frontiers | Frontiers in Signal Processing

TYPE Review
PUBLISHED 27 June 2023
DOI 10.3389/frsip.2023.1193523

Check for updates

OPEN ACCESS

EDITED BY
Feng Yang,
Google, United States

REVIEWED BY
Frederic Dufaux,
Université Paris-Saclay, France
Junjie Ke,
Google, United States
Xinwei Yao,
Google, United States

*CORRESPONDENCE
Avinab Saha,
✉ avinab.saha@utexas.edu

†These authors have contributed equally to this work

RECEIVED 25 March 2023
ACCEPTED 15 June 2023
PUBLISHED 27 June 2023

CITATION
Saha A, Pentapati SK, Shang Z, Pahwa R, Chen B, Gedik HE, Mishra S and Bovik AC (2023), Perceptual video quality assessment: the journey continues!. *Front. Sig. Proc.* 3:1193523.
doi: 10.3389/frsip.2023.1193523

COPYRIGHT
© 2023 Saha, Pentapati, Shang, Pahwa, Chen, Gedik, Mishra and Bovik. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Perceptual video quality assessment: the journey continues!

Avinab Saha*†, Sai Karthikey Pentapati†, Zaixi Shang†, Ramit Pahwa†, Bowen Chen†, Hakan Emre Gedik†, Sandeep Mishra† and Alan C. Bovik†

Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, United States

Perceptual Video Quality Assessment (VQA) is one of the most fundamental and challenging problems in the field of Video Engineering. Along with video compression, it has become one of two dominant theoretical and algorithmic technologies in television streaming and social media. Over the last 2 decades, the volume of video traffic over the internet has grown exponentially, powered by rapid advancements in cloud services, faster video compression technologies, and increased access to high-speed, low-latency wireless internet connectivity. This has given rise to issues related to delivering extraordinary volumes of picture and video data to an increasingly sophisticated and demanding global audience. Consequently, developing algorithms to measure the quality of pictures and videos as perceived by humans has become increasingly critical since these algorithms can be used to perceptually optimize trade-offs between quality and bandwidth consumption. VQA models have evolved from algorithms developed for generic 2D videos to specialized algorithms explicitly designed for on-demand video streaming, user-generated content (UGC), virtual and augmented reality (VR and AR), cloud gaming, high dynamic range (HDR), and high frame rate (HFR) scenarios. Along the way, we also describe the advancement in algorithm design, beginning with traditional hand-crafted feature-based methods and finishing with current deep-learning models powering accurate VQA algorithms. We also discuss the evolution of Subjective Video Quality databases containing videos and human-annotated quality scores, which are the necessary tools to create, test, compare, and benchmark VQA algorithms. To finish, we discuss emerging trends in VQA algorithm design and general perspectives on the evolution of Video Quality Assessment in the foreseeable future.

KEYWORDS

**video quality assessment, subjective quality database, quality of experience, streaming, VR/AR, cloud gaming, HDR**

## 1 Introduction

Perceptual Video Quality Assessment (VQA) is a field in Video Engineering that has attained increasing importance in the last few decades due to the proliferation of video-based applications and services. Along with video compression, VQA has become one of two dominant theoretical and algorithmic technologies in television streaming and social media. Video content is ubiquitous, from streaming movies and TV shows over the internet to video conferencing, social media, cloud gaming, and virtual reality applications. Increasing internet speeds, high-speed data transmission, low-latency wireless connectivity, faster encoding technologies, and rapid smartphone user growth have together contributed to an exponential rise in video content generation and consumption.

According to a report by PR-Newswire (2023), video traffic on the internet grew by 24% in 2022 and contributed to 65% of the internet traffic. Delivering high-quality video content and retaining users' interest is of prime importance. Insufficient video quality often leads to dissatisfaction, reduced engagement, and a negative user experience. Thus, Video Quality Assessment (VQA) is a crucial component of video engineering pipelines, as it ensures that the video content meets viewers' expectations.

Video Quality Assessment aims to objectively assess the perceived quality of videos based on human perception. The performance of VQA algorithms is generally measured by the correlation of their predictions with human judgments. Video quality can be affected by many factors, such as spatial and temporal resolution, frame rate, compression, color depth, and contrast, and other visual impairments, such as artifacts, noise, blur, compression, and distortion. The intricate interplay among the subjective nature of human perception, the content in the video, and the video distortions contribute to the unique challenges in the Video Quality Assessment task. Thus, to understand such complex phenomena, it is imperative to conduct human studies where volunteers watch videos, and their judgments of video quality are recorded in the form of opinion scores. In most cases, the feedback is limited to opinion scores, with some subjective studies involving physiological measures such as eye-tracking Liu and Heynderickx (2011). Objective VQA algorithms developed using the data from these subjective studies aim to mimic human judgments closely. In recent years, there has been a significant advancement in the field of Video Quality Assessment (VQA), as algorithms initially created for general 2D videos have evolved to cater to the need of specific applications such as virtual and augmented reality (VR and AR), cloud gaming, high dynamic range (HDR), high frame rate (HFR), on-demand video streaming, and user-generated content (UGC). Creating psychometric video quality databases is crucial to develop new and improved algorithms for these specialized applications. This survey aims to comprehensively review classical and recent developments in the VQA field by discussing and comparing the salient features of various psychometric subjective video quality databases and VQA algorithms.

## 1.1 Related surveys

Survey papers have proved crucial in advancing research across various disciplines, including Image and Video Quality Assessment. These papers serve as a valuable resource for researchers as they offer an extensive summary of existing research and enable them to identify research gaps and avenues for further exploration. By providing a comprehensive overview of the existing literature and upcoming trends, survey papers facilitate a deeper understanding of the subject matter and help identify research objectives. This section discusses the relevant surveys in the Image and Video Quality Assessment domain. Early survey works include Wang and Bovik (2009), where they provided an initial analysis of Full-Reference (FR) image fidelity measure, with a pivot on mean square error (MSE). Later, Wang and Bovik (2011) provided a more general introduction to Reduced-Reference (RR) and No-Reference (NR) image quality assessment methods. In 2011, the survey on perceptual visual quality metrics by Lin and Kuo (2011) discussed several key aspects, including signal decomposition, just-noticeable distortions, visual attention, feature and artifact detection, feature pooling, viewing conditions, computer-generated signals, and visual attention. In 2011; Moorthy and Bovik (2011) presented their vision for the future of VQA research. They postulated that we could achieve better performance by leveraging our growing understanding of the human visual system into the development of quality assessment algorithms; Chikkerur et al. (2011) discussed FR and RR objective evaluation metrics by dividing them based on visual characteristics. In 2014; Mohammadi et al. (2014) reviewed subjective and objective image quality assessment techniques, focusing on Full-Reference Image Quality Assessment (FR IQA) measures and two emerging directions: high dynamic range (HDR) and 3D IQA. Another survey by Shahid et al. (2014) discusses classical and well-known NR VQA methods. A recent survey by Zhou et al. (2022) focuses on Quality of experience (QoE) assessment for adaptive video streaming.

## 1.2 Relevance of our survey

This survey provides a comprehensive and up-to-date review of the new domains and methods which have emerged in the past decade. Most of the existing surveys in VQA discuss classical VQA techniques. Classical techniques are important, but recently, there has been a gradual shift to more deep learning-based techniques, greatly impacting all facets of VQA. The current surveys also fail to address progress in application-specific VQA, such as HDR, HFR, VR/AR, Cloud Gaming, and QoE. Our survey aims to address these shortcomings and provide a comprehensive discussion covering these trends.

## 1.3 Scope and organization of the survey

Our survey highlights the salient characteristics and novelty of various subjective quality databases and VQA algorithms. Although we thoroughly compare and analyze these algorithms and databases, we refrain from benchmarking them. This is because VQA is a specialized domain with varied application-specific use cases, and algorithms are generally developed to cater to a specific use case. While VQA algorithms developed for generic VQA tasks are often used to demonstrate usability in application-specific subjective quality databases, their performance is generally inferior to algorithms developed for specific use cases. Additionally, it is worth noting that not all VQA algorithm results are available publicly across all databases, further limiting the feasibility of benchmarking. Our survey is organized as follows. Section 2 discusses the advancements in generic video quality assessment tasks and includes the television, online video streaming, and UGC videos use case. Section 3 introduces application-specific VQA and is organized into further sub-sections covering HDR,

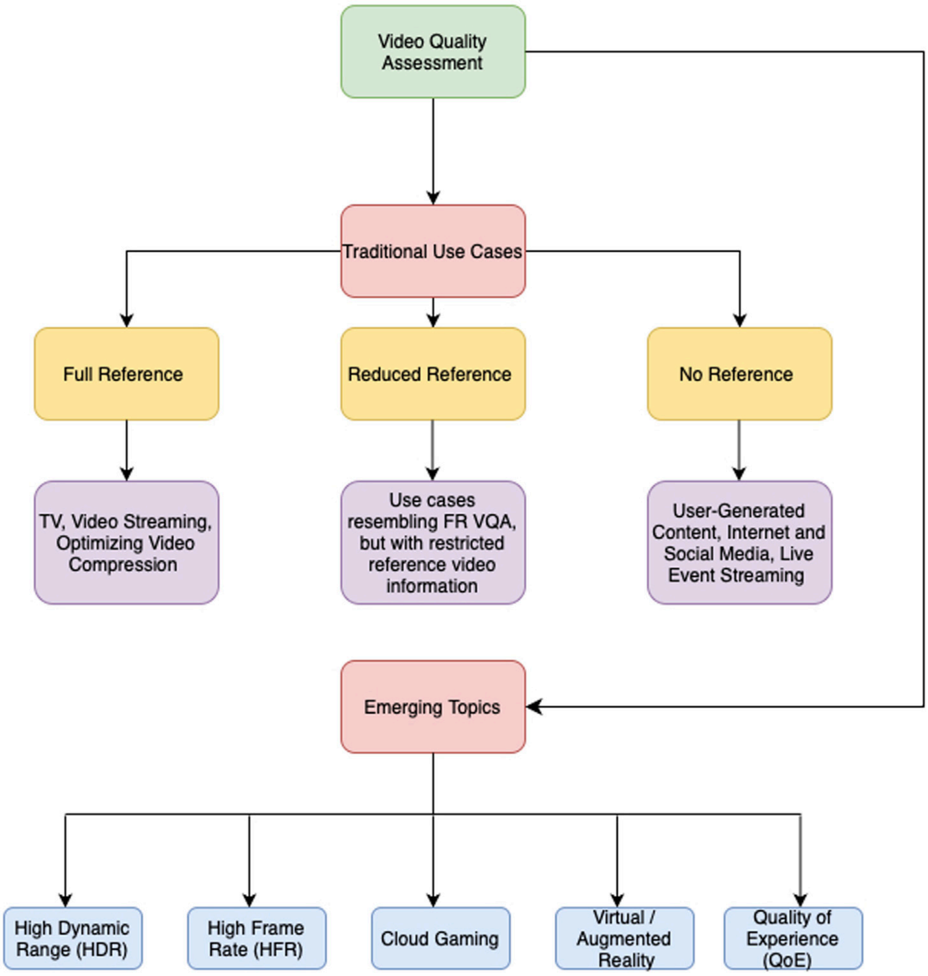
Video Quality Assessment
Traditional Use Cases
Full Reference
Reduced Reference
No Reference
TV, Video Streaming, Optimizing Video Compression
Use cases resembling FR VQA, but with restricted reference video information
User-Generated Content, Internet and Social Media, Live Event Streaming
Emerging Topics
High Dynamic Range (HDR)
High Frame Rate (HFR)
Cloud Gaming
Virtual / Augmented Reality
Quality of Experience (QoE)

FIGURE 1
Organization of our Survey.

HFR (High-Frame Rate), VR/AR, Cloud Gaming, and Quality of Experience. Figure 1 illustrates this categorization pictorially. We conclude in Section 4 discussing the upcoming trends in VQA algorithm design and general perspectives on the evolution of Video Quality Assessment in the foreseeable future.

## 2 Generic video quality assessment

Video Quality Assessment can be broadly classified into three categories based on the amount of information obtained from the original reference (pristine) video in the test video.

- **Full-Reference (FR-VQA)** involves comparing the entire reference video with the test video to evaluate its quality. This method is widely considered to be the most precise way to assess video quality. Essentially, FR-VQA involves measuring the signal or information fidelity of the test video with respect to the reference video. FR-VQA is widely used in the television and online streaming industry and in developing video compression algorithms.
- **Reduced Reference (RR-VQA)** methods are used when the entire reference video signal is unavailable. Only a subset of information from the reference video is available for comparison with the test video. Subsets of information may include but are not limited to motion vectors, edge information, texture, and color histograms. By comparing the test video to the reference video and considering the subset of available information, the quality of the test video can be assessed. While RR-VQA methods require less information about the reference video than FR-VQA, it may not be as accurate as FR-VQA. The practical use cases of the RR-VQA are limited as modern video quality monitoring workflows at the source typically use FR-VQA. In contrast, applications requiring real-time monitoring at the client use NR-VQA methods. Hence, we exclude this family of algorithms from our analysis.
- **No Reference Video (NR-VQA)** involves assessing the quality of the test video independent of the reference video or the distortions applied to the reference video. As a result, it is more complex and challenging than the FR-VQA and RR-VQA methods. NR-VQA is relevant for practical "in-the-wild"

scenarios and is widely used to assess video quality for social media, user-generated content, and live-streaming applications where a reference video is unavailable. NR-VQA is also crucial in determining algorithmic performance for various real-life applications, including Super-Resolution, Novel View Synthesis, and Video Enhancement methods where a reference is unavailable. NR-VQA is also essential for computational resource and latency constraint applications like Cloud Gaming, Live Video Streaming.

**Generic Training Pipeline of VQA Algorithms:** Most VQA algorithms use a combination of spatial and temporal feature extractors, utilizing distorted (test) videos and reference videos (only in the case of FR-VQA). These feature extractors are designed to capture relevant spatial and temporal information relating to video quality. Spatial feature extractors analyze the visual content in individual frames, focusing on attributes such as spatial attributes including colors, textures, shapes, and object representations. Temporal feature extractors, on the other hand, consider the changes and motion patterns between consecutive frames, capturing motion information and temporal dynamics. The spatial and temporal features extracted are combined and utilized to train a regressor that maps these features to the MOS. With the emergence of deep learning techniques and the availability of larger VQA databases, there has been a shift towards training many algorithms end-to-end. This leverages the power of deep neural networks to directly learn the mapping between visual inputs and quality assessments without explicit feature extraction and regression stages. By training the entire VQA model end-to-end, these algorithms can learn complex representations and capture intricate relationships between the input visual data and the corresponding MOS. The integration of feature extraction and regression in an end-to-end framework has demonstrated promising results, showcasing the potential of deep learning in advancing the field of VQA.

## 2.1 General VQA datasets

One of the most reliable ways to evaluate video quality is by conducting subjective studies involving human subjects. Typically a group of subjects watch the videos and provide quality ratings based on a standardized scale. These ratings are then compiled across all human subjects to create a comprehensive database of subjective scores corresponding to each video in the database. However, developing a subjective video quality database can be costly and time-consuming. This process involves several steps, such as recruiting subjects, designing a user-friendly software interface to display the videos and capture the human responses, selecting appropriate video content, and conducting the study in a controlled laboratory or on a crowdsourcing platform such as Amazon Mechanical Turk. Despite these challenges, developing subjective video quality databases is essential for several reasons. Firstly, they provide ground truth data for VQA algorithms, allowing video quality engineers and researchers to compare the performance of various objective quality assessment methods against a reference standard of scores obtained from the group of human subjects. Furthermore, developing VQA models that can predict the subjective quality of videos aids in automating the video quality assessment process, thus saving time and cost. Apart from VQA, subjective video quality databases are used to benchmark video compression codecs. Overall, subjective video quality databases play a crucial role in VQA and are essential for comparing and evaluating the performance of objective quality assessment algorithms.

