Supplementary: Efficient Space-time Video Super Resolution using Low-Resolution Flow and Mask Upsampling

Saikat Dutta  
IIT Madras, India

Nisarg A. Shah  
IIT Jodhpur, India

Anurag Mittal  
IIT Madras, India

1. Details of architectures

1.1. Gridnet

GridNet [1] is an encoder-decoder architecture. In GridNet, encoder and decoder blocks are arranged in a grid-like structure which allows network to fuse information from different scales. Instead of using Deconvolution for upsampling, we use bilinear upsampling as deconvolution produces checkerboard artifacts in generated images [3, 2]. We have tried a 3-level Gridnet in this work. Channel sizes used in each level are 32, 64, 96 from top to bottom.

Architecture of Gridnet is shown in Fig. 1. Lateral blocks consist of two convolutional layers and a skip connection. Upsampling blocks contain one bilinear upsampling layer followed by two convolution layers. Downsampling blocks consist of one stride pool layer and a convolutional layer.

2. UNet

We have used similar UNet architecture as in [4]. The encoder part has 12 convolutional layers and 5 average pooling layers and the decoder part has 10 convolutional layers and 5 bilinear upsampling layers. The first two convolutional layers in the encoder have $7 \times 7$ kernels and the following two layers use $5 \times 5$ kernels. Rest of the convolutional layers use $3 \times 3$ kernels. Features from encoder layers are concatenated in corresponding decoder layer as shown in Fig. 2.

3. UNet++

UNet++ is a Nested UNet architecture originally used for Medical Image Segmentation [5]. UNet++ bridges the semantic gap between encoder and decoder feature maps using dense skip connections. Instead of using 5-level pyramid in the original paper, we use 4-level pyramid to reduce the number of parameters and model complexity. Max-pooling is used for downsampling and bilinear upsampling is used for upsampling the feature maps. Convolutional blocks consist of two convolutional layers with kernel size of $3 \times 3$, except the output block which has only one convolutional layer with kernel size $1 \times 1$. Channel sizes in pyramid levels are 32, 64, 96, 128 from top to bottom respectively. Architecture diagram of UNet++ is shown in Fig. 3.
References


