2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) | 978-1-6654-5824-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/WACVW54805.2022.00013

Subjective Quality Assessment of User-Generated Content Gaming Videos

Xiangxu Yu, Zhengzhong Tu, Zhenqiang Ying, Alan Bovik
The University of Texas at Austin
{yuxiangxu,zhengzhong.tu,zqying,bovik}@utexas.edu

Neil Birkbeck, Yilin Wang, Balu Adsumilli
YouTube, Google Inc.
{birkbeck,yilin,badsumilli}@google.com

Abstract

Benefited from the rapid development of the digital game industry, the growing popularity of online user-generated content (UGC) videos for games has accelerated the development of perceptual video quality assessment (VQA) models specifically for gaming videos. As a novel UGC type, gaming videos are recorded by gamers and uploaded to major streaming media platforms such as YouTube and Twitch, and have been extremely popular among the audience. However, there is little work on VQA research related to gaming videos and understanding their characteristics. In order to promote the development of the gaming VQA model, we created a new UGC gaming video VQA resource, named LIVE-YouTube Gaming video quality (LIVE-YT-Gaming) database, composed of 600 authentic UGC gaming videos and 18,600 subjective quality ratings collected from an online subjective study. We also compared and analyzed several state-of-the-art (SOTA) VQA models on the new database. To support work in this field, the new database will be publicly available through the link: `https://live.ece.utexas.edu/research/LIVE-YT-Gaming/index.html`.

1. Introduction

In recent years, the vigorous development of the digital game industry has boosted the popularity of a novel type of digital video - gaming video, which has made countless game fans ecstatic. The world-famous online video sharing platforms such as YouTube, Vimeo, and Twitch, have a huge amount of user-original-content videos, usually shot and uploaded by the untrained general public. Among them, the share of gaming videos is growing rapidly. In 2020, YouTube Gaming reached a milestone of 100 billion hours of watch time and 40 million active gaming channels [4]. The enthusiastic response of online gaming videos from users has attracted the attention of online video providers. In order to provide users with a better gaming video viewing experience, perceptual video quality assessment (VQA) research has become particularly important.

(a) High quality UGC video (b) Low-quality UGC video

(c) High-quality gaming UGC (d) Low-quality gaming UGC

Figure 1. Challenges in gaming video quality perception: the distortion of gaming UGC videos is highly content-dependent ((c) high quality and (d) low quality), and exhibits different properties compared to normal UGC videos ((a) high quality and (b) low quality). Thus, new subjective studies as well as video quality models need to be developed for analyzing gaming UGC videos.

According to whether there exists a high-quality pristine video for reference, VQA algorithms are divided into three categories: Full-Reference (FR) algorithms that require access to a complete reference video; Reduced-Reference (RR) algorithms that only require partial information passed from the reference video, and No-Reference (NR) algorithms, that completely discard the reference video and directly score the quality of the test video. For user-generated content (UGC) videos without reference videos, the NR VQA algorithm is the only choice.

General NR-VQA Models: General-purpose NR-VQA models involve extracting handcrafted quality-aware nat-

978-1-6654-5824-5/22/$31.00 ©2022 IEEE
DOI 10.1109/WACVW54805.2022.00013

ural scene statistics (NSS) features. Noteworthy examples include NIQE [21], BRISQUE [20], V-BLIINDS [23], HIGRADE [14], GM-LOG [41], DESIQUE [48], and FRIQUEE [5]. More recent models that employ efficiently optimized NSS/NVS features, and/or combined with deep features, include VIDEVAL [30] and RAPIQUE [31]. Regarding data-driven deep learning methods, VSFA [16] makes use of a pre-trained Convolutional Neural Network (CNN) as a deep feature extractor, while PVQ [42] makes use of local-to-global quality predictions to improve overall VQA performance. Other top performed deep models include V-MEON [19], NIMA [28], PQR [47], and DLIQA [11].

Gaming VQA Models: As Fig. 1 suggests, gaming video quality varies in different ways as compared to natural UGC videos. Thus, there is a pressing need for designing gaming-oriented VQA models. The quality assessment research for gaming videos of newly released games has started recently. NR-GVQM [45], Nofu [7], NR-GVSQI [1] were all developed early by extracting image or video processing features and then training regression models. As one of the two recently published algorithms, DEMI [44] is a deep learning algorithm that focuses on predicting blockiness and blurriness of two types of distortions. DEMI first trains a CNN model, then fine-tunes the parameters in an image database, and finally trains a regression model to output the prediction results. Another algorithm, NDNetGaming [34], uses VMAF scores as proxy ground truth training, and proposes a new temporal pooling method. In [38], the authors propose an algorithm based on deep learning for quality assessment of mobile gaming videos.

General UGC VQA Database: VQA research has always been supported by databases that contain videos with sufficient subjective quality rating data. Large-scale VQA databases usually contain thousands of UGC videos collected and sampled from online video sources, on which large amounts of subjective data can be collected through crowdsourced subjective study. Some representative crowdsourced video quality databases are CVD2014 [22], LIVE-In-Capture [6], KoNViD-1k [10], YFCC100M [29], LIVE-VQC [27], and YouTube-UGC [37].

