Low Complexity Neural Network-Based In-loop Filtering with Decomposed Split Luma-Chroma Model for Video Compression

Tong Shao [1] **Jay N. Shingala** [2] **Ajay Shyam** [2] **Peng Yin** [1] **Arjun Arora** [1] **Sean McCarthy** [1]

Abstract

In this paper, a novel low complexity split luma-chroma model is proposed for in-loop filtering in video compression. The basic block of the model adopts the decomposed regular 3x3 convolutional layer, which is replaced by 1x1 pointwise convolutions and 3x1/1x3 separable convolutions via CP decomposition to reduce complexity. It's further proposed to fuse the two adjacent 1x1 convolutional layers into one. To efficiently exploit the dependencies between luma and chroma while modeling the independent characteristics of luma/chroma component, a novel split luma-chroma architecture within one CNN model is proposed. The input layer and the first hidden layers serving as the common path jointly process luma-chroma inputs. Then the output feature maps are split into luma and chroma feature maps, and they are independently processed using the same basic block as in common path, i.e., one luma path with 24 channels and one chroma path with 8 channels. Experimental results show that the model has 5.66% BD-Rate luma gain over NNVC-4.0 under Random Access (RA) while the chroma gains are also greatly improved, at the complexity of 17.7 kMAC/Pixel. The BD-Rate and kMAC/Pixel plot also shows the superior trade-off between complexity and coding gain compared to state-of-the-art filters. And the subjective results demonstrate improved visual quality. Moreover, the split luma-chroma architecture also possesses the flexibility to get arbitrary luma-chroma rate-distortion distribution by adjusting the number of channels in each path.

[1]Dolby Laboratories, Inc., Sunnyvale, CA, USA [2]Ittiam Systems Pvt. Ltd., Bengaluru, India. Correspondence to: Tong Shao <tong.shao@dolby.com>.

Published at ICML 2023 Workshop Neural Compression: From Information Theory to Applications.

1. Introduction

In recent years, the fast development of video related applications, e.g., video social media and video conference, has brought about large amount video data for transmission and storage. And the demand for more efficient video compression is increasing. There have been lots of research efforts on video compression and the state-of-the-art video coding standard Versatile Video Coding (VVC) (Bross et al., 2021) has achieved up to 40% bit-rate reduction compared to the previous High Efficiency Video Coding (HEVC) (Sullivan et al., 2012).

In a typical block-based hybrid video compression framework such as VVC, the raw video frames will be partitioned into blocks, and then experience Intra/Inter prediction, residual transform, quantization and entropy coding before forming the bitstreams. Each reconstructed frame will be processed by in-loop filters to alleviate coding artifacts and enhance quality. Since the filtered frame will also be used as reference for future frames, the benefit from in-loop filtering will improve not only current frame, but also greatly increase the overall coding efficiency, making the in-loop filter an essential tool in video compression. In VVC, the deblocking filter (DBF) is utilized to reduce the blocking artifacts by smoothing the boundaries. Then the Sample Adaptive Offset (SAO) is used to alleviate the ringing artifacts by analyzing the pixels in the block and adding an offset accordingly. And another Adaptive Loop Filter (ALF) is adopted to minimize the mean square error between the original pixels and the reconstructed pixels by adaptively updating the filter coefficients.

To further alleviate the artifacts and improve the coding efficiency, neural networks, most of which are convolutional neural network (CNN), have been introduced for in-loop filtering. With the original uncompressed samples as the targeted output, the CNN models are designed to learn the features of the coding artifacts and reconstruct high quality frame samples in an end-to-end style. Dai et al. (2017) first introduced a variable-filter-size residue-learning CNN (VRCNN) model to replace the standard DBF and SAO in intra frame filtering. Zhang et al. (2018) proposed a residual highway CNN (RHCNN) model with a key technique of direct link between layers. Temporal information, i.e., the

previously reconstructed frames are used as extra inputs to the model with method to select high quality reference frames (Li et al., 2019; Shao et al., 2022). Meanwhile, some proposals on neural network-based loop filter in JVET Exploration Experiments studies have shown state-of-the-art coding performance, which is around 10% gain (Random Access) with 500 kMAC/Pixel (number in 1000 of Multiply-Accumulate operations per pixel) complexity. For example, in JVET-AA0088 (Wang et al., 2022), ResBlocks are used as basic modules, while the predicted YUV is processed and concatenated with reconstructed YUV. And in JVET-AA0111 (Li et al., 2022), Attention Residual Blocks are used and the luma and chroma are processed separately.

