

Study of Subjective and Objective Quality Assessment of Mobile Cloud Gaming Videos

Avinab Saha, Yu-Chih Chen, Chase Davis, Bo Qiu, Xiaoming Wang, Rahul Gowda, Ioannis Katsavounidis, Alan C. Bovik, *Fellow, IEEE*

Abstract—We present the outcomes of a recent large-scale subjective study of Mobile Cloud Gaming Video Quality Assessment (MCG-VQA) on a diverse set of gaming videos. Rapid advancements in cloud services, faster video encoding technologies, and increased access to high-speed, low-latency wireless internet have all contributed to the exponential growth of the Mobile Cloud Gaming industry. Consequently, the development of methods to assess the quality of real-time video feeds to end-users of cloud gaming platforms has become increasingly important. However, due to the lack of a large-scale public Mobile Cloud Gaming Video dataset containing a diverse set of distorted videos with corresponding subjective scores, there has been limited work on the development of MCG-VQA models. Towards accelerating progress towards these goals, we created a new dataset, named the LIVE-Meta Mobile Cloud Gaming (LIVE-Meta-MCG) video quality database, composed of 600 landscape and portrait gaming videos, on which we collected 14,400 subjective quality ratings from an in-lab subjective study. Additionally, to demonstrate the usefulness of the new resource, we benchmarked multiple state-of-the-art VQA algorithms on the database. The new database will be made publicly available on our website: <https://live.ece.utexas.edu/research/LIVE-Meta-Mobile-Cloud-Gaming/index.html>

Index Terms—Mobile Cloud Gaming, No-Reference Video Quality Assessment, Cloud Gaming Video Quality Database.

I. INTRODUCTION

THE last decade has witnessed the growth of cloud gaming services as an emergent technology in the digital gaming industry, and many major technology companies such as Meta, Google, Apple, NVIDIA and Microsoft have aggressively invested in building cloud gaming infrastructure. According to a survey by Allied Market Research [1], the cloud gaming industry is projected to grow at a compounded annual growth rate of 57.2% from 2021 to 2030. This astronomical growth may be attributed to multiple factors. Cloud gaming services are a cost-effective alternative to traditional physical gaming consoles and PC (personal computer) based digital video

games, a critical factor contributing to their rapid growth. Cloud gaming subscribers are able to access large and diverse libraries of games playable on any device anywhere without downloading or installing them. Cloud gaming aims to provide high-quality gaming experiences to users by executing complex game software on powerful cloud gaming servers, and streaming the computed game scenes over the internet in real-time, as depicted in Fig 1. Gamers use lightweight software that can be executed on any device to view real-time video game streams while interacting with the games. Cloud gaming services also facilitate rapid video game development processes by eliminating support requirements on multiple user systems, leading to lower overall production costs. This alleviates the need to upgrade consoles and PCs to maintain the gaming experiences of the end-users, as newer and more complex games are made available. Other notable factors contributing to the growth of cloud gaming services include the development of hardware-accelerated video compression methods, access to inexpensive high-speed, lower latency wireless internet services facilitated by the introduction of global 5G services, and the availability of more efficient and affordable cloud platform infrastructures like AWS, Google Cloud, and Microsoft Azure. Another significant contributor to the acceleration of the cloud gaming market since 2019 has been COVID-19 induced restrictions and lockdowns. Indeed, the amount of time spent playing video games increased by more than 71% during the COVID-19 lockdown, as reported in [1].

Recent trends suggest that smartphones have begun to dominate the global cloud gaming industry, and this uptrend is expected to continue. Mobile Cloud Gaming differs from generic Cloud Gaming in various important ways. First, Mobile Cloud Gaming services generally render video game scenes at 720p resolution and 30 frames per second (fps) to accommodate the current gamut of mobile devices, while helping to stabilize delivery and ensuring smoother connections. By comparison, non-mobile Cloud Gaming applications, which are typically played on PCs and televisions, are usually rendered at 1080p/4K resolution and 30-120 fps. Second, Mobile Gaming experiences support gameplay in both portrait and landscape orientations on mobile devices, unlike PCs and television games, which are only playable in landscape mode. Third, Mobile Cloud Gaming services allow users to play over the wireless internet, and must contend with variable internet connections and transmission speeds, unlike cloud gaming services played on PCs and televisions having stable, high-bandwidth wired internet access. This raises significant

This work was supported by Meta Platforms, Inc. A.C. Bovik was supported in part by the National Science Foundation AI Institute for Foundations of Machine Learning (IFML) under Grant 2019844. (Corresponding author: Avinab Saha.)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB), University of Texas, Austin, under FWA No. 00002030 and Protocol No. 2007-11-0066.

Avinab Saha, Yu-Chih Chen, Alan C Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, TX 78712 USA (e-mail: avinab.saha@utexas.edu, berriechen@utexas.edu, bovik@ece.utexas.edu). Chase Davis, Bo Qiu, Xiaoming Wang, Rahul Gowda, Ioannis Katsavounidis are with Meta Platforms Inc., Menlo Park, CA 94025, USA (e-mail: chased@fb.com, qiub@fb.com, xmwang@fb.com, rahulgowda@fb.com, ikatsavounidis@fb.com).

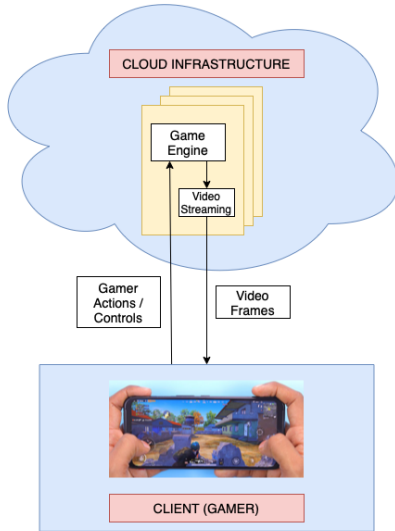


Fig. 1. Exemplar Mobile Cloud Gaming system. Video games scenes are rendered in the Cloud servers of service providers, then the gaming video frames are sent over the Internet to end-users' Mobile devices. The game players' interactions are sent back to the Cloud server over the same network.

technical challenges that must be met to deliver acceptable levels of perceived game video quality.

In a cloud gaming setup, video artifacts can severely impair the perceptual quality of delivered gaming videos. Because of this, there is heightened interest in developing perceptual Video Quality Assessment (VQA) models for gaming videos. However, there have been limited advancements in this direction for two reasons. First, VQA algorithms that have been trained on generic VQA databases generally do not perform well on content-specific gaming videos, which exhibit different appearances and statistical properties than naturalistic camera-captured videos.

Second, building those models inevitably requires the construction of psychometric VQA databases containing large numbers of representative gaming videos that have been labeled with human-annotated scores. Unfortunately, there are very few VQA databases dedicated to Cloud Gaming VQA research, and none are public databases focused on MCG-VQA. Towards advancing progress in this domain, we created a new resource that we call the LIVE-Meta Mobile Cloud Gaming (LIVE-Meta MCG) database, composed of 600 landscape and portrait gaming videos, and targeted explicitly towards mobile cloud gaming. The new database contains 600 videos drawn from 30 source sequences obtained from 16 different games, impaired by varying degrees of video compression and resizing distortions. We then conducted a sizeable human subjective study on these videos. To demonstrate the usefulness of the new database, we also performed a rigorous evaluation of current state-of-the-art VQA models on it, and compared their performance.

The remaining parts of the paper are organized as follows. Section II presents prior work relevant to our mobile cloud gaming video quality. In Section III, we discuss the relevance of the new mobile gaming VQA dataset and highlight the

novelty and significance of our work. Section IV explains the data acquisition process and the design of the human study protocol. Section V compares the performances of various state-of-the-art (SOTA) No-Reference VQA models on the LIVE-Meta Mobile Cloud Gaming (LIVE-Meta MCG) database. Section VI studies the performances of popular Full Reference VQA algorithms originally developed for natural videos, from the perspective of their possibly being used as proxy-MOS or pre-training targets in the development of deep-learning based NR-VQA models for Mobile Cloud Gaming. We conclude in Section VII by summarizing the paper and discussing possible directions of future work.

II. RELATED WORK

Video Quality Assessment research over the last decade has been elevated by the availability of large, comprehensive databases containing videos labeled by subjective quality scores obtained by conducting either laboratory or online studies. Given the explosive growth of the digital gaming industry over the last few years, there is an urgent need to develop gaming-specific VQA algorithms that can be used to monitor and control the quality of video gaming streams transmitted throughout the global internet, towards ensuring that millions of users will experience holistic, high-quality gameplay. Consequently, VQA researchers have begun to develop subjective VQA databases that are focused on gaming videos, as tools for the development of Gaming VQA algorithms. Early work has produced the GamingVideoSET [2] and the Kingston University Gaming Video Dataset (KUGVD) [3]. However, these databases are quite limited in the number of videos having associated subjective quality ratings and in the variety of source content. Both databases [2], [3] were built on only six source sequences, each used to create 15 resolution-bitrate distortion pairs, yielding a total of only 90 videos rated by human subjects. These data limitations are a bottleneck to the development of reliable and flexible VQA models. Towards bridging this gap, a more extensive Cloud Gaming Video Dataset (CGVDS) dataset was introduced in [4]. This dataset includes subjective quality ratings on more than 360 gaming videos obtained from 15 source sequences, collected in a laboratory human study. However, all of the videos in the CGVDS dataset were rendered in landscape mode; hence training a VQA model on them could result in unreliable performance on portrait gaming videos. The other two datasets in the Gaming VQA domain are the Tencent Gaming Video (TGV) dataset [5] and the LIVE-YT-Gaming dataset [6]. The TGV dataset contains 1293 landscape gaming videos drawn from 150 source sequences. However, this dataset is not available in the public domain. The LIVE-YT-Gaming video dataset contains 600 original user-generated content (UGC) gaming videos harvested from the internet. Since these UGC videos were obtained by downloading after-the-fact user-generated gameplay videos from a variety of websites, they are not good candidates for training Cloud Gaming VQA algorithms. Instead, it is desirable to be able to train MCG-VQA models on multiple distorted versions of high-quality source videos, so that they can be used to choose optimal

TABLE I
A SUMMARY OF EXISTING GAMING VQA DATABASES AND THE NEW LIVE-META MOBILE CLOUD GAMING DATABASE

Database	# Videos	# Source Sequences	Pristine Source Sequences	# Ratings per Video	Public	Resolution	Distortion Type	Duration	Display Device	Display Orientation	Study Type
GamingVideoSET	90	6	Yes	25	Yes	480p, 720p, 1080p	H.264	30 sec	24" Monitor	Landscape	Laboratory
KUGVD	90	6	Yes	17	Yes	480p, 720p, 1080p	H.264	30 sec	55" Monitor	Landscape	Laboratory
CGVDS	360 + anchor stimuli	15	Yes	Unavailable	Yes	480p, 720p, 1080p	H.264 NVENC	30 sec	24" Monitor	Landscape	Laboratory
TGV	1293	150	No	Unavailable	No	480p, 720p, 1080p	H.264, H.265, Tencent codec	5 sec	Unknown Mobile Device	Landscape	Laboratory
LIVE-YT-Gaming	600	600	No	30	Yes	360p, 480p, 720p, 1080p	UGC distortions	8-9 sec	Multiple Devices	Landscape	Online
LIVE-Meta Mobile Cloud Gaming	600	30	Yes	24	Yes	360p, 480p, 540p, 720p	H.264 NVENC	20 sec	Google Pixel 5	Landscape, Portrait	Laboratory

streaming settings for given network conditions, to deliver the best possible viewing experiences to gaming end-users.

Other than the LIVE-YT-Gaming dataset, the source videos in gaming databases are of very high pristine quality. They have generally been played using powerful hardware devices, under high-quality game settings and recorded with professional-grade software. The source sequences are then typically processed with resizing and video compression operations to generate a corpus of the distorted videos. We summarize the characteristics of existing gaming VQA databases along with the new LIVE-Meta Mobile Cloud Gaming video quality database in Table I.

Along with the development of Gaming Video Quality databases, several methods have been proposed for Gaming VQA tasks. NR-GVQM [7] trains an SVR model to evaluate the quality of gaming content videos by extracting 9 frame-level features, using VMAF [8] scores as proxy ground-truth labels. In [9], the authors introduced “nofu”, a lightweight model that uses only a center crop of each frame, to speed up the computation of 12 frame-based features, followed by model training and temporal pooling. Recent gaming VQA models based on deep learning include NDNNet-Gaming [10], DEMI [11], and GAMIVAL [12]. Both NDNNet-Gaming and DEMI use Densenet-121 [13] deep learning backbones. Because of the limited amount of subjective scores available to train deep-learning backbones, the Densenet-121 in NDNNet-Gaming is pre-trained on VMAF scores that serve as proxy ground truth labels, then fine-tuned using MOS scores. A temporal pooling algorithm is finally used to compute video quality predictions. DEMI uses a CNN architecture similar to NDNNet-Gaming, while addressing artifacts that include blockiness, blur, and jerkiness. GAMIVAL combines features computed under distorted natural scene statistics model with features computed by the pre-trained CNN backbone used in NDNNet-Gaming, to predict gaming video quality. The ITU-T G.1072 [14] planning model determines gaming video quality based on using objective (non-perceptual) video parameters such as bitrate, framerate, encoding resolution, game complexity, and network parameters.

