Preprint: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, June 2021

AUTOENCODER FOR VIBROTACTILE SIGNAL COMPRESSION

Zhuoran Li[1], *Rania Hassen*[2,3], *Zhou Wang*[1]

[1]Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
[2]Human-Machine Interaction (HMI) Lab, Toronto Research Centre,
HUAWEI Technologies Canada CO., LTD.
[3]Computer Science Department, Assiut University, Egypt

ABSTRACT

Vibrotactile signals contain rich haptic information about textured surfaces but their large data volume makes it a challenging task to transmit such signals to remote locations to create immersive and realistic user experiences. Inspired by the recent success of deep neural network (DNN) based autoencoder, we make the first attempt to apply autoencoder for lossy compression of haptic vibrotactile signals, where a convolutional neural network (CNN) and a rate-distortion (RD) function are used as the transform and cost functions, respectively. Performance comparisons with state-of-the-art methods using both peak signal-to-noise ratio (PSNR) and perceptually motivated spectral temporal similarity (ST-SIM) measures show that the proposed end-to-end vibrotactile autoencoder (EVA) is highly competitive at preserving signal quality while keeping the data rate low.

Index Terms— haptic communication, vibrotactile signals compression, autoencoder, deep neural network, BD-rate

1. INTRODUCTION

Humans excel at gathering and sensing tactile feedback during interactions with the surrounding environment [1]. Exploring a textured surface using a mediated tool or bare finger, results in induced high-frequency vibrations (referred to as vibrotactile signals) that pertain to the roughness (macro- and micro-roughness) submodality of this surface [2]. These vibrations are thought to convey useful information about the identity of the surface [3, 4, 5, 6]. Points of interaction with a textured surface collectively create an immense amount of tactile information that needs to be processed, let alone to be transmitted and multiplexed with audio and video signals. Therefore, efficient tactile data compression techniques become a pressing choice to deliver tactile textures essential for creating immersive and realistic user experiences. Owing to the recent standardization efforts of Tactile Internet (TI) [7], haptic codecs activities [8], and the wide variety of vibrotactile rendering tools with different technology flavors that create the illusion of textured surfaces [9, 10, 11], there has been a growing demand to store/transmit/recreate tactile interactions over media with limited capacity. Accordingly, efforts have been devoted to the critical need of vibrotactile signal compression techniques that ensure not only low bit rates but also minimum perceived loss of signal quality [12, 13, 14, 15, 16].

The Haptic Codec subgroup of the IEEE P1918.1 Tactile Internet Group part of IEEE-SA has devoted great effort to define data reduction algorithms and schemes for the communication of kinesthetic, tactile, or combination of kinesthetic/tactile information. Tactile codec requirements and call for contributions were published and described in [8]. Two competing vibrotactile codecs were submitted in response to the call for contributions. The first is based on traditional linear transform coding using discrete Wavelet transform [16], in which the Wavelet coefficients are quantized by optimizing the SNR in each Wavelet band using a bit allocation procedure guided by a psychohaptic model. The second codec called PVC-SLP, initially presented in [12] and further evolved to [13], is based on sparse linear prediction coding and perceptual quantization using novel tactile sensitivity model. Subjective experiments conducted by psychology experts over a set of materials that vary across the tactile microscopic roughness dimension shows the superiority of PVC-SLP in preserving high perceptual quality while keeping the bit-rate low.

Inspired by the recent success of deep neural network (DNN) based autoencoder in image compression [17], here we make the first attempt to apply autoencoder for lossy compression of vibrotactile signals. Our performance comparison shows that the proposed end-to-end vibrotactile autoencoder (EVA) is highly competitive, especially at low bit rates.

2. RELATED WORK

In the literature, few efforts had been devoted to vibrotactile signal compression [14, 15, 18]. Early methods focused on optimizing mathematically tractable measures such as the *mean-square error* (MSE) and the *signal-to-noise ratio* (SNR), which are known to poorly predict perceptual quality. In [14], an ITU standard speech coding technique, namely CS-ACELP, was adopted for vibrotactile signal coding. The

codec is based on linear predictive modeling of speech segments followed by algebraic code approximation of the associated LP filter excitation. The codec is developed based on the assumption that maintaining high SNR values across the entire frequency spectrum will ideally keep the coding distortion imperceptible. In [15] and [16] the masking phenomena within the Pacinian channel was developed by assimilating the auditory masking properties of a human ear. However, unlike what we know in audition and vision, vibrotactile perception is mediated by more than one channel (four-channels categorized to P-channel and three non-P-channels). For a complex vibrotactile stimuli with more than one frequency, it is important to take into consideration the masking effect across channels in perceptual modeling.

