FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Special Issue NGWCN | Pages 20-24 | e-ISSN: 2347-470X

Adaptive Video Coding Framework with Spatial-Temporal Fusion for Optimized Streaming in Next-Generation Networks

Pranob Kumar Charles[1], Habibulla Khan[2*], and K S Rao[3]

[1]*Research Scholar, Dept of ECE, JNTUH, India; Email: pranob2005@gmail.com*
[2*]*Professor - Dept of ECE, K L University India; Email: habibulla@kluniversity.in*
[3]*Principal - Jyothishmathi Institute of Technology and Science (JITS), India; Email: drksraodir@gmail.com*

Correspondence: Habibulla Khan; habibulla@kluniversity.in

ABSTRACT- Predicting future frames and improving inter-frame prediction are ongoing challenges in the field of video streaming. By creating a novel framework called STreamNet (Spatial-Temporal Video Coding), fusing bidirectional long short-term memory with temporal convolutional networks, this work aims to address the issue at hand. The development of STreamNet, which combines spatial hierarchies with local and global temporal dependencies in a seamless manner, along with sophisticated preprocessing, attention mechanisms, residual learning, and effective compression techniques, is the main contribution. Significantly, STreamNet claims to provide improved video coding quality and efficiency, making it suitable for next-generation networks. STreamNet has the potential to provide reliable and optimal streaming in high-demand network environments, as shown by preliminary tests that show a performance advantage over existing methods.

Keywords: Video Coding, Temporal Convolutional Networks, Next-Generation Networks, Spatial-Temporal Fusion, Optimized Streaming, STreamNet, BiLSTM.

ARTICLE INFORMATION

Author(s): Pranob Kumar Charles, Habibulla Khan, and K S Rao;

Received: 10/07/2023; **Accepted**: 29/08/2023; **Published**: 30/12/2023;
E- ISSN: 2347-470X;
Paper Id: IJEER NG-WCN-015;
Citation: 10.37391/IJEER.11ngwcn04
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-11/ijeer-11ngwcn04.html

Crossref member
CROSSREF.ORG THE CITATION LINKING BACKBONE

1. INTRODUCTION

Video coding is fundamental for efficient storage and transmission of video data, especially in a world demanding high-quality content across various platforms [1]. Conventional techniques often fail to capture the complex spatial and temporal interactions in video data, leading to suboptimal compression and quality loss [2]. This necessitates innovative solutions that align with developing network capabilities and user expectations for high-quality, low-latency video [3]. STreamNet, a proposed framework, addresses this by combining Temporal Convolutional Networks (TCNs) and Bi-directional Long Short-Term Memory (BiLSTM) for precise spatial-temporal fusion, fast coding, optimized inter-frame prediction, and accurate future frame prediction. Unique features of STreamNet include a novel combination of loss functions accounting for numerical accuracy and perceptual quality, and a Temporal Self-Attention Mechanism focusing on relevant temporal features for each frame prediction. Built upon contemporary machine learning and video compression techniques, and strengthened by advanced preprocessing, attention mechanisms, residual learning, and optimal quantization, STreamNet aims to redefine video coding standards and adapt to next-generation networks' needs, providing reliable performance under various circumstances [4-9].

The article explains the importance of video coding and the problems solved by a proposed solution called STreamNet. It includes a review of past research, a description of how STreamNet works, and an experiment to test its effectiveness. The conclusion summarizes the findings, discusses the potential uses of STreamNet, and suggests ideas for future research. The article includes a well-organized list of references.

2. RELATED RESEARCH

The review includes various research efforts employing machine learning (ML) and deep learning (DL) techniques for video coding and compression optimization. A ML-based coding unit (CU) depth decision method for High Efficiency Video. Deep learning is explored for video coding quality analysis [3]. Coding efficiency is enhanced using a Squeeze-and-Excitation Filtering CNN (SEFCNN) structure [10]. A low-complexity in-loop filter model is presented for mobile multimedia [11]. A neural network-based inter-prediction scheme is introduced for video compression [12], and deep learning is employed for low-complexity error resilient video coding [13]. ML-based solutions are proposed for HTTP adaptive streaming [14], and a reinforcement learning framework is introduced for frame-level bit allocation in HEVC/H.265 [15]. A deep convolutional neural network (DCNN) is employed for enhancing video quality in versatile video coding (VVC) [16], and human vision models and ML are leveraged for H.266/VVC encoding [17]. Deep learning is used for video streaming over the next-generation network,

