Decoder Side Color Image Quality Enhancement using a Wavelet Transform based 3-stage Convolutional Neural Network

Kai Cui and Eckehard Steinbach
Chair of Media Technology, Technical University of Munich
Munich, Germany
kai.cui@tum.de, eckehard.steinbach@tum.de

Abstract

In this paper, we describe our submission to the workshop and challenge on learned image compression (CLIC) hosted at CVPR 2019. Lossy compressed images usually suffer from unpleasant artifacts, especially when the bit-rate is low. In order to improve the image quality without spending extra bit-rate, decoder side quality enhancement becomes necessary. Most approaches focus on spatial information exploration and the quality enhancement is usually only performed on the luminance component, which leads to the neglect of inter-channel correlation. In addition, since compressed images mainly lose the high-frequency components, high-frequency and low-frequency components show different characteristics. Motivated by the characteristics of compressed images, a wavelet transform based 3-stage CNN is proposed in this paper. With the RGB image as input, the proposed network exploits the latent inter-channel correlations and enhances the low-frequency and high-frequency sub-band separately. Both objective and subjective evaluations show the noticeable quality improvements compared to Better Portable Graphics (BPG) and previous approaches.

1. Introduction

Most of the modern lossy image and video codecs (e.g. JPEG, WebP, BPG[1], H.264, HEVC) are block based. The compressed images and videos often suffer from visible distortion (e.g. block and ringing artifacts) for areas with rich texture and sharp edges, especially when the bit-rate is relatively low. For some image and video codecs, there are built-in filters in the decoder to mitigate this problem. In HEVC, in-loop filtering is adopted, including a deblocking filter (DBF) and sample adaptive offset (SAO), to alleviate the block and ringing artifacts, respectively. However, the results are still not satisfactory when the bit-rate is low.

A number of approaches have been proposed to reduce these artifacts. Conventional approaches design filters based on image priors (low-rank, non-local similarity, sparsity). But most of these priors are hand-crafted and not optimal in some cases.

With the success of convolutional neural networks (CNN) in image processing, CNN based algorithms have also been proposed. In [5], a compression artifact reduction CNN (ARCNN) is proposed, which achieves significant improvement compared to conventional approaches. In [7], a reconstruction network is proposed, which solves both super-resolution and enhancement problems at the same time. In [11], a decoder-side HEVC quality enhancement using a scalable CNN is proposed, which enhances the quality of Intra-frames and Inter-frames with different subnetworks. In [12], a residual learning based denoising network is proposed, which can also be adopted to solve multiple image restoration problems like deblocking or super-resolution when trained with corresponding data.

However, most of these approaches exploit only spatial information, they are typically applied on the luminance component only, and the inter-channel correlation is not exploited. In image and video compression, the YUV color format (YUV420, YUV444) is usually adopted, based on the assumption that the human visual system is not so sensitive to color differences compared to brightness changes. When the decoded RGB images are obtained, usually the G channel has the best quality, R and B have lower quality. In [4], it has been shown that for both Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity (MS-SSIM), the G channel of compressed images shows higher quality than the other two channels, even when the bit-rate is low. The R, G and B channels captured with a single sensor have very strong inter-channel correlations [3], which means the reconstruction of one channel can benefit from the samples of the other two channels. Based on this observation, we reconstruct the RGB channels instead of the luminance channel only.

Most compression algorithms achieve high compression ratios by discarding high-frequency components, while at the same time preserving the low-frequency components. Therefore, the major loss of the compressed image should be the high-frequency component, the low-frequency part should be similar as the original image.

In order to prove this, a simple test is performed with BPG image compression and the kodak dataset. The 24 images of the kodak dataset are compressed with the BPG codec with a quantization parameter (qp) of 40, using the YUV444 format. Then, the Haar wavelet transform is adopted to perform frequency decomposition for both the original images and the compressed images. Four sub-bands LL, HL, LH, HH are obtained. Pearson correlation coefficient is adopted to evaluate the similarity of the coefficients of different sub-bands between the original images and the BPG compressed images. The mean values and the variance of the correlation coefficients for each sub-band are shown in Tab. 1.

