A MULTISCALE GRADIENT-BACKPROPAGATION OPTIMIZATION FRAMEWORK FOR DEFORMABLE CONVOLUTION BASED COMPRESSED VIDEO ENHANCEMENT

Yanbo Gao^1, Menghu Jia^2, Shuai Li^3, Xun Cai^1, Mao Ye^2, Frédéric Dufaux^4

^{1}School of Software, Shandong University, China
^{2}Sch. Info. & Comm. Eng., University of Electronic Science and Technology of China, China
^{3}School of Control Science and Engineering, Shandong University, China
^{4}Laboratoire des signaux et Systèmes, CentraleSupélec, CNRS, Université Paris-Saclay

ABSTRACT

Deep learning based compressed video quality enhancement has raised lots of interest recently. To explore the information over multiple frames, deformable convolution has been used for temporal alignment. However, in the existing methods, the deformable convolution is used in a relatively naïve way, without differing the characteristics of offset and features, and their behavior in gradient backpropagation. In this paper, a multiscale gradient-backpropagation optimization framework is proposed for the deformable convolution based compressed video quality enhancement. By analyzing the gradient backpropagation mechanism of deformable convolution, a multi-scale deformable convolution alignment structure is developed to facilitate the gradient backpropagation at all scales. Moreover, a progressive offset prediction module is developed, which decouples the offset prediction from the feature up-sampling, thus reducing the noise flow over scales. Experimental results show that the proposed method achieves the state-of-the-art performance, with 25.6% BD-rate saving compared to the HEVC reference software (HM).

Index Terms— Video coding, Quality enhancement, Deformable convolution

1. INTRODUCTION

With the popularity of online videos and high-resolution videos, video encoding standards are also evolving to achieve high compression ratio, including the High Efficiency Video Coding (HEVC/H.265) [1] and the newest Versatile Video Coding (VVC/H.266) [2]. These encoding standards are developed based on the hybrid block-based encoding architecture. While these compression methods greatly reduce bit rate, they also introduce various artifacts, such as blocking, blurring and ringing artifact, which deteriorates the quality of the compressed video.

In order to reduce these artifacts, a large number of quality enhancement algorithms have emerged, especially methods based on deep learning [3-15] due to its great success in image/video denoising. While there are methods working on enhancing each frame solely based on the spatial information and treating video only as a collection of pictures, a large portion of the work directly focuses on enhancing the video with multiple frames considering the temporal information. One crucial step in the multi-frame enhancement algorithm is the temporal alignment, i.e., locating the relevant information for the current frame from the neighboring frames. Some earlier works proposed to use optical flow to align the reference frames to the current frame, such as the MFQE [12], SpyNet [16], TOFlow [17] and PWC-Net [18] etc. However, for the compressed video enhancement, it has been observed that the aligned frames warped by optical flow are usually not very accurate and even appear unwelcomed artifacts, damaging the performance. In addition, Kernel Prediction Network (KPN) [19] has also been used to deal with multi-frame offset without explicit alignment. It generates per-pixel filtering kernels, and registers, averages and denoises the sequence images at the same time. In [20], a spatio-temporal network based on filter adaptive convolutional (FAC) layers was proposed for video deblurring. The generated filters and FAC layers can achieve temporal alignment and deblurring.

Inspired by the superior performance of deformable convolution [21] in alignment, many video enhancement and restoration tasks have utilized deformable convolution to align and fuse temporal information. For example, in [13], a novel Spatio-Temporal Deformable Fusion (STDF) scheme was proposed to aggregate temporal information. In STDF, Unet was employed in offset prediction network and the offset was predicted in one step. For videos with large motion, such simple architecture may not be able to locate the correct motion. In [22], a pyramid cascading structure based deformable convolution are introduced into the alignment module to effectively align videos with large motion. However, the offset residuals at each level are learned from original features, without the participation of any upper-level warped features.

