents and distortions, labeled with adequate amounts of subjective data. There are a number of legacy VQA databases, including LIVE VQA [10], LIVE Mobile [11], CDVL [12] and MCL-V [13], MCL-JCV [14] and VideoSet [15], each containing about 10-20 pristine video contents, and many distorted versions of them. The distortions in all of these databases were synthetically generated in isolation. While these databases have successfully driven the development of early reference VQA models, they are of limited value with respect to the UGC video quality assessment problem. More recently, a few novel databases have been developed, including LIVE VQC [8], KoNViD-1k [16] and YouTube UGC [17], which contain many UGC videos of very diverse contents and distortions. These databases were originally designed for the development of NR VQA models, but we will also find them useful here. The source videos were collected from diverse users and online repositories. These databases are not of direct value, and so it was necessary for us to generate a new and dedicated subject VQA database, containing UGC videos of highly diverse real-world content and distortions as source "reference," as well as additionally compressed versions of them, to help us develop and evaluate our model.

The rest of the paper is organized as follows: we introduce related work in Section II. We present our newly built database in Section IV. We describe our proposed model in Section V. We examine our model performance in Section VI. Finally we summarize the paper in Section VII.

II. Related Work

Early work on VQA models largely focused on reference-based models, beginning with the simple PSNR and advancing to more sophisticated FR frame-based models that better account for visual perception, including SSIM, MS-SSIM [18], VSNR [19], FSIM [20], and DOG-SSIM [21], and spatio-temporal models like VMAF and ST-MAD [22]. There are also successful RR models, such as ST-RRED and SpEED-QA [23].

The development of NR models which do not require any reference videos has been a hot topic in recent years, driven by the need to evaluate UGC videos and older, lower-quality content. Early models were often distortion-specific, targeting only one or more specific distortions. Later, more general approaches have involved training quality prediction models to learn mappings from features to subjective judgments, e.g., BRISQUE [24] and V-BLIINDS [25], which make use of natural scene statistics (NSS) models; TLVQM [26], which deploys a simplified motion estimator, and more recent deep learning models. VSFA [27] and its enhanced version [28] are state-of-art deep learning VQA models which claim superior performance on publicly available databases. PVQ [29] is a recent published local-to-global region-based NR VQA architecture. There are other deep learning approaches including NIMA [30], PaQ-2-PiQ [31], PQR [32], DLIQA [33], [34]. There are also completely blind (unsupervised) models, like NIQE [35] and IL-NIQE [36].

The approach that we take to the problem of assessing the quality of distorted videos undergoing compression relies on spatial-temporal NSS models. NSS models are the foundation of many FR, RR and NR models. In particular, "distorted NSS" models seek to quantify losses of statistical regularity in visual signals.

In these approaches, bandpass picture or video coefficients are modeled by Gaussian Scale Mixture (GSM) distributions. These models typically operate in a bandpass (wavelet, DCT scale-space, etc.) domain. In the absence of distortion, perceptually relevant conditioning or divisive normalization nonlinearities tend to Gaussianize and further reduce the redundancy of these coefficients. Departures from uncorrelated Gaussianity of the coefficients arising from distortion are measured or learned by Image Quality Assessment (IQA)/VQA models, and mapped to perception.

One of the most commonly used spatial-NSS models, which is closely related to methods used here, is BRISQUE. When applied on videos, BRISQUE operates by computing Mean Subtracted Contrast Normalized (MSCN) coefficients on a frame basis. The statistics of MSCN of high quality video

frames tend to follow a Gaussian distribution, but this characteristic is generally disturbed by the presence of distortions. The MSCN of distorted video frames are commonly modeled as following a parametric Generalized Gaussian Distribution (GGD), the parameters of which tend to vary with distortion type and severity. During training, MSCN features are computed on a set of distorted videos, from which model parameters are extracted and then fed into a machine learning tool along with human quality labels, to learn mappings from model space to perceived quality. The trained model can then be used to blindly predict the quality of videos.

In [37], the authors devised a spatial-temporal VQA model, which utilizes the statistics of directional frame differences. These are defined as differences between adjacent frames that have been spatially displaced (by ± 1 pixel) relative to one another, thereby capturing directional space-time bandpass behavior that exhibits high statistical regularity. We build on this model in defining the quality-aware feature set in 1stepVQA.

The new model we introduce here, called 1stepVQA, utilizes similar underlying models as BRISQUE, but also employs temporal features. Moreover, we created a new, special-purpose subjective database to enable a unique training process, whereby two types of human subjective labels are trained on: (i) pre-compressed quality, and (ii) post-compressed quality. In this way, the 1stepVQA model is able to learn to predict post-compression video quality, while *also* accounting for pre-compressed quality.

Previously, we proposed an IQA model [38] to solve the same problem (predicting the quality of still pictures with pre-existing distortions that are then JPEG compressed). Unlike 1stepVQA, this prior approach, called 2stepQA, operates in two distinct, sequential, quality assessment stages [39]. The first stage conducts traditional NR IQA (e.g., using NIQE [35]) on the source picture, to form a prediction of the pre-existing picture quality. In the second stage of 2stepQA, a standard reference IQA model (e.g., SSIM [5]) is used to predict the post-compressed quality *relative* to the source picture. These quality predictions are then combined, using, for example, a simple weighted product of FR/RR and NR stages. Instead, 1stepVQA attempts to predict the quality of compressed videos by processing both the NR features from the reference video and FR features on both reference and compressed videos.

The contributions that we make are summarized as follows:

- We created a first-of-a-kind subjective VQA database, by first collecting 55 UGC videos that were selected from the existing LIVE VQC database, along with human quality opinions of them from that database. To avoid the usual process of hand-selecting compression levels as has historically been done in the creating of VQA databases, which introduces human bias, we instead implemented a highly realistic perceptual RDO protocol using VMAF based Pareto optimization, to create multiple compressed versions of each video, on which we also collected a large number of human subjective opinions.
- We used this psychometric resource to learn a new space-time NSS feature-based VQA model, called 1stepVQA, which utilizes both traditional spatial NSS features, as well as new displaced space-time difference NSS features. This new model does not require motion estimation, or transformation to another domain. We show that using displaced frame differences, rather than non-displaced differences, boosts prediction accuracy. 1stepVQA is relatively fast, uses fewer features than most other VQA models, yet it generates highly competitive results against other top performing models on this important practical problem. We show that a combination of 1stepVQA with a reference module leads to a further enhanced version of 1stepVQA, dubbed 1stepVQA-R, which provides more accurate prediction performance.
- We show that although reference VQA models are able to guide traditional encoding optimization schemes when there are available high quality pristine reference videos, they fail to perceptually optimize the compression of UGC videos with pre-existing distortions. However, the new 1stepVQA model is able to handle these situations and can be used to improve the performance of UGC encoding optimization processes.

We are making the source code of the 1stepVQA model publicly available at https://github.com/xiangxuyu/1stepVQA, and we are also providing the dedicated new database free of charge at https://live.ece.utexas.edu/research/onestep/index.html.

III. Predicting the Quality of Distorted-Then-Compressed Videos

Traditional FR or RR models require access to a pristine reference video which has no visible distortions and is of very high quality. They predict the quality of a distorted video by comparing computed perceptual differences between it and its pristine reference. Thus, traditional FR or RR models provide a *relative* quality score which indicates the degree of perceptual deviation of the distorted video from its pristine reference. If the reference video is of very high quality, then the resulting quality score may be regarded as a reliable *absolute* quality score. However, if the reference video is distorted, then any reference measurement loses meaning.

We exemplify the unreliability of FR/RR VQA models in this context in Fig. 1. Given a pristine video I_h of high quality, an FR or RR model (e.g., SSIM or VMAF) can be used to predict the quality of its compressed version I'_h by measuring perceptual deviations between them. However, if the quality of the reference video is degraded, such as the video I_l, then FR/RR VQA models become unreliable. As shown in Fig. 1, the ranking of the four videos in quality (Mean Opinion Score, or MOS) from high to low is I_h, I'_h, I_l, and I'_l. The MOS values indicate that the perceptual quality of the video I'_h, which is the compressed version of the high quality reference I_h, is significantly better than that of the low quality reference I_l and its compressed version I'_l. However, both the Difference Mean Opinion Scores (DMOS) and the VMAF scores *reverse* the (predicted) quality order! The DMOS of I'_l is less than that of I'_h, indicating that the former video is of better quality than the latter. The VMAF score of I'_l is greater than that of

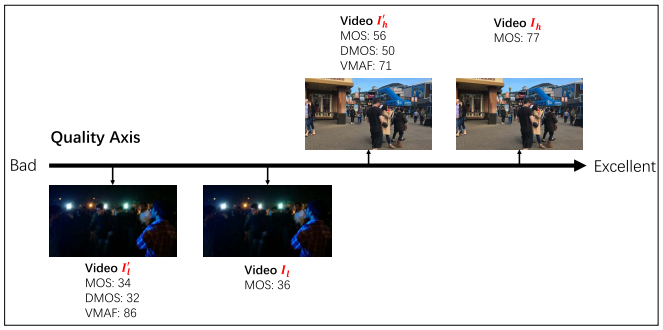


Fig. 1. An example of two reference videos, one of high quality (I_h) and the other of low quality (I_l), along with compressed versions of them (I'_h and I'_l, respectively) from the new LIVE Wild Compressed Video Database. The quality axis spans low to high quality from left to right.

