

Flow incorporated neural network based lightweight video compression architecture

Sangeeta, Preeti Gulia, Nasib Singh Gill

Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India

Article Info

Article history:

Received Mar 10, 2021

Revised Jan 30, 2022

Accepted Mar 14, 2022

Keywords:

Autoencoder

Deep learning

Peak signal-to-noise ratio

Structural similarity index

Video compression

ABSTRACT

The sudden surge in the video transmission over internet motivated the exploration of more promising and potent video compression architectures. Though the frame prediction based hand designed techniques are performing well and widely used but the recent deep learning based researches in this domain provided further directions of pure deep learning based next generation codecs. As the bandwidth over the internet is varying, adaptive bit rate representation is more suitable for video quality adjustment in tune with bandwidth variation. The proposed architecture comprises of end to end trainable video compression network consisting of majorly three modules namely-motion extension network, flow autoencoder and frame autoencoder. Frame autoencoder generates the individual compressed frames, flow autoencoder is used for optical flow based motion compensation chore and next frame is predicted by the motion extension network. The network is designed and evaluated in incremental manner. The analysis of the outcomes demonstrates the promising performance of the network quantitatively and qualitatively. Moreover, the results reveal that inclusion of optical flow based motion compensation network to the MotionNet architecture has enhanced the performance.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Sangeeta

Department of Computer Science and Applications, Maharshi Dayanand University

Rohtak, India

Email: sangirao5228@gmail.com

1. INTRODUCTION

A superfluous rise in video content over the internet has been observed over the past few years. As the bandwidth is limited and cannot be scaled rapidly, the growing voluminous video content has put a significant stress over the bandwidth. The sudden increase in the demand of streaming services has majorly contributed to the growing video content. The introduction of enhanced video and image standards has also contributed to its complexity. To meet the high video quality requirements of the people, the service providers have to look for more potent compression techniques. The widely used existing video compression schemes are improved and enhanced using diverse AI based approaches. As after some significant improvements to traditional compression schemes, current enhanced techniques are performing at their par, only marginal improvements can be obtained for these schemes with further enhancements. Such outcomes encouraged the researchers to explore next generation of deep learning based video compression schemes. The promising results of image compression achieved using deep learning based techniques have encouraged the experimentation of such techniques in video compression domain also. Afterwards, a number of models comprising of convolutional neural networks (CNN), recurrent neural networks and autoencoders were proposed and experimented. The proposed model also comprises of CNN and convolutional gated recurrent

unit (ConvGRU) layers for modelling the deep learning based end to end video compression architecture and are an extension of the MotionNet architecture as presented in [1].

A number of video related chores like object and action detection and classification are directly affected by the compression quality and efficiency [2]. Moreover, compression architecture has a direct impact on the compression rate [3]. The widely used video compression schemes primarily comprise of hand designed techniques. In the traditional schemes, the redundancies among the frames are addressed using discrete cosine transform and block based motion estimation techniques [4], [5]. During the last few years, the emergence of deep learning as a powerful tool in image and video compression has also led to further explorations to improve the efficiency and performance. A lot of image compression strategies have been proposed and experimented. These techniques resulted in significant and competent results [6]-[15] when compared with widely used existing methods like joint photographic experts group (JPEG) or better portable graphics (BPG) [16], [17]. The deep neural network (DNN) has the capability to address large non-linear transform and large scale end to end training as they comprises of a number of multiple hidden layers. But traditional image codecs does not comprise of such scheme.

