OPTE: ONLINE PER-TITLE ENCODING FOR LIVE VIDEO STREAMING

Vignesh V Menon[1]*, *Hadi Amirpour*[1]*, *Mohammad Ghanbari*[1,2], *and Christian Timmerer*[1]

[1]Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität, Klagenfurt, Austria
[2] School of Computer Science and Electronic Engineering, University of Essex, UK
*These authors contributed equally to this work

ABSTRACT

Current per-title encoding schemes encode the same video content at various bitrates and spatial resolutions to find an optimized bitrate ladder for each video content in *Video on Demand* (VoD) applications. However, in live streaming applications, a bitrate ladder with fixed bitrate-resolution pairs is used to avoid the additional latency caused to find optimum bitrate-resolution pairs for every video content. This paper introduces an online per-title encoding scheme (OPTE) for live video streaming applications. In this scheme, each target bitrate's optimal resolution is predicted from any pre-defined set of resolutions using *Discrete Cosine Transform* (DCT)-energy-based low-complexity spatial and temporal features for each video segment. Experimental results show that, on average, OPTE yields bitrate savings of 20.45% and 28.45% to maintain the same PSNR and VMAF, respectively, compared to a fixed bitrate ladder scheme (as adopted in current live streaming deployments) without any noticeable additional latency in streaming.

***Index Terms*—** Per-title encoding, live streaming, bitrate ladder, resolution prediction.

1. INTRODUCTION

Video on Demand (VoD) and live streaming are widely embraced in video services, and their applications have attracted tremendous attention in recent years. Since streaming services continuously adapt the video delivery to the end user's network conditions and device capabilities, *HTTP Adaptive Streaming* (HAS) continues to grow and has become the *de-facto* standard in recent years for delivering video over the Internet. In HAS, each video is encoded at a set of bitrate-resolution pairs, referred to as *bitrate ladder*. Traditionally, a fixed bitrate ladder, *e.g.*, *HTTP Live Streaming* (HLS) bitrate ladder [1], is used for all video contents. However, due to the vast diversity in video content characteristics and network conditions, the "one-size-fits-all" can be optimized per *title* to increase the *Quality of Experience* (QoE) [2].

Per-title encoding schemes are based on the fact that each resolution performs better than others in a specific bitrate region for a given bitrate range. These bitrate regions depend on the *video content*. As shown in Fig. 1, for the *Beauty* sequence, the cross-over bitrate between the quality of 540p and 1080p resolutions happens at approximately $b_1 = 1.2Mbps$, which means at bitrates lower than b_1, 540p resolution outperforms 1080p, while at bitrates higher than b_1, 1080p resolution outperforms 540p. On the other hand, for the *Golf* sequence, 1080p remains superior at the entire bitrate range, which means 1080p should be selected for the bitrate ladder for the entire bitrate range. This *content-dependency* to select the optimal bitrate-resolution pairs is the basis for introducing the per-title encoding scheme [2]. Each video segment is encoded at several quality levels in this scheme, and bitrate-resolution pairs per each quality level and *convex-hull* is determined. The bitrate-resolution pair with the highest quality (*i.e.*, closer to the convex-hull) is selected for each quality level. For example, the *Lake* sequence is encoded at a set of bitrates and resolutions, and their bitrate-quality pairs are shown in Fig 2. For each requested bitrate, the resolution with bitrate-quality pair closer to the convex-hull is selected for the bitrate ladder.

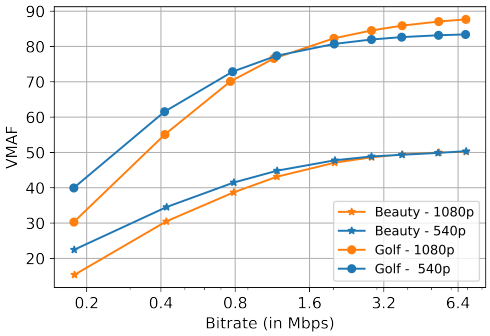


Fig. 1: Rate-Distortion (RD) curves using VMAF as the quality metric of *Beauty* and *Golf* sequences of UVG and BVI datasets [3, 4] encoded at 540p and 1080p resolutions.

