Video Prediction Recalling Long-term Motion Context via Memory Alignment Learning
- *Supplementary Material* -

Sangmin Lee[1] Hak Gu Kim[2] Dae Hwi Choi[1] Hyung-Il Kim[3] Yong Man Ro[1]
[1] Image and Video Systems Lab, KAIST, South Korea
[2] EPFL, Switzerland [3] ETRI, South Korea
{sangmin.lee,sinjh1796,ymro}@kaist.ac.kr hakgu.kim@epfl.ch hikim@etri.re.kr

1. Effects of Memory Size

We conduct experiments to observe the effects of the memory size s on the video prediction performance. The memory size s indicates the number of slots in the memory. In the experiments, s is varied with an exponential scale (10, 50, 100, 500, and 1000) on the KTH Action dataset. Figure 1 shows the experiments results for s. As shown in the figure, when the memory size exceeds 50, it shows relatively stable results for memory size changing. This result represents the robustness to the setting of memory size.

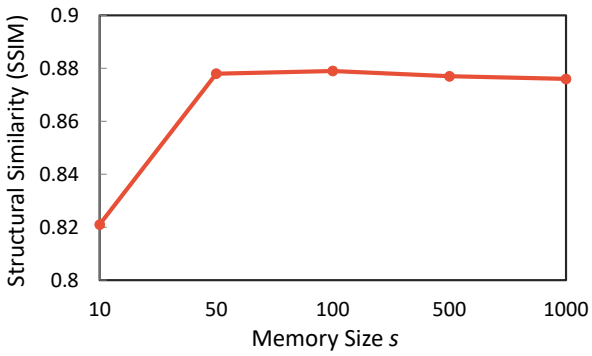

Figure 1: Effects of the memory size s on the SSIM performance. Memory size is varied with 10, 50, 100, 500, and 1000 on the KTH Action dataset.

2. Network Structure Details

Spatial Encoder / ConvLSTMs / Decoder / Attention		
Layer	**Filter / Stride**	**Output Size** $(width \times height \times channel)$
2D-Conv 1	3×3 / (2, 2)	$(W/2)\times(H/2)\times64$
2D-Conv 2	3×3 / (1, 1)	$(W/2)\times(H/2)\times64$
2D-Conv 3	3×3 / (2, 2)	$(W/4)\times(H/4)\times128$
2D-Conv 4	3×3 / (1, 1)	$(W/4)\times(H/4)\times128$
ConvLSTM 1	3×3 / (1, 1)	$(W/4)\times(H/4)\times128$
ConvLSTM 2	3×3 / (1, 1)	$(W/4)\times(H/4)\times128$
ConvLSTM 3	3×3 / (1, 1)	$(W/4)\times(H/4)\times128$
ConvLSTM 4	3×3 / (1, 1)	$(W/4)\times(H/4)\times128$
2D-DeConv 1	3×3 / (1, 1)	$(W/4)\times(H/4)\times128$
2D-DeConv 2	3×3 / (2, 2)	$(W/2)\times(H/2)\times64$
2D-DeConv 3	3×3 / (1, 1)	$(W/2)\times(H/2)\times64$
2D-DeConv 4	3×3 / (2, 2)	$W\times H\times C$
Avg Pooling	Global×Global / (Global, Global)	1×1×256
FC 1	-	1×1×16
FC 2	-	1×1×128

Table 1: Network structure details of the spatial encoder E_{sp}, the ConvLSTMs, the frame decoder D, and the attention in the motion context-aware video prediction P. Suppose the input video frame has size of $\mathbb{R}^{W\times H\times C}$.

