Supplementary: Stochastic Dynamics for Video Infilling

Qiangeng Xu[1] Hanwang Zhang[2] Weiyue Wang[1] Peter N. Belhumeur[3] Ulrich Neumann[1]
[1]University of Southern California [2]Nanyang Technological University [3]Columbia University
{qiangenx,weiyuewa,uneumann}@usc.edu hanwangzhang@ntu.edu.sg belhumeur@cs.columbia.edu

Appendices

A. Architecture and Training Details

$Encoder$, $Extractor$ and $Decoder$ use the same architecture of DCGAN. For step S and T, feature maps of all layers in $Encoder$ will be gathered as a multi-scale residuals ctn_S and ctn_S to help reconstruct the static content. $LSTM_{infr}$, $LSTM_{pst}$ and $LSTM_{dyn}$ use the structure of one layer ConvLSTM. The output dimensions of our modules are listed in Table 1. Our reported result is created using one extended references frame on each side (totally 2 extended reference frames). These two extra references bring extra long-term information into Reference module. However, frames too far away from the interval would contain too much unrelated information. We also tested on 4 extended reference frames and find the benefit is insignificant.

For the evaluation presented in our paper, all datasets have been trained with a input frame dimension of 64×64 with a interval of 7 frames. We also train on KTH with a input frame dimension of 128×128 (See Section 2 in the video web page) and BAIR with intervals of 9 frames (See Section 1 in the video web page). We use standard Adam optimizer with 0.5 as the first momentum decay rate. All settings of the hyper parameters for different datasets are shown in 2. The β is initially set to 1 and gradually reduce to 0.4. To prevent the accumulation of errors, we will first roll back the cell state of $LSTM_{infr}$ to $t-1$, and then input h_t after inferring $N_{infr}(\mu_t, \sigma_t)$. This operation has been proved to be crucial to our result. On the early stage of the training, a less meaningful $\hat{h}_t$ would accumulatively disturb the cell state of $LSTM_{infr}$ and lead to a slow convergence.

B. Dataset Details

SMMNIST: Sequences were generated on the fly by randomly choosing two digits from MNIST: 50k digits from MNIST training set for training, 10k digits for validation, and 10k in MNIST testing set for testing. We create the ground truth video frames as 16 fps.

KTH: We used person 1-14 (1337vids) for training, 15-16(190vids) for validation and 17-25 (863vids) for testing. We sample the ground truth video frames as 12 fps.

Bair: By default, we use 40000 scenes for training, 3863 for validation and 256 for testing. We sample the ground truth video frames as 16 fps.

UCF101: The dataset contains 101 realistic human actions taken in a wild and exhibits various challenges, such as background clutter, occlusion, and complicated motion. The training set contains 3223 video sequences with varying length, and the test set contains 557 video sequences. We sample the video as 16 fps so that the input reference frames for the network are 2 fps.

Feature	Dim 1	Dim 2	Dim 3
C_{start}	4	4	256
C_{end}	4	4	256
h_t	4	4	256
$\hat{h}_t$	4	4	256
σ_t SMMNIST & KTH	4	4	32
μ_t SMMNIST & KTH	4	4	32
σ_t BAIR	4	4	64
μ_t BAIR	4	4	64

Table 1: The dimensionalities of different features

Training Parameters	SMMNIST	BAIR	KTH
α	0.002	0.0002	0.0002
β	1 to 0.4	1 to 0.4	1 to 0.4
Map Weight η	N/A	2	3

Table 2: Hyper parameters for training on different datasets

C. More Results

Figure 3 provides another UCF101 comparision. Results of more conditions and more promising results can be found in the "video_result.html". Please open the webpage in your browser to see (the web page contains gif videos such as: Figure 2). We use the full datasets and train all actions together. We will include more videos generated by our ablation studies and the comparative models in our project website.

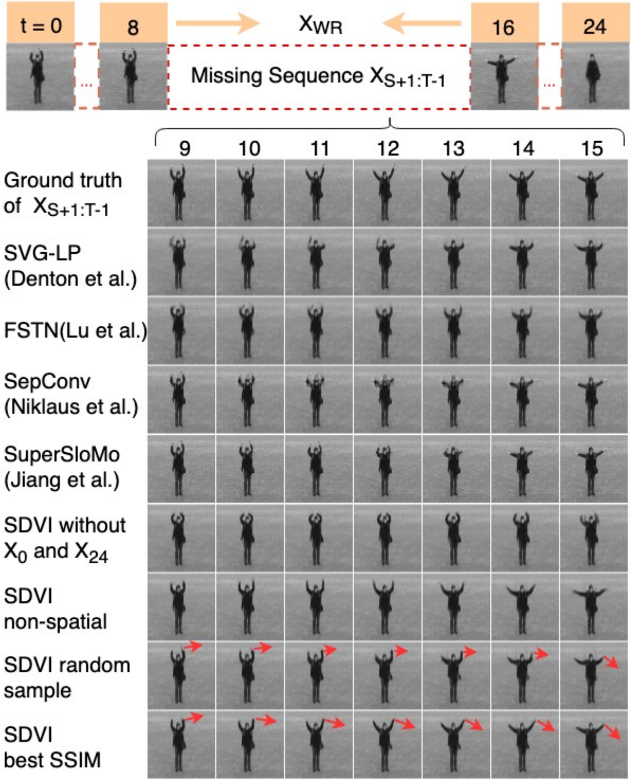

Figure 1: shows the full comparisons for he wave action.