Though per-title encoding schemes [2, 5, 6] enhance the quality of video delivery, determining convex-hull is computationally very expensive, making it suitable for only VoD streaming applications. Considering R resolutions and B bitrates, finding the optimal per-title bitrate ladder requires $R \times B$ test encodings to determine the convex-hull. Some methods pre-analyze the video contents to avoid brute force en-

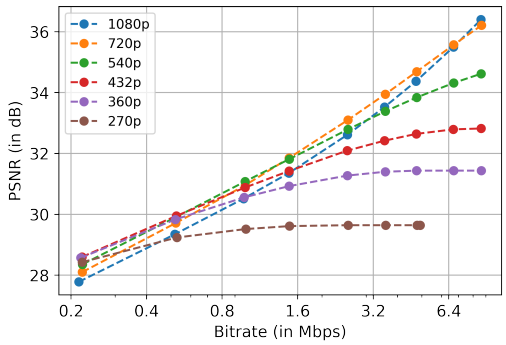


Fig. 2: The *Lake* sequence is encoded at a set of bitrates and resolutions to determine the convex-hull.

coding of all bitrate-resolution pairs [7]. Katsenou *et al.* [8] introduced a content-gnostic method that employs machine learning to find the bitrate range for each resolution that outperforms other resolutions. Bhat *et al.* [9, 10] proposed a Random Forest (RF) classifier to decide which encoding resolution is best suited over different quality ranges and study machine learning-based adaptive resolution prediction. However, these approaches yield *latency* much higher than the accepted latency in live streaming. According to the Bitmovin Video Developer Report 2021 [11], *live streaming at scale* has the highest scope for innovation in video streaming services. Therefore, in this paper, a low-latency **O**nline **P**er-**T**itle **E**ncoding (OPTE) scheme is proposed that improves bitrate ladders for live video streaming applications. OPTE exploits content-aware features, *i.e.*, *Discrete Cosine Transform* (DCT)-energy-based low-complexity spatial and temporal features that are extracted to determine video segments' characteristics. Based on these features, a low-complexity bitrate ladder construction algorithm is proposed, which *predicts the optimized resolution for each video segment at every target bitrate* from a pre-defined set of resolutions.

2. OPTE ARCHITECTURE

The architecture of the proposed Online Per-title Encoding (OPTE) scheme is shown in Fig. 3, according to which the input video sequence is split into multiple segments. The bitrate ladder for each segment is predicted using the spatial and temporal features of the segment and the set of pre-defined resolutions (R) and bitrates (B) of the bitrate ladder. The encoding process is carried out only for the predicted bitrate-resolution pairs for each segment, thereby eliminating the need to encode in all bitrates and resolutions to find the optimal bitrate-resolution pairs. The bitrate ladder prediction in OPTE is classified into two steps: *(i)* feature extraction and *(ii)* resolution prediction, which are explained in Section 2.1 and Section 2.2, respectively.

2.1. Feature extraction

For online prediction in live streaming applications, selecting low-complexity features is critical to ensure low-latency

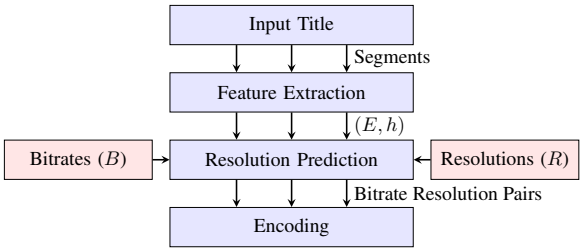


Fig. 3: OPTE architecture.

