The Dissertation Committee for Zaixi Shang
certifies that this is the approved version of the following dissertation:

Subjective and Objective Quality Assessment for Advanced Videos

Committee:

Alan C. Bovik, Supervisor

Hai Wei

Hyeji Kim

Zhangyang (Atlas) Wang

Edison Thomaz

Subjective and Objective Quality Assessment for Advanced Videos

by
Zaixi Shang

Dissertation

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin
August 2023

Dedication

To Dad, Mom, my family and North.

Epigraph

日就月将，学有缉熙于光明。

——《诗经·周颂·敬之》

With sun's retreat, the moon ascends, each day a gain, each month amends. In endless learning we engage, to find ourselves in a brighter age.

—The Carol of the Zhou Dynasty, the Book of Odes

Acknowledgments

I would like to express my profound gratitude to Professor Alan Bovik for offering me the invaluable opportunity to switch my major and work in the LIVE lab. Your unwavering faith in my potential, understanding and forgiveness of my mistakes during my PhD journey have allowed me to learn, grow, and evolve. I am deeply grateful for your mentorship and for believing in me.

My appreciation extends to my colleagues and alumni of LIVE, who have contributed immensely to my personal and professional development. To Praful, Janice, Xiangxu, Zhengzhong, Li-Heng, Dae-Yeol, Pavan, Zhenqiang, Somdyuti, Abhinau, Haoran, Yize, Meixu, Xiongkuo, Zhaolin thank you for your generous assistance, engaging discussions, and unwavering support. Our intellectual exchanges have been an integral part of my journey. My gratitude extends to the energetic, new generation of the lab - Avinab, Sandeep, Berrie, Shresth, Seobin, Philip, Karthik, Hakan, Cheng-Han, Bowen, Asvin, Ramit, and Krishna. Your vigor and enthusiasm have breathed new life into the lab, and I am eager to witness the great strides you will make in the future. Among this remarkable group, I extend a special thanks to my closest colleague, Josh. Your unwavering support, constructive feedback, and shared dedication towards our work have been instrumental to my growth and the successful completion of our projects. The shared intellectual exc

A special note of thanks to my collaborators at Amazon Prime Video, especially Yongjun, Hai, Sriram, and Yixu, for their guidance on my projects and during my internship. The experiences and insights I have gained under your expert mentorship have been invaluable.

I would be remiss if I didn't extend my appreciation to my Basspass group. Your camaraderie and shared enthusiasm for musicals have provided much-needed respites from my daily work. Our shared experiences have been not just an escape,

but a source of joy and laughter that made my journey brighter. Thank you for your friendship and the memorable times we've had together.

To my cat, North, your quiet presence and tolerance of my temper during times of high pressure have been more comforting than words can express. You have been a constant source of solace, providing me with a sense of calm and companionship, particularly during challenging times. Thank you, North, for being there, and for helping me maintain my balance amidst the highs and lows of this journey.

My sincere thanks go to the dedicated staff at the Wireless Networking and Communications Group (WNCG) and the Department of Electrical and Computer Engineering (ECE). Your tireless efforts behind the scenes have ensured the smooth running of our programs and projects, providing us with an environment conducive to our research.

I also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC, visualization, database, and grid resources that have contributed to the research results reported in this paper. URL: http://www.tacc.utexas.edu.

Finally, my deepest gratitude goes to my parents. Your love, support, and unwavering faith in me have been my driving forces. Your sacrifices and unyielding support have made this milestone possible. I am forever grateful.

Abstract

Subjective and Objective Quality Assessment for Advanced Videos

Zaixi Shang, PhD
The University of Texas at Austin, 2023

SUPERVISOR: Alan C. Bovik

The surge of video streaming services, particularly for high motion content such as sporting events, necessitates advanced techniques to maintain video quality, facing challenges such as capture artifacts and distortions during coding and transmission. The advent of High Dynamic Range (HDR) content, offering a broader and more accurate representation of brightness and color, poses additional complexities due to increased data volume. The critical need for robust Video Quality Assessment (VQA) models arises from these challenges.

To meet this need, we conducted three substantial subjective quality studies and constructed corresponding databases. The Laboratory for Image and Video Engineering (LIVE) Livestream Database comprises 315 videos of 45 source sequences from 33 original contents impaired by six types of distortions. This database facilitated the gathering of over 12,000 human opinions from 40 subjects. The LIVE HDR Database, the first of its kind dedicated to HDR10 videos, includes 310 videos from 31 distinct source sequences, processed with ten different compression and resolution combinations. This resource was instrumental in amassing over 20,000 human quality judgments under two different illumination conditions. An additional LIVE HDR AQ was developed with 400 videos from 40 unique source sequences. These videos were

processed using varied compression, resolution combinations, and AQ-mode settings, to study the effects of adaptive quantization (AQ) and rate-distortion optimization techniques on HDR video perceptual quality.

Building on these invaluable databases, we developed two innovative objective quality models: HDRMAX and HDRGREED. HDRMAX, a pioneering framework designed to create HDR quality-sensitive features, augments the widely-deployed Video Multimethod Assessment Fusion (VMAF) model, yielding significantly improved performance on both HDR and SDR videos. HDRGREED, a novel model leveraging localized histogram equalization and Difference of Gaussian filters, employs the Generalized Gaussian Distribution to model the bandpass responses and measure the entropy variations between reference and distorted videos. This model is particularly sensitive to banding and blocking artifacts introduced by inappropriate AQ settings.

In conclusion, the comprehensive subjective quality studies and databases, along with the state-of-the-art objective quality models, HDRMAX and HDRGREED, significantly contribute to the advancement of future VQA models. These tools cater specifically to challenges posed by live streaming and HDR content, providing critical resources for the development, testing, and comparison of future VQA models. These databases, publicly available for research purposes, and the innovative models offer valuable insights to improve and control the perceptual quality of streamed videos.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Video traffic now occupies more than 70% of all total downstream Internet traffic and is still expected to grow For; glo (2018). Major content providers such as Amazon Prime Video, YouTube, Netflix, and Hulu are providing increasing amounts of video on demand (VoD) content, as well as live streaming videos, to an expanding audience. live streaming, which is real-time audio and video transmission of live events, is gaining popularity very rapidly, especially for sporting events like the Super Bowl Chen and Lin (2018).

Although significant efforts have been made to enable the delivery of high-quality, high-resolution VoD, little effort has focused on live, high motion video streaming. In live streaming, there are still a variety of factors that can adversely affect the quality of live streaming videos. For example, bandwidth and stability may affect the received video source quality, causing distortion like blocking, banding, deinterlacing motion mismatches, local flicker Ni et al. (2011), aliasing and interpolation artifacts Keating (1993). If the network connection is unstable or the bitrate inadequate, then frame drops may also occur. The videos may be distorted by stutter or motion blur, especially when there is rapid motion. By contrast with VoD streaming, a large portion of live streamed content is still interlaced and then deinterlaced, causing combing effects, flicker or noticeable line movements.

Video impairments like these can severely impair the delivered video quality and users' holistic levels of visual satisfaction. This is a pressing problem for high motion, action content such as sports videos. high motion videos generally contain richer temporal information and are harder to compress, hence compression artifacts are often more severe in sports videos. Other distortions can also be exacerbated by high motion. For example, at lower frame rates, high motion sports may appear discontinuous over time, and may exhibit obvious judder. Likewise, high motion can worsen the visual appearance of interlacing, causing jagged moving edges.

The human visual system (HVS) is able to perceive luminance levels between 10^{-6} cd/m^2 and 10^8 cd/m^2 using various mechanical, photochemical, and neuronal adaptive processes Kunkel et al. (2016). Traditional imaging and display systems produce content having much narrower ranges of luminance values than the vision system is able to perceive, due to limitations on sensor technology, processing, transmission, bandwidths, and display depths. These older content formats are commonly referred to as Standard Dynamic Range (SDR), and have specifications on brightness, contrast, and color that were originally designed for display on cathode ray tube (CRT) devices ITU (2011). Although CRTs are obsolete, a considerable fraction of content continues to be produced according to SDR specifications. A device that displays SDR content, which has a bit depth of 8 bits/channel, can represent a maximum luminance of 100 nits (1 nit = 1 candela/meter^2) and a minimum luminance of 0.1 nits, using the Rec. 709/sRGB color gamut ITU (2011), which covers 35.6

High Dynamic Range (HDR) is a set of techniques that extend the ranges of luminances and color that can be represented and displayed. "HDR" pictures are sometimes synthesized by combining photographs taken at multiple exposures into a single picture, then tone-mapping it to the 8 bit range that is compatible with SDR displays. What we will refer to as "true HDR" video content is captured using single exposures with advanced sensors, and compatible with HDR displays having wider dynamic ranges and higher average and peak brightness levels. True HDR content has a bit depth of at least 10 bits/channel. HDR10 is an open HDR standard announced by the Consumer Technology Association in 2015 CTA and remains the most widely used HDR format. HDR10 content must have a bit-depth of 10 bits, use the Rec. 2020 ITU color primaries (which cover 75.8% of the CIE 1931 color space), and must apply the SMPTE ST 2084 Standard (2014) opto-electronic Transfer Function (OETF) to the linear RGB signals, also known as the Perceptual Quantizer (PQ).

HDR10 has seen increasing adoption over the past few years. Streaming and video hosting services such as Amazon Prime, Netflix, and YouTube now offer content in HDR10. HDR10 is also used as the default standard for UHD Blu-Rays. Major

TV manufacturers such as LG, Samsung, and Panasonic support HDR10 content, and manufacturers such as Lenovo and Apple have also recently released laptops that can display HDR10 content. HDR10 is now part of live broadcast and film production workflows and is progressing rapidly into an industry standard.

The adoption of HDR10 has created challenges related to the quality of user experience and the performance of compression algorithms. The increases in bit depth and the use of nonlinear transfer functions in HDR can change the visibility and severity of compression distortions. Being able to measure and control perceptual quality is a critical element of video compression and communication workflows. However, there are few video quality assessment (VQA) models that address the compression of HDR videos. Most existing VQA models can only operate on 8 bit luminance and color data, let alone account for HDR transfer functions and expanded color gamuts. For example, one of the most successful VQA models, the Video Multimethod Assessment Fusion (VMAF) algorithm Li et al. (2017) can be applied to 10 bit data, but it does not take into account the extended luminance range or transfer function of HDR10.

An important consideration is the nonlinear visual response to brightness. Because the vision system is more sensitive to luminance ratios than to absolute brightness values, the perception of differences between luminances is governed by the Weber-Fechner law Cornsweet and Pinsker (1965). The exponential function or "gamma," as specified in the industry standard BT. 709, has been traditionally applied to nonlinearly encode SDR images, but it fails to work with HDR imaging, due to the mismatch of quantization and human perception. Therefore, SDR VQA models, which operate under the assumption of gamma, are less effective on HDR content. This does not imply that SDR VQA models, developed under the assumption of gamma, are always ineffective for HDR content. Several studies, such as Sugito et al. (2022); Krasula et al. (2023), have demonstrated that these models can perform competitively even when applied to HDR content, depending on other aspects of the content, or the device it is displayed on, suggesting a nuanced landscape Mantiuk and

Azimi (2021); Mikhailiuk et al. (2022). Furthermore, the perception of brightness distortions is influenced by the viewing conditions, including the image background, the environmental light, the peak luminance, and the dynamic range of the display.

Most HDR videos are encoded using the High Efficiency Video Coding (HEVC) standard. Adaptive quantization (AQ) is a technique used in HEVC to improve the quality of the encoded videos. This is accomplished by adjusting the quantization parameter for each coding block in the video frame, typically allowing for more data to be allocated to areas of the frame that contain complex visual information, and less data to be allocated to areas with less complex information. While this can result in significant savings in terms of the amount of data required to represent the video, it can also result in worse video quality. This is because the smooth areas of the video, which are allocated fewer bits, can suffer from distortions such as banding and blocking. These distortions are particularly noticeable to the human eye because of the lack of contrast masking and can be quite annoying, damaging the overall quality of the video.

The aq-mode option in most HEVC encoder, such as libx265, enables the use of AQ in the encoder, which adjusts the quantization level on a per-block basis based on the complexity of the source image. This means that more quantization is applied on complex areas of the video, and less quantization on smooth areas. This can help to offset the tendency of the encoder to spend too many bits on complex areas and not enough in flat areas, which can cause distortions such as banding and blocking. By enabling this option, the encoder can better balance the allocation of bits across the video frame, resulting in improved visual quality and reduced artifacts.

In this paper, we introduce a new resource to enhance the understanding and development of video quality assessment (VQA) models for high motion, live-streamed sports content. This novel database, named the LIVE Livestream Database, is specifically built with a diverse collection of high definition videos, featuring a total of 315 videos derived from 45 source sequences from 33 original contents. What distin-

guishes this database from prior VQA databases is its inclusion of Full High Definition (FHD) and Ultra High Definition (UHD) content captured by professional videographers, which are impaired by six types of common processing distortions reflective of real-world scenarios.

With the aim of determining the perceptual quality of these live-streamed videos, we engaged a large pool of volunteers in a human subjective study. By presenting the aforementioned videos to these volunteers, we obtained Mean Opinion Scores (MOS) –a vital component in comprehending viewer experiences. Moreover, we capitalized on this newly created database to carry out a comprehensive evaluation of the performance of current state-of-the-art VQA models. Not only did this provide a comparative performance analysis but it also offered insights into potential future challenges in live streaming VQA.

Furthermore, our exploration extends to the domain of high dynamic range (HDR) video quality prediction. We present a new HDR-specific video feature framework, HDRMAX, which is used to modify the extensively validated and commercially successful VMAF model. By supplementing the VMAF model with HDRMAX features, we enhance its sensitivity to expanded luminance ranges, transfer functions, and large color gamuts inherent in HDR video formats.

Lastly, we propose a novel HDR-VQA model, coined as HDR-GeneRalizEd Entropic Difference (GREED). The HDR-GREED model exploits localized histogram equalization (LHE) and difference of Gaussian (DoG) filters to discern distortions on smooth areas, thereby augmenting the model's sensitivity to compression artifacts. Specifically, these bandpass filters are tailored to respond to banding distortions, which paves the way for a more precise and comprehensive assessment of video quality.

Chapter 2: Study of the Subjective and Objective Quality of High Motion Live Streaming Videos

2.1 Related Work

Over the past decade, there have been many efforts to build subjective video quality databases. Among those, the LIVE VQA Database Seshadrinathan et al. (2010) includes 10 pristine videos processed with compression and packet loss distortions. Similarly, the later database in De Simone et al. (2010) contains 156 videos modified by H.264 compression artifacts and wireless packet losses. The LIVE QoE Database for HTTP-based Video Streaming Chen et al. (2014) studies the quality of experience (QoE) of users who viewed compressed videos with simulated video stalls, which can arise when there is low channel throughput. This database models the perception of video quality on mobile devices, and the human study was performed on mobile phones and tablets. Another QoE database proposed in Duanmu et al. (2017) aims to motivate QoE prediction in video streaming, with different bitrate levels and stalling events. Among 20 1080p source sequences, 5 videos contain high motion content. Another databaseDe Simone et al. (2011) studied H.264 compressed videos transferred through an error-prone network, including 156 sequences at CIF and 4CIF spatial resolutions. The LIVE Mobile Video Quality DatabaseMoorthy et al. (2012b) consists of 200 distorted videos created from 10 RAW HD reference videos, including compression and wireless packet-losses, with dynamically varying distortions. The MCL-V databaseLin et al. (2015) was designed for streaming video quality assessment, and contains 12 source video clips and 96 distorted video clips impaired by H.264 compression, as well as compression followed by spatial scaling. The TUM databasesKeimel et al. (2010); Keimel et al., contain several synthesized videos with H.264 compression. Other exemplars include the MCL video quality databaseWang et al. (2017), ECVQ and EVVQVranješ et al. (2013), and the Poly@NYU Video

Quality DatabasesOu et al. (2010, 2014).

More recently, novel databases have been introduced that contain user-generated-content (UGC) videos with authentic distortions. The LIVE-VQC databaseSinno and Bovik (2018) contains 585 videos, all of unique contents captured by a large group of users deploying various camera devices, including smartphones of all brands. The LIVE-VQC videos cover a wide range of qualities, and include complex, often commingled authentic distortions. The large KoNViD-1k Hosu et al. (2017) video quality database contains 1,200 video sequences, covering a wide variety of contents and authentic distortions. The YouTube UGC DatasetWang et al. (2019) contains 1500 20-second video clips covering popular UGC video categories, including gaming and sports.

A number of deficiencies limit the usefulness of all of these databases for the study of the quality of live video streams. Older, legacy databases contain only limited numbers of SD source contents, which are not representative of current high-resolution live streaming. Although most databases consider compression distortions and packet loss, other prevalent distortions common to live streaming videos are rarely found in them. Given exploding interest in live streaming video, a comprehensive database that includes both ample video content and representative live streaming distortions is needed.

UGC video quality databases usually include a large number of contents, but there is a lack of professionally captured content, and the distortions encountered in live streaming often significantly differ from those caused by typical casual social media users. The only existing publicly available VQA database designed for live streaming is the LIMP Video Quality Database Vega et al. (2016). The LIMP database consists of nine high-quality videos taken from the LIVE Video Quality Video DatabaseSeshadrinathan et al. (2010), with simulated compression modeling transmitted in a controlled network. However, it suffers from the same problems mentioned above. Motivated by an apparent dearth of live streaming databases containing

enough high-resolution video contents and sufficiently representative live streaming distortions, we have created a large new resource intended to address modern aspects of the live sports streaming video quality problem.

2.2 Relevance and Novelty

In recent years, the streaming of live high motion video content such as sports has explodedFor. Live streaming high motion videos often suffer from severe distortions less often encountered in the streaming of generic content. In live streaming, considerations of network instability and bandwidth limitations imply greater challenges when attempting to control video quality. Moreover, the real-time requirement greatly limits the time available for post-processing to compensate for defects. The unique nature of live streaming introduces many obstacles that differ from those encountered in generic on-demand video streaming. For example, sports videos usually include content containing complex, large motions. Rapid and irregular camera motions occurs frequently, when tracking moving objects, such as balls or players. Temporal distortions often arise that are annoying and that adversely affect the viewer experiences.

The new psychometric database that we describe here has a number of unique attributes. It contains a larger number of unique source contents and distortion types. We summarize the attributes of public video quality databases in Table **??**. The new database includes 45 source sequences token from 33 unique contents. All of the videos contain complex, fast motions, which are rarely included in existing databases. The new resource contains a wide variety of distortion classes common to sports live streaming content that is not found in existing VQA databases. Although LIVE-Flicker and Live-Mobile include specific temporal distortions such as flicker or frame-freeze, neither contains a holistic collection of high motion, live streaming distortion types.

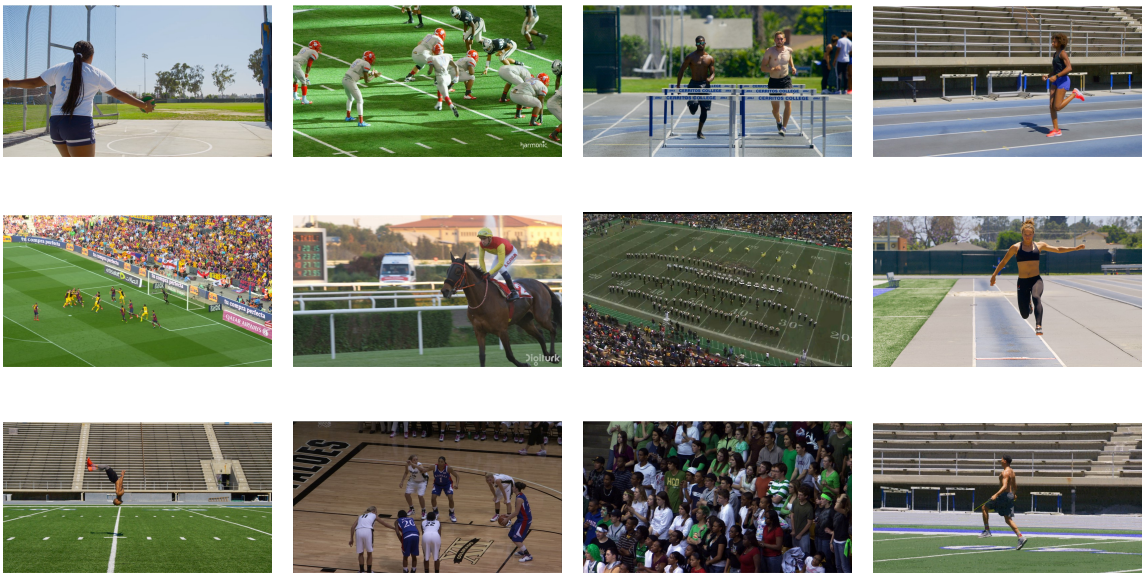
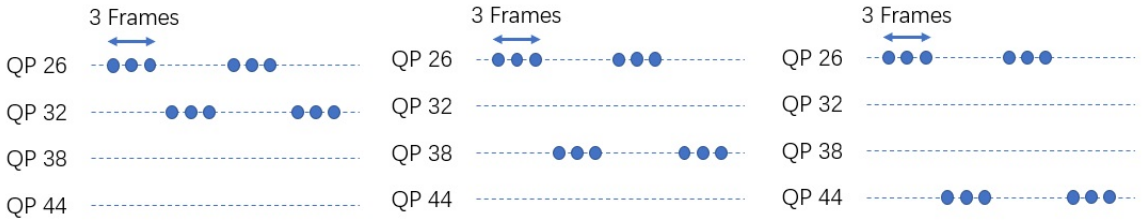


Figure 2.1: Exemplar screenshots of frames from source videos in the LIVE Livestream Database.

(a) (b) (c) (d)

Figure 2.2: Simulation of motion judder from 3:2 pulldown. (a) Original frames at 23.94 fps. (b) Odd video fields. (c) Even video fields (d) Resulting frames at 29.97 fps



(a) Flicker level 1 (b) Flicker level 2 (c) Flicker level 3

Figure 2.3: Three levels of flicker synthesis.

2.3 Details of subjective study

We constructed a new video quality database that consistent of 315 video sequences including 45 reference videos and 6 copies of synthetically distorted version of each reference video. Those videos are used as stimuli in the subjective study.

2.3.1 Source Sequences

We collected 33 uncompressed, high-quality source videos with sports content. These videos are freely available online from multiple sources, including from Tampere UniversityMercat et al., the MCML GroupCheon and Lee (2017), the Netflix Public DatasetLi et al. (2016), the VQEG HD3 DatasetVideo Quality Experts Group (2000 (accessed October 31,2020), the Consumer Digital Video Library (CDVL)Yodel (2011), and the SJTU Media LabSong et al. (2013). All of the selected videos were captured with professional, high-end camera equipment and are distortion-free. The original pristine videos all have resolutions of 1920x1080 or 3840x2160 pixels, and were progressively scanned in YUV 4:2:0 format with audio components removed. The videos have frame rates at 30 fps. The video contents include 10 different types of sports, including running, football, and soccer, and one video of the audience in a stadium, as exemplified in Fig. 2.1.

The original 33 videos that we collected are of durations ranging from 5s to 26s. However, since viewing videos of such differences of durations could cause biases

in subjective and objective judgments, longer videos may exhibit visible changes of distortion over time. While the effects of video duration is interesting and worthy of study, this also would increase the dimensionality of the study. Thus, we manually cropped the longer videos along the temporal dimension into one or two shorter clips of about 7 seconds with no overlap or close proximity between the clips. Based on internal studies at UT-LIVE, it has been observed that very short videos of sports videos may cause annoying content disruptions, such as incomplete "play," but these events are usually shorter than 8s. To avoid unpleasant cuts during action scenes, we allowed some flexibility of the video durations, hence the final set of original videos had lengths in the range 5s-8s, averaging 7.88s with a standard deviation of 1.36s. In this way, 45 video clips were created from the 33 originals, of which 22 clips are of resolution 1920x1080 and 23 clips are of resolution 3840x2160.

2.3.2 Synthetic Distortions

We created 6 distorted video sequences from each of the pristine sequences, using six different distortion processes. These included H.264 compression, aliasing, judder, flicker, frame drops, and interlacing. Since our primary goal is to model the visual quality high motion live sports videos, the distortions chosen were judged to be the most common and salient ones that are encountered during live sports events. During live streaming of high motion contents, certain distortions may produce more severe effects than on more generic video content. For example, a moving object may cause large pixel offsets between neighboring frames or fields. If the video is interlaced, then severe edge combing and blur may occur. If the frame rate is too slow, then judder from 3:2 conversionDaly et al. (2015); Oh et al. may be visible in high motion regions, which can seriously and adversely impact the appearances of sports videos. Purely temporal distortions, such as frame drops, which cause discontinuities and motion stalls, are difficult to detect.

When applying different levels of each distortion type, we sought to ensure that

the distorted videos would be both perceptually separable and also cover a wide range of perceptual qualities, following successful practice in numerous previous studies Seshadrinathan et al. (2010); Moorthy et al. (2012a); Sinno and Bovik (2018). However, given the large number of source sequences, it is not practical to include multiple copies of the same content, which can greatly increase the duration of the human study. Moreover, having larger number of unique contents can contribute to improved model building. Hence, given the fairly large number of source videos, we dictated that each would only have a single level of severity of each distortion type applied to it. For example, four levels of H.264 compression, corresponding to different constant rate factors (CRF) were defined. This was accomplished in a "round robin" sequential manner: the first reference video could only be compressed using the first CRF level, the second reference was only compressed using the second CRF level, and so on. The fifth source video then had the first level of distortion applied. However, to ensure that there would be no content-related quality bias, the first video in the quality level cycle was also sequenced as subsequent distortions were applied. In this way, each of the 45 clips taken from the original 33 pristine source videos has 6 associated distorted versions of it, yielding 315 videos including the 45 reference videos.

2.3.2.1 H.264 Compression

H.264 remains the most widely-accepted and used video compression standard. A 2020 streaming industry survey202 (2020) found that 91% of streaming services use H.264. Although newer codecs exist, such as HEVC, VP9 and AV1, they are not yet as widely adopted. Browsers and devices also don't have full support for all codecs. The Apple Safari browser supports HEVC, but not VP9, while Chrome and Firefox support VP9 and AV1, but not HEVC. All browsers support H.264. Hence, when designing this VQA database, we deemed H.264 to be most representative of current practice. Moreover, even emerging standards still follow the basic hybrid codec method of distortion, viz., quantization of DCT blocks, while several distortions are

not compression-related. Hence, we believe that the new database will retain usefulness as the compression standards evolve. We fixed four levels of H.264 compression using the criteria described earlier, by varying the CRF values. Similar to other successful VQA databasesSeshadrinathan et al. (2010); De Simone et al. (2010); Lin et al. (2015), we included a wide range of compression CRFs to ensure that the distorted videos cover a wide range of perceptual qualities, while also ensuring perceptual difference between the applied compression levels, to allow for improved model-building. Since in practice, the compression parameters differ on videos of different resolution, we selected different sets of CRFs for the 4K videos and the 1080p videos. The CRF values selected for the 4K videos were 9, 27, 39, and 43, while those for 1080p videos were 9, 25, 35, and 39. All of the compressed videos were generated using FFmpeg.

2.3.2.2 Aliasing

Aliasing was simulated by first downscaling each video, then upscaling it back to its original dimensions. The downscaling was performed by spatially downsampling the video to half the original size without the use of an anti-aliasing filter, while the upscaling was performed using a Lanczos filter.

2.3.2.3 Judder

Motion judder is an artifact that is introduced when scenes shot at 23.94 fps are converted to 29.97 fps by a process called 2:3 pulldown. The ratio of these frame rates is 4:5: for every 4 input frames, 5 output frames were created by temporally downsampling the video to 23.94 fps, then converting the frame rate to 29.97 by 2:3 pulldown. The odd video field of every 2^{nd} frame, and the even video field of every 3^{rd} frame of each group of 4 frames were combined to form an additional frame, for each group of 4 frames. This process is shown in Fig. 2.2. Classic 2:3 pulldown followed a slightly different pattern where the 2nd and 3rd frames of the original video would be interlaced to form the 3rd frame of the juddered video, and the 4th

and 5th frames of the original video would be interlaced to form the 4th frame of the juddered video. This had the disadvantage of producing two "dirty" frames, which were the 3rd and 4th frames in each group, but was used in legacy systems where the buffer could not hold fields from more than one frame at a time. The version we use here is a more advanced pulldown, supported by cameras released after 2000 such as the Panasonic DVX100 pan or the Canon XL2 can. The more advanced version of pulldown generates only one "dirty" frame and also allows for better compression and easier conversion back to 23.94 fps.

