This is the accepted version of this paper. The version of record is available at https://doi.org/10.1109/MMSP48831.2020.9287055

A Large-scale Evaluation of the bitstream-based video-quality model ITU-T P.1204.3 on Gaming Content

Rakesh Rao Ramachandra Rao*, Steve Göring*, Robert Steger*, Saman Zadtootaghaj†
Nabajeet Barman‡, Stephan Fremerey*, Sebastian Möller† and Alexander Raake*
*Audio Visual Technology; Technische Universität Ilmenau; Germany
Email: [rakesh-rao.ramachandra-rao, steve.goering, robert.steger, stephan.fremerey, alexander.raake]@tu-ilemnau.de
†Quality and Usability Lab; TU Berlin; Germany
Email: [saman.zadtootaghaj, sebastian.moeller]@qu.tu-berlin.de
‡Kingston University; London; UK; Email: n.barman@kingston.ac.uk

***Abstract*—The streaming of gaming content, both passive and interactive, has increased manifolds in recent years. Gaming contents bring with them some peculiarities which are normally not seen in traditional 2D videos, such as the artificial and synthetic nature of contents or repetition of objects in a game. In addition, the perception of gaming content by the user is different from that of traditional 2D videos due to its pecularities and also the fact that users may not often watch such content. Hence, it becomes imperative to evaluate whether the existing video quality models usually designed for traditional 2D videos are applicable to gaming content. In this paper, we evaluate the applicability of the recently standardized bitstream-based video-quality model ITU-T P.1204.3 on gaming content. To analyze the performance of this model, we used 4 different gaming datasets (3 publicly available + 1 internal) not previously used for model training, and compared it with the existing state-of-the-art models. We found that the ITU P.1204.3 model out of the box performs well on these unseen datasets, with an RMSE ranging between $0.38 - 0.45$ on the 5-point absolute category rating and Pearson Correlation between $0.85 - 0.93$ across all the 4 databases. We further propose a full-HD variant of the P.1204.3 model, since the original model is trained and validated which targets a resolution of 4K/UHD-1. A 50:50 split across all databases is used to train and validate this variant so as to make sure that the proposed model is applicable to various conditions.**

***Index Terms*—video quality, gaming, bitstream models, pixel models**

I. Introduction

Gaming-video streaming can be broadly classified into two scenarios: An interactive gaming scenario represented by cloud gaming, where the gaming scenes are rendered on a server and streamed to the players with ideally lowest possible latency, and a passive gaming scenario, in which a viewer can watch the gameplay of other players [20]. Owing to the advances made in both passive and interactive online gaming services, there has been a tremendous increase in gaming-video popularity in recent years. For the particular case of the passive gaming scenario, the popularity and attention gained by gaming video streaming services such as Twitch.tv or YouTubeGaming has led to widespread viewing of passive gaming content. For example, Twitch.tv alone is, with its nine million subscribers and about 800k active viewers at the same time, responsible for the 4th highest peak Internet traffic in the US [6].

Both scenarios come with their own requirements and associated effects in terms of user-perceived quality of experience (QoE). For example: Passive gaming via popular streaming services such as Twitch.tv and YouTubeGaming encounter quality-related effects typical of HAS (HTTP-based adaptive streaming), the streaming technology used by these platforms. These issues include quality switching, initial loading delay and stalling, with playout stopping until the playout buffer is filled again. In turn, cloud-gaming sessions are affected by impairments due to the delays caused by coding and transmission, or impairments due to packet loss, low network bandwidth etc. As a consequence, any video-quality model to be used in one of the two scenarios should be sensitive to the corresponding scenario-specific requirements.

In addition to this, a video-quality model suitable for gaming should be able to take into account the uniqueness of this type of video such as the artificial and synthetic nature of contents and repetitions of objects in a game. Classical video codecs like H.264, H.265 or VP9 are usually not optimized for the specific properties of gaming contents, which leads to additional challenges for compression [2]. Moreover, the unique nature and the lack of exposure to gaming contents results in these videos being perceived differently by the average users [3].