video streaming without disruptions. For each video segment, we calculate two features, *i.e.*, the average texture energy and the average gradient of the texture energy. In our previous works [12–14], a DCT-based energy function was introduced to determine the block-wise texture of each frame, which is defined as:

$$H_{p,k} = \sum_{i=0}^{w-1} \sum_{j=0}^{w-1} e^{|(\frac{ij}{w^2})^2 - 1|} |DCT(i,j)| \tag{1}$$

where k is the block address in the p^{th} frame, $w \times w$ pixels is the size of the block, and $DCT(i,j)$ is the $(i,j)^{th}$ DCT component when $i + j > 1$, and 0 otherwise [15]. Exponentially higher costs are assigned to higher DCT frequencies since it is expected that a mixture of objects causes the higher frequencies. Additionally, the Direct Current (DC) value is not included since it does not affect the texture. The texture is averaged to determine the *spatial energy* feature denoted as E as follows:

$$E = \sum_{p=0}^{P-1} \sum_{k=0}^{C-1} \frac{H_{p,k}}{P \cdot C \cdot w^2} \tag{2}$$

where C represents the number of blocks per frame, and P denotes the number of frames in the segment. Furthermore, the block-wise SAD of the texture energy of each frame compared to its previous frame is computed and then averaged for each frame of the segment to obtain the average *temporal energy* (h) as follows:

$$h = \sum_{p=1}^{P-1} \sum_{k=0}^{C-1} \frac{SAD(H_{p,k}, H_{p-1,k})}{(P-1) \cdot C \cdot w^2} \tag{3}$$

In the following, E and h feature values are used to predict the resolution for every target bitrate.

2.2. Resolution Prediction

This section explains the resolution prediction algorithm of the OPTE scheme. Resolution scaling factor (s) for every resolution r is defined as:

$$s = \frac{W}{W_{max}}; r \in R \tag{4}$$

where W denotes the width (in pixels) of the resolution r and W_{max} is the width of the maximum resolution of the input bitrate ladder. S denotes the set of s corresponding to resolutions defined in R. An exponentially decaying (increasing) function is modelled to determine s as a function of target bitrate b as shown below:

$$s(b) = 1 - s_0 e^{-Kb} \tag{5}$$

where $1 - s_0$ is the value at $b = 0$, K determines the rate of decay. The decay rate, K, is directly proportional to the temporal characteristics of the segment and the target bitrate for encoding. At the same time, it is inversely proportional to the spatial characteristics of the segment. Using this information, K is modeled as shown below:

$$K = \frac{\Gamma_{MA}(r_{max}, f) \cdot h}{E} \tag{6}$$

$\Gamma_{MA}(r_{max}, f)$ is the proportion constant named the *Menon Amirpour resolution scaling coefficient* which depends on the video framerate (f) and the original resolution (r_{max}). Hence, the final equation for computing the optimized resolution scaling factor ($\hat{s}$) is:

$$\hat{s}(b) = 1 - s_0 e^{-\frac{\Gamma_{MA}(r_{max}, f) \cdot h \cdot b}{E}} \tag{7}$$

Determining Γ_{MA}**:** In the bitrate ladder, R and S are predefined. To determine Γ_{MA}, the *half-life* of the $(1 - s)$ function is evaluated, *i.e.*, the bitrate when $(1 - s)$ becomes $\frac{1}{2}$. $(1 - s)$ is an exponentially decaying (decreasing) function whose half-life is given as:

$$b_{\frac{1}{2}} = \frac{ln(2)}{K} \tag{8}$$

Γ_{MA} can thus be determined as:

$$\Gamma_{MA} = \frac{ln(2) \cdot E}{h \cdot b_{\frac{1}{2}}} \tag{9}$$

Γ_{MA} values obtained for the training sequences of each framerate and the original resolution is averaged to determine $\Gamma_{MA}(r_{max}, f)$. After determining Γ_{MA} for each framerate and original resolution, the model predicts the optimized resolution scaling factor ($\hat{s}$) for any segment using the extracted E and h features and the target bitrate as shown in Algorithm 1.

Algorithm 1: Bitrate ladder construction

Inputs:
r_{max} : original resolution
f : number of video frames per second
S : set of all resolution scaling factors $\tilde{s}$
B : set of all target bitrates b

Output: $\hat{s}(b)\ \forall\ b \in B$

Step 1: Compute E and h features.

Step 2: Determine $\hat{s}(b)$.

$$\hat{s}(b) = 1 - s_0 e^{-\frac{\Gamma_{MA}(r_{max}, f) \cdot h \cdot b}{E}}$$

Step 3: Map $\hat{s}(b)$ to the closest $\tilde{s} \in S$.