**Calculation of Mean Opinion Score (MOS)**: The Subjective studies are conducted using human subjects, following which scores from all participants are aggregated to compute the Mean Opinion Score. In the past, the most commonly used method for computing the MOS was the one outlined in ITU-R BT.500-13 (ITU-R, 2012). The process involved calculating Z-scores and then subjecting them to rejection based on inconsistencies identified by scrutinizing the scores received from each human participant. However, an improved method for computing MOS has been proposed in the latest ITU-T BT P.910 (ITU-T, 2022) recommendation. This newer method proposes obtaining subjective quality scores from raw measurements that may be affected by noisy measurements. It involves using maximum likelihood estimation to simultaneously estimate the subjective quality of impaired videos, human participants' consistency and bias, and video content's ambiguity. We refer the reader to Li and Bampis (2017) for more details. Subjective video quality is typically represented using MOS, a reliable indicator of perceived quality. In the absence of reference undistorted videos, MOS is essential for developing and evaluating No-Reference (NR) Video Quality Assessment (VQA) algorithms. On the other hand, Difference MOS (DMOS) is commonly used in developing and evaluating Full-Reference (FR) VQA algorithms as it reduces content dependence. For a given distorted video, DMOS is determined by subtracting its own MOS value from the MOS of the original, undistorted version of the video.

**Categories of Video Quality Databases**: Depending on the video contents and the types of distortions present, subjective video quality databases can be divided into two broad categories.

- **Type 1: Synthetically Distorted Databases**: These are created by applying known types and levels of distortions to high-quality pristine source videos. Distortions may include introducing compression artifacts, noise, blurring, judder, aliasing, and other transmission relation distortions. The advantage of synthetic distortions is that they provide control over the type and level of distortion and can be adjusted to match specific research objectives, allowing researchers to evaluate the performance of video quality assessment algorithms under controlled conditions and to study the impact of different distortions on video quality. As high-quality pristine videos are available, these databases can be used to develop both FR-VQA and NR-VQA algorithms.
- **Type 2: Authentically Distorted Databases**: These databases are created aggregating real-world "in-the-wild" videos that have been degraded using unknown distortions and for which the corresponding high-quality versions are unavailable. These videos may contain distortions introduced during recording, transmission, or storage. The advantage of authentic distortion databases is that they provide a more realistic evaluation of video quality assessment algorithms. The distortions are similar to those encountered in real-world video applications. As the reference pristine video is unavailable corresponding to each "in-the-wild" video, these databases can be used to develop only NR-VQA algorithms.

**TABLE 1 Summary of popular generic VQA databases.**

| Database | # Videos | # Pristine source sequences | Source (type 1)/ Video (type 2) characteristics | # Ratings per video | Public | Distortions | Duration | Display device |
|---|---|---|---|---|---|---|---|---|
| LIVE-VQA (Type 1) | 150 | 10 | 768 × 432/25-50fps | 38 | Yes | H.264, MPEG-2 Simulated transmission errors | 8.68–10 s | 1,024 × 768 CRT Monitor |
| EPFL-PoliMI (Type 1) | 156 | 12 | CIF-4CIF/30fps | 23 | Yes | H.264/AVC Simulated transmission errors | 10 s | 30″, WQXGA, LED Monitor, 19″, SXGA, LED Monitor |
| VQEG HDTV (Type 1) | 675 | 45 | 1080p/25-30fps | 24 | Yes | H.264, MPEG-2 Simulated transmission errors | 10 s | 24.1″, UXGA, LED Monitor |
| MCL-JCV (Type 1) | 1,560 | 30 | 1080p/Varying fps | 50 | Yes | H264/AVC | 5 s | 65-inch, 4K TV |
| CVD-2014 (Type 2) | 234 | UGC Content | 480p-720p/Varying fps | 30 | Yes | Authentic distortions | 10–25 s | 28″ 4K LED Monitor |
| LIVE-Qualcomm Mobile In-Capture (Type 2) | 208 | UGC Content | 1080p/30fps | 39 | Yes | Authentic distortions | 15 s | 24″, 1080p, LED Monitor |
| KoNViD-1k (Type 2) | 1,200 | UGC Content | 540p/Varying fps | 50 | Yes | Authentic distortions | 8 s | Online Study |
| KoNViD-150 k (Type 2) | 153,841 | UGC Content | 540p/Varying fps | 5/89 | Yes | Authentic distortions | 5 s | Online Study |
| LIVE-VQC (Type 2) | 585 | UGC Content | 240p-1080p/Varying fps | 240 | Yes | Authentic distortions | 10 s | Online Study |
| YouTube-UGC (Type 2) | 1,500 | UGC Content | 360p-4K/15–60 | 100 | Yes | Authentic distortions | 20 s | Online Study |
| LIVE-FB LSVQ (Type 2) | 39,000 | UGC Content | Varying Resolution/ Varying fps | 35 | Yes | Authentic distortions | 5–12 s | Online Study |
| Waterloo IVC 4K (Type 1) | 1,200 | 20 | 540p,1080p,4K/24-30fps | 30 | Yes | HEVC, H.264/AVC, VP9, AV1, AVS2 | 10 s | 28″, 4K, LED Monitor |
| LIVE-APV Livestream (Type 1) | 315 | 45 | 1080p-4K/25-30fps | 38 | Yes | H264, Aliasing, Judder, Flicker, Frame drops, Interlacing | 5–7 s | 65″, 4K, LED Monitor |
| LIVE-SJTU A/ V-QA (Type 1) | 336 | 14 | 1080p/24-30fps | 35 | Yes | HEVC, Compression, Scaling AAC (Audio) | 8 s | 23.8″, 1080p, LED Monitor |
| LIVE Wild Compressed (Type 1 + 2) | 275 | 55 (Sampled from LIVE-VQC) | 360p-1080p/25–30 fps | 40 | Yes | Authentic Distortions 2nd Stage: Scaling and H.264 | 10 s | 23.8″, 1080p, LED Monitor |

Each database type has advantages and limitations and is used for a specific purpose in VQA research. Table 1 lists the popular subjective video quality datasets along with their salient features like the number of videos, source characteristics, distortion types, number of ratings per video, and display devices used during the human study for each dataset.

Initial efforts aimed at developing subjective VQA databases include LIVE-VQA Seshadrinathan et al. (2010), EPFL-PoliMI De Simone et al. (2009), and VQEG-HDTV VQE (2010). These databases typically contain a very small number of pristine, high-quality videos, which are then synthetically degraded using video compression and simulated transmission channel distortions to generate degraded versions for the human study. As videos in these datasets contain synthetically introduced distortions, they belong to the category of the Type 1 databases, as described above. The LIVE-VQA database comprises 150 videos from 10 pristine video sequences created using four different distortions: MPEG-2 and H.264 video compression and simulated transmission errors over IP and Wireless networks. EPFL-PoliMI VQA database was developed to investigate the effects of transmission channel distortions on video quality. It consists of 156 video streams, encoded with H.264/AVC and corrupted by simulating the packet loss due to transmission over an error-prone network. The Video Quality Experts Group developed VQEG-HDTV (VQEG), a set of 5 databases specifically designed for HD videos. The database has a total of 675, including the 45 source videos, and they are encoded with MPEG-2 and H.264 video compression standards at various bitrate. Transmission errors were also included. Error types include packet errors (IP and Transport Stream) such as packet loss, packet delay variation, jitter, overflow and underflow, bit errors, and over-the-air transmission errors.

MCL-JCV (Wang et al., 2016), is another Type 1 database designed to explore “Just Noticeable Differences” (JNDs) in visual distortion perception. The dataset comprises 30 pristine video sequences encoded using H.264/AVC with quantization parameters (QP) ranging from integer values of 1–51. A unique aspect of the MCL-JCV database as compared to other databases is that as opposed to subjectively rating the quality of a particular video, participants were presented with two video sequences, which were distorted versions (at different QP values) of the same pristine video sequence, and were asked to determine if they could be distinguished. With 1,560 subjectively quality-rated videos, MCL-JCV introduced a considerably larger database than the existing LIVE-VQA, EPFL-PoliMI, and VQEG-HDTV databases in 2016.

In CVD 2014 (Nuutinen et al., 2016), the first subjective quality database in the Type 2 category was introduced. It contains a total of 234 videos that are captured using 78 cameras. The motivation for developing the database was to study the distortions induced during video acquisition using various cameras. However, it is worth noting that the database has limited scene diversity, as it only included five unique ones. LIVE-Qualcomm Mobile In-Capture Database (Ghadiyaram et al., 2018) was introduced to alleviate this issue. It comprises 208 1080p videos captured using eight mobile cameras across 50 scenes.

The subjective databases we have discussed above have been developed by conducting in-lab human studies. However, these studies have limitations, including a limited number of subjects and ratings per video. We will next discuss the popular Type 2 databases developed using online studies that address this issue. These databases KoNViD-1k (Hosu et al., 2017), KoNVID-150 k (Götz-Hahn et al., 2021), LIVE-VQC (Sinno and Bovik, 2019), YouTube-UGC (Wang et al., 2019), and LIVE-FB LSVQ (Ying et al., 2020) provide an opportunity to gather data from a more extensive and diverse pool of participants, which can help increase the study’s reliability and generalizability. KoNViD-1k consists of a total of 1,200 videos, sampled from YFCC100m (Thomee et al., 2016). Later, an extended version of the database, KoNViD-150 k, was introduced. Most videos in KoNViD-150 k are evaluated much more coarsely than in KoNViD-1k, with only 5 evaluations per video. LIVE-VQC is another closely related database. It contains 585 videos captured from 101 devices by 80 users covering various resolutions, layouts, and frame rates. In 2019, Google introduced the YouTube-UGC database, which contains diverse user-generated content (UGC) videos. The database comprises videos sampled from 1.5 million YouTube videos from various categories, including animations, cover songs, news clips, sports, and more. Moreover, the dataset is representative of millions of videos uploaded to YouTube, making it a valuable resource for VQA. LIVE-FB LSVQ is another significant Type-2 database that follows a methodology similar to the KoNViD database by sampling videos from the YFCC-100m database. However, LIVE-FB LSVQ also samples videos from the Internet Archive to create a more authentic representation of real-world conditions. Moreover, the uniqueness of the LIVE-FB LSVQ database is its inclusion of subjective ratings not only for the entire videos but also for the 117 k space-time localized patches sampled from the 39,000 videos in the database.

As the demand for 4K video streaming and the emergence of newer compression standards such as HEVC, VVC (Bross et al., 2021) continues, there was a growing need for subjective quality databases containing videos with these characteristics. Thus, Waterloo IVC 4K (Li Z. et al., 2019) and LIVE-APV (Shang et al., 2022b) were introduced. Both these databases belong to the Type 1 category, as distorted videos in the database are generated using synthetic distortions. Waterloo IVC 4K database was developed to compare the performance of modern video encoders on 4K videos, which require higher data rates. Twenty reference video sequences are encoded with H.264/AVC, VP9, AV1, AVS2, and HEVC at four distortion levels and three spatial resolutions. LIVE-APV Livestream Video Quality Assessment Database was built to investigate the aggravated effects of common distortions such as H.264 compression, aliasing, judder, flicker, frame drops, and interlacing on high-motion live-streamed videos 1080p and 4K videos.

Unlike other subjective VQA databases, the LIVE-SJTU Audio and Video Quality Assessment (A/V-QA) Database (Min et al., 2020) is unique as it includes both video and audio components. The accompanying subjective study recognizes that videos are typically presented with audio in real-life situations. Distortions in visual or auditory signals can impact the overall quality of experience (QoE), discussed in Section 3.5 of this review. The database comprises 336 audio-video sequences generated by applying synthetic audio and video compression to 14 pristine sources. The final database we discuss is the LIVE Wild Compressed (Yu et al., 2021) database. The main focus of this study is to analyze the quality of videos that undergo two stages of distortions. The first stage involves authentic distortion, commonly seen in user-generated content videos. The second stage involves synthetic compression using H.264 codec. This study is relevant for use cases where a “Video-in-the-Wild” is further compressed using controlled streaming settings. To evaluate the impact of these two stages of distortions, and the study creates four versions of each UGC video, i.e., “Video-in-the-Wild”, each with varying degrees of H.264 compression applied. In the following two sub-sections, we discuss the popular FR-VQA, and NR-VQA algorithms developed using the databases discussed in this section.

## 2.2 Full-reference objective video quality assessment

FR-VQA algorithms are designed to assess the fidelity between the distorted and the reference frames in the videos. The peak-signal-to-noise ratio (PSNR) has been widely used to measure image fidelity. However, it has been proven to correlate poorly with human visual distortion perception (Wang and Bovik, 2009). One of the earliest attempts to create an image fidelity metric that was motivated by human perception was made through the development of the Structural Similarity Index Measure (SSIM) (Wang et al., 2004). Over the last 2 decades, SSIM has been widely used to predict the perceived quality of digital television, cinematic pictures, and other kinds of digital images and videos. SSIM has been extended for videos by utilizing an average pooling strategy across all the frames in a video. Seshadrinathan and Bovik (2011) demonstrated that the perceptual video quality could also be measured using the hysteresis effect of Human Visual System (HVS) by extending IQA methods. Motion-compensated SSIM (MC-SSIM) was proposed by Moorthy and Bovik (2010a) inspired by video compression and proposes to compute video quality by evaluating structural retention between motion-compensated regions. MOVIE (Seshadrinathan and Bovik, 2009) is a spatio-temporally localized multi-scale framework for evaluating dynamic video fidelity where space, time, and space-time distortions are considered. ST-MAD (Vu et al., 2011) is an extension based upon the most apparent distortion (MAD) model (Larson and Chandler, 2010) and incorporates visual perception of motion artifacts. The Optical Flow-based VQA model (Manasa and Channappayya, 2016)is based on the hypothesis that distortions affect local flow statistics, and the extent of deviation from the original flow statistics is directly proportional to the amount of distortion present. AFViQ (You et al., 2013) exploits visual perceptual mechanisms in VQA by proposing an advanced foveal imaging model to generate a perceived video representation.

Similar to SSIM, VIF (Sheikh et al., 2005) is another important FR-IQA model that has been suitably adapted and used in many FR-VQA models. VIF employs a Gaussian Scale Mixture (GSM) to model the statistical properties of bandpass wavelet coefficients, along with a local Gaussian channel to model distortions. The principles underlying VIF have been applied in the development of many reduced-reference quality models, including ST-RRED (Soundararajan and Bovik, 2013) and SpEED-QA (Bampis et al., 2017a). It is important to note that the reduced reference models ST-RRED and SpEED-QA models can also be used in FR-VQA use cases. The most popular and widely used FR-VQA method is the Video Multi-Method Assessment Fusion (VMAF) (Li Z. et al., 2016), developed by Netflix and academic collaborators. The fundamental principle of VMAF is to combine “weaker” quality models, referred to as “atoms,” to create a higher-performing quality model. Essentially, VMAF functions as an ensemble model composed of several “weaker” quality models. The specific atoms utilized by VMAF include DLM (Li et al., 2011), VIF calculated on a Gaussian pyramid at four different scales, and a temporal difference feature designed to capture motion. Further extensions to VMAF were proposed in the form of Spatio Temporal VMAF (ST-VMAF), and Ensemble VMAF (E-VMAF) in Bampis et al. (2018a). VMAF models are known to have high computational requirements due to the need for calculating a diverse set of atom features. FUNQUE proposed in Venkataramanan et al. (2022) addresses this issue. FUNQUE aims to simplify the process by unifying the atom quality features through an HVS-aware decomposition. This decomposition produces a 10% improvement in correlation against subjective scores compared to VMAF while also reducing the computational cost by a factor of 8.

Recently, there has been a notable trend toward leveraging deep learning features to predict video quality. For instance, DeepVQA (Kim et al., 2018) applies a CNN-based feature extractor to quantify spatiotemporal visual perception, then aggregates the frame-wise quality scores over time. C3DVQA (Xu et al., 2020) employs 3D convolutional layers to address temporal aliasing issues that could arise from frame-wise score aggregation. In Zhang et al. (2021a), the authors suggest a method that involves integrating DenseNet with spatial pyramid pooling and RankNet. This approach enables the extraction of high-level distortion representation and global spatial information from samples while also allowing the characterization of temporal correlation among frames. Vision Transformer (Khan et al., 2022), which has demonstrated effectiveness in various vision tasks, has also been adapted for use in VQA. Li et al. (2021) combines CNN-based feature extraction with a Transformer-based encoder to enhance video quality prediction. In the next sub-section, we discuss NR-VQA methods.

