A Perceptual Quality Enhancement Method for Video Coding

Xiaoran Qin, Chenghua Li, Xiaoyi Dong, Yu Zhu, Zhenyu Guo, Cong Leng, Jian Cheng
Institute of Automation, Chinese Academy of Sciences
xiaoran.qin@ia.ac.cn

Abstract

This paper describes the technical scheme of team Foo in the video compression track of Workshop and Challenge on Learned Image Compression (CLIC) 2022. Our method includes a VVC/H.266 codec and a quality enhancement network. Firstly, considering the coding efficiency and target bitrates of this challenge, we use the VVC codec and adopt adaptable configuration parameters for each video. Then, we propose a quality enhancement network supervised by multiple objective and perceptual losses to postprocess the decoded frames for high quality video restoration. Compared to the VVC codec configured with default parameters, the proposed method improves both PSNR and SSIM.

1. Introduction

VVenC/VVdeC software [11, 12] is an open-source encoder and decoder implementation of VVC/H.266 [1], the newest generation of video coding standards. Due to additions and improvements of various coding tools, VVC achieves about a 50% bit-rate reduction over its predecessor, HEVC/H.265 [10]. Consequently, the VVC codec with high coding efficiency is well suitable to serve as the basis of our solution. In the meanwhile, considering that the compressed videos are judged by human raters, VVenC including subjective quality enhancement techniques are chosen for higher perceptual quality instead of VTM [9], which is the official test model of VVC.

Similar to previous video coding standards since H.26x [10], VVC is based on the hybrid video coding principle, combining prediction and transform coding of the quantized prediction residual to reduce redundancy in the video signal. When encoding a frame in a specific video sequence, the encoder first divides the frame into blocks with various sizes adaptively according to local statistics, and then for each block, the predictions of the samples are generated by inter-picture prediction or intra-picture prediction referring to the samples in the temporally colocated block or spatial neighboring samples respectively. After that, the difference between the prediction and the original input video signal will be transformed, quantized and eventually encoded into the binary bitstream together with other necessary coding information using Context Adaptive Binary Arithmetic Coding (CABAC).

2. Proposed Method

The proposed method includes a VVC codec and a quality enhancement network. For the VVC codec, the VVenC/VVdeC software is adopted. In order to obtain decoded videos with high quality under bitrate requirements, we select adaptable configuration parameters and an appropriate quantization parameter (QP) for each video. Subsequently, a quality enhancement network is proposed to improve the reconstruction quality of the decoded videos. This quality enhancement network is based on EDVR [13] model and supervised by multiple objective and perceptual losses, which is inspired by [15].

2.1. Searching the optimal QP for each video

To satisfy the bitrate constraint of the challenge and improve quality of the decoded videos as much as possible, it is crucial to assign an appropriate value to the QP that determines the step size of the quantization process and controls the fidelity and bitrate of video compression. As a larger QP lowers the bitrate but also deteriorates the quality, the optimal QP is the minimal one that fulfils the bitrate requirement. At the same time, it is notable that allocating an identical QP value to each video in the video set is unsensible, because complex and dynamic videos need more bitrate to be encoded than simple and static videos at the same compression quality. To maximize the overall quality of the entire video set under the bitrate constraint, we apply a dynamic programming algorithm inspired by the optimal bit allocation algorithm in [4] to search the optimal bitrate and QP for each video.

First, we calculate the weight of each video w_k using

$$w_k = \frac{n_k}{\sum_k n_k}, \tag{1}$$

where n_k is the number of pixels for each frame in the k_{th} video in the video set.

