Quality Enhancement of VVC Intra-frame Coding based on HGRDN

Ting Fu, Zhengxue Cheng, Jiapeng Hu, Li Guo,
Shihao Wang, Xiongxin Zhao, Dajiang Zhou, Yang Song
Ant Group, Xihu District, Hangzhou, China.
{ft224739, zhengxue.czx, libby.hjp,
li.gl, shihao.wsh, xiongxin.zxx, dajiang.zdj, zhaoshan.sy}@antgroup.com

Abstract

This paper presents a detailed description on our submitted method ANTxNN_PSNR to Workshop and Challenge on Learned Image Compression (CLIC) 2021. Our method mainly incorporates the Enhanced Spatial Attention Block (ESA) to previous hierarchical grouped residual dense networks (HGRDN) as a post-processing quality enhancement for VVC intra-frame coding. Besides, we use a combination of multiple perceptual losses in RaLSGAN, for perceptual quality enhancement. Experimental results have demonstrated that our approach achieves 30.137dB, 32.473dB, 35.307dB in terms of PSNR at the rate constraint of 0.075bpp, 0.15bpp and 0.30bpp on CLIC validation dataset, respectively.

1. Introduction

Image compression is a fundamental technique to realize efficient data transmission and data storage. After great efforts of many decades, conventional compression algorithms have been finalized into standards and are in a wide use in industries. Recently, Joint Video Experts Team (JVET) has released a new-generation video coding standard named Versatile Video Coding (VVC/266) [1], which has just been finalized in July 2020. VVC/266 is expected to provide around 50% bit-rate saving at the same subjective visual quality over its predecessor, High Efficiency Video Coding (H.265/HEVC)[2].

However, due to the hand-crafted designs of conventional codecs, high compression ratio inevitably results to visible artifacts such as blocking artifacts, ringing effects and blurring. Recent research has started to apply deep learning techniques to address the issue of artifact reduction and quality enhancement. Yu [3] designed an AR-CNN to reduce the coding artifacts and showed a rate-distortion improvement in terms of PSNR and SSIM. In [4], hierarchical grouped residual dense network (HGRDN) architecture is proposed to efficiently remove artifacts from VVC intra coding. Our method further improves the perceptual quality and PSNR on the basis of HGRDN by incorporating the Enhanced Spatial Attention Block (ESAB) [5] to HGRDN. We use the network in [6] as our baseline and the main revisions of this paper can be summarized as follows:

- In the architecture of GRDB [6], features are concatenated, while we use dense connection to preserve all the feature-maps in preceding layers. Dense aggregation is flexible for multi-level feature learning [7]. In our experiment, using dense GRDB can improve PSNR.
- We integrate ESA blocks at the end of the residual blocks to force the features to put more attention on the regions of interest for quality enhancement.
- We use a combination of multiple perceptual losses in GAN learning, and each loss function provides an unique perspective on GAN-based perceptual quality enhancement.

2. Proposed Method

2.1. Network and Loss

In our proposed framework, a raw image is encoded using VVC intra coding and then the network based quality enhancement process is followed. The networks used in the quality enhancement are trained toward the coding artifact reduction.

The network architecture is shown in Fig. 1. The HGRDN mainly consists of several grouped residual dense connections (GRDB), a down-sampling layer, a up-sampling layer, and a convolutional block attention module (CBAM) [8] layer. The GRDB is consisted of four residual dense block (RDB). In order to make the residual features to be focused on spatial contents of key importance, we utilize the enhanced spatial attention (ESA) block, referring to [5]. The ESA mechanism works at the end of the residual dense block to force the features to put more attention on the regions of interest.

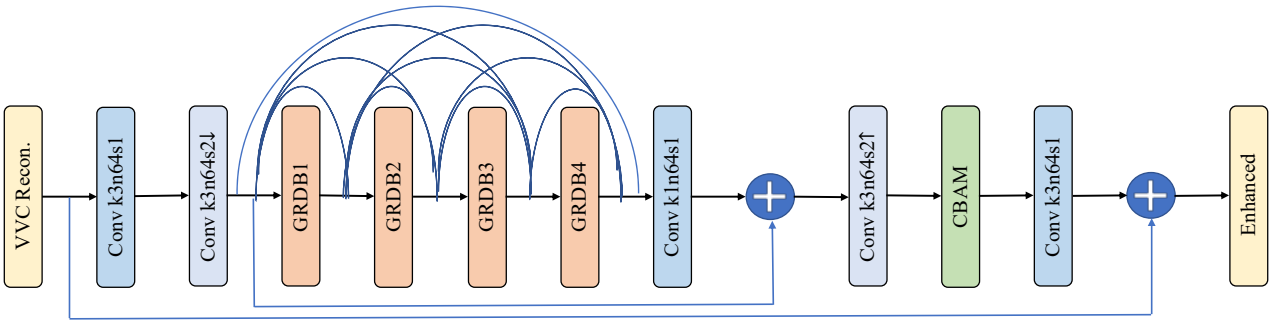

Figure 1: The HGRDN architecture, where k, n, s represent the size of convolution kernel, the number of feature maps and the size of convolution stride, respectively. We use dense HGRDB here. The GRDB module is depicted in Fig. 2.