## 2.3 No reference objective video quality assessment

Early works in NR-VQA utilized well-known No Reference Image Quality Assessment (NR-IQA) approaches such as NIQE (Mittal et al., 2012c), and BRISQUE (Mittal et al., 2012). These methods were extended by pooling them temporally to produce a single video quality rating. A comparative study of various temporal pooling strategies can be found in Tu et al. (2020). Although effective, these methods have limitations since they fail to fully leverage temporal information, an essential factor in accurately predicting video quality. Consequently, there arose a need for developing models designed explicitly for VQA. V-CORNIA (Xu et al., 2014) predicts video quality based on frame-level unsupervised feature learning and hysteresis temporal pooling. Video BLIINDS (Saad et al., 2014) combines natural video statistics (NVS)-based spatial-temporal features, spatial naturalness, and motion-related features to train a support vector regressor (SVR) model for NR-VQA. The VIIDEO model (Mittal et al., 2015) incorporates models of intrinsic statistical patterns found in natural videos, which are further utilized to quantify the disturbances caused by distortions. TLVQM (Korhonen, 2019) is a widely used NR-VQA model that evaluates the quality of consumer videos, usually affected by capture artifacts, such as sensor noise, motion blur, and camera shakiness. It employs a two-level approach where low complexity features are computed for every frame, while high complexity features are computed for only certain selected frames in the video. Similar to VMAF in FR-VQA, VIDEVAL (Tu et al., 2021a) is an NR-VQA model developed to evaluate the quality of UGC videos by combining subsets of statistical features extracted from various existing NR-IQA and NR-VQA models. It employs an SVR model that utilizes selected features to regress from the feature space to the final MOS scores. Chip-QA (Ebenezer et al., 2021) uses space-time slices to capture motion in localized regions of videos and applies the parametric model of natural video statistics to measure distortions. It can quantify distortions in a video by measuring deviations from the natural video statistics. The model performs better than existing feature learning methods while maintaining a low computational cost.

Deep learning-based feature learning approaches have greatly improved the performance of NR-VQA. To take advantage of pre-trained CNN architectures, VSFA (Li D. et al., 2019), and RIRNet (Chen et al., 2020) have utilized them to extract features from the video frames. Then, they modeled the temporal sequence of frames using GRU/RNN to aggregate the features into a video quality score, resulting in significant benefits for NR-VQA. Although models pre-trained for classification tasks can be beneficial, they often neglect low-level visual features crucial for the VQA task. This limitation is demonstrated in DeepSIM (Gao et al., 2017) and MC360IQA (Sun et al., 2019), where the authors highlight the importance of low-level visual and semantic features for image and video quality assessment. On the other hand, traditional methods (Saad et al., 2014) tend to extract low-level vision quality-aware features without considering the video content. To address these issues, RAPIQUE (Tu et al., 2021b) proposes a novel approach that combines quality-aware scene statistics features with content-aware deep convolutional features, leveraging the power of pre-trained deep networks for content understanding. This approach results in superior performance to traditional natural scene statistics (NSS) feature-based models, showcasing the potential to incorporate low-level visual features and deep learning-based approaches for VQA.

Recent NR-VQA works using deep networks include Patch-VQ (Ying et al., 2021), FAST-VQA (Wu et al., 2022a), HVS-5M (Zhang et al., 2022), DOVER (Wu et al., 2022b). Patch-VQ introduced a novel approach for VQA by creating a large-scale UGC video database with subjective ratings for full videos and spatiotemporal crops of the videos. The framework involves extracting patches along space, time, and space-time, followed by feature extraction using 2D and 3D CNNs. These extracted features are then pooled to obtain a quality score for the video. HVS-5M proposes a novel NR framework for video quality assessment by dividing the learning task into multiple modules. The framework utilizes ConvNext (Li X. et al., 2016) to extract spatial features, while the SlowFast Network (Feichtenhofer et al., 2019) is used to obtain dynamic temporal features. The extracted features are then pooled to obtain a quality score for the video by training a low-complex regressor. FAST-VQA (Wu et al., 2022a) significantly reduced the computation cost and improved performance by designing a quality-preserving video sampling scheme. FAST-VQA proposes Grid Mini-patch Sampling (GMS), which considers local quality by taking patches at raw resolution and captures global quality using Attention which helps build contextual relations between the sampled patches. DOVER aims to disentangle different aspects of the video during training to help the deep networks focus on different aspects of the video. They propose two separate Quality Evaluators: Aesthetic Quality Evaluator (AQE), which utilizes spatial down-sampling (Keys, 1981) and temporal sparse frame sampling (Wang et al., 2018) to learn semantic and contextual information, and a Technical Quality Evaluator (TQE), which uses sampled raw resolution patches to form fragments similar to what was introduced in FAST-VQA. These quality estimates are then fused together to output a final quality score. Next, we discuss the evaluation metrics used in benchmarking NR-VQA and FR-VQA algorithms.

**Evaluation Metrics**: The performance of objective FR-VQA/NR-VQA algorithms is evaluated using the following metrics: Spearman's Rank Order Correlation Coefficient (SROCC), Kendall Rank Correlation Coefficient (KRCC), Pearson's Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE). The metrics SROCC and KRCC measure the monotonicity of the objective model prediction (DMOS for FR-VQA/MOS for NR-VQA) concerning human scores, while the metrics PLCC and RMSE measure prediction accuracy. For PLCC and RMSE measures, the predicted quality scores were passed through a logistic non-linearity function as shown in Seshadrinathan et al. (2010) to further linearize the objective predictions and to place them on the same scale as MOS/DMOS:

$$f(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp\left(-x + \beta_3 / |\beta_4|\right)}$$

## 3 Application specific video quality assessment

Video quality assessment has evolved over the years to become increasingly specific. With technological advancements, researchers focus on developing methods catering to specific video types. In the following subsections, we discuss emerging topics in VQA, providing an in-depth overview of associated databases and algorithms specific to the application.

### 3.1 High dynamic range (HDR) videos

HDR adoption has risen recently, especially in video streaming services like Amazon Prime, Netflix, and YouTube. It is the standard for UHD Blu-rays and is supported by major TV manufacturers such as LG, Samsung, and Panasonic. HDR has become essential to live broadcast and film production workflows, with increasing adoption as an industry standard. HDR videos represent a significant technological advancement, offering a more realistic and immersive viewing experience. HDR techniques expand the range of luminance, color representation, and display, creating more realistic and immersive visual experiences. With HDR, the captured image contains a wider range of brightness and color, resulting in increased detail in both bright and dark areas of the image. This wider range of luminance values and expanded color gamut is achieved through different nonlinear transfer functions like HLG and PQ (ITU, 2018), designed to replace the legacy gamma Electro-Optical Transfer Function (EOTF) defined in BT. 1886 (ITU, 2011). Furthermore, modern HDR videos use the Rec. BT. 2020 (ITU, 2015) color primaries cover 75.8% of the CIE 1931 color space, much larger than the Rec. 709/sRGB color gamut, which covers 35.6% of the CIE 1931 color space. However, the new techniques also present several unique challenges due to the distinct features of HDR videos.

One of the primary challenges in VQA for HDR videos is the expanded luminance range compared to Standard Dynamic Range (SDR) videos, with peak brightness levels reaching 10,000 nits. This expanded luminance range allows for brighter highlights and deeper shadows, resulting in a more realistic and immersive viewing experience. However, evaluating this expanded luminance range requires specialized VQA models and subjective assessments. Another challenge is the increased bit-depth utilized in HDR. Most existing VQA models are designed to operate on 8-bit luminance and color data, making it difficult to evaluate the quality of HDR videos utilizing 10 or 12-bit data. This increase in bit-depth allows for a more precise representation of HDR's expanded luminance and color ranges but poses a significant challenge for VQA. Additionally, utilizing non-linear EOTFs in

HDR can greatly alter the visibility and severity of compression distortions. The relationship between the intensity of the light as the physical magnitude measured on display and how bright it appears to a human observer has long been known to be non-linear (Cornsweet and Pinsker, 1965). The perceived brightness heavily depends on both the image stimulus and the viewing environment, including the image background, peak luminance, and the display's dynamic range (Billock and Tsou, 2011; Bertalmío, 2020). As a result, the same magnitude of contrast may appear vastly different under different circumstances, making it difficult to predict the perception of distortions. Finally, a significant challenge in VQA for HDR videos is the scarcity of high-quality subjective VQA databases. As HDR standards such as HDR10, HDR10+, and Dolby Vision are relatively recent, publicly available HDR content is limited, and few subjective VQA databases are dedicated to HDR. This database scarcity makes it difficult to accurately evaluate the quality of HDR videos and to develop and test new VQA models. Furthermore, several existing databases have limited utility as they have been rendered obsolete by recent HDR standards, while a few are unavailable publicly.

### 3.1.1 HDR databases

The earliest work in HDR subjective VQA was VPQM-MPEGHDR (Rerabek et al., 2015), which included a subjective study of five source HDR videos, each distorted by four compression levels, to compare objective HDR VQA algorithms. The videos were tone-mapped to 8 bits before displaying to the human subjects. The HDR-VQM Video dataset (Narwaria et al., 2015b) contains 10 HDR video sequences that were compressed at eight different bitrate using a backward-compatible compression method. The videos were tone mapped to an 8-bit SDR version and compressed before being displayed. When the videos were decompressed, they were inverse tone mapped to HDR. DML-HDR (Azimi et al., 2018) used 30 videos displayed on a non-standard HDR device supporting the older BT. 709 gamut. The dataset comprises indoor and outdoor video sequences with different brightness, motion levels, and representative distortions. Pan et al. (2018) conducted a study of the effects of compression on HDR quality using six source videos encoded using PQ and HLG, and BT.2020 color space, but used the AVS2 compression, which has seen little industry adoption. Waterloo UHD-HDR-WCG (Athar et al., 2019) conducted a subjective study of HDR10 content. This study includes 14 HDR10 source contents using H.264 and HEVC to generate 140 distorted videos. Recent works in HDR subjective VQA include LIVE-HDR (Shang et al., 2022a) and HDR Sports (Shang et al., 2023) databases. The LIVE-HDR database comprises 310 HDR10 videos with various compression and scaling distortions. The videos are displayed on an HDR TV under two different ambient light conditions. The study concluded that the ambient condition tested in the study has an insignificant effect on the perception of video quality. It is worth noting that the LIVE HDR Database is the only publicly available HDR VQA database that complies with contemporary HDR standards. The HDR Sports study conducted a subjective quality study with 42 source content to benchmark the performance of leading FR VQA models on common streaming problems, including compression, scaling, and quality crossovers among resolutions and frame rates. The study also attempted to reveal the effect of various encoding parameters, such as encoding mode and adaptive quantization. The salient features of the above-mentioned databases can be found in Table 2.

### 3.1.2 HDR VQA algorithms

The research on predicting the quality of High Dynamic Range (HDR) videos is still in its early stages, with few approaches proposed in the literature. HDR-VDP Mantiuk et al. (2005) is one of the earliest works in HDR objective VQA, which considers the non-linear response to light of high-contrast content and the full range of luminances. An improved version, HDR-VDP-2 (Mantiuk et al., 2011), uses a model based on contrast sensitivity measurements to account for all luminance conditions. Subsequent developments of HDR-VDP-2 include implementing improved pooling methods (HDR-VDP2.2 (Narwaria et al., 2015a; 2014)). Another approach to predicting the quality of HDR videos, proposed in Aydın et al. (2008), involves using a non-linear transformation to extend traditional Standard Dynamic Range (SDR) quality metrics to the HDR domain. This approach aims to make traditional SDR quality metrics more applicable to HDR videos by considering the expanded luminance range of HDR. In addition, some researchers have focused on the chromatic aspects of HDR video quality, such as color fidelity (Abebe et al., 2015), HDR Uniform Color Spaces (Rousselot et al., 2019), and color difference models (Choudhury et al., 2021). HDR-VQM (Narwaria et al., 2015b) utilizes spatiotemporal analysis to simulate human perception. The HDRMAX model (Ebenezer et al., 2023) uses a set of features designed by applying nonlinear transforms to enhance the estimation of the distortions in the brightest and darkest regions in the frames, which are often difficult to measure using traditional SDR-based metrics. These features improve the performance of state-of-the-art VQA models on HDR 10-bit videos.

To conclude, the research on HDR VQA is still in its early stages. Few approaches have been proposed in the literature. However, the scarcity of high-quality subjective VQA databases for HDR is a challenge that needs to be addressed to improve the accuracy and reliability of VQA models for HDR videos.

## 3.2 High frame rate (HFR) videos

In recent years, the integration of High Dynamic Range (HDR), 4K resolution, and High Frame Rate (HFR) videos has greatly improved the viewing experience for users. While the standard frame rate for movies is 24 fps and progressive television formats are typically 30 fps, HFR videos are displayed at 50 fps or more, effectively reducing temporal distortions such as motion blur and stutter. With improved communication technologies and the advent of powerful GPUs, consumers can now stream, share, and view content in high-resolution HFR format. It is crucial to develop VQA algorithms for HFR videos to enhance the storage and streaming efficiency of these videos, which will ultimately result in an improved user viewing experience.

### 3.2.1 High frame rate video quality databases

The BVI-HFR (Mackin et al., 2019) database is one of the earliest HFR VQA databases. It comprises 22 source sequences and their corresponding temporally down-sampled version

**TABLE 2** Summary of popular HDR VQA Databases. All databases are Type-1 as described in Section 2.1.

| Database | # Videos | # Pristine source sequences | Source characteristics | # Ratings per video | Public | Distortions | Duration | Display device |
|---|---|---|---|---|---|---|---|---|
| VPQM-MPEGHDR | 20 | 5 | 944 × 1080/24-60fps | 24 | No | HEVC, Tone Mapped to 8 bit for display | 15–40s | 47″ SIM2 HDR LCD TV |
| HDR-VQM | 90 | 10 | 1920 × 1,080/25fps | 25 | No | H.264/AVC, Tone Mapped to 8 bit for Compression Inverse Tone Mapped for display | N/A | 47″ SIM2 Solar HDR LCD TV |
| DML-HDR | 30 | 5 | 2048 × 1,080/30fps | 18 | Yes | HEVC, AWGN, Low Pass Filter, Salt-Pepper Noise, Mean Intensity Shift | 10 s | 40”” full HD LCD Prototype HDR Monitor |
| Pan et al. (2018) | 144 | 6 | 4K/50fps OETF: HLG and PQ, Color Gamut: BT.2020 | 22 | No | AVS2 | 10 s | Sony 30″ OLED 4K HDR TV, Sony 75″ LCD 4K HDR TV, LG 65″ OLED 4K HDR TV |
| Waterloo UHD-HDR-WCG | 140 | 14 | 4K/24–60 fps OETF: PQ Color Gamut: BT.2020 | 51 | No | H.264 and HEVC | 10 s | 31″ 4K HDR ReferenceMonitor |
| LIVE HDR | 310 | 31 | 4K/50-60fps OETF: PQ, Color Gamut: BT.2020 | 33 | Yes | HEVC and Scaling | 7–10 s | 65″ QLED 4K UHD HDR TV |
| HDR Sports | 1,002 | 42 | 1080p-4K/50fps OETF: Gamma and PQ, Color Gamut: BT.709 and BT.2020 | 30 | No | HEVC, Temporal Sampling | 6–9 s | Samsung 55″ 4K HDR TV, LG 55″ 4K HDR TV |

created by frame averaging, which can cause motion blur but eliminates the strobing artifacts due to frame dropping. Frame-averaged videos require fewer bits than frame-dropped videos since averaging results in a loss of edge sharpness, thus reducing high-frequency components. A mini-study by the authors found that frame-averaged videos were preferred over frame-dropped videos at very low frame rates. In an attempt to capture the effects of multiple distortions on video quality, the AVT-VQDB-UHD-1 (Rao et al., 2019) database was released. The authors conducted multiple tests, where Test #4 was specifically designed to understand the impact of frame rate on subjective video quality. Results from the study concluded that popular FR VQA models such as VMAF, performed poorly on videos with very low frame rates. Another work in the HFR VQA domain is LIVE YT-HFR (Madhusudana et al., 2020b), which includes 16 source videos with a native frame rate of 120 fps and temporally downsampled versions created by frame dropping to 24, 30, 60, 82, and 98 fps. The videos were distorted using the CRF (Constant Rate Factor) parameter in the VP9 encoder. The study found that the perceived visual quality of a video is significantly impacted by its frame rate and that the preferred frame rate is dependent on the content of the video. Participants showed a strong preference for higher frame rates in videos with excessive motion, and the authors observed that the effect of frame rate on subjective quality decreases beyond 60 fps. The ETRI LIVE STSVQ (Lee et al., 2021) database is a recent addition to the HFR VQA landscape, containing 437 videos that have undergone spatial and temporal downsampling and compression. Unlike the LIVE YT-HFR videos, the downsampled videos in this database were interpolated to match the frame rate of the original video before being displayed. The study reveals that reducing the bitrate budget and applying spatial and temporal downsampling can lead to better quality scores by minimizing compression artifacts by reducing the amount of data that needs to be compressed. However, the optimal balance between spatial and temporal downsampling depends on the content. For videos with fast and large movements, temporal downsampling can reduce quality at lower bitrate, so spatial downsampling alone may be more effective in limited bitrate budgets. Conversely, for videos with minimal or no motion but plenty of spatial detail, temporal downsampling is preferable at low bitrate compared to spatial downsampling. The salient features of the above-mentioned databases can be found in Table 3.