Several deep learning based proposed image compression schemes were enhanced and extended for their applicability in video compression. These schemes cannot be simply applied or extended to videos but encounters some challenges pertaining to incorporation, generation, representation and compression of motion information. The adjacent video frames comprising of temporal redundancies. The schemes related to removal of such redundancies must be incorporated in compression architecture to achieve efficient compression. The technique chosen for motion information processing has a direct impact on the compression architecture. Optical flow is found to be an efficient motion representation scheme. The optical flow is based on the production of flow fields. Hence, the efficiency of learning based optical flow techniques can be improved by improving the accuracy of produced flow fields. Sometimes the accurate flow fields may not resulted in optimal performance [18]. Optical flow also makes the data voluminous and when this data is compressed by direct application of existing compression scheme, it requires more number of bits for motion representation [19], [20]. Some of the major recent developments in the field of video compression comprises of end to end designed and trainable complex compression networks employing frame prediction schemes with different optimization modules. One such hybrid network comprising of Spatio-temporal coherence with predictive coding named pixel-motion CNN (PMCNN) has been proposed by Zhibo *et al.* Though this network possesses high complexity but resulted in promising outcomes. Another efficient optical flow based end to end trained network for prediction and regeneration of frames has been proposed and experimented by Lu *et al.* Video compression architecture can also be designed using image interpolation. It is observed that pure outperforming deep learning based approaches have rigorous computation requirements and high complexity. These challenges motivated the researchers to explore and look for more lightweight and competing deep learning based networks for video compression.

The proposed model represents a deep learning based end to end video compression architecture which is an extension of the MotionNet architecture as presented in [1]. The network comprises of three sub networks namely flow autoencoder, frame autoencoder and motion extension network. One of the layers of frame autoencoder is ConvGRU based, a convolutional recurrent neural network, comprises productive edges of both recurrent neural network (RNN) and CNN. The whole network is trained, learned and optimized in a joint manner. This model presents one to one correspondence with traditional architecture. The random number of emission steps has been used for the training. The visual quality of regenerated frames have been is measured by structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) whereas flow end point error (EPE) and time per frame (TPF) are used for efficiency measurements. The results show marginal improvement in frame visual quality in tradeoff with reconstruction time. Some of the works related to deep learning based compression techniques has been discussed below in this section. The architectural details of the proposed network have been detailed out in section 2. The experimental results and its comparative analysis are given in section 3. Section 4 presents the conclusion of the whole work.

2. RELATED WORK

A video comprises of sequence of image frames. Several image compression methods are extended for their applicability to videos. The widely used image compression standards like JPEG are manually designed employing quantization, discrete cosine transform (DCT) and entropy encoding [16]. Such traditional image compression standards are evolved in an integrated manner and hence instead of end to end optimization, only individual modules can be optimized to attain overall optimization. This limits the optimal performance of the traditional compression schemes.

The researchers proposed and experimented several different deep learning based compression techniques for images. Some of the initial image compression architectures were designed using recurrent

neural networks [8], [11], [21]. These networks were optimized by minimizing the mean square error between the real and regenerated frames while the number of bits used being ignored. A number of convolutional neural networks based image compression networks were also proposed and experimented [7], [10], [12] and their performance was analyzed in comparison with the existing image codecs. The initial techniques employed non-adaptive arithmetic coding but the advanced compression networks, proposed later, comprises of rate distortion optimization schemes and adaptive coding to enhance the compression efficiency and the performance. An efficient discrete cosine based image compression technique for JPG, PNG and BMP formats have also been proposed and experimented [21]. Several prominent works were also done in the field of color image compression [22], [23]. Further researches resulted in invention of some new image compression techniques like adversarial training, multi scale image decomposition and importance maps. A significant improvement in image compression has been observed and their promising outcomes motivated the application of these techniques in the design of deep learning based video compression architecture.