Gaming VQA Database: Four gaming VQA databases that contain subjective ratings have been created in recent years: GamingVideoSET [3], KUGVD [1], CGVDS [46], and TGV [38]. A comparison of the four existing gaming video quality databases is given in Table 1. The original reference videos used in these databases are of perfect quality, recorded by powerful hardware devices, high-quality in-game settings, and professional recording software. We refer these kinds of videos as PGC gaming videos. In addition, the distorted videos on these four databases are only contaminated by compression artifacts, limiting the development of UGC gaming VQA algorithms.

To We summarize the contributions we make as follows:

- **We constructed a first-of-its-kind subjective UGC gaming video database, which we call the LIVE-YouTube Gaming video quality database (LIVE-YT-Gaming).** The new database contains 600 UGC gaming videos of unique content, from 59 different games, making it the largest UGC gaming video database.

- **We conducted a large-scale online human study for the UGC gaming database whereby we collected a large number of subjective quality labels on gaming videos**, yielding 18,600 human quality ratings recorded by 61 human subjects.

- **We conducted a benchmark study on the newly established LIVE-YouTube Gaming databse,** setting a reliable baseline to be compared against. We also give some empirical observations on the results of gaming VQA models.

2. LIVE-YT-Gaming Database

2.1. Video Collection

Many recent general UGC VQA databases [10, 43, 42, 17] were created by collecting a large number of source videos from one or more large free public video repositories, followed by a statistical sampling process. However, this does not apply to the creation of UGC gaming VQA databases because gaming videos are characterized by the type and content of the original game, which affects the statistical structure of the video signals [2]. Therefore, we adopt the following method to collect UGC gaming video.

We chose the Internet Archive (IA) [12] to be the source of gaming videos. Unlike other UGC databases that downloaded source videos entirely randomly, we first selected games being included in our database based on the popularity of the games on YouTube and the wide diversity of types of games, and then searched for gaming videos on IA according to the game titles. According to the resolution and frame rate constraints, videos of the same game type were downloaded randomly. Four video resolutions were selected: 360p, 480p, 720p and 1080p, and two frame rates were selected: 30 fps and 60 fps. The video resolution was selected based on the YouTube video display resolution and aspect ratio standards [35]. In addition to downloading videos from the Internet, we also used the Windows 10 Xbox game bar [40] to record the gameplay of some games. In the end, we obtained gameplay videos of 59 games, each of them having dozens to hundreds of source videos, as the data corpus for the video selection.

Table 1. Evaluation of Four Existing Gaming Video Quality Databases: GamingVideoSET, KUGVD, CGVDS, and TGV, and The Proposed New Database: LIVE-YT-Gaming

Database	Year	Content No	Video No	Game No	Subjective Data	Public	Resolution	FPS	Duration	Format	Distortion Type	Subject No	Rating No	Data	Study Type
GamingVideoSET	2018	24	600	12	90	Yes	480p, 720p, 1080p	30	30 sec	mp4, yuv	H.264 compression	25	25	MOS	In-lab study
KUGVD	2019	6	150	6	90	Yes	480p, 720p, 1080p	30	30 sec	mp4, yuv	H.264 compression	17	17	MOS	In-lab study
CGVDS	2020	15	225	15	360 + anchor stimuli	Yes	480p, 720p, 1080p	20, 30, 60	30 sec	mp4, yuv	H.264 compression	over 100	Unavailable	MOS	In-lab study
TGV	2021	150	1293	17	Unavailable	No	480p, 720p, 1080p	30	5 sec	Unavailable	H264, H265, Tencent codec	19	Unavailable	Unavailable	In-lab study
LIVE-YT-Gaming	2021	600	600	59	600	Yes	360p, 480p, 720p, 1080p	30, 60	8-9 sec	mp4	UGC distortions	61	30	MOS	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: Framerate per second. Subject No: Total number of participating subjects. Rating No: Average number of ratings per video.

Figure 2. Examples from the new LIVE-YouTube Gaming video quality database. The game titles and MOS of videos from left to right are: Fallout (MOS: 4), Dota (MOS: 24), Forza Horizon (MOS: 51), Sekiro:Shadows Die Twice (MOS: 76), Super Smash Bros. (MOS: 95).

Table 2. Distribution of Video Resolutions in LIVE-YT-Gaming Database

Resolution	1080p	720p	480p	360p
30 fps	137	187	36	55
60 fps	129	51	0	5

2.2. Video Selection

We randomly cut several clips of approximately 10 sec from each source video, totaling about 3000 gaming video clips. Taking into account the expected scale of the online human study to be conducted and the number of subjects available, as well as ensuring sufficient game diversity, we finally selected 600 videos to be included in the database. Considering the limitations of online human study, including avoiding video stalls and limiting the duration of the subjects' sessions, we further cropped the videos to a duration in the range of 8-9 sec. A summary of the distributions of the video resolutions present in the LIVE-YT-Gaming database is tabulated in Table 2. Fig. 2 shows a few example videos from the new database. The examples were ranked according to their quality, with the leftmost video having the worst quality and the rightmost one having the best quality.