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Special Issue NGWCN | Pages 20-24 | e-ISSN: 2347-470X

and a learning-based video compression framework achieves a 30.1% BD-rate reduction compared to HEVC. A deep CNN-LSTM framework is presented for fast video coding, and an end-to-end deep video codec is proposed for efficient video compression in 5G/B5G. A DL-based method is proposed for intra mode derivation in VVC, and a novel "convLSTM" approach is introduced for video prediction and compression. A DL-based methodology is proposed for predicting video streaming quality, power consumption, and bandwidth requirements.

Overall, the reviewed articles highlight the trend towards using ML and DL techniques to address challenges in video coding and compression, such as error resilience, real-time encoding, network integration, and optimizing complexity in HEVC. The innovation in video coding has expanded with the introduction of advanced techniques like generative adversarial networks (GANs) and on-the-fly decision-making algorithms, indicating the necessity of the proposed comprehensive platform, STreamNet, for video compression, quality enhancement, and adaptive streaming.

3. METHODS AND MATERIALS

The STreamNet framework addresses the need for innovative video coding solutions suitable for next-generation networks. It uses Temporal Convolutional Networks (TCNs) and Bi-directional Long Short-Term Memory (BiLSTM) for effective and efficient video compression by optimizing inter-frame and future frame prediction, and enabling spatial-temporal fusion. Features like the Temporal Self-Attention Mechanism enhance its capabilities. STreamNet is designed for modern digital communication needs, emphasizing effectiveness, efficiency, and quality, and offers novel opportunities in video streaming. This following description examines the framework's design that shown in *figure 1* in response to the evolving technological landscape. The STreamNet video coding framework includes several key steps:

1. *Preprocessing*: Pixel values are scaled using Min-Max Normalization, and the video frames' color space is converted to YUV to enhance compression efficiency.
2. *Spatial Feature Extraction*: TCNs with dilated convolutions and MAX POOLING are used to capture spatial features and reduce dimensionality.
3. *Temporal Feature Extraction:* TCNs, BiLSTM networks, and a Temporal Self-Attention Mechanism are employed to model local and global temporal dependencies and focus on relevant temporal features.
4. *Residual Connections*: Used to enable direct learning of inter-frame differences.
5. *Prediction Translation*: The combined spatial and temporal features are translated to the required output format using fully connected layers, activation functions, and a loss function that combines MSE and SSIM.
6.*Quantization & Entropy Coding*: Uniform Scalar Quantization and Huffman Coding are applied for optimal compression and efficient lossless compression, exploiting parallelism in both TCNs and BiLSTM for efficient computation.

Preprocessing of video data in StreamNet involves Min-Max Normalization and YUV color space transformation. The normalization scales pixel values to a standardized range between 0 and 1, using the formula, where is the input pixel value, and 'min' and 'max' are the minimum and maximum pixel values, respectively. The YUV color space transformation separates video frames into luminance and chrominance (U and V) components using the formulas $Y = 0.299R + 0.587G + 0.114B$, $U = -0.147R - 0.289G + 0.436B$, and $V = 0.615R - 0.515G - 0.100 * B$, where are the red, green, and blue channels of the original pixel.

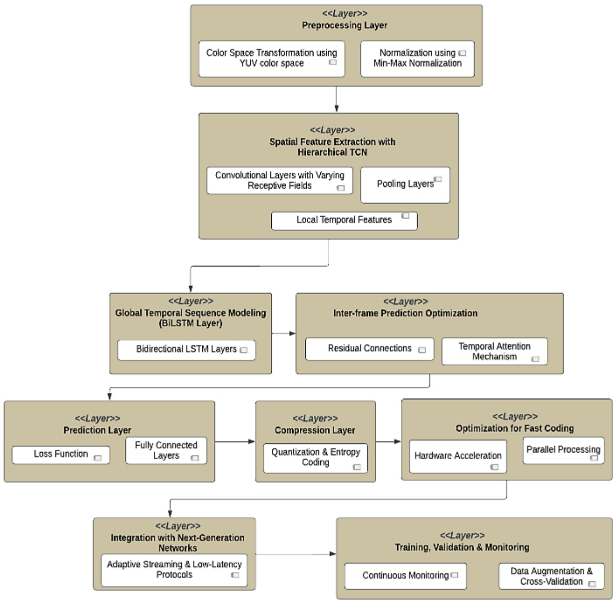

Figure 1: Flow Diagram of STreamNet Framework

Spatial Feature Extraction involves using Temporal Convolutional Networks (TCNs) with dilated convolutions and MAX POOLING to extract spatial features and reduce dimensionality while maintaining hierarchical information.

