Perceptual Video Quality Prediction Emphasizing Chroma Distortions

Li-Heng Chen, *Graduate Student Member, IEEE*, Christos G. Bampis, *Member, IEEE*, Zhi Li, Joel Sole, *Senior Member, IEEE*, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*— Measuring the quality of digital videos viewed by human observers has become a common practice in numerous multimedia applications, such as adaptive video streaming, quality monitoring, and other digital TV applications. Here we explore a significant, yet relatively unexplored problem: measuring perceptual quality on videos arising from both luma and chroma distortions from compression. Toward investigating this problem, it is important to understand the kinds of chroma distortions that arise, how they relate to luma compression distortions, and how they can affect perceived quality. We designed and carried out a subjective experiment to measure subjective video quality on both luma and chroma distortions, introduced both in isolation as well as together. Specifically, the new subjective dataset comprises a total of 210 videos afflicted by distortions caused by varying levels of luma quantization commingled with different amounts of chroma quantization. The subjective scores were evaluated by 34 subjects in a controlled environmental setting. Using the newly collected subjective data, we were able to demonstrate important shortcomings of existing video quality models, especially in regards to chroma distortions. Further, we designed an objective video quality model which builds on existing video quality algorithms, by considering the fidelity of chroma channels in a principled way. We also found that this quality analysis implies that there is room for reducing bitrate consumption in modern video codecs by creatively increasing the compression factor on chroma channels. We believe that this work will both encourage further research in this direction, as well as advance progress on the ultimate goal of jointly optimizing luma and chroma compression in modern video encoders.**

***Index Terms*— Subjective study, video quality assessment, video codec optimization.**

I. Introduction

MODELING the human perception of video quality has become a crucial research problem owing to the tremendous boom of shared and streaming video content, accessible mobile video devices, and diverse video services such as Netflix, Facebook, Youtube, Hulu, and so on. Perceptual optimization of multimedia workflows is important, since humans are the direct consumers of visual information. Videos have become the dominant portion of Internet traffic, and it is predicted that pictures and videos will comprise 80% of the moving bits in broadband and mobile networks in the near future [1]. Given this tremendous, growing, network bandwidth demand, improving the rate-distortion performance of video compression in perceptual ways has become a critical issue, with the potential for significant social, ecological, and economic impact.

Because of significant efforts applied to developing algorithms that can accurately predict perceptual video quality, many promising models have been demonstrated. One successful example is the trained Netflix video quality prediction model called Video Multimethod Assessment Fusion (VMAF) [2], [3], which is used to optimize a significant percentage of Internet video traffic at a level only matched by SSIM [4]. However, most practically deployed video quality models, including VMAF, only extract visual information from the video luma channel, disregarding color entirely. Yet, neglecting color/chroma components may be detrimental to achieving the most accurate evaluations of video quality. As a practical example, one may intentionally manipulate the chroma fidelity of a video: before encoding a pristine source video, it is common to decimate the chroma components to reduce bitrate consumption, with little impact on predicted or perceived quality. Of course, chroma quality features have been used in some simple ways. For example, averaging the PSNR values across luma and chroma channels has been used in multimedia tools such as FFmpeg. Similarly, the Moving Picture Experts Group (MPEG) community often evaluates the performance of video codecs using a weighted mean of Bjøntegaard-Delta bitrates (BD-rate) [5] across per-channel PSNR values. However, rather than taking human perception into account, the linear weights were empirically derived as a proportion of color channels defined by pixel format.

One potentially high-impact application of perceptually quantifying chroma distortion is the optimization of video compression. For example, in Fig. 1 we applied several different quantization factors settings on noisy chroma channels. Here, distortions of Cb and Cr caused by maximizing the *chroma_qp_offset* parameter to K, thereby increasing quantization of Cb and Cr, are clearly visible. However, while the visual differences caused by the two settings are subjectively and objectively noticeable when displaying chroma channels independently, it is important to observe that these considerable defects in chroma do not alter the visual quality

Manuscript received April 18, 2020; revised September 18, 2020 and November 27, 2020; accepted November 28, 2020. Date of publication December 17, 2020; date of current version December 29, 2020. This work was supported by Netflix. The work of Li-Heng Chen was supported by the summer internship at Netflix. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Giuseppe Valenzise. *(Corresponding author: Li-Heng Chen.)*

Li-Heng Chen 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: lhchen@utexas.edu; bovik@ece.utexas.edu).

Christos G. Bampis, Zhi Li, and Joel Sole are with Netflix, Inc., Los Gatos, CA 95032 USA (e-mail: christosb@netflix.com; zli@netflix.com; jsole@netflix.com).

Digital Object Identifier 10.1109/TIP.2020.3043127

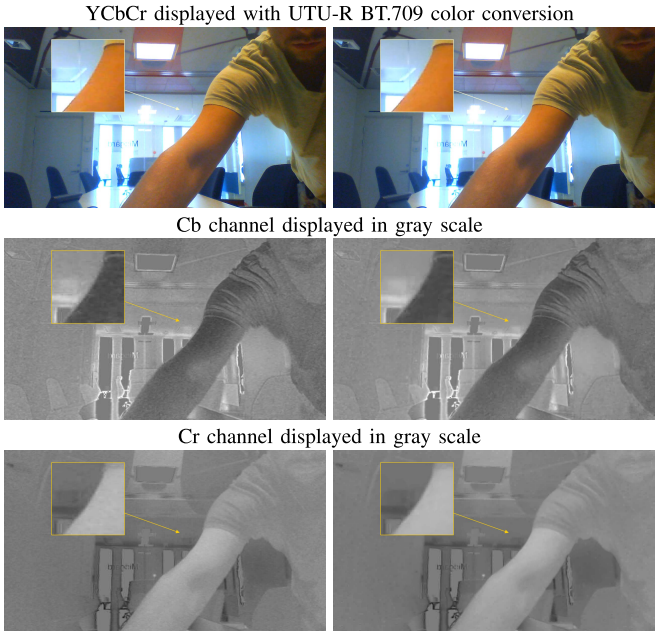


Fig. 1. A bitrate-saving example of encoding test sequence *dark720p_120f* using HEVC. The two videos were encoded at a fixed QP of 17 with different *chroma_qp_offset* settings, resulting in (a) $\text{PSNR}_{\text{Cb}} = 47.77$ db; $\text{PSNR}_{\text{Cr}} = 49.01$ db; bitrate = 20631.65 Kbps, and (b) $\text{PSNR}_{\text{Cb}} = 39.50$ db; $\text{PSNR}_{\text{Cr}} = 41.92$ db; bitrate = 13112.41 Kbps.

when YCbCr are displayed together. Yet, in terms of bit consumption, the encoded video in Fig. 1(b) only occupies 6.5 Mbytes, which is 36% less than the video in Fig. 1(a). Unfortunately, the potentially significant benefits of trading off chroma fidelity against bitrate cannot be captured by the conventional per-channel PSNR, and cannot contribute to improved PSNR-based rate-distortion optimization schemes.

We summarize the main idea behind this work in Fig. 2. Most video encoders operate near the region defined by the black dotted diagonal line, where the amount of compression between chroma and luma is tightly coupled. Using a similar working assumption, most previous subjective quality databases and most quality models do not consider the perceptual effects of decoupling luma and chroma compression. In these scenarios, as shown in Figs. 2(c), (d) and (e), chroma artifacts are typically observed only when the luma quantization factor is high enough (as in Fig. 2(e)). Notably, the upper left region provides room for further bitrate savings, by exploiting heavier chroma quantization while fixing the luma quantization level, without coupling the two. Figure 2(a) shows an example of severely degrading chroma, while luma quantization is kept to a much lower level. As a result, it exhibits significant loss of chroma information (desaturation), but the scene structure remains intact. Moving across Fig. 2 in the horizontal direction, luma quantization increases, until significant loss of detail and texture are observed (as in Fig. 2(b)). Notably, the bottom-right of Fig. 2 is a region where luma is more severely quantized than chroma, an approach that should be less efficient from the rate-distortion point of view, given that human perception is more sensitive to loss of detail.

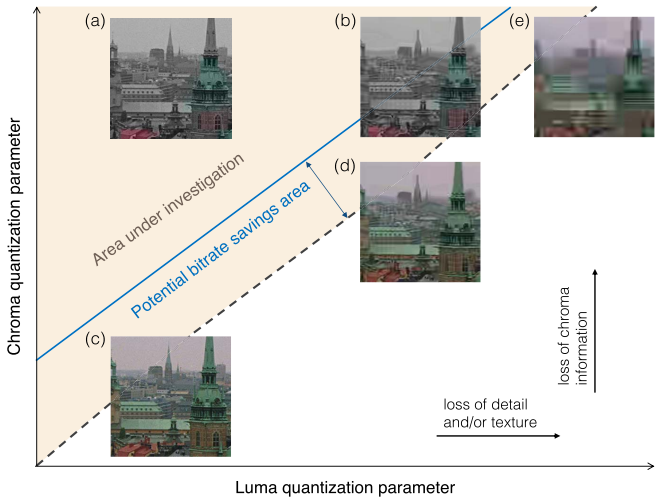


Fig. 2. Conceptual overview of this paper. Each patch shows distortion from a specific combination of luma and chroma quantization. The black dotted line denotes the default setting *chroma_qp_offset* = 0. The area between the black dotted diagonal and the blue solid diagonal illustrates where modern codecs are currently allowed to vary quantization in chroma ($0 \leq$ *chroma_qp_offset* $\leq K$).

Motivated by the significant potential of perceptual chroma quality prediction and optimization, we have attempted to improve the current VMAF model by accounting for the perceptual quality attributes of the chroma components of videos. To help us advance progress in this direction, we built a new chroma distortion-specific video quality database, on which we conducted a subjective quality study to better understand how different levels of compression distortion in chroma affect quality perception, and how they relate to luma distortions. This study seeks to examine ways of exploiting the upper-left area of Fig. 2, to better understand the perceptual tradeoffs that should mediate luma and chroma compression. Using the collected data, we developed and tested new chroma quality-aware features which can be used to improve existing learning based video quality predictors. We also studied the problem of color-sensitized video quality prediction, resulting in the design of an improved VMAF model, that we introduce here, and which can be used to improve the perceptual optimization of compression.