3. EVALUATION

This section introduces the test methodology used in this paper, and then the experimental results are presented.

3.1. Test Methodology

The encodings are generated using x265 v3.5 [16] with the *veryfast* preset and the Video Buffering Verifier (VBV) rate control mode. The peak bitrate is set to 110% of the target bitrate, and the maximum buffer size is set to 300% of the peak bitrate. The segment length is set to 4 seconds. All experiments are run on a dual-processor server with Intel Xeon Gold 5218R (80 cores, frequency at 2.10 GHz). The E-h feature extraction [17] and bitrate ladder encoding are thus scheduled as parallel processes. The size of the block to determine block-wise texture is set as 32x32 pixels. The test sequences from *(i)* JVET [18], *(ii)* MCML [19], *(iii)* SJTU [20], and *(iv)* VQEG HDTV SVT datasets are used for evaluating the proposed scheme. The resolutions specified in Apple HLS authoring specifications [1] are considered in the evaluation. In the HLS ladder [1], R and S are defined as $\{360p, 432p, 540p, 720p, 1080p, 1440p, 2160p\}$ and $\{\frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\}$, respectively. $s = 1$ denotes 2160p spatial resolution. $s = \frac{1}{6}$ denotes 360p spatial resolution for the segment. Since the minimum s is $\frac{1}{6}$, the initial condition, $s(b = 0) = \frac{1}{6}$ is assumed. Thus, $s_0 = \frac{5}{6}$. The lower resolution sources are generated from the original video source by applying bi-cubic scaling using FFmpeg [21]. In this paper, $\Gamma_{MA}(2160p, 30)$, $\Gamma_{MA}(2160p, 50)$, and $\Gamma_{MA}(2160p, 60)$ are trained as 0.06, 0.03, and 0.02, respectively.

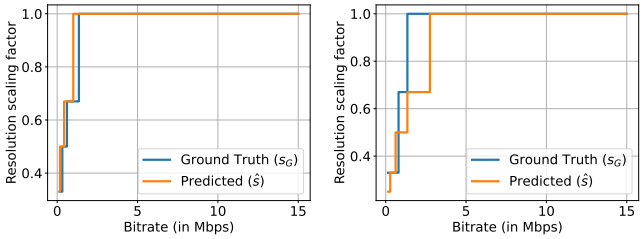
The resulting quality in PSNR and VMAF [22] and the achieved bitrate are compared for each test sequence. Since the content is assumed to be displayed on the highest resolution (2160p), the encoded content is scaled (bi-cubic) to 2160p resolution, and VMAF and PSNR [23] are calculated. Bjøntegaard delta rates [24] BDR_P and BDR_V refer to the average increase in bitrate of the representations compared with that of the fixed bitrate ladder encoding to maintain the same PSNR and VMAF, respectively. A negative BDR indicates a gain in encoding efficiency of the proposed method compared to the fixed bitrate ladder encoding.

3.2. Experimental Results

This section presents the results of the proposed scheme. The optimal resolution scaling factor s_G is determined manually using the brute force approach for each target bitrate for every segment, which is used as the ground truth. The prediction accuracy of the bitrate ladder construction algorithm is determined by the L_2 norm of the s_G and the selected opti-

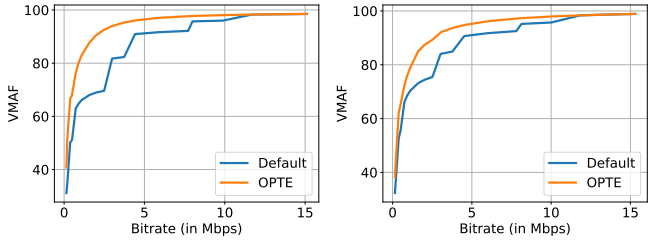
Table 1: Results of OPTE against fixed bitrate ladder approach.

