Enhancing VVC with Deep Learning based Multi-Frame Post-Processing

Duolikun Danier* Chen Feng* Fan Zhang David Bull
University of Bristol
{duolikun.danier, chen.feng, fan.zhang, dave.bull}@bristol.ac.uk

Abstract

This paper describes a CNN-based multi-frame post-processing approach based on a perceptually-inspired Generative Adversarial Network architecture, CVEGAN. This method has been integrated with the Versatile Video Coding Test Model (VTM) 15.2 to enhance the visual quality of the final reconstructed content. The evaluation results on the CLIC 2022 validation sequences show consistent coding gains over the original VVC VTM at the same bitrates when assessed by PSNR. The integrated codec has been submitted to the Challenge on Learned Image Compression (CLIC) 2022 (video track), and the team name associated with this submission is BVI_VC.

1. Introduction

Video compression is one of the most important and popular topics in the image and video processing research field. It plays an essential role to trade off the tension between the large amount of bitrate required for transmitting immersive and high quality video content and the limited bandwidth available [5]. The efficiency of video codecs have been significantly improved over the past few decades, with the latest MPEG video coding standard, Versatile Video Coding (VVC) [4], achieving nearly 50% coding gains over its predecessor Higher Efficiency Video Coding (HEVC) [12].

More recently, inspired by the advances of machine learning techniques, in particular with deep convolutional neural networks, a number of deep learning based video coding methods have been proposed. Some of these are designed to offer alternative solutions to the conventional coding framework using auto-encoder type architectures associated with end-to-end optimization [1, 7], while another group of methods focus on the enhancement of individual coding tools for standard video codecs [14, 15]. All these method have demonstrated great potential to outperform conventional hybrid video coding algorithms. On the other hand, we noted that the aim of video compression is to offer optimal visual quality with a given bitrate rather than to minimize the absolute difference between the coded content and its corresponding original. This concept can be integrated with the deep learning based coding methods using a perceptually-inspired loss function for training and optimization [10].

In this paper, a deep learning based multi-frame post processing approach is presented, which has been submitted to the Challenge on Learned Image Compression (CLIC) 2022 (video track). This method is based on a previously developed perceptual-inspired Generative Adversarial Network (GAN) architecture, CVEGAN [9]. It allows multiple frames (rather than a single frame) as input, which further improves the overall enhancement performance. This approach has been integrated with the Versatile Video Coding Test Model, VTM 15.2, and it achieves consistent coding gains based on the assessment of PSNR when tested on the CLIC validation video sequences.

The rest of the paper is organized as follows. Section 2 describes the multi-frame post-processing method, the integrated coding framework and the training process. The coding results are then presented in Section 3. Finally, Section 4 concludes the paper and outlines the future work.

2. Proposed Algorithm

The coding framework with the multi-frame post-processing approach is shown in Fig 1. The encoder process is identical to that in standard video codecs, and we use VVC VTM 15.2 [3] as the host encoder. The CNN-based post-processing is applied at the decoder after the host decoder reconstructs video frames from the compressed bitstream. The employed network architecture for multi-frame post-processing and its training process are described below.

2.1. Employed Network Architecture

In this work, we used the same generator architecture of the Generative Adversarial Network for Compressed Video quality Enhancement (CVEGAN), which was originally developed for single frame post-processing and spatial resolution adaptation. CVEGAN has been reported of offer supe-

*Equal contribution.

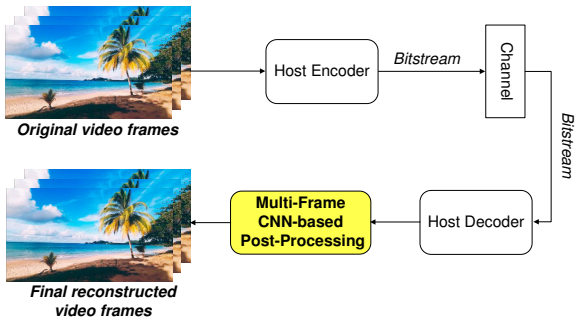


Figure 1. The coding framework with a CNN-based multi-frame post-processing module.

rior coding gains compared to other state-of-the-art network structures when integrated into various coding modules and host codecs [9].