H.264 or high efficiency video coding (HEVC) are the widely used traditional video codecs. Though being hand designed, they have good efficiency and widely used by various online platforms. These schemes basically employ predictive coding architecture and evolved by integration of multiple modules. Such network cannot be end to end optimized but overall optimization is achieved by optimization of individual modules. Several different techniques based on deep neural networks have been proposed as enhancements to existing traditional for the improvement in their performance like mode decision, entropy coding, intra prediction and residual coding. These enhancements are also block based and improve the corresponding block only. The models based on block based learning strategies are found to have block artifact issues. In addition, the models employing traditional block estimation schemes based motion information transfer do not perform optimally. Some autoencoder based extensions to the existing H.264 codec have also been proposed but they were limited to particular domains only. The processing of motion data in these works are not deep learning based. Frame interpolation based RNN architecture for video compression has also been proposed and experimented. But this architecture is also not end to end optimized and does not employ deep learning strategy for motion compensation. A notable deep learning based end to end video compression framework, deep video compression (DVC), has been proposed in [24]. An effective adversarial video compression based on soft edge detection has also been proposed and experimented [25]. It has been observed that the notable outperforming pure deep learning based schemes have high complexity and computation issues. Such challenges motivated the further exploration of pure deep learning based end to end trainable and optimizable lightweight video compression networks. This paper also presents a lightweight compression network for video compression comprising of three sub-networks.

3. METHOD

The proposed video compression architecture is an extension of the MotionNet architecture as presented in [1]. Optical flow is found to be an efficient representation for motion estimation. Hence, a flow autoencoder has been added to the MotionNet architecture for better motion compensation. The whole network presents a deep tool employed end to end trained and learned video compression architecture. The motion extension network along with flow and frame auto encoders collectively forms the compression network. The frame autoencoder comprises of encoder and decoder units for individual frame compression. For the motion compensation, flow autoencoder has been employed. This sub network predicts the motion information, compress it and finally transmit it to the motion extension network for next frame prediction. The motion extension network makes the next frame prediction based on current frame, previous predicted frame and flow information. The different network components of the architecture have been discussed below in detail.

3.1. Frame autoencoder

The frame autoencoder comprises of encoder and decoder networks for encoding the frame and then decoding the original frame form compressed format. The encoder network comprises of five layers comprising of four convolutional layers and one ConvGRU layer. The input frame goes through these varying five layers and compressed in binary format. This binary format has been quantized before decompression. The decoder has been designed in same fashion as of the encoder. Hence, the frame is reconstructed using four convolutional and one ConvGRU based layers of the decoder network. Both the encoder and decoder networks are trained together. The design of the frame autoencoder has been given in Figure 1.