Several studies have assessed the suitability of the existing state-of-the-art (SoA) full-reference (FR) quality metrics and models such as VMAF [15], VIFP, SSIM [17] or PSNR for evaluating gaming content [2, 5, 3]. However, full-reference (FR) models are not suitable for gaming contents since the game sessions are recorded by the player and then streamed, thereby lacking the pristine quality reference video that is required to compute quality scores. Here, no-reference (NR) quality models may be more suitable for gaming video streaming, as they provide video-quality predictions solely from the processed pixel information. This has led to studies

evaluating the performance of existing no-reference metrics namely BRISQUE [12], NIQE [13] and BIQI [14] on gaming content [19], showing promising results. Moreover, several gaming-specific NR models with good prediction performance have been developed and evaluated, e.g. nofu [8], NR-GVQM [19] and NR-GVSQI [1].

Besides pixel-based NR models, bitstream-based NR models have also been studied for gaming content. For example, prediction performance of the standardized P.1203 series of models [10] have been analyzed in recent studies [21], showing good performance.

In this paper, we evaluate the recently standardized bitstream-based video-quality model P.1204.3 [16, 11] on gaming content. For this purpose, we consider four different datasets, namely, GamingVideoSET [1], KUGVD [4] and CGVDS [21], which are open datasets, and a Twitch dataset which is an in-house created dataset using game recordings from Twitch.tv. All four considered datasets use full-hd (FHD) as target resolution. Since the P.1204.3 model was initially developed for 4K/UHD as target screen resolution, an adaptation or mapping to the gaming datasets is required. A first approach is based on a per-database linear mapping of the P.1204.3 scores (out of the box) to the subjective test scores [9]. A second approach consists in the development of a dedicated adaptation of P.1204.3 targeting FHD resolution video. Both approaches are evaluated and compared in the paper.

The remaining part of the paper is organized as follows. Sec. II provides an overview of the evaluation of "traditional" and gaming-specific video-quality models on gaming content. Following this, in Sec. III, the four gaming datasets that were used to evaluate the standardized P.1204.3 model on gaming content are described. The evaluation of the P.1204.3 model and the comparison with the existing SoA metrics on the four gaming-video datasets is presented in Sec. IV. Finally, in Sec. V we conclude with a discussion and an outlook on future work.

II. Related Work

As described in Sec. I, there have been several studies evaluating existing video-quality models developed for classical 2D video content on gaming content, or proposing newer gaming-specific video quality models. In this section, we focus on analyzing the advantages and drawbacks of the SoA models on gaming content, and the added value of the gaming-specific video quality models. Here, we focus on both pixel- and bitstream-based models.

One of the currently most widely used pixel-based FR models for predicting the quality of classical 2D videos is VMAF [15]. In one of these studies, Barman et al.[5] analyzed the performance of VMAF on gaming content, reporting a Pearson Corelation Coefficient (PCC) of 0.87 with the subjective test data used. In addition, other models were evaluated for gaming video, namely the FR metrics SSIM and PSNR, the reduced-reference (RR) models ST-RREDOpt and SpEEDQA, and the following NR metrics, BRISQUE, NIQE and BIQI. In the corresponding papers, it was found that the FR metrics showed the best performance. Other studies [8, 21], performed by Göring et al., Zadtootaghaj et al., also consider VMAF in case of gaming content and confirm the values reported by Barman et al. [5]. Barman et al. [3] report a high performance of 0.927 PCC for VMAF, with VIFP being the next best-performing FR metric, with a PCC of 0.859. A similar analysis was performed for FR metrics such as VIFP, SSIM and PSNR by Barman et al. [2] for videos encoded with H.264, H.265 and VP9. Although VMAF shows good performance on gaming content, it suffers from two main drawbacks, namely, a) it is an FR model, however a reference is usually unavailable in case of gaming video streaming and b) it does not have explicit features to take into account the uniqueness of gaming content such as the artificial/synthetic nature of the content, repetition of objects in the scene and more.