2.3.2.4 Flicker

We simulated flicker distortion from compression by alternating the H.264 quantization parameter (QP) on the video. The QP is fixed at a constant value by passing this parameter to libx264. These QP values were applied to each frame, regardless of the frame type, content and motion. Three pairs of QPs were chosen to form three flicker distortion levels: QP26 and QP32, QP26 and QP 38, and QP26 and QP44. The flicker rate, which is the number of QP alternations per second, was kept a constant roughly 5 Hz i.e. by alternating the QP every 3 frames. This process is depicted in Fig. 2.3.

2.3.2.5 Frame Drops

We simulated video frame losses that occur when a source video is transmitted over a channel, such as a wireless network. We simulated frame drop clusters of adjacent frames to account for 10%-30% of a group of pictures (GOP). When a cluster of frames was removed from a video, the previous frame was repeated as many times as needed so that the total video duration remained unchanged. Three levels of frame drop densities were chosen: 3, 6 and 9 frames per cluster, yielding a slight to severe impact on the perceptual qualities of the videos.

2.3.2.6 Interlacing

On each frame of the video, the even and odd lines were separated to form two fields, field A and field B. Field B from each current frame and field A from each next frame were then combined to create interlaced frames. In the presence of motion, combing effects become evident. Since interlaced video fields are captured at different moments in time, interlaced frames often exhibit motion combing artifacts, when objects move quickly enough to be at different positions in each field.

2.3.3 Subjective Testing Environment and Display

The human study was carried out in the LIVE Subjective study room at The University of Texas at Austin. The Lab was arranged to simulate a living room environment. The windows were covered, and background distractions were removed. A Samsung UN65RU7100FXZA Flat 65-Inch 4K UHD TV was used to display all of the videos. All advanced motion optimization options on the TV, including the anti-judder and anti-flicker functions, were disabled. The viewing distance was about 2H, where H is the height of the TV so that the subjects could comfortably view the videos and assess the video distortions. The level of illumination was set to be similar to a living room, using one stand-up incandescent lamp and two indirect white LED studio lights behind the viewer. The lights were positioned to eliminate reflections from the lights on the screen.

Since the TV is able to upscale 1080p content using an unknown algorithm, all of the 1080p videos were instead upscaled using the Lanczos resizing function in OpenCV Bradski (2000), to avoid any unpredictable effects. The 1080p videos were upscaled to 4K, after the distortions were applied. To ensure perfect playback, all of the videos were stored as raw YUV 4:2:0 files. The powerful Venueplayer application developed by VideoClarity was used to guarantee smooth playback of the 4K videos, without introducing any additional artifacts that could impact the perception of video quality.

After displaying each of the test videos, a continuous rating bar was displayed on the screen with a randomly placed cursor. The quality bar was marked with labels "Bad," "Poor," "Fair," "Good," and "Excellent" quality to facilitate the subjects in making decisions. The scores given by the subjects were sampled as integers from [0, 100] although numerical values were not made visible to the subjects. A Palette gear console was provided to enable the subjects to move the cursor without distraction. After moving the cursor to each desired scoring position, the subject depressed the button next to the sliding bar to confirm the score, which was then recorded without any further change. After each score was stored, the system immediately began to play the next video on the playlist.

2.3.4 Subjective Testing Design

In the human study, a single-stimulus (SS) method was employed, as described in the ITU-R BT 500.13 recommendationITU (2012). The reference videos are included as "hidden reference", not explicitly marked as "distorted" or "reference." The subjects used a rating bar to record their subjective opinion scores. Video rating scores were given after watching each video on an (invisible) scale ranging from 0 to 100, where 0 indicates the worst quality and 100 indicates the best quality. Due to the large number of video sequences, each subject participated in two sessions. The 45 contents associated with the pristine videos were divided into two sessions, where the reference videos and their corresponding distorted versions were grouped into the same session. The playlists within each of the two sessions were placed in randomized order for each subject, where videos of the same content, were separated by at least one video. This was done to counter any visual memory effects that might affect the subjective quality judgments, or any bias caused by playing the videos in a particular order. Each session required about 40 minutes.

2.3.5 Subjects and Training

A total of 40 human subjects were recruited from the student population at The University of Texas at Austin. The male/female gender ratio of the subject pool was 4.0. The mean and standard deviation of the ages of the participants was 23.47 and 1.78. Each subject participated in two sessions separated by at least 24 hours. Two of the subjects finished only one of the two sessions, while the rest of the 38 human subjects finished both sessions. 180 of the videos were rated by 40 subjects, while 187 videos were rated by 38 subjects. The subject pool was inexperienced with the topic of video quality assessment and video distortions.

The Snellen test and the Ishihara test were performed to validate each subject's vision. Two subjects were found to have 20/30 visual acuity, while one subject was found to have a color deficiency. However, these subjects were allowed to participate since the overall subject pool was deemed to be a good representation of the general population, following our common practice lan. We conducted the tests as a screen against an unusual percentage of deficient subjects. Before the study, each subject was presented with a brief introduction to the study. The introduction described the study's goals, and gave detailed instructions on how to operate the system and assign scores. Each subject was asked to rate each video by quality only, without regard to the appeal of the content. Before the actual study commenced, each subject participated in a training session on two videos, to familiarize themselves with the system. The training videos and their scores were not included in the final database.

2.4 PROCESSING OF SUBJECTIVE SCORES

Subjective Mean Opinion Scores (MOS) were computed using the formulas below: Let s_{ij} denote the score by subject i for the video j. The subject scores were then converted into Z-scores z_{ij} for each subject. Subject rejection was performed based on the ITU-R BR 500.11 recommendation ITU (2012). The scores z_{ij} for each video

were tested against the normal distribution using the β_2 test:

$$\beta_{2j} = \frac{m_4}{(m_2)^2}, \tag{2.1}$$

where

$$m_x = \frac{\sum_{i=1}^{N}(z_{ij} - \overline{z}_{ij})^x}{N_j} \tag{2.2}$$

for subject i and video j, where N_j is the number of subjects that viewed video j. A score was regarded as normally distributed if β_{2j} fell between 2 and 4. We calculated the quantities P_i and Q_i for each subject i, by comparing z_{ij} with the mean $\bar{z_j}$ standard deviation σ_j of video j: If the score for video j was found to be normally distributed then:

if $z_{ij} \geq \bar{z_j} + 2\sigma_j$, then $P_i = P_i + 1$

if $z_{ij} \leq \bar{z_j} - 2\sigma_j$, then $Q_i = Q_i + 1$

If the score for video j was found to not be normally distributed, then:

if $z_{ij} \geq \bar{z_j} + \sqrt{20}\sigma_j$, then $P_i = P_i + 1$

if $z_{ij} \leq \bar{z_j} - \sqrt{20}\sigma_j$, then $Q_i = Q_i + 1$.

A subject i was rejected if the following two conditions held:

$$\frac{P_i + Q_i}{N} > 0.05, \tag{2.3}$$

and

$$\left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3. \tag{2.4}$$

In our study, 8 of the 40 subjects satisfied these two conditions. However, since most of the rejected subjects fell close to the decision boundaries, we decided to revisit how the rejection criteria should be used. Given that the intent of subject rejection is to eliminate the outcomes of less engaged, distracted, or otherwise deficient subjects, we believed it worth considering whether any of the high-deviation subjects were actually representative, as we have done in other recent studies Madhusudana

Table 2.1: Internal Consistency

	PLCC mean	PLCC median
All subjects	0.9616	0.9632
Reject Group 2 & 3	0.9603	0.9618
Reject Group 3	0.9635	0.9647

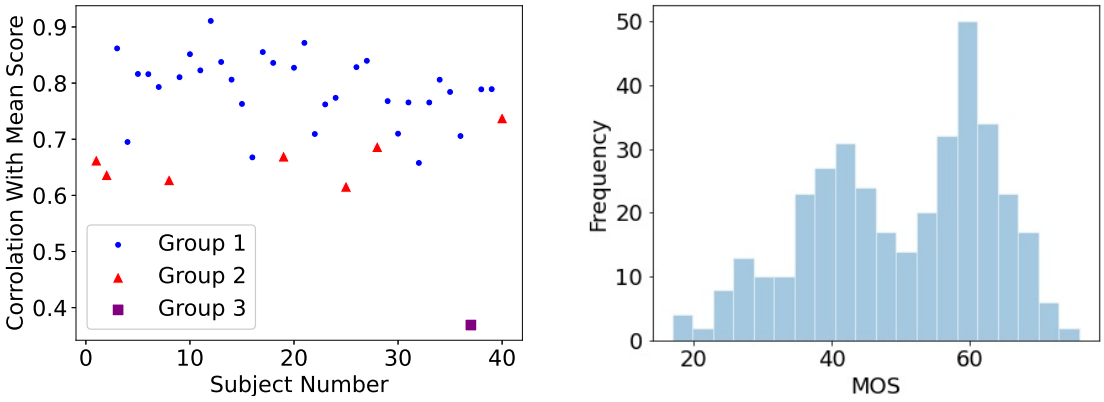
et al. (2020). We therefore computed the correlations between each subject's score and the MOS calculated using three different variations of the rejection criterion: 8 rejected, none rejected, and 1 (most anomalous) subject rejected, as shown in Fig. 2.4a. Specifically, the subjects were divided into three groups: Group 1 included all subjects not excluded by the ITU method. Group 2 and Group 3 included only the 8 subjects that were rejected, while Group 3 considered only of the single subject having the worst correlation against MOS. In the end, we chose to report all of the foregoing results by only excluding the single subject in Group 3.

Table 2.1 shows our analysis of the data's internal consistency. Our modification of the typical outlier rejection criterion finds support in the analysis, and allows for a larger amount of likely representative data for model-building. We randomly divided the subjects into two equally sized groups and computed the Pearson correlation coefficient (PLCC) between the two groups' scores. We repeated this calculation over 1000 results, and report the mean and median correlations in Table. 2.1. As may be seen, the best results were attained by removing the single very anomalous subject. We also observed negligible effect of the choice of rejection criteria on the objective algorithm performances reported later.

The Z-scores were then linearly rescaled from [-3,3] to [0,100]:

$$z'_{ij} = \frac{100(z_{ij} + 3)}{6}. \tag{2.5}$$

Finally the Mean Opinion Score (MOS) of each video was calculated:



(a) (b)

Figure 2.4: (a) The correlations against MOS of all subjects. Group 1 are subjects not rejected by the ITU method. Group 2 are all subjects that were rejected by the ITU method. Group 3 consists of the single lowest-correlating subject. (b) Distribution of MOS over all videos in the LIVE Livestream Database.

$$MOS_j = \frac{1}{N_j} \sum_{i=1}^{N_j} z'_{ij}. \tag{2.6}$$

The converted MOS score is shown in Fig. 2.4b.

Fig. 2.5 shows the distributions of scores for each individual video distortion class. The shapes of the MOS distributions of the reference videos are more Gaussian-like. The distorted video classes exhibit different distribution shapes, since they reflect different types and levels of distortion. Further, the MOS of the different levels of compression, flicker, and frame drop distortions are shown in Fig. 2.6. Generally, the MOS ranges of different distortion levels are mostly well-separated, but there are overlaps between distortion levels, largely because of the different interactions that occur between content and distortion. The perceptual quality of distorted (compressed videos) is affected by content masking, e.g. in regions containing significant high frequency spatial energy or high motion. While spatial masking is well-understood,

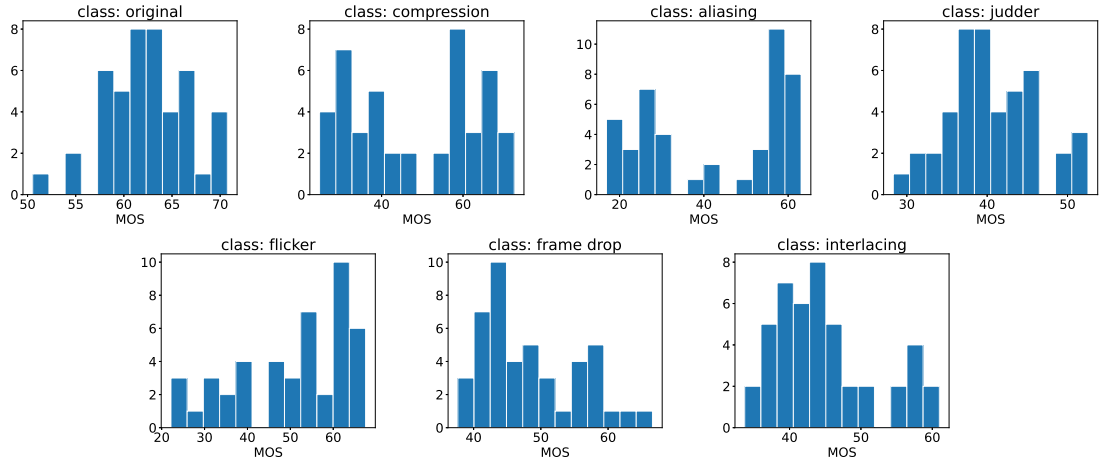
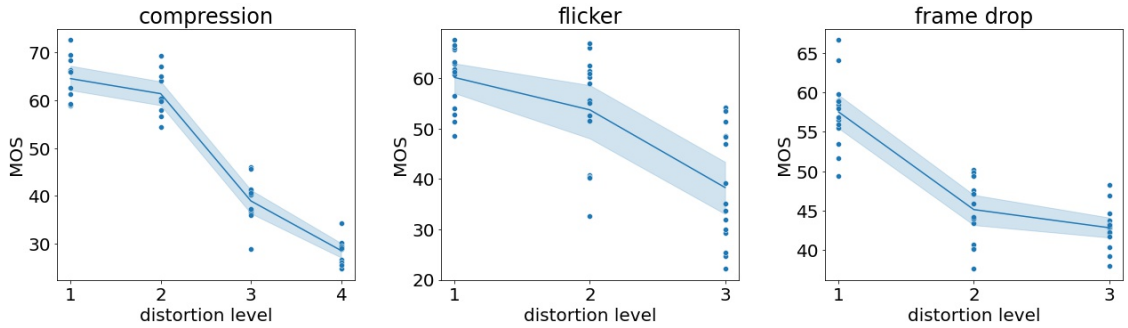


Figure 2.5: Distribution of MOS of original, and synthetically distorted videos.



(a) (b) (c)

Figure 2.6: MOS of videos affected by compression, flicker, and frame drops. Each dot in the plot is a video and the line and the shadowed regions indicate the average MOS and the 95% confident interval. (a)Compression, (b) Flicker, and (c) Frame drops.

temporal masking is less so, although it is known that motion has a silencing effect on flicker Ni et al. (2011).

Fig. 2.7 plots the MOS against spatial resolution for each distortions class. The purely temporal distortions: judder and frame drops yielded similar ranges of MOS for 1080p and 4K videos. However, aliasing resulted in very different MOS ranges,

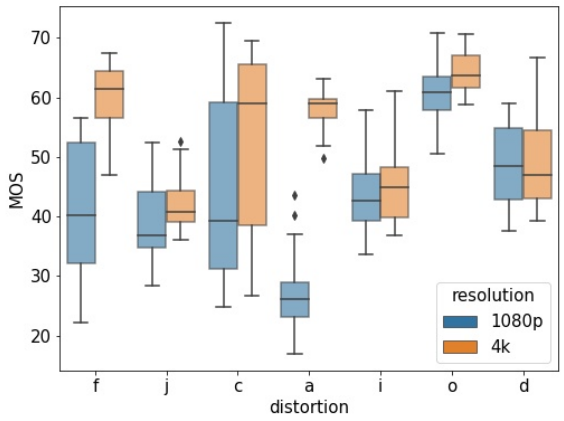


Figure 2.7: Box plot comparing MOS against distortion type for both considered video resolutions. The labels on the horizontal axis represent: f: flicker; j: judder; c: compression; a: aliasing; i: interlacing; d: frame drop and o: original (reference videos).

likely because of the additional upscaling of 1080p videos when displayed on the 4K TV.

Tables 2.2 and 2.3 show measurements of the consistency of human scores for each of the different distortion types. The Tables list the Spearman' s Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC) computed on the entire database and for each distortion type, again by randomly dividing the subjects into two groups. It may be observed that the SROCC was slightly lower than the PLCC, which might be explained by subjects having difficulty supplying correctly ordered ratings of videos of very similar quality. but still generally able to make predictions in a linear manner. Overall, the results of the results indicate a very high degree of internal consistency and agreement amount the human subjects on all of the distorted video types.

Although MOS is a good representation of the subjective quality of videos and is necessary for the development and evaluation of NR VQA algorithms, the Difference

Table 2.2: Min, Median, and Max SROCC of Human Scores Divided Into Two Groups.

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
MIN	0.9326	0.8285	0.8097	0.9092	0.8071	0.8543	0.8908
MEDIAN	0.9552	0.8967	0.8714	0.9425	0.8860	0.9151	0.9283
MAX	0.9685	0.9470	0.9250	0.9701	0.9373	0.9565	0.9651

Table 2.3: Min, Median, and Max PLCC of Human Scores Divided Into Two Groups.

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
MIN	0.9292	0.9547	0.9688	0.9334	0.9287	0.8891	0.9253
MEDIAN	0.9648	0.9728	0.9792	0.9633	0.9607	0.9383	0.9524
MAX	0.9741	0.9870	0.9875	0.9795	0.9763	0.9644	0.9725

Table 2.4: SROCC of the Compared FR VQA Models. The Scores of the Top Performing Algorithm Are Boldfaced

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
PSNR	0.3760	0.8750	0.4012	0.2117	0.5264	0.3024	**0.7507**
SSIM	**0.6976**	0.9171	0.7341	**0.3933**	0.8758	0.5291	0.6623
MS-SSIM	0.6757	0.9154	0.7335	0.3622	0.8652	0.5997	0.6179
SpEEDQA	0.6894	0.8979	0.8124	0.3165	0.8780	0.5993	0.6130
ST-RRED	0.6564	0.8943	0.8269	0.2968	0.8653	0.5635	0.7121
FAST	0.6192	**0.9283**	0.7269	0.2769	**0.9391**	0.7733	0.5960
VMAF	0.6434	0.9135	**0.9153**	0.3039	0.9243	**0.7843**	0.5346

MOS (DMOS) is more commonly used in the development and evaluation of FR VQA models, since it allows a way to reduce content dependencies of quality labels. Since we are supplying this resource for the study of both NR and FR models, we also calculated the DMOS of the videos with references. We calculated the DMOS according to:

$$DMOS_j = MOS_j^{ref} - MOS_j, \tag{2.7}$$

where MOS_j is the MOS of video j, and MOS_j^{ref} is the MOS of the reference video j, which is regarded as a "hidden reference," since it is not identified as such to the subjects.

Table 2.5: PLCC of the Compared FR VQA Models. The Scores of the Top Performing Algorithm Are Boldfaced

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
PSNR	0.4192	0.9586	0.4885	0.3452	0.5723	0.5886	**0.7840**
SSIM	0.7107	0.9659	0.7483	**0.5679**	0.8308	0.5770	0.7460
MS-SSIM	0.6907	**0.9690**	0.7696	0.5259	0.8589	0.6421	0.7105
SpEEDQA	**0.7235**	0.9526	0.9234	0.5037	0.8432	0.6183	0.7806
ST-RRED	0.6694	0.9483	**0.9425**	0.3952	0.8358	0.5915	0.7465
FAST	0.6520	0.9587	0.8329	0.4142	**0.9391**	0.8298	0.6978
VMAF	0.6355	0.9675	0.9296	0.3043	0.9242	**0.8654**	0.6242

Table 2.6: RMSE of the Compared FR VQA Models. The Scores of the Top Performing Algorithm Are Boldfaced

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
PSNR	10.3355	4.2304	13.6253	4.3601	9.5390	5.2311	**3.9024**
SSIM	8.0082	3.8493	10.3588	**3.8237**	6.4740	5.2852	4.1864
MS-SSIM	8.2324	**3.6708**	9.9705	3.9510	5.9578	4.9605	4.4235
SpEEDQA	**7.8589**	4.5223	5.9924	4.0129	6.2531	5.0856	3.9294
ST-RRED	8.4573	4.7155	**5.2190**	4.2673	6.3858	5.2174	4.1832
FAST	8.6315	4.2267	8.6452	4.2282	**3.7669**	3.6114	4.5028
VMAF	8.7894	3.7600	5.7557	4.4251	4.4430	**3.2435**	4.9154

Table 2.7: SROCC of the Compared NR VQA Models. The Scores of the Top Performing Algorithm Are Boldfaced

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
NIQE	0.3232	0.2775	0.2860	0.2863	0.2832	0.2842	0.2780
BRISQUE	0.6381	0.6409	0.7482	0.8039	0.6440	0.4180	0.8720
CORNIA	0.6778	0.7399	**0.8142**	0.7049	0.7193	0.0000	**0.8782**
HIGRADE	0.6916	0.7234	0.7337	0.5784	0.6429	0.5748	0.8060
V-BLIINDS	0.7330	0.7131	0.7482	**0.8679**	0.5769	0.7513	0.7936
TLVQM	0.7503	0.6574	0.7915	0.8246	0.6966	**0.8927**	0.8369
ChipQA	**0.7994**	**0.7482**	0.7998	0.8514	**0.7668**	0.7874	0.8111

2.5 Objective VQA Model Comparison

We evaluated several publicly available objective VQA algorithms on the LIVE Livestream Database to demonstrate the usefulness of the new resource. Given MOS and DMOS, we are able to test and compare both FR and NR VQA models. The performances of the objective VQA algorithms were evaluated using three standard

Table 2.8: PLCC of the Compared NR VQA Models. The Scores of the Top Performing Algorithm Are Boldfaced.

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
NIQE	0.4962	0.2805	0.2865	0.2860	0.2848	0.2849	0.2850
BRISQUE	0.6698	0.7616	0.9415	0.8362	0.7166	0.4265	0.9185
CORNIA	0.7257	0.8197	0.9595	0.7409	0.7841	0.0000	**0.9234**
HIGRADE	0.6990	0.8395	0.9426	0.6310	0.6938	0.5806	0.8528
V-BLIINDS	0.7477	0.8313	0.9277	**0.9239**	0.6238	0.7850	0.8826
TLVQM	0.7513	0.6991	0.9550	0.8850	0.8037	**0.9153**	0.8648
ChipQA	**0.8156**	**0.8408**	**0.9613**	0.9040	**0.8608**	0.8470	0.8587

Table 2.9: RMSE of the Compared NR VQA Models. The Scores of the Top Performing Algorithm Are Boldfaced

ALGORITHM	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
NIQE	50.4055	45.7805	45.8797	50.8770	49.9032	50.8316	50.9243
BRISQUE	9.6376	9.1434	5.5712	7.1173	8.4869	9.1254	4.3474
CORNIA	9.6960	8.0173	4.6140	8.5343	7.3121	9.7778	**4.3074**
HIGRADE	9.6469	**7.7381**	5.2704	9.9567	8.6036	8.0093	5.8163
V-BLIINDS	8.4058	7.7836	6.1912	**4.7971**	9.3751	5.8211	5.1704
TLVQM	8.7217	10.0801	4.8209	6.1302	7.1367	**4.0113**	5.5803
ChipQA	**7.2874**	7.7510	**4.3791**	5.3599	**5.6626**	5.0459	5.6679

metrics: the Spearman' s Rank Order Correlation Coefficient (SROCC), the Pearson Linear Correlation Coefficient (PLCC), and the Root Mean Square Error (RMSE).

2.5.1 Performances of FR VQA Models

Here we present the results for the following seven popular FR VQA models: PSNR, SSIM, MS-SSIM, SpEEDQA, ST-RRED, FAST, and VMAF. The distorted versions of the 45 reference contents (270 videos in total) were processed to produce predictions that were cast against the DMOS. Note that most FR VQA models require that there be an equal number of frames between each reference video and its corresponding compared distorted video. However, the videos subjected to interlacing distortions have one less frame than the originals they derive from. Hence, the final frame of the interlaced video is duplicated to match the reference. The predicted scores s were passed though a five-parameter nonlinear logistic regression function

before the PLCC and MSE were computed:

$$f(s) = \beta_1(\frac{1}{2} - \frac{1}{(1 + exp(\beta_2(s - \beta_3))}) + \beta_4 s + \beta_5, \tag{2.8}$$

where s are the predicted scores produced by the tested algorithm and $f(s)$ is the mapped score. By fitting parameters β_i $(i = 1, 2, 3, 4, 5)$, the MSE between the mapped and subjective scores is minimized. The SROCC, PLCC, and RMSE for each category of distortions are calculated by comparing the predictions made by the FR models and the ground truth for each of those distortions separately. Table 2.4, 2.5, and 2.6 show the performance metrics of the compared algorithms, which will be discussed shortly.

2.5.2 Performance of NR VQA Models

We compared the quality predictions made by a variety of NR models against the MOS. The NR VQA algorithms that were tested include NIQEMittal et al. (2012b), BRISQUEMittal et al. (2012a), HIGRADEKundu et al. (2017), CORNIAYe et al., TLVQMKorhonen (2019), V-BLIINDSSaad et al. (2012), and ChipQAEbenezer et al. (2020b, 2021). BRISQUE, HIGRADE, CORNIA, TLVQM, V-BLIINDS, and ChipQA are supervised learning algorithms that use a support vector regressor (SVR) to learn mappings from 'quality-aware' features to mean opinion scores. These algorithms were tested on 1000 random train-test splits. On each split, 80% of the data was used for training, and 20% for testing. Follow common practice, 5-fold cross-validation was applied within each training set to find the best parameters for the SVR. Care was taken to ensure that no content could appear in both the training and testing set, or the training and validation set.

NIQE, BRISQUE and HIGRADE are image quality assessment (IAQ) algorithms, so they were used to extract features frame by frame, followed by temporal average pooling.