Table 1. Correlation coefficients of different sub-bands between original image and compressed BPG image

Correlation	LL	HL	LH	HH
mean	0.9901	0.7983	0.8311	0.5412
variance	0.000045	0.0055	0.0028	0.0119

From Tab. 1, it can be seen that the LL sub-band is highly correlated, which means that the image codec maintains the low-frequency part well. The correlation of the HH sub-band is much lower than for the other three, which means that high-frequency information in the compressed images is lost. The results here are in accordance with the aforementioned analysis. Existing approaches are usually performed on the original pixel values. Some approaches operate in the frequency domain [10, 2, 6], however, they just perform the enhancement jointly for all sub-bands and do not discriminate between different frequency components. In comparison, we design a scheme which adopts different models to deal with different frequency components. Since the Discrete Wavelet Transform (DWT) is widely used in image processing, we adopt the DWT as the frequency decomposition approach in our work.

Based on these observations, the inter-channel correlations and the frequency components characteristics, in this paper, we propose a wavelet transform based 3-stage CNN approach to enhance the compressed color image quality. First, since the R/G/B channels have high correlations, we adopt the RGB image as the input of the network instead of using only the luminance Y. The network exploits the latent inter-channel correlations and the reconstruction of samples from one channel benefits from the samples of other channels. Second, due to the frequency components characteristics, we design the network structure to make it perform enhancement for different frequency components separately, which allows the network to better adapt to the characteristics of the compressed image.

The major contributions of this work can be summarized as follows: First, RGB images are fed to the network to exploit the latent inter-channel correlation. Second, a wavelet transform based 3-stage structure is proposed where the low-frequency sub-band and high-frequency sub-band are enhanced separately. Third, the proposed scheme is a pure post-processing approach and hence compatible with any existing image and video codecs, which makes the approach applicable in practice.

2. Proposed scheme

In image and video compression, YUV is the most commonly used format. The U and V chrominance components are usually compressed more aggressively than Y because of the characteristics of the human visual system. This leads to the aforementioned characteristics of compressed images when transformed back to RGB domain. The R, G, and B channels exhibit strong inter-channel correlation, which means that the samples from other channels can be used to enhance the quality of the current channel.

The compressed images usually lose more high-frequency components while maintaining most of the low-frequency components. Dealing with the low-frequency and high-frequency components with different models is an intuitive approach to mitigate this kind of difference. Based on these characteristics, we propose the wavelet transform based 3-stage CNN structure shown in Fig. 1 for compressed color image quality enhancement.

First, the compressed image is decoded with a standard image codec. The obtained RGB images are then transformed into the wavelet domain. For easy implementation and computational complexity consideration, we adopt the simple Haar wavelet. After the Haar wavelet transform, the image is reshaped to a 12-channel image as shown in Fig. 2, this 12-channel image is the input of the first stage network.

After getting the output of the first stage, in the second stage, two parallel networks are designed, which enhance the LL component and the other three high-frequency components separately. In the third stage, the enhanced low-frequency components and the high-frequency components are concatenated and fed to the third stage network to perform the final refinement. Finally, the enhanced RGB images are obtained with reshaping and inverse wavelet transform.

Fig. 3 shows the detailed structure of the network unit for each stage. In the first layer, 64 filters of size $3 \times 3 \times d$ are used to generate feature maps, the last convolutional layer adopts d filters of size $3 \times 3 \times 64$ to generate the corresponding output. For the hidden layers, 64 filters of size $3 \times 3 \times 64$ are adopted. The number of the layers in each unit K is set to 10 and d is set to 12, 3 and 9, 12 in the three stages, respectively. Stride is set to 1, and zero-padding of size 1 is used to ensure that each feature map has the same size as