The only difference from the original CVEGAN, where the network processes an input block segmented from a single frame, is that a 96×96×9 input patch is accepted by the generator network in this work. It is obtained by cropping three 96×96×3 YCbCr 4:4:4 blocks from three consecutive reconstructed frames (at the same spatial location), and combining them as a nine channel patch. The network output is in the same format, targeting their uncompressed counterpart.

The architecture of the discriminator also remains the same as the CVEGAN, which takes the output of the generator and the ground truth patch as the input, and outputs a set of feature points for calculating the discriminator loss. More details on the CVEGAN architecture and its training methodology can be found in [9].

2.2. Training Configuration

We also follow the same training strategy as for the original CVEGAN [9], which consists of two stages. First, the generator is trained using a combined perceptual loss function to obtain the preliminary model. The used loss function is given as below.

$$\mathcal{L}_p = 0.3\mathcal{L}_{L1} + 0.2\mathcal{L}_{SSIM} + 0.1\mathcal{L}_{L2} + 0.4\mathcal{L}_{MSSSIM} \quad (1)$$

The generator is then trained jointly with the discriminator using the ReSphereGAN training methodology [9].

The employed network was implemented based on the PyTorch platform version 1.10 [11]. The training process was performed based on the following configurations: Adam optimization [6] with the hyper-parameters: β_1=0.9 and β_2=0.999.; batch size of 16; 200 training epochs (100 for both Stage 1 and 2); initial learning rate (0.0001); weight decay of 0.1 for every 100 epochs.

2.3. Training Content

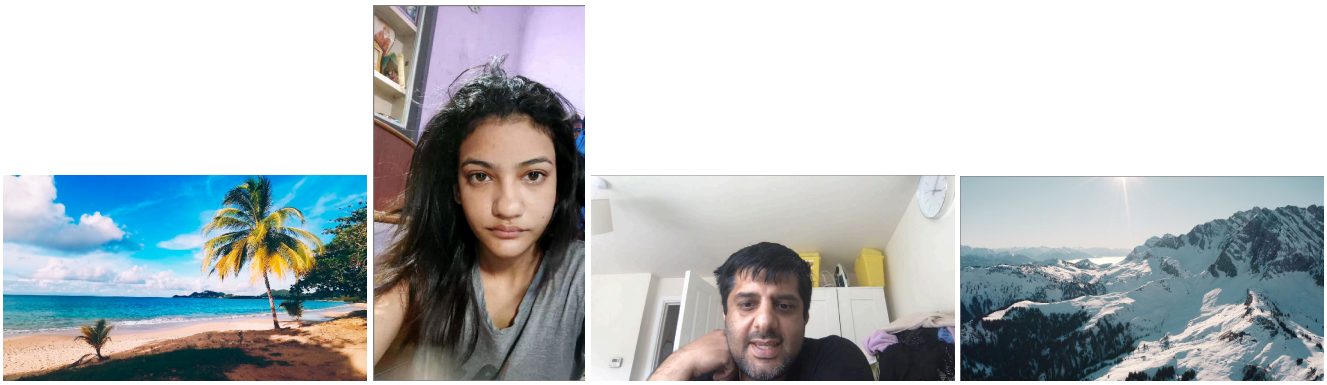
The training data was generated from 200 HD source sequences in the BVI-DVC [8] database, and 562 videos clips (with a spatial resolution of 720p) from the YouTube User Generated Content (UGC) dataset [13]. BVI-DVC has been used by MPEG JVET as a training database for optimizing neural network based coding tools of VVC, while YouTube UGC contains diverse content which has similar characteristics to the CLIC validation set.

All the original sequences were encoded using VVC VTM 15.2 Random Access mode with two quantization parameter (QP) values (32 and 46). These two QP values were selected to simulate the scenarios for two target bitrates (1 Mbps and 0.1 Mbps) set up by the CLIC 2022. All the compressed sequences and their original counterparts were then cropped into 96×96×9 patches and randomly selected as the training material. Rotation and flips were also used for data augmentation. This results in 80,000 pairs of patches in total. After training, two CNN models are obtained for two bitrate scenarios (1 Mbps and 0.1 Mbps).

