DCNGAN: A Deformable Convolution-Based GAN with QP Adaptation for Perceptual Quality Enhancement of Compressed Video

Saiping Zhang, Luis Herranz, Marta Mrak, Marc Gorriz Blanch, Shuai Wan and *Fuzheng Yang*

Email: spzhang@stu.xidian.edu.cn

Paper ID: 1772

Backgrounds

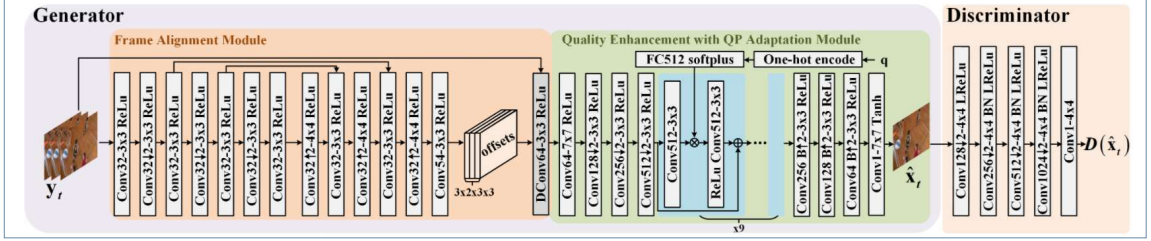
1.Video compression methods, such as HEVC and VVC, are indispensable to remove the spatial and temporal redundancy in videos and reduce bit-rates, due to the limitation of bandwidth.
2.Compressed videos, especially at low bit rate, suffer from the degraded quality due to compression artifacts.
3.It is crucial to enhance the quality of compressed videos.
4.Previous works mainly focus on enhancing the objective quality of compressed videos. However, sometimes the objective quality is inconsistent with the perceptual quality.
5.Although some of previous works can also improve the perceptual quality, they either introduce new artifacts or have poor performance.
6.Previous compressed video quality methods require training and storing various models to enhance videos compressed at different QPs, which sets high demand on the memory.

*It is of great importance to develop a compressed video perceptual quality enhancement method which can not only **achieve advanced performance** but also be **QP adaptive**.*

Contributions

1.A GAN framework based on deformable convolutions to enhance the perceptual quality of compressed videos.
2.A single adaptive model to enhance videos compressed at various QPs.
3.We compare the proposed DCNGAN with state-of-the-art compressed video quality enhancement networks, showing its superior performance.

Network Architecture

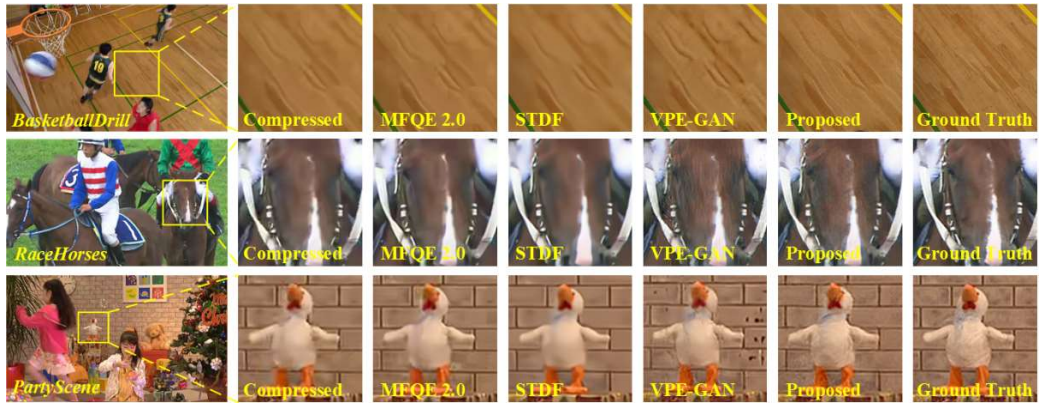


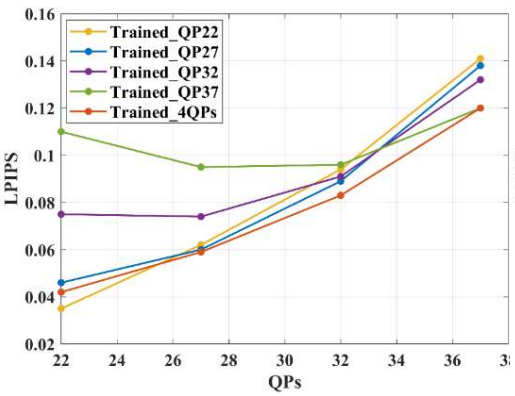
Proposed Method

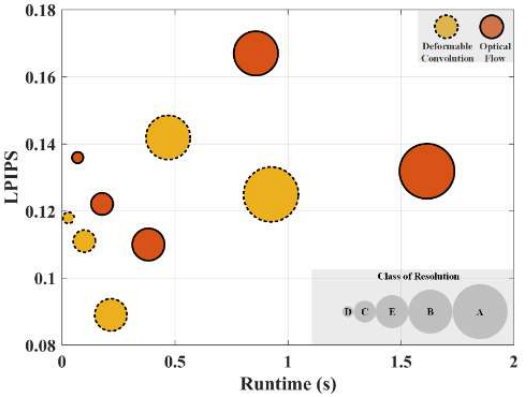
Frame Alignment Module: Three consecutive frames (i.e., previous, current, and next frames) are the **input**. The **deformable convolution** is used to **align the input**. The key to the correct alignment is the computation of the **offsets**, which are predicted by a network based on **U-net**. A single representation which already integrates information from the three frames is the **output**.

Quality Enhancement with QP Adaptation Module: When aligning frames, deformable convolutions are **more efficient** than optical flows. The single representation which is outputted by the deformable convolution is the **input**. The module is based on an **encoder-decoder structure with 9 residual blocks**. Encoded QP information q is embedded into each residual block to make the network **modulated by QP values**. The enhanced frame is the **output**.

Results







Discriminator: Real/fake frames are the **input** of the **patch discriminator**. It is implemented in a **fully convolutional fashion**. The average probabilities over patches in the frame being real or fake is the **output**.

$$Loss = \min_G \max_D \left(L_{gan}(G,D) + L_{vgg}(G) + L_{fm}(G,D)\right)$$

where

$$\begin{cases} L_{gan}(G,D) = \mathbb{E}_{(\mathbf{y},\mathbf{q})}\left[\left(D\left(G\left(\mathbf{y},\mathbf{q}\right)\right)-1\right)^2\right] + \mathbb{E}_{\mathbf{x}}\left[D(\mathbf{x})^2\right] \\ L_{vgg}(G) = \mathbb{E}_{(\mathbf{y},\mathbf{q},\mathbf{x})} \sum_{i=1}^{N_f} \frac{1}{M_i} \sum_{j=1}^{M_i} \left\| f_j^i(\mathbf{x}) - f_j^i\left(G\left(\mathbf{y},\mathbf{q}\right)\right)\right\|_1 \\ L_{fm}(G,D) = \mathbb{E}_{(\mathbf{y},\mathbf{q},\mathbf{x})} \sum_{i=1}^{y} \frac{1}{M_i^g} \sum_{j=1}^{x} \left\| g_j^i(\mathbf{x}) - g_j^i\left(G\left(\mathbf{y},\mathbf{q}\right)\right)\right\|_1 \end{cases}$$