III. RELEVANCE AND NOVELTY OF LIVE-META MOBILE CLOUD GAMING DATABASE

The new psychometric data resource that we describe here has multiple unique attributes that address most of the shortcomings of existing gaming databases.

First, it includes the largest number of unique source sequences of any non-UGC public gaming VQA database. While the LIVE-YT-Gaming dataset does contain more unique contents, it is directed towards a different problem - VQA of low-quality, user-generated, user-recorded gaming videos. The TGV dataset [5] also has more source sequences, but none of the data is publicly available, making it impossible to independently verify the integrity and modeling value of the videos. Moreover, the video durations are only 5 seconds, heightening the possibility that the subjective quality ratings on the gaming videos, which often contain much longer gameplay scenes, might be less reliable, as explained in [15]. The videos that comprise the LIVE-Meta MCG dataset include a wide range of gameplay and game-lobby video shots. The level of activity in the videos include low, medium, and high motion scenes, a diversity not present in other public gaming databases.

Second, the new data resource can be used to design reliable and robust VQA algorithms, suitable for analyzing high-quality gaming videos subjected to wide ranges and combinations of resizing and compression distortions characteristic of modern streaming workflows. A salient feature of the dataset is that we include videos for all possible resolution-bitrate pairs that are currently relevant to mobile cloud gaming. We believe that VQA tools designed on this data will enable better decision making when selecting streaming settings to deliver perceptually optimized viewing experiences.

Third, not only does the corpus of videos that we assembled target the mobile device scenario, we also conducted the human study using a modern mobile device, unlike any other gaming VQA resource.

Lastly, another unique and differentiating aspect of the new LIVE-Meta MCG is that it includes gaming videos presented in both portrait and landscape orientations. A summary of unique attributes of the new dataset with comparisons against existing gaming VQA datasets is given in Table I.

IV. DETAILS OF SUBJECTIVE STUDY

The LIVE-Meta MCG Database contains 600 video sequences generated from 30 high-quality (pristine) reference source videos by compressing each video using 20 different resolution-bitrate protocols. These videos served as the stimuli that were quality-rated by the humans who participated in our laboratory subjective experiments. Sample frames of landscape



Fig. 2. Sample frames of landscape gaming videos in the LIVE-Meta Mobile Cloud Gaming Database.

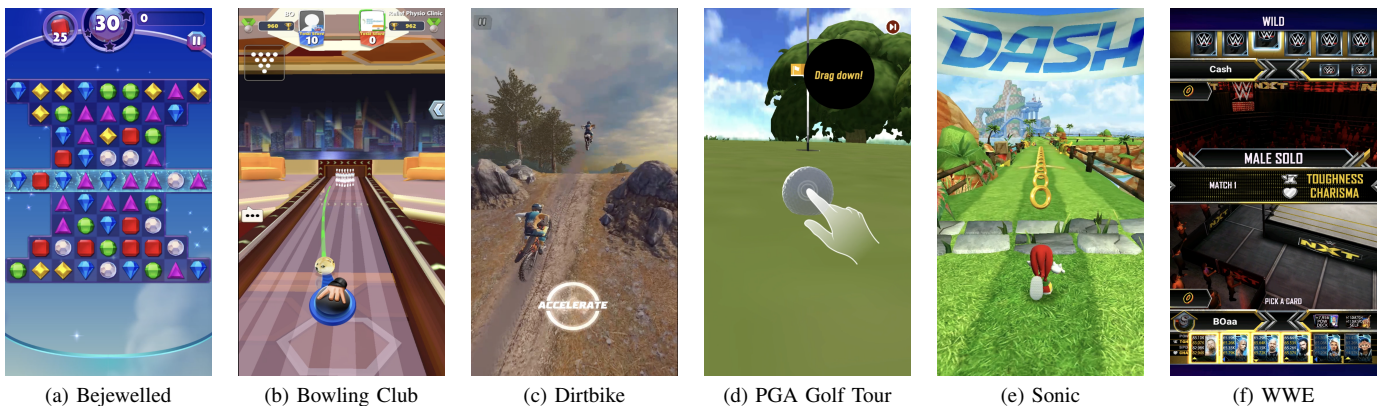


Fig. 3. Sample frames of portrait gaming videos in the LIVE-Meta Mobile Cloud Gaming Database.

and portrait mode gaming video contents in the database are shown in Figs. 2 and 3, respectively.

A. Source Sequences

We collected 16 uncompressed, high-quality source gameplay videos from the Facebook Cloud Gaming servers. We recorded the raw YUV 4 : 2 : 0 video game streams, which were rendered at the cloud servers without any impairments, i.e., before the cloud gaming application pipeline distorted the video stream during gameplay sessions. All of the obtained videos were of original 720p resolution and framerate 30 frames per second, in raw YUV 4 : 2 : 0 format, with their audio components removed. Since, we included both portrait and landscape games in the dataset, by 720p resolution we mean that either the width or the height is 720 pixels, with the other dimension being at least 1280 pixels and often larger. The video contents include 16 different games encompassing diverse contents. Section VIII-A details the games present in the dataset along with their original resolutions as rendered by the Cloud Game engine.

The original 16 reference videos we collected ranged from 58 seconds to 3 minutes which were clipped to lengths that were practical for the human study. Deciding the clip durations presents decisions that depend on several factors. For example, using videos of varying lengths could lead to biases in the subjective ratings provided by the human volunteers. Using

longer videos could limit the data diversity in human studies of necessarily limited participant duration. Moreover, long videos often exhibit distortion changes over time. While it would be worthwhile to investigate time varying distortions of gaming videos, that topic falls outside the scope of the current study, being more appropriate for “Quality of Experience” (QoE) studies similar to those presented in [16], [17], [18].

The goal of our study is to conduct a passive viewing test that will enable us to annotate the video quality of gaming videos. The results from the study [15] illustrated that no significant differences were observed in video quality ratings obtained on the viewing of interactive and passive games that were of 90 seconds duration. However, passive tests of duration 10 seconds yielded significantly higher quality ratings on videos than longer passive tests, indicating that time-varying QoE factors play little role in short-duration tests. The ITU-T P.809 [19] standard recommends using 30-second videos when conducting passive human evaluation of gaming video quality. However, we conducted a trial study involving 20 human participants, each of whom were shown gaming videos of durations ranging from 5 to 35 seconds and asked to provide subjective video quality ratings. The human participants’ feedback led us to conclude that gaming videos of durations no more than 15-20 seconds were needed in order to comfortably provide subjective quality ratings. The feedback received generally indicated that it was sometimes difficult

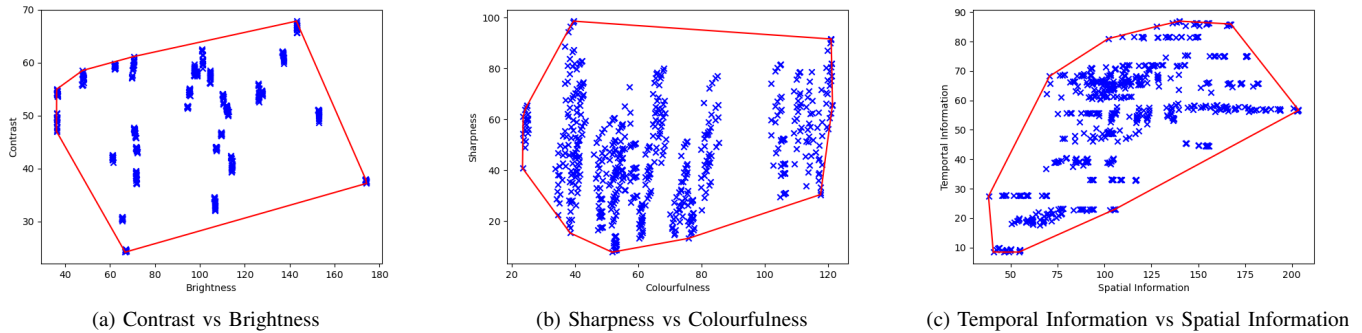


Fig. 4. Source content (blue ‘x’) distribution in paired feature space with corresponding convex hulls (red boundaries). Left column: Contrast x Brightness, middle column: Sharpness x Colourfulness, right column: Temporal Information vs Spatial Information.

TABLE II
RESOLUTION AND BITRATES VALUES OF THE VIDEOS IN THE
LIVE-META MOBILE CLOUD GAMING DATABASE

Encoding Parameter	Value
Resolution	360p, 480p, 540p, 720p
Bitrate	250kbps, 500kbps, 800kbps, 2mbps, 50mbps

to comfortably rate videos that were 10 seconds or shorter, especially on those containing significant motion typical of gaming videos. On the other hand, videos that were 25 seconds or longer were reported to feel too lengthy, and that quality could have been accurately assessed within the initial 15-20 seconds. Moreover, some participants observed the video quality to change over the course of the 25-35 seconds, making it challenging to assign a single quality score. Since the focus of the current study is not to study the time varying (QoE) effects sometimes observed on longer duration videos, we selected between one and three clips from each reference video, each of 20 seconds duration, yielding a total of 30 video clips drawn from the 16 reference videos, all of 720p resolution. We took care that each clip did not include annoying disruptions of otherwise interesting gameplay, and also that clips from the same game presented different scenarios. By distorting the 30 video clips as described in Section IV-B, we obtained 600 videos.

To illustrate the diversity of the video contents in the database, we calculated the following objective features: Brightness, Contrast, Colorfulness [20], Sharpness, Spatial Information and Temporal Information as recommended in [21], [22] for all 600 videos in the database. We calculated the first four objective features on each video frame, then averaged them across all frames to obtain the final feature values. For each frame, brightness and contrast were determined as the mean and standard deviation of the pixel luminance values. We calculated the sharpness of each frame by computing the mean sobel gradient magnitudes at each frame coordinate. We superimposed the convex hulls of the scatter plots of pairs of these features, illustrating the broad feature coverage of the videos in Fig. 4. In Fig. 12, we compare the coverage of our proposed database against other existing Cloud Gaming databases.

B. Mobile Cloud Gaming Pipeline

From each of the 30 reference sequences, 20 distorted video sequences were generated using a combination of resizing and compression distortion processes. Fig. 5 shows a simplified model of the mobile cloud gaming pipeline. The encoding settings we used are similar to those employed in the CGVDS database [4]. We used the Constant Bit Rate (CBR) encoding mode in the hardware accelerated NVIDIA NVENC H.264 encoder [23], with preset set to low latency and high quality. The videos were spatially resized using FFmpeg’s default bicubic interpolation.

We processed each of the 30 reference videos using all 20 possible combinations of resolutions and bitrates listed in Table II. The bitrates range from 250 kbps to 50 mbps, and the resolutions range from 360p to 720p. The reference videos were first spatially resized to 360p, 480p, or 540p or they were maintained at the original 720p resolution, followed by encoding in CBR mode at different bitrates. The selected combinations broadly emulate generic mobile cloud gaming services and available wireless network bandwidths. Most mobile cloud gaming service providers render games at 720p resolution and then, depending on network conditions, either downscale the games to resolutions 360p, 480p, or 540p, or maintain the original resolution before encoding the videos at constant bitrates. Based on our experiments, we generally observed that 250 kbps was the lowest threshold of bandwidth for which acceptable levels of video quality were observed for most of the games in the dataset. We also encoded the videos at higher bitrates typical of common encoding scenarios: 500 kbps, 800 kbps, and 2 mbps, in addition to 250 kbps. Our choice of bitrates ensured that we observed a wide range of perceptual qualities across these bitrates and contents.