Moving forward, section II reviews related literature and motivates the need for a new dataset. Section III details the new database's source contents, the creation of chroma distortions, and the perceptual study design and outcomes. Section IV discusses data analysis of the subjective study results. Section V explores different chromatic features and appropriate design methodologies for constructing quality prediction models, while section VI offers an analysis of the performance of a variety of objective VQA algorithms on the new database. Finally, section VII concludes with a discussion of future directions of research.

II. Background

We begin with a background review of studies related to subjective video quality. Following that, currently available objective video quality assessment algorithms also discussed.

A. Subjective Video/Picture Quality Databases

Numerous Video Quality Assessment (VQA) datasets have been designed over the past decade. The early LIVE VQA Database [6] and the LIVE Mobile Video Quality Database [7] remain among the most widely used public-domain video quality databases. These databases model post-acquisition distortions, such as H.264 compression and transmission errors, to match real world scenarios. Other similar databases, of videos having larger resolutions, deeper bit-depths, or encoded by more advanced codecs have been proposed, including the CSIQ-VQA Database [8], the SHVC Database [9], and the VQEG Database [10]. We refer the reader to [11], [12] for a comprehensive analysis and collection of publicly available databases through 2012. In recent years, more video quality databases [13]–[19] have been proposed to address different scenarios. Although these databases have undoubtedly accelerated the development of VQA algorithms, generally distortions of the chroma channels were not specifically designed. In most cases, chroma distortion are hightly coupled with luma distortion.

To the best of our knowledge, we are aware of only a few studies of chromatic distortions. Shang *et al.* [20] introduced different levels of chromatic distortions than conventional codecs. Source videos were divided into three planes, then independently encoded at different quantization levels. The independently encoded channels were then combined into a single distorted video. Although this methodology produces combinations of Y/Cb/Cr components having different levels of distortions, it is not naturally produced by a video codec. Here, we instead propose a different framework whereby distorted videos are generated directly from an encoder, as we will discuss.

There are also two databases that studied *image* quality, including certain distortions of chroma components: TID2013 [21] comprises various color distortions in addition to 17 other distortion types. However, the four synthesized color distortions are peculiar and not likely to be present in practical situations. In [22], Sinno *et al.* studied the quality of billboard and thumbnail images displayed by streaming services. This database contains JPEG-distorted still pictures in 444/420 formats, representative of modern internet applications. They found that chromatic distortions, such as color bleeding or jaggies, were induced by chroma subsampling. However, these artifacts are often negligible, and are generally only noticeable when viewed in side-by-side pairwise comparisons. They tend to be even more subtle when appearing on video, where temporal masking effects often suppress faint distortions.

B. Objective Video Quality Assessment Algorithms

Another closely related topic, objective video quality assessment, has been a long-standing and fundamental research problem. Generally, video quality prediction models are classified as full-reference (FR), reduced-reference (RR), or no-reference (NR), based on the availability of the high-quality reference data. Here we only focus on the FR scenario, since it may be assumed that ground-truth data is available in our target applications. Beyond the popular SSIM [4], [23] and MS-SSIM [24] models, which are computationally simple, numerous other powerful perceptual models have also been proposed. These include the VIF index [25] which supplies most of the features in VMAF, the MOVIE index [26], ST-MAD [27], the VQM-VFD models [28]–[30], and many others [31]–[38]. A few still picture (IQA) models [39]–[41] have been designed to tackle chroma distortions, yielding improvements on the color-distorted images in the TID2013 database.

Despite the commercial success of many perceptual quality predictors, the pixel-wise PSNR is still a mainstream model, especially in the testing of video codecs. Indeed, variants of PSNR have been used in the context of codec comparison. During the coding process three PSNR values $\mathrm{PSNR_Y}$, $\mathrm{PSNR_{Cb}}$ and $\mathrm{PSNR_{Cr}}$ are obtained. They are usually combined to produce $\mathrm{PSNR_{611}}$ [42] or $\mathrm{PSNR_{411}}$[1] as the final quality score:

$$\mathrm{PSNR_{k11}} = \frac{\mathrm{k} \cdot \mathrm{PSNR_Y} + \mathrm{PSNR_{Cb}} + \mathrm{PSNR_{Cr}}}{\mathrm{k}+2}. \tag{1}$$

PSNR-HVS-M [43], perceptually-oriented extension of PSNR, has been integrated into the Daala codec [44] for performance evaluation. A recently proposed color-sensitivity-based combined PSNR (CS-PSNR) [20], uses perceptually optimized weights on Y/Cb/Cr based on a subjective analysis of color sensitivity. The MSE weight for each channel is inversely proportional to just-noticeable unit area of a checkerboard pattern. However, this model might be limited by its straightforward linear fusion of per-channel MSE's. Again, chromatic information is often neglected, or naively exploited, both in the design of databases and in quality prediction algorithms.

III. Subjective Experiment Design

Next we describe the key characteristics of the subjective study that we conducted to capture human judgments of compression distortion on both luma and chroma channels.

A. Generating Chroma Compression Distortions

In most modern video coding standards, the quantization parameters (QP) for chroma components are not explicitly designated. Instead, it is derived from the QP value of the luma channel with a parameter *chroma_qp_offset* (the syntax name may vary across different codec standards) that gives a certain degree of flexibility. For example, syntaxes **cb_qp_offset** and **cr_qp_offset** are defined in the High Efficiency Video Coding (HEVC) standard. The offsets are first clipped to the range $[-12, 12]$, then added to the luma QP (denoted by $\mathrm{QP_Y}$):

$$\begin{aligned} \mathrm{QP_{i,Cb}} &= \mathrm{QP_Y} + \mathrm{clip}_{[-12,12]}(\textbf{cb\_qp\_offset}), \\ \mathrm{QP_{i,Cr}} &= \mathrm{QP_Y} + \mathrm{clip}_{[-12,12]}(\textbf{cr\_qp\_offset}). \end{aligned} \tag{2}$$

Then, $\mathrm{QP_{i,Cb}}$ and $\mathrm{QP_{i,Cr}}$ are mapped to the QP values for Cb and Cr:

$$\begin{aligned} \mathrm{QP_{Cb}} &= f\left(\mathrm{QP_{i,Cb}}\right), \\ \mathrm{QP_{Cr}} &= f\left(\mathrm{QP_{i,Cr}}\right), \end{aligned} \tag{3}$$

[1] used in libavfilter of ffmpeg

(a) $(\mathrm{QP_Y}, \mathrm{QP_c}) = (15, 15)$ (b) $(\mathrm{QP_Y}, \mathrm{QP_c}) = (15, 51)$

(c) $(\mathrm{QP_Y}, \mathrm{QP_c}) = (35, 33)$ (d) $(\mathrm{QP_Y}, \mathrm{QP_c}) = (35, 51)$

Fig. 3. Exemplar frames encoded using different QP values on luma and chroma. (a) and (c) are encoded without increasing chroma QP values; (b) and (d) are encoded with extreme chroma QP's.

TABLE I

SPECIFICATION OF QP FOR CHROMA ($\mathrm{QP_C}$) AS A FUNCTION OF THE TRANSITIONAL VALUE $\mathrm{QP_I}$ IN HEVC

$\mathrm{QP_i}$	< 30	30	31	32	33	34	35	36
$\mathrm{QP_c}$	$\mathrm{QP_i}$	29	30	31	32	33	33	34
$\mathrm{QP_i}$	> 43	43	42	41	40	39	38	37
$\mathrm{QP_c}$	$\mathrm{QP_i}$-6	37	37	36	36	35	35	34

where $f(.)$ is a nonlinear mapping function normally implemented as a look-up table. Table I describes this function in the HEVC standard. Consequently, a tremendous challenge in generating realistic chroma compression distortions is that the *chroma_qp_offset* parameter in all of the video compression standards are limited to the range $[-12, 12]$, hence one cannot directly produce compressed videos with arbitrary chroma quantization levels distinct from the luma quantization parameter.

As a preliminary step, we experimented with a "stitching" method that generates two videos encoded at different QP values. These are decoded, then the luma of the first is combined with the chroma from the second. For example, one can combine a luma component encoded with $\mathrm{QP} = 15$, with chroma components encoded with $\mathrm{QP} = 51$ to create extreme chroma distortion. While this approach is intuitive and easy, it comes with a number of disadvantages. First, it does not produce bitstreams for every distorted video, hence RD performance is difficult or impossible to analyze. Moreover, we noticed that distortions created by this approach can appear very different from a real encoding result. Despite having the same quantization in chroma channels, their predictor/residual are considerably different. To address these shortcomings, we decided to create true encoding results by removing the clipping function (2) defined in the HEVC standard. An example of distortion generated in this way can be found in Fig. 3. As may be observed, both Figs. 3(b) and 3(d) show severe color shiftings on the runners. They are not exactly the same because of the different encoding outcomes from using different $\mathrm{QP_Y}$ values. These unpleasant artifacts cannot be created without modifying the video codec standards.

Fig. 4. Sample video frames extracted from the 21 video contents in the new database.

B. Source Sequences

We selected 21 pristine Full High Definition (FHD) video sources from amongst the Netflix movies and TV shows in the Netflix catalog. These video frames were stored using visually lossless JPEG2000 compression. When needed, they were decoded to YUV420 format (YCbCr color space with 4:2:0 sampling). None of the videos contain audio components. Fig. 4 shows examples of some of the selected content. The source videos have a variety of characteristics, such as dark/bright scenes, static/action scenes, close ups of human faces, and so on. All videos are about 8–10 second long and all of them have a frame rate of 24 frames per second.

As a way of quantifying the low-level content of the videos, we adopted the popular Spatial Information (SI), Temporal Information (TI), and Colorfulness (CF) indices [11], [15], [45], [46]. To obtain SI, the luma channel is processed by horizontal and vertical Sobel filters to obtain responses s_h and s_v. Let $s_r = \sqrt{s_h^2 + s_h^2}$ denote the edge magnitude at each pixel location. The SI value is then simply the maximum standard deviation of s_r

$$\mathrm{SI} = \max_t \left(\sigma_{s_r}\right), \tag{4}$$

where σ_{s_r} are calculated on a per-frame basis, and t denotes frame index. The TI value is computed as the standard deviation of the differences between adjacent frame pairs

$$\mathrm{TI} = \max_t \left(\sigma_{\Delta I_t}\right), \tag{5}$$

where $\Delta I_t = I_t - I_{t-1}$ is the pixel-wise difference between the t^{th} frame and the $(t-1)^{\mathrm{th}}$ frame. Finally, color information is captured using CF index in [47], which is formulated as

$$\mathrm{CF} = \max_t \left(\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}\right), \tag{6}$$

where $rg = R - G$ and $yb = 0.5(R+G) - B$ represent opponent color spaces, while μ_x and σ_x are the mean and standard deviation of a plane x. Before calculating CF, we transformed YUV420 to RGB444. As shown in Fig. 5, the reference videos of our database widely span the SI-TI-CF space.

