

1

Quality Assessment in Media and Entertainment: Challenges and Trends

Abhinav K. Venkataramanan, Zaixi Shang, Joshua P. Ebenezer,
Meixu Chen, Zhengzhong Tu and Alan C. Bovik

CONTENTS

1.1	Standard Streaming Videos	4
1.1.1	Overview of Video Compression	5
1.1.2	Scaling and Rate-distortion Optimization	6
1.1.3	Subjective Quality Databases	7
1.1.4	Objective Quality Assessment	8
1.1.4.1	Full-Reference Quality Assessment	8
1.1.4.2	No-Reference Quality Models	10
1.1.4.3	Bitstream-based Quality Models	10
1.1.5	Challenges and Trends	10
1.2	User-Generated Content (UGC) Videos	11
1.2.1	Introduction to UGC Video Quality	12
1.2.2	Subjective UGC Video Quality Databases	13
1.2.3	Objective UGC Video Quality Assessment	15
1.2.4	Challenges and Future Trends	16
1.3	High Dynamic Range Video	16
1.3.1	Challenges Facing Adoption of HDR	17
1.3.2	Publicly Available Subjective Quality Databases and Studies	18
1.3.3	Video Quality Algorithms	19
1.3.3.1	HDR-VDP	19
1.3.3.2	PU-SSIM and PU-PSNR	19
1.3.3.3	HDR-VQM	20
1.3.3.4	HDRMAX and HDR-ChipQA	20
1.4	High Frame Rate Video	20
1.4.1	Challenges Facing Adoption of HFR	21
1.4.2	Subjective Quality Databases	21
1.4.2.1	LIVE YT HFR	21
1.4.2.2	ETRI LIVE STSVQ	22
1.4.2.3	AVT-VQDB-UHD-1	22

4 *Quality Assessment in Media and Entertainment: Challenges and Trends*

1.4.2.4	BVI HFR	23
1.4.3	Video Quality Algorithms	23
1.4.3.1	STGREED	23
1.4.3.2	VSTR	24
1.4.3.3	FAVER	24
1.5	Virtual Reality	24
1.5.1	Challenges Facing Adoption of VR	25
1.5.2	Subjective Quality Databases	25
1.5.2.1	LIVE 3D VR IQA Database	25
1.5.2.2	IVQAD 2017 Database	26
1.5.2.3	VRQ-TJU Database	26
1.5.2.4	LIVE-FBT-FCVR Database	26
1.5.3	VR Quality Algorithms	27
1.6	Conclusion	28

Assessing and optimizing the perceptual quality of media plays an important role in enhancing the subjective experiences of hundreds of millions of television, streaming, and social media viewers. Quality assessment methods may be broadly categorized into “subjective” and “objective” methods. Subjective quality assessment consists of conducting studies where participants manually rate the quality of media. On the other hand, objective quality models consist of algorithms designed to accurately predict subjective quality. In this chapter, we will discuss the challenges in assessing the quality of popular visual media, including streaming video, virtual reality, high-dynamic-range, high-framerate, and user-generated video. In particular, we will discuss subjective databases and state-of-the-art (SOTA) quality models that have been developed to address the problem of quality assessment. Finally, we will discuss future trends and open problems that are interesting to tackle.

1.1 Standard Streaming Videos

Video-On-Demand (VOD) and streaming videos comprise the most ubiquitous video modality consumed by users on the Internet. VOD includes a broad range of platforms such as Netflix, Amazon Prime Video, Hulu, Disney+, HBO Max, etc. In this chapter, we will use VOD to refer to professionally generated content, which is typically produced with high-quality equipment that deliver pristine source contents. User-generated content will be discussed separately in Section 1.2 and is considered a separate video modality in this chapter, even when it is “streamed” on the Internet.

In order to stream videos under the bandwidth constraints of the Internet, lossy compression is used to significantly reduce file sizes. However, since com-

pression is lossy, compressed videos are “distorted” versions of their respective source contents. Common distortions associated with video compression are blocking artifacts, temporal incoherency, and color bleeding.

In this section, we will first review video compression and popular SOTA codecs. Then, we will discuss rate-distortion optimization (RDO), which refers to the common paradigm used to “optimally” encode videos. Then, we will provide an overview of commonly used subjective quality databases that study the impact of compression on subjective quality. Finally, we will discuss SOTA quality assessment algorithms for compressed videos and outline interesting future work.

1.1.1 Overview of Video Compression

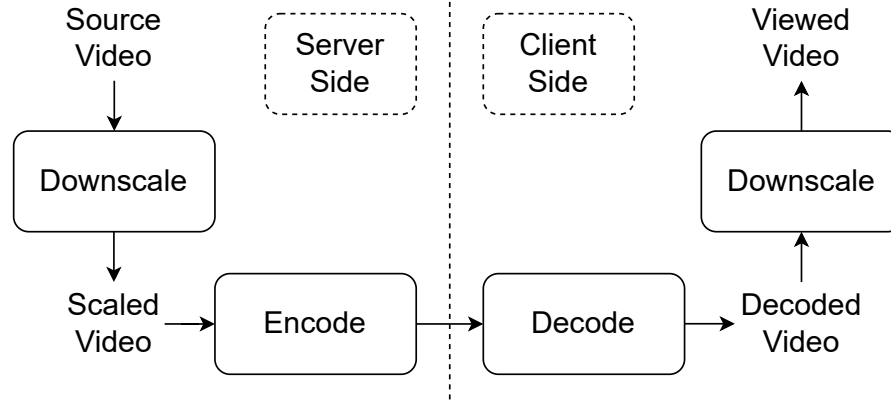
The widespread use of streaming and social media platforms has led to an explosion in the amount of video content that is being generated and consumed on the Internet. In order to reliably store and deliver such large volumes of bandwidth-hungry video content, lossy compression is a necessity. In general, the term “lossy compression” refers to “encoding” videos in order to reduce their file sizes and/or bandwidth requirements, without guaranteeing perfect reconstruction at the decoder.

Modern video compression algorithms may be broadly summarized as comprising the following sequence of operations, whose order may vary.

1. Converting pixel values to an entropy-reducing or sparsifying transform domain, such as the Discrete Cosine Transform [1] or the Integer Transform [2].
2. Coding “intra” frames by computing residuals between neighboring blocks from the same frame.
3. Coding “inter” frames using block motion estimation and compensation to compute motion-compensated residuals.
4. Quantization of transformed residuals, which is the source of “loss” that may be tuned based on user preference.
5. Entropy coding of transformed residuals, motion vectors, and other data using algorithms such as Context Adaptive Variable Length Coding (CAVLC) [3] and Context Adaptive Binary Arithmetic Coding (CABAC) [4] to yield the compressed bitstream.

As described above, the amount of quantization determines the amount of “loss” in the bitstream, thereby affecting the file size/compression ratio/bitrate. The amount of compression (and so, quantization) is typically tuned by selecting values of the “quantization parameter” (QP), “constant rate factor” (CRF), or the “bitrate.”

Popular video compression algorithms include H.264/Advanced Video Coding (AVC) [3], H.265/High Efficiency Video Coding (HEVC) [5], and the

**FIGURE 1.1**

Combining scaling and encoding for optimal compression

recent H.266/Versatile Video Coding (VVC) [6], which have all been developed by the Moving Picture Experts Group (MPEG). In addition to such proprietary algorithms, the popularity of open-source compression algorithms has grown in recent years with the development of algorithms such as VP9 [7] by Google and AOMedia Video 1 (AV1) [8] by the Alliance for Open Media (AOM).

1.1.2 Scaling and Rate-distortion Optimization

As mentioned in Section 1.1.1, compression algorithms are equipped with a user-selected parameter (such as CRF, QP, or directly setting the bitrate) to control the trade-off between the bitrate of the encoded video and the amount of “loss,” which can be characterized by distortion functions. Traditionally, metrics such as Mean Square Error (MSE) or Peak signal to noise ratio (PNSR) have been used to measure compression-induced distortion. However, with the development of perceptual video quality algorithms, models such as Structural Similarity (SSIM) [9, 10] and Visual Multimethod Assessment Fusion (VMAF) [11] have become widely adopted in the video industry.

In practice, videos are typically encoded at a lower encoding resolution compared to their native resolution in order to exploit spatial redundancy and to achieve higher compression ratios. This is particularly useful for source videos that do not contain a great amount of detail. For example, a 1080p source content may be scaled down to 480p prior to encoding by downscaling and compression. At the user’s end, the encoded bitstream is decoded and upscaled to 1080p for display. This process is illustrated in Figure 1.1.

While the relationship between bitrate and distortion is generally monotonic for a given content, the perceptual effects of compression significantly

Database	Source Videos	Test Videos
BVI-CC [16]	9	198
LIVE-SJTU A/VQA [17]	14	336
VQEG-HDTV [18]	45	828
LIVE Streaming [19]	45	245

TABLE 1.1

Summary of Subjective Video Databases

vary with and depend on content, and the combined effects of scaling and further complicates analysis and modeling of distortion. RDO algorithms such as the Dynamic Optimizer [12] tackle this problem by performing a convex-hull analysis to identify a set of “Pareto-optimal” encoding “recipes” (such as a combination of encoding scale and bitrate) from the given set. The Pareto-optimal set essentially consists of a set of encoding recipes that may be ordered in terms of their rates and distortions, thereby restoring monotonicity. During deployment, the best encoding recipe is chosen based on a rate/distortion constraint.

