# Supplementary Material: Improved Conditional VRNNs for Video Prediction

# **A. Hierarchical VRNN**

## A.1. ELBO Derivation

We start from the ELBO for VRNNs [3]:

$$\log p(\mathbf{x}|\mathbf{c}) \ge \sum_{t=1}^{T} \mathbb{E}_{q(z_t|\mathbf{z}_{<\mathbf{t}},\mathbf{x}_{\le\mathbf{t}},\mathbf{c})} \log p(x_t|\mathbf{z}_{\le\mathbf{t}},\mathbf{x}_{<\mathbf{t}},\mathbf{c}) - D_{KL}(q(z_t|\mathbf{z}_{<\mathbf{t}},\mathbf{x}_{\le\mathbf{t}},\mathbf{c}))||p(z_t|\mathbf{z}_{<\mathbf{t}},\mathbf{x}_{<\mathbf{t}},\mathbf{c}))$$
(6)

Recall we defined  $\mathbf{z_t} = (z_t^1, ..., z_t^L)$  and factorized the prior as:

$$p(\mathbf{z}_{\mathbf{t}}|\mathbf{z}_{<\mathbf{t}}, \mathbf{x}_{<\mathbf{t}}, \mathbf{c}) = \prod_{l=1}^{L} p(z_{t}^{l}|\mathbf{z}_{\mathbf{t}}^{<\mathbf{l}}, \mathbf{z}_{<\mathbf{t}}^{l}, \mathbf{x}_{<\mathbf{t}}, \mathbf{c}).$$
(7)

And the posterior:

$$q(\mathbf{z}_{\mathbf{t}}|\mathbf{z}_{<\mathbf{t}}, \mathbf{x}_{\le \mathbf{t}}, \mathbf{c}) = \prod_{l=1}^{L} q(z_{t}^{l}|\mathbf{z}_{\mathbf{t}}^{<\mathbf{l}}, \mathbf{z}_{\le \mathbf{t}}^{l}, \mathbf{x}_{\le \mathbf{t}}, \mathbf{c}).$$
(8)

We then substitute these terms in the VRNN ELBO, first looking at the reconstruction term inside the summation over time:

$$\sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z}_{t}|\mathbf{z}_{< t}, \mathbf{x}_{\le t}, \mathbf{c})} \log p(x_{t}|\mathbf{z}_{\le t}, \mathbf{x}_{< t}, \mathbf{c}) = \sum_{t=1}^{T} \mathbb{E}_{q(z_{t}^{1}, \dots, z_{t}^{L}|\mathbf{z}_{< t}, \mathbf{x}_{\le t}, \mathbf{c})} \log p(x_{t}|z_{t}^{1}, \dots, z_{t}^{L}, \mathbf{z}_{< t}, \mathbf{x}_{< t}, \mathbf{c})$$

$$= \sum_{t=1}^{T} \mathbb{E}_{q(z_{t}^{1}|\mathbf{z}_{< t}, \mathbf{x}_{\le t}, \mathbf{c}) \dots q(z_{t}^{L}|\mathbf{z}_{t}^{< L}, \mathbf{z}_{< t}, \mathbf{x}_{\le t}, \mathbf{c})} \log p(x_{t}|z_{t}^{1}, \dots, z_{t}^{L}, \mathbf{z}_{< t}, \mathbf{x}_{< t}, \mathbf{c})$$

$$(9)$$

And then looking at the summation of KL divergences:

$$-\sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z}_{t}|\mathbf{z}_{

$$= -\sum_{t=1}^{T} \mathbb{E}_{q(z_{t}^{1}|\mathbf{z}_{

$$= -\sum_{t=1}^{T} \mathbb{E}_{q(z_{t}^{1}|\mathbf{z}_{(10)$$$$$$

(by definition of conditional KL divergence)

$$= -\sum_{t=1}^{T}\sum_{l=1}^{L}D_{KL}((q(z_t|\mathbf{z}_{<\mathbf{t}}, \mathbf{x}_{\leq \mathbf{t}}, \mathbf{c})||p(z_t|\mathbf{z}_{<\mathbf{t}}, \mathbf{x}_{<\mathbf{t}}, \mathbf{c}))$$

Adding both terms together we obtain the ELBO defined in eq. 5.

#### A.2. Posterior Dense Connectivity

Fig A.1 illustrates the dense connection of the approximate posterior. For each latent variable has a deterministic connection to  $x_{t-1}$  (red arrows in Fig 2), in addition to all the latent variables from the layers below (green arrow in in Fig 2). Finally, each latent variable has a direct connection to the output variables  $x_t$ , corresponding to the inference path.



Figure A.1: Schematic view of the approximate posterior with the dense-connectivity pattern. Arrows in red show the connections from the input at the previous timestep to current latent variables. Arrows in green highlight skip connections between latent variables to outputs. Arrows in **black** indicate recurrent temporal connections. We empirically observe that this dense-connectivity pattern eases the training of latent hierarchy.