The TCN operation for a given layer $'l'$ and dilation factor $'d'$ is expressed as $output[i] = sum(input[i + jd]weight[j]$ for j in $range$ $(0, kernel_{size}))$, and the MAX POOLING operation for a given pool size $'p'$ and stride $'s'$ is expressed as $output[i, j] = max(input[k, l]$ for k in $range(is, is + p)$ and l in $range(js, js + p))$.

Temporal Feature Extraction involves using TCNs, Bi-directional Long Short-Term Memory (BiLSTM) networks, and a Temporal Self-Attention Mechanism to capture local and global temporal patterns in video sequences. The TCN operation is expressed as $output[t] = activation(sum(F[t + k]weight[k]$ for k in $range(-kernel_{size} / 2, kernel_{size} / 2)))$, the BiLSTM operation involves processing the data both forward and backward to capture long-term dependencies, and the Temporal Self-

Attention Mechanism is expressed as Attention $Scores: scores = softmax(QK^{T} / sqrt(d))\, and\, Output: Y = scores * V$, where $Q, K\, and\, V$ are the query, key, and value matrices, and d is the dimensionality of the keys.

Direct Learning of Inter-Frame Differences involves using residual connections to promote faster convergence, solve the vanishing gradient issue, and achieve more precise predictions of upcoming frames. The residual connection operation is expressed as output=F(input)+input.

Prediction Translation involves translating the learned spatial and temporal features into the desired output format. This is achieved by combining the spatial and temporal features (Combined $Features: C = concatenate(S, T)$), passing them through fully connected layers with activation functions (Fully Connected $Operation: Y = activation(WC + b)$ and Reshaping $Operation: final_{output} = reshape(Y, output_{shape})$), and using a loss function that incorporates both Mean Squared Error (MSE) and Structural Similarity Index (SSIM) $(Loss = alphaMSE + beta*(1 - SSIM))$.

Quantization & Entropy Coding involves using Uniform Scalar Quantization and Huffman Coding for effective video data compression. The quantization operation is expressed as Quantized $Value: Q = round((value - min) / step_{size}) * step_{size} + min$, and the Huffman Coding involves calculating symbol frequencies, building a priority queue based on frequencies, dequeuing two nodes with the lowest frequency and enqueuing a new node with the sum of their frequencies, and traversing the tree to assign binary codes to symbols based on their path from the root.

4. EXPERIMENTAL STUDY

The experimental study evaluates the performance of deep learning-based video coding models, including STreamNet, DLIMD [18], and convLSTM [19], using five different datasets [20]: UCF10, HMDB51, Hollywood2, ImageNet VID, and 20BN. The evaluation metrics include BPP, Bit Rate, PSNR [21], MS-SSIM [22], MSE, Compression Ratio, Encoding/Decoding Time, and BDBR [23]. Experiments use 10-fold cross-validation with 80% data for training, 10% for validation, and 10% for testing. Each dataset focuses on different aspects, like human movements, actions, facial expressions, human-object interactions, object detection and recognition, and cinematic scenarios. NVIDIA RTX 2080Ti GPU, Python, and its companion libraries were used for implementation.

4.1 Dataset

The UCF101 dataset covers 101 action categories, allowing an assessment of STreamNet's efficiency in handling various motions and activities. The HMDB51 dataset includes human actions, facial expressions, and human-object interactions, assessing STreamNet's proficiency in capturing and encoding human movements. The Hollywood2 dataset comprises video clips from movies, examining STreamNet's adaptability to cinematic scenarios. The ImageNet VID dataset emphasizes object detection and recognition, evaluating STreamNet's capabilities in object recognition tasks. The 20BN-something-something Dataset [31] underlines human-object interactions, analyzing STreamNet's competence in grasping multifaceted relationships.