I'_h, resulting in the same conclusion with respect to DMOS. In this case, both DMOS and VMAF fail to indicate the level of subjective quality of two compressed videos, since they are only able to make *relative* quality predictions, and they do not at all account for the quality of the reference videos [38].

Motivated by these observations, one might instead attempt to apply NR VQA models to simply predict the absolute quality of the distorted-then-compressed videos using an NR model. Unfortunately, the performances of current existing NR VQA models are quite limited. Towards making progress on this problem, we propose a new 1stepVQA model, which, unlike traditional FR/RR VQA models, also makes use of information descriptive of the intrinsic quality of the reference videos, thereby improving prediction accuracy. Section VI-B further explains and illustrates why FR/RR VQA models fail in this scenario, and why traditional VQA performance metrics may not reflect this.

IV. New Database

In recent years, the number of UGC videos that are uploaded daily has increased to an incredible degree. Most of these contributors are amateur videographers having limited skills and uncertain hands, hence, the qualities of these video uploads varies significantly over a very wide range. These UGC videos are then uploaded onto Internet servers, and pass through multiple processing stages, before being streamed to potentially millions of clients. These uploaded UGC videos will often be corrupted by any of many possible distortions (exposure, shake, multiple types of noise or blur and many more). The stages of processing may introduce further defects, including any additional video coding. For example, when a video is recorded, it may be compressed within a device before being uploaded. After being uploaded to a server, the video may be further compressed for storage and transmission.

Under this scenario, the uploaded video may already suffer from mixed in-capture distortions, against which there are no reference videos of 'pristine' quality, hence FR/RR VQA models may fail if used to predict the quality of the pre-distorted/multiply-compressed video. While it would be desirable to directly apply an NR VQA model, to the ultimate compressed output, current NR algorithms are insufficiently general to effectively conduct this very complex task. There is also a lack of dedicated databases designed to model this scenario. There appears to be a database that may relate to this problem [40], but it is not publicly available. Towards filling this gap, we created a new database, called the LIVE Wild Compressed Video Quality Database, which contains hundreds of compressed UGC videos, including the source (reference) videos.

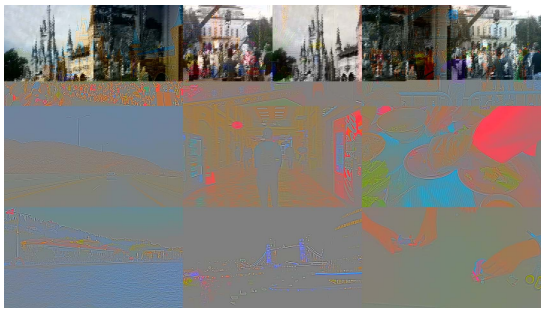


Fig. 2. Example reference videos from the new LIVE Wild Compressed Video Database.

A. Database Content

We randomly selected 55 different reference videos (contents) from among the 110 1080p videos contained in the LIVE VQC Database, each of duration 10 seconds. These are all UGC videos captured with highly diverse mobile cameras, covering a wide range of contents and qualities. Most of these videos are corrupted by diverse authentic, mixed in-capture distortions.

Since the LIVE VQC database provides subjective scores, specifically MOS, we randomly sampled the 55 reference videos to match the MOS distribution of the 110 source videos. In this way, we ensured that the set of reference videos spanned a wide quality range. Fig. 2 shows a few example reference videos from the new database.

B. Compressed Video Generation

We adopted the 'Per-Title Encode Optimization' [2] method introduced by Netflix, which is based on VMAF and deployed globally, to create and select compressed videos. Each reference video was subjected to two processing steps to simulate stages that may occur when, for example, a video is uploaded to a social media site. First, each video was spatially down-scaled using bicubic interpolation (if needed) to four resolutions: 1080p, 720p, 540p, and 360p. Second, we generated these compressed videos at each down-scaled resolution using H.264 at 17 compression levels using CRF: $1, 4, 7, \ldots, 49$. Thus, for each video content, there is one 1080p reference video and 68 altered versions of it (four resolutions) x (17 compression levels). This yielded a total of 3740 videos with downsample/compression distortions, which is far too many to all be viewed by the limited number of subjects that typically subscribe to a psychometric quality study in a laboratory. Therefore, we adopted a strategy similar to that used by video service providers to determine bitrate allocation vs. predicted quality. To do this, we calculated the VMAF scores of all the videos along with their bitrates, and used these

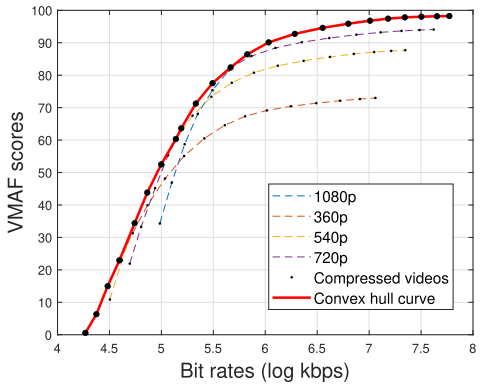


Fig. 3. Convex hull curve used to select compressed videos of an examplar video content.

to plot the convex hull curve of each video content. Based on this curve, we selected four videos on the curve and included them in the database. An example of the convex hull curve of one video content is shown in Fig. 3. We chose VMAF scores to draw the convex hull curve because it is a state-of-art FR VQA model that is already used in this way in the streaming industry [2], and its scores map approximately linearly against perceptual quality, which is convenient for video selection and distinction. To create downscaled/compressed versions of each content that are perceptually separable from each other, we selected four compressed videos on the curve having VMAF adjacent score differences approximately between 10 and 20. The compressed videos at the four compression levels have VMAF scores within the ranges [80,90] (compression level 1), [60,70] (compression level 2), [40,50] (compression level 3), and at the heaviest [20,30] (compression level 4), respectively. We applied the same process on all contents, yielding four perceptually different downscaled/compressed versions of each, or 220 intentionally impaired UGC videos overall, in addition to the original 55 (275 total). Since the convex hull curve of each content is unique, likewise the parameter combinations selected for each content are as well. During the human study, the compressed videos were upscaled to 1080p before being displayed on the monitor.

Although using the convex hull curve to (Pareto) optimize compression is highly effective (given pristine high quality reference videos), it is not reliable if the reference videos are not of high quality. For example, in Fig. 1, the lightly compressed video I_l' receives a high VMAF score, even though both it and its reference video I_l are both of very low quality. Indeed, using VMAF to conduct perceptual RDO might lead to even higher compression (and good VMAF scores), but even lower perceptual quality.

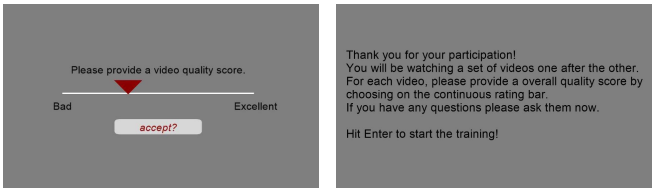
It is important to note that using VMAF to create the compressed content introduces a *database bias* so that VMAF (or models using the same features as VMAF, like VIF [6] and ST-RRED [7]), unfairly benefit from when applied on the new database, and indeed, this bias has been observed. Hence VMAF is separately analyzed in the later comparisons.

We adopted the H.264 video compression format in our experiments. While more recent formats, such as HEVC, VP9 and AVI exist, H.264 remains the most widely used compression format. Moreover, H.264 is largely representative of modem compression standards, which are sophistications of the classic block DCT hybrid codec. In any case, the framework taught here is extensible to arbitrary compression schemes. We used the ffmpeg version 4.0.2 libx264 encoder to apply H.264 compression. The compression speed was set to slow mode, and the output video type was yuv420p.

C. Human Study

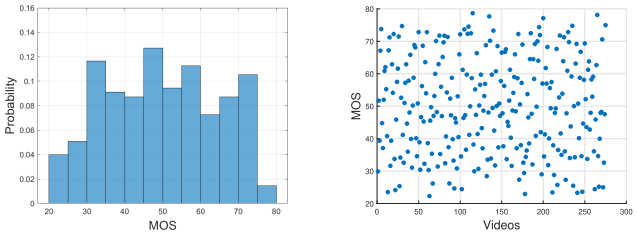
In order to obtain subjective quality labels on each video (for model training, testing, and comparison), we conducted a human study in the LIVE subjective study laboratory. The participating subjects were mostly UT-Austin students without a background in video quality. Each subject completed two test sessions of duration about 45 mins, separated by at least 24 hours. The database of 275 videos was divided randomly into two parts in each session, one containing 27 contents and another containing 28 contents, including both reference videos and their respective four downscaled/compressed versions, hence each subject viewed 135 and 140 videos in consecutive sessions on different days. We adopted a single stimulus continuous-scale quality evaluation protocol, as detailed in [41] and used in [10], [11], [38], [42]. The videos were played in a randomized order with each video shown only once during each session, and where different distorted versions of each unique content were separated by at least 5 videos. The source videos were included as references. The total number of subjects that took part in the study was 40, and all of them successfully finished both sessions.

