A Swin Transformer Based Restoration Scheme for VVC Compressed Images

Zhenchao Ma, Yixiao Wang, Hamid Reza Tohidypour, Panos Nasiopoulos, Victor C. M. Leung
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC, Canada
email: {zhenchaoma, yixiaow, htohidyp, panos, vleung}@ece.ubc.ca

***Abstract*—The Versatile Video Coding (VVC) standard is shown to significantly outperform the High Efficiency Video Coding (HEVC), the previous compression standard image/video codecs. More complex structures and advanced prediction techniques are behind this improved performance, leading to reduced visual artifacts. Deep learning-based image restoration algorithms have been proposed and are increasingly used for further reducing VVC generated artifacts. In this paper, we propose a Swin Transformer based image restoration model for VVC compression artifacts reduction that employs a self-attention mechanism to explore both global and local features to better understand the relation between existing and missing information. Performance evaluations showed that our proposed method outperforms existing state-of-the-art approaches yielding 0.884 dB quality improvement or 15.95% bitrate savings.**

Keywords-image restoration; VVC; vision transformer; multi scale window; artifacts reduction.

I. INTRODUCTION

Digital image and video compression standards played an important role in the rise of digital communications and entertainment technologies and are still of enormous importance in the emerging worlds of social media, Virtual Reality (VR) and metaverse. Lossy image compression approaches utilize the characteristics of the human visual system and its varying sensitivity to certain frequencies, brightness, contrast, and colors to achieve a high compression while guaranteeing acceptable visual image quality. The intrinsic characteristics of lossy image compression have increased the efficiency of sharing and viewing ultra-high resolution personal images, while inevitably introducing some undesired artifacts. The Versatile Video Coding (VVC) [1] is the latest generation of international coding standards jointly developed by the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG), and which achieves about a 50% bitrate reduction over its predecessor the High Efficiency Video Coding (HEVC) [2]. This improved performance is due to a series of new coding improvements, such as the Quad-Tree with nested Multi-type Tree (QTMT) structure of Coding Unit (CU) partition, Wide-Angle Intra Prediction (WAIP) and Matrix-based Intra Prediction (MIP).

Lately, with the evolution of convolutional neural networks, Convolutional Neural Networks (CNNs) based image restoration algorithms have been widely proposed for enhancing the quality of images compressed by VVC. Li *et al*. [3] proposed a convolutional neural network based filter which leverages auxiliary information to enhance image quality. Lu *et al*. [4] combined convolutional layers with multi-scale spatial priors to effectively reduce VVC caused artifacts. Bonnineau *et al*. [5] proposed multitask learning to perform both VVC quality enhancement and super-resolution. Different from the previous approaches that employ multiple neural networks, this method uses an optimized single shared network to achieve both tasks.

Lately, the Vision Transformer (ViT) network architecture has demonstrated improved performance on a variety of computer vision tasks such as image classification [6], semantic segmentation [7] and object detection [8]. Particularly, Swin Transformer proposed a shifted window scheme that limits the self-attention computation to non-overlapping local windows to improve computational efficiency [9]. SwinIR [10], an image restoration algorithm based on Swin Transformer, was developed for image denoising, image super-resolution, and achieved state of the art [10]. The same network was used for reducing visual artifacts of Joint Photographic Experts Group (JPEG) compressed images and showed to achieve improved performance compared to other existing approaches. However, VVC employs a complex and sophisticated partitioning structure and other advanced features that lead to better compression performance and better image quality which is much more difficult to improve.