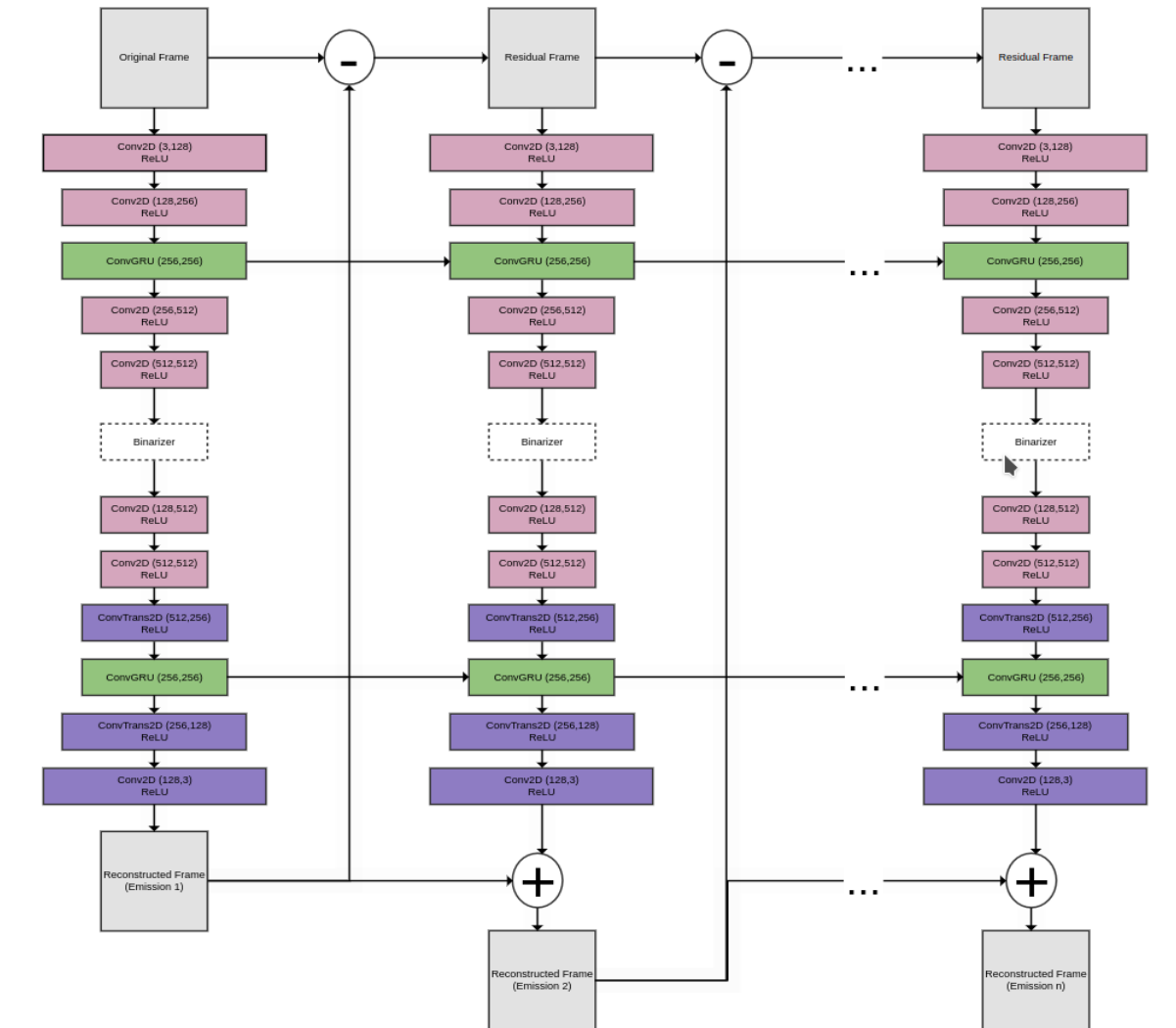


Figure 1. Frame autoencoder

3.2. Flow autoencoder

A video comprises of frames with motion information among them. The motion information is required for the next frame prediction. Optical flow is one of the efficient ways of motion representation. Farneback based flow estimation has been used. Flow autoencoder is used to efficiently compress and represent the flow information. The flow autoencoder comprises of five convolutional layers both in encoder and decoder networks. Generalized divisive normalization has been between each two convolutional layers to expedite non-linearity. The encoder encodes the flow information into binary format which then reconstructed using the similarly designed decoder network. The details of flow autoencoder have been given in Figure 2.

3.3. Motion extension network

The motion extension network is an integral part of video compression architecture. The motion extension network reconstructs the frame from the compressed format. The network has been designed the in incremental manner. Firstly, motion extension network was used along with frame autoencoder only, named as MotionNet (R), as described in [1] where R represents that the network is trained with randomized emission step training strategy. The number of emission steps has been chosen from one to ten. Later, in this paper, the same network has been extended to incorporate the flow autoencoder to expedite motion information, named as extended-MotionNet (R). In extended network, the frame is reconstructed using current compressed frame, previous reconstructed frame and flow information by flow autoencoder. The details of the extended architecture have been given in Figure 3.

