Spatio-Temporal Deformable Convolution for Compressed Video Quality Enhancement

Jianing Deng,[1] Li Wang,[2] Shiliang Pu,[2] Cheng Zhuo[1]*
[1]College of Information Science and Electronic Engineering, Zhejiang University
[2]Hikvision Research Institute
{dengjn, czhuo}@zju.edu.cn, {wangli7, pushiliang}@hikvision.com

Abstract

Recent years have witnessed remarkable success of deep learning methods in quality enhancement for compressed video. To better explore temporal information, existing methods usually estimate optical flow for temporal motion compensation. However, since compressed video could be seriously distorted by various compression artifacts, the estimated optical flow tends to be inaccurate and unreliable, thereby resulting in ineffective quality enhancement. In addition, optical flow estimation for consecutive frames is generally conducted in a pairwise manner, which is computational expensive and inefficient. In this paper, we propose a fast yet effective method for compressed video quality enhancement by incorporating a novel Spatio-Temporal Deformable Fusion (STDF) scheme to aggregate temporal information. Specifically, the proposed STDF takes a target frame along with its neighboring reference frames as input to jointly predict an offset field to deform the spatio-temporal sampling positions of convolution. As a result, complementary information from both target and reference frames can be fused within a single Spatio-Temporal Deformable Convolution (STDC) operation. Extensive experiments show that our method achieves the state-of-the-art performance of compressed video quality enhancement in terms of both accuracy and efficiency.

1 Introduction

Nowadays, video content has become a major fraction of digital network traffic and is still growing (Wien 2015). To transmit video under limited bandwidth, video compression is indispensable to significantly reduce the bit-rate. However, compression algorithms, such as H.264/AVC (Wiegand et al. 2003) and H.265/HEVC (Sullivan, Ohm, and Wiegand 2013), often introduce various artifacts in the compressed video, especially at low bit-rate. As shown in Figure 1, such artifacts may considerably diminish video quality, resulting in degradation of Quality of Experience (QoE). The distorted contents in low-quality compressed video may also reduce performance of subsequent vision tasks (e.g., recognition, detection, tracking) in low-bandwidth applications (Galteri et al. 2017; Lu et al. 2019). Thus, it's crucial to study on compressed video quality enhancement (VQE).

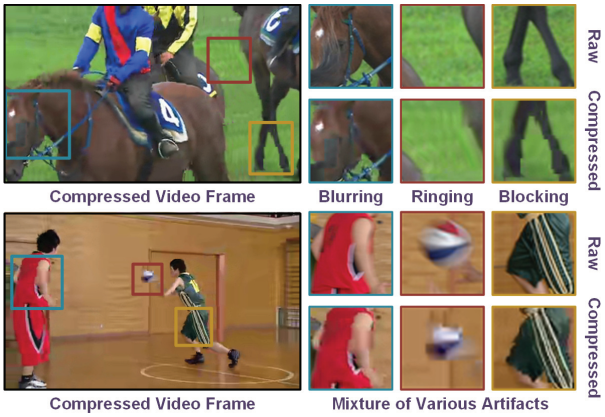


Figure 1: Illustration of compression artifacts. Videos are compressed by the latest H.265/HEVC coding algorithm.

During the past decades, extensive works have been conducted on artifacts removal or quality enhancement for single compressed image. Traditional methods (Foi, Katkovnik, and Egiazarian 2007; Zhang et al. 2013) reduced artifacts by optimizing the transform coefficients for specific compression standard, thus they are hard to extend to other compression schemes. With the recent advances in Convolutional Neural Networks (CNNs), CNN-based methods (Dong et al. 2015; Tai et al. 2017; Zhang et al. 2017; 2019) have also emerged for image quality enhancement. They usually learn a non-linear mapping to directly regress the artifact-free image from a large amount of training data, leading to impressive results with high efficiency. However, these methods cannot be directly extended to compressed video since they treat frames independently and thus fail to exploit temporal information.

On the other hand, there is only limited study on quality enhancement for compressed video. Yang et al. first proposed Multi-Frame Quality Enhancement (MFQE 1.0) approach to leverage temporal information for VQE (Yang et al. 2018). Specifically, high quality frames in compressed video are utilized as *reference* frame to help enhancing quality of neighboring low quality *target* frame via a novel Multi-Frame CNN (MF-CNN). Recently, an upgraded version MFQE 2.0 (Guan et al. 2019) was introduced to further

*Corresponding author.

improve the efficiency of MF-CNN, and achieved state-of-the-art performance. In order to aggregate information from target frame and reference frames, both MFQE methods adopt a widely used temporal fusion scheme that incorporates dense optical flow for motion compensation (Kappeler et al. 2016; Caballero et al. 2017; Xue et al. 2017). However, this temporal fusion scheme may be suboptimal in the context of VQE task. Since compression artifacts could seriously distort video contents and break pixelwise correspondances between frames, the estimated optical flow tends to be inaccurate and unreliable, thereby resulting in ineffective quality enhancement. In addition, optical flow estimation needs to be repeatedly performed for different reference-target frame pairs in a pairwise manner, which involves substantially increased computational cost to explore more reference frames.

To address the aforementioned issues, we introduce a Spatio-Temporal Deformable Fusion (STDF) scheme for VQE task. Specifically, we propose to learn a novel Spatio-Temporal Deformable Convolution (STDC) to aggregate temporal information while avoiding explicit optical flow estimation. The main idea of STDC is to adaptively deform the spatio-temporal sampling positions of convolution so as to capture the most relevant context and exclude the noisy content for quality enhancement of the target frame. To this end, we adopt a CNN-based predictor to jointly model the correspondance across target and reference frames, and accordingly regress those sampling positions within a single inference pass. The main contributions of this paper are summarized as follows:

- We propose an end-to-end CNN-based method for VQE task, which incorporates a novel STDF scheme to aggregate temporal information.
- We analytically and experimentally compare the proposed STDF to prior fusion schemes, and demonstrate its higher flexibility and robustness.
- We quantitatively and qualitatively evaluate the proposed method on VQE benchmark dataset and show that it achieves state-of-the-art performance in terms of accuracy and efficiency.