While the traditional approach to scaling involves the use of classical interpolation algorithms such as Bicubic/Lanczos interpolation, modern approaches use Deep Neural Networks (DNNs) to learn scaling algorithms from data by optimizing the rate-distortion tradeoff. Deep Video Precoding [13] and Deep Perceptual Optimization (DPO) [14] are two recent advancements in this nascent area of research. In addition to the aforementioned algorithms, “progressive” neural network architectures have been used to support fractional scale factors [15].

1.1.3 Subjective Quality Databases

The effect of video compression on perceptual quality has been an area of extensive research over the last decade. As a result, several subjective databases of compressed HD videos exist in the literature. In this section, we will describe some exemplary subjective databases that have aided the development of powerful quality assessment algorithms. A summary of the databases described below is provided in Table 1.1.

1. **Bristol Vision Institute Codec Comparison Database (BVI-CC)**: The BVI-CC [16] database compares the performance of three popular video codecs, namely HEVC, VVC, and AV1, when compressing High-Definition (HD) and Ultra-High-Definition (UHD) content. The database consists of nine source sequences subjected to both scaling and compression distortions, leading to a total of 198 distorted test videos.

The study found that according to human viewers, AV1 and HEVC

performed similarly on average, while VVC generally outperformed both AV1 and HEVC. In addition, the study investigated the effect of the Dynamic Optimizer, which was found to improve the performance of HEVC over AV1.

2. **LIVE-SJTU Audio and Video Quality Assessment Database (LIVE-SJTU A/VQA):** The LIVE-SJTU A/VQA [17] database studies the joint impact of audio and video compression on subjective quality, which is representative of streaming use-cases. The database consists of 14 reference and 336 distorted sequences, generated by applying both video compression (with and without scaling) using the HEVC codec, and audio compression using the Advanced Audio Coding (AAC) codec.

The study found that subjects reported higher difficulty in assessing audio quality compared to video quality, with joint audio video quality being reported as the most difficult. Furthermore, the authors observed that subjective scores varied non-monotonically with combined audio and video bitrates, thereby illustrating the complexity of the multi-modal quality assessment task.

3. **Video Quality Experts Group - High Definition Television Database (VQEG-HDTV):** VQEG-HDTV [18] is a collection of five publicly available databases created by VQEG to validate objective quality models that may be used to evaluate the perceptual quality of HD videos. In total, the set of databases consists of 45 diverse source contents compressed using the MPEG-2 [20] and AVC codecs at various bitrates to yield 828 test contents. In addition to compression, the database also includes simulated IP network errors.
4. **LIVE Livestreaming Database:** The LIVE Livestreaming database [19] is a large subjective database that studies the impact of several streaming-related distortions such as compression, aliasing, judder, and frame drops. The database consists of 45 HD and UHD source contents, distorted to produce 245 test contents. The compression artifacts were produced using the AVC codec applied at four quality levels.

The study observed that UHD videos were typically rated higher than HD videos. Furthermore, aliasing due to scaling affected the quality of HD videos to most, while judder had the greatest impact on the subjective quality of UHD videos.

1.1.4 Objective Quality Assessment

In this section, we review three classes of objective quality models, namely, full-reference (FR), no-reference (NR), and bitstream-based quality models.

1.1.4.1 Full-Reference Quality Assessment

The journey of modern image and video quality assessment (I/VQA) begins with the development of the SSIM [9] index. Prior to SSIM, the standard measure of “distortion” was PSNR, which depends logarithmically on the mean squared error (MSE). However, PSNR was found to correlate poorly with subjective judgments, which is the primary goal of objective quality models.

SSIM introduced a much-needed perceptual element to the measurements of distortions by using similarity metrics that embody Weber masking of both luminance and contrast. Weber masking refers to the phenomenon that human perception is sensitive to relative changes, rather than absolute changes. As a result, in the presence of a large “background value,” such as a high luminance or contrast, small changes in the value are not perceptible, that is, are “masked.” SSIM and multi-scale SSIM (MS-SSIM) have successfully been used in the quality control of most cable, satellite, and fiber optic Internet streaming television providers and by most social media outlets. Recently, a variant of SSIM called Enhanced SSIM [10] has been shown to achieve even higher accuracy at lower computational cost.

Another breakthrough in quality assessment occurred with the development of the Visual Information Fidelity (VIF) quality model. VIF was the first algorithm to use powerful statistical models of natural scenes, modified to model distortions, to conduct FR quality assessment. VIF uses a Gaussian Scale Mixture (GSM) to model the statistics of bandpass wavelet coefficients, and a local Gaussian channel to model distortions. Versions of VIF have been deployed to conduct the quality control of videos on the Internet, most notably by contributing most of the features in the SOTA VMAF quality model. Furthermore, the philosophy behind VIF has been used to develop reduced-reference quality models such as ST-RRED [21] and SpEED-QA [22].

The Detail Loss Metric (DLM) [23] is another quality model that has emerged as a powerful tool to measure the quality of compressed videos. This is particularly because the primary spatial artifact introduced during compression is blocking, which leads to loss in detail that DLM is designed to capture. DLM uses a dyadic wavelet decomposition of input frames and a model of the contrast sensitivity function (CSF) of the Human Visual System (HVS) to decompose distortions into detail losses and additive impairments. Due to its high performance, DLM contributes one part of the SOTA VMAF quality model.

As alluded to above, the SOTA video quality models used by the streaming industry are MS-SSIM and VMAF, which was jointly developed by Netflix and academic collaborators. The core principle of VMAF is to fuse smaller “atom” quality models to obtain a better-performing quality model. VMAF can be viewed as an ensemble model consisting of “weaker” quality models. In particular, VMAF uses DLM, VIF computed at four scales on a Gaussian pyramid, and a temporal difference feature to capture Motion, as its atom

features. VMAF and its anti-hacking variant [24] have been widely adopted in the industry for quality assessment, RDO, and codec comparison.

1.1.4.2 No-Reference Quality Models

All the aforementioned algorithms are said to be “full-reference” since they assume the availability of a “reference video” for each test video. The reference video is a “pristine” version of the test video that is used as a gold standard against which the test video is evaluated. On the server-side, this assumption generally holds since the pristine source video is used to generate the encoded videos. As a result, the majority of quality algorithms used in the streaming video industry are FR.

Nevertheless, “no-reference” quality models have been an active area of research, finding applications in moderating the quality of source contents prior to encoding, and for any client-side quality assessment. As the name suggests, “no-reference” models operate on test videos directly without using a paired reference video. The classical approach to NR quality assessment has been using natural scene statistics (NSS).

“Natural” images have been observed to exhibit regular statistical properties when subjected to band-pass decompositions, and distorted videos exhibit deviations from ideal behavior. Modeling the statistics of band-pass coefficients yields quality-aware features that are used to predict the subjective quality of videos. This approach has formed the basis of popular NR image quality models such as DIIVINE [25], BRISQUE [26], and NIQE [27]. When assessing the quality of videos, these models are applied frame-wise and the quality scores are aggregated across frames.

Extending NSS-based quality models to video typically involves introducing temporal information in the form of spatio-temporal band-pass decompositions [28], or using optical flow/motion vectors. Quality models that utilize motion information have studied the use of optical flow statistics [29, 30], motion-coherency [31], and localized space-time slices called “chips” [32].

1.1.4.3 Bitstream-based Quality Models

A recent novel approach to assessing the quality of compressed videos is “bitstream-based” quality assessment. While traditional FR and NR models operate on decoded frames, that is, pixel values, bitstream-based models operate directly on information extracted from the encoded bitstream. This alleviates the additional computational burden of decoding videos prior to quality assessment. Common “features” obtained from the bitstream include values of frame resolution, frame rate, frame type (I, B, or P-frame), encoded frame size, bitrate, and QP.

The SOTA for bitstream-based quality assessment is the ITU P.1204.3 [33] standard, which uses a combination of a parametric model and a machine learning model to predict quality scores.

1.1.5 Challenges and Trends

Despite the success of FR models such as VMAF, there remains a variety of challenges to be tackled. Firstly, the SOTA FR model VMAF demands a high computational load due to the calculation of a heterogeneous set of atom features. Given the development of hardware encoders for popular codecs, quality assessment (using VMAF) has emerged as the bottleneck operation in RDO. In order to address these concerns and to advance the SOTA, the FUNQUE [34] model has been proposed. FUNQUE “unifies” the atom quality features by computing them from a common HVS-aware decomposition. The HVS-aware decomposition provides for an 10% improvement in correlation against subjective scores compared to VMAF, while the computation sharing reduces the computational load by a factor of 8!

Despite the advances made by FUNQUE, both VMAF and FUNQUE operate only on the luma (grayscale) channel and are almost entirely applied framewise. The only temporal information that has been included in the model is the temporal difference between successive reference frames. A common solution is to introduce optical-flow-based quality models and apply quality models on various color channels [35, 36]. However, such methods often drastically increase the computational burden of quality assessment. So, incorporating color and temporal quality information into SOTA FR models in an efficient manner has proven to be an open practical challenge.