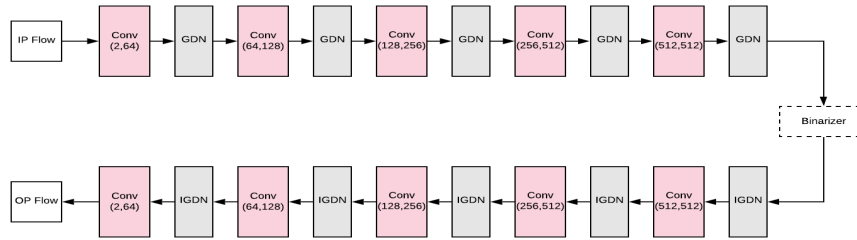


Figure 2. Flow autoencoder

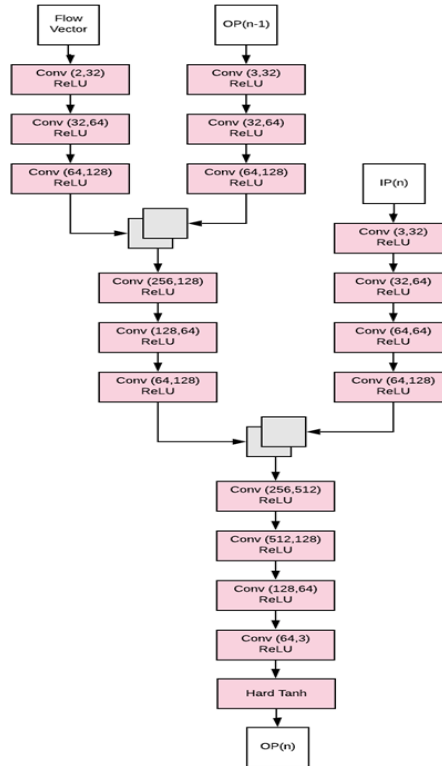


Figure 3. Extended motion extension network

3.4. Proposed video compression network

The video compression network comprises of all three sub networks. The entire network has been end to end trained. To measure the structural distortion among the reconstructed and the original frame, the following loss function has been used.

$$F(x_t, x't) = \lambda_1 \text{SSIM}(x_t, x't) + \lambda_2 \text{MSE}(x_t, x't)$$

Here, the structural similarity index metric loss has been measured by SSIM ($x_t, x't$) and the mean square error between the input and the output has been represented by MSE ($x_t, x't$). λ_1 and λ_2 are the multipliers. A number of emission steps, ranging from one to ten, have been used to refine the output. Each additional emission step improves the reconstruction quality, subsequently leading to declined compression efficiency. The randomized emission steps have been used to train the network.

3.5. Video decoding

The binary coded output of frame encoder and flow encoder are required for video reconstruction. The number of emission step will control the size of frame encoder data and so the bit-rate of signal. The Intermediate frame is calculated by passing binary code frame data through frame decoder. Flow information is also decoded via flow decoder network. Finally motion extension network takes previous image, merges

this with decoded flow information to create intermediate representation of current image. It merges this intermediate representation on already decoded intermediate frame to result in high quality current frame.

$$I_{\text{decoded}} = f_{\text{decoder}}(I_{\text{encoded}}, F_{\text{encoded}}, I_{\text{decoded}_{\text{prev}}})$$

where I_{encoded} and F_{encoded} are binaried encoding of current frame and flow vectors, $I_{\text{decoded}_{\text{prev}}}$ is previously decoded frame and f_{decoder} is representation of decoder neural network.

3.6. Dataset

The randomized training strategy has been used to train the proposed network. The number of emission steps varies from one to ten. MotionNet (R) represents the results obtained without flow autoencoder and extended-MotionNet (R) represents the outcomes after incorporating the flow values.

Dataset	YouTube UGC
Video quality	A mix of 360p, 480p, 720p and rescaled to 64x64 pixels
Total video samples	826
Training set	571 examples, each 20s long
Training	With a batch size of 10, randomly selected from the video frames
Test data set	96 examples and selected from the beginning.

