Multi-scale Grouped Dense Network for VVC Intra Coding

Xin Li, Simeng Sun *, Zhizheng Zhang, Zhibo Chen [†]

CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China

{lixin666, smsun20, zhizheng} @mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

Versatile Video Coding (H.266/VVC) standard achieves better image quality when keeping the same bits than any other conventional image codec, such as BPG, JPEG, and etc. However, it is still attractive and challenging to improve the image quality with high compression ratio on the basis of traditional coding techniques. In this paper, we design the multi-scale grouped dense network (MSGDN) to further reduce the compression artifacts by combining the multi-scale and grouped dense block, which are integrated as the post-process network of VVC intra coding. Besides, to improve the subjective quality of compressed image, we also present a generative adversarial network (MSGDN-GAN) by utilizing our MSGDN as generator. Across the extensive experiments on validation set, our MSGDN trained by MSE losses yields the PSNR of 32.622 on average with teams "IMC" and "haha" at the bit-rate of 0.15 in Low-rate track. Moreover, our MSGDN-GAN could achieve the better subjective performance.

1. Introduction

Image/video compression algorithms such as JPGE, JPEG2000 and BPG are of practical importance for multimedia data storage and transmission. As the next generation of video coding technology, Versatile Video Coding (H.266/VVC) standard are developed by Video Coding Experts Group (VCEG) of ITU-I and Moving Picture Experts Group (MPEG) of ISO/IEC, which achieves up to 30% bitrate reduction compared with H.265/HEVC. However, various coding artifacts, such as blocking and blurring, are still serious when the compression ratio gets higher. To further eliminate compression artifacts, some CNN-based filtering methods [12, 13] have been used in in-loop filters of VVC.

With the development of deep neural network, learning based image compression algorithms [1, 7, 8, 2, 16, 3, 4]

---

*Equal Contribution
[†]Corresponding Author

Figure 1. Example of CTU partition in VVC.

have drawn huge attention. Belie *et al*. [1] implemented the end-to-end optimized image compression by utilizing the general divisive normalization and soft quantization with additive uniform noise. To further optimize the entropy coding, Lee *et al*. [7] proposed the context adaptive entropy model to further improve the compression efficiency, which exceeded BPG. Moreover, with the image compression and image enhancement network in cascade, Lee *et al*. [8] achieved the comparable performance to VVC intra coding. Despite it, learning based codecs are not very efficient as conventional codecs in their computational complexity. Hence, in this paper, we propose a learning based approach as a post-process technique to improve the performance of VVC intra coding.

To this regard, some studies [15, 10] have applied the network of image restoration to the post process of compression frameworks. However, these methods are not designed according to the distortion characteristics of VVC. As shown in Fig. 1, the size of coding units in VVC are different, which caused the compression artifacts distributed in different spatial size. Moreover, the compression arti-
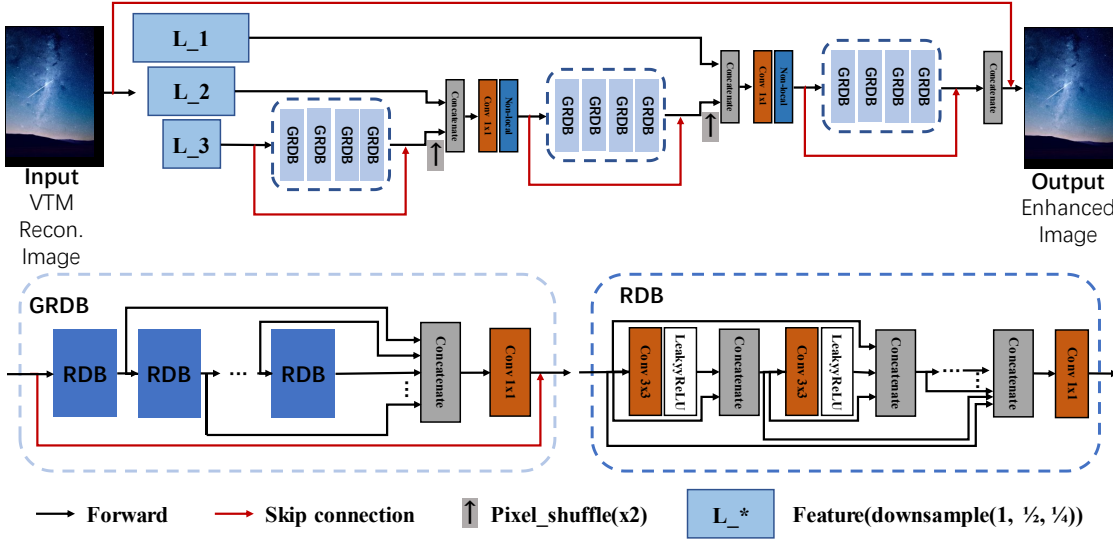
Figure 2. Overview of our Multi-scale grouped dense network (MSGDN).