Underlying mechanism for effective alignment of deformable convolution is still unclear, especially considering its better performance over the optical flow. In [23], it is argued that the mechanism of deformable convolution is the same as that of optical flow, and its

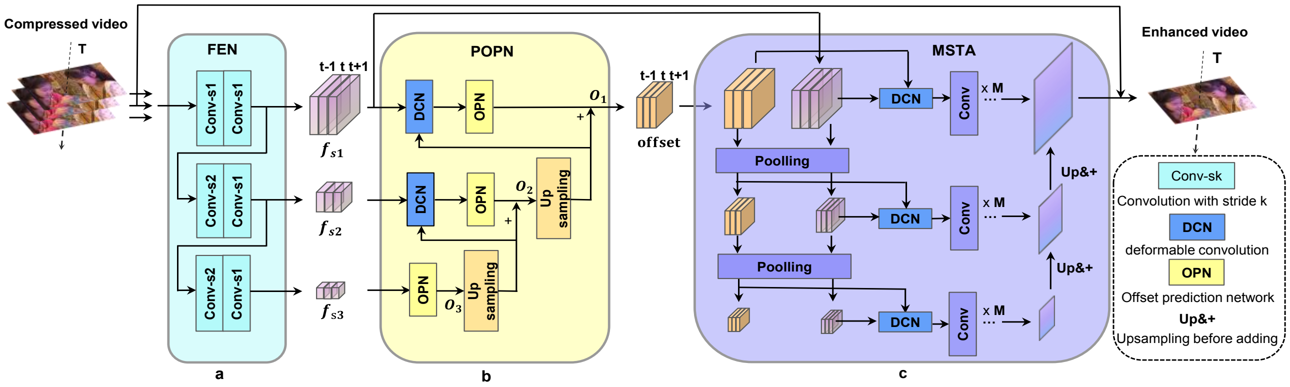

Fig. 1 Framework of the proposed MGO-VEN. (a) FEN: feature extraction network; (b) POPN: progressive offset prediction network; (c) MSTA: multi-scale temporal alignment network.

advantage lies in the diversity of offset. Accordingly, an offset-fidelity loss has been proposed to train deformable convolution alignment networks. Inspired by this, a flow-guided deformable alignment module was proposed to overcome the training instability in [24]. It shows that the features of optical flow before alignment can be used to reduce the difficulty in predicting the offset of deformable convolution. As a result, the training is more stable and the predicted offset locates in a reasonable range. In order to stabilize the training of deformable convolution networks, these methods seek solutions from the correlation between the offset and optical flow, but not from the mechanism of deformable convolution itself.

In order to solve the above problem, we propose a Multiscale Gradient-backpropagation Optimization framework for the deformable convolution based compressed Video Enhancement Network (MGO-VEN). The contributions can be summarized as: (1) A progressive offset prediction sub-network (POPN) is proposed to predict the offset of deformable convolution, which decouples the offset prediction from the feature up-sampling. (2) By analyzing the gradient backpropagation mechanism of deformable convolution, a multi-scale deformable convolution temporal alignment sub-network (MSTA) is proposed, which stabilizes the offset prediction through gradient propagation optimization. The multi-scale features are then used to progressively reconstruct the target frame from coarse to fine scales. (3) Experimental results, with ablation study on each module, have validated the effectiveness of the proposed method.

2. PROPOSED METHOD

2.1. Overview

The overall framework of the proposed method is shown in Fig. 1. It is composed of three sub networks: feature extraction network (FEN), progressive offset prediction network (POPN), and multi-scale temporal alignment network (MSTA).

