

Subjective and Objective Quality Assessment of High-Motion Sports Videos at Low-Bitrates

Joshua P. Ebenezer , Yixu Chen, Yongjun Wu, Hai Wei, Sriram Sethuraman
Amazon Prime Video, Seattle, WA

Abstract—Videos often have to be transmitted and stored at low bitrates due to poor network connectivity during adaptive bitrate streaming. Designing optimal bitrate ladders that would select the perceptually-optimized resolution, frame-rate, and compression level for low-bitrate videos for adaptive streaming across the internet is therefore a task of great interest. Towards that end, we conducted the first large-scale study of medium and low-bitrate videos from live sports for two codecs (Elemental AVC and HEVC) and created the Amazon Prime Video Low-Bitrate Sports (APV LBS) dataset. The study involved 94 participants and 742 videos, with more than 23,000 human opinion scores collected in total. We analyzed the data obtained and we also conducted an extensive evaluation of objective Video Quality Assessment (VQA) algorithms and benchmarked their performance, and make recommendations on bitrate ladder design. We're making the metadata and VQA features available at <https://github.com/JoshuaEbenezer/lbmfr-public>.

Index Terms—Video quality assessment, Database, Low-Bitrate

I. INTRODUCTION

Live video is expected to have reached 13.2% of global internet traffic by the end of 2021, up from 3.3% in 2016 [1]. Sports' streaming is a subset of live video traffic and is expected to grow from a 18 billion USD market in 2020 to an 87 billion USD market by 2028 [2]. Streaming services have to make decisions on what resolution, frame rate, and compression level must be used to stream videos to a customer with a certain available bitrate such that the delivered video is perceptually optimal. Live Sports streaming services such as Amazon Prime Video and NBC Peacock are rapidly growing in areas with limited or developing internet connectivity. Even for users with connections that are high-bandwidth on average, temporary slowdowns in internet speed can occur which may necessitate livestreaming sports' events at lower bitrates. However, perceptual data on low-bitrate high-motion contents is scarce and to our knowledge no subjective video quality database exists that is designed for this use-case.

At higher bitrates and higher resolutions (more than 1080p), studies [3], [4] have found that increasing the frame rate is preferred over increasing the resolutions for a fixed bitrate budget. For example, for a wide variety of content, a video streamed at 1080p at 60 fps will be perceptually better than a video streamed at 4K at 30 fps. However, it is unknown whether a transition bitrate-budget point occurs below which higher resolutions are preferred over higher frame rates. In

addition, contents with sports or high-motion will need to be treated differently since a higher frame rate might be preferred for a larger range of bitrate-budgets since important events (like a tennis ball being hit) are occurring momentarily in such videos. The gold standard in studying such variations are the perceptual opinion scores provided by humans for different videos of the same content whose parameters such as resolution, frame-rate, and bitrate are varied. We created such a database of 742 videos rated by a total of 94 participants in order to provide an analysis of the low-bitrate regime and also as a resource to train and test objective video quality assessment algorithms.

II. RELATED WORK

A number of databases have been created for generic video quality tasks. Crowdsourced databases such as Konvid [5], YT-UGC [6], and VQC [7] are typically used to design video quality metrics for user-generated content. In-lab databases such as LIVE-ETRI [8] and LIVE Livestream [9] are created for professionally created content in highly controlled settings and as a result are typically more internally consistent. However, there are very few databases that consider the effect of resolution, compression level, and frame-rate at the same time, and to the best of our knowledge there are no databases specifically targeting the low-bitrate regime for high-motion contents. The LIVE-YT-HFR [10] database contains 480 videos at 6 frame rates and 5 VP9 constant rate factor levels, but only two resolutions (1080p and 4K). This limits its use for low-bitrate scenarios, where resolutions are typically lower than 720p. The LIVE-ETRI [8] database contains 437 HEVC compressed videos spanning 3 frame rates (30, 60, 120 fps), four resolutions (960×540, 120×720, 1920×1080, and 3840×2160) and 5 QP values. Again, the limited number of low resolutions make the database unsuited for low-bitrate ladder design. The AVT-VQDB-UHD1 [11] database contains 756 videos rated by 104 participants, with resolutions spanning 360p to 4K, framerates from 15 fps to 60 fps, and bitrates from 300 kbps to 15 Mbps. However, the contents in this database are not high-motion sports contents. High-motion contents typically have temporal video artifacts that may not manifest as strongly in low-motion contents. This affects bitrate ladder decisions since the interplay of compression, resolution, and frame-rate is affected by temporal artifacts. Towards filling this important need, we created the first large-scale, in-lab database of high-motion live sports contents encoded at medium and low bitrates.