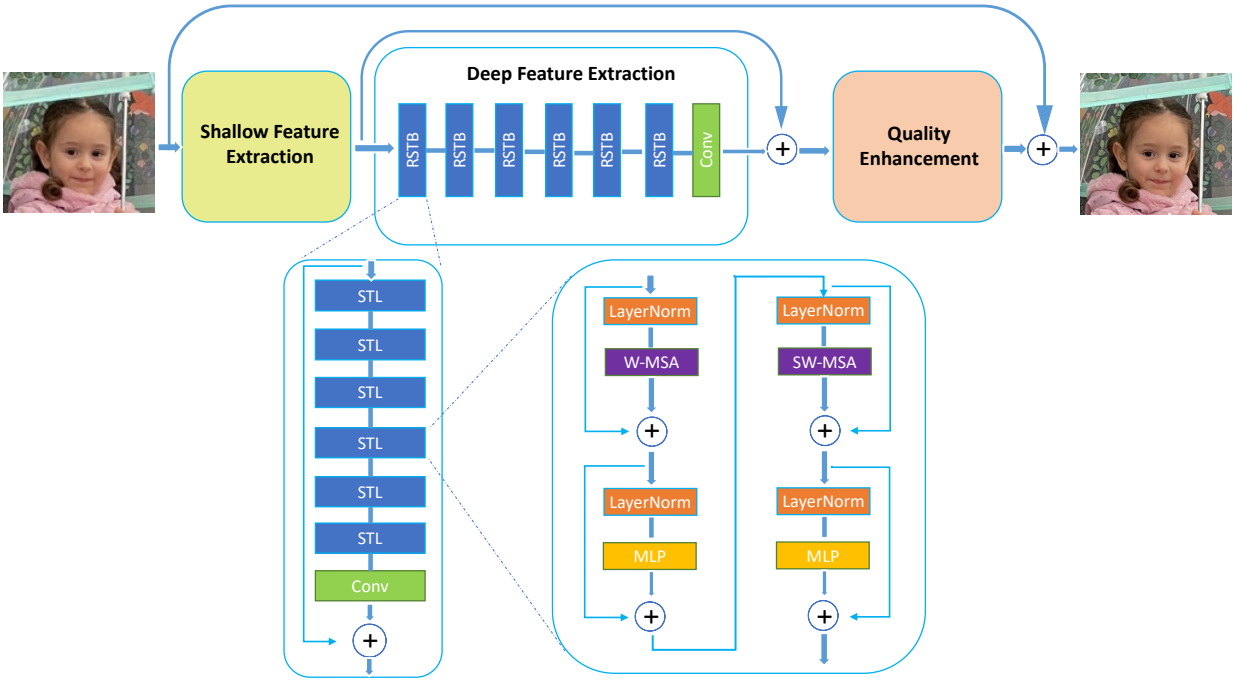
In this paper, we propose a Swin Transformer based image restoration model for reducing VVC compression artifacts. In our approach we propose a 16x16 pixels attention window which is shown to best capture the local features in images/frames compressed by the VVC complex tree structure coding unit architecture. Performance evaluations have shown that our proposed network outperforms existing state-of-the-art approaches, yielding 0.884 dB quality improvement or 15.95% bitrate savings.

The rest of the paper is structured as follows. Section II describes our proposed network. Section III presents the experimental results and discussion. Finally, Section IV concludes the paper.

II. PROPOSED METHOD

We propose a Swin Transformer based image restoration network for VVC compression artifacts reduction, which follows the architectural design outlined in [10]. Figure 1 shows the overall architecture, which consists of three main modules. The shallow feature extraction module uses a 3x3

 ISBN: 978-1-68558-072-8



(a) Residual Swin Transformer Block (RSTB) (b) Swin Transformer Layer (STL)

Figure 1. The architecture of the proposed deep learning model for reducing visual artifacts of VVC compressed frames.