Furthermore, the deep-learning revolution has arrived in the quality assessment space, with the development of models such as LPIPS [37], NIMA [38], Patch-VQ [39], and CONTRIQUE [40]. The current SOTA deep architecture for vision tasks is the Transformer [41], which “tokenizes” images into patches and applies attention mechanisms to identify salient features. The success of Transformers in computer vision suggests that their application to quality assessment could break new ground. A common challenge that hinders the widespread adoption of deep models is the computational burden of applying them at scale. However, the development of hardware-accelerated deep learning platforms and large server farms with dedicated tools such as Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs) aims to bridge the gap between research and the industry.

Finally, in practice, watching most videos involves listening to the audio, which encounters many of the same compression/transmission challenges that video faces. As a result, end users often simultaneously encounter distorted audio and video. However, the effect of audio on subjective quality is less understood when compared to videos. In fact, human subjects report greater difficulty in rating audio and audio-visual quality when compared to visual quality. Due to these factors, despite recent advances [17], multi-modal audio-visual quality assessment remains a challenging problem.

1.2 User-Generated Content (UGC) Videos

In January 1996, Bill Gates stated in a published article that “content is king” [42]. The television industry during the past century has arguably validated the value of his words, but it has become even more true in the 2020s as we enter a new phase of the Internet where overwhelming user creation weighs over professional artifacts made by traditional media. The rise of social media platforms such as Facebook, YouTube, and TikTok, coupled with the increased availability of affordable mobile capturing devices, has led to UGC being created, uploaded, and shared almost instantly and ubiquitously across the globe every day. Among the vast amount of shared video content, UGC videos undoubtedly account for the largest portion of Internet streams. It has been reported that YouTube viewers watch over 1 billion hours of video every day and generate billions of views [43]. Thus, it is of great importance for streaming companies to understand and analyze this new type of content, and accordingly, design optimized UGC processing pipelines to improve the encoding and streaming efficiency of UGC videos.

1.2.1 Introduction to UGC Video Quality

VQA is one of the core research topics in video compression and streaming pipelines that are designed to monitor, control, and calibrate video processing algorithms. Most VQA models may be classified into two categories: full-reference models (FR-VQA) and no-reference or blind (NR-VQA or BVQA) metrics. FR-VQA models compare the perceptual difference against pristine videos, while BVQA models directly predict the quality without references. For UGC videos, the “pristine” video is never available, and hence, only NR-VQA models can be applied, which makes the UGC-VQA problem [44] very challenging. Moreover, unlike relatively simple synthesized distortions simulated in traditional VQA databases, UGC video distortions arise from realistic user captures. Thus, UGC videos suffer from unknown, highly diverse, and comingled impairments. These authentic and intermixed visual defects include, but are not limited to, bad lighting conditions, capturing by shaky hands, imperfect shooting devices, processing and editing effects, compression and streaming artifacts, etc. Figure 1.2 shows some examples of UGC videos having various degrees of degradation. Indeed, apart from the technical distortions such as noise, blur, ringing, and banding that have been studied for decades, there exist even more complicated and subjective factors that affect how viewers perceive regarding aesthetics, interestingness, and content preference. Finally, many current video databases dedicated to UGC quality prediction are far smaller in size compared to mainstream video databases [45]. Moreover, none of them are capable of modeling the full complexity and diversity of real-world UGC distortions. Due to these challenging aspects, predicting the

perceptual quality of UGC videos has remained largely unsolved for several years. So, it is of great interest to both industry and academia to investigate this most general and impactful problem in video quality research.

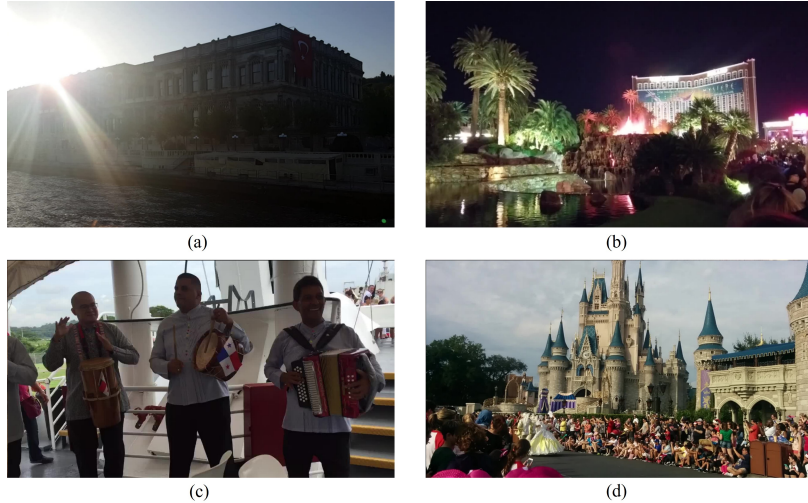


FIGURE 1.2

Exemplar frames of UGC video contents sampled from the LIVE-VQC database [46]. These natural videos captured from mobile devices are subject to unknown, highly diverse video distortions, and making it harder for objective video quality models to accurately predict their perceived qualities.

1.2.2 Subjective UGC Video Quality Databases

Subjective video quality research usually requires extensive resources invested in time- and label-consuming human studies on a selected video corpus to obtain reliable subjective data. These human-labeled videos are then used to build a video quality database for the training, benchmarking, and calibration of video quality prediction (VQA) models. Traditional subjective video quality databases are typically constructed by first selecting a small set of high-quality sources, dubbed “pristine” or “reference,” then synthetically distorting them using simulated approaches. Some examples of databases constructed in this manner are LIVE-VQA [47], EPFL-PoliMI [48], CSIQ [49], and MCL [50]. These databases are limited in their diversity of contents and distortions, making them incapable of modeling more complicated, real-world UGC impairments.

Recently, researchers have started to build “in-the-wild” databases that include authentically distorted contents captured under realistic settings. We list and briefly summarize these efforts as follows.

1. **CVD2014 [51]**. The Camera Video Database (CVD2014) is perhaps the first authentically distorted video quality database in the literature. It consists of 234 480p and 720p videos recorded using 78 different cameras that are prone to distortions induced by video acquisition. However, the database only covers five unique scenes.
2. **LIVE-Qualcomm Mobile In-Capture Database [52]**. The LIVE-Qualcomm database proposed in 2017 is another camera-captured database containing 208 realistic consumer contents with smartphones in full HD (1080p) resolution. Compared to CVD2014, this dataset contains more than 50 different scenes, thus having better diversity in contents.
3. **KoNViD-1k [53]**. The KoNViD-1k dataset is a subjective UGC video quality database consisting of 1,200 public-domain video sequences. The 1,200 sequences were sampled from a large public video dataset named YFCC100m [54], based on a selected feature space of blur, contrast, colorfulness, Spatial Information (SI), Temporal Information (TI), and NIQE scores [27]. All the videos were trimmed and resized to 540p for eight seconds.
4. **LIVE-VQC [46]**. The LIVE-VQC database is another large-scale UGC database containing videos that were captured using a diverse set of mobile devices. It consists of 585 contents that span various resolutions, layouts, and framerates. Subjective scores were collected from 4,776 unique participants via online crowdsourcing on Amazon Mechanical Turk.
5. **YouTube UGC [55]**. In 2020, Google released a large-scale UGC dataset called YouTube UGC, whose videos were sampled from uncurated YouTube uploads. It contains a total of 1,380 unique 20-second-long video clips, having resolutions ranging from 360p to 4k and framerates ranging from 15 to 60. Over 8,000 subjects were hired to rate these videos in a crowdsourcing environment.
6. **KonVid-150k [56]**. The KoNViD-150k dataset follows similar design principles as KonViD-1k, but was orders-of-magnitude larger. It consists of two subsets, a coarsely annotated set of 153,841 videos having five quality ratings each, and a densely annotated set of 1,596 videos with over 89 ratings each. Notably, the first subset of KonVid-150k can be used as a weakly labeled pre-training set for various downstream VQA tasks.
7. **LIVE-FB LSVQ [39]**. LSVQ is yet another large-scale “in-the-wild” UGC database containing 39,000 real-world videos and 117k space-time localized video patches, labeled by 5.5 million human perceptual quality annotations. It is constructed from web-scale crawled video sources including Internet Archive [57] and YFCC-

100M [54] to make the dataset more representative of real-world content.

8. **LIVE-YT Gaming [58].** LIVE-YouTube Gaming dataset is the first UGC video dataset dedicated for a specific type of UGC – gaming contents. It consists of 600 real-streaming UGC gaming videos spanning 59 various types of games.

1.2.3 Objective UGC Video Quality Assessment

The goal of subjective video quality studies or databases is to help develop automatic objective VQA metrics. As described before, UGC video quality problems can only be solved by no-reference or blind VQA algorithms since no “pristine” video is available for these UGC uploads. Thus, here we briefly review the evolution of BVQA models, from traditional feature-based approaches to recent deep learning-based models.

