Towards the Perceptual Quality Enhancement of Low Bit-rate Compressed Images

Younhee Kim[1*], Seunghyun Cho[2], Jooyoung Lee[1], Se-Yoon Jeong[1], Jin Soo Choi[1], Jihoon Do[1],
[1]Broadcasting and Media Research Laboratory
Electronics and Telecommunications Research Institute
218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Korea
[1]{kimyounhee, leejy1003, jsy, jschoi, jhdo}@etri.re.kr

[2]Department of Information and Communication Engineering
7 Kyungnamdaehak-ro, Masanhappo-gu,
Changwon-si, Gyeongsangnam-do, 51767, Republic of Korea
[2]scho@kyungnam.ac.kr

Abstract

In this paper, a low bit-rate compressed image quality enhancement framework is presented. A recent image/video coding method and a deep learning based quality enhancement method are integrated to improve the perceptual quality of compressed images. The proposed architecture is designed to reduce the coding artifact and restore the blurred texture details. The experimental results presents that the proposed framework yields a 33% improvement in the Perceptual Index score which is consistent with visual evaluation on a sample of results.

1. Introduction

Image compression has been a long research topic and the compression ratio is improved continuously. As we consume the massive multimedia data everyday, the traditional image compression techniques do not provide the sufficient compression ratio.

Joint Video Experts Team (JVET) is ready to release a new video coding standard named Versatile Video Coding (VVC). VVC aims to provide two times higher compression ratio than the recent standard High Efficient Video Coding (HEVC).

High compression inevitably comes with artifacts such as blocking artifacts, ringing effects and blurring. Recent research has started to apply deep learning to artifact reduction problem. Yu [1] designed AR-CNN to reduce the coding artifact and showed an improvement in terms of PSNR and SSIM. Kim [2] proposed GRDN, which is based on the residual dense network (RDN), and won the image denoising challenge, NTIRE 2019. Both methods still optimize the CNN based network using pixel difference between the original images and the network output.

In this paper, in order to provide the good perceptual quality at low bit-rate compressed images, a perceptual image quality enhancement framework is proposed in pursuit of reducing coding artifact and restoring the texture details.

2. Proposed Methods

In the proposed framework, an 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 and the texture details restoration.

2.1. Encoding using VVC intra coding

We utilize the VVC intra coding scheme to obtain high bit saving. VVC intra coding has a 23% BD-rate gain compared to the previous HEVC intra coding. Various tools were newly adopted to achieve the high compression in VVC. The number of directional intra mode was extended to 65 [3]. Multi-line intra prediction [4] used non-adjacent lines as a reference for the prediction. Position Dependent Prediction Combination (PDPC) [5] [6] utilized both left and above reference samples to reduce the prediction error. Block size and intra mode dependent filter selection [7], and Cross component linear model (CCLM) [8] [9] were adopted to improve the prediction. In the proposed framework, we utilizes VTM 7.3 (all intra configuration in the common condition setting) to encode the original images with the quantization parameter range between 35 and 38 for the targeted low bit-rate.

* Corresponding author

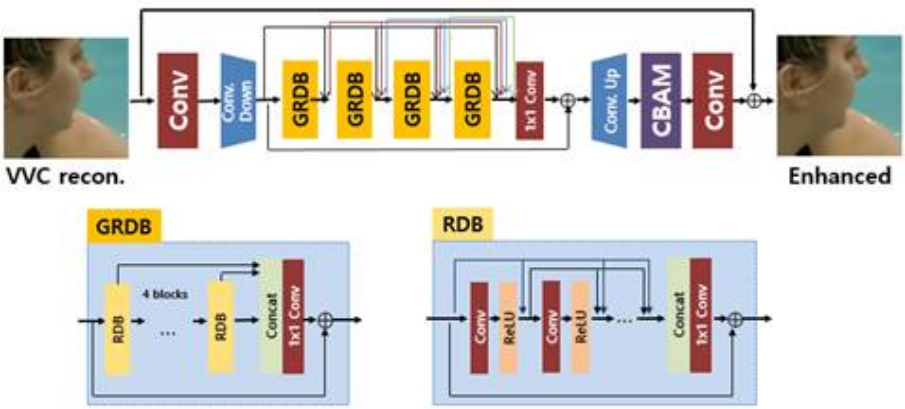

Figure 1: HGRDN architecture.