facts including blurring, blocking and texture loss are complex, which require the network to have strong representation ability. Therefore, in this paper, we propose the multi-scale grouped dense network (MSGDN) as the post-process of VVC intra coding based on these analysis.

Specifically, inspired by [10], we utilize multi-scale architecture to remove the compression artifacts distributed in different spatial sizes. Moreover, to increase the representation ability of network, we utilize the GRDB [6] as backbone in each scale. By combining the multi-scale and grouped residual dense block, we propose a post-process network named MSGDN, which achieves the state-of-the-art compression performance with the same training datasets. Moreover, to further improve the subjective quality of the compressed image at low bit-rate, we also propose a generative adversarial network (MSGDN-GAN) by utilizing our MSGDN as generator.

We conduct extensive experiments on validation sets and our MSGDN trained by MSE losses achieves the highest PSNR of 32.62 on average at bit-rate of 0.15. Moreover, our MSGDN-GAN could achieve the better subjective performance as shown in Fig. 4.

2. Approaches

2.1. Multi-scale grouped dense network

Across different scaled feature, the network could fuse the coarse and fine information to enhance the representational ability. To process the compression artifacts across from different spatial dimensions, we utilize the multi-scale feature extraction as Fig. 2. We apply the convolution layer

with stride 2 to implement the down-sample of feature. The multi-scale features are composed of three layers. The resolution of first layer is the same as original image and other two layers are down-sampled by 2 and 4 respectively. To enhance the representation power of each layer, we utilize the GRDB [6] in each layer. The channels of GRDB in three layers from low to high are set as 128, 128, and 64. Each GRDB contains 4 RDBs and each RDB contains 8 convolution layers. To fusion multi-scale feature representation, we first up-sample the low representation and concatenate it with higher representation followed by one 1x1 convolution layer and non-local module.

2.2. Generative adversarial network for VVC

Generative adversarial network has been used in image restoration, image-to-image translation and style transfer, which could improve subjective quality efficiently. Inspired by [14], we utilize our MSGDN as generator. Then we enhance the discriminator with Relativistic average Discriminator (RaD) [5], which focus on whether input data are more relative realistic or not.

Here, we formula real image and fake image as x_r and x_f respectively. The discriminator loss could be formulated as Equ. 1.

$$L_D^{Ra} = -E_{x_r}[\log(D(x_r, x_f))] \\ - E_{x_f}[\log(1 - D(x_f, x_r))], \quad (1)$$

where E_{x_r} means the average of all real data in the mini-batch. $D(x_f, x_r)$ is defined in [5] as Equ. 2.

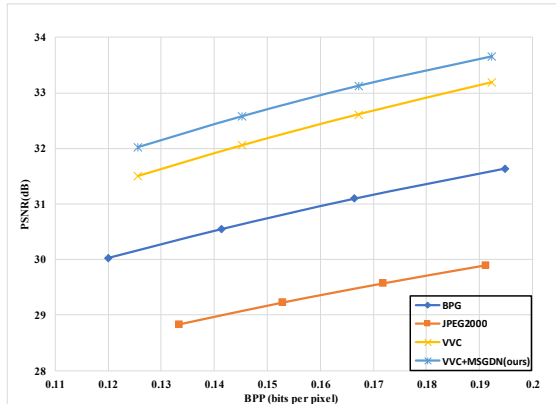$$D(x_r, x_f) = f(C(x_r) - E_{x_f}[C(x_f)]), \quad (2)$$

