QUALITY ENHANCEMENT FOR INTRA FRAME CODING VIA CNNS: AN ADVERSARIAL APPROACH

Zhipeng Jin[1,2,3], Ping An[1,2,3], Chao Yang[2,3], Liquan Shen[1,2,3]

1 Shanghai Institute for Advance Communication and Data Science, 2 Key Laboratory of Advance Displays and System Application, Ministry of Education, 3 School of Communication and Information Engineering, Shanghai University, Shanghai, 200444, China

ABSTRACT

Lossy compression is an indispensable technique in image/video processing, due to its highly desirable ability of reducing the huge data volume. However, lossy compression introduces complex compression artifacts. To reduce these artifacts, post-processing techniques have been extensively studied. In this paper, we propose a novel post-processing technique using multi-level progressive refinement network via an adversarial training approach, called MPRGAN, for artifacts reduction and coding efficiency improvement in intra frame coding. Furthermore, our network generates multi-level residues in one feed-forward pass through the progressive reconstruction. This coarse-to-fine work fashion, which makes our network have high flexibility, can make trade-off between enhanced quality and computational complexity. Thereby facilitates the resource-aware applications. Extensive evaluations on benchmark datasets verify the superiority of our proposed MPRGAN model over the latest state-of-the-art methods with fast deployment running speed.

Index Terms—Convolutional neural network (CNN), compression artifacts reduction, compression post-processing, image enhancement, High Efficiency Video Coding (HEVC).

1. INTRODUCTION

High Efficiency Video Coding (HEVC) [1] is the state-of-the-art video coding technology, which is able to provide similar subjective quality at half the bitrate of H.264/AVC [2]. Thanks to its outstanding coding efficiency, HEVC has been increasingly applied to generate video streams for saving bandwidth, to avoid network congestion and to reduce costs in multimedia applications.

However, we pay for their high compression rate with complex compression artifacts, such as blocking artifacts, ringing artifacts, and blurring artifacts [3]. Block-based hybrid video coding will inevitably result in blocking artifacts, especially at the low bit rate or acute motions are contained in the input video. Ringing effects along the edges occur resulted from the coarse quantization of the high-frequency components. The removal of high frequencies causes blurring as well, but the blurring is less noticeable compared to the ringing artifacts. All these artifacts not only cause severe degradation on Quality of Experience (QoE), but also adversely affect various low-level image processing routines that take compressed images as input [4]. How to reduce compression artifacts has attracted more and more attention.

In order to alleviate unpleasant artifacts, HEVC employ in-loop filtering techniques to post-process the reconstructed images, where a de-blocking filter (DF) is done followed by a sample adaptive offset (SAO) filter [5]. DF is specifically designed to reduce blocking artifacts, which is a predefined low-pass characteristics non-linear filter, without signaling any bit to decoder sides. Unlike DF, SAO is designed for general compression artifacts, which correct the quantization errors by sending offset values to decoders, and reconstructs image by adding an offset. Zhang et al. [6] proposed a compression artifact reduction approach that utilizes both the spatial and the temporal correlation to form multi-hypothesis predictions from spatio-temporal similar blocks. Recently, Zhang et al. [7] incorporated the low rank regularization into HEVC in-loop filtering algorithm, and develop a non-local adaptive in-loop filter. However, those manually designed methods are insufficient for modeling compression artifacts, leaving a space for further improvement.

Deep convolutional neural networks (CNN) have become a widespread tool to address high-level computer vision tasks very successfully. Recently, they also achieved great success in low-level computer vision tasks, such as super-resolution [8], de-haze [9] and edge detection [10]. Inspired by these successes, it was also proposed utilizing CNN for artifacts reduction. Dong et al. first proposed an artifact reduction CNN (ARCNN) for reducing artifacts in JPEG compressed images [11]. ARCNN consists of four convolutional layers, feature extraction, feature enhancement, mapping and reconstruction layers, jointly in an end-to-end framework, and reported a remarkable jump compared to the previous most successful de-blocking oriented methods.

Based on ARCNN, Park and Kim proposed a new in-loop filtering technique using CNN (IFCNN), to replace the DF or SAO in HEVC [12]. IFCNN predicts the residues between the original image and the reconstructed image, and decoders only need to add IFCNN output to the reconstructed images. However, HEVC supports different sizes of discrete cosine transform (DCT), including 4×4, 8×8, 16×16 and 32×32, while JPEG adopts a uniform 8 × 8 DCT. Moreover, HEVC intra coding has 33 directional prediction models, much more complex than that in JPEG. In short, post-processing of video coding is much more complicated than JPEG images. But IFCNN is just a modification of ARCNN for video coding artifacts reduction, which leads to IFCNN not achieving very good results.