For each frame I_t at time t from compressed video I to be enhanced, its temporal adjacent frames in the neighborhood of length r from both forward and backward directions are used. In other words, the continuous frame I_{t-r}-I_{t+r} are used as input for the network while frame I_t is to be enhanced. For simplicity, r is set to 1 in Fig. 1. First, FEN is used to extract the multi-scale features: f_{S1} , f_{S2} , f_{S3} of each frame, respectively, where f_{Sn} represents the feature with $1/2^{n-1}$ size of the original scale (both height and width). The specific process of FEN is shown in Fig. 1(a) which use a six-layer convolution network to extract features, and convolution with stride 2 is used to achieve down-sampling. After FEN processing, features with three scales are obtained. Deformable convolution is then used to align temporal features, where POPN and MSTA are proposed to predict offset and perform temporal alignment, respectively. These two subnetworks are developed to optimize the gradient backpropagation in the deformable convolution. MSTA will output the enhanced target frame. The two subnetworks are described in detail in the following subsections.

2.2. Progressive offset prediction network

The proposed POPN is developed based on the classic pyramid and cascading structure to assist the prediction of offsets under scenes with large motions. In POPN, instead of up-sampling the aligned features, the generated offset is up-sampled and progressively predicted. In the existing pyramid structure [22], the aligned features using the small-scale features and offsets are up-sampled for the final prediction. However, such procedure introduces noise and information loss in the up-sampled features considering the error in the small-scale offset and the distortion in the feature up-sampling process. To address this problem, we propose to up-sample the small-scale offset, and using the original large-scale feature together with the up-sampled offset for the following prediction as shown in Fig. 1 (b). Offset is in the form of positions or geometric distances, and up-sampling offset introduces much less noise than up-sampling the features. More importantly, the loss in the offset up-sampling

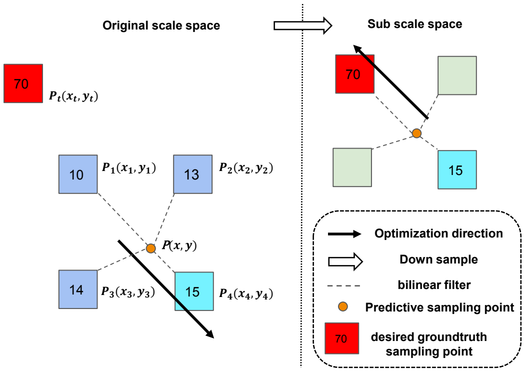

Fig. 2 Gradient-backpropagation optimization for deformable convolution.

process do not affect the overall process, since the large-scale offset is the value to be predicted and the loss can be compensated in the prediction. Meanwhile, the input large-scale features are obtained from the original input without any up-sampling noise and information loss. Additionally, only up-sampling the offset (with less channels) costs much less computation than up-sampling the whole features.

There are three stages in the proposed POPN to predict the offsets from the small scales to the original scale. The input of each stage comes from the feature map of the corresponding scale generated by FEN. The deformable convolution alignment is performed on the input using the up-sampled offset generated in the previous stage, and then the offset value of the next scale is predicted by OPN, together with the initial up-sampled offset. OPN uses a three-layer convolution to predict the offset. The specific offset prediction process of each stage is formulated as follows:

$$O_k = U(O_{k+1}) + OPN(DCN(f_{Sk}, U(O_{k+1}))) \qquad (1)$$

where O_k represents the offset predicted in stage k, and U represent the up-sampling process. f_{Sk} represents the input feature map at the scale of the to-be-predicted offset. The prediction process of the smallest scale can be simply written as $O_k = OPN(f_{Sk})$.

When predicting the offset in each stage, the input error only comes from the offset predicted in the previous stage, and the offset value can be progressively refined in the form of learning residuals in multiple stages. The offset of the original scale is generated at the final stage and used for the following MSTA module.

2.3. Multi-scale temporal alignment network

Deformable convolution adds a learnable offset to the standard convolution kernel, which makes the block to-be-convolved deformed. In deformable convolution, offset and features are of different natures, position indicated by distance, versus feature values coming from the color value. The offset relates to the feature values via the bilinear interpolation process, thus enabling the gradient backpropagation from color values to geometric distance. As shown in Fig. 2, the value of the current sampling point (P) is obtained by bilinear interpolation from the values of the surrounding four integer pixels (P_1, P_2, P_3, P_4), which is formulated as follows:

$$x(P) = \delta \cdot \sum_{k=1}^{4} \frac{x(P_k)}{d(P, P_k)} \qquad (2)$$

where P represents the value of the sampling point, and $P_k \in \{1,2,3,4\}$ represents the value of the corresponding integer pixel point. $d(P, P_k)$ represents the distance between P and P_k. δ represents the normalization coefficient. After all the sampling point values are obtained, the convolution operation can be carried out. Deformable convolution can be formulated as follows:

$$y(p) = \sum_{i=1}^{n^2} w(p_i) \cdot x(p + p_i + \Delta p_i) \qquad (3)$$

$$P_i = p + p_i + \Delta p_i, i \in 1 \sim n^2 \qquad (4)$$

where y represents the output of the convolution, and x represents the input feature value. p represents the current convolution position, and $p_i \in \{(-1,-1), (0,-1), \cdots, (1,1)\}$ is the offset of conventional convolution. $\Delta p_i$is the predicted offset in the deformable convolution. P_i is the sampling point with offset, and its value can be obtained via Eq. (2).

In the process of training the network, the current sampling point indicated by the current offset is moved according to the gradient of the offset. The gradient of the offset is backpropagated through the features, then via the distance-value relationship in the bilinear interpolation, finally to the offsets. That is to say, for the offset that interacts with the feature values via the bilinear interpolation, its update is only determined by the values of the surrounding four pixels. Therefore, in relatively smooth regions, the offset may be stuck without updating to the desired location. Specifically, as shown in Fig. 2, assume the desired groundtruth sampling point is P_t . When the network parameters are updated, the current sampling point is updated along the gradient direction, that is, the direction where the value close to the value of the groundtruth sampling point. The updated sampling point will be shifted to the direction of P_4, which results to a larger error of the predicted offset.

To alleviate this problem and optimize the gradient backpropagation for the offset, a multi-scale alignment module is proposed in this paper. The features to be aligned and the offsets are synchronously down-sampled using average pooling to one-half and one-quarter of the original size. Then the features at each scale are processed to produce the enhanced image at different scales. As shown in Fig.2, at a smaller scale, the groundtruth information can be largely incorporated in the interpolation and thus the gradient can be greatly optimized.

Since the offsets of the three scales are obtained from the same source, the gradients can be backpropagated to larger scales. While the multiscale output images can be supervised individually, in this paper, they are progressively combined in a similar way as in POPN for simplicity. A shortcut connection is added from the input to the output as in Fig. 1

Table 1. Ablation study under three frames

Groups	PSNR (dB)
STDF [13]	0.65
MGO-VEN w/o POPN	0.75
MGO-VEN w/o MSTA	0.80
MGO-VEN	**0.84**

and thus the MSTA only needs to produce a residual image to the original one.

The specific implementation is shown in the subgraph MSTA of Fig. 1(c). The offset to be optimized is O, and the output residual image is R. The output can be obtained as follows:

$$R = R_1 + U(R_2 + U(R_3)) \tag{5}$$

$$R_k = C_M(C_{M-1}(\cdots C_1(DCN(O_k, f_k) \cdots) \tag{6}$$

$$\{O_k, f_k\} = Pool(\{O_{k-1}, f_{k-1}\}), k \in \{2,3\} \tag{7}$$

where C_m represents the m-th convolution layer, and $Pool$ represents pooling. O_1 is O , and f_1 is f_{S1} .The overall network can be trained end-to-end using MSE as the objective function.

3. EXPERIMENTS

Extensive experiments are conducted to demonstrate the effectiveness of the proposed MGO-VEN method.

Experiment setup. The dataset used in MFQE [12], containing 108 videos, is used for training and evaluation. Among them, 100 videos are used for training and the other 8 videos are used for evaluation. The reconstructed video is obtained using HM16.9 with LDP configuration under QP ∈{22,27,32,37}. The common test conditions suggested by JVT is used for test [23]. Blocks of size 128x128 are used as the input, that is, for each group of inputs of the training set, the image block of a 128 * 128 area in the same position is randomly drawn from multiple consecutive frames for training. Adam is used as the optimizer with an initial learning rate of 10^{-4}. The learning rate is reduced according to the result of the validation set. M in Fig. 1, the number of convolutional layers in MSTA, is set to 8.