Figure 3. Compression performance on validation sets, compared with JPEG2000, BPG and VVC.

where f is sigmoid function and C represents the output of discriminator. We also adopt perceptual loss in [14] based on the convolution layer of VGG-19 to further constrain the subjective performance. As shown in Fig. 4, our MSGDN-GAN achieves the better performance compared with traditional compression methods. However, it will cause the reduction of objective performance.

2.3. Loss function

We adopt different loss function for different optimization goals. To improve objective performance, we first utilize the L1 loss as our loss function. Then we fine-tune the MSGDN with MSE loss. In this way, we achieve the highest PSNR of 32.622 on validation set. On the other hand, in order to improve the subjective performance, we first train our generator with L1 loss, and then we replace the L1 loss with a hybrid loss, which is composed of L1 loss, adversarial loss and perceptual loss as Equ. 3.

$$Loss = L1loss + 0.01L_{adv} + 0.0001L_p, \quad (3)$$

where L_{adv} represents the adversarial loss and L_p represents the perceptual loss. L_{adv} could be computed with Equ. 4.

$$L_{adv} = -E_{x_r}[\log(1 - D(x_r, x_f))] \\ - E_{x_f}[\log(D(x_f, x_r))], \quad (4)$$

which is symmetrical form of Equ. 1.

3. Experiment

3.1. Datasets

For training our MSGDN, we utilize 1633 images from Dataset P (professional) and M (Mobile) and 30000 images which are selected randomly from COCO2014 datasets.
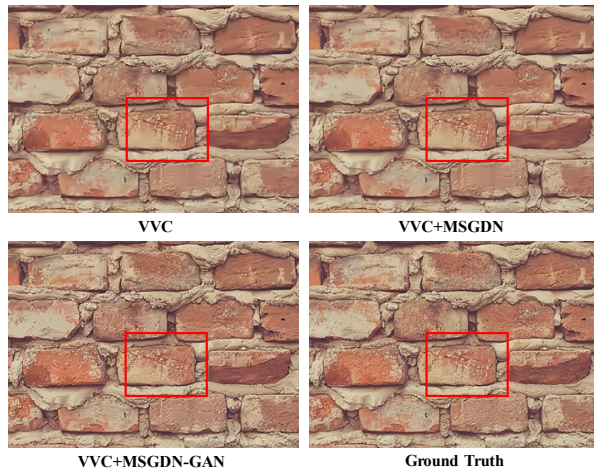

Figure 4. Comparison of image qualities between VVC, VVC+MSGDN, VVC+GAN and Ground Truth, which verify the superiority of our algorithms.

Then we transform these images from RGB to YUV444 and use VTM8.0 to compress these images with QP from 37 to 39. Finally, we transform these images from YUV444 to RGB. In this way, we get 31633 pairs of images at each QP.

3.2. Implementation details

The MSGDN is implemented based on PyTorch framework with four NVIDIA 1080Ti GPUs. In the process of training, we set the number of mini-batch as 8 and we utilize Adam optimizer with a initial learning rate of 0.0001. The learning rate will decay by a factor valued 0.5 every 100 epochs. To ensure that the average bpps is 0.15, we conduct image-level bit allocation by selecting images from coded images with QP from 37 to 39. However, in this paper, we only train one MSGDN for these QPs. We find that it cannot bring significant improvement by training different network.

3.3. Comparision with traditional methods

In this section, we compare our approach with traditional coding techniques including JPEG2000, BPG and VTM8.0 respectively. As shown in Fig. 3, with the same 0.15bpp, the PSNR of our MSGRD is 0.6dB higher than VVC. To compare the subjective quality, we visualize the processed images at about 0.15bpp. As shown in Fig. 4, the blurring and texture loss are serious in images coded with VVC. However, our MSGDN can remove these artifacts efficiently and our MSGRD-gan achieves the best subjective quality in texture details compared with other methods.