For the unsupervised methods (NIQE), the scores s were passed through the

Table 2.10: Computation time on a single 3840x2160 video with 210 frames from the LIVE Livestream VQA database

Algorithm	Time (s)	GFLOPS	Complexity
NIQE	1008	3094	$\mathcal{O}(k^2NT)$
BRISQUE	301	352	$\mathcal{O}(k^2NT)$
TLVQM	1002	477	$\mathcal{O}(k_1^2NT+$ $(log(N)+k_2^2)NT_2))$
CORNIA	2056	4480	$\mathcal{O}(k^2MNT)$
VBLIINDS	3086	465	$\mathcal{O}((k^2N+$ $log(w)N+w^2d^3)T)$
HIGRADE	16240	9604	$\mathcal{O}(3(2k^2+k_2)NT)$
ChipQA	814	700	$\mathcal{O}((\frac{k^2}{D^2}+\frac{Q}{RD^2}+$ $\frac{Q\log Q}{R^3D^2})NT)$

k: window size; N pixel number per frame; T: number of frames;

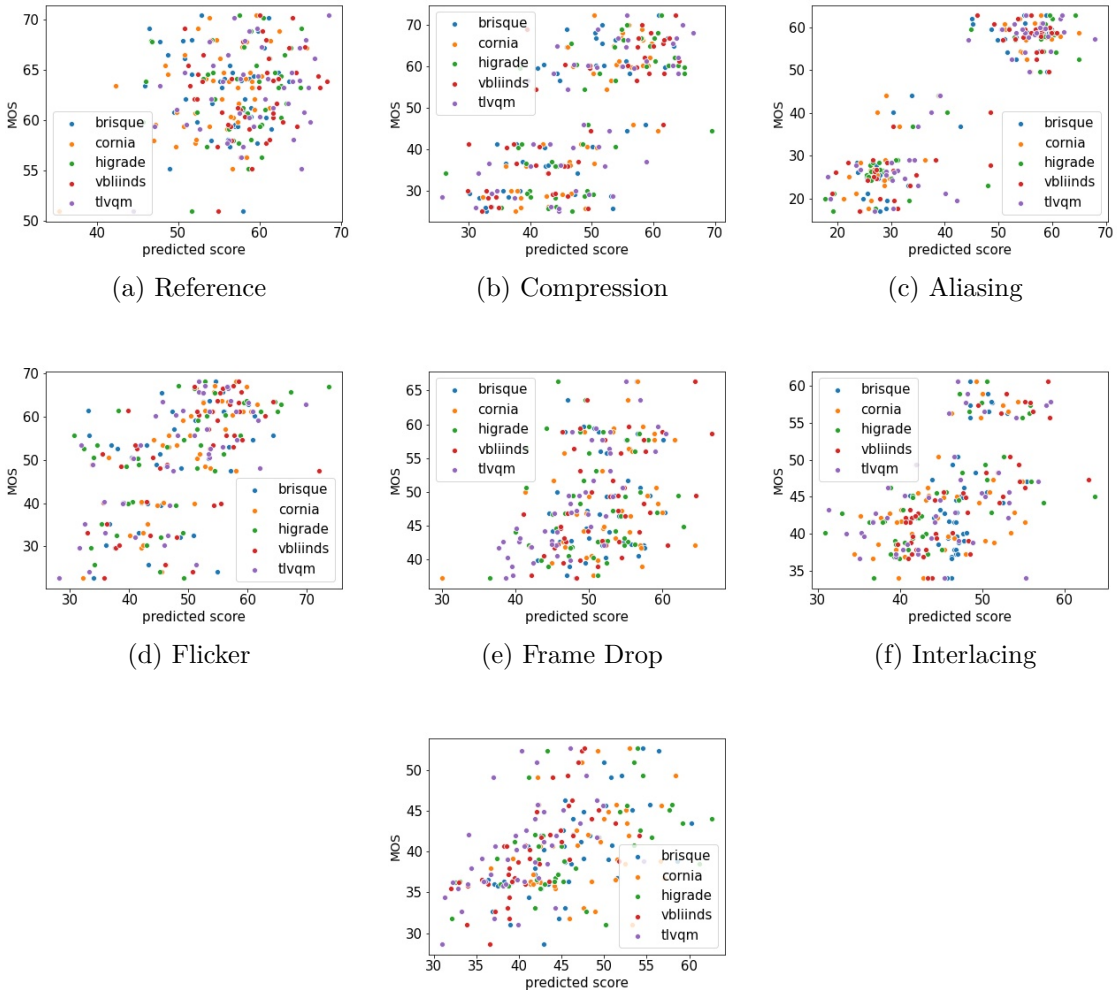
TLVQM: k_1,k_2: filter size, T_2: number of representative frames;

CORNIA, M: codebook size;

V-BLIINDS, w: window size, d: motion vector tensor size;

HIGRADE, k_2: gradient kernel size;

ChipQA, D: downsampling factor Q: chip search's quantization factor, R: size of each dimension of a chip



(a) Reference (b) Compression (c) Aliasing

(d) Flicker (e) Frame Drop (f) Interlacing

(g) Judder

Figure 2.8: Scatter plots of the predicted scores produced by several NR VQA models against MOS for each class of distorted videos.

same nonlinear logistic regression process before the PLCC and MSE were computed, as described earlier. The performances of the compared VQA models on the entire database, as well as for each synthetic distortion, are shown in Tables 2.7, 2.8, and 2.9, where the best performing model on each distortion category is boldfaced. The results for each specific distortion were acquired by training the SVR on the reference

Table 2.11: Results of One-Sided T-Test Performed Between SROCC Values of Various Algorithms on the Live Livestream Database. Each Cell Contains 7 Entries: the Entire Database, 6 Distortions in the Order: Compression, Aliasing, Judder , Flicker, Frame Drop, and Interlacing. A Value of ' 1' Indicates That the Row Is Statistically Superior (Better Visual Quality) Than the Column, While a Value of ' 0' Indicates that the Column Is Statistically Superior Than the Row. A Value of ' -' Indicates Statistical Equivalence Between Row and Column.

ALGORITHM	NIQE	BRISQUE	CORNIA	HIGRADE	V-BLIINDS	TLVQM	ChipQA
NIQE	-------	0000000	0000000	0000000	0000000	0000000	0000000
BRISQUE	1111111	-------	00-101-	0011-01	00-0101	01--101	0000001
CORNIA	1111111	11-010-	-------	011-001	01-0001	01-0101	0-10001
HIGRADE	1111111	1100-10	100-110	-------	0000101	010010-	0000000
V-BLIINDS	1111111	11-1010	10-1110	1111010	-------	11-1-0-	0001000
TLVQM	1111111	10--010	10-1010	101101-	00-0-1-	-------	0000011
ChipQA	1111111	111110	1-01110	1111111	1110111	1111100	-------

sequences and the specific distorted sequences. Scatter plots of some selected objective VQA models against MOS are shown in Fig. 2.8.

2.5.3 Statistical Evaluation

A one-sided t-test was performed on the 1000 SROCC scores of the NR VQA models computed on the LIVE Livestream Database, using the 95% confidence level to evaluate whether one VQA algorithm was statistically superior to another. The results are shown in Table 2.11. Results on the entire database and on individual distortions are both included. Each entry in the table consists of 7 symbols corresponding to the entire database, and the 6 distortions, in the order of compression, aliasing, judder, flicker, frame drop, and interlacing. A symbol '1' indicates using the performance of the algorithm on the row was statistically superior to that of the column, while a symbol '0' indicates that the column algorithms was statistically better than the row algorithm. A symbol of '-' indicates that the performances of the row and the column algorithms were statistically equivalent.

2.5.4 Computational Cost

Since we are interested in live streaming use scenarios, we studied the computational costs, the number of giga floating point operations (GFLOPS), and complexity of the compared models, as shown in Table 2.10. The $\mathcal{O}(\cdot)$ figures make clear that all of the compared algorithms could be implemented as real-time hardware realizations. To measure computation time, we used a single 4K video having 210 frames. Of the compared algorithms, V-BLIINDS, and ChipQA were implemented in Python. All other algorithms were implemented in MATLAB®. All the algorithms were run on an Intel Xeon E5-2620 CPU with a maximum frequency of 3 GHz.

While none of the tested algorithms runs in real time in their current implementations, they may be optimized to do so. In most of the algorithms, the most expensive step is filtering. For example, in BRISQUE the largest computation is computing the mean subtracted contrast normalized (MSCN) coefficients. However, filtering scales up linearly and is highly parallelizable. Frame based algorithms can be applied at a lower frame rate with little loss of prediction efficacy Tu et al. (2021b). While V-BLIINDS expends considerable computation on motion computation, motion vectors can be re-used from those produced by the involved codec. The complexity of CORNIA, which computes dot-products between local descriptors and visual codewords, is affected by the codebook size, which can be quite large.

2.5.5 Discussion of Results

The results presented in Tables 2.4, 2.5 and 2.6 suggest that, other than PSNR, the compared FR VQA models generally delivered similar overall performances on the entire database, but some algorithms yielded better performances on certain distortions. For example, SSIM, which performed well overall, obtained the highest correlation against DMOS on the compressed videos, but low correlation on the judder videos. The main reason that SSIM delivers low performance on judder videos is that it is a frame-based model. Judder is a temporal distortion that arises when

high motion is present in a video. The greater the magnitude of the motion, the more apparent the distortion is likely to be. While SSIM effectively captures spatial distortions (like compression), it is unable to capture the temporal effects of judder. ST-RRED does include limited temporal information expressed as NSS features from adjacent frame differences, which is inadequate to model complex or longer-duration temporal distortions, hence it does not outperform the other compared FR models. VMAF yielded the highest correlation on the aliased and flicker videos, but low correlations on the interlaced videos. The FR VQA models tended to deliver decent performances on common distortions found in other VQA databases, such as compression, and also flicker, which is compression based. These distortions are better studied and easier to catch with the presence of the reference. However, when tested on the purely temporal distortions, all of the compared FR VQA models delivered low correlations against DMOS. This suggests ample room for research on developing better models of temporal and motion-related distortions.

From Tables 2.7, 2.8, and 2.9, it may be observed that ChipQA performed the best among the compared NR VQA algorithms, while TLVQM and V-BLIINDS also achieved relatively higher correlations against the human judgments. TLVQM achieved the top performance on flicker and frame drops, likely because of the large number of temporal features it uses. ChipQA builds a statistical representation of local spatiotemporal data that is attuned to local orientations of motion over large spatial fields, motivated by processes in areas V1 and MT of the brain. The explicit modeling of deviations from statistical regularity in the spatiotemporal domain allows it to perform well on both spatial and temporal distortions. NIQE and BRISQUE are similar methods, but BRISQUE is trained while NIQE is completely blind, hence BRISQUE usually can deliver predictions having higher correlations against human quality judgments. Similar statistical features are used in V-BLIINDS and HIGRADE. The frame-based models NIQE, BRISQUE, HIGRADE, and CORNIA do not access any motion information, which greatly limits their performance. CORNIA yielded top performances on compression, aliasing, and interlacing, all of

which present strong spatial aspects of distortion. However, the overall performance of CORNIA was lower than that of V-BLIINDS, TLVQM, and ChipQA, due to the lack of temporal information.

Chapter 3: A Study of Subjective and Objective Quality Assessment of HDR Videos

3.1 Related work

3.1.1 Subjective HDR Video Quality Databases

Over the past few years, a number of efforts have been made to create video quality datasets for HDR, but all of these have limited usefulness, either because they have been rendered obsolete by the rapid pace of HDR standard development, or by the inability of authors to publicly release their data owing to copyright issues. Azimi *et al.* Azimi et al. (2018) conducted a study using 18 human subjects who viewed 5 different 12-bit YUV contents captured by a RED Scarlet-X Camera and afflicted by compression and four other types of distortion, yielding 30 videos. The videos were displayed on a non-standard HDR device the authors designed themselves, supporting the older, more limited BT. 709 gamut, rather than the HDR10 compliant BT. 2020 gamut, and the PQ OETF was not applied prior to compression. Moreover, the videos were of maximum resolution 1920×1080 (1080p), while most current HDR content is 4K. Pan *et al.* Pan et al. (2018) conducted a study of the effects of compression on HDR quality using 6 source videos encoded using PQ and HLG and the BT. 2020 color space, but the codec used for compression was AVS2, which has seen little industry adoption. The study included 144 videos that were rated by 22 subjects, but unfortunately none of the video or subjective data has been made publicly available. Baroncini *et al.* Baroncini et al. (2016) conducted a study of 12 compressed HDR videos evaluated by 40 human subjects. The source contents did not follow ITU Rec. BT 2020, the PQ OETF was not applied on the video data, and again, none of the data was made publicly available. Moreover, the resolution of all the videos was 1080p. Rerabek *et al.* Rerabek et al. (2015) conducted a study of 5 HDR videos, each distorted by 4 compression levels, with the aim of comparing objective HDR VQA

algorithms, but the data was not made publicly available. The videos were all only of resolution 944×1080, and the data was tone-mapped to 8-bit format before being displayed to the subjects. Athat *et al.* Athar et al. (2019) conducted a subjective study of HDR10 content, but none of the data was publicly released because of copyright issues. The authors compressed 14 HDR10 source contents using H.264 and HEVC to generate 140 distorted videos, which were viewed and rated by 51 subjects.

The study that we report here advances the field in several ways: first, all of the source videos are compliant with the most widely used modern HDR standard (HDR10) and include wide color gamut (WCG) and high frame rate (HFR) videos. Second, the new dataset contains almost twice as many videos as any prior HDR VQA dataset, and more than double the number of collected subjective opinion scores. Third, we conducted the largest and most contemporaneous HDR VQA study on it to date. Fourth, we compared the performances of leading HDR VQA models on it to validate the usefulness of the collected data. Lastly, unlike nearly all of the prior datasets, we are making the LIVE HDR dataset publicly available at `http://live.ece.utexas.edu/research/LIVEHDR/LIVEHDR_index.html`.

3.1.2 Objective Video Quality Assessment Algorithms

Objective VQA algorithms aim to automatically predict the perceptual quality of videos. There are three categories of objective VQA models: full-reference (FR), reduced reference (RR), and no-reference (NR). FR VQA models operate by comparing pristine reference videos against distorted versions of them using perceptually motivated features and/or training data Wang et al. (2004); Sheikh and Bovik (2005). Reduced reference VQA models use only partial reference information to achieve efficiencies Wang and Simoncelli (2005); Wang et al. (2006); Soundararajan and Bovik (2011); Wu et al. (2015). NR VQA models require no information regarding any reference videos, and instead predict perceptual video quality based only on information extracted form distorted videos Mittal et al. (2012a, 2015); Moorthy and Bovik (2011);

Saad et al. (2012). We use the new psychometric HDR VQA database to compare leading HDR VQA models that fall into the FR VQA category. The MSE (or equivalently, the PSNR) has long been used as a basic index of video quality. More recent popular VQA models include Structural Similarity (SSIM) Wang et al. (2004), Multi-scale SSIM (MS-SSIM) Wang et al. (2003), Gradient Magnitude Similarity Deviation (GMSD) Xue et al. (2013), most apparent distortion (MAD) Larson and Chandler (2010), visual information fidelity (VIF) Sheikh and Bovik (2005), and FSIM Zhang et al. (2011), among others Vu et al. (2011); Vu and Chandler (2014); Bampis et al. (2017a); Seshadrinathan and Bovik (2009). More recently, machine learning-based FR-VQA frameworks have become quite popular. For example, VMAF Li et al. (2017) combines features from two VQA models, using a Support Vector Regressor (SVR) to map their feature sets to video quality predictions. FR VQA models that rely on deep learning have recently achieved competitive performance, such as Deep-VQA Kim et al. (2018), and some even use unsupervised deep learning (UDL) Vega et al. (2017).

HDR quality prediction research is still a nascent field, and there is only a small literature on the subject. Wang et al. (2020) discusses HDR visual quality impairments and efforts at developing dedicated objective HDR video quality metrics. An early algorithm was HDR-VDP Mantiuk et al. (2005), which considers the nonlinear response to light of high contrast content and the full range of luminances. An improved version called HDR-VDP-2 Mantiuk et al. (2011) uses a model of all luminance conditions derived from contrast sensitivity measurements. Further improvements of HDR-VDP-2 include HDR-VDP2.2 Narwaria et al. (2015a, 2014)) and HDR-VDP3 Mantiuk et al. (2023). The author of Aydın et al. (2008) proposed PU, a nonlinear transform to extend normal SDR quality metrics to HDR. Recent developments such as the PU21 encoding function have further refined the field, providing an enhanced methodology for designing quality metrics specific to HDR images Mantiuk and Azimi (2021). Other authors have focused on the chromatic aspects of HDR video quality by focusing on color fidelity Abebe et al. (2015), using HDR Uniform Color Spaces

Rousselot et al. (2019), and using color difference models Choudhury et al. (2021). Another method called HDR-VQM utilizes spatio-temporal analysis that simulates human perception Narwaria et al. (2015b).

Each of these prior methods has shortcomings. Most of them rely on simple transforms that map video features to quality predictions, such as, the root mean square error (RMSE) used in color difference models, spatial pooling in HDR-VDP-2, or the PU-SSIM and PU-PSNR models proposed in Aydın et al. (2008). While these methods are effective on their intended applications, they were primarily designed for legacy HDR videos or HDR images. The modern HDR10 standard, however, introduces several significant changes, including the use of the Perceptual Quantizer (PQ) curve for encoding luminance information, the adoption of the BT.2020 color space, and the inclusion of metadata for accurate display of HDR content. Furthermore, our focus on a video database inherently includes temporal distortions, a factor not present in image databases. Given these changes, it is likely that the reliability of legacy-based quality metrics is reduced when applied to HDR10 content. Therefore, it is necessary to evaluate these existing methods within the context of HDR10 content to ensure their continued relevance and accuracy. Additionally, our study also emphasizes the use of the HEVC codec, which aligns with modern practice. This new codec may introduce different types of distortions, and the visibility of these distortions may also be different, further underscoring the need for evaluation.

3.2 Subjective Experiment Design

3.2.1 HDR Video Contents

We gathered a collection of high-quality, distortion-free HDR10 sequences from CDV; Song et al. (2016) and nearly distortion-free content from 4km (2020). These videos were captured by professionals using high-end cinematic HDR video cameras. These sequences were all progressively captured at resolution 3840×2160 with the audio signal removed. The sequences from CDV; Song et al. (2016) were captured

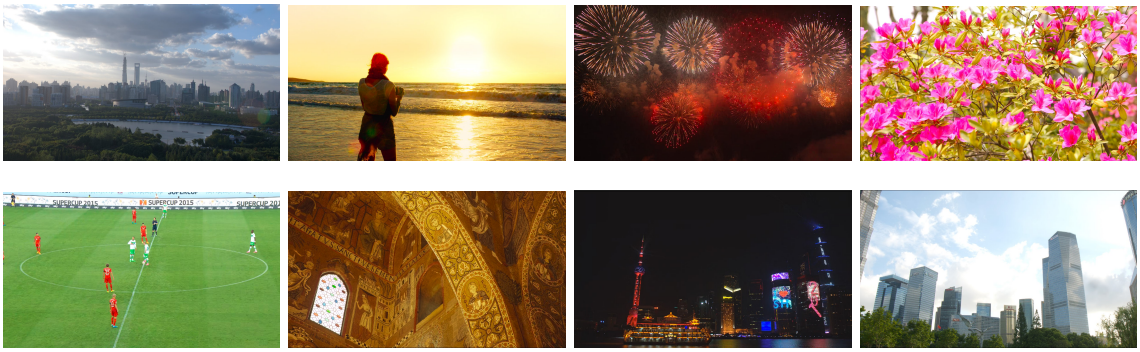


Figure 3.1: Exemplar screenshots of frames from source sequences.

using Sony F55 or Sony F65 cameras with the dynamic range fixed to the S-log3 profile and are then transformed to PQ EOTF in the post-production process. The videos from 4km (2020) were provided in HDR10 format. The videos from CDV; Song et al. (2016) have frame rates of 60 frames per second (fps) and those from 4km (2020) include both 50 fps and 60 fps. All of the source sequences are HDR-WCG-HFR videos. Following recent studies Mercer Moss et al. (2016); Mercer Moss et al. (2016); Zhang et al. (2018); Paudyal et al. (2019), we segmented all of the video sequences into one or more clips of 7-10 seconds duration. This range was chosen to balance data collection efficiency and maintaining the integrity of the depicted scenes. The 31 source clips were generated from 19 different sources. When clipping the videos, care was taken to avoid awkward interruptions of content and to prevent similar clips from being taken from the same segments, ensuring a more coherent, diverse, and representative set of visual experiences for studying quality assessment.

Fig. 4.1 shows several sample frames from the source sequences we acquired. The videos span a wide range of contents. We directly applied the spatial information (SI), or integrated Sobel magnitude, and the temporal information (TI), or absolute average frame difference, both defined in ITU (2008), to the 10-bit HDR data. Similarly, the colorfulness measure denoted as CF was computed as in Hasler and Suesstrunk (2003). Fig 4.2 plots the SI, TI, and CF of all of the source sequences in the LIVE-HDR database, indicating wide coverage of low-level content and activity in space

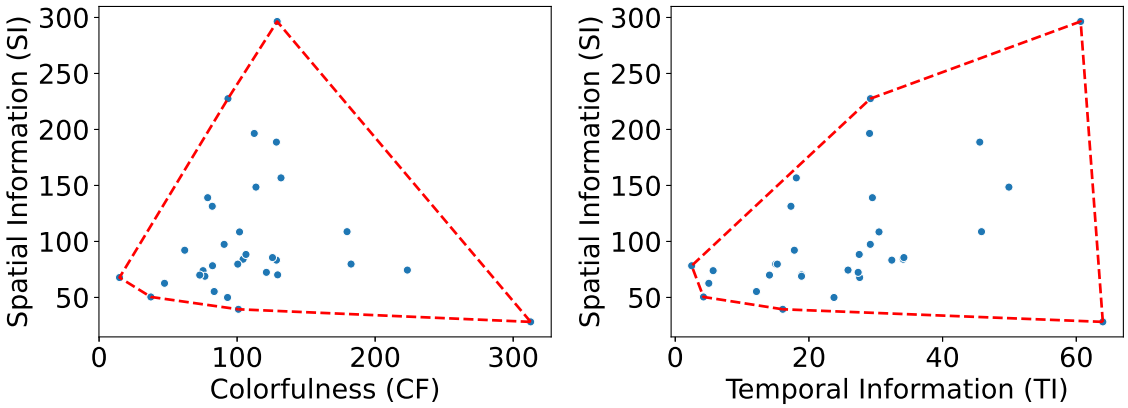


Figure 3.2: Spatial Information (SI) versus (a) colorfulness (CF) and (b) Temporal Information (TI), measured on all of the source sequences in the new LIVE-HDR Database. The corresponding convex hulls are plotted by red lines.

and time.

Moreover, we included additional characteristics of the HDR content: min, max, mean, and median luminance, and the portion of pixels outside of the sRGB color gamut. These new metrics, visualized in Figs. 3.3 and 3.4, provide further insights into the diversity and coverage of the color and luminance in the HDR videos of our database.

3.2.2 Test Sequences

We collected 9 distorted video sequences from each source sequence using the High Efficiency Video Coding (HEVC) Codec. The selection process was subjective but systematic, aiming to ensure that the videos are perceptually distinguishable while spanning a broad range of perceptual qualities. We initially generated a substantial set of videos using a range of bitrates and spatial resolutions, including but extending beyond common settings in the streaming industry. We manually reviewed all the videos and progressively reduced their number to make the total playback duration suitable for our human subjective study. The final bitrate and resolution settings that

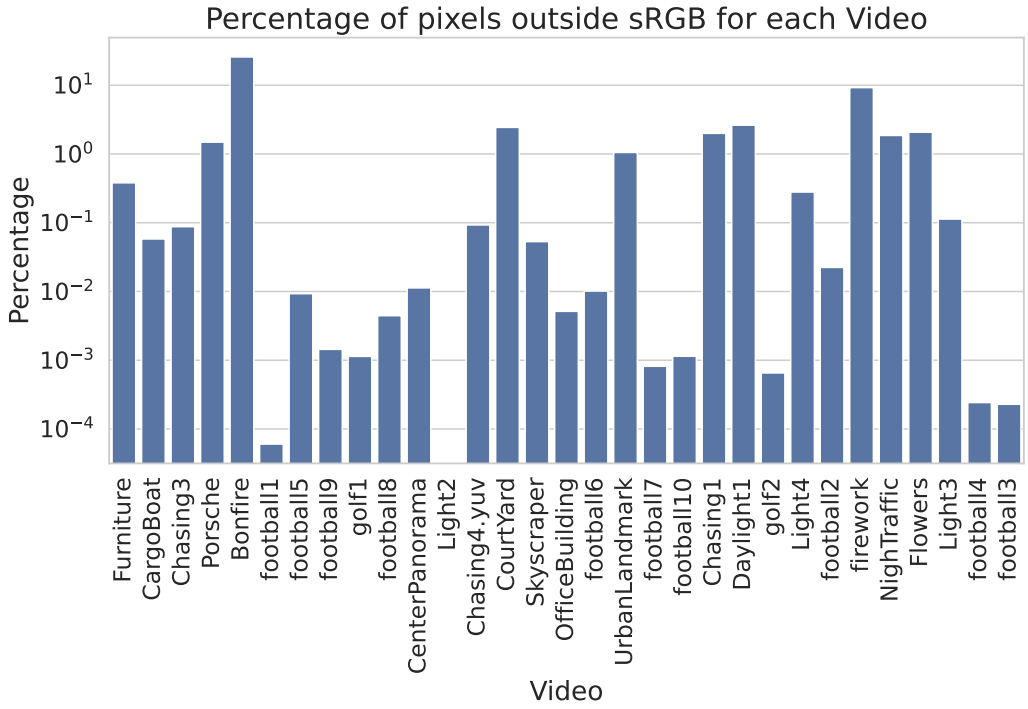


Figure 3.3: Proportion of pixels outside of the sRGB color gamut, measured on all of the source sequences in the new LIVE-HDR Database.

we used are listed in Table 4.1.

As for the encoding parameters, we used the libx265 encoder in constant bitrate mode with single-pass encoding, which is most commonly used in industrial streaming applications, owing to its simplicity and efficiency. While certain bitrates and resolutions may be less prevalent in practical applications, their inclusion remains advantageous. For instance, a 540p video with a 2.2 Mbps bitrate may exceed those encountered in real-world situations, yet it exemplifies a scenario with pronounced scaling artifacts and reduced compression artifacts. Conversely, the 2160p video at 3 Mbps exhibits significant compression artifacts, devoid of any scaling issues. Lastly, the 720p video at 2.6 Mbps represents a confluence of both compression and scaling artifacts. In numerous past studies Shang et al. (2021c,b) we have found this approach to be an effective way to cover the distortion space, helping to ensure subsequent

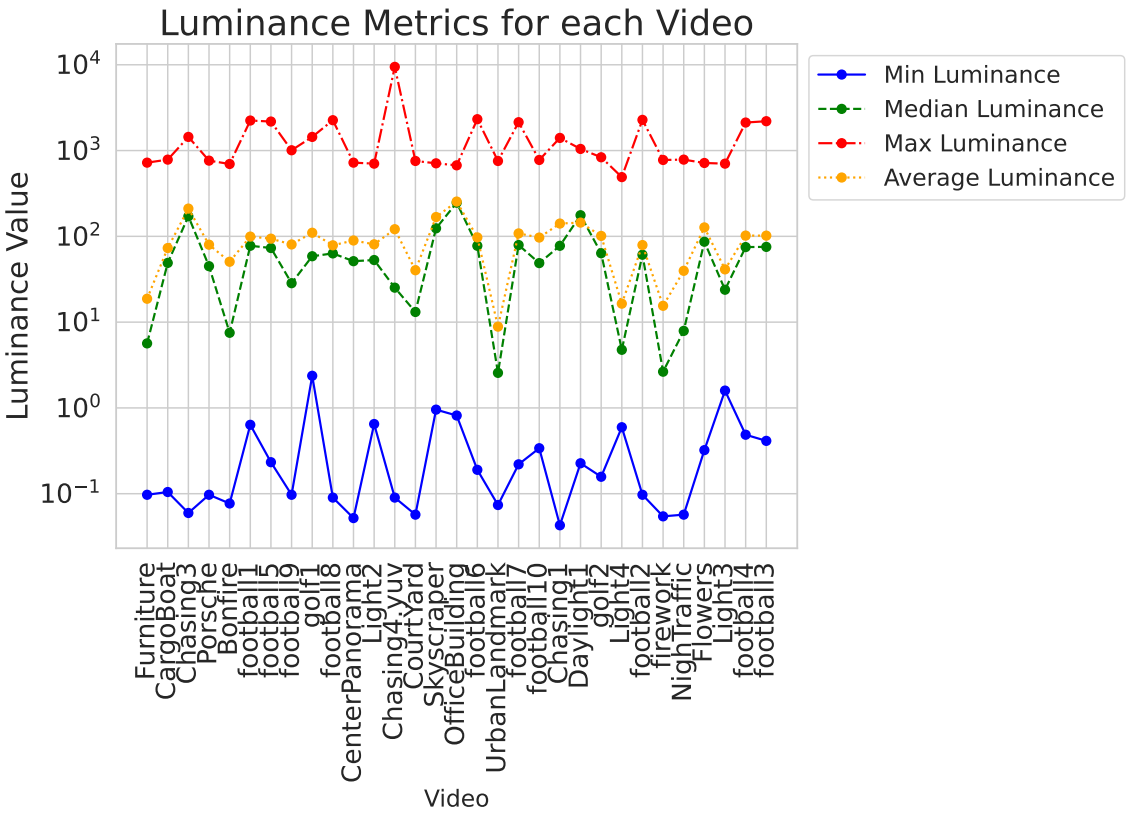


Figure 3.4: Min, max, mean, and median luminance metrics measured on all of the source sequences in the new LIVE-HDR Database.

model learning. The source videos were included in the database and subsequent psychometric study, to serve as labeled reference videos against which difference mean opinion scores (DMOS) can be calculated. The videos include four practical spatial resolutions. The higher-resolution 4K and 1080p videos were compressed using four and three bitrate targets, respectively, mimicking the bitrate ladders used in HDR video streaming. The videos compressed at the highest bitrate may be observed to present only slightly visible compression artifacts, while the videos compressed to the lowest bitrates exhibit obvious blocking, banding, temporal and scaling artifacts. The 1080p, 720p and 540p videos were all upscaled to 4K resolution when displayed to the human subjects, using bicubic interpolation. This method was selected for its

Table 3.1: Bitrate and Resolution Settings Used to Create the Distorted videos.