To show the wide diversity of spatial and temporal richness of the video contents in the database, we calculated Spatial Information (SI) [39] and Temporal Information (TI) [33] of 600 videos. SI and TI are defined as follows:

$$SI = \max_{time} \left\{ std_{space} \left[Sobel(F_n(i,j)) \right] \right\}, \qquad (1)$$

$$TI = \max_{time} \left\{ std_{space} \left[M_n(i,j) \right] \right\}, \qquad (2)$$

where F_n denotes the luminance component of a video frame at instant n, (i,j) denotes spatial coordinates, and $M_n = F_n - F_{n+1}$ is the frame difference operation. $Sobel(F_n)$ denotes Sobel filtering [33]. Fig. 3 shows the distributions of SI and TI for the video contents we selected.

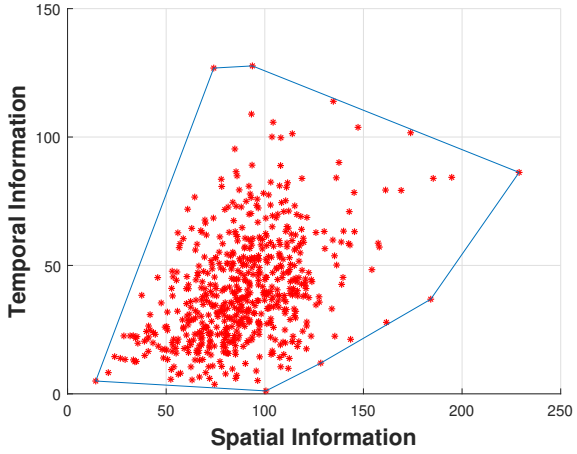


Figure 3. Scatter plot of SI against TI on the 600 gaming videos in the LIVE-YouTube Gaming Video Quality Database.

3. Subjective Study

3.1. Study Protocol

As an online human study, it was important to ensure that all videos were played normally on the subject's client device. We thus stored all videos on Amazon S3 cloud server, providing a secure cloud storage service at high and stable Internet speeds, with enough bandwidth to ensure that the video loading speeds on the client devices of the study participants are satisfactory. We recruited 61 volunteers who had no experience in VQA research from the students of The University of Texas at Austin to participate and completed the entire study. We designed this study in this way, as an online study with fewer subjects, but providing reliable data. Before the study, we randomly and equally divided 61 subjects and 600 into six groups respectively, and each subject watched three groups of videos, or 300 videos. We adopted a round-robin presentation order [15] to cross-assign the video groups to different subject groups so that each video was watched by approximately 30 subjects

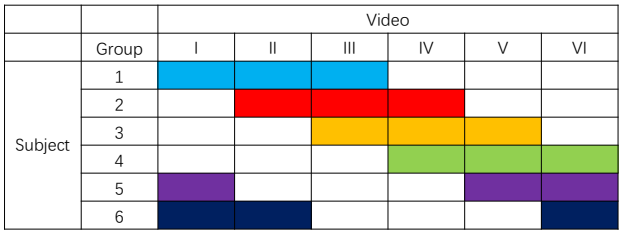


Figure 4. Illustration of the round-robin approach used to allocate video groups and subject groups. Grids having the same color indicate video groups watched by subjects in the same group.

Figure 5. Flow chart of the online study.

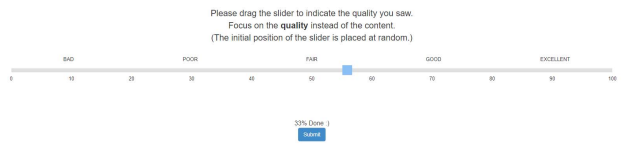


Figure 6. Screenshot of the rating bar used in the online human study.

from three different subject groups to avoid possible bias caused by a same group of subjects watching a same group of videos. Fig. 4 illustrates the structure of this round-robin approach.

Fig. 3.1 shows the flow chart of the steps of the online study. The volunteer subjects first received a detailed description of the study purpose, procedures, and display device configuration instructions in advance, and signed a consent form describing the nature of the human study after reading it. The subjects were required to complete the entire study using a desktop or laptop that had been remotely checked by us and met the configuration requirements. The study included a training session and three test sessions, each session being released at a two-day interval. Before the start of each session, the subject received the corresponding weblink and completed the entire session following the instructions on the webpage. After the subjects completed a session, the recorded data was sent to us. After the subjects completed the entire study, they were asked to complete a short questionnaire regarding their opinions on the study.

3.2. Training And Testing Session

The steps of the training session are introduced as follows After opening the link, the subjects first read four web pages of instructions: (p1) Study purpose and basic process, (p2) Description of video scoring process, (p3) Research timetable and data submission process, (p4) Other details, such as recommended viewing distance, required resolution settings, and so on. The subjects were required to read the instructions on each page for at least 30 sec before proceeding to the next page.

