

Bit-rate aware effective inter-layer motion prediction using multi-loop encoding structure

Sandeep Gowdra Siddaramappa, Gowdra Shivanandappa Mamatha

Department of Information Science and Engineering, R. V. College of Engineering,

Affiliated to Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received May 22, 2024

Revised Sep 11, 2024

Accepted Sep 29, 2024

Keywords:

Bit-rate

Deep learning

Entropy coding

Inter-layer

Learned video compression

Motion prediction

ABSTRACT

Recently, there has been a notable increase in the use of video content on the internet, leading for the creation of improved codecs like versatile-video-coding (VVC) and high-efficiency video-coding (HEVC). It is important to note that these video coding techniques continue to demonstrate quality degradation and the presence of noise throughout the decoded frames. A number of deep-learning (DL) algorithm-based network structures have been developed by experts to tackle this problem; nevertheless, because many of these solutions use in-loop filtration, extra bits must be sent among the encoding and decoding layers. Moreover, because they used fewer reference frames, they were unable to extract significant features by taking advantage from the temporal connection between frames. Hence, this paper introduces inter-layer motion prediction aware multi-loop video coding (ILMPA-MLVC) techniques. The ILMPA-MLVC first designs an multi-loop adaptive encoder (MLAE) architecture to enhance inter-layer motion prediction and optimization process; second, this work designs multi-loop probabilistic-bitrate aware compression (MLPBAC) model to attain improved bitrate efficiency with minimal overhead; the training of ILMPA-MLVC is done through novel distortion loss function using UVG dataset; the result shows the proposed ILMPA-MLVC attain improved peak-signal-to-noise-ratio (PSNR) and structural similarity (SSIM) performance in comparison with existing video coding techniques.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sandeep Gowdra Siddaramappa

Department of Information Science and Engineering, R. V. College of Engineering

Affiliated to Visvesvaraya Technological University

Belagavi-590018, India

Email: yashasgowdaniharika@gmail.com

1. INTRODUCTION

The proliferation of portable electronic devices like smartphones and tablets has boosted innovation in the video sector, making video content accessible from anywhere, at any time. High-resolution [1] movies are becoming more integrated into the daily lives of individuals because of technical advancements. To create a more real-life experience, new video codecs are being developed, such as high-frame rate (HFR), high-dynamic-range (HDR) and ultra-high definition (UHD) having 8K and 4K resolutions [2]. Therefore, it is crucial to find ways to compress video information more efficiently without losing quality [3]. Moreover, improved quality of transmission and decreased transmission and storage costs are also possible outcomes of technological advancements in video encoding and decoding [4]. Several global video-coding standards are being developed in order to improve performance even further, increase adaptability, and ease the burden on the Internet. The latest of the video-coding standards is versatile-video-coding (VVC), which is a variant of

high-efficiency video-coding (HEVC). Despite the high compressing performance attained by these standards, they still face a number of issues. Hence, these methods are not capable of meeting the requirements for low latency, high frame-rate, and high resolution [5].

The latest innovation in video and image compression is through end-to-end learning which has been brought about by the development of deep-learning (DL) [6], [7]. When comparing DL-based video and image compression approaches to standard block-based video-compressing approaches, it has been observed that the DL-based approaches can accomplish higher data compressing rates without compromising the image or video quality. Du *et al.* [7] established a significant link among the compression of images along with the hyperprior-based approach, resulting in the development of end-to-end image-compression. Moreover, numerous cutting-edge neural video compression (NVC) approaches [7], [8] use compensation networks and motion estimation to forecast between frames to achieve a short delay configuration. These include pixel-motion convolutional neural network (PMCNN) [8], which relies on hybrid forecasting networks and movement extension; NVC [9], which aggregates connected spatial-temporal assumptions; and deep video compression (DVC) [10], which uses end-to-end deep video encoding instead of the conventional video-coding architectures. These techniques employed variational auto-encoder (VAE) for compressing optical-flow and residuals, and utilized motion vectors estimates, like optical-flow, for representing temporal data contained within the video.

Additionally, several solutions for B-frame compression have been developed by various researchers. For example, Jia *et al.* [11] developed allocation methods and recurrent improvement techniques. Hu *et al.* [12] focused on designing interpolation-based video encoding networks and used optical-flow encoding networks which were capable of decoding both interpolation coefficients and optical-flow simultaneously. Nevertheless, the presented DL-based video encoding [13]-[17] methods have restrictions on the total quantity of references that may be utilized to encode more frames, which limits their capacity to investigate temporal correlations and eliminate redundancies. Furthermore, earlier DL-based method's [18], [19], i.e., entropy coding similarly estimated the probability associated with the hidden feature representations upon every frame separately, neglecting the relationship among adjacent frames hidden feature representations. From a researcher's perspective, it is worth noting that time correlations in the hidden feature area can additionally be investigated using a hybrid learning approach combining convolutional and recurrent networks. Hence, this paper proposes an inter-layer motion prediction-aware multi-loop video coding (ILMPA-MLVC) which is composed of multiloop adaptive encoder and multiloop probabilistic bitrate aware compression model. The presented ILMPA-MLVC method utilizes recurrent networks to represent a given input; reconstructs encoded outputs and models probabilistic function during entropy coding. For more specific information, the multi-layer adaptive-encoder (MLAE) network has been presented where all preceding frames can be considered as points of reference for compressing the present frame. In addition, the given multi-loop probabilistic-bitrate aware compression (MLPBAC) network iteratively constructs the video frame's probability function based on all past hidden feature representations; thus, has the capability of achieving a lower bit-rate for compressing video frames. The main contribution of the is given below:

- The proposed ILMPA-MLVC utilized convolution and recurrent framework within learned video compression to effectively utilize the temporal relationship between a wide range of video frames.
- To increase the number of possible reference frames, ILMPA-MLVC presents a method using an auto-encoder designed combining convolutional and recurrent networks. For estimating the hidden feature representational temporally conditioned probability functions, the ILMPA-MLVC utilized a recurrent probabilistic approach to obtain good bit-rate performance.
- The ILMPA-MLVC method outperformed state-of-the-art video compression techniques throughout the tests, and the performance of ILMPA-MLVC was confirmed through the results.