Okamoto et al. presented a compression scheme for vibrotactile material-like texture [18]. A surface height profile is constructed as a waveform of texture height as a function of lateral distance, which is subsequently converted to discrete cosine transform (DCT) domain, thresholded, and quantized based on differential thresholds for the vibratory amplitudes [19]. The algorithm is reported to reduce data size up to 75% while preserving subjective quality. It should be noted that the surface height profiles do not match the complex vibrations produced when the finger strokes a textured surface. Moreover, the height profile of each surface needs to be available, which restricts this algorithm to offline applications only.

In [12], a vibrotactile compression method based on sparse linear prediction coding was introduced. The work was further evolved into the PVC-SLP vibrotactile codec [13]. PVC-SLP perceptually quantizes the prediction residuals in the DCT domain using a cutaneous sensitivity model inspired from the four-channel mediation of tactile sensation in the glabrous skin of the human somatosensory periphery.

3. VIBROTACTILE SIGNAL COMPRESSION

The goal of signal compression is to achieve the best quality under limited data rate budget, which can be expressed mathematically as a rate-distortion optimization (RDO) problem. The rate-distortion (RD) cost function is given by $R + \lambda D$, where R represents the data rate, usually measured in the unit of bits per second or bits per sample, D represents the distortion, usually measured in mean square error or perceptually inspired quality metrics, and the Lagrange multiplier λ balances the trade-off between rate and distortion. The proposed general nonlinear transform coding framework directly optimizes for the RD cost function in and end-to-end manner.

Fig.1[17] illustrates the proposed coding framework, where x and $\hat{x}$ represent the original pristine and reconstructed signals in data space, respectively, y and $\hat{y}$ represent the corresponding coded signals in the continuous code space, and z and $\hat{z}$ represent the transformed signals in the perceptual space. In this framework, a vector of signal values $x \in \mathbb{R}^N$ is mapped to the latent code space by a parametric analysis transform, $y = g_a(x, \phi)$, where ϕ represents the vector of parameters that need to be optimized. After the analysis transform, the coding space representation y is quantized, producing a discrete-valued vector $q \in \mathbb{Z}^m$, which is compressed afterwards using context-adaptive binary arithmetic coding (CABAC)[20]. The rate R of the discrete code is lower-bounded by the entropy of the quantized vector $H[P_q]$. To reconstruct the signal, the quantized signal is mapped back to the continuous code space. Then the parametric synthesis transform is applied on the dequantized signal $\hat{y}$, $\hat{x} = g_s(\hat{y}, \theta)$, where θ represents another vector of parameters that need to be optimized. The distortion is computed by transforming to a perceptual space using the transform $\hat{z} = g_p(\hat{x})$, followed by evaluating a distortion metric $d(z, \hat{z})$. The perceptual transform in the proposed method is the identity transform, other perceptually meaningful transform can be applied as well. We optimize the parameter vectors ϕ and θ for a weighted sum of the rate and distortion measures, $R + \lambda D$, over a set of vibrotactile signals. As in [17], CNNs are used to implement the analysis and synthesis transforms (but with 2D convolution filters replaced by 1D filters), allowing for end-to-end learning and testing. There are three layers in the analysis transform, each consisting of a convolutional layer (with 128 kernels of sizes 9, 5, 5), followed by a down-sample layer (of downsampling factors 4, 2, 2) and a generalized divisive normalization (GDN) layer. The synthesis transform is the inverse of the analysis transform.

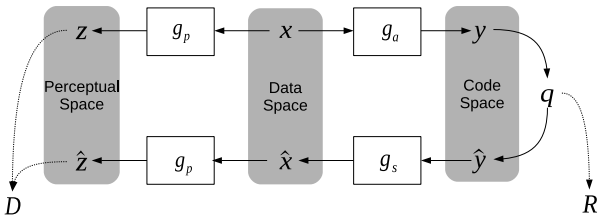


Fig. 1. Autoencoder in a transform coding framework.