### 3.2.2 High frame rate VQA algorithms

FR VQA algorithms such as VMAF and SSIM are typically designed to compare videos with the same frame rate. Thus in the HFR VQA scenario, they are only applicable in assessing video quality after the SFR video has been temporally upsampled to match the HFR source video. However, these models often exhibit inconsistencies when compared to human subjective opinions as shown in (Madhusudana et al., 2020a). Thus, it is crucial to develop and employ algorithms that consider quality variations arising from changes in frame rate.

**TABLE 3** Summary of popular HFR VQA databases. All databases are Type-1 as described in Section 2.1.

| Database | #Videos | #Pristine source sequences | Source characteristics | #Ratings per video | Public | Distortions | Duration | Display device |
|---|---|---|---|---|---|---|---|---|
| BVI HFR | 88 | 22 | 4K/120fps | 51 | Yes | Temporal Downscaling: Frame Averaging | 10 s | 27″ LCD monitor, Refresh rate: 120 hz |
| AVT-VQDB-UHD-1 Test4 | 120 | 8 | 4K/60fps | 24 | Yes | Temporal Downscaling, H.264, Scaling | 8–10 s | N/A |
| LIVE YT HFR | 480 | 16 | 4K/120fps | 42 | Yes | Temporal Downscaling: Frame Dropping, VP9 Compression | 10 s | 27″ LCD Monitor, Refresh rate: 120 hz |
| ETRI LIVE STSVQ | 437 | 15 | 4K/120fps 10 bit | 34 | Yes | Temporal Downscaling: Frame Dropping, HEVC, Scaling | 4.5–7 s | 27″ LCD Monitor, Refresh rate: 120 hz |

Both Full-Reference and No Reference VQA algorithms are employed for evaluating videos with different frame rates. STGREED (Madhusudana et al., 2020a), and VSTR (Lee et al., 2020) are the top-performing FR VQA algorithms. STGREED, derived from the principles of natural video statistics, models the statistics of the temporal bandpass video coefficients as GGDs (Generalized Gaussian Distributions). The entropies of these coefficients are computed for both the distorted and source videos and then compared. It was observed that the entropy remains relatively constant for the same frame rate but varies across frame rates. An entropy bias term was introduced to remove the effect of frame rate bias when comparing videos with the same frame rate. Meanwhile, VSTR models the statistics of the most-Gaussian frame difference and then computes the entropic difference between the distorted and source sequence to quantify quality, where the most-Gaussian frame difference is the one with a minimum KL distance from a Normal Distribution. The authors establish that the statistics of pristine videos are highly Gaussian along directions of motion while unpredictable along other directions and for distorted videos.

Although FR-VQA algorithms perform well, there is a need for reliable NR-VQA algorithms to evaluate the quality of distorted videos without their pristine version. Framerate Aware Video Evaluator w/o Reference (FAVER) (Zheng et al., 2022) is an NR-VQA method developed for the HFR use-case, where the statistics of temporal and spatial bandpass videos are modeled as GGDs, and the generated features are used for quality analysis.

STGREED achieves state-of-the-art correlation scores against human judgments on LIVE YT HFR and ETRI LIVE STSVQ, with VSTR as a close competitor. Although FAVER achieves the highest performance, among other NR methods, on the LIVE YT HFR database with a correlation of 0.6 against human judgments, it is still a relatively low value. It highlights the challenges in HFR NR VQA. As HFR videos are adopted more by consumers, it is crucial to utilize perceptual analysis for optimal streaming and storage decisions. NR VQA models are especially vital for User Generated Content (UGC). The development of such models will allow the faster adoption of HFR videos.

## 3.3 Cloud gaming videos

Cloud gaming is another popular streaming video application where monitoring the video feed is essential to ensure a high-quality gaming experience. Cloud gaming services have grown in popularity within the digital gaming industry over the last decade. Several major technology companies have aggressively invested in cloud gaming infrastructure, including Meta Platforms, Google, Apple, Sony, NVIDIA, and Microsoft. With cloud gaming, users can play a wide range of games on their devices connected to the internet by viewing gameplay scenes as videos, while the compute-intensive game scene rendering process is performed on powerful cloud servers. Client devices like laptops, smartphones, and tablets capture users' interactions and transmit them to cloud servers. A simplified block diagram of an exemplar Cloud gaming setup is shown in Figure 2. This section discusses the recent advancements made toward accurately predicting Cloud Gaming Video Quality.

To begin, we discuss the challenges of assessing Video Quality for Cloud Gaming videos and why algorithms specifically designed for Cloud Gaming are necessary. The convenience of playing complex video games on a resource-constrained mobile device presents significant challenges due to engineering limitations in the Cloud Gaming video streaming pipeline. Video rendered on powerful Cloud Gaming servers is typical of very high quality requiring high bandwidth for streaming to the client's devices. Like other streaming platforms, a Cloud Gaming provider resizes and compresses the rendered gaming videos to ensure that videos can be transmitted over limited available bandwidth to multiple Cloud Gaming clients, resulting in reduced video quality. Another critical factor affecting gameplay in cloud gaming is network latency. As the video and control input data must be transmitted over the internet, high network latency can reduce the gameplay experience of the clients. Other factors affecting video quality include resolution, frame rate, and the display capabilities of the device playing the video. Additionally, complex games and a large number of players can also negatively impact video quality. It is also pertinent to note that generic VQA algorithms trained on natural scene databases have suboptimal performance when used to

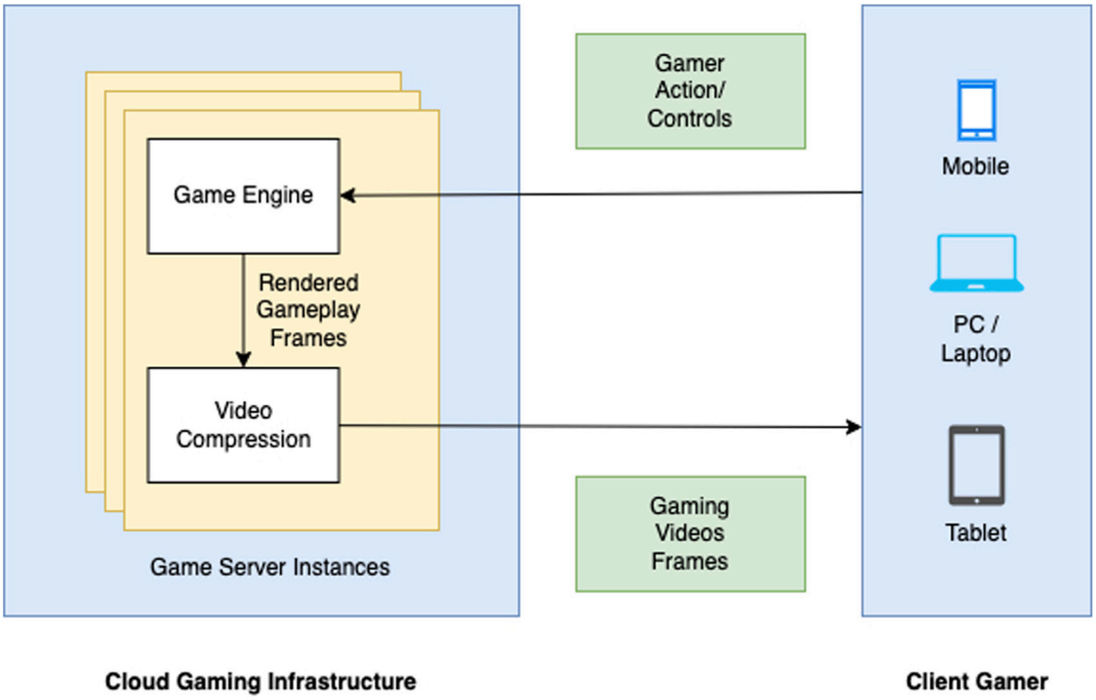

FIGURE 2
Example of a simplified cloud gaming system. The video frames of video games are rendered on service providers' cloud servers and sent over the Internet to end users' devices. Players' interactions are sent back to the cloud server over the same network.

estimate video quality in Cloud Gaming videos. This is due to the fact that rendered scenes of Cloud Gaming videos have very different underlying statistical properties from natural scenes. Cloud gaming also requires low latency and high frame rate in order to provide a smooth and responsive gaming experience. These requirements may not be adequately captured by traditional VQA algorithms. Thus, in order to provide seamless gameplay for Cloud Gaming clients and to reduce streaming costs for Cloud Gaming providers, the development of algorithms that accurately estimate Cloud Gaming video quality is crucial. The rest of this sub-section discusses subjective Video Quality databases developed for Cloud Gaming and the VQA algorithms used to accurately calculate Video Quality for Cloud Gaming.

### 3.3.1 Cloud gaming video quality databases

As part of the early works, two relatively small databases were developed. GamingVideoSET was introduced in Barman et al. (2018), while the work by Barman et al. (2019) resulted in the Kingston University Gaming Video Dataset (KUGVD). These databases, however, were severely limited in terms of the number of videos with subjective quality ratings and the variety of content. In both databases, 15 resolution-bitrate distortion pairs were created from each of the six source sequences, resulting in a total of only 90 human-annotated videos. The lack of abundant annotated data made the development of reliable Cloud Gaming-specific VQA algorithms difficult. A more comprehensive Cloud Gaming Video Dataset (CGVDS) database was developed by Zadtootaghaj et al. (2020b) to close this gap. With over 360 gaming videos collected from 15 source sequences, the CGVDS included subjective quality ratings, significantly increasing the number and variety of Cloud Gaming videos that could be used to develop reliable algorithms for VQA that were specific to Cloud Gaming use-cases. The three databases GamingVideoSET, KUGVD, and CGVDS involved rendering pristine gaming videos at 1080p resolution in the Cloud Gaming servers and applying predetermined synthetic distortions to generate lower-quality videos that were used in the subjective studies. Human subjects watched the gaming videos on monitors/TVs and provided quality ratings. The recent rise of Mobile Cloud Gaming in the past few years and the increasing number of Cloud Games in portrait mode necessitated the development of a video quality database focused on requirements of Mobile Cloud Gaming like including portrait videos in addition to landscape videos, 720p resolution pristine videos and the human ratings obtained with videos watched on small screen sized devices. Such a database would be necessary to develop accurate VQA algorithms focused on Mobile Cloud Gaming scenarios. In order to bridge the gap, the LIVE-Meta Mobile Cloud Gaming (LIVE-Meta MCG) database was introduced by Saha et al. (2023a). Using 20 pairs of resolution-bitrate distortions, 600 videos were obtained from 30 source sequences of resolution 720p. A Google Pixel 5 was used as the display device to display the videos, making it the only Cloud Gaming subjective video quality study conducted on a mobile device. Another notable resource on Cloud Gaming is the Tencent Gaming Video (TGV) dataset presented in Wen et al. (2021), however, the database is not publicly available. We summarize the characteristics of all the Cloud Gaming databases in Table 4.

TABLE 4 Summary of popular Cloud Gaming databases. All databases are Type-1 as described in Section 2.1.

| Database | # Videos | # Pristine source sequences | Source characteristics | # Ratings per video | Public | Distortions | Duration (sec) | Display device |
|---|---|---|---|---|---|---|---|---|
| GamingVideoSET | 90 | 6 | 1080p/30fps | 25 | Yes | H.264 Compression and Scaling | 30 | 24″ Monitor |
| KUGVD | 90 | 6 | 1080p/30fps | 17 | Yes | H.264 Compression and Scaling | 30 | 55″ Monitor |
| CGVDS | 360 + anchor stimuli | 15 | 1080p/30-60fps | Unavailable | Yes | H.264 NVENC Compression and Scaling | 30 | 24″ Monitor |
| TGV | 1,293 | 150 | 1080p/30fps | Unavailable | No | H.264, H.265, Tencent codec Compression and Scaling | 5 | Unknown Mobile Device |
| LIVE-Meta Mobile Cloud Gaming | 600 | 30 | 720p/30fps | 24 | Yes | H.264 NVENC Compression and Scaling | 20 | Google Pixel 5 |

### 3.3.2 Cloud gaming VQA algorithms

Cloud Gaming scenarios make it much more practical to measure Gaming Video Quality at the client without reference video. Therefore, No-Reference Video Quality Assessment algorithms are preferred over Full-Reference Video Quality algorithms for estimating video quality for Cloud Gaming workflows. Several generic No-Reference VQA algorithms trained on Cloud Gaming databases have been used. A few of these algorithms are BRISQUE, VIDEVAL, RAPIQUE, and VSFA. However, due to the distinct requirements of Cloud Gaming that distinguish it from generic video streaming, generic NR-VQA algorithms perform poorly when benchmarked on Cloud Gaming databases. Consequently, new NR-VQA algorithms focused on Cloud Gaming videos have been developed to improve the performance on cloud gaming databases. These algorithms are specially designed considering the Cloud Gaming requirements, resulting in enhanced performance. In Zadtootaghaj et al. (2018), the authors introduce NR-GVQM that trains an SVR model to evaluate the quality of gaming content videos by extracting nine frame-level features, using VMAF (Li Z. et al., 2016) scores as proxy ground-truth labels. The nine frame-level features include perceptually motivated and objective features. The perceptually motivated features include BRISQUE, NIQE, and BIQI (Moorthy and Bovik, 2010b), while spatial and temporal information, blockiness, blurriness, noise, and contrast constitute objective features. A closely related work to NR-GVQM is presented in "nofu" (Göring et al., 2019). It uses only the 360p center crop of each frame to speed up the computation of twelve frame-based features, followed by temporal pooling. More recent NR-VQA Cloud Gaming shows that deep learning-based models can boost performance for Cloud Gaming VQA task. These include NDNet-Gaming (Utke et al., 2020), DEMI (Zadtootaghaj et al., 2020a), and GAMIVAL (Chen et al., 2023). NDNet-Gaming and DEMI employ a complex Densenet-121 (Huang et al., 2016) deep learning backbone. NDNet-Gaming pre-trains the Densenet-121 with ground truth labels obtained from VMAF scores, then fine-tunes it with MOS scores due to the limited availability of subjective scores. The final step involves computing video quality predictions using a temporal pooling algorithm. The CNN architecture used in DEMI architecture is similar to that of NDNet-Gaming, but specifically addresses artifacts that include blockiness, blur, and jerkiness. GAMIVAL combines modified spatial and temporal natural scene statistic models and the pre-trained Densenet-121 backbone used in NDNet-Gaming, to predict gaming video quality. GAMIVAL's superior performance on LIVE-Meta MCG shows the benefit of using a dual path approach by deploying distortion-sensitive natural scene features on one side and content-aware deep features on the other.