Furthermore, Wu et al. [13] proposed a variable-filter-size residue-learning CNN (VRCNN), which variable filter size is proposed to suit for HEVC variable block size transform. By concatenating different-sized filters, network can provide multi-level information of the input image. VRCNN adopts residual-learning and gradient clipping, and makes a fast convergence speed.

VRCNN outperforms the previously studied ARCNN or VDSR [14], in achieving lower memory cost, and speedup deployment. More recently, Wang et al. [15] propose a decoder-side scalable CNN (DSCNN) approach to achieve quality enhancement for HEVC, which does not require any modification of the encoder. In DSCNN, a scalable structure with two sub-networks is designed to make a trade-off between computational complexity and quality enhancement.

All these works seem to open up a new direction that adopts CNN into video coding to further improve the coding efficiency. Despite the remarkable progress, existing state-of-the-art methods cannot generate intermediate predictions at multiple refining levels. As a result, one needs to train a large variety of models for various applications with different desired quality enhancement and computational loads. To address these drawbacks, we propose the Multi-level Progressive Refinement Networks through an adversarial training approach (MPRGAN). The main contributions of this study as follows:

1) We formulate a multi-level progressive refinement network (MPRGAN) for solving video post-processing tasks, through an adversarial training process, which can efficiently suppress various compression artifacts.

2) A scalable structure is explicitly included in our MPRGAN, through progressively predicts the sub-band residues in a coarse-to-fine fashion. Hence, the computational complexity and quality enhancement of our MPRGAN model is adjustable by simply bypassing the computation of residues at finer levels. This makes our MPRGAN can accommodate the computational resources of the hardware.

3) Extensive experiments demonstrate the effectiveness of our MPRGAN over state-of-the-art methods [13, 15] both subjective and objective visual quality of the reconstructed intra frame through HEVC.

The rest of the paper is organized as follows: Section 2 illustrates our proposed MPRGAN architecture, with details of the implementation. In section 3, extensive experiments are conducted to evaluate MPRGAN. Finally, we conclude this work with some future directions in section 4.

2. FRAMEWORK OF MPRGAN

2.1 Network architecture

We propose MPRGAN through an adversarial training approach, for artifacts reduction and coding efficiency improvement in intra frame coding. The MPRGAN does not require signaling bits by using the same trained weight in both encoder and decoder. Fig. 1 illustrates the framework of the proposed MPRGAN, which is composed of two CNNs, i.e., the enhancement network E and the discriminative network D. We train the MPRGAN via adversarial training approach, enabling it has capable of modeling complex multi-modual distribution, thus to boost its performance, and to be able to generate sharper images.

The enhancement network E is proposed to predict the residues between the original frame y (ground truth) and the input frame x (reconstructed through HEVC). The residues predicted by the multi-level modules are added back to the input, generate the enhanced frame $e = E(x)$. Following the design principle of VGG net [16], network E using a stack of Convolution-ReLU layers without BatchNorm. We pad zeros before convolutions to keep the sizes of all feature maps (including the output image) the same. Note that the output layer of the network E, 1×1 convolutional layers, does not use activation functions.

The discriminative network D employs popular ResNet architecture that consists of Convolution-BatchNorm-LeakyReLU layers. E and D are trained in a competing fashion, discriminator D is trained to distinguish positive example $x|y$ and negative example $e|y$, where $\cdot|\cdot$ indicates channel-wise concatenation. In contrary, the enhancement network E tries to confuse the discriminative network D by generating more and more "realistic" samples. This minimax game can be formulated as

$$\min_E \max_D U(E,D) = \mathrm{E}_y[\log D(x|y)] + \mathrm{E}_x[\log(1 - D(E(x)|y))] \quad (1)$$