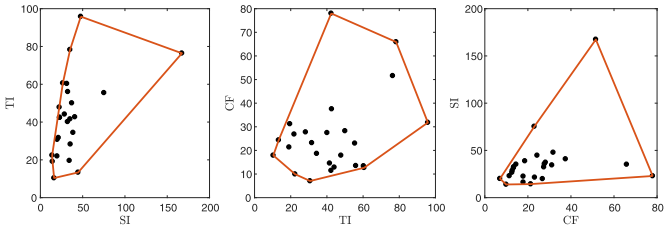


Fig. 5. Scatter plots and corresponding convex hulls of Spatial Information (SI), Temporal Information (TI), and Colorfulness (CF) of the video contents in our database.

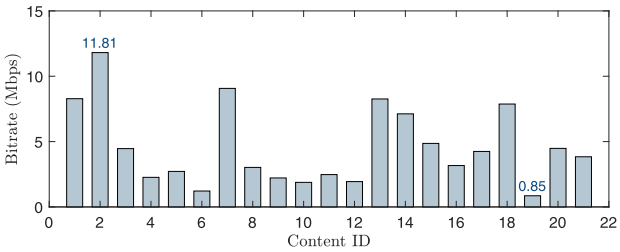


Fig. 6. Encoding complexity of each video content in the new database, expressed as the bitrate derived by encoding them using libx264 with a fixed CRF of 23.

In addition to the typical SI-TI-CF plots, we used another strong indicator of video complexity. Specifically, we encoded all video contents using libx264[2] at a fixed Constant Rate Factor (CRF) setting of 23. Then, the bitrate of each encoded video was measured as an empirical indication of encoding complexity [48] (intuitively, complex contents consume more bits than simple contents at the same quantization level). Figure 6 demonstrates that our video contents span a large range of the bitrate spectrum at constant CRF, ranging from less than 1.0 Mbps up to around 11.8 Mbps. Due to content licensing restrictions, the video contents of the database cannot be made available.

C. Encoding Space and Experimental Design

In subjective experiments of video quality, it is important to carefully design the distortion space, such that the video artifacts are perceptually well-separated and the number of video distortions is not too large. A large number of video distortions will either increase the average viewing time for participants or the number of required participants, both of which are undesirable in subjective lab tests.

We carefully selected three luma QP values such that their VMAF distributions were separated, as demonstrated in Fig. 7. In fact, the QP values give visually separated distortion levels on most of contents. The quantization parameters we used are summarized in Table II, where QP_Y represents luma QP, and QP_c represents chroma QP. For each QP_Y employed on a content, three **cb_qp_offset / cr_qp_offset** settings, ranging from 0 to 51, were assigned resulting in three different QP_c s. Aside from the two extreme values, we uniformly covered the range of QP_c for each content. For example, a content encoded with $QP_Y = 15$ can bracket one set of QP_c from the sets $\{15, 21, 51\}$, $\{15, 27, 51\}$, or $\{15, 39, 51\}$. After the QP selection process, the resulting distorted videos were stored for later display to the human participants.

[2]https://trac.ffmpeg.org/wiki/Encode/H.264

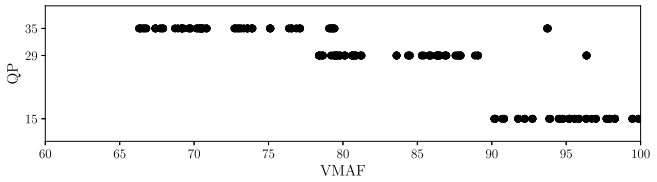


Fig. 7. VMAF scores of distorted videos against their luma QP. Each point represents a distorted video in the dataset.

TABLE II

DESIGN OF QUANTIZATION DISTORTION LEVELS IN THE NEW DATABASE. FOR EACH CONTENT, ONE OF THE THREE VALUES IN THE BRACKETS WAS SELECTED FOR THE STUDY

QP_Y	**cb_qp_offset / cr_qp_offset**	QP_c
15	0, {6, 12, 30}, 51	15, {21, 27, 39}, 51
29	0, {6, 12, 20}, 51	29, {33, 36, 43}, 51
35	0, {6, 12, 17}, 51	33, {36, 41, 46}, 51

Following standard practice, the subjective study sessions were divided into several separate 30-minute viewing sessions, to avoid users' and/or visual fatigue. In addition, every subject's sessions were separated by at least 24 hours, for the same reason [49]. Hence, we designed a two-session study as follows. The videos were assumed to have durations of 10 seconds each (actually 8–10s), and both sessions presented the same 21 contents.

In the first session for each content, five distorted versions of it and one pristine video were included for each content. Therefore, the corresponding viewing time of the first session was about

$$21.0 \text{ min} = 10 \text{ sec} \times 21 \times \overbrace{\underbrace{(5+1)}_{\text{distorted + pristine}}}^{\text{\# of videos}}. \tag{7}$$

Similarly, the viewing time of the second session was about

$$17.5 \text{ min} = 10 \text{ sec} \times 21 \times (4+1). \tag{8}$$

To allow for variations, 30 minutes were allocated to each session, including the instructions given each subject in the first session.

D. Viewing Conditions

During the study, all the videos and graphical user interface (GUI) were displayed using a 15 inch MacBook Pro having a 0.391 m diagonal length and a 2880 × 1800 native resolution retina display. During the subjective test, we set the display resolution to 1920 × 1200. The display was positioned at a viewing distance of about 0.610 m, which is approximately equivalent to three times the height of the display monitor. Also, the subjects were told not to modify their viewing positions very much, while still remaining comfortable.

The laboratory environment for the subjective study followed the recommended viewing conditions in section 2.1.1 of BT.500-11 [49]. The ambient illumination was fixed at a low light level, the display brightness was held constant at 50% of maximum throughout the study, and the automatic brightness

adjustment feature was disabled to guarantee consistency of brightness.

E. Study Instructions, Training, and Subjects

Each participant was asked to take a Snellen visual acuity test and an Ishihara color test, to understand their state of vision. If a subject normally used corrective lenses when watching videos, they were asked to wear them to achieve normal vision when participating in the study. Moreover, a set of instructions were given to each subject explaining the subjective testing process. They were asked to report opinion scores expressive of their viewing experience on every video. The detailed written instructions were given, followed by a brief verbal exchange to ensure that the subjects understood their tasks. In order to remove any possible rating biases, the participants were requested not to base their judgments on whether or not any video content was interesting. Finally, the subjects were informed that there were no right or wrong answers in the experiment.

As a practice, four “practice" videos were presented before the actual study began. These videos were designed to be broadly representative of the range of distortion types and levels that the participants might experience during the actual test. This **“training session"** facilitated the subject being familiar with the study, and was only given in the first session. The contents used for training were the same for all subjects and were distinct from the actual test contents. The **“testing session"** will be described in the next subsection.

The study was carried out over a three-week period at Netflix. In total, 34 subjects were voluntarily recruited. Roughly half of the subjects were experts in the field of image/video engineering while the others were average viewers. In terms of visual acuity, 13 subjects (38.2%) possessed approximately 20/25 vision (slightly worse than ideal) while the others had 20/20 or better vision after correction. Only two participants did not accurately recognize Ishihara Plates during the color vision test, but were allowed to continue as being typical of the populace. Interestingly, the two subjects who failed the color vision test were able to perceive the chroma distortions, and were not rejected as outliers.

F. User Interface and Test Methodology

The user interface was designed using PsychoPy3 [50], which is an open source software often used in experimental psychology and neuroscience research. We followed the single stimulus procedure described in [51], whereby the videos were displayed one after the other. Each video was displayed at native resolution to prevent additional scaling artifacts. Two 1920×60 black bars were rendered at the top and the bottom of the screen, respectively, to fill the 16:10 aspect ratio of the MacBook Pro display.

At the end of each displayed video, a continuous Likert scale of possible video quality scores was displayed on the screen, as demonstrated in Fig. 8. The cursor was reset to the center of the rating bar after each rating, to avoid bias. Five equally spaced labels “Bad," “Poor," “Fair," “Good," and “Excellent," were marked below the rating bar to help the subject understand the range and types of ratings they could apply.

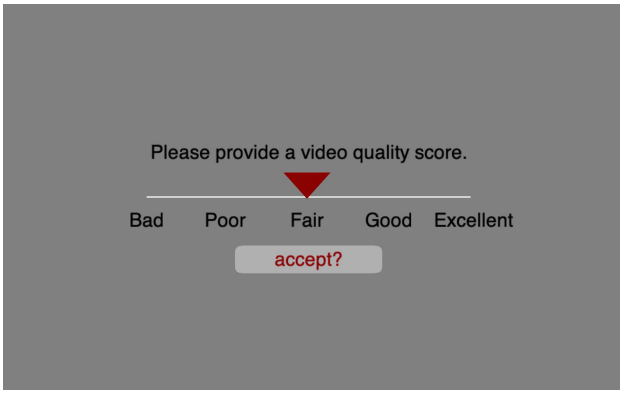


Fig. 8. User interface employed in the subjective test: Likert rating bar for the subject to submit a quality score for the video they completed viewing.

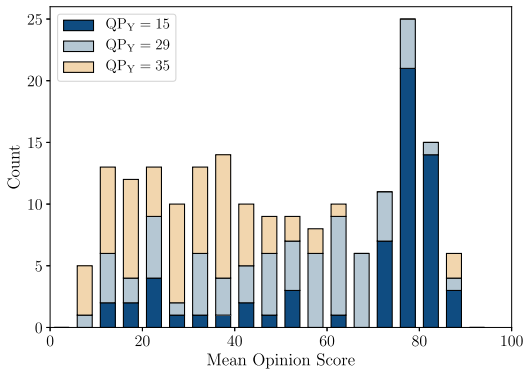


Fig. 9. MOS distribution across the NFLX-Color database. A further breakdown with respect to QP_Y were shown in each bin.