At the start of the first session, each subject participated in a visual acuity (Snellen) test, and were asked whether they had any uncorrected visual deficiency. A viewing distance of approximately two feet was maintained during the test. The video sequences were displayed on an HP VH240a 23.8-inch 1080p monitor at their native 1920 x 1080 resolution, and the subjects were asked to adjust the height and angle of the monitor to find their best position. Before starting the experiment, each subject was required to read and sign a consent form including general information about the human study, then the procedures and requirements of the test were explained. A short training session was also presented at the beginning of the first session, using a different set of videos than in the test experiment, to help the subjects become familiar with the procedures. After watching each video, each subject was asked to provide an overall opinion score of the video's quality by dragging a slider along a continuous rating bar. As shown in Fig. 4(a), the quality range was labeled from low to high with two adjectives: Bad and Excellent. The subjective scores obtained from the subjects were converted to numerical quality scores in [0, 100]. We believe this is superior to the ITU-R Absolute Category Rating scale, which uses a five-category quality scale, since using a continuous scale can expand the range of scores while providing richer, denser data to train learning based models [10], [11], [38], [42]. A screenshot of the subjective study instruction interface is shown in Fig. 4(b). The interface was developed on a Windows PC using the PsychoPy software [43].



(a) (b)

Fig. 4. (a) Screenshot of the rating bar shown to the subjects. (b) Screenshot of the instruction interface shown to the subjects.



(a) (b)

Fig. 5. (a) MOS distribution across the entire LIVE Wild Compressed Video Quality Database. (b) Scatter plot of the MOS obtained on all the videos in the database.

As described above, each subject was required to complete a short training session before participating in the first test session. The subject was given enough time to read instructions which explained the purpose and procedures of the study. The subject was allowed to ask questions after reading the instructions, which were answered by one of the researchers conducting the study. Then, the subject was instructed to open the study interface, and begin the training session. There were six training videos in total, broadly covering the quality range, allowing the subject to obtain a sense of the possible video qualities they would encounter during the test session. A researcher was present during the training session to ensure the subject understood the procedures. During training, the subject was encouraged to rate the quality of the videos, rather than aesthetic aspects. Each subject was also encouraged to utilize the full range and continuity of the rating bar, which was labeled with Likert markings only at the end points, to avoid biases caused by intermediate markings.

D. Data Processing

The subjective MOS were then computed following the procedures described below. The collected raw scores were first converted into Z-scores. Let s_{ijk} denote the score rated by the i-th subject on the j-th video in the session $k = \{1, 2\}$. The Z-scores were computed as follows:

$$z_{ijk} = \frac{s_{ijk} - \bar{s}_{ik}}{\sigma_{ik}}, \tag{1}$$

where $\bar{s}_{ik}$ is the mean of the collected raw scores over all videos assessed by subject i in session k, and σ_{ik} is the standard deviation. Conversion to z-scores removes subject biases in the form of differences in the location and range of values used by the subjects. Details of data processing such as z-score computation can be found in [10].

While we applied the subject rejection procedure described in [41], [44] to remove outlier subjects, we found that it was not necessary. First, the criteria in [41] was designed for high-quality television studies, rather than analysis UGC video. Human subjects' expectations of, and perception of UGC video quality are naturally much broader than of television. Moreover, we also found the inter-group correlations with and without rejection to be nearly identical, as were the correlations against the various tested objective models. Subject scores of UGC content tend to be noisy, and hence include more outliers that occur by chance, rather than by malintent. Indeed, the largest outlier ratio was only 8.36%, from one of the subjects. We are elaborating this point since we believe it is worth re-thinking the use of some of the traditional data-cleaning techniques when analyzing UGC content. Indeed, it may result in an excess of data smoothing, reduced relevance to reality, and a loss of valuable data. While 10 of the 40 subjects were marginally rejected, we elected to retain all of the data, while noting that this changed the results negligibly. Of course, with all subjects scores available, users of the database may set their own rejection criteria.

The Z-scores of the 40 subjects were subsequently linearly rescaled to [0, 100]. Finally, the MOS were obtained by computing the mean of the rescaled Z-scores of each video.

E. Analysis

We first conducted a consistency test. We randomly split the subjective ratings gathered on each video into two disjoint equal groups, and computed and compared the MOS on each video. We repeated the random splits 1000 times and the median SROCC between the two groups was **0.9849**, which is excellent. We also compared the correlation of the MOS of the 55 reference videos used in our new study, against the previously obtained (crowdsourced) MOS on the same 55 videos from the LIVE VQC database, and found SROCC = **0.8390**. The SROCC results show that our data are reliable.

The overall MOS distribution and a scatter plot of the MOS of the LIVE Wild Compressed Video Quality Database are plotted in Fig. 5. Fig. 6(a) is a box plot of the MOS of the reference videos and the four compression levels (but across downscaling). The MOS decreases as the compression is increased. Box plots indicate points as outliers if they are greater than $q_3 + w \times (q_3 - q_1)$ or less than $q_1 - w \times (q_3 - q_1)$, where w is the maximum whisker length 1.5, and $q1$ and $q3$ are the 25th and 75th percentiles of the sample data, respectively. The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier. A higher index indicates heavier compression. Fig. 6(b) plots the MOS across all contents, each color coded at a different compression level. While the curves are nicely separated by content, it is important to observe the mixing of MOS across contents, which is largely caused by the reference distortions, rather than the applied downscaling/compression. Fig. 6(c) shows the 95% confidence intervals of MOS for the five different compression levels (including uncompressed reference). As the compression was increased, the confidence intervals became narrower, suggesting that the perceived qualities of the videos became slightly more concentrated within a narrower range.

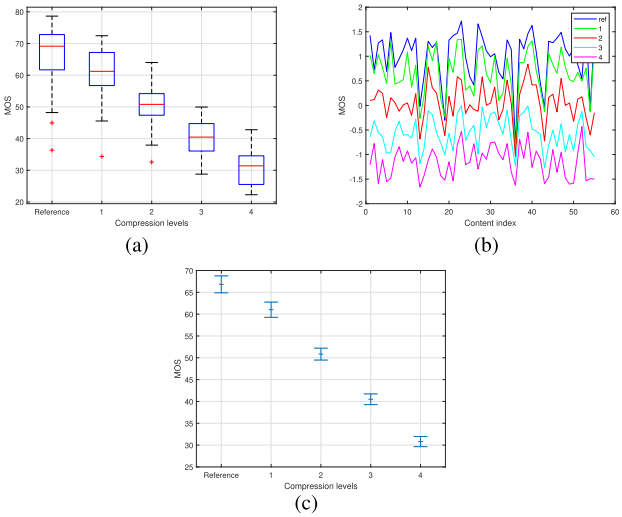


Fig. 6. (a) Box plot of MOS of videos in the LIVE Wild Compressed Video Quality Database for different compression levels. The central red mark represents the median, while the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, while outliers are plotted individually using the '+' symbol. (b) MOS of all contents for four different compression (distortion) levels and reference videos, coded by color. (c) 95% confidence intervals of MOS of videos for five different compression levels (including reference videos).

V. 1stepVQA Model

Research on the relationship between visual perception and the statistics of natural images has aroused widespread attention [45]. These models become well-behaved when expressed in bandpass domain, e.g., wavelets, DCT and Gabor transforms, etc [46]. The 1stepVQA model utilizes an extended set of video NSS models, that capture directional bandpass space-time attributes and uses them to predict quality. In addition, we developed an extended version of 1stepVQA, dubbed 1stepVQA-R, which combines the 1stepVQA features with a reference module. An overview of the 1stepVQA and 1stepVQA-R models is shown in Fig. 7. We discuss the spatial features and spatial-temporal features next.

A. Natural Frame Statistics

Here forward to distinguish them from more general spatial-temporal NSS models, we shall refer to spatial (frame) NSS as Natural Frame Statistics (NFS). The application of NFS models in 1stepVQA is quite similar to BRISQUE [24]. To capture quality-aware NFS features, apply a local non-linear bandpass operation on the luminance component of each video frame, in the form of local mean subtraction followed by divisive normalization. Given a video having T luminance frames $I_1, I_2 \ldots, I_t \ldots, I_T$, define spatial coordinates (i, j), where $i \in 1, 2 \ldots M$, $j \in 1, 2 \ldots N$ are spatial indices of I_t, then the mean subtracted contrast normalized coefficients (MSCN) $\hat{I}_t$ are:

$$\hat{I}_t(i,j) = \frac{I_t(i,j) - \mu_t(i,j)}{\sigma_t(i,j) + C} \tag{2}$$

where

$$\mu_t(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} I_t(i-k, j-l) \tag{3}$$

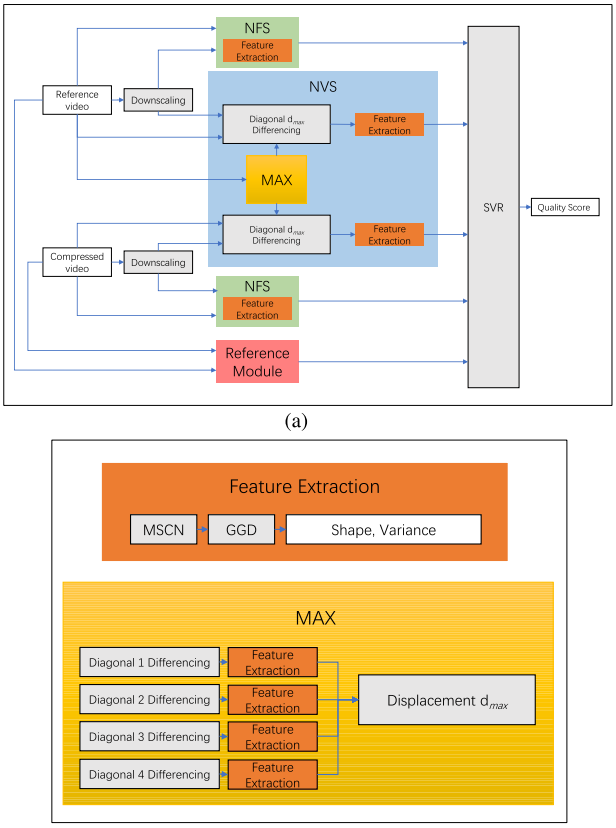


Fig. 7. (a) Flow diagram of 1stepVQA and 1stepVQA-R. (b) Feature Extraction and MAX processing flows in 1stepVQA.

and

$$\sigma_t(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} \big(I_t(i-k, j-l) - \mu_t(i,j)\big)^2} \tag{4}$$

where $C = 1$ is a stability constant, and $w = \{w_{k,l} | k = -K, \ldots, K, l = -L, \ldots, L\}$ is a 2D circularly-symmetric Gaussian weighting function with $K = L = 3$.