After reading the instructions, the subjects entered the experiential training phase. After the subjects watched a gaming video, they were required to rate the quality of the video using the rating bar that appeared on the webpage. On the rating page, a continuous Likert scale [18] was displayed, as shown in Fig. 6. The subjects rated the overall quality of the video by dragging the marker along the continuous rating bar. The more to the right of the rating bar, the higher the quality score. After the subject clicked the "Submit" button at the bottom of the rating bar, the final position of the mark was considered as the scoring response. Then the subject watched the next video and repeated the process until the end of the session. All the videos presented in each session were displayed in a random order, each appears only once, and the order of the videos viewed by different subjects was different.

The steps of the test sessions were similar to the training session, except that the time limit for viewing the instruction page was removed, so that subjects could quickly browse and skip the instructions. Each subject participated in three test sessions in total. Each test session lasted about 30 minutes, which was provided to subjects on alternating days to reduce the influence of fatigue and memory bias.

3.3. Data Recording

In addition to the subject's subjective quality score, we also recorded the subject's computing device (desktop or laptop allowed), operating system (three types allowed: Windows, Linux, and macOS), monitor (resolution), network status, real-time play log, and other information. We also recorded random initial values of the rating cursor for each displayed video and compared them with the final score, to ensure that the subject moved and responded the cursor. These data helped us ensure the reliability of the ratings collected. After checking the collected data and the subjects' feedback, it was found that there were no issues worthy of action.

3.4. Post Questionnaire

53 of the participants completed our questionnaire.

3.4.1 Video Duration

Table 3 summarized the results of the subjects' opinions on the durations of video playback.

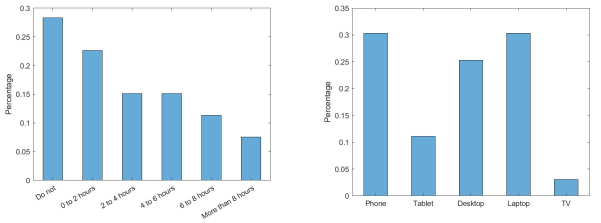
Among the subjects participating in the questionnaire, four-fifths believed that the durations of the video observed (8-9 seconds) was sufficient for them to give an accurate quality score of the played video. Another 5% of the subjects believed that the video duration could be shorter, and

Table 3. The Opinion of Study Participants About Video Duration

	Long enough	Not long enough	Could be shorter
No.	42 (79.2%)	8 (15.1%)	3 (5.7%)

Table 4. Opinions of Study Participants Regarding Video-Induced Dizziness

	None	<30%	30%∼50%	50%∼75%	>75%
No.	34 (64.2%)	16 (30.2%)	2 (3.8%)	1 (1.9%)	0



(a) (b)

Figure 7. Demographic details of the participants (a) Typical number of total hours watching gaming videos each week. (b) Device used to watch gaming videos (multiple choice question).

only 15% needed a longer video duration to rate the quality of the video. The results show that video duration is generally considered satisfactory.

3.4.2 Dizziness

Some gaming videos contain fast motions, which may cause discomfort to some subjects. From the survey results in Table 4, about two-thirds of the subjects experienced dizziness to varying degrees during the study, which may affect the reliability of the final data.

3.4.3 Demographics

We plot in Fig. 7 the statistics of subjects' answers to two questions we designed: the total amount of time they usually spent watching gaming videos each week, and the devices they use to watch gaming videos. Approximately 70% of the participants watched gaming videos in daily life, while 50% of them watched at least 2 hours of gaming videos a week. Most of the subjects watched gaming videos on computers, while 30% of them also watched gaming videos on mobile devices.

3.5. Data Processing

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 approximately half of the participated subjects, let $\delta(i, j)$ be the indicator function

$$\delta(i,j) = \begin{cases} 1 & \text{if subject } i \text{ rated video } j \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

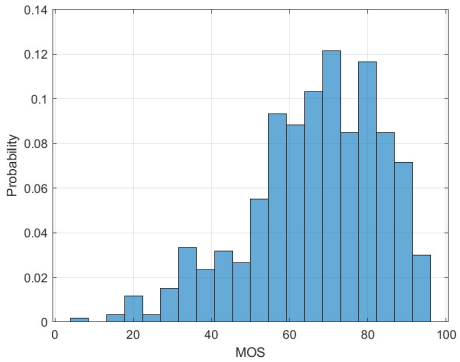


Figure 8. MOS distribution across the entire LIVE-YouTube Gaming Video Quality Database.

The z-scores were then computed from raw subjective data as follows:

$$z_{ijk} = \frac{s_{ijk} - \bar{s}_{ik}}{\sigma_{ik}}, \tag{4}$$

where $\bar{s}_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} s_{ijk}$, and $\sigma_{ik} = \sqrt{\frac{1}{N_{ik}-1} \sum_{j=1}^{N_{ik}} (s_{ijk} - \bar{s}_{ik})^2}$, where N_{ik} 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 $\delta(i, j) = 0$. Following the recommended subject rejection procedure described in ITU-R BT 500.13 [32], we removed five subject outliers of the 61 subjects. The z-scores z_{ij} of the remaining 56 subjects were then linearly rescaled to $[0, 100]$. Finally, the Mean Opinion Score (MOS) of each video was calculated by averaging the rescaled z-scores:

$$MOS_j = \frac{1}{N_j} \sum_{i=1}^{N} z'_{ij} \delta(i, j), \tag{5}$$

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 $[4.52, 95.95]$.