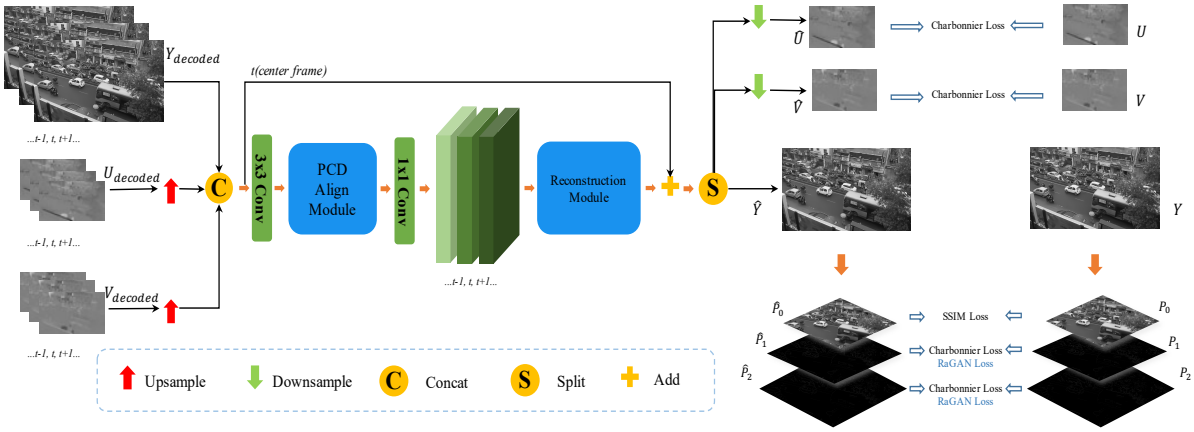

Figure 1. The architecture of the quality enhancement network, which is optimized by multiple objective and perceptual losses

Second, we define $dp_{i,j}$ as the maximum weighted SSIM of the first i videos with the total bitstream size no more than j given by

$$dp_{i,j} = \max\left\{\sum_{1\le k\le i} w_k * SSIM_k\right\} s.t. \sum_{1\le k\le i} size_k \le j, \tag{2}$$

where w_k, $SSIM_k$ and $size_k$ are the weight, SSIM and bitstream size of the k_{th} video respectively.

From the definition mentioned above, it is not difficult to gain the state transition equation as

$$dp_{i,j} = \begin{cases} 0 & i * j = 0 \\ \max\{dp_{i-1,\ j-size_{i,m}} & \\ +w_i * SSIM_{i,m} | m \in M\} & \text{else,} \end{cases} \tag{3}$$

where $SSIM_{i,m}$ and $size_{i,m}$ represent the SSIM and bitstream size of the i_{th} video when using the m_{th} QP value (QP_m) to encode, and m belongs to a predefined QP range M. And the optimal QP choice $choice_i$ for the i_{th} video is calculated as

$$choice_i = QP_{\arg\max_m \{dp_{i-1,j-size_{i,m}} + w_i * SSIM_{i,m} | m \in M\}}. \tag{4}$$

Finally, to obtain the unknown $SSIM_{i,m}$ and $size_{i,m}$ in Eqs. (3) and (4), we select a specific QP range M to guarantee a relatively large search space, and then each video in the video set is encoded by each QP in this range. In this way, we acquire corresponding bitstream size and SSIM under different QP settings for each video. After that, all data needed for the algorithm have been ready and the optimal QP for each video can be readily calculated by means of the dynamic programming algorithm.

2.2. VVenC coding configuration

In addition to QP, some other configuration parameters are also adapted to the specific video:

- Input, output and size are set to the corresponding values of the video;
- Gopsize is set to 32 for static videos and this value changes to 16 when encoding dynamic videos since there is more difference between the adjacent frames;
- Intraperiod is set to 64 after a tradeoff between bitrate (longer intraperiod) and picture quality (shorter intraperiod);
- Internal-bitdepth is set to 8.

The remaining configuration parameters maintain the default values in VVenC.

2.3. The quality enhancement network

The network architecture is shown in Fig. 1. Based on the EDVR model, we maintain its Pyramid, Cascading and Deformable (PCD) alignment module and reconstruction module, and remove other modules considering a tradeoff between efficiency and model size. Given $2N + 1$ consecutive decoded frames $I_{[t-N,t+N]}$, the quality enhancement network aims to enhance the center frame I_t referring to its neighboring frames. For each input frame, its two chrominance channels are first upsampled and then concatenated with its luminance channel to form a three-channel input frame. A 3×3 convolution layer extracts features for each three-channel input frame. Then, the PCD alignment module aligns features of each neighboring frames to that of the target center frame following pyramidal structure and cascading refinement principles, which is designed to process large and complex motions and replace optical flow network. Then the aligned features of each frame are fused by a 1×1 convolution layer. Next, a reconstruction module composed of several residual blocks is performed and the last convolution layer generates a three-channel frame residual. Finally, the enhanced frame is obtained by adding

the predicted frame residual to the three-channel input of center frame. The enhanced frame is split, where luminance channel $\hat{Y}$ is generated and chrominance channels ($\hat{U}$ and $\hat{V}$) need to be downsampled.