3.4. Comparison with other post-processing works

To further valid the effectiveness of our MSGDN, we retrain a series of post-process network including GRDN [6], DHDN [11], and CLIC2019_MS [10], DR_US [9], . As shown in Fig. 5, our approach achieves the highest PSNR at the point of 0.15bpp, which is 0.3dB than DHDN.
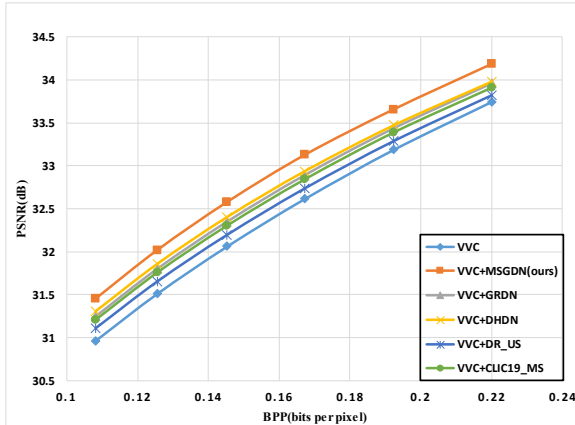

Figure 5. Compression performance on validation sets, compared with other post-process networks.

4. Conclusion

In this paper, we propose the multi-scale grouped dense network (MSGDN) as the post-process module of VVC intra coding. By utilizing the multi-scale feature representation and grouped dense block, our MSGDN trained on MSE achieves the PSNR of 32.622 with team "IMC" and "haha" at low bits of 0.15bpp. However, the texture details cannot be reconstructed effectively only with MSE. Therefore, we utilize our MSGDN as a generator, and implement the MSGDN-GAN by combining the loss of RaGAN [5] and perceptual loss. Extensive experiments have demonstrated the effectiveness of our MSGDN and MSGDN-GAN. Our MSGDN-GAN achieves better subjective quality in texture details.

5. Acknowledgement

References

[1] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

[2] Zhibo Chen and Tianyu He. Learning based facial image compression with semantic fidelity metric. *Neurocomputing*, 338:16–25, 2019.

[3] Tianyu He, Simeng Sun, Zongyu Guo, and Zhibo Chen. Beyond coding: Detection-driven image compression with semantically structured bit-stream. In *2019 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2019.

[4] Xin Jin, Runchun Ye, and Zhibo Chen. Multiscale progressive image compression network guided by learnable just noticeable distortion. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2018.

[5] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[6] Dong-Wook Kim, Jae Ryun Chung, and Seung-Won 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 Workshops*, pages 0–0, 2019.

[7] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. *arXiv preprint arXiv:1809.10452*, 2018.

[8] Jooyoung Lee, Seunghyun Cho, and Munchurl Kim. A hybrid architecture of jointly learning image compression and quality enhancement with improved entropy minimization. *arXiv preprint arXiv:1912.12817*, 2019.

[9] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019.

[10] Ming Lu, Tong Chen, Haojie Liu, and Zhan Ma. Learned image restoration for vvc intra coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[11] Bumjun Park, Songhyun Yu, and Jechang Jeong. Densely connected hierarchical network for image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[12] Mingze Wang, Shuai Wan, Hao Gong, Yuanfang Yu, and Yang Liu. An integrated cnn-based post processing filter for intra frame in versatile video coding. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1573–1577. IEEE, 2019.

[13] Ming-Ze Wang, Shuai Wan, Hao Gong, and Ming-Yang Ma. Attention-based dual-scale cnn in-loop filter for versatile video coding. *IEEE Access*, 7:145214–145226, 2019.

[14] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[15] Yuyang Xue and Jiannan Su. Attention based image compression post-processing convolutional neural network. *arXiv preprint arXiv:1905.11045*, 2019.

[16] Zhizheng Zhang, Zhibo Chen, Jianxin Lin, and Weiping Li. Learned scalable image compression with bidirectional context disentanglement network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1438–1443. IEEE, 2019.