4.2 Performance Analysis

The STreamNet model's performance is evaluated using key metrics such as BPP, Compression Ratios, Bit Rates, Encoding/Decoding Times, PSNR, and BDBR across five datasets, as detailed in *table 1*, *table 2* and *table 3* provide the mean and deviation of key metrics for DLIMD and convLSTM models across the same datasets and the results are shown in *figure 2*. The comparative study section indicates that STreamNet consistently outperforms DLIMD and convLSTM across multiple datasets, often leading in BPP, compression ratio, bit rate, PSNR, and BDBR. Rate-Distortion (R-D) curves comparing normalized bit-rate and PSNR for STreamNet, DLIMD, and convLSTM across diverse datasets, reinforcing that the superior performance of STreamNet in terms of efficiency, effectiveness, and adaptability across various content and datasets.

Table 1: mean and deviation values of results observed from the experiments conducted on STreamNet

Dataset	**UCF101**	**HMDB51**	**Hollywood2**	**VID**	**20BN**
BPP	0.17±0.01	0.019±0.001	0.19±0.01	0.15±0.01	0.25±0.01
Compression Ratio	20:01±1:00	16:01±11:00	30:01±1:00	20:01±0:30	20.5±0.5
Bit Rate (kbps)	600±15	297.5±5	300±7	650±10	695±15
Coding Time (ms)	15±1	14.4±0.5	12.2±0.6	12±0.8	24.1±0.6
PSNR	38.2±0.6	39±0.8	38.6±0.5	38.6±0.5	35.5±0.5
BDBR	-3.5±0.2%	-6±0.8%	-5.5±0.5%	-4±1%	-3.3±0.3%

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Special Issue NGWCN | Pages 20-24 | e-ISSN: 2347-470X

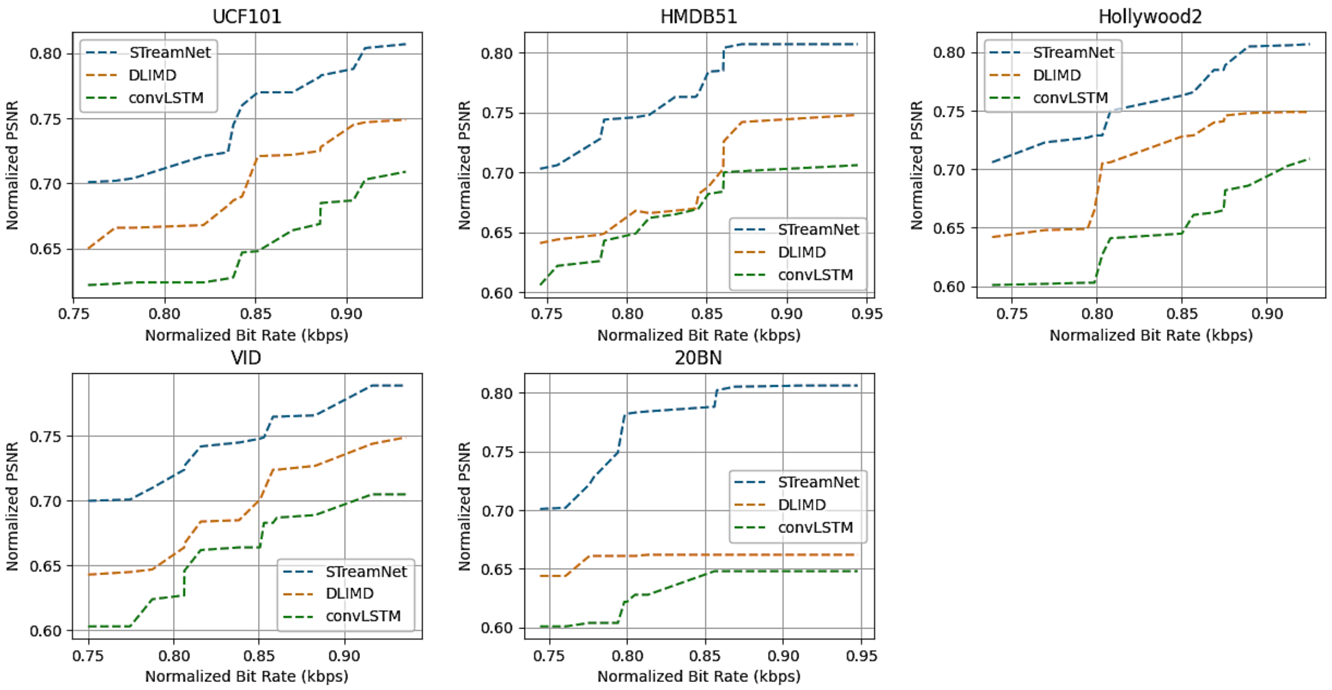

Figure 2: Rate-Distortion Curves Comparing Normalized Bit-Rate and PSNR for STreamNet, DLIMD, and convLSTM Across Diverse Datasets