2.4. Objective and perceptual losses

During training, the quality enhancement network is optimized by multiple objective and perceptual losses as shown in Fig. 1, which is inspired by [15]. Because luminance channel contains most of the texture information and details, it is reasonable to add more supervision on luminance channel than that on chrominance channels. The predicted $\hat{Y}$ is decomposed to a Laplacian pyramid [2] consisting of three components, where low-frequency component $\hat{P}_0$ contains global luminance and image structure and high-frequency components $\{\hat{P}_1, \hat{P}_2\}$ contain multi-scale details. Correspondingly, the ground truth luminance channel Y is also decomposed to a three-layer Laplacian pyramid to generate three components $\{P_0, P_1, P_2\}$. To reduce the difference between $\hat{P}_0$ and P_0, we adopt the SSIM loss [14] $\mathcal{L}_s$ which focuses on structure similarity, which is defined as

$$\mathcal{L}_s = 1 - SSIM(\hat{P}_0, P_0). \tag{5}$$

To reconstruct multi-scale details, we adopt the Charbonnier loss [8] on high-frequency components as

$$\mathcal{L}_d = \sqrt{\|\hat{P}_2 - P_2\|^2 + \epsilon^2} + \sqrt{\|\hat{P}_1 - P_1\|^2 + \epsilon^2}, \tag{6}$$

where $\epsilon = 10^{-3}$. For chrominance channels, we also use the Charbonnier loss on each channel as

$$\mathcal{L}_{uv} = \sqrt{\|\hat{U} - U\|^2 + \epsilon^2} + \sqrt{\|\hat{V} - V\|^2 + \epsilon^2}. \tag{7}$$

To generate realistic texture and improve subjective quality of enhanced frames, we use generative adversarial network [3] during training. The quality enhancement network is known as the generator, and a PatchGAN [5] discriminator is proposed. Additionally, we enhance the discriminator based on the Raletivistic average GAN (RaGAN) [6], which predicts the probability that a real data is relatively more realistic than a fake data. The RaGAN loss is proposed on high-frequency components of the Laplacian pyramid of luminance channel. The adversarial loss is defined as

$$L_G = \sum_i \{-\mathbb{E}_{P_i}[\log(1 - D_{Ra}(P_i, \hat{P}_i))] - \mathbb{E}_{\hat{P}_i}[\log(D_{Ra}(\hat{P}_i, P_i))]\}, \tag{8}$$

and the discriminator loss is defined as

$$L_D = \sum_i \{-\mathbb{E}_{P_i}[\log(D_{Ra}(P_i, \hat{P}_i))] - \mathbb{E}_{\hat{P}_i}[\log(1 - D_{Ra}(\hat{P}_i, P_i))]\}, \tag{9}$$

where $i \in \{1, 2\}$ and D_{Ra} is the Raletivistic average Discriminator defined as $D_{Ra}(P_i, \hat{P}_i) = \sigma(C(P_i) - \mathbb{E}_{\hat{P}_i}[C(\hat{P}_i)])$, σ is the sigmoid function and $C(x)$ is the non-transformed layer.

Therefore, the final loss is calculated as

$$L_{final} = L_s + L_d + L_{uv} + \lambda(L_G + L_D), \tag{10}$$

where λ is a hyper-parameter to weight the RaGAN loss.

3. Experiments

3.1. Implementation Details

The original $.mp4$ video sequences are first converted to $.yuv$ sequences, and then are encoded with the VVenC software to produce target bitstreams. To determine the optimal QP for each video, we select a QP range from 31 to 37 for the high bitrate track (1mbps) and a QP range from 46 to 52 for the low bitrate track (0.1mbps).

The quality enhancement network takes three ($N = 1$) frames as input. The PCD alignment module adopts 5 residual blocks to extract features before constructing its pyramidal structure. The reconstruction module adopts 10 residual blocks. To train the network, we utilize 262 videos from the CLIC2022 validation set (excluding the 30 videos needed for the validation phase) as training dataset and produce compressed frames by the VVenC/VVdeC software with adaptable coding configuration in Sec. 2.2. The QPs are set to 32, 34 and 36 (1mbps) and 47, 49 and 51 (0.1mbps) during the training dataset compression. Therefore, we train two quality enhancement networks with different weights for 1mbps and 0.1mbps tracks.