The optimization process aims to minimize the RD cost over the parameters of forward, inverse and perceptual transforms. The Lagrange multiplier λ is set to govern the trade-off between rate and distortion. A key difference between our method and previous vibrotactile encoding methods is to directly optimize the RD cost in an end-to-end manner. Furthermore, we rely on the nonlinear transform to warp the space appropriately instead of searching for the optimal quantization scheme over the high dimensional signal space which is nearly intractable. The warping process allows us to use a fixed uniform scalar quantizer in code space, and largely simplifies the coding process. The objective function is defined in terms of entropy as

$$L[g_a, g_s, P_q] = -\mathbb{E}[\log_2 P_q] + \lambda \mathbb{E}[d(z, \hat{z})] \qquad (1)$$

where P_q is the probability mass function of the quantized output vector of the analysis transform.

A technical difficulty is that that the derivatives of the quantization function are zero almost everywhere, making it impossible to execute any gradient descent based optimization methods. As in [17], we replace the quantizer with an additive i.i.d uniform noise source $\triangle y$, which has the same width as the quantization bins (one). Consequently, the continuous relaxation density function of $\tilde{y} = y + \triangle y$ can be used in the gradient descent process

$$p_{\tilde{y}}(n) = P_q(n), \text{ for all } n \in \mathbb{Z}^M \tag{2}$$

With the continuous approximation of the quantized coefficient distribution, the loss function for parameters θ and ϕ across all training samples i is

$$\begin{aligned} L(\theta, \phi) = \mathbb{E}_{x, \triangle y}[-\sum_i \log_2 p_{\tilde{y}_i}(g_a(x;\phi) + \triangle y; \psi^{(i)}) \\ +\lambda d(g_p(g_s(g_a(x;\phi)) + \triangle y; \theta), g_p(x))] \end{aligned} \tag{3}$$

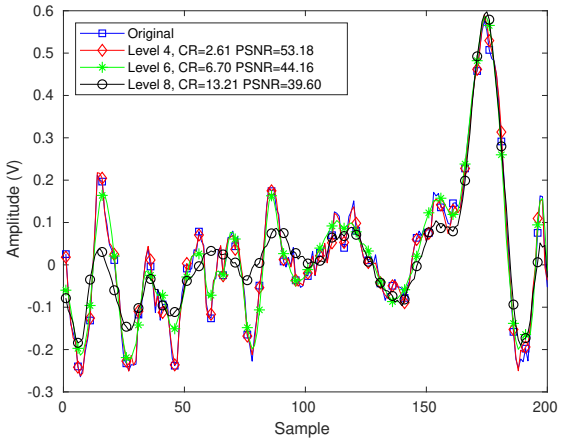
We choose mean squared error (MSE) as the distortion measure d, though any other differentiable quality metric can be adopted in the general framework.

4. EXPERIMENTAL RESULT

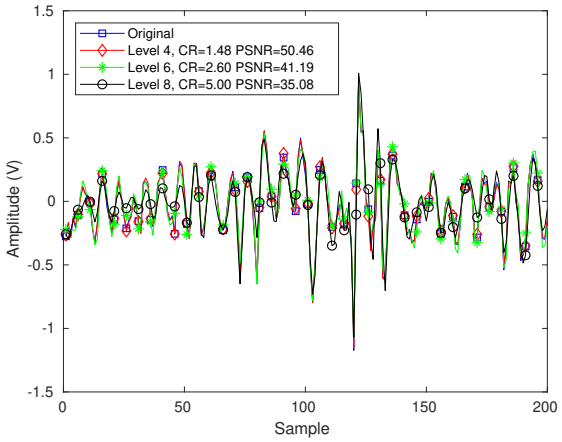
The proposed EVA model is trained on the tactile reference data traces provided along with the IEEE P1918.1.1. standard [21]. The training process is repeated for 8 sets of parameters, corresponding to 8 different λ values, which lead to different levels of compression. The reference tactile database is split into 80/20 training/testing sets. Few vibrotactile signal compression methods are available in the public domain, and PVC-SLP is the top performer in recent IEEE-SA haptic codec subgroup test. Therefore, we compare the RD performance of the proposed EVA model against the standardized PVC-SLP codec. For comparison, the original signal traces are compressed into 22 different compression levels using PVC-SLP codec. In addition to PSNR, we also adopt a novel quality model named ST-SIM [22], which was designed to reflect the perceptual quality of vibrotactile signals.