3. Results and Discussion

To evaluate the performance of the proposed coding framework, four sequences from the CLIC 2022 validation set was used here for testing the proposed method (the CLIC 2022 test set was not available when the paper was submitted). Their indices and example frames are shown in Figure 2. During evaluation, these sequences are first encoded using VVC VTM 15.2 Random Access mode [2] with a QP value of 46. The bitstreams are then decoded using the VVC VTM decoder and converted to YCbCr 4:4:4 format. Each frame together with its temporally previous and subsequent neighbors are then segmented into 96×96×9 overlapping patches (96×96×3 from each frame at the same spatial location) with a spatial overlap size of 4 pixels as network input. The middle three channels of generator output patch (96×96×3) are then converted to RGB format (required by the CLIC 2022) and aggregated following the same pattern to form the final reconstructed current frame. In the cases when processing the first or the last frame of a sequence, we input 96×96×9 patches cropped from this and two subsequent (or previous) frames, and take the first (or the last) three channels of the generator output to form the final reconstructed frame. The training and evaluation operations were executed on a cluster computer with 32 GPU nodes with 2.4GHz Intel CPUs and NVIDIA P100 GPUs.

The proposed method is benchmarked against the original VVC VTM 15.2, using PSNR for quality assessment. Table 1 summarizes the performance of the proposed post-processing method for five different sequences, with an average PSNR gain of 0.09 dB over the original VTM. Here the bitrate remains the same for both codecs (VTM and the



S1 S2 S3 S4

Figure 2. Example frames of the four test sequences.

proposed method) in each test case.

Sequence No	S1	S2	S3	S4
PSNR Gain	0.08dB	0.04dB	0.13dB	0.1dB

Table 1. PSNR gains achieved by the proposed method over the original VVC VTM 15.2.

4. Conclusion

In this paper, we present a CNN-based multi-frame post processing approach for enhancing the visual quality of compression content. This method is based on the perceptual-inspired CVEGAN, and has been integrated with the Versatile Video Coding Test Model (VTM) 15.2 as a submission (BVI_VC) to the Challenge on Learned Image Compression (CLIC) 2022 (video track). This approach has been evaluated on the CLIC 2022 validate sequences, and the results show consistent coding gains based on the assessment of PSNR. Future work should focus on the complexity reduction of the employed network architecture and more advanced structures for multi-frame processing.

Acknowledgement

Duolikun Danier was funded by the China Scholarship Council, University of Bristol, and the UKRI MyWorld Strength in Places Programme (SIPF00006/1).

References

[1] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 1

[2] F. Bossen, J. Boyce, X. Li, V. Seregin, and K. Sühring. JVET common test conditions and software reference configurations for SDR video. In *the JVET meeting*, number JVET-M1001. ITU-T and ISO/IEC, 2019. 2

[3] B. Bross, J. Chen, S. Liu, and Y.-K. Wang. Versatile video coding (draft 7). In *the JVET meeting*, number JVET-P2001. ITU-T and ISO/IEC, 2019. 1

[4] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 1

[5] David R. Bull and Fan Zhang. *Intelligent image and video compression: communicating pictures*. Academic Press, 2021. 1

[6] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[7] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019. 1

[8] Di Ma, Fan Zhang, and David Bull. BVI-DVC: a training database for deep video compression. *IEEE Transactions on Multimedia*, 2021. 2

[9] Di Ma, Fan Zhang, and David R Bull. CVEGAN: A perceptually-inspired gan for compressed video enhancement. *arXiv preprint arXiv:2011.09190*, 2020. 1, 2

[10] Di Ma, Fan Zhang, and David R. Bull. Gan-based effective bit depth adaptation for perceptual video compression. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020. 1

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv:1912.01703 [cs, stat]*, Dec. 2019. arXiv: 1912.01703 version: 1. 2

[12] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video

coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 1

[13] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019. 2

[14] Ning Yan, Dong Liu, Houqiang Li, Bin Li, Li Li, and Feng Wu. Convolutional neural network-based fractional-pixel motion compensation. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):840–853, 2018. 1

[15] Fan Zhang, Chen Feng, and David R Bull. Enhancing VVC through CNN-based post-processing. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 1