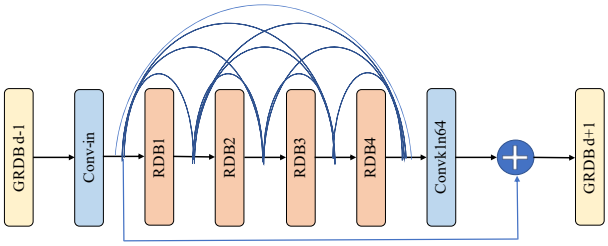

Figure 2: The dense GRDB architecture. The structure of RDB is exactly same as the structure of GRDB, except that the RDB is replaced by ESA module. ESA module is depicted in Fig. 3

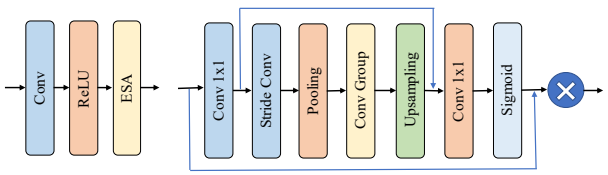

Figure 3: Left: The enhanced spatial attention (ESA) block. Right: Details of the ESA mechanism.

To train our quality-enhancement network, the loss function is defined as follows:

$$\mathcal{L}_p = \mathbb{E}\left[\|(\hat{x}_i) - (x_i)\|\right] \tag{1}$$

where $\hat{x}_i$ and x_i are compressed images by the VVC encoder and uncompressed raw images, respectively. $E[\cdot]$ represents the expectation that has applied to the batch data. Instead of mean square errors, we use L1-norm to be consistent with [6].

Meanwhile, to improve perceptual quality, we also optimize our enhanced HGRDN model using RaLSGANs, referring to [9]. The loss function will be accordingly changed to:

$$\begin{aligned}\mathcal{L}_g =& \alpha\mathcal{L}_{\text{charbonnier}} + \beta\mathcal{L}_{\text{vgg}} + \gamma\mathcal{L}_{\text{gram}} + \delta(\\ & \mathbb{E}[(d(\hat{\boldsymbol{x}}) - \mathbb{E}[d(\boldsymbol{x})] - 1)^2] + \mathbb{E}[(d(\boldsymbol{x}) - \mathbb{E}[d(\hat{\boldsymbol{x}})] + 1)^2]) \\ \mathcal{L}_d =& \mathbb{E}[(d(\boldsymbol{x}) - \mathbb{E}(d(\hat{\boldsymbol{x}})) - 1)^2] + \mathbb{E}[(d(\hat{\boldsymbol{x}}) - \mathbb{E}(d(\boldsymbol{x})) + 1)^2]\end{aligned}$$

where $\mathcal{L}_{\text{charbonnier}}$ is a Charbonnier loss between enhanced images and raw images. $\mathcal{L}_{\text{vgg}}$ denotes the L2-norm between features from VGG-19 model. $\mathcal{L}_{\text{gram}}$ represents the gram matrix of feature layers multiplied on transposed self and features are also from VGG-19 model. Other parts are corresponding to Relativistic average Least Squares GANs and $d(\cdot)$ denotes the output of discriminator .

2.2. Implementation Details

To implement the proposed framework, we integrated the VVC Test Mode (VTM) [10] version 10.1 with the optimized HGRDN. The Adam [11] optimizer is employed with β_1 = 0.9 and β_2 = 0.999. We initialize the learning rate to 1×10^{-4} and halve over every 50000 iterations. Total iteration is 300k.

The training dataset we used is a 2K resolution high-quality *DIV2K* dataset [12] with 800 images. The original images are first converted into YUV420 and encoded by VTM with all intra configuration setting. For the targeted low bit-rate (0.075 bpp) network training, the quantization parameter (QP) is ranging from 42 to 41. For the targeted medium bit-rate (0.15 bpp) network training, the QP is ranging from 38 to 36. For the targeted high bit-rate (0.3 bpp) network training, the QP is ranging from 32 to 30. Next, data augmentation with random $90°$, $180°$ rotations and horizontal flips is performed to these reconstructed images. After data augmentation, the reconstructed images are converted again into RGB format. Then image patches with size $96 \times 96 \times 3$ are randomly cropped from these samples and fed into the network.