2 Related Work

Image and Video Quality Enhancement. Over the past decade, an increasing number of works have focused on quality enhancement for compressed image (Foi, Katkovnik, and Egiazarian 2007; Jancsary, Nowozin, and Rother 2012; Chang, Ng, and Zeng 2013; Zhang et al. 2013; Dong et al. 2015; Guo and Chao 2016; Zhang et al. 2017; 2019). Among them, CNN-based end-to-end methods achieve the recent state-of-the-art performance. Specifically, Dong et al. first introduced a 4-layer AR-CNN (Dong et al. 2015) to remove various JPEG compression artifacts. Later, Zhang et al. managed to learn a very deep DnCNN (Zhang et al. 2017) with residual learning scheme for several image restoration tasks. Most recently, Zhang et al. proposed an even deeper network RNAN (Zhang et al. 2019) with residual non-local attention mechanism to capture long-range dependencies between pixels and set up a new state-of-the-art of image quality enhancement. These methods tend to apply large CNNs to capture discriminative features within an image, resulting in a large amount of computations and parameters. On the other hand, MFQE 1.0 (Yang et al. 2018) pioneered on applying multi-frame CNN to take advantage of temporal information for compressed video quality enhancement, where high quality frames are utilized to help enhancing quality of the adjacent low quality frames. To exploit long range temporal information, Yang et al. later introduced a modified convolutional long short-term memory network (Yang et al. 2019) for video quality enhancement. Most recently, Guan et al. proposed MFQE 2.0 (Guan et al. 2019) to upgrade several key components of MFQE 1.0 and achieved state-of-the-art performance in terms of accuracy and speed.

Leveraging Temporal Information. It is crucial to leverage complementary information across multiple frames for video related tasks. Karpathy et al. first introduced several convolution based fusion schemes to combine spatio-temporal information for video classification (Karpathy et al. 2014). Kappeler et al. later investigated those fusion schemes for low-level vision tasks (Kappeler et al. 2016), and managed to improve accuracy by compensating motion across consecutive frames with a Total Variation (TV) based optical flow estimation algorithm. Caballero et al. further replaced the TV based flow estimator with CNN to enable end-to-end training (Caballero et al. 2017). Since then, temporal fusion with motion compensation has been widely adopted for various vision tasks (Xue et al. 2017; Yang et al. 2018; Kim et al. 2018; Guan et al. 2019). However, these methods heavily rely on accurate optical flow which is hard to obtain due to general problems (e.g., occlusion, large motion) or task-specific problems (e.g., compression artifacts). To cope with this, works have been conducted to bypass explicit optical flow estimation. Niklaus, Mai, and Liu proposed AdaConv (Niklaus, Mai, and Liu 2017) to adaptively generate convolution kernels by implicitly utilizing motion cues for video frame interpolation. Shi et al. introduced ConvLSTM network to exploit contextual information from a long range of adjacent frames (Shi et al. 2015). In this work, we propose to combine motion cues with convolution to efficiently aggregate spatio-temporal information, which also omits the explicit estimation of optical flow.

Deformable Convolution. Dai et al. first proposed to augment regular convolution with learnable sampling offsets to model complex geometric transformations for object detection (Dai et al. 2017; Zhu et al. 2019). Later, several works (Bertasius, Torresani, and Shi 2018; Tian et al. 2018; Wang et al. 2019) extended it along temporal extent to implicitly capture motion cues for video-related applications, and achieved better performance than traditional methods. However, these methods perform deformable convolution in a pairwise manner, thus fail to fully explore temporal correspondances across multiple frames. In this work, we propose STDC to jointly consider a video clip rather than splitting it into several reference-target frame pairs, leading to more effective use of contextual information.