J.P. Ebenezer worked on this project during his internship at Amazon. His current affiliation is with the University of Texas at Austin.

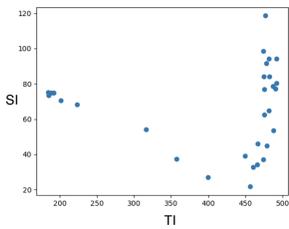


Fig. 1: Plot of SI vs TI

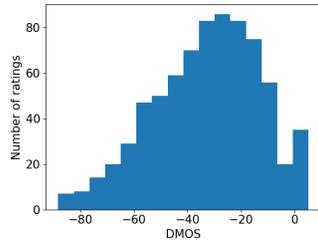


Fig. 2: Histogram of DMOS

III. DETAILS OF SUBJECTIVE STUDY

A. Source Sequences

The 30 source videos were all of sports contents captured from live feeds of sports broadcasting partners. Nine videos were of American football events, eight videos were of soccer, seven videos were of tennis, and six videos were of cricket. The videos covered diverse scenes, including of players running, touchdowns, goals, rallies, slow-motion replays, commentators. The American football source videos had a frame rate of 59.94 frames per second (fps). The soccer, tennis, and cricket videos had a frame rate of 50 fps. Each source video sequence was cut from a longer video. In order to cut the videos accurately, the longer videos were first converted to the y4m format and the frames corresponding to the desired source sequence were cut out. By doing so, we were able to avoid issues related to cutting encoded bitstreams. All the source clips were cut to a length of 8s. The diversity of the 8s source contents can also be captured through the Spatial Information (SI) and the Temporal Information (TI) indices, described in ITU Rec. 910 [12]. The SI is plotted against the TI in Fig. 1. Note that the values of TI span a large range and are clustered around high values (> 450), indicating the high-motion nature of the contents. For reference, the highest value of TI for the LIVE-ETRI database was reported to be 250 and the highest TI in the LIVE-YT-HFR database was reported to be 35.

B. Encoding Settings

The frame rates of the source videos (50/59.94 fps) were considered as High Frame Rates (HFR). Each source video was temporally subsampled by dropping alternate frames to create a Standard Frame Rate (SFR) version of 25 fps or 29.97 fps. Both the HFR and SFR versions were encoded at different bitrates and resolutions using the AVC [13] and HEVC [14] codecs with the Elemental hardware encoder [15] in constant bitrate mode. The bitrate-resolution (BR) combinations are listed in Table I for each frame rate. The same combinations were used for the AVC and HEVC codecs. Each of the 30 source videos was used to generate 24 distorted videos, and hence 750 videos were used in the study. However, after the study began 8 AVC-encoded SFR videos of soccer were found to have unintended black-borders due to a bug in the encoder software, and hence had to be discarded for the final analysis. 6 AVC encoded HFR videos of soccer were also found to have the black-border, but were corrected in time for 24 subjects in

TABLE I: Bitrate-resolution-framerate combinations for encoding

HFR	SFR
288p at 300 kbps	288p at 200 kbps
396p at 600 kbps	396p at 450 kbps
396p at 900 kbps	396p at 800 kbps
540p at 1300 kbps	540p at 1200 kbps
540p at 1900 kbps	540p at 1800 kbps
720p at 2200 kbps	720p at 2000 kbps

group 2 to watch the correct version, and were hence retained for the final analysis with 24 scores.