Though compression efficiency is greatly improved, most of the models above are of very high complexity, making it difficult to be deployed in real word applications. Therefore, this paper is aimed at a low complexity neural network-based in-loop filter, which should provide good gain with much less complexity, while the trade-off is improved. Two low complexity models provide the best performance. In JVET-X0140 (Wang et al., 2021), a simple CNN model is introduced with the Boundary Strength and quantization step as additional inputs. In (Shingala et al., 2022), a CNN model using fused CP decomposition is proposed and the complexity is further reduced.

The above two models provide good trade-off and compression efficiency at low complexity operation point. And both are processing the luma and chroma components jointly using one single CNN model. However, though they indeed exploit cross-channel dependencies between luma and chroma, luma and chroma components also have very different signal characteristics and coding error induced distortions. Therefore, it's hard for one single model to efficiently process both luma and chroma. To be more specific, in video compression, this may result in the unbalanced rate-distortion performances and compression efficiency between luma and chroma. For a commonly used YUV420 format video, chroma components are only half resolution of luma, and they require less fidelity during compression as human perception is more sensitive to luma in motion pictures. Thus, the single joint model has very little flexibility to adjust to these characteristics.

To exploit both cross-channel dependencies and independent characteristics of luma-chroma components, we propose a split luma-chroma model with decomposed convolutional neural network. Inspired by the baseline CNN model in JVET-X0140 (Wang et al., 2021) and using the fused CP decomposed convolution in (Shao et al., 2023) as the baseline, the proposed model split the luma and chroma feature maps into two independent groups of channels in the intermediate layers, and achieves improved compression efficiency, better performance-complexity trade-off as well as good subjective results. Firstly, the regular 3x3 convolutional layers are replaced by 1x1 convolutional layers and separable convolutional layers via CP decomposition (Lebedev et al., 2014) to reduce complexity. And the adjacent two 1x1 convolutional layers after decomposition are fused into one single 1x1 convolutional layer. In the proposed model, the input layer and the first hidden layers jointly process luma-chroma samples. And the output feature maps are split into luma and chroma feature maps, which are then independently processed using two groups of channels and output for reconstructed luma-chroma samples. The model is implemented in Tensorflow and tested on top of NNVC-4.0 (Alshina et al., 2023). Experimental results have proved the good trade-off between complexity and coding gain of the proposed model. And the split luma-chroma architecture provides great flexibility to get any desired rate-distortion point and luma-chroma performance distribution by adjusting the split channels number and/or hidden layers numbers. To the best of our knowledge, this is the first work to use split luma-chroma channels in one single CNN model for in-loop filtering and achieves good performance trade-off.

2. Proposed method

The overall architecture of proposed low complexity NNLF model is illustrated in Figure 1. Assuming video frames of YUV420 format are being processed, the input to the model consists of 10 2-dimensional tensors. Four of them are luma reconstructed samples before deblocking, while two are chroma reconstructed samples, one for the quantization step and three for the boundary strength of YUV respectively. The quantization step and the boundary strength provide the location and the intensity information of the compression artifacts, as the artifacts are mainly introduced during transform and quantization of the prediction residuals. In the input, the dimension of each tensor is 72x72, which corresponds to a 128x128 luma block and 4 neighboring samples from left, right, top and bottom regions of luma block reorganized into 4 sub-blocks. These tensors are input into an 3x3 convolutional layer first, followed by a Leaky ReLU. Then, the feature maps are input to a basic block (black dotted rectangle), which consists of a 1x1 convolutional layer, a Leaky ReLU, a 1x1 convolutional layer and a 3x3 convolutional layer that will actually be decomposed to 1x1 point-wise conv. layers and 3x1 separable conv. layer, and 1x3 separable con. layer for complexity reduction. The 3x3 conv. layer performing CP decomposition is depicted in the yellow dotted rectangle in Figure 1. The input layer and the first hidden layers (basic block) jointly process luma-chroma samples, exploiting the dependencies between luma and chroma.