4. RESULTS & DISCUSSION

4.1. Experimental results and analysis

The performance of the network has been evaluated in terms of both qualitative and quantitative parameters. Qualitative performance relates to the visual quality of the frames generated by the network. SSIM and PSNR are the widely used parameters for visual perception measurement. SSIM, structural similarity index measure, is used to measure the structural similarity among the original and reconstructed frames where PSNR, peak signal to noise ratio is measured in terms of mean square error. SSIM is better parameter than PSNR for visual perception. The SSIM and PSNR values with addition of each emission step have been given in Figures 4 and 5 respectively. The inclusion of flow autoencoder has improved the SSIM by 0.009 means indicates a small improvement in frame quality. Their comparative graphical representation has been given in Figures 4 and 5 respectively representing the pattern of change in SSIM and PSNR values during the varying emission steps. The qualitative performance or efficiency of the network has been measured using flow end point error and time per frame parameters. Keeping in view the motion perspective, Flow EPE represents the error in the reconstructed frame and the time taken by the network to reconstruct an individual frame has been given by time per frame parameter. The individual values of both parameters have been given in Figures 6 and 7 respectively. The graphical representation of the same to show the variation with each emission step wise has been presented in Figures 6 and 7 respectively. The results show that addition of flow autoencoder has marginally increased the reconstruction time.

The performance of the video compression architecture over varying bandwidth can better be measured with the average values of parameters used and the same has been given in Table 1. The inclusion of flow autoencoder has significantly increased the SSIM value from 0.8955 to 0.9036, which presents improvement in visual quality of frames. A small improvement in end point error and reconstruction time has also been observed.