To overcome the problem of the unavailability of the reference video, the evaluation of several NR models has been reported in the literature. One of these studies was conducted by Barman et al. [5]. BRISQUE, NIQE, BIQI are the NR metrics that are considered in their analysis, and it was concluded that they show a bad performance when applied to gaming content out-of-the-box and a corresponding retraining is required to improve the prediction accuracy.

Another analysis of NR models was performed by Zadtootaghaj et al. [21]. In their study they report similary poor performance of the considered NR metrics such as BRISQUE and NIQE on gaming video as reported by Barman et al.[5].

Göring et al. [8] propose a retrained model using BRISQUE and NIQE as the underlying features and report performance comparable to FR metrics, and use the retrained NR-model as their baseline comparison for their novel model introduced in that paper. This need for retraining of existing NR models for gaming video was also indicated by Barman et al. [5].

To tackle both the problems of lack of availability of reference video and gaming-specific adaptation of the models, several NR gaming-specific video quality models have been reported in the literature. NR-GVQM proposed by Zadtootaghaj et al. [19] and NR-GVSQE, NR-GVSQI proposed by Barman et al. [1] are examples of such models. All these models use existing NR metrics such as BRISQUE, NIQE and BIQI as features and combine them with additional features to estimate impairments such as e.g blockiness, blurriness, contrast, exposure. NR-GVQM considers the GamingVideoSet dataset [4] for training and validation and the performance is comparable to VMAF. The NR-GVQSE and NR-GVQSI models use the GamingVideoSet [4] and KUGVD [1] for model training and validation. NR-GVQSI achieved a performance of 0.87 PCC on the GamingVideoSet and 0.89 on KUGVD. NR-GVQSE was designed as the NR equivalent of VMAF (i.e. using VMAF as groundtruth) and showed a correlation of 0.97 when compared with VMAF.

Furthermore, Göring et al. [8] describe gaming-video-specific features to estimate staticness, blockiness, blockmotion etc. and propose the NR model "nofu", reporting a performance of 0.96 PCC using 10-fold cross-validation on

the publicly available GamingVideoSet [4].

In addition to studying the efficacy of the pixel-based metrics, the suitability of bitstream models for predicting the quality of gaming content has also been analyzed. As an example, Zadtootaghaj et al. [21] perform an analysis of the P.1203 [10] series of models on the CGVDS dataset and report that mode 3 shows good performance with 0.88 PCC and 0.48 RMSE on the 5-point scale. They also propose a bitstream model using perceptual dimensions such as video discontinuity, video fragmentation and video unclearness and report a PCC of 0.78 and RMSE of 0.39. However, the P.1203 series was originally developed for H.264 encoded videos with a maximum resolution of Full HD, whereas the recently standardized P.1204.3 model covers modern video codecs (H.264, H.265, VP9) and is trained for UHD-1/4K content.

To summarize, existing FR models like VMAF and VIFP perform well even on gaming videos despite not being developed for this particular use case. However, they suffer from the drawback that they are slow in terms of computation time and the lacking availability of the reference video. The newly proposed gaming-specific NR models such as NR-GVQM [19], NR-GVQSI, NR-GVQSE [1] and nofu [8] show a significant improvement in performance compared to the traditional NR metrics such as BRISQUE, NIQE and BIQI. Also, bitstream-based NR metrics can be used to ensure good performance if applied to the particular codecs the model was developed for.

III. Datasets

We consider four different datasets, namely the three publicly available datasets GamingVideoSet, KUGVD, CGVDS, and a self-developed proprietary Twitch dataset. In the following, the datasets are described in more detail.

A. GamingVideoSet (GVS)

The GamingVideoSet (GVS) [4] consists of 24 reference videos of 30 s duration from 12 different games with a framerate of 30 fps. Three different resolutions, namely, 480p, 720p and 1080p were considered with a total of 24 different bitrates across these resolutions. This resulted in a total of 576 different processed video sequences (PVS). A subjective test with 90 PVSs based on 6 source contents and 15 different resolution-bitrate pairs was conducted using the absolute category rating (ACR) scale, following the procedure outlined in ITU-T BT.500 with FHD as target resolution.The videos were encoded at Constant Bitrate (CBR) at a fixed resolution with **veryfast** preset using the ffmpeg x264 encoder. A total of 25 subjects participated in this test.