Contemporary subjective video quality databases commonly include reference videos. However, since Android mobile devices cannot play lossless (QP=0 encoded) videos, we could not directly incorporate true reference videos in the human study. As an alternative, we encoded the videos at a very high bitrate of 50 mbps to produce “visually lossless” alternatives to uncompressed videos. We will refer to these videos as “proxy reference videos.” We conducted a thorough visual inspection, comparing each reference video to its proxy reference, and

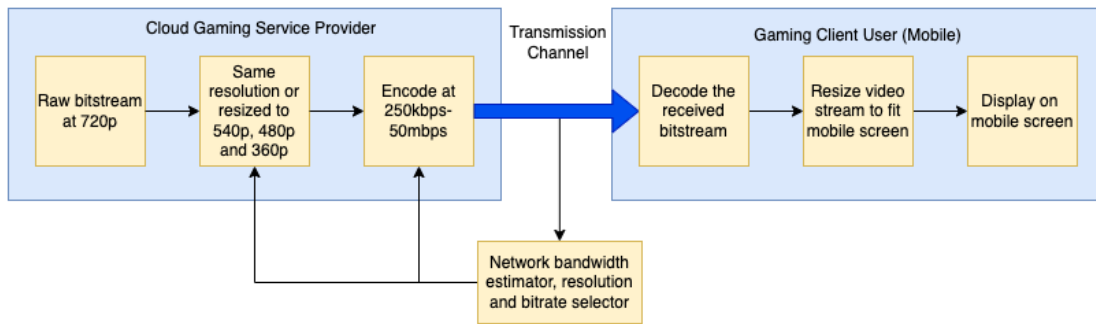


Fig. 5. High-level flow diagram of the mobile cloud gaming pipeline used in the creation of LIVE-Meta Mobile Cloud Gaming database.

concluded that the 50 mbps bitrate was sufficiently high to preserve all visual information in the videos and prevent the introduction of visible artifacts, particularly when taking into account the maximum resolution of the videos was 720p. To further support the conclusions obtained by visual inspection, we also encoded the source videos using QP=0 and observed that the average bit rate of those videos across all the contents was less than that of the proxy reference videos (50 mbps). This strengthens our earlier claim of preserving the visual information in the proxy reference videos since more bits were allocated in the encoding process than would be required for lossless compression. We were also unable to include videos with only resizing distortions (i.e., without video compression) because of the same device limitation. However, following our observation that the proxy reference videos were “visually lossless” when encoded at a bitrate of 50 mbps, we used the same bitrate to encode the videos with only resizing distortions.

C. Subjective Testing Environment and Display

We conducted the large-scale human study in the Subjective Study room in the Laboratory of Image and Video Engineering at The University of Texas at Austin. A Google Pixel 5, running on the Android 11 operating system, was used to display all videos using a custom-built android application. We chose the popular and affordable mid-tier Google Pixel 5 mobile phone as a reasonably representative device that Cloud Gaming clients may often use. The device’s compatibility with the Android operating system also provided us with great flexibility when developing the interface application for the subjective study. The Pixel 5’s high-quality OLED display is renowned for its excellent color accuracy in the brightness range of 60 - 80% of peak brightness [24], making it an excellent choice.

The mobile device was interfaced with a wireless mouse and keyboard to enable the subjects to easily record video quality ratings. The Google Pixel 5 has a 6-inch OLED panel with a 19.5 : 9 aspect ratio Full HD+ (2340 × 1080) resolution and up to a 90Hz refresh rate. The adaptive brightness feature of the mobile device was disabled, and the brightness was set to 75% of the maximum to prevent fluctuations during the study sessions. We utilized the mobile device’s ability to automatically resize incoming video streams using its hard-

ware scaler during cloud gaming, by up-scaling the videos displayed on the mobile device to fit the mobile screen during playback to the subjects. The Android application was memory and compute optimized to ensure smooth playback during the human study.

We arranged the lighting and environment of the LIVE Subjective Study room to simulate a living room. The room’s glass windows were covered with black paper to prevent volunteers from being distracted by any outside activities. To achieve a similar level of illumination as one found in a typical living room, we used two stand-up incandescent lamps, and also placed two white LED studio lights behind where the viewer was seated. We positioned all the lights so that there were no reflections of the light sources from the display screen visible to the subjects. The incident luminance on the display screen was measured by a lux meter and found to be approximately 200 Lux.

A sturdy smartphone mount similar to those found on car dashboards was deployed to secure the mobile device onto the subjects’ desktop. The mount is telescopic, with adjustable viewing angles and heights of the mobile device. The study participants sat comfortably in height-adjustable chairs and were asked to adjust the viewing angle and the height of the mount so they could observe the videos played on the mobile device at approximately arm’s length, similar to the experience of typical gameplay sessions.

We created a video playlist for each participant. After each video was played, a continuous rating bar appeared with a cursor initialized to the extreme left. With the mouse connected wirelessly to the device, the volunteers could freely move the cursor to finalize the quality ratings they gave. There were five labels on the quality bar indicating Bad, Poor, Fair, Good and Excellent to help guide the participants when making their decisions. The subjects’ scores were sampled as integers on [0, 100] based on the final position of the cursor, where 0 indicated the worst quality and 100 the best. However, numerical values were not shown to the volunteers. To confirm the final score of each video, the volunteer pressed the NEXT button below the rating bar, and the score was then stored in a text file. The application then played the following video on the playlist. Fig. 13 in the Appendix Section VIII demonstrates the steps involved in the video quality rating process in the Android application.

TABLE III

ILLUSTRATION OF THE ROUND-ROBIN APPROACH USED TO ALLOCATE VIDEO GROUPS TO SUBJECT GROUPS. SESSIONS A, B REFER TO THE TWO SESSIONS OF THE HUMAN STUDY FOR EVERY SUBJECT. GRID LOCATIONS MARKED AS X INDICATE THE VIDEO GROUP IN THE COLUMN WAS NOT RATED BY THE SUBJECT GROUP IN THE ROW. EACH VIDEO GROUP CONTAINED 100 VIDEOS AND EACH SUBJECT GROUP HAS 12 SUBJECTS

GROUP	Video Group : I	Video Group : II	Video Group : III	Video Group : IV	Video Group : V	Video Group : VI
Subject Group : 1	Session A	Session B	X	X	X	X
Subject Group : 2	X	Session A	Session B	X	X	X
Subject Group : 3	Session B	X	Session A	X	X	X
Subject Group : 4	X	X	X	Session A	Session B	X
Subject Group : 5	X	X	X	X	Session A	Session B
Subject Group : 6	X	X	X	Session B	X	Session A

D. Subjective Testing Protocol

We followed a single-stimulus (SS) testing protocol in the human study, as described in the ITU-R BT 500.13 recommendation [25]. As explained in Section IV-B, we could not include the actual reference videos due to limitations of the Mobile device, but we did include 50 mbps, and 720p resolution encoded versions of each source video as reasonable proxy reference videos.

As explained in Section IV-B, we generated the 600 processed videos by combinations of resizing and compression of the 30 reference videos. The reference (and hence the distorted) videos include equal numbers of portrait and landscape videos. We divided the 30 reference videos into six groups in such a way that groups I, II, III were comprised only of portrait videos while groups IV, V, VI comprised only of landscape videos. In addition, we ensured that no two reference videos in a video group came from the same game. Since we generated 20 distorted versions of each reference video, each video group contained $5 * 20 = 100$ videos. We evenly split the 72 human participants into six groups. Using a round-robin method, we assigned two video groups to each subject group across two sessions, A and B. The exact allocation of video groups for each subject group can be found in Table III. As shown in the Table III, since two subject groups rated each video group, we obtained $2 * 12 = 24$ ratings per video. We designed the study protocol as shown in Table III in a manner such that all the subjects watched either portrait or landscape orientation in both sessions, and never viewed both portrait and landscape videos. We used this approach to eliminate biases caused by any difference in subject preferences for one or the other orientation by any subject.

For the human study, we developed a unique playlist for each session. The order of the videos in the playlist was randomized, with the constraint that videos generated from a reference video were separated by at least one video generated from another reference video. The randomized ordering of the videos reduced the possibility of visual memory effects or any bias caused by playing the videos in a particular order. Each human study session involved rating 100 videos, and required approximately 38 – 40 minutes of each participant’s time.

E. Subject Screening and Training

Seventy-two human student volunteers were recruited from various majors at The University of Texas at Austin to take part in the study. The pool of subjects had little/no experience in

image and video quality assessment. Each subject participated in two sessions separated by at least 24 hours to avoid fatigue.

At the beginning of a volunteer’s first session, we administered the Snellen and Ishihara tests to validate each subject’s vision. Two subjects were found to have a color deficiency, while three volunteers had 20/30 visual acuity. These tests were performed to ensure there was no abnormally high percentage of deficient subjects. All subjects, regardless of their vision deficiencies, were allowed to participate in the study, following our standard goal of designing more realistic psychometric video quality databases [26]. In Section IV-G, we study impact of participants having imperfect vision on the study, by analysing the individual bias and consistency scores obtained using the maximum likelihood estimation algorithm described in [27].

We explained the study objectives to each volunteer before they engaged in the experiment. Volunteers were instructed to rate the gaming videos only on quality, and not on the appeal of the content, such as how boring or exciting the game content was or how well or poorly the player had performed on the recorded gaming video they were rating. Additionally, we demonstrated how the setup could be used to view and rate gaming videos. At the beginning of each test session, volunteers were shown three versions of a same video, which were of perceptually separated qualities to familiarize themselves with the system and to experience the ranges of video quality they would be rating. The scores subjects gave the training videos were not included in the psychometric database.

F. Post Study Questionnaire

The subjects were asked to fill out a questionnaire at the end of each video quality rating session. The data were collected to ensure the reliability of the subjective ratings collected during the human study sessions. Within this sub-section, we present a summary of answers to those questions and demographic information about the subjects.

In Section IV-A, we deliberated on how to determine the optimal duration of each video in our database. To reinforce the result from our pre-study trial (that 20 seconds was long enough to comfortably rate the perceptual quality of each video), we asked every volunteer, as part of the post-study questionnaire, whether the duration of the videos was long enough. Out of the 144 sessions (72 subjects, with 2 sessions per subject) we conducted, in 97.9% (141/144) of the sessions, the human subjects felt that the 20-second

duration was adequate to subjectively judge the video quality. Furthermore, we investigated observer bias and consistency among the three volunteers who deemed the allocated 20 seconds to be inadequate to evaluate subjective video quality in Section IV-G. Section VIII-E summarizes the answers given to the questions regarding the difficulty of rating the videos, and any uneasiness/dizziness induced during the rating process. It also includes the demographic data of the human subjects.

G. Processing of Subjective Scores

To ensure the reliability of the subjective data acquisition process, we first examined the inter-subject and intra-subject consistency of the data using the raw video quality ratings obtained from the human subjects. As explained earlier, we divided the 72 subjects into 6 groups as shown in Table III. We report the inter-subject consistency scores for each group. In order to determine inter-subject consistency, we randomly grouped the scores received for the videos rated by each subject group into two equal but disjoint subgroups, and computed the correlations of the mean opinion scores between the two sub-groups. The random groupings were performed over 100 trials and the medians of both the Spearman's Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC) between the two sub-groups were computed for each of the subject groups and are listed in Table XIII in the Appendix Section VIII. Overall, the average SROCC and PLCC for inter-subject consistency across all subject groups was 0.912 and 0.929, respectively. Furthermore, we calculated intra-subject consistency measurements which provide insight into the behavior of individual subjects [28] on the videos they rated. To do this, we measured the SROCC and PLCC between the individual opinion scores and MOS calculated using all the subjects within each subject group. This process was repeated for every human subject within all the subject groups. The medians for each of the subject groups for both SROCC and PLCC are listed in Table XIII in the Appendix Section VIII. The average SROCC and PLCC over all subject groups was respectively 0.848 and 0.860. These high correlation scores from the above analysis indicate that we can assign a high degree of confidence to the obtained opinion scores.

We employed the method described in [27] to compute the final subjective quality scores on the videos using the raw subjective scores acquired from the human participants. The authors of [27] demonstrate that a maximum likelihood estimate (MLE) method of computing MOS offers advantages to traditional methods, by combining Z-score transformations and subject rejections [25]. The MLE method is less susceptible to subject corruption, provides tighter confidence intervals, better handles missing data, and can provide information on test subjects and video contents.

In [27], the raw opinion scores of the videos are modeled as random variables $\{X_{e,s}\}$. Decompose every rating of a video

in the following way :

$$\begin{aligned} X_{e,s} &= x_e + B_{e,s} + A_{e,s}, \\ B_{e,s} &\sim \mathcal{N}(b_s, v_s^2), \\ A_{e,s} &\sim \mathcal{N}(0, a_{c:c(e)=c}^2), \end{aligned} \quad (1)$$

where $e = 1, 2, 3, \dots, 600$ refer to the indices of the videos in the database and $s = 1, 2, 3, \dots, 72$ refers to the unique human participants. In the above model, x_e represents the quality of the video e as perceived by a hypothetical unbiased and consistent viewer. $B_{e,s}$ are i.i.d gaussian variables representing the human subject s parameterized by a bias (i.e., mean) b_s and inconsistency (i.e., variance) v_s^2 . The human subject bias and inconsistency are assumed to remain constant across all the videos rated by the subject s . $A_{e,s}$ are i.i.d gaussian variables representing a particular video content parameterized by the ambiguity (i.e., variance) a_c^2 of the content c , and $c = 1, 2, \dots, 30$ indexes the unique source sequences in the database. All of the distorted versions of a reference video are presumed to contain the same level of ambiguity, and the video content ambiguity is assumed to be consistent across all users. In this formulation, the parameters $\theta = (\{x_e\}, \{b_s\}, \{v_s\}, \{a_c\})$ denote the variables of the model. To estimate the parameters θ using MLE, the log likelihood function L is defined as :

$$L = \log P(\{x_{e,s}\}|\theta) \quad (2)$$

Using the data obtained from the psychometric study, we derive a solution for $\hat{\theta} = \arg \max_{\theta} L$ using the Belief Propagation algorithm, as shown in [27].

