Subjective and Objective Analysis of Streamed Gaming Videos

Xiangxu Yu, Zhenqiang Ying, Neil Birkbeck, Yilin Wang, Balu Adumilli, and Alan C. Bovik, Fellow, IEEE

Abstract—The rising popularity of online user-generated-content (UGC) in the form of streamed and shared videos has hastened the development of perceptual video quality assessment (VQA) models, which can be used to help optimize their delivery. Gaming videos, which are a relatively new type of UGC videos, are created when skilled and casual gamers post videos of their gameplay. These kinds of screenshots of UGC gameplay videos have become extremely popular on major streaming platforms, such as YouTube and Twitch. Synthetically generated gaming content presents challenges to existing VQA algorithms, including those based on natural scene/video statistics models. Synthetically generated gaming content presents different statistical behavior than naturalistic videos. A number of studies have been directed toward understanding the perceptual characteristics of professionally generated gaming videos arising in gaming video streaming, online gaming, and cloud gaming. However, little work has been done on understanding the quality of UGC gaming videos, and how it can be characterized and predicted. Toward boosting the progress of gaming video VQA model development, we conducted a comprehensive study of subjective and objective VQA models on UGC gaming videos. To do this, we created a novel UGC gaming video resource, called the LIVE-YouTube Gaming video quality database, comprised of 600 real UGC gaming videos. We conducted a subjective human study on this data, yielding 18,600 human quality ratings recorded by 61 human subjects. We also evaluate a number of state-of-the-art VQA models on the new database, including a new one, called GAME-VQP, based on both natural video statistics and CNN-learned features. To help support work in this field, we are making the new LIVE-YouTube Gaming Database, along with code for GAME-VQP, publicly available through the link: https://live.ece.utexas.edu/research/LIVE-YT-Gaming/index.html.

Index Terms—Database, gaming video, multimedia, user-generated content (UGC), video quality assessment (VQA).

I. INTRODUCTION

O

VER the past few years, Internet traffic has continued to grow dramatically. According to Cisco Visual Networking Index [1], video streaming now occupies the majority of Internet bandwidth and is the main driver of this growth. Streaming service providers, such as Netflix, Hulu, and Amazon Prime Video, generate and stream a substantial portion of this visual traffic in the form of professionally generated content (PGC) videos. At the same time, online video sharing platforms, such as YouTube, Vimeo, and Twitch, collect and make available vast numbers of user-generated-content (UGC) videos, generally uploaded by less skilled videographers. Driven by the rapid development of the digital game industry and by consumer desire to see skilled gamers in action, a large number of online videos of gameplay content have been uploaded and made public on the Internet. In 2020, YouTube Gaming reached a milestone of 100 billion hours of watch time and 40 million active gaming channels [2]. The rapid expansion of the gaming market has also given birth to many streaming-related services, such as game live broadcasts, online games, and cloud games. At the same time, increasingly powerful computing, communication, and display hardware and software technologies have continuously uplifted the quality of these services. In order to provide users with better quality streaming gaming video services and to improve users’ viewing experiences, perceptual video quality assessment (VQA) research has become particularly important. Both general-purpose algorithms able to handle diverse distortion scenarios, as well as VQA models specifically designed for gaming videos need to be studied and developed.

VQA research efforts can generally be divided into two categories: conducting subjective quality assessment experiments, and developing objective quality assessment models and algorithms. If they were feasible, subjective quality assessment would best meet the needs of streaming providers and users, but the cost is prohibitive and does not allow for real-time quality-based decisions or processing steps, such as perceptually optimized compression. Therefore, automatic perception-based objective video quality prediction algorithms are of great interest. Objective VQA models can be classified into three categories, based on whether there is a reference video available. The first, full-reference (FR) models require the availability of an entire source video as a reference against which to measure visual differences, i.e., video fidelity. Reduced-reference (RR)
models only require partial reference information, often only a small amount. No-reference (NR) or “blind” models predict the quality of a test video without using any reference information. It is generally believed that when predicting the quality of the test video, if there is an available PGC video as a reference, that is, an original, “pristine” video of high quality, shot by experts with professional equipment, then FR or RR VQA algorithms are able to achieve superior prediction results, as compared with NR models. However, there is often a lack of accessible reference videos, e.g., for UGC videos captured by naïve users having limited capture skills or technology. These can only be evaluated using NR VQA models. Since these kinds of videos are quite pervasive on social media and sharing platforms, the NR VQA problem is of high practical importance. Since they also typically encompass much wider ranges of quality and of distortion types which are often simultaneously present and mixed, the NR VQA problem is much more challenging.

A. Motivation

Unlike real-world videos taken by videographers with optical cameras, gaming videos are generally synthetically generated. Because of underlying statistical differences in the structure of synthetic and artificial videos [3, 4], popular NR VQA models may not achieve adequate quality prediction performance on them.

Research on the assessment of gaming video quality mostly commenced very recently. Barman et al. [3] studied and analyzed differences between gaming and nongaming content videos, in terms of compression and user quality of experience (QoE). Some VQA models designed for gaming videos have been proposed, e.g., [5, 6]. However, this work has been limited to the study of the characteristics of PGC gaming videos and how they are affected by compression. None of these have discussed or analyzed the quality assessment of UGC gaming videos recorded by amateur users. We briefly elaborate on the characteristics of PGC and UGC gaming videos in Section II-D.

Advancements of research on video quality have always relied on freely available databases of distorted videos and subjective judgments of them. Over the past two decades, many VQA databases have been created and shared with researchers, which has significantly driven the development of VQA models. Databases created in the early stages of VQA research usually contained only a small number of PGC reference videos and a small number of synthetically distorted versions of them. In recent years, as UGC VQA research has gained more attention, more advanced and specialized databases have been created, containing hundreds or thousands of unique UGC videos with numerous, complex authentic distortions. Recently, databases dedicated to the development of gaming VQA research have also been proposed. However, these databases only contain a limited number of PGC gaming videos as references, along with compressed versions of them. Therefore, toward advancing progress on understanding streamed UGC gaming video quality, we have created a new UGC gaming video resource, which we call the LIVE-YouTube Gaming video quality (LIVE-YT-Gaming) database. This new dataset contains subjective rating scores on a large number of unique gaming videos from a study that was conducted online.

B. Contribution

We summarize the contributions we make as follows.

1) Using this new resource, we conducted a novel online human study whereby we collected a large number of subjective quality labels on gaming videos. By contrast, with previous studies that have extensively utilized laboratory and crowdsourcing environments, our study required participants to use their personal devices and log in to our custom-designed website. Specific details regarding this methodology can be found in Section IV of this article. To demonstrate the utility of the new database, we compared and contrasted the performance of a wide range of state-of-the-art (SOTA) VQA models, including one of our own design that attains top performance.