This Likert scale is similar to the ACR scale documented in ITU-R [49]. To rate the videos, the subjects moved the red inverted triangle cursor above the rating bar using the touchpad on the MacBook Pro. After the subject decided a quality score that corresponded to his or her perception, the subject was asked to click on the *Accept* button to record the score and to view the next video. After submitting the score, the position of the sliding bar was converted to an integer quality score in the range $[1, 100]$. Then, the next video clip was presented. It should be noted that once a score was submitted, the name and score of that video were written to file and could not be changed. Of course, the video could not be viewed again.

In this way, we collected 6468 scores over 56 sessions from 34 subjects. In the following, we will refer to the videos and the data collected during the subjective study as the NFLX-Color ($NFLX_c$) video quality database.

IV. Analysis of the Subjective Experiment

The following subsections present an analysis of the subjective experiment results.

A. Data Processing

Given the raw scores collected from the study, subjective Mean Opinion Scores (MOS) were then computed according to the procedures detailed in [6]. Let s_{ijk} denote the raw score assigned by the i-th subject on the video j in session k,

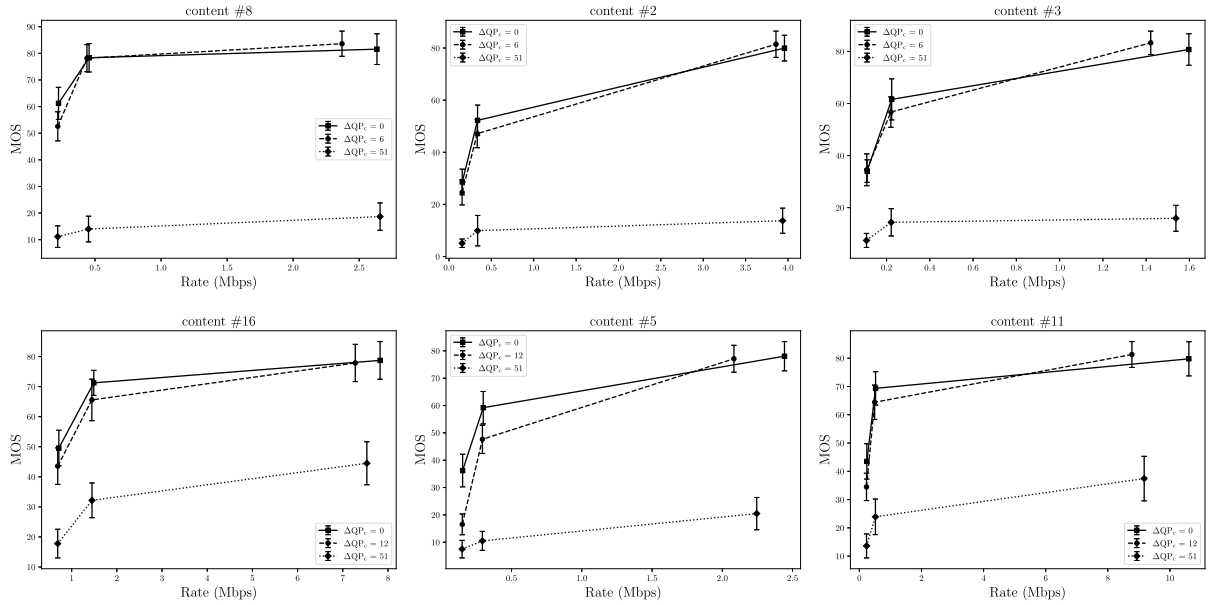


Fig. 10. Examples of MOS Rate-Distortion curves for different *chroma_qp_offset* (denoted by $\Delta\mathrm{QP_c}$) settings. The error bars indicate the 95% confidence interval of each RD point.

where $k \in \{1, 2\}$. The raw scores were then converted into z-scores for each session:

$$z_{ijk} = \frac{s_{ijk} - \mu_{ik}}{\sigma_{ik}}, \tag{9}$$

where μ_{ik} and σ_{ik} are the mean and the standard deviation of the raw scores over all videos assessed by subject i in the k^{th} session. Then, the subject rejection procedure described in ITU-R BT500.13 [52] was conducted to remove outliers. After this procedure, we linearly scaled the remaining z-scores to $[0, 100]$. The MOS of each video was the mean value of the rescaled z-scores from the remaining subjects. In our study, we used the SUREAL[3] python package to calculate MOS with subject rejection following the procedure described above, and only 6 out of 34 subjects were rejected.

B. Analysis of Subjective Quality Scores

Figure 9 plots the distribution of MOS obtained by the aforementioned procedures. Clearly, the MOS of the distorted videos span the entire range of quality. A peak may be observed at a high MOS region, since slightly increasing the chroma QP (e.g. $\Delta\mathrm{QP_c} = 6, 12$) on the videos with $\mathrm{QP_Y} = 15$ did not affect subjective quality much. We also computed the standard error of the subjective scores on each video, obtaining an standard error ranging from 0.822 to 4.414, with an average value of 2.877. This indicates the difficulty of the quality prediction task on specific contents or distortions. We found bifurcated opinions from the subjects occurred on combinations of a "clean" luma plane and severely distorted chroma channels. This is understandable: the recipe $(\mathrm{QP_Y}, \mathrm{QP_c}) = (15, 51)$ is an encoding setting that does not usually appear in daily life; hence they received more inconsistent scores. On the contrary, consensus emerged on the other corner, where both luma and chroma were distorted at the highest quantization level.

[3]https://github.com/Netflix/sureal

C. MOS Rate-Distortion Curves

The Rate-Distortion (or Rate-Quality) curve is a common tool for comparing different encoders or encoding settings in lossy compression. In our study, we were able to collect the bitstream and subjective quality score for each distorted video. Given these results, we compared different *chroma_qp_offset* ($\Delta\mathrm{QP_c}$) settings, as shown in Fig. 10, by plotting MOS against bitrate. A key result in these RD-curves is that, when increasing $\Delta\mathrm{QP_c}$ by 6 or 12 at a high bitrate, roughly 2.5%–17.2% of bits can be reduced without suffering losses of perceptual quality. Occasionally, it may be observed that a video with small $\Delta\mathrm{QP_c}$ was slightly preferred over the setting $\Delta\mathrm{QP_c} = 0$. This could be because both distorted videos were of very high quality, whereby some subjects were unable to distinguish differences in subjective quality, i.e., this may be attributed to statistical noise. On the other hand, the MOS sometimes dropped drastically even with slight increments of chroma quantization, at low bitrate settings. Towards understanding this, we observed more severe color shifts due to the additional quantization that affects the low frequency chroma components. It is possible that human perception is sensitive to this distortion type. Additionally, the percentage of bitrate reduction was not as significant as at high bitrates. This observation has important implication for encoding recipes: there is still room for perceptually optimizing coding efficiency, by better configuring VQA models to align with human percepts of chroma distortion.

D. Limitations of the Current Study

Introducing excessive chroma distortion is relatively new to the construction of a video quality dataset. Despite our best

TABLE III

PERFORMANCE COMPARISON OF CANDIDATE CHROMATIC FEATURES: EACH CELL SHOWS THE SROCC VALUES OF A FEATURE APPLIED ON Cb AND Cr CHANNELS, EXPRESSED AS SROCC_{Cb} / SROCC_{Cr}. THE BEST RESULTS AMONG EACH DATASET ARE DENOTED BY BOLDFACE. PLEASE REFER TO THE DATASET ACRONYMS IN SECTION VI-A.1

Dataset	LIVE-VQA	LIVE-MBL	NFLX	$\mathbf{NFLX_c}$	BVI-HD	CSIQ-VQA	EPFL	VQEG	SHVC
PSNR	0.074 / 0.146	0.559 / 0.548	0.580 / 0.669	0.689 / 0.705	0.408 / 0.398	0.405 / 0.401	0.271 / 0.251	0.557 / 0.587	0.569 / 0.519
SSIM	0.109 / 0.048	0.441 / 0.435	0.662 / 0.702	0.367 / 0.355	0.454 / 0.406	0.412 / 0.373	0.206 / 0.190	0.636 / 0.636	0.579 / 0.536
VIF	0.152 / 0.125	0.379 / 0.360	0.506 / 0.564	0.151 / 0.133	0.326 / 0.239	0.392 / 0.393	0.228 / 0.208	0.451 / 0.449	0.417 / 0.453
VIF_{s0}	0.105 / 0.109	0.367 / 0.348	0.484 / 0.539	0.118 / 0.102	0.307 / 0.216	0.380 / 0.381	0.228 / 0.182	0.432 / 0.427	0.362 / 0.418
VIF_{s1}	0.234 / 0.269	0.466 / 0.398	0.776 / 0.788	0.458 / 0.381	0.547 / 0.473	0.457 / 0.432	0.444 / 0.391	0.618 / 0.588	0.753 / 0.717
VIF_{s2}	0.298 / 0.301	0.489 / 0.414	0.807 / 0.811	0.546 / 0.462	0.611 / 0.533	0.443 / 0.415	0.573 / 0.517	0.675 / 0.647	0.790 / 0.746
VIF_{s3}	**0.381 / 0.362**	0.531 / 0.457	0.818 / 0.819	0.609 / 0.525	0.649 / 0.581	0.439 / 0.407	**0.640 / 0.606**	0.716 / 0.701	**0.807 / 0.756**
ADM	0.212 / 0.176	0.628 / 0.709	0.858 / 0.906	0.758 / 0.782	0.690 / 0.659	0.446 / 0.432	0.380 / 0.356	0.668 / 0.668	0.752 / 0.668
ADM_{s0}	0.184 / 0.092	0.304 / 0.352	0.569 / 0.616	0.155 / 0.199	0.238 / 0.170	0.187 / 0.165	0.143 / 0.134	0.301 / 0.269	0.496 / 0.452
ADM_{s1}	0.161 / 0.161	0.458 / 0.501	0.782 / 0.823	0.432 / 0.462	0.538 / 0.488	0.377 / 0.373	0.224 / 0.218	0.508 / 0.496	0.781 / 0.662
ADM_{s2}	0.211 / 0.183	0.646 / 0.705	**0.883 / 0.918**	0.772 / 0.807	**0.717 / 0.684**	0.469 / 0.441	0.387 / 0.257	0.646 / 0.643	0.702 / 0.617
ADM_{s3}	0.277 / 0.196	**0.757 / 0.723**	0.841 / 0.869	**0.877 / 0.901**	0.692 / 0.668	**0.482 / 0.450**	0.465 / 0.368	**0.706 / 0.731**	0.682 / 0.605

efforts, there are several limitations of this database. As with most the lab-controlled subjective studies, the experiment was hardly exhaustive. Although we designed a database having reasonably comprehensive content and distortion types, it is still circumscribed by several factors, such as the number of subjects and the quantity of subjective quality labels. As a result, constraints had to be made on the combinations of contents and distortions.