After the common path, the K output feature maps of the first hidden layers are split into luma (K_Y) and chroma

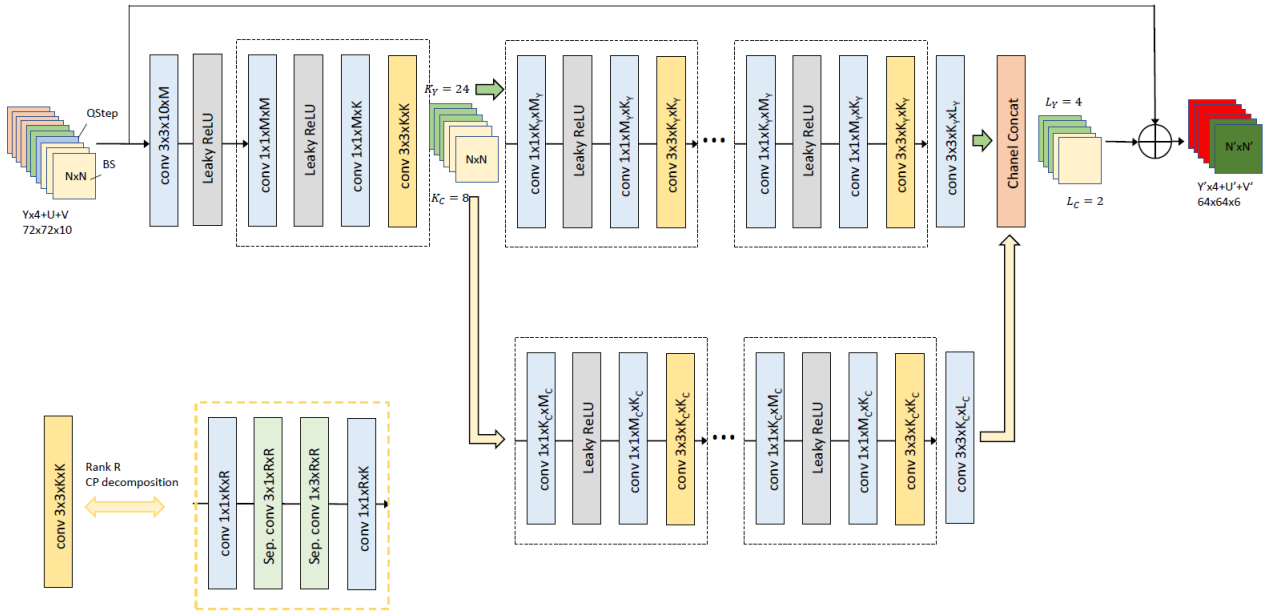

Figure 1. Architecture of the proposed Split Luma-Chroma CNN model with fused CP decomposition for in-loop filtering. For the conv AxBxCxD, AxB represents the kernel size, and C and D represent the input channel number and the output channel number respectively.

(K_C) feature maps, where K_Y and K_C represent the number of channels in luma and chroma paths respectively. Then the split luma and chroma features maps are independently processed using n_Y basic blocks for luma and n_C basic blocks for chroma. The basic block's components are the same as the one in the common path. Similarly, the 3x3 conv. layer is replaced by 4 conv. layers, among which are two separable conv. layers. To further reduce complexity, the first 1x1 decomposed layer can be conveniently fused with the 1x1 conv. layer before, resulting in one single 1x1 conv. layer. And the last 1x1 decomposed layer can be fused with the 1x1 conv. layer after. The idea of fused CP decomposition in CNN model is derived from (Shao et al., 2023). The key idea of split luma-chroma paths is to provide better modeling of luma and chroma's distortion characteristics independently, and by adjusting the number of channels (K_Y, K_C) in each path, there is a flexibility to get arbitrary luma-chroma rate-distortion distribution/balance.

In each path, the last layer before output is another 3x3 convolutional layer. The luma and chroma samples before input are added to the NNLF model's output and hence the final output of the filtered samples consists of 6 2-dimensional 64x64 tensors, which represent 4 planes of luma samples and 2 planes of chroma samples respectively.

In our model parameter settings, M=72, K=32, K_Y=24, K_C=8, n_Y=n_C=10, R_Y=24, R_C=24. The number of channels in luma path is 24, while for chroma path it's 8. This corresponds to the fact that chroma components are only half resolution of luma and its fidelity requirement is less than luma's. In practice, we found that for a simple low complexity CNN, it's very hard to let the model automatically learn the distribution of channels for luma or chroma components. R_Y and R_C represent the Rank of CP decomposition, i.e., the intermediate channel numbers for basic blocks in luma and chroma paths respectively. Finally, the kMAC/Pixel of this split luma-chroma model is calculated to be 17.7.