2) We constructed a novel subjective UGC gaming video database, which we call the LIVE-YT-Gaming. The new database contains 600 UGC gaming videos of unique content, from 59 different games. This is the largest video quality database addressing UGC gaming, and it includes the most exemplar games. Unlike the few existing gaming video quality databases, where PGC videos were professionally recorded, the LIVE-YT-Gaming videos were collected from online uploads by casual users (UGC videos) and are generally afflicted by highly diverse, mixed distortions.

C. Article Structure

The rest of this article is organized as follows. We discuss related work in Section II. We present the new distorted gaming video database in Section III. We provide details on the subjective study in Section IV, and an analysis of the collected subjective data in Section V. We compare the performances of several current VQA models on the new database in Section VI, and introduce a new model, called GAME-VQP. Finally, we summarize this article and discuss possible future work in Section VII.

II. RELATED WORK

A. Subjective Video Quality Database

Existing VQA databases differ by the types of video contents and distortions that are included, as well by the volume and methods of procuring subjective data on them. Two different methods are commonly used to collect subjective ratings on videos in the databases: laboratory studies and online crowdsourced studies.

1) Laboratory VQA Studies: Most of the early VQA databases collected subjective data in the laboratory. A small set of unique reference videos (usually 10 to 20), along with distorted versions of them, was presented to a relatively small (<100) set of volunteer subjects. We will refer to these synthetically distorted datasets as legacy VQA databases. Some representative legacy databases are LIVE VQA [7], LIVE Mobile [8], CDVL [9], MCL-V [10], MCL-JCV [11], and VideoSet [12].
Laboratory studies have advantages: the human data that are obtained are generally more reliable, and the scientist has significant control over the experimental environment. However, it is difficult to recruit a large number of subjects, hence much less data can be collected. Because of this, legacy databases are data-poor although they are valuable test beds because of the high quality of their data.

2) Crowdsourced VQA Studies: In recent years, crowdsourcing as a tool for conducting subjective studies has been used to create a variety of large-scale VQA databases containing UGC videos. Crowdsourcing is usually conducted on online platforms, such as Amazon Mechanical Turk (MTurk) and Crowd-Flower, whereby it is possible to collect a large amount of subjective data from participants around the world. Crowdsourced databases usually contain hundreds to thousands of UGC videos selected and sampled from online video sources, on which large amounts of subjective data can be collected by crowdsourcing. Some representative crowdsourced video quality databases are CVD2014 [13], LIVE-In-Capture [14], KoNViD-1k [15], YFCC100M [16], LIVE-VQC [17], and YouTube-UGC [18].

Crowdsourcing makes it possible to collect a large volume of subjective data from thousands of workers. However, collected data are less reliable than quality scores collected in the laboratory, and disingenuous subjects must be addressed. Further, the researcher must ensure that the subjects have adequate displays and network bandwidths to properly participate. Since a very large number of paid workers are required, online studies can be very expensive.

3) Gaming VQA Databases: As far as we are aware, there are four video quality databases that have been designed for gaming video quality research purposes: GamingVideoSET [19], KUGVD [20], CGVDS [21], and TGV [22].

The GamingVideoSET database contains 24 gaming contents recorded from 12 different games, along with 576 compressed versions of these videos using H.264 compression. However, only 90 of these have associated subjective data. Three different resolutions are included: 480p, 720p, and 1080p, all with frame rates of 30 frames/second (fps). The KUGVD database is similar to GamingVideoSET, but only has six gaming contents and 144 compressed versions of them, 90 of which have subjective data available. Both GamingVideoSET and KUGVD databases are, therefore, limited by having small amounts of subjective data, which hinders model development. All of the reference videos were recorded under professional conditions without visible distortion, unlike UGC videos. Moreover, the distorted videos were all compressed using the same H.264 standard, making quality prediction less difficult. The CGVDS database was specifically developed for analyzing and modeling the quality of gaming videos compressed using hardware accelerated implementations of H.264/MPEG-AVC. It has 360 distorted videos with subjective quality ratings, compressed from 15 PGC reference videos, at framerates of 20, 30, and 60 fps. Finally, the TGV database is a mobile gaming video database containing 1293 gaming video sequences compressed from 150 source videos. Unlike the aforementioned databases, which only contain computer or console games, the videos in the TGV database were all recorded from 17 different mobile games. All four of the abovementioned gaming video databases contain only PGC videos that have been impaired by a single distortion type (compression). Prior to our effort here, there has been no database designed for conducting VQA research on real UGC gaming videos created by casual users.

A comparison of the four existing gaming video quality databases and the proposed LIVE-YT-Gaming database is given in Table 1. There is one more recently published database called GamingHDRVideoSET [23], which includes HDR gaming videos; however, it does not include any subjective data, so we excluded it from our evaluations.

B. Objective Quality Assessment Model

The following is a brief introduction to the development of modern NR VQA models, including both general-purpose models and gaming video quality models.

1) General NR VQA Models: While the earliest NR VQA models were developed to address specific types of distortion, the focus of VQA research shifted toward the development of more general-purpose NR algorithms, which extract various “quality-aware” hand-crafted features, which are fed to simple regressors, such as SVRs, to learn mappings to human subjective quality labels.

The most successful features derive from models of natural scene statistics (NSS) [24] and natural video statistics (NVS) [25], under which high-quality optical images, when subjected to perceptually relevant bandpass processes obey certain statistical regularities that are predictably altered by distortions. Noteworthy examples include NIQE [26], BRISQUE [27], V-BLIINDS [28], HIGRADE [29], GM-LOG [30], DESIQUE [31], and FRIQUEE [32]. More recent models that employ efficiently optimized NSS/NVS features, and/or combined with deep features, include VIDEAL [33], which leverages a hand-optimized selection of statistical features, and RAPIQUE [34], which combines a large number of easily-computed statistics with semantic deep features in an extremely efficient manner, yielding both excellent quality prediction accuracy and rapid computation.

Unlike NSS-based models, data-driven methods, such as CORNIA [35], begin by constructing a codebook via unsupervised learning techniques, instead of using a fixed set of features, followed by temporal hysteresis pooling. A recent published NR VQA model called TLVQM [36] makes use of a two-level feature extraction mechanism to achieve efficient computation of a set of distortion-relevant features that measure motion, specific distortion artifacts, and aesthetics.