Long-term Motion Context Encoder / Motion Matching Encoder / Motion Context Embedding		
Layer	**Filter / Stride**	**Output Size** $(time \times width \times height \times channel)$
3D-Conv 1	3×3×3 / (1, 1, 1)	$T\times W\times H\times64$
Max Pooling 1	1×2×2 / (1, 2, 2)	$T\times(W/2)\times(H/2)\times64$
3D-Conv 2	3×3×3 / (1, 1, 1)	$T\times(W/2)\times(H/2)\times128$
Max Pooling 2	2×2×2 / (2, 2, 2)	$(T/2)\times(W/2)\times(H/2)\times128$
3D-Conv 3	3×3×3 / (1, 1, 1)	$(T/2)\times(W/4)\times(H/4)\times256$
3D-Conv 4	3×3×3 / (1, 1, 1)	$(T/2)\times(W/4)\times(H/4)\times256$
Max Pooling 3	2×2×2 / (2, 2, 2)	$(T/4)\times(W/8)\times(H/8)\times256$
3D-Conv 5	3×3×3 / (1, 1, 1)	$(T/4)\times(W/8)\times(H/8)\times512$
3D-Conv 6	3×3×3 / (1, 1, 1)	$(T/4)\times(W/8)\times(H/8)\times512$
Max Pooling 4	2×2×2 / (2, 2, 2)	$(T/8)\times(W/16)\times(H/16)\times512$
Avg Pooling	Global×1×1 / (Global, 1, 1)	$1\times(W/16)\times(H/16)\times512$
2D-DeConv 1	3×3 / (2, 2)	$(W/8)\times(H/8)\times256$
2D-DeConv 2	3×3 / (2, 2)	$(W/4)\times(H/4)\times128$

Table 2: Network structure details of the long-term motion context encoder E_{LMC}, the motion matching encoder E_{MM}, and the motion context embedding for the memory. E_{LMC} and E_{MM} have the same network structure, typical C3D structure. Suppose the input video sequence has size of $\mathbb{R}^{T\times W\times H\times C}$.

3. Additional Qualitative Results

3.1. Results on Moving-MNIST

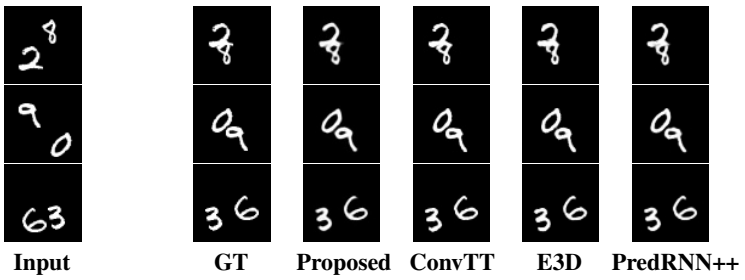

Figure 2: Animation qualitative results of video frame prediction up to 99-th frame with given 10 frames on the Moving-MNIST dataset. Each digit moves independently and sometimes overlaps with each other. Other results are obtained from official sources. Adobe reader is required to play the animation.