3. Experimental Results

3.1. Implementations and Simulations

The proposed model is placed before the SAO and ALF stages in the NNVC-4.0 (Alshina et al., 2023), and is implemented in Tensorflow. The training is conducted on Nvidia A100, and the inference is performed only in CPU. The models are trained using DIV2K (Agustsson & Timofte, 2017) dataset for AI and BVI-DVC (Ma et al., 2021) dataset for RA. The raw images of DIV2K were compressed using NNVC-4.0 to generate the AI training samples. For RA, two-stage training is performed: 1st stage RA training dataset was generated using NNVC-4.0 plus AI trained models without deblock RDO; 2nd stage RA training dataset was generated using NNVC-4.0 plus AI and first stage RA models with deblock RDO enabled. 2 models are trained for Intra and 2 models are trained for Inter.

Table 1. BD-Rate (%) of the proposed split luma-chroma model compared to NNVC-4.0 anchor, under RA configuration. Negative value means coding gain.

CLASS	Y	U	V
A1	-5.85%	-6.90%	-6.23%
A2	-6.01%	-9.04%	-5.40%
B	-5.28%	-9.85%	-8.88%
C	-5.74%	-10.59%	-8.78%
D	-6.94%	-10.04%	-10.44%
OVERALL	-5.66%	-9.30%	-7.63%

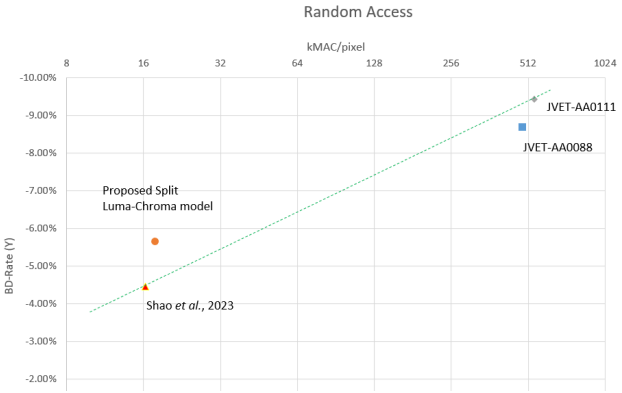

Figure 2. Complexity vs. gain trade-off comparisons of the state-of-the-art models under RA.

Inference simulations are performed under common test conditions for neural network-based video coding technology (Alshina et al., 2023). The test sequences include class A (3840x2160), class B (1920x1080), class C (832x480), class D (416x240) and class E (1280x720), under All Intra (AI) and Random Access (RA) configurations. And 5 quantization parameters (QPs) 22, 27, 32, 37, 42 are tested to calculate the BD-Rate (Bjøntegaard, 2001), which represents the bit saving of a model compared to anchor with same PSNR quality.

3.2. Rate-distortion Performance

Table 1 illustrates the BD-Rate performance of the proposed NNLF model compared to NNVC-4.0 anchor under Random Access (RA). The proposed split luma-chroma model achieves the BD-Rate savings of 5.66%, 9.30%, 7.63% (Y, U, V, respectively) for RA compared to NNVC-4.0 anchor. And it also shows the coding gain of 4.99%, 7.46%, 7.03% (Y, U, V, respectively) for All Intra (AI) compared to the same anchor.

Compared to the results of the model in (Shao et al., 2023), our split luma-chroma model shows the coding gain of 1.21%, 3.85%, 2.35% (Y, U, V, respectively) for RA. The complexity increases by only 1.5 kMAC/Pixel (ours 17.7 vs. (Shao et al., 2023) 16.2). This clearly demonstrates the improved coding efficiency of compared to the state-of-the-art low complexity model. And both luma and chroma components have improved performance, while the U and V's are even more obvious, proving the split luma-chroma architecture's effectiveness in balancing performances between components. To better demonstrate the complexity vs. coding gain trade-off of the proposed model compared to other state-of-the-art ones, the BD-Rate vs. kMAC/Pixel plot is provided in Figure 2. From the distribution of plot locations around the dotted line, it can be concluded that the proposed model has an obviously better trade-off than the existing models. And compared with the two with best RD performance, the proposed model can provide more than half of the coding gain with 3% of the complexity.

3.3. Subjective Results

Subjective results comparison is provided in Figure 3 in Appendix A. Filtering examples under RA from our split luma-chroma model and the VVC default filters are provided. The QP is set around 37, and the bits from our model is equal or less than the VVC codec. It demonstrates that with the same bits, our model can compress and reconstruct the video frames with higher subjective quality. To be more specific, the model can preserve the boundaries better, especially in the areas with lines and clear structures.