Traditional feature-based BVQA models. Almost all the earliest blind quality prediction models were developed for a single or at most two types of distortions. Some examples are models for analyzing noise [59], blockiness [60], blur [61], ringing [62], banding artifacts [63] in streamed, distorted videos, or to assess multiply concurrent compression and transmission defects [64, 65], to name a few. More recently, general-purpose BVQA models that are capable of capturing versatile distortions have been developed based on natural scene statistics (NSS) [66] and its deduced simple but highly effective low-level features [26, 27] equipped with machine learning techniques. Among these NSS-based models, BRISQUE [26] and NIQE [27] operate in the divisively-normalized image domain, BLIINDS [67] examines statistics in the DCT domain, and Video-BLIINDS [31] extended BLIINDS to analyze DCT transforms of frame differences. TLVQM [29] leverages a bag of well-defined quality-related features extracted from multiple visual spaces and then trains a shallow support vector regressor. VIDEVAL [44] is the most recent BVQA model, and it uses an ensemble of features based on NSS inspired by multiple prior BVQA models. VIDEVAL has been shown to achieve strong performance on UGC databases.

Deep learning-based BVQA models. Despite performing well on recent UGC-VQA databases, these “handcrafted” feature-based models have been observed to underperform on recently released larger-scale datasets such as LSVQ. On the other hand, deep convolutional neural networks (ConvNet) that are purely data-driven have gained strength rapidly. VSFA [68] is a strong CNN-based model that applies a pre-trained ResNet backbone as a feature extractor and integrates the deep spatial features using recurrent layers and temporal pooling. PatchVQ [39] models the relationship between local and global perceptual quality from space, time, and space-time video patches. Finally, PatchVQ spatiotemporally pools quality scores to make the final predictions. Li et al. [69] leveraged transfer learning for the BVQA task by mixed

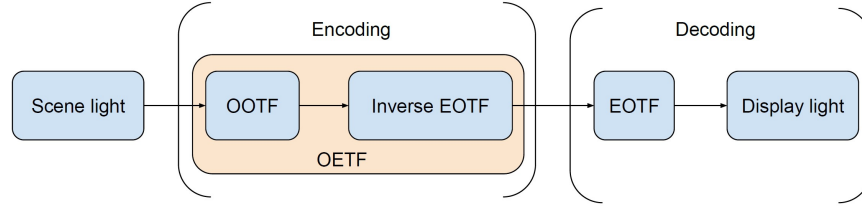
dataset pre-training on both authentic IQA databases and large-scale action recognition sets. Some researchers have also proposed taking advantage of both shallow (statistical) features and deep-learning features to make the best prediction, such as CNN-TLVQM [70], RAPIQUE [71], and GAME-VQP [72].

1.2.4 Challenges and Future Trends

Although promising progress has been made on the UGC-VQA problem, several challenges remain. Firstly, compared to mainstream vision tasks such as image/video classification [73, 74], existing video databases dedicated to video quality are quite limited in scale, thus unable to fully exploit the power of deep learning approaches. Building massive-scale, diverse, and unbiased video datasets for video quality and compression research continues to be a critical research topic. Secondly, current deep learning-based models, despite showing standout performance, require pretty heavy floating-point operations. Thus, in their current form, they are neither efficient nor practical to be widely deployed. Exploring accurate and also efficient VQA models that can make predictions almost real-time for streaming services is still very challenging and largely unsolved. Finally, thanks to the evolution of smartphones and portable cameras, users can capture, curate, and share advanced media like high dynamic range (HDR), high frame-rate (HFR), and even virtual reality (VR) sources. Most current VQA research involving these types of media modalities mainly focuses on small-scale compression video studies, whereas future trends in UGC video quality research must extensively embrace such “advanced-for-now” but “normal-to-be” multimedia sources.

1.3 High Dynamic Range Video

The human vision system (HVS) can perceive luminance levels between 10^{-6} cd/m² and 10^8 cd/m² using various mechanical, photochemical, and neuronal adaptive processes [75]. However, traditional imaging and display systems produce content having much narrower ranges of luminance values than the vision system is able to perceive. These older content formats are commonly referred to as Standard Dynamic Range (SDR). HDR is a set of techniques that extend the ranges of luminances and colors that can be represented and displayed. HDR has seen increasing adoption over the past few years, particularly in the area of streaming and video hosting, where services such as Amazon Prime, Netflix, and YouTube now offer content in HDR. HDR is also used as the default standard for UHD Blu-Rays. Major TV manufacturers such as LG, Samsung, and Panasonic support HDR content, and manufacturers such as Lenovo and Apple have also recently released laptops that can display HDR

**FIGURE 1.3**

A representation of the PQ system. The initial step involves the transformation of the scene light into an electrical signal through the application of the OETF, followed by the conversion of the electrical signal into display light utilizing the EOTF.

content. HDR is now part of live broadcast and film production workflows and is progressing rapidly into an industry standard.

1.3.1 Challenges Facing Adoption of HDR

The challenges of HDR VQA arise from the distinct features of HDR videos. HDR videos contain much wider luminance ranges than SDR. HDR video content represents peak brightness up to 10,000 nits ($1 \text{ nit} = 1 \text{ candela/meter}^2$), enabling the display of brighter highlights. Its lower limits are also below those of SDR content, enabling the display of darker black and deeper shadows. In order to represent such a wide range of luminance with limited data volume, while introducing minimal perceivable distortion due to quantization, the HLG and PQ electro-optical transfer functions (EOTFs) are designed to replace the legacy gamma EOTF defined in BT.-1886 [76], which was originally designed to be compatible with old cathode ray tube (CRT) displays, as shown in Figure 1.3. Furthermore, the ranges of colors that can be represented by HDR video are also broader and richer than those captured by SDR. Modern HDR videos use the Rec. 2020 [77] color primaries, which cover 75.8% of the CIE 1931 color space, much larger than the Rec. 709/sRGB color gamut [78], which covers 35.6% of the CIE 1931 color space.

The use of different nonlinear transfer functions in HDR can greatly change the visibility and severity of compression distortions. The relationship between the intensity of the light as the physical magnitude measured on the display and how bright it appears to a human observer has long been known to be not linear [79]. The perceived brightness heavily depends on both the image stimulus and the viewing environment, including the image background, the peak luminance and the dynamic range of the display [80, 81]. As a result, the same magnitude of contrast may appear very different under different circumstances, making it difficult to predict the perception of distortions.

The increases in bit depth and the use of nonlinear transfer functions in HDR provide different challenges for the task of HDR VQA. 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 [11] can be applied to 10-bit data, but it does not take into account the extended luminance range or transfer function of HDR10. However, there are a few VQA models that address the compression of HDR videos, even working with 10-bit data. These models will be discussed later in the section.

Finally, one more challenge is the scarcity of high-quality subjective VQA databases. As the HDR standards HDR10, HDR10+, and Dolby Vision are still fairly recent, publicly available HDR content is difficult to obtain, and there are few subjective VQA databases dedicated to HDR. Several existing databases have limited usefulness because they have been rendered obsolete by the latest HDR standards. Several other databases do comply with new HDR standards but are not publicly available due to copyright issues.

1.3.2 Publicly Available Subjective Quality Databases and Studies

1. **DML-HDR Video Database:** The DML-HDR video database [82] is designed to evaluate various HDR image and video quality metrics. The database consists of five videos captured by a RED SCARLET-X camera. The videos are approximately ten seconds long with a framerate of 30 frames per second (fps) and resolution of 2048×1080 . The videos are stored in “.hdr” and 12 bit “.yuv” raw files using the Rec. BT.709 color primaries. The videos were then compressed using H.264/AVC and HEVC, displayed on a custom-built HDR display system, and viewed by 17 subjects.
2. **HDR-VQM Video Database:** The HDR-VQM Video Dataset [83] is a subjective database that contains ten source HDR video sequences with a resolution of 1920×1080 and the frame rate are 25 fps. These videos were compressed at eight different bitrates using a backward-compatible compression method. The videos were tone mapped to an 8-bit SDR version and compressed. When the videos were decompressed, they were inverse tone mapped to HDR. The videos were displayed on an HDR display with a peak luminance of 4,000 nits in a relatively bright environment (200 cd/m^2).
3. **Laboratory for Image and Video Engineering (LIVE) HDR Database):** The LIVE HDR Database [84] is a subjective database that targets HDR10 videos. It contains 310 videos from 31 HDR10 source contents. All the source videos have a resolution of 3840×2160 and frame rates of 50 or 60 fps. The average duration of

videos in the database is approximately eight seconds. The source sequences were processed by ten different compression and resolution combinations using the HEVC codec, simulating standard industrial practices. A subjective quality study was conducted using the database with 66 subjects. The subjective study was conducted under two different illumination conditions to test the effects of ambient illumination on the perceived quality of HDR content. It is worth noting that the LIVE HDR Database is the only publicly available HDR VQA database that complies with contemporary HDR standards.

Other than these three databases, there are a few other HDR subjective quality studies that have not made their data publicly available. Pan *et al.* [85] created a VQA database using 12 pristine HDR videos compressed with 12 quantization parameters (QP) levels. The videos are generated by both PQ and HLG and the BT. 2020 color space, compressed using the AVS2 codec, developed by the Audio and Video Coding Standard (AVS) Working Group of China. A subjective quality study was conducted with 22 subjects using this dataset. Baroncini *et al.* [86] conducted a study of 12 compressed HDR videos evaluated by 40 human subjects. Rerabek *et al.* [87] conducted a study of 20 HDR videos with the resolution of 944×1080 , with four distortion levels. Athar *et al.* [88] conducted a subjective study of 14 HDR10 content compressed using H.264 and HEVC to generate 140 distorted videos viewed and rated by 51 subjects.