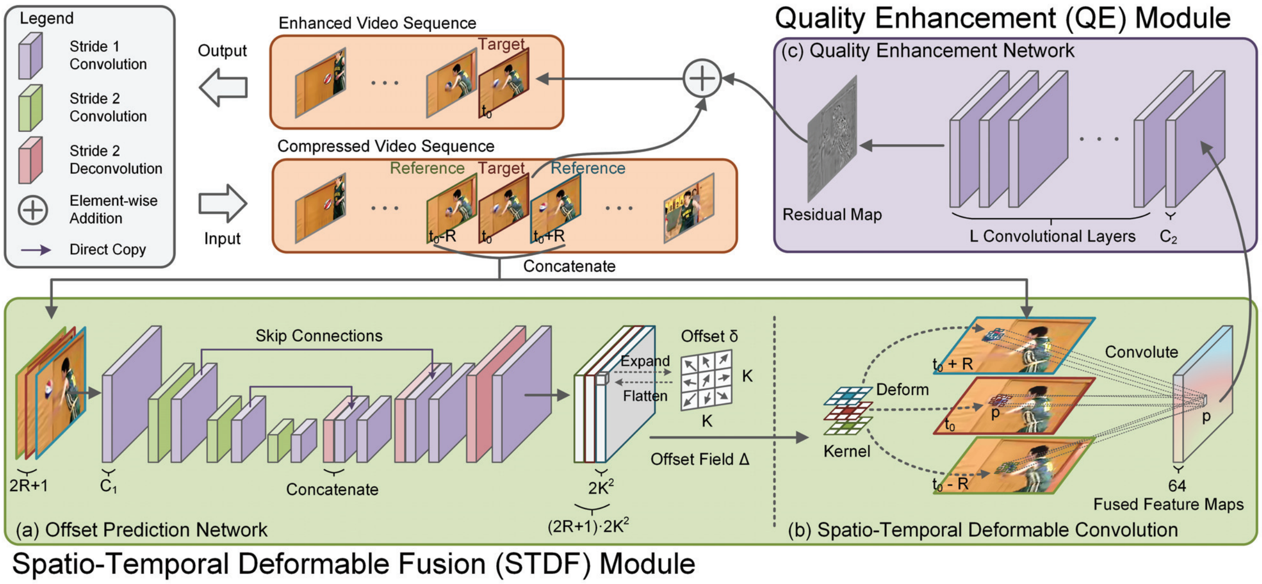


Figure 2: Overview of the proposed framework for compressed video quality enhancement. Given a compressed video clip with $2R+1$ concatenated frames, an offset prediction network is first adopted to generate deformable offset field. With this offset field, spatio-temporal deformable convolution is then performed to fuse temporal information and produce fused feature maps. At last, QE network is used to compute the enhancement residual map, and the final enhanced result can be obtained by adding the residual map back to the compressed target frame. Herein, temporal radius R=1, and deformable kernel size K=3.

3 Proposed Method

3.1 Overview

Given a compressed video which is distorted by compression artifacts, the goal of our method is to remove those artifacts and accordingly enhance the video quality. To be specific, we conduct the enhancement separately for each compressed frame $I_{t_0}^{LQ} \in \mathbb{R}^{H\times W}$ at time t_0 [1]. In order to leverage temporal information, we take the preceding and succeeding R frames as reference to help enhancing quality of each target $I_{t_0}^{LQ}$. The enhanced solution $\hat{I}_{t_0}^{HQ} \in \mathbb{R}^{H\times W}$ can then be expressed as

$$\hat{I}_{t_0}^{HQ} = \mathcal{F}_\theta(\{I_{t_0-R}^{LQ}, \cdots, I_{t_0}^{LQ}, \cdots, I_{t_0+R}^{LQ}\}) \tag{1}$$

where $\mathcal{F}_\theta(\cdot)$ represents the proposed quality enhancement model and θ are the learnable parameters.

Figure 2 demonstrates the framework of our method, which is composed of a **Spatio-Temporal Deformable Fusion (STDF)** module and a **Quality Enhancement (QE)** module. The STDF module takes both target frame and reference frames as input, and fuse contextual information via a spatio-temporal deformable convolution, where the deformable offsets are adaptively generated by an offset prediction network. Then, with the fused feature maps, the QE module incorporates a fully convolutional enhancement network to compute the enhanced result. Since both STDF module and QE module are convolutional, our unified framework can be trained in an end-to-end manner.

3.2 STDF Module

Spatio-Temporal Deformable Convolution. For a compressed video clip $\{I_{t_0-R}^{LQ}, \cdots, I_{t_0}^{LQ}, \cdots, I_{t_0+R}^{LQ}\}$, the most straightforward temporal fusion scheme, i.e., Early Fusion (EF) (Karpathy et al. 2014), can be formulated as multi-channel convolution applied directly on the compressed frames as

$$F(\mathbf{p}) = \sum_{t=t_0-R}^{t_0+R} \sum_{k=1}^{K^2} W_{t,k} \cdot I_t^{LQ}(\mathbf{p}+\mathbf{p}_k) \tag{2}$$

where F is the resulting feature map, K represents the size of convolution kernel, $W_t \in \mathbb{R}^{K^2}$ is the kernel for t-th channel, $\mathbf{p}$ indicates arbitrary spatial position and $\mathbf{p}_k$ represents the regular sampling offsets. For example, $\mathbf{p}_k \in \{(-1,-1),(-1,0),\cdots,(1,1)\}$ for K=3. Despite the high efficiency, EF may easily introduce noisy content and reduce the performance of subsequent enhancement due to temporal motion, as shown in Figure 3. Inspired by Dai et al. (Dai et al. 2017), we address this issue by introducing a novel Spatio-Temporal Deformable Convolution (STDC) to augment the regular sampling offset with extra learnable offset $\boldsymbol{\delta}_{(t,\mathbf{p})} \in \mathbb{R}^{2K^2}$ as

$$\mathbf{p}_k \leftarrow \mathbf{p}_k + \boldsymbol{\delta}_{(t,\mathbf{p}),k} \tag{3}$$

It is worth noting that the deformable offset $\boldsymbol{\delta}_{(t,\mathbf{p})}$ are position-specific, i.e., individual $\boldsymbol{\delta}_{(t,\mathbf{p})}$ will be assigned for each convolution window centered at spatio-temporal position $(t,\mathbf{p})$. Thus, spatial deformations as well as temporal

[1] For simplicity, we assume enhancement is performed on luminance channel only. Thus we represent all frames as 2D matrices.

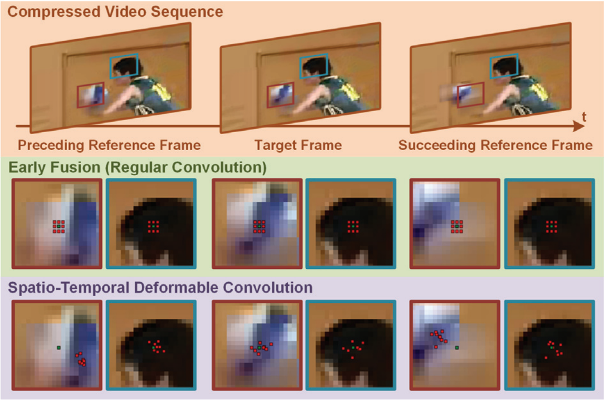

Compressed Video Sequence
t
Preceding Reference Frame
Target Frame
Succeeding Reference Frame
Early Fusion (Regular Convolution)
Spatio-Temporal Deformable Convolution

Figure 3: Visualization of EF and our STDC. Herein, red points represent the sampling positions of 3×3 convolution window centered at green points. The STDC can adapt to both large temporal motion (ball) and the small one (head), and accordingly captures the relevant context for quality enhancement.

dynamics within the video clip can be simultaneously modeled, as shown in Figure 3. Since the learnable offsets can be fractional, we follow Dai et al. (Dai et al. 2017) to apply the differentiable bilinear interpolation to sample sub-pixel $I_t^{LQ}(\mathbf{p} + \mathbf{p}_k)$.