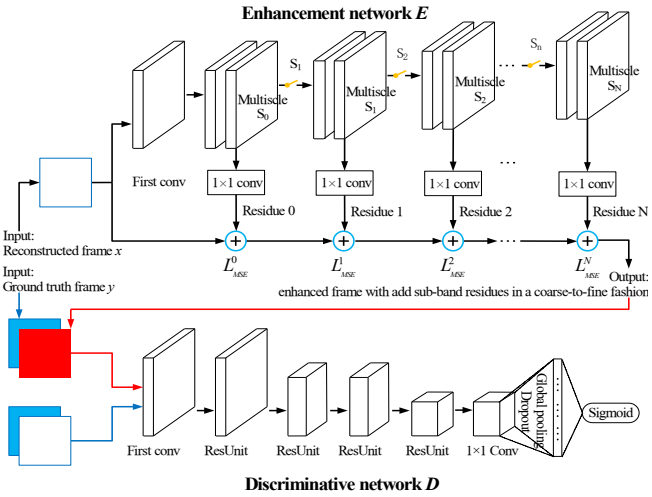

Fig. 1: An overview of our MPRGAN, through progressively predicting the sub-band residuals in a coarse-to-fine fashion.

2.2 Scalable structure of enhancement network *E*

To meet the varying computational resources of different hardware, we propose MPRGAN, explicitly includes a scalable structure. Only a single complete model is trained, enhancement network E can achieve scalable quality enhancement, through progressively predicting the sub-band residues in a coarse-to-fine fashion.

As shown in Fig. 1, we use a two-layer 3×3 convolution across 64 channels as the basic multi-level building module, and adopt switches $\{S_i\}$ to control the computational complexity of enhancement network E in deployment testing. At each level, we first apply a cascade of convolutional layers to extract feature maps. Then, we use 1×1 convolutional layers to predict the sub-band residues. Finally, the predicted residues at each level are utilized to enhance the reconstructed frame through additional operations.

Note that switches $\{S_i\}$ decide whether to enable the convolutional layers of multi-level S_i. Once the computational resources are not sufficient, the deeper multi-level layers switches $\{S_i\}_{i=L}^{N}$ are turned off and bypassed, thereby only shallower convolution layers $\{S_i\}_{i=0}^{L-1}$ are used for refining the reconstructed frame. That is, the final enhance frame is:

$$e = E(x) = x + \sum_{i=0}^{L-1} residue_i, \quad L \in [1, N] \quad (2)$$

When the computational resources are sufficient, $\{S_i\}_{i=1}^{N}$ are turned on, MPRGAN starts to progressively predict the sub-band residues in a coarse-to-fine fashion work, based on the multi-level residue $\{residue_i\}_{i=0}^{N}$ which is output from the 1×1 convolutional layers of enhancement network E. That is to say, once

computational resources are sufficient, the quality of refined images can be further enhanced.

2.3 Loss functions

By penalizing the discrepancy between the enhanced output frame and ground-truth frame, optimal MPRGAN can be trained to discover the mapping from the input compressed image with artifacts to the quality enhanced image.

In our enhancement network *E*, multi-level mean squared error (MS-MSE) loss function is employed to calculate the discrepancy of the enhanced image and the original image, i.e.,

$$L^i_{MSE} = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}(e^i_{wh} - y_{wh})^2 \quad (3)$$

$$L_{MSE} = \sum_i \lambda_i L^i_{MSE} \quad (4)$$

Where, function L^i_{MSE} measures the discrepancy between the ground truth frame y and multi-level enhanced frame e^i that generated by the shallower convolution layers $\{S_i\}$ of the enhancement network *E*. $\{\lambda_i\}_{i=0}^{N}$ are hyper-parameters balancing the influence of *N* different sub-band residues.

Based on the aforementioned adversarial loss and MSE loss, we develop the MPRGAN adversarial framework. Network *E* and *D* play a minimax game by optimizing different loss functions. The loss function of enhancement network L_E and the loss function of discriminative network L_D are formally defined as

$$L_E = -\beta \times \log(D(E(x)\,|\,y)) + \sum_i \lambda_i L^i_{MSE} \quad (5)$$

$$L_D = -\log(D(x\,|\,y)) - \log(1 - D(E(x)\,|\,y)) \quad (6)$$

These loss functions are minimized by stochastic gradient decent algorithm with the standard back-propagation. We learn the network *E* and *D* parameters by minimizing $L_E(\theta)$ and $L_D(\phi)$ alternately.

