Supplementary Material of Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation

Huaizu Jiang[1] Deqing Sun[2] Varun Jampani[2]
Ming-Hsuan Yang[3,2] Erik Learned-Miller[1] Jan Kautz[2]
[1]UMass Amherst [2]NVIDIA [3]UC Merced
{hzjiang,elm}@cs.umass.edu,{deqings,vjampani,jkautz}@nvidia.com, mhyang@ucmerced.edu

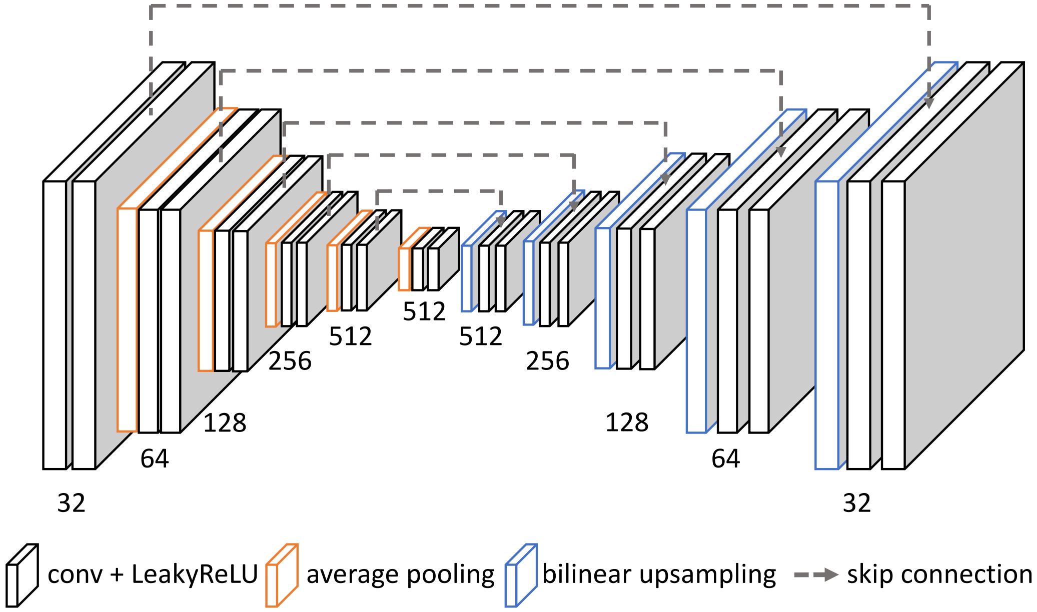


Figure 1: Illustration of the architecture of our flow computation and flow interpolation CNNs.

1. Network Architecture

Our flow computation and flow interpolation CNNs share a similar U-Net architecture, shown in Figure 1. The U-Net is a fully convolutional neural network, consisting of an encoder and a decoder, with skip connections between the encoder and decoder features at the same spatial resolution. For both networks, we have 6 hierarchies in the encoder, consisting of two convolutional and one Leaky ReLU ($\alpha = 0.1$) layers. At the end of each hierarchy except the last one, an average pooling layer with a stride of 2 is used to decrease the spatial dimension. There are 5 hierarchies in the decoder part. At the beginning of each hierarchy, a bilinear upsampling layer is used to increase the spatial dimension by a factor of 2, followed by two convolutional and Leaky ReLU layers.

For the flow computation CNN, it is crucial to have large filters in the first few layers of the encoder to capture long-range motion. Therefore, we use 7×7 kernels in the first two convolutional layers and 5×5 in the second hierarchy. For the rest layers in the flow computation CNN, we use 3×3 convolutional kernels.

Similarly, for the flow interpolation CNN, 7×7 and 5×5 kernels are used in the first and second hierarchies, respectively. For the rest layers, we use 3×3 convolutional kernels.

2. Visual Comparisons on UCF101

Figure 2 and Figure 3 show visual comparisons of single-frame interpolation results on the UCF101 dataset. For more visual comparisons, please refer to our supplementary video http://jianghz.me/projects/superslomo/superslomo_public.mp4.