Given the emergence of large-scale video quality databases in recent years, the availability of sufficient quantities of training samples has made it possible to train high-performance deep learning models. VSFA [37] makes use of a pretrained convolutional neural network (CNN) as a deep feature extractor, then integrates the framewise features using a gated recurrent unit and a temporal pooling layer. It is one of several SOTA deep learning VQA models that attain superior performance on publicly available UGC video quality databases. PaQ-2-PiQ [38] is a recent frame-based local-to-global deep learning-based
TABLE I
EVALUATION OF FOUR EXISTING GAMING VIDEO QUALITY DATABASES AND LIVE-YT-GAMING DATABASE

| Database         | GamingVideoSEIT | KUGVD | CGVDS | TGVS   | LIVE-YT-Gaming |
|------------------|------------------|-------|-------|--------|----------------|
| Year             | 2018             | 2019  | 2020  | 2021   | 2021          |
| Content No       | 24               | 6     | 15    | 15     | 150           |
| Video No         | 600              | 150   | 255   | 1293   | 600           |
| Game No          | 12               | 6     | 15    | 17     | 59            |
| Subjective Data  | 90               | 90    | 360 + anchor stimuli | Unavailable | 600 |
| Public           | Yes              | Yes   | Yes   | No     | Yes           |
| Resolution       | 480p, 720p, 1080p | 480p, 720p, 1080p | 480p, 720p, 1080p | 480p, 720p, 1080p | 360p, 480p, 720p, 1080p |
| FPS              | 30               | 30    | 20, 30, 60 | 30   | 30, 60        |
| Duration         | 30 s             | 30 s  | 30 s  | 5 s    | 8–9 s         |
| Format           | mp4, yuv         | mp4, yuv | mp4, yuv | Unavailable | mp4 |
| Distortion Type  | H.264 compression | H.264 compression | H.264 compression | H.264, H.265, Tencent codec | UGC distortions |
| Subject No       | 25               | 17    | Over 100 | 19   | 61            |
| Rating No        | 25               | 17    | Unavailable | 19 | 30            |
| Data             | MOS              | MOS   | MOS   | Unavailable | MOS |
| Study Type       | In-lab study     | In-lab study | In-lab study | In-lab study | Online study |

Content No: Total number of unique contents. Video No: Total number of videos. Game No: Total number of source games. Subjective Data: Total number of videos with subjective ratings available. FPS: Frame rate per second. Rating No: Average number of ratings per video.

VQA architecture, while PatchVQ [39] extends the idea of using local quality predictions to improve global space–time VQA performance. Other deep models include V-MEON [40], NIMA [41], PQR [42], and DLlQA [43], which deliver high performance on legacy and UGC video quality databases.

2) Gaming VQA Models: Several NR VQA models have been proposed for gaming. Zadtootaghaj et al. [44] proposed a blind model called NR-GVQM, which trains an SVR model to evaluate the quality of gaming content by extracting nine frame-level features, and by using VMAF scores as proxy ground truth labels. The Nofu model proposed in [45] is a learning-based VQA model, which applies center cropping to rapidly compute 12 frame-based features, followed by model training and temporal pooling. Barman et al. [20] proposed two models. Their NR-GVSQI model also uses VMAF scores as training targets, whereas their NR-GVSQE model uses MOS as the training target. They both make use of basic distortion features that measure blockiness, blurriness, contrast, exposure, etc., as well as quality scores generated by the NSS-based IQA models BIQI, BRISQUE, and NIQE. The ITU-T standard G.1072 [46] describes a gaming video QoE model based on three quality dimensions: spatial video quality, temporal video quality, and input quality (interaction quality). Two recently released models based on deep learning, called NDNetGaming [6] and DEMI [5], were both developed on a DenseNet framework. NDNetGaming is based on a CNN structure trained on VMAF scores as proxy ground truth, then finetuned using MOS. Quality prediction phase is then accomplished via temporal pooling. DEMI uses a CNN architecture similar to NDNetGaming, but focuses on only two types of distortions, blockiness and blurriness.

However, each of these systems were designed and developed on databases containing only a small number of PGC reference videos impaired by a single type of manually applied compression. None have been systematically trained and tested on UGC gaming videos. One reason for this is that there is no UGC gaming video quality database currently available.

C. Video Game and Gaming Video

Video games generally refer to interactive games that run on electronic media platforms. Popular mainstream games may take the form of computer games, console games, mobile games, handheld games, VR games, cloud games, and so on. Many games are available on multiple platforms. Games also fall into various genres, including role-playing games, adventure games, action games, first-person shooters, real-time strategy games, fighting games, board games, massive multiplayer online role-playing games, and others. At present, most gaming video quality research has focused on streamed gaming, such as interactive cloud games and online games, as well as passive live broadcast and recorded gameplay [47].

D. Recording of PGC and UGC Gaming Video

Next, we describe ways by which gaming videos are created, leading to a discussion of PGC and UGC gaming videos.

Unlike natural videos captured by optical cameras, gaming videos, as opposed to live gameplay, are usually obtained by recording the screen of the devices on which the games are being played. A variety of factors affect the output quality of recorded gaming videos. The most basic are the graphics configuration settings of the operation system, and the settings built into each game. The display quality of games is usually affected by settings, such as spatial resolution, frame rate, motion blur, control of screen-space ambient occlusions,
vertical synchronization, variation of texture/material, grass quality, anisotropic filtering, depth of field, reflection quality, water surface quality, shadow quality/style, type of tessellation, and use of triple buffering and antialiasing. Professional players enjoy the ultimate gaming experience by optimizing these settings; however, they present formidable choices for most players, so games provide more intuitive and convenient options. Players can simply set the game quality to high, medium, or low, letting the game program automatically set the appropriate parameters.

Of course, whether or not, the video quality determined by the game settings can be correctly displayed depends heavily on the hardware configuration. A sufficiently powerful GPU can provide complete real-time calculation support to ensure that the game display quality is stable at a high level, but ordinary GPUs may not meet all the quality requirements of all users. For example, high motion scenes may present with noticeable frame drops or lagging. Some games require very large data calculations, so if the GPU cannot process them quickly enough, delays in the display may occur. For those games which require connection to the Internet, poor network conditions may also cause the game screen to delay or even freeze.

The quality of recorded gaming videos is also affected by the recording software used. Professional software can accurately reproduce the original gameplay, but substantial system resources are required, increasing the burden on the gameplay device. The software may provide options for faster and easier use, such as automatic resolution selection, frame rate reduction, or further compression during recording, which may degrade the output quality.

When we refer to PGC gaming videos, we refer to videos captured using professional-grade screen recording software, where the video settings during the gameplay are adjusted to a high level, processed and displayed using professional hardware equipment. By contrast, UGC gaming videos refer to videos recorded by ordinary, casual nonprofessional users on ordinary computers or consoles, using diverse game graphics settings, and various types of recording software. Thus, unlike PGC videos, recorded UGC gaming videos may present with a very wide range of perceptual qualities.

E. Extensions to Prior Work

Here, we more completely explain the design of the database and the protocol of the human study, and we conduct a deep analysis and comparison of existing models on the data. We also introduce a new gaming-video-specific VQA model in Section VI, and compare it against the prior SOTA.

III. LIVE-YOUTUBE GAMING VIDEO QUALITY DATABASE

We present a detailed description of the new LIVE-YT-Gaming database in this section. The new database contains 600 UGC gaming videos harvested from online sources. It is the largest public-domain real UGC gaming video quality database having associated subjective scores. Our main objective has been to create a resource that will support and boost gaming video quality research. By providing this useful and accessible tool, we hope more researchers will be able to engage in research on the perceptual quality of gaming videos. Fig. 1 shows a few example videos from the new database.

