Post-Processing Network Based on Dense Inception Attention for Video Compression

Hao Tao[1], Jian Qian[1], Li Yu[1]*, Hongkui Wang[1], Wenhao Zhang[2], Zhengang Li[2], Ning Wang[2], Xing Zeng[2]
[1]School of Electron. Inf. & Commun., Huazhong Univ. of Sci. & Tech., [2]ZTE Corporation
{husthtao, qianjian, hustlyu, hkwang}@hust.edu.cn[1], {zhang.wenhao, li.zhengang1, wangning, zeng.xing1}@zte.com.cn[2]

Abstract

Traditional video coding standards, such as HEVC and VVC, have achieved significant compression performance. To further improve the coding efficiency, a post-processing network is proposed to enhance the compressed frames in this paper. Specifically, the proposed network, namely DIA_Net, contains multiple inception blocks, attention mechanism and dense residual structure. The DIA_Net can efficiently extract information of multiple scale and fully exploit the extracted feature to improve image quality. In addition, the DIA_Net is integrated into the latest test model of VVC (VTM-8.0) to post-process the reconstructed frames of the decoder for better compression performance. The proposed scheme has achieved the best performance in the sense of PSNR at the similar bitrate in the validation sets of challenge on learned image compression (CLIC), which demonstrates the superiority of our approach.

1. Introduction

With the blossom of consumer electronics industry, there is an exponential increase in demand of digital image and video services, which leads to an ever stronger demand for high efficiency video compression techniques. Unlike images, in addition to spatial correlation, there is also temporal correlation between successive video frames. Consequently, extensive researches are devoted to fully exploiting temporal correlation to improve compression efficiency.

The past few decades has witnessed the great progress in video compression and many video coding standards have been released. Among them, the standards released by Moving Picture Experts Group (MPEG), eg., Advanced Video Coding (AVC)[1], High Efficiency Video Coding (HEVC)[2] and Versatile Video Coding (VVC)[3] are widely applied to video compression and transmission. In order to remove temporal redundancy in videos, many coding techniques aiming at representing the moves between the continuous video frames in an efficient way have been adopted in the standards. Multiple novel coding techniques have been integrated into the latest VVC standard, which leads to its superb coding performance at present.

With the booming of deep learning in recent years, many researchers focus on improving video compression performance with the help of neural network. Video coding standardization organizations, such as MPEG and AVS, have set up intelligent video coding group to promote the development of learning based video compression. Meanwhile, extensive works have demonstrated the superiority of applying deep learning to video coding and the learning based post-process technique is an outstanding branch of them. Dai *et al*. [4] designed a Variable-filter-size Residue-learning convolutional neural network (VRCNN) to improve the post-processing performance of HEVC and achieved on average 4.6% bit-rate reduction. Liu *et al*. [5] proposed a frame-based post-processing scheme for HEVC, which adopts a 20-layers CNN aided with metadata to reduce reconstruct error and to improve the post-processing perfoamance. Yet the methods have improved compression performance in HEVC to a certain extent, post-processing performance in VCC still can be improved.

In order to further enhance the quality of compressed frames in VVC, a novel video compression scheme based on post-processing network is proposed in this paper. To reduce the compression artifacts and obtain compressed frames of better quality, we design a dense inception attention based neural network (DIA_Net) to post-process the reconstructed frames at the decoding side. The DIA_Net contains multiple inception blocks which utilize kernels of different size to dig out features in different scale. Meanwhile, attention mechanism including spatial attention and channel attention is proposed to fully exploit feature information and dense residual structure is adopted to deepen the network and increase model capacity. In addition, the DIA_Net is integrated into VTM-8.0 to serve as a post-processing module for better compression quality. Exper-

*This work was supported in part by the NSFC under Grant 61871437 and the Natural Science Foundation of Hubei Province of China under Grant 2019CFA022

imental results demonstrate that the proposed video compression approach can achieve superb performance in the validation sets of CLIC[6].