In our experiment design, we only included distortion types from luma and chroma compression. However, in real world applications, such as adaptive streaming, resolution changes are often used to optimize RD performance and can also cause distortions (scaling artifacts). Also, studying videos having distorted luma channels with pristine chroma channels may be of interest. Due to the lack of capacity, we did not include this distortion scenario in our dataset. Furthermore, recalling the preceding RD-curve examples in Fig. 10, it was observed that bitrate reductions from heavier quantization of chroma only occurred in high bitrate regions. Yet, the granularity of luma distortion levels was limited, making it difficult to find a precise sweet spot to optimize a codec.

V. OBJECTIVE VIDEO QUALITY MODEL DESIGN

VMAF is a data driven video quality prediction framework that extracts a number elementary VQA features then nonlinearly fuses these features by training a Support Vector Regressor (SVR). The current version of VMAF extracts the Additive Distortion Metric (ADM) [53] feature and four Visual Information Fidelity (VIF) [25] features computed on different oriented frequency bands. The Temporal Information (TI) feature of VMAF is simply the average difference between consecutive frames. This is used to capture temporal distortions associated with, or possibly causing motion or change. These six features are all derived on the luminance component only. Of course, this current limitation hinders VMAF from capturing chromatic distortions.

A. Chromatic Features

Clearly, our goal is to find features that are expressive of chromatic distortions and that can be used to benefit the training of VQA models. To this end, we selected the VMAF features, along with the well-known PSNR and SSIM algorithms as candidates for the experiments on chroma distortion. We then conducted a systematic evaluation to understand the performances of these existing luma features and algorithms when applied on chroma channels. It should be noted that the chroma channels only have halved horizontal and vertical resulutions compared to luma, since we used videos with YUV420 format for all the experiments. Table III tabulates the obtained Spearman Rank Order Correlation Coefficients (SROCC) obtained on nine databases, including NFLX_c. The first thing to notice is that the two most widely-used perceptual quality models (SSIM and VIF) performed far below desired levels. This was possibly due to the fact that they have been designed and extensively tested only on luminance signals. However, it must also be recognized that chromatic components possess different statistical and scaling properties than luma components. Interestingly, the third scale of the ADM features ($\text{ADM}_{s3}^{(Cb)}$ and $\text{ADM}_{s3}^{(Cr)}$) achieved standout performance across most of the databases. In particular, on NFLX_c, it achieved close to 0.9 of SROCC. This suggests that this feature is highly consistent with the human perception context of chromatic distortion. We also illustrate the multiscale framework of the ADM feature in Fig. 11 for better understanding.

The second observation that can be made is with regards to the multiscale behavior of VIF and ADM: performance was improved as the scale index was increased (lower frequency subband). This is not surprising, since the human chromatic contrast sensitivity functions (CSF) (red-green and blue-yellow color opponents) [54]–[56] pass much lower ranges of frequencies than the luminance CSF. The human visual system neglects higher frequencies when processing chromatic information.

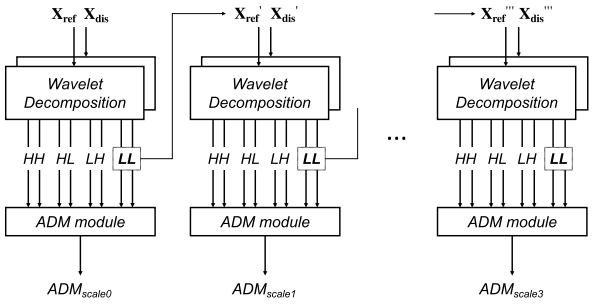


Fig. 11. Framework of multiscale ADM features. The "LL" components from Wavelet transform in each scale are the input of the next, which produces different scales of ADM score iteratively. Larger scale index represents lower frequency subband.

B. Learning a Color VMAF Model

We began by empirically including $\text{ADM}_{\text{s3}}^{(\text{Cb})}$ and $\text{ADM}_{\text{s3}}^{(\text{Cr})}$ as two additional features for training chroma-aware VMAF models, given its exceptional performance on NFLX_c. However, while the trained model significantly boosted correlation on NFLX_c, we discerned considerable performance degradation on other databases, especially LIVE-VQA. In fact, when tested on LIVE-VQA, all of the chromatic features yielded unsatisfactory SROCC, as may be seen in Table III. This may have been because nearly half of the LIVE-VQA database videos contain transient distortions produced by simulated network losses, which may not be effectively captured by chromatic features. It is important to note that these distinct distortions are different from our training data (see Section VI-A), posing additional challenges. In any case, without careful training, the additional chroma features can lower performance.

In order to tackle this issue, we regularized the additional chromatic features using a uniform quantization function. A uniform quantizer with a parameterized quantization step size $\Delta_N = 1/N$ that maps a real value $x \in (0, 1]$ to N discrete values is given by

$$\tilde{x} = Q_N(x) = \Delta_N \lceil \frac{x}{\Delta_N} \rceil. \tag{10}$$

The spirit behind this approach is simple: when a feature is excessively quantized, say, with $N = 1$, the feature becomes a constant to the regression model. In this case, the quantized variable cannot contribute to the model learning, resulting in the same feature set as VMAF 0.6.1; conversely, as N grows, the result becomes closer to the original feature. Thus, the quantization function allows a flexible trade off in performance on NFLX_c and the other databases. We studied the effect of using different step sizes Δ_N, and report the results in Table IV. It may be seen that the SROCC performance on LIVE-VQA and EPFL decreased as N was increased, whereas a reversed trend was observed on NFLX_c. Some databases, such as VQEG, were less affected by varying step size. Based on the results of Table IV, we chose $N = 8$ to quantize the chroma ADM features.

TABLE IV

SROCC PERFORMANCE WITH RESPECT TO DIFFERENT QUANTIZATION STEP SIZES REPRESENTED BY N. THE LAST ROW ($N = \infty$) DENOTES THE FEATURES WITHOUT QUANTIZATION. PLEASE REFER TO THE DATASET ACRONYMS IN SECTION VI-A.1

N	L-VQA	NFLX_c	EPFL	VQEG	Overall
4	0.744	0.772	0.879	0.829	0.829
8	0.740	0.822	0.873	0.834	**0.838**
16	0.706	0.862	0.858	0.829	0.836
32	0.701	0.875	0.845	0.830	0.831
64	0.703	0.877	0.841	0.828	0.831
∞	0.692	0.880	0.841	0.827	0.830

Ultimately, the feature vector used to learn the color VMAF model, which we refer to as VMAF_c, is expressed by

$$\begin{aligned}\mathbf{f}_{\text{VMAF}_\text{c}} = [&\text{VIF}_{\text{s0}}, \text{VIF}_{\text{s1}}, \text{VIF}_{\text{s2}}, \text{VIF}_{\text{s3}}, \text{TI}, \\ &\text{ADM}, \tilde{\text{ADM}}_{\text{s3}}^{(\text{Cb})}, \tilde{\text{ADM}}_{\text{s3}}^{(\text{Cr})}].\end{aligned} \tag{11}$$

The terms $\tilde{\text{ADM}}_{\text{s3}}^{(\text{Cb})}$ and $\tilde{\text{ADM}}_{\text{s3}}^{(\text{Cr})}$ denote the $\text{ADM}_{\text{s3}}^{(\text{Cb})}$ and $\text{ADM}_{\text{s3}}^{(\text{Cr})}$ features quantized by (10). The finalized VMAF_c model was trained on the VMAF+ dataset [48], as we will describe in section VI-B.

VI. OBJECTIVE VIDEO QUALITY ASSESSMENT EXPERIMENTS

A. Experimental Setup

1) Evaluation Datasets: In addition to NFLX_c, we tested the models on a variety of subjective VQA databases, including LIVE VQA [6], LIVE Mobile (LIVE-MBL)[4] [7], NFLX [2], CSIQ VQA [8], BVI-HD [14], VQEG-HD3 [10], EPFLPolimi [58], and SHVC [9]. These publicly available datasets are commonly used to evaluate VQA models. They contain a large variety of contents, resolutions, and distortion types, such as MPEG4/H.264/HEVC compression, resolution changes, transmission errors, frame rate adaptations, and so on.

2) Evaluation Criteria: To evaluate the performance of a VQA model, we used the SROCC and the Pearson linear correlation coefficient (PLCC), which are calculated between the ground truth MOS and the predicted scores. The SROCC measures the degree of monotonic relationship between two variables, while the PLCC is computed after a logistic mapping [59] to measure the degree of linear correlation against MOS. Larger values of SROCC/PLCC indicate better performance in terms of correlation with human perception. To compute the overall correlation, we applied Fisher's z-transform [60] to each correlation coefficient value r

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}, \tag{12}$$

then averaged them over all the test databases [48]. This average value of transformed correlation coefficients is then transformed back using the inverse function $r = \tanh(z)$.