Many recent (nongaming) UGC video quality databases [15], [38], [39], [49] were created by harvesting a large number of source videos from one or more large free public video repositories, such as the Internet Archive [50] or YFCC-100M [16]. These are typically “winnowed” to a set of videos that are representative of a category of interest, such as social media videos. This is accomplished by a statistical matching process based on low-level video features, such as blur, colorfulness [51], contrast, spatial information (SI) [52] and temporal information (TI) [53]. Statistical sampling based on matching is not suitable for harvesting gaming videos, however, because of a dearth of large gaming video databases available for free download. While there are numerous and diverse gaming videos that have been uploaded by users onto video websites such as YouTube and Twitch, these generally cannot be downloaded because of copyright issues. Furthermore, gaming videos are strongly characterized by the type and content of the original games, which impacts the statistical structure of the video signals, such as their bandpass properties [3]. Overall, data collection is more difficult and complex than it is for real-world UGC videos.

A. Video Collection

We found the Internet Archive, a free digital library, to be a good source of gaming videos. Taking into account the popularity of games on YouTube, as well as the wide variety of types of games (as described in Section II-C), we selected 59 games to be included in our database. Unlike the completely random downloading of videos used by ourselves and others to create many PGC and UGC databases, we found that the best approach was to search the videos by their game titles, then to ensure the diversity of sources and to reduce bias, videos of each same game title were downloaded randomly, subject to wild resolution and frame rate constraints. Four video resolutions were allowed: 360p (640 x 360), 480p (854 x 480), 720p (1280 x 720), and 1080p (1920 x 1080), and two frame rates were allowed: 30 and 60 fps. In addition, we used the Windows 10 Xbox game bar [54] to capture the gameplay of some games, at frame rates of 30 and 60 fps, and resolutions of 720p and 1080p. The video resolutions were selected based on the YouTube video display resolution and aspect ratio standard [55]. In the end, we downloaded dozens to
hundreds of source videos of each game to use as a data corpus for the next step, video selection.

### B. Video Selection

After obtaining the source video resources, we randomly extracted a few 10-s clips from each video, thereby obtaining about 3000 gaming video clips. Considering the desired scale of the online human study to be conducted, and the number of subjects available (to be explained), we further randomly selected 600 videos from amongst those clips. We then observed the remaining videos, and replaced videos of some popular games that were overrepresented, with videos of less popular games to ensure diversity, and we removed videos containing any inappropriate content. Through this approach, we ensure the representativeness and reliability of the final database. For a variety of reasons, including avoiding video stalls and limiting the subjects’ session durations (Section IV-F), we cropped the videos to durations in the range of 8–9 s.

A summary of the distributions of the video resolutions present in the LIVE-YT-Gaming database is tabulated in Table II.

We also computed SI and TI on all of the remaining videos. These quantities roughly measure the spatial and temporal richness and variety of the video contents. SI and TI are defined as follows:

\[
SI = \max_{\text{time}} \{\text{std}_{\text{space}}[\text{Sobel}(F_n(i,j))]\} \tag{1}
\]

\[
TI = \max_{\text{time}} \{\text{std}_{\text{space}}[M_n(i,j)]\} \tag{2}
\]

where \(F_n\) denotes the luminance component of a video frame at instant \(n\), \((i, j)\) denote spatial coordinates, and \(M_n = F_n - F_{n+1}\) is the frame difference operation. \(\text{Sobel}(F_n)\) denotes Sobel filtering [53]. Fig. 2 shows the distributions of SI and TI for the video contents we selected, indicating that in these aspects, the selected videos contain richer contents than many other UGC video and gaming video databases (see [4], [20], [33]).

## IV. Subjective Quality Assessment

Next we describe the design and implementation of our online study. We stored all of the videos on the Amazon S3 cloud server, providing a safe cloud storage service at high Internet speeds, with sufficient bandwidth to ensure satisfactory video loading speed at the client devices of the study participants. We recruited 61 volunteer naive subjects who participated in and completed the entire study. All were students at The University of Texas at Austin (UT-Austin) with no background knowledge of VQA research. We designed the study in this way, as an online study with fewer, but very reliable subjects, because of Covid-19 regulations. Before the study, we randomly divided the subjects into six groups, each containing about ten subjects. We also randomly divided the 600 videos into six groups, each containing 100 videos. Each subject watched three groups of videos, or 300 videos. In order to avoid any possible bias caused by a same group of videos being watched by a fixed group of subjects, we adopted a round-robin presentation ordering to cross-assign video groups to different subject groups. In this way, each video would be watched by about 30 subjects from the three different subject groups. Fig. 3 illustrates the structure of this round-robin approach.

### A. Study Protocol

Fig. 4 shows the flowchart of the steps of the online study. The volunteer subjects first registered by signing a consent form describing the nature of the human study, after which they received an instruction sheet explaining the purpose, procedures, and display device configuration required for the study. The subjects were required to use a desktop or laptop computer to complete the experiment, and needed to complete a computer configuration check before the study, to meet the study requirements. Each subject received a web link to the training session.

| Video Group Number | 1 | 2 | 3 | 4 | 5 | 6 |
|--------------------|---|---|---|---|---|---|
| Subject Group Number | 1 | S1 | S2 | S3 | 4 | 5 | 6 |
| 2 | S1 | S2 | S3 | 1 | 6 | 5 | 3 |
| 3 | S1 | S2 | S3 | 5 | 6 | 4 | 3 |
| 4 | S1 | S2 | S3 | 1 | S5 | S2 | S3 |
| 5 | S1 | S2 | S3 | S6 | 5 | 6 | 4 |
| 6 | S1 | S2 | S3 | S5 | S6 | 4 | 3 |

Fig. 2. Scatter plot of SI against TI on the 600 gaming videos in the LIVE-YouTube Gaming Video Quality Database.

Fig. 3. Illustration of the round-robin approach used to allocate video groups and subject groups. Grids having the same color indicate video groups watched by subjects in the same group. S1, S2, and S3 are session indices.

Fig. 4. Flowchart of the online study.
Based on the data records captured from the training session, we analyzed whether the subjects' hardware configurations met our requirements. Following that, the subjects received a link to the three rounds of testing sessions at two-day intervals, three rounds overall. After the subjects finished the entire study, they were asked to complete a short questionnaire regarding their opinions of the study.

B. Training Session

The training session was conducted as follows. After opening the link, the subjects first read four webpages of instructions. The first webpage introduced the purpose and basic flow of the study. There were five sample videos at the bottom of the page, labeled as bad, poor, fair, good, and excellent, respectively, exemplifying the possible quality levels they may encounter in the study. This was intended to give subjects a better understanding of the quality of the videos they might encounter and prevent any biases that may arise from our specific quality level definitions. The second webpage provided an explanation of how to use the rating bar. The third webpage introduced the study schedule and how to submit data. The last webpage explained other particulars, such as suggested viewing distance, desired resolution settings, the use of corrective lenses if normally worn, and so on. The subjects were required to display each instructional webpage for at 30 s before proceeding to the next page.