Fig. 2 shows two testing data traces compressed and then reconstructed by the proposed EVA model at 3 compression levels, in comparison with the original signals on the first 200 signal samples. The signal traces selected are AluminumGrid-Slow and BalticBrown-Slow from the testing dataset, respectively. As can be seen, when the compression level goes up, the fine details of the original signal are gradually lost, while the reconstructed data traces generally follow the smoothed version of the original data trace.

The RD performance of the proposed EVA model outperforms the PVC-SLP codec on almost all testing data traces at all compression levels, especially at the low bit rates. Fig. 3 shows the RD performance comparison between PVC-SLP and EVA models using the scatter plots of compression ratio versus PSNR/ST-SIM. The fitting curves show that the proposed EVA model outperforms PVC-SLP by a large margin, especially at low bit rate levels.



(a) AluminumGrid-Slow

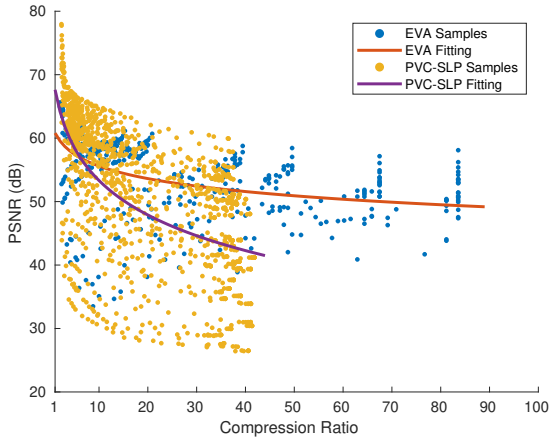


(b) BalticBrown-Slow

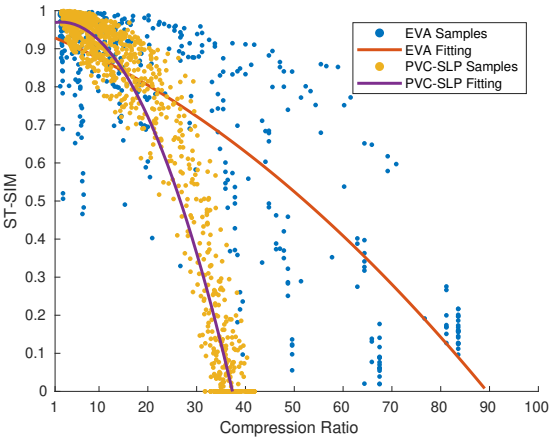
Fig. 2. Comparison of original versus EVA compressed testing signal traces.

In order to quantify the coding gain or rate savings achieved by the proposed model, we borrow the Bjøntegaard Delta (BD) method [23][24], which is widely used in the video compression field to evaluate the relative coding efficiency of one codec against a reference codec [25] over a range of quality-bit rate data points. Given two RD curves produced by two codecs, we compute the BD-rate metric, which estimates the average bit rate savings for the same video quality (in terms of PSNR or ST-SIM). The bit rate saving for a given level of quality is calculated as

$$\Delta R(Q) = \frac{R_B(Q) - R_A(Q)}{R_A(Q)} \tag{4}$$



(a) PSNR over CR



(b) ST-SIM over CR

Fig. 3. ST-SIM and PSNR versus compression ratio (CR) for 56 testing traces encoded at different bit rates.

where $R_A(Q)$ and $R_B(Q)$ are the bitrates for quality level Q on the reference and test RD curves, respectively. Since the logarithmic scale $r = \log R$ is used in the BD model on the bit rate axis, the bitrate saving can be expressed as

$$\Delta R(Q) = 10^{r_B(Q) - r_A(Q)} - 1 \tag{5}$$

Considering both the actual RD points and the fitted RD curves $\hat{r}(Q)$, the BD-rate can be approximated by

$$\Delta R_{Overall} \approx 10^{\frac{1}{Q_H - Q_L} \int_{Q_L}^{Q_H} [\hat{r}_B(Q) - \hat{r}_A(Q)] dQ} - 1 \tag{6}$$

where Q_H is the maximum of the minimum quality that the two curves could reach, and Q_L is the minimum of the maximum quality that the two curves could reach. The region of integration is exemplified as the blue region in Fig. 4.