3.6. Subject-Consistency Test

We computed the following two indicators, inter-subject and intra-subject consistency analysis, to evaluate the reliability of the collected subjective scores.

Inter-Subject Consistency We randomly divided the subjective score obtained on each video into two disjoint equal groups, calculated the MOS of each video, one for each group, and calculated the SROCC values between the two randomly divided groups. A median SROCC of **0.9400** was

obtained after conducting 100 such random splits, showing a high degree of internal consistency.

Intra-Subject Consistency Intra-subject reliability testing was used to measure the degree of consistency of individual subjects [9]. Therefore, we measured the SROCC between the personal opinion scores and the MOS. A median SROCC value of **0.7804** was obtained over all subjects.

3.7. Dataset Statistics

The last row of Table 1 lists the detailed information of the LIVE-YT-Gaming database. The overall MOS histogram of the database is drawn in Fig. 8. The MOS of most videos were fell in range $[50, 90]$, showing a right-skewed distribution.

4. Performance and Analysis

Model Baselines: To show the practicability of the new LIVE-YT-Gaming database, we compared the quality prediction performance of several leading public domain NR VQA algorithms on the new database. We selected four popular NR VQA models that are based on feature training: BRISQUE, TLVQM [13], VIDEVAL, and RAPIQUE, a training-free model, NIQE, and a deep learning based model, VSFA. We used the Support Vector Regression (SVR) [24] for regressor training, except for VSFA which uses end-to-end training. We also tested the performance of two pre-trained networks VGG-16 [26] and Resnet-50 [8], by extracting the output of the fully connected layer and average pooling layer, respectively, from videos to train an SVR model. We also include one gaming video quality model for comparison, NDNetGaming, the code of which is publicly available.

Evaluation Metrics: 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). Before computing the LCC and RMSE measures, the predicted quality scores were passed through a logistic non-linearity function [36] to further linearize the objective predictions to be on the same scale as MOS: [25]:

$$f(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp\left(-x + \beta_3/|\beta_4|\right)}. \tag{6}$$

Larger values of both SROCC and LCC imply better performance, while larger values of RMSE indicate worse performance.

Evaluation Procedure: We randomly divided the database into non-overlapping 80% training and 20% test sets. 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.

4.1. Main Evaluation Results

The performances of all models are shown in Table 5. The Table 5 showed that several algorithms perform similarly: VIDEVAL, RAPIQUE and VSFA. To determine whether there exists significant differences between the performances of them, we conducted a statistical significance test using the distributions of the obtained SROCC and LCC values computed over the 100 random train-test iterations. The statistical significance results are tabulated in Tables 6 and 7. When comparing the distribution of SROCC values, VIDEVAL and RAPIQUE behave statistically the same. However, RAPIQUE performed significantly better than VIDEVAL in regarding of LCC distributions, indicating the great potential of the fusion design of NSS and CNN features.

The purely NSS feature-based algorithms, NIQE and BRISQUE, did not perform well on the database. TLVQM, that emphasizes the characteristics of video motions, delivered better performance than NIQE and BRISQUE. RAPIQUE, as an algorithm mainly based on NSS features, supplemented by deep features, achieved top performance than other existing models. VIDEVAL adopted a sampling method to select the best features from a bag of different types of features, including NSS and motion-related features, showing the comparable performance compared with RAPIQUE regarding of SROCC distributions, and only fallen back in a slight range when comparing LCC distributions. The deep model, VSFA, based on the Resnet-50 model, has achieved close performance to that of VIDEVAL and RAPIQUE, and is significantly better than the pre-trained VGG-16 and Resnet-50 models.

Fig. 9 shows box plots of the SROCC and LCC correlations obtained over 100 iterations for the algorithms compared in Table 5. A lower standard deviation with a higher median SROCC or LCC values indicates better and more robust performance. Both VIDEVAL and RAPIQUE exceeded the performance of all the other algorithms, both in terms of stability and performance results.

4.2. Scatter Plot

The correlation comparison of VQA model predictions are visualized through scatter plots. To calculate scatter plots over the entire LIVE-YT-Gaming database, we applied 5-fold cross validation and aggregated the predicted scores obtained from each fold. Scatter plots of model quality predictions of six models are given in Fig. 10. As can be shown in the Fig. 10(a), the correlation between the predicted NIQE scores and MOS was very poor, as well as NDNetGaming shown in Fig. 10(f). The other three models, TLVQM (Fig. 10(b)), RAPIQUE (Fig. 10(c)), and Resnet-

Table 5. Performance Comparison of Various No-Reference VQA Models on The LIVE-YouTube Gaming Video Quality Database Using Non-Overlapping 80% Training And 20% Test Sets. The Numbers Denote Median Values Over 100 Iterations of Randomly Chosen Non-Overlapping 80% Training And 20% Test Sets (Subjective MOS vs Predicted MOS). The Boldfaces Indicate The Top Performing Model. The Italics Indicate Deep Learning VQA Models. The Underline Indicates The Prior VQA Model Designed for Gaming Videos.