3. Results

In this section, experimental results and ablation studies are presented. We have shown the results of our submitted methods and HGRDN method in [6] on CLIC validation dataset in Table. 1. Experimental results in Table. 1 are optimized by loss function L_p only. Table. 2 shows the results optimized by loss function $\mathcal{L}_\mathrm{g}$ and $\mathcal{L}_\mathrm{d}$. The first three rows in Table. 2 is optimized by loss function $\mathcal{L}_\mathrm{g}$ and $\mathcal{L}_\mathrm{d}$ and the last three rows in Table. 2 is optimized by loss function L_p. Besides, we conducted an experiment to find out best setting of our proposed method. Table. 3 shows the experimental results on different settings, including the number of GRDB, the architecture of GRDB, number of RDB, and image patch size. The rate we used in ablation study is 0.15bpp. As shown in Table. 3, the dense GRDB architecture is better than merge GRDB (i.e. concatenated features). Meanwhile, deepening the network does not significantly improve the quality.

Table 1: Results of our submitted methods *ANTxNN_PSNR* using L_p.

Model	PSNR	Rate (bpp)
ANTxNN_PSNR	35.307	0.30
ANTxNN_PSNR	32.473	0.15
ANTxNN_PSNR	30.137	0.075
HGRDN	35.255	0.30
HGRDN	32.456	0.15
HGRDN	30.108	0.075
VTM10.0	35.145	0.30
VTM10.0	32.356	0.15
VTM10.0	29.989	0.075

Table 2: Results of our submitted methods *ANTxNN_PSNR* using $\mathcal{L}_\mathrm{g}$ and $\mathcal{L}_\mathrm{d}$.

Loss	MS-SSIM	FID	Rate (bpp)
$\mathcal{L}_g, \mathcal{L}_d$	0.974	161.698	0.30
$\mathcal{L}_g, \mathcal{L}_d$	0.953	182.386	0.15
$\mathcal{L}_g, \mathcal{L}_d$	0.927	201.466	0.075
$\mathcal{L}_p$	0.979	212.613	0.30
$\mathcal{L}_p$	0.965	237.817	0.15
$\mathcal{L}_p$	0.942	276.100	0.075

Table 3: Ablation studies.

C_{GRDB}	N_{GRDB}	N_{RDB}	patch size	PSNR
merge	4	4	96	32.455
dense	4	4	96	32.473
dense	8	4	96	32.396
dense	4	8	96	32.474
dense	4	4	128	32.470

4. Conclusion

This paper presents a detailed description on our submitted method ANTxNN_PSNR to Workshop and Challenge on Learned Image Compression (CLIC) 2021. Our method mainly incorporates the Enhanced Spatial Attention Block (ESA) to previous hierarchical grouped residual dense networks (HGRDN) as a post-processing quality enhancement of VVC intra-frame coding. Besides, we use a combination of multiple perceptual losses in GAN for perceptual quality enhancement. Experimental results can further improve PSNR compared to HGRDN.

References

[1] G. J. Sullivan and J. R. Ohm, *"Versatile video coding Towards the next generation of video compression"*, Picture Coding Symposium, Jun. 2018. 1

[2] J. Brandenburg et al., "Towards Fast and Efficient VVC Encoding," 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 2020, pp. 1-6, doi: 10.1109/MMSP48831.2020.9287093. 1

[3] Yu, Ke, et al. "Deep convolution networks for compression artifacts reduction." arXiv preprint arXiv:1608.02778 (2016). 1

[4] Cho S, Kim D W, Jung S W. Quality enhancement of VVC intra-frame coding for multimedia services over the Internet[J]. International Journal of Distributed Sensor Networks, 2020, 16(5): 1550147720917647. 1

[5] Liu J, Zhang W, Tang Y, et al. Residual feature aggregation network for image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2359-2368. 1

[6] Y. Kim, S. Cho, J. Lee, S-Y Jeong, J. S. Choi, J. Do, *"Towards the Perceptual Quality Enhancement of Low Bit-rate Compressed Images"*, CVPR Workshop CLIC, 2020. 1, 2, 3

[7] Zhu Z, Bian Z P, Hou J, et al. When Residual Learning Meets Dense Aggregation: Rethinking the Aggregation of Deep Neural Networks[J]. arXiv preprint arXiv:2004.08796, 2020. 1

[8] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19. 1

[9] S. Zhang, D. Cheng, D. Jiang, Q. Kou, *"Least Squares Relativistic Generative Adversarial Network for Perceptual Super-Resolution Imaging"*, IEEE Access, Oct. 21, 2020. 2

[10] VVC Official Test Model VTM, https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM. 2

[11] D. P. Kingma and J. Ba, *"Adam: A method for stochastic optimization"*, arXiv:1412.6980, pp.1-15, Dec. 2014. 2

[12] Agustsson, Eirikur and Timofte, Radu, *"NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study"*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017. 2