[4]We only used the mobile subset in LIVE Mobile

TABLE V

OVERALL PERFORMANCE COMPARISON OF VQA ALGORITHMS: EACH CELL SHOWS THE SROCC AND PLCC VALUES OF AN VQA MODEL EVALUATED ON A SPECIFIC DATABASE, EXPRESSED AS SROCC / PLCC. BOTH VMAF 0.6.1 AND VMAF_c WERE TRAINED ON THE VMAF+ DATASET [48]. THE THREE BEST SROCC RESULTS AMONG EACH DATASET ARE DENOTED BY BOLDFACE

Dataset	LIVE-VQA	LIVE-MBL	NFLX	**NFLX_c**	BVI-HD	CSIQ-VQA	EPFL	VQEG	SHVC	Overall
PSNR_Y	0.523 / 0.549	0.687 / 0.717	0.705 / 0.706	0.412 / 0.424	0.588 / 0.600	0.579 / 0.565	0.753 / 0.754	0.770 / 0.776	0.755 / 0.761	0.655 / 0.664
PSNR_{411}	0.434 / 0.465	0.663 / 0.690	0.703 / 0.698	0.571 / 0.585	0.565 / 0.577	0.545 / 0.536	0.598 / 0.613	0.734 / 0.728	0.738 / 0.746	0.626 / 0.635
PSNR_{611}	0.459 / 0.494	0.672 / 0.698	0.704 / 0.702	0.529 / 0.546	0.573 / 0.585	0.555 / 0.544	0.644 / 0.653	0.745 / 0.743	0.749 / 0.755	0.635 / 0.644
CSPSNR	0.498 / 0.530	0.685 / 0.714	0.704 / 0.705	0.588 / 0.605	0.581 / 0.592	0.572 / 0.558	0.727 / 0.724	0.765 / 0.767	0.750 / 0.757	0.661 / 0.670
PSNR_HVS	0.662 / 0.689	0.757 / 0.784	0.819 / 0.810	0.516 / 0.534	0.739 / 0.748	0.599 / 0.641	0.904 / 0.909	0.798 / 0.799	0.831 / 0.872	0.758 / 0.776
SSIM	0.694 / 0.704	0.757 / 0.767	0.788 / 0.790	0.555 / 0.560	**0.784** / 0.786	0.698 / 0.712	0.712 / 0.703	0.907 / 0.909	0.754 / 0.817	0.754 / 0.766
MS-SSIM	0.732 / 0.739	0.748 / 0.761	0.741 / 0.745	0.524 / 0.525	0.747 / 0.752	**0.749** / 0.746	**0.931** / 0.934	0.898 / 0.902	0.715 / 0.787	0.780 / 0.791
ST-RRED	**0.802** / 0.801	0.881 / 0.903	0.762 / 0.761	0.616 / 0.612	**0.781** / 0.797	**0.801** / 0.786	**0.950** / 0.952	**0.921** / 0.708	**0.888** / 0.899	**0.846** / 0.829
SpEED-QA	0.767 / 0.763	**0.883** / 0.905	0.780 / 0.781	**0.622** / 0.626	0.770 / 0.784	0.746 / 0.747	**0.941** / 0.869	**0.908** / 0.659	0.880 / 0.887	0.834 / 0.798
ST-MAD	**0.825** / 0.830	0.663 / 0.686	0.768 / 0.746	0.549 / 0.550	0.757 / 0.758	0.735 / 0.740	0.901 / 0.908	0.847 / 0.840	0.611 / 0.619	0.760 / 0.761
VQM-VFD	**0.804** / 0.823	0.816 / 0.847	**0.931** / 0.942	0.597 / 0.624	**0.792** / 0.802	**0.839** / 0.830	0.850 / 0.847	**0.939** / 0.943	0.863 / 0.888	**0.847** / 0.859
iCID	0.552 / 0.569	0.763 / 0.773	0.776 / 0.769	**0.890** / 0.892	0.709 / 0.714	0.679 / 0.682	0.778 / 0.774	0.887 / 0.893	0.717 / 0.728	0.766 / 0.773
$\text{VMAF}_{0.6.1}$	0.752 / 0.759	**0.905** / 0.924	**0.931** / 0.944	0.612 / 0.627	0.772 / 0.785	0.615 / 0.624	0.844 / 0.858	0.857 / 0.866	**0.901** / 0.922	0.826 / 0.844
VMAF_c	0.740 / 0.751	**0.881** / 0.901	**0.932** / 0.949	**0.821** / 0.832	0.759 / 0.766	0.597 / 0.623	0.873 / 0.883	0.834 / 0.852	**0.912** / 0.925	**0.838** / 0.855

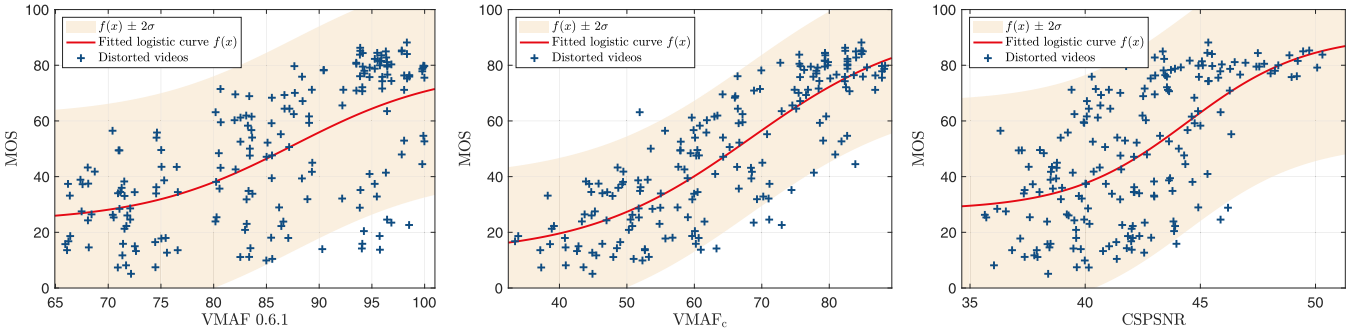


Fig. 12. Scatter plots and logistic regression curves of VMAF 0.6.1 / VMAF_c / CSPSNR versus MOS on NFLX_c dataset. The σ denotes the standard deviation of the residuals between MOS values and $f(x)$. We follow the style in [57].

3) Implementation Details: We used the *libsvm* package [61] to implement ν-SVR [62]. The effectiveness of SVR depends on the selection of kernel, for which we chose the radial basis function (RBF) with penalty parameter $C = 2^3$ and kernel parameter $\gamma = 2^{-3}$. The parameters were empirically selected from a 2D grid of values $(C, \gamma) \in \{2^{-5}, 2^{-4}, \ldots, 2^5\} \times \{2^{-5}, 2^{-4}, \ldots, 2^5\}$. Instead of maximizing the overall performance, we selected a sub-optimal parameter set satisfying a monotonicity property (see Section VI-F), as well as ensuring an appropriate trade-off between NFLX_c and the other datasets. Before training or inferencing, all of the features were linearly rescaled to $[0, 1]$ using min-max normalization. For the sake of simplicity and stability, we used the arithmetic mean [63] to aggregate the per-frame scores.

B. Overall Comparison

We evaluated the optimized VMAF_c against a number of popular FR IQA/VQA models: PSNR, PSNR_{411}, PSNR_{611}, CSPSNR [20], PSNR-HVS-M [43], SSIM [4], MS-SSIM [24], ST-RRED [33], SpEED-QA [34], ST-MAD [27], VQM-VFD [30], iCID [40], and VMAF 0.6.1 [2]. Since separate databases cannot be combined into one in a simple way, most learning-based IQA/VQA models are train/test on multiple databases independently with cross validation. Under this setting, each dataset is typically split into 80%-20% portions with respect to content for training-testing, and the model parameters are searched to maximize individual correlations. Nonetheless, this does not give a general model that can be practically used and is prone to database-specific biases. To avoid these problems, we followed the benchmark criteria of VMAF, where models are trained on a *distinct* dataset VMAF+ [48], then evaluated on the others. VMAF+ is a large VQA dataset designed for training VMAF, containing 522 distorted videos subjected to different levels of scaling and compression artifacts. The performance results are shown in Table V and plots of model predictions versus MOS are shown in Fig. 12. With further investigation, we observed that the outliers in the scatter plot of VMAF 0.6.1 are mostly the videos with $(\text{QP}_\text{Y}, \text{QP}_\text{c}) = (15, 51)$, where chroma distortions were decoupled from luma. These are the instances which VMAF 0.6.1 failed the most.

From these results, we can draw a number of interesting conclusions. The first thing to notice is the performance on the NFLX_c dataset. The best performer among all the VQA models, except VMAF_c and iCID, only reaches SROCC

TABLE VI

RESULTS OF SIGNIFICANCE TESTS BETWEEN THE PERFORMANCES OF VQA MODELS ON THE $NFLX_C$ DATASET. EACH CELL SHOWS THE STATISTICAL SIGNIFICANCE OF SROCC AND PLCC. A VALUE OF '1' INDICATES THAT THE ROW HAS STATISTICALLY HIGHER CORRELATION COEFFICIENT THAN THE COLUMN, WHILE '0' SIGNIFIES THAT THE COLUMN HAS STATISTICALLY LOWER CORRELATION COEFFICIENT THAN THE ROW. A SYMBOL OF '-' INDICATES NO STATISTICAL DIFFERENCE BETWEEN THE CORRELATION COEFFICIENTS OF ROW AND COLUMN

	$PSNR_Y$	$PSNR_{411}$	$PSNR_{611}$	CSPSNR	$PSNR_{HVS}$	SSIM	MS-SSIM	ST-RRED	SpEED-QA	ST-MAD	VQM-VFD	iCID	$VMAF_{0.6.1}$	$VMAF_c$
$PSNR_Y$	--	0 0	--	0 0	--	--	--	0 0	0 0	--	0 0	0 0	0 0	0 0
$PSNR_{411}$	1 1	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
$PSNR_{611}$	--	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
CSPSNR	1 1	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
$PSNR_{HVS}$	--	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
SSIM	--	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
MS-SSIM	--	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
ST-RRED	1 1	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
SpEED-QA	1 1	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
ST-MAD	--	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
VQM-VFD	1 1	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
iCID	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	--	1 1	1 1
$VMAF_{0.6.1}$	1 1	--	--	--	--	--	--	--	--	--	--	--	0 0	--	0 0
$VMAF_c$	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	0 0	1 1	--

slightly higher than 0.6, whereas $VMAF_c$ outperformed most of the other methods, achieving 0.82 SROCC. This improvement over VMAF 0.6.1 is understandable, since the chromatic features were efficiently integrated. It may also be found that $VMAF_c$ performed marginally worse than VMAF 0.6.1 on some databases. Overall, a gain of about 0.01 in both SROCC/PLCC was achieved by $VMAF_c$. Regarding the training set of $VMAF_c$, it should be noted that the VMAF+ dataset does not incorporate any videos with the setting of ΔQP_c, yet a remarkable performance improvement is still attained on $NFLX_c$. This is quite significant, since unlike other databases, $NFLX_c$ allows the measurement of performance on compressed videos with independent chroma compression, which as we have shown, can result in significantly improved perceptual rate-distortion optimization. Despite being the top-performer on $NFLX_c$, the iCID model failed on many of the other datasets with an overall SROCC performance of 0.766, making it hard to justify its use in practical applications.