## 3.4 Virtual reality

A rapidly growing technology, Virtual Reality (VR) allows users to interact in an immersive environment in a digital world. To ensure a satisfactory user experience, VR content must be evaluated rigorously as the market expands. Virtual Reality (VR) aims to create a sense of "presence," i.e., of being physically present in a virtual environment. Omnidirectional/VR videos are particularly well-suited for VR use cases. A VR video's wide field of view (FOV) allows viewers to view the scene from all directions in an immersive first-person perspective. Unlike regular videos viewed on flat screens, VR videos require a headset with a gyroscopic sensor to adjust the video to the viewer's head movements. In addition to factors such as resolution and compression level, VR VQA models must consider the larger FOV and viewing angle. When viewing VR content, the viewer's eyes are much closer to the headset screen than the typical viewing distance in the case of regular videos, causing a significant reduction in the number of pixels per angle. Spherical distortions can also occur when large FOV videos are projected onto planar surfaces to make it easier to encode and transmit. VR videos are modified differently to minimize transmission bandwidth and storage requirements, as shown in Figure 3. Foveated compression is one such method that exploits the fact that the human eye is more sensitive to detail in the center of the visual field (the fovea) than detail at the periphery by reducing the resolution of the regions non-central relative to the viewing direction. This achieves the reduction of the size of the video

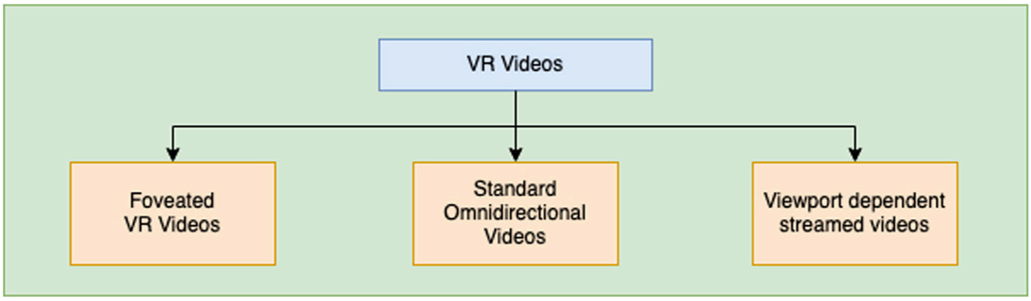
VR Videos
Foveated VR Videos
Standard Omnidirectional Videos
Viewport dependent streamed videos

FIGURE 3
Broad categories of VR Videos.

**TABLE 5 Summary of popular VR-VQA databases. All databases are Type-1 as described in Section 2.1.**

| Database | #Videos | #Pristine source sequences | Source characteristics | #Ratings per video | Public | Distortions | Duration | Display device |
|---|---|---|---|---|---|---|---|---|
| VR-VQA48 | 48 | 12 | 4,096 × 2048, 25 fps | 40 | Yes | 3 HEVC compression QPs. Also contains head tracking data | 12 s | HTC Vive |
| Zhang et al. (2018b) | 50 | 10 | 8K/4K, 30/60 fps | 30 | Yes | 5 HEVC compression QPs | 10 s | HTC Vive |
| IVQAD2017 | 160 | 10 | 4,096 × 2048, 30 fps | 13 | Yes | 3 resolutions, 3 frame rates and 3 bitrate | 15 s | HTC Vive |
| VRQ-TJU | 377 | 13 | 2560 × 2560, 30 fps | 30 | Yes | 4 JPEG2000 bit rates, 4 H264 QPs, stereo compression. Symmetric and asymmetric | 17 s | HTC Vive |
| Zhang et al. (2017) | 384 | 16 | 4,096 × 2048, 30 fps | 23 | Yes | 2 Noise levels, Gaussian blur, Box blur, H264 encoding (6 levels) | 10 s | HTC Vive |
| VQA-ODV | 540 | 60 | 4K–8K, 24–30 fps | ~22 | Yes | 3 map projections, 3 H265 QP levels, contains head and eye tracking data | 10–23 s | HTC Vive |
| VOD-VQA | 774 | 18 | 4K, 30 fps | ~20 | Yes | 4 viewport resolutions, 3 frame rates, 4 H264 compression QPs | 10 s | HTC Vive |
| Xie et al. (2020) | 1,608 | 30 | 3,840 × 2048, 30 fps | ~68 | Yes | 5 H264 QP levels, 3 viewport resolutions 10 viewport refinement durations | 10 s | HTC Vive |
| LIVE-FBT-FCVR 2D | 180 | 10 | 7680 × 3840, 30 fps | 36 | Yes | 3-region radially foveated samples built using random selection from 4 radii and 5 VP9 QPs | 10 s | HTC Vive |
| LIVE-FBT-FCVR 3D | 180 | 10 | 5376 × 5376, 30 fps, 3D | 34 | Yes | 3-region radially foveated samples built using random selection from 4 radii and 5 VP9 QPs | 10 s | HTC Vive |

while maintaining similar overall video quality. Another notable method is the viewport-based transmission. By dissecting the overall available FOV into smaller viewports, only the content within the viewport along the viewing direction is streamed in high quality. In contrast, lower quality is used elsewhere to reduce network bandwidth consumption. Quality refinement occurs after the user moves their focus to a new viewport. These factors complicate VR video quality assessment (VR VQA) and render the generic VQA algorithms ineffective for VR VQA.

### 3.4.1 VR video quality databases

Over the past few years, several subjectively rated VR videos have been released covering a plurality of distortions as listed in Table 5. The datasets cover various distortions like compression artifacts, foveation, viewport, projection, blur, and noise. VR-VQA48 dataset (Xu et al., 2019) has been the most popular dataset for benchmarking over the past few years. It contains VR videos distorted with HEVC compression with 3 QP levels. Zhang Y. et al. (2018) provides another dataset of VR videos with five quality levels of HEVC compression. Duan et al. (2017) provides the Immersive Video Quality Assessment Database 2017 (IVQAD 2017) with VR videos compressed with different resolutions, bit rates, and frame rates. VRQ-TJU dataset (Yang et al., 2018) provides videos by compressing stereo pairs of VR videos in both symmetric and asymmetric manner with a combination of four bit-rates and two compression algorithms. In addition to distorting VR videos with compression algorithms, Zhang et al. (2017) developed a VQA database by adding noise, Gaussian, and box blurs at various levels. VQA-Omnidirectional Video (VQA-ODV) database (Li et al., 2018) contains head and eye tracking data of subjects that were recorded as they evaluated videos consisting of compression and projection distortions. For viewport-dependent streamed videos, Viewpoint-based Omni-Directional Video Quality Assessment (VOD-VQA) (Meng and Ma, 2022) was introduced. It comprises videos with varying combinations of viewport resolutions, compression QPs, and refinement time. Xie et al. (2020) provide over 1,600 subjectively rated samples for the same use case. Foveation distortions are covered in the LIVE-Facebook Technologies-Foveated/Compressed VR (LIVE-FBT-FCVR) database (Jin et al., 2021a), which contains 2D and 3D videos. We summarize the VR VQA databases in Table 5.

### 3.4.2 Full-reference virtual reality video quality assessment (FR VR VQA)

The FR VR VQA algorithms range from simple modifications to the popular single image IQA metrics to deep neural network based metrics. The earliest FR VR VQA developed was Spherical PSNR (S-PSNR) (Yu et al., 2015). For computing S-PSNR, the panoramic ground truth, and distorted images are projected onto two unit spheres. The spheres are uniformly sampled, and the mean square error of signal values at those points is used for computing the S-PSNR. CPP-PSNR (Zakharchenko et al., 2016) proposed to first convert the images to Craster's Parabolic Projection (CPP) format before comparing the usual PSNR. Weighted-to-spherically-uniform PSNR (WS-PSNR) (Sun et al., 2017) uses squared differences of signal values that are weighted according to the corresponding mapped spherical area in observation space. Similar to WS-PSNR, spherical modification to the SSIM metric, S-SSIM, is proposed in Chen et al. (2018). Xu et al. (2019) proposed to incorporate the user's probable viewing direction and estimates the empirical distribution of viewing angle from subjective experiments and weigh each pixel accordingly to obtain non-content-based perceptual PSNR (NCP-PSNR). Further, the metric is improved to develop content-based perceptual PSNR (CP-PSNR), for which they estimate the viewing direction based on the content of the panoramic images. The viewing direction is estimated using a random forest model trained using image saliency heat maps. The same framework can be extended to develop CP-SSIM. Croci et al. (2020) proposed a framework to modify all available standard VQA metrics and adapt them to VR VQA use cases by including attention information. They propose decomposing the sphere into Voronoi patches of equal areas and averaging the VQA metric across each patch.

The methods mentioned to this point fail to account for temporal distortions as they all are modifications of single image quality estimation metrics. M2OVQA (Chai and Shao, 2022) is a statistical FR VR VQA metric that proposes a novel approach to combine features in the spatial domain for image content (SDIC), frequency domain for image content (FDIC), and frequency domain for video content (FDVC). The SDIC features include contrast invariant phase congruent structures, features extracted using filters inspired by the human visual system, and color-sensitive similarity features. The FDIC feature is computed by measuring the similarity between the power spectral density of the distorted and the reference frame. Similarly, the FDVC feature is the similarity of power spectral densities of frame differences of distorted and original videos. SDIC, FDIC, and FDVC metrics are combined for each viewport to get a consolidated quality score. Lastly, the quality scores for each viewport are summed with normalized weights proportional to their saliency to obtain the unified quality metric.

Next, we discuss a few deep learning-based models for FR VR VQA. Li et al. (2018) proposed a deep learning-based FR VR VQA method that leverages head and eye movement data captured along with the VQA-ODV dataset. The difference between the reference and distorted videos is fed as the input to the network along with the distorted frame. The distorted and error frames are sampled into n patches using the head motion heatmap as a probability distribution. Each patch is processed through a deep learning network trained on the VQA-ODV dataset that estimates the quality of the patch to obtain 'n' local quality scores. Next, the scores are combined by weighting them to the normalized eye motion heatmaps. V-CNN (Li C. et al., 2019) is another popular FR VR VQA model that first determines the likeliness of viewports based on the frame's contents and uses it to weigh and sum the quality of each viewport. The viewport prediction network takes in the current frame and the current frame difference. It employs spherical convolutions to output a spherical heatmap of the importance of viewports, followed by the shortlisting of the most important viewports. A softening filter follows this and is applied to merge the proposed viewports that are too close to each other. The selected viewports are used for the quality estimation step. Each viewport of the distorted frame and its difference from the reference viewport are fed to a deep neural network of stacked dense blocks to estimate the local quality score. The overall VQA score is computed by averaging the local scores according to viewport importance weights. 3D-360 VQA (Guo et al., 2022) is another deep learning-based method that consists of a deep neural network containing 3D convolutions to better capture the temporal structure of the video. They propose a viewport projection method to reduce the spherical distortion when the spherical frame is projected on a 2D viewport. The deep neural network inputs the distorted and reference frames and outputs the local quality score. The local scores of each patch are averaged to output the overall quality score of the VR video.

### 3.4.3 No reference virtual reality video quality assessment (NR VR VQA)

Compared to FR VR VQA methods, NR VR VQA methods are limited. The method proposed by Zhang et al. (2021b) involves computing frame-level features in the spherical domain before pooling them temporally to obtain video-level features. The frame-level features include sharpness, blockiness, blurriness, and spatial and temporal information. Each feature is pooled temporally by calculating the root mean squares. Polynomial kernels with four bases are used to fit the features to the quality scores using multiple kernel learning (MKL) regression.

In Wu et al. (2019), a deep learning-based NR VR VQA method is proposed that leverages 3D convolutions. The network takes ten consecutive 128 × 128 RGB patches as input and is trained on data collected by the authors. The dataset contains seven panoramic videos distorted with 20 variations, including two projections, two compression algorithms, and 5 bitrate. The scores of different patches are averaged in a positionally dependent manner to obtain the final quality score of the panoramic video. Another 3D convolution-based deep learning method is proposed by Yang et al. (2018) that operates on stereo pairs of the panoramic video. The difference between the two stereo image patches of sizes 32 × 32 from 10 consecutive frames is fed as input to the network. This method allows for estimating quality when the distortions of the pair of images are not the same. The scores of the patches are then averaged positionally to obtain the cumulative score for the VR video. Yang et al. (2021) proposed another deep learning network based on 3D spherical convolutions (Cohen et al., 2018) for NR VR VQA. The input to the network is three stereo difference frames (sampled at an interval of frames) resized to 1280 × 1280 pixels. This network employs spherical convolutional kernels to directly operate in the spherical domain and an attention-based non-local block to better capture long-ranged non-local structures in the video.

### 3.4.4 Foveated VR VQA

The assessment of foveated content requires sophisticated tools and is more complicated than that of typical VR videos. The VR VQA methods mentioned earlier work well for content with consistent resolution and detail. However, they do not account for the reduced perception of content in peripheral vision due to the cone sensors' exponential decline. Thus, it is essential to determine how the strength and gradient of the applied foveation affect the perceived quality of foveated content to develop accurate quality assessment models. There exist only a few FR-IQA models for foveated images like Foveated Wavelet Quality Index (FWQI) (Wang et al., 2001), foveated PSNR (FPSNR) and foveated weighted SNR (FWSNR) (Lee et al., 2002), Foveation-based Content Adaptive SSIM (FA-SSIM) (Rimac-Drlje et al., 2011). Foveated Entropic Differencing (FED) (Jin et al., 2021c) is a recently developed FR foveated VQA algorithm which employs the natural scene statistics of bandpass responses by applying differences of local entropies weighted by a foveation-based error sensitivity function.

Spatially Varying BRISQUE (SV-BRISQUE) (Jin et al., 2021b), is an NR foveated VQA tool that proposes a spatially-varying version of natural scene statistics (NSS) (Mittal et al., 2012) and natural video statistic (NVS) features for estimating the video quality. These parameters are assumed to be stationary over concentric regions and computed for "K" concentric regions. Finally, the extracted features are used for training a support vector regressor (SVR) to map to the MOS scores. Foveated Video Quality Assessment (FOVQA) (Jin et al., 2022) improves the SV-BRISQUE algorithm by considering the gradient of the quality fall-off in the radial direction by building upon the space-varying NSS and NVS features to provide a more accurate model for NR foveated VQA. After obtaining space-variant NSS and NVS feature maps as in Jin et al. (2021b), each parameter map is weighted and aggregated using a set of "K" toroidal Gaussian functions with different eccentricities and variance to obtain a set of radial basis features. In addition, based on mean ranked opinion scores, it was observed that sudden spatial increases in compression could be quite noticeable, especially on moving content. In order to measure this, a model that has been trained separately is utilized to analyze the input video and generate a QP map that estimates the compression level in each specific region. The fall-off gradient is then calculated by applying Gaussian derivatives that have been smoothed radially to the QP map. The gradient map is obtained by convolving the predicted QP map with the Gaussian derivatives. The gradient maps are combined using toroidal Gaussian, similar to the process of processing NSS feature maps, to create radial basis derivative features. These features, along with radial basis NSS and NVS features, are used to train an SVR to predict video quality. FOVQA surpasses all traditional VQA algorithms and SV-BRISQUE for NR-foveated VR VQA.

### 3.4.5 VQA for viewport adaptive streaming

Quality assessment of videos streamed in a viewport adaptive manner has to account for the changes in quantization parameter (QP) and spatial resolution (SR) to the refinement duration (RD) when switching from an arbitrary Low Quality (LQ) scale to an arbitrary High Quality (HQ) one. To this end, a couple of specialized VQA models have been developed. Meng and Ma (2022) formulate an analytical model to connect the perceptual quality of a compressed viewport video with the triplet of variables mentioned above. The QP, SR, and RD variables are estimated using linearly weighted content features. VOD-VQA dataset is created and used for developing this model. Xie et al. (2020) also construct a dataset of subjective quality scores on VR videos by modifying the QP, SR, and RD parameters and using it to build an analytical model for perceptual quality. The video quality is modeled as a product of separable exponential functions that measure the QP and SR-induced perceptual impacts in terms of the RD and a perceptual index measuring the subjective quality of the corresponding viewport video after refinement.

## 3.5 Quality of experience (QoE)

Content creators and distributors have traditionally focused on video quality, but Quality of Experience (QoE) has emerged as a key performance indicator as users' experience transcends technical parameters. Initially, the focus was on improving technical aspects of video quality, such as resolution, frame rate, and compression. With the shift from traditional broadcasting to online streaming buffering, latency, startup delay, playback stalling events, and playback continuity challenges emerged as equally significant factors. In response, content creators and distributors began paying more attention to the quality of experience, user interface design, and user engagement. Machine learning and artificial intelligence algorithms are increasingly used to improve QoE and make it more personalized. Today, QoE is a crucial metric for content creators, service providers, and advertisers. By ensuring a high QoE, they can increase engagement, reduce churn, and ultimately increase revenues.