Dataset	Video	f	SI	TI	E	h	$\| s_G - s \|_2$	BDR_V	BDR_P
MCML	Bunny	30	23.38	6.43	23.03	4.88	0.01	-39.48%	-32.25%
MCML	Characters	30	50.43	29.85	41.44	29.21	0.04	-51.90%	-68.81%
MCML	Crowd	30	33.76	10.13	33.11	12.22	0.01	-29.82%	-14.18%
MCML	Dolls	30	16.88	19.91	10.47	0.27	0.02	-1.43%	-8.49%
SJTU	BundNightScape	30	48.82	7.06	54.90	11.62	0.02	-61.22%	-60.86%
SJTU	Fountains	30	43.37	11.42	60.90	23.02	0.02	-32.93%	-8.49%
SJTU	TrafficFlow	30	33.57	13.80	58.93	15.83	0.02	-50.54%	-40.90%
SJTU	TreeShade	30	52.88	5.29	80.19	8.83	0.01	-47.76%	-38.55%
VQEG	CrowdRun	50	50.77	22.33	96.55	33.33	0.01	-8.50%	-1.90%
VQEG	DucksTakeOff	50	47.77	15.10	119.12	30.88	0.01	-2.99%	-2.79%
VQEG	IntoTree	50	24.41	12.09	74.45	21.95	0.03	-26.50%	-5.75%
VQEG	OldTownCross	50	29.66	11.62	92.75	22.06	0.02	-30.91%	-22.53%
VQEG	ParkJoy	50	62.78	27.00	102.80	52.15	0.02	-12.08%	-2.62%
JVET	CatRobot	60	44.45	11.84	56.36	14.25	0.01	-13.43%	-5.95%
JVET	DaylightRoad2	60	40.51	16.21	66.40	20.13	0.02	-27.52%	-9.35%
JVET	FoodMarket4	60	38.26	17.68	50.71	20.71	0.02	-18.11%	-3.74%
Average							0.02	**-28.45%**	**-20.45%**



(a) *BundNightScape* (b) *Bunny*

Fig. 4: Optimized resolution scaling factor ($\hat{s}$) prediction (*cf.* Eq. 7) results for the first segment of *BundNightScape* and *Bunny* sequences.



(a) *BundNightScape* (b) *Bunny*

Fig. 5: Comparison of RD curves for encoding the first segment of *BundNightScape* and *Bunny* sequences using the fixed bitrate ladder and OPTE.

mized resolution scaling factor ($\hat{s}$), *i.e.*, $|| \ s_G - \hat{s} \ ||_2$. The prediction accuracy is depicted graphically in Fig. 4 using the results of the *BundNightScape* and *Bunny* sequences. Table 1 summarizes the $|| \ s_G - \hat{s} \ ||_2$, BDR_P and BDR_V results of various test sequences using OPTE scheme. The average $|| \ s_G - \hat{s} \ ||_2$ is observed as 0.02. The prediction error ranges from 0.01 to 0.04, which is negligible. The average speed of bitrate ladder prediction for each segment is 370 fps using VCA [17] with 8 CPU threads. On average, BDR_P and BDR_V of OPTE compared to the fixed bitrate ladder approach are -20.45% and -28.45%, respectively. The *BundNightScape* sequence yields the highest BDR_P and BDR_V of -60.86% and -61.22%, respectively. Fig. 5 shows the comparison of RD curves of encoding the first segment encodings of *BundNightScape* and *Bunny* sequences using the fixed bitrate ladder and OPTE scheme.

4. CONCLUSIONS AND FUTURE WORK

This paper proposed OPTE, an online per-title encoding scheme for live streaming applications. The proposed scheme includes a bitrate ladder prediction algorithm that predicts the optimal resolution for a given bitrate for each segment, which helps improve the QoE of live streaming. DCT-energy-based features are used to determine segments' spatial and temporal complexity, which is fast and effective. The performance of OPTE is analyzed using the x265 open-source HEVC encoder for an HLS compliant ABR ladder. It is observed that live streaming using OPTE requires 20.45% fewer bits to maintain the same PSNR and 28.45% fewer bits to maintain the same VMAF.

In the future, since the content in a segment does not vary significantly, the bitrate ladder can be predicted with a portion of the segment (*e.g.*, a group of pictures (GOP)), reducing the prediction time without significant error.