The paper organization is as follows. In section 2 discusses various current methodologies designed to attain enhanced bitrate performance and highlight metric and dataset used to validate the model and its limitations. In section 3 presents a novel deep learning-based video encoder architecture that is efficient in effectively predicting inter-layer motion prediction and optimizing it for better temporal representation. In section 4, the performance efficiency of proposed approach is studied using UVG dataset in terms of peak-signal-to-noise-ratio (PSNR) and structural similarity (SSIM) and compared with current baseline video compression models. In section 5 discusses the significance of proposed approaches and future research direction.

2. LITERATURE SURVEY

This section studies various existing video coding method to attain improved video quality with better perception and limited computation overhead. Hu *et al.* [12], presented an approach called hyperprior

deep-video compression (HDVC). This work considered end-to-end compression instead of block-based compression using deep learning. They presented a residual-channel attention intermediary module for both decoding and encoding for enhancing the frames during reconstruction. The evaluation was done using MS-SSIM and PSNR on UVG dataset. Findings show that HDVC achieved 32.93 average PSNR and 40.92 MS-SSIM on UVG dataset. Further evaluation was conducted on MCL-JCV dataset where it achieved PSNR of 6.03 and MS-SSIM of 21.85. Das *et al.* [13], presented a neural-network which was based on post-processing approach for enhancing decoded frames. Here, they used quantization-parameter (QP) mapping initially for achieving high-quality images. Further, this work utilized a convolutional neural network (CNN) where the QP map output along with the reconstructed frame from the dataset were used as an input for the CNN layer. The CNN layer extracted n features from the images and finally provided a quality-enhanced frame. This work utilized the videos from MCL-JCV dataset [14], and two metrics were used for evaluation, i.e., peak-signal-to-noise-ratio (PSNR) and Bjøntegaard Delta (BD Δ). Findings from the results showed that, the neural-network achieved BD Δ values of 5.21% for all-intra configurations, 4.13% for low-delay configurations and 4.54% for random access (RA) configuration. The PSNR values also showed better values in comparison with a CNN post-processing approach [15].

Thang and Bang [16], presented an interpolation network approach for improving the prediction of inter-frames. Here, they presented a RA configuration within a neural video coding and a frame network for predicting the next frame. Finally, they presented an end-to-end-hierarchical deep video-compression network and a loss function was used for evaluating the losses for maximizing the quality of frame during reconstruction. This work utilized the UVG dataset [17] and HEVC dataset [18] for evaluating their work. Also, this work utilized the BD Δ and running time as performance metrics. The findings show that for UVG dataset, the interpolation network achieved 48.01% of BD Δ and for HEVC-class B dataset, the interpolation network achieved 50.96% of BD Δ when compared with deep-video-compression (DVC), DVC Pro [19], and scale-space flow (SSF) compression [20] approaches. Yoon *et al.* [21], first presented a textural-detail preservation-network (TDPNet) for analysing the details, noise and texture of the frame in a video. From the analysis, this work presented an approach called perceptual-training method (PTM) for solving issue of PSNR for preserving high number of textural details. Further, they presented a multi-scale resolution-training method (MRTM) for solving problem of poor efficiency when evaluating on different datasets. This work evaluated their work on UVG, SNU-FILM [22], HD [23], and Vimeo90K [24] datasets. Further, the evaluation was done using PSNR, perceptual similarities, DISTS [25], and LPIPS [26]. The findings showed PTM achieved an PSNR of 29.3. He *et al.* [27], presented an approach called deep neural representation for videos (D-NeRV) for encoding various kinds of videos by decoupling specified frame visual-content using motion data, providing temporal decision making for implicit neural networks and using a task-oriented approach for the intermediary output for reducing spatial redundancy. The evaluation was done on two datasets, i.e., UVG, DAVIS [28], and UCF101 [29] dataset. The evaluation was done using multi-scale structural similarity (MS-SSIM) and PSNR. The D-NeRV achieved average PSNR of 35.52 for UVG, 30.06 and 0.951 for PSNR and MS-SSIM for UCF-101 dataset and average PSNR of 21.3 for DAVIS dataset.

Choi *et al.* [30], presented an approach called Bi-directional optical-flow (BDOF) for refining bi-prediction block. Further, this work presented a design called an attention-based Bi-prediction-network (ABPN) for substituting the entire current existing bi-prediction approaches for achieving better video reconstruction. The evaluation was conducted using DVI-DVC [31] dataset. The findings show that the proposed approach achieved better results, i.e., reduced BD Δ by 4.91% and 5.89% for LDB and RA in comparison with existing approach. Kwan *et al.* [32], presented an approach called HiNeRV which contains light-weight layers having layered wise positional-encodings. This work used deep CNN approach, multi-layer-perceptron (MLP) and interpolation networks for encoding the frames of videos and offering a better approach for solving the issue of patches in videos. The evaluation was done using MCL-JCV and UVG dataset. Also, evaluation was conducted using PSNR. Findings show that the HiNeRV approach achieved better bit-rate saving of 72.3% and 43.4% PSNR in comparison with existing approaches. The HiNeRV achieved average PSNR of 36.27 on UVG dataset. Wang *et al.* [33], presented an edge-oriented compressed-video super-resolution network (EOCVSR) for reconstruction of video frames providing better quality details. In this work, they first presented a motion-guided alignment-module (MGAM) for achieving Bi-direction motions in a multi-scale way. Further, presented an edge-oriented recurrent-block (EOB) for reconstructing edges of the frames using implicit and explicit edge extraction approach. This work considered Vimeo, MCL-JCV, UVG, and CTC [34] dataset. Evaluation of EOCVSR was done by considering quality-enhancement PSNR and rate-distortion (BD Δ). The EOCVSR achieved average PSNR of 30.262 for CTC and Vimeo and average PSNR of 31.76 for UVG and MCL-JCV dataset.