B. Kingston University Gaming Video Dataset (KUGVD)

The Kingston University Gaming Video Dataset (KUGVD) [1] was developed using 6 out of the 24 source videos from the GamingVideoSet. All 24 different resolution-bitrate pairings defined in the GamingVideoSet paper [4] were used to generate the PVSs. This resulted in a total of 144 PVSs and finally 90 PVSs with the same resolution-bitrate pairs as in GamingVideoSET were selected for the subjective test. As in the GamingVideoSet, CBR encoding was used in this dataset, too. The reason to go for the same encoding settings as in the GamingVideoSet was to make sure not to introduce any new impairments in this dataset since these two datasets were used in the training and validation of the NR-GVQSI and NR-GVQSE models. In total, 17 subjects participated in this test. Subjective test was conducted with FHD as the target resolution.

C. Cloud Gaming Video Dataset (CGVDS)

Compared to the aforementioned datasets, the Cloud Gaming Video Dataset (CGVDS) [21] consists of a larger number of games, i.e. 15, and also includes videos captured at 60fps. Similar to the previously discussed two datasets, three different resolutions, namely, 480p, 720p and 1080p are considered at three different framerates of 20, 30 and 60fps. A total of 17 bitrate conditions spread across all the resolutions are used in the design of theis dataset. Unlike the GamingVideoSet and KUGVD datasets, this dataset uses a hardware accelerated implementation of H.264/MPEG-AVC (NVENC) because most of the cloud providers use these for delay-sensitive cloud gaming services. A CBR mode of encoding with the preset of **llhq** (low latency, high quality) was used to encode the videos. 5 different subjective tests were conducted to make sure all 15 games were addressed, using 3 video sequences as anchor conditions. Each subjective test had a total of 72 PVSs using a display with FHD resolution. Over 100 subjects participated across all tests with a minimum of 20 subjects for each test.

D. Twitch Dataset

The last considered dataset, referred to as Twitch Dataset, was created with the initial aim of using it for genre classification, hence, due effort was spent to make sure that the dataset comprises gaming videos of different genres. This dataset consists of a total of 36 different games, with each genre being represented by 6 games. The genres were chosen based on their relevance and popularity on Twitch. Three different streamers were recorded three times per game to maintain high diversity for each game. A total of 351 video sequences with a duration of approximately 50 s spanning all representation levels were downloaded from Twitch. This was done to ensure the usage of real-world encodings in the subjective test. A subset of 90 sequences out of the 351 sequences were used in the test. Only the first 30 s of each video in the chosen subset were shown to the test subjects to maintain a fixed duration of one hour for the test. All 36 games from the original dataset are represented in the test with either two or three streamers. Resolutions of 160p, 360p, 480p, 720p, 900p and 1080p and framerates of 30 and 60 fps were used. The encoding scheme was the one used in Twitch.tv since the encoded representations were directly downloaded from Twitch. A total of 29 subjects participated in the test. One outlier was detected using a criteria of 0.75 PCC and was removed from further analysis.

TABLE I
OVERVIEW OF THE DATASETS

Parameter	GVS	KUGVD	CGVDS	Twitch
No. of sources	6	6	15	36
No. of PVS's	90	90	72 * 5	90
Resolution	480p, 720p, 1080p	480p, 720p, 1080p	480p, 720p, 1080p	160p, 360p, 480p, 720p, 900p, 1080p
Framerate (fps)	30	30	20, 30, 60	30, 60
Duration (s)	30	30	30	30
Encoder	ffmpeg x264	ffmpeg x264	ffmpeg NVENC (H.264)	H.264
Encoding mode	CBR	CBR	CBR	Twitch default
Preset	veryfast	veryfast	llhq	Twitch default
No. of subjects	25	17	>100 (5 tests)	29

IV. EVALUATION

THe following section is divided into two parts. In the first part, we describe two approaches to develop a FHD-mapped version of the P.1204.3 model. In the second part, we compare the performance of these two model variants with the existing SoA FR, RR and NR models.

