Low Bit-rate Image Compression based on Post-processing with Grouped Residual Dense Network

Seunghyun Cho*, Jooyoung Lee, Jongho Kim, Younhee Kim
Broadcasting and Media Research Laboratory
Electronics and Telecommunitcaions Research Institute
218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Korea
{shcho, leejy1003, pooney, kimyounhee}@etri.re.kr

Dong-Wook Kim, Jae Ryun Chung, Seung-Won Jung
Department of Multimedia Engineering
Dongguk University
30, Pildong-ro 1-gil, Jung-gu, Seoul, 04620, Korea
{spnova12, wjdwofus1004, swjung83}@gmail.com

Abstract

In this paper, an image compression method implemented for CVPR 2019 Challenge on Learned Image Compression (CLIC) is introduced. It is designed to satisfy both requirements of image compression, "higher compression ratio" and "better quality", at the same time. To this end, a neural network based image quality enhancement is incorporated into the most recent traditional image/video coding technique. The decoders, ETRIDGU, ETRIDGUlite, and ETRIDGUfast, which implement the proposed image compression method are designed to have different degrees of complexity and compression efficiency. ETRIDGU, which provides the highest compression efficiency, is reported to achieve the 2nd highest PSNR in the lowrate track of CLIC. ETRIDGUlite, which compromises between the compression efficiency and the complexity, is reported to be the fastest one among the decoders with high mean opinion score (MOS) in the same track.

1. Introduction

Image compression technology is everywhere in our lives. It is inevitably used in almost all online services such as information retrieval, education, shopping, and distribution services. In addition, individual users produce massive amounts of image data every day with advanced mobile devices, and consume them through online social network services. Although widely used image compression technologies such as JPEG, WebP, and BPG already exist, it is still demanded to improve the image compression efficiency to reduce data traffic and improve online service quality.

ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG) have constituted the Joint Video Experts Team (JVET) to develop a new video coding standard called Versatile Video Coding (VVC). VVC is the next generation of video coding technology that is the successor to the existing High Efficient Video Coding (HEVC).

On the other hand, compression artifacts such as blockiness, ringing, and contouring appear in reconstructed images in lossy image/video compression based on block-based prediction, transform, and quantization. To eliminate these compression artifacts, VVC employs three in-loop filters; deblock filter, sample adaptive offset (SAO), and adaptive loop filter (ALF). However, at low bitrates, it is very difficult to completely remove compression artifacts through these traditional filters. Through successful studies over the last few years, it has been confirmed that image/video restoration using deep neural network (DNN) can be a breakthrough in solving this problem. JVET is exploring its applicability to the VVC standard by studying the performance and complexity of neural network based post- and in-loop filters through Core Experiment (CE) [6].

In this work, we have attempted to improve image compression efficiency by incorporating the intra frame coding of VVC with grouped residual dense network (GRDN) [5] as a post-processing filter. As far as we know, this is a combination of the best-performing traditional video coding algorithm that can also be used for image compression

*Corresponding author

(a) Original (b) JPEG (26.23 dB, 0.12 bpp)

(c) BPG (29.56 dB, 0.11 bpp) (d) Proposed (30.83 dB, 0.10 bpp)

Figure 1. Comparison of image qualities between original, JPEG, BPG, and the proposed. The differences are more evident when zoomed in on the electronic version of this paper.

and one of the state-of-the-art DNN-based filters for image restoration.

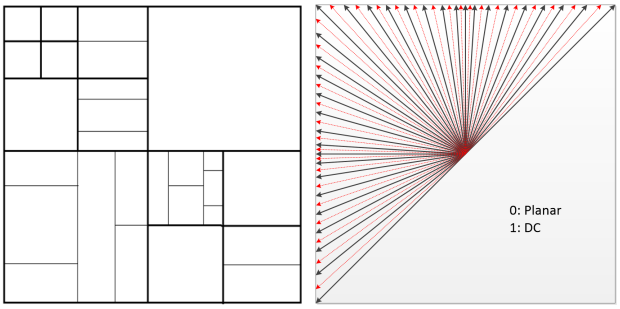
Fig. 1 compares qualities of compressed images from JPEG, BPG, and the proposed method to the original image. For the purpose of comparison, an image was selected from the CLIC testset and then compressed using each method. While Fig. 1 is showing only a part of the selected image, both PSNR and bpp are measured for the entire region. The proposed method gives the highest PSNR and the best visual perception although it is compressed at the lowest bpp.