3. PROPOSED METHOD

This section introduces a novel multi-loop video coding scheme through effective inter-layer motion prediction employing novel multi-loop encoder architecture combined with multiloop probabilistic bitrate aware encoder-decoder model, as presented in Figure 1. Finally, shows the training optimization adopted using proposed video encoder model to attain improved PSNR and SSIM.

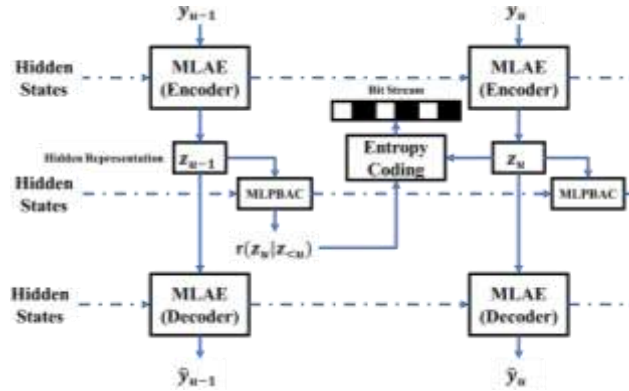


Figure 1. The architecture of proposed multi-loop inter-layer motion prediction aware video coder

3.1. System framework

The architecture of ILMPA-MLVC is presented in Figure 2. This work has drawn inspiration from conventional video codecs and uses motion-optimization, which has already been shown to be extremely useful in learning compression [34], and to decrease video frame redundancies. To pinpoint the temporal motion difference among the preceding compressed frame along with the earlier compressed frame, for instance, g_u and \hat{g}_{u-1} , we utilize an optical-flow network [34].

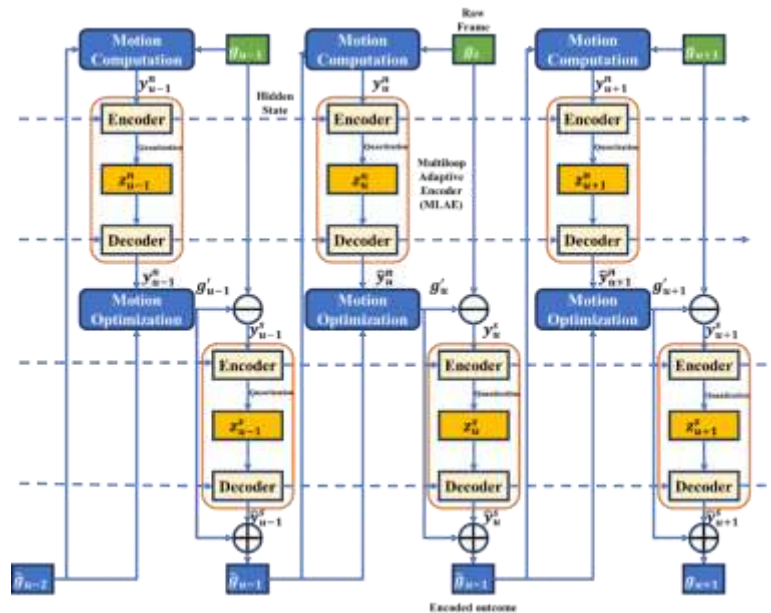


Figure 2. Framework of proposed multi-loop adaptive video coding

In this work the compressed and raw frames are represented as $\{\hat{g}_u\}_{u=1}^U$ and $\{g_u\}_{u=1}^U$ respectively. Further, the suggested MLAE compresses the predicted motion y_u^n , and the resulting compression motion \hat{y}_u^n is used as motion-optimization. The motion-optimization approach used in our work is similar to [34].

In this study, the residual (y_u^s) among the original frame g_u along with the motion- optimized frame g'_u was determined and compressed using an additional MLAЕ. Considering the compressed-residual as \hat{y}_u^s , the compressed frame $\hat{g}_u = g'_u + \hat{y}_u^s$ is capable of being reconstructed. The residual and motion compression, represented as z_u^s and z_u^n respectively, the proposed MLAЕ utilizes two layer-encoders (LEs) in every frame to generate latent-representations. Further, for compressing z_u^s and z_u^n into bit-streams, this work presents MLPBAC network for recursively predicting temporal condition probability-function of $\{z_u^s\}_{u=1}^U$ and $\{z_u^n\}_{u=1}^U$. The conditioned cross-entropy is anticipated to be less compared to the independent cross-entropy employed in standard learning techniques because there exists a temporal link between video frames. Therefore, bit-rate for entropy-coding is successfully decreased by employing conditional probability-function calculated by proposed MLPBAC network [12].

3.2. Multi-loop adaptive encoder architecture

As previously stated, two LEs are utilized for compressing y_u^n and y_u^s . To keep things simple, we will refer to both y_u^n and y_u^s as y_u in the following subsections since the structure of the two LEs is identical. In standard learning-based video compression techniques [12], the u^{th} frame is compressed by mapping the input y_u into a hidden feature-representation using an encoder F , which is parameterized with φ_F . Next, the continuous-valued \hat{z}_u is converted into a discrete-valued $z_u = \lfloor \hat{z}_u \rfloor$ by quantization. The \hat{z}_u can be represented using (1). Further, the output which has been compressed is built again using decoder by utilization of quantized hidden feature-representations as presented in (2). By providing the decoder and encoder with solely the present frame’s y_u and z_u as inputs, they are unable to make use of the temporal connection between succeeding frames.

$$\hat{z}_u = E(y_u; \varphi_F) \tag{1}$$

$$\hat{y}_u = E(z_u; \varphi_E) \tag{2}$$

On the other hand, the suggested MLAЕs incorporate recurring cells within the decoder and encoder. The structure for the MLAЕs network is depicted in Figure 3. In this work, we implemented an encoder for MLAЕ using four $2 \times$ down-sampling layers of convolution, following the approach described in [12]. We also used the enhanced GDN [35] as an activating function for the encoder. Within the MLAЕ architecture consisting of 4 layers of convolution, a ConvGRU [36] cell is incorporated to establish a multi-loop encoder framework. Consequently, any information coming from preceding frames is transmitted across the encoding network for the present frame using the hidden configurations of ConvGRU. Thus, the MLAЕs suggested in this study produce hidden feature-representations by considering both present and prior inputs.