2.2. Network based quality enhancement

Low bit-rate compression results in blocking artifacts, ringing effects, and blurring. Traditionally, specific filters to reduce such artifacts have been designed. Various deep learning based image restoration methods such as super resolution (SR) recently have been proposed. Since artifacts reduction problem is a similar problem to SR, the state-of-the-art SR networks are widely adopted in the artifact reduction.

First approach is designed to optimize the pixel difference between the input and output, which results in improving PSNR value. These PSNR-oriented approaches are effective in suppressing blocking artifact while retaining edges. However, the PSNR-oriented approaches results in over blurred images.

Second approach is designed to improve the perceptual quality. These perception-oriented approaches are effective in restoring texture details and naturalness.

In the proposed framework, we adopt the PSNR-oriented approach for noise reduction and the perception-oriented approach for texture restoration.

2.2.1 Noise reduction network

In this paper, HGRDN [10] is utilized to reduce the coding artifact, particularly blocking artifact and ringing effect. The HGRDN has improved GRDN [2], which ranked first place for real image denoising in terms of PSNR and SSIM. HGRDN structure is shown in Figure 1. HGRDN is consisted of four grouped residual dense connections (GRDB), a down-sampling layer, a up-sampling layer, and a CBAM [11] layer. The loss function of the noise reduction network denoted as L_{GN} is:

$$L_{GN} = \mathbb{E}[\|G_n(x_i) - y_i\|], \quad (1)$$

where x_i and y_i are input image and ground truth image, respectively, and $G_n(\cdot)$ represents the output of the noise reduction network. $\mathbb{E}[\cdot]$ represents the expectation that has applied to the batch data.

2.2.2 Texture restoration network

We also employed HGRDN as the basic architecture for texture restoration. We applied the loss functions borrowed from the well known perception-oriented SR network in order to obtain the naturalness, particularly restoration of the texture details. Three losses defined in ESRGAN [12] are applied.

The fist loss is a feature loss. Johnson et al. [13] proposed this feature loss using VGG [14] features to measure the perceptual similarity, and many SR methods such as SRGAN [15] and ESRGAN [12] have employed the feature loss. The feature loss in the texture restoration network is defined as follows:

$$L_f = \mathbb{E}[\|VGG19_{54}(x_i) - VGG19_{54}(y_i)\|] \quad (2)$$

where x_i and y_i are input image and ground truth image, respectively, and $VGG19_{54}$ represents the features obtained by the 4th convoluation before the 5th layer of 19-layer VGG network. $\mathbb{E}[\cdot]$ represents the expectation that has applied to the batch data.

The second loss is an adversarial loss. The GAN [16] framework has been known to be able to generate the realistic images. The generator generates an image and the discriminator distinguishes if the generated image looks real. We adopted the adversarial loss of relative GAN (RaGAN) that ESRGAN [12] designed. The discriminator and generator losses are defined as follows:

$$L_D = -\mathbb{E}_{x_r}\left[\log(\mathrm{D}(x_r, x_f))\right] - \mathbb{E}_{x_f}\left[\log(\mathrm{D}(x_f, x_r))\right], \quad (3)$$

$$L_G = -\mathbb{E}_{x_r}\left[\log(1 - \mathrm{D}(x_r, x_f))\right] - \mathbb{E}_{x_f}\left[\log(\mathrm{D}(x_f, x_r))\right], \quad (4)$$

$$\mathrm{D}(x_r, x_f) = \mathrm{D}(x_r) - \mathbb{E}[D(x_f)], \quad (5)$$

$$\mathrm{D}(x_f, x_r) = \mathrm{D}(x_f) - \mathbb{E}[D(x_r)], \quad (6)$$

where x_r and x_f denote real data and fake data, respectively, and $\mathbb{E}$ denotes the expectation of all mini-

batch data.