1.3.3 Video Quality Algorithms

Existing VQA algorithms mostly consider only 8-bit data and without much consideration regarding the nonlinear transform. Some VQA algorithms that have color components use color space-conversion methods that are specific to Rec. BT. 709 color spaces. This makes most of them inapplicable to HDR VQA scenarios. In this section, a few VQA algorithms dedicated to the HDR VQA problem are reviewed.

1.3.3.1 HDR-VDP

HDR-VDP is a modification proposed to extend the Visual Difference Predictor (VDP) to improve the prediction of perceivable differences in HDR applications [89]. HDR-VDP operates on linear luminance, and it relies on three stages to model the behavior of the optics and retina. HDR-VDP models the light scattering in the human eye using the Optical Transfer Function (OTF), scales the amplitude of the signal in units of Just Noticeable Differences (JND) to account for the nonlinear response of the HVS, and finally filters the data using a model of the CSF. Subsequently, improved versions of HDR-VDP have been proposed by including models of several luminance conditions, improved pooling methods, and better calibrations [90, 91].

1.3.3.2 PU-SSIM and PU-PSNR

Perceptually Uniform (PU)-SSIM and PU-PSNR have been proposed as an extension to the popular quality metrics [92]. A nonlinear transform has been proposed, called the perceptually uniform (PU) encoding, to ensure that the distortion visibility is approximately uniform across all encoded values. By using PU encoding to transform luminance values into perceptually uniform pixel values, SDR VQA algorithms can be directly applied to HDR images and videos.

1.3.3.3 HDR-VQM

HDR-VQM is an objective HDR video quality measure based on signal pre-processing, transformation, and subsequent frequency-based decomposition [83]. It uses the PU transform from PU-SSIM and PU-PSNR as a pre-processing step, then extracts the error signal in each subband in log-Gabor filters, and then the final result is obtained via a spatial-temporal pooling step.

1.3.3.4 HDRMAX and HDR-ChipQA

HDRMAX [84] and HDR-ChipQA [93] are two VQA methods based on a similar idea as PU-SSIM and PU-PSNR. They rely on nonlinear transforms to extend the base SDR algorithm to HDR-enhanced regions, that is, the bright and dark areas in the frames. HDR-VMAF is a FR method, while HDR-ChipQA is a NR method that uses statistical quality-aware features that are extracted in parallel from the original frames and the transformed frames.

1.4 High Frame Rate Video

Considerable advancement has been made in recent years in video technologies that improve the quality of experience (QoE) for end users. These improved capture, storage, and display technologies have facilitated the adoption of high resolution 4K videos, HDR videos, and HFR videos. The standard frame rate for movies is 24 fps and progressive television formats are typically 30 fps, while HFR videos are displayed at 50 fps or more. The increased frame rate of HFR videos alleviates temporal distortions such as motion blur and stutter. Powerful GPUs have made it possible to view high-resolution, HFR videos on consumer devices, and expansions in network capabilities now allow customers to stream live events in a HFR format. Quality assessment of HFR videos is an important and open problem in order to stream and store these videos

efficiently and improve QoE. In this section, we will discuss subjective quality databases and objective quality algorithms created for HFR.

1.4.1 Challenges Facing Adoption of HFR

A number of studies [94, 95, 96] have shown that for some contents quality saturates as frame rate is increased, but this is highly content dependent and HFR indeed makes a positive and significant difference in quality for high motion content. The perceptual benefits of HFR content are still relatively unknown due to a dearth of HFR content, and consumers are so accustomed to the aesthetics of SFR content that HFR content may be negatively received as being “too smooth” or having a “soap opera” effect, as happened in the case of the movie release of the “The Hobbit.” However, this is a question of aesthetics, and since display and capture technologies continue to develop the desire to view content that is as close to reality as possible will almost certainly lead to faster adoption of HFR. In fact, some studies [97, 98] have found that frame rates of above 900 fps may be necessary for certain types of contents to imitate optical reality. The prohibitive cost of producing and distributing HFR content and aesthetic concerns remains a significant challenge to its adoption in the present.

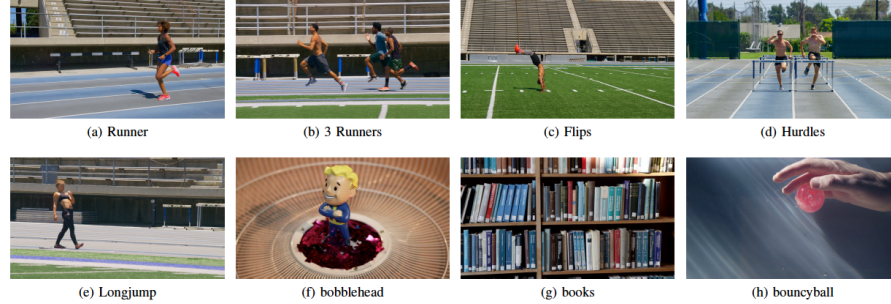
Increasing the frame rate of videos forces encoders to compress a greater number of bits, and hence the trade-off between bitrate budgets and quality is an important one to consider. Very few NR or FR VQA models exist that are specifically designed to predict the quality of videos with different frame rates, but such models are necessary for automated, content-dependent, and data-driven decisions on the best combinations of resolution, frame rate, and bit depth for a fixed bitrate budget. The lack of these models poses a significant challenge to incorporating HFR into production workflows in an efficient and cost-friendly way that still delivers the best QoE.

1.4.2 Subjective Quality Databases

In this subsection, recent publicly available subjective quality databases that study the effect of frame rate on video quality are reviewed.

1.4.2.1 LIVE YT HFR

The LIVE YT HFR [99] database consists of 480 videos having six different frame rates. The temporally downsampled videos were obtained from the source videos by frame dropping to 24 fps, 30 fps, 60 fps, 82 fps, and 98 fps. All of the videos were by varying the CRF in the VP9 encoder. Each video was viewed by a minimum of 42 subjects. The videos in this database were displayed at their native frame rate by adjusting the refresh rate of the display to match the frame rate for each video. Screenshots of some contents from this database are shown in Figure 1.4.

**FIGURE 1.4**

Screenshots of contents from the LIVE YTHFR database. Figure taken from [99].

The analysis of the subjective scores yielded the conclusion that the frame rate has a considerable impact on human subjective judgments of video quality but that the extent of the preference towards higher frame rates was content-dependent. Higher frame rates were strongly preferred for videos with significant camera motion. Another important conclusion drawn from this study was that the improvement in subjective quality diminishes above 60 fps.

1.4.2.2 ETRI LIVE STSVQ

The ETRI LIVE STSVQ [100] database contains 437 videos generated by spatial downsampling, temporal downsampling, and compression of 15 4K 10 bit source sequences. The resolutions of the downsampled videos are 1080p, 720p, and 540p and the frame rates represented in the database are 30 fps, 60 fps, and 120 fps, respectively. The videos were compressed using the x265 encoder at 5 bitrates. The temporally downsampled videos were created by frame dropping and were temporally interpolated linearly to the frame rate of their source videos before they were displayed, in contrast to LIVE YTHFR where the videos were displayed at their native frame rates. Each video was rated by 34 subjects.

The study concluded that as the bitrate budget was reduced, spatial and temporal downsampling led to higher quality scores as compression artifacts are minimized when there is less video data to compress. However, the tradeoff between spatial and temporal downsampling is highly content-dependent. For contents with rapid and large motions, the temporal downsampling can cause drops in quality even at lower bitrates and performing spatial downsampling alone may produce better quality judgments at limited bitrate budgets. For contents with little or no motion but a large amount of spatial detail, the rate-distortion curve would favor temporal downsampling over spatial downsampling at low bitrates.

1.4.2.3 AVT-VQDB-UHD-1

AVT-VQDB-UHD-1 [101] Test 4 consists of 120 videos with resolutions ranging from 360p to 2160p and frame rates ranging from 15 fps to 60 fps. The videos were compressed using HEVC. Each video was watched by at least 24 participants. The study showed that leading FR VQA models such as VMAF performed poorly on videos with very low frame rates compared to the source video, and that low frame rates had a strong negative impact on quality scores, but quality saturated as frame rate was increased.

1.4.2.4 BVI HFR

BVI HFR [102] contains 22 source sequences and their temporally downsampled versions. The frame rates represented in the database are 15 fps, 30 fps, 60 fps, and 120 fps. The temporally downsampled videos were generated by frame-averaging. Frame dropping may result in strobing artefacts, while frame-averaging can cause motion blur. Frame-averaged videos requires fewer bits than frame-dropped videos because edges lose their sharpness. The authors found in a mini-study that frame-averaged videos were preferred over frame-dropped videos for very low frame rates. Fifty-one subjects participated in this study. The study showed that increasing frame rates had a greater impact on visual quality when camera motion was present.

1.4.3 Video Quality Algorithms

Existing FR VQA algorithms such as VMAF and SSIM can only compare videos that have the same number of frames and can hence only be used to evaluate video quality across frame rates after the SFR video is temporally upsampled to the same frame rate as the HFR source video. Such models have been observed to perform poorly against human subjective opinions. We briefly review algorithms that have specifically designed for assessing quality across frame rates.