convolutional layer to obtain the shallow features of the input image, which are then fed directly into the deep feature extraction module. The deep feature extraction model consists of 6 residual Swin Transformer blocks (RSTB) and a 3x3 convolutional layer. Each RSTB layer has six Swin Transformer Layers (STL), which involve normalization and a self-attention module that may focus on what is important in a local area. Self-attention is computed only within local windows that are arranged to divide the image evenly in a non-overlapping manner. Small size windows tend to extract local detailed information, while larger windows focus more on high level conceptual information. We evaluated different window sized for our task of removing VVC compression artifacts to determine the size that yields the best results. It is worth mentioning that for the case of improving JPEG compression artifacts, the best self-attention window size was shown to be equal to 7x7 pixels [10]. When the size of the window was increased to 8x8 pixels, a size that equals the size of the blocks used in JPEG, the performance dropped significantly. However, our analysis (details presented in the following section) showed that larger sizes are more effective for the case of VVC compressed images. In fact, size 16x16 yielded the best visual results. The reason for this may be that unlike JPEG that uses only 8x8 blocks, VVC uses a much more complex tree structure unit architecture with multiple and larger sizes, resulting in block boundaries very different than those of JPEG (which are always 8x8). It worth mentioning here that during the first stage, W-MSA (Window Multi-head Self-Attention) finds the relationship between pixels in the original image distribution, while the second stage, SW-MSA (Shift Window MSA) calculates the relationship of pixels when they are shifted. The Quality Enhancement module involves only a 3x3 convolutional layer to fuse the shallow and deep features together. Finally, the input image is added to the fused information of the quality enhancement layer to form the output reconstructed image.

In our model, we use the Charbonnier loss function [11] to improve the performance of our network:

$$\mathcal{L} = \sqrt{||I_{RHQ} - I_{HQ}||^2 + \varepsilon^2}\,, \qquad (1)$$

where I_{RHQ} is the restored image generated by our model, I_{HQ} is the corresponding ground-truth image, and ε is a constant that is empirically set to 10^{-3}.

III. Experimental Studies

A. Datasets

For training the image restoration model, we built our training dataset using a total of 8194 images from the DIV2K [12], Flickr2K [13] and WED [14] datasets. First, all the RGB images were converted to YUV 420 format. Then, we encoded these images using VVC's latest software VTM 18.2, at four different Quantization Parameters (QPs), 22, 27, 32 and 37. For our encoding, we followed the "encoder_intra_vvc" configuration recommended by the JVET General Test Conditions, where the "CUTsize" is set to 128 and the "InputChomaFormat" to 420. For each image, we ended up having four compressed images with different visual quality, one for each QP which were decompressed and converted to the RGB format. Finally, we generated four training datasets, each consisting of images with similar visual quality, i.e., each dataset corresponded to one QP. The original uncompressed images were used as the target images for the training phase.

 ISBN: 978-1-68558-072-8

TABLE I. PSNR (DB) COPMPARISON OF OUR METHODS AGAINST VVC COMPRESSION ON SET5 AND LIVE1 BENCHMARK DATASETS

		Methods			
Dataset	**Quality**	VVC PSNR	Ours (win = 4) PSNR (Δ%)	Ours (win = 8) PSNR (Δ%)	**Ours (win = 16) PSNR** (Δ%)
Set5	QP = 22	36.14	38.05 (+5.28%)	38.22 (+5.76%)	**38.24 (+5.81%)**
	QP = 27	33.86	35.13 (+3.75%)	35.27 (+4.16%)	**35.32 (+4.31%)**
	QP = 32	31.52	32.27 (+2.38%)	32.55 (+3.27%)	**32.66 (+3.62%)**
	QP = 37	29.51	29.95 (+1.49%)	30.31 (+2.71%)	**30.35 (+2.85%)**
LIVE1	QP = 22	37.04	38.78 (+4.69%)	38.93 (+5.10%)	**38.95 (+5.16%)**
	QP = 27	34.70	35.73 (+2.96%)	35.95 (+3.60%)	**35.99 (+3.72%)**
	QP = 32	32.06	32.74 (+2.12%)	32.92 (+2.68%)	**32.97 (+2.84%)**
	QP = 37	29.46	29.81 (+1.18%)	30.02 (+1.90%)	**30.06 (+2.04%)**

B. Model Training

We trained a total of 12 different restoration models. For each dataset (or each quality level QP), we trained 3 models, one with attention window 4x4, one with attention window 8x8 and one with window 16x16. Training was done on a compute cluster with 2 Intel Silver 4216 Cascade Lake CPUs, 2 NVIDIA V100 Volta GPUs and 32G of memory. The Adam optimizer was used with an initial learning rate of $2e^{-4}$, and a batch size of 4.