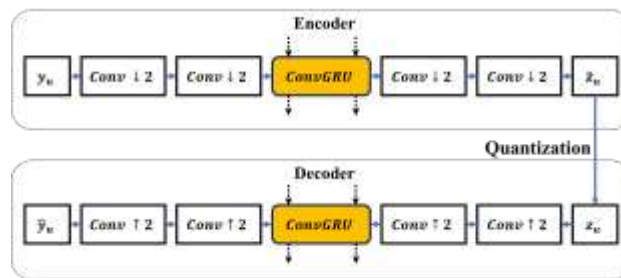


Figure 3. Multiloop adaptive encoder architecture

Furthermore, the decoder present in MLAЕ utilizes ConvGRU cell in between of four $2 \times$ up-sampling layers of convolutional using IGDN [35], and further reconstruction of \hat{y}_u takes place using previous and current latent-representations. In conclusion, the MLAЕ network is formulated using (3). From both the (3) and (4), it is possible to consider all prior frames as frames of reference for compressing the present one. This means that the present method can utilize data from an extensive variety of frames, which is a significant improvement compared to the small number of reference frames used in up-to-date baseline techniques [12].

$$z_u = \lfloor F(y_1, \dots, y_u; \varphi_F) \rfloor \tag{3}$$

$$\hat{y}_u = E(z_1, \dots, z_u; \varphi_E) \tag{4}$$

3.3. Multi-loop probabilistic bitrate aware compression model

In compressing the complete sequences of hidden feature-representations, represented as $\{z_u\}_{u=1}^U$, in this work a network called as MLPBAC is presented for entropy-coding is presented in Figure 4. Initially, this work utilizes $r(z_u)$ and $q(z_u)$ for representing estimated and true independent probability functions of z_u .

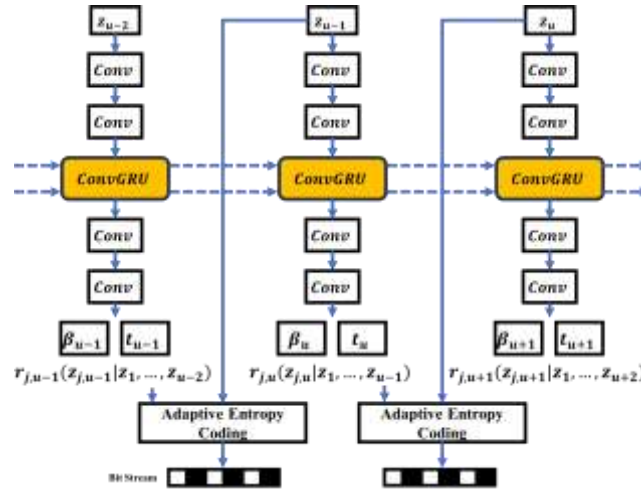


Figure 4. Architecture of multiloop probabilistic bitrate aware compression model

The anticipated bit-rate for z_u is represented as cross-entropy presented in (5). It should be taken into account that entropy-coding [12] has the capability of encoding z_u with a bit-rate which is equivalent to that of cross-entropy, with minimal additional data. It is evident from (5) that when z_u possesses a higher level of certainty, the bit-rate is likely to be reduced. Moreover, distribution of z_u in successive frames is associated because of the temporal link between video frames. Thus, given the knowledge of prior latent-representations z_1, \dots, z_{u-1} , it is anticipated that the present representation z_u will exhibit a higher level of certainty. In other words, we can define $r_u(z_u|z_1, \dots, z_{u-1})$ and $q_u(z_u|z_1, \dots, z_{u-1})$ as estimated and true temporal conditioning probabilistic functions of z_u , and it has to be noted that the condition cross-entropy could be less in comparison with independent cross-entropy presented in (5). Hence, for achieving the anticipated bit-rate, (6) is presented. The (6) represents the MLPBAC network which is used as recurrent model for the condition probabilistic functions $r_u(z_u|z_1, \dots, z_{u-1})$.

$$H(q, r) = \mathcal{E}_{z_u \sim q}[-\log_2 r(z_u)] \tag{5}$$

$$H(q_u, r_u) = \mathcal{E}_{z_u \sim q_u}[-\log_2 q_u(z_u|z_1, \dots, z_{u-1})] \tag{6}$$

To be more precise, adaptable entropy-coding [12] enables us to modify the probability functions associated with every component in z_u , resulting in the estimation of various conditional probability functions $r_{ju}(z_{ju}|z_1, \dots, z_{u-1})$, for various components z_{ju} . In this study, z_{ju} has been defined as the component located within the j^{th} 3D position in z_u . The condition probabilistic functions of z_u are mathematically expressed as (7) where O represents overall 3D positions present in z_u . In this work, the $r_{ju}(z_{ju}|z_1, \dots, z_{u-1})$ is modeled for every component as separated logistic-distribution. Given that the quantification functioning in LE quantifies all \tilde{y}_{ju} into a single value of z_{ju} , hence, we can determine the condition probabilistic function of the quantified z_{ju} through integration of a continuous logistic-distribution from $(z_{ju} + \tau)$ to $(z_{ju} - \tau)$.

$$r_u(z_u|z_1, \dots, z_{u-1}) = \prod_{j=1}^O r_{ju}(z_{ju}|z_1, \dots, z_{u-1}) \tag{7}$$

$$r_{ju}(z_{ju}|z_1, \dots, z_{u-1}) = \int_{z_{ju}-\tau}^{z_{ju}+\tau} \text{Logistic}(z; \beta_{ju}, t_{ju}) dz \tag{8}$$

Where the logistic-distribution is evaluated using (9). Also, the sigmoid-distribution for (9) can be denoted using (10). From (8)-(10), the overall condition probabilistic function can be represented using (11).