A. FHD-mapped P.1204.3 model

Before we report the performance of the P.1204.3 and the FHD-mapped P.1204.3 models, we will present and motivate the FHD-mapped P.1204.3 version that we propose. The standardized P.1204.3 model was trained and validated on two different target devices, namely, a TV/PC monitor with 3840×2160 and a mobile/tablet (MO/TA) with 2560×1440 as the two target resolutions. Hence, the corresponding $scale\_factor$ that is used in P.1204.3 to determine the "upscaling degradation" is specified differently for PC/TV and MO/TA, as given in Equations 1 and 2, respectively [16]:

$$scale\_factor = \frac{coding\_resolution}{3840 \cdot 2160} \quad for\ PC/TV \quad (1)$$

$$scale\_factor = \frac{coding\_resolution}{2560 \cdot 1440} \quad for\ MO/TA \quad (2)$$

All described gaming datasets use PC/TV as target device, and hence only the PC/TV case was used for the FHD-mapped P.1204.3 version. As can be seen from Equation 1, the normalization of the $coding\_resolution$ is done w.r.t the display resolution of 3840×2160. This is expected to lead to over-predicting the upscaling degradation when a lower resolution video is considered, that in the actual test was presented on an FHD screen rather than a 4K/UHD screen. For example: If a Full HD video is considered, the upscaling degradation should be 0 since the coding resolution of the video matches the display resolution used in the tests. But, if we use the original $scale\_factor$ definition, this would result in a finite non-zero upscaling degradation which is actually not the case. In a similar manner, the relative perception of other lower resolutions changes with the target display resolution.

A first approach for FHD-mapping consists in a per-database linear mapping of the P.1204.3 scores (out of the box) to the subjective test scores as proposed in [9].

In the second approach, we focus on developing a dedicated adaptation of P.1204.3 targeting FHD resolution. Here, we propose a correction factor to account for the overly strong handling of the upscaling degradation part by the original model when applying it to FHD resolution. This correction factor is referred to as $D_{u\_corr\_fac}$ and is defined in Equation 3

$$D_{u\_corr\_fac} = a * log\left(b * \left(\frac{coding\_resolution}{1920 * 1080}\right)\right) \quad (3)$$

where $coding\_resolution = coding\_height * coding\_width$ and log is the natural logarithm.

Hence, the final prediction of the P.1204.3 model is adjusted using $D_{u\_corr\_fac}$ as defined in Equation 3 to obtain a final FHD-mapped prediction. This is represented in Equation 4

$$pred_{hd\_mapped} = pred_{p1204\_3} + D_{u\_corr\_fac} \quad (4)$$

where, $pred_{p1204\_3}$ is the output of the standardized P.1204.3 model. The additive term reflects the overall architecture of the P.1204.3 model, considering the overly strong handling of the upscaling effect when applying P.1204.3 to FHD. With this approach, the original P.1204.3 model could be kept unchanged.

For training this correction factor $D_{u\_corr\_fac}$, we split the 4 datasets into a training and a validation set. We consider GamingVideoSet and KUGVD as the training datasets, which have a total of 24 encoding conditions (i.e. bitrate and resolutions). These two datasets consider 12 different sources in total which are encoded at 3 different resolutions (480p, 720p and 1080p) to result in a combined total of 180 PVSs (90 + 90). The remaining two datasets, namely, the CGVDS and Twitch dataset were used as validation datasets.

The final coefficient values (cf. Equation (3)) after the training procedure are: $a = -0.10756695$ and $b = 0.08303269$.

B. Performance Analysis

In this section, we present the performance evaluation of both variants of the P.1204.3 model, namely, the standardized version and the FHD-mapped version, and compare these with the SoA FR, RR and NR models. A common question regarding applicability of the model arises considering the sequence duration used in the 4 datasets of 30 s, because the P.1204.3 model was trained and validated on shorter video sequences of 7-9 s duration [16]. However, the study by Fröhlich et al.[7] on the effect of duration on quality ratings indicates that there is no significant difference for sequences of between 10-30 s duration. Thus, we chose to use the average across the entire 30 s duration to obtain the final quality score, instead of a more sophisticated aggregation of the per-1-second quality scores that is estimated by the P.1204.3 model in addition to the per-segment score.