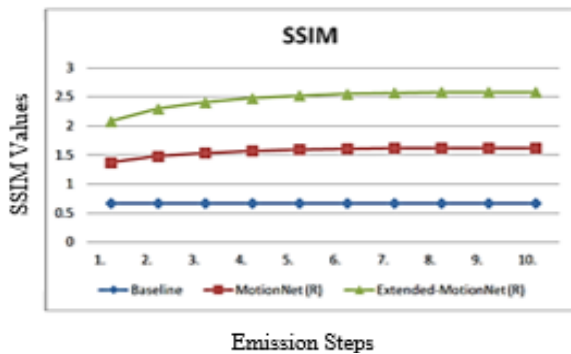


Figure 4. Variation of SSIM per emission

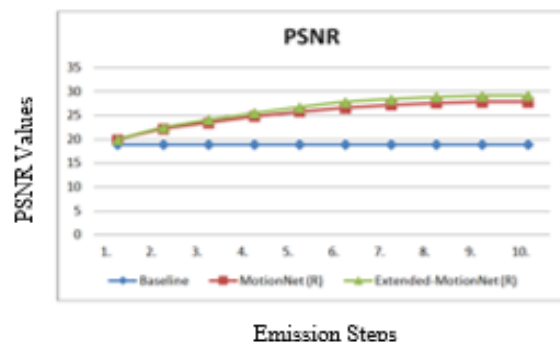


Figure 5. Variation of PSNR per emission

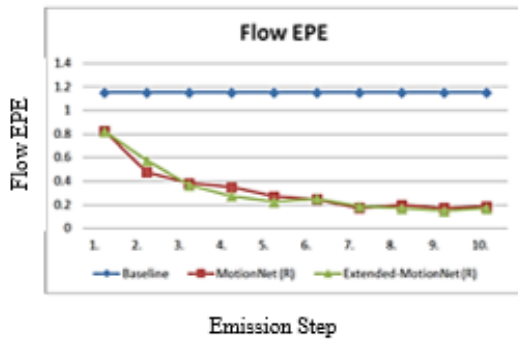


Figure 6. Variation of flow EPE per emission

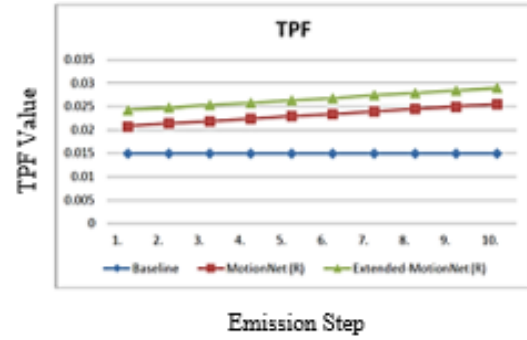


Figure 7. Variation of TPF per emission

Table 1. Average performance values

	Baseline	MotionNet (R)	Extended-MotionNet (R)
Avg. SSIM	0.67	0.8955	0.9036
Avg. EPE	1.154	0.3293	0.3199
Avg. PSNR	18.9	25.33	26.22
Avg. TPF	0.015	0.02317	0.02662

4.2. Comparison with state of art results

During the last few years, numerous deep learning based models and architectures have been invented and studied in tune with already existing widely used traditional codecs. The quality of reconstructed frames by the proposed network has been compared with some of the prominent traditional and deep learning schemes. The MS-SSIM value of the proposed network is comparable to these codecs and the comparative results of the same are given in Table 2. The limitation of the comparison is that the other factors of compression have not been considered and results obtained over a small dataset. But the network shows a promising direction of further exploration in deep learning based video compression domain.

Table 2. Comparative MS-SSIM results of proposed network

Architecture	H.264	H.265	[24]	[25]	Proposed Extended-MotionNet(R)
MS-SSIM	0.955	0.96	0.955	0.9476	0.963

5. CONCLUSION

The proposed architecture presents an end to end trained and learned lightweight deep learning based video compression architecture. The MotionNet network has been extended to incorporate motion compensation autoencoder with farneback based flow estimation. As the whole network is designed in incremental fashion, the experimental results show that the use of flow autoencoder has enhanced the performance of the simple MotionNet architecture comprising of frame autoencoder only. The increased value of SSIM reflects improvement in reconstructed frame but in tradeoff with efficiency measured in terms of frame reconstruction time. This network can be further scaled and enhanced with the inclusion of optimization networks, and entropy coding.




REFERENCES

- [1] Sangeeta and P. Gulia, "MotionNet Architecture with randomized emission training strategy for adaptive bitrate video compression," in *PDGC*, Nov 2020, doi: 10.1109/PDGC50313.2020.9315852.
- [2] C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Kr'ahenb'uhl, "Compressed video action recognition," in *CVPR*, pp. 6026-6035, 2018.
- [3] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," 2015, *arXiv: 1510.00149*.
- [4] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h. 264/avc video coding standard," in *TCSVT*, vol. 13, no. 7, pp. 560-576, 2003, doi: 10.1109/TCSVT.2003.815165.
- [5] G. J. Sullivan *et al.*, "Overview of the high efficiency video coding (hevc) standard," in *TCSVT*, vol. 22, no. 12, pp. 1649-1668, 2012, doi: 10.1109/TCSVT.2012.2221191.
- [6] G. Toderici *et al.*, "Variable rate image compression with recurrent neural networks," 2015, *arXiv: 1511.06085*.
- [7] J. Ball'e, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," 2016, *arXiv: 1611.01704*.
- [8] G. Toderici *et al.*, "Full resolution image compression with recurrent neural networks," in *CVPR*, pp. 5435-5443, 2017, doi: 10.1109/CVPR.2017.577.
- [9] E. Agustsson *et al.*, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *NIPS*, pp. 1141-1151, 2017.