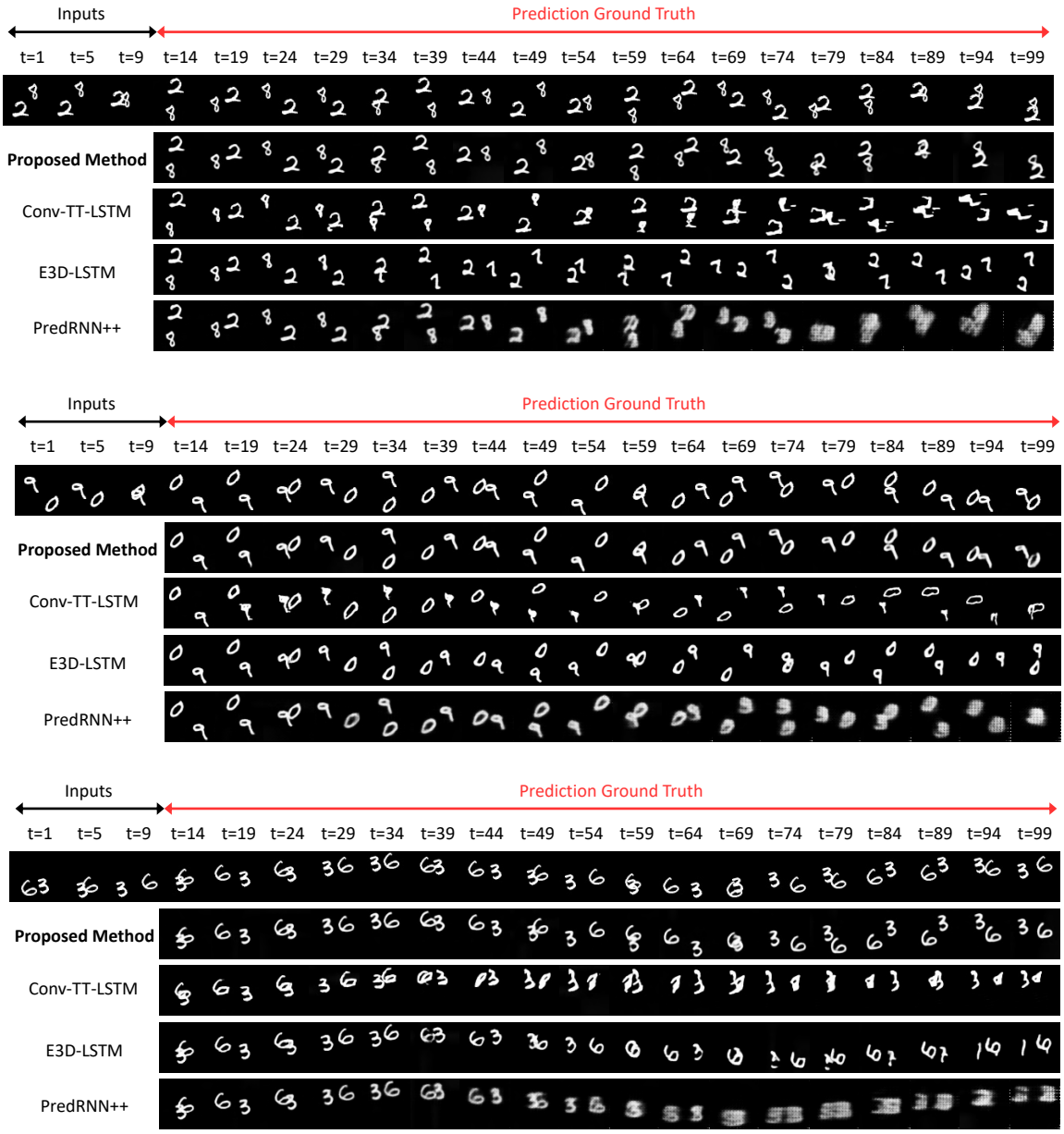

Figure 3: Static qualitative results of frame prediction up to 99-th frame with given 10 frames on the Moving-MNIST dataset.

3.2. Results on KTH Action

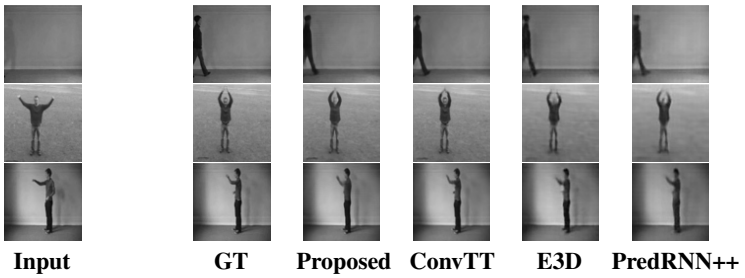

Figure 4: Animation qualitative results of video frame prediction up to 99-th frame with given 10 frames on the KTH Action dataset. The first one shows action for walking, the second shows handwaving including camera zoom in/out, and the last one shows periodic boxing action. Other results are obtained from official sources. Adobe reader is required to play the animation.