Further, from (11), the conditional probabilistic function for every location is defined using variables t_{ju} (12) and β_{ju} (13) which keep on vary with respect to the various location present in z_u . Hence, the MLPBAC network persistently estimates the logistic-distribution using (10).

$$\text{Logistic}(z; \beta, t) = \frac{\exp(-(z-\beta)/t)}{t(1+\exp(-(z-\beta)/t))^2} \tag{9}$$

$$\text{Logistic}(z; \beta, t)dz = \text{Sigmoid}(z; \beta, t) + C \tag{10}$$

$$r_{ju}(z_{ju}|z_1, \dots, z_{u-1}) = \text{Sigmoid}(z_{ju} + 0.5; \beta_{ju}, t_{ju}) - \text{Sigmoid}(z_{ju} - 0.5; \beta_{ju}, t_{ju}) \tag{11}$$

$$t_u = \{t_{ju}\}_{j=1}^o \tag{12}$$

$$\beta_u = \{\beta_{ju}\}_{j=1}^o \tag{13}$$

Figure 4 presents the framework that makes up the MLPBAC network, showcasing the inclusion of a network of recurrence Q with layers of convolution along with a ConvGRU cell positioned in the center. Further, the generation of β_u and t_u is dependent upon all prior hidden feature-representations, which is caused by the recurrent architecture. This is represented using (14) where, φ_Q denotes the training variable in MLPBAC. As P considers prior hidden feature -representations z_1, \dots, z_{u-1} as an input, β_u along with t_u trains the model using the probability function of every z_u on the basis of z_1, \dots, z_{u-1} using (11). Further, condition probabilistic function $r_{ju}(z_{ju}|z_1, \dots, z_{u-1})$ are subjected to the adaptable entropy-coding [12] for encoding z_u into a bit-stream.

$$\beta_u, t_u = Q(z_1, \dots, z_{u-1}; \varphi_Q) \tag{14}$$

3.4. Training optimization

In this study, the PSNR and MS-SSIM index are utilized for evaluation of compression quality through optimization of distortion D . The bit-rates for z_1^n and z_1^s are denoted as $S_1(z_1^n)$ and $S_1(z_1^s)$ respectively. The P-frames provided in the dataset have been compressed using the MLPBAC network suggested in our research. For $u \geq 2$, the researcher is able to determine the exact bit-rate. The relationship between $r_u(z_{ju}|z_1, \dots, z_{u-1})$ is represented by the suggested MLPBAC approach, as described in (7) to (14). It should be noted that, under the assumption that distribution for the set used for training is the same as the accurate distribution. Further, the actual bit-rate $S_{MLPBAC}(z_u)$ is anticipated to correspond with the conditioned cross-entropy in (6). In our research strategy, we utilize a pair MLPBAC networks to process the hidden feature-representations of residual and motion. The bit-rates for both of these networks are denoted as $S_{MLPBAC}(z_u^n)$ and $S_{MLPBAC}(z_u^s)$ respectively. The ILMPA-MLVC approach used in our research was developed on the traditional UVG [37] dataset. In this dataset, the initial frames were compressed as the I-frame, while all the remainder the remaining frames were considered as P-frames. Therefore, the training process was improved by progressively incorporating P-frame g_1 . Initially, the inter-layer estimation of motion network underwent training using the loss function as presented in (16) where, y_1^n represents output for motion estimated network as presented in Figure 2, and X represents warping process. Moreover, when the \mathbb{L}_{ILME} was converged, the MLAЕ network was incorporated for motion compression and for motion-optimization network in training process, utilizing the loss-function presented in (17).

$$S_{MLPBAC}(z_u) = -\log_2(r_u(z_u|z_1, \dots, z_{u-1})) = \sum_{j=1}^o \log_2(r_{ju}(z_{ju}|z_1, \dots, z_{u-1})) \tag{15}$$

$$\mathbb{L}_{ILME} = E(g_1, X(g_0, y_1^n)) \tag{16}$$

$$\mathbb{L}_{ILMO} = \delta \cdot E(g_1, g_1') + S_1(z_1^n) \tag{17}$$

Further, after convergence of \mathbb{L}_{ILMO} , the overall network was merged and trained on g_1 by using the loss represented in (18). Next, we use a loss function for training recurrent model end-to-end using the sequential-training frames as presented in (19). Throughout the training process, the relaxation of quantization was implemented using the method described in [12]. This is done in order to prevent the

occurrence of zero gradients. In our research, we followed the methodology outlined in [12] to determine the values of δ . The Adam optimizer was employed for training. For each loss-functions presented in (16)-(19), the initial rate at which they are learned was set to 10^{-4} . When training the entire model using the final loss presented in (19), the learning rate was decreased by a factor of 10 following convergence till it reaches 10^{-6} . The suggested ILMPA-MLVC approach demonstrates a favorable equilibrium among accomplishing higher PSNR and improved MS-SSIM, as illustrated in the below section.

$$\mathbb{L}_1 = \delta \cdot E(g_1, \hat{g}_1) + S_1(z_1^n) + S_1(z_1^s) \quad (18)$$

$$\mathbb{L} = \delta \cdot \sum_{u=1}^{MAX} E(g_1, \hat{g}_1) + S_1(z_1^n) + S_1(z_1^s) + \sum_{u=1}^{MAX} (S_{MLPBAC}(z_u^n) + S_{MLPBAC}(z_u^s)) \quad (19)$$

4. RESULTS AND DISCUSSION

This section studies the performance of proposed ILMPA-MLVC over various existing video coding methodologies like hierarchical random-access coding for deep neural video compression (HRAC) deep neural video compression (DNVC) [4], HDVC [25]. The video compression performance is studied using UVG [37] dataset; the dataset is generated with video resolution of 3840×2160 , a total of 7 videos, out of which like [16] the first 300 frames are used in each video for experiment analysis. Different case studies and methodologies have been considered and performance is measured in terms of PSNR and MS-SSIM.