Number	resolution	bitrate (Mbps)
1	3840×2160	15
2	3840×2160	6
3	3840×2160	3
4	1920×1080	9
5	1920×1080	6
6	1920×1080	1
7	1280×720	4.6
8	1280×720	2.6
9	960×540	2.2

balance between computational efficiency and performance, which minimizes distortion and delay during video playback, thereby maintaining the integrity of the HDR content. The overall video database contains 279 distorted videos and 31 reference videos, yielding a total of 310 videos that were presented to the human subjects.

3.2.3 Subjective Testing Design

The human study was conducted in the Laboratory for Image and Video Engineering (LIVE) subjective study room at The University of Texas at Austin. A 65 inch Samsung Class Q90T QLED 4K UHD HDR Smart TV TV was used to display the HDR content to the participating subjects. The TV was calibrated for HDR by an Imaging Science Foundation (ISF) certified professional using a Calman Calibration kit.

After calibration, the TV had a peak luminance of approximately 1033 cd/m^2, and a minimum luminance below the measurement threshold of 0.7 cd/m^2. Color gamut coverages were 99.88% for BT.709, 88.86% for P3, and 66.33% for BT.2020. All

the measurement was made with a SpectraScan® Spectroradiometer PR-655. It was crucial to ascertain that the TV detected and displayed HDR input correctly, thereby avoiding any unintended tone mapping processes that might introduce distortions. To accomplish this, we made specific configurations and settings adjustments.

First, we enabled the "input signal plus function" in the TV settings, allowing the Samsung TV to receive an extended input signal range and enable HDR input. Subsequently, in the Windows 10 operating system, we activated HDR functionality in the Display settings. Additionally, in the Nvidia Control Panel, we modified the output format to yuv420p and 10-bit depth, while setting the refresh rate at 60Hz. These settings were meticulously reviewed and ensured to remain consistent throughout the entire study. The TV was connected to a workstation having a 12 GB Titan X Graphics Processing Unit (GPU), via an HDMI 2.0b cable allowing for smooth playback of the videos.The Potplayer Video Player with the MadVR renderer was used for playback. In the MadVR settings, we took additional measures to guarantee an authentic HDR viewing experience for the subjects. Specifically, we configured MadVR to pass through HDR content directly to the display. Moreover, we ensured that the "Send HDR metadata to the display" option was enabled. We also used the test pattern in TECHNICAL (2020) to verify the display. All advanced temporal processing options on the TV were disabled to avoid the introduction of any processing artifacts.

For all the subjects the viewing distance was about $1.5H$, where H is the height of the display. During a session, the subject would watch each video, then see a screen where they were asked to record a quality judgment on the video that they had just seen, using a visible slider on the screen they controlled with their mouse. While the rating scale was continuous, the user was guided by five Likert-like markers placed at uniform intervals labeled as "Bad," "Poor," "Fair," "Good," and "Excellent." The scores given by the subjects were sampled as integers on the interval [0, 100], although numerical values were not made visible to the subjects. In order to prevent bias due

to initial positioning of the rating indicator, it would not appear on the sliding scale until the subject placed the cursor on the slider and clicked on it.

The first session shown to each subject was preceded by a briefer training session that presented six exemplar videos of two contents (different from those that followed) that generally spanned the range of distortions that would be seen. For each of the two contents, one reference video and two compressed versions were displayed. All of the training videos were played in a randomized order, each followed by the interactive rating screen, to allow the subjects to become familiar with the overall rating protocol. We utilized the Absolute Category Rating with Hidden Reference (ACR-HR) protocol ITU (2008) when displaying the training and test videos, hence the videos shown in each session were displayed in randomized order. Each subject viewed the videos in a different random order.

3.2.4 Ambient Conditions

Two different lighting conditions were used to test the effects of ambient illumination on the perceived quality of HDR content. The first was a dark viewing condition, where the incident illumination on the television was measured to be 5 lux, following the recommendation in ITU (2018) for critical viewing of HDR content, and the recommendation in ITU describing general viewing conditions for a subjective study conducted in a laboratory environment. An incandescent table lamp and floor lamp were used to create the light necessary for this environment.

The second ambient condition was illuminated by a pair of yellow-filtered Neewer LED lights to produce an incident illumination on the TV of 200 lux, following the recommendation in ITU for general viewing conditions in a home environment. In this environment, a set of studio LED lights and a 95 W studio compact fluorescent light were placed behind and below the television in order to create a uniform, diffuse ambient illumination. In both environments, the lights were positioned so that their reflections off the television would not be visible to the viewers. The incident

luminance on the TV was measured by a Dr. Meter LX1330B luxmeter.

3.2.5 Subjects

A total of 66 human subjects were recruited from the student population at The University of Texas at Austin. Each subject participated in two sessions separated by at least 24 hours. The subjects were divided into two groups, one for each ambient condition. Hence 33 subjects watched the videos in the darker environment and 33 watched the videos in the brighter environment. No subject was given any information about the ambient conditions. We applied the Snellen and Ishihara tests of test each subject's visual acuity and color perception, respectively. One subject was found to have a color deficiency, but no subjects had less than 20/30 visual acuity on the Snellen test, when wearing their corrective lenses (if needed). The color deficient subject was not rejected from the study following our common practice of promoting a more realistic subject pool, as explained on our website liv.

3.3 Processing of Subjective Scores

There are a number of ways in which subjective scores can be converted into Mean Opinion Scores (MOS). We computed MOS as the average of subjective scores given by subjects (MOS), the average of z scores (ZMOS), and we also computed MOS using the statistical method proposed in Li et al. (2020).

3.3.1 MOS

Let i_d index those subjects that viewed videos in the dark environment, and i_b index the subjects who viewed the videos in the bright environment. MOS is calculated as the average of the scores given by a set of subjects, in ITU (2012). We will also define separate MOS values for the dark and light environments. Let the scores given by a subject i_k on video j be $s_{i_k j}$. We will refer to the MOS of a video

j whose scores were collected under the darker (brighter) ambient conditions as the respective average scores given under each condition: MOS_{dj} and MOS_{bj}, where

$$MOS_{kj} = \sum_{i_k=1}^{S_k} s_{i_k j}, \tag{3.1}$$

for $k = d, b$ (dark, bright), and $j = 1, 2 \dots N$.

3.3.2 ZMOS

We also define MOS calculated as the average of the z scores Shang et al. (2021a); Madhusudana et al. (2021a), given by

$$z_{i_k j} = \frac{s_{i_k j} - \mu_{i_k}}{\sigma_{i_k}} \tag{3.2}$$

for $k = b, d$, where the subjects under dark (bright) conditions are indexed $i_d = 1, 2 \dots S_d$ ($i_b = 1, 2 \dots S_b$) when rating videos indexed $j = 1, 2 \dots N$. In our database, $S_d = 33, S_b = 33$ and $N = 310$. In (3.2), μ_{i_k} and σ_{i_k} are the mean and standard deviation of the scores given by subject i_k across all videos:

$$\mu_{i_k} = \frac{\sum_{j=1}^{N} s_{i_k j}}{N} \tag{3.3}$$

and

$$\sigma_{i_k} = \sqrt{\frac{\sum_{j=1}^{N} (s_{i_k j} - \mu_{i_k})^2}{N}}. \tag{3.4}$$

Since there are two ambient conditions, for each video $j = 1, \dots, N$ we will refer to the MOS calculated from scores that were collected under darker (brighter) ambient conditions as $ZMOS_{dj}$ and $ZMOS_{bj}$, respectively, where

$$ZMOS_{kj} = \sum_{i_k=1}^{S_k} z_{i_k j} \tag{3.5}$$

for $k = d, b$ (dark, bright).

Table 3.2: Consistency Analysis of the Subjective Data.

	Correlations before ITU BT 500.11 outlier removal.	Number of outliers according to ITU BT 500.11.	Correlations after ITU BT 500.11 outlier removal.
MOS_d	0.9481	0	0.9481
MOS_b	0.9528	2	0.9492
$ZMOS_d$	0.9636	7	0.9581
$ZMOS_b$	0.9669	6	0.9665

3.3.3 Consistency Analysis

We studied the internal consistency of the scores as follows. We randomly partitioned the subjects who participated under each ambient condition into two approximately equal sized groups and computed the correlations between the mean MOS computed separately from the two groups over 100 random divisions. We then computed the correlation across the 100 splits. As expected, the internal consistency of the $ZMOS$ was better than that of MOS. We applied the outlier rejection method suggested by ITU Rec. BT 500.11 on both the MOS and $ZMOS$, separately for each ambient condition. However, we found that the internal correlations did not improve when the outliers were removed, as shown in Table 3.2. We also examined the scores of the color-deficient subject, and found that his scores correlated more highly against the other subjects who participated under the same ambient condition (0.88) than the average correlation between individual scores and group scores (0.82). In our analysis, we therefore chose not to remove the outliers when conducting the subsequent statistical analysis.

3.3.4 SUREAL Scores

A number of deficiencies in the ITU BT 500.11 outlier removal method have been observed in Li et al. (2020), along with an improved method called SUREAL that finds a Maximum Likelihood (ML) estimate of the scores. Using this method,

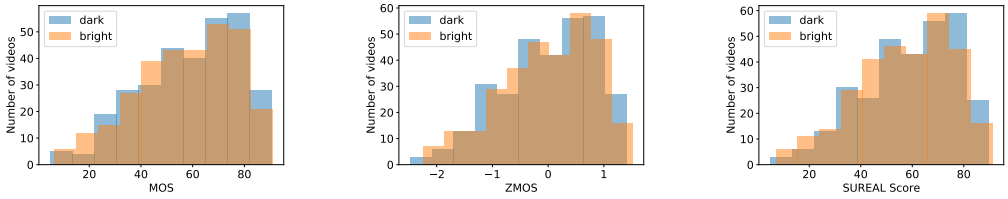


Figure 3.5: Histograms showing distributions of MOS, $ZMOS$, and SUREAL scores.

represent the opinion scores $s_{i_k j}$ as random variables $S_{i_k r}$

$$S_{i_k j} = \psi_{kj} + \Delta_{i_k} + \nu_{i_k} X, \tag{3.6}$$

where ψ_{kj} is the true quality of video j under ambient condition k, Δ_{i_k} represents the bias of subject i_k, the non-negative term ν_{i_k} represents the inconsistency of subject i_k, and $X \sim N(0,1)$ are i.i.d. Gaussian random variables. The quantities $\psi_{kj}, \Delta_{i_k}, \nu_{i_k}$ are estimated by computing the log-likelihood of the observed scores, using the Newton-Raphson method to solve for the values of $\psi_{kj}, \Delta_{i_k}, \nu_{i_k}$ that maximize the log-likelihood. We plotted the estimated subject biases in Fig. 14 and their inconsistencies in Fig. 12 in the supplementary material. It may be observed that both the subject biases and inconsistencies are quite dispersed. In this way, subject biases are accounted for when estimating the true qualities ψ_{kj}, and the method is robust against subject inconsistencies.

3.4 Effect of ambient illumination

We used all three types of summary subjective opinion scores to analyze the effects of ambient illumination on impressions of quality. It is worth noting that the MOS and SUREAL scores preserve the differences between the absolute values of the scores under the two ambient conditions, while the $ZMOS$ scores do not, since they are normalized. The distributions of MOS, $ZMOS$, and SUREAL are shown in Fig. 3.5. The MOS and SUREAL values under each of the ambient conditions cover a wide range, and it may be observed that the overall distributions of scores

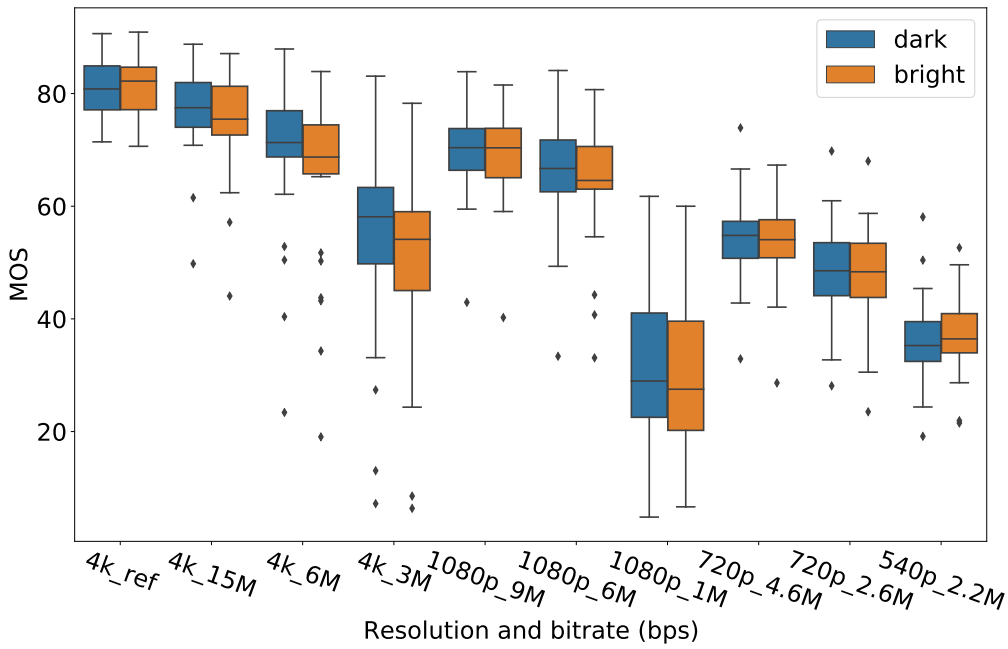


Figure 3.6: A box plot showing the distribution of MOS under two ambient illumination settings for each distortion combination.

under the two ambient conditions are similar. Since SUREAL and MOS are absolute scores, one may deduce from Fig. 3.5 that the videos watched under darker ambient conditions were rated as being of slightly higher qualities than those watched under bright ambient conditions. The same conclusions cannot be drawn regarding $ZMOS$, which is a normalized score, suggesting that these results reflect a slight preference for viewing under the darker conditions, but the relative ratings remain largely unaffected. Fig. 3.6 plots MOS against spatial resolution and bitrate. It may be observed that the MOS recorded under both ambient conditions fell in similar ranges for each spatial resolution and bitrate combination, but the MOS recorded under brighter conditions were slightly lower than under darker conditions at most resolution and bitrate settings. These differences, however, were more pronounced at lower bitrates and resolutions.

To assess the possible significance of the differences that we observed in Fig. 3.6,

we conducted Welch's two-sided t-test on the MOS under both ambient illumination settings. We compared the MOS at each resolution and bitrate setting, obtaining the p-values shown in Table 3.3. As may be seen, none of the resolution and bitrate combinations yielded a p-value less than 0.05, indicating that, while differences may be discerned between the MOS obtained under the two different illumination settings, these differences were not statistically significant. Separately, we also tested the raw (non-averaged) scores that were recorded by the individual subjects under the two ambient conditions. From among 310 labeled videos, only 17 were associated with differences in quality judgments that were statistically significant.

We further investigated the influence of ambient illumination on perceived video quality through a permutation test as outlined in Li et al. (2019). Despite 17 videos showing statistically significant differences in mean scores under different viewing conditions in our initial t-test analysis ($D = 17$), we sought to examine whether this could occur by chance. In the permutation test, subjects were randomly divided into two groups and mean scores for each video were recalculated. A paired t-test was then executed for each video. This process was replicated 10,000 times to construct a distribution of counts of significant differences, D', under random group assignment.

For ambient illumination to be considered significant, it must satisfy $\Pr(D' < D) \geq 0.95$. Our analysis revealed that the 95th percentile of the D' distribution was 41, greater than observed $D = 17$, leading to the conclusion that differences between bright and dark conditions were not statistically significant. The D' distribution, observed D, and the 95th percentile are shown in Fig. 3.7, illustrating the lack of significant impact of the ambient illumination on video quality ratings.

Further, we calculated the average luminances of each video which does not depend on the illumination. Fig. 3.8 shows a scatter plot of the p-values of videos in the raw score comparisons against the computed average luminances. There was no clear tendency of p-values against the average luminance. Indeed, the Pearson's correlation coefficient between the p-values and the average luminances were essentially

Table 3.3: The P-value of Each Bitrate and Resolution Settings for the Distorted Videos.

Number	resolution	bitrate (Mbps)	p-value
1	3840×2160	ref	0.5987
2	3840×2160	15	0.1539
3	3840×2160	6	0.1750
4	3840×2160	3	0.1538
5	1920×1080	9	0.3422
6	1920×1080	6	0.2856
7	1920×1080	1	0.3105
8	1280×720	4.6	0.4361
9	1280×720	2.6	0.3645
10	960×540	2.2	0.7095

nil (0.03).

We also used the confidence intervals of the SUREAL scores to study the effects of ambient illumination. The SUREAL method provides 95% confidence intervals on the subjective scores using the Cramer-Rao bound. The values of ψ_{dj} and ψ_{bj} are plotted in Fig. 15 in the supplementary material. We found that for 10 of the 310 videos, the confidence intervals did not overlap, indicating statistically significant differences. We also computed the 95% confidence intervals of the MOS (assuming normality) and plotted the scores and their confidence intervals in Fig. 16 in the supplementary material.

3.5 Objective Video Quality Model Design

The goal of our model design is to find features that are expressive of distortions that are more noticeable in HDR videos. As compared to SDR videos, HDR videos

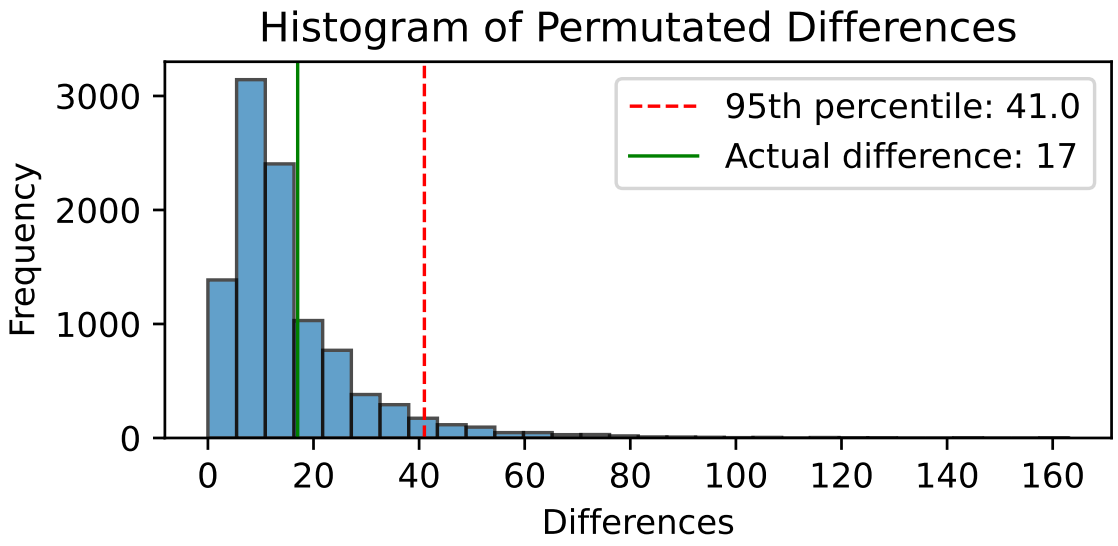


Figure 3.7: Distribution of significant differences (D') under random group assignment for the permutation test. The observed value $D = 17$ and the 95th percentile of the D' distribution are also shown, indicating that the observed differences in scores under bright and dark ambient conditions are not statistically significant.

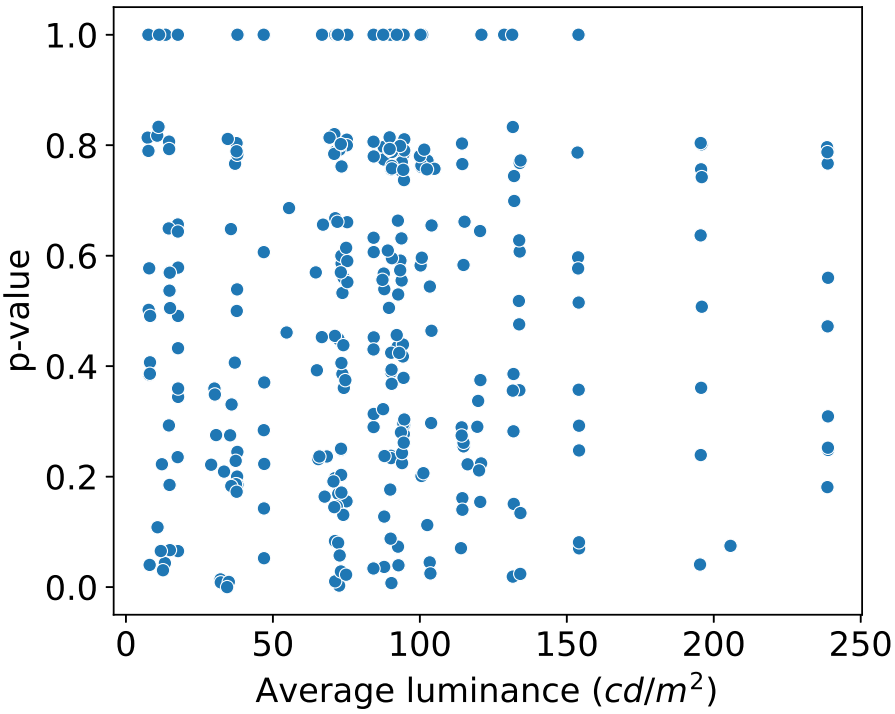


Figure 3.8: Scatter plot of the p-values of the raw score comparison against the average luminances of each video .

contain lower black levels, higher peak luminances, and more brilliant colors. Rich visual information, and visible distortions, can be observed in the dark and bright zones, both affecting subjective quality; however, conventional SDR VQA models have difficulty capturing this information.

The reason for this is that the responses of conventional VQA feature sets are dominated by, or at least strongly affected by distortions on regions that are "SDR-like," *i.e.*, occupying the mid-range of brightnesses. The feature responses to very dark and bright regions become dilute, greatly reducing the sensitivity of standard VQA models to highly conspicuous "HDR" distortions. Moreover, the visual response to luminance is highly nonlinear. The visual system is able to map large ranges of luminances onto much smaller ranges of perceived lightness, thereby achieving a high

degree of compression Radonjić et al. (2011). For a distortion of fixed magnitude, the Weber ratio of luminance is higher on dark and is reduced as the luminance increases. Thus, small changes in luminance in dark regions will be more noticeable than in bright regions.

Because of these reasons, distortions on the darkest and brightest areas have distinct perceptual responses and contributions to perceptual quality. The perceptual distortion information in these areas is not effectively captured by conventional VQA feature sets. Thus, we introduce additional feature computation pathways to capture "HDR-specific" features in parallel with the traditional "SDR" features, to better account for perceived distortions in these areas.

Specifically, we introduce HDRMAX, a simple but effective way to process bright and dark regions separately, and computing HDR-aware quality features on them, while avoiding complicated computations such as image segmentation. Instead, we define a pair of nonlinear transforms that expand the luminance ranges of very dark and bright regions, at the expense of the mid-range, which effectively amplifies the impact of "HDR" distortions on VQA feature responses. Following the transforms, we define separate and parallel feature extraction paths, to drive the quality-aware features specific to each of the areas, so that features computed on the nonlinearly altered frames can be used to augment conventional SDR VQA features.

3.5.1 Double Exponential Nonlinearity

The main characteristic of the nonlinear transform is to stretch the brightness values near the minimal (darker) and maximum (lighter) values, thereby enhancing the contrast there.

Neural responses are adequately modelled as sigmoidal functions Billock and Tsou (2011):

$$R = R_{max} \frac{I^n}{I^n + I_s^n}, \tag{3.7}$$

where R is the response to an input signal I, R_{max} is a maximum response, I_s is a semisaturation constant, and n depends on the type of neuron, but usually falls in the range $[1, 2]$ Bertalmío (2020). The sigmoidal function has the greatest slope for the smallest input magnitudes, gradually decreasing as the input increases.

We selected an exponential functions as a simple and effective way to amplify the brightness values at the extreme ends of the dynamic range in a nonlinear fashion, while gradually compressing the mid-range brightness values. This choice was guided by the simplicity of an exponential function's form and the control it provides over the degree of expansion through its parameters. The numerical stability it offers also contributed to its selection. While we do not claim that it accurately models the perceptual response, its use is quite perceptually relevant to VQA model design. The reason is that it is making perceptually relevant distortion information more available to VQA algorithms. It does this in a way that is copacetic with theories of distortion-sensitive natural video statistics. In this sense it may be viewed as a pooling preprocessing step that can remedy the defects of current learning-based VQA models. Since it is not meant to model a biological perceptual process, there may be other functional forms that are as effective, or more so, but our choice is a simple one. Moreover, HDRMAX incorporates a local adaptation operation, a process fundamental to vision, facilitating adjustment to a wide range of brightness values. Local adaptation adjusts the sensitivity of the visual system based on the local luminance level, acting specifically on each region of the retina Ledda et al. (2004). A refined model of this process, building upon the Naka-Rushton equation, has been proposed to simulate the physiological adaptations of the retina. Particularly, it modifies the half-saturation parameter, depending on the local luminance level. Inspired by this model of local brightness adaptation, we integrated a mean debiasing operation into HDRMAX. This operation precedes the exponential transform, its purpose being to adjust the nonlinearity based on the local mean luminance, thereby preserving sensitivity across different local luminance levels within each frame.

In the context of the HDRMAX augmentation, the mean debiasing operation

is positioned before the input into conventional SDR VQA models. This reflects the local adaptation model that simulates the initial stages of visual processing in the retina. Implementing this operation before later stages of the visual pathway modeled by existing SDR VQA models aligns with the natural flows of visual processing. As a result, HDRMAX ensures that the local nonlinear operation maintains sensitivity and responsiveness across varying local luminance levels.

The basic goal of HDRMAX is to address the inability of conventional SDR VQA models to capture some HDR distortion characteristics. Our method makes better available distortion information in the extreme range of luminance and color that are highly visible but not well accessed by current VQA models such as VMAF. We do this by introducing a separate processing pathway that expands the extreme ends of the dynamic range. This is accomplished by introducing an expansive nonlinearity whose outputs are nicely analyzable using natural video statistics model.

The nonlinearities are applied on the perceptually uniform PQ-encoded luma. However, their inherent flexibility also enables their use with linear luminance. An advantage of applying the nonlinearities on perceptually uniform luma is that it allows for predictable modifications to video content. This predictability enables a clear understanding of how the nonlinearities stretch or compress bright/dark regions, providing a greater level of control over the quality assessment process.

Assume that the brightness values $I(x, y, t)$ fall within the range $[0, 1]$. If they don't, linearly scale the brightness range $[A, B] \mapsto [0, 1]$, where A and B represent the minimum and maximum brightness values within each frame, respectively. This scaling operation aligns the dynamic range of each frame's brightness values with the $[0, 1]$ interval, while controlling the strength of the applied exponential function, maintaining uniformity across each frame and avoiding extreme values. We then apply point operations on the scaled brightness values, with the goal of nonlinearly expanding the dynamic ranges of the extreme high (bright) and dark ends. Once these operations are applied, feature extraction is conducted in parallel on three

videos at each frame instant - two nonlinearly transformed, and the original. The nonlinear transformations are adaptive, since it includes local mean debiasing. The two nonlinearly transformed videos are given by:

$$\widetilde{I}_1^l(x,y,t) = exp[\delta_1(I(x,y,t) - \bar{I}^l(x,y,t))], \tag{3.8}$$

and

$$\widetilde{I}_2^l(x,y,t) = exp[-\delta_2(I(x,y,t) - \bar{I}^l(x,y,t))]. \tag{3.9}$$

The parameters $\delta_1, \delta_2 \in 0.5, 1, 2, 5$ in equations 8 and 9 control the expansion strength in the bright and dark areas, respectively. This choice, akin to a log grid search, offers a balance between model complexity and computational feasibility, and appropriately captures the inherent data patterns.