The subjects then entered the experiential training phase and started watching videos. After a subject watched a gaming video, a rating bar appeared after it was played. It was emphasized that they were to provide ratings of video quality, rather than of content or other aspects. On the rating page, a continuous Likert scale [56] was displayed, as shown in Fig. 5. The quality range was labeled from low to high with five guide markers: BAD, POOR, FAIR, GOOD, and EXCELLENT. The initial position of the rating cursor was randomized. The subject was asked to provide an overall opinion score of video quality by dragging the marker anywhere along the continuous rating bar. After the subject clicked the “Submit” button, the marker’s final position was considered to be the rating response. Then, the next video was presented, and the process repeated until the end of the session. The rating scores received from all the subjects were linearly mapped to a numerical quality score in [0, 100].

All of the videos presented in each session were displayed in a random order, and each appeared only once. The order of presentation was different across subjects.

C. Test Session

Each subject participated in a total of three test sessions, each about 30 min. The subjects were allowed to take a break in the middle of the session, if they desired. The sessions were provided to each subject on alternating days to avoid fatigue and memory bias. The steps of the test session were similar to the training session, except that there was no time limit on viewing of the instruction page, so subjects could quickly browse and skip the instructions.

D. Post Questionnaire

We received feedback from 53 of the participants regarding aspects of the study.

1) Video Duration: We asked the subjects for their opinions on the video playback duration. A summary of the results is given in Table III. Among the subjects who participated in the questionnaire, 79% believed that the durations of the observed videos (8–9 s) was long enough for them to accurately rate the video quality. The results in the table indicate that the video durations were generally deemed to be satisfactory.

2) Dizziness: Another issue is that some participants may feel dizzy when watching some gaming videos, especially those that contain fast motion. From the survey results in Table IV, about two-thirds of the subjects did not feel any discomfort during the test, while the remaining one-third suffered varying degrees of dizziness. This is an important issue that should be considered in other subjective studies of gaming, since it may affect the reliability of the final data.

3) Demographics: All of the participants were between 18 and 22 years old. Fig. 6 plots the statistics from the answers to two questions: the total time typically spent watching gaming videos each week, and the devices they used to watch gaming videos. Approximately 30% of the participants did not watch gaming videos, while 50% of them watched at least 2 h of gaming videos a week. Most of the subjects watched gaming videos on
computers including laptops and desktops, while 30% of them watched gaming videos on mobile phones. This suggests that it is of interest to conduct additional research on the quality of gaming videos on mobile devices.

E. Data Recording

In addition to the subject’s subjective quality scores, we also recorded information on the subject’s computing equipment, display, network conditions, and real-time playback logs. These data helped us to guarantee the reliability of the ratings collected. Examination of the collected data, along with feedback from the subjects, revealed no issues worth acting upon.

We also recorded the random initial values of the rating cursor for each displayed video and compared them against the final scores. This was done to ensure that the subjects were responsive and moved the cursor. We recorded the operating system used by each subject, of the three allowed types: Windows, Linux, and MacOS, and the model and version of the browser.

F. Challenges

During our study, we encountered two issues which had to be addressed to ensure the reliability of the collected data. The first was video stalls caused by poor Internet conditions. If a subject’s Internet was unstable, a video could be delayed at startup or paused during playback. This must be avoided, since the subjects might account for any delays, pauses, or rebuffering events when giving their quality ratings, leading to inaccurate results. The second problem was artifacts arising from automatic rescaling by the client device. For example, if a subject was to not set the browser to full-screen mode, their device may spatially scale the videos to fit the current window size, which introducing rescaling artifacts that may affect the subjects’ video quality scores. We took the following steps to deal with these issues.

1) Video Stalls: To avoid stalls, we applied several protocols. First, each video was required to download entirely before playback. As mentioned, the videos were of 8–9 s duration. While most of the videos had volumes less than 20 Megabytes (MB), if any exceeded this, we applied very light visually lossless compression to reduce their size. In this way, we were able to reduce the burden of the Internet download process. Likewise, as each video was playing, download of the next two videos would commence in the background. By preloading the videos to be played next, the possibility of video playback problems arising from network instabilities was reduced.

We also recorded a few relevant parameters of each video playback to determine whether each subject’s Internet connection was stable and whether the videos played correctly. We recorded the playing time of each video on the subject’s device, then compared it with the actual duration of the video, to detect and measure any playback delays. We calculated the total (summed) delay times of all videos played in each session, and found that the accumulated delay time over all sessions of all subjects was less than 0.5 s, indicating that all of the videos played smoothly. We attributed this to the fact that all subjects participated in the same local geographic area (Austin, Texas), avoiding problems encountered in large, international online studies.

2) Rescaling Effects: On most devices, videos are automatically rescaled to fit the screen size, which can introduce rescaling artifacts that may significantly alter the perceived video quality, resulting in inaccurate quality ratings. To avoid this difficulty, we recorded the resolution setting of each subjects’ device as they opened the study webpage. We asked all subjects to set their system resolution to 1080 x 1920, and to display the webpage in full-screen mode throughout the study, so the videos would be played at their original resolutions. Before each session, we recorded the display resolution of the subject’s system and the actual resolution of their browser in full-screen mode. We also checked the zoom settings of the system and browser and required the subjects to set it to default (no zoom). The subjects had to pass these simple criteria before beginning each session.

G. Comparison of Our Study With Laboratory and Crowdsourced Studies

There are interesting similarities and differences between our online study and conventional laboratory and crowdsourced studies.

1) Volume of Collected Data: The amount of data obtained in the study depends on the number of available videos and of participating subjects. Laboratory studies usually accommodate only dozens to hundreds of both subjects and original video contents. By contrast, crowdsourced studies often recruit thousands of workers who may collectively rate thousands of videos. While our online study was not as large as crowdsourced studies, it was larger than most laboratory VQA studies in regards to both video volume and subject subscription.

2) Study Equipment: The experimental equipment used in laboratory studies is provided by the researchers, and is often of professional/scientific grade. Because of this, laboratory studies can address more extreme scenarios, such as very high resolutions or frame rates, and high dynamic display ranges. Crowdsourced studies rely on the subjects’ own equipment, which generally implies only moderate playback capabilities, which limits the research objectives. In large crowdsourced studies, many of the participants may be quite resource-light. These are also constrained by the tools available in the crowdsourcing platform being used. For example, MTurk does not allow full-screen video playback. Therefore, 1080p videos are often automatically downscaled to adapt to the platform page size, introducing rescaling distortions.

By comparison, the subject pool in our experiment was composed of motivated university students who already are required to possess adequate personal computing resources and who generally have access to gigabit Internet in the Austin area.