1.4.3.1 STGREED

STGREED [103] is a FR VQA algorithm that is based on principles of natural video statistics. The statistics of temporally band-passed video coefficients are modelled as generalized Gaussian distributions. The entropies of these coefficients are computed and it was observed that the entropies were relatively constant for the same frame rate for different bitrates, but varied across frame rates. An entropic difference term was designed that would remove the frame rate bias while accounting for quality variations if the distorted video had the same frame rate as the source video. STGREED's performance on LIVE YTHFR is shown in Table 1.2 against that of other algorithms, using Spearman's Rank Ordered Correlation Coefficient (SROCC), Pearson's Linear

TABLE 1.2

Performance Comparison of STGREED Against Different FR Algorithms on the LIVE-YT-HFR Database.

	SROCC $\uparrow$	PLCC $\uparrow$	RMSE $\downarrow$
PSNR	0.7802	0.7481	7.75
SSIM [9]	0.5566	0.5418	9.99
MS-SSIM [104]	0.4135	0.5512	10.01
FSIM [105]	0.6528	0.6332	9.34
ST-RRED [106]	0.4516	0.6073	9.58
SpEED [107]	0.6051	0.5206	10.28
FRQM [108]	0.5133	0.5017	10.38
VMAF [11]	0.7782	0.7419	8.10
deepVQA [109]	0.4331	0.3996	10.87
STGREED	0.8822	0.8869	5.48

Correlation Coefficient (PLCC), and the Root Mean Square Error (RMSE) between the predictions and the subjective scores as the metrics.

1.4.3.2 VSTR

“nobreak

1.4.3.3 FAVER

frame rate-Aware Video Evaluator w/o Reference (FAVER) [110] is an NR VQA model designed for variable frame rate videos. FAVER models the statistics of temporally and spatially band-passed videos as GGDs. The temporal filtering is similar to that used in ST-GREED. Statistical features are extracted from luma and chromatic difference channels for QA. FAVER achieves the highest performance on the YTHFR database, with a correlation of approximately 0.6 against human quality scores. The relatively low value of the SOTA shows the difficulty and challenging nature of NR VQA for videos with different frame rates.

1.5 Virtual Reality

VR and its applications have attracted significant and increasing attention. However, the requirements of much larger file sizes, different storage formats, and immersive viewing conditions pose significant challenges to the goals of acquiring, transmitting, compressing and displaying high quality VR content. Towards meeting these challenges, it is important to understand the distor-

tions that arise when delivering VR content and the effect they have on the perceived quality of displayed VR content.

1.5.1 Challenges Facing Adoption of VR

VR images are usually captured using a 360 camera equipped with multiple lenses that capture the entire 360 degrees of a scene. The recent Insta360 Titan is a professional 360 camera with eight 200° fisheye lenses that can capture both 2D and 3D images of resolution up to 11K. After the images are captured simultaneously by separate lenses, they are stitched together to generate a spherical image.

Unlike traditional viewing conditions where people watch images and videos on flat-panel computer and mobile displays, VR offers a more immersive viewing environment. Since the image can cover the entire viewing space, users are free to view the image in every direction. Usually, only a small portion of the image is displayed as they gaze in any given direction, so the content that a user sees is highly dependent on the spatial distribution of image content, the object being fixated on, and the spatial distribution of visual attention. The free-viewing of high resolution, immersive VR implies significant data volume, which leads to challenges when storing, transmitting and rendering the images that can affect the viewing quality. Therefore, it is important to be able to analyze and predict the perceptual quality of immersive VR image.

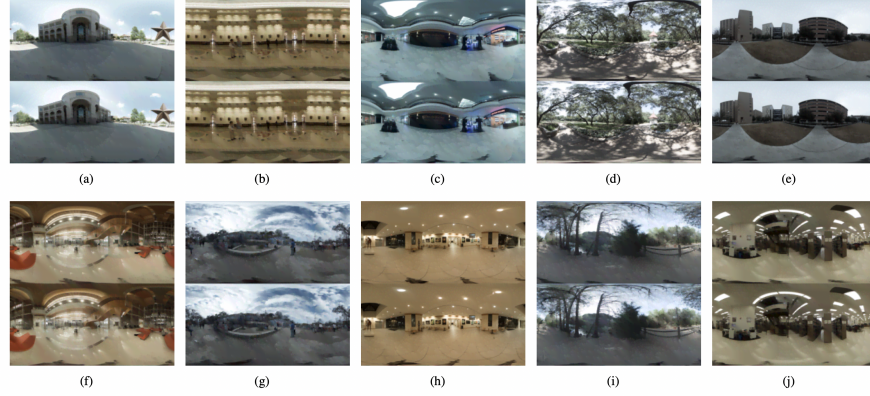
1.5.2 Subjective Quality Databases

In this subsection, recent publicly available subjective quality databases that study VR images and videos are reviewed.

1.5.2.1 LIVE 3D VR IQA Database

The LIVE 3D VR IQA Database [111] contains 450 distorted 3D VR images obtained from 15 pristine images modified by six types of distortions. The images were evaluated by 42 subjects using an HTC Vive headset with Tobii eye-tracking. This dataset includes traditional distortions like Gaussian noise, Gaussian blur, downsampling, as well as VR-specific stitching distortions. The dataset also includes recent compression distortions, including VP9 and H.265, to study and model the way they compress and perceptually distort VR images.

The study concluded that applying a reprojection weight to modify traditional IQA methods can help their performance on VR images. The study also showed that training on the subject data of these 3D VR images was an important step of the NR models to capture the unique perceptual peculiarities of the distorted VR image viewing experience. From the gaze maps of the study, it was shown that there exists an equator bias when viewing VR images. Subjects were more likely to view the center of the image (center bias), but

**FIGURE 1.5**

Screenshots of contents from the LIVE 3D VR IQA database. Figure taken from [111].

this also depended on the content and whether there were objects of interest near the center. Sample images from the dataset are shown in Figure 1.5.

1.5.2.2 IVQAD 2017 Database

The IVQAD 2017 database [112] contains 10 raw videos and 150 distorted videos degraded by three quality factors: bit rate, frame rate, and resolution. To simulate quality degradation, three resolutions are picked as 4096×2048 , 2048×1024 , and 1024×512 , and under each resolution, different bit rates and frame rates are selected to simulate different bandwidth. All the videos are encoded with MPEG-4 and 13 subjects were involved in this study. The study concluded that resolution and frame rate have almost the same degree of influence on quality assessment, whereas bit rate has the least influence.

1.5.2.3 VRQ-TJU Database

The VRQ-TJU database [113] contains 13 reference 3D VR videos subjected to two distortions: H.264 compression and JPEG 2000 compression, both symmetrically and asymmetrically on the original videos. In total, 104 symmetrically distorted videos and 260 asymmetrically distorted videos were obtained and evaluated by 30 subjects.

1.5.2.4 LIVE-FBT-FCVR Database

The LIVE-FBT-FCVR database [114] consists of two separate databases, one for 2D foveated compressed videos and the other one for 3D foveated compressed videos. Each database includes 10 reference videos and 180 foveated videos, which were processed by three levels of foveation on the pristine videos.

In the 2D study, each video was of resolution 7680×3840 and was viewed by 36 subjects, while in the 3D study, each video was of resolution 5376×5376 and rated by 34 subjects. Both studies were conducted on top of a foveated video player having a motion-to-photon latency of around 50ms.

The study showed that in the 2D study, subjective quality was more affected by peripheral quality, while in the 3D study, subjective quality was largely affected by foveal quality.

1.5.3 VR Quality Algorithms

It is important to develop algorithms to assess the quality of VR content based on its characteristics. In this section, we briefly introduce different types of VR quality assessment algorithms, namely quality algorithms for VR non-foveated content and VR foveated QA.

VR Quality Algorithms. Several VR-specific QA models have been proposed over the years, with the increasing interest in VR. The very first attempts towards developing VR-specific algorithms were based on existing algorithms for flat panel displays such as SSIM and PSNR. Yu *et al.* [115] proposed a spherical PSNR model called S-PSNR that averages quality over all viewing directions. In [116], a craster parabolic projection-based PSNR (CPP-PSNR) VR-IQA model was introduced. Xu *et al.* [117] proposed two kinds of perceptual VQA (P-VQA) methods: a non content-based PSNR (NCP-PSNR) algorithm and a content-based PSNR (CP-PSNR) method. WS-PSNR [118] is yet another PSNR based VR-IQA method, which reweights pixels according to their location in space. SSIM has also been extended in a similar manner, as exemplified by S-SSIM [119].

Yang *et al.* [120] proposed a content-aware algorithm designed specifically to assess stitched VR images, by combining a geometric error metric with a locally constructed guided IQA method. An NR-IQA method designed to assess stitched panoramic images using convolutional sparse coding and compound feature selection was proposed in [121]. Given the explosive popularity of deep learning, many more recent VR-IQA methods have been learned to analyze immersive images and videos, often achieving impressive results. For example, in [113], the authors deployed an end-to-end 3D convolutional neural network to predict the quality of VR videos without reference. In [122] and [123], the power of adversarial learning was utilized to successfully predict the quality of images.