These parameters help modulate the representation of HDR details in dark and bright regions. Extreme values could lead to under-detailed or unnaturally contrasted images, emphasizing the need for careful selection of these parameters. In the experiments, we fixed $\delta_1 = 0.5$ and $\delta_2 = 5$ but we discuss these choices and how performance varies with them in the performance evaluation section. $\bar{I}^l(x,y,t)$ is the local mean brightness estimate:

$$\bar{I}^l(x,y,t) = \sum_{k=-K}^{K} \sum_{j=-L}^{L} w_{k,l} I_{k,l}(i,j), \tag{3.10}$$

where $w = \{w_{k,l} | k = -K, ..., K, l = -L, ..., L\}$ is a 2D circularly-symmetric unit-volume Gaussian weighting function sampled out to 3 standard deviations. We used $K = L = 31$ in our experiments and we discuss the choice of the parameter later in the performance evaluation section.

We show plots of the exponential transforms in Fig. 3.9, illustrating the expansion of the extreme dark and bright ranges. We use separate transformations, because it allows flexibility when accessing information at the bright and dark ends. For example, we assume throughout that the brightness values are expressed as luma, rather than luminance. In most HDR streaming video workflows, the PQ OETF is

applied to the linear luminance signals received by RGB sensors to convert them to nonlinear color R'G'B', which are then weighted and summed to compute luma and color-difference channels ($Y'C'_BC'_R$, sometimes referred to as YUV.) The nonlinearities (3.8)-(3.10) are flexible enough to be used either on luma or on luminance, the latter of which has already been transformed by an asymmetric nonlinearity. Moreover, the relative sensitivities of the human eyes to distortions in bright and dark areas is at least partly determined by Weber's Law, which states that the visibility of signal perturbations is affected by the local brightness.

Two sample reference frames taken from the 'flower' and the 'firework' videos, as well as the result of applying the nonlinear transformations to the 'flower' and 'firework' frames, are shown in Fig. 3.10. The 'flower' video frame contains areas containing mostly mid-range brightness values, while the 'firework' video frame contains very bright areas on a very dark background. As such, the nonlinearly processed 'firework' video will contain more heavily enhanced areas. Of course, these printed representations are not HDR and are being shown to give an idea of the applied effects. To illustrate the effects on distortion visibility, we also show magnified areas of 'flower' and 'firework' before and after compression and with nonlinearities applied in Fig. 3.11 and Fig. 3.12. To demonstrate the amplification of distortions on the bright areas, we also show the result of applying transformation (3.8) and (3.9). As may be observed, application of the nonlinear transformation greatly enhances the distortions in the bright regions of 'firework,' and less so on the mid-range distortions in 'flower.'

3.5.2 Modifying VMAF Using HDRMAX Features

VMAF is a data driven video quality framework that extracts several highly successful VQA features, then uses a trained SVR to map the features to human judgments. The features used in VMAF 2.3.0 include the Detail Loss Metric (DLM), four Visual Information Fidelity (VIF) features computed on different oriented fre-

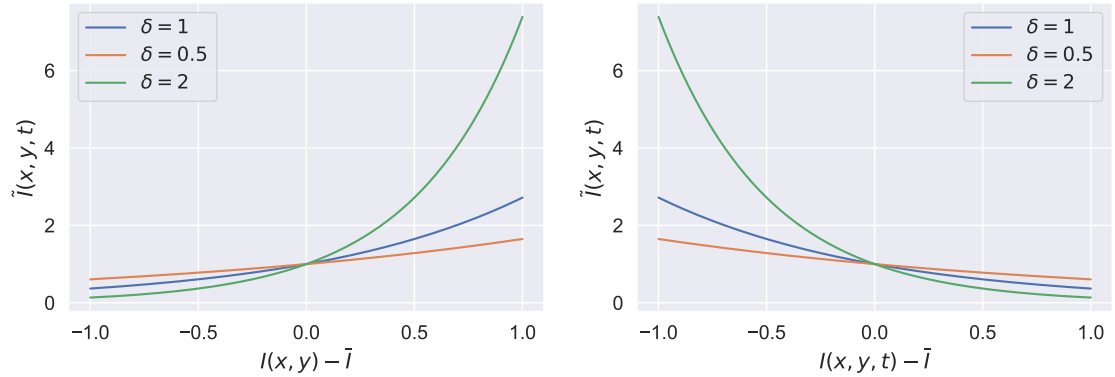


Figure 3.9: The two exponential transforms in (3.8) (left) and (3.9) (right) plotted for several values of the expansion parameters δ_1 and δ_2.

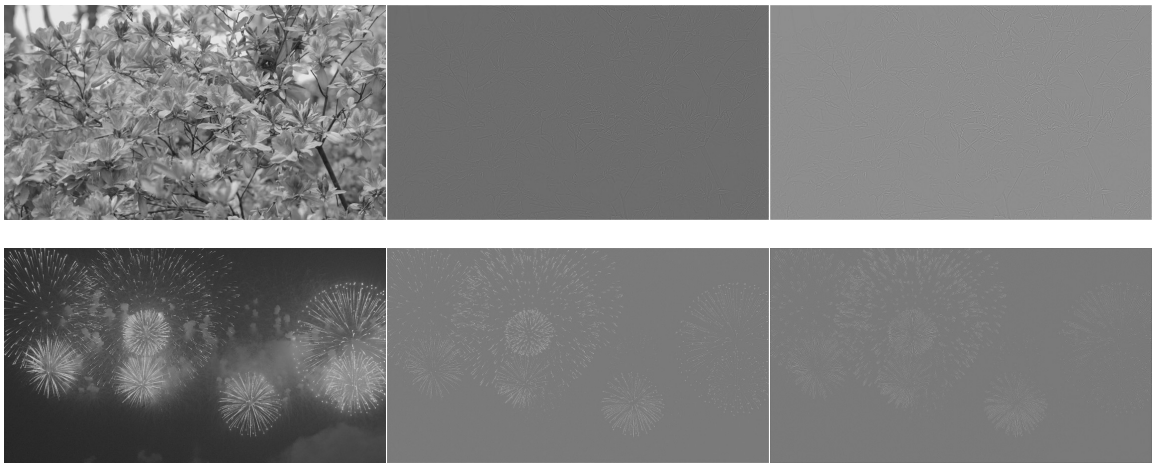
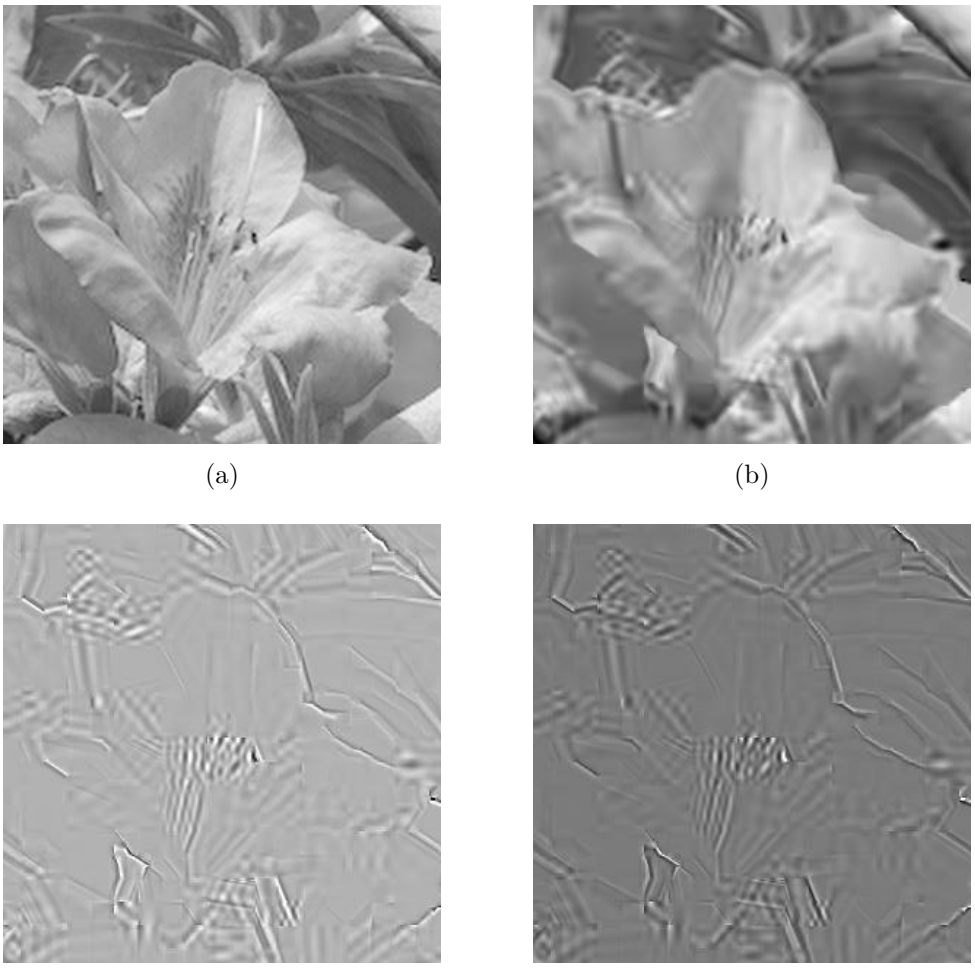


Figure 3.10: The reference frames 'flower' and 'firework' (left), the transformed reference frames after processing with (3.8) (middle) and (3.9) (right).

quency bands, and a simple frame difference feature, all of which are applied on the PQ luma component only. Modifying VMAF to include HDRMAX features is quite simple. On the brightness component of each video frame, also compute the nonlinearly transformed frames $\widetilde{I}_1^l$ and $\widetilde{I}_2^l$, along with the usual VMAF features computed on I. Table 4.2 summarizes the features used.



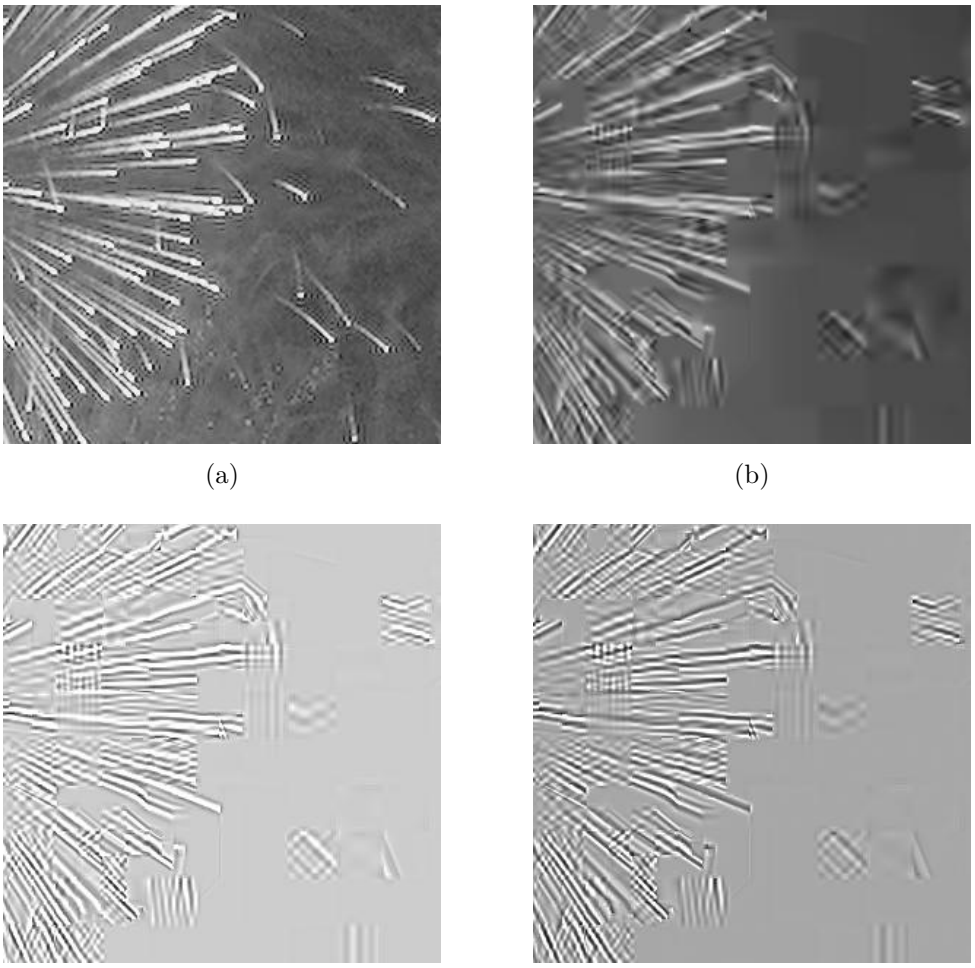
(a) (b)

(c) (d)

Figure 3.11: A patch from 'flower'. (a) from the reference frame; (b)-(d) from the compressed frame. (b) before nonlinear transformation; (c) after nonlinear transformation (3.8); (d) after nonlinear transformation (3.9).

3.6 Objective Video Quality Assessment Experiments

As a way of demonstrating the usefulness of the new LIVE HDR Database, we used it to study the performance of several existing HDR VQA models, as well as state-of-the-art (SOTA) SDR VQA models. We also studied the performance of VMAF augmented by HDRMAX features as its parameters were varied.



(a) (b)

(c) (d)

Figure 3.12: A patch from 'firework'. (a) from the reference frame; (b)-(d) from the compressed frame. (b) before nonlinear transformation; (c) after nonlinear transformation (3.8); (d) after nonlinear transformation (3.9).

3.6.1 Evaluation Criteria

We used the SUREAL scores owing to their statistical reliability. Since they are absolute quality scores, we obtained quality differences referred to as difference MOS (DMOS). Given a video indexed j with SUREAL score ψ_{dj}, compute the difference

Table 3.4: Descriptions of Features

Feature index	Description
$f_1 - f_5$	VIF and DLM features from the original frame.
f_6	Motion feature
$f_7 - f_{16}$	VIF and DLM features from the frames following the non-linear transformation.

score

$$D\psi_{dj} = \psi_{dj}^{ref} - \psi_{dj}. \tag{3.11}$$

The performances of the compared algorithms, including VMAF+HDRMAX, were evaluated using three standard metrics: the Spearman' s Rank Order Correlation Coefficient (SROCC), the Pearson Linear Correlation Coefficient (PLCC), and the Root Mean Square Error (RMSE). Following common practice Sheikh et al. (2006), we fit the predicted scores to the real scores using a logistic function

$$f(s) = \beta_1(\frac{1}{2} - \frac{1}{(1 + \exp(\beta_2(s - \beta_3))}) + \beta_4 s + \beta_5 \tag{3.12}$$

before computing the PLCC and the RMSE.

3.6.2 Evaluation Protocol

We used an SVR to learn the mappings from features to DMOS. The SVR was implemented using the linear kernel. All of the compared algorithms were evaluated using 1000 random train-test splits. On each split, 80% of the data was used for training, and the other 20% for testing, while not allowing any sharing of content between training and testing subsets. Notably, the new dataset includes several videos derived from the same longer clips, specifically, the football videos (football 1-8) and golf videos (golf 1-2). We diligently ensured that these videos were not split between the training and testing sets, to avoid any potential leakage of similar content between

the sets. We applied 5-fold cross-validation to find the optimal SVR parameters for each training set.

3.6.3 Performance Evaluation of VMAF+HDRMAX

We tested the performance of VMAF+HDRMAX against different choices of the expansion parameters δ_1 and δ_2. For each parameter combination, we computed the 16 features in Table 4.2, on the LIVE HDR Database and conducted 1000 train-test splits. The median values of the obtained performance metrics SROCC, PLCC and RMSE are given in Table 3.5. For better visualization, a heatmap of the SROCC as the parameters δ_1, δ_2 were varied is shown in Fig. 3.13. As may be observed, smaller values of δ_1 and larger value of δ_2 generally resulted in higher SROCC, while $(\delta_1, \delta_2) = (0.5, 5.0)$ yielded the best SROCC. One possible explanation for this is that HDR10 videos extend the original SDR luminance range from 0.01-100 nits to 0.0001-10000 nits. The difference between the darkest blacks of SDR and HDR is much less than between the brightest SDR and HDR values, suggesting that greater expansion is required on the darker end. However, although the choice of the parameter selection does influence the measured model efficacy, the differences are not large, and every choice and combination resulted in excellent performance relative to other, prior models. This demonstrates the efficacy of the nonlinear transformation and HDR features.

We also conducted experiments on the patch size W used in transformation (3.8) and (3.9). The results for $W = 9, 17, 31$ and 63 are reported in Table 3.6 using $\delta_1 = 0.5$ and $\delta_2 = 5$. We avoided W values that are multiples of 4 to avoid alignments of the transformation window edges with compression block boundaries. The choice of window size had a minor effect on performance, but we chose the one giving the highest degree of correlation between predicted quality against human judgments.

We also studied other design choices. First, we extended the nonlinear transformation to the components of three color spaces: the BT.2020 RGB color space, the

Table 3.5: Performance of Luma VMAF+HDRMAX as the Expansion Parameters δ_1 and δ_2 Varied, for Using the Nonlinear Transform (3.8)-(3.10). The Top Performing Combination is Boldfaced.

δ_1	δ_2	SROCC	PLCC	RMSE
0.5	0.5	0.8470	0.8056	11.9296
0.5	1	0.8238	0.7918	11.4521
0.5	2	0.8610	0.8167	10.8815
0.5	5	**0.8755**	**0.8397**	**10.1410**
1	0.5	0.8516	0.8099	11.6104
1	1	0.8500	0.8125	11.5388
1	2	0.8628	0.8303	10.9217
1	5	0.8584	0.8213	11.2416
2	0.5	0.8335	0.7861	11.5870
2	1	0.8282	0.7953	11.5907
2	2	0.8433	0.8200	10.1993
2	5	0.8540	0.8268	10.1404
5	0.5	0.8378	0.8003	11.8099
5	1	0.8422	0.8086	11.3328
5	2	0.8203	0.8081	11.6327
5	5	0.8216	0.7958	11.6869

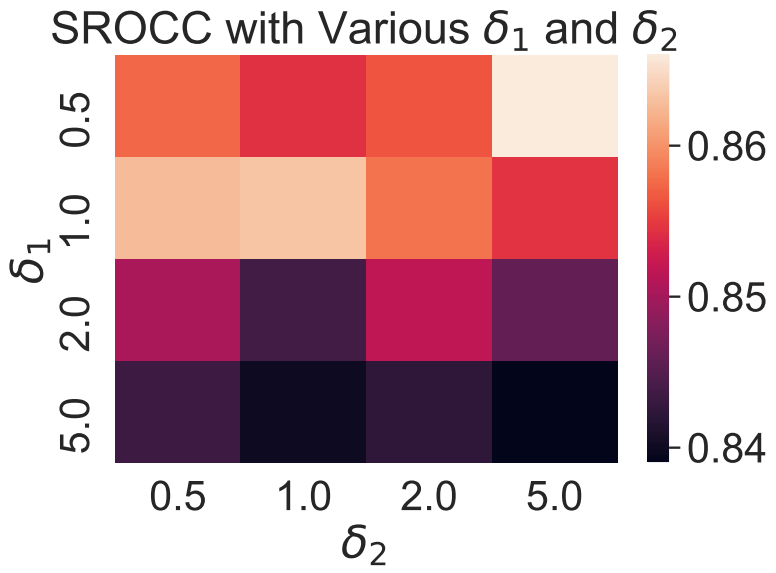


Figure 3.13: A heatmap visualizing median SROCC as (δ_1, δ_2) are varied for the nonlinear transformation (3.8)-(3.10).

YC_BC_R ITU color space, and the $HDR-Lab$ Fairchild and Chen (2011) color space. The RGB space is associated with acquisition and display. YC_BC_R is a common format for HDR videos. In $HDR-Lab$, the $L*$ component captures the perceived lightness of a color as compared to a white reference. The $a*$ and $b*$ components represent the position of the color between red/magenta and green, yellow and blue respectively. For each variant model, we extracted the original six VMAF features on each channel, and also extracted the four VIF features and the DLM feature on the nonlinearly transformed frames of each component. Thus, each color variant of VMAF+HDRMAX utilizes 46 features. As a final comparison model, we applied the nonlinearity (3.8)-(3.10) on the linear luminances instead of the PQ luma values, but without any color components. The performance results for these four variants of HDRMAX are shown in Table 3.7. The results for all models were quite good, but not as high as for the luma-only VMAF+HDRMAX results. Since the database contains videos that have excellent color diversity and coverage, this suggests that most of the distortion artifacts can be captured and analyzed within the luma channel, while

Table 3.6: Performance of the Nonlinear Transform for Various Window Sizes. Top Performance is Boldfaced.

W	SROCC	PLCC	RMSE
9	0.8601	0.8265	11.1056
17	0.8552	0.8354	11.0654
31	**0.8755**	**0.8397**	**10.1410**
63	0.8675	0.8205	11.1852

Table 3.7: Performance of Color Variants of VMAF+HDRMAX. The "setting" column indicates the color space. "Linear" Indicates the Two-Exponential Transform and Features are Performed on the Linear Luminance Values. The Top Performance in Each Domain is Boldfaced.

Setting	SROCC	PLCC	RMSE
$HDR-Lab$	0.7850	0.7348	14.3641
RGB	0.7986	0.7477	13.3448
YC_BC_R	0.8025	0.7502	13.7340
linear	**0.8355**	**0.8068**	**11.3307**

increasing the dimension of the feature space slightly reduces the model performance.

3.6.4 Comparison Against Other VQA Models

We also evaluated several other FR HDR and FR SDR VQA models on the new database and compared them against the VMAF+HDRMAX. The existing HDR algorithms we studied are the latest PU21 enhanced models, including PSNR, SSIM, MS-SSIM, FSIM Zhang et al. (2011) and VSI Zhang et al. (2014), HDR-VDP2.2, HDR-VDP3, and HDR-VQM, while the compared SDR methods are PSNR, SSIM,

MS-SSIM, STRRED, SpEED-QA, and VMAF. Most of these models are not trained. We listed both the pre-trained and retrained VMAF for comparison. The results of the comparison are shown in Table 3.8 and Table 3.9 against the DMOS obtained from the dark environment and bright environment respectively. It may be seen that VMAF modified using HDRMAX was able to significantly outperform the other models, including retrained VMAF. The fact that VMAF+HDRMAX outperforms VMAF by a large margin implies that the unmodified VMAF largely captures distortions from the usually dominant mid-range of brightness.

To further substantiate our claim, we performed a one-sided t-test for statistical analysis. For each model, we used 1000 SROCC values, obtained from individual train-test splits. In the case of models that do not require training, we randomly selected a 20% video sample to calculate a comparable SROCC sample. The single-sided t-test was then performed on the SROCCs between our proposed VMAF+HDRMAX method and the rest of the models, under both bright and dark conditions. The details of these t-test analyses can be found in Table 3.10. It may be observed that the SROCC values for pretrained and retrained VMAF appear to be similar in Tables 3.8 and 3.9, but show some difference in Table 3.10. This minor difference arises from the fact that we sample 20% of the videos for the pretrained VMAF in the process of t-test, leading to slightly varied SROCC values obtained from these samples. This provided statistical evidence of our method's superior performance, with all p-values below the threshold of 0.05, denoting statistical significance.

3.6.5 Evaluation on SDR Database

We also trained and evaluated VMAF+HDRMAX on the SDR-only LIVE Livestream Database Shang et al. (2022, 2021b) to study the efficacy of the nonlinear transformation prior to conducting SDR VQA. We also re-trained the original (SDR) VMAF in a similar manner for a fair comparison. The LIVE Livestream Database was selected because it is both modern and very diverse. It contains 315 videos of varying

Table 3.8: Performance of the Compared HDR and SDR Quality Models Evaluated Using the Scores from the Dark Environment. The Top Performance is Boldfaced.

Method		SROCC	PLCC	RMSE
SDR Quality Models	PSNR	0.5798	0.6229	13.6735
	SSIM	0.4982	0.4925	15.2124
	MS-SSIM	0.5139	0.5252	14.8741
	STRRED	0.5670	0.5506	14.5913
	SpEED-QA	0.5716	0.5685	14.6258
	VMAF (original)	0.7628	0.7492	12.2953
	VMAF (retrained)	0.7940	0.7679	11.4522
HDR Quality Models	HDR-VDP2.2	0.5868	0.5128	15.0052
	HDR-VDP3.0.7	0.7363	0.7307	11.9332
	HDR-VQM	0.5543	0.5450	14.3890
	PU21-PSNR	0.5841	0.5767	14.2798
	PU21-SSIM	0.6019	0.6065	13.8971
	PU21-MSSSIM	0.6593	0.6564	13.1868
	PU21-FSIM	0.6470	0.6372	13.4705
	PU21-VSI	0.6795	0.6667	13.0284
	VMAF+ HDRMAX	**0.8755**	**0.8397**	**10.1410**

Table 3.9: Performance of the Compared HDR and SDR Quality Models Evaluated Using the Scores from the Bright Environment. The Top Performance is Boldfaced.

Method		SROCC	PLCC	RMSE
SDR Quality Models	PSNR	0.6268	0.6621	13.0476
	SSIM	0.5493	0.5406	14.6461
	MS-SSIM	0.5740	0.5831	14.1442
	STRRED	0.6373	0.6167	13.7048
	SpEED-QA	0.6435	0.6254	13.6944
	VMAF (original)	0.8184	0.7947	11.0224
	VMAF (retrained)	0.8133	0.7890	11.0915
HDR Quality Models	HDR-VDP2.2	0.6472	0.6254	13.9861
	HDR-VDP3.0.7	0.8080	0.8098	10.2139
	HDR-VQM	0.6315	0.6144	13.5114
	PU21-PSNR	0.6117	0.5963	13.9762
	PU21-SSIM	0.6403	0.6301	13.5188
	PU21-MSSSIM	0.7120	0.6969	12.4859
	PU21-FSIM	0.7116	0.6904	12.5951
	PU21-VSI	0.7290	0.7058	12.3334
	VMAF+ HDRMAX	**0.8693**	**0.8256**	**10.6864**

resolutions (1080p and 4K) multiple types of distortions and significant high-motion temporal content. It offers professional-quality videos captured under controlled lab conditions, similar to the anticipated application scenarios of the HDRMAX model. Moreover, there is no content overlap with the LIVE-HDR database, ensuring independent evaluation.

Our findings, displayed in Table 3.11, indicate that HDRMAX notably enhances performance on SDR content as well, underscoring the value of focusing on dark and bright regions during VQA. This improvement does not merely result from an increase in the size of the feature space. In the context of machine learning, it is widely recognized that adding more features does not inherently enhance model performance. Instead, the efficacy of a feature lies in its discriminative power and its relevance to the task at hand. The features added by HDRMAX are both discriminitive and highly sensitive to video quality characteristics, thus contributing to improved performance. Recognizing potential interest in the contribution of HDRMAX features, we also include the standalone performance of these features.

I removed the blue font color from the table.

3.6.6 Evaluation on HDR Inage Database

To better illustrate the generalizability of our method, we conducted additional testing on the Unified Photometric Image Quality dataset (UPIQ) Mikhailiuk et al. (2022). UPIQ is an expansive collection of over 4000 HDR and SDR images, and has proven to be a valuable resource for developing and validating HDR metrics. However, given the scope of our study, we focused exclusively on the 380 HDR images in UPIQ.

It is noteworthy that the images in UPIQ are represented in absolute photometric and colorimetric units, reflecting light emitted from a display. To make these images compatible with our method, we transformed the pixel values into PQ before applying our models. We show the results in Table 3.12. Although our model didn't outperform all of the existing HDR metrics on this dataset, it still demonstrated commendable

performance. This extra evaluation indicates the potential of our approach on diverse HDR contents and highlights its applicability to real-world scenarios.