Ablation study. Firstly, in order to verify the effectiveness of the proposed POPN and MSTA modules, a group of ablation studies were conducted, as shown in Table 1. The first group replaced POPN with the conventional OPN network. The second group removed the multi-scale branches of MSTA and retained only the original scale branch. The third group retains all proposed modules. The number of frames is set to 3 for simplicity and the experiments are conducted under QP=37. As can be seen from the results in Table 1, the proposed POPN and MSTA modules are both effective and the proposed method with both of them performs the best.

Comparison with the state-of-the-art methods. Experiments on the compressed videos under common test conditions are also conducted. In this experiment, the number of reference frames is set to 6, that is, r = 3, to explore more temporal reference information, which also applied to STDF. Table 2 shows the results comparison between the proposed method and other state-of-the-art methods. It can be seen that the proposed method achieves the best performance on all test sequences, and provides 25.6% BD-rate saving on average.

Table 2. BD-BR reduction (%) of the proposed method compared with the state-of-the-art methods

	Sequence	ARCNN [3]	MFQE 2.0[12]	STDF [13]	MGO-VEN
A	Traffic	7.4	17.0	20.9	**24.9**
	PeopleOnStreet	7.0	15.1	18.0	**23.2**
B	Kimono	6.1	13.3	18.6	**26.7**
	ParkScene	4.5	13.7	20.2	**24.7**
	Cactus	6.2	14.8	23.0	**28.6**
	BasketballDrive	5.8	11.9	15.6	**23.7**
	BQTerrace	6.9	14.7	26.2	**30.5**
C	BasketballDrill	4.7	12.6	15.0	**19.5**
	BQMall	5.6	13.5	21.1	**25.2**
	PartyScene	1.9	11.3	20.5	**25.0**
	RaceHorses	5.2	9.6	11.1	**16.8**
D	BasketballPass	5.1	13.4	20.0	**24.6**
	BQSquare	0.7	11.0	31.1	**36.9**
	BlowingBubbles	3.2	15.2	19.5	**24.9**
	RaceHorses	5.6	11.6	14.2	**21.3**
E	FourPeople	8.4	17.5	21.2	**24.8**
	Johnny	7.7	18.6	25.0	**30.8**
	KristenAndSara	8.9	18.3	23.3	**27.3**
	Average	5.6	14.1	20.4	**25.6**

4. CONCLUSION

In this paper, we have presented a multiscale gradient-backpropagation optimization framework for the deformable convolution based compressed video quality enhancement network. First, a progressive offset prediction network is developed to predict the offset more accurately without introducing the distortion in up-sampling the features with low-scale offsets. Then, by analyzing the gradient backpropagation mechanism of deformable convolution, a multi-scale deformable convolution alignment network is proposed, where the gradients of the offsets are optimized. Experiments, with ablation study on each module, have demonstrated the effectiveness of the proposed method, and the proposed method achieves the state-of-the-art performance with a BD-rate reduction of 25.6% on average compared with HM.

5. ACKNOWLEDGMENT

This work was partly supported by the National Key R&D Program of China (2018YFE0203900), the National Natural Science Foundation of China (No. 61901083, 62001092) and SDU QILU Young Scholars program.

6. REFERENCES

[1] G.J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

[2] B. Bross, J. Chen and S. Liu, "Versatile Video Coding (Draft 3)," Document JVET-L1001, Oct. 2018.

[3] C. Dong, Y. Deng, C.C. Loy and X. Tang. "Compression Artifacts Reduction by a Deep Convolutional Network," *IEEE International Conference on Computer Vision*, 2015, pp. 576-584.

[4] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5457-5466.