4.1. Case-study 1

In the first case the performance of ILMPA-MLVC is compared with HRAC-DNVC [16], DVC-Pro [19], DVC [10], and SSF [38] using UVG dataset. Figure 5 shows the PSNR outcome considering varied BPP i.e., rate distortion curve; The curve with higher value indicates superior performance. The result shows SSF experience poor PSNR and HRAC-DNVC attains much better PSNR in comparison with DVC and DVCPro. On the other hand, the ILMPA-MLVC attains much improved PSNR in comparison with HRAC-DNVC. Thus, are tolerable to varying distortion levels in comparison with all current video compression methods.

4.2. Case-study 2

This section studies the performance of flow-based methods like CAIN [1], AdaCoF [2], ABME [3], RIFE [4], IFRNet [5], TDPNet [20], and proposed ILMPA-MLVC methods. Figure 6 shows the PSNR outcome considering varied Bpp i.e., rate distortion curve; The curve with higher value indicates superior performance. The result shows the proposed ILMPA-MLVC methods attain improved PSNR performance considering varied BPP in comparison with CAIN, AdaCoF, ABME, RIFE, IFRNet [5], and TDPNet [10] methods. In Table 1, comparative study between proposed ILMPA-MLVC and TDPNet methods in terms of PSNR considering two UVG video sequence such as Honeybee and Beuty. In both cases the ILMPA-MLVC attains much higher PSNR in comparison to TDPNet. Thus, shows the efficiency of the proposed video compression model.

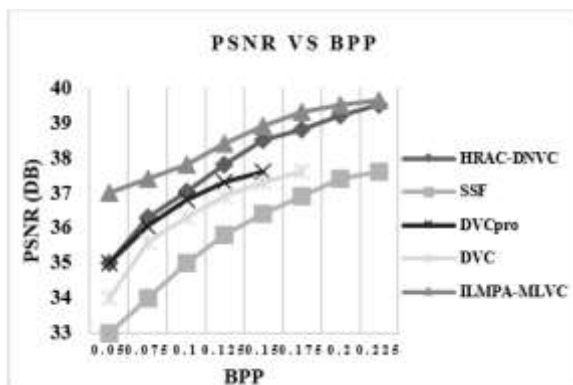


Figure 5. PSNR performance attained by different models with varying BPP levels

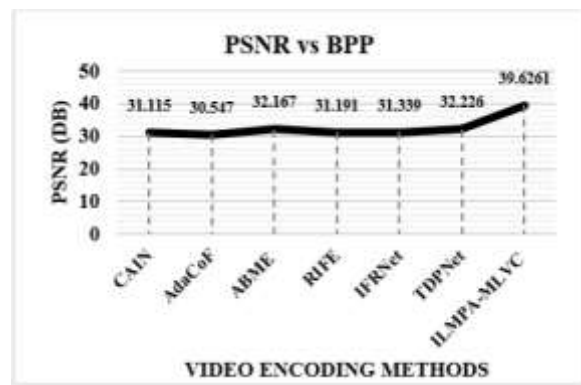


Figure 6. PSNR performance attained by different models with varying BPP levels

Table 1. PSNR performance summary of literature survey

Models	PSNR [TDPNet]	PSNR [ILMPA-MLVC]
HoneyBee	38.672	40.3297
Beauty	30.294	35.2415

4.3. Case-study 3

In this case the performance of ILMPA-MLVC is compared with deep learning-based entropy coding methods like DVC [17], and HDVC [34] using UVG dataset. Figure 7 shows the PSNR outcome considering varied BPP i.e., rate distortion curve; the curve with higher value indicates superior performance. The result shows DVC experience poor PSNR and HDVC attains much better PSNR in comparison with DVC. On the other hand, the ILMPA-MLVC attains much improved PSNR in comparison with DVC and HDVC. Figure 8 shows the MS-SSIM outcome considering varied BPP i.e., rate distortion curve; The curve with higher value indicates superior performance. The result shows DVC experience poor MS-SSIM and HDVC attains much better MS-SSIM in comparison with DVC. On the other hand, the ILMPA-MLVC attains much improved MS-SSIM in comparison with DVC and HDVC. Thus, are tolerable to varying distortion levels in comparison with all current video compression methods.