If there are spatial distortions present, then the statistical distribution of frame MSCN coefficients tend to become predictably altered [24]. For example, Fig. 8 plots the GGD model of the MSCN coefficients of a frame of a UGC source video, along with various compressed versions of it.

Following [24], [35], we use a Generalized Gaussian Distribution (GGD) model of the MSCN coefficients. A GGD with zero mean is given by:

$$f(x; \alpha, \sigma^2) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} exp(-(\frac{|x|}{\beta})^{\alpha}) \tag{5}$$

where

$$\beta = \sigma \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}} \tag{6}$$

and $\Gamma(\cdot)$ is the gamma function:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt \quad a > 0. \tag{7}$$

There are two GGD parameters: a shape parameter α which controls the 'shape' of the distribution, and a variance

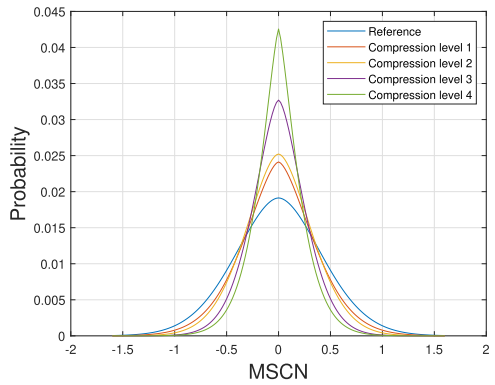


Fig. 8. GGD models of MSCN coefficients of a frame of a UGC source video and various compressed versions of it. A higher index indicates heavier compression.

parameter which is controlled and represented by σ. These are estimated using the popular moment-matching approach in [47]. We include both parameters as part of the feature set.

B. Natural Video Statistics

Here we refer to the spatial-temporal NSS of videos as Natural Video Statistics (NVS), which are also significantly affected by distortions, such as jitter, ghosting, and motion compensation mismatches, as well as compression and transmission artifacts. Moreover, UGC videos are often afflicted by mixed in-capture distortions, such as camera shake, under- or over-exposure, sensor noise, and color distortions, which cannot be easily modeled.

While it is difficult to find regularities in the statistics of motion, there are strong regularities of temporal bandpass videos, such as frame difference signals [7]. However, our model goes further, and also utilizes models of the statistics of displaced frame differences. Given two adjacent frames, perform a spatial translation of one of them, then compute the displaced frame difference between the two frames. By doing so, we seek to capture space-time statistics of videos without computing motion. As it turns out, bandpass processing of displaced frame differences are also very regular, and predictive of quality. For simplicity, we used four diagonal directions to compute displaced frame differences. Given a video with T luminance frames $I_1, I_2 \ldots, I_t \ldots, I_T$ and spatial coordinates (i, j), $i \in 1, 2 \ldots M$, $j \in 1, 2 \ldots N$, the four diagonal displaced frame differences between each pair of adjacent shifted frames are defined and depicted in Fig. 9. The four displaced frame differences are:

$$D_{t1}(i, j) = I_t(i, j) - I_{t+1}(i - 1, j - 1) \quad (8)$$

$$D_{t2}(i, j) = I_t(i, j) - I_{t+1}(i + 1, j - 1) \quad (9)$$

$$D_{t3}(i, j) = I_t(i, j) - I_{t+1}(i - 1, j + 1) \quad (10)$$

$$D_{t4}(i, j) = I_t(i, j) - I_{t+1}(i + 1, j + 1) \quad (11)$$

The reason that we capture NVS in opposite directions is because they provide complementary, and not redundant distortion information relative to the direction of any motion field. One of two opposite directions will generally be more in the direction of local motion, and the other less so, presenting

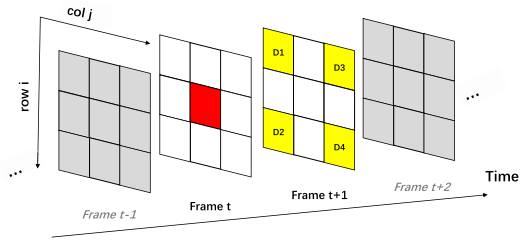


Fig. 9. Depiction of a pixel in frame t and the non-displaced and four displaced pixels in frame $t + 1$ it is differenced with.

different (more vs. less) statistical regularity in the corresponding frame differences, which can be affected by distortion. Then compute the MSCN coefficients of each of the four diagonal displaced frame differences, $\hat{D}_{tk}(i, j)$, $k = 1, \ldots, 4$, using (2). The corresponding values of $\mu_{tk}(i, j)$ and $\sigma_{tk}(i, j)$ are computed in the same way as in (3) and (4). As with NFS, the GGD model (5) - (7) is a good fit to the spatial-temporal MSCN histograms. Thus the histograms of each directional difference $\hat{D}_{tk}(i, j)$ also nicely fit the GGD model, each fit yielding two GGD parameters.

Motion occurs along diverse different directions and over different time durations, hence, the corresponding displaced frame differences are statistically different, each expressive of possible (or lacking) motion information. We assume that the true local motion will best match one of the displacement directions. In separate work, we have shown that the MSCN coefficients of frame differences displaced in the direction of motion strongly tend towards Gaussianity [48]. Hence, we find the displacement direction where the largest GGD shape value is obtained among the MSCN coefficients of the four displaced frame differences, and only make use of the shape and variance parameters obtained from the frame difference in this direction. We will refer to this as the 'MAX' operation. We apply this process on every frame of the reference video to obtain four averaged shape features α_k, $k \in 1, 2, 3, 4$ from the four displaced differences. However, we only utilize the displacement d_{max} where the shape value is the largest:

$$d_{max} = argmax_{k \in \{1,2,3,4\}} (\alpha_k) \quad (12)$$

We thus compute the frame differences in the displacement direction corresponding to d_{max} on both the reference and compressed videos, and collect the maximum shape parameter along with the associated variance parameters as features of the 1stepVQA model.

Figs. 10 and 11 show frames from two exemplar UGC source videos, containing high and low degrees of motion, respectively. We will use these two contents to illustrate the results of MSCN processing and the construction of GGD models of displaced frame differences, and to compare these with the MSCN and GGD models of non-displaced frame differences.

Figs. 10(b) and 10(c) plot the GGD models of histograms of MSCN coefficients of non-displaced and displaced frame differences computed on the high motion video, and also of several compressed versions of it, while Figs. 11(b) and 11(c) plot similar GGD models of the same low motion video. While the GGD models of the MSCN histograms of the source

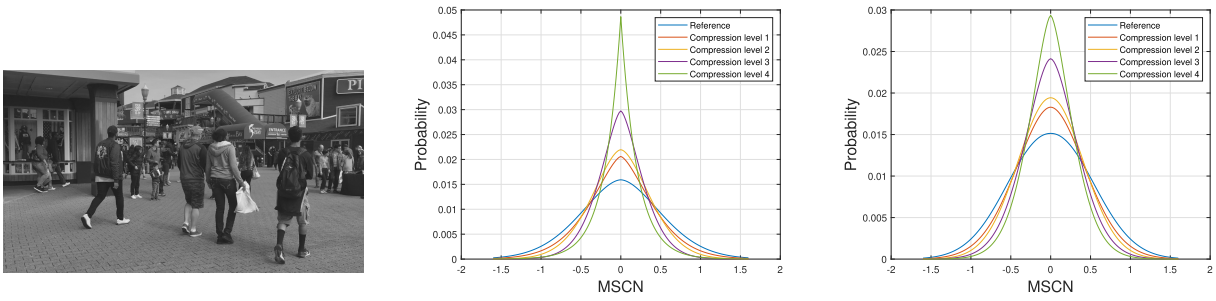


Fig. 10. (a) A luminance frame of a UGC source video containing high motion. (b) Histogram of the MSCN coefficients of the non-displaced frame difference and several compressed versions of it. (c) Histogram of the MSCN coefficients of the displaced frame difference $\hat{D}_{t1}(i, j)$ and compressed versions of it. A higher index indicates heavier compression.