We train the network in two steps. For the first step, only objective losses $L_s + L_d + L_{uv}$ are used to pretrain a model focusing on details reconstruction, and for the second step, we initialize the model weights by those from the pretrained one and utilize L_{final} to improve the perceptual quality. During training, we use patches of size 256×256 as input and augment the training data with random cropping, flipping, 90° rotations and the CutBlur [16] technique. The batch size is set to 32. The network is optimized by the Adam [7] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate starts at $2e^{-4}$ for the first step and $5e^{-5}$ for the second step, and decays with a cosine learning rate decay strategy. All the networks are trained for 150000 iterations. During inference for the validation phase, we divide frames of 1080p to patches of size 1080×1080 with an overlap of size 240×240 because of the limit of GPU memory. Particularly, for frames of 720p, we adjust the size of patches to 720×1080 according to the length of the shortest side of each frame.

3.2. Results

We evaluate our proposed method on the 30 videos needed for the validation phase. As shown in Tabs. 1 and 2, the proposed method achieves higher PSNR and SSIM compared to the VVenC/VVdeC configured with identical QP for each video and default parameters. The optimal QP

Method	Datasize	PSNR	SSIM
VVenC/VVdeC (QP=34)	33170549	35.557	0.9613
Ours (optimal QP)	37465824	36.283	0.9657
Ours (optimal QP + S_{fst})	37465824	**36.412**	**0.9673**
Ours (optimal QP + S_{sec})	37465824	36.162	0.9641

Table 1. Performance of the submitted 30 videos in 1mbps track using different methods. S_{fst} means the first step training of the quality enhancement network with objective losses $L_s+L_d+L_{uv}$. S_{sec} means the second step training with L_{final} loss.

Method	Datasize	PSNR	SSIM
VVenC/VVdeC (QP=49)	3297603	27.797	0.8396
Ours (optimal QP)	3648583	28.565	0.8652
Ours (optimal QP + S_{fst})	3648583	**28.660**	**0.8705**
Ours (optimal QP + S_{sec})	3648583	28.583	0.8670

Table 2. Performance of the submitted 30 videos in 0.1mbps track using different methods.

strategy makes better use of avaliable bitstream sizes under challenge requirements and obtains more reasonable allocation of bitrates between different videos. The first step training of the quality enhancement network with objective losses $L_s + L_d + L_{uv}$ gains better performance. The second step training with L_{final} loss leads to a relatively lower PSNR score but achieves better subjective quality. Specifically, S_{fst} produces frames that are relatively smooth and miss details, while S_{sec} produces frames that preserve more textures. And because PSNR score reflects accuracy in pixel level, the optimization in perception level often leads to a decrease in PSNR score. For the validation phase, our team Foo achieves PSNR scores of 36.412 in 1mbps track and 28.660 in 0.1mbps track.

4. Conclusion

In this paper, we propose a perceptual quality enhancement method for video coding. The video coding standard VVC is adopted to generate decoded frames by leveraging both the optimal QP searching strategy and adaptable coding configuration. Then the quality enhancement network postprocesses the VVC-decoded frames to produce reconstructed frames with high subjective quality. Experiments in both 1mbps and 0.1mbps demonstrate the effectiveness of the proposed method.

References

[1] B. Bross, J. Chen, J. R. Ohm, G. J. Sullivan, and Y. K. Wang. Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc). In *Proceedings of the IEEE*, pages 1–31, Jun 2021. 1

[2] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987. 3

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[4] Z. Huang, K. Lin, C. Jia, S. Wang, and S. Ma. Beyond vvc: Towards perceptual quality optimized video compression using multi-scale hybrid approaches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1866–1869, Jun 2021. 1

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[6] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 3

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[8] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 3

[9] VTM reference software for VVC. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM. 1

[10] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 1

[11] Fraunhofer Versatile Video Decoder (VVdeC). https://github.com/fraunhoferhhi/vvdec. 1

[12] Fraunhofer Versatile Video Encoder (VVenC). https://github.com/fraunhoferhhi/vvenc. 1

[13] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[14] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[15] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4781–4790, 2021. 1, 3

[16] Jaejun Yoo, Namhyuk Ahn, and Kyung-Ah Sohn. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8375–8384, 2020. 3