2. Pursuit of higher compression ratio and better image quality

2.1. Intra frame coding in VVC

VVC extended the maximum Coding Tree Unit (CTU) size and changed the block coding structure of those used for HEVC to achieve higher coding efficiency from larger coding blocks with more flexible partitioning; it supports 256x256 CTUs that can split using quadtree plus binary tree (QTBT). Fig. 2(a) visualizes an example of QTBT block partitioning of VVC.

The maximum transform size has also been extended to 64×64. Mode dependent non-separable secondary transforms and explicit multiple core transform for intra frame coding are adopted in VVC. In addition, the following coding tools are introduced for more accurate intra prediction:



(a) (b)

Figure 2. Features of VVC for intra frame coding: (a) Example of QTBT block partitioning, (b) 67 intra prediction modes; both (a) and (b) are copied from [4]

- 67 intra prediction modes; visualized in Fig. 2(b)
- Wide-angle intra prediction for non-square blocks
- Block size and mode dependent 4 tap interpolation filter
- Position dependent intra prediction combination
- Cross component linear model intra prediction
- Multi-reference line intra prediction
- Intra sub-partition

Thanks to above changes, it is reported that VVC outperforms HEVC in terms of Bjontegaard (BD) rate gain for Y-PSNR by 21.23% when the latest VVC test model (VTM) [2] is compared to the latest HEVC test model (HM) model (HM) [1] for intra frame coding [3].

2.2. Coding artifact reduction with GRDN

GRDN [5] is employed in the proposed image compression method for effective removal of compression artifacts. It showed the best performance in the NTIRE 2019 Real Image Denoising Challenge - Track 2: sRGB. Fig. 3 visualizes the architecture of GRDN used for this work. A GRDN consists of cascading grouped residual dense blocks (GRDBs) followed by a convolutional block attention module (CBAM) [8]. To enable effective learning of a deeper and wider network, the proposed GRDN employs down-sampling and up-sampling, respectively before and after passing through GRDB, besides a global skip connection. In the bottom left of Fig. 3, a GRDB consists of a number of residual dense blocks (RDBs) [9] followed by a concatenation and an 1×1 convolution layer. Not only are the GRDBs

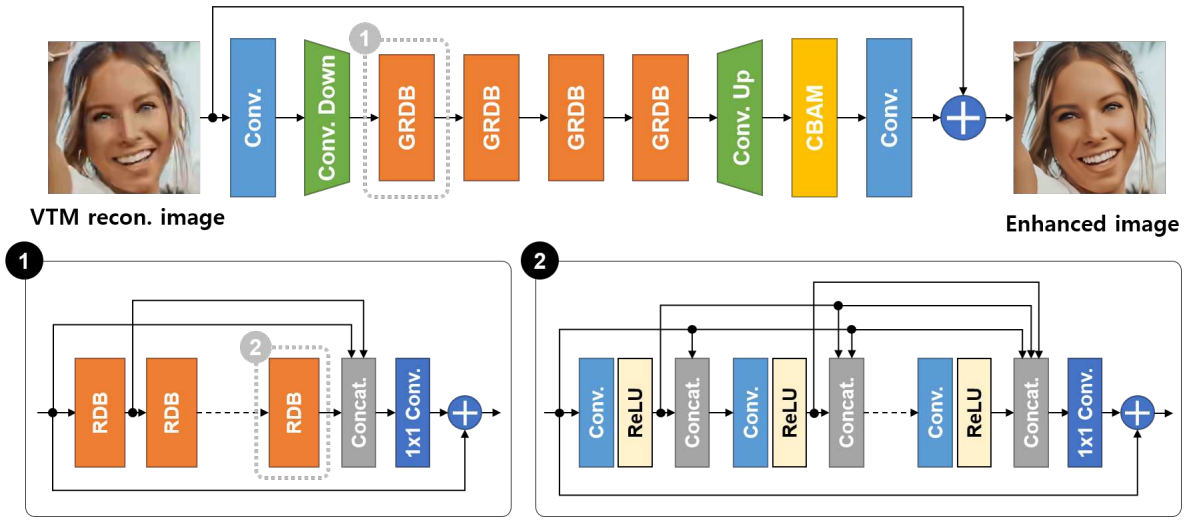


Figure 3. The architecture of the GRDN used for compression artifact reduction. This figure is redrawn from its original version [5].