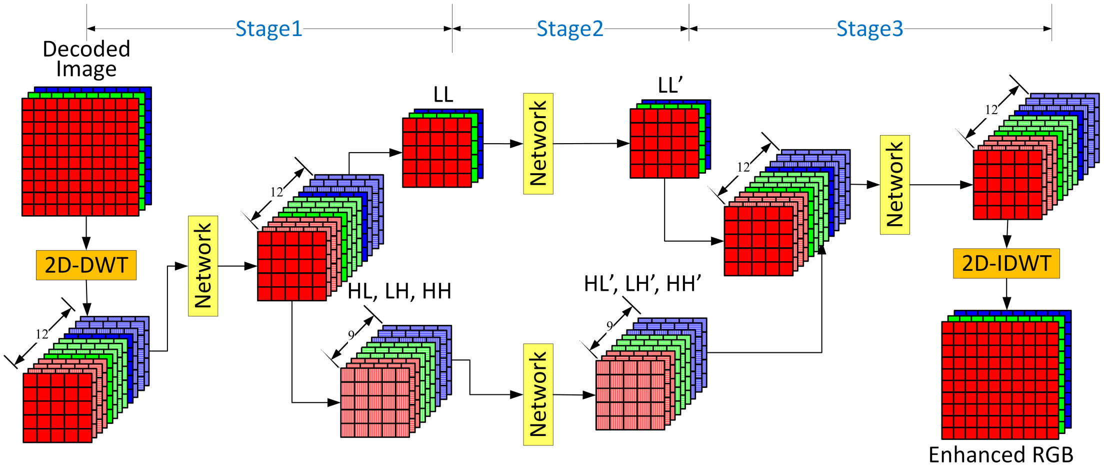


Figure 1. Structure of the proposed wavelet transform based 3-stage CNN scheme

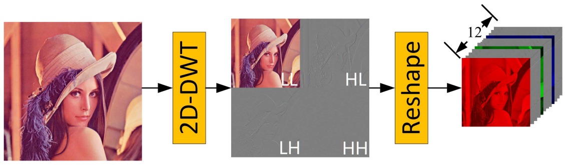


Figure 2. Color image 2D wavelet transform and reshaping to 12-channel images

the input. A shortcut connection is used for each stage to boost the training process, which is similar as the residual learning structure used in [8] and [12].

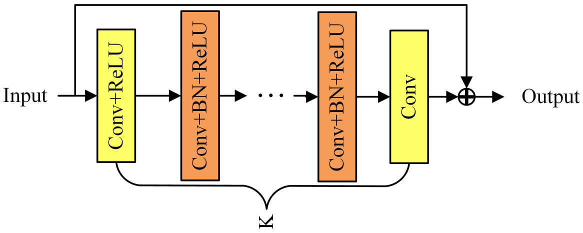


Figure 3. Structure of Network Unit

Consider the training dataset $(\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^{N}$, where $\mathbf{X}_i$ is the i-th decoded compressed image, $\mathbf{Y}_i$ is the corresponding ground-truth RGB image, and N is the number of images in the training data. During training, a loss function is defined to optimize the parameters of the networks. The mean squared error (MSE) function is used as the loss function which is defined as follows.

$$\begin{aligned} &L(\omega_1, \omega_{21}, \omega_{22}, \omega_3) = \\ &= \frac{1}{N}\sum_{i=1}^{N}\Big(\|\mathcal{F}(\mathbf{X}_i; \omega_1, \omega_{21}, \omega_{22}, \omega_3) - \mathbf{O}_i\|^2\Big) \end{aligned} \tag{1}$$

where ω_j represents the corresponding network parameters of the j-th stage. $\mathcal{F}(\mathbf{X}_i; \omega_1, \omega_{21}, \omega_{22}, \omega_3)$ is the i-th output of the 3-stage network. In order to ensure the generalizability of the trained model, a regularization term is also adopted in the loss function during training. The overall loss is defined in Eq. 2. The regularization coefficient λ is set to 0.0005.