- [10] J. Ball'e, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," 2018 *arXiv*: 1802.01436.
- [11] N. Johnston *et al.*, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks," in *CVPR*, Jun. 2018, doi: 10.1109/CVPR.2018.00461.
- [12] L. Theis, W. Shi, A. Cunningham, and F. Husz'ar, "Lossy image compression with compressive autoencoders," 2017, *arXiv* 1703.00395.
- [13] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in *ICML*, 2017.
- [14] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *CVPR*, Jun. 2018, doi: 10.1109/CVPR.2018.00339.
- [15] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," 2018, *arXiv*: 1804.02958, doi: 10.1109/ICCV.2019.00031.
- [16] G. K. Wallace, "The jpeg still picture compression standard." *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. 18-34, 1992, doi: 10.1109/30.125072.
- [17] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The jpeg 2000 still image compression standard," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36-58, 2001, doi: 10.1109/79.952804.
- [18] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," 2017, *arXiv*: 1711.09078.
- [19] G. Lu, X. Zhang, W. Ouyang, L. Chen, Z. Gao, D. Xu, "An end-to-end learning framework for video compression", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Apr. 2020, doi: 10.1109/TPAMI.2020.2988453.
- [20] N. Johnston *et al.*, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks," in *Computer Vision and Pattern Recognition (cs.CV)*. 2017.
- [21] R. A. Hamzah, M. Md Roslan, A. F. bin Kadmin, S. F. bin Abd Gani, and K. A. A. Aziz, "JPG, PNG and BMP image compression using discrete cosine transform." *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 19, no. 3, pp. 1010-1016, 2021, doi: 10.12928/telkommika.v19i3.14758.
- [22] B. A. Sultan and L. E. George, "Color image compression based on spatial and magnitude signal decomposition," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 5, pp. 4069-4081, 2021, doi: 10.11591/ijece.v11i5.pp4069-4081.
- [23] W. M. Abd-Elhafiez, W. Gharibi, and M. Heshmat, "An efficient color image compression technique," *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 18, no. 4, pp. 2371-2377, 2020, doi: 10.12928/telkommika.v18i5.8632.
- [24] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," 2019, *arXiv*: 1812.00101v3, doi: 10.1109/CVPR.2019.01126.
- [25] S. Kim *et al.*, "Adversarial Video Compression Guided by Soft Edge Detection," 2018, *arXiv*:1811.10673v1.

BIOGRAPHIES OF AUTHORS






Sangeeta    received the B.Tech. and M.Tech. degrees in computer science and engineering from Deenbandhu Chhottu Ram University of Science & Technology, Murthal, Sonapat, India. Currently, she is pursuing her Ph.D. from Maharshi Dayanand University, Rohtak, India. Previously, she had worked as Assistant Professor in Central University of Haryana, Mahendergarh, India for 4.5 years. She has been awarded UGC NET-JRF fellowship to pursue her doctoral degree. She has also qualified IIT-GATE in 2013. Her area of research includes machine learning, artificial intelligence, image and video processing and deep learning. She has authored many research papers of International indexing. She can be contacted at email: sangirao5228@gmail.com.



Preeti Gulia    is currently Associate Professor in the Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India. She is serving the Department since 2009. She earned her doctoral degree in 2013. She has published more than 65 research papers and articles in journal and conferences of National/International repute including ACM, Scopus. Her area of research includes Data Mining, Big Data, Machine Learning, Deep Learning, IoT, and Software Engineering. She is an active professional member of IAENG, CSI and ACM. She is also serving as Editorial Board Member Active Reviewer of International/National Journals. She has guided two research scholars as well as guiding four Ph.D. research scholars from various research areas. She can be contacted at email: preeti.gulia81@gmail.com.



Nasib Singh Gill    received his Post Doctoral in Computer Science from Brunel University, United Kingdom. He is currently working as Professor and Head, Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India. He has authored or coauthored more than 263 refereed journal and conference papers and 5 books. He has been awarded the Commonwealth Fellowship Award by ACU, London at Brunel University, United Kingdom (U.K.) for the year 2001-2002 (01.10.2001-30.09.2002) and. He has also received the Best Paper/Article Award by Computer Society of India in 1994. His research interests include the Software Metrics, Component-based Metrics, Testing, Reusability, Data Mining and Data Warehousing, NLP, Network Security, Information Security. He can be contacted at email: nasibsgill@gmail.com.