	ETRIDGU	**~lite**	**~fast**
Down-sampling	×1/2	×1/2	×1/4
# of RDBs	16	16	4
# of 3×3 conv.	8	8	2
# of filters	80	64	8
Patch size	96×96	48×48	48×48
PSNR (dB)	**31.18**	31.16	30.82
MS-SSIM	**0.9566**	0.9565	0.9535
MOS	-	**3.6360**	-
Dec. time (sec.)	2532.30	1891.83	**930.95**
Dec. size (MB)	487.86	312.53	**1.07**

Table 1. Comparison of specifications and test results between the submitted models.

connected in series, their output features are also concatenated together before they are passed to the 1×1 convolution. As shown in the bottom right of Fig. 3, an RDB has densely connected 3×3 convolutions each of which is activated by a rectified linear units (ReLU). Finally, a CBAM introduced in [8] is employed to improve the compression artifact removal performance of GRDN just before the output convolutional layer. Further information on GRDN can be found in [5].

2.3. Implementation

To implement the proposed image codec, we integrated the VVC Test Model (VTM) [2] version 4.0 with GRDN. For image encoding, the original image is first converted its color format from RGB to YUV420 and fed into the VTM for intra frame coding. Then, the reconstructed image obtained from the VTM is converted its color format from YUV420 back to RGB and transferred to a GRDN. Finally, compression artifact reduction through the GRDN is performed to obtain the final result image. A GRDN is used for one or more quantization parameter (QP) values in VVC; two images encoded with different QPs can be processed by the same GRDN.

Three GRDNs with different complexities are implemented and trained using PyTorch-1.0.1 running on two Intel Intel Xeon CPUs E5-2643 v4 @ 3.40 GHz with 64 GB of DDR4 and four NVIDIA 1080ti GPUs. Table 1 shows implementation details of the GRDN used for each submitted decoder, ETRIDGU, ETRIDGUlite, and ETRIDGUfast. Relative comparisons among the three decoders for required amounts of memory and complexities can be made from Table 1. They are designed to have four GRDBs as depicted in Fig. 3. ETRIDGUfast performs 1/4 down-sampling to speedup GRDB processing while the others performs 1/2 down-sampling. For the same purpose, ETRIDGUfast employs only four RDBs with two 3×3 convolutional layers each of which has eight output channels. ETRIDGU and ETRIDGUlite use 16 RDBs with eight 3×3 layers each of them has 80 and 64 output channels, respectively.

3. Experiments

For training the GRDN, 1633 CLIC training images from Dataset P (professional) and M (mobile) and 118,287 images of Microsoft COCO training dataset [7] are used together. Original images were first encoded through VTM-4.0 [2] and then split into non-overlapping image patches and used for GRDN training. As shown in Table 1, 96×96 patch size is used for training ETRIDGU while 48×48 is used for training ETRIDGUlite and ETRIDGUfast; 96×96 patches result in slightly higher PSNR than 48×48 patches at the cost of longer training time. The initial learning rate

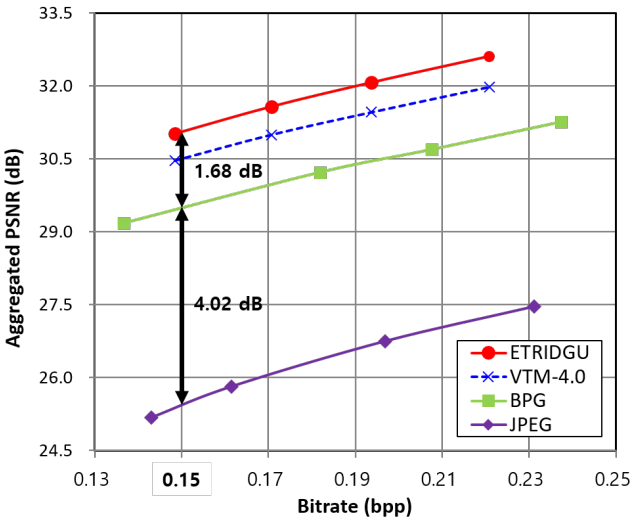
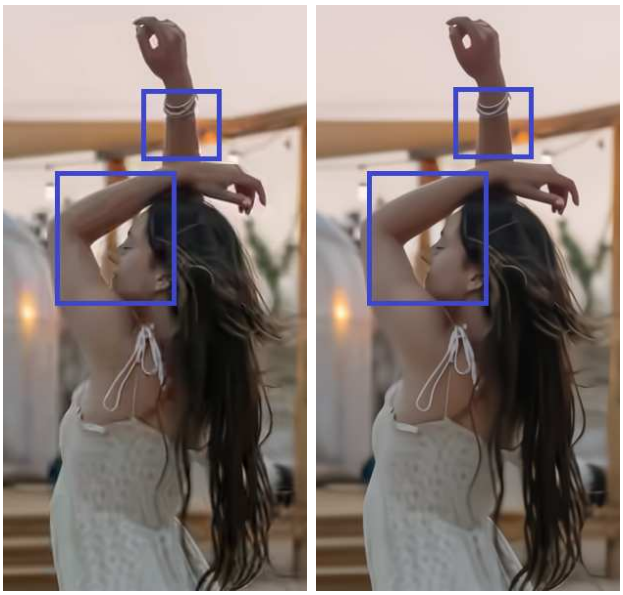