To remove the bias between the subjective tests across different datasets without changing the predicted rank-order, we apply a linear (i.e. first-order) mapping [9] per dataset to the objective scores obtained from all objective models to the subjective scores before computing the performance evaluation metrics.

We report the performance in terms of five different metrics, namely, RMSE, PCC, Spearman Rank Order Correlation Coefficient (SROCC), Kendall rank correlation, and the $R^2 - score$

(R^2). In the following sections, this performance evaluation is reported for each individual dataset. It should be noted that for the GamingVideoSet and KUGVD datasets, the performance of the FHD-mapped P.1204.3 model (using Equation 4) are self-validated results since these two datasets were used to train the correction part of the model.

1) GamingVideoSet: Table II shows the performance of the P.1204.3 model along with the SoA FR, RR and NR metrics. The considered SoA metrics correspond to the ones presented in [1]. The performance measures for these metrics were calculated using the data that has been open-sourced as part of GamingVideoSet [4]. It can be observed from the results that the NR P.1204.3 model outperforms all the existing SoA metrics. The FR video-quality model VMAF performs almost on-par with these two models. As expected, the pixel-based NR models are the worst-performing ones.

TABLE II
PERFORMANCE EVALUATION OF P.1204.3 AND FHD-MAPPED P.1204.3 USING GAMINGVIDEOSET; (* USING EQUATION 4)

Model	RMSE	PCC	SROCC	Kendall	R^2
P.1204.3	0.45	0.88	0.87	0.69	0.77
FHD-mapped P.1204.3*	0.43	0.89	0.88	0.70	0.79
PSNR	0.63	0.74	0.74	0.57	0.55
SSIM	0.57	0.80	0.80	0.61	0.62
VMAF	0.47	0.87	0.86	0.69	0.75
STRREDopt	0.65	0.71	0.74	0.55	0.51
SPEEDQA	0.65	0.71	0.74	0.56	0.51
BRISQUE	0.84	0.49	0.46	0.31	0.20
BIQI	0.84	0.43	0.45	0.30	0.18
NIQE	0.64	0.77	0.71	0.53	0.52
MEON	0.87	0.35	0.30	0.20	0.13

2) KUGVD: The results reported in Table III show the performance of the P.1204.3 model in comparison with the other SoA metrics. Again, the same SoA models are shown for comparison. Like for the GamingVideoSet, the P.1204.3 model is the best performing when compared to the other metrics.

TABLE III
PERFORMANCE EVALUATION OF P.1204.3 AND FHD-MAPPED P.1204.3 USING KUGVD; (* USING EQUATION 4)

Model	RMSE	PCC	SROCC	Kendall	R^2
P.1204.3	0.39	0.93	0.92	0.77	0.86
FHD-mapped P.1204.3*	0.35	0.94	0.93	0.78	0.88
PSNR	0.62	0.80	0.84	0.67	0.64
SSIM	0.48	0.89	0.91	0.74	0.79
VMAF	0.41	0.92	0.92	0.77	0.85
STRREDOpt	0.71	0.73	0.84	0.66	0.53
SpeedQA	0.74	0.70	0.85	0.67	0.49
BRISQUE	0.81	0.63	0.60	0.42	0.39
BIQI	0.83	0.60	0.59	0.40	0.36
NIQE	0.55	0.85	0.84	0.66	0.72
MEON	0.94	0.44	0.39	0.26	0.19

3) CGVDS: In comparison to the previous dataset, this dataset is unique among the four ones in terms on the considered encoder, since it uses a hardware-accelerated H.264 encoder (NVENC) to ensure the requirement of fast encoding in cloud gaming services.Although the standardized model was not trained for the hardware-accelerated encoder with preset "llhq", it still performs very well and is on par with the best-performing FR metric VMAF, as can be seen in Table IV.