VR Foveated Quality Algorithms. The Foveated Wavelet Image Quality Index (FWQI) utilizes wavelets to extract position-dependent spatial quality information [124]. Several factors are taken into consideration, including the spatial CSF, which is used to determine local visual cutoff frequencies, which guides modeling of human visual sensitivity across the available wavelet subbands, when combined with assumptions on viewing distance and the display resolution. Lee *et al.* [125] proposed a foveal signal-to-noise ratio (FSNR) to evaluate the quality of picture or video streams. In this method, a foveated

image is obtained by a foveated coordinate transformation on the original image(s) whose quality is to be assessed. Riouac-Drlje *et al.* [126] proposed a foveated mean squared error (FMSE) that models the effects of spatial acuity reduction due to motion. Another model called the foveation-based content Adaptive Structural SIMilarity index (FA-SSIM), which is based on the popular IQA model SSIM [127] combines SSIM with a foveation-based sensitivity function. You *et al.* [128] proposed a full reference attention-driven foveated video quality metric (AFViQ) that accounts for the localization of fixations in images and videos. A recently developed NR FVQA model called Space-Variant BRISQUE (SVBRISQUE) achieves SOTA performance using NSS features and a neural noise model to predict the quality of immersive videos [114].

1.6 Conclusion

In this chapter, we have discussed various video modalities in media and entertainment, including popular modes such as video streaming and social media, and emerging modalities such as high dynamic range (HDR) and high frame rate (HFR) video, and 3D virtual reality (VR). While all visual media rely on the same core perceptual principles, understanding domain-specific artifacts and distortions requires the development of dedicated subjective databases, which are in turn used to develop objective quality models. These models are then used to validate and enhance the quality of videos that are delivered to the end-user.

Moreover, we have identified key challenges and opportunities for future research in these various domains. Some challenges are domain-specific, such as the lack of color-aware or framerate-aware quality models for streaming and HFR videos. On the other hand, we have also identified broader challenges that affect the field of video quality assessment as a whole. One such challenge is the relatively small size of subjective databases, which is a consequence of the difficulty in obtaining subjective ratings. Recently, large-scale databases such as KonVid-150k and LIVE-FB LSVQ have been developed for user-generated social media content (UGC), but still more progress is required to usher in the deep learning era for video quality assessment. At the same time, the high data requirements of emerging video modalities such as HDR, HFR, and VR video demand the development of efficient encoding and quality assessment algorithms. This presents an interesting opportunity to turn a challenge in one domain (high data volume) into a solution in another domain (higher data requirements for quality assessment).

In summary, the development of perceptual quality models has fueled the explosion of video streaming and consumption over the last decade, and the

future holds exciting possibilities and interesting challenges for emerging visual media.



References

- [1] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.
- [2] H.S. Malvar, A. Hallapuro, M. Karczewicz, and L. Kerofsky. Low-complexity transform and quantization in h.264/avc. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):598–603, 2003.
- [3] Thomas Wiegand. Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H. 264—ISO/IEC 14496-10 AVC). *JVT-G050*, 2003.
- [4] D. Marpe, H. Schwarz, and T. Wiegand. Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):620–636, 2003.
- [5] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012.
- [6] B. Bross, Y. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J. R. Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.
- [7] D. Mukherjee, J. Bankoski, A. Grange, J. Han, J. Koleszar, P. Wilkins, Y. Xu, and R. Bultje. The latest open-source video codec VP9 - An overview and preliminary results. In *Picture Coding Symposium (PCS)*, pages 390–393, 2013.
- [8] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi, C. Chiang, Y. Wang, P. Wilkins, J. Bankoski, L. Trudeau, N. Egge, J. Valin, T. Davies, S. Midtskogen, A. Norkin, and P. de Rivaz. An overview of core coding tools in the AV1 Video Codec. In *Picture Coding Symposium (PCS)*, pages 41–45, 2018.
- [9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

- [10] A. K. Venkataramanan, C. Wu, A. C. Bovik, I. Katsavounidis, and Z. Shahid. A Hitchhiker’s guide to structural similarity. *IEEE Access*, 9:28872–28896, 2021.
- [11] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward a practical perceptual video quality metric. *Netflix TechBlog*, Apr 2017.
- [12] I. Katsavounidis. Dynamic optimizer - A perceptual video encoding optimization framework. *The Netflix Tech Blog*, 2018.
- [13] E. Bourtsoulatz, A. Chadha, I. Fadeev, V. Giotsas, and Y. Andreopoulos. Deep video precoding. *arXiv preprint arXiv:1908.00812*, 2019.
- [14] A. Chadha, R. Anam, I. Fadeev, V. Giotsas, and Y. Andreopoulos. Escaping the complexity-bitrate-quality barriers of video encoders via deep perceptual optimization. In *Applications of Digital Image Processing XLIII*, volume 11510, pages 38 – 52. International Society for Optics and Photonics, SPIE, 2020.
- [15] L. Chen, C. G. Bampis, Z. Li, J. Sole, and A. C. Bovik. A progressive architecture for learned fractional downsampling. In *2021 Picture Coding Symposium (PCS)*, pages 1–5, 2021.
- [16] A. Katsenou, F. Zhang, M. Afonso, G. Dimitrov, and D. R. Bull. BVI-CC: A dataset for research on video compression and quality assessment. *Frontiers in Signal Processing*, 2, 2022.
- [17] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Transactions on Image Processing*, 29:6054–6068, 2020.
- [18] Report on the validation of video quality models for high definition video content, 2010.
- [19] Z. Shang, J. P. Ebenezer, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik. Study of the subjective and objective quality of high motion live streaming videos. *IEEE Transactions on Image Processing*, 31:1027–1041, 2022.
- [20] A. Puri. Video coding using the MPEG-2 compression standard. In Barry G. Haskell and Hsueh-Ming Hang, editors, *Visual Communications and Image Processing ’93*, volume 2094, pages 1701 – 1713. International Society for Optics and Photonics, SPIE, 1993.
- [21] R. Soundararajan and A. C. Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):684–694, 2013.

- [22] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik. SpEED-QA: Spatial efficient entropic differencing for image and video quality. *IEEE Signal Processing Letters*, 24(9):1333–1337, 2017.
- [23] S. Li, F. Zhang, L. Ma, and K. N. Ngan. Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia*, 13(5):935–949, 2011.
- [24] Z. Li, K. Swanson, C. Bampis, L. Krasula, and A. Aaron. Toward a better quality metric for the video community. *The Netflix Tech Blog*, page 2, 2020.
- [25] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011.
- [26] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.
- [27] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.
- [28] S. V. Reddy Dendi and S. S. Channappayya. No-reference video quality assessment using natural spatiotemporal scene statistics. *IEEE Transactions on Image Processing*, 29:5612–5624, 2020.
- [29] J. Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019.
- [30] K. Manasa and S. S. Channappayya. An optical flow-based no-reference video quality assessment algorithm. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2400–2404, 2016.
- [31] M. A. Saad, A. C. Bovik, and C. Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014.
- [32] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik. ChipQA: No-reference video quality prediction via space-time chips. *IEEE Transactions on Image Processing*, 30:8059–8074, 2021.
- [33] R. R. R. Rao, S. Göring, P. List, W. Robitza, B. Feiten, U. Wüstenhagen, and A. Raake. Bitstream-based model standard for 4K/UHD: ITU-T P.1204.3 — Model details, evaluation, analysis and open source implementation. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2020.

- [34] A. K. Venkataramanan, C. Stejerean, and A. C. Bovik. Funque: Fusion of unified quality evaluators. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2147–2151, 2022.
- [35] F. Zhang, A. Katsenou, C. Bampis, L. Krasula, Z. Li, and D. Bull. Enhancing VMAF through new feature integration and model combination. In *2021 Picture Coding Symposium (PCS)*, pages 1–5, 2021.
- [36] L. Chen, C. G. Bampis, Z. Li, J. Sole, and A. C. Bovik. Perceptual video quality prediction emphasizing chroma distortions. *IEEE Transactions on Image Processing*, 30:1408–1422, 2021.
- [37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.
- [38] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.
- [39] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik. Patch-VQ: ‘patching up’ the video quality problem. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021.
- [40] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.
- [41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.
- [42] B. Gates. Content is king. *Retrieved October, 29:2017*, 1996.
- [43] M. Mohsin. 10 youtube statistics that you need to know in 2021, May 2022.
- [44] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik. Ugcvqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021.
- [45] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.
- [46] Z. Sinno and A. C. Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2019.
- [47] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441, 2010.

- [48] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi. A H.264/AVC video database for the evaluation of quality metrics. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2430–2433, 2010.
- [49] P. V. Vu and D. M. Chandler. ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, 23(1):013016, 2014.
- [50] J. Y. Lin, R. Song, C. Wu, T. Liu, H. Wang, and C-C J. Kuo. MCL-V: A streaming video quality assessment database. *Journal of Visual Communication and Image Representation*, 30:1–9, 2015.
- [51] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen. CVD2014—A database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, 2016.
- [52] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K. Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077, 2018.
- [53] 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 *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017.
- [54] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [55] Y. Wang, S. Inguva, and B. Adsumilli. YouTube UGC dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019.
- [56] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe. KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. *IEEE Access*, 9:72139–72160, 2021.
- [57] Internet Archive: Digital Library of Free & Borrowable Books, Movies, Music & Wayback Machine.
- [58] X. Yu, Z. Ying, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C Bovik. Subjective and objective analysis of streamed gaming videos. *arXiv preprint arXiv:2203.12824*, 2022.
- [59] C. Chen, M. Izadi, and A. Kokaram. A perceptual quality metric for videos distorted by spatially correlated noise. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1277–1285, 2016.