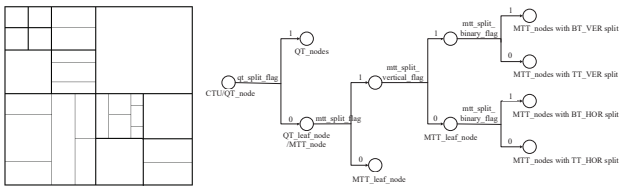
The rest of this paper is organized as follow: inter prediction methods in VVC will be reviewed in section II and our DIA-Net will be concretely described in section III. Experimental results will be presented and analyzed in Section IV and the conclusion will be given in the Section V.

2. Inter Prediction in VVC

In video compression, current frame can be predicted from previous compressed frames and only the residual need to be transmitted. In this section, we will concisely review the key techniques for inter prediction in VVC.

2.1. Partitioning of the CTUs in VVC

Video frames are divided into a sequence of coding tree units (CTUs) and a CTU is further split into coding units (CUs) to adapt to various local characteristics. A quad-tree with nested multi-type tree using binary and ternary splits segmentation structure is proposed to split the CTUs in VVC, which replaces the concepts of multiple partition unit types in HEVC, eg., prediction units (PUs) and transform units (TUs). As shown in Figure 1 (a), a CTU is split into multiple CUs with a quad-tree and nested multi-type tree coding block structure, where the bold block edges represent quad-tree partitioning and the remaining edges represent multi-type tree partitioning. Besides, the split mode decision is made by the procedure shown in Figure 1 (b).



(a) (b)

Figure 1. Example of quad-tree with nested multi-type tree coding block structure and split mode decision procedure.

2.2. Key Inter Prediction Techniques in VVC

Generally, in order to obtain a prediction picture of the current frame, motion estimation and motion compensation will be performed on the reference picture. Multiple techniques aiming at improving the aforementioned process are adopted to obtain a prediction picture of better quality. To achieve higher-precision motion estimation, subblock-based temporal motion vector prediction (SbTMVP) is proposed to obtain the optimal motion vector in sub-CU level. Adaptive motion vector resolution (AMVR) scheme and higher-precision motion field storage are proposed to more efficiently process the motion vectors of high precision. Besides, a block-based affine transform motion model is applied in motion estimation and motion compensation. Compared with traditional translation motion model, motion vector can be derived from more directions and the compensated block can be more similar with the original block, which leads to better coding performance. Owing to these inter prediction methods, the coding performance is significantly improved. More details can be found in [7].

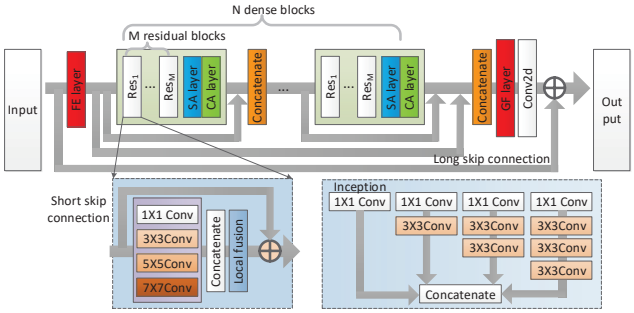


Figure 2. The architecture of the proposed DIA_Net.

3. Dense Inception Attention Based Post-Processing Network

Since the block based hybrid coding structure is adopted in VCC, major operations such as quantization, intra-prediction, inter-prediction are performed block by block. Consequently, the coding parameters vary by the blocks, which leads to blocking effects. In addition, high frequency components of the video will be lost during quantization and transform process, which results in ringing and blurring effects. Aiming at eliminating these compression artifacts, a post-processing network is designed to improve the quality of compressed videos. The proposed post-processing network, namely DIA_Net, contains inception structure, attention mechanism and dense residual structure, which makes it capable of processing different kinds of compression artifacts. More details about the DIA_Net will be introduced below.