The rate saving computed by the BD-rate metric varies for different signal content. For the Ceramic-Slow testing

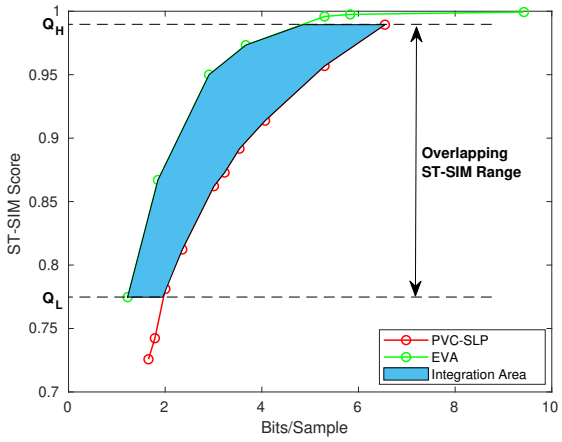


Fig. 4. Sample RD curve comparison (Ceramic-Slow) and BD-rate computation.

data trace shown in Fig. 4, a bitrate saving of 38.7% of EVA against PVC-SLP is achieved. Similarly, we have calculated the BD-rate saving of every testing data trace using both PSNR and ST-SIM as the quality metric. The results show that EVA's performance advantage holds for most of the testing data traces. Using the PSNR and ST-SIM as the quality metrics, the proposed EVA model achieves 14.98% and 13.08% bit rate savings on average when compared against the PVC-SLP codec, respectively.

For computational complexity assessment, our tests on a Intel(R) Core i7-8700 (x64) CPU at 3.20GHz at a compression ratio of 10 show that the encoding time of our current implementations of the PVC-SLP and EVA algorithms is 7.14 seconds and 24.99 seconds per million signal samples, respectively, and the corresponding decoding time is 3.57 seconds and 21.42 seconds per million signal samples, respectively.

5. CONCLUSION

In recent years, tactile digital content has become a trending media source of rapidly growing demand. To handle its large data volume, it is critically important to find efficient data compression methods that can deliver better quality-of-experience under limited data transmission capacity. In this work, we make the first attempt to design a novel haptic vibrotactile signal compression method built upon end-to-end training of a DNN-based autoencoder structure. The proposed EVA model achieves superior performance against state-of-the-art algorithms based on both PSNR and ST-SIM quality measures. In the future, perceptually more meaningful quality measures may be incorporated into the RD optimization framework. Other autoencoder structures may be exploited. It is also desirable to develop precise rate control mechanisms to fulfill any given bit rate requirement.

6. REFERENCES

[1] R. L. Klatzky and S. J. Lederman, "Touch," in *Handbook of psychology, Volume 4: Experimental psychology*, chapter 6, pp. 147–176. John Wiley &Sons, Inc, 2003.

[2] S. Okamoto, H. Nagano, and Y. Yamada, "Psychophysical dimensions of tactile perception of textures," *IEEE Transactions on Haptics*, vol. 6, no. 1, pp. 81–93, 2013.

[3] L. R. Manfredi, H. P. Saal, K. J. Brown, M. C. Zielinski, J. F. Dammann III, V. S. Polashock, and S. J. Bensmaïa, "Natural scenes in tactile texture," *Journal of Neurophysiology*, vol. 111, no. 9, pp. 1792–1802, 2014.

[4] A. I. Weber, H. P. Saal, J. D. Lieber, J. W. Cheng, L. R. Manfredi, J. F. Dammann, and S. J. Bensmaia, "Spatial and temporal codes mediate the tactile perception of natural textures," *Proceedings of the National Academy of Sciences*, p. 201305509, 2013.

[5] S. J. Bensmaïa and M. Hollins, "The vibrations of texture," *Somatosensory & Motor Research*, vol. 20, no. 1, 2003.