Table 3.10: Statistical Analysis of Model Comparisons

	Test Condition	Dark		Bright	
	Model	t-statistic	p-Value	t-statistic	p-Value
SDR Quality Models	PSNR	7.32	1.78E-13	3.72	1.01E-04
	SSIM	23.67	1.07E-109	21.42	4.43E-92
	MS-SSIM	31.76	6.76E-180	25.71	1.72E-126
	ST-RRED	14.20	5.87E-44	12.60	2.18E-35
	SpEED-QA	19.32	1.02E-76	14.31	1.30E-44
	VMAF (original)	29.59	3.34E-160	17.93	4.18E-67
	VMAF (retrained)	2.63	4.30E-03	3.52	2.23E-04
HDR Quality Models	HDR-VDP 2.2	13.46	6.78E-40	12.46	1.19E-34
	HDR-VDP3.0.7	40.53	5.86E-263	24.05	9.69E-113
	HDR-VQM	405.60	0	444.17	0
	PU21-PSNR	74.90	0	74.11	0
	PU21-FSIM	61.62	0	48.35	0
	PU21-MSSSIM	57.80	0	74.55	0
	PU21-SSIM	73.21	0	69.63	0
	PU21-VSI	53.53	0	45.73	2.44E-313

Table 3.11: Performance of the Evaluated Algorithms on LIVE Livestream Database. The Top Performance is Boldfaced.

Algorithms	SROCC	PLCC	RMSE
PSNR	0.3760	0.4192	10.3355
SSIM	0.6976	0.7107	8.0082
MS-SSIM	0.6757	0.6907	8.2324
STRRED	0.6564	0.6694	8.4573
SpEED-QA	0.6894	0.7235	7.8589
VMAF (original)	0.6434	0.6355	8.7894
VMAF (retained)	0.6836	0.6912	8.2712
HDRMAX	0.6613	0.6755	8.9744
VMAF+HDRMAX	**0.7632**	**0.7743**	**7.2468**

Table 3.12: Performance of the Evaluated Algorithms on UPIQ Database. The Top Performance is Boldfaced.

	SROCC	PLCC	RMSE
HDR-VDP 3.0.7	0.8448	0.8426	0.3528
HDR-VQM	**0.8893**	**0.8824**	**0.3082**
PU21-FSIM	0.7358	0.71944	0.4551
PU21-MSSSIM	0.8192	0.8193	0.3757
PU21-PSNR	0.4903	0.4192	0.5950
PU21-SSIM	0.7215	0.7270	0.4499
PU21-VSI	0.6792	0.6713	0.4857
VMAF+HDRMAX	0.8485	0.8417	0.3680

Chapter 4: A Subjective and Objective Study of Adaptive Quantization of HDR Videos

4.1 Related work

4.1.1 Adaptive Quantization

There have been numerous efforts in recent years to improve the effectiveness of AQ for video coding. Xiang *et al.* proposed a novel AQ algorithm that analyzes several factors influencing the efficacy of AQ while accounting for the temporal characteristics leading to visually pleasing quantization parameter (QP) offset distributions Xiang et al. (2018). He *et al.* proposed an adaptive frame-level QP selection algorithm for H.265/HEVC random access coding that considers inter-frame dependencies He et al. (2018). Bichon *et al.* designed per-block optimal quantizers that achieve global rate-distortion optimization, which was incorporated into the HEVC reference model Bichon et al. (2019). Dai *et al.* proposed a perceptual AQ technique based on a convolutional neural network (CNN) and HEVC Dai et al. (2022), which adaptively determines CTU-level QP values in HEVC intra-coding using high-level features extracted by the CNN. Vu *et al.* used the Video Multimethod Assessment Fusion (VMAF) algorithm to find the optimal QP for the x.264 codec, resulting in bitrate savings Vu et al. (2022). In addition to these efforts to optimize the performance of AQ algorithms, Somdyuti *et al.* proposed a method for estimating the contrast masking threshold on natural scene patches, using these estimates to enhance AQ for AV1 encoding Paul et al. (2021). This method produces fewer visible compression artifacts at lower bitrates as compared to a variance-based AQ approach.

4.1.2 Subjective HDR Video Quality Databases

A number of studies have been conducted in the past to create perceptual video quality datasets for HDR content. All the databases used in these studies are com-

prised of professional contents. However, many of these datasets have limited usefulness due to either the rapid pace of development of the HDR standards, or copyright issues that prevent those authors from releasing the data publicly. For example, Azimi *et al.* conducted a study using 30 videos that were displayed on a non-standard HDR device supporting the older BT. 709 gamut, rather than the modern HDR10-compliant BT. 2020 gamut, and the PQ OETF was not applied prior to compression. The videos were also only 1080p resolution Azimi et al. (2018). Pan *et al.* Pan et al. (2018) conducted a study of the effects of compression on HDR quality using 6 source videos encoded using PQ, HLG, and the BT. 2020 color space, but the codec used for compression was AVS2, which has seen little industry adoption. The study included 144 videos that were rated by 22 subjects. Baroncini *et al.* Baroncini et al. (2016) conducted a study on 12 compressed HDR videos that were evaluated by 40 subjects, but the source contents did not follow the ITU Rec. BT 2020 standard, and the PQ OETF was not applied on the video data. Rerabek *et al.* Rerabek et al. (2015) conducted a study of 5 HDR videos, each distorted by 4 compression levels, with the aim of comparing objective HDR VQA algorithms. The videos were all of only resolution 944×1080, and the data was tone-mapped to 8-bit format before being displayed to the subjects. Athat *et al.* Athar et al. (2019) conducted a subjective study on 14 HDR10 source contents compressed by H.264 and HEVC to generate 140 distorted videos. More recently, Shang *et al.* conducted a subjective quality study on 42 source contents to benchmark the performance of leading FR VQA models on common streaming problems, including compression, scaling, and quality crossovers among resolutions and frame rates Shang et al. (2023).

By contrast with these previous HDR video quality studies, our investigation includes source videos that adhere to the widely used HDR10 standard. It also significantly expands upon previous efforts as it contains almost twice as many videos and more than double the number of subjective scores collected. Moreover, we systematically consider the important perceptual effects of AQ on resulting compressed HDR VQA study to date, and demonstrate its usefulness by evaluating the performance of

leading VQA models.

4.1.3 Objective VQA Algorithms

There have been substantial efforts dedicated to the development of objective VQA models aiming to automatically predict the perceptual quality of videos. FR VQA models operate by comparing pristine reference videos against distorted versions of them using perceptually motivated features and/or training data Wang et al. (2004); Sheikh and Bovik (2005). The MSE (or equivalently, the peak signal-to-noise ratio (PSNR)) has long been used as a basic index of video quality. More recent popular VQA models include Structural Similarity (SSIM) Wang et al. (2004), Multiscale SSIM (MS-SSIM) Wang et al. (2003), Gradient Magnitude Similarity Deviation (GMSD) Xue et al. (2013), most apparent distortion (MAD) Larson and Chandler (2010), visual information fidelity (VIF) Sheikh and Bovik (2005), and FSIM Zhang et al. (2011), among others Vu et al. (2011); Vu and Chandler (2014); Bampis et al. (2017a); Seshadrinathan and Bovik (2009). In recent years, machine learning-based FR VQA models have gained widespread popularity. One example is VMAF Li et al. (2017), which leverages features from two VQA models to drive a Support Vector Regressor (SVR) to predict video quality scores. FR VQA models that employ deep learning techniques have also demonstrated impressive performance, such as DeepVQA Kim et al. (2018). Additionally, some FR VQA models utilize unsupervised deep learning (UDL) methods, as in Vega et al. (2017).

Research on predicting the quality of HDR videos is in relatively early stages. One of the earliest algorithms, HDR-VDP Mantiuk et al. (2005), takes into account the nonlinear response to light of high contrast content and the full range of luminances. An improved version, called HDR-VDP-2 Mantiuk et al. (2011), uses a model based on contrast sensitivity measurements to account for all luminance conditions. Further developments of HDR-VDP-2 include the implementation of improved pooling methods (HDR-VDP2.2 Narwaria et al. (2015a, 2014)). Another approach,

proposed in Aydın et al. (2008), involves using a nonlinear transform to extend traditional SDR quality metrics to the HDR domain. Other researchers have focused on the chromatic aspects of HDR video quality, such as color fidelity Abebe et al. (2015), the use of HDR Uniform Color Spaces Rousselot et al. (2019), and color difference models Choudhury et al. (2021). Another method, called HDR-VQM, utilizes a spatio-temporal analysis to simulate human perception Narwaria et al. (2015b). The HDRMAX model Ebenezer et al. (2023); Shang et al. (2018), is a set of features that was designed by applying nonlinear transforms to enhance the measurability of the distortions in the brightest and darkest local regions of video frames. The features are used to improved the performance of state-of-the-art (SOTA) VQA models on HDR and 10-bit videos.

The majority of existing HDR quality prediction algorithms rely on simple transforms to map video features to quality predictions, such as the root mean square error (RMSE) used in color difference models, and spatial pooling in HDR-VDP-2. These approaches often target legacy HDR videos or simply HDR images, which differ significantly from the modern HDR10 standard. Additionally, many of these models lack sensitivity to distortions in smooth areas, which are particularly susceptible to banding and blocking artifacts. As such, there is a need for more advanced algorithms that can accurately predict the quality of modern HDR videos while taking into consideration these kinds of visual distortions.

4.2 Subjective Experiment Design

4.2.1 HDR Video Contents

We gathered 40 high-quality, distortion-free source HDR10 video clips. The videos include both Video-on-Demand (VoD) videos as well as live streaming contents acquired from one of the streaming service provider and thus, they are a good representation of real-world videos. Nevertheless, due to copyright restrictions, the video contents will not be made publicly available. However, we will extract quality-

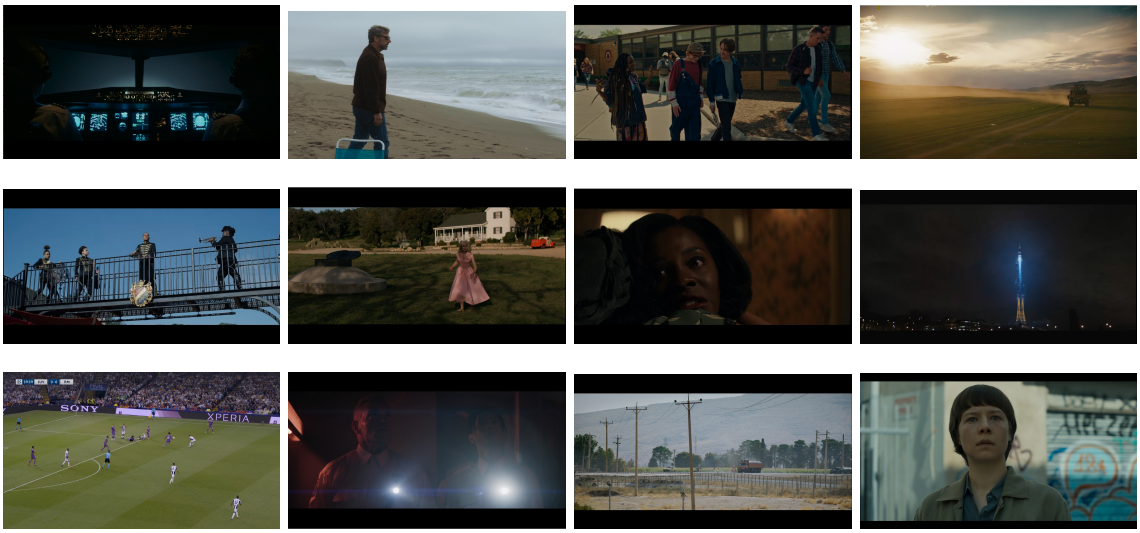


Figure 4.1: Exemplar frames from the source sequences.

aware features as done in Bampis et al. (2017b). All of the videos had acquisition, grading and production processes performed on them by industry professionals. The source sequences all have resolution 3840x2160 pixels, frame rate 25-30 fps and were progressively scanned with audio removed. Although four of the contents are sports-related and originally 50 fps, we temporally downsampled them to 25 fps to maintain consistency and avoid introducing an extra variable in our dataset. These contents include at least 10 videos having large smooth areas as well as texture-rich regions, on which the effects of compression are often more apparent, and which may benefit by the use of AQ. We carefully segmented the video sequences into clips of 7-10 seconds, varying the durations and endpoints to prevent awkward or annoying scene cuts.

Figure 4.1 presents several sample frames from the source sequences that were acquired for this study. These videos encompass a wide range of content. Following prior work Chen et al. (2021); Tu et al. (2021a), we calculated low-level descriptors on each source sequence, including spatial information (SI) (integrated Sobel magnitude), temporal information (TI) (absolute average frame differences) defined in ITU (2008), and a popular measure of colorfulness (CF) described in Hasler and Suesstrunk (2003).

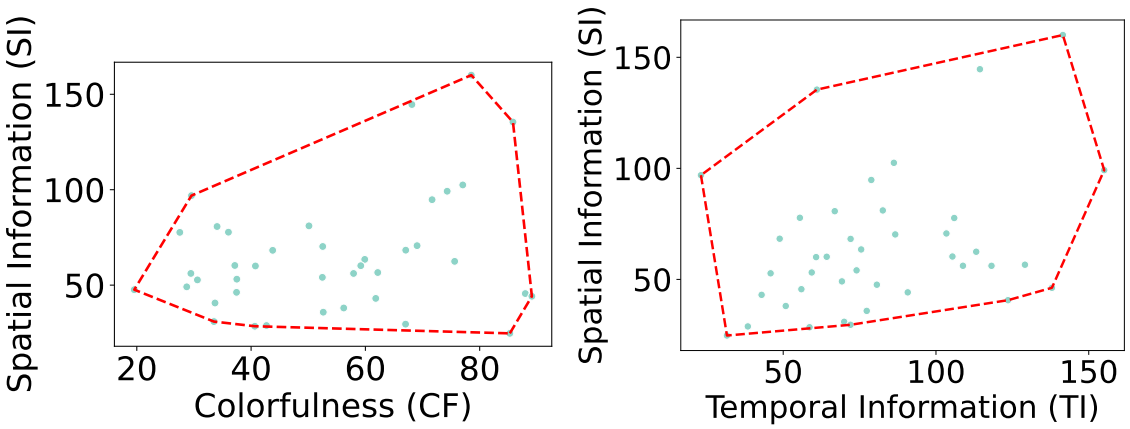


Figure 4.2: Spatial Information (SI) versus (a) colorfulness (CF) and against (b) Temporal Information (TI), measured on all of the source sequences in the new LIVE AQ-HDR Database. The corresponding convex hulls are plotted by red lines.

Figure 4.2 plots the SI, TI, and CF of all of the source sequences in the new database, illustrating the wide range of low-level content and activity present in these sequences in the spatial, temporal, and color dimensions.

4.2.2 Test Sequences

The new HDR VQA database contains a total of 400 videos, which is significantly larger than any prior HDR VQA dataset. It includes 30 source videos processed to obtain 8 different combinations (indexed 1-8 in Table 4.1) of bitrates and resolutions with the AQ option turned on. Additionally, we have 10 "AQ content" compressed with 12 different combinations, *i.e.*, combinations indexed 1-8 in Table 4.1 with the AQ option turned on, and another 4 combinations processed with level 2,5,7,8 in Table 4.1 with the AQ option turned off. These bitrates and resolutions were chosen to cocompass common HDR video streaming practice, using the HEVC Codec. To ensure a broad range of quality levels in the dataset, we introduced bitrate variations by sampling from Gaussian distributions centered on the specified mean bitrates. The means and standard deviations, as well as the AQ options for every considered

Table 4.1: Bitrate and Resolution Settings Used to Create the Distorted videos. Four of the Sports Videos are Compressed at 1.7 Mbps instead of 1 Mbps at Number 3.

Number	resolution	bitrate mean (Mbps)	bitrate stand deviation (Mbps)	AQ option
1	3840×2160	11.2	0.3	on
2	3840×2160	3	0.3	on/off
3a	3840×2160	1	0.1	on
3b	3840×2160	1.7	0.1	on
4	1920×1080	5.4	0.3	on
5	1920×1080	1.2	0.1	on/off
6	1920×1080	0.5	0.1	on
7	1280×720	1.5	0.3	on/off
8	960×540	0.8	0.2	on/off

bitrate and resolution combination indexed 1-8 are shown in Table 4.1. Every video was subjected to every bitrate/resolution combination 1-8, with the exception of combination 3, where the bitrates of four high motion videos were increased to 1.7 Mbps to avoid strong motion artifacts. These four processed videos were processed using combination 3(b), while the other 36 videos were processed as in combination 3(a).

4.2.3 Subjective Testing Design

We conducted a subjective human study in the Laboratory for Image and Video Engineering (LIVE) at The University of Texas at Austin. The study was conducted in a controlled environment, where participants viewed the videos on a 65-inch Samsung Class Q90T QLED 4K UHD HDR Smart TV TV connected to a workstation

with a 12 GB Titan X Graphics Processing Unit (GPU) via an HDMI 2.0b cable. The VLC player VideoLAN was used for HDR video playback, and all advanced temporal processing options on the TV were disabled to ensure the integrity of the test conditions and to avoid the introduction of any additional artifacts.

The participants were each asked to view every video, providing a quality judgments on each using a visible slider on the display, which was controlled with a mouse. The rating scale was continuous, with five Likert-like markers labeled "Bad," "Poor," "Fair," "Good," and "Excellent" placed at uniform intervals to guide the subjects when giving ratings. Scores were recorded as integers on the interval [0, 100], although numerical values were not made visible to the participants. In order to prevent bias due to initial positioning of the rating indicator, it did not appear on the sliding scale until the participant placed the cursor on the slider and clicked, where upon it became visible at that location, but remained visible and available to be moved and repositioned as the subject desired.

The first session for each subject was preceded by a training session where instructions were given, followed by six exemplar videos of two different content types that generally spanned the range of distortions that would be seen in the subsequent videos. The Absolute Category Rating with Hidden Reference (ACR-HR) protocol ITU (2008) was used for the training and test videos, and the order in which the videos were presented was randomized for each subject. Participants viewed the videos from a distance of approximately 1.5 times the height of the display.

4.2.4 Subjects

A total of 42 human subjects were recruited from the student population at The University of Texas at Austin. Each subject participated in two sessions separated by at least 24 hours. We applied the Snellen and Ishihara tests of each subject's visual acuity and color perception, respectively. No subject was found to have a color deficiency, and no subject had less than 20/30 visual acuity on the Snellen test, when

wearing their corrective lenses (if needed).

4.3 Processing of Subjective Scores

4.3.1 Computing of Mean Opinion Score

We computed Mean Opinion Score (MOS) using the statistical method proposed in Li et al. (2020). An improved method to recover MOS from noisy data called SUREAL finds a Maximum Likelihood (ML) estimate of the scores. Using this method, the opinion scores s_{ij} of video j from subject i are regarded as random variables

$$S_{ij} = \psi_j + \Delta_i + \nu_i X, \tag{4.1}$$

where ψ_j is the true quality of video j, Δ_i represents the bias of subject i, the non-negative term ν_i represents the inconsistency of subject i, and $X \sim N(0,1)$ are i.i.d. Gaussian random variables. The quantities $\psi_j, \Delta_{i_k}, \nu_i$ are jointly estimated in an iterative manner. By adjusting the weights according to subjects' inconsistency ν_i, the recovered MOS converges to a final value. Since this method assigns a smaller weights to subjects having higher inconsistencies, subject biases are accounted for when estimating the true qualities ψ_j, and the method is robust against subject inconsistencies. The inconsistency of each subject can also be used to reject the scores obtained from a subject.

The distribution of human subjective scores is depicted in Figure 4.3, using a box plot. As may be seen, the distortion combinations provided a wide range of perceived qualityies. The box at 2160p 1.7Mbps represents the four afore mentioned high motion videos, which exhibit a similar range of quality as the other contents compressed at 1Mbps despite their increased bitrate.

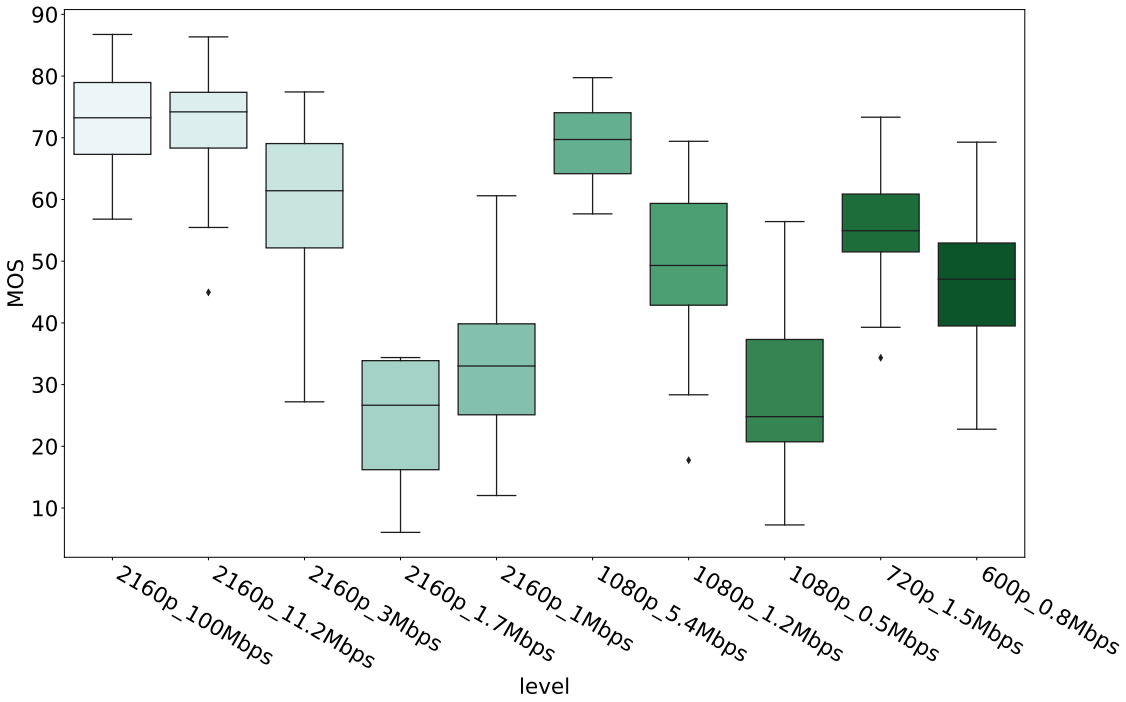


Figure 4.3: Box plots of the distributions of the MOS at each bitrate and resolution combination.

4.3.2 Effect of AQ on MOS

We also plotted the SUREAL MOS of the ten contents for the two AQ options (on and off) in Fig 4.4. It may be observed that the quality of the videos is more affected by bitrate, and less so by AQ. Intriguingly, some contents, such as content 2 and content 6, which feature particularly large smooth sky areas, appear to benefit from AQ across the entire bitrate range. Conversely, other contents exhibit minimal improvement from AQ or only show benefits at specific bitrates, contingent upon the nature of the content as shown in Fig. 4.5. We performed a t-test on the raw scores obtained from the subjective study to study the statistical significance of the scores. The resulted p-values are also plotted in Fig. 4.5. The computed p-values were usually above 0.05, indicating that although AQ produces differences in quality, these are relatively subtle, and rarely significant.

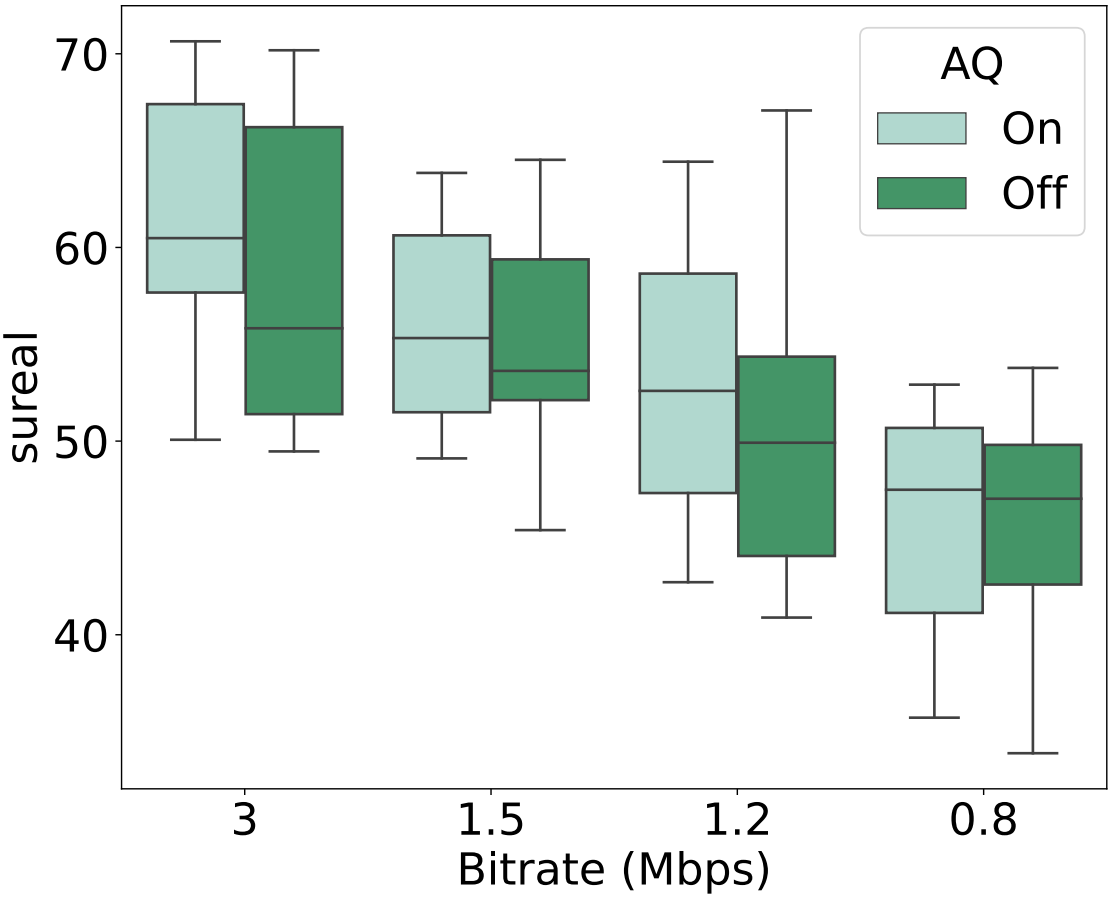


Figure 4.4: Box plots of the distributions of the MOS of the videos with both AQ options.

4.4 Objective Video Quality Model Design

Taking inspiration from models of the human visual pathway, we have developed new quality-aware features that we use to define a new VQA model that is able to accurately identify and analyze distortions related to AQ option, such as banding. This could be used in video compression workflows to affect decisions regarding whether to deploy AQ.

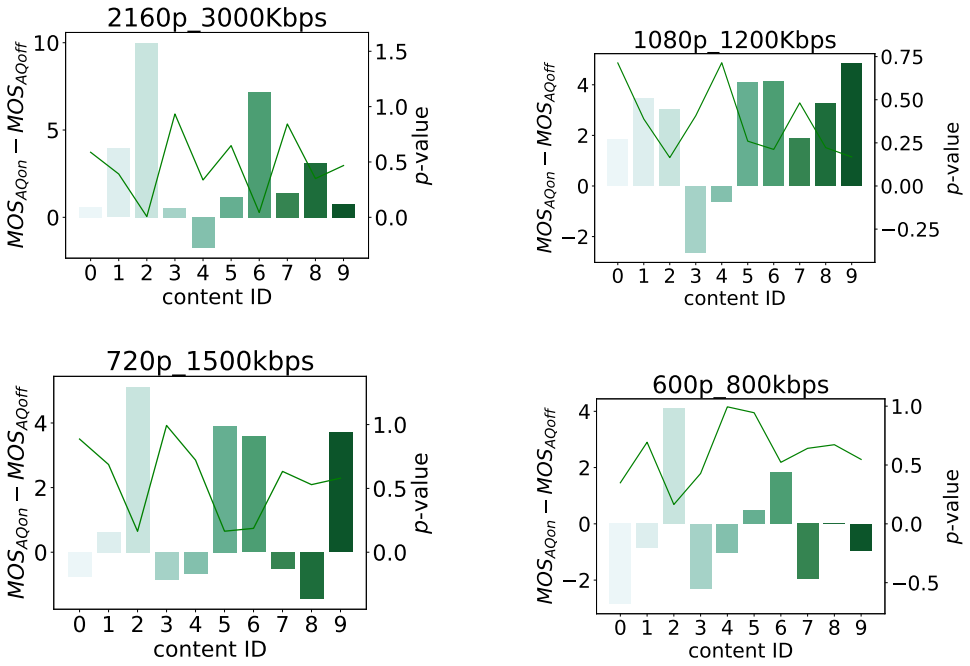


Figure 4.5: A group of bar plots showing differences of MOS between AQ enabled and AQ disabled contents. The left vertical axis of each plot express the MOS difference between the video scores with AQ on and AQ off, while the vertical axis on the right side of the plots are the p-values obtained by the t-test.