3) Reliability: The reliability of the study mainly depends on control over the experimental environment and on the veracity and focus of the subjects. Laboratory studies are conducted in a professional scientific setting, and the recruited subjects generally derive from a known, reliable pool. They also personally interact with and are accompanied by the researchers. The online study that we conducted was similar, since although
it was carried out remotely, a similar reliable subject pool was subscribed, who were in remote communication with the research team in regards to the study instructions and to receive assistance to help them set the environment variables. This was felt to be an optimal approach to address COVID restrictions at the time. A crowdsourced study is limited by the test platform, and by the often questionable reliability of the remote, unknown subjects, with whom communication is inconvenient. Moreover, substantial subject validation and rejection protocols must be desired to address the large number of inadequately equipped, distracted, or even frankly dishonest subjects [38], [39], [57].

V. SUBJECTIVE DATA PROCESSING

We describe the processing of the collected mean opinion score (MOS) next. The raw MOS data were first converted into z-scores. Let $s_{ijk}$ denote the score provided by the $i$th subject on the $j$th video in session $k = \{1, 2, 3\}$. Since each video was only rated by half of the subjects, let $\delta(i, j)$ be the indicator function

$$\delta(i, j) = \begin{cases} 
1 & \text{if subject } i \text{ rated video } j \\
0 & \text{otherwise.}
\end{cases}$$

(3)

The z-scores were then computed as follows:

$$z_{ijk} = \frac{s_{ijk} - \bar{s}_k}{\sigma_k}$$

(4)

where

$$\bar{s}_k = \frac{1}{N_k} \sum_{j=1}^{N_k} s_{ijk}$$

(5)

and

$$\sigma_k = \sqrt{\frac{1}{N_k - 1} \sum_{j=1}^{N_k} (s_{ijk} - \bar{s}_k)^2}$$

(6)

where $N_k$ is the number of videos seen by subject $i$ in session $k$. The z-scores from all subjects over all sessions were computed to form the matrix $\{z_{ij}\}$, where $z_{ij}$ is the z-score assigned by the $i$th subject to the $j$th video, $j \in \{1, 2 \ldots 600\}$. The entries of $\{z_{ij}\}$ are empty when $\delta(i, j) = 0$. By converting values to z-scores, subject biases in the form of differences in the locations and ranges of values are eliminated. Further information regarding data processing, such as the computation of z-scores, can be found in [7]. Subject rejection was conducted to remove outliers, following the recommended procedure described in ITU-R BT 500.13 [58], resulting in 5 of the 61 subjects being rejected. The z-scores $z_{ij}$ of the remaining 56 subjects were then linearly rescaled to $[0, 100]$. Finally, the MOS of each video was calculated by averaging the rescaled z-scores

$$\text{MOS}_j = \frac{1}{N_j} \sum_{i=1}^{N} z_{ij}' \delta(i, j)$$

(7)

where $z_{ij}'$ are the rescaled z-scores, $N_j = \sum_{i=1}^{N} \delta(i, j)$, and $N = 600$. The MOS values all fell in the range [4.52, 95.95],

![Fig. 7. MOS distribution across the entire LIVE-YouTube Gaming Video Quality Database and other three publicly available databases. (a) LIVE-YT-Gaming. (b) GamingVideoSET. (c) KUGVD. (d) CGVDS.](image)

A. Subject-Consistency Test

To assess the reliability of the collected subjective ratings, we performed inter and intra-subject consistency analysis in the following two ways.

1) **Intersubject consistency:** We randomly divided the subjective ratings obtained on each video into two disjoint equal groups, calculated the MOS of each video, one for each group, and computed the SROCC values between the two randomly divided groups. We conducted 100 such random splits, obtaining a median SROCC of 0.9400, indicating a high degree of internal consistency.

2) **Intrasubject consistency:** The intrasubject reliability test provides a way to measure the degree of consistency of individual subjects [59]. We, thus, measured the SROCC between the individual opinion scores and the MOS. A median SROCC of 0.7804 was obtained over all of the subjects.

Both the inter and intra-consistency experiments illustrate the high degree of reliability and consistency of the collected subjective ratings.

B. Analysis

Table I lists the particulars of the LIVE-YT-Gaming database. Fig. 7 shows the overall MOS histograms of the LIVE-YT-Gaming database as well as three other publicly available databases: GamingVideoSET, KUGVD, and CGVDS. To ensure consistency, the MOS values of the other three databases were linearly rescaled to the [0, 100] range before plotting. Fig. 8 shows the boxplots of the MOS of each game, demonstrating the diverse content and quality distribution of the new database. Compared with the other four databases, the LIVE-YT-Gaming database contains a larger variety of 59 different games, compared with fewer than 20 games in other databases. Moreover, all

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
600 videos in the LIVE-YT-Gaming database have associated MOS data, ensuring a comprehensive database. We conducted an online study, with slightly weaker control over the experimental environment compared with the laboratory studies used in other databases. However, we had a larger number of participants, second only to CGVDS, and collected 30 rating scores on each video, ensuring the reliability of the data. The video durations in the LIVE-YT-Gaming database range from 8 to 9 s, which is shorter compared with the videos in three of the four previous databases. This decision was made based on various considerations, as detailed in Section IV-F. As shown in the results presented in Section IV-D, the selected video durations were sufficient for participants to rate video quality.

VI. PERFORMANCE AND ANALYSIS

To demonstrate the usefulness of the new LIVE-YT-Gaming database, we evaluated the quality prediction performance of a variety of leading public-domain NR VQA algorithms on it. We selected four popular feature-based NR VQA models: BRISQUE, TLVQM, VIDEVAL, and RAPIQUE, all based on training a support vector regression (SVR) [60], a “completely blind,” training-free model, NIQE, and two deep learning based models, VSFA and PatchVQ. We used the LIBSVM package [61] to implement the SVR for all algorithms that we retrained on the new databases, with a radial basis function kernel. We implemented VSFA and PatchVQ using the code released by the authors. We also tested the performance of two pretrained networks VGG-16 [62] and Resnet-50 [63], by applying the network to the videos and using the output of the fully connected layer to train an SVR model for. Since our main goal is to evaluate the quality prediction performance of different algorithms for gaming videos, we also include one gaming video quality model, NDNetGaming, the code of which is publicly available. We evaluated the performance between predicted quality scores and MOS using three criteria: Spearman’s rank order correlation coefficient (SROCC), Pearson’s (linear) correlation coefficient (LCC) and the Root Mean Squared Error (RMSE). SROCC measures the ranked correlation of two sample distributions without refitting. Before computing the LCC and RMSE measures, the predicted quality scores were passed through a logistic nonlinearity, as described in [64]. Larger values of both SROCC and LCC imply better performance, whereas larger values of RMSE indicate worse performance.

We randomly divided the database into nonoverlapping 80% training and 20% test sets. To determine the best performing model, the training set was split into an 80% set and a 20% validation set, and the cross-validation method was employed. The best model selected through this process was then applied to the test set. We repeated the above process over 100 random splits, and report the median performances over all iterations. At each iteration, the number of samples in the training set and test set were 480 and 120, respectively. For all models, we made use of the publicly available source code from the authors, and used their default settings.