Table 2: mean and deviation values of results observed from the experiments conducted on DLIMD

Dataset	UCF101	HMDB51	Hollywood2	VID	20BN
BPP	0.25±0.01	0.025±0.001	0.29±0.01	0.18±0.01	0.27±0.01
Compression Ratio	15:01±0.58	17:01±1.11	25:01±1.05	17:01±0.44	18.3±0.46
Bit Rate (kbps)	808±14	320±5	348±8	721±6	743±14
Coding Time (ms)	15.1±0.7	16.4±0.7	15.2±0.8	14.6±0.5	29.7±0.9
PSNR	34.5±0.4	36±0.9	36±0.9	36.9±0.7	34.3±0.2
BDBR	-2.9±0.2%	-3.2±0.8%	-3±1%	-1.6±0.5%	-2.2±0.3%

Table 3: mean and deviation values of results observed from the experiments conducted on convLSTM

Dataset➔	UCF101	HMDB51	Hollywood2	VID	20BN
BPP	0.25±0.01	0.023±0.001	0.25±0.01	0.18±0.01	0.27±0.01
Compression Ratio	15:01±0.60	19:01±1.07	26:31:00±1.04	18:01±0.63	18.5±0.5
Bit Rate (kbps)	749±15	311±5	327±5	713±10	738±12
Coding Time (ms)	15.0±0.7	15.6±0.5	14.2±0.7	14.5±0.6	28.8±0.8
PSNR	35±0.6	36±0.8	36±0.8	37.4±0.7	34.3±0.2
BDBR	-3.0±0.2	-3.6±1.1	-3±1%	-2.2±0.8%	-2.3±0.3%

5. CONCLUSION

The article introduces STreamNet, a novel video coding framework designed to address spatial and temporal complexities in video data. It incorporates Temporal Convolutional Networks (TCNs), Bi-directional Long Short-Term Memory (BiLSTM), Temporal Self-Attention, and specialized loss functions to improve inter-frame prediction and ensure numerical accuracy and perceptual quality. Experiments on five datasets show STreamNet outperforms contemporary models in BPP, compression ratio, PSNR, and BDBR, demonstrating its flexibility, effectiveness, and visual quality maintenance. The framework's design and strong

Research Article | Volume 11, Special Issue NGWCN | Pages 20-24 | e-ISSN: 2347-470X

performance indicate its potential to optimize streaming in next-generation networks and transform the digital landscape.

REFERENCES

[1] Zhang, Yun, Sam Kwong, Xu Wang, Hui Yuan, Zhaoqing Pan, and Long Xu. (2015), "Machine learning-based coding unit depth decisions for flexible complexity allocation in high efficiency video coding." IEEE Transactions on Image Processing 24, no. 7 (2015): 2225-2238.

[2] Puri, Saurabh. (2017), "Learning, selection and coding of new block transforms in and for the optimization loop of video coders." PhD diss., Nantes, 2017.

[3] Topiwala, Pankaj, Madhu Krishnan, and Wei Dai. (2018), "Deep learning techniques in video coding and quality analysis." In Applications of Digital Image Processing XLI, vol. 10752, pp. 353-367. SPIE, 2018.

[4] Wiegand, Thomas, Gary J. Sullivan, et al., (2003), "Overview of the H. 264/AVC video coding standard." IEEE Transactions on circuits and systems for video technology 13, no. 7 560-576.

[5] Wang, Tingting, Mingjin Chen, and Hongyang Chao. (2017), "A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC." In 2017 data compression conference (DCC), pp. 410-419. IEEE, 2017.

[6] Zhang, Yun, Sam Kwong, and Shiqi Wang. (2020), "Machine learning based video coding optimizations: A survey." Information Sciences 506 (2020): 395-423.