4.4.1 Banding Distortions

Banding is a distortion that commonly affects images and videos. It is characterized by the appearance of visible bands or stripes in the image, with abrupt transitions between adjacent colors, brightness, or tones. These bands can be caused by a variety of factors, including lossy compression, limited bit depth, and other types of quantization errors. To detect and measure the impact of banding artifacts on the perceived quality of videos, we propose a new set of features combines difference-of-Gaussian (DoG) filters, statistical features, and local histogram processing with the Space-Time GeneRalized Entropic Difference (ST-GREED) model Madhusudana et al. (2021b).

We design the DoG filters to target specific frequency bands in videos that may

be associated with banding artifacts. We show that this allows us to more accurately detect and measure the impact of banding. An example of the effect of the DoG filters is depicted in Fig. 4.6 and 4.7. Fig. 4.6 shows a frame of a video and a compressed version of it. The latter exhibits very noticeable banding artifacts in the dark sky regions. When the DoG filter is applied, as shown in Fig. 4.7, the banding artifacts become more pronounced.

Figure 4.6: The 'taipei' video. Top: an original frame; Bottom: the compressed frame exhibiting apparent banding artifacts. (The contrasts have been enhanced for visualization.)

In the ST-GREED model, the statistics of spatial and temporal bandpass video coeficients are analyzed to measure the perceived quality of videos of diverse frame rates. We replace the spatial filters in ST-GREED with specially designed DoG filters, customizing their frequency responses to better detect and isolate banding artifacts.

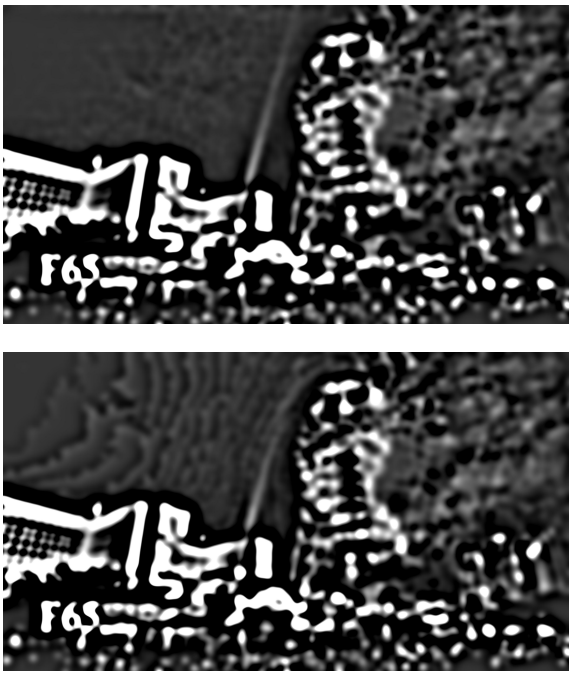


Figure 4.7: Application of DoG filters to the video frames in Fig 4.6. Top: original; Bottom: compressed showing enhanced banding artifacts. (The contrasts have been enhanced for visualization.)

4.4.1.1 Generalized Gaussian Distribution

The Generalized Gaussian Distribution (GGD) is a reliable model of the statistics of bandpass-processed natualistic videos Wang and Simoncelli (2005); Madhusudana et al. (2021b); Bampis et al. (2017a). Here, we deploy the GGD to model the bandpass statistics of reference and distorted videos denoted by R and D, having frames R_t and D_t, where t is the temporal index. The responses of a suitable band-pass filter to the reference and distorted videos will be denoted by B_t^R and B_t^D, respectively, both of which are assumed to follow a GGD model, the parameters of which may vary with t: B_t^R and B_t^D follows a GGD model, i.e., $B_t^R \sim GGD(\mu_t^R, \alpha_t^R, \beta_t^R)$ and $B_t^D \sim GGD(\mu_t^D, \alpha_t^D, \beta_t^D)$.

Here the location parameters μ_t are the distribution means, while the scale parameters α_t determine the variances, and the parameters β_t control the shapes of the

distributions (tail weight and peakiness). Let the bandpass coefficients at frame t be partitioned into non-overlapping patches/blocks of size $\sqrt{M} \times \sqrt{M}$ and indexed by $p \in \{1, 2, ...P\}$. The vectors of bandpass responses in patch p of subband k on frame t of the reference and distorted videos will be denoted B_{pt}^R and B_{pt}^D. Following Sheikh and Bovik (2006, 2005); Madhusudana et al. (2021b), we model any perceptual imperfections using an additive neural noise model:

$$\tilde{B}_{pt}^R = B_{pt}^R + W_{pt}^R, \quad \tilde{B}_{pt}^D = B_{pt}^D + W_{pt}^D \tag{4.2}$$

where B_{pt}^R and B_{pt}^D are independent of W_{pt}^R and W_{pt}^D, respectively, and where W_{pt}^R and W_{pt}^D are Gaussian distributed noise fields with zero mean and variance $\sigma_W^2 \mathbf{I_M}$, where $\mathbf{I_M}$ is the identity matrix of dimensions $M \times M$. The presence of distortion, such as banding, causes the statistics of videos to be altered in ways that can be predictive of quality. One way of capturing these statistical changes is by comparing (differencing) their entropies before and after distortions. The entropy of a GGD random variable $X \sim GGD(0, \alpha, \beta)$ has a closed form expression given by:

$$h(X) = \frac{1}{\beta} - \log(\frac{\beta}{2\alpha\Gamma(1/\beta)}) \tag{4.3}$$

To obtain the values of α and β, the bijective mapping between the GGD parameters and kurtosis is applied; interested readers can refer to Madhusudana et al. (2021b) for complete details.

4.4.1.2 DoG GREED Measure

We obtain spatial bandpass responses using a DoG filter

$$G(x, y) = G_1(x, y) - G_2(x, y), \tag{4.4}$$

where $G_1(x, y)$ and $G_2(x, y)$ are Gaussian kernels of the form

$$G_i(x, y) = \frac{1}{2\pi\sigma_i^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_i^2}\right) \tag{4.5}$$

where $\sigma_2 > \sigma_1$. The fourier transform of (4.4) is of the form

$$\tilde{H}(w) = \exp\left[-2(\pi\sigma_1)^2\left(\frac{w}{N}\right)^2\right] - \exp\left[-2(\pi\sigma_2)^2\left(\frac{w}{N}\right)^2\right]. \tag{4.6}$$

where we let $\sigma_1 = \sigma$ and $\sigma_2 = k\sigma$. The peak response of (4.6) occurs at

$$w = w_p = \pm\frac{N}{\pi\sigma}\sqrt{\frac{\ln\sqrt{k}}{k^2-1}}. \tag{4.7}$$

Following numerous other authors we take $k = \sqrt{2}$. The upper and lower half-peak cutoff frequencies of the DoG filter may be found by solving

$$\tilde{H}(w) = \frac{1}{2}\tilde{H}(w_p) \tag{4.8}$$

which yields $w_{LOW} \approx \frac{N}{4\pi\sigma}$ and $w_{HIGH} \approx \frac{N}{3\sigma}$. We explain the role of W_{LOW} and W_{HIGH} on model performance in Section 4.5.2.

The bandpass coefficients by filtering the reference and distorted videos using (4.4) are

$$R_t^{DoG}(x,y) = R_t * G(x,y), D_t^{DoG}(x,y) = D_t * G(x,y). \tag{4.9}$$

The spatial entropies $h(\tilde{R}_t^{DoG})$ and $h(\tilde{D}_t^{DoG})$ of the responses in (4.9) are calculated using equation (4.3), where under the additive noise model (4.2). The scaling factors and modified entropies are defined as in Bampis et al. (2017a); Madhusudana et al. (2021b); Soundararajan and Bovik (2011):

$$\begin{aligned} \eta_{pt}^R = \log(1+\sigma^2(\tilde{R}_t^{DoG})), \quad \eta_{pt}^D = \log(1+\sigma^2(\tilde{D}_t^{DoG})) \\ \theta_{pt}^R = \eta_{pt}^R h(\tilde{R}_t^{DoG})), \quad \theta_{pt}^D = \eta_{pt}^D h(\tilde{D}_t^{DoG}). \end{aligned} \tag{4.10}$$

Finally, we define the DoG-GREED index as:

$$\text{SGREED}_t = \frac{1}{P}\sum_{p=1}^{P}|\theta_{pt}^D - \theta_{pt}^R|. \tag{4.11}$$

4.4.2 Localized Histogram Equalization Features

Histogram equalization is commonly used to improve the contrast or dynamic range of an image by stretching the range of intensities or luminances values in an image, while approximately equalizing their frequencies of occurrence. This is done by calculating the empirical cumulative distribution function (CDF) then using it to remap the pixel values. LHE operates by applying histogram equalization on local spatial regions, defined either by image partition or a moving window Hummel (1977). This allows the brightness distribution to be modified in a more localized and selective manner. In the simplest form of LHE, each pixel is transformed by equalizing the histogram within a neighborhood of each pixel.

An example of the effect of LHE is depicted in 4.8, which shows the same frames as in Fig. 4.6. As depicted in Fig. 4.8, the banding artifacts become even more noticeable. When the LHE is applied, the frame exhibiting the banding effect is significantly enhanced on the same compressed frame as Fig. 4.6, especially along the edges of the banding, as shown in Fig. 4.8 (bottom). However, no such effect is observed on the original frame although many other features are highlighted, as may be seen in Fig. 4.8 (top). Given reference and distorted frames, R_t and D_t, denote the LHE processed frames by R'_t and D'_t. The next stage is to quantify the statistical differences between the LHE-processed videos. To do this, one can compute the local mean-subtracted, contrast-normalized (MSCN) coefficients R'^{MSCN}_t and D'^{MSCN}_t Ebenezer et al. (2020a); Mittal et al. (2012a); Kundu et al. (2017) often used to model contrast-gain masking processes in early human vision Carandini et al. (1997); Rao et al. (2001). The MSCN coefficients R'^{MSCN}_t and D'^{MSCN}_t of an LHE transformed video frame R'_t and D'_t are:

$$R'^{MSCN}_t[i,j] = \frac{R'_t[i,j] - \mu_{Rt}[i,j]}{\sigma_{Rt}[i,j] + C}, \tag{4.12}$$

and

$$D'^{MSCN}_t[i,j] = \frac{D'_t[i,j] - \mu_{Dt}[i,j]}{\sigma_{Dt}[i,j] + C}, \tag{4.13}$$

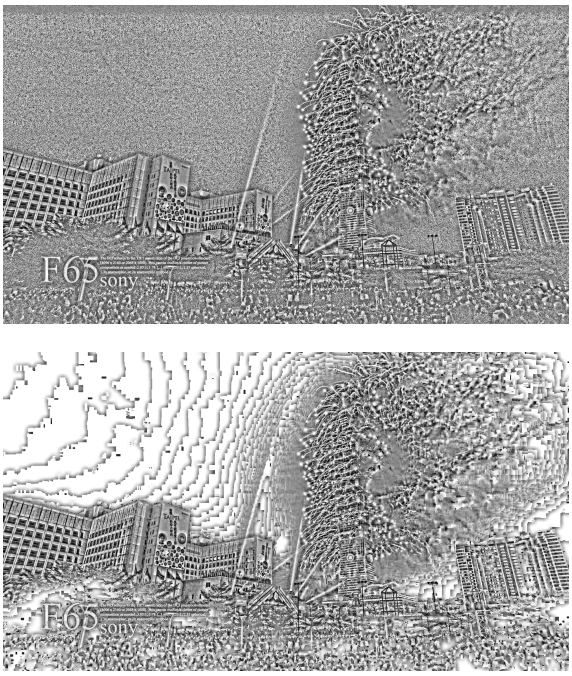


Figure 4.8: Application of local histogram equalization (LHE) to the video frames in Fig 4.6. Top: original; Bottom: compressed showing enhanced banding artifacts.

where $i \in \{1, 2.., M\}$, $j \in \{1, 2.., N\}$ are spatial indices, M and N are the frame height and width, respectively, the constant $C = 0.01$ imparts numerical stability, and where

$$\mu[i,j] = \sum_{k=-K}^{k=K} \sum_{l=-L}^{l=L} w[k,l] R'_t[i+k, j+l] \tag{4.14}$$

and

$$\begin{aligned} \sigma[i,j] = (\sum_{k=-K}^{k=K} \sum_{l=-L}^{l=L} w[k,l] (R'_t[i+k, j+l] \\ - \mu_{Rt}[i,j])^2)^{\frac{1}{2}} \end{aligned} \tag{4.15}$$

for the reference frame. The statistics μ_{Dt} and σ_{Dt} are computed in identical manner on the distorted frames. The weights $w = \{w[m,l] | m = -L, \ldots, L, l = -L, \ldots, L\}$ are a 2D circularly-symmetric Gaussian weighting function sampled out to 3 standard deviations and rescaled to unit volume, where $K = L = 3$. The same spatial entropies

$h(\tilde{R}_t'^{MSCN})$ and $h(\tilde{D}_t'^{MSCN})$ are calculated using (4.3) under the additive neural noise model. The scaling factors and modified entropies computed as

$$\begin{aligned} \gamma_{pt}^R = \log(1+\sigma^2(\tilde{D}_t'^{MSCN})), \quad \gamma_{pt}^D = \log(1+\sigma^2(\tilde{D}_t'^{MSCN})) \\ \epsilon_{pt}^R = \gamma_{pt}^R h(\tilde{D}_t'^{MSCN})), \quad \epsilon_{pt}^D = \gamma_{pt}^D h(\tilde{D}_t'^{MSCN}). \end{aligned} \tag{4.16}$$

Finally, define the LHE-GREED index as:

$$\text{LHE-GREED}_t = \frac{1}{P}\sum_{p=1}^{P} |\epsilon_{pt}^D - \epsilon_{pt}^R|. \tag{4.17}$$

4.4.3 PSNR features

PSNR was included as an additional component to provide a complementary perspective on the evaluation of video quality. Although PSNR is a simple and widely used metric, it can offer insights into differences between original and distorted videos on a per-pixel basis. By incorporating PSNR into the HDR-GREED algorithm, the model can benefit in cases where banding artifacts lead to significant pixel value differences. Combining PSNR with the more advanced features and human visual system-inspired components of HDR-GREED allows for a more comprehensive assessment of HDR video quality.

4.4.4 Implementation Details

For simplicity, we implemented our model only in the luminance domain. When calculating entropies we used spatial patches of size 5×5 (*i.e.* $\sqrt{M} = 5$). The neural noise variance was always fixed at $\sigma_W^2 = 0.1$ in (4.2), matching those employed in Soundararajan and Bovik (2011) and Bampis et al. (2017a). Similar to Bampis et al. (2017a); Madhusudana et al. (2021b), we found that out model is most effective when DoG-GREED and LHE-GREED are calculated over multiple spatial scales. Hence, we computed all features over 6 spatial scales, by bicubically downsampling the frames 2^s times, $s \in \{0, 1, 2, ..., 5\}$ Madhusudana et al. (2021b); Soundararajan and Bovik

Table 4.2: Descriptions of Features Used In HDR-GREED

Feature index	Description
$f_1 - f_{16}$	Original ST-GREED features
$f_{17} - f_{22}$	DoG-GREED features
$f_{23} - f_{28}$	LHE-GREED features
f_{29}	PSNR feature

(2011); Bampis et al. (2017a); Li et al. (2017). Finally, the entropy terms are pooled over all temporal indices by averaging them:

$$\bar{h} = \frac{1}{T} \sum_{t=1}^{T} h_t \tag{4.18}$$

where T is the number of frames for each video.

4.4.5 Regression

The DoG-GREED, LHE-GREED, and original ST-GREED features form a powerful set of quality aware features, which we used to train a support vector regressor (SVR) on the new database to map the features to human judgments of quality. These features are summarized in Table 4.2.

4.5 Objective VQA Experiments

To demonstrate the usefulness of the new LIVE AQ Video Quality Database, we conducted a series of experiments to evaluate the performances of several leading HDR VQA models, as well as SOTA standard dynamic range (SDR) VQA models. We also examined the effects of varying the parameters of the HDR-GREED model, which is defined by the feautres in Table 4.2. These experiments allowed us to assess the relative capabilities and limitations of these different models.

4.5.1 Evaluation Protocol

To evaluate the performances of the compared VQA algorithms, we employed an SVR model with a linear kernel to learn the mappings from features to difference mean opinion scores (DMOS). We conducted 1000 random train-test splits of the data, where in each, 80% of the data was used for training and 20% was used for testing, with no overlap between the training and testing subsets, nor of the original content the videos derived from. We applied 5-fold cross-validation to determine the optimal SVR parameters for each training set. This allowed us to robustly assess the compared algorithms and study their generalization capabilities. We used three standard performance metrics: the Spearman's Rank Order Correlation Coefficient (SROCC), the Pearson Linear Correlation Coefficient (PLCC), and the RMSE.

4.5.2 Selection of DOG-GREED Parameters

We first studied the performance of HDR-GREED against different choices of parameters of the DoG filter used to generate DoG-GREED features. From the previous section $w_{LOW} \approx \frac{N}{4\pi\sigma}$ and $w_{HIGH} \approx \frac{N}{3\sigma}$, where units of σ are in $pixels$ and frequencies are in $cycles/frame$. Let $N = 3840$ for 4K videos, we studied candidate DoG frequency falling in the range $2 - 20$ $cycles/frame$. The reason we selected this frequency range is empirical, yet it is unique and based on sound observation. Banding artifacts are usually repetitive, especially in critical, large sky regions. These quasi-periodic degradations are typically quite low-frequency and it is convenient to consider them in units of $cycles/frame$, where we use "frame" to mean the longer frame dimension. We have found that repetitive bands nearly always have reciprocal periods (fundamental frequencies) in the range of $2 - 20$ $cycles/frame$. The DoG bandpass parameters are determined by σ; we used grid search over w_{LOW} to find an optimal parameter for the DoG filter. The parameters sets that we experimented are shown in Table 4.3.

The results of using the parameter values in Table 4.3 are shown in Table 4.4.

Table 4.3: The Parameters Used in the DoG filter.

w_{LOW}	w_{HIGH}	σ
2	8.3773	152.7933
4	16.7546	76.3966
6	25.1320	50.9310
8	33.5093	38.1983
10	41.8866	30.5586
12	50.2640	25.4655
14	58.6413	21.8276
16	67.0186	19.0991
18	75.3960	16.9770
20	83.7733	15.2793

Plots that showing the performance against different choices of the parameters are provided in Fig. 4.9. As may be observed, the DoG-GREED features generally performed well, obtaining high correlations against human judgments, and significantly better than when the DoG-GREED features were removed ("None" in Table 4.4). From Fig. 4.9, the correlation first increases and then decreases as the cutoff frequency increases. This clear pattern provides valuable insights into the relationship between the cutoff frequency and the performance of the DoG-GREED features forfor predicting human judgments. The peak performance was obtained at $w_{LOW} = 6$ $cycles/frame$, corresponding to a $\sigma = 50.93$ and the bandpass filter spanning approximately $6 - 25.13$ $cycles/frame$.

Table 4.4: Correlations against Human Score Obtained by HDR-GREED as σ Is Varied in DoG-GREED. The Top Performing Parameter is Boldfaced.

W_low	SROCC	PLCC	RMSE
None	0.8572	0.8477	10.0364
2	0.8735	0.8699	9.4725
4	**0.8804**	0.8751	9.4199
6	0.8797	**0.8763**	**9.3201**
8	0.8782	0.8752	9.3795
10	0.8767	0.8709	9.4203
12	0.8725	0.8678	9.5515
14	0.8726	0.8669	9.4585
16	0.8709	0.8667	9.5849
18	0.8719	0.8660	9.5583
20	0.8729	0.8672	9.6743

4.5.3 Performance comparison and benchmark

We conducted a comparison of various leading FR VQA models designed for both HDR and SDR videos on the new LIVE AQ-HDR Database. We included PSNR, SSIM, MS-SSIM Wang et al. (2003), SpEED-QA Bampis et al. (2017a), and ST-RRED Soundararajan and Bovik (2011), VMAF Li et al. (2017), HDR-VDP 2.2 Mantiuk et al. (2011), and HDR-GREED. Since many FR VQA algorithms directly compute video quality predictions without using machine learning when mapping features to human opinion scores, we modified some of the leading FR algorithms to extract individual quality-aware features, then applied the same machine learning protocol as the other trained models to map them to predictions of perceptual quality, to ensure fair comparisons. Specifically, we decomposed SSIM into three

features: those representing luminance, contrast, and structural similarity. Similarly, we decomposed MS-SSIM into eleven features, comprising two SSIM features from each of four spatial scales, and three from the coarsest scale. For SpEED-QA, we extracted both the reduced-reference and single-number versions of the spatial and temporal SpEED-QA values, yielding a total of four features. The ST-RRED features were obtained from five levels of the steerable pyramid used in that algorithm, and the HDR-VDP features were obtained by pooling quality features over nine spatial scales. We then trained an SVR to map these features to human quality judgments on the new LIVE AQ-HDR database. We also conducted an ablation study with HDR-GREED, removing each feature set and evaluating the performance of the rest of the feature sets in the algorithm. The results of the comparison are shown in Table 4.5 and the ablation study is shown in Table 4.6. We also calculated the performance of the FR VQA algorithms on the AQ videos and non-AQ videos separately in the LIVE AQ-HDR Database. The results are shown in Table 4.7. It may be observed from the table that HDR-GREED outperformed on the non-AQ videos in the LIVE AQ-HDR Database. Although it doesn't obtain top performance on the AQ videos, it is competitive and improves on the performance by a great amount comparing to ST-GREED. Overall, HDR-GREED outperformed all of the other algorithms by healthy margins. This superior performance can be attributed to the predictive ability of the new features, which are able to capture HDR-relevant distortions, including when AQ is varied to attempt to ameliorate banding effects. It is noteworthy that both HDR-VDP and PSNR effectively capture distortions in banding videos, likely because of their pixel-based video differencing, which can accurately capture subtle differences present in banding videos.

We also evaluated the performances of the same VQA models on the recent LIVE-HDR database, which is a new database dedicated to the study of HDR perception but without AQ variations. This database consists of 310 HDR10 videos that were viewed by 66 subjects, where the distorted videos were created by applying compression and spatial downscaling using the x265 encoder. By testing the models

Table 4.5: Performances of the Compared HDR and SDR Quality Models When Evaluated on the LIVE AQ-HDR Database. The Top Performing Models Are Boldfaced.

Algorithm	SROCC	PLCC	RMSE
PSNR	0.7330	0.7176	12.5456
SSIM	0.5337	0.5491	14.6435
MS-SSIM	0.5706	0.5914	14.6072
SpEED-QA	0.7810	0.7781	11.6660
ST-RRED	0.6960	0.6886	13.2177
HDR-VDP2.2	0.8094	0.8073	11.0161
VMAF	0.8186	0.8224	10.8991
ST-GREED	0.8092	0.8186	10.8458
HDR-GREED	**0.8815**	**0.8763**	**9.3047**

Table 4.6: Ablation Study on the LIVE AQ-HDR Database.

Features removed	SROCC	PLCC	RMSE
LHE-GREED	0.8633	0.8615	9.8531
DoG-GREED	0.8572	0.8477	10.0364
ST-GREED	0.8737	0.8698	9.3622
PSNR	0.8439	0.8404	10.1536
None	0.8815	0.8763	9.3047

Table 4.7: Performances of the Compared HDR and SDR Quality Models When Evaluated on the LIVE AQ-HDR Database, Separated by AQ and Non-AQ Videos. The Top Performing Models Are Boldfaced.

	Non-AQ videos			AQ videos		
Algorithm	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
PSNR	0.7644	0.7613	12.2169	**0.8470**	**0.8524**	**8.3270**
SSIM	0.5122	0.5144	16.2522	0.7026	0.7391	12.3005
MS-SSIM	0.5995	0.6178	15.4960	0.6061	0.6057	13.1560
ST-RRED	0.6817	0.7000	13.6990	0.6522	0.5468	14.9204
SpEED-QA	0.7858	0.7917	11.9264	0.7148	0.6892	13.2230
HDR-VDP2.2	0.8379	0.8413	10.6262	0.8317	0.8020	10.4441
VMAF	0.8259	0.8391	10.8275	0.8043	0.8147	9.7939
ST-GREED	0.8377	0.8357	10.9461	0.6504	0.6661	13.1795
HDR-GREED	**0.8745**	**0.8819**	**9.4293**	0.8026	0.786	11.2271

Table 4.8: Performances of the Compared HDR and SDR Quality Models When Evaluated on the LIVE HDR Database. The Top Performing Model Is Boldfaced.

Algorithm	SROCC	PLCC	RMSE
PSNR	0.6242	0.6357	14.2095
SSIM	0.5208	0.4898	16.9252
MS-SSIM	0.6007	0.5810	15.5789
ST-RRED	0.6863	0.6569	13.2224
SpEED-QA	0.6110	0.6196	14.3763
HDR-VDP2.2	0.7041	0.6722	13.7727
VMAF	0.6753	0.6086	14.8758
ST-GREED	0.6456	0.6180	14.6678
HDR-GREED	**0.7398**	**0.7241**	**11.6438**

on this database, we sought to study the generalizability of the compared models. As shown in Table 4.8, HDR-GREED was able to deliver significant improvement and outperforance, on an independent dataset containing different applied distortion processes.

4.5.4 Evaluation on SDR Database

We also trained and evaluated HDR-GREED on the SDR-only LIVE ETRI Database Lee et al. (2021) to study the efficacy of the new models. The LIVE ETRI database consists of 437 videos that have undergone compression, spatial aliasing, and temporal subsampling. Human subjective scores are also provided for these videos. The videos in this database are standard dynamic range (SDR) and were encoded using 10 bits per pixel, with a BT 709 gamma curve and color gamut. As shown in Table 4.9, the HDR-GREED model achieved comparable performance to SOTA VQA

Table 4.9: Performances of the Compared Algorithms on the LIVE ETRI Database. The Top Performing Model is Boldfaced.

Algorithm	SROCC	PLCC	RMSE
PSNR	0.4941	0.4289	13.3712
SSIM	0.3568	0.3358	14.0594
MS-SSIM	0.5234	0.5319	13.2987
ST-RRED	0.7500	0.7587	10.9945
SpEED-QA	0.7461	0.7676	9.8110
VMAF	0.6439	0.6415	11.8306
ST-GREED	0.7245	0.7613	9.9702
HDR-GREED	**0.7772**	**0.7847**	**9.8033**

methods on this dataset. This suggests that the features employed by HDR-GREED are not restricted to HDR content and generalize well to other types of artifacts that may occur in SDR streaming scenarios, even in the presence of another different class of applied distortions.

Figure 4.9: Exploring the accuracy of HDR-GREED: the impact of w_{LOW} on correlation and RMSE against human scores.

Chapter 5: Discussion

This research journey embarked on addressing critical aspects in the field of Video Quality Assessment (VQA) by developing comprehensive databases and advanced models. In line with this objective, we established two large-scale video quality databases designed specifically for high-motion, live-streaming scenarios and HDR10 video format. The former includes 45 source sequences from 33 original contents with six different distortion types, while the latter contains 310 videos with subjective evaluations under two lighting conditions.