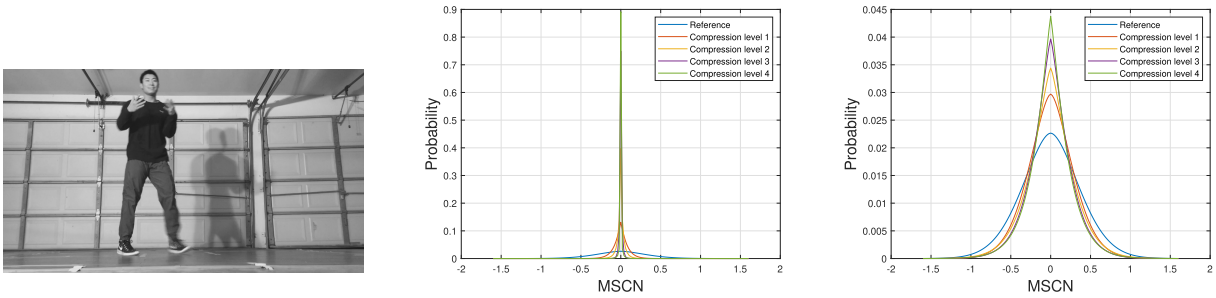


Fig. 11. (a) A luminance frame of a UGC source video containing low motion. (b) Histogram of the MSCN coefficients of the non-displaced frame difference and several compressed versions of it. (c) Histogram of the MSCN coefficients of the displaced frame difference $\hat{D}_{t1}(i, j)$ and compressed versions of it. A higher index indicates heavier compression.

videos follow a Gaussian distribution, the GGD models of the compressed videos deviate towards a heavy-tailed distribution shape, especially at heavier compressions. It is interesting to observe that the GGD models of the displaced frame differences are less peaked than those of the non-displaced frame difference.

However, there are significant differences between the GGD models of non-displaced and displaced MSCN frame-difference signals, those of the former being much more peaked than those of the latter. In other words, displaced MSCN frame differences tend to retain Gaussianity as the compression parameter is varied. Although they remain monotonic and are well separated, they are also more regular (and as well we shall see, more predictable). One reason for this is that in motion regions, where the motion direction is similar to the frame-level displacement, redundancy is better exploited. In directions opposite to motion, displaced MSCN frame differences will produce very marked localized changes, which tend to broaden the MSCN histograms. Even on low-motion or static videos, the displaced frame differences become similar to spatial directional gradient operators, which have been previously observed to enhance bandpass statistical regularity and to enhance quality prediction [49], [50]. However, as shown in Fig. 11, on low-motion video, non-displaced frame differences fail to be predictive, as the corresponding GGD models are not well separated across different compression levels. On heavily compressed videos, a high percentage of the MSCN coefficients of non-displaced frame differences tend towards zero, tending to sharply peaked distributions.

C. Reference VQA Module

In addition to the NFS and NVS features described above, we also combined 1stepVQA with a reference VQA module, forming a more generalized version, which we called 1stepVQA-R. The reference VQA component aims to capture perceptual quality differences, mainly caused by compression artifacts, between a compressed video and a reference video. The quality prediction of compression distortion has been well studied, and many VQA models are able to measure the degree of compression artifacts quite well. We include the predicted quality scores from any reference VQA model as one feature of 1stepVQA-R. We show in Section VI-D that this additional feature boosts the performance of 1stepVQA.

D. Feature Summary and Training

Table I summarizes the features that comprise 1stepVQA. Images and videos are naturally multiscale, and it has been observed and demonstrated that incorporating multiscale information in both space and time (and space-time) enhances quality prediction performance [7], [18], [24]. We extract all of the listed features in Table I, at two scales. Scale 1 is the original video resolution, while scale 2 is the downscaled (by a factor of 2) resolution. This is accomplished by Gaussian pre-filtering before down-sampling, as with BRISQUE. 1stepVQA extracts features from both the source videos and compressed videos, while the displacement d_{max} is computed only from the source videos. While 1stepVQA is a reference VQA model, it is unique since it deploys feature sets designed to capture both NR aspects (of the source video) and FR aspects

TABLE I
SUMMARY OF 1STEPVQA FEATURES

	Scale		Type		Source			
Feature Index	1	2	NFS	NVS	Reference	Compressed	Feature Description	Computation Procedure
$f_1 - f_8$	✓	✓	✓		✓	✓	Shape, variance	Fit GGD to spatial MSCN coefficients
$f_9 - f_{16}$	✓	✓		✓	✓	✓	Shape, variance	Fit GGD to d_{max} displaced difference MSCN coefficients

(of both). A total of 16 1stepVQA features are used, none of which require motion estimation. In 1stepVQA-R version, the predicted quality score of a reference VQA model supplies one more feature.

Feature pooling in 1stepVQA and 1stepVQA-R is also simple: all features are computed on a per frame basis, then averaged across all frames. In our implementation, a Support Vector Machine Regressor (SVR) is used to learn a regression model from the extracted feature space to quality scores. Specifically, we conducted a cross-validation study on 1000 train-tests splits of each model on the new LIVE Wild Compressed Video Quality Database, as detailed in Section VI. Similar approaches are implemented in [24]–[26].

In Section VI-C, we study the performance of 1stepVQA, and compare it against other models and feature combinations.

VI. PERFORMANCE AND ANALYSIS

We used the newly built LIVE Wild Compressed Video Quality Database to evaluate and compare the performance of the 1stepVQA model against other FR, RR and NR VQA models. It is first worth noting that while DMOS labels are typically supplied with existing quality research databases, those have available pristine undistorted videos as references. As discussed in Section III, DMOS only captures the quality degradations of a test video with respect to a reference video, which is a *relative* subjective quality score. However, the ultimate goal of objective VQA models is to predict the *absolute* quality of a tested video as represented by MOS. Although VQA models are often evaluated against DMOS, this is inappropriate unless the reference video is pristine. Also, as discussed in [38], given imperfect references, DMOS is unable to capture absolute video quality, as for example, on UGC reference (original) videos suffering from pre-existing distortions. Thus, when conducting subjective (or objective) VQA on UGC videos, only MOS is appropriate.

We evaluated the relationships between predicted quality scores and MOS using SROCC, Pearson's (linear) correlation coefficient (LCC) and the Root Mean Squared Error (RMSE). To compute LCC and RMSE, the predicted scores are passed through a logistic non-linearity before computing performance. SROCC measures the ranked correlation of the given samples, and does not require any remapping. Larger values of SROCC and LCC indicate better performance, while larger values of RMSE imply worse performance.

The best way to evaluate the generality of training based algorithms is to conduct cross-database training and testing. However, in our situation, there is no similar database that is publicly available. Hence we were only able to test on the new database. To do this, we randomly divided the database into non-overlapping 80% training and 20% test sets, with no overlap of original content between sets. In each iteration, the number of samples in the training and test sets was 176 and 44, respectively. The same randomized splits were repeated over 1000 iterations to avoid biased results, and we report the median values of the results. During each iteration, we conducted a cross-validation test using grid search to find the best parameters C and g of the SVR model. All of the compared VQA models, including 1stepVQA, were trained (if needed) and tested on the new database using the same process as described above, for fair comparison. When testing the models that do not need training, we only report the median values of each such model's results on the 1000 test sets, as with the training based models.

For comparison, we tested against prominent FR and RR models: PSNR, MS-SSIM, FSIM, ST-MAD, VSI, and our previously developed 2stepQA, and NR models: NIQE, BRISQUE, V-BLIINDS, and TLVQM. We also included the results of the modern deep learning NR VQA model VSFA. We made use of the source codes released publicly by the authors of the compared VQA models, and used their default settings, including original code for chroma components, if any.

Since we found that the 70th feature of TLVQM was the same on all of the videos, we excluded it when training the model.

A. Comparisons Against Mainstream VQA Methods

We first computed and compared the performance of 1stepVQA and 1stepVQA-R against several reference and NR VQA models, with the results shown in Table II. The reference VQA component of 1stepVQA-R can be any reference VQA model. We chose FSIM because of its good performance and singular phase feature, and included its scores as one feature in 1stepVQA-R. We also show its performance in Table II. To determine whether there exists significant differences between the performances of the compared models, we conducted a statistical significance test. We used the distributions of the obtained SROCC scores computed over the 1000 random train-test iterations. The non-parametric Wilcoxon Rank Sum Test [51], which compares the rank of two lists of samples, was used to conduct hypothesis testing. The null hypothesis was that the median for the row model was equal to the median of the column model 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 VQA algorithms using SROCC are tabulated in Table V. The significance tests

TABLE II

PERFORMANCES OF 1STEPVQA AGAINST VARIOUS FULL-REFERENCE, REDUCED-REFERENCE AND NO-REFERENCE VQA MODELS ON THE LIVE WILD COMPRESSED VIDEO QUALITY DATABASE USING NON-OVERLAPPING 80% TRAINING AND 20% TEST SETS. ITALICS INDICATE NO-REFERENCE ALGORITHMS. BOLDFACE INDICATES THE BEST PERFORMING MODEL

	PSNR	MS-SSIM	FSIM	ST-MAD	VSI	2stepQA	*NIQE*	*BRISQUE*	*V-BLIINDS*	*TLVQM*	*VSFA*	1stepVQA	**1stepVQA-R**
SROCC	0.5084	0.7856	0.8778	0.8197	0.7813	0.8493	0.7150	0.7877	0.8276	0.8381	0.8519	0.8918	**0.9236**
LCC	0.5074	0.7744	0.8776	0.8141	0.7806	0.8455	0.7149	0.7887	0.8228	0.8303	0.8339	0.8902	**0.9224**
RMSE	11.2782	8.3797	6.3419	7.7239	8.1939	7.1600	9.1012	8.0702	7.8482	7.5766	7.0889	6.0275	**5.1551**

TABLE III

PERFORMANCES OF 1STEPVQA AGAINST VARIOUS FULL-REFERENCE, REDUCED-REFERENCE AND NO-REFERENCE VQA MODELS ON THE LIVE WILD COMPRESSED VIDEO QUALITY DATABASE USING NON-OVERLAPPING 70% TRAINING AND 30% TEST SETS. ITALICS INDICATE NO-REFERENCE ALGORITHMS. BOLDFACE INDICATES THE BEST PERFORMING MODEL