C. Subjective Study Settings

The study was conducted at Amazon Prime Video’s lab at Seattle over a duration of 4 weeks. Two TVs were placed more than 15 feet apart such that subjects could watch the videos sitting back-to-back without interference. Subjects were seated at a distance of $1.5H$ from the TV, where H was the height of the TV. The TVs were identically calibrated 55 inch 4K LG C9 OLED TVs which are capable of playing 4K videos at 60 fps. They were attached through HDMI 2.1 as external displays to two Alienware M15 R4 Gaming Laptops with NVIDIA GeForce RTX 3070 GPUs. The windows of the lab were covered with black paper to block outside light and the lights were dimmed to create a low-light setting compliant with ITU Rec 500 [16]. 94 participants volunteered to take part in the study. The 750 videos were divided into 3 groups of 250 videos each such that videos of the same content appeared in the same group. The contents were divided across the 3 groups in such a way that each group had an equal representation of each sport. The participants were divided into 3 groups of 32, 30, and 32 participants each. Each group of participants watched a different group of 250 videos. The study was divided into two 30 minute sessions for each participant to minimize fatigue, and participants were allowed to choose when to schedule their sessions. The age range of the participants was 19-40, all participants had normal or corrected-to-normal vision. The first session began with a training session where subjects were shown 3 videos of poor, medium, and good quality, and told about the quality of the videos, with the intention of ensuring that the score distribution would be reasonable. The scores from the training session were discarded. After that, the study would begin and scores for each video would be collected according to the ACR-HR protocol described in ITU Rec. 910 [12]. At the end of each video, a rating scale from 0-100 with increments of 1 was presented with 5 equally spaced verbal markers of Poor, Bad, Fair, Good, and Excellent. Subjects could choose any value on the scale using a mouse.

IV. ANALYSIS OF SCORES

A. Internal Consistency

For each video, all the subjects who watched that video were divided into two equal-sized groups randomly, and Z-score for that video was computed for each group separately. This was done for all the videos, generating two sets of Z-scores. The Spearman Rank Ordered Correlation Coefficient

(SRCC) was computed between the two sets. This procedure was repeated 100 times, each time doing a different random split of all the subjects who watched a video. The median SRCC of the 100 randomized splits was 0.9543, indicating a high degree of internal consistency.

B. Recovery of Quality Scores

We used the Sural [17] method to estimate the quality scores from the study using maximum-likelihood estimation. Each opinion score U_{ij} is modelled as

$$U_{ij} = \psi_j + \Delta_i + \nu_i X \quad (1)$$

where ψ_j is the true quality score of video j , Δ_i represents the bias of subject i , the non-negative term ν_i represents the inconsistency of subject i , and $X \sim N(0, 1)$ are i.i.d. Gaussian random variables. The true score, subject bias, and subject inconsistency are estimated through maximum-likelihood estimation with a Newton-Raphson solver. ψ was treated as the Mean Opinion Score (MOS) and the Differential MOS (DMOS) was computed as the difference between the MOS of the distorted video and the MOS of the pristine video. A histogram of DMOS from the study is shown in Fig. 2 and shows that the distribution is broad. We also plot DMOS according to different categories in Fig. 3. The scores in each category have a high degree of overlap because of the interplay of the other factors, though the broadly monotonic relationships between resolution and DMOS and bitrate and DMOS are expected. HEVC encoded videos have a slightly higher median DMOS than AVC encoded videos, but with a high degree of overlap, as can be seen from Fig. 3d.

A few DMOS vs bitrate plots are shown in Fig. 4 for 4 of the 30 contents, with each sport being represented. Each figure is therefore plotted using data for 25 videos and represents quality scores for all the BR combinations for HEVC and AVC codecs for both HFR and SFR. The BR combinations are from Table I. Each plot is quite different and this is indicative of the highly content-dependent nature of perceptual quality and spatio-temporal artifacts. A number of important observations can be made from each plot for each content. When a video's frame rate is doubled at the same bitrate while keeping the resolution the same, spatial artifacts increase because the codec has to encode double the number of frames with the same bitrate budget. If the deleterious effect of these spatial artifacts is greater than that of the temporal artifacts caused by high-motion contents played at SFR, the perceptual quality of the SFR version will be greater. For the soccer video encoded with HEVC, HFR is preferred over SFR only after the 540p at 1200 kbps point, indicating that there are indeed such transition points in the low-bitrate regime below which SFR is preferred. On the other hand, for the AVC-compressed tennis video and the HEVC and AVC compressed football videos, the SFR version is always rated better than the HFR version, indicating that the transition point may only appear beyond the 720p at 2000 kbps point. The HEVC-encoded version of the tennis video, however, does have a transition point at 540p at 1300 kbps beyond which HFR is preferred over SFR. The

content-dependent nature of the ratings indicates the need for per-title or per-shot bitrate-ladder decisions.