[7] Liu, Dong, Yue Li, Jianping Lin, Houqiang Li, and Feng Wu. (2020), "Deep learning-based video coding: A review and a case study." ACM Computing Surveys (CSUR) 53, no. 1 (2020): 1-35.

[8] Dharwadkar, Shri N., and Nabegha Masood. (2007), "Next Generation Network." In 2007 IEEE International Symposium on Consumer Electronics, pp. 1-4. IEEE, 2007.

[9] Modarressi, Abdi R., and Seshadri Mohan. (2000), "Control and management in next-generation networks: challenges and opportunities." IEEE Communications Magazine 38, no. 10 (2000): 94-102.

[10] Ding, Dandan, Lingyi Kong, Guangyao Chen, Zoe Liu, and Yong Fang. (2019), "A switchable deep learning approach for in-loop filtering in video coding." IEEE Transactions on Circuits and Systems for Video Technology 30, no. 7 (2019): 1871-1887.

[11] Liu, Chao, Heming Sun, Jiro Katto, Xiaoyang Zeng, and Yibo Fan. (2020), "A learning-based low complexity in-loop filter for video coding." In 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1-6. IEEE, 2020.

[12] Murn, Luka, Saverio Blasi, Alan F. Smeaton, and Marta Mrak. (2021), "Improved CNN-based learning of interpolation filters for low-complexity inter prediction in video coding." IEEE Open Journal of Signal Processing 2 (2021): 453-465.

[13] Wang, Taiyu, Fan Li, Xiaoya Qiao, and Pamela C. Cosman. (2020), "Low-Complexity Error Resilient HEVC Video Coding: A Deep Learning Approach." IEEE Transactions on Image Processing 30 (2020): 1245-1260.

[14] Çetinkaya, Ekrem. (2021), "Machine Learning Based Video Coding Enhancements for HTTP Adaptive Streaming." In Proceedings of the 12th ACM Multimedia Systems Conference, pp. 418-422. 2021.

[15] Ho, Yung-Han, Yun Liang, Chia-Hao Kao, and Wen-Hsiao Peng. (2022), "Action-Constrained Reinforcement Learning for Frame-Level Bit Allocation in HEVC/H. 265 through Frank-Wolfe Policy Optimization." arXiv preprint arXiv:2203.05127 (2022).

[16] Bouaafia, Soulef, Randa Khemiri, Seifeddine Messaoud, Olfa Ben Ahmed, and Fatma Ezahra Sayadi. (2022), "Deep learning-based video quality enhancement for the new versatile video coding." Neural Computing and Applications 34, no. 17 (2022): 14135-14149.

[17] Chen, Mei-Juan, Cheng-An Lee, Yu-Hsiang Tsai, Chieh-Ming Yang, Chia-Hung Yeh, Lih-Jen Kau, and Chuan-Yu Chang. (2022), "Efficient partition decision based on visual perception and machine learning for H. 266/versatile video coding." IEEE Access 10 (2022): 42141-42150.

[18] Zhu, Linwei, Yun Zhang, Na Li, Gangyi Jiang, and Sam Kwong. (2023), "Deep Learning-Based Intra Mode Derivation for Versatile Video Coding." ACM Transactions on Multimedia Computing, Communications and Applications 19, no. 2s (2023): 1-20.

[19] Liu, Bowen, Yu Chen, Shiyu Liu, and Hun-Seok Kim. (2021), "Deep learning in latent space for video prediction and compression." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 701-710. 2021.

[20] Zhang, Yun, et al. (2023), "A survey on perceptually optimized video coding." ACM Computing Surveys 55.12 (2023): 1-37.

[21] Korhonen, Jari, and Junyong You. (2012), "Peak signal-to-noise ratio revisited: Is simple beautiful?." In 2012 Fourth International Workshop on Quality of Multimedia Experience, pp. 37-38. IEEE, 2012.

[22] Nasr, M. Abdel-Salam, Mohammed F. et al. (2017), "multi-scale structural similarity index for motion detection." Journal of King Saud University-Computer and Information Sciences 29, no. 3, 399-409.

[23] Barman, Nabajeet, Maria G. Martini, and Yuriy Reznik. (2022), "Revisiting Bjontegaard delta bitrate (BD-BR) computation for codec compression efficiency comparison." In Proceedings of the 1st Mile-High Video Conference, pp. 113-114.