	PSNR	MS-SSIM	FSIM	ST-MAD	VSI	2stepQA	*NIQE*	*BRISQUE*	*V-BLIINDS*	*TLVQM*	*VSFA*	1stepVQA	**1stepVQA-R**
SROCC	0.4887	0.7594	0.8409	0.8148	0.7640	0.8271	0.7078	0.7917	0.7930	0.8166	0.8590	0.8811	**0.9218**
LCC	0.4870	0.7540	0.8521	0.8087	0.7706	0.8284	0.7118	0.7884	0.7920	0.7983	0.8665	0.8779	**0.9190**
RMSE	11.3557	8.6399	6.8834	7.7415	8.2753	7.3767	9.1093	8.2302	8.4115	8.2279	6.8167	6.2683	**5.2089**

TABLE IV

PERFORMANCES OF 1STEPVQA AGAINST VARIOUS FULL-REFERENCE, REDUCED-REFERENCE AND NO-REFERENCE VQA MODELS ON THE LIVE WILD COMPRESSED VIDEO QUALITY DATABASE USING NON-OVERLAPPING 60% TRAINING AND 40% TEST SETS. ITALICS INDICATE NO-REFERENCE ALGORITHMS. BOLDFACE INDICATES THE BEST PERFORMING MODEL

	PSNR	MS-SSIM	FSIM	ST-MAD	VSI	2stepQA	*NIQE*	*BRISQUE*	*V-BLIINDS*	*TLVQM*	*VSFA*	1stepVQA	**1stepVQA-R**
SROCC	0.4862	0.7634	0.8347	0.8124	0.7606	0.8287	0.7046	0.7826	0.7762	0.8085	0.8487	0.8705	**0.9167**
LCC	0.4795	0.7547	0.8494	0.8048	0.7672	0.8303	0.7074	0.7762	0.7769	0.7827	0.8411	0.8673	**0.9130**
RMSE	11.4160	8.6159	6.9338	7.7745	8.3794	7.3153	9.1760	8.4480	8.7190	8.5602	6.8744	6.5475	**5.3804**

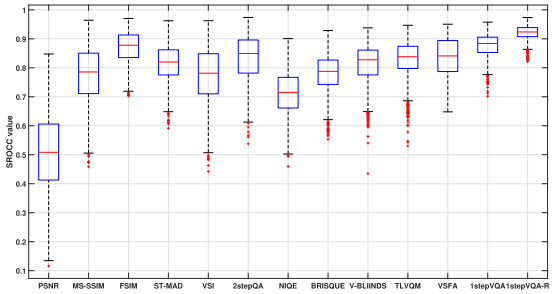


Fig. 12. Box plot of the SROCC distributions of the compared algorithms in Table II over 1000 randomized trials on the LIVE Wild Compressed Video Quality Database.

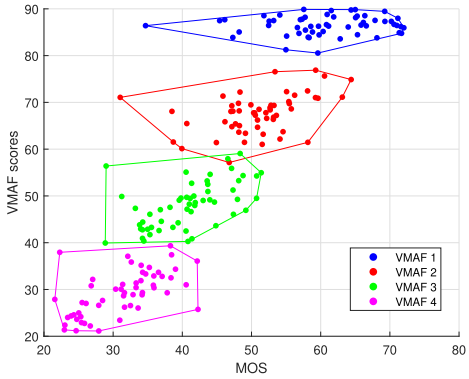


Fig. 13. Scatter plots and convex hulls of VMAF scores and MOS of videos at four compression levels. 'VMAF 1', 'VMAF 2', 'VMAF 3', and 'VMAF 4' are the VMAF scores of videos compressed at levels 1, 2, 3, and 4 respectively. A higher index indicates heavier compression.

show that 1stepVQA significantly outperformed most of the other models. One exception was FSIM, which did nearly as well using an expensive phase measurement model. Thus, we included FSIM as the reference component of 1stepVQA-R. By integrating with FSIM, 1stepVQA-R achieves much better performance as compared with all other models. Fig. 12 shows a box plot of the attained SROCC correlations computed over 1000 iterations, for each of the compared algorithms in Table II. The 1stepVQA and 1stepVQA-R models yielded more reliable results, with higher median SROCC values and lower standard deviations than did the other compared models. The median value of SROCC for 1stepVQA on the training set, over 1000 randomized iterations, was 0.9271, which is not much higher than the performance on the test set (0.8918) as shown in Table II, strongly indicating that the trained model was not overfitted.

We also repeated the same training and test procedures but instead divided the database into 70% training and 30% test sets, and 60% training and 40% test sets, with the results given in Table III and IV respectively, to show the robustness and generalization of the proposed 1stepVQA model. The performance of most models was decreased, but 1stepVQA and 1stepVQA-R were still the best performers among all models. FSIM performed well using 80% training and 20% test set divisions, but its performance decreased in Table III and IV, showing that it is not as robust as 1stepVQA.

B. Limits of Reference Models

As mentioned in Section III, reference VQA models are only able to measure the perceptual fidelity of a compressed video relative to a possibly distorted reference video, but existing reference models do not account for the pre-existing quality of the reference. We use the popular FR VQA model VMAF as an example to show this. As mentioned earlier, the database is biased towards VMAF, since VMAF was used to create a realistic RDO optimization schedule to enhance the authenticity of the dataset. Since the compressed videos were selected using VMAF RDO curves, VMAF is able to perfectly distinguish between the compression levels of each set of the four compressed versions of the same video content

TABLE V

RESULTS OF ONE-SIDED WILCOXON RANK SUM TEST PERFORMED BETWEEN SROCC VALUES OF THE VQA ALGORITHMS COMPARED IN TABLE II. 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. BOLDFACE INDICATES THE BEST PERFORMING MODEL

	PSNR	MS-SSIM	FSIM	ST-MAD	VSI	2stepQA	*NIQE*	*BRISQUE*	*V-BLIINDS*	*TLVQM*	*VSFA*	1stepVQA	**1stepVQA-R**
PSNR	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
MS-SSIM	1	0	-1	-1	0	-1	1	0	-1	-1	-1	-1	-1
FSIM	1	1	0	1	1	1	1	1	1	1	1	-1	-1
ST-MAD	1	1	-1	0	1	-1	1	1	0	-1	-1	-1	-1
VSI	1	0	-1	-1	0	-1	1	0	-1	-1	-1	-1	-1
2stepQA	1	1	-1	1	1	0	1	1	1	1	0	-1	-1
NIQE	1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1
BRISQUE	1	0	-1	-1	0	-1	1	0	-1	-1	-1	-1	-1
V-BLIINDS	1	1	-1	0	1	-1	1	1	0	-1	-1	-1	-1
TLVQM	1	1	-1	1	1	-1	1	1	1	0	-1	-1	-1
VSFA	1	1	-1	1	1	0	1	1	1	1	0	-1	-1
1stepVQA	1	1	1	1	1	1	1	1	1	1	1	0	-1
1stepVQA-R	1	1	1	1	1	1	1	1	1	1	1	1	0

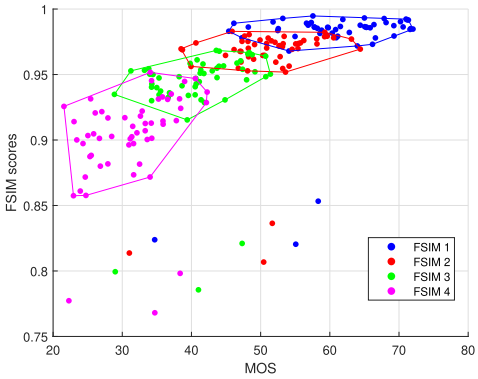


Fig. 14. Scatter plots and convex hulls of FSIM scores and MOS of videos at four compression levels. 'FSIM 1', 'FSIM 2', 'FSIM 3', and 'FSIM 4' are the FSIM scores of videos compressed at levels 1, 2, 3, and 4 respectively. A higher index indicates heavier compression. The outlier points were excluded when drawing convex hulls.

in the new database, as shown in Fig. 13. As expected, the database bias significantly elevates the overall performance of VMAF. To show that this limitation generalizes to other FR models, we also show scatter plots of another FR VQA model FSIM in Fig. 14. For better visualization, we excluded outliers when drawing convex hulls. FSIM also effectively distinguishes between compression levels, but the correlation against subjectivity decreases within each compression level, similar to the VMAF. For example, at compression level 1, both VMAF and FSIM predict all videos to be of good quality, but their predicted MOS spans a much wider range. This also shows the inaccuracy and limits of reference models.

We also compared the performance of two reference models, VMAF and FSIM, against 1stepVQA and 1stepVQA-R at compression levels 2 and 3, using videos compressed at level 1 as references in Table VI. We chose FSIM as the reference VQA module of 1stepVQA-R. FSIM, which is the second best model in the Section VI-A, performed the worst in this instance. At compression level 2, 1stepVQA and 1stepVQA-R significantly outperformed VMAF. At compression level 3, VMAF performed the best, while 1stepVQA was better than FSIM. When combined with FSIM, 1stepVQA-R performed much better at compression level 3. When comparing overall performance at compression levels 2 and 3, 1stepVQA and 1stepVQA-R outperformed VMAF and FSIM.