V. RESULTS ON OBJECTIVE QUALITY ASSESSMENT

A. Metrics

We report the SRCC between the raw quality predictions made by the objective VQA algorithms and the DMOS. We also transform the predictions to the range of the DMOS by using a logistic function

$$f(s) = \beta_1 \left(\frac{1}{2} - \frac{1}{(1 + \exp(\beta_2(s - \beta_3)))} \right) + \beta_4 x + \beta_5. \quad (2)$$

The parameters are found by fitting $f(s)$ to the DMOS. We then compute Pearson's Linear Correlation Coefficient (LCC) and the Root Mean Square Error (RMSE) between the transformed predictions and the DMOS.

B. Full-Reference Algorithms

Full-Reference (FR) VQA algorithms compare the distorted video to the pristine video and predict the difference in quality. We tested a number of FR algorithms with our dataset. All the FR algorithms expected the input distorted and pristine videos to be of the same dimensions. Since the distorted videos in our dataset were not all the same resolution and frame rate as the pristine videos, they had to be spatially and temporally upsampled to match the resolution and frame rate of the pristine videos. Spatial upscaling was performed with fast bilinear upscaling. Temporal upscaling was performed with both frame duplication and motion interpolation, and we found that FR algorithms had an average increase in SRCC of 12% when their inputs were the motion-interpolated distorted videos when compared to when the inputs were frame duplicated videos. In Table II, we report results obtained for FR algorithms that are either pre-trained or that do not need to be trained for motion-interpolated distorted videos.

We trained and evaluated ST-GREED [18], an FR algorithm that generates 4 features that need to be trained by a Support Vector Regressor (SVR). We also trained and evaluated VMAF from scratch with the motion-interpolated videos. The training procedure was to divide the videos into a training set and a test set with an 80:20 ratio. 5-fold cross-validation was performed on the training set with a grid search to find the best hyperparameters for the SVR. The resulting SVR was evaluated on the test set. This procedure was repeated 100 times with different train-validation-test splits. The median metrics and their standard deviations are presented in Table III.

C. No-Reference Algorithms

We evaluate a completely blind No-Reference (NR) algorithm, NIQE, and also train and evaluate ChipQA [19], a leading NR VQA algorithm. The results are presented in Table V. We also evaluate two algorithms that rely on additional metadata: ITU P1204.3 [20], a no-reference algorithm that requires the bitstream information of the distorted video in addition to the video itself, and another feature set that