2.3 Implementation and training details

Our training and validation sets are the same as VRCNN and DSCNN, which are selected from BSD500 database [17]. Specifically, VRCNN and DSCNN [15] were trained on the 400 train and test images of the BSD500 dataset and tested on the 100 remaining validation images. All training images are encoded by HEVC all intra (AI) mode, using HM 16.0 [18]. We decompose the ground-truth and HEVC reconstructed frames into image patches with the size of 32×32, using the stride of 16. Note, only the luminance channel is considered for training.

We use the publicly available code of Caffe [19] for training MPRGAN, on a NVIDIA GeForce GTX 1060 graphical processing unit (GPU). For each QP, a separate network is trained out. We optimize the network parameters with Adam [20] for a total of 200 epochs. Weight initialization using MSRA [21], momentum is set to 0.9, and weight decay is 0.0001. We start from a learning rate of 10^{-3}, and divided it following a linear decay over training iteration.

3. EXPERIMENTAL RESULTS

In this subsection, we evaluate the performance of quality enhancement of our MPRGAN, comparing with the latest state-of-the-art methods [13][15]. We don't evaluate the ARCNN [11] and VDSR [14], which have been beaten by the VRCNN and DSCNN.

We integrate the enhancement network *E* into HM 16.0, and test our approach on test sequences from JCT-VC database [22]. Here, we trained an enhancement network *E* consists of 4 multi-level modules $\{S_i\}_{i=0}^{3}$. Note that non-overlapping with the training and testing sets, this can more objectively reflect the generalizability of the trained network. For fair comparison, we use the same test set and test method as that in [15], and comparing with the original experimental results in [15]. The test set includes *BQTerrace*, *BasketballDrive* and *Cactus* from Class B, *BQMall* from Class C, *BasketballPass* from Class D and *FourPeople*, *Johnny*, *Vidyo1* from Class E. Quality enhancement is measured by Y-PSNR improvement (ΔPSNR).

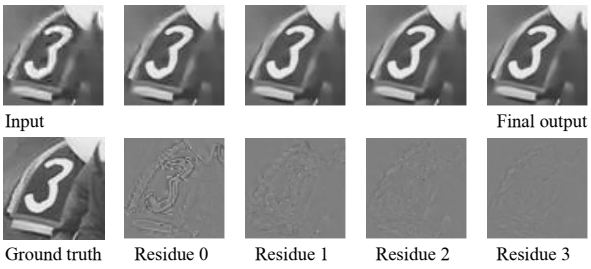

Fig. 2. Multi-level progressive refinement processing of network *E*. The top row presents the HEVC reconstructed image (left) and the multi-level enhanced result obtained by adding the sub residues estimated up to the input image. The second row presents the ground truth image (left) and the sub residuals predicted by intermediate multi-level layers. The magnitude of residue images has been scaled ×3 for display purpose. We encourage the reader to zoom-in onto the images to best view the fine details.

Table 1. The quality enhancement (ΔPSNR dB at QP42) and deployment time (seconds per CTU) of various methods.

Network	ΔPSNR	Deployment time	
		CPU time	GPU time
VRCNN	0.2928	0.1160	0.0190
DSCNN	0.3504	0.5089	0.0908
MPRGAN S_0	0.2805	0.0630	0.0100
MPRGAN S_{01}	0.3501	0.1131	0.0183
MPRGAN S_{012}	0.3886	0.1652	0.0285
MPRGAN S_{0123}	0.4025	0.2270	0.0370

3.2 Multi-level coarse-to-fine work fashion

Enhancement network *E* employs a scalable structure, through progressively predicting the sub-band residues in a coarse-to-fine fashion. Fig. 2 illustrates the multi-level progressive prediction process of enhancement network *E*. One insight is that each layer in the network *E* removes part of the artifacts in the image, rather than removing it all at once at the end of the network *E*. It can be observed from Fig. 2 that, the multi-level layer S_0 can deal with most of obvious artifacts. While deeper multi-level layers, e.g. S_3, mainly focus on recovering and enhancing the details and textures. This may be explained by the fact that the deeper layers correspond to a larger receptive field, and therefore may refine in a better way of global pattern. Such as details that may be indistinguishable from artifacts if viewed just in the context of a small local patch.