When comparing $PSNR_Y$, $PSNR_{411}$, $PSNR_{611}$, and CSPSNR, it may be observed that the levels of performance attained by the PSNR family are quite poor. However, when tested on $NFLX_c$, CSPSNR did better than $PSNR_{411}$ and $PSNR_{611}$, which promotes the result reported in [20].

C. Significance Test

We further analyzed the statistical significance of model performances as expressed by the SROCC and PLCC values reported in Table V, following the recommended procedure in section 7.6.1 of ITU-T Rec. P.1401 [60]. The test uses statistics derived from Fisher's-z transformed correlation coefficients in each comparison, compared with the 95% two-tailed Student's t-test critical value. Table VI shows the results of the statistical significance tests.

From the results shown in the table, it may be observed that iCID and $VMAF_c$ statistically surpassed all the other models, since most of them only utilize luminance information. Unsurprisingly, $PSNR_Y$ performed significantly worse than most of the models. It may also be noticed that there was no statistical difference observed when comparing the other models. This is likely because the test methodology was too conservative, given the limited sample size used to calculate correlation coefficients. However, the $VMAF_c$ model was still statistically better than most of the other models under this protocol.

TABLE VII

COMPARISON AGAINST DEEP LEARNING BASED VQA MODEL

	LIVE-VQA		CSIQ-VQA	
	SROCC	PLCC	SROCC	PLCC
DeepVQA-*4ch*	0.891	0.881	0.904	0.901
DeepVQA-*CNAN*	0.915	0.895	0.912	0.914
VMAF	0.939	0.915	0.635	0.575
$VMAF_c$	0.934	0.918	0.683	0.632

D. Comparison With a Deep Learning Based VQA Model

Recently, deep convolutional neural networks have been shown to deliver standout performance on a wide variety of applications. In the field of video quality, a full-reference model called DeepVQA [64] has been proposed, that achieves state-of-the-art performance on the LIVE-VQA and CSIQ-VQA datasets. Unfortunately, the authors of DeepVQA could not provide a trained model that can be tested on all the VQA datasets. To fairly compare DeepVQA against VMAF/$VMAF_c$, we followed the train-test split that yielded the median performance of the DeepVQA-*4ch* model provided by the authors in [64]. We re-trained the VMAF models on each dataset with the same parameters (C, γ) as the original model for simplicity. The results of the performance comparison are shown in Table VII. We report both the DeepVQA-*4ch* model and the best-performing DeepVQA-*CNAN* model. The values for the DeepVQA models are taken from Tables III and IV in the original paper [64]. Overall, DeepVQA yielded slightly worse SROCC/PLCC performance than VMAF and $VMAF_c$ on LIVE-VQA, while performing much better on CSIQ-VQA. This is because CSIQ-VQA mostly contains legacy distortions, such as additive white noise (AWGN) and simulated wireless transmission loss,

TABLE VIII

CROSS DATASET COMPARISON OF THE VMAF_C MODEL. EACH CELL SHOWS THE SROCC PERFORMANCE OF TRAINING ON THE DATASET IN THE ROW AND TESTING ON THE DATASET IN THE COLUMN. THE BEST OVERALL PERFORMANCE IS HIGHLIGHTED IN BOLDFACE

Dataset	LIVE-VQA	LIVE-MBL	VMAF+	NFLX	NFLX_c	BVI-HD	CSIQ-VQA	EPFL	VQEG	SHVC	Overall
LIVE-VQA	–	0.814	0.796	0.877	0.572	0.775	0.644	0.828	0.737	0.830	0.778
LIVE-MBL	0.654	–	0.849	0.915	0.859	0.763	0.627	0.853	0.786	0.889	0.832
VMAF+	0.706	0.891	–	0.932	0.871	0.757	0.626	0.840	0.828	0.923	**0.854**
NFLX	0.723	0.916	0.903	–	0.814	0.784	0.575	0.870	0.844	0.886	0.851
NFLX_c	0.643	0.845	0.899	0.886	–	0.715	0.649	0.749	0.784	0.827	0.818
BVI-HD	0.648	0.841	0.821	0.906	0.797	–	0.642	0.876	0.797	0.863	0.813
CSIQ-VQA	0.527	0.832	0.823	0.799	0.874	0.738	–	0.775	0.799	0.829	0.785
EPFL	0.706	0.763	0.647	0.854	0.629	0.769	0.636	–	0.726	0.860	0.766
VQEG	0.694	0.886	0.880	0.933	0.772	0.779	0.621	0.854	–	0.907	0.832
SHVC	0.608	0.797	0.767	0.886	0.857	0.749	0.632	0.686	0.696	–	0.772

which are of little interest in modern video streaming scenarios. The results also indicate that deep neural networks have the potential to learn good features, including chroma, for assessing video quality.

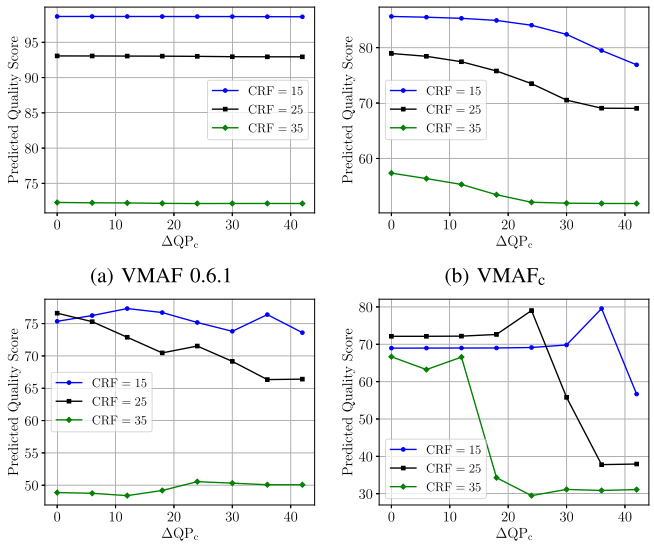
E. Cross-Database Comparison

In addition to analyzing model performance on one training dataset, we investigated the effects of using different datasets to train VMAF_c, with results reported in Table VIII. Using the 10 available VQA databases, we trained the SVR model (with feature $\mathbf{f}_{\text{VMAF}_\text{c}}$) on each dataset, then tested on the others. Due to the differences between the datasets, using the same parameters $(C, \gamma) = (2^3, 2^{-3})$ yielded unsatisfactory performance on some training sets. Therefore, for fair comparison, we separately searched the SVR parameters on a 7×7 grid on each dataset to optimize the overall performance. The experimental results clearly shows the outstanding performance attained by using VMAF+ as the training data. Also, the SROCC attained when testing on NFLX_c was generally greater than 0.7, among the different training sets. This strongly suggests the robustness of the added chromatic features. It should be noted that the performance results on the VMAF+ training set differ slightly from the results reported in Table V, due to different optimization objectives.

F. Monotonicity Analysis

Lastly, we study the monotonicity property of the trained models. Ideally, a VQA model should satisfy the following property: when the *chroma_qp_offset* parameter is increased, while fixing the other parameters, the predicted quality score M should be monotonically non-increasing. That is, given two *chroma_qp_offset* values $\Delta\text{QP}_{\text{c},1} \leq \Delta\text{QP}_{\text{c},2}$, then we desire that $M_1 \geq M_2$. Similarly, if *chroma_qp_offset* is fixed, the predicted quality score should decrease or maintain at the same level, as the CRF is increased. This would allow the model to be used for constructing bitrate ladders and for calculating BD-rate.

We collected 15 test video contents of 1080p resolution and YUV420 format from Xiph Video Test Media[5] to conduct the experiment. The source videos were encoded at 3 different CRF values: 15, 25, and 35. At each CRF level, we further assigned 8 equally separated ΔQP_c steps, ranging from 0 to 42. As shown in Fig. 13, VMAF 0.6.1 delivered a flat result with respect to ΔQP_c, as should be expected, since its features only ingest luma information. By contrast, the perfectly monotonic plot given by VMAF_c indicates that the additional chromatic features were properly integrated. It may be observed in Fig. 13(b) that the plot for CRF = 35 saturates fast. This is because the quantization parameter in (3) already reaches its maximum value of 51 at $\Delta\text{QP}_\text{c} = 24$. We also demonstrated the merit of this analysis by a fail case in Figs. 13(c), 13(d): this model achieves 0.730 of SROCC on the NFLX_c dataset, which is a substantial improvement as compared with the 0.612 of SROCC of VMAF 0.6.1. Unfortunately, the monotonicity property does not hold individually or on average, which suggests the possibility of overfitting.



(a) VMAF 0.6.1 (b) VMAF_c

(c) Fail model (avg.) (d) Fail model (*blue_sky_1080p25*)

Fig. 13. Monotonicity analysis for variants of trained VMAF models. (a) and (b) are the predictions from VMAF 0.6.1 and VMAF_c (averaged over 15 sequences), respectively. (c) and (d) show failed cases of a model trained with $(C, \gamma) = (2^3, 2^2)$ using the same features as VMAF_c, but without quantization.

[5]https://media.xiph.org/video/derf/

VII. Conclusion and Future Work

We constructed a subjective video quality study and database to support the design of algorithms that can better predict the quality of chromatically distorted videos. This new resource contains HEVC-compressed video contents, spanning wide ranges of quality levels applied differently on a per-channel basis. We also improved an existing high-performing, learning based VQA model (VMAF) by integrating two simple features extracted from chroma channels, and compared the performance of the new chroma-sensitized model against several leading objective VQA algorithms. The new VMAF_c model was found to be the top performer on the new chroma distortion dataset NFLX_c.

We hope that this work encourages increased awareness of ways to address chromatic distortions in the design of quality models, databases, and future compression protocols. The results from our human study indicate that there is room for improvement in the perceptual coding efficiency of modern video codecs, by increasing the compression factor on chroma channels. With the improved video quality model, it is possible to jointly optimize luma and chroma compression in video encoders. For example, the chroma components could be further subsampled or quantized without suffering perceptual fidelity.

