Learned Image Restoration for VVC Intra Coding

Ming Lu† Tong Chen† Haojie Liu† Zhan Ma‡
Nanjing University, Nanjing, China
†{luming, tong, haojie}@smail.nju.edu.cn ‡mazhan@nju.edu.cn

Abstract

We propose a learned image restoration network as the post-processing module for emerging Versatile Video Coding (VVC) Intra Profile (https://jvet.hhi.fraunhofer.de) based image coding to further improve the reconstructed image quality. The image restoration network is designed using multi-scale spatial priors to effectively alleviate compression artifacts in the decoded images induced by the quantization based lossy compression algorithms. Experimental results demonstrate the performance gains of our proposed post-porcessing network with VVC Intra coding, offering about 6.5% Bjontegaard-Delta Rate (BD-Rate) reduction for YUV 4:4:4 and 12.2% for YUV 4:2:0, against the VVC Intra without our restoration network on the Test Dataset P/M released by the Computer Vision Lab of ETH Zurich, where the distortion is Peak Signal to Noise Ratio (PSNR).

1. Introduction

Image compression algorithms (e.g., JPEG and its successor JPEG 2000 [12]) are often used to compress the raw images to ensure the smooth network delivery and guarantee the satisfactory Quality of Experience (QoE) to some extent for end users. Meanwhile, some alternatives such as WebP, BPG (an image compression method uses the modified High-Efficiency Video Coding (HEVC) [4] Intra Profile) and other intra coding modes of video codecs also show impressive image compression performance. However, the lossy image compression is always accompanied by undesired artifacts, such as blocky, motion blurring and ringing, especially at high compression ratios (or equivalent low bittate), resulting in unpleasant visual experience. With the exponential growth of the image applications over the Internet (e.g., sharing, exchange, storage, etc), it is necessary to develop more efficient lossy image compression algorithms with higher performance.

Recent works have revealed the great potential in lossy image compression using deep learning. Liu *et al.*, Minnen *et al.*, Rippel *et al.*, etc [5, 7, 9] propose end-to-end image compression frameworks using stacked convolutional neural networks (CNNs) based auto-encoders. These methods, completely relying on deep learning technologies, present impressive performance gains with considerable BD-Rate reduction against existing traditional image compression algorithms, such as JPEG, JPEG2000, and BPG. Despite of noticed performance improvement obtained by these learned methodologies, it would take times to use it in practices, particularly for mobile applications that require dedicated hardware accelerations.

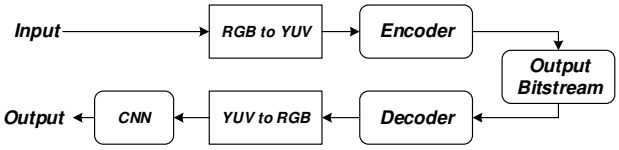

Figure 1. The flowchart of the proposed image compression framework, which is consisted of the emerging Versatile Video Coding (VVC) Intra Profile based image coding and a post-processing module using an image restoration network.

On the other hand, in-loop filters (e.g., deblocking, and/or Sample Adaptive Offset (SAO)) are incorporated in popular HEVC standard to reduce the compression artifacts for quality improvement. Inspired by this, a number of learning based methods are utilized as the post-processing module for compressed images to further improve the visual quality [6, 2, 8, 11]. These methods are focusing on eliminating the artifacts of the decoded images, which are caused by the lossy compression algorithms.

In this work, we propose a learning based image restoration network as the post-processing module for emerging VVC to further improve its reconstruction quality. VVC Intra coding is adopted and appropriate color conversion for RGB raws is enforced to use VVC compliant YUV source. Correspondingly, the decoded image in YUV space is transformed back to RGB signal prior to our post-processing network, shown in Fig. 1. We design the network in Fig. 2 using multi-scale spatial priors to effectively reduce compression artifacts in the decoded image.