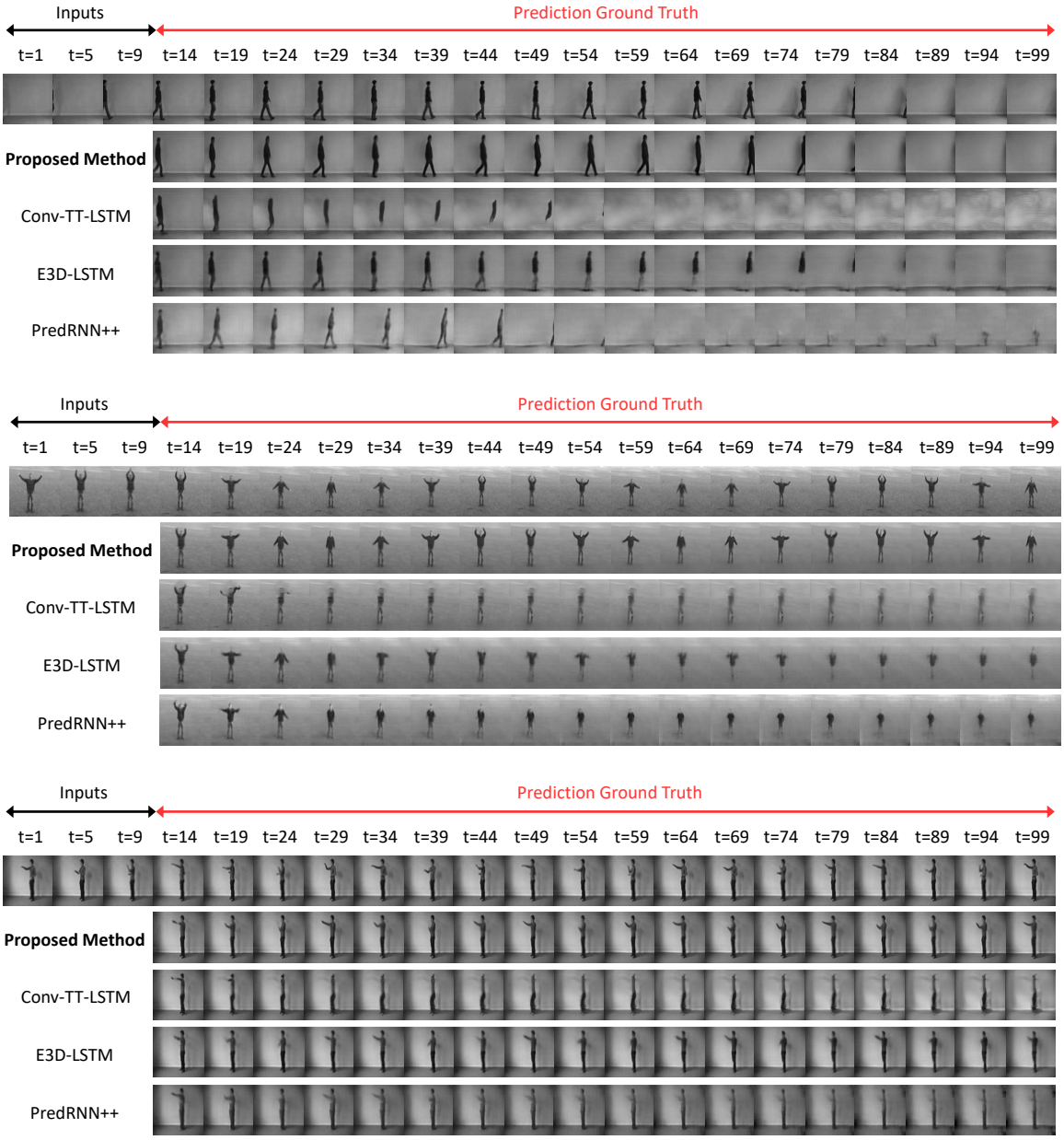

Figure 5: Static qualitative results of frame prediction up to 99-th frame with given 10 frames on the KTH action dataset.

3.3. Results on Human 3.6M

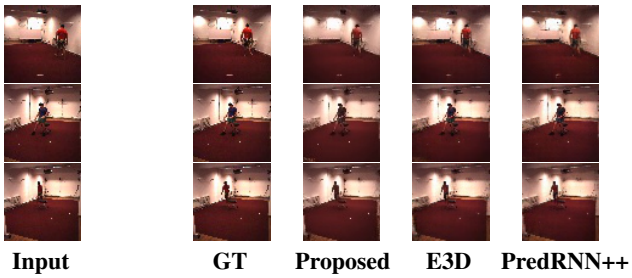

Figure 6: Animation qualitative results of video frame prediction up to 99-th frame with given 10 frames on the Human 3.6M dataset. The first one shows turning at the corner, the second shows action of picking up on the floor, and the last one shows stopping walking with phoning. Other results are obtained from official sources. Adobe reader is required to play the animation.