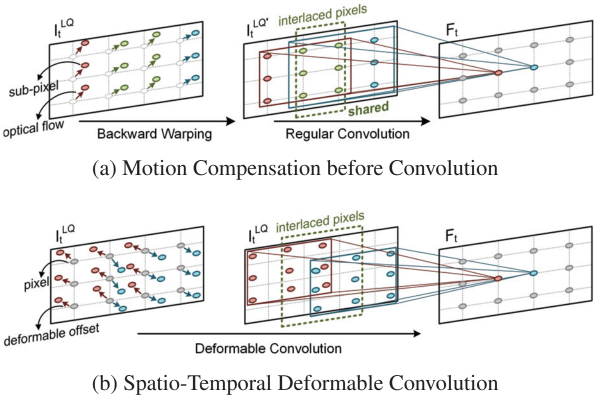
Unlike previous VQE methods (Yang et al. 2018; Guan et al. 2019) which perform explicit motion compensation before fusion to alleviate the effect of temporal motion, STDC implicitly combines motion cues with position-specific sampling while fusion. This leads to higher flexibility and robustness because adjacent convolution windows can sample contents independently, as shown in Figure 4.

Joint Deformable Offset Prediction. Different from optical flow estimation that solely handles one reference-target frame pair at a time, we propose to take the whole clip into consideration and jointly predict all deformable offsets at once. To this end, we apply an offset prediction network $\mathcal{F}_{\theta_{op}}(\cdot)$ to predict an offset field $\Delta \in \mathbb{R}^{(2R+1)\times 2K^2 \times H \times W}$ for all spatio-temporal positions in the video clip as

$$\Delta = \mathcal{F}_{\theta_{op}}([I_{t_0-R}^{LQ}, \cdots, I_{t_0}^{LQ}, \cdots, I_{t_0+R}^{LQ}]) \quad (4)$$

where frames are concatenated together as input. Since consecutive frames are highly correlated, offset prediction for one frame can benefit from the other frames, leading to more effective use of temporal information than pairwise scheme. In addition, joint prediction is more computational efficient because all deformable offsets can be obtained in a single inference pass.

As shown in Figure 2-(a), we adopt a U-Net based network (Ronneberger, Fischer, and Brox 2015) for offset prediction to enlarge receptive field so as to capture large temporal dynamics. Convolutional and deconvolutional layers (Zeiler and Fergus 2014) with stride of 2 are used for downsampling and upsampling respectively. For convolutional layer with stride of 1, zero padding is used to retain feature size. For simplicity, we set the filter number of all (de)convolutional layers to C_1. Rectified Linear Unit (ReLU) is adopted as activation function for all layers except the last one which is followed by linear activation to regress the offset field Δ. We do not use any normalization layer in the network.



(a) Motion Compensation before Convolution

(b) Spatio-Temporal Deformable Convolution

Figure 4: Comparison between motion compensation based convolution and spatio-temporal deformable convolution. Herein, 3×3 convolution is used for demonstration.

3.3 QE Module

The main idea of QE module is to fully explore complementary information from fused feature maps F and accordingly generate the enhanced target frame $\hat{I}_{t_0}^{HQ}$. In order to take advantage of residual learning (Kim, Lee, and Lee 2016), we first learn a non-linear mapping $\mathcal{F}_{\theta_{qe}}(\cdot)$ to predict the enhancement residual as

$$\hat{\mathcal{R}}_{t_0}^{HQ} = \mathcal{F}_{\theta_{qe}}(F) \quad (5)$$

The enhanced target frame can then be generated as

$$\hat{I}_{t_0}^{HQ} = \hat{\mathcal{R}}_{t_0}^{HQ} + I_{t_0}^{LQ} \quad (6)$$

As illustrated in Figure 2-(c), we implement $\mathcal{F}_{\theta_{qe}}(\cdot)$ through another CNN which consists of L convolutional layers of stride 1. All layers except the last one have C_2 convolutional filters followed by ReLU activation. The last convolutional layer outputs the enhancement residual. Without bells and whistles, such plain QE network is able to achieve satisfactory enhancement results.

3.4 Training Scheme

Since STDF module and QE module are fully-convolutional and thus differentiable, we jointly optimize θ_{op} and θ_{qe} in an end-to-end fashion. The overall loss function $\mathcal{L}$ is set to the Sum of Squared Error (SSE) between the enhanced target frame $\hat{I}_{t_0}^{HQ}$ and the raw one $I_{t_0}^{HQ}$ as

$$\mathcal{L} = \|\hat{I}_{t_0}^{HQ} - I_{t_0}^{HQ}\|_2^2 \quad (7)$$

Note that, as there is no ground-truth for deformable offsets, learning for offset prediction network $\mathcal{F}_{\theta_{op}}(\cdot)$ is totally unsupervised and fully driven by the final loss $\mathcal{L}$, which is different from previous works (Yang et al. 2018; Guan et al. 2019) that incorporate auxiliary losses to constrain optical flow estimation.