We also compare the computational complexity of different networks. The quality enhancement and the networks deployment time are summarized in Table 1; these results are the average results of all test sequences. "MPRGAN S_0" indicates the network *E* with deeper multiple layers switches $\{S_i\}_{i=1}^{3}$ are turned off, and only shallower convolution layers S_0 are used for refining the

reconstructed frame. "MPRGAN S_{0123}" indicate the network E with deeper multiple layers $\{S_i\}_{i=0}^{3}$ are used for refining the reconstructed frame. Overall, only needing to train a single complete model "MPRGAN S_{0123}", enhancement network E can achieve scalable quality enhancement, through turn off and bypassing the deeper multi-level layers. It can be seen from Table 1 and Fig. 3 that, from "MPRGAN S_0" to "MPRGAN S_{0123}", with the increase of computing resources, our MPRGAN network can boost the quality of enhancement continually.

It can be seen from Table 1, our MPRGAN S_{01} achieves 19.5% extra PSNR improvement comparing with VRCNN, with the similar computational complexity. While our MPRGAN S_{0123} achieves up to 37.5% extra PSNR improvement comparing with VRCNN, at the cost of ~1.0 times increment of computational complexity. Comparing with DSCNN, our MPRGAN S_{01} saves 77.6% computational complexity, with the analogous quality enhancement. While our MPRGAN S_{0123} outperforms the DSCNN by a large margin both in quality enhancement and deploy speed.

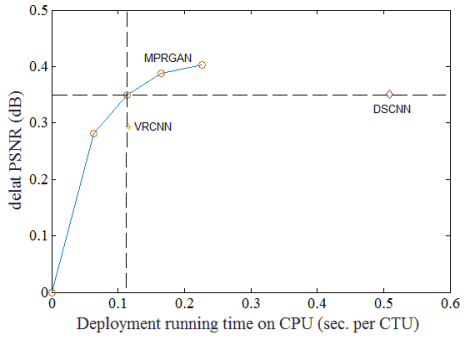

Fig. 3. Deployment speed vs. refinement quality.

Table 2. Performance of various methods (ΔPSNR dB)

QP	Class	Sequence	VRCNN	DSCNN	MPRGAN
42	B	BQTerrace	0.3127	0.3789	**0.3978**
	B	Cactus	0.1754	0.2001	**0.2298**
	B	BaskeballDrive	0.1776	0.2281	**0.2849**
	C	BQMall	0.2946	0.3433	**0.3971**
	D	RaceHorses	0.4117	0.4320	**0.5808**
	E	FourPeople	0.4060	0.4791	**0.5124**
	E	Johnny	0.2823	0.3363	**0.3881**
	E	Vidyo1	0.3619	0.4086	**0.4290**
		Average	0.2928	0.3504	**0.4025**
47		Average	0.2940	0.3413	**0.4001**

3.3 Performance of quality enhancement for HEVC

Performance of objective quality. It can be seen from Table 2, the proposed MPRGAN obviously outperforms DSCNN and VRCNN over all test sequences. At QP = 42, the averaged ΔPSNR (0.4025 dB) of our MPRGAN is 14.9% higher than DSCNN (0.3504 dB) and 37.5% higher than VRCNN (0.2928 dB). In particular, MPRGAN achieves up to 0.5808 dB, while the highest improvements of DSCNN and VRCNN are 0.4791 dB and 0.3612 dB, respectively. Similar results also can be found at QP = 47. In summary, our proposed MPRGAN performs best in quality enhancement of HEVC intra coding among three approaches.

Performance of subjective quality. We also compare the visual quality of refined images as shown in Fig. 4. In contrast to Fig. 4(a), (b) and (c), we can observe that image processed by the original DF and SAO filter in HEVC baseline, greatly reduces blocking at smooth regions, but in-sufficient largely along edges and high-frequency regions. From Fig. 4(d) and (e), we can observe that VRCNN and DSCNN effectively reduce the most artifacts caused by HEVC compression, and produces better visual quality than HEVC baseline. From Fig. 4(f), we could see that the result of our MPRGAN could produce much sharper edges with much less blocking and ringing artifacts compared with VRCNN and DSCNN, contrast along edges is better enhanced. Hence, the effectiveness of our MPRGAN for video coding quality enhancement can be validated.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel multi-level progressive refinement network with an adversarial training approach (MPRGAN) for video coding post-processing. We also designed a scalable structure in our MPRGAN for achieving the trade-off between quality enhancement and computational complexity, which improves the flexibility of our MPRGAN and makes it adjustable to the resource-aware applications. The proposed network MPRGAN shown to reduce artifacts in a coarse-to-fine work fashion, and outperforms other state-of-the-art quality enhancement approaches on the HEVC standard test dataset. Our future work is planned to further simplify the network while boosting its visual quality, and aim to extend the methodology to more related applications.