The third loss is a pixel loss. The pixel loss is incorporated to reduce the unpleasant noise created by the GAN-based methods [17]. The pixel loss is defined as follows:

$$L_P = \mathbb{E}[\|G_t(x_i) - y_i\|], \quad (7)$$

where x_i and y_i are input image and ground truth image, respectively, $G_t(x_i)$ represents the output of the texture restoration network.

2.3. Implementation

To implement the proposed framework, we integrated the VVC Test Mode (VTM) [18] version 7.3 with the noise reduction network (NR) and texture restoration network (TR). For image encoding, the original image is first converted into YUV420 and encode it using VTM with all intra configuration setting. The reconstructed image is converted again into RGB format and fed into the network based quality enhancement process which includes the NR and/or TR. We defined the quality enhancement types according to the applying of NR and TR as Table 1.

We used 1633 CLIC training images and 36,000 images of Microsoft COCO training dataset [19]. The training dataset images are encoded using VTM and then randomly cropped with the size 96×96 for training. We trained using Adam [20] with $\beta_1 = 0.9, \beta_2 = 0.999$. The initial learning rate was set to 1×10^{-4}, and then decreased to half at [50k, 100k, 200k, 300k] iterations. We used 64 filters, 8 3x3 convolutions and 16 RDBs in HGRDN. We trained the noise reduction network using L_1 loss as eq. (1). The texture restoration network loss is defined as follows:

$$L_{GT} = \alpha L_p + \gamma L_f + \delta L_G, \quad (8)$$

where the hyper parameter for α, γ, δ were set as 2×10^{-2}, $1, 5 \times 10^{-4}$, respectively.

Table 1.

Type	Method
Type 1	NR
Type 2	TR
Type 3	NR + TR
Type 4	CAS

3. Experimental results

We defined four types of the image enhancement process. In Table 1, NR and TR refer to the noise reduction network and the texture restoration network, respectively. CAS refers that cascade training of NR and TR. The type 1 enhancement process means that the VVC reconstructed image is fed into only the NR network. The type 2 enhancement process means that the VVC reconstructed image is fed into only the TR network. The type 3 enhancement process means that the reconstructed image is fed into pre-trained NR first and then fed into pre-trained TR. The type 4 enhancement process means that the reconstructed image is fed into the cascade trained NR and TR.

We employed the perceptual index metric to measure the perceptual quality of our framework. The PI (Perceptual Index) was used from the PIRM-SR Challenge [21] to judge the perceptual quality of the SR algorithms. The PI is calculated as follows:

$$\text{PI} = \frac{(10 - \text{Ma}) + \text{NIQE}}{2}, \quad (9)$$

where Ma [22] and NIQE [23] are two different well known non-reference quality metrics. The PI value is the lower, when the perceptual quality is better.

Figure 2 and Figure 3 show the PSNR values and PI values for each type according to the bit rate in BPP. The rank order of the type in the PSNR measure and PI measure are different. For example, Type 1 is the best in PSNR but Type 4 is the best in PI.

As the second row (IMG_20170504_183130) in Figure 4 showed, Type 1 method reduces significantly the ringing artifact around the elbow. Type 2 method tends to add noticeable noise. The Type 3 and Type 4 methods show the trade off between the artifacts and the texture detail restoration. The Type 3 applied image remains the ringing artifact but recovers the realistic elbow, not too smoothed. The Type 4 image has the noticeable grain noise in the elbow.