Experimental results have demonstrated the effective

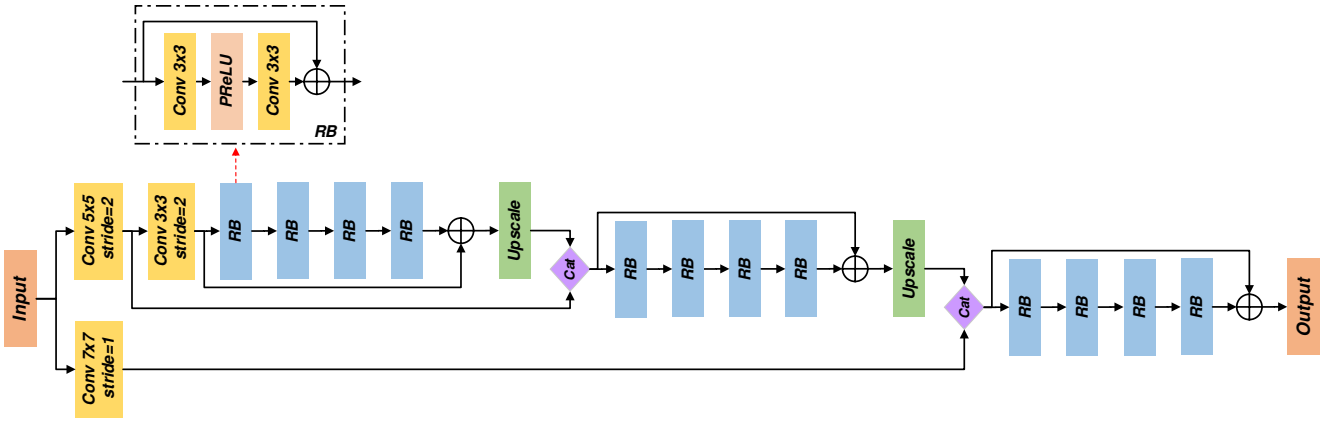

Figure 2. Pipeline of our learned image restoration network using multi-scale priors.

performance improvement of our network on top of the VVC Intra with about 0.3 dB PSNR gain for YUV 4:4:4, and about 0.5 dB PSNR gain for YUV 4:2:0 respectively at the same bit rate.

2. Image Restoration Neural Network

As depicted in Fig. 1, the input images in RGB format are transformed to another color space (e.g., YUV 4:4:4) using the well-known FFmpeg (https://www.ffmpeg.org/), to reduce the inter color redundancy for better compression and more bitrate reduction without visual quality degradation. This is mainly because that VVC Intra now accepts YUV sources, rather raw RGBs. YUV 4:4:4 samples are then encoded by VVC Intra to generate the compressed bitstream that will be decoded at the receiver. YUV to RGB conversion is involved prior to applying the image restoration based quality enhancement with better visual quality. The VVC is an emerging International Standard developed by the joint forces of ISO/IEC MPEG and ITU-T VCEQ experts, promising another 2× efficiency (e.g., 50% bitrate reduction at same quality) compared with the state-of-the-art HEVC.

2.1. Multi-Scale Spatial Priors

As shown specifically in Fig. 2, the image restoration network mentioned above is designed using multi-scale spatial priors. Different from [11], by setting the different stride sizes of the convolutional operations respectively, which can be seen clearly in the figure, the input image is resized into three scales. Scale-wise convolution kernel sizes are utilized to capture the multi-scale priors spatially, namely 3×3 for 1/16 of the original image, 5×5 for 1/4 of the original image and 7×7 for the original image. Such operation can extract features from different scales more precisely with suitable convolutional patch sizes which coincides with the variable-size Coding Unit (CU) idea utilized in video codecs to well exploit the regional content characteristics (i.e., rich texture area with small-size convolution and CU, and stationary background with large-size convolution and CU). Four modified Residual Blocks (RB in the figure) [3] with kernel size at 3×3 are applied at each scale for acquisition of high-dimensional features. We adopt 256 output channels for each convolutional layer at 1/16 of the original dimension, 128 channels at 1/4 of the original dimension and 64 channels at the original dimension. The skip connection operation which is used in the residual networks is also adopted between the start and the end positions of the residual blocks for better convergency results. The pixel-shuffle layer [10] is utilized to upsample the feature maps to next scale for concatenation with the maps generated by the previous convolutional layer using residual connections. With such architecture, spatial information of each scale can be fused together closely for final quality enhancement of the image, which helps restore the block-to-block correlation effectively.