TABLE VI

SROCC BETWEEN VMAF, FSIM, 1STEPVQA AND 1STEPVQA-R SCORES AND MOS AT COMPRESSION LEVELS 2 AND 3. VMAF, FSIM, 1STEPVQA AND 1STEPVQA-R SCORES WERE COMPUTED USING VIDEOS COMPRESSED AT LEVEL 1 AS REFERENCES. FSIM WAS USED AS THE REFERENCE VQA MODULE IN 1STEPVQA-R. BOLDFACE INDICATES THE BEST PERFORMING MODEL

Compression level	2	3	All
VMAF	0.3308	**0.5951**	0.7398
FSIM	0.3149	0.3748	0.6780
1stepVQA	0.5682	0.4136	0.7549
1stepVQA-R	**0.5727**	0.5136	**0.7877**

C. Feature Combination Performance Evaluation

We also conducted a series of feature studies to evaluate other possible 1stepVQA feature combinations. First, we studied the efficacy of only selecting the displacement direction yielding the largest shape parameter value among four diagonal directions, against using GGD features from all eight adjacent directions, and also against using non-displaced frame differences. Second, we tested performance including displaced frame differences along the same four diagonal directions, but using larger displacements (than single pixel). Finally, we also studied the efficacy of using eight displacement directions (instead of four) to compute the direction yielding the largest GGD shape value. To summarize:

- 1stepVQA-I: NVS features computed along all four diagonal directions without the MAX operation.
- 1stepVQA-II: NVS features computed along eight adjacent directions, without the MAX operation.
- 1stepVQA-III: Along with 1stepVQA features, include features of frame differences spatially displaced by two pixels.
- 1stepVQA-IV: Same as 1stepVQA, except applying the MAX operation on the shape features from eight directions instead of four.
- 1stepVQA-V: NFS features only.

We show the number of features in each of these extended versions in Table IX. We tested the four extended models and report the obtained results in Table VII. Table VIII

TABLE VII

PERFORMANCES OF THE 1STEPVQA MODEL AND FIVE EXTENDED 1STEPVQA MODELS, USING DIFFERENT FEATURE GROUPS. BOLDFACE INDICATES THE BEST PERFORMING MODEL

	1stepVQA	I	II	III	IV	V
SROCC	**0.8918**	0.8831	0.8629	0.8840	0.8901	0.8597
LCC	**0.8902**	0.8801	0.8595	0.8805	0.8881	0.8576
RMSE	**6.0275**	6.3045	6.8603	6.2451	6.0777	6.7978

TABLE VIII

RESULTS OF ONE-SIDED WILCOXON RANK SUM TEST PERFORMED BETWEEN SROCC VALUES OF THE VQA ALGORITHMS COMPARED IN TABLE VII. 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

	1stepVQA	I	II	III	IV	V
1stepVQA	0	1	1	1	0	1
I	-1	0	1	0	-1	1
II	-1	-1	0	-1	-1	1
III	-1	0	1	0	-1	1
IV	0	1	1	1	0	1
V	-1	-1	-1	-1	-1	0

TABLE IX

COMPARISON OF THE NUMBER OF FEATURES OF VARIOUS EXTENDED 1STEPVQA VQA MODELS

	1stepVQA	I	II	III	IV	V
Number of Features	16	40	80	24	16	8

shows the results of the obtained statistical significance tests comparing the performance of 1stepVQA against those of the four extended versions of 1stepVQA. It is evident that using the MAX operation to obtain the shape parameter yielded improved performance (while eliminating the need to compute multiple variances) as compared with including all features from all of the displacement direction. Moreover, many fewer features are required to train on. The performance of 1stepVQA-IV was statistically equivalent to that of 1stepVQA, indicating that the additional data was not useful, but requires significantly more computation. Overall, 1stepVQA was statistically superior to all the feature combinations other than 1stepVQA-IV.

D. 1stepVQA-R Model

The combination of 1stepVQA and various reference models provides a more general and robust approach which boosts performance. Table X plots the performances of various 1stepVQA-R models incorporating 1stepVQA with different reference VQA models. We considered five reference VQA models: PSNR, MS-SSIM, FSIM, ST-MAD and VSI. The high-performing reference model FSIM was able to boost performance the most. PSNR and VSI provided limited improvement to the integrated model. While 1stepVQA is able to predict quality considering both authentic distortions on reference videos and compression artifacts, incorporating a reliable reference model provided better prediction of quality degradation between reference and compressed videos.

TABLE X

PERFORMANCES OF 1STEPVQA-R WITH REFERENCE VQA MODULES PSNR, MS-SSIM, FSIM, ST-MAD. BOLDFACE INDICATES THE BEST PERFORMING MODEL

	PSNR	MS-SSIM	**FSIM**	ST-MAD	VSI
SROCC	0.8928	0.9093	**0.9236**	0.9004	0.8914
LCC	0.8896	0.9082	**0.9224**	0.9008	0.8908
RMSE	6.0644	5.6349	**5.1551**	5.8216	6.0976

TABLE XI

PERFORMANCES OF 1STEPVQA MODEL APPLIED ON LUMINANCE AND CHROMA SPACE RESPECTIVELY, AND THEIR COMBINATION

	Luminance	Chroma	Combination
SROCC	0.8918	0.7439	0.8166
LCC	0.8902	0.7426	0.8217
RMSE	6.0275	8.8879	7.6652

TABLE XII

SROCC PERFORMANCES OF TWO-STEP COMBINATIONS OF REFERENCE AND NO-REFERENCE VQA MODELS ON THE LIVE WILD COMPRESSED VIDEO QUALITY DATABASE. BOLDFACE INDICATES THE BEST PERFORMING MODEL

	NIQE	BRISQUE	V-BLIINDS	TLVQM	VSFA
PSNR	0.6882	0.6814	0.6073	0.5615	0.6439
MS-SSIM	0.8349	0.8257	0.7822	0.7753	0.7967
FSIM	**0.8804**	0.8748	0.8444	0.8412	0.8569
ST-MAD	0.8351	0.8585	0.8345	0.8157	0.8373
VSI	0.8067	0.7863	0.7620	0.7594	0.7758

E. Chroma Feature Analysis

Most existing VQA models only make use of features from the luminance domain. However, we also explored the influence of chroma features defined in other color spaces. Specifically, we analyzed the performance of 1stepVQA features applied in chroma space. The LAB chroma space that we used is defined below:

$$C^*_{ab} = \sqrt{a^{*2} + b^{*2}} \tag{13}$$

where a^* and b^* are the two chrominance components of a given frame [52].

Table XI shows the results of applying the 1stepVQA model on the LAB chroma space. As compared to luminance, the influence of the chroma features was weak, even when combined with the luminance features. This might suggest that luminance features are enough when predicting the perceptual quality of UGC videos, but it might also suggest that other, more chroma-sensitive features could be designed to obtain better performance.

F. 2stepQA and General Two-Step Model

The previous 2stepQA model [38] was designed to address the quality assessment problem of distorted-then-compressed images by a simple product combination of separate Reference and NR modules. It performs well, as shown in [38]. However, since it makes use of scores from two models, its performance depends on the efficacy of both of the models (NR and FR). When attempting to extend 2stepQA to the VQA problem, we have been unable to find a sufficiently good NR VQA model that can adequately boost performance within a 2stepQA framework, as shown in Table XII. Only a few combinations are able to boost performance compared

TABLE XIII

COMPARISON OF TOTAL COMPUTE TIMES OF COMPARED VQA MODELS ON THE SAME 300 FRAME 1080P VIDEO, ON A XEON E5 1620 v3 3.5GHz PC WITH 64 GB OF RAM. ALGORITHMS WERE RUN IN MATLAB. ITALICS INDICATE NR ALGORITHMS

	PSNR	MS-SSIM	FSIM	ST-MAD	VSI	2stepQA	*NIQE*	*BRISQUE*	*V-BLIINDS*	*TLVQM*	*VSFA*	1stepVQA
Time (sec.)	22	64	95	3096	203	356	292	72	2746	246	982	155

TABLE XIV

COMPARISON OF THE NUMBER OF FEATURES OF THE COMPARED FEATURE-BASED VQA MODELS. ITALICS INDICATE NR ALGORITHMS

	NIQE	*BRISQUE*	*V-BLIINDS*	*TLVQM*	1stepVQA
Number of Features	36	36	46	75	16

TABLE XV

SROCC PERFORMANCES OF 1STEPVQA, MS-SSIM, AND VMAF ON THE H.264 AND MPEG-2 COMPRESSED VIDEOS FROM THE LIVE VIDEO QUALITY DATABASE

	Combination	H.264	MPEG-2
1stepVQA	0.7206	0.7381	0.9048
MS-SSIM	0.7529	0.7857	0.7619
VMAF	0.8000	0.9048	0.8982

with reference models alone. Naturally, this will change as NR VQA models continue to evolve. Another recent work [53] also discussed this limitation of 2stepQA.

G. Computational Complexity

As compared with other VQA models, 1stepVQA computes the fewest features, and hence is very efficient. We also compared the overall computational complexity of 1stepVQA with those of the FR VQA model ST-MAD and the NR VQA models V-BLIINDS and TLVQM. Table XIII lists the time required (in sec.) to compute each quality model on a single compressed video and its reference video from the LIVE Wild Compressed Video Quality Database. All algorithms were run in Matlab using publicly available online source code. 1stepVQA proved to be faster than the other spatio-temporal VQA models (ST-MAD, V-BLIINDS, TLVQM, VSFA). While some frame-based (spatial) IQA models are faster than 1stepVQA, they do not compute any temporal features and perform significantly worse than 1stepVQA. In Table XIV, we list the number of features used by each of the training-based VQA models. As Tables XIII and XIV show, 1stepVQA is quite efficient.