Next, we plan to seek new chromatic features that can be used to both effectively capture chroma distortions, as well as preserve performance on luminance distortions. Investigating more sophisticated machine learning models is also of interest. Although compression engines create the most prevalent and common artifacts in streaming video scenarios, there exist other important sources of chromatic distortions, such as subsampling and chroma noise. Creating databases and VQA algorithms that address different kinds of realistic chromatic distortions would be quite valuable. New distortion types emerging from deep video processing problems [65], [66] are also worthy of study. Looking further ahead, developing protocols to optimize video encoders by exploiting improved models of chromatic distortion perception is an intriguing topic of potentially high practical impact.

Acknowledgment

The authors would like to thank A. Norkin and M. Afonso for fruitful discussions on subjective study design. They would also like to thank the Video Algorithms team at Netflix for supporting this work.

References

[1] Cisco Visual Networking Index, "Global mobile data traffic forecast update, 2017–2022," Cisco, San Francisco, CA, USA, White Paper, Feb. 2019.

[2] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," The NETFLIX Tech Blog, Tech. Rep., 2016. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[3] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. D. Cock, "VMAF: The journey continues," The NETFLIX Tech Blog, Tech. Rep., 2018. [Online]. Available: https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12

[4] 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.

[5] G. Björntegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, 13th Meeting ITU-T Video Coding Experts Group (VCEG), Austin, TX, USA, Apr. 2001.

[6] 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.

[7] 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.

[8] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, Feb. 2014, Art. no. 013016.

[9] *SHVC Verification Test Results*, document ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Joint Collaborative Team on Video Coding (JCT-VC), Feb. 2012.

[10] *VQEG HDTV Phase I, Video Quality Experts Group (VQEG)*. Accessed: Apr. 15, 2020. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

[11] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[12] (2012). *Image and Video Quality Resources*. [Online]. Available: http://stefan.winkler.net/resources.html

[13] M. Cheon and J.-S. Lee, "Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 7, pp. 1467–1480, Jul. 2018.

[14] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2620–2630, Oct. 2018.

[15] D. Ghadiyaram, J. Pan, and A. C. Bovik, "A subjective and objective study of stalling events in mobile streaming videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 183–197, Jan. 2019.

[16] R. R. R. Rao, S. Göring, W. Robitza, B. Feiten, and A. Raake, "AVT-VQDB-UHD-1: A large scale video quality database for UHD-1," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 17–177.

[17] P. C. Madhusudana and R. Soundararajan, "Subjective and objective quality assessment of stitched images for virtual reality," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5620–5635, Nov. 2019.

[18] X. Yu, N. Birkbeck, Y. Wang, C. G. Bampis, B. Adsumilli, and A. C. Bovik, "Predicting the quality of compressed videos with pre-existing distortions," 2020, *arXiv:2004.02943*. [Online]. Available: http://arxiv.org/abs/2004.02943

[19] D. Y. Lee *et al.*, "A subjective and objective study of space-time subsampled video quality," *IEEE Trans. Image Process.*, to be published.

[20] X. Shang, J. Liang, G. Wang, H. Zhao, C. Wu, and C. Lin, "Color-sensitivity-based combined PSNR for objective video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1239–1250, May 2019.

[21] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[22] Z. Sinno, A. K. Moorthy, J. De Cock, Z. Li, and A. C. Bovik, "Quality measurement of images on mobile streaming interfaces deployed at scale," *IEEE Trans. Image Process.*, vol. 29, pp. 2536–2551, 2020.

[23] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[24] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, Nov. 2003, pp. 1398–1402.

[25] 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.

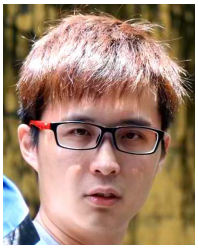
[26] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[27] 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.

[28] M. 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.

[29] S. Wolf, "Variable frame delay (VFD) parameters for video quality measurements," U.S. Dept. Commer., Nat. Telecommun. Inf. Admin., Boulder, CO, USA, Tech. Rep. TM-11-475, Apr. 2011.
[30] M. H. Pinson, L. K. Choi, and A. C. Bovik, "Temporal video quality model accounting for variable frame delay distortions," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 637–649, Dec. 2014.
[31] A. Hekstra *et al.*, "PVQM—A perceptual video quality measure," *Signal Process. Image Commun.*, vol. 17, no. 10, pp. 781–798, Nov. 2002.
[32] P. Tao and A. M. Eskicioglu, "Video quality assesment using M-SVD," *Proc. SPIE*, vol. 6494, Jan. 2007, Art. no. 649408.
[33] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.
[34] 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.
[35] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.
[36] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, and C.-C.-J. Kuo, "Objective video quality assessment based on perceptually weighted mean squared error," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 9, pp. 1844–1855, Sep. 2017.
[37] 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.
[38] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," 2020, *arXiv:2005.14354*. [Online]. Available: http://arxiv.org/abs/2005.14354
[39] I. Lissner, J. Preiss, P. Urban, M. S. Lichtenauer, and P. Zolliker, "Image-difference prediction: From grayscale to color," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 435–446, Feb. 2013.
[40] J. Preiss, F. Fernandes, and P. Urban, "Color-image quality assessment: From prediction to optimization," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1366–1378, Mar. 2014.
[41] 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.
[42] G. Sullivan and K. Minoo, *Objective Quality Metric and Alternative Methods for Measuring Coding Efficiency*, document JCTVC-H0012, ITU-T/ISO/IEC Joint Collaborative Team Video Coding (JCT-VC), 2012.
[43] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Proc. 3rd Int. Workshop Video Process. Quality Metrics Consum. Electron.*, Jan. 2007, pp. 1–4.
[44] J.-M. Valin *et al.*, "Daala: Building a next-generation video codec from unconventional technology," in *Proc. IEEE 18th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2016, pp. 157–162.
[45] M. Buczkowski and R. Stasiński, "Comparison of effective coverage calculation methods for image quality assessment databases," *Intl. J. Elect. Telecomm.*, vol. 64, no. 3, pp. 307–313, Aug. 2018.
[46] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. C. Bovik, "Study of 3D virtual reality picture quality," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 89–102, Jan. 2020.
[47] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Proc. SPIE*, vol. 5007, Jun. 2003, pp. 87–95.
[48] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2019.
[49] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Recommendation BT.500-11, International Telecommunication Union, 2002.
[50] J. W. Peirce, "PsychoPy—Psychophysics software in python," *J. Neurosci. Methods*, vol. 162, nos. 1–2, pp. 8–13, May 2007.
[51] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T Recommendation P.910, International Telecommunication Union, 2008.
[52] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Recommendation BT.500-13, International Telecommunication Union, 2012.
[53] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.
[54] K. T. Mullen, "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings," *J. Physiol.*, vol. 359, no. 1, pp. 381–400, 1985.
[55] N. Sekiguchi, D. R. Williams, and D. H. Brainard, "Efficiency in detection of isoluminant and isochromatic interference fringes," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 10, no. 10, pp. 2118–2133, Oct. 1993.
[56] J. M. Rovamo, M. I. Kankaanpää, and H. Kukkonen, "Modelling spatial contrast sensitivity functions for chromatic and luminance-modulated gratings," *Vis. Res.*, vol. 39, no. 14, pp. 2387–2398, Jun. 1999.
[57] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik, "BBAND Index: A no-reference banding artifact predictor," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2712–2716.
[58] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 2430–2433.
[59] VQEG. (Apr. 2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*. [Online]. Available: http://www.vqeg.org/
[60] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, document ITU-T Recommendation P.1401, International Telecommunication Union, 2012
[61] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.
[62] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, May 2000.
[63] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2020, pp. 141–145.
[64] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 219–234.
[65] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, and Y.-Y. Chuang, "Deep video frame interpolation using cyclic frame generation," in *Proc. AAAI*, 2019, pp. 8794–8802.
[66] H. Ko, D. Y. Lee, S. Cho, and A. C. Bovik, "Quality prediction on deep generative images," *IEEE Trans. Image Process.*, vol. 29, pp. 5964–5979, 2020.

↓ This photo was quantized at QP_c = 51

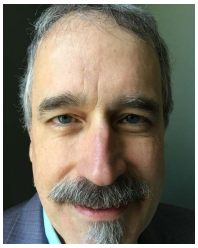


Li-Heng Chen (Graduate Student Member, IEEE) received the B.S. degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2012, and the M.S. degree in communication engineering from National Taiwan University, Taipei City, Taiwan, in 2014. He is currently pursuing the Ph.D. degree with the Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin. He was a Senior Engineer with the Video Encoder Team, MediaTek Inc., from 2014 to 2018. His research interests include perceptual image and video quality assessment, image and video compression, and machine learning.

Christos G. Bampis (Member, IEEE) 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 the 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.

Zhi Li received the B.Eng. and M.Eng. degrees in electrical engineering from the National University of Singapore, Singapore, in 2005 and 2007, respectively, and the Ph.D. degree from Stanford University, Stanford, CA, USA, in 2012. He is currently with the Encoding Technologies, Netflix, Inc. His current research interests include improving video streaming experience for consumers through understanding how humans perceive video and applying that knowledge to encoding/streaming design and optimization. He has broad interests in applying mathematics in solving real-world engineering problems. His doctoral work focuses on source and channel coding methods in multimedia networking problems. He was a recipient of the Best Student Paper Award at IEEE ICME 2007 for his work on cryptographic watermarking, and a co-recipient of Multimedia Communications Best Paper Award from IEEE Communications Society in 2008 for a work on multimedia authentication.

Joel Sole (Senior Member, IEEE) received the M.Sc. degree in telecommunications from the Technical University of Catalonia, Barcelona, Spain, the M.Sc. degree from the Telecom ParisTech, Paris, France, and the Ph.D. degree from UPC. From 2006 to 2010, he was with Technicolor Corporate Research, Princeton, NJ, USA, initially as a Post-Doctoral Fellow and later as a Staff Member and a Senior Scientist. Then, he was a Manager/Senior Staff Engineer with Qualcomm, San Diego, CA, USA, focusing on video coding standards and leading screen content and HDR coding projects. Since 2018, he has been a Senior Research Scientist with Netflix, Inc., Los Gatos, CA, USA.



Alan C. Bovik (Fellow, IEEE) is currently the Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. His research interests include image processing, digital photography, digital television, digital streaming video, and visual perception. For his work in these areas, he was a recipient of the 2019 Progress Medal from The Royal Photographic Society, the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from The Optical Society, 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 received about ten 'best journal paper' awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. His 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, TX, USA, in 1994.