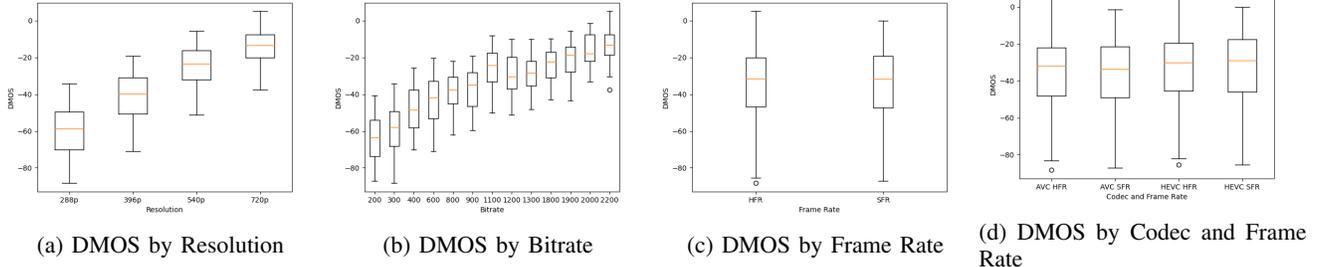


Fig. 3: DMOS by different categories

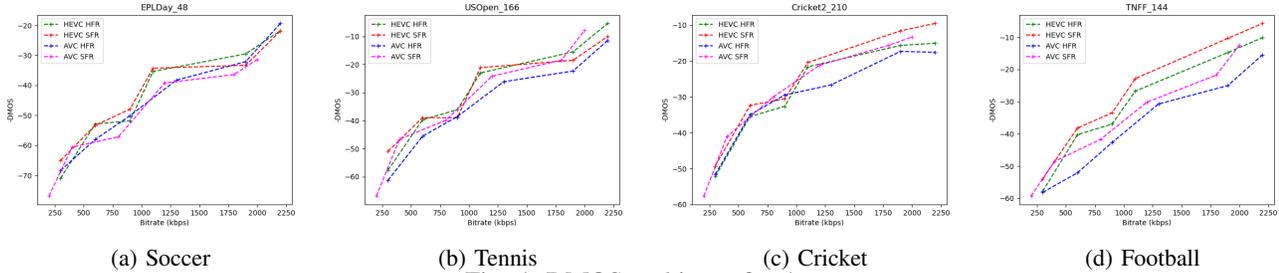


Fig. 4: DMOS vs bitrate for 4 contents

TABLE II: SRCC, LCC, and RMSE for untrained/pre-trained FR

METHOD	SRCC	LCC	RMSE
PSNR	0.3544	0.3119	17.8795
SSIM [21]	0.4953	0.3750	17.4286
MS-SSIM [22]	0.3416	0.0357	19.6043
SPEED [23]	0.5749	0.5904	15.1737
V-SPEED [23]	0.3904	0.0996	18.7071
Spatial VIF [24]	0.4581	0.4735	16.5590
DLM [25]	0.8430	0.8275	10.5541
VMAF [26] (pre-trained)	0.8518	0.8325	10.4156

TABLE III: SRCC, LCC, and RMSE for trained FR

METHOD	SRCC	LCC	RMSE
ST-GREED [18]	0.57±0.08	0.59±0.10	14.56±1.84
VMAF (trained) [26]	0.84±0.10	0.84±0.09	9.63±2.19

just contains the bitrate, frame-rate, resolution, and codec, which we simply refer to as "Vid. Level metadata". The training procedure described in Section V-B was used, with the exception that the algorithms were trained to predict the MOS and not the DMOS, and the results are presented in Table V. The metadata based algorithms perform well though such metadata is not always available. NR algorithms are especially relevant for livestreaming since the video obtained at the source may not be free of artifacts, as discussed in detail in [9].

TABLE IV: SRCC, LCC, and RMSE for blind NR VQA

METHOD	SRCC	LCC	RMSE
NIQE [27]	0.4195	0.4592	15.7479
ChipQA [19]	0.83±0.09	0.84±0.09	9.31±2.40

TABLE V: SRCC, LCC, and RMSE for trained metadata-based NR VQA

METHOD	SRCC	LCC	RMSE
P 1204.3 [20]	0.93±0.14	0.93±0.01	6.18±0.75
Vid. Level Metadata	0.94±0.02	0.94±0.02	6.10±0.62

VI. DISCUSSION AND CONCLUSION

It appears from the data that there isn't a uniform recommendation suitable for all high-motion contents with different properties. Each content has a different concave envelope that represents the best BR combination and frame rate that has to be used for perceptually-optimized viewing. This further underscores the need for VQA algorithms that can predict the quality of videos encoded at different settings and make a decision on the perceptually optimal setting for each content or shot separately. For livestreamed sports, making such content-dependent decisions on the fly, ideally with an automated system, requires FR or NR algorithms with very low latencies. VMAF performs the best among the FR algorithms we tested. Among NR algorithms, both ChipQA and P1204.3 perform very well. P1204.3 performs better than ChipQA, but the difference in their performance is within the standard deviations of their results. If the video is re-encoded in some way at any stage after encoding P1204.3 will fail, whereas ChipQA is a pixel-based method that can work without bitstream information. We also make the observation that comparing motion-interpolated SFR distorted videos to HFR pristine videos allows FR algorithms to make better judgments than using frame duplication. We are releasing the metadata, VQA features, and scores for our study for researchers to study the effect of different encoding and frame-rate parameters on quality and envision that this database will be an important resource for the community.