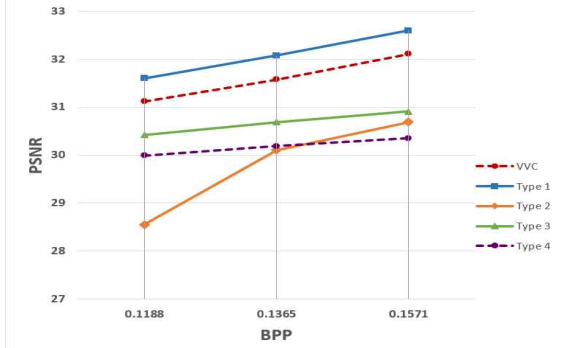

Figure 2: Comparison of the quality enhancement types in PSNR

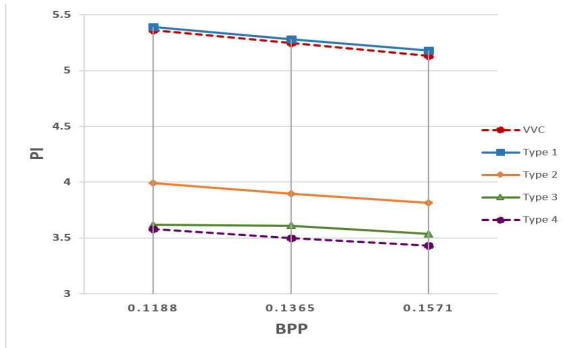

Figure 3: Comparison of the quality enhancement types in PI

4. Conclusion

In this paper, a low bit-rate image compression framework towards high perceptual quality is presented. An image is encoded using VVC, and then fed into a network-based quality enhancement process. We employed two networks for the post processing and defined four types of the enhancement process according to the combination of the post processing networks. The proposed method (Type 3 and Type 4) has shown that the coding artifact is reduced and the texture detail is recovered in the visual quality. The enhancement of the quality is also measured using the PI and a 33% improvement has achieved. We remain various network combinations (e.g., changing the order of NR and TR) for further study.

Acknowledgement

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2017-0-00072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media).