QP	Test Videos[†]		Image QE Methods			Video QE Methods				
			Dong et al.	Zhang et al.	Zhang et al.	Yang et al.	Guan et al.	Ours	Ours	Ours
			AR-CNN	DnCNN	RNAN[‡]	MFQE 1.0	MFQE 2.0	STDF-R1	STDF-R3	STDF-R3L
37	Class A	*Traffic*	0.27 / 0.50	0.35 / 0.64	0.40 / 0.86	0.50 / 0.90	0.59 / 1.02	0.56 / 0.92	0.65 / 1.04	0.73 / 1.15
		PeopleOnStreet	0.37 / 0.76	0.54 / 0.94	0.74 / 1.30	0.80 / 1.37	0.92 / 1.57	1.05 / 1.66	1.18 / 1.82	1.25 / 1.96
	Class B	*Kimono*	0.20 / 0.59	0.27 / 0.73	0.33 / 0.98	0.50 / 1.13	0.55 / 1.18	0.66 / 1.32	0.77 / 1.47	0.85 / 1.61
		ParkScene	0.14 / 0.44	0.17 / 0.52	0.20 / 0.77	0.39 / 1.03	0.46 / 1.23	0.41 / 1.05	0.54 / 1.32	0.59 / 1.47
		Cactus	0.20 / 0.41	0.28 / 0.53	0.35 / 0.76	0.44 / 0.88	0.50 / 1.00	0.59 / 1.06	0.70 / 1.23	0.77 / 1.38
		BQTerrace	0.23 / 0.43	0.33 / 0.53	0.42 / 0.84	0.27 / 0.48	0.40 / 0.67	0.55 / 0.89	0.58 / 0.93	0.63 / 1.06
		BasketballDrive	0.23 / 0.51	0.33 / 0.63	0.43 / 0.92	0.41 / 0.80	0.47 / 0.83	0.60 / 0.99	0.66 / 1.07	0.75 / 1.23
	Class C	*RaceHorses*	0.23 / 0.49	0.31 / 0.70	0.39 / 0.99	0.34 / 0.55	0.39 / 0.80	0.41 / 0.98	0.48 / 1.09	0.55 / 1.35
		BQMall	0.28 / 0.69	0.38 / 0.87	0.45 / 1.15	0.51 / 1.03	0.62 / 1.20	0.75 / 1.44	0.90 / 1.61	0.99 / 1.80
		PartyScene	0.14 / 0.52	0.22 / 0.69	0.30 / 0.98	0.22 / 0.73	0.36 / 1.18	0.52 / 1.49	0.60 / 1.60	0.68 / 1.94
		BasketballDrill	0.23 / 0.48	0.42 / 0.89	0.50 / 1.07	0.48 / 0.90	0.58 / 1.20	0.64 / 1.19	0.70 / 1.26	0.79 / 1.49
	Class D	*RaceHorses*	0.26 / 0.59	0.34 / 0.80	0.42 / 1.02	0.51 / 1.13	0.59 / 1.43	0.63 / 1.51	0.73 / 1.75	0.83 / 2.08
		BQSquare	0.21 / 0.30	0.30 / 0.46	0.32 / 0.63	-0.01 / 0.15	0.34 / 0.65	0.75 / 1.03	0.91 / 1.13	0.94 / 1.25
		BlowingBubbles	0.16 / 0.46	0.25 / 0.76	0.31 / 1.08	0.39 / 1.20	0.53 / 1.70	0.53 / 1.69	0.68 / 1.96	0.74 / 2.26
		BasketballPass	0.26 / 0.63	0.38 / 0.83	0.46 / 1.08	0.63 / 1.38	0.73 / 1.55	0.80 / 1.54	0.95 / 1.82	1.08 / 2.12
	Class E	*FourPeople*	0.40 / 0.56	0.54 / 0.73	0.70 / 0.97	0.66 / 0.85	0.73 / 0.95	0.83 / 1.01	0.92 / 1.07	0.94 / 1.17
		Johnny	0.24 / 0.21	0.47 / 0.54	0.56 / 0.88	0.55 / 0.55	0.60 / 0.68	0.65 / 0.71	0.69 / 0.73	0.81 / 0.88
		KristenAndSara	0.41 / 0.47	0.59 / 0.62	0.63 / 0.80	0.66 / 0.75	0.75 / 0.85	0.84 / 0.83	0.94 / 0.89	0.97 / 0.96
	Average		0.25 / 0.50	0.36 / 0.69	0.44 / 0.95	0.46 / 0.88	0.56 / 1.09	0.65 / 1.18	0.75 / 1.32	0.83 / 1.51
32	Average		0.19 / 0.17	0.33 / 0.41	0.41 / 0.62	0.43 / 0.58	0.52 / 0.68	0.64 / 0.77	0.73 / 0.87	0.86 / 1.04
27	Average		0.16 / 0.09	0.33 / 0.26	- / -	0.40 / 0.34	0.49 / 0.42	0.59 / 0.47	0.67 / 0.53	0.72 / 0.57
22	Average		0.13 / 0.04	0.27 / 0.14	- / -	0.31 / 0.19	0.46 / 0.27	0.51 / 0.27	0.57 / 0.30	0.63 / 0.34

[†] Video resolution: Class A (2560×1600), Class B (1920×1080), Class C (832×480), Class D (480×240), Class E (1280×720).

[‡] Patch-wise enhancement is performed for RNAN method due to memory restriction.

Table 1: Quantitative results of ΔPSNR (dB) / ΔSSIM ($\times 10^{-2}$) on test videos at 4 different QPs.

Method	Processing Speed @ Different Resolution					#Param(K)
	120p	240p	480p	720p	1080p	
DnCNN	191.8	54.7	14.1	6.1	2.6	556
RNAN	5.6	-	-	-	-	8957
MFQE 1.0	-	12.6	3.8	1.6	0.7	1788
MFQE 2.0	-	25.3	8.4	3.7	1.6	255
STDF-R1	141.9	38.9	9.9	4.2	1.8	330
STDF-R3	132.7	36.4	9.1	3.8	1.6	365
STDF-R3L	96.6	23.8	5.9	2.5	1.0	1275

Table 2: Quantitative results of speed (FPS) and amount of parameters. Results of speed are measured on Nvidia GeForce GTX 1080 Ti GPU.