[6] S. Bensmaïa and M. Hollins, "Pacinian representations of fine surface texture," *Perception & Psychophysics*, vol. 67, no. 5, pp. 842–854, 2005.

[7] O. Holland, E. Steinbach, R. V. Prasad, Q. Liu, Z. Dawy, A. Aijaz, N. Pappas, K. Chandra, V. S. Rao, S. Oteafy, M. Eid, M. Luden, A. Bhardwaj, X. Liu, J. Sachs, and J. Araújo, "The IEEE 1918.1 "Tactile Internet" Standards Working Group and Its Standards," *Proceedings of the IEEE*, pp. 1–24, 2019.

[8] E. Steinbach, M. Strese, M. Eid, X. Liu, A. Bhardwaj, Q. Liu, M. Al-Ja'afreh, T. Mahmoodi, R. Hassen, A. El Saddik, and O. Holland, "Haptic codecs for the tactile internet," *Proceedings of the IEEE*, pp. 1–24, 2018.

[9] S. Choi and K. J. Kuchenbecker, "Vibrotactile display: Perception, technology, and applications," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2093–2104, 2013.

[10] J. M. Romano and K. J. Kuchenbecker, "Creating realistic virtual textures from contact acceleration data," *IEEE Transactions on Haptics*, vol. 5, no. 2, pp. 109–119, April 2012.

[11] K. J. Kuchenbecker, J. Romano, and W. McMahan, "Haptography: Capturing and recreating the rich feel of real surfaces," in *Robotics Research*, pp. 245–260. Springer, 2011.

[12] R. Hassen and E. Steinbach, "Vibrotactile signal compression based on sparse linear prediction and human tactile sensitivity function," in *IEEE World Haptics Conference (WHC)*, 2019.

[13] R. Hassen, B. Gulecyuz, and E. Steinbach, "PVC-SLP: Perceptual Vibrotactile-Signal Compression based-on Sparse Linear Prediction," *IEEE Transaction on Multimedia (Early Access)*, 2020.

[14] R. Chaudhari, B. Çizmeci, K. J. Kuchenbecker, S. Choi, and E. Steinbach, "Low bitrate source-filter model based compression of vibrotactile texture signals in haptic teleoperation," in *Proceedings of the 20th ACM International Conference on Multimedia*, 2012, pp. 409–418.

[15] R. Chaudhari, C. Schuwerk, M. Danaei, and E. Steinbach, "Perceptual and bitrate-scalable coding of haptic surface texture signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 462–473, April 2015.

[16] A. Noll, B. Gulecyuz, A. Hofmann, and E. Steinbach, "A rate-scalable perceptual wavelet-based vibrotactile codec," in *IEEE Haptics Symposium*, March 2020.

[17] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Int'l Conf on Learning Representations (ICLR)*, Toulon, France, April 2017.

[18] S. Okamoto and Y. Yamada, "Lossy data compression of vibrotactile material-like textures," *IEEE Transactions on Haptics*, vol. 6, no. 1, pp. 69–80, 2013.

[19] S. Okamoto and Y. Yamada, "Perceptual properties of vibrotactile material texture: Effects of amplitude changes and stimuli beneath detection thresholds," in *IEEE International Symposium on System Integration*, 2010, pp. 384–389.

[20] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 620–636, 2003.

[21] IEEE Communications Society, "P1918.1-Tactile Internet: Application Scenarios, Definitions and Terminology, Architecture, Functions, and Technical Assumptions," 2016.

[22] R. Hassen and E. Steinbach, "Subjective Evaluation of the Spectral Temporal SIMilarity (ST-SIM) Measure for Vibrotactile Quality Assessment," *IEEE Transactions on Haptics*, vol. 13, no. 1, pp. 25–31, 2020.

[23] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," in *ITU-T Q. 6/SG16, 33th VCEG Meeting*, 2001.

[24] G. Bjontegaard, "Improvements of the BD-PSNR model, VCEG-AI11," in *ITU-T Q. 6/SG16, 34th VCEG Meeting*, 2008.

[25] T. K. Tan, R. Weerakkody, M. Mrak, N. Ramzan, V. Baroncini, J. Ohm, and G. J. Sullivan, "Video quality evaluation methodology and verification testing of hevc compression performance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 76–90, 2016.