TABLE IV
PERFORMANCE EVALUATION OF P.1204.3 AND FHD-MAPPED P.1204.3 USING CGVDS; (* USING EQUATION 4)

Model	RMSE	PCC	SROCC	Kendall	R^2
P.1204.3	0.38	0.85	0.84	0.65	0.72
FHD-mapped P.1204.3*	0.40	0.84	0.83	0.62	0.70
PSNR	0.60	0.64	0.65	0.47	0.41
SSIM	0.59	0.67	0.78	0.60	0.45
MS-SSIM	0.53	0.74	0.83	0.63	0.55
VMAF	0.38	0.88	0.87	0.69	0.77
BRISQUE	0.67	0.52	0.49	0.34	0.27
PIQE	0.67	0.51	0.52	0.35	0.27
NIQE	0.66	0.54	0.56	0.41	0.29
NDNetGaming	0.35	0.90	0.90	0.73	0.80

4) Twitch Dataset: As mentioned earlier, the Twitch dataset is unique, since the video segments are directly downloaded from the gaming streaming service Twitch.tv in all possible representations and hence the encoding is Twitch.tv platform specific.During the training and validation of the P.1204.3 model, encodings from different online streaming services such as YouTube and Bitmovin have also been included [16]. It is known that every streaming service optimizes their encoding pipeline based on different criteria like content complexity, for example dynamic encoding by Netflix and per-scene adaptation by Bitmovin. Hence, this dataset coming from an online streaming service that was not considered during model development was expected to pose a unique challenge to the P.1204.3 model. The results reported in Table V indicate that the P.1204.3 model and the FHD-mapped P.1204.3 model adapt themselves very well to this new encoding scenario and have a very good performance both in terms of RMSE and PCC. For this dataset, other than the performance of the P.1204.3 model variants, only the NR metrics such as BRISQUE and NIQE are calculated, and not any other FR or RR metrics since we had no access to the reference videos. This reflects a real-life gaming monitoring context, where access to a reference video may be difficult. As can be seen from Table V, both the NR metrics show rather poor results.

TABLE V
PERFORMANCE EVALUATION OF P.1204.3 AND FHD-MAPPED P.1204.3 USING TWITCH DATASET; (* USING EQUATION 4)

Model	RMSE	PCC	SROCC	Kendall	R^2
P.1204.3	0.40	0.93	0.93	0.77	0.87
FHD-mapped P.1204.3*	0.45	0.91	0.92	0.75	0.83
BRISQUE	0.97	0.237	0.275	0.197	0.016
NIQE	0.96	0.241	0.114	0.17	0.043

V. CONCLUSION AND FUTURE WORK

In light of the increasing popularity of gaming video streaming, we evaluated the newly standaradized P.1204.3 mode 3 bitstream-based video-quality model on gaming content. In addition to this evaluation, we proposed a full-HD-mapped variant of the standardized model, to explicitly take into account a target display of Full HD resolution. For the purpose of both evaluation and proposing the FHD-variant, we considered four different databases which enabled us to evaluate the models on a wide-range of gaming content, scenario and

encoding settings. To develop the FHD-mapped variant of P.1204.3, we made a 50:50 split of datasets and used the GamingVideoSet and KUGVD datasets for training the new model and validated this with two other datasets, namely, the CGVDS and Twitch datasets. We evaluated the performance of these two models on all the datasets, and compared them with the SoA FR, RR and NR models. Both the variants performed either on-par with or better than the best performing FR metric, VMAF, for all datasets. For all models, we applied a first-order mapping following [9]. It is notable, that the bitstream-based model and also the FHD-mapped version perform very well on unknown databases of gaming video streaming and encoding scenarios such as hardware-based encoding as in the case of the CGVDS dataset, or a proprietary encoding setting as in the case of the Twitch dataset. Neither of these two encoding scenarios were tested during the development of P.1204.3.

As future work, an in-depth feature analysis of the P.1204.3 model will be performed to identify the features best suited for gaming-video streaming quality prediction. In addition, new features that more explicitly take into account the uniqueness of gaming content can be developed and used to improve the prediction performance of P.1204.3 and its FHD version. Also, a more thorough validation of the proposed FHD mapping will be performed, using more datasets along with testing on "traditional" non-gaming video. A comparison between the average quality score over the 30 s duration and a more sophisticated aggregation with temporal weighting that e.g. accounts the peak-end-effect and recency [18] will be performed, too, to invesitage possible improvements in prediction.