A. New Blind Gaming VQA Model

We recently created a blind VQA model designed for the quality prediction of UGC gaming videos, which overcomes the limitations of existing VQA models when evaluating both synthetic and authentic distortions. We will refer to this model as the Game Video Quality Predictor, or GAME-VQP. Our proposed model utilizes a novel fusion of NSS features and deep learning features to produce reliable quality prediction performance. In the design of GAME-VQP, a bag of spatial and spatio-temporal NSS features is extracted over several color spaces, supported by the assumption that NSS features from different spaces capture distinctive aspects of perceived quality. A widely used pre-trained CNN model, Resnet-50, was used to extract deep learning features. The extracted NSS features and CNN features were each used to train an independent SVR model, then the final prediction score was obtained as the average prediction score of the two models. We also include results that compare the performance of GAME-VQP with other leading VQA models. The algorithm details, analysis, and performance of GAME-VQP can be found in [65].

B. Evaluation Results

The performances of all models are given in Table V. To determine whether there exists significant differences between the performances of the compared models, we conducted a statistical significance test. We used the distributions of the obtained SROCC and LCC values computed over the 100 random train-test iterations. The nonparametric Wilcoxon Rank Sum Test [66], which compares the rank of two lists of samples, was used to conduct hypothesis testing. The null hypothesis was that the median for the row model was equal to the median of the column model at the 95% significance level. The alternate hypothesis was that the median of the row was different from the median of the column. A value of “1” in the table means that the row algorithm was statistically superior to the column algorithm, whereas a value of “-1” means the counter result. A value of “0” indicates that the row and column algorithms were statistically indistinguishable (or equivalent). The statistical significance results are tabulated in Tables VI and VII. As may be observed, our GAME-VQP performed significantly better than the other algorithms. VIDEVAL and RAPIQUE performed well, but still fell short despite their excellent performance on UGC videos.
Table V: Performance Comparison of Various No-Reference VQA Models on the LIVE-YouTube Gaming Video Quality Database Using Nonoverlapping 80% Training and 20% Test Sets

| Model     | SROCC | BRISQUE | TLVQM | VIDEVAL | RAPIQUE | VSFA | PatchVQ | VGG-16 | ResNet-50 | NDNetGaming | GAME-VQP |
|-----------|-------|---------|-------|---------|---------|------|---------|--------|-----------|-------------|----------|
| NIQE      | 0.2801| 0.6037  | 0.7484| 0.8071  | 0.8028  | 0.7762| 0.6803  | 0.5768  | 0.7290    | 0.4640      | 0.8451   |
| LCC       | 0.3037| 0.6383  | 0.7564| 0.8118  | 0.8248  | 0.8014| 0.7334  | 0.6429  | 0.7677    | 0.4682      | 0.8649   |
| RMSE      | 16.208| 13.268  | 11.134| 10.093  | 9.661   | 10.396| 13.240  | 11.083  | 15.108    | 8.878       |

The numbers denote median values over 100 iterations of randomly chosen nonoverlapping 80% training and 20% test sets (subjective MOS versus predicted MOS). The boldface indicates a top-performing model. The italics indicate deep learning VQA models. The underline indicates the prior VQA model designed for gaming videos.

Table VI: Results of One-Sided Wilcoxon Rank Sum Test Performed Between SROCC Values of the VQA Algorithms Compared in Table V

Table VII: Results of One-Sided Wilcoxon Rank Sum Test Performed Between LCC Values of the VQA Algorithms Compared in Table V

The completely blind NR VQA model (NIQE) did not perform well in the new database. This is understandable because the pristine model used by NIQE was created using natural pristine images, while gaming videos are synthetically generated and have very different statistical distributions. One could imagine a NIQE created on pristine gaming videos. Although BRISQUE extracts features similar to NIQE, training it on the UGC gaming videos produced much better results. The relative results of NIQE and BRISQUE suggest that the statistical structures of the gaming videos are different from those of natural videos, they nevertheless possess regularities that may be learned using NSS features. TLVQM captures motion characteristics from the videos and delivered better performance than BRISQUE, indicating the importance of accounting for motion in gaming videos. The performance of the deep learning VSFA model was close to that of RAPIQUE and VIDEVAL, and better than the pretrained Resnet-50 and VGG-16 models. Another deep learning model, PatchVQ, delivered performance falling between those of the two aforementioned models. These results show that deep models are able to capture the characteristics of synthetic videos, suggesting the potential of deep models for gaming VQA. VIDEVAL and RAPIQUE both performed well, showing the effectiveness of combining NSS features with CNN features. This also shows that not all NSS and CNN features successfully contribute to UGC gaming video quality prediction, since the performance of GAME-VQP exceeded those models. By selecting fewer features that deliver high performance, our gaming VQA model produced results that were better than all of the other models.

Fig. 9 shows box plots of the SROCC and LCC distributions of the algorithms compared in Table V over 100 randomized trials on the LIVE-YouTube Gaming Video Quality Database. The central red mark represents the median, while the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, while the outliers are individually plotted using the “.” symbol. (a) SROCC distributions, (b) LCC distributions.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Scatter plots of predicted quality scores versus MOS trained with an SVR using fivefold cross validation on all videos are given in Fig. 10, which shows the correlations between predicted and ground truth scores. Scatter plots of MOS versus the quality predictions for different models are given in Fig. 10(b)–(d), followed regular trends against the MOS. As may be clearly seen in Fig. 10(e), the distribution of the proposed model predictions are more compact than those of the other models/modules, in agreement with the higher correlations against MOS that it attains.

C. Scatter Plot

Scatter plots of VQA model predictions are a good way to visualize model correlations. To calculate scatter plots over the entire LIVE-YT-Gaming database, we applied fivefold cross validation and aggregated the predicted scores obtained from each fold. Scatter plots of MOS versus the quality predictions produced by NIQE, TLVQM, RAPIQUE, Resnet-50, and our proposed model are given in Fig. 10. As may be observed in Fig. 10(a), the predicted NIQE scores correlated poorly with the MOS. The other three models, as shown in Fig. 10(b)–(d), followed regular trends against the MOS. As may be clearly seen in Fig. 10(e), the distribution of the proposed model predictions are more compact than those of the other models/modules, in agreement with the higher correlations against MOS that it attains.

VII. CONCLUSION

UGC gaming videos are receiving significant consumer interest. We have presented a new gaming VQA database, called LIVE-YT-Gaming, containing 600 videos of unique user generated gaming contents labeled by 18 600 subjective ratings. We designed and conducted a new online study to collect subjective data from a smaller group of reliable subjects. As compared against laboratory and crowdsourced studies, our online study has several advantages. We also tested several popular general-purpose and gaming-specific VQA models on the new database, and compared their performances with that of a new VQA model, called GAME-VQP designed for gaming videos. We show that while some existing VQA models performed well, GAME-VQP significantly outperformed the others. We believe this new subjective data resource will help other researchers advance their work on the VQA problem for gaming videos.