4 Experiments

4.1 Datasets

Following MFQE 2.0 (Guan et al. 2019), we collect a total of 130 uncompressed videos with various resolutions and contents from two databases, i.e., Xiph (Xiph.org) and VQEG (VQEG), where 106 of them are selected for training and the rest are for validation. For testing, we adopt the dataset from Joint Collaborative Team on Video Coding (Ohm et al. 2012) with 18 uncompressed videos. These testing videos are widely used for video quality assessment with around 450 frames per video. We compress all the above videos by the latest H.265/HEVC reference software HM16.5 [2] under Low Delay P (LDP) configuration, as in previous work (Guan et al. 2019). The compression is conducted at 4 different Quantization Parameters (QPs), i.e., 22, 27, 32, 37, in order to evaluate performance under different compression levels.

4.2 Implementation Details

The proposed method is implemented based on PyTorch framework with reference to MMDetection toolbox (Chen et al. 2019) for deformable convolution. For training, we randomly crop 64×64 clips from raw and the corresponding compressed videos as training samples. Data augmentation (i.e., rotation or flip) is further used to better exploit those training samples. We train all models using Adam optimizer (Kingma and Ba 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. Learning rate is initially set to 10^{-4} and retained throughout training. We train 4 models from scratch for the 4 QPs respectively. For evaluation, as with previous works, we only apply quality enhancement on Y-channel (i.e., luminance component) in YUV/YCbCr space. We adopt incremental Peak Signal-to-Noise Ratio (ΔPSNR) and Structural Similarity (ΔSSIM) (Wang et al. 2004) to evaluate quality enhancement performance, which measure the improvement of the enhanced video from the compressed one. We also evaluate the complexity of a quality enhancement approach in terms of parameters and computational cost.

4.3 Comparison to State-of-the-arts

We compare the proposed method with state-of-the-art image/video quality enhancement methods, **AR-CNN** (Dong et al. 2015), **DnCNN** (Zhang et al. 2017), **RNAN** (Zhang

[2] https://hevc.hhi.fraunhofer.de

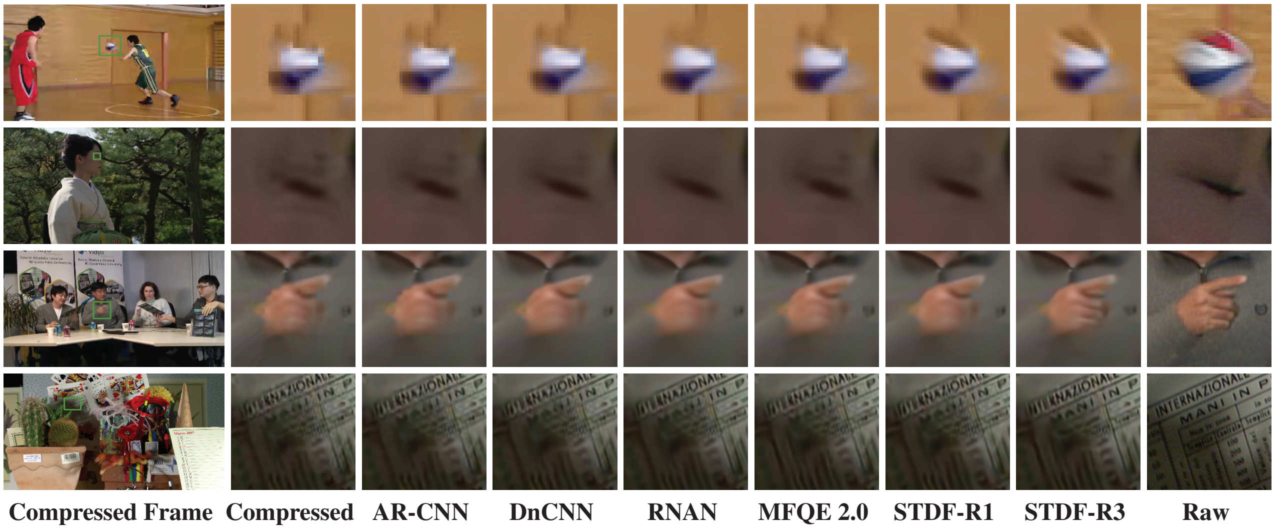


Figure 5: Qualitative results at QP 37. Note that enhancement is only conducted on luminance component for all methods. Video index (from top to bottom): *BasketballPass*, *Kimono*, *FourPeople*, *Cactus*.

et al. 2019), **MFQE 1.0** (Yang et al. 2018) and **MFQE 2.0** (Guan et al. 2019). For fair comparison, all image quality enhancement methods are retrained on our training set. Results of video quality enhancement methods are cited from (Guan et al. 2019). Three variants of our method with different configurations (refer to previous section for details) are evaluated. 1) **STDF-R1** with R=1, C_1=32, C_2=48, L=8. 2) **STDF-R3** with R=3, C_1=32, C_2=48, L=8. 3) **STDF-R3L** with R=3, C_1=64, C_2=64, L=16.

Quantitative Results. Table 1 and Table 2 present the quantitative results of accuracy and model complexity respectively. As can be observed, our method consistently outperform all compared methods in terms of average ΔPSNR and ΔSSIM on the 18 test videos. More specifically, at QP 37, our STDF-R1 outperforms MFQE 2.0 on most of the videos, with faster processing speed and comparable parameters. We note that our STDF-R1 simply takes the preceding and succeeding frames as reference, unlike MFQE 2.0 which utilizes high quality neighboring frames, thereby saving the computational cost for searching those high quality frames in advance. As temporal radius R increases to 3, our STDF-R3 manages to leverage more temporal information, and thus further improves the average ΔPSNR to 0.75 dB, which is 34% higher than MFQE 2.0, 63% higher than MFQE 1.0 and 70% higher than RNAN. Due to the high efficiency of the proposed STDF module, the overall speed of STDF-R3 is still faster than that of MFQE 2.0. Furthermore, ΔPSNR of the enlarged model STDF-R3L reaches 0.83 dB, showing there is still room for improvement of our method by optimizing network architecture. Similar results can be found for ΔSSIM as well as other QPs.