[5] C. Li, L. Song, R. Xie and W. Zhang. "CNN based post-processing to improve HEVC," *IEEE International Conference on Image Processing*, Sep. 2017, pp. 4577-4580.

[6] Y. Zhang, T. Shen, X. Ji, R. Xiong and Q. Dai, "Residual Highway Convolutional Neural Networks for in-loop Filtering in HEVC," *IEEE Transactions on Image Processing*, 2018, pp. *3*827-3841.

[7] M. Jia, Y. Gao, S. Li, J. Yue, and M. Ye, "An explicit self-attention-based multimodality CNN in-loop filter for versatile video coding," *Multimedia Tools and Applications*, 2021.

[8] C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu and S. Ma, "Content-Aware Convolutional Neural Network for In-Loop Filtering in High Efficiency Video Coding," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3343-3356, July 2019.

[9] J. Yue, Y. Gao, S. Li, and M. Jia, "A Mixed Appearance-based and Coding Distortion-based CNN Fusion Approach for In-loop Filtering in Video Coding," *IEEE International Conference on Visual Communications and Image Processing*, 2020, pp. 487-490.

[10] D. Ding, L. Kong, G. Chen, Z. Liu and Y. Fang, "A Switchable Deep Learning Approach for In-Loop Filtering in Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1871-1887, July 2020.

[11] S. Zhang, Z. Fan, N. Ling and M. Jiang, "Recursive Residual Convolutional Neural Network- Based In-Loop Filtering for Intra Frames," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1888-1900, July 2020.

[12] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu and Z. Wang, "MFQE 2.0: A New Approach for Multi-Frame Quality Enhancement on Compressed Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, pp. 949-963.

[13] J. Deng, L. Wang, S. Pu and C. Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," *AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10696-10703, April. 2020.

[14] Z. Pan, W. Yu, J. Lei, N. Ling and S. Kwong, "TSAN: Synthesized View Quality Enhancement via Two-Stream Attention Network for 3D-HEVC," *IEEE Transactions on Circuits and Systems for Video Technology*, doi: 10.1109/TCSVT.2021.3057518.

[15] Z. Pan, F. Yuan, J. Lei, W. Li, N. Ling and S. Kwong, "MIEGAN: Mobile Image Enhancement via A Multi-Module Cascade Neural Network," *IEEE Transactions on Multimedia*, doi: 10.1109/TMM.2021.3054509.

[16] A. Ranjan and M.J. Black, "Optical flow estimation using a spatial pyramid network," *IEEE conference on computer vision and pattern recognition,* 2017, pp. 4161-4170.

[17] T. Xue, B. Chen, J. Wu, D. Wei and W.T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol.*127, no.* 8, pp. 1106-1125, 2019.

[18] D. Sun, X. Yang, M.Y. Liu and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," *IEEE conference on computer vision and pattern recognition,* 2018, pp. 8934-8943.

[19] B. Mildenhall, J.T. Barron, J. Chen, D. Sharlet, R. Ng, and R.E. Carroll, "Burst Denoising with Kernel Prediction Networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2502-2510.

[20] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J.S. Ren, "Spatio-Temporal Filter Adaptive Network for Video Deblurring," *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2482-2491.

[21] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu and Y. Wei, "Deformable convolutional networks," *IEEE international conference on computer vision*, 2017, pp. 764-773.

[22] X. Wang, K.C. Chan, K. Yu, C. Dong and C.C. Loy, "EDVR: Video Restoration With Enhanced Deformable Convolutional Networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp.1954-1963.

[23] K.C. Chan, X. Wang, K. Yu, C. Dong and C.C. Loy, "Understanding deformable alignment in video super-resolution," *arXiv preprint arXiv:2009.07265*, *4*, 3, 2020.

[24] K.C. Chan, S. Zhou, X. Xu and C.C. Loy, "BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment," *arXiv preprint arXiv:2104.13371, 2021.*

[25] F. Bossen, "Common test conditions and software reference configurations," Document *JCTVC-L1100, Jan. 2013.*