3.1. Network Architecture

As shown in Figure 2, the proposed DIA_Net is mainly composed of three modules, namely inception module, dense residual block and attention mechanism. The input frame is first fed to the feature extraction layer to generate a feature map, which is utilized as the input of dense residual blocks then. After the feature map got through the dense blocks which include multiple residual blocks orderly attached with spatial attention layer and channel attention layer, the features are finally sent to the global fusion layer. The final output frame is generated by adding the input frame and the output of global fusion layer. Due to the well-designed network architecture, the DIA_Net can capture features of multiple scale and fully exploited the spatial

and channel correlation.

Inception Structure: As illustrated in [8], kernels with different size are sensitive to information of different scale. Specifically, smaller kernels are more sensitive to subtle information like dense contours while kernels with larger size focus on coarse outlines. In view of this, the kernel size of the convolution layers in the DIA_Net are set to be 1×1, 3×3, 5×5 and 7×7, which is shown in the bottom squares of Figure 2. Consequently, the convolution operation can be preformed in different scale and then all outputs are concatenated together. To generate the final output, the concatenated features are fed into a local fusion layer and a short skip connection is also applied. The aforementioned process can be formulated as fellow:

$$f_i^m = Res_m(f_i^{m-1}) = f_i^{m-1} + ReLU \\ (cat[Conv_1(f_i^{m-1}), ..., Conv_k(f_i^{m-1})]) \tag{1}$$

Where f_i^{m-1} and k denote input frame and kernel size, respectively. The activation function utilized here is rectifier linear unit (ReLU). The processed features are then added with input f_i^{m-1}.

Attention Mechanism: Fei *et al*. [9] have indicated that network performance can be significant improved by attention mechanism. Therefore, we also applied it to our DIA_Net. In order to weight the importance of different channels and spatial regions to the quality of final reconstruction frame, channel attention (CA) and spatial attention (SA) are introduced to the post-processing network in a learning manner. Inspired by the fact that some channels take no effects to the final output while others have great impact on that, an intuitive idea is to distribute weights to different channels according to its importance, which is the main principle of CA. Similarly, there also exits importance difference in different spatial regions. Consequently, a weight map aiming at allocating weights to each pixel is learned to represent the SA.

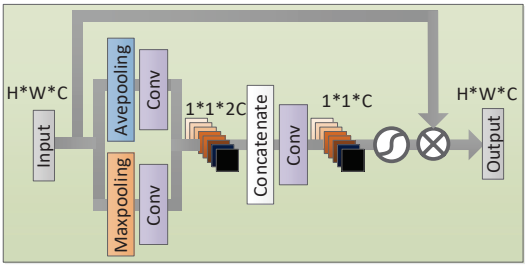
As shown in Figure 3, features with size $H \times W \times C$ are fed into CA layer where C denotes the channel number and the weights of channels are extracted with average and max pooling. The outputs are then concatenated and activated by sigmoid function. In addition, the weights are squashed within [0,1] and a channel-wise multiplication is employed between input and squashed weights. The aforementioned process is formulated as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} x_c(i,j) \tag{2}$$

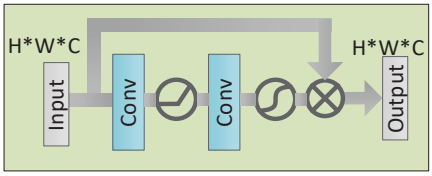
$$\hat{z}_c = \max(x_c(i,j)) \tag{3}$$

$$y = [x \times \sigma(z), x \times \sigma(\hat{z})] \tag{4}$$

Where x_c and y denote input and output of CA layer. σ is sigmoid function, $z_c, \hat{z}_c$ represent output of average and max pooling. As for SA layer, features are fed to the convolution layers and activated by ReLU function. As formulated in equation (5), the final output is the element-wise product of input features and activated outputs.