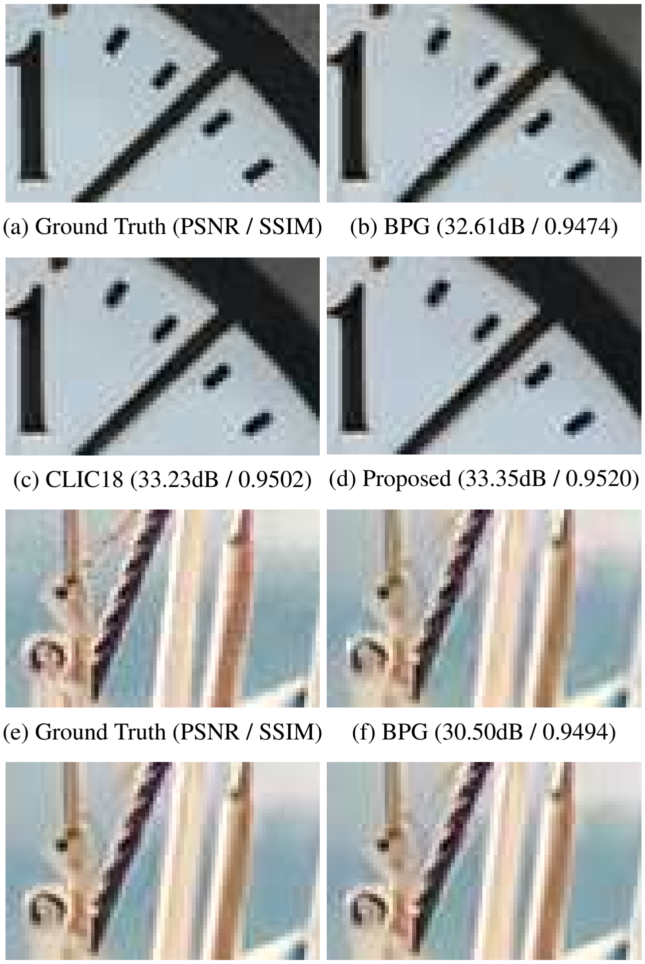
$$L_{all}(\omega_1, \omega_{21}, \omega_{22}, \omega_3) = L + \frac{1}{2}\lambda\sum_{j}(\|\omega_j\|^2) \tag{2}$$

3. Experiments and results

The provided training dataset is adopted in our experiments as training data. In this dataset, there are 1633 high-resolution natural images of various scenes shot by mobile devices and professional cameras.

The BPG codec is used to generate the compressed training images. The qp is set to 39 to meet the bit-rate constraint, jctvc option is enabled to achieve the best compression results, the level is set to 9, and bit-depth is set to 12 to achieve more accurate internal calculation. YUV444 format and color space YCgCo are adopted because they slightly improve the image quality compared to the default settings. The patch size is set to 160×160, and the patches are non-overlapping. The mini-batch size is set to 64. The weights of the networks are initialized according to [8] and the Adam solver is used to optimize the parameters. The starting learning rate is 0.001, and divided by 5 every 5 epochs. There are 45 epochs in total. Other hyper-parameters are using the default settings from [9].

First, two example images from the test dataset are depicted in Fig. 4 to show the visual quality of the proposed method. Usually the texture-rich and sharp edge area are the challenging cases. We zoom in some parts of these images to show the details. It can be seen that for BPG compression, block artifacts, false-color pixels and shadows can be


(a) Ground Truth (PSNR / SSIM) (b) BPG (32.61dB / 0.9474)

(c) CLIC18 (33.23dB / 0.9502) (d) Proposed (33.35dB / 0.9520)

(e) Ground Truth (PSNR / SSIM) (f) BPG (30.50dB / 0.9494)

(g) CLIC18 (30.93dB / 0.9516) (h) Proposed (31.17dB / 0.9541)

Figure 4. Visual Quality Comparison (Best seen on a computer monitor. *8d978b17742f8a06d429a3ab82fa9068e8a1989f.png* and *196690d0d3c97d21bfa659891f07e925c67d8b20.png* from the test dataset)

observed along the edges of the objects. With the proposed method, these artifacts can be well eliminated and the visual quality is improved. Not only the image quality is improved, due to performing the convolution in the wavelet domain, the spatial resolution is just a quarter compared to the original resolution with the same amount of parameters, the proposed method is 3 times faster than the CLIC18 [4] approach. We also notice that the groundtruth images are not always perfect, some of them suffer from slight noise or other issues, because these images are captured with mobile and professional cameras and processed with conventional image processing pipelines in the cameras.