**TABLE 6** Summary of popular QoE databases. All databases are Type-1 as described in Section 2.1.

| Database | #Videos | #Pristine source sequences | Source characteristics | #Ratings per video | Public | Distortions | Duration | Display device |
|---|---|---|---|---|---|---|---|---|
| LIVE Mobile VQA Database | 200 | 10 | 2K/30fps, 60fps | 27 | Yes | Compression rate, packet-loss, frame-freezes,<br>H.264 Compression | 15 s | Smartphone (960 × 540), Tablet (1,280 × 800) |
| HTTP-based Video Streaming Database | 15 | 3 | 720p/30fps | 25 | Yes | Bitrate,<br>H.264 Compression | 300 s | 58 inch HDTV |
| LIVE Mobile Stall Video Database-I | 180 | 24 | 720p, 360p | 27 | Yes | Frequency and length of stalls, length of initial delays,<br>H.264 Compression | 29–134 s | Apple iPhone 5 |
| LIVE Mobile Stall Video Database-II | 174 | 24 | 720p, 360p/30fps | 27 | Yes | Start-up delay length, stall lengths, stall positions, and the number of stalls,<br>H.264 Compression | 29–134 s | Laptop-size monitor |
| WaterlooSQoE-I | 200 | 20 | 1080p/24-30fps | 25 | Yes | Initial buffering, stalling and compression rate,<br>H.264 Compression | 10 s | LCD monitor (2,560 × 1,600) |
| LIVE-NFLX-I | 112 | 14 | 1080p/24,<br>25 and 30f | 19 | Yes | Compression rate, rebuffering events,<br>H.264 Compression | About 60 s | 5.1 inch Samsung S5 |
| WaterlooSQoE-II | 168, 588 | 12 | 1080p/30fps | 35 | Yes | Compression rate, spatial resolution, frame rate, H.264 Compression | 4 s, 8 s | LCD monitor (1920 × 1,080) |
| WaterlooSQoE-III | 450 | 20 | 1080p/24-30fps | 34 | Yes | initial buffering, stalling, bitrate switching<br>(13 network traces and 6 ABR algorithms),<br>H.264 Compression | 13 s | LCD monitor (1920 × 1,080) |
| LIVE-NFLX-II | 420 | 15 | 1080p/30fps | More than 22 | Yes | initial buffering, stalling, bitrate switching<br>(7 network traces and 4 ABR algorithms),<br>H.264 Compression | 25 s | 24 inch LCD monitor (1920 × 1,080) |
| WaterlooSQoE-IV | 1,350 | 5 | 4K/24 and 30 fps | 33 for Phone, 32 for HDTV, 32 for UHDTV | Yes | initial buffering, stalling, bitrate switching<br>(9 network traces and 5 ABR algorithms),<br>H.264 and HEVC Compression | 30 s | Phone (5.8 inch Apple iPhone XS Max),<br>HDTV (24 inch View SonicVA2452SM),<br>UHDTV (55 inch Sony XBR55 × 800H) |

Dynamic adaptive streaming over HTTP (DASH) provides a technology that adapts to the viewer's bandwidth and processing capabilities, allowing the video to be streamed efficiently without requiring the viewer to adjust the video quality manually. In this work, we limit our focus to VQA databases and algorithms tailored to analyze common challenges in video streaming, such as buffering, latency, startup delay, playback stalling events, and playback continuity. We do not study other effects encompassing QoE like immersion, as discussed in Perkis et al. (2020).

### 3.5.1 QoE databases

Similar to traditional VQA, subjective QoE databases serve as the benchmarks for objective QoE assessment models. A summary of the popular QoE databases can be found in Table 6. The early QoE databases are all of Type 1 (as discussed in 2.1), comprising synthetically QoE distorted videos from a handful of pristine videos. The LIVE Mobile VQA database (Moorthy et al., 2012) comprises 200 distorted videos manually created from 10 HD original videos with distortion types including compression, packet-loss, and temporal dynamics, like compression rates and frame-freezes. A small database of 18 distorted videos was proposed in a study by Chen et al. (2014), which created 15 quality-varying videos of relatively long duration by varying encoding bitrate. The LIVE Mobile Stall Video Database-I (Ghadiyaram et al., 2014) and LIVE Mobile Stall Video Database-II (Ghadiyaram et al., 2019) are two with simulated patterns of stalling events in length, position, and frequency of occurrence. WaterlooSQoE-I (Duanmu et al., 2017b) is another QoE video database containing 200 video sequences focusing on the combined effect of the network artifacts, like compression, startup delay, and stalling events occurring at the beginning or the middle point of the video sequence, while WaterlooSQoE-II (Duanmu et al., 2017a) contains 588 videos with varying levels of compression, spatial resolution, and frame rates. LIVE-NFLX-I (Bampis et al., 2017b) comprises 112 distorted videos generated using H.264 encoder with eight payout patterns, including different compression rates and rebuffering events.

Next, we discuss Type-2 databases that contain realistic QoE distortions and include real-world network traces to record actual network changes. WaterlooSQoE-III (Duanmu et al., 2018) studies QoE of streaming videos transmitted with six different adaptive bitrate streaming algorithms: BBA (Huang et al., 2014), AIMD (Liu et al., 2011), ELASTIC (De Cicco et al., 2013), QDASH (Mok et al., 2012), and FESTIVE (Jiang et al., 2014) with 13 bandwidth profiles. LIVE-NFLX-II (Bampis et al., 2021) consists of 420 adaptive streaming videos focusing on low bandwidth conditions with actual network measurements and a pragmatic client buffer simulator. WaterlooSQoE-IV (Duanmu et al., 2020) contains 1,350 streaming videos generated from various source video material, video codecs, network setups, adaptive bitrate (ABR) algorithms, and viewing screens. These subjective quality databases can be used to develop objective QoE algorithms for adaptive video streaming, enabling researchers to develop objective methods more closely aligned with human visual perception.

### 3.5.2 QoE evaluation algorithms

A common QoE evaluation model is a hybrid model that combines QoS-driven user QoE evaluation and FR-VQA-based signal fidelity measurements to quantify human visual perception. Some common representative VQA models used for QoE evaluation are SSIM, MS-SSIM (Wang et al., 2003), ST-RRED, and VMAF. The other part of a typical QOE evaluation model is the QoE assessment model. Streaming Quality Index (SQI) (Duanmu et al., 2017b) combines video presentation quality, which applies FR-IQA algorithms like SSIM, MS-SSIM, and rebuffering information. Each stalling event divides the streaming session into three parts, including the time interval before, during, and after the stall. The QoE loss during the stalling event and the decline of memory retention after the stall event is approximated with an exponential decay function and Hermann Ebbinghaus forgetting curves. The instantaneous QoE drop due to stall events is computed by aggregating the QoE drop caused by each stall event.

QoE assessment algorithm proposed in Singh et al. (2012) involves using a random Neural Network to estimate QoE scores based on inputs of QP value and rebuffering-related features. Video assessment of temporal artifacts and stalls (Video ATLAS) (Bampis and Bovik, 2017) is a machine learning framework that combines several QoE-related features, including objective quality features, rebuffering-aware features, and memory-driven features, to make QoE predictions. Unlike Video ATLAS, which can only provide overall QoE scores, Bampis et al. (2018b) suggested a range of recurrent dynamic neural networks that carry out continuous-time QoE prediction. It incorporated VQA scores, playback status, and memory data to predict QoE scores. A time-series forecasting ensemble aggregating two or more continuous QoE forecasts were used to provide more reliable prediction performance. Deep learning models have demonstrated excellent performance in recent years in comprehending the growth of human sensory cortex processing. D-DASH (Gadaleta et al., 2017) is a framework that integrated deep learning and reinforcement learning methods to improve the quality of dynamic adaptive streaming over HTTP (DASH). DeepQoE (Zhang H. et al., 2018), an end-to-end QoE prediction framework, uses features obtained from a deep neural network trained on classification or regression tasks. In Tao et al. (2019), a data-driven strategy was proposed to predict QoE scores by developing a novel deep neural network (DNN) approach that analyzes the correlation between mobile video transmission network parameters and subjective QoE scores. In Huang et al. (2022), a model-assisted deep learning technique was employed to predict channel route loss, which was then used to predict video streaming MOS. Yan et al. (2019) conducted a comparative analysis of multiple ABR algorithms on the QoE using data from the deployed live TV streaming website Puffer 1. In addition, an enhanced ABR algorithm was proposed, trained using the collected data, that helped achieve superior video quality while minimizing time spent on video stalls during streaming.

## 4 Conclusion and future work

Video Quality Assessment is a highly challenging and significant issue within the realm of Video Engineering and has garnered significant attention. Recent advances in deep learning and learning theory, as well as the emergence of newer video technologies such as High Dynamic Range and High Frame Rate videos, the growth of Cloud Gaming and VR/AR applications, and the advent of hardware-accelerated video compression tools, have significantly impacted the evolution of VQA. This paper provides a comprehensive survey of the development of VQA over the last 2 decades, tracing its journey from the introduction of perceptual image quality assessment metrics such as SSIM and VIF in 2003–04 to the modern-day VQA algorithms using deep learning, which has expanded beyond traditional videos to encompass contemporary video applications. In conclusion, this survey culminates in discussing the most recent and upcoming trends in VQA and our perspective on how the VQA domain will evolve in the next decade.

**Deep Learning** models have already achieved remarkable progress in most sub-domains in VQA, surpassing traditional methods. The superior performance of Deep Learning models may be attributed to their ability to automatically learn high-level representations from raw data, thereby minimizing the need for manual feature engineering. As the field of VQA continues to make strides in deep learning, it is anticipated that the growth of VQA databases will expand in quantity and scale. These more extensive databases will provide a richer and more comprehensive data source for training, testing, and benchmarking deep learning models, enabling them to attain greater accuracy and reliability. We also foresee considerable work involving neural network architecture design innovations using CNNs and Transformers similar to the recent work MUSIQ (Ke et al., 2021) that are specifically suited for IQA/VQA tasks.

**Vision Science** models have continued to dominate VQA (Mantiuk et al., 2021) even with the deep learning revolution. We hypothesize that vision science-based models will continue to play a crucial role in further advancing VQA by providing a more comprehensive understanding of the perceptual processes involved in human visual perception. By leveraging insights from vision science through a better understanding of human visual systems, researchers can develop models that more accurately simulate human visual processing and attention mechanisms, which will further improve the VQA system's accuracy and reliability. Vision Science models offer another significant advantage of being inherently explainable, as they are developed based on the workings of the human visual system. This makes them particularly valuable in developing robust VQA systems, as they avoid the uncertainties associated with deep learning-based black box models. By providing clear explanations for how VQA models arrive at their decisions, Vision Science models can enhance the interpretability and trustworthiness of VQA systems, making them more useful in industrial deployment applications.

**Representation Learning** has emerged as another promising area for advancing VQA, as it addresses the current limitations of labeled data in IQA/VQA. The success of Unsupervised feature learning for high-level vision tasks has propelled the development of representation learning for IQA/VQA (Madhusudana et al., 2022a; b; Saha et al., 2023b). We anticipate that the trend towards developing representation learning frameworks from the perspective of low-level vision tasks such as VQA will continue to gain momentum. This approach has the potential to significantly improve the accuracy and reliability of VQA systems and may lead to breakthroughs in Video Engineering.

**Emerging Domains** such as HDR, HFR, Cloud Gaming, and VR/AR are expected to grow steadily, closely following the increasing wireless internet speeds enabled by the introduction of 5G and 6G technologies. As these technologies become more widely adopted by consumers, newer and innovative applications of VQA will emerge, creating more exciting opportunities for future VQA research and development.

**Generative AI** has revolutionized machine learning research recently. The success of Generative AI can be attributed to evolving deep learning techniques and the availability of large-scale datasets. These advances have enabled researchers to create previously impossible, realistic, and detailed images and videos. As a result, generative AI has opened up new avenues of research and innovation across various fields, including VQA. As more Generative AI models gradually progress from research into industrial workflows, technical quality control of the generated images and videos will be crucial. We hypothesize that current VQA algorithms will not be able to predict the technical quality of images and videos obtained from Generative AI workflows, potentially opening up the scope for VQA research focused on these directions.

## Author contributions

AS: Introduction, Cloud Gaming, Conclusion; SP: Virtual Reality ZS: HDR; RP and HG: Generic Video Quality Assessment; BC: Quality of Experience; SM: HFR AB: Overall organization and review. All authors contributed equally to the article and approved the submitted version.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Abebe, M. A., Pouli, T., and Kervec, J. (2015). Evaluating the color fidelity of itmos and hr color appearance models. *ACM Trans. Appl. Percept.* 12, 1–16. doi:10.1145/2808232

Athar, S., Costa, T., Zeng, K., and Wang, Z. (2019). "Perceptual quality assessment of UHD-HDR-WCG videos," in *Ieee int. Conf. Image process.*, 1740–1744.

Aydın, T. O., Mantiuk, R., and Seidel, H.-P. (2008). "Extending quality metrics to full luminance range images," in *Human vision and electronic imaging XIII*. Editors B. E. Rogowitz and T. N. Pappas (International Society for Optics and Photonics (SPIE), 6806, 109–118. doi:10.1117/12.765095

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

Bampis, C. G., and Bovik, A. C. (2017). *Learning to predict streaming video qoe: Distortions, rebuffering and memory. CoRR abs/1703.00633*.

Bampis, C. G., Gupta, P., Soundararajan, R., and Bovik, A. C. (2017a). Speed-qua: Spatial efficient entropic differencing for image and video quality. *IEEE Signal Process. Lett.* 24, 1333–1337. doi:10.1109/LSP.2017.2726542

Bampis, C. G., Li, Z., and Bovik, A. C. (2018a). Spatiotemporal feature integration and model fusion for full reference video quality assessment. *IEEE Trans. Circuits Syst. Video Technol.* 29, 2256–2270. doi:10.1109/tcsvt.2018.2868262

Bampis, C. G., Li, Z., Katsavounidis, I., and Bovik, A. C. (2018b). Recurrent and dynamic models for predicting streaming video quality of experience. *IEEE Trans. Image Process.* 27, 3316–3331. doi:10.1109/tip.2018.2815842

Bampis, C. G., Li, Z., Katsavounidis, I., Huang, T.-Y., Ekanadham, C., and Bovik, A. C. (2021). Towards perceptually optimized adaptive video streaming-a realistic quality of experience database. *IEEE Trans. Image Process.* 30, 5182–5197. doi:10.1109/TIP.2021.3073294

Bampis, C. G., Li, Z., Moorthy, A. K., Katsavounidis, I., Aaron, A., and Bovik, A. C. (2017b). Study of temporal effects on subjective video quality of experience. *IEEE Trans. Image Process.* 26, 5217–5231. doi:10.1109/TIP.2017.2729891

Barman, N., Jammeh, E., Ghorashi, S. A., and Martini, M. G. (2019). No-reference video quality estimation based on machine learning for passive gaming video streaming applications. *IEEE Access* 7, 74511–74527. doi:10.1109/ACCESS.2019.2920477

Barman, N., Zadtootaghaj, S., Schmidt, S., Martini, M. G., and Möller, S. (2018). "Gamingvideoset: A dataset for gaming video streaming applications," in *2018 16th annual workshop on network and systems support for games (NetGames)*, 1–6.

Bertalmío, M. (2020). "Chapter 5 - brightness perception and encoding curves," in *Vision models for high dynamic range and wide colour gamut imaging*. Editor M. Bertalmío (Academic Press), 95–129. Computer Vision and Pattern Recognition. doi:10.1016/B978-0-12-813894-6.00010-7

Billock, V. A., and Tsou, B. H. (2011). To honor fechner and obey stevens: Relationships between psychophysical and neural nonlinearities. *Psychol. Bull.* 137, 1–18. doi:10.1037/a0021394

Bross, B., Wang, Y.-K., Ye, Y., Liu, S., Chen, J., Sullivan, G. J., et al. (2021). Overview of the versatile video coding (vvc) standard and its applications. *IEEE Trans. Circuits Syst. Video Technol.* 31, 3736–3764. doi:10.1109/TCSVT.2021.3101953

Chai, X., and Shao, F. (2022). M2ovqa: Multi-space signal characterization and multi-channel information aggregation for quality assessment of compressed omnidirectional videos. *J. Vis. Comun. Image Represent* 82, 103419. doi:10.1016/j.jvcir.2021.103419

Chen, C., Choi, L. K., de Veciana, G., Caramanis, C., Heath, R. W., and Bovik, A. C. (2014). Modeling the time—Varying subjective quality of http video streams with rate adaptations. *IEEE Trans. Image Process.* 23, 2206–2221. doi:10.1109/TIP.2014.2312613

Chen, P., Li, L., Ma, L., Wu, J., and Shi, G. (2020). "Rirnet: Recurrent-in-recurrent network for video quality assessment," in *Proceedings of the 28th ACM international conference on multimedia*, 834–842.