(a) Channel attention layer.



(b) Spatial attention layer.

Figure 3. Details of spatial attention layer and channel attention layer.

$$y = x \times \sigma(M_{SA}(x)) \tag{5}$$

Dense Residual Structure: As illustrated in [10] and [11], dense architecture and residual learning can deepen the model and exploit hierarchical information. Considering that memory consumption increases when dense network is utilized, we propose a dense residual (DR) structure to implement a dense network with low memory cost. Specifically, as shown in Figure 2, the DR blocks are orderly connected and each block includes M residual blocks. Since we just concatenate the output of each DR block as formulated in equation (6), memory consumption can be greatly reduced compared with a fully connected dense network.

$$f_i^n = cat[f_i^{n-1}, DR(f_i^{n-1})] \tag{6}$$

Where f_i^{n-1} and f_i^n denote the input and output of DR block. The f_i^{n-1} is fed to DR and processed by M cascaded residual blocks. Besides, N DR blocks are employed in our DIA_Net.

3.2. The DIA_Net-Based Post-Processing Scheme

The proposed DIA_Net is mainly utilized in the decoding end of the VTM-8.0 to enhance the quality of reconstruction frame. Specifically speaking, immediately after the decoder has generated the reconstruction frame, it will be fed to the DIA_Net to obtain the final enhanced frame. In particular, the proposed post-processing scheme can be applied for different kinds frames including I-frame, P-frame and B-frame.

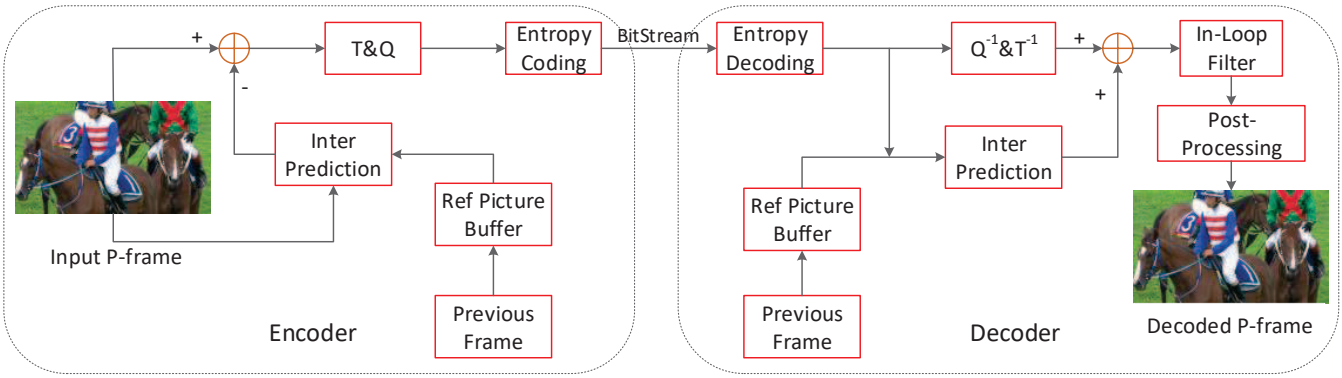


Figure 4. The framework of the proposed SmartPCodec.

4. Experimental Results

4.1. Implementation

Special P-frame Codec for CLIC: In view of the superb coding performance of VVC, a special codec, namely SmartPCodec, is developed from the VTM-8.0. As shown in Figure 4, the proposed codec is customized for the P-frame track of CLIC. Since we need to compress a video frame conditioned on the previous uncompressed frame, we modified VTM-8.0 by skipping the encoding process of the first frame and directly adding the previous frame to the reference picture list for P-frame compression. In addition, we also remove many syntax which are not utilized in P-frame compression to reduce bit cost and complexity. Meanwhile, corresponding modifications are made in the decoder to keep consist with the encoder.