	NIQE	BRISQUE	TLVQM	VIDEVAL	**RAPIQUE**	*VSFA*	*VGG-16*	*Resnet-50*	NDNetGaming
SROSS	0.2801	0.6037	0.7484	0.8071	**0.8028**	0.7762	0.5768	0.7290	0.4640
LCC	0.3037	0.6383	0.7564	0.8118	**0.8248**	0.8014	0.6429	0.7677	0.4682
RMSE	16.208	13.268	11.134	10.093	**9.661**	10.396	13.240	11.083	15.108

Table 6. Results of One-Sided Wilcoxon Rank Sum Test Performed Between SROCC Values of The VQA Algorithms Compared In Table 5. A Value Of "1" Indicates That The Row Algorithm Was Statistically Superior to The Column Algorithm; " − 1" Indicates That the Row Was Worse Than the Column; A Value Of "0" Indicates That the Two Algorithms Were Statistically Indistinguishable. The Boldfaces Indicate The Top Performing Model. The Italics Indicate Deep Learning VQA Models. The Underline Indicates A Prior VQA Model Designed for Gaming Videos.

	NIQE	BRISQUE	TLVQM	**VIDEVAL**	**RAPIQUE**	*VSFA*	*VGG-16*	*Resnet-50*	NDNetGaming
NIQE	0	-1	-1	**-1**	**-1**	-1	-1	-1	-1
BRISQUE	1	0	-1	**-1**	**-1**	-1	1	-1	1
TLVQM	1	1	0	**-1**	**-1**	-1	1	1	1
VIDEVAL	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**
RAPIQUE	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**
VSFA	1	1	1	**-1**	**-1**	0	1	1	1
VGG-16	1	-1	-1	**-1**	**-1**	-1	0	-1	1
Resnet-50	1	1	-1	**-1**	**-1**	-1	1	0	1
NDNetGaming	1	-1	-1	**-1**	**-1**	-1	-1	-1	0

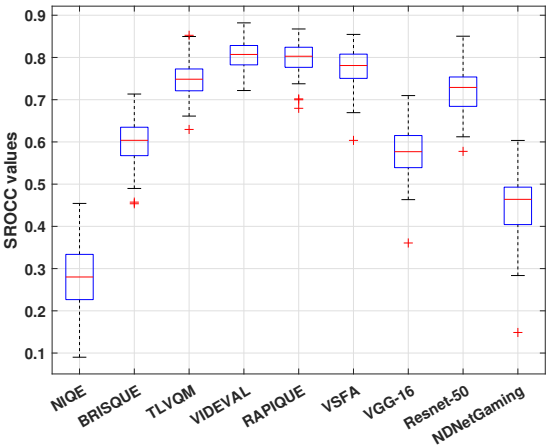


Figure 9. Box plots of the SROCC distributions of the algorithms compared in Table 5 over 100 randomized trials on the LIVE-YouTube Gaming Video Quality Database. The central red mark represents the median, while the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, while the outliers are individually plotted using the '+' symbol.

50 (Fig. 10(e)), followed the better trends against the MOS, indicating stronger correlation. Among them, the distribution of RAPIQUE is the most compact, in line with its best performance as shown in Table 5.