REFERENCES

[1] 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". In: *IEEE Access* 7 (2019), pp. 74511–74527.

[2] N. Barman and M. G. Martini. "H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 codec comparison for live gaming video streaming". In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. 2017, pp. 1–6.

[3] N. Barman, M. G. Martini, S. Zadtootaghaj, S. Möller, and S. Lee. "A Comparative Quality Assessment Study for Gaming and Non-Gaming Videos". In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. 2018, pp. 1–6.

[4] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller. "GamingVideoSET: A Dataset for Gaming Video Streaming Applications". In: *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*. 2018, pp. 1–6.

[5] N. Barman, S. Schmidt, S. Zadtootaghaj, M. G. Martini, and S. Möller. "An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming". In: *Proceedings of the 23rd Packet Video Workshop*. PV '18. Amsterdam, Netherlands: Association for Computing Machinery, 2018, 7–12. URL: https://doi.org/10.1145/3210424.3210434.

[6] D. Fitzgerald and D. Wakabayashi. *Apple Quietly Builds New Networks*. 2014. URL: http://www.wsj.com/articles/apple-quietly-builds-new-networks-1391474149.

[7] P. Fröhlich et al. "QoE in 10 seconds: Are short video clip lengths sufficient for Quality of Experience assessment?" In: *2012 Fourth International Workshop on Quality of Multimedia Experience*. 2012, pp. 242–247.

[8] S. Göring, R. R. R. Rao, and A. Raake. "nofu — A Lightweight No-Reference Pixel Based Video Quality Model for Gaming Content". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. 2019, pp. 1–6.

[9] ITU-T. *P.1401 : Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*. Tech. rep. Int. Telecommunication Union, 2014.

[10] ITU-T. *Recommendation P.1203 - Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport*. Tech. rep. International Telecommunication Union, 2016.

[11] ITU-T. *Recommendation P.1204.3 - Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information*. Tech. rep. International Telecommunication Union, 2019.

[12] A. Mittal, A. K. Moorthy, and A. C. Bovik. "No-Reference Image Quality Assessment in the Spatial Domain". In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708.

[13] A. Mittal, R. Soundararajan, and A. C. Bovik. "Making a "Completely Blind" Image Quality Analyzer". In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 209–212.

[14] A. K. Moorthy and A. C. Bovik. "A two-stage framework for blind image quality assessment". In: *2010 IEEE International Conference on Image Processing*. 2010, pp. 2481–2484.

[15] Netflix. *Netflix VMAF*. URL: https://github.com/Netflix/vmaf (visited on 12/08/2018).

[16] R. Rao Ramachandra Rao et al. "Bitstream-based Model Standard for 4K/UHD: ITU-T P.1204.3 – Model Details, Evaluation, Analysis and Open Source Implementation". In: *Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. Athlone, Ireland, May 2020.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE trans. on image processing* 13.4 (2004), pp. 600–612.

[18] B. Weiss et al. "Temporal development of quality of experience". In: *Quality of experience*. Springer, 2014, pp. 133–147.

[19] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, and S. Möller. "NR-GVQM: A No Reference Gaming Video Quality Metric". In: *2018 IEEE International Symposium on Multimedia (ISM)*. 2018, pp. 131–134.

[20] S. Zadtootaghaj, S. Schmidt, N. Barman, S. Möller, and M. G. Martini. "A Classification of Video Games based on Game Characteristics linked to Video Coding Complexity". In: *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*. 2018, pp. 1–6.

[21] S. Zadtootaghaj, S. Schmidt, S. Shafiee Sabet, S. Moeller, and C. Griwodz. "Quality Estimation Models for Gaming Video Streaming Services Using Perceptual Video Quality Dimensions". In: *Proceedings of the 11th International Conference on Multimedia Systems*. ACM. 2020.