Real-time video super resolution network using recurrent multi-branch dilated convolutions – 2021

The paper discusses an approach that addresses drawbacks of approaches like VESPCN and 3DVSRnet, RNN architecture.

Named – Efficient Recurrent Video Super Resolution Network, henceforth mentioned as ERSVR, it uses Multi-branch dilate (MBD) mode, that extracts spatial-temporal features at different scales in parallel resulting in superior performance.

ERSVR is based on depth-wise separable convolution to reduce overall computational complexity, so that proposed network can reconstruct video clip up to 50 fps when inferencing on a single NVIDIA 980 Ti GPU.

Methodology used:

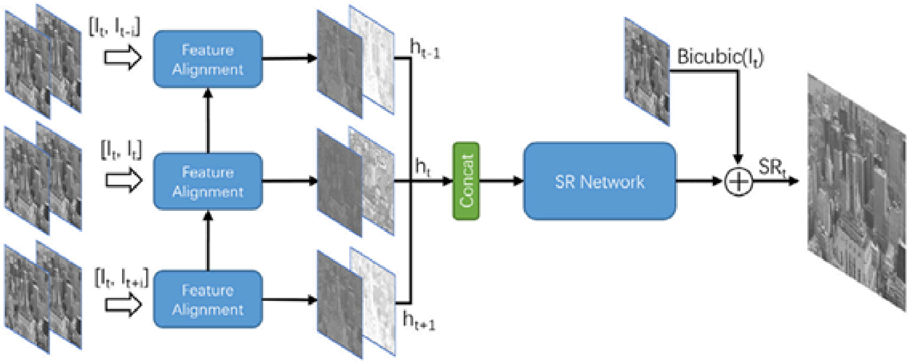


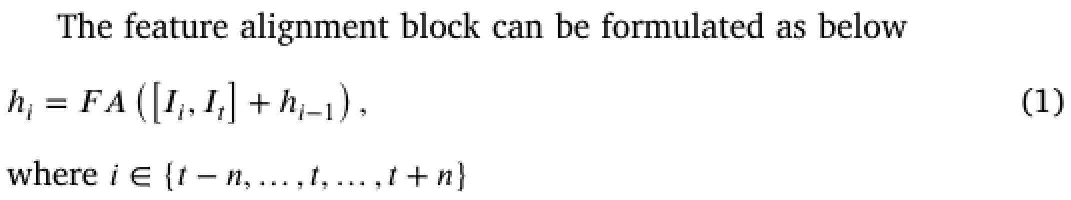
Fig. 1. Architecture of our proposed video super resolution network, where *I* denotes input LR images, *h* are featuremaps from feature alignment block and subscripts $t-1, t$ *and* $t+1$ are temporal position. Note that intensities of feature maps are normalized for better visual effects.

Concatenate input frames at I_{t-1}, I_t and I_{t+1} concatenated with reference frame I_t.

There are 2 blocks that are important:
The feature alignment block – generates aligned maps h_{t-1}, h_t, h_{t+1} with respect to each input.

The second is the SR Network, that focuses all concatenated feature maps and processes spatial upscaling

At the rear end of network, add the bicubic upscaled reference frame to the output of the SR network as global residual connection so network can focus on refining high frequency details and textures. In MBD module, input feature maps are embedded with a pointwise convolution and processed in parallel with convolutional operations of different dilation factors.

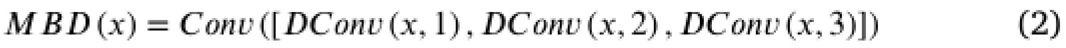


The feature alignment block can be formulated as below

$$h_i = FA\left(\left[I_i, I_t\right] + h_{i-1}\right), \quad (1)$$

where $i \in \{t-n, \ldots, t, \ldots, t+n\}$

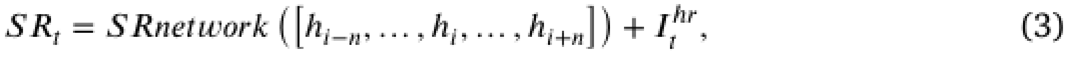
Then MBD module fuses all extracted feature with a convolutional layer. The MBD module can be formulated as follow:



$$MBD(x) = Conv([DConv(x,1), DConv(x,2), DConv(x,3)]) \quad (2)$$

where *Conv*(*x*) denotes standard convolution of input *x*, *DConv*(*x*,*d*) denotes a pointwise convolution and a dilated convolution of input *x* with dilation factor *d*.

For SR network, we adopt a 9-layer ESPCN backbone network. SR network can be formulated as:



$$SR_t = SRnetwork\left(\left[h_{i-n}, \ldots, h_i, \ldots, h_{i+n}\right]\right) + I_t^{hr}, \quad (3)$$

Where h_i à feature map at temporal position t

I_t^{hr} à Bicubic Interpolation of low-res input image

Introduced sub-pixel convolution (also known as periodic shift operation) for efficient spatial up-sampling. Let *r* to be the upscaling factor and *H*,*W*,*C* represent height, width, depth of tensor respectively. For upscaling factor*r*=4, we adopt pyramid up-sampling which operates twice up-sampling of scale factor of 2.

ERVSR can be decomposed into two parts, feature alignment module and up-sampling network. Feature alignment block consists of 5convolutional layers and the MBD module and Up-sampling Network consists of 10 convolutional layers. Number of convolutional Filters=32.

Trained on – Vimeo 90k, learning rate 10^{-3}, ADAM optimizer with momentum of 0.9. Network converges at 800 epochs.

ERSVR requires 3.8 ms per frame.