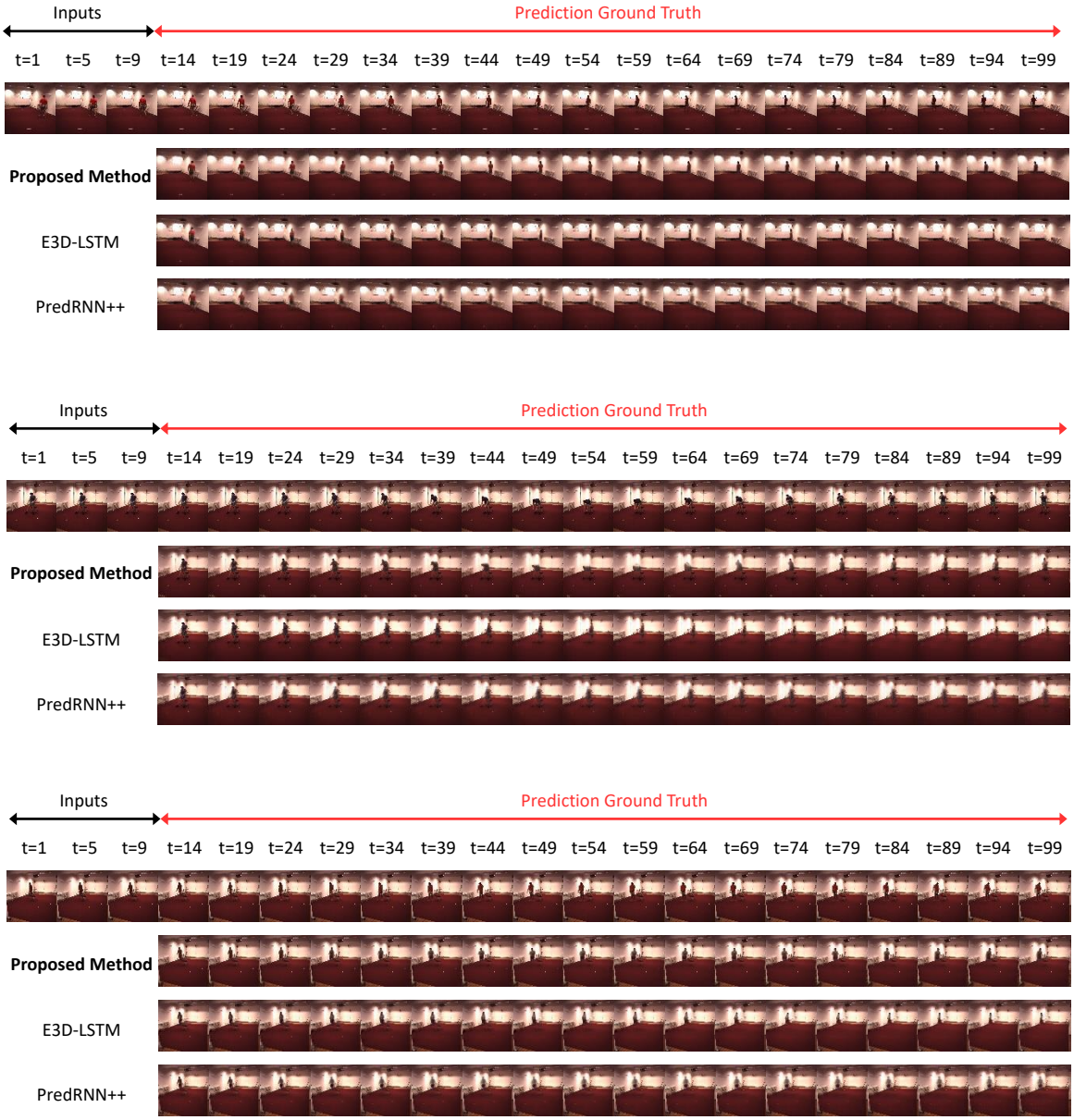

Figure 7: Static qualitative results of frame prediction up to 99-th frame with given 10 frames on the Human 3.6M dataset.