2.2. Loss Function

The MSE is adopted as the loss function of the image restoration network for its positive correlation with PSNR. In addition, we also use $\mathcal{L}_1$ norm to replace the MSE for fine-turning, which achieves another improvement on top of the model pre-trained with MSE.

3. Experimental Studies

We train the image restoration network using the training dataset called DIV2K [1] with the images compressed by the intra coding filters of the VVC as the inputs and the original images as the labels. The input sources sampled at YUV 4:4:4 and YUV 4:2:0 are generated with fixed quantization parameters (QPs) respectively for training. Not that several QPs (e.g., 25, 30, 35, etc) are adopted as the variables instead to fit different segments of bit rate so that the

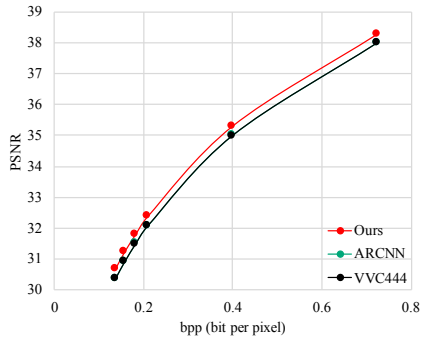

Figure 3. Compression performance on the Test Dataset, compared with VVC (with input source sampled at YUV 4:4:4)

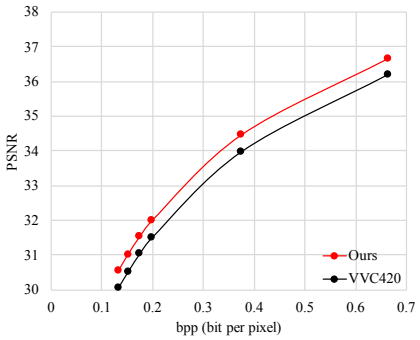

Figure 4. Compression performance on the Test Dataset, compared with VVC (with input source sampled at YUV 4:2:0)

network can learn the compression artifacts with certain QP accordingly.

3.1. Model Training

Our experiments are performed on a desktop with an i7-7700K CPU and a NVIDIA Quadro P5000 GPU. We choose PyTorch platform to implement the proposed model. The model is trained using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate is 10^{-4} initially, which is divided by 10 after 200 epochs. The batch size is set at 16 with the input size at 192×192 basically, which is cropped from the dataset randomly every time to avoid the overflow of the memory. It is worth to mention that the results can be better when we increase the crop size. It is mainly because the multi-scale architecture can benefit from the larger image size with more pixels.

We train the network using a transfer learning manner. Models of higher QPs are trained based on parameters from models of lower QPs. (e.g., network parameters at QP 22 is used to derive network models at QP 27). Such stepwise procedure leads to faster convergence of parameters and better results than training the network at different QPs independently.

3.2. Performance Evaluation

We evaluate our network on the Test Dataset P/M with 330 images totally released by the Computer Vision Lab of ETH Zurich, and compare with the VVC using the model correspondingly trained respectively. Fig. 3 shows the PSNR performance on the Test Dataset with the input sources of VVC sampled at YUV 4:4:4 and our network achieves 0.3 dB gains at each bit rate point and average 6.5% BD-Rate reduction over default VVC Intra. We also replace our image restoration network using the ARCNN [2] which is optimized using the same settings as our network, and the BD-Rate curve of the ARCNN is almost overlapped with the VVC Intra. It further evident the effectiveness of our image restoration network.

(a) (b) (c) (d)

Figure 5. Four image snapshots from the Test Dataset released by the Computer Vision Lab of ETH Zurich.

Figure 4 shows the PSNR performance on the Test Dataset for YUV 4:2:0 source and our network achieves 0.5 dB gains at each bit rate point and corresponding average 12.2% BD-Rate reduction.

Moreover, we select four typical images with different types and complex scenes which are usually difficult for efficient compression from the dataset, as shown in Fig. 5. The PSNR gains are achieved by 0.2 dB, 0.2 dB, 0.25 dB and 0.15 dB, respectively, when using our proposed network as the post-processing on top of VVC Intar. The BD-Rate has been respectively reduced by 4.35%, 4.03%, 4.56% and 2.99% against VVC with YUV 4:4:4 input, as illustrated in Fig. 6.