The most compelling aspect of these new databases is their accessibility to the public. By offering these resources to the wider research community, we facilitate testing, comparison, and development of both No-Reference (NR) and Full-Reference (FR) VQA models. This represents a significant leap forward in making high-quality datasets available for the development and improvement of VQA models.

The unique focus of our study on the HDR10 format, a commonly used standard in contemporary video technology, addresses a critical gap in the field. While our current efforts are centered around HDR10, this research has the potential to stimulate further work in other HDR formats such as HDR10+, Dolby Vision, and Hybrid Log-Gamma (HLG), opening up new avenues for future investigation.

Our work significantly advances the HDR VQA discipline by leveraging the HDR10 standard, incorporating a diverse selection of both Video on Demand (VoD) and live videos from actual streaming sources. This strategy results in a dataset that is much more representative of modern HDR content than previous attempts. Furthermore, the introduction of an evaluation database encoded with and without Adaptive Quantization (AQ) options allows for an in-depth analysis of AQ's impact on perceived HDR video quality, an aspect often overlooked in previous research.

In the course of this research, we designed and implemented two distinct yet com-

plementary models: a framework for defining HDR quality-aware features and a novel HDR VQA model called HDR-GREED. The latter integrates Laplacian of Gaussian (LoG) and Difference of Gaussian (DoG) filters to enhance the model's sensitivity to spatial distortions such as banding and blocking. These distortions are particularly relevant in HDR context, further improving the practicality and effectiveness of HDR-GREED.

Despite the enhanced performance and the additional feature extraction stages in our new models, we ensured that computational complexity remains minimal relative to conventional models such as VMAF. Notably, these feature extraction stages can occur in parallel on the transformed and original frames, suggesting that parallel computation can substantially accelerate feature extraction, thereby presenting a significant improvement in the computational efficiency of VQA models.

Finally, the HDR-GREED model exhibited impressive performance, surpassing state-of-the-art (SOTA) Full-Reference models on the new LIVE AQ-HDR Database. It also demonstrated its robustness and effectiveness across previous HDR and SDR databases, underscoring its potential as a versatile tool for evaluating video quality across diverse formats and applications. This achievement symbolizes a culmination of our efforts, and we look forward to seeing how our contributions will propel the field of HDR VQA further into the future.

Works Cited

The Consumer Digital Video Library. https://www.cdvl.org/. URL https://www.cdvl.org/.

Cisco global cloud index forecast and methodology 2015-2020. URL https://www.cisco.com/c/dam/m/en_us/service-provider/ciscoknowledgenetwork/files/622_11_15-16-Cisco_GCI_CKN_2015-2020_AMER_EMEAR_NOV2016.pdf.

65" class Q90T QLED 4K UHD HDR smart TV 2020 TVs - QN65Q90TAFXZA: Samsung US. URL https://www.samsung.com/us/televisions-home-theater/tvs/qled-4k-tvs/65-class-q90t-qled-4k-uhd-hdr-smart-tv-2020-qn65q90tafxza.

Canon xl2 manual. URL https://www.usa.canon.com/internet/portal/us/home/support/details/camcorders/support-professional-camcorders/xl2/xl2?tab=manuals.

Laboratory for image and video engineering. URL https://live.ece.utexas.edu/research/Quality/visualScreening.htm.

Laboratory for image amp; video engineering. URL https://live.ece.utexas.edu/research/Quality/visualScreening.htm.

Panasonic dvx-100b manual. URL https://tfma.temple.edu/sites/tfma/files/site-pdfs/DVX100aManual.pdf.

2018 global internet phenomena report", 2018. URL https://www.sandvine.com/phenomena.

2020 Bitmovin video developer report, Sep 2020. URL https://f.hubspotusercontent30.net/hubfs/3411032/Bitmovin%20Developer%20Report%202020-21/bitmovin-developer-rep
pdf.

Free Ultra-HD / HDR / HLG / Dolby Vision 4K video demos. https://4kmedia.org/, 2020. URL https://4kmedia.org/.

Mekides Assefa Abebe, Tania Pouli, and Jonathan Kervec. Evaluating the color fidelity of itmos and hdr color appearance models. *ACM Trans. Appl. Percept.*, 12(4), sep 2015. ISSN 1544-3558. doi: 10.1145/2808232. URL https://doi.org/10.1145/2808232.

S. Athar, T. Costa, K. Zeng, and Z. Wang. Perceptual quality assessment of UHD-HDR-WCG videos. In *IEEE Int. Conf. Image Process.*, pages 1740–1744, 2019.

Tunç O. Aydın, Rafal Mantiuk, and Hans-Peter Seidel. Extending quality metrics to full luminance range images. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging XIII*, volume 6806, pages 109 – 118. International Society for Optics and Photonics, SPIE, 2008. doi: 10.1117/12.765095. URL https://doi.org/10.1117/12.765095.

Ma. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M .T. Pourazad, and P. Nasiopoulos. Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content. *arXiv preprint arXiv:1803.04815*, 2018.

Christos G. Bampis, Praful Gupta, Rajiv Soundararajan, and Alan C. Bovik. Speed-qa: Spatial efficient entropic differencing for image and video quality. *IEEE Signal Processing Letters*, 24(9):1333–1337, 2017a. doi: 10.1109/LSP.2017.2726542.

Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, 26(11):5217–5231, 2017b.

V. Baroncini, K. Andersson, A.K. Ramasubramonian, and G. Sullivan. Verification test report for HDR/WCG video coding using HEVC main 10 profile. In *Proc. JCTVC-X1018 24th JCT-VC Meeting*, 2016.

Marcelo Bertalmío. Chapter 5 - brightness perception and encoding curves. In Marcelo Bertalmío, editor, *Vision Models for High Dynamic Range and Wide Colour Gamut Imaging*, Computer Vision and Pattern Recognition, pages 95–129. Academic Press, 2020. ISBN 978-0-12-813894-6. doi: https://doi.org/10.1016/B978-0-12-813894-6.00010-7. URL `https://www.sciencedirect.com/science/article/pii/B9780128138946000107`.

Maxime Bichon, Julien Le Tanou, Michael Ropert, Wassim Hamidouche, and Luce Morin. Optimal adaptive quantization based on temporal distortion propagation model for hevc. *IEEE Transactions on Image Processing*, 28(11): 5419–5434, 2019. doi: 10.1109/TIP.2019.2919180.

Vincent A Billock and Brian H Tsou. To honor fechner and obey stevens: relationships between psychophysical and neural nonlinearities. *Psychological bulletin*, 137(1):1, 2011.

G. Bradski. The OpenCV Library. *Dr. Dobb's J. Software Tools*, 2000.

Matteo Carandini, David J Heeger, and J Anthony Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21):8621–8644, 1997.

Chao Chen, Lark Kwon Choi, Gustavo De Veciana, Constantine Caramanis, Robert W Heath, and Alan C Bovik. Modeling the time—varying subjective quality of http video streams with rate adaptations. *IEEE Trans. Image Process.*, 23(5):2206–2221, 2014.

Chia-Chen Chen and Yi-Chen Lin. What drives live-stream usage intention? the perspectives of flow, entertainment, social interaction, and endorsement. *Telemat. Inform.*, 35(1):293–303, 2018.

L.H. Chen, C. G. Bampis, Z. Li, J. Sole, and A. C. Bovik. Perceptual video quality prediction emphasizing chroma distortions. *IEEE Trans. Image Process.*, 30:1408–1422, 2021. doi: 10.1109/TIP.2020.3043127.

Manri Cheon and Jong-Seok Lee. Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience. *IEEE Trans. Circuits Syst. Video Technol*, 28(7):1467–1480, 2017.

Anustup Choudhury, Robert Wanat, Jaclyn Pytlarz, and Scott Daly. Image quality evaluation for high dynamic range and wide color gamut applications using visual spatial processing of color differences. *Color Research & Application*, 46(1):46–64, 2021. doi: https://doi.org/10.1002/col.22588. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/col.22588`.

Tom N Cornsweet and HM Pinsker. Luminance discrimination of brief flashes under various conditions of adaptation. *The Journal of Physiology*, 176(2):294, 1965.

CTA. Television technology consumer definitions. https://cdn.cta.tech/cta/media/media/memb
technology-consumer-definitions.pdf.

Yuqi Dai, Changbin Xue, and Li Zhou. Visual saliency guided perceptual adaptive quantization based on hevc intra-coding for planetary images. *Plos one*, 17(2):e0263729, 2022.

Scott Daly, Ning Xu, James Crenshaw, and Vikrant J Zunjarrao. A psychophysical study exploring judder using fundamental signals and complex imagery. *SMPTE Motion Imaging J.*, 124(7):62–70, 2015.

Francesca De Simone, Marco Tagliasacchi, Matteo Naccari, Stefano Tubaro, and Touradj Ebrahimi. A H. 264/AVC video database for the evaluation of quality metrics. In *IEEE ICASSP*, pages 2430–2433, 2010.

Francesca De Simone, Matteo Naccari, Marco Tagliasacchi, Frederic Dufaux, Stefano Tubaro, and Touradj Ebrahimi. Subjective quality assessment of H. 264/AVC video streaming with packet losses. *EURASIP J. Image Video Process*, 2011(1):1–12, 2011.

Zhengfang Duanmu, Kai Zeng, Kede Ma, Abdul Rehman, and Zhou Wang. A quality-of-experience index for streaming video. *IEEE J. of Sel. Topics Signal Processing*, 11(1):154–166, 2017. doi: 10.1109/JSTSP.2016.2608329.

Joshua P Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, and Alan C Bovik. No-reference video quality assessment using space-time chips. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2020a.

Joshua P. Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, and Alan C. Bovik. No-reference video quality assessment using space-time chips, 2020b. URL `https://arxiv.org/abs/2008.00031`.

Joshua P. Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, and Alan C. Bovik. Making video quality assessment models robust to bit depth. In *SPL*, pages 556–564, January 2023.

Joshua Peter Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. Chipqa: No-reference video quality prediction via space-time chips. *IEEE Trans. Image Process.*, 30:8059–8074, 2021. doi: 10.1109/TIP.2021.3112055.

Mark D. Fairchild and Ping-Hsu Chen. Brightness, lightness, and specifying color in high-dynamic-range scenes and images. In Susan P. Farnand and Frans

Gaykema, editors, *Image Quality and System Performance VIII*, volume 7867, pages 233 – 246. International Society for Optics and Photonics, SPIE, 2011. doi: 10.1117/12.872075. URL https://doi.org/10.1117/12.872075.

D. Hasler and S.E. Suesstrunk. Measuring colorfulness in natural images. In *Human Vis. Electr. Imaging VIII*, volume 5007, pages 87–95. Intl. Soc. Opt. Photon., 2003.

Jing He, En-Hui Yang, Fuzheng Yang, and Kehu Yang. Adaptive quantization parameter selection for h.265/hevc by employing inter-frame dependency. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(12):3424–3436, 2018. doi: 10.1109/TCSVT.2017.2751519.

V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe. The Konstanz natural video database (KoNViD-1k). In *Int. Conf. Qual. Multim. Exp. (QoMEX)*, pages 1–6. IEEE, 2017.

R Hummel. Image enhancement by histogram transformation. comp. graph. 1977.

ITU. BT.500 : Methodologies for the subjective assessment of the quality of television images. Technical report.

ITU. ITU. 910 : Subjective video quality assessment methods for multimedia applications. Technical report, Intl. Telecomm. Union, 2008.

ITU. BT.1886 : Reference electro-optical transfer function for flat panel displays used in HDTV studio production. Technical report, Intl. Telecomm. Union, 2011.

ITU. BT.709 : Parameter values for the hdtv standards for production and international programme exchange. Technical report, Intl. Telecomm. Union, 2011.

ITU. Methodology for the subjective assessment of the quality of television pictures ITU-R recommendation BT. 500-13. Technical report, 2012.

ITU. BT.2020 : Parameter values for ultra-high definition television systems for production and international programme exchange, 2015. URL `https://www.itu.int/rec/R-REC-BT.2020`.

ITU. BT.2100 : Image parameter values for high dynamic range television for use in production and international programme exchange. Technical report, Intl. Telecomm. Union, 2018.

Stephen M Keating. Image signal process. with digital filtering to minimize aliasing caused by image manipulation, Apr. 1993.

Christian Keimel, Arne Redl, and Klaus Diepold. The TUM high definition video datasets. In *QoMEX 2012*, pages 97–102.

Christian Keimel, Julian Habigt, Tim Habigt, Martin Rothbucher, and Klaus Diepold. Visual quality of current coding technologies at high definition IPTV bitrates. In *IEEE MMSP*, pages 390–393, 2010.

Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *the 15th ECCV*, pages 219–234, 2018.

Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.*, 28(12):5923–5938, 2019.

Lukáš Krasula, Anustup Choudhury, Scott Daly, Zhi Li, Robin Atkins, Ludovic Malfait, and Aditya Mavlankar. Subjective video quality for 4k hdr-wcg content using a browser-based approach for" at-home" testing. *Electronic Imaging*, 35: 263–1, 2023.

Debarati Kundu, Deepti Ghadiyaram, Alan C Bovik, and Brian L Evans. No-reference quality assessment of tone-mapped hdr pictures. *IEEE Trans. Image Process.*, 26(6):2957–2971, 2017.

T. Kunkel, S. Daly, S. Miller, and J. Froehlich. Perceptual design for high dynamic range systems. In *High Dynamic Range Video*, pages 391–430. Elsevier, 2016.

Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imaging*, 19(1):011006, 2010.

Patrick Ledda, Luis Paulo Santos, and Alan Chalmers. A local model of eye adaptation for high dynamic range images. In *Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, pages 151–160, 2004.

Dae Yeol Lee, Somdyuti Paul, Christos G Bampis, Hyunsuk Ko, Jongho Kim, Se Yoon Jeong, Blake Homan, and Alan C Bovik. A subjective and objective study of space-time subsampled video quality. *arXiv preprint arXiv:2102.00088*, 2021.

Jing Li, Lukáš Krasula, Yoann Baveye, Zhi Li, and Patrick Le Callet. Accann: A new subjective assessment methodology for measuring acceptability and annoyance of quality of experience. *IEEE Transactions on Multimedia*, 21(10): 2589–2602, 2019. doi: 10.1109/TMM.2019.2903722.

Z. Li, C.G. Bampis, L. Janowski, and I. Katsavounidis. A simple model for subject behavior in subjective experiments. *Electron. Imag.*, 2020(11):131–1, 2020.

Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *Netflix Tech Blog*, 6:2, 2016.

Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric, Apr 2017. URL https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9

Joe Yuchieh Lin, Rui Song, Chi-Hao Wu, TsungJung Liu, Haiqiang Wang, and C-C Jay Kuo. MCL-V: A streaming video quality assessment database. *J. Vis. Commun. Image Represent.*, 30:1–9, 2015.

P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C Bovik. Subjective and objective quality assessment of high frame rate videos. *IEEE Access*, 9:108069–108082, 2021a.

Pavan C. Madhusudana, Xiangxu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Subjective and objective quality assessment of high frame rate videos, 2020. URL https://arxiv.org/abs/2007.11634.

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Transactions on Image Processing*, 30:7446–7457, 2021b. doi: 10.1109/TIP.2021.3106801.

Rafal Mantiuk, Scott J. Daly, Karol Myszkowski, and Hans-Peter Seidel. Predicting visible differences in high dynamic range images: model and its calibration. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly, editors, *Human Vision and Electronic Imaging X*, volume 5666, pages 204 – 214. International Society for Optics and Photonics, SPIE, 2005. doi: 10.1117/12.586757. URL https://doi.org/10.1117/12.586757.

Rafal Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4), jul 2011. ISSN 0730-0301. doi: 10.1145/2010324.1964935. URL https://doi.org/10.1145/2010324.1964935.

Rafal K. Mantiuk, Dounia Hammou, and Param Hanji. Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content, 2023.

Rafa K. Mantiuk and Maryam Azimi. Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In *2021 Picture Coding Symposium (PCS)*, pages 1–5, 2021. doi: 10.1109/PCS50896.2021.9477471.

Alexandre Mercat, Marko Viitanen, and Jarno Vanne. Uvg dataset: 50/120fps 4K sequences for video codec analysis and development. In *ACM MMSys 2020*, pages 297–302.

Felix Mercer Moss, Ke Wang, Fan Zhang, Roland Baddeley, and David R. Bull. On the optimal presentation duration for subjective video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(11):1977–1987, 2016. doi: 10.1109/TCSVT.2015.2461971.

Felix Mercer Moss, Chun-Ting Yeh, Fan Zhang, Roland Baddeley, and David R. Bull. Support for reduced presentation durations in subjective video quality assessment. *Signal Processing: Image Communication*, 48:38–49, 2016. ISSN 0923-5965. doi: https://doi.org/10.1016/j.image.2016.08.005. URL `https://www.sciencedirect.com/science/article/pii/S0923596516301126`.

Aliaksei Mikhailiuk, María Pérez-Ortiz, Dingcheng Yue, Wilson Suen, and Rafał K. Mantiuk. Consolidated dataset and metrics for high-dynamic-range image quality. *IEEE Transactions on Multimedia*, 24:2125–2138, 2022. doi: 10.1109/TMM.2021.3076298.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. image Process.*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett*, 20(3):209–212, 2012b.

Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Trans. Image Process.*, 25(1):289–300, 2015.

Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Trans. Image Process.*, 20(12):3350–3364, 2011.

Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE J. Sel. Topics Signal Process.*, 6(6):652–671, 2012a.

Anush Krishna Moorthy, Lark Kwon Choi, G de Veciana, and AC Bovik. Mobile video quality assessment database. In *IEEE ICC Workshop Realizing Advanced Video Optimized Wireless Netw.*, pages 7055–7059, 2012b.

Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet, and Romuald Pepion. On improving the pooling in HDR-VDP-2 towards better HDR perceptual quality assessment. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder, editors, *Human Vision and Electronic Imaging XIX*, volume 9014, pages 143 – 151. International Society for Optics and Photonics, SPIE, 2014. doi: 10.1117/12.2045436. URL `https://doi.org/10.1117/12.2045436`.

Manish Narwaria, Rafal Mantiuk, Mattheiu P. Da Silva, and Patrick Le Callet. HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 24(1):1 – 3, 2015a. doi: 10.1117/1.JEI.24.1.010501. URL `https://doi.org/10.1117/1.JEI.24.1.010501`.

Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46–60, 2015b. ISSN 0923-5965. doi: https://doi.org/10.1016/j.image.2015.04.009. URL `https://www.sciencedirect.com/science/article/pii/S0923596515000703`.

Pengpeng Ni, Ragnhild Eg, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen. Flicker effects in adaptive video streaming to handheld devices. In *19th ACM Int. Conf. Multimed.*, pages 463–472, 2011.

Se Ri Oh, Dongchan Kim, Pyeong Gang Heo, and HyunWook Park. A new metric for judder in high frame-rate video. In *ICIP 2016*, pages 3802–3806. IEEE.

Yen-Fu Ou, Yan Zhou, and Yao Wang. Perceptual quality of video with frame rate variation: A subjective study. In *IEEE ICASSP*, pages 2446–2449, 2010.

Yen-Fu Ou, Yuanyi Xue, and Yao Wang. Q-star: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions. *IEEE Trans. Image Process.*, 23(6):2473–2486, 2014.

X. Pan, J. Zhang, S. Wang, S. Wang, Y. Zhou, W. Ding, and Y. Yang. HDR video quality assessment: Perceptual evaluation of compressed HDR video. *J. Vis. Comm. Image Rep.*, 57:76–83, 2018.

Pradip Paudyal, Federica Battisti, and Marco Carli. Reduced reference quality assessment of light field images. *IEEE Transactions on Broadcasting*, 65(1): 152–165, 2019. doi: 10.1109/TBC.2019.2892092.

Somdyuti Paul, Andrey Norkin, and Alan C. Bovik. On visual masking estimation for adaptive quantization using steerable filters. *Signal Processing: Image Communication*, 96:116290, 2021. ISSN 0923-5965. doi: https:

//doi.org/10.1016/j.image.2021.116290. URL https://www.sciencedirect.com/science/article/pii/S0923596521001235.

Ana Radonjić, Sarah R Allred, Alan L Gilchrist, and David H Brainard. The dynamic range of human lightness perception. *Current Biology*, 21(22):1931–1936, 2011.

R Rao, B Olshausen, M Lewicki, Martin J Wainwright, Odelia Schwartz, and Eero P Simoncelli. Natural image statistics and divisive normalization: modeling nonlinearities and adaptation in cortical neurons. 2001.

M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi. Subjective and objective evaluation of HDR video compression. In *9th Intl. Workshop Video Process. Qual. Metrics Consum. Electron. (VPQM)*, 2015.

Maxime Rousselot, Olivier Le Meur, Rémi Cozot, and Xavier Ducloux. Quality assessment of hdr/wcg images using hdr uniform color spaces. *Journal of Imaging*, 5(1), 2019. ISSN 2313-433X. doi: 10.3390/jimaging5010018. URL https://www.mdpi.com/2313-433X/5/1/18.

Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process.*, 21(8):3339–3352, 2012.

Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. image Process.*, 19(2):335–350, 2009.

Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE Trans. Image Process.*, 19(6):1427–1441, 2010.

Z. Shang, J. P. Ebenezer, A. C. Bovik, Y. Wu, H. Wei, and S. Sethuraman. Assessment of subjective and objective quality of live streaming sports videos. In *Picture Coding Symposium (PCS)*, pages 1–5, 2021a. doi: 10.1109/PCS50896.2021.9477502.

Zaixi Shang, Joshua Peter Ebenezer, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. A study of subjective and objective quality assessment of hdr videos. *Multimedia Tools and Applications*, 77(12):14817–14840, 2018.

Zaixi Shang, Joshua P. Ebenezer, Alan C. Bovik, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Assessment of subjective and objective quality of live streaming sports videos. In *2021 Picture Coding Symposium (PCS)*, pages 1–5, 2021b. doi: 10.1109/PCS50896.2021.9477502.

Zaixi Shang, Joshua Peter Ebenezer, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C Bovik. Study of the subjective and objective quality of high motion live streaming videos. *IEEE Transactions on Image Processing*, 31: 1027–1041, 2021c.

Zaixi Shang, Joshua Peter Ebenezer, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. Study of the subjective and objective quality of high motion live streaming videos. *IEEE Transactions on Image Processing*, 31: 1027–1041, 2022. doi: 10.1109/TIP.2021.3136723.

Zaixi Shang, Yixu Chen, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Subjective and objective video quality assessment of high dynamic range sports content. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 556–564, January 2023.

Hamid R Sheikh and Alan C Bovik. A visual information fidelity approach to video quality assessment. In *Int. Workshop Video Process. Quality Metrics for Consumer Electron.*, volume 7, page 2. sn, 2005.

Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3440–3451, 2006.

H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. doi: 10.1109/TIP.2005.859378.

Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Trans. Image Process.*, 28(2):612–627, 2018.

L. Song, Y. Liu, X. Yang, G. Zhai, R. Xie, and W. Zhang. The SJTU HDR video sequence dataset. In *Proc. Int. Conf. Qual. Multim. Exp. (QoMEX)*, page 100, 2016.

Li Song, Xun Tang, Wei Zhang, Xiaokang Yang, and Pingjian Xia. The SJTU 4K video sequence dataset. In *5th QoMEX*, pages 34–35. IEEE, 2013.

Rajiv Soundararajan and Alan C Bovik. RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Trans. Image Process.*, 21(2):517–526, 2011.

SMPTE Standard. High dynamic range electro-optical transfer function of mastering reference displays. *SMPTE ST*, 2084(2014):11, 2014.

Yasuko Sugito, Javier Vazquez-Corral, Trevor Canham, and Marcelo Bertalmío. Image quality evaluation in professional hdr/wcg production questions the need for hdr metrics. *IEEE Transactions on Image Processing*, 31:5163–5177, 2022. doi: 10.1109/TIP.2022.3190706.

EBU TECHNICAL. Eotf chart for calibration and monitoring. Technical Report Tech. 3374, European Broadcasting Union, Dec 2020. URL `https://tech.ebu.ch/publications/tech3374`.

Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE Trans. Image Process.*, 30:4449–4464, 2021a.

Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Trans. Image Process.*, 30:4449–4464, 2021b. doi: 10.1109/TIP.2021.3072221.

Maria Torres Vega, Vittorio Sguazzo, Decebal Constantin Mocanu, and Antonio Liotta. An experimental survey of no-reference video quality assessment methods. *Int. J. Pervasive Comput. Commun.*, pages 66–86, 2016.

Maria Torres Vega, Decebal Constantin Mocanu, Jeroen Famaey, Stavros Stavrou, and Antonio Liotta. Deep learning for quality assessment in live video streaming. *IEEE Signal Process. Lett*, 24(6):736–740, 2017.

Video Quality Experts Group. *"Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment"*, 2000 (accessed October 31,2020). URL `http://www.its.bldrdoc.gov/vqeg/projects/frtvphaseI`.

VideoLAN. Vlc media player. URL `https://www.videolan.org/vlc/`.

Mario Vranješ, Snježana Rimac-Drlje, and Krešimir Grgić. Review of objective video quality metrics and performance comparison using different databases. *Signal Process. Image Commun.*, 28(1):1–19, 2013.

Phong V Vu and Damon M Chandler. Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *J. Electron. Imaging*, 23(1):013016, 2014.

Phong V Vu, Cuong T Vu, and Damon M Chandler. A spatiotemporal most-apparent-distortion model for video quality assessment. In *IEEE ICIP*, pages 2505–2508, 2011.

Tien Huu Vu, Huy Phi Cong, Thipphaphone Sisouvong, Xiem HoangVan, Sang NguyenQuang, and Minh DoNgoc. Vmaf based quantization parameter prediction model for low resolution video coding. In *2022 International Conference on Advanced Technologies for Communications (ATC)*, pages 364–368, 2022. doi: 10.1109/ATC55345.2022.9942982.

Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al. Videoset: A large-scale compressed video quality dataset based on JND measurement. *J. Vis. Commun. Image Represent.*, 46:292–302, 2017.

Y. Wang, Sasi I., and Balu A. YouTube UGC dataset for video compression research. In *IEEE Int. Workshop Multim. Signal Process. (MMSP)*, pages 1–5, 2019.

Zhou Wang and Eero P Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human Vision and Electronic Imaging X*, volume 5666, pages 149–159, 2005.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The 37th ACSSC*, volume 2, pages 1398–1402, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

Zhou Wang, Guixing Wu, Hamid R Sheikh, Eero P Simoncelli, En-Hui Yang, and Alan C Bovik. Quality-aware images. *IEEE Trans. image Process.*, 15 (6):1680–1689, 2006.

Zhou Wang, Hojatollah Yeganeh, Kai Zeng, and Jiheng Wang. Diagnosing visual quality impairments in high dynamic-range/wide-color-gamut videos. *Journal of Digital Video*, 5:74–83, 2020.

Jinjian Wu, Weisi Lin, Guangming Shi, Yazhong Zhang, Weisheng Dong, and Zhibo Chen. Visual orientation selectivity based structure description. *IEEE Trans. Image Process.*, 24(11):4602–4613, 2015.

Guoqing Xiang, Huizhu Jia, Mingyuan Yang, Yuan Li, and Xiaodong Xie. A novel adaptive quantization method for video coding. *Multimedia Tools and Applications*, 77(12):14817–14840, 2018.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Trans. Image Process.*, 23(2):684–695, 2013.

Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *CVPR 2012*, pages 1098–1105. IEEE.

Web Yodel. *The consumer digital video library*, 2011. URL `https://cdvl.org/`.

Fan Zhang, Felix Mercer Moss, Roland Baddeley, and David R. Bull. Bvi-hd: A video quality database for hevc compressed and texture synthesized content. *IEEE Trans. Multimed.*, 20(10):2620–2630, 2018. doi: 10.1109/TMM.2018.2817070.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20 (8):2378–2386, 2011.

Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10):4270–4281, 2014.

Vita

Zaixi Shang received the B.S. degree in biomedical engineering from Shanghai Jiao Tong University in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of Texas at Austin. He joined the Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin, in 2019. He was a recipient of the Cockrell Graduate Engineering Fellowship from The University of Texas at Austin from 2018 to 2022.

Address: zxshang@utexas.edu

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.