Supplementary Material
Zooming Slow-Mo: Fast and Accurate One-Stage Space-Time Video Super-Resolution

Xiaoyu Xiang[1,*], Yapeng Tian[2,*], Yulun Zhang[3], Yun Fu[3], Jan P. Allebach[1,†], Chenliang Xu[2,†]
[1]Purdue University, [2]University of Rochester, [3]Northeastern University
{xiang43,allebach}@purdue.edu, {yapengtian,chenliang.xu}@rochester.edu,
yulun100@gmail.com, yunfu@ece.neu.edu

In this supplementary material, we have a demo to provide video results of our STVSR method and also compare it to the best performing two-stage network: DAIN [1]+EDVR [2] among all compared methods. In addition, we further clarify the implementation details of our network architecture.

Network Architecture

We further illustrate the feature temporal interpolation network in Figure 1 and the proposed STVSR framework in Figure 2 to help readers better understand the overall structure of our proposed network.

To make our paper be concise and easy to follow, we use a simple version of deformable sampling to introduce the proposed feature temporal interpolation and deformable ConvLSTM. However, in our implementation, as stated in Section 3.4 of the paper, we adopt a Pyramid, Cascading and Deformable (PCD) structure as in [2] to implement the deformable sampling, which can exploit multi-scale contexts with a feature pyramid. The official PyTorch implementation of the PCD can be found in https://github.com/xinntao/EDVR.

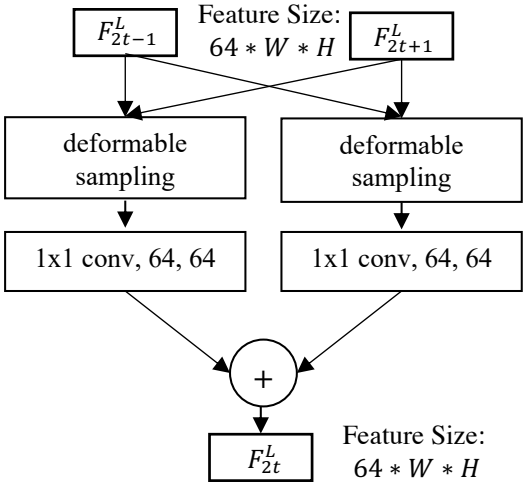

Figure 1: Feature temporal interpolation for intermediate LR frames. It will predict an intermediate LR frame feature map F^L_{2t} from two neighboring feature maps: F^L_{2t-1} and F^L_{2t+1}, where t = 1, 2, ..., n. Note that the deformable sampling module on the left samples features from F^L_{2t-1} with generated sampling parameters from both F^L_{2t-1} and F^L_{2t+1}; on the contrary, the deformable sampling module on the right samples features from F^L_{2t+1}.

*Equal contribution; †Equal advising.

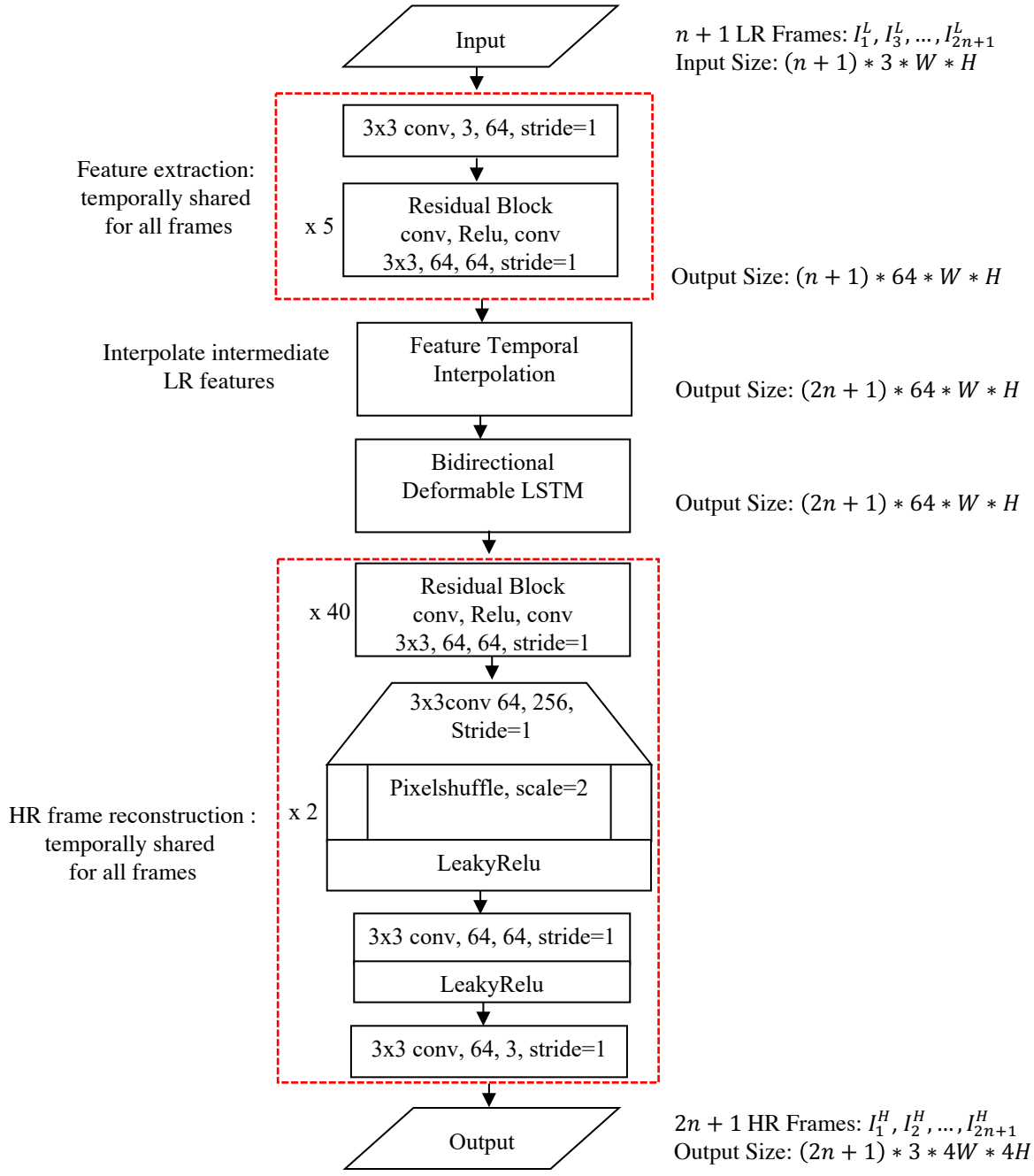

Figure 2: Flowchart of the proposed one-stage STVSR framework. The feature extraction and HR frame reconstruction networks are temporally shared for all frames, in which different frames are processed independently.

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 1

[2] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1