Chen, S., Zhang, Y., Li, Y., Chen, Z., and Wang, Z. (2018). "Spherical structural similarity index for objective omnidirectional video quality assessment," in *2018 IEEE international conference on multimedia and expo (ICME)*, 1–6. doi:10.1109/ICME.2018.8486584

Chen, Y.-C., Saha, A., Davis, C., Qiu, B., Wang, X., Gowda, R., et al. (2023). Gamival: Video quality prediction on mobile cloud gaming content. *IEEE Signal Process. Lett.* 1, 324–328. doi:10.1109/LSP.2023.3255011

Chikkerur, S., Sundaram, V., Reisslein, M., and Karam, L. J. (2011). Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. Broadcast.* 57, 165–182. doi:10.1109/tbc.2011.2104671

Choudhury, A., Wanat, R., Pytlarz, J., and Daly, S. (2021). Image quality evaluation for high dynamic range and wide color gamut applications using visual spatial processing of color differences. *Color Res. Appl.* 46, 46–64. doi:10.1002/col.22588

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. (2018). *Spherical cnns. CoRR abs/1801.10130.*

Cornsweet, T. N., and Pinsker, H. (1965). Luminance discrimination of brief flashes under various conditions of adaptation. *J. Physiology* 176, 294–310. doi:10.1113/jphysiol.1965.sp007551

Croci, S., Ozcinar, C., Zerman, E., Knorr, S., Cabrera, J., and Smolic, A. (2020). Visual attention-aware quality estimation framework for omnidirectional video using spherical voronoi diagram. *Qual. User Exp.* 5, 4. doi:10.1007/s41233-020-00032-3

De Cicco, L., Caldaralo, V., Palmisano, V., and Mascolo, S. (2013). "Elastic: A client-side controller for dynamic adaptive streaming over http (dash)," in *2013 20th international packet video workshop*, 1–8. doi:10.1109/PV.2013.6691442

De Simone, F., Naccari, M., Tagliasacchi, M., Dufaux, F., Tubaro, S., and Ebrahimi, T. (2009). "Subjective assessment of h.264/avc video sequences transmitted over a noisy channel," in *2009 international workshop on quality of multimedia experience*, 204–209. doi:10.1109/QOMEX.2009.5246952

Duan, H., Zhai, G., Yang, X., Li, D., and Zhu, W. (2017). "Ivqad 2017: An immersive video quality assessment database," in *2017 international conference on systems, signals and image processing (IWSSIP)*, 1–5. doi:10.1109/IWSSIP.2017.7965610

Duanmu, Z., Liu, W., Li, Z., Chen, D., Wang, Z., Wang, Y., et al. (2020). *Assessing the quality-of-experience of adaptive bitrate video streaming.*

Duanmu, Z., Ma, K., and Wang, Z. (2017a). *Quality-of-experience of adaptive video streaming: Exploring the space of adaptations*, 1752–1760. doi:10.1145/3123266.3123418

Duanmu, Z., Rehman, A., and Wang, Z. (2018). A quality-of-experience database for adaptive video streaming. *IEEE Trans. Broadcast.* 64, 474–487. doi:10.1109/TBC.2018.2822870

Duanmu, Z., Zeng, K., Ma, K., Rehman, A., and Wang, Z. (2017b). A quality-of-experience index for streaming video. *IEEE J. Sel. Top. Signal Process.* 11, 154–166. doi:10.1109/JSTSP.2016.2608329

Ebenezer, J. P., Shang, Z., Wu, Y., Wei, H., and Bovik, A. C. (2023). Making video quality assessment models robust to bit depth. *SPL* 30, 488–492. doi:10.1109/lsp.2023.3268602

Ebenezer, J. P., Shang, Z., Wu, Y., Wei, H., Sethuraman, S., and Bovik, A. C. (2021). Chipqa: No-Reference video quality prediction via space-time chips. *IEEE Trans. Image Process.* 30, 8059–8074. doi:10.1109/tip.2021.3112055

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gadaleta, M., Chiariotti, F., Rossi, M., and Zanella, A. (2017). D-Dash: A deep q-learning framework for dash video streaming. *IEEE Trans. Cognitive Commun. Netw.* 3, 703–718. doi:10.1109/TCCN.2017.2755007

Gao, F., Wang, Y., Li, P., Tan, M., Yu, J., and Zhu, Y. (2017). Deepsim: Deep similarity for image quality assessment. *Neurocomputing* 257, 104–114. doi:10.1016/j.neucom.2017.01.054

Ghadiyaram, D., Bovik, A. C., Yeganeh, H., Kordasiewicz, R., and Gallant, M. (2014). "Study of the effects of stalling events on the quality of experience of mobile streaming videos," in *2014 IEEE global conference on signal and information processing (GlobalSIP)*, 989–993. doi:10.1109/GlobalSIP.2014.7032269

Ghadiyaram, D., Pan, J., and Bovik, A. C. (2019). A subjective and objective study of stalling events in mobile streaming videos. *IEEE Trans. Circuits Syst. Video Technol.* 29, 183–197. doi:10.1109/TCSVT.2017.2768542

Ghadiyaram, D., Pan, J., Bovik, A. C., Moorthy, A. K., Panda, P., and Yang, K.-C. (2018). In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Trans. Circuits Syst. Video Technol.* 28, 2061–2077. doi:10.1109/TCSVT.2017.2707479

Göring, S., Ramachandra Rao, R. R., and Raake, A. (2019). *Nofu -a lightweight no-reference pixel based video quality model for gaming content*. doi:10.1109/QoMEX.2019.8743262

Götz-Hahn, F., Hosu, V., Lin, H., and Saupe, D. (2021). Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. *IEEE Access* 9, 72139–72160. doi:10.1109/access.2021.3077642

Guo, J., Huang, L., and Chien, W.-C. (2022). Multi-viewport based 3d convolutional neural network for 360-degree video quality assessment. *Multimedia Tools Appl.* 81, 16813–16831. doi:10.1007/s11042-022-12073-1

Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., et al. (2017). "The konstanz natural video database (konvid-1k)," in *2017 ninth international conference on quality of multimedia experience (QoMEX)*, 1–6. doi:10.1109/QoMEX.2017.7965673

Huang, G., Ercetin, O., Gokcesu, H., and Kalem, G. (2022). "Deep learning-based qoe prediction for streaming services in mobile networks," in *2022 18th international conference on wireless and mobile computing, networking and communications (WiMob)*, 327–332. doi:10.1109/WiMob55322.2022.9941672

Huang, G., Liu, Z., and Weinberger, K. Q. (2016). *Densely connected convolutional networks. CoRR abs/1608.06993.*

Huang, T.-Y., Johari, R., McKeown, N., Trunnell, M., and Watson, M. (2014). "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," in *ACM SIGCOMM computer communication review*. doi:10.1145/2619239.2626296

ITU (2011). *BT.1886: Reference electro-optical transfer function for flat panel displays used in HDTV studio production. Tech. rep., Intl. Telecomm. Union.*

ITU (2015). *BT.2020: Parameter values for ultra-high definition television systems for production and international programme exchange.*

ITU (2018). *BT.2100: Image parameter values for high dynamic range television for use in production and international programme exchange. Tech. rep, Intl. Telecomm. Union.*

ITU-R (2012). *Methodology for the subjective assessment of the quality of television pictures document ITU-R recommendation BT 500-13 2012.*

ITU-T (2022). *Subjective video quality assessment methods for multimedia applications Document ITU-T Recommendation P.910, 2022.*

Jiang, J., Sekar, V., and Zhang, H. (2014). Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive. *IEEE/ACM Trans. Netw.* 22, 326–340. doi:10.1109/TNET.2013.2291681

Jin, Y., Chen, M., Goodall, T., Patney, A., and Bovik, A. C. (2021a). Subjective and objective quality assessment of 2d and 3d foveated video compression in virtual reality. *IEEE Trans. Image Process.* 30, 5905–5919. doi:10.1109/TIP.2021.3087322

Jin, Y., Goodall, T., Patney, A., Webb, R., and Bovik, A. C. (2021b). "A foveated video quality assessment model using space-variant natural scene statistics," in *2021 IEEE international conference on image processing* (ICIP), 1419–1423. doi:10.1109/ICIP42928.2021.9506032

Jin, Y., Patney, A., and Bovik, A. (2021c). *Evaluating foveated video quality using entropic differencing*. doi:10.48550/ARXIV.2106.06817

Jin, Y., Patney, A., Webb, R., and Bovik, A. C. (2022). Fovqa: Blind foveated video quality assessment. *IEEE Trans. Image Process.* 31, 4571–4584. doi:10.1109/TIP.2022.3185738

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. (2021). *Musiq: Multi-scale image quality transformer*. doi:10.48550/ARXIV.2108.05997

Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. speech, signal Process.* 29, 1153–1160. doi:10.1109/tassp.1981.1163711

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* 54, 1–41. doi:10.1145/3505244

Kim, W., Kim, J., Ahn, S., Kim, J., and Lee, S. (2018). "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proceedings of the European conference on computer vision* (ECCV), 219–234.

Korhonen, J. (2019). Two level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.* 28, 5923–5938. doi:10.1109/TIP.2019.2923051

Larson, E., and Chandler, D. (2010). Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* 19, 011006. doi:10.1117/1.3267105

Lee, D. Y., Ko, H., Kim, J., and Bovik, A. C. (2020). "Video quality model for space-time resolution adaptation," in *2020 IEEE 4th international conference on image processing, applications and systems (IPAS)*, 34–39. doi:10.1109/IPAS50080.2020.9334940

Lee, D. Y., Paul, S., Bampis, C. G., Ko, H., Kim, J., Jeong, S. Y., et al. (2021). *A subjective and objective study of space-time subsampled video quality. arXiv preprint arXiv:2102.00088.*

Lee, S., Pattichis, M. S., and Bovik, A. C. (2002). Foveated video quality assessment. *IEEE Trans. Multim.* 4, 129–132. doi:10.1109/6046.985561

Li, C., Xu, M., Du, X., and Wang, Z. (2018). "Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *Proceedings of the 26th ACM international conference on multimedia* (New York, NY, USA: Association for Computing Machinery), MM '18), 932–940. doi:10.1145/3240508.3240581

Li, C., Xu, M., Jiang, L., Zhang, S., and Tao, X. (2019a). "Viewpoint proposal cnn for 360° video quality assessment," in *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 10169–10178. doi:10.1109/CVPR.2019.01042

Li, D., Jiang, T., and Jiang, M. (2019b). "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM international conference on multimedia*, 2351–2359.

Li, S., Zhang, F., Ma, L., and Ngan, K. N. (2011). Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Trans. Multimedia* 13, 935–949. doi:10.1109/TMM.2011.2152382

Li, X., Guo, Q., and Lu, X. (2016a). Spatiotemporal statistics for video quality assessment. *IEEE Trans. Image Process.* 25, 3329–3342. doi:10.1109/tip.2016.2568752

Li, Y., Feng, L., Xu, J., Zhang, T., Liao, Y., and Li, J. (2021). "Full-reference and no-reference quality assessment for compressed user-generated content videos," in *2021 IEEE international conference on multimedia and expo workshops (ICMEW)* (IEEE), 1–6.

Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A., and Manohara, M. (2016b). Toward a practical perceptual video quality metric. *Netflix Tech Blog* 6. doi:10.1117/12.320105

Li, Z., and Bampis, C. G. (2017). "Recover subjective quality scores from noisy measurements," in *2017 data compression conference (DCC)*, 52–61. doi:10.1109/DCC.2017.26

Li, Z., Duanmu, Z., Liu, W., and Wang, Z. (2019c). "Avc, hevc, vp9, avs2 or av1? — A comparative study of state-of-the-art video encoders on 4k videos," in *Image analysis and recognition*. Editors F. Karray, A. Campilho, and A. Yu (Cham: Springer International Publishing), 162–173.

Lin, W., and Kuo, C.-C. J. (2011). Perceptual visual quality metrics: A survey. *J. Vis. Commun. image Represent.* 22, 297–312. doi:10.1016/j.jvcir.2011.01.005

Liu, C., Bouazizi, I., and Gabbouj, M. (2011). "Rate adaptation for adaptive http streaming," in *Proc. ACM conf. On multimedia systems*. doi:10.1145/1943552.1943575

Liu, H., and Heynderickx, I. (2011). Visual attention in objective image quality assessment: Based on eye-tracking data. *Circuits Syst. Video Technol. IEEE Trans.*21, 971–982. doi:10.1109/TCSVT.2011.2133770

Mackin, A., Zhang, F., and Bull, D. R. (2019). A study of high frame rate video formats. *IEEE Trans. Multimedia* 21, 1499–1512. doi:10.1109/TMM.2018.2880603

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. (2022a). *Conviqt: Contrastive video quality estimator*. doi:10.48550/ARXIV.2206.14713

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. (2022b). Image quality assessment using contrastive learning. *IEEE Trans. Image Process.* 31, 4149–4161. doi:10.1109/tip.2022.3181496

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. (2020a). *ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction. CoRR abs/2010.13715*.

Madhusudana, P. C., Yu, X., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. (2020b). *Subjective and objective quality assessment of high frame rate videos. CoRR abs/2007.11634.*

Manasa, K., and Channappayya, S. S. (2016). An optical flow-based full reference video quality assessment algorithm. *IEEE Trans. Image Process.* 25, 2480–2492. doi:10.1109/tip.2016.2548247

Mantiuk, R., Daly, S. J., Myszkowski, K., and Seidel, H.-P. (2005). "Predicting visible differences in high dynamic range images: Model and its calibration," in *Human vision and electronic imaging X*. Editors B. E. Rogowitz, T. N. Pappas, and S. J. Daly (International Society for Optics and Photonics (SPIE), 5666, 204–214. doi:10.1117/12.586757

Mantiuk, R. K., Denes, G., Chapiro, A., Kaplanyan, A., Rufo, G., Bachy, R., et al. (2021). Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Trans. Graph.* 40, 1–19. doi:10.1145/3450626.3459831

Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. (2011). Hr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.* 30, 1–14. doi:10.1145/2010324.1964935

Meng, Y., and Ma, Z. (2022). Viewpoint-based omnidirectional video quality assessment: Database, modeling and inference. *IEEE Trans. Circuits Syst. Video Technol.* 32, 120–134. doi:10.1109/TCSVT.2021.3057368

Min, X., Zhai, G., Zhou, J., Farias, M. C. Q., and Bovik, A. C. (2020). Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans. Image Process.* 29, 6054–6068. doi:10.1109/TIP.2020.2988148

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Trans. image Process.* 21, 4695–4708. doi:10.1109/tip.2012.2214050

Mittal, A., Saad, M. A., and Bovik, A. C. (2015). A completely blind video integrity oracle. *IEEE Trans. Image Process.* 25, 289–300. doi:10.1109/tip.2015.2502725

Mittal, A., Soundararajan, R., and Bovik, A. C. (2012c). Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* 20, 209–212. doi:10.1109/lsp.2012.2227726

Mohammadi, P., Ebrahimi-Moghadam, A., and Shirani, S. (2014). *Subjective and objective quality assessment of image: A survey. arXiv preprint arXiv:1406.7799.*

Mok, R., Luo, X., Chan, E., and Chang, R. (2012). *Quash: A qoe-aware dash system*, 11–22. doi:10.1145/2155555.2155558

Moorthy, A. K., and Bovik, A. C. (2010b). A two-step framework for constructing blind image quality indices. *IEEE Signal Process. Lett.* 17, 513–516. doi:10.1109/LSP.2010.2043888

Moorthy, A. K., and Bovik, A. C. (2010a). Efficient video quality assessment along temporal trajectories. *IEEE Trans. circuits Syst. video Technol.* 20, 1653–1658. doi:10.1109/tcsvt.2010.2087470

Moorthy, A. K., and Bovik, A. C. (2011). Visual quality assessment algorithms: What does the future hold? *Multimedia Tools Appl.* 51, 675–696. doi:10.1007/s11042-010-0640-x

Moorthy, A. K., Choi, L. K., Bovik, A. C., and de Veciana, G. (2012). Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE J. Sel. Top. Signal Process.* 6, 652–671. doi:10.1109/JSTSP.2012.2212417

Narwaria, M., Mantiuk, R., Silva, M. P. D., and Callet, P. L. (2015a). HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images. *J. Electron. Imaging* 24, 1–3. doi:10.1117/1.JEI.24.1.010501

Narwaria, M., Perreira Da Silva, M., and Le Callet, P. (2015b). HDR-VQM: An objective quality measure for high dynamic range video. *Signal Process. Image Commun.* 35, 46–60. doi:10.1016/j.image.2015.04.009

Narwaria, M., Silva, M. P. D., Callet, P. L., and Pepion, R. (2014). "On improving the pooling in HDR-VDP-2 towards better HDR perceptual quality assessment," in *Human vision and electronic imaging XIX*. Editors B. E. Rogowitz, T. N. Pappas, and H. de Ridder (International Society for Optics and Photonics (SPIE), 9014, 143–151. doi:10.1117/12.2045436

Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P., and Häkkinen, J. (2016). Cvd2014—A database for evaluating no-reference video quality assessment algorithms. *IEEE Trans. Image Process.* 25, 3073–3086. doi:10.1109/TIP.2016.2562513

Pan, X., Zhang, J., Wang, S., Wang, S., Zhou, Y., Ding, W., et al. (2018). HDR video quality assessment: Perceptual evaluation of compressed HDR video. *J. Vis. Comm. Image Rep.* 57, 76–83. doi:10.1016/j.jvcir.2018.10.016

Perkis, A., Timmerer, C., Baraković, S., Husić, J. B., Bech, S., Bosse, S., et al. (2020). *Quaint white paper on definitions of immersive media experience (imex).*

Pr-Newswire (2023). *Sandvine's 2023 global internet phenomena report shows 24% jump in video traffic, with Netflix volume overtaking YouTube [online; accessed 4th-march-2023].*

Rao, R. R. R., Göring, S., Robitza, W., Feiten, B., and Raake, A. (2019). "Avt-vqdb-uh-1: A large scale video quality database for uh-1," in *2019 ieee ism*, 1–8.