Qualitative Results. Figure 5 provides the qualitative results on 4 test videos. It can be seen that compressed frames are seriously distorted by various compression artifacts (e.g., ringing in *Kimono* and blurring in *FourPeople*). Although image quality enhancement methods can decently reduce those artifacts, the resulting frames usually become overblurred and lack of details. On the other hand, video quality enhancement methods achieve better enhancement results with the help of reference frames. Compared to MFQE 2.0, our STDF models are more robust to compression artifacts and can better explore spatio-temporal information, thereby leading to better restoration of structural details.

Quality Fluctuation. It is observed that dramatic quality fluctuation exists in compressed video (Guan et al. 2019), which may severely break temporal consistency and degrade QoE. To investigate how our method can help with this, we plot PSNR curves of 2 sequences in Figure 6. As can be seen, our STDF-R1 model can effectively enhance most of the low quality frames and alleviate quality fluctuation. By enlarging the temporal radius R to 3, our STDF-R3 model manages to take advantage of adjacent high quality frames, leading to better performance than other compared methods.

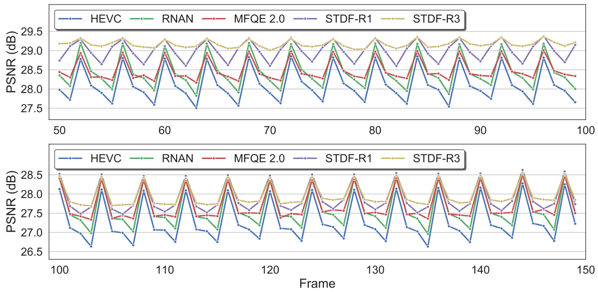


Figure 6: PSNR curves of 2 test sequences at QP 37. Top: *BQSquare*. Bottom: *PartyScene*.

4.4 Analysis and Discussions

In this section, we conduct ablation study and further analysis. For fair comparison, we only ablate the fusion scheme and fix the QE network to L=8, C_2=48. All models are trained following the same protocol and ΔPSNR/ΔSSIM are averaged over all test videos from Class B to E at QP 37. Float-point operations (FLOPs) computed on 480p video is used to evaluate the computational cost.

Effectiveness of STDF. To demonstrate the effectiveness of STDF for temporal fusion, we compare it with two previous fusion schemes, i.e., early fusion (EF) (Karpathy et al. 2014) and early fusion with motion compensation (EFMC) [3] (Yang et al. 2018; Guan et al. 2019). Specifically, for EFMC scheme, we select two CNN-based optical flow estimators for motion compensation. 1) **EF_STMC**, a lightweight Spatial Transformer Motion Compensation (STMC) network used in MFQE 2.0 (Guan et al. 2019). 2) **EF_FlowNetS**, a larger network used in FlowNet (Dosovitskiy et al. 2015). We train models under different temporal radius R to evaluate the scalability.

Figure 7 presents the comparative results. We can observe that all methods with reference frames outperform the single frame baseline, demonstrating the effectiveness of using temporal information. For R=1, STDF significantly outperforms both EF and EFMC in terms of ΔPSNR with comparable computational cost, which suggests STDF can make better use of temporal information. As R further increases, it is intriguing that ΔPSNR of EF_STMC deteriorates instead, and that of EF_FlowNetS only has marginal improvement. We think the reason is twofold. First, it is difficult for optical flow estimator to capture large temporal motion, which results in ineffective use of the added reference frames. Second, the training samples with different motion intensity may confuse the optical flow estimator, especially for EF_STMC which has relatively low capacity. In contrast, the proposed STDF takes the whole video clip into consideration, forcing the offset prediction network to simultaneously learn motion with various intensity. Thus, ΔPSNR of STDF consistently improves as R increases. In addition, the computation of STDF increases much slower than that of EF_STMC and EF_FlowNetS as R increases, which demonstrates the higher efficiency of STDF.

Effectiveness of STDC. The proposed STDC features position-specific sampling for temporal fusion, which enables higher flexibility and robustness than traditional fusion scheme with motion compensation (EFMC). To verify this, we introduce a variant of our method that replaces STDC with EFMC. Specifically, the optical flow estimator in EFMC is modified from the offset prediction network, where the output layer of the network is revised for flow estimation instead of offset prediction. According to Table 3, although parameters and FLOPs slightly improve when replacing STDC with EFMC, the overall ΔPSNR at R=1 and R=3 drops by 0.04 dB and 0.08 dB respectively. This demonstrates the effectiveness of the proposed STDC.

[3] EFMC scheme applies optical flow based motion compensation to mitigate temporal motion before early fusion.

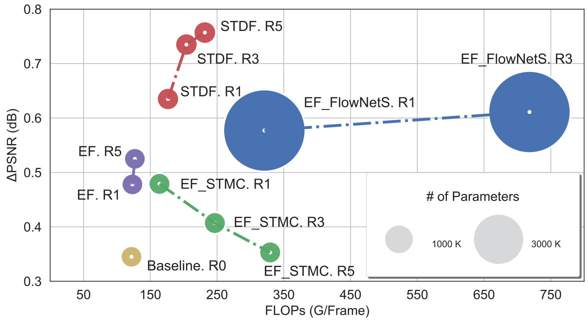


Figure 7: Comparison of temporal fusion schemes.