C. Performance Evaluation

We evaluated the performance of our models on two public test datasets, the set5 that consists of images with resolution varying from 288x288 to 512x512 and LIVE1 with two resolutions, 768x512 and 512x768 pixels. All the test data were compressed using the VTM 18.2 software at four QPs (22, 27, 32 and 37), with the same configuration parameters used for preparing our training datasets. Table I shows the objective performance of VVC and that of all our models in terms of PSNR. For each QP, we have a different model while the columns show the combination of our models using 3 different attention windows. First, we observe that for both datasets, Set5 and LIVE1, the models using attention window of 16x16 outperform those using 4x4 and 8x8 windows. Second, we observe that for both datasets, our model trained using frames compressed at QP=22 and using attention window 16x16 outperform the models trained with QPs = 27, 32 and 37. More specifically, the improvement in visual quality in terms of PSNR is 5.81% and 5.16% for the Set5 and LIVE1, respectively.

Figure 2 shows the original uncompressed image, the image compressed using VVC at QP=32 and the image reconstructed by our model (16x16 window and QP=32). The reason for using QP=32 is that the visual quality of images compressed at QP=22 is very high (PSNR of VVC = 36.14 dB). We observe from the zoomed in portion of the image (bottom left side) that our model improved the visual quality of the image resulting from VVC.

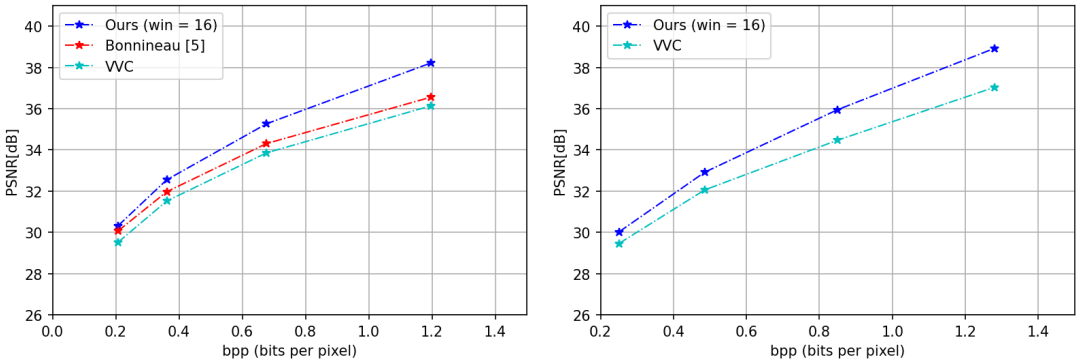
We also compare the performance of our method against the Bonnineau's state-of-the-art approach presented in [5]. Figure 3 shows the rate distortion plots for the Set5 dataset of our model trained with 16x16 window and all QPs, Bonnineau's model and original VVC encoder.

Table II shows the Bjøntegaard Delta rate (BD-rate) and Peak Signal Noise Ratio (BD-PSNR) using the piecewise cubic fitting from the above results for our model and the Bonnineau's relative to VVC. We observe that our approach outperforms Bonnineau's approach [5] by 15.95% on BD-Rate and 0.884 dB on BD-PSNR.

(a) Original Image (b) VVC at QP = 32 (c) Ours (win =16) at QP = 32

Figure 2. Visual compression (a) original image, (b) VVC (QP = 32), and (c) Ours (win = 16, QP = 32).

 ISBN: 978-1-68558-072-8



(a) (b)

Figure 3. Rate distortion plots for (a) Set5 dataset of our model trained with 16x16 window and all QPs, Bonnineau's model and original VVC encoder, and (b) LIVE1 dataset for our model and VVC.