Fig. 6 shows a visual representation of the estimated parameters describing the recovered scores, the subject bias, and the inconsistency and content ambiguity. Fig. 6a shows the recovered quality scores for the 600 videos in the database. The video files are indexed by increasing bitrate values, and further sorted by resolution within each bitrate group. The order of the presented video content is consistent across all resolutions and bitrates. According to our expectations, the average predicted quality scores of videos generally increased as bitrate was increased. Fig. 6a roughly identifies five clusters of videos based on predicted quality scores corresponding to the five bitrate values. Based on the parameter estimates obtained, the lowest bias value $b_s = -20.21$ was found for subject #19, whereas the highest bias value $b_s = 15.43$ was found for subject #59, indicating subject #19's quality scores were, on average, on the low side, while those of subject #59 were, on average, on the high side, as compared to the other human subjects. The median bias value obtained was 0.77. Subject #65 exhibited the greatest variability $v_s = 23.33$ when assigning quality judgements as indicated by the inconsistency estimates v_s , while subject #19 exhibited the lowest level of variability $v_s = 2.06e^{-51}$. The median of the inconsistency estimates was 9.49. Fig. 6c shows the ambiguity in the 30 source videos. A source video from the State of Survival game had the lowest ambiguity $a_c = 4.73$, while a source video from the Sonic game had the highest ambiguity $a_c = 9.99$ among the 30 source videos. We denote the final opinion scores recovered using the above parameters as MLE-MOS.

We analysed both observer bias and inconsistency among

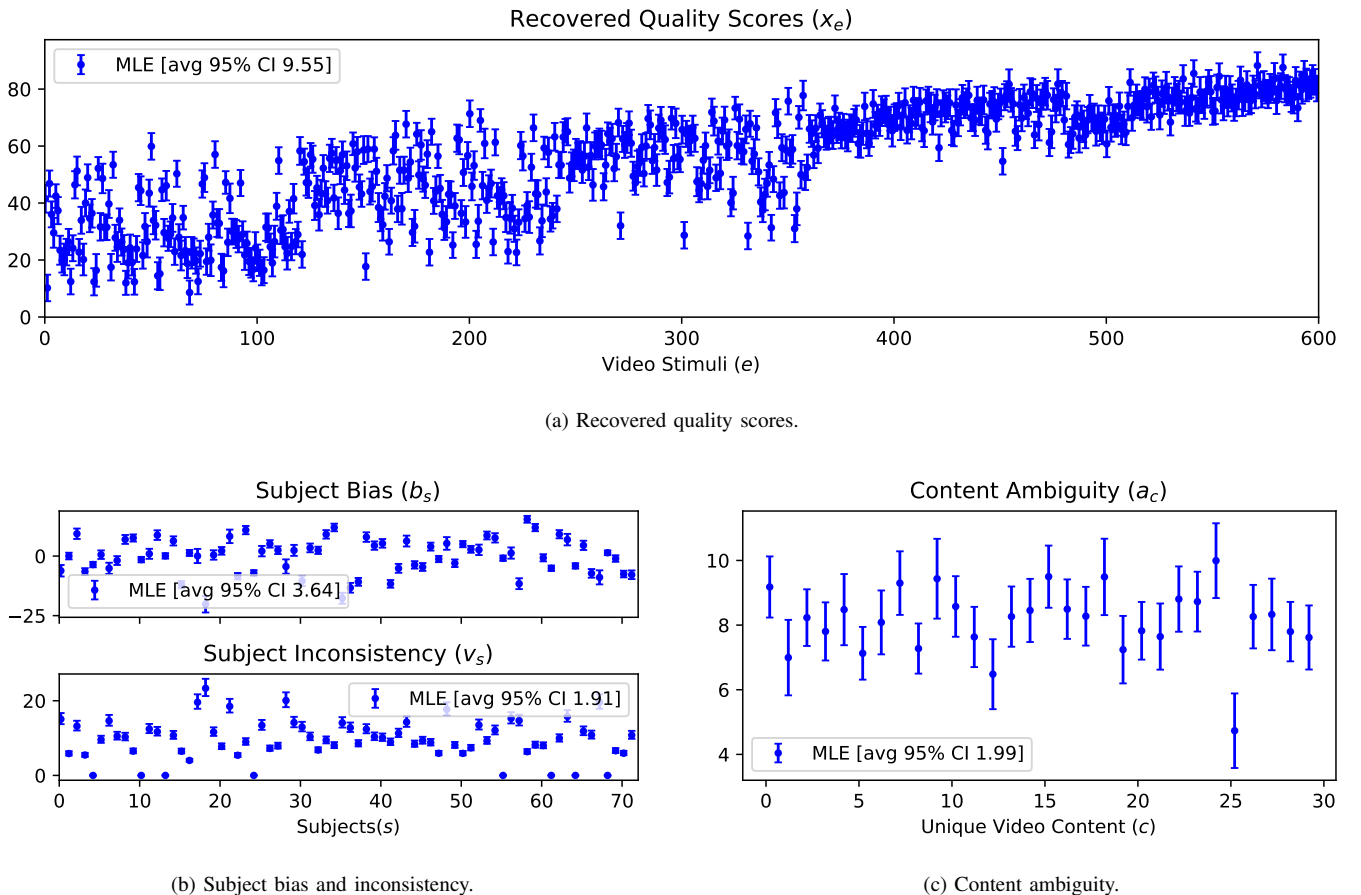


Fig. 6. The result of the MLE formulation to estimate final opinion scores and associated information about subjects and contents. Both the estimated parameters and their 95% confidence intervals are shown.

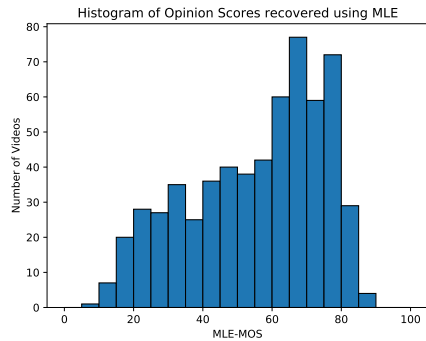
individuals having imperfect vision. We first consider observer bias. Earlier in this section, we reported that the minimum, median, and maximum of observer bias values across all subjects were -20.21 , 0.77 , and 15.43 , respectively. The two subjects, #32 and #49, having color deficiencies, had estimated observer biases of 3.43 and 5.30 , respectively, while the three subjects, #29, #58, and #64, with 20/30 Snellen acuity had estimated observer bias values of -11.59 , 6.90 , and -4.39 , respectively. Since these bias values were not extrema, it is difficult to conclude that visual deficiencies had any impact on the subjective ratings. The minimum, median, and maximum subject inconsistencies across all subjects were estimated to be $2.06e^{-51}$, 9.49 , and 23.33 , respectively. The observer inconsistencies for #32 and #49 were estimated to be 10.35 and 17.67 , respectively, while those for #29, #58, and #64 were estimated to be 14.68 , 15.78 , and 20.06 , respectively. Although some inconsistency values were notably higher than the median, they were not extrema across all the subjects. Thus, we could not conclude that there was any induced observer inconsistency. A more detailed study, with subjects equally sampled with and without visual deficiencies, could better help reveal any impacts of color deficiencies and of slightly reduced visual acuity on video quality ratings. A similar analysis of observer bias and consistency was conducted for subjects #2, #47 and #60, who deemed the 20-second duration insufficient to rate

video quality in one of their sessions. The estimated observer bias values for these subjects were 0.01 , 3.96 , and 11.96 , respectively, and their estimated observer inconsistency values were 5.85 , 8.80 , and 8.16 , respectively. Again, the observer bias and inconsistency values for this group of individuals were not the highest or lowest values among all the subjects in our study. Hence, we could not make any significant conclusions or derive any notable insights from the analysis.

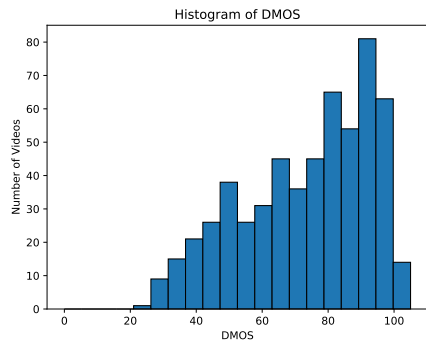
MLE-MOS or MOS in general, is a reliable representation of subjective video quality and is required for the development and evaluation of No-Reference (NR) VQA algorithms, because reference undistorted videos are not available. The Difference MOS (DMOS) is more commonly used in the development and evaluation of Full Reference (FR) VQA algorithms because it allows the reduction of content-dependent quality labels. As discussed earlier, we use the 50 mbps encoded versions of the source videos at 720p resolution as the proxy reference videos when calculating the DMOS scores. The DMOS score of the i^{th} video in the dataset is :

$$DMOS(i) = 100 - (MOS(ref(i)) - MOS(i)), \quad (3)$$

where $MOS(i)$ refers to the MLE-MOS of the i^{th} distorted video obtained using the MLE formulation, and $ref(i)$ refers to the proxy reference video generated from the same source video sequence as the distorted video.



(a) Histogram of MLE-MOS of the human subjects using 20 equally spaced bins.



(b) Histogram of DMOS of the human subjects using 20 equally spaced bins.

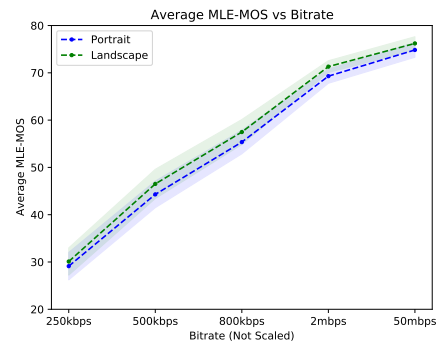
Fig. 7. (a) MLE-MOS (b) DMOS for the LIVE-Meta Mobile Cloud Gaming Database.

H. Analysis and Visualization of the Opinion Scores

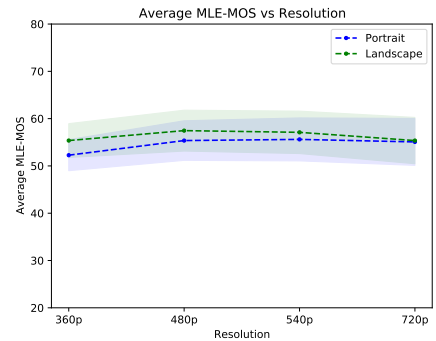
Fig. 7a plots a histogram of the mean opinion scores recovered using the maximum likelihood estimator. The MLE-MOS of the videos in the database ranged from $[8.558, 88.29]$. The MLE-MOS distribution shown in Fig. 7a is slightly right-skewed, typical of other VQA databases. Fig. 7b plots the histogram of DMOS computed using equation 3. The DMOS of the videos in our database ranged from $[21.94, 104.04]$. The distribution of DMOS has a strong resemblance to that of MLE-MOS, with the only difference being a slight shift to the right.

Since our new dataset contains videos in both of the common display orientations (portrait and landscape), we also examined the statistics of the MLE-MOS on each of these two video categories. While the average MLE-MOS rating on all videos was 55.45, it dropped to 54.578 on the portrait videos, and rose to 56.322 on the landscape video. Before reaching any conclusions, we conducted a two-sample one-sided t-test at the 95% confidence interval, to determine whether the differences in the population means of the two video categories were statistically significant. The outcome of the test led us to conclude that the ratings on the two categories of oriented videos were statistically equivalent. We also plotted the average MLE-MOS scores as function of bitrate and resolution after partitioning the videos by orientation category in Fig. 8. Fig. 8a plots the average MLE-MOS for portrait

and landscape videos against bitrate. Although the curve for landscape videos is slightly elevated above the one for portrait videos across all bitrates, applying a two sample one-sided t-tests at each bitrate concluded that the differences between were statistically insignificant. We observed that the average MLE-MOS increased monotonically against bitrate, as expected. A similar analysis was done on the average MLE-MOS of the portrait and landscape videos against resolution, as shown in Fig. 8b. Again, the plot of average MLE-MOS for landscape videos was higher than that of portrait videos across all resolutions, with the separation decreasing with increased resolution. Again, the differences were statistically insignificant across all resolutions.



(a) Average MLE-MOS vs Bitrate for Portrait and Landscape Videos.



(b) Average MLE-MOS vs Resolution for Portrait and Landscape Videos.

Fig. 8. Comparison of the effect of Bitrate and Resolution on MLE-MOS for Landscape and Portrait Videos.

