# Spatial-Temporal Fusion Convolutional Neural Network for Compressed Video enhancement in HEVC

Xiaoyu Xu*, Jian Qian*, Li Yu*, Hongkui Wang*, Hao Tao*, Shengju Yu*

*School of Electron. Inf. & Commun., Huazhong Univ. of Sci. & Tech., Wuhan, China
{hustxyxu,qianjian,hustlyu,hkwang,husthtao,shengju_yu}@hust.edu.cn

Most existing methods for compressed video enhancement focus on single frame processing where copious temporal and spatial information is neglected. In this paper, we propose a spatial-temporal fusion convolutional neural network (STEF-CNN) to employ spatial and temporal information to improve the performance of in-loop filter in HEVC.

In the proposed method, we use two stage enhancement which is to pre-denoise the compressed videos firstly and enhance them later with the STEF-CNN. The pre-denoising part employs residual learning and dense network to enhance its performance. The STEF-CNN contains three parts: image alignment (IA), spatial-temporal fusion (STF) and quality enhancement (QE). Specifically speaking, the pre-denoising and QE parts have similar structure which cascade $M$ residual blocks together and forms a dense block. Then all output of dense blocks are concatenated together. The IA part takes frame pair as input, then up-samples and down-samples the frame pair to different scales. After, the deformable convolutions are utilized to generate aligned frames. The process is formulated as: $F_o = GConv([F_a^0, F_a^{0d}, F_a^{0u}])$, $F_a^{ju} = GConv((F_a^{(j+1)u})^{\downarrow 2}, DConv(F_{t+i}^{ju}, \Theta^j))$,$F_a^{jd} = GConv((F_a^{(j+1)d})^{\uparrow 2}, DConv(F_{t+i}^{jd}, \Theta^j))$. $F_o$ is the final aligned image, $F_a^0$ denotes aligned feature from the initial image pair, $F_a^{jd}, F_a^{ju}$ represent aligned features from down-sampled and up-sampled image pairs respectively. $GConv(\cdot)$ and $DConv(\cdot)$ denote general Convolution and deformable convolution operations. The $\Theta^j$ is offset parameter. In STF part, a spatial attention feature is firstly extracted from each frame. With element-wise multiplication between each frame and its attention feature, useful regions are maintained and useless regions are masked. The process is formulated as: $y = \sigma(M_{SA}(x)) \times \sigma(\frac{1}{H \times W} \sum_W \sum_H (GConv(x)))$. Where $\sigma(\cdot)$ is sigmoid activation function, $M_{SA}$ denotes spatial attention model. $H, W$ denotes the size of input frame $x$. Extensive experimental results demonstrate the effectiveness of the proposed method. The STEF-CNN achieves 11.53% BD-BR reduction in all-intra (AI) configuration and 10.20% BD-BR reduction in random-access (RA) configuration.

## Acknowlegement

This work is supported in part by the National Natural Science Foundation of China under Grant 61871437 and in part by the Natural Science Foundation of Hubei Province of China under Grant 2019CFA022.