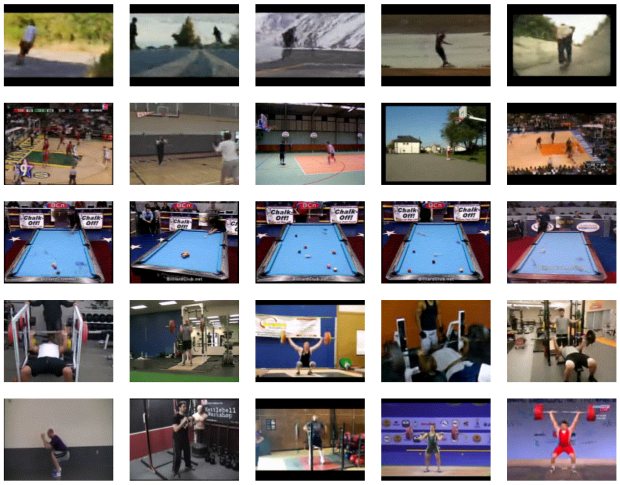

Figure 2: a snapshot of the gifs in the "video_result.html"

References

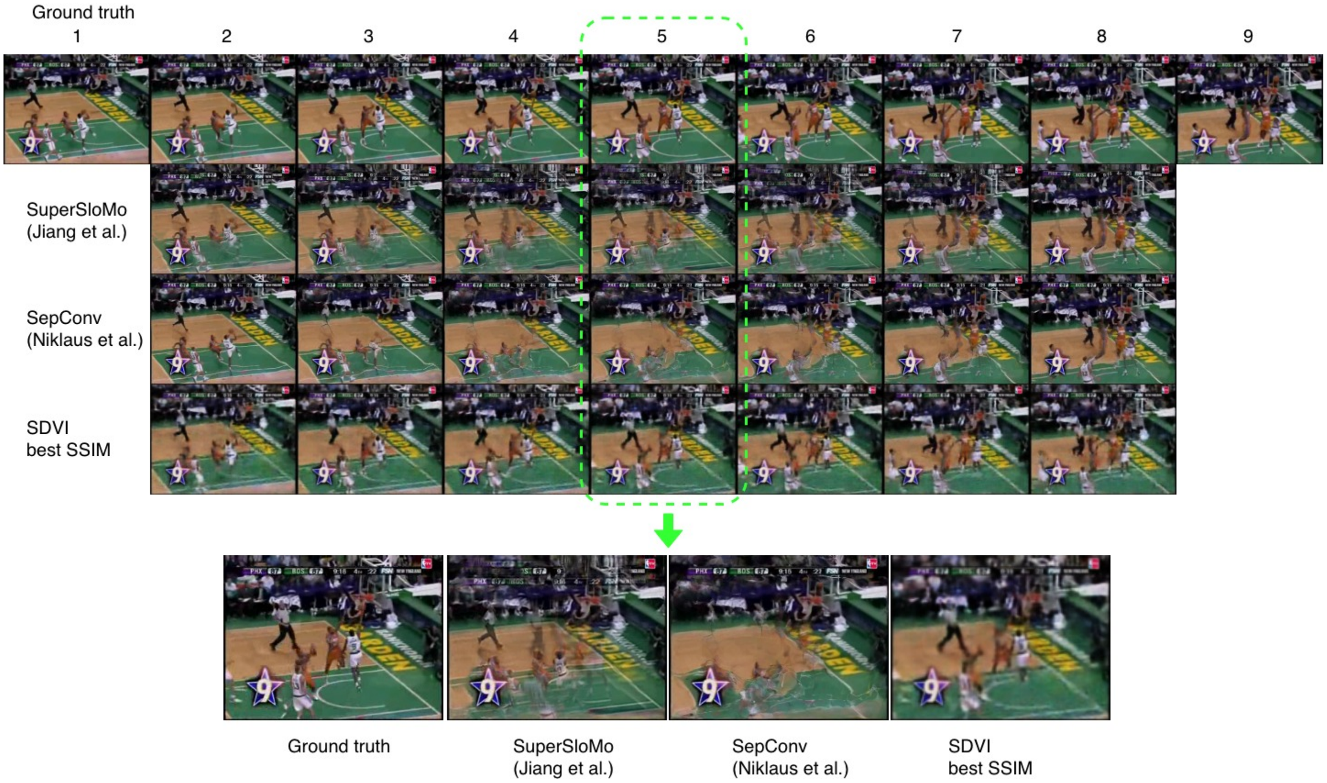

Figure 3: A more complicated UCF101 example: a real basketball video sequence involving multiple objects. Our method can model the dynamic correctly and generate better moving objects than SuperSloMo and SepConv.