REFERENCES

[1] Cisco Visual Networking Index, “Cisco visual networking index: Forecast and methodology,” 2017–2022, Accessed: Dec. 1, 2022. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html
[2] Cisco Visual Networking Index, “2020 is YouTube Gaming’s biggest year, ever: 100B watch time hours,” Dec. 2020. Accessed: Dec. 1, 2020. [Online]. Available: https://blog.youtube/news-and-events/youtube-gaming-2020/
[3] N. Barman, M. G. Martini, S. Zadtootaghaj, S. Möller, and S. Lee, “A comparative quality assessment study for gaming and non-gaming videos,” in Proc. 10th Int. Conf. Qual. Multimedia Exp., 2018, pp. 1–6.
[4] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, “An objective and subjective quality assessment study of passive gaming video streaming,” Int. J. Netw. Manag., vol. 30, no. 3, 2018, Art. no. e2054.
[5] S. Zadtootaghaj et al., “DEMI: Deep video quality estimation model using perceptual video quality dimensions,” in Proc. IEEE 22nd Int. Workshop Multimedia Signal Process., 2020, pp. 1–6.
[6] M. Utke, S. Zadtootaghaj, S. Schmidt, S. Bosse, and S. Möller, “NDNetGaming-development of a no-reference deep CNN for gaming video quality prediction,” Multimedia Tools Appl., vol. 81, pp. 3181–3203, 2020.
[7] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” IEEE Trans. Image Process., vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[8] A. K. Moorothy, L. K. Choi, A. C. Bovik, and G. De Veciana, “Video quality assessment on mobile devices: Subjective, behavioral and objective studies,” IEEE J. Sel. Topics Signal Process., vol. 6, no. 6, pp. 652–671, Oct. 2012.
[9] M. H. Pinson, “The consumer digital video library [best of the web],” IEEE Signal Process. Mag., vol. 30, no. 4, pp. 172–174, Jul. 2013.
[10] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo, “MCL-V: A streaming video quality assessment database,” J. Vis. Commun. Image Representation, vol. 30, pp. 1–9, 2015.
[11] H. Wang et al., “MCL-JCV: A JND-based H. 264/AVC video quality assessment dataset,” in Proc. IEEE Int. Conf. Image Process., 2016, pp. 1509–1513.
[12] H. Wang et al., “Videoset: A large-scale compressed video quality dataset based on JND measurement,” J. Vis. Commun. Image Representation, vol. 46, pp. 292–302, 2017.
[13] M. Nuutila, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, “CVD2014—a database for evaluating no-reference video quality assessment algorithms,” IEEE Trans. Image Process., vol. 25, no. 7, pp. 3073–3086, Jul. 2016.
[14] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorothy, P. Panda, and K.-C. Yang, “In-capture mobile video distortions: A study of subjective behavior and objective algorithms,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 9, pp. 2061–2077, Sep. 2018.
[15] Y. Hosu et al., “The Konstanz natural video database (KoNViD-1k),” in Proc. 9th Int. Conf. Qual. Multimedia Experience, 2017, pp. 1–6.
[16] B. Thomee et al., “YFCC100M: The new data in multimedia research,” Commun. ACM, vol. 59, no. 2, pp. 64–73, 2016.
[17] Z. Sinno and A. C. Bovik, “Large-scale study of perceptual video quality,” IEEE Trans. Image Process., vol. 28, no. 2, pp. 612–627, Feb. 2019.
[18] Y. Wang, S. Inguva, and B. Adsumilli, “YouTube UGC dataset for video compression research,” in Proc. IEEE Int. Workshop Multimedia Signal Process., 2019, pp. 1–5.
[19] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, “GamingVideoSET: A dataset for gaming video streaming applications,” in Proc. 16th Annu. Workshop Netw. Syst. Support Games., 2018, pp. 1–6.
Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Zhenqiang Ying received the B.E. degree in computer science and technology from the Dalian University of Technology, Dalian, China, in 2015, the M.S. degree in computer science from Peking University, Beijing, China in 2018, and the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin, TX, USA, in 2022. He is currently a Research Scientist with Meta Platforms, Inc., Menlo Park, CA, USA. His research interests include image quality assessment, image enhancement, and image analysis.

Dr. Ying was a Member of the Laboratory of Image and Video Engineering and Wireless Networking and Communications Group.

Neil Birkbeck received the Ph.D. degree in computing science from the University of Alberta, Edmonton, AB, Canada, in 2011. He was a Research Scientist with Siemens Corporate Research, Princeton, NJ, USA, working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Software Engineer with the Media Algorithms Team, YouTube/Google, Mountain View, CA, USA. He is working on computer vision, graphics, and robotics, with a specific focus on image-based modeling and rendering. His research interests include perceptual video processing, video coding, and video quality assessment.

Yilin Wang received the B.S. and M.S. degrees in computer science from Nanjing University, Nanjing, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, in 2014. After graduation, he joined the Media Algorithm team with YouTube/Google, Mountain View, CA, USA. He is working on topics in computer vision and image processing. His research interests include video processing infrastructure, video quality assessment, and video compression.

Balu Adsumilli received the master’s degree from the University of Wisconsin Madison, Madison, WI, USA, and the Ph.D. degree in watermark-based error resilience in video communications from the University of California, Santa Barbara, CA, USA. He is currently the Head of Media Algorithms Group with YouTube/Google, Mountain View, CA, USA. Prior to YouTube, he was a Senior Manager Advanced Technology with GoPro, San Mateo, CA, USA, and before that, he was a Senior Research Scientist with Citrix Online, Santa Barbara, CA, USA, at both places developing algorithms for images/video quality enhancement, compression, capture, and streaming. He has authored or coauthored more than 70 papers and 80 granted patents. His research interests include image/video processing, audio and video quality, video compression and transcoding, spherical capture/AR/VR, and visual effects.

Dr. Adsumilli is on both the Television Academy and NATAS committees, IEEE MMSP Technical Committee, and ACM MHV Steering Committee. He is on TPCs and organizing committees for various conferences and held numerous workshops. He is an active member of IEEE, ACM, SPIE, and Visual Effects Society.

Alan C. Bovik (Fellow, IEEE) is currently the Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin, Austin, TX, USA. He is the author of *The Essential Guides to Image and Video Processing*. His research interests include image processing, digital photography, digital television, digital streaming video, social media, and visual perception.

Prof. Bovik was the recipient of the IEEE Edison Medal, IEEE Fourier Award, Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, Technology and Engineering Emmy Award from the National Academy for Television Arts and Sciences, Progress Medal from The Royal Photographic Society, Edwin H. Land Medal from Optica, Norbert Wiener Society Award and Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society, and ten best journal paper awards, including the IEEE Signal Processing Society Sustained Impact Award. He co-founded and was longest serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING and created/Chaired the IEEE International Conference on Image Processing, which was first held in Austin, Texas, 1994. He is an Elected Member of the United States National Academy of Engineering, the National Academy of Inventors, and Academy Europaea.