TABLE II. BD-RATE AND BD-PSNR COMPARISION FOR OUR MODEL AND BONNINEAU MODEL RELATIVE TO VVC ON SET5

Dataset	Metrics	Bonnineau [5]	Ours (win = 16)
Set5	BD-Rate (%)	-11.48%	**-27.44%**
	BD-PSNR (dB)	0.458	**1.342**

TABLE III. BD-RATE AND BD-PSNR FOR OUR MODEL RELATIVE TO VVC ON LIVE1

Dataset	Metrics	Bonnineau [5]	Ours (win = 16)
LIVE1	BD-Rate (%)	-	**-20.35%**
	BD-PSNR (dB)	-	**1.129**

Table III shows the Bjøntegaard Delta rate (BD-rate) using the piecewise cubic fitting from the above results for our model relative to VVC on the LIVE1 dataset. We observe that our approach improves the visual quality of the VVC compressed images by 1.129 dB or saving the bitrate by 20.35%.

IV. CONCLUSION AND FUTURE WORK

In this paper, we propose a Swin Transformer based image restoration model for VVC compression artifacts reduction that employs a self-attention mechanism to explore both global and local features to better understand the relation between existing and missing information. Compared to other models, our model with a window size of 16x16 is shown to best capture local features in images/frames compressed by the VVC complex tree-structured coding unit architecture and achieve state-of-the-art performance for all four different QPs on two benchmark datasets. Performance evaluations showed that our proposed model achieves an average of 27.44% and 20.35% BD-Rate reduction over the original VVC standard on two benchmark datasets.

ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC – PG 11R12450), and TELUS (PG 11R10321).

REFERENCES

[1] B. Bross *et al.*, "Overview of the Versatile Video Coding (VVC) Standard and its Applications," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736-3764, Oct. 2021, doi: 10.1109/TCSVT.2021.3101953.

[2] B. Bross, J. Chen, J. -R. Ohm, G. J. Sullivan, and Y. -K. Wang, "Developments in International Video Coding Standardization After AVC, With an Overview of Versatile Video Coding (VVC)," in *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463-1493, Sept. 2021, doi: 10.1109/JPROC.2020.3043399.

[3] Y. Li, L. Zhang, and K. Zhang, "Convolutional Neural Network Based In-Loop Filter For VVC Intra Coding," *2021 IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, USA, 2021, pp. 2104-2108, doi: 10.1109/ICIP42928.2021.9506027.

[4] M. Lu, T. Chen, H. Liu, and Z. Ma, "Learned Image Restoration for VVC Intra Coding," in *Proceedings of CVPR Workshops*, 2019, pp. 1–4.

[5] C. Bonnineau, W. Hamidouche, J.-F. Travers, N. Sidaty, and O. Deforges, "Multitask Learning for VVC Quality Enhancement and Super-resolution," in *Proceedings of Picture Coding Symposium (PCS)*, 2021, pp. 1-5.

[6] 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*, 2016, pp. 770-778.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention*. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Procedings of the 28th International Conference on Neural Information Processing Systems*, vol. 1, December 2015, pp. 91-99.

[9] Z. Liu *et al.*, "Swin transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 10012–10022.

[10] J. Liang *et al.*, "SwinIR: Image Restoration using Swin Transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, October 2021, pp. 1833–1844.

[11] W. -S. Lai, J. -B. Huang, N. Ahuja and M. -H. Yang, "Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks," in *IEEE Transactions on Pattern Analysis*

 ISBN: 978-1-68558-072-8

and Machine Intelligence, vol. 41, no. 11, pp. 2599-2613, 1 Nov. 2019, doi: 10.1109/TPAMI.2018.2865304.

[12] R. Timofte *et al*., "NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Honolulu, HI, USA, 2017, pp. 1110-1121.

[13] B. Lim, S. Son, H. Kim, S. Nah and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Honolulu, HI, USA, 2017, pp. 1132-1140.

[14] K. Ma et al., "Waterloo Exploration Database: New Challenges for Image Quality Assessment Models," in *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004-1016, Feb. 2017.

 ISBN: 978-1-68558-072-8