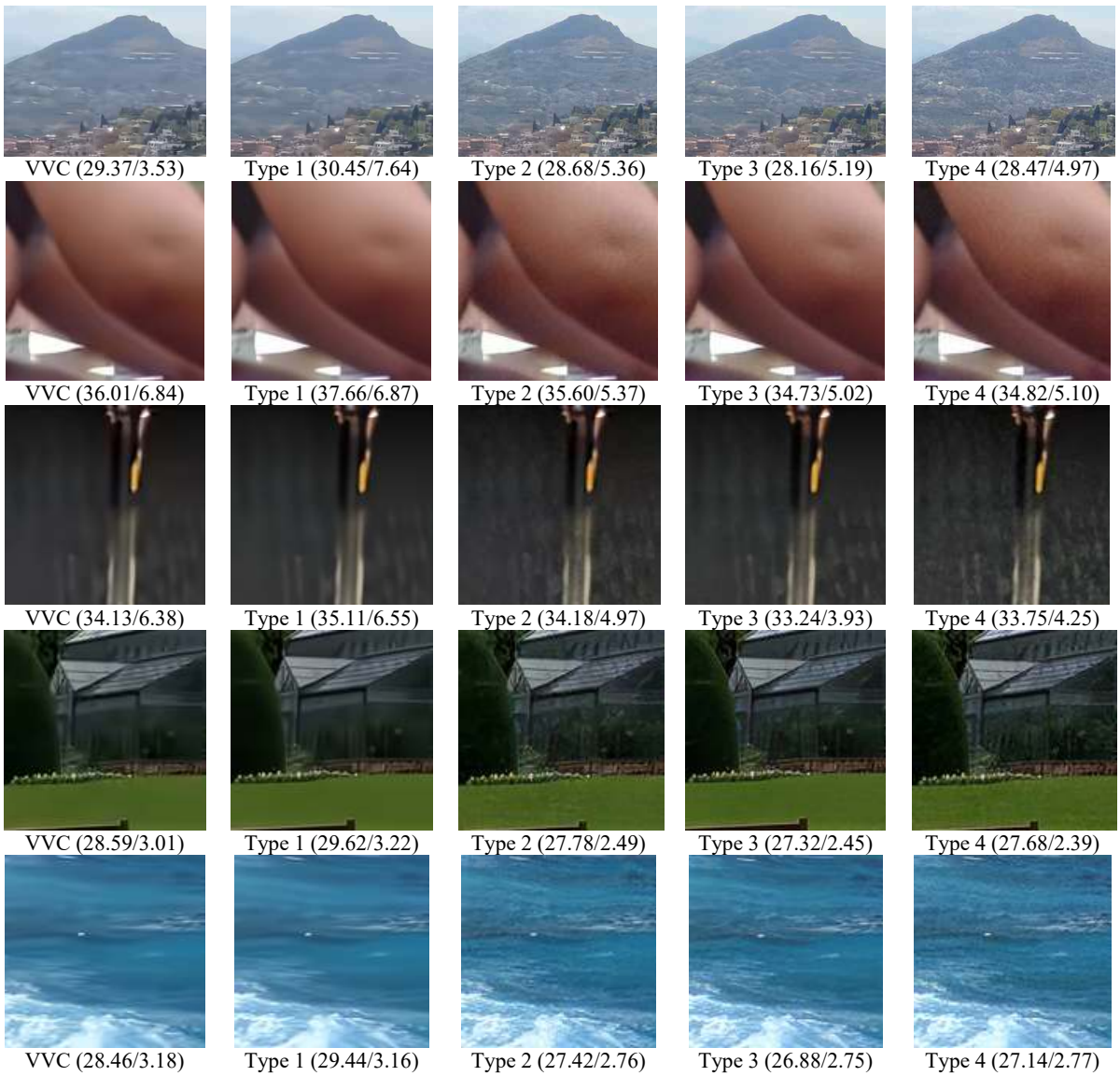

Figure 4: PSNR and PI results after Type 1, Type 2, Type 3, and Type 4 processing. The images are selected from validate set, which names are IMG_0470_1, IMG_20170504_183130, IMG_20170114_204505, IMG_20170721_103913, IMG_20170730_133144 from the top row.

References

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

[2] Kim, Dong-Wook, et al. "Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[3] Bross, B., et al. "CE3: Multiple reference line intra prediction." JVETK0051, Ljubljana, SI (2018).

[4] Zhao, Xin, et al. "Low-Complexity Intra Prediction Refinements for Video Coding." 2018 Picture Coding Symposium (PCS). IEEE, 2018.

[5] Said, Amir, et al. "Position dependent prediction combination for intra-frame video coding." 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016.

[6] Filippov, A., et al. "CE3: A combination of tests 3.1. 2 and 3.1. 4 for intra reference sample interpolation filter." document JVET-L0628, in Proc. of 12th JVET meeting. 2018.

[7] X. Ma, et al. "CE3: Tests of cross-component linear model in BMS, " document JVET-K0190, in Proc. of 13th JVET meeting. 2018.

[8] G. Laroche, et al. "CE3-5.1: On cross-component linear model simplification." document JVET-L0191, in Proc. of 12th JVET meeting. 2018.

[9] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[10] S. Cho, D.-W. Kim, and S.-W. Jung, "Quality Enhancement of VVC Intra-frame Coding for Multimedia Services over the Internet, " International Journal of Distributed Sensor Networks, 2020 (accepted for publication)

[11] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[12] Wang, Xintao, et al. "Esrgan: Enhanced super-resolution generative adversarial networks." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[13] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." European conference on computer vision. Springer, Cham, 2016.

[14] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[15] Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[16] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

[17] Kim, Dong-Wook, et al. "Constrained adversarial loss for generative adversarial network-based faithful image restoration." ETRI Journal 41.4 (2019): 415-425.

[18] Versatile video coding reference software version 7.3 (VTM-7.3)"https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/tags/VTM-7.3/

[19] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[20] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[21] Blau, Yochai, et al. "The 2018 PIRM challenge on perceptual image super-resolution." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[22] Ma, Chao, et al. "Learning a no-reference quality metric for single-image super-resolution." Computer Vision and Image Understanding 158 (2017): 1-16.

[23] Mittal, Anish, Rajiv Soundararajan, and Alan C. Bovik. "Making a "completely blind" image quality analyzer." IEEE Signal Processing Letters 20.3 (2012): 209-212.