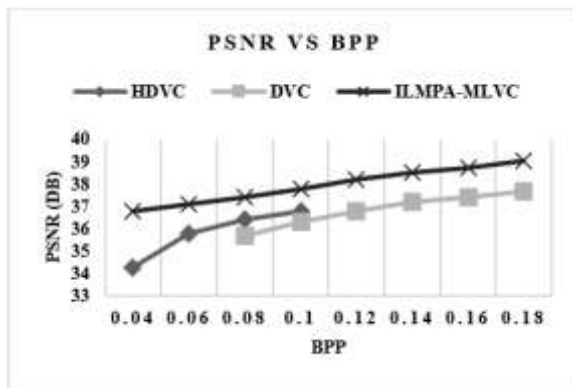


Figure 7. PSNR performance attained by different models with varying BPP levels

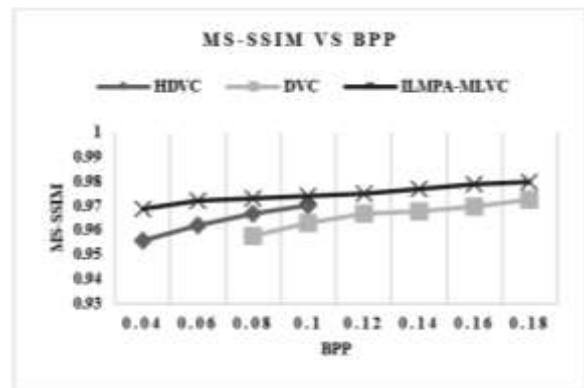


Figure 8. PSNR performance attained by different models with varying BPP levels

5. CONCLUSION




In this paper a novel video compression method is introduced by leveraging the benefit of convolution neural network and recurrent neural network. A multi-loop adaptive encoder and multi-loop probabilistic bitrate aware compression model is devised to design effective adaptive entropy-based coding techniques. The study shows the current method uses lesser number of frames to perform encoding operation; on contrary, the proposed video compression method uses significant amount of reference frames with minimal overhead. The adoption of motion detection and optimization aided in improving models PSNR and MS-SSIM. The proposed ILMPA-MLVC is compared with various existing methodologies which includes flow-based, learning-based, and deep learning-based coding scheme; in all the cases the proposed methods attains much improved PSNR and MS-SSIM; however, in future work more optimization will be done during training process considering improved motion detection and distortion optimization considering more diverse dataset and performance metrics.

REFERENCES





- [1] X. Jin, H. Sun, and Y. Zhang, "Research on VVC intra-frame bit allocation scheme based on significance detection," *Applied Sciences (Switzerland)*, vol. 14, no. 1, p. 471, Jan. 2024, doi: 10.3390/app14010471.
- [2] C. Bonnineau, W. Hamidouche, J. F. Travers, N. Sidaty, and O. Deforges, "Multitask learning for VVC quality enhancement and super-resolution," in *2021 Picture Coding Symposium, PCS 2021 - Proceedings*, Jun. 2021, pp. 1–5, doi: 10.1109/PCS50896.2021.9477492.
- [3] X. Sheng, L. Li, D. Liu, and H. Li, "Spatial Decomposition and temporal fusion based inter prediction for learned video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6460–6473, Jul. 2024, doi: 10.1109/TCSVT.2024.3360248.
- [4] W. Xiao, Y. Hao, J. Liang, L. Hu, S. A. Alqahtani, and M. Chen, "Adaptive compression offloading and resource allocation for edge vision computing," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2024, doi: 10.1109/TCCN.2024.3400820.

