## Learning to Decompose and Disentangle Representations for Video Prediction (Supplementary)

Jun-Ting Hsieh Stanford University junting@stanford.edu Bingbin Liu Stanford University bingbin@stanford.edu **De-An Huang** Stanford University dahuang@cs.stanford.edu

Li Fei-Fei Stanford University feifeili@cs.stanford.edu Juan Carlos Niebles Stanford University jniebles@cs.stanford.edu

## **A** Implementation Details

For our image encoder and decoder, we use the DCGAN architecture [2] as the image encoder and decoder in our model. The number of layers are set based on the input or output image size, 5 layers for  $64 \times 64$  images and 6 layers for  $128 \times 128$ . All recurrent neural networks are LSTMs with hidden size 64. The dimension of the content vector  $z_C$  is 128, and the dimension of the pose vectors  $z_{t,P}$  is 3, containing the parameters of a spatial transformer. We train our model for 200k iterations with the Adam optimizer [1] with initial learning rate 0.001, which is decayed to 0.0001 halfway through training. For all experiments, we optimize both the reconstruction and prediction losses during the first half of training, and optimize only the prediction loss in the second half, though we found that training with both losses throughout the entire training process produces similar results.

We assume our random latent variables,  $z_{0,P}^i$ ,  $\beta_t^i$ , and  $z_C^i$ , to be Gaussian. Thus, our model outputs the mean and standard deviation for these variables. The prior distributions are  $p(\beta_t^i) \sim \mathcal{N}(0, 0.1)$ , and  $p(z_C^i) \sim \mathcal{N}(0, 1)$ . The prior for initial pose is  $p(z_{0,P}^i) \sim \mathcal{N}([2,0,0], [0.2,1,1])$  for Moving MNIST, and  $p(z_{0,P}^i) \sim \mathcal{N}([4,0,0], [0.2,1,1])$  for Bouncing Balls.

## **B** Qualitative Results

In this section, we show more qualitative results on Moving MNIST (Figure 1) and Bouncing Balls (Figure 2). Figure 2 shows more examples where our model predicts the collision.

Below we present some failure cases for Bouncing Balls, where the balls fail to be separated. If the balls are too close together for all input frames, our model may produce blurry results. For collisions, since the trajectories after collision are highly sensitive to the collision surface, our model may identify the collision but produce incorrect trajectories.

## References

- [1] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. ICLR, 2015.
- [2] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

Input	<b>6</b> 5	6 5	<b>6</b> 5	<b>6</b> 5	<b>6</b> 5	65	ક	ۆ	Jo	ولد
Ground truth	هري	56	5	5 6	5 6	56	56	56	56	56
DRNet	Se	56	50	56	5	5	56	56	56	á 6
Ours w/o disentanglement	$\tilde{\mathbf{x}}$	10 °	و. ())	40 <b>6</b>	8	8	84	86	86	86 <b>G</b>
Ours (DDPAE)	50	56	5 6	5 6	5 6	5 6	56	56	56	56
Ours - 1st component	6	6	6	6	6	6	6	6	6	6
Ours - 2nd component	5	5	5	5	5	5	5	5	5	5
Input	58	58	£8	ଜ୍ୟ	وكل	ھک	ર્ક	85	<b>8</b> 5	85
Ground truth	85	85	85	85	\$5	Ş	Ser la	<u>8</u>	5	500
DRNet	85	85	85	85	5	6	5	500	30	ŝ
Ours w/o disentanglement	35	8 <sub>6</sub>	3 <sub>5</sub>	B	35	Þ	5	ş	Sec.	ŝ
Ours (DDPAE)	85	85	85	85	8	8	5	500	Labo	Libo
Ours - 1st component	Б	Б	Б	Б	Б	Б	Б	Б	Б	Б
Ours - 2nd component	8	8	8	8	8	8	8	8	8	8
Input	Ì	B	3	3	충	భ	\$	Ŕ	Ŕ	చ్
Ground truth	ίŊ	intro.	52	ſX	5	52	57	57	F	F
DRNet	(J)	S) (	ഹ്	ŝ	ŝ	5	5	52	్ర	്റ്റ
Ours w/o disentanglement	S	d)	5	5	62	52	52	62	57	5
Ours (DDPAE)	J,	dx dx	5	52	52	52	52	52	52	Ś
Ours - 1st component	5	5	5	5	5	5	5	5	5	5
Ours - 2nd component	z	z	z	z	z	z	z	z	z	z

Figure 1: Qualitative results on Moving MNIST.

Input	••••	•	•	••••	••••	•••	•••	•••	::	::
Ground truth	::	• •	• •	• •	• •	•••	•••	•••	•••	•••
Ours w/o dependencies	: :	: :	• .•	٠.	۰.		• .•	• .'	• •	• • •
Ours (DDPAE)	: :	: :	: .	÷ .•	• .•	• •	• •	• •	• • •	•••
Ours - 1st component	•	•		•					•	
Ours- 2nd component										
Ours - 3rd component	•					•				•
Ours - 4th component	•	•	•		٠	٠	٠	٠	•	
Input	••••	••••	•••	•••	•••	•	•	•	•••	•••
Ground truth	:.	•••	:	•••	•••	•	•	•	•	•
Ours w/o dependencies	:	•	•	:	:	8	•	•	•	•
Ours (DDPAE)	•••	•••	•••	•••	•••	•••	•••	•••	•••	•••
Ours -										
1st component Ours-				•	•	•		•		
2nd component										
Ours -							•			
3rd component										
Ours -	•	•		•	•	•	•	•	•	•
4th component										
Input	•••	••	•••	•••	•••	•••	•••	•••	• •	• :
Ground truth	• •	. :	. :	••		. :	. :	• :	•	••
Ours w/o dependencies					. :	. 1	. 1		. •	. •
Ours (DDPAE)					. :	:	. :			:
Ours -	٠									
1st component				•		•			•	
Ours-					_					
2nd component		•		•				•		
Ours - 3rd component	-			-					_1	
Ours -							•	•		
4th component	•	٠	٠	۰	•	٠	٠	•	•	٠

Figure 2: Qualitative results on Bouncing Balls.

Input	•	•	•		•	•	•	•	•	•
Ground truth	•	•	•	••••	•	•	••••	••••	••••	•••
Ours (DDPAE)						4	-	4		
Ours - 1st component Ours- 2nd component	•			•						
Ours - 3rd component Ours - 4th component	•			•	•	•				•
Input	•••	•••	•••	• •	• •	•	•	•	• ••	• .:
Input Ground truth	•••	••••	••••	••••	••••	••••	••••	•	• .:	• .:
-	•	••••	••••	•••••	••••	• ;•	•	•	•	••
Ground truth Ours (DDPAE) Ours -	•	••••	••••	••••	•••••	•;	•;	• ;	•	•
Ground truth Ours (DDPAE)	•			••••	••••			• ;	•	• .: •• ••
Ground truth Ours (DDPAE) Ours - 1st component Ours-	•								• •• •	• .: •• •.• •

Figure 3: Bouncing Balls failure cases.