5. Conclusion

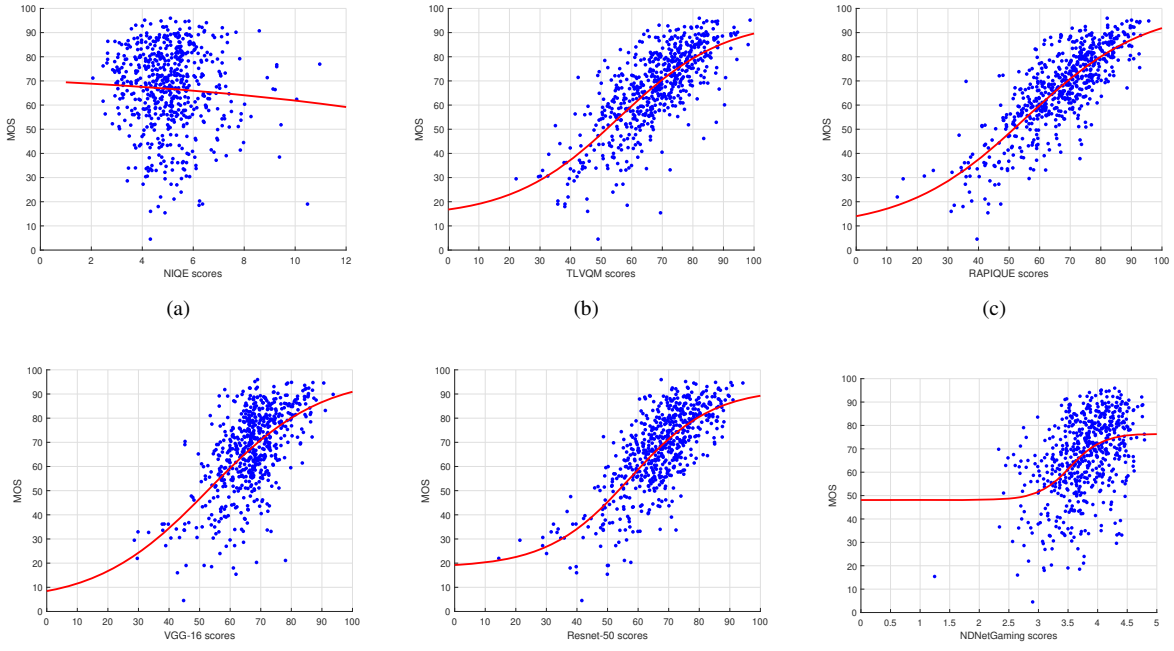
UGC gaming videos have gained more attention in recent years. In response to this, we conducted subjective study on the VQA problem of UGC gaming videos. We have created a new gaming video quality assessment database, called LIVE-YT-Gaming, containing 600 videos of unique user generated gaming contents from 59 different games. We presented a new online study to collect subjective data labeled by 18,600 subjective ratings from 61 subjects for videos included in LIVE-YT-Gaming database. We also tested several popular general-purpose and gaming-specific VQA models on the new database, compared and analyzed their performance from several aspects. The results showed the potential of both NSS features and deep features on VQA research of gaming videos. This database fills the gaps in the research of UGC gaming video quality, and aims to provide researchers with free public resources with subjective quality labels. We believe that this new subjective data resource of UGC gaming videos will help other researchers to further expand their work on gaming VQA problems, such as better study and analysis of the characteristics of UGC gaming videos and the difference between them and general UGC videos for development of VQA algorithms targeted on gaming videos.

Acknowledgment

The human study was approved by the Institutional Review Board of UT-Austin. This research was supported by a gift from YouTube, and by grant number 2019844 for the

Table 7. Results of One-Sided Wilcoxon Rank Sum Test Performed Between LCC Values of The VQA Algorithms Compared In Table 5. A Value Of "1" Indicates That The Row Algorithm Was Statistically Superior to The Column Algorithm; " − 1" Indicates That the Row Was Worse Than the Column; A Value Of "0" Indicates That the Two Algorithms Were Statistically Indistinguishable. The Boldfaces Indicate The Top Performing Model. The Italics Indicate Deep Learning VQA Models. The Underline Indicates A Prior VQA Model Designed for Gaming Videos.

	NIQE	BRISQUE	TLVQM	VIDEVAL	**RAPIQUE**	*VSFA*	*VGG-16*	*Resnet-50*	NDNetGaming
NIQE	0	-1	-1	-1	**-1**	-1	-1	-1	-1
BRISQUE	1	0	-1	-1	**-1**	-1	1	-1	1
TLVQM	1	1	0	-1	**-1**	-1	1	1	1
VIDEVAL	1	1	1	0	**-1**	1	1	1	1
RAPIQUE	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**
VSFA	1	1	1	-1	**-1**	0	1	1	1
VGG-16	1	-1	-1	-1	**-1**	-1	0	-1	1
Resnet-50	1	1	-1	-1	**-1**	-1	1	0	1
NDNetGaming	1	-1	-1	-1	**-1**	-1	-1	-1	0



(a) (b) (c) (d) (e) (f)

Figure 10. Scatter plots of predicted quality scores versus MOS trained with an SVR using 5-fold cross validation on all videos in the LIVE-YouTube Gaming Video Quality Database. (a) NIQE, (b) TLVQM, (c) RAPIQUE, (d) VGG-16, (e) Resnet-50, (f) NDNetGaming.

National Science Foundation AI Institute for Foundations of Machine Learning (IFML).

References

[1] Nabajeet Barman, Emmanuel Jammeh, Seyed Ali Ghorashi, and Maria G Martini. No-reference video quality estimation based on machine learning for passive gaming video streaming applications. *IEEE Access*, 7:74511–74527, 2019.

[2] Nabajeet Barman, Maria G Martini, Saman Zadtootaghaj, Sebastian Möller, and Sanghoon Lee. A comparative quality assessment study for gaming and non-gaming videos. In *2018 Tenth Int. Conf. Quality Multimedia Exp. (QoMEX)*, pages 1–6. IEEE, 2018.

[3] Nabajeet Barman, Saman Zadtootaghaj, Steven Schmidt, Maria G Martini, and Sebastian Möller. GamingVideoSET: a dataset for gaming video streaming applications. In *Proc. 16th Annu. Workshop Netw. Syst. Support Games (NetGames),*, pages 1–6. IEEE, 2018.