- [5] Y. Li, B. Chen, B. Chen, M. Wang, S. Wang, and W. Lin, "Perceptual quality assessment of face video compression: a benchmark and an effective method," *IEEE Transactions on Multimedia*, vol. 26, pp. 8596–8608, 2024, doi: 10.1109/TMM.2024.3380260.
- [6] Y. Tian, G. Lu, Y. Yan, G. Zhai, L. Chen, and Z. Gao, "A coding framework and benchmark towards low-bitrate video understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5852–5872, Aug. 2024, doi: 10.1109/TPAMI.2024.3367879.
- [7] P. Du, Y. Liu, and N. Ling, "CGVC-T: contextual generative video compression with transformers," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 14, no. 2, pp. 209–223, Jun. 2024, doi: 10.1109/JETCAS.2024.3387301.
- [8] Z. Chen, T. He, X. Jin, and F. Wu, "Learning for video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 566–576, Feb. 2020, doi: 10.1109/TCSVT.2019.2892608.
- [9] H. Liu *et al.*, "Neural video coding using multiscale motion compensation and spatiotemporal context model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3182–3196, Aug. 2021, doi: 10.1109/TCSVT.2020.3035680.
- [10] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: an end-to-end deep video compression framework," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 10998–11007, doi: 10.1109/CVPR.2019.01126.
- [11] D. Jia, K. Wang, S. Luo, T. Liu, and Y. Liu, "BRAFT: recurrent all-pairs field transforms for optical flow based on correlation blocks," *IEEE Signal Processing Letters*, vol. 28, pp. 1575–1579, 2021, doi: 10.1109/LSP.2021.3099076.
- [12] Y. Hu, C. Jung, Q. Qin, J. Han, Y. Liu, and M. Li, "HDVC: deep video compression with hyperprior-based entropy coding," *IEEE Access*, vol. 12, pp. 17541–17551, 2024, doi: 10.1109/ACCESS.2024.3350643.
- [13] T. Das, K. Choi, and J. Choi, "High quality video frames from VVC: a deep neural network approach," *IEEE Access*, vol. 11, pp. 54254–54264, 2023, doi: 10.1109/ACCESS.2023.3281975.
- [14] "USC media communications Lab – MCL-JCV Dataset." <https://mcl.usc.edu/mcl-jcv-dataset/>.
- [15] F. Zhang, D. Ma, C. Feng, and D. R. Bull, "Video compression with CNN-based postprocessing," *IEEE Multimedia*, vol. 28, no. 4, pp. 74–83, Oct. 2021, doi: 10.1109/MMUL.2021.3052437.
- [16] N. V. Thang and L. V. Bang, "Hierarchical random access coding for deep neural video compression," *IEEE Access*, vol. 11, pp. 57494–57502, 2023, doi: 10.1109/ACCESS.2023.3283277.
- [17] "Ultra video group (UVG) dataset," *ultravideo.fi*. <http://ultravideo.fi/#main> (accessed Apr. 29, 2024).
- [18] J. J. Quinlan, A. H. Zahran, and C. J. Sreenan, "Datasets for AVC (H.264) and HEVC (H.265) evaluation of dynamic adaptive streaming over HTTP (DASH)," in *Proceedings of the 7th International Conference on Multimedia Systems, MMSys 2016*, May 2016, pp. 386–391, doi: 10.1145/2910017.2910625.
- [19] G. Lu, X. Zhang, W. Ouyang, L. Chen, Z. Gao, and D. Xu, "An end-to-end learning framework for video compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3292–3308, 2021, doi: 10.1109/TPAMI.2020.2988453.
- [20] E. Agustsson, D. Minnen, N. Johnston, J. Ballé, S. J. Hwang, and G. Toderici, "Scale-space flow for end-to-end optimized video compression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp. 8500–8509, doi: 10.1109/CVPR42600.2020.00853.
- [21] K. Yoon, J. Huh, Y. H. Kim, S. Kim, and J. Jeong, "Textural detail preservation network for video frame interpolation," *IEEE Access*, vol. 11, pp. 71994–72006, 2023, doi: 10.1109/ACCESS.2023.3294964.
- [22] M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, "Channel attention is all you need for video frame interpolation," *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10663–10671, Apr. 2020, doi: 10.1609/aaai.v34i07.6693.
- [23] W. Bao, W. S. Lai, X. Zhang, Z. Gao, and M. H. Yang, "MEMC-Net: motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 933–948, Mar. 2021, doi: 10.1109/TPAMI.2019.2941941.
- [24] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019, doi: 10.1007/s11263-018-01144-2.
- [25] K. ding, K. ma, S. wang, and E. P. simoncelli, "Image quality assessment: unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022, doi: 10.1109/TPAMI.2020.3045810.
- [26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 586–595, doi: 10.1109/CVPR.2018.00068.
- [27] B. He *et al.*, "Towards scalable neural representation for diverse videos," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 6132–6142, doi: 10.1109/cvpr52729.2023.00594.
- [28] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: a dataset of 101 human actions classes from videos in the wild," *arXiv (Cornell University)*, 2012, [Online]. Available: <http://arxiv.org/abs/1212.0402>.
- [29] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2016, vol. 2016-December, pp. 724–732, doi: 10.1109/CVPR.2016.85.
- [30] Y. J. Choi, Y. W. Lee, J. Kim, S. Y. Jeong, J. S. Choi, and B. G. Kim, "Attention-based Bi-prediction network for versatile video coding (VVC) over 5G Network," *Sensors*, vol. 23, no. 5, p. 2631, Feb. 2023, doi: 10.3390/s23052631.
- [31] D. Ma, F. Zhang, and D. R. Bull, "BVI-DVC: a training database for deep video compression," *IEEE Transactions on Multimedia*, vol. 24, pp. 3847–3858, 2022, doi: 10.1109/TMM.2021.3108943.
- [32] H. M. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "HiNeRV: video compression with hierarchical encoding-based neural representation," *Advances in Neural Information Processing Systems*, vol. 36, 2023, doi: 10.5555/3666122.3669299.
- [33] Z. Wang, G. Quan, and G. He, "Edge-oriented compressed video super-resolution," *Sensors*, vol. 24, no. 1, p. 170, Dec. 2024, doi: 10.3390/s24010170.
- [34] Frank Bossen, "Common test conditions and software reference configurations," *JCT-VC Document, JCTVC-G1200*, vol. 12, no. 1, 2010.
- [35] A. Ravi and F. Karray, "Exploring convolutional recurrent architectures for anomaly detection in videos: a comparative study," *Discover Artificial Intelligence*, vol. 1, no. 1, p. 6, Dec. 2021, doi: 10.1007/s44163-021-00004-2.
- [36] H. Shao, B. Liu, Z. Li, C. Yan, Y. Sun, and T. Wang, "A high-throughput processor for GDN-based deep learning image compression," *Electronics (Switzerland)*, vol. 12, no. 10, p. 2289, May 2023, doi: 10.3390/electronics12102289.
- [37] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120fps 4K sequences for video codec analysis and development," in *MMSys 2020 - Proceedings of the 2020 Multimedia Systems Conference*, May 2020, pp. 297–302, doi: 10.1145/3339825.3394937.
- [38] R. Pourreza, H. Le, A. Said, G. Sautiere, and A. Wiggers, "Boosting neural video codecs by exploiting hierarchical redundancy," in *Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023*, Jan. 2023, pp. 5344–5353, doi: 10.1109/WACV56688.2023.00532.

BIOGRAPHIES OF AUTHORS

Sandeep Gowdra Siddaramappa     working as Assistant Professor in the Department of Computer Science and Engineering, G M Institute of Technology, Davanagere, Karnataka, India. Completed B.E. (IS&E) in Adichunchanagiri Institute of Technology, Chikmagalore in 2004 Under V.T.U. M. Tech in UBBDT College of Engineering, Davanagere in 2010 Kuvempu University. Pursuing Ph.D. in Dept. of ISE, RV College of Engineering, Bangalore, and Karnataka. He has more than 14 years of experience in Teaching and 2 years in Industry. Published more than 5 research international journals, 3 papers in international conferences, and more than 20 papers in national journals/conferences. Received 6 awards for best paper and best projects. Attended and conducted more than 20 Workshops/FDP/Activities. Lifetime member for LMISTE. He can be contacted at email: yashasgowdaniharika@gmail.com.



Gowdra Shivanandappa Mamatha     Professor and Associate Dean (PG Studies) in RV College of Engineering, M.Tech from B.I.E.T, Davanagere in 2005, VTU, Belagavi, First Class with Distinction. Ph.D. from Avinashilingam University, Coimbatore, in 2014. Teaching experience: 19 years and R&D: 09 years. Research interest are: ad-hoc networks, routing and security, IoT, cloud computing, artificial intelligence. Published more than 18 papers in international journal, 10 papers in international conferences. Conducted and attended more than 30 events in national and international levels. Awards: Life Member of “The Indian Society for Technical Education” (ISTE) organization for promoting the quality and standards in technical education since 2006. Membership No.: LM-48853. Achievements: 3 best paper awards and 1 reviewer award. R&D projects are executed for around Rs. 10 lakhs, consultancy projects are undertaken for around Rs.15 lakhs. She can be contacted at email: mamathags@rvce.edu.in.