Fusion Scheme	R	ΔPSNR	ΔSSIM	#Param(K)	FLOPs(G)
EFMC & Joint	1	0.60	0.0104	308.17	159.22
	3	0.66	0.0116	313.94	163.82
STDC & Pairwise	1	0.59	0.0105	313.95	245.10
	3	0.65	0.0117	316.25	409.49
STDC & Joint	1	0.64	0.0117	329.84	176.47
	3	0.74	0.0133	364.51	204.08

Table 3: Ablation study on convolution and offset prediction.

Effectiveness of Joint Offset Prediction. Joint prediction scheme is introduced to generate deformable offsets for STDC. To demonstrate its effectiveness, we replace it with pairwise prediction scheme. Specifically, we modify the input and output layers of offset prediction network, and conduct offset prediction separately for each reference-target pair. From Table 3 we can see that ΔPSNR and ΔSSIM with pairwise scheme are reduced, while FLOPs is greatly increased, which shows the proposed joint prediction scheme can better exploit temporal information with high efficiency.

5 Conclusion

We have presented a fast yet effective method for compressed video quality enhancement, which incorporates a novel spatio-temporal deformable convolution to aggregate temporal information from consecutive frames. Our method performs favorably against previous methods in terms of both accuracy and efficiency on benchmark dataset. We believe the proposed spatio-temporal deformable convolution can also extend to other video related low-level vision tasks, including super-resolution, restoration and frame synthesis, for efficient temporal information fusion.

Acknowledgments. This work was partially supported by NSFC (Grant No.61601406, No.61974133), and Guangdong Province (Grant No.2018B030338001).

References

Bertasius, G.; Torresani, L.; and Shi, J. 2018. Object detection in video with spatiotemporal sampling networks. In *ECCV*, 342–357.

Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 4778–4787.

Chang, H.; Ng, M. K.; and Zeng, T. 2013. Reducing artifacts in jpeg decompression via a learned dictionary. *TSP* 62(3):718–728.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *ICCV*, 764–773.

Dong, C.; Deng, Y.; Change Loy, C.; and Tang, X. 2015. Compression artifacts reduction by a deep convolutional network. In *ICCV*, 576–584.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2758–2766.

Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. *TIP* 16(5):1395–1411.

Galteri, L.; Seidenari, L.; Bertini, M.; and Bimbo, A. D. 2017. Deep generative adversarial compression artifact removal. In *ICCV*, 4836–4845.

Guan, Z.; Xing, Q.; Xu, M.; Yang, R.; Liu, T.; and Wang, Z. 2019. Mfqe 2.0: A new approach for multi-frame quality enhancement on compressed video. *TPAMI*.

Guo, J., and Chao, H. 2016. Building dual-domain representations for compression artifacts reduction. In *ECCV*, 628–644.

Jancsary, J.; Nowozin, S.; and Rother, C. 2012. Loss-specific training of non-parametric image restoration models: A new state of the art. In *ECCV*, 112–125.

Kappeler, A.; Yoo, S.; Dai, Q.; and Katsaggelos, A. K. 2016. Video super-resolution with convolutional neural networks. *TCI* 2(2):109–122.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*, 1725–1732.

Kim, T.; Sajjadi, M. S. M.; Hirsch, M.; and Bernhard, S. 2018. Spatio-temporal transformer network for video restoration. In *ECCV*, 111–127.

Kim, J. W.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 1646–1654.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Lu, G.; Ouyang, W.; Xu, D.; Zhang, X.; Cai, C.; and Gao, Z. 2019. Dvc: An end-to-end deep video compression framework. In *CVPR*, 11006–11015.

Niklaus, S.; Mai, L.; and Liu, F. 2017. Video frame interpolation via adaptive separable convolution. In *ICCV*, 261–270.

Ohm, J.-R.; Sullivan, G. J.; Schwarz, H.; Tan, T. K.; and Wiegand, T. 2012. Comparison of the coding efficiency of video coding standards—including high efficiency video coding (hevc). *TCSVT* 22(12):1669–1684.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 802–810.

Sullivan, G. J.; Ohm, J. R.; and Wiegand, T. 2013. Overview of the high efficiency video coding (hevc) standard. *TCSVT* 22(12):1649–1668.

Tai, Y.; Yang, J.; Liu, X.; and Xu, C. 2017. Memnet: A persistent memory network for image restoration. In *ICCV*, 4549–4557.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2018. Tdan: Temporally deformable alignment network for video super-resolution. *arXiv:1812.02898*.

VQEG. Vqeg video datasets and organizations. https://www.its.bldrdoc.gov/vqeg/video-datasets-and-organizations.aspx.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP* 13(4):600–612.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Loy, C. C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*.

Wiegand, T.; Sullivan, G. J.; Bjontegaard, G.; and Luthra, A. 2003. Overview of the h.264/avc video coding standard. *TCSVT* 13(7):560–576.

Wien, M. 2015. *High Efficiency Video Coding*. Springer.

Xiph.org. Xiph.org video test media (derf's collection). https://media.xiph.org/video/derf/.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2017. Video enhancement with task-oriented flow. *IJCV* 127(1):1106–1250.

Yang, R.; Xu, M.; Wang, Z.; and Li, T. 2018. Multi-frame quality enhancement for compressed video. In *CVPR*, 6664–6673.

Yang, R.; Sun, X.; Xu, M.; and Zeng, W. 2019. Quality-gated convolutional lstm for enhancing compressed video. In *ICME*, 532–537.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*, 818–833.

Zhang, X.; Xiong, R.; Fan, X.; Ma, S.; and Gao, W. 2013. Compression artifact reduction by overlapped-block transform coefficient estimation with block similarity. *TIP* 22(12):4613–4626.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP* 26(7):3142–3155.

Zhang, Y.; Li, K.; Li, K.; Zhong, B.; and Fu, Y. 2019. Residual non-local attention networks for image restoration. In *ICLR*.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In *CVPR*, 9308–9316.