SmartPCodec Configuration: Since the SmartPCodec is a modified version of VTM-8.0, the general test condition is consistent with the Common Test Conditions[12] of VVC. It's worth noting that we simplified the hierarchical P-frame coding structure for just one P-frame per sequence. In addition, the quantization parameter (QP) is set according to the target bit-rate.

Training Dataset: The UGC dataset, released by CLIC, contains 739 videos, with a total of 466684 frames. Since the designed DIA_Net is aimed at enhancing the reconstructed frame of SmatrPCodec, we first pre-process the videos by encoding them with SmartPCodec and utilize the reconstructed video frames as the input of DIA_Net during the training phase. It should be noted that the post-processing module of SmartPCodec is disabled when pre-processing the dataset.

Training Settings: We train our network using Adam[13] with β_1 = 0.9, β_2 = 0.999, a learning rate of 0.001 and a mini-batch size of 16 samples. The network is optimized in the sense of L1 loss and one patch with size 64×64 is randomly cropped from each frame as input during the training. Besides, we also augment the input data by randomly rotating the patches.

Method	Data Size(Bytes)	PSNR(dB)	MS-SSIM
VTM-8.0	38,891,706	42.081	0.9958
SmartPCodec	38,701,571	42.344	0.9960

Table 1. Compression performance in the validation sets of CLIC.

4.2. Performance

In order to evaluate the proposed SmartPCodec, the codec is tested in the validation sets of CLIC. Meanwhile, the VTM-8.0 is also tested in the aforementioned sets for comparing with our SmartPCodec. As shown in Table 1, compared with original VTM-8.0, the improvement of P-SNR is 0.263 dB and the improvement of MS-SSIM is 0.0002, respectively. Since we applied the post-processing network to enhance the compressed frames and remove the syntax that are not related to P-frame compression, the bit consumption of our codec is lower while the compression performance is better. Besides, the performance of SmartP-Codec ranks first among all the participants of CLIC in the sense of PSNR, which demonstrates the superiority of the proposed scheme.

5. Conclusion

In this paper, we propose a novel video compression scheme based on post-processing network. In order to improve the quality of compressed frames in VVC, we design a dense inception attention based post-processing network. The proposed post-processing network is composed of inception structure, dense residual blocks and attention mechanism, which makes it capable of extracting feature of multiple scale and process various kinds of compression artifacts. In addtion, the proposed scheme is applied in P-frame track of CLIC and a special codec is developed from VTM-8.0 to meet the conditions of P-frame compression. Evaluation results demonstrate that our scheme outperforms others in the validation sets of CLIC.

References

[1] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003.

[2] 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.

[3] Bross Benjamin, Chen Jianle, and Liu Shan. Versatile video coding (draft 5). In *JVET-N1001*, 2019.

[4] Yuanying Dai, Dong Liu, and Feng Wu. A convolutional neural network approach for post-processing in hevc intra coding. In *International Conference on Multimedia Modeling*, pages 28–39. Springer, 2017.

[5] Chen Li, Li Song, Rong Xie, and Wenjun Zhang. Cnn based post-processing to improve hevc. In *2017 IEEE International Conference on Image Processing*, pages 4577–4580. IEEE, 2017.

[6] Workshop and challenge on learned image compression(CLIC). http://www.compression.cc.

[7] Chen Jianle, Ye Yan, and Kim SeungHwan. Algorithm description for versatile video coding and test model 5 (vtm 5). In *JVET-N1002*, 2019.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[9] Wang Fei, Mengqing Jiang, Qian Chen, Shuo Yang, Li Cheng, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[12] Bossen Frank, Boyce Jill, Suehring Karsten, Li Xiang, and Seregin Vadim. Jvet common test conditions and software reference configurations for sdr video. In *JVET-N1010*, 2019.

[13] Kingma Diederik P and Ba Jimmy. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*, 2014.