References

[1] E. Herbst, S. Seitz, and S. Baker. Occlusion reasoning for temporal interpolation using optical flow. Technical report, August 2009. 3, 4

[2] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 3, 4

[3] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, 2017. 3, 4

[4] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung. Phase-based frame interpolation for video. In *CVPR*, 2015. 3, 4

[5] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, 2017. 3, 4

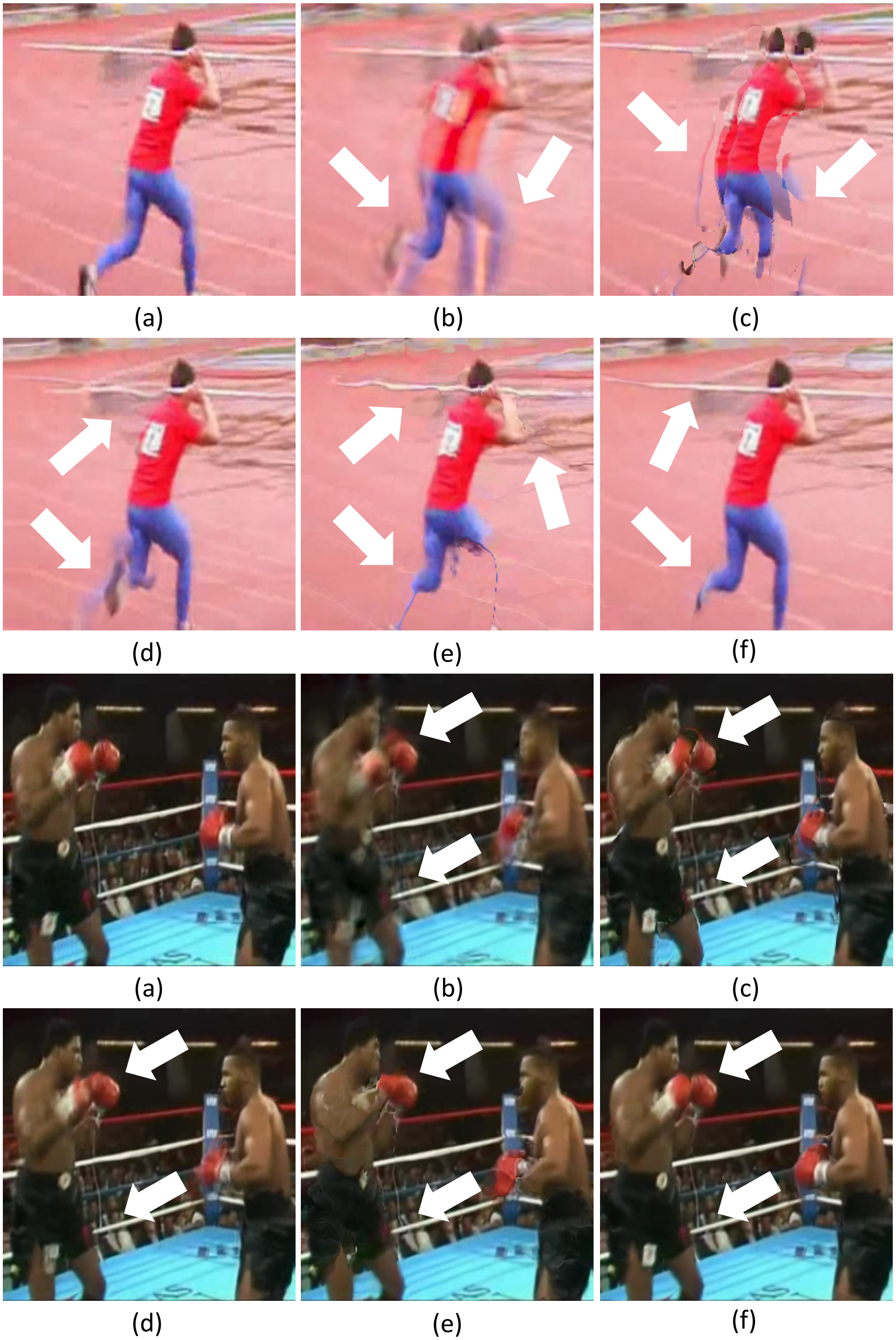
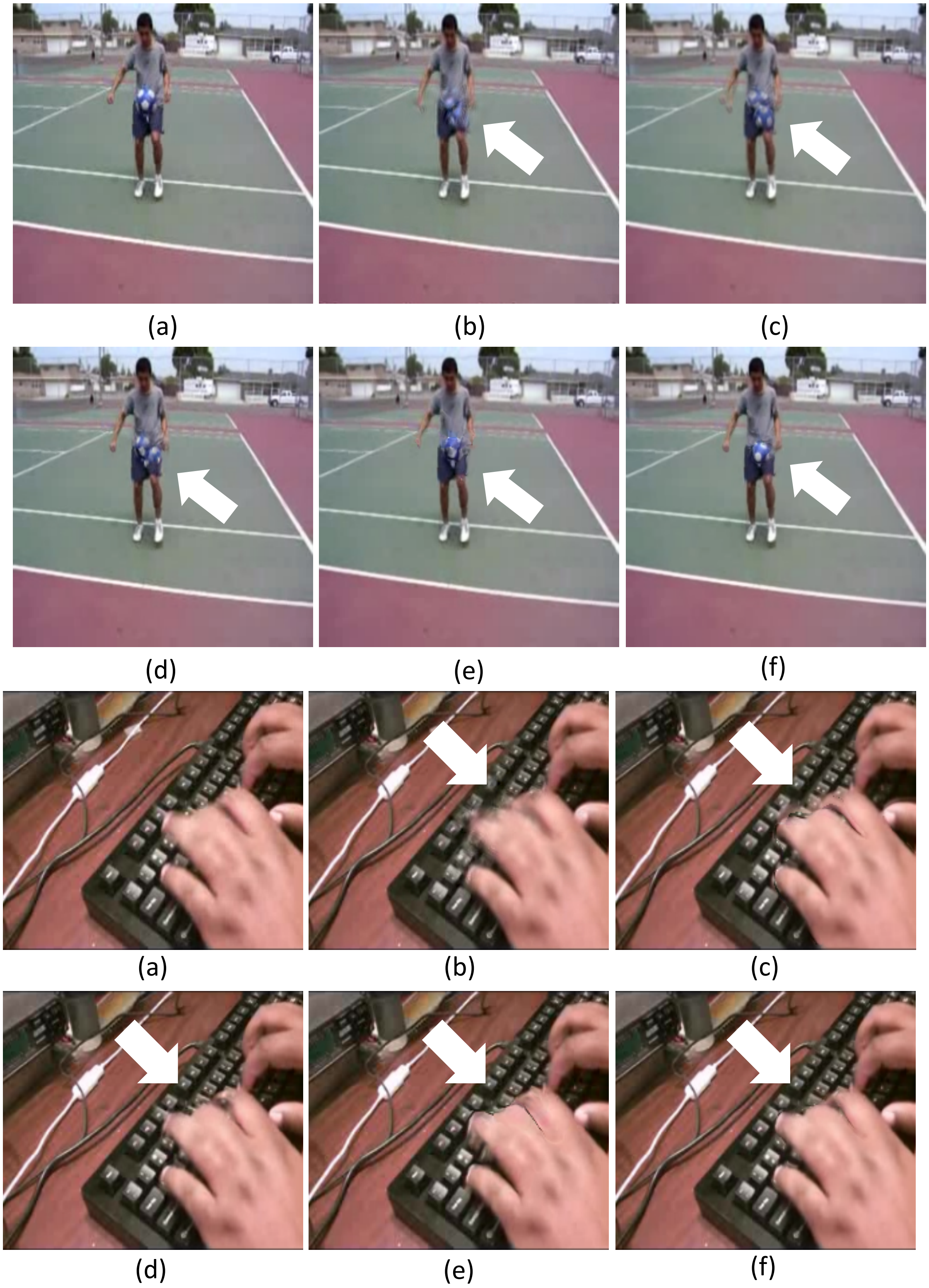


Figure 2: Visual comparisons on the UCF101 dataset. (a) Ground truth in-between frame, interpolation results from (b) PhaseBased [4], (c) FlowNet2 [1, 2], (d) SepConv [5], (e) DVF [3], and (f) Ours.



‘’

Figure 3: Visual comparisons on the UCF101 dataset. (a) Ground truth in-between frame, interpolation results from (b) PhaseBased [4], (c) FlowNet2 [1, 2], (d) SepConv [5], (e) DVF [3], and (f) Ours.