The average PSNR, composite PSNR (CPSNR) and MS-SSIM are adopted to evaluate the objective quality of the proposed approaches. A weighted PSNR is also adopted in this challenge, which computes a single Mean Squared Error (MSE) value by averaging across all RGB channels of all pixels of the whole dataset. From that value a PSNR value is calculated, which is marked as the WPSNR in the table. The results are listed in Table 2. They are all under the constraint 0.15bpp which is required by the challenge.

Table 2. PSNR (in dB) and MS-SSIM results for the proposed wavelet transform based 3-stage approach on the validation dataset

Evaluation	BPG	C18-2	C18-3	Proposed
PSNR-R	31.22	31.85	31.94	32.02
PSNR-G	32.25	32.75	32.81	32.66
PSNR-B	30.98	31.78	31.90	32.53
CPSNR	31.43	32.08	32.17	32.37
WPSNR	30.85	31.48	31.57	31.72
MS-SSIM	0.948	0.954	0.955	0.957

From the results, it can be seen that the proposed method leads to an extra 0.2dB PSNR and 0.002 MS-SSIM improvement on the validation datasets in comparison to the two approaches (C18-2, C18-3) proposed in CLIC2018 [4]. In terms of the PSNR of each color component, the G channel is slightly lower, but obvious improvements for B channel can be observed. Compared to the original BPG codec with default parameter settings and the same bit constraint, our proposed method can achieve about 0.9dB PSNR improvement and 0.009 MS-SSIM improvement, especially for B channel, there is about 1.5dB PSNR improvement.

This approach is proposed for the CLIC2019 challenge. The BPG decoder is implemented by python binding with the shared objects libbpg.so compiled from BPG source code. The enhancement network is implemented with Tensorflow. For encoding, we use the standard BPG encoder compiled from the source code to encode all the images to the compressed format. Because there's an overall bit-rate constraint of 0.15bpp, we set the qp of most of the images to 39 and some to 38 to make the most of the bit budget. Regarding the running time, for the validation dataset, the proposed approach takes less than 2000s to decode all 102 images with a single CPU, which is faster than the 7000s for C18-3 proposed in CLIC2018 [4].

4. Conclusion

This paper presents a wavelet transform based 3-stage CNN for decoder side color image enhancement. The experimental results on both validation and test datasets show that the proposed scheme leads to noticeable quality and speed improvements compared to the previous approaches. Also, the proposed method is a post-processing approach, which makes it compatible with any existing image codec. As a part of the future work, we will investigate other wavelet bases and the potential gains applying the proposed scheme to video codecs.

References

[1] Fabrice Bellard. The BPG image format. http://bellard.org/bpg/. 1

[2] Honggang Chen, Xiaohai He, Linbo Qing, Shuhua Xiong, and Truong Q Nguyen. DPW-SDNet: Dual pixel-wavelet domain deep CNNs for soft decoding of JPEG-compressed images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 711–720, June 2018. 2

[3] Kai Cui, Zhi Jin, and Eckehard Steinbach. Color image demosaicking using a 3-stage convolutional neural network structure. In *IEEE International Conference on Image Processing*, pages 2177–2181, October 2018. 1

[4] Kai Cui and Eckehard Steinbach. Decoder side image quality enhancement exploiting inter-channel correlation in a 3-stage CNN: Submission to CLIC 2018. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2571–2574, June 2018. 1, 4

[5] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *IEEE International Conference on Computer Vision*, pages 576–584, December 2015. 1

[6] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 104–113, July 2017. 2

[7] Feng Jiang, Wen Tao, Shaohui Liu, Jie Ren, Xun Guo, and Debin Zhao. An end-to-end compression framework based on convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3007–3018, 2018. 1

[8] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, June 2016. 3

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, May 2015. 3

[10] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 773–782, June 2018. 2

[11] Ren Yang, Mai Xu, and Zulin Wang. Decoder-side HEVC quality enhancement with scalable convolutional neural network. In *IEEE International Conference on Multimedia and Expo*, pages 817–822, July 2017. 1

[12] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 1, 3