(a) Ground-truth (b) Before post-processing, 27.5849 dB

(c) HEVC baseline, 27.7746 dB (d) VRCNN, 28.2565 dB

(e) DSCNN, 28.3021 dB (f) MPRGAN (ours), 28.3554dB

Fig. 4. The first frame of class D RaceHorses, encoded by HEVC at QP 42, and post-processed by HEVC baseline as well as different learning methods. This figure is best zoom in to see details on the screen.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China, under Grants U1301257, 61571285, and 61422111, and project of Shanghai Science and Technology Commission under Grant 17DZ2292400.

REFERENCES

[1] G. Sullivan, J. Ohm, W. Han, T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," IEEE Transactions on Circuits and Systems for Video Technology, 22 (12), pp. 1649–1668, 2012.

[2] T. Tan, R. Weerakkody, M. Mrak, et al., "Video quality evaluation methodology and verification testing of HEVC compression performance," IEEE Transactions on Circuits and Systems for Video Technology, 26 (1), pp. 76–90, 2016.

[3] M. Yuen, H. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," Signal Processing, 70(3), pp. 247-278, 1998.

[4] C. Dong, C. Chen, K. He, X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," IEEE Transactions on Pattern Analysis & Machine Intelligence, 38 (2), pp. 295, 2016

[5] C. Fu, E. Alshina, A. Alshin, et al., "Sample adaptive offset in the HEVC standard," IEEE Transactions on Circuits and Systems for Video Technology, 22(12) , pp. 1755-1764, 2012.

[6] Zhang X., Xiong R., Lin W., Ma S., Liu J., et al., "Video Compression Artifact Reduction via Spatio-Temporal Multi-Hypothesis Prediction," IEEE Transactions on Image Processing, vol. 24(12), pp. 6048-6061, 2015.

[7] Zhang X., Xiong R., Lin W., Zhang J., Wang S., Ma S., et al., "Low-Rank based Nonlocal Adaptive Loop Filter for High Efficiency Video Compression," IEEE Transactions on Circuits & Systems for Video Technology, vol. 27, pp. 2177-2188, 2017.

[8] W. Shi, J. Caballero, F. Huszár, et al., "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1874–1883, 2016.

[9] K. Zhang, W. Zuo, Y. Chen, "DehazeNet: An End-to-End System for Single Image Haze Removal," IEEE Transactions on Image Processing, 25 (11), pp. 5187-5198, 2016

[10] S. Sarkar, V. Venugopalan, K. Reddy, J. Ryde, "Deep Learning for Automated Occlusion Edge Detection in RGB-D Frames," Journal of Signal Processing Systems, 88 (2), pp. 1-13, 2017

[11] C. Dong, Y. Deng, C. C. Loy, et al., "Compression artifacts reduction by a deep convolutional network," IEEE International Conference on Computer Vision, 71 (2), pp. 576-584, 2015.

[12] W. Park, M. Kim, "CNN-based in-loop filtering for coding efficiency improvement," IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). pp. 1-5, 2016

[13] Y. Dai, D. Liu, F. Wu, "A Convolutional Neural Network Approach for Post-Processing in HEVC Intra Coding," International Conference on Multimedia Modeling, pp.28-39, 2017

[14] J. Kim, J. Lee, K. Lee, "Accurate image super-resolution using very deep convolutional networks," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646-1654, 2016

[15] R. Yang, M. Xu, Z. Wang, "Decoder-side HEVC quality enhancement with scalable convolutional neural network," IEEE International Conference on Multimedia & Expo, 2017

[16] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," In International Conference on Learning Representations (ICLR), 2015.

[17] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, "Contour detection and hierarchical image segmentation," IEEE Transactions on Pattern Analysis & Machine Intelligence, 33(5), PP. 898–916, 2011.

[18] https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.0

[19] Y. Jia, E. Shelhamer, J. Donahue, et al., "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint* arXiv: 1408.5093, 2014.

[20] D. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint* arXiv: 1412.6980, 2014.

[21] K. He, X. Zhang, S. Ren, J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *arXiv preprint* arXiv: 1502.01852, 2015

[22] F. Bossen et al., "Common test conditions and software reference configurations," Joint Collaborative Team on Video Coding (JCT-VC), JCTVC-F900, 2011