H. 1stepVQA Performance on LIVE Video Quality Database

We also tested the 1stepVQA model on the H.264 and MPEG-2 compressed videos and corresponding subjective scores in the LIVE Video Quality Database (LIVE VQA), which contains 10 contents, each having four H.264 compressed versions and four MPEG-2 compressed versions, yielding 80 compressed videos overall. Since 1stepVQA requires training, for fair comparison we divided the 10 contents into non-overlapping 80% training and 20% test sets, and iterated training and testing over all 45 combinations, and report the median obtained SROCC in Table. XV, along with those of MS-SSIM and VMAF. However, the reference videos in LIVE VQA are all of very high pristine quality, hence convey little or no intrinsic distortion information to 1stepVQA. Even so, 1stepVQA is able to provide results comparable to those of MS-SSIM, and similar to those of VMAF on the MPEG-2 compression group, although VMAF (which is specifically trained on H.264), outperformed on those compressed contents, and hence overall.

VII. CONCLUSION

We have presented a new VQA model, called 1stepVQA, that is designed to tackle the problem of assessing the quality of compressed videos given reference videos afflicted by pre-existing, authentic distortions. 1stepVQA is computed using information from both the reference and compressed videos, but does not require pristine reference videos. It uses features that capture the deviations of spatial and spatial-temporal statistical regularities caused by the presence of pre-existing in-capture distortions as well as post-capture compression artifacts. Our method utilizes both NFS features, as well as NVS features extracted from displaced frame differences. We find that 1stepVQA is able to outperform mainstream VQA models, while requiring relatively fewer features, making it more efficient and easy to implement. We show that an extended version called 1stepVQA-R is able to achieve higher performance by adding a reference module. We also developed a significant new resource for researchers working on this problem, called the LIVE Wild Compressed Video Quality Database, which contains both UGC reference videos and many compressed versions of them, along with subjective quality labels assigned to them as the outcome of a human study. 1stepVQA also has potential industrial value as a practical and fast way to improve the performance of video encoding optimization on UGC videos.

REFERENCES

[1] Cisco Visual Networking Index. (Feb. 2019) *Cisco Visual Networking Index: Forecast and Methodology, 2017–2022*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[2] A. Aaron, Z. Li, M. Manohara, J. De Cock, and D. Ronca. (Dec. 2015). *Per-Title Encode Optimization*. [Online]. Available: https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2

[3] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. (Jun. 2016). *Toward a Practical Perceptual Video Quality Metric*. [Online]. Available: https://medium:com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[4] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

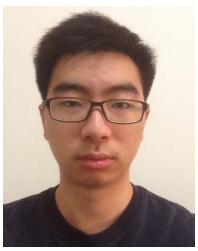
[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

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

[7] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2012.

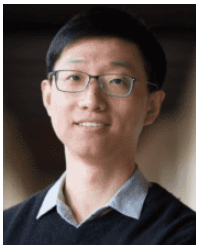
[8] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, Feb. 2019.

[9] D. L. Ruderman, "The statistics of natural images," *Netw., Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.
[10] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[11] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012.
[12] M. H. Pinson, "The consumer digital video library [best of the web]," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 172–174, Jul. 2013.
[13] J. Y. Lin, R. Song, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *J. Vis. Commun. Image Represent.*, vol. 30, pp. 1–9, Jul. 2015.
[14] H. Wang *et al.*, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1509–1513.
[15] H. Wang *et al.*, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 292–302, Jul. 2017.
[16] V. Hosu *et al.*, "The Konstanz natural video database (KoNViD-1k)," in *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2017, pp. 1–6.
[17] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2019, pp. 1–5.
[18] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.
[19] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.
[20] 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.
[21] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual DOG model fused with random forest," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3282–3292, Nov. 2015.
[22] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2505–2508.
[23] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.
[24] 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.
[25] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.
[26] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, Dec. 2019.
[27] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2351–2359.
[28] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *Int. J. Comput. Vis.*, vol. 129, pp. 1–20, Apr. 2021.
[29] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ: 'Patching up' the video quality problem," 2020, *arXiv:2011.13544*. [Online]. Available: http://arxiv.org/abs/2011.13544
[30] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.
[31] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," 2019, *arXiv:1912.10088*. [Online]. Available: http://arxiv.org/abs/1912.10088
[32] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 609–613.
[33] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015.
[34] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.
[35] 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, Nov. 2013.
[36] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.
[37] Z. Sinno and A. C. Bovik, "Spatio-temporal measures of naturalness," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1750–1754.
[38] 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.
[39] A. Bovik, "Assessing quality of images or videos using a two-stage quality assessment," U.S. Patent 10 529 066, Jan. 7, 2020.
[40] Y. Li, S. Meng, X. Zhang, S. Wang, Y. Wang, and S. Ma, "UGC-VIDEO: Perceptual quality assessment of user-generated videos," 2019, *arXiv:1908.11517*. [Online]. Available: http://arxiv.org/abs/1908.11517
[41] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Rec. BT.500.13, 2012.
[42] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 587–599, Apr. 2010.
[43] J. W. Peirce, "Generating stimuli for neuroscience using psychopy," *Frontiers Neuroinform.*, vol. 2, p. 10, Jan. 2009.
[44] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 52–61.
[45] B. A. Wandell, *Foundations of Vision*, vol. 8. Sunderland, MA, USA: Sinauer Associates, 1995.
[46] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *J. Math. Imag. Vis.*, vol. 18, no. 1, pp. 17–33, 2003.
[47] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, Feb. 1995.
[48] D. Y. Lee, H. Ko, J. Kim, and A. C. Bovik, "On the space-time statistics of motion pictures," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 38, no. 7, p. 908, Jul. 2021.
[49] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.
[50] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.
[51] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.
[52] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, 2016.
[53] Y. Li, S. Meng, X. Zhang, S. Wang, Y. Wang, and S. Ma, "User-generated video quality assessment: A subjective and objective study," 2020, *arXiv:2005.08527*. [Online]. Available: http://arxiv.org/abs/2005.08527



Xiangxu Yu received the B.Eng. degree in electronic and information engineering from The Hong Kong Polytechnic University, Hong Kong, China, in 2015, and the M.S. degree in electrical and computer engineering from The University of Texas at Austin, Austin, in 2018, where he is currently pursuing the Ph.D. degree with the Laboratory for Image and Video Engineering. His research interests focus on image and video processing and machine learning.

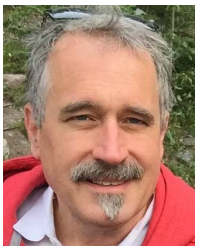
Neil Birkbeck received the Ph.D. degree in computer vision, graphics and robotics from the University of Alberta in 2011, with a specific focus on image-based modeling and rendering. He went on to become a Research Scientist at Siemens Corporate Research working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Software Engineer at Media Algorithms Team, YouTube/Google, with research interests in perceptual video processing, video coding, and video quality assessment.



Yilin Wang (Member, IEEE) received the B.S. and M.S. degrees in computer science from Nanjing University, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from The University of North Carolina at Chapel Hill in 2014, with a focus on computer vision and image processing. After graduation, he joined Media Algorithm Team, Youtube/Google. His research fields include video processing infrastructure, video quality assessment, and video compression.

Christos G. Bampis received the M.Eng. (Diploma) degree in electrical engineering from the National Technical University of Athens in 2014 and the Ph.D. degree in electrical engineering from The University of Texas at Austin in 2018. He is currently a Senior Software Engineer with Video Algorithms Team, Netflix, Inc. His work focuses on research and development of perceptual video quality prediction systems. His research interests include image and video quality assessment, computer vision, and machine learning.

Balu Adsumilli received the master's degree in wireless video communication from the University of Wisconsin-Madison and the Ph.D. degree in watermark-based error resilience in video communications from the University of California Santa Barbara. He is currently the Head of Media Algorithms Group, YouTube/Google. Prior to YouTube, he was a Senior Manager of advanced technology at GoPro, and before that, he was a Senior Research Scientist at Citrix Online, at both places developing algorithms for images/video quality enhancement, compression, capture, and streaming. He has coauthored more than 70 papers and 80 granted patents. His fields of research include image/video processing, audio and video quality, video compression and transcoding, spherical capture/AR/VR, visual effects, and related areas. He serves on both the Television Academy and NATAS committees, the IEEE MMSP Technical Committee, and the ACM MHV Steering Committee. He is on TPCs and organizing committees for various conferences and held numerous workshops. He is an Active Member of IEEE, ACM, SPIE, and Visual Effects Society.



Alan C. Bovik (Fellow, IEEE) is currently Cockrell Family Regents Endowed Chair Professor at The University of Texas at Austin. His books include *The Essential Guides to Image and Video Processing*. His research interests include image processing, digital photography, digital television, digital streaming video, social media, and visual perception. For his work in these areas, he has been the recipient of the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, the 2017 Edwin H. Land Medal from The Optical Society, the 2019 Progress Medal from The Royal Photographic Society, the 2019 IEEE Fourier Award, the 2020 Technology and Engineering Emmy Award from the National Academy for Television Arts and Sciences, Norbert Wiener Society Award, and Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. He has received about ten best journal paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. He co-founded and was the longest-serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING and created/chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, in 1994.