- [60] Z. Wang, A.C. Bovik, and B.L. Evan. Blind measurement of blocking artifacts in images. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, volume 3, pages 981–984 vol.3, 2000.
- [61] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. A no-reference perceptual blur metric. In *Proceedings. International Conference on Image Processing*, volume 3, pages III–III, 2002.
- [62] X. Feng and J. P. Allebach. Measurement of ringing artifacts in JPEG images. In Jan P. Allebach and Hui Chao, editors, *Digital Publishing*, volume 6076, pages 74 – 83. International Society for Optics and Photonics, SPIE, 2006.
- [63] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik. Bband index: A no-reference banding artifact predictor. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2712–2716. IEEE, 2020.
- [64] J. E. Caviedes and F. Oberti. No-reference quality metric for degraded and enhanced video. In *Visual Communications and Image Processing 2003*, volume 5150, pages 621–632. SPIE, 2003.
- [65] C. Keimel, T. Oelbaum, and K. Diepold. No-reference video quality evaluation for high-definition video. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1145–1148, 2009.
- [66] D. L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517, 1994.
- [67] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012.
- [68] D. Li, T. Jiang, and M. Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019.
- [69] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
- [70] J. Korhonen, Y. Su, and J. You. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3311–3319, 2020.

- [71] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C Bovik. RAPIQUE: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021.
- [72] X. Yu, Z. Tu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C Bovik. Perceptual quality assessment of UGC gaming videos. *arXiv preprint arXiv:2204.00128*, 2022.
- [73] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [74] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [75] 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.
- [76] ITU. BT.1886 : Reference electro-optical transfer function for flat panel displays used in HDTV studio production. Technical report, Intl. Telecomm. Union, 2011.
- [77] ITU. BT.2020 : Parameter values for ultra-high definition television systems for production and international programme exchange, 2015.
- [78] ITU. BT.709 : Parameter values for the HDTV standards for production and international programme exchange. Technical report, Intl. Telecomm. Union, 2011.
- [79] T. N. Cornsweet and H. M. Pinsker. Luminance discrimination of brief flashes under various conditions of adaptation. *The Journal of Physiology*, 176(2):294, 1965.
- [80] M. 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.
- [81] V. A. Billock and B. H. Tsou. To honor Fechner and obey Stevens: Relationships between psychophysical and neural nonlinearities. *Psychological Bulletin*, 137(1):1, 2011.
- [82] 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. *TBD*, 2022.
- [83] 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, 2015.
 - [84] Zaixi Shang, Joshua P. Ebenezer, Abhinav K. Venkataramanan, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. A study of subjective and objective quality assessment of HDR videos. *IEEE Transactions on Image Processing*, 33:42–57, 2024.
 - [85] 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. *Journal of Visual Communication and Image Representation*, 57:76–83, 2018.
 - [86] V. Baroncini, K. Andersson, A.K. Ramasubramonian, and G. Sullivan. Verification test report for HDR/WCG video coding using HEVC main 10 profile. In *Proceedings of the JCTVC-X1018 24th JCT-VC Meeting*, 2016.
 - [87] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi. Subjective and objective evaluation of HDR video compression. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2015.
 - [88] S. Athar, T. Costa, K. Zeng, and Z. Wang. Perceptual quality assessment of UHD-HDR-WCG videos. In *IEEE International Conference on Image Processing*, pages 1740–1744, 2019.
 - [89] 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.
 - [90] 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 Transactions on Graphics*, 30(4), jul 2011.
 - [91] 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.

- [92] 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.
- [93] Joshua P. Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. HDR-ChipQA: No-Reference quality assessment on high dynamic range videos, 2023.
- [94] R. Selfridge, K. C. Noland, and M. Hansard. Visibility of motion blur and strobing artefacts in video at 100 frames per second. In *Proceedings of the 13th European Conference on Visual Media Production (CVMP 2016)*, pages 1–10, 2016.
- [95] P. J. Bex, G. K. Edgar, and A. T. Smith. Multiple images appear when motion energy detection fails. *Journal of Experimental Psychology: Human Perception and Performance*, 21(2):231, 1995.
- [96] R. M. Nasiri, J. Wang, A. Rehman, S. Wang, and Z. Wang. Perceptual quality assessment of high frame rate video. In *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2015.
- [97] A. Mackin, K. C. Noland, and D. R. Bull. High frame rates and the visibility of motion artifacts. *SMPTE Motion Imaging Journal*, 126(5):41–51, 2017.
- [98] A. B. Watson. High frame rates and human vision: A view through the window of visibility. *SMPTE Motion Imaging Journal*, 122(2):18–32, 2013.
- [99] 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, 2021.
- [100] D. Y. Lee, S. Paul, C. G. Bampis, H. Ko, J. Kim, S. Y. Jeong, B. Homan, and A. C. Bovik. A subjective and objective study of space-time sub-sampled video quality. *IEEE Transactions on Image Processing*, 2021.
- [101] R. R. R. Rao, S. Göring, W. Robitza, B. Feiten, and A. Raake. AVT-VQDB-UHD-1: A large scale video quality database for UHD-1. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 17–177. IEEE, 2019.
- [102] A. Mackin, F. Zhang, and D. R. Bull. A study of high frame rate video formats. *IEEE Transactions on Multimedia*, 21(6):1499–1512, 2018.

- [103] P. C Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik. ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Transactions on Image Processing*, 30:7446–7457, 2021.
- [104] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402 Vol.2, Nov 2003.
- [105] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.
- [106] R. Soundararajan and A. C. Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):684–694, April 2013.
- [107] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik. SpEED-QA: Spatial efficient entropic differencing for image and video quality. *IEEE Signal Processing Letters*, 24(9):1333–1337, Sep. 2017.
- [108] F. Zhang, A. Mackin, and D. R. Bull. A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling. In *IEEE International Conference on Image Processing*, pages 300–304, Sep. 2017.
- [109] 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 *European Conference on Computer Vision*, pages 219–234, September 2018.
- [110] Q. Zheng, Z. Tu, Y. Fan, X. Zeng, and A. C. Bovik. No-reference quality assessment of variable frame-rate videos using temporal bandpass statistics. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1795–1799, 2022.
- [111] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. C. Bovik. Study of 3D virtual reality picture quality. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):89–102, 2019.
- [112] H. Duan, G. Zhai, X. Yang, D. Li, and W. Zhu. IVQAD 2017: An immersive video quality assessment database. In *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–5. IEEE, 2017.
- [113] J. Yang, T. Liu, B. Jiang, H. Song, and W. Lu. 3D panoramic virtual reality video quality assessment based on 3D convolutional neural networks. *IEEE Access*, 6:38669–38682, 2018.

- [114] Y. Jin, M. Chen, T. Goodall, A. Patney, and A. C Bovik. Subjective and objective quality assessment of 2D and 3D foveated video compression in virtual reality. *IEEE Transactions on Image Processing*, 30:5905–5919, 2021.
- [115] M. Yu, H. Lakshman, and B. Girod. A framework to evaluate omnidirectional video coding schemes. In *2015 IEEE International Symposium on Mixed and Augmented Reality*, pages 31–36. IEEE, 2015.
- [116] V. Zakharchenko, K. P. Choi, and J. H. Park. Quality metric for spherical panoramic video. In *Optics and Photonics for Information Processing X*, volume 9970, page 99700C. International Society for Optics and Photonics, 2016.
- [117] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan. Assessing visual quality of omnidirectional videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.
- [118] Y. Sun, A. Lu, and L. Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing letters*, 24(9):1408–1412, 2017.
- [119] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang. Spherical structural similarity index for objective omnidirectional video quality assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.
- [120] L. Yang, Z. Tan, Z. Huang, and G. Cheung. A content-aware metric for stitched panoramic image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2487–2494, 2017.
- [121] S. Ling, G. Cheung, and P. Le Callet. No-reference quality assessment for stitched panoramic images using convolutional sparse coding and compound feature selection. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.
- [122] H. T. Lim, H. G. Kim, and Y. M. Ra. VR IQA net: Deep virtual reality image quality assessment using adversarial learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6737–6741. IEEE, 2018.
- [123] H. G. Kim, H. T. Lim, and Y. M. Ro. Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
- [124] Z. Wang, A. C. Bovik, L. Lu, and J. L. Koulouheris. Foveated wavelet image quality index. 4472:42–53, 2001.

- [125] S. Lee, M. S. Pattichis, and A. C. Bovik. Foveated video quality assessment. *IEEE Transactions on Multimedia*, 4(1):129–132, 2002.
- [126] S. Rimac-Drlje, M. Vranješ, and D. Žagar. Foveated mean squared error - A novel video quality metric. *Multimedia Tools and Applications*, 49(3):425–445, 2010.
- [127] Snježana Rimac-Drlje, Goran Martinović, and Branka Zovko-Cihlar. Foveation-based content adaptive structural similarity index. In *2011 18th International Conference on Systems, Signals and Image Processing*, pages 1–4, 2011.
- [128] J. You, T. Ebrahimi, and A. Perkins. Attention driven foveated video quality assessment. *IEEE Transactions on Image Processing*, 23(1):200–213, 2014.