5. ACKNOWLEDGMENT

The financial support of the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and the Christian Doppler Research Association is gratefully acknowledged. Christian Doppler Laboratory ATHENA: https://athena.itec.aau.at/.

6. REFERENCES

[1] Apple Inc., "HLS Authoring Specification for Apple Devices," Jun. 2020. [Online]. Available: https://developer.apple.com/documentation/http_live_streaming/hls_authoring_specification_for_apple_devices

[2] J. De Cock, Z. Li, M. Manohara, and A. Aaron, "Complexity-based consistent-quality encoding in the cloud," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016.

[3] A. Mercat, M. Viitanen, and J. Vanne, *UVG Dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development*. New York, NY, USA: Association for Computing Machinery, 2020, p. 297–302. [Online]. Available: https://doi.org/10.1145/3339825.3394937

[4] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 3407–3411.

[5] V. P. K. Malladi, C. Timmerer, and H. Hellwagner, "MiPSO: Multi-Period Per-Scene Optimization For HTTP Adaptive Streaming," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[6] H. Amirpour, C. Timmerer, and M. Ghanbari, "PSTR: Per-Title Encoding Using Spatio-Temporal Resolutions," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[7] Bitmovin, "White Paper: Per Title Encoding," 2018. [Online]. Available: https://bitmovin.com/whitepapers/Bitmovin-Per-Title.pdf

[8] A. V. Katsenou, J. Sole, and D. R. Bull, "Content-gnostic bitrate ladder prediction for adaptive video streaming," in *2019 Picture Coding Symposium (PCS)*, 2019.

[9] M. Bhat, J.-M. Thiesse, and P. L. Callet, "Combining Video Quality Metrics To Select Perceptually Accurate Resolution In A Wide Quality Range: A Case Study," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2164–2168.

[10] ——, "A Case Study of Machine Learning Classifiers for Real-Time Adaptive Resolution Prediction in Video Coding," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[11] Bitmovin, "5th Annual Video Developer Report," Dec. 2021. [Online]. Available: https://go.bitmovin.com/video-developer-report

[12] V. V. Menon, H. Amirpour, M. Ghanbari, and C. Timmerer, "Efficient Content-Adaptive Feature-Based Shot Detection for HTTP Adaptive Streaming," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2174–2178.

[13] ——, "CODA: Content-aware Frame Dropping Algorithm for High Frame-rate Video Streaming," in *2022 Data Compression Conference (DCC)*, 2022.

[14] V. V. Menon, H. Amirpour, C. Timmerer, and M. Ghanbari, "INCEPT: INTRA CU Depth Prediction for HEVC," in *2021 IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2021.

[15] M. King, Z. Tauber, and Z.-N. Li, "A New Energy Function for Segmentation and Compression," in *2007 IEEE International Conference on Multimedia and Expo*, 2007, pp. 1647–1650.

[16] MulticoreWare Inc., "x265 HEVC Encoder/H.265 Video Codec." [Online]. Available: http://x265.org/

[17] V. V. Menon, C. Feldmann, H. Amirpour, M. Ghanbari, and C. Timmerer, "VCA," Feb. 2022. [Online]. Available: https://github.com/cd-athena/VCA

[18] J. Boyce, K. Suehring, X. Li, and V. Seregin, "JVET-J1010: JVET common test conditions and software reference configurations," Jul. 2018.

[19] M. Cheon and J.-S. Lee, "Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 7, pp. 1467–1480, 2018.

[20] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K Video Sequence Dataset," *Fifth International Workshop on Quality of Multimedia Experience (QoMEX2013)*, Jul. 2013.

[21] FFmpeg, "Ffmpeg documentation." [Online]. Available: https://ffmpeg.org/ffmpeg.html

[22] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. Cock, "VMAF: The journey continues," *Netflix Technology Blog*, vol. 25, 2018.

[23] Netflix, "VMAF 4K model." [Online]. Available: https://github.com/Netflix/vmaf/tree/master/model/other_models/vmaf_4k_v0.6.1.pkl

[24] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *VCEG-M33*, 2001.