The standard deviations of the estimated MLE-MOS were in the range $[2.023, 2.917]$ with an average of 2.435. The corresponding 95% confidence intervals of MLE-MOS estimates were in the range $[7.93, 11.433]$ with an average of 9.546. We also separately computed the mean of 95% confidence intervals of the MLE-MOS estimates for the portrait and landscape videos. The 95% confidence intervals for the portrait videos were found to fall in the range $[8.421, 11.433]$ with an average of 9.843, while the landscape videos confidence intervals were in the range $[7.93, 10.011]$ with an average of 9.25. We verified that differences in the means of the 95% confidence intervals of the MLE-MOS estimates between the portrait and landscape videos were statistically significant, by conducting a two-

sample one-sided t-test. We also observed that the six source contents contributing to the highest magnitudes of the 95% confidence interval in MLE-MOS estimates were all portrait videos. Based on this evidence, it may be hypothesized that landscape videos provide a more immersive experience than portrait videos, thanks to the horizontal alignment of the eyes. This may contribute to the tighter confidence intervals when measuring video quality.

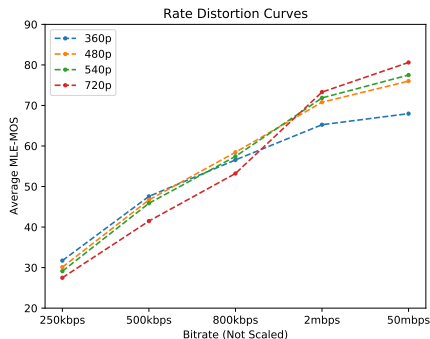


Fig. 9. Rate distortion curves at fixed resolutions.

Fig. 9 plots rate-distortion curves for all four resolutions of videos in the dataset. A plot of this type can supply clues regarding the selection of optimal streaming video resolutions as a function of bandwidth. We observed considerable overlap among the rate-distortion curves around the middle of the bitrate range (500 kbps to 2 mbps). Towards both lower and higher bitrates, the amount of overlap reduced, with 360p being the most preferred resolution at bandwidths of 500 kbps or less, and 720p the preferred resolution at 2 mbps or higher. We provide additional analysis of the mean opinion scores in Section VIII-E of the Appendix.

V. BENCHMARKING OBJECTIVE NR-VQA ALGORITHMS

To demonstrate the usefulness of the new data resource, we evaluated a number of publicly available No-Reference (NR-VQA) algorithms on the LIVE-Meta MCG database. We selected six well-known general-purpose NR-VQA models to test : NIQE [29], BRISQUE [30], TLVQM [31], VIDEVAL [32], RAPIQUE [33], and VSFA [34], as well as three NR-VQA models that were specifically developed for gaming video quality assessment tasks : NDNNet-Gaming [10], GAME-VQP [35] and GAMIVAL [12]. NIQE and BRISQUE are frame-based, and operate by extracting quality-aware features on each frame, then average pooling them to obtain quality feature representations. For the unsupervised, training-free model NIQE, the predicted frame quality scores were directly pooled, yielding the final video quality scores. For the supervised methods (BRISQUE, TLVQM, VIDEVAL, RAPIQUE, GAME-VQP and GAMIVAL), we used a support vector regressor (SVR) with the radial basis function kernel to learn mappings from the pooled quality-aware features to the ground truth MLE-MOS. VSFA uses a Resnet-50 [36] deep learning backbone to obtain quality-aware features, followed by a single layer Artificial Neural Network (ANN) and Gated

Rectified Unit (GRU) [37] to map features to MLE-MOS. The NDNNet-Gaming model however, regressed the video quality scores directly using a Densenet-121 [13] deep learning backbone. GAMIVAL modifies RAPIQUE’s natural scene statistics model and replaces its Imagenet [38] pretrained Resnet-50 CNN feature extractor with the Densenet-121 backbone used in NDNNet-Gaming

We evaluated the performance of the objective NR-VQA algorithms using the following metrics: Spearman’s Rank Order Correlation Coefficient (SROCC), Kendall Rank Correlation Coefficient (KRCC), Pearson’s Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE). The metrics SROCC and KRCC measure the monotonicity of the objective model prediction with respect to human scores, while the metrics PLCC and RMSE measure prediction accuracy. As stated earlier for the PLCC and RMSE measures, the predicted quality scores were passed through a logistic non-linearity function [39] to further linearize the objective predictions and to place them on the same scale as MLE-MOS :

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

We tested the algorithms mentioned above on 1000 random train-test splits using the four metrics. For each split, the training and validation set consisted of videos randomly selected from 80% of the contents, while videos from the remaining 20% constituted the test set. We also ensured that the contents of the training and validation sets were always mutually disjoint. We separated the contents in the training, validation, and test sets to ensure that the content of the videos would not influence the performance of the NR-VQA algorithms. Other than NIQE and NDNNet-Gaming, all of the algorithms were trained on one part of the dataset, then tested using the other, using the aforementioned train-test dataset split. Since NIQE is an unsupervised model, we evaluated its performance on all 1000 test sets, without any training. We also evaluated NDNNet-Gaming using the available pre-trained model on all of the 1000 tests sets, since training code was not available from the authors. We applied five-fold cross-validation to the training and validation sets of BRISQUE, TLVQM, VIDEVAL, RAPIQUE, GAME-VQP and GAMIVAL to find the optimal parameters of the SVRs they were built on. When testing VSFA, for each of the 1000 splits, the train and validation videos were used to select the best performing ANN-GRU model weights on the validation set.

A. Performance of NR-VQA Models

Table IV lists the performances of the aforementioned NR-VQA algorithms on the LIVE-Meta Mobile Cloud Gaming database. In addition, we used the 1000 SROCC and PLCC scores produced by the NR VQA models to run one-sided t-tests, using the 95% confidence level, to determine whether one VQA algorithm was statistically superior to another. Each entry in Table V consists of two symbols, where the first symbol corresponds to the t-test done using the SROCC values, and the second symbol corresponds to the t-test done using the PLCC values. We found that NIQE performed poorly, which is unsurprising since it was developed using natural

TABLE IV
MEDIAN SROCC, KRCC, PLCC, AND RMSE ON THE LIVE-META MOBILE CLOUD GAMING DATABASE OF NR-VQA ALGORITHMS OVER 1000 TRAIN-TEST SPLITS (SUBJECTIVE MLE-MOS VS PREDICTED MLE-MOS). STANDARD DEVIATIONS ARE SHOWN IN PARENTHESES. THE BEST PERFORMING ALGORITHM IS BOLD-FACED

Metrics	SROCC(\uparrow)	KRCC(\uparrow)	PLCC(\uparrow)	RMSE(\downarrow)
NIQE	-0.3900 (0.1816)	-0.2795 (0.1366)	0.4581 (0.2165)	16.5475 (1.9996)
BRISQUE	0.7319 (0.1358)	0.5395 (0.1154)	0.7394 (0.1285)	12.5618 (2.5135)
TLVQM	0.6553 (0.1428)	0.4777 (0.1166)	0.6889 (0.1464)	13.5413 (2.6724)
VIDEVAL	0.7621 (0.1061)	0.5756 (0.0982)	0.7763 (0.1105)	11.7520 (2.2783)
RAPIQUE	0.8740 (0.0673)	0.6964 (0.0759)	0.9039 (0.0565)	8.0242 (1.6755)
GAME-VQP	0.8709 (0.0616)	0.6885 (0.0714)	0.8882 (0.0560)	8.5960 (1.7621)
NDNet-Gaming	0.8382 (0.1227)	0.6485 (0.1009)	0.8200 (0.1227)	10.5757 (3.0354)
VSFA	0.9143 (0.0435)	0.7484 (0.0572)	0.9264 (0.0380)	7.1316 (1.6082)
GAMIVAL	0.9441 (0.0281)	0.7964 (0.0474)	0.9524 (0.0290)	5.7683 (1.429)

TABLE V
RESULTS OF ONE-SIDED T-TEST PERFORMED USING THE 1000 (SROCC, PLCC) VALUES OF THE COMPARED NR-VQA ALGORITHMS COMPUTED ON THE LIVE-META MCG DATABASE. EACH CELL CONTAINS 2 SYMBOLS: THE FIRST SYMBOL CORRESPONDS TO THE T-TEST DONE USING THE SROCC VALUES, AND THE SECOND CORRESPONDS TO THE T-TEST DONE USING THE PLCC VALUES. WHEN A SYMBOL '1' APPEARS, IT DENOTES THAT THE ALGORITHM ON THE ROW WAS STATISTICALLY SUPERIOR TO THAT ON THE COLUMN, WHEREAS '0' INDICATES THAT THE ALGORITHM ON THE COLUMN WAS STATISTICALLY SUPERIOR. A '-' SYMBOL INDICATES THAT THE COLUMN AND ROW ALGORITHMS PERFORMED EQUALLY WELL

ALGORITHM	NIQE	BRISQUE	TLVQM	VIDEVAL	RAPIQUE	GAME-VQP	NDNet-Gaming	VSFA	GAMIVAL
NIQE	(-, -)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
BRISQUE	(1, 1)	(-, -)	(1, 1)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
TLVQM	(1, 1)	(0, 0)	(-, -)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
VIDEVAL	(1, 1)	(1, 1)	(1, 1)	(-, -)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
RAPIQUE	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(-, -)	(-, 1)	(1, 1)	(0, 0)	(0, 0)
GAME-VQP	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(-, 0)	(-, -)	(1, 1)	(0, 0)	(0, 0)
NDNet-Gaming	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(0, 0)	(0, 0)	(-, -)	(0, 0)	(0, 0)
VSFA	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(-, -)	(0, 0)
GAMIVAL	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(-, -)

TABLE VI
MEDIAN SROCC, KRCC, PLCC, AND RMSE OF THE COMPARED NR-VQA MODELS ON THE LIVE-META MOBILE CLOUD GAMING DATABASE, DIVIDED BY DISPLAY ORIENTATIONS, OVER 400 TRAIN-TEST SPLITS. STANDARD DEVIATIONS ARE SHOWN IN PARENTHESES. THE BEST PERFORMING ALGORITHM IS BOLD-FACED

Metrics	Landscape Videos				Portrait Videos			
	RAPIQUE	GAME-VQP	VSFA	GAMIVAL	RAPIQUE	GAME-VQP	VSFA	GAMIVAL
SROCC(\uparrow)	0.876 (0.120)	0.885 (0.087)	0.927 (0.084)	0.955 (0.035)	0.851 (0.122)	0.850 (0.111)	0.903 (0.076)	0.900 (0.062)
KRCC(\uparrow)	0.701 (0.117)	0.715 (0.093)	0.774 (0.090)	0.829 (0.056)	0.680 (0.124)	0.673 (0.109)	0.732 (0.087)	0.735 (0.083)
PLCC(\uparrow)	0.919 (0.103)	0.912 (0.069)	0.946 (0.071)	0.969 (0.023)	0.882 (0.122)	0.876 (0.103)	0.916 (0.075)	0.912 (0.068)
RMSE(\downarrow)	7.294 (2.811)	7.470 (2.630)	5.873 (2.226)	4.547 (1.525)	8.723 (2.632)	8.706 (2.504)	7.371 (2.822)	7.417 (2.576)

TABLE VII
COMPUTATION COMPLEXITY EXPRESSED IN TERMS OF TIME AND FLOATING POINT OPERATIONS (FLOPS) ON 600 FRAMES OF A 360×720 VIDEO UPSCALED TO 1080×2160 FRAMES FROM THE LIVE-META MCG DATABASE

ALGORITHM	Platform	Time (seconds)	FLOPS ($\times 10^9$)
NIQE	MATLAB	728	1965
BRISQUE	MATLAB	205	241
TLVQM	MATLAB	588	283
VIDEVAL	MATLAB	959	2334
RAPIQUE	MATLAB	103	322
GAME-VQP	MATLAB	2053	11627
NDNet-Gaming	Python, Tensorflow	779	126704
VSFA	Python, Pytorch	2385	229079
GAMIVAL	Python, Tensorflow, MATLAB	201	8683

images, while gaming videos are rendered synthetically and have different statistical structures. However, the performance of the same NIQE features improved when we extracted them and used an SVR to regress from the features to the MLE-

MOS in the BRISQUE algorithm. The gap in performance between NIQE and BRISQUE points to the differences in the statistics of camera-captured videos of the real world as compared to graphical rendered synthetic gaming video scenes. However, BRISQUE was able to adapt to these synthetic scene statistics. The performance of TLVQM was average, probably because that model uses many hand-tuned hyper-parameters that were selected to optimize the prediction of video quality on general purpose content and do not generalize well to gaming videos. A similar scenario occurs with VIDEVAL. Although VIDEVAL had slightly boosted performance relative to BRISQUE, its performance may be limited since it uses 60 features selected from more than 700 to maximize performance on in-the-wild UGC videos. Models that use deep learning like VSFA and NDNet-Gaming, and others that use hybrids of deep-learning-based features and handcrafted perceptual features, like RAPIQUE, GAME-VQP and GAMIVAL exhibit considerably improved performance, showing that they are able to capture the statistical structure of synthetically generated gaming videos, suggesting their poten-

tial as VQA algorithms targeting Cloud Gaming applications. The NR-VQA algorithms GAME-VQP and RAPIQUE use a combination of traditional NSS and deep-learning features to considerably improve performance relative to BRISQUE, VIDEVAL, and TLVQM on the LIVE-Meta MCG database. The superior performance of the VSFA model over GAME-VQP and RAPIQUE using only deep-learning features might indicate a reduced relevance of NSS features in the context of NR-VQA for cloud gaming. However, the GAMIVAL model, which uses adaptations of traditional NSS features, similar to the use of neural noise models in [40], along with deep-learning features, produced superior performance on synthetic gaming video content, suggesting the relevance of appropriately modified NSS features for synthetic rendered content. Fig. 10 shows boxplots of the SROCC values computed on the predictions produced by each NR-VQA models, visually illustrating the results reported in Table IV. The two top-performing algorithms VSFA and GAMIVAL exhibit very low variances of SROCC values, suggesting the reliability of these algorithms across multiple train-test splits.