REFERENCES

- [1] Cisco, *Cisco Global Forecast*, 2021 (accessed December 20, 2021). [Online]. Available: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf
- [2] V. M. Research, *Sports Online Live Video Streaming Market Size And Forecast*, 2021 (accessed December 20, 2021). [Online]. Available: <https://www.verifiedmarketresearch.com/product/sports-online-live-video-streaming-market/>
- [3] K. Debattista, K. Bugeja, S. Spina, T. Bashford-Rogers, and V. Hulusic, "Frame rate vs resolution: A subjective evaluation of spatiotemporal perceived quality under varying computational budgets," in *Computer Graphics Forum*, vol. 37, no. 1. Wiley Online Library, 2018, pp. 363–374.
- [4] E. B. Union, *Subjective testing confirms importance of frame rate for UHDTV*, 2013 (accessed December 28, 2021). [Online]. Available: <https://tech.ebu.ch/news/subjective-testing-confirms-importance-o-07aug13>
- [5] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupé, "The Konstanz natural video database (Konvid-1k)," in *Int. Conf. Quality of Multimedia Experience*, 2017, pp. 1–6.
- [6] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube ugc dataset for video compression research," in *IEEE Int. Workshop Multimed. Signal Process.* IEEE, 2019, pp. 1–5.
- [7] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, 2018.
- [8] D. Y. Lee, S. Paul, C. G. Bampis, H. Ko, J. Kim, S. Y. Jeong, B. Homan, and A. C. Bovik, "A subjective and objective study of space-time subsampled video quality," *arXiv preprint arXiv:2102.00088*, 2021.
- [9] Z. Shang, J. P. Ebenezer, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "Study of the subjective and objective quality of high motion live streaming videos," *IEEE Transactions on Image Processing*, pp. 1–1, 2021.
- [10] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108 069–108 082, 2021.
- [11] R. R. Ramachandra Rao, S. Göring, W. Robitza, B. Feiten, and A. Raake, "Avt-vqdb-uhd-1: A large scale video quality database for uhd-1," in *2019 IEEE International Symposium on Multimedia (ISM)*, 2019, pp. 17–177.
- [12] ITU, *Subjective video quality assessment methods for multimedia applications*, 2008 (accessed December 28, 2021). [Online]. Available: <https://www.itu.int/rec/T-REC-P.910>
- [13] ITU, *H.264 : Advanced video coding for generic audiovisual services*, 2021 (accessed December 28, 2021). [Online]. Available: <https://www.itu.int/rec/T-REC-H.264>
- [14] ITU, *H.265 : High efficiency video coding*, 2021 (accessed December 28, 2021). [Online]. Available: <https://www.itu.int/rec/T-REC-H.265>
- [15] Elemental, "Elemental hardware encoder," June 2021. [Online]. Available: "https://aws.amazon.com/elemental-live/"
- [16] ITU, *BT.500 : Methodologies for the subjective assessment of the quality of television images*, 2019 (accessed December 28, 2021).
- [17] Z. Li, C. G. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," *Electronic Imaging*, vol. 2020, no. 11, pp. 131–1, 2020.
- [18] 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," *IEEE Transactions on Image Processing*, vol. 30, pp. 7446–7457, 2021.
- [19] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "ChipQA: No-Reference Video Quality Prediction via Space-Time Chips," *IEEE Transactions on Image Processing*, vol. 30, pp. 8059–8074, 2021.
- [20] ITU, *P.1204.3 : Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information*, 2008 (accessed December 28, 2021). [Online]. Available: <https://www.itu.int/rec/T-REC-P.1204.3>
- [21] 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.
- [22] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, vol. 2. Ieee, 2003, pp. 1398–1402.
- [23] 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.
- [24] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [25] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.
- [26] Netflix, *VMAF: The Journey Continues*, 2018 (accessed December 28, 2021). [Online]. Available: <https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12>
- [27] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2012.