For the bitrate budget imposed by the Workshop and Challenge on Learned Image Compression (CLIC) of CVPR 2019, i.e., each encoded image can't exceed the bitrate 0.15 bpp (bits per pixel), four different QP levels (i.e., QP $\in$ 35, 36, 37, 38) are selected in Rate Distortion Opti-

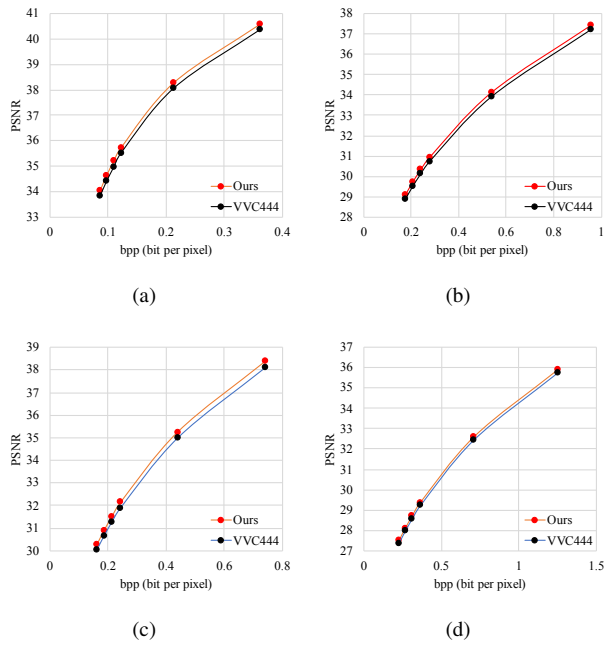

Figure 6. Compression performance for selected four images, compared with VVC with YUV 4:4:4 input.

mization (RDO). In the beginning, all images are encoded at the highest QP (QP 38) as the initial state. Then image with the maximum $\frac{PSNR_{i-1}-PNSR_i}{filesize_{i-1}-filesize_i}$ will be encoded at a lower QP iterativeely until the overall file size reaches to the file size limitation. The final results could be found on the leaderboard (http://www.compression.cc/leaderboard/) and our team name is NJUVisionPSNR.

4. Conclusion

In this work, we propose an image restoration network as the post-processing module combining with the intra coding profile of the VVC to further improve the quality of reconstructed image. Experiments demonstrate noticeable improvements in both subjective and objective quality measurement at the same bit rate. The designed restoration network has utilized multi-scale spatial priors to alleviate or eliminate compression artifacts induced by the traditional codecs. Given that our method is served as a post-processing module, it can be easily patched to existing video codecs for general applications. As for future studies, temporal prior is worth for exploration to further improve the restoration efficiency.

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 2

[2] C. Dong, Y. Deng, C. Change Loy, and X. Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584, 2015. 1, 3

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] J. Lainema, F. Bossen, W. Han, J. Min, and K. Ugur. Intra coding of the hevc standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1792–1801, Dec 2012. 1

[5] H. Liu, T. Chen, P. Guo, Q. Shen, X. Cao, Y. Wang, and Z. Ma. Non-local attention optimized deep image compression. *arXiv preprint arXiv:1904.09757*, 2019. 1

[6] M. Lu, M. Cheng, Y. Xu, S. Pu, Q. Shen, and Z. Ma. Learned quality enhancement via multi-frame priors for hevc compliant low-delay applications. *arXiv preprint arXiv:1905.01025*, 2019. 1

[7] D. Minnen, J. Ballé, and G. D. Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pages 10794–10803, 2018. 1

[8] S. Nah, T. Hyun Kim, and K. Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017. 1

[9] O. Rippel and L. Bourdev. Real-time adaptive image compression. *arXiv preprint arXiv:1705.05823*, 2017. 1

[10] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2

[11] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 1, 2

[12] D. Taubman and M. Marcellin. *JPEG2000 image compression fundamentals, standards and practice: image compression fundamentals, standards and practice*, volume 642. Springer Science & Business Media, 2012. 1