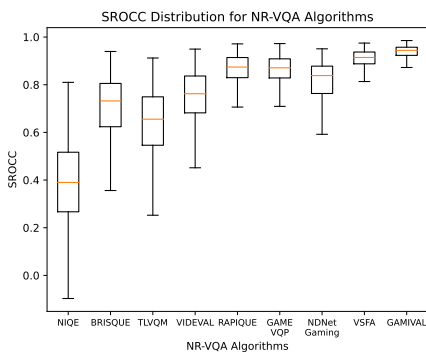


Fig. 10. Boxplots of SROCC distributions of the compared NR-VQA algorithms.

B. Effects of Display Orientation on VQA Performance

The new LIVE-Meta MCG database contains both portrait and landscape videos, allowing us to test the performances of NR-VQA algorithms on different display orientations. We tested the performance of the top-performing algorithms RAPIQUE, GAME-VQP, VSFA, and GAMIVAL on videos of both orientations over 400 train-test splits each. We may conclude from the results shown in Table VI that the NR-VQA algorithms performed slightly better when trained on landscape videos, than on portrait videos. Further, we performed one-sided t-tests using the 400 SROCC and PLCC scores used to report the results in Table VI. We were able to conclude from the results of the tests that the performances of the NR-VQA algorithms were statistically superior when trained on landscape videos than on portrait videos. This could be attributed to the tighter 95% confidence intervals of the MLE-MOS estimates obtained on landscape videos as compared to portrait videos, as discussed in Sec. IV-H. From Tables IV and VI, one may observe that although overall GAMIVAL is the best performing algorithm on the LIVE-Meta MCG database,

VSFA delivered slightly superior performance on the portrait gaming videos.

C. Comparison of Computational Requirements and Runtime

This section analyzes the performance vs. complexity trade-off of the NR-VQA algorithms studied in Section V-A. All of the algorithms were run on a standalone computer equipped with an Intel Xeon E5-2620 v4 CPU running at a maximum frequency of 3 GHz. We used one of the videos from the LIVE-Meta MCG database of 360x720 resolution, upscaled it to the display resolution (1080x2160), and applied the algorithms on it. We report the execution time and the floating-point operations used by each algorithm in Table VII. The algorithms VSFA and NDNet-Gaming were implemented in Python, GAMIVAL was implemented partly in MATLAB and partly in Python, while all the other algorithms were implemented in MATLAB. During the evaluation of deep NR-VQA algorithms, we ensured that the GPU was not used for fair comparison against other algorithms implemented on the CPU. From the results reported in Table VII, none of the tested algorithms implemented in high level prototyping languages like MATLAB/Python run in real-time in their current implementations, however, they may be optimized for specific hardware using low-level languages like C/C++ by effectively exploiting their parallel processing capabilities in an application-specific setup. Based on the arguments presented above, we plotted the performance versus complexity trade-off (SROCC versus FLOPS) for each of the algorithms in Fig. 11. Different orders of magnitude of FLOPS of the NR-VQA algorithms are indicated by distinct colors. The figure shows that the top four algorithms, RAPIQUE, GAME-VQP, VSFA and GAMIVAL, are computationally complex in varying degrees, with RAPIQUE having the lowest computational complexity and VSFA the highest. In addition to being the top-performing algorithm, GAMIVAL is also computationally efficient compared to VSFA and NDNet-Gaming, making it a viable option when evaluating the video quality of Mobile Cloud Gaming.

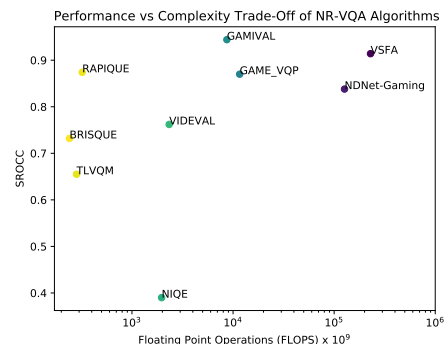


Fig. 11. Comparison of Performance vs Computational Requirement of NR-VQA Algorithms. FLOPS are shown in GigaFlops and shown in log scale.

VI. PERFORMANCE OF FR-VQA ALGORITHMS

In this section, we examine the performances of various Full Reference (FR) VQA models originally developed for

TABLE VIII

MEDIAN SROCC, KRCC, PLCC, AND RMSE OF FR-VQA ALGORITHMS ON THE LIVE-META MOBILE CLOUD GAMING DATABASE OVER 1000 TRAIN-TEST SPLITS (SUBJECTIVE DMOS VS PREDICTED DMOS). STANDARD DEVIATIONS ARE SHOWN IN PARENTHESES. THE BEST PERFORMING ALGORITHM IS BOLD-FACED

Metrics	SROCC(\uparrow)	KRCC(\uparrow)	PLCC(\uparrow)	RMSE(\downarrow)
PSNR	0.7093 (0.0681)	0.5329 (0.0616)	0.7172 (0.0676)	13.1194 (1.2216)
SSIM	0.9235 (0.0301)	0.7647 (0.0435)	0.9332 (0.0313)	6.7599 (1.5737)
MS-SSIM	0.9069 (0.0360)	0.7396 (0.0495)	0.9115 (0.0357)	7.7878 (1.5813)
ST-RRED	-0.8840 (0.0406)	-0.7071 (0.0508)	0.9012 (0.1028)	8.2752 (2.1837)
SpEED-QA	-0.9171 (0.0283)	-0.7528 (0.0389)	0.9070 (0.3196)	8.0244 (4.3767)
ST-GREED	0.8573 (0.0556)	0.6642 (0.0667)	0.8776 (0.0514)	8.9718 (1.8265)
VMAF (v0.6.1)	0.9347 (0.0210)	0.7773 (0.0328)	0.9362 (0.0261)	6.6705 (1.3785)
Gaming VMAF	0.9410 (0.0407)	0.7913 (0.0544)	0.9428 (0.0420)	6.2562 (1.9643)

TABLE IX

RESULTS OF ONE-SIDED T-TEST PERFORMED USING THE 1000 (SROCC, PLCC) VALUES OF THE COMPARED FR-VQA ALGORITHMS COMPUTED ON THE LIVE-META MCG DATABASE. EACH CELL CONTAINS 2 SYMBOLS: THE FIRST SYMBOL CORRESPONDS TO THE T-TEST DONE USING THE SROCC VALUES, AND THE SECOND CORRESPONDS TO THE T-TEST DONE USING THE PLCC VALUES. WHEN A SYMBOL '1' APPEARS, IT DENOTES THAT THE ALGORITHM ON THE ROW WAS STATISTICALLY SUPERIOR TO THAT ON THE COLUMN, WHEREAS '0' INDICATES THAT THE ALGORITHM ON THE COLUMN WAS STATISTICALLY SUPERIOR. A '-' SYMBOL INDICATES THAT THE COLUMN AND ROW ALGORITHMS PERFORMED EQUALLY WELL

ALGORITHM	PSNR	SSIM	MS-SSIM	ST-RRED	SpEED-QA	ST-GREED	VMAF (v0.6.1)	Gaming VMAF
PSNR	(-, -)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
SSIM	(1, 1)	(-, -)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(0, -)	(0, 0)
MS-SSIM	(1, 1)	(0, 0)	(-, -)	(1, 1)	(0, 1)	(1, 1)	(0, 0)	(0, 0)
ST-RRED	(1, 1)	(0, 0)	(0, 0)	(-, -)	(0, 0)	(1, 1)	(0, 0)	(0, 0)
SpEED-QA	(1, 1)	(0, 0)	(1, 0)	(1, 1)	(-, -)	(1, 1)	(0, 0)	(0, 0)
ST-GREED	(1, 1)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(-, -)	(0, 0)	(0, 0)
VMAF (v0.6.1)	(1, 1)	(1, -)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(-, -)	(-, 0)
Gaming VMAF	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(-, 1)	(-, -)

natural videos on our proposed database. Our goal is to assess whether they can be utilized as suitable replacements for mean-opinion scores, or serve as pre-training targets when developing deep NR-VQA models for Mobile Cloud Gaming. Deep learning-based algorithms proposed in [33], [34], [41], [42], [43], [44] have been successfully used for generic No-Reference Video Quality tasks. Most of these deep learning backbones are pre-trained on one of the large natural image and video classification databases like ImageNet, Imagenet-22K [38], Kinetics-400 [45] or benefit from dedicated large databases as in [34]. Developing dedicated deep learning-based models similar to those that involve pre-training on a classification database is complicated in niche VQA sub-domains like Cloud Gaming, due to the absence of large-scale classification datasets comprising rendered gaming content. Furthermore, existing Cloud Gaming VQA databases are too small to support the training of deep learning backbones. To overcome these challenges, researchers working in the Cloud Gaming VQA domain have frequently employed Full Reference VQA algorithms originally developed for generic VQA tasks as substitutes for MOS scores when pre-training complex deep networks for NR-VQA [7], [11], [10]. They achieve this by selecting a popular VQA metric, like VMAF, using it to predict the FR-VQA scores using a pristine gaming video and a synthetically distorted version of the pristine video. The low expense of producing synthetically distorted videos and estimating proxy MOS scores in the form of FR-VQA outputs makes it feasible to create large databases for pre-training deep networks. Once a deep network backbone is pre-trained, most authors [11], [10] fine-tune the pre-trained backbone with a small amount of human-annotated data to achieve better

performance than traditional handcrafted feature-based models on the Cloud Gaming NR-VQA task. It is worth noting that using deep learning backbones pre-trained on natural images and videos may not lead to optimal performance on Cloud Gaming NR-VQA task. This is because the visual content generated by computer graphics, as in Cloud Gaming videos, typically has fewer details and is smoother than naturalistic videos or images, which alters the bandpass statistics of Cloud Gaming videos relative to those of naturalistic videos [12].

Cloud Gaming NR-VQA algorithms [7], [11], [10] usually employ VMAF scores as their pre-training targets. Here, we comprehensively compare the performances of seven FR-VQA algorithms: PSNR, SSIM [46], MS-SSIM [47], ST-RRED [48], SpEED-QA [49], ST-GREED [50], and VMAF on the LIVE-Meta Mobile Cloud Gaming database to explore for their suitabilities as Proxy-MOS or intermediate pre-training targets for the development of NR-VQA models focused on Mobile Cloud Gaming. We calculated the DMOS using equation (3) and the proxy reference videos in our database were used as reference videos when computing the FR-VQA scores. To ensure consistency, we utilized the same 1000 train-test split used for the NR-VQA algorithms in our evaluation of FR-VQA algorithms.

PSNR, SSIM, and MS-SSIM are computed per-frame between the reference and distorted videos, then averaged across all frames. The FR-VQA algorithms PSNR, SSIM, MS-SSIM, ST-RRED, and SpEED-QA algorithms do not require training, and therefore, were directly evaluated on the 1000 test sets. ST-GREED features were obtained from the proxy reference and distorted videos in the training and test sets. The features from the training set and the corresponding DMOS were then

used to train an SVR similar to the NR-VQA algorithms. Once the SVR model was obtained, the features from the test set and the corresponding DMOS scores were used to obtain the performance of the overall algorithm. We also present two versions of VMAF: VMAF (v0.6.1), the pre-trained open source version widely used for generic VQA tasks, and our version of VMAF which we call Gaming VMAF, which uses the same features as VMAF (v0.6.1) but with the SVR trained on the LIVE-Meta MCG database using the same evaluation strategy as ST-GREED. Table VIII summarizes the results obtained for all the FR-VQA algorithms. It may be observed that the VMAF models outperformed the other models, while the computationally less expensive SSIM model also demonstrated competitive performance. Similar to the evaluation of NR-VQA algorithms, we used the 1000 SROCC and PLCC scores produced by the FR VQA models to run one-sided t-tests, using the 95% confidence level to determine whether the performance of one FR-VQA algorithm was statistically superior to another. Each entry in Table IX consists of two symbols, corresponding to the t-tests conducted using the SROCC and PLCC values. Based on the results, we conclude that when comparing the two VMAF models, the use of SROCC as a performance metric did not show statistically significant differences. However, using PLCC revealed statistically significant differences, with Gaming VMAF exhibiting slightly better performance. It may also be concluded that a statistically significant difference exists between the performances of the Gaming VMAF and SSIM models when evaluated using both performance metrics.