Figure 4. R-D curve comparisons between ETRIDGU, VTM-4.0, BPG, and JPEG.



(a) VTM-4.0 recon. (32.97dB)

(b) Enhanced by GRDN (33.31 dB)

Figure 5. Image quality enhancement for a test image encoded in 0.33 bpp. The differences are more evident when zoomed in on the electronic version of this paper.

for training GRDN was set to 0.0001 and decayed by half for every 5 epochs after first 20 epochs while the training was conducted up to 50 epochs.

Table 1 shows the test results of the submitted decoders in terms of aggregated PSNR, average MS-SSIM, MOS, and decoding time. The test results are measured for the 330 images of CLIC test dataset which are encoded at 0.15 bpp. ETRIDGU provides the highest coding efficiency at a cost of the longest decoding time, and ETRIDGUfast provides the fastest decoding time at a cost of the lowest PSNR. ETRIDGUlite can be regarded as a tradeoff between coding efficiency and decoding time.

Fig. 4 compares the rate-distortion (R-D) curve of ETRIDGU for the CLIC test dataset with those of BPG and JPEG, based on aggregated PSNRs measured at a bitrate range including 0.15 bpp. ETRIDGU results in 1.68 dB and 5.7 dB higher PSNR compared to BPG and JPEG, respectively. In addition, Fig. 4 also shows the RD-curve of VTM-4.0 with the blue dashed line. ETRIDGU results in 0.6∼0.7dB higher PSNRs compared to VTM-4.0 over the bitrate range; this improvement comes from the compression artifact reduction by the GRDN.

For the subjective image quality evaluation of the proposed image compression method, Fig. 5 compares the image obtained from ETRIDGU with the image obtained from VTM-4.0 [2] without GRDN. Fig. 5(a) and Fig. 5(b) show the same area of the two compressed images. The original image was selected from the CLIC test dataset and then compressed respectively at a bitrate of 0.33 bpp. Looking at the blue boxes on Fig. 5(a), contouring and blocky artifacts are noticeable, even though the experimental image is encoded to have a relatively high bpp. On the other hand, in Fig. 5(b), it can be seen that these artifacts are removed so that they are not visually perceived.

4. Conclusion

In this paper, a DNN-based image compression method submitted to the CLIC 2019 lowrate compression track is introduced. The proposed method efficiently removes the coding artifact of a VVC intra-coded image by post-processing through GRDN. This ensures a high image compression ratio and improved image quality in PNSR at the same time. Implementation details and experimental results of the three different decoders implemented using the proposed method are described in this paper. Among them, one with the best image compression performance achieved the 2nd place in the highest PSNR while another one with lower complexity took the 1st place in the fastest decoder providing high MOS in the CLIC 2019.

5. Acknowledgement

This work was supported by Institute for Information and communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) 2017-0-00072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media.

References

[1] "High efficiency video coding reference software version 16.20 (HM-16.20)". https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.20/, 2018.

[2] "Versatile video coding reference software version 4.0 (VTM-4.0)". https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/tags/VTM-4.0/, Feb. 2019.

[3] F. Bossen, X. Li, A. Norkin, and K. Sühring. Ahg report: Test model software development (AHG3), JVET-N0003, Geneva, CH, Mar. 2019.

[4] J. Chen, Y. Ye, and S. H. Kim. Algorithm description for versatile video coding and test model 4 (VTM 4), JVET-M1002, Marrakech, MA, Jan. 2019.

[5] D.-W. Kim, J. R. Chung, and S.-W. Jung. GRDN:Grouped residual dense network for real image denoising and gan-based real-world noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. submitted.

[6] Y. Li, S. Liu, and K. Kawamura. Description of core experiment 13 (CE13): Neural network based filter for video coding, JVET-M1033, Marrakech, MA, Jan. 2019.

[7] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, Sept. 2014.

[8] S. Woo, J. Park, J.-Y. Lee, and S. K. In. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[9] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018.