Rerabek, M., Hanhart, P., Korshunov, P., and Ebrahimi, T. (2015). "Subjective and objective evaluation of HDR video compression," in *9th intl. Workshop video process. Qual. Metrics consum. Electron. (VPQM).*

Rimac-Drlje, S., Martinović, G., and Zovko-Cihlar, B. (2011). "Foveation-based content adaptive structural similarity index," in *2011 18th international conference on systems, signals and image processing*, 1–4.

Rousselot, M., Le Meur, O., Cozot, R., and Ducloux, X. (2019). Quality assessment of hr/wcg images using hr uniform color spaces. *J. Imaging* 5, 18. doi:10.3390/jimaging5010018

Saad, M. A., Bovik, A. C., and Charrier, C. (2014). Blind prediction of natural video quality. *IEEE Trans. Image Process.* 23, 1352–1365. doi:10.1109/tip.2014.2299154

Saha, A., Chen, Y.-C., Davis, C., Qiu, B., Wang, X., Gowda, R., et al. (2023a). Study of subjective and objective quality assessment of mobile cloud gaming videos. *IEEE Trans. Image Process.* 1, 3295–3310. doi:10.1109/TIP.2023.3281170

Saha, A., Mishra, S., and Bovik, A. C. (2023b). "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 5846–5855.

Seshadrinathan, K., and Bovik, A. C. (2009). Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. image Process.* 19, 335–350. doi:10.1109/tip.2009.2034992

Seshadrinathan, K., and Bovik, A. C. (2011). "Temporal hysteresis model of time varying subjective video quality," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 1153–1156.

Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K. (2010). Study of subjective and objective quality assessment of video. *Trans. Img. Proc.* 19, 1427–1441. doi:10.1109/TIP.2010.2042111

Shahid, M., Rossholm, A., Lövström, B., and Zepernick, H.-J. (2014). No-Reference image and video quality assessment: A classification and review of recent approaches. *EURASIP J. image Video Process.* 2014, 40–32. doi:10.1186/1687-5281-2014-40

Shang, Z., Chen, Y., Wu, Y., Wei, H., and Sethuraman, S. (2023). "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* (IEEE), 556–564.

Shang, Z., Ebenezer, J. P., Bovik, A. C., Wu, Y., Wei, H., and Sethuraman, S. (2022a). "Subjective assessment of high dynamic range videos under different ambient conditions," in *2022 IEEE international conference on image processing (ICIP)* (IEEE), 786–790. doi:10.1109/ICIP46576.2022.9897940

Shang, Z., Ebenezer, J. P., Wu, Y., Wei, H., Sethuraman, S., and Bovik, A. C. (2022b). Study of the subjective and objective quality of high motion live streaming videos. *IEEE Trans. Image Process.* 31, 1027–1041. doi:10.1109/TIP.2021.3136723

Sheikh, H., Bovik, A., and de Veciana, G. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Process.* 14, 2117–2128. doi:10.1109/TIP.2005.859389

Singh, K. D., Hadjadj-Aoul, Y., and Rubino, G. (2012). "Quality of experience estimation for adaptive," in *2012 IEEE consumer communications and networking conference (CCNC)*, 127–131. doi:10.1109/CCNC.2012.6181070

Sinno, Z., and Bovik, A. C. (2019). Large-scale study of perceptual video quality. *IEEE Trans. Image Process.* 28, 612–627. doi:10.1109/TIP.2018.2869673

Soundararajan, R., and Bovik, A. (2013). Video quality assessment by reduced reference spatio-temporal entropic differencing. *Circuits Syst. Video Technol. IEEE Trans.*23, 684–694. doi:10.1109/TCSVT.2012.2214933

Sun, W., Min, X., Zhai, G., Gu, K., Duan, H., and Ma, S. (2019). Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment. *IEEE J. Sel. Top. Signal Process.* 14, 64–77. doi:10.1109/jstsp.2019.2955024

Sun, Y., Lu, A., and Yu, L. (2017). Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Process. Lett.* 24, 1–1412. doi:10.1109/LSP.2017.2720693

Tao, X., Duan, Y., Xu, M., Meng, Z., and Lu, J. (2019). Learning qoe of mobile video transmission with deep neural network: A data-driven approach. *IEEE J. Sel. Areas Commun.* 37, 1337–1348. doi:10.1109/JSAC.2019.2904359

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., et al. (2016). YFCC100m. *Commun. ACM* 59, 64–73. doi:10.1145/2812802

Tu, Z., Chen, C.-J., Chen, L.-H., Birkbeck, N., Adsumilli, B., and Bovik, A. C. (2020). "A comparative evaluation of temporal pooling methods for blind video quality assessment," in *2020 IEEE international conference on image processing (ICIP)* (IEEE), 141–145.

Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., and Bovik, A. C. (2021a). Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Trans. Image Process.* 30, 4449–4464. doi:10.1109/tip.2021.3072221

Tu, Z., Yu, X., Wang, Y., Birkbeck, N., Adsumilli, B., and Bovik, A. C. (2021b). Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open J. Signal Process.* 2, 425–440. doi:10.1109/ojsp.2021.3090333

Utke, M., Zadtootaghaj, S., Schmidt, S., Bosse, S., and Moeller, S. (2020). "NDNetGaming - development of a No-reference deep CNN for gaming video quality prediction," in *Multimedia tools and applications* (Springer).

Venkataramanan, A. K., Stejerean, C., and Bovik, A. C. (2022). *Funque: Fusion of unified quality evaluators*. doi:10.48550/ARXIV.2202.11241

VQEG-HDTV VQE (2010). *Report on the validation of video quality models for high definition video content [online; accessed 30-january-2022]*.

Vu, P. V., Vu, C. T., and Chandler, D. M. (2011). "A spatiotemporal most-apparent-distortion model for video quality assessment," in *2011 18th IEEE international conference on image processing* (IEEE), 2505–2508.

Wang, H., Gan, W., Hu, S., Lin, J. Y., Jin, L., Song, L., et al. (2016). "Mcl-jcv: A jnd-based h.264/avc video quality assessment dataset," in *2016 IEEE international conference on image processing (ICIP)*, 1509–1513. doi:10.1109/ICIP.2016.7532610

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., et al. (2018). Temporal segment networks for action recognition in videos. *IEEE Trans. pattern analysis Mach. Intell.* 41, 2740–2755. doi:10.1109/tpami.2018.2868668

Wang, Y., Inguva, S., and Adsumilli, B. (2019). "Youtube ugc dataset for video compression research," in *2019 IEEE 21st international workshop on multimedia signal processing (MMSP)*.

Wang, Z., Bovik, A. C., Lu, L., and Kouloheris, J. L. (2001). "Foveated wavelet image quality index," in *Applications of digital image processing XXIV*. Editor A. G. Tescher (International Society for Optics and Photonics (SPIE), 4472, 42–52. doi:10.1117/12.449797

Wang, Z., and Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE signal Process. Mag.* 26, 98–117. doi:10.1109/msp.2008.930649

Wang, Z., and Bovik, A. C. (2011). Reduced-and no-reference image quality assessment. *IEEE Signal Process. Mag.* 28, 29–40. doi:10.1109/msp.2011.942471

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Trans. image Process.* 13, 600–612. doi:10.1109/tip.2003.819861

Wang, Z., Simoncelli, E., and Bovik, A. (2003). "Multiscale structural similarity for image quality assessment," in *The thrity-seventh asilomar conference on signals, systems and computers, 2003*, 2, 1398–1402. doi:10.1109/ACSSC.2003.1292216

Wen, S., Ling, S., Wang, J., Chen, X., Fang, L., Jing, Y., et al. (2021). *Subjective and objective quality assessment of mobile gaming video. ArXiv* abs/2103.05099.

Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., et al. (2022a). "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *Computer vision–ECCV 2022: 17th European conference, tel aviv, Israel, october 23–27, 2022, proceedings, Part VI* (Springer), 538–554.

Wu, H., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., et al. (2022b). *Disentangling aesthetic and technical effects for video quality assessment of user generated content. arXiv preprint arXiv:2211.04894.*

Wu, P., Ding, W., You, Z., and An, P. (2019). "Virtual reality video quality assessment based on 3d convolutional neural networks," in *2019 IEEE international conference on image processing (ICIP)*, 3187–3191. doi:10.1109/ICIP.2019.8803023

Xie, S., Xu, Y., Shen, Q., Ma, Z., and Zhang, W. (2020). Modeling the perceptual quality of viewport adaptive omnidirectional video streaming. *IEEE Trans. Circuits Syst. Video Technol.* 30, 3029–3042. doi:10.1109/TCSVT.2019.2934136

Xu, J., Ye, P., Liu, Y., and Doermann, D. (2014). "No-reference video quality assessment via feature learning," in *2014 IEEE international conference on image processing (ICIP)* (IEEE), 491–495.

Xu, M., Chen, J., Wang, H., Liu, S., Li, G., and Bai, Z. (2020). "C3dvqa: Full-reference video quality assessment with 3d convolutional neural network," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 4447–4451.

Xu, M., Li, C., Chen, Z., Wang, Z., and Guan, Z. (2019). Assessing visual quality of omnidirectional videos. *IEEE Trans. Circuits Syst. Video Technol.* 29, 3516–3530. doi:10.1109/TCSVT.2018.2886277

Yan, F. Y., Ayers, H., Zhu, C., Fouladi, S., Hong, J., Zhang, K., et al. (2019). *Continual learning improves internet video streaming. CoRR abs/1906.01113.*

Yang, J., Liu, T., Jiang, B., Lu, W., and Meng, Q. (2021). Panoramic video quality assessment based on non-local spherical cnn. *IEEE Trans. Multimedia* 23, 797–809. doi:10.1109/TMM.2020.2990075

Yang, J., Liu, T., Jiang, B., Song, H., and Lu, W. (2018). 3d panoramic virtual reality video quality assessment based on 3d convolutional neural networks. *IEEE Access* 6, 38669–38682. doi:10.1109/ACCESS.2018.2854922

Ying, Z., Mandal, M., Ghadiyaram, D., and Bovik, A. C. (2020). *Patch-vq: 'patching up' the video quality problem. CoRR abs/2011.13544.*

Ying, Z., Mandal, M., Ghadiyaram, D., and Bovik, A. (2021). "Patch-vq:'patching up'the video quality problem," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14019–14029.

You, J., Ebrahimi, T., and Perkis, A. (2013). Attention driven foveated video quality assessment. *IEEE Trans. Image Process.* 23, 200–213. doi:10.1109/TIP.2013.2287611

Yu, M., Lakshman, H., and Girod, B. (2015). "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE international symposium on mixed and augmented reality*, 31–36. doi:10.1109/ISMAR.2015.12

Yu, X., Birkbeck, N., Wang, Y., Bampis, C. G., Adsumilli, B., and Bovik, A. C. (2021). Predicting the quality of compressed videos with pre-existing distortions. *IEEE Trans. Image Process.* 30, 7511–7526. doi:10.1109/TIP.2021.3107213

Zadtootaghaj, S., Barman, N., Rao, R. R. R., Göring, S., Martini, M. G., Raake, A., et al. (2020a). "Demi: Deep video quality estimation model using perceptual video quality dimensions," in *2020 IEEE 22nd international workshop on multimedia signal processing (MMSP)*, 1–6. doi:10.1109/MMSP48831.2020.9287080

Zadtootaghaj, S., Barman, N., Schmidt, S., Martini, M. G., and Möller, S. (2018). "Nr-gvqm: A no reference gaming video quality metric," in *2018 IEEE international symposium on multimedia (ISM)*, 131–134.

Zadtootaghaj, S., Schmidt, S., Sabet, S. S., Möller, S., and Griwodz, C. (2020b). "Quality estimation models for gaming video streaming services using perceptual video quality dimensions," in *Proceedings of the 11th ACM multimedia systems conference* (New York, NY, USA: Association for Computing Machinery), 213–224. MMSys '20. doi:10.1145/3339825.3391872

Zakharchenko, V., Choi, K. P., and Park, J. H. (2016). "Quality metric for spherical panoramic video," in *Optics and photonics for information processing X*. Editors K. M. Iftekharuddin, A. A. S. Awwal, M. G. Vázquez, A. Márquez, and M. A. Matin (International Society for Optics and Photonics (SPIE), 9970, 99700C. doi:10.1117/12.2235885

Zhang, A.-X., Wang, Y.-G., Tang, W., Li, L., and Kwong, S. (2022). *Hvs revisited: A comprehensive video quality assessment framework. arXiv preprint arXiv:2210.04158.*

Zhang, B., Zhao, J., Yang, S., Zhang, Y., Wang, J., and Fei, Z. (2017). "Subjective and objective quality assessment of panoramic videos in virtual reality environments," in *2017 IEEE international conference on multimedia and expo workshops (ICMEW)*, 163–168. doi:10.1109/ICMEW.2017.8026226

Zhang, H., Hu, H., Gao, G., Wen, Y., and Guan, K. (2018a). "Deepqoe: A unified framework for learning to predict video qoe," in *2018 IEEE international conference on multimedia and expo (ICME)*, 1–6. doi:10.1109/ICME.2018.8486523

Zhang, Y., He, L., Lu, W., Li, J., and Gao, X. (2021a). Video quality assessment with dense features and ranking pooling. *Neurocomputing* 457, 242–253. doi:10.1016/j.neucom.2021.06.026

Zhang, Y., Liu, Z., Chen, Z., Xu, X., and Liu, S. (2021b). "No-reference quality assessment of panoramic video based on spherical-domain features," in *2021 picture coding symposium (PCS)*, 1–5. doi:10.1109/PCS50896.2021.9477498

Zhang, Y., Wang, Y., Liu, F., Liu, Z., Li, Y., Yang, D., et al. (2018b). Subjective panoramic video quality assessment database for coding applications. *IEEE Trans. Broadcast.* 64, 461–473. doi:10.1109/TBC.2018.2811627

Zheng, Q., Tu, Z., Madhusudana, P. C., Zeng, X., Bovik, A. C., and Fan, Y. (2022). *Faver: Blind quality prediction of variable frame rate videos.*

Zhou, W., Min, X., Li, H., and Jiang, Q. (2022). A brief survey on adaptive video streaming quality assessment. *J. Vis. Commun. Image Represent.* 86, 103526. doi:10.1016/j.jvcir.2022.103526