The high correlations obtained on the VMAF models suggest that the VMAF models could be reasonably used as proxy-MOS scores or as pre-training targets for MCG NR-VQA models. By pre-training a deep learning model on VMAF scores, a model could potentially learn to extract useful “gaming quality-aware” features on a small human-annotated database like ours, potentially improving performance on the MCG NR-VQA task. However, it is important to note that while pre-training can be beneficial, it may not always result in improved performance. Therefore, it is crucial to exercise caution when selecting a pre-training dataset, the synthetic distortions applied, and the proxy FR-VQA algorithm to ensure that pre-training boosts the performance of the target MCG NR-VQA task. Furthermore, relying on pre-training using a single FR-VQA model presents the potential danger of NR-VQA models adopting the strengths and limitations of that FR-VQA model, leading to reduced NR-VQA generalization. One possible solution would be to convert the pre-training to a Multi-Task Learning problem [51], using multiple FR-VQA algorithms as different tasks. For example, in case of Mobile Cloud Gaming, a combination of VMAF, SSIM and SpEED-QA could be used as multiple tasks to pre-train the deep network backbone. This approach could enable more generalized “quality-aware” representations, which might further enhance performance on the MCG NR-VQA task.

VII. CONCLUSION AND FUTURE WORK

In this work, we have introduced a new psychometric database that we call the LIVE-Meta Mobile Cloud Gaming

(LIVE-Meta MCG) video quality database. It is our hope that this resource helps advance the development of No Reference VQA algorithms directed towards Mobile Cloud Gaming. The new database will be made publicly available to the research community at <https://live.ece.utexas.edu/research/LIVE-Meta-Mobile-Cloud-Gaming/index.html>. We have also demonstrated the usability of the database for comparing, benchmarking and designing NR VQA algorithms. As a next step, algorithms based on traditional natural scene statistics (NSS) models and/or deep-learning methods could be developed to further improve the accuracy of NR-VQA algorithms. In addition, since cloud gaming applications require real-time video quality prediction capability, it is also of utmost interest to develop algorithms capable of running at least in real-time.

We also demonstrated that tighter 95% confidence intervals were obtained on the MLE-MOS estimates of landscape videos than those of portrait videos. A possible research direction could be to explore this dichotomy in further detail. Future work could also focus on development of “Quality of Experience” (QoE) databases comprised of subjective QoE responses to various designs dimensions such as changing bitrates, content-adaptive encoding, network conditions and video content which would further help in the development of perceptually-optimized cloud video streaming strategies, leading to improved mobile cloud gaming experiences.

VIII. APPENDIX

A. Gaming Video Contents in LIVE-META Mobile Cloud Gaming Database

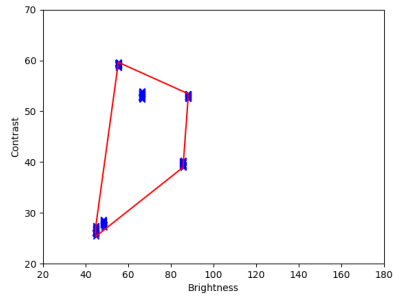
Table X lists the games present in the dataset along with their original resolutions as rendered by the Cloud Game engine. Fig. 12 compares the coverage of a number of objective features, including contrast, brightness, sharpness, colorfulness, spatial information, and temporal information of the videos in our database against the same features computed from other existing Cloud Gaming databases. The content distribution in the paired feature space shows that the coverage of our proposed database is significantly better than all the other three existing cloud gaming databases.

B. Android Application

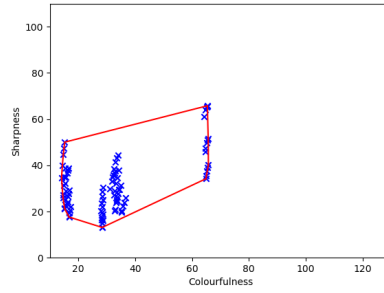
We used a custom developed Android Application to conduct the in-lab subjective study for the development of the LIVE-Meta MCG database. The code will be made publicly available at <https://github.com/avinabsaha/LIVE-Meta-MCG-SubjectiveStudySetup>. Fig. 13 demonstrates the steps involved in the video quality rating process in the Android application.

C. Additional Post Study Questionnaire & Demographics

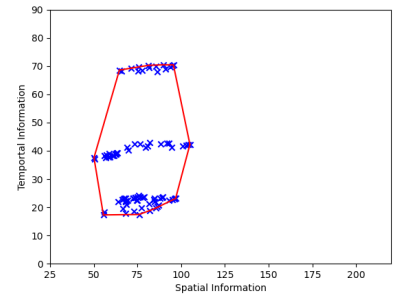
As a part of the post-study questionnaire, we also asked the human subjects about the distribution of videos, the difficulty of rating the videos, and whether they experienced any sort of dizziness or uneasiness while viewing and rating the videos. In the end, in 74.3% (107/144) sessions, the subjects felt that the distribution of quality was uniform with an equal number



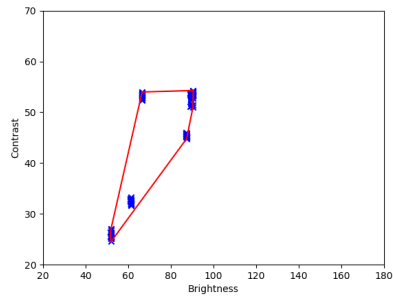
(a) GamingVideoSET : Contrast vs Brightness



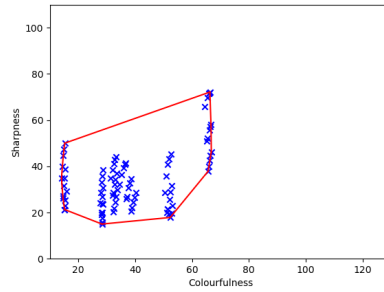
(b) GamingVideoSET : Sharpness vs Colourfulness



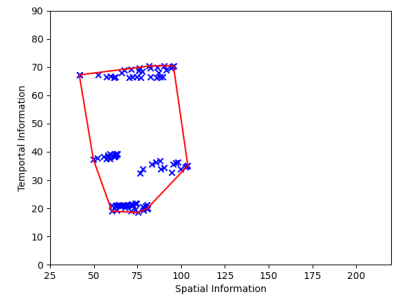
(c) GamingVideoSET : TI vs SI



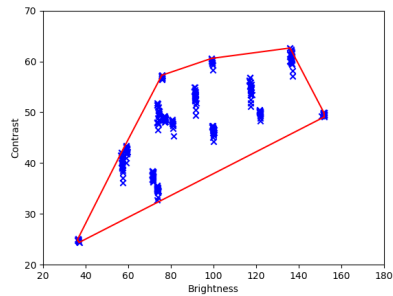
(d) KUGVD : Contrast vs Brightness



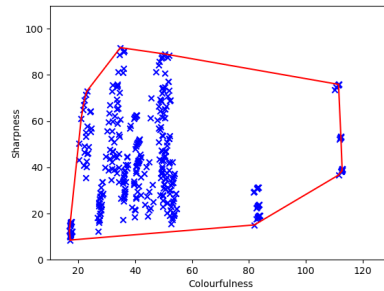
(e) KUGVD : Sharpness vs Colourfulness



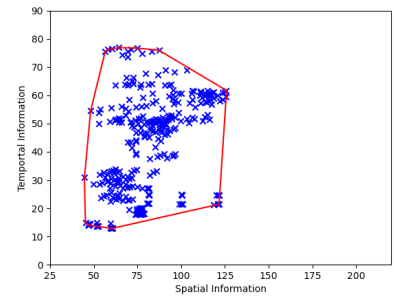
(f) KUGVD : TI vs SI



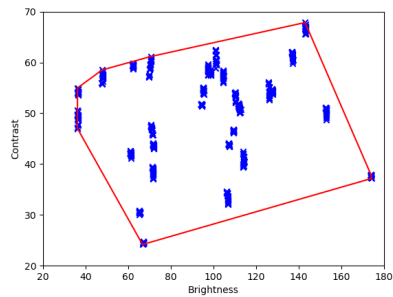
(g) CGVDS : Contrast vs Brightness



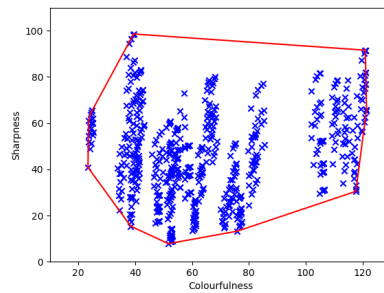
(h) CGVDS : Sharpness vs Colourfulness



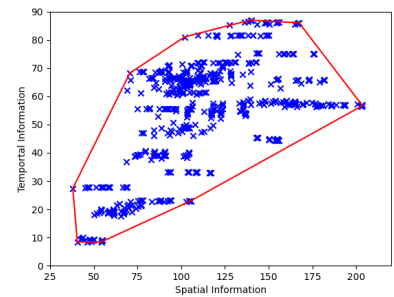
(i) CGVDS : TI vs SI



(j) LIVE-Meta MCG : Contrast vs Brightness



(k) LIVE-Meta MCG : Sharpness vs Colourfulness



(l) LIVE-Meta MCG : TI vs SI

Fig. 12. Source content (blue 'x') distribution in paired feature space with corresponding convex hulls (red boundaries). Left column: Contrast x Brightness, middle column: Sharpness x Colourfulness, right column: Temporal Information (TI) vs Spatial Information (SI) across four Cloud Gaming Databases.

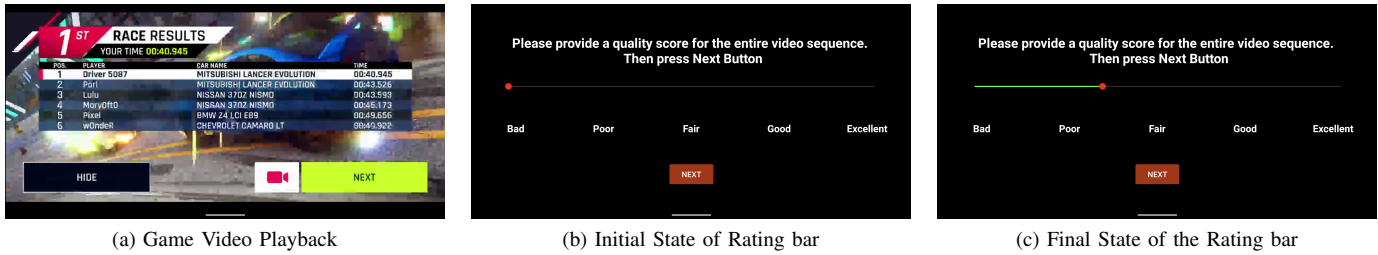


Fig. 13. Video Quality Rating process in our custom-developed Android Application. Left column: A game video playback of duration 20 seconds, Middle Column: Initial state of the rating bar initialized to extreme left, Right Column: Exemplar final state of the rating bar when the user records their final score.

TABLE X
DETAILS OF GAMES PRESENT IN THE PROPOSED LIVE-META MOBILE CLOUD GAMING (LIVE-META MCG) DATABASE

Cloud Games	Original Resolution	Display Orientation
Asphalt	1664 x 720	Landscape
Bejewelled	720 x 1280	Portrait
Bowling Club	720 x 1440	Portrait
Design Island	1664 x 720	Landscape
Dirt Bike	720 x 1440	Portrait
Dragon Mania Legends	1440 x 720	Landscape
Hungry Dragon	1512 x 720	Landscape
Mobile Legends Adventure	1440 x 720	Landscape
Monument Valley 2	720 x 1280	Portrait
Mystery Manor	1728 x 720	Landscape
PGA Golf Tour	720 x 1280	Portrait
Plants vs Zombies	1280 x 720	Landscape
Solitaire	1664 x 720	Landscape
Sonic	720 x 1280	Portrait
State of Survival	1664 x 720	Landscape
WWE	720 x 1440	Portrait

of good, intermediate and bad quality videos. In the other sessions, the subjects felt that the majority of the videos were either of very good or very bad quality, and few, if any of the videos were of intermediate quality. On a scale from 0 to 100, we asked the subjects to rate the difficulty of judging the perceptual quality of the video after each session, with 0 being very difficult and 100 being reasonably easy to judge. All of the subjects were able to provide subjective quality ratings without much difficulty, as reflected by the mean and median scores of difficulty, which were 72.1 and 77.5, respectively. The human subjects reported that they felt slight dizziness or uneasiness in approximately 11% of the sessions, however the percentage of dizziness or uneasiness inducing videos was much lower. More detailed results from the survey regarding dizziness and uneasiness can be found in Table XI.

The demographic data of age and gender were collected only at the end of the first session. The mean, median, and standard deviation of the ages of the participants were found to be 23.57, 23.0, and 3.04. We summarize the gender distribution among the participants in Table XII.