[4] Cisco Visual Networking Index. 2020 is YouTube Gaming's biggest year, ever: 100B watch time hours, Dec 2020.

[5] Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *J. Vis.*, 17(1):32–32, 2017.

[6] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Trans. Circuits Syst. Video Technol.*, 28(9):2061–2077, 2017.

[7] Steve Göring, Rakesh Rao Ramachandra Rao, and Alexander Raake. nofu—a lightweight no-reference pixel based video quality model for gaming content. In *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, pages 1–6. IEEE, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, 2016.

[9] Tobias Hossfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *IEEE Trans. Multimedia*, 16(2):541–558, 2013.

[10] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The Konstanz natural video database (KoNViD-1k). In *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, pages 1–6. IEEE, 2017.

[11] Weilong Hou, Xinbo Gao, Dacheng Tao, and Xuelong Li. Blind image quality assessment via deep learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 26(6):1275–1286, 2015.

[12] Internet Archive. Internet archive.

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

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

[15] Dae Yeol Lee, Hyunsuk Ko, Jongho Kim, and Alan C Bovik. On the space-time statistics of motion pictures. *JOSA A*, 38(7):908–923, 2021.

[16] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proc. ACM Int. Conf. Multimedia*, pages 2351–2359, 2019.

[17] Yang Li, Shengbin Meng, Xinfeng Zhang, Shiqi Wang, Yue Wang, and Siwei Ma. UGC-VIDEO: perceptual quality assessment of user-generated videos. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 35–38. IEEE, 2020.

[18] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[19] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *Proc. ACM Multimedia Conf. (MM)*, pages 546–554, 2018.

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

[21] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

[22] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. CVD2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Trans. Image Process.*, 25(7):3073–3086, 2016.

[23] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Trans. Image Process.*, 23(3):1352–1365, 2014.

[24] Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural Comput.*, 12(5):1207–1245, 2000.

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

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[28] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Trans. Image Process.*, 27(8):3998–4011, 2018.

[29] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[30] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021.

[31] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. RAPIQUE: Rapid and accurate video quality prediction of user generated content. *IEEE Open J. Signal Process.*, pages 1–1, 2021.

[32] I. T. U. Methodology for the subjective assessment of the quality of television pictures. ITU-R recommendation BT.500.13, Tech. Rep., 2012.

[33] I. T. U. Subjective video quality assessment methods for multimedia applications. ITU-T recommendation P.910, 2008.

[34] Markus Utke, Saman Zadtootaghaj, Steven Schmidt, Sebastian Bosse, and Sebastian Möller. NDNetGaming-development of a no-reference deep cnn for gaming video quality prediction. *Multimedia Tools Appl.*, pages 1–23, 2020.

[35] Video resolution and aspect ratios. Google.

[36] VQEG. Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment, 2000.

[37] Yilin Wang, Sasi Inguva, and Balu Adsumilli. YouTube UGC dataset for video compression research. In *Proc. IEEE Int. Workshop Multimedia Signal Process.*, pages 1–5. IEEE, 2019.

[38] Shaoguo Wen, Suiyi Ling, Junle Wang, Ximing Chen, Lizhi Fang, Yanqing Jing, and Patrick Le Callet. Subjective and objective quality assessment of mobile gaming video. *arXiv preprint arXiv:2103.05099*, 2021.

[39] S. Winkler. Analysis of public image and video databases for quality assessment. *IEEE J. Sel. Topics Signal Process.*, 6(6):616–625, Oct 2012.

[40] Xbox game bar. Microsoft corporation.

[41] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Trans. Image Process.*, 23(11):4850–4862, 2014.

[42] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-VQ:'patching up'the video quality problem. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14019–14029, 2021.

[43] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[44] Saman Zadtootaghaj, Nabajeet Barman, Rakesh Rao Ramachandra Rao, Steve Göring, Maria G Martini, Alexander Raake, and Sebastian Möller. Demi: Deep video quality estimation model using perceptual video quality dimensions. In *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, pages 1–6. IEEE, 2020.

[45] Saman Zadtootaghaj, Nabajeet Barman, Steven Schmidt, Maria G Martini, and Sebastian Möller. NR-GVQM: A no reference gaming video quality metric. In *Proc. IEEE Int. Symp. Multimedia (ISM)*, pages 131–134. IEEE, 2018.

[46] Saman Zadtootaghaj, Steven Schmidt, Saeed Shafiee Sabet, Sebastian Möller, and Carsten Griwodz. Quality estimation models for gaming video streaming services using perceptual video quality dimensions. In *Proc. 11th ACM Multimedia Syst. Conf.*, pages 213–224, 2020.

[47] Hui Zeng, Lei Zhang, and Alan C Bovik. Blind image quality assessment with a probabilistic quality representation. *IEEE Int'l Conf. on Image Process.*, pages 609–613, 2018.

[48] Yi Zhang and Damon M Chandler. No-reference image quality assessment based on log-derivative statistics of natural scenes. *J. Electron. Imag.*, 22(4):043025, 2013.