4. Conclusions

A novel low complexity split luma-chroma model is proposed for in-loop filtering in video compression. The basic block of the model adopts the CP decomposed 3x3 conv. layers, which is replaced by 1x1 conv. and 3x1/1x3 separable conv. to reduce complexity. It's further proposed to fuse the two adjacent1x1 conv. layers into one. To efficiently exploit the dependencies between luma and chroma while modeling their independent characteristics, a novel split luma-chroma architecture is proposed. The input layer and the first hidden layers serving as the common path jointly process luma-chroma inputs. And the output feature maps are split into luma and chroma feature maps, which are independently processed using the same basic block as in common path, i.e., one luma path with 24 channels and one chroma path with 8 channels. Experimental results show improved BD-Rate, subjective quality, the state-of-the-art gain vs. complexity trade-off and more balanced luma-chroma rate-distortion distribution. And the model also possesses the flexibility to get arbitrary luma-chroma rate-distortion distribution by adjusting the number of channels in each path.

References

Agustsson, E. and Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017.

Alshina, E., Liao, R.-L., Liu, S., and Segall, A. JVET common test conditions and evaluation procedures for neural network-based video coding technology. *JVET-AC2016, Joint Video Experts Team (JVET), January*, 2023.

Bjøntegaard, G. Document VCEG-M33: Calculation of average psnr differences between rd-curves. *Proceedings of the ITU-T Video Coding Experts Group (VCEG) Thirteenth Meeting, April*, 2001.

Bross, B., Wang, Y.-K., Ye, Y., Liu, S., Chen, J., Sullivan, G. J., and Ohm, J.-R. 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.

Dai, Y., Liu, D., and Wu, F. A convolutional neural network approach for post-processing in hevc intra coding. In *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I 23*, pp. 28–39. Springer, 2017.

Lebedev, V., Ganin, Y., Rakhuba, M., Oseledets, I., and Lempitsky, V. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*, 2014.

Li, T., Xu, M., Zhu, C., Yang, R., Wang, Z., and Guan, Z. A deep learning approach for multi-frame in-loop filter of hevc. *IEEE Transactions on Image Processing*, 28(11): 5663–5678, 2019.

Li, Y., Zhang, K., Li, J., Zhang, L., Wang, H., Reuze, K., Kotra, A. M., Karczewicz, M., Galpin, F., Andersson, K., Ström, J., Liu, D., and Sjöberg, R. EE1-1.6: Deep in-loop filter with fixed point implementation. *JVET-AA0111, Joint Video Experts Team (JVET), October*, 2022.

Ma, D., Zhang, F., and Bull, D. R. Bvi-dvc: A training database for deep video compression. *IEEE Transactions on Multimedia*, 24:3847–3858, 2021.

Shao, T., Liu, T., Wu, D., Tsai, C.-Y., Lei, Z., and Katsavounidis, I. Ptr-cnn for in-loop filtering in video coding. *Journal of Visual Communication and Image Representation*, 88:103615, 2022.

Shao, T., Shingala, J. N., Yin, P., Arora, A., Shyam, A., and McCarthy, S. A low complexity convolutional neural network with fused cp decomposition for in-loop filtering in video coding. In *2023 Data Compression Conference (DCC)*, pp. 238–247. IEEE, 2023.

Shingala, J. N., Kadaramandalgi, S, Shyam, A., Shao, T., Arora, A., Yin, P., Pu, F., Lu, T., and McCarthy, S. AHG11: Complexity reduction on neural-network loop filter. *JVET-AA0080, Joint Video Experts Team (JVET), July*, 2022.

Sullivan, G. J., Ohm, J.-R., Han, W.-J., and Wiegand, T. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012.

Wang, H., Chen, J., Reuze, K., Kotra, A. M., and Karczewicz, M. EE1-1.4: Tests on neural network-based in-loop filter with constrained computational complexity. *JVET-X0140, Joint Video Experts Team (JVET), October*, 2021.

Wang, L., Xu, X., and Liu, S. EE1-1.5: neural network based in-loop filter with a single model. *JVET-AA0088, Joint Video Experts Team (JVET), October*, 2022.

Zhang, Y., Shen, T., Ji, X., Zhang, Y., Xiong, R., and Dai, Q. Residual highway convolutional neural networks for in-loop filtering in hevc. *IEEE Transactions on image processing*, 27(8):3827–3841, 2018.