D. Group-wise Inter-Subject and Intra-Subject Consistency

We report the inter-subject and intra-subject consistency scores for each of the subject groups in Table XIII using the methodology described in Section IV-G of the main paper. Across subject groups, the SROCC scores for inter-subject consistency ranged from 0.900 to 0.936 with an average of

TABLE XI
OPINIONS OF STUDY PARTICIPANTS REGARDING THE PERCENTAGE OF GAMING VIDEOS THAT INDUCED DIZZINESS/UNEASINESS

% of Gaming videos inducing dizziness/uneasiness	None	<10%	10-20%	20-40%	>40%
# of sessions	128 (88.89%)	6 (4.16%)	7 (4.86%)	3 (2.08%)	0 (0%)

TABLE XII
DEMOGRAPHICS OF HUMAN STUDY PARTICIPANTS BASED ON GENDER

Gender	Male	Female	Others	Prefer Not to Say
Count(%)	58(80.55%)	11(15.27%)	2(2.72%)	1(1.36%)

0.912, while PLCC scores ranged from 0.915 to 0.949 with an average of 0.929. The SROCC scores for intra-subject consistency ranged from 0.827 to 0.866 with an average of 0.848, while PLCC scores ranged from 0.844 to 0.870 with an average of 0.860. These scores reflect the consistency of our data acquisition process across all the subject groups.

E. Additional Analysis and Visualization of Opinion Scores

Fig. 14, examines the interplay of source video content and bitrate and how these together affect MLE-MOS. To obtain the plot, we separately calculated the average MLE-MOS ratings of each of the 30 source sequences on a per-bitrate basis across all available resolutions. Fig. 14 shows a clear separations between the MLE-MOS curves of all the contents, except at very high bitrates. Across contents, however, the curves are commingled, which is a good illustration of the difficulty of the VQA problem (it is not just about bitrate). The variation of MLE-MOS for all contents was greatly reduced at bitrates of 2 mbps or higher as compared to lower bitrates. Clearly, as shown in prior studies, the effect of video

TABLE XIII
SUBJECT CONSISTENCY

Subject Group	Inter-Subject Consistency		Intra-Subject Consistency	
	SROCC	PLCC	SROCC	PLCC
1	0.901	0.915	0.850	0.870
2	0.900	0.917	0.840	0.854
3	0.905	0.920	0.849	0.870
4	0.913	0.941	0.827	0.844
5	0.916	0.933	0.866	0.859
6	0.936	0.949	0.854	0.865

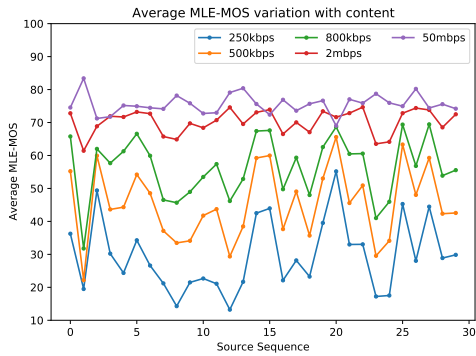


Fig. 14. Variation of average MLE-MOS against content for five fixed bitrates.

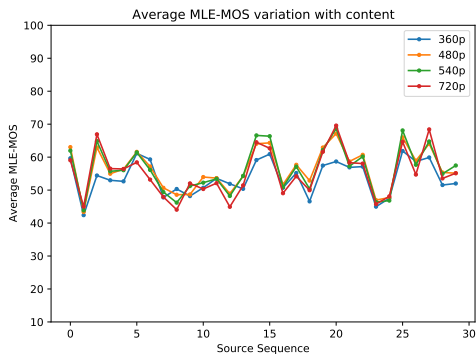


Fig. 15. Variation of average MLE-MOS against content for four fixed resolutions.

compression induced distortions on perceptual video quality is highly content-dependent because of perceptual masking and similar processes.

Fig. 15 shows the effects of video source content on MLE-MOS, across all bitrates for each of the fixed four resolutions. Specifically, we plotted the average MLE-MOS scores of the encoded videos over the five different bitrates associated with each resolution in the database. As may be observed, there was no strong separation between the MLE-MOS curves, although the content did cause notable differences in the reported video qualities. A salient takeaway from these two analyses is that video compression has a heavier impact on the visual perception of video quality than does resizing, at least on gaming videos. This further suggests the efficacy of resizing to achieve data efficiencies with little perceptual loss in the context of mobile gaming video streaming.

ACKNOWLEDGMENT

The authors would thank all the volunteers who took part in the human study. The authors also acknowledge the Texas Advanced Computing Center (TACC), at the University of Texas at Austin for providing HPC, visualization, database, and grid resources that have contributed to the research results reported in this paper. URL: <http://www.tacc.utexas.edu>

CHANGE LOG

- v1 Uploaded to Arxiv on 26th May, 2023.

REFERENCES

- [1] “Cloud Gaming Market by offering (Infrastructure and Gaming Platform Service), Device Type (Smartphones, Tablets, Gaming Consoles, PCs & Laptops, Smart TVs, and HMDs), and Solution (File streaming and video streaming): Global Opportunity Analysis and Industry Forecast, 2021–2030.” <https://www.alliedmarketresearch.com/cloud-gaming-market-A07461>, 2021, [Online; accessed 30-January-2022].
- [2] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, “Gamingvideaset: A dataset for gaming video streaming applications,” *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, pp. 1–6, 2018.
- [3] N. Barman, E. Jammeh, S. A. Ghorashi, and M. G. Martini, “No-reference video quality estimation based on machine learning for passive gaming video streaming applications,” *IEEE Access*, vol. 7, pp. 74 511–74 527, 2019.
- [4] S. Zadtootaghaj, S. Schmidt, S. S. Sabet, S. Möller, and C. Griwodz, “Quality estimation models for gaming video streaming services using perceptual video quality dimensions,” in *Proceedings of the 11th ACM Multimedia Systems Conference*, ser. MMSys ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 213–224. [Online]. Available: <https://doi.org/10.1145/3339825.3391872>
- [5] S. Wen, S. Ling, J. Wang, X. Chen, L. Fang, Y. Jing, and P. L. Callet, “Subjective and objective quality assessment of mobile gaming video,” *ArXiv*, vol. abs/2103.05099, 2021.
- [6] X. Yu, Z. Tu, Z. Ying, A. C. Bovik, N. Birkbeck, Y. Wang, and B. Adsumilli, “Subjective quality assessment of user-generated content gaming videos,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 74–83.
- [7] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, and S. Möller, “Nr-gvqm: A no reference gaming video quality metric,” *2018 IEEE International Symposium on Multimedia (ISM)*, pp. 131–134, 2018.
- [8] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward a practical perceptual video quality metric,” vol. 6, 2016, p. 2.
- [9] S. Göring, R. R. Ramachandra Rao, and A. Raake, “nofu -a lightweight no-reference pixel based video quality model for gaming content,” 06 2019.
- [10] M. Utke, S. Zadtootaghaj, S. Schmidt, S. Bosse, and S. Moeller, “NDNetGaming - Development of a No-Reference Deep CNN for Gaming Video Quality Prediction,” in *Multimedia Tools and Applications*. Springer, 2020.
- [11] S. Zadtootaghaj, N. Barman, R. R. R. Rao, S. Göring, M. G. Martini, A. Raake, and S. Möller, “Dem: Deep video quality estimation model using perceptual video quality dimensions,” in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSp)*, 2020, pp. 1–6.
- [12] Y.-C. Chen, A. Saha, C. Davis, B. Qiu, X. Wang, R. Gowda, I. Katsavounidis, and A. C. Bovik, “Gamival: Video quality prediction on mobile cloud gaming content,” *IEEE Signal Processing Letters*, vol. 30, pp. 324–328, 2023.
- [13] G. Huang, Z. Liu, and K. Q. Weinberger, “Densely connected convolutional networks,” *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: <http://arxiv.org/abs/1608.06993>
- [14] *Opinion model predicting gaming quality of experience for cloud gaming services*, document ITU-T recommendation G.1072, 2020.
- [15] S. Schmidt, S. Möller, and S. Zadtootaghaj, “A comparison of interactive and passive quality assessment for gaming research,” in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, 2018, pp. 1–6.
- [16] D. Ghadiyaram, J. Pan, and A. C. Bovik, “A subjective and objective study of stalling events in mobile streaming videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 183–197, 2019.
- [17] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, “Study of temporal effects on subjective video quality of experience,” *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5217–5231, 2017.
- [18] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, “Towards perceptually optimized end-to-end adaptive video streaming,” 2018. [Online]. Available: <https://arxiv.org/abs/1808.03898>
- [19] *Subjective evaluation methods for gaming quality*, document ITU-T Recommendation P.809, 2018.
- [20] D. Hasler and S. E. Suesstrunk, “Measuring colorfulness in natural images,” in *Human Vision and Electronic Imaging VIII*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 5007, International Society for Optics

- and Photonics. SPIE, 2003, pp. 87 – 95. [Online]. Available: <https://doi.org/10.1117/12.477378>
- [21] S. Winkler, “Analysis of public image and video databases for quality assessment,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.
- [22] *Subjective video quality assessment methods for multimedia applications*, document ITU-T recommendation P.910, 2008.
- [23] “NVENC Video Encoder API Programming Guide,” <https://docs.nvidia.com/video-technologies/video-codec-sdk/nvenc-video-encoder-api-prog-guide/>, 2021, [Online; accessed 30-January-2022].
- [24] “Google Pixel 5 Display Review: Worthy of a Flagship,” https://www.xda-developers.com/google-pixel-5-display-review/#color_accuracy, 2021, [Online; accessed 19-February-2023].
- [25] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Recommendation BT. 500-13, 2012.
- [26] “Visual Screening, Laboratory of Image and Video Engineering,” <https://live.ece.utexas.edu/research/Quality/visualScreening.htm>, [Online; accessed 30-January-2022].
- [27] Z. Li and C. G. Bampis, “Recover subjective quality scores from noisy measurements,” *CoRR*, vol. abs/1611.01715, 2016. [Online]. Available: <http://arxiv.org/abs/1611.01715>
- [28] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Dieopold, and P. Tran-Gia, “Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing,” *Multimedia, IEEE Transactions on*, vol. 16, pp. 541–558, 02 2014.
- [29] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [30] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [31] J. Korhonen, “Two-level approach for no-reference consumer video quality assessment,” *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.
- [32] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, “Ugc-vqa: Benchmarking blind video quality assessment for user generated content,” *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.
- [33] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, “RAPIQUE: rapid and accurate video quality prediction of user generated content,” *CoRR*, vol. abs/2101.10955, 2021. [Online]. Available: <https://arxiv.org/abs/2101.10955>
- [34] D. Li, T. Jiang, and M. Jiang, “Quality assessment of in-the-wild videos,” *CoRR*, vol. abs/1908.00375, 2019. [Online]. Available: <http://arxiv.org/abs/1908.00375>
- [35] X. Yu, Z. Ying, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “Subjective and objective analysis of streamed gaming videos,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.12824>
- [36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [37] K. Cho, B. van Merriënboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *CoRR*, vol. abs/1406.1078, 2014. [Online]. Available: <http://arxiv.org/abs/1406.1078>
- [38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [39] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.
- [40] Y. Jin, A. Patney, R. Webb, and A. C. Bovik, “FOVQA: blind foveated video quality assessment,” *CoRR*, vol. abs/2106.13328, 2021. [Online]. Available: <https://arxiv.org/abs/2106.13328>
- [41] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, “Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild,” *IEEE Access*, vol. 9, pp. 72 139–72 160, 2021.
- [42] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, “Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling,” in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*. Springer, 2022, pp. 538–554.
- [43] A.-X. Zhang, Y.-G. Wang, W. Tang, L. Li, and S. Kwong, “Hvs revisited: A comprehensive video quality assessment framework,” 2022. [Online]. Available: <https://arxiv.org/abs/2210.04158>
- [44] Z. Ying, M. Mandal, D. Ghadiyaram, and A. C. Bovik, “Patch-vq: ‘patching up’ the video quality problem,” *CoRR*, vol. abs/2011.13544, 2020. [Online]. Available: <https://arxiv.org/abs/2011.13544>
- [45] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” *CoRR*, vol. abs/1705.06950, 2017. [Online]. Available: <http://arxiv.org/abs/1705.06950>
- [46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [47] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.
- [48] R. Soundararajan and A. C. Bovik, “Video quality assessment by reduced reference spatio-temporal entropic differencing,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2012.
- [49] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, “Speed-qa: Spatial efficient entropic differencing for image and video quality,” *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1333–1337, 2017.
- [50] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “ST-GREED: space-time generalized entropic differences for frame rate dependent video quality prediction,” *CoRR*, vol. abs/2010.13715, 2020. [Online]. Available: <https://arxiv.org/abs/2010.13715>
- [51] M. Crawshaw, “Multi-task learning with deep neural networks: A survey,” *CoRR*, vol. abs/2009.09796, 2020. [Online]. Available: <https://arxiv.org/abs/2009.09796>