LAYERS	DETAILS
Conv2D	input $\rightarrow 64$
<b>ResNet Block</b>	64  ightarrow 64
MaxPool	2x2, s = 2
<b>ResNet Block</b>	64  ightarrow 128
<b>ResNet Block</b>	128  ightarrow 128
MaxPool	2x2, s = 2
<b>ResNet Block</b>	128  ightarrow 256
<b>ResNet Block</b>	256  ightarrow 256
MaxPool	2x2, s = 2
<b>ResNet Block</b>	$256 \rightarrow 512$
<b>ResNet Block</b>	512  ightarrow 512
MaxPool	2x2, s = 2
<b>ResNet Block</b>	$512 \rightarrow 512, ks = 4$
<b>ResNet Block</b>	512  ightarrow 512

Figure B.1: Frame Encoder architecture.

# **B. Model Specification**

We specify the architecture used for the 64x64 model. Convolutional layers in our model use 3x3 kernels with stride s = 1 and padding p = 1 unless otherwise specified. We use modified Resnet blocks made up of two groups of ReLU + Conv2D + GroupNorm. GroupNorm layers use g = 16 groups. Transposed Convolutions use 4x4 kernels with stride s = 2 and padding p = 1, which upscales 2x the input tensor. ConvLSTM layers use 3x3 kernels with stride s = 1 and padding p = 1 and GroupNorm.

## **C. Additional Samples**

See the figures below.

#### **D. PredNet**

We additionally compare to PredNet on Cityscapes using the official implementation. Note that PredNet is deterministic and can't model future uncertainty. The model is only able to correctly predict a few timesteps before becoming blurry, as the uncertainty increases with time. PredNet obtained a FVD score of 1079.19 and a LPIPS score of 0.397, while ours with the hierarchical model are 567.51 and 0.264 respectively (lower is better).

LAYERS	DETAILS
ConvLSTM	$512 \rightarrow 512, ks = 4$
UpConv	$512 \rightarrow 512$ , scale = 4
ConvLSTM	$512 \rightarrow 512$
UpConv	$512 \rightarrow 512$
ConvLSTM	512  ightarrow 512
UpConv	512  ightarrow 256
ConvLSTM	256  ightarrow 256
UpConv	256  ightarrow 128
ConvLSTM	128  ightarrow 128
UpConv	128  ightarrow 64
ConvLSTM	64  ightarrow 64
Conv2D + GroupNorm + ReLU	64  ightarrow 64
Conv2D	64  ightarrow input
Figure B.2: LIkelihood/Deco	der architecture.

re
]

Level	LAYERS	DETAILS								
1x1	Conv2D + GroupNorm ConvLSTM Conv2D + GroupNorm	$\begin{array}{c} 128 \rightarrow 128,  ks = 1 \\ 128 \rightarrow 128 \\ 128 \rightarrow 128 \text{x2},  ks = 1 \end{array}$								
8x8	Conv2D + GroupNorm ConvLSTM Conv2D + GroupNorm	$512 \rightarrow 512, ks = 1$ $512 \rightarrow 1512$ $512 \rightarrow 512x2, ks = 1$								
32x32	Conv2D + GroupNorm ConvLSTM Conv2D + GroupNorm	$512 \rightarrow 512, ks = 1$ $512 \rightarrow 1512$ $512 \rightarrow 512x2, ks = 1$								
	Figure B.3: Prior/Posterior architecture.									

LEVEL	LAYERS	DETAILS
1x1	Conv2D + GroupNorm + ReLU Conv2D + GroupNorm	$512 \rightarrow 512, ks = 1$ $512 \rightarrow 512x2, ks = 1$
4x4	Conv2D + GroupNorm + ReLU Conv2D + GroupNorm	$512 \rightarrow 512, ks = 1$ $512 \rightarrow 512x2, ks = 1$
8x8	Conv2D + GroupNorm + ReLU Conv2D + GroupNorm	$512 \rightarrow 512, ks = 1$ $512 \rightarrow 512x2, ks = 1$
16x16	Conv2D + GroupNorm + ReLU Conv2D + GroupNorm	$\begin{array}{c} 256 \rightarrow 256,  ks = 1 \\ 256 \rightarrow 256 \text{x2},  ks = 1 \end{array}$
32x32	Conv2D + GroupNorm + ReLU Conv2D + GroupNorm	$128 \rightarrow 128, ks = 1$ $128 \rightarrow 128x2, ks = 1$
64x64	Conv2D + GroupNorm + ReLU Conv2D + GroupNorm	$64 \rightarrow 64, ks = 1$ $64 \rightarrow 64x2, ks = 1$

Figure B.4: Initial State network architecture.

Context	I				Predicte	d Frames				
t = 1 $t = 2$	t = 3	t = 4	t = 6	t = 8	t = 10	t = 12	t = 15	t = 18	t = 20	t = 25
5										
<b>SVG</b> [5]										
Ours w/o Hier										
Ours w/ Hier										
GT										
SVG [5]										150
Ours w/o Hier										
OURS W/ HIER										
GT										
SVG [5]										
Ours w/o Hier										
OURS W/ HIER										
5										
SVG [5]								1		
Ours w/o Hier										
OURS W/ HIER	E Sand		- Cand	- Stand		a Stand			and the second	

Context	1				Predicted	l Frames				
t = 1 $t = 2$	t = 3	t = 4	t = 6	t = 8	t = 10	t = 12	t = 15	t = 18	t = 20	t = 25
E										
SVG [5]										
OURS W/O HIER	C. C	C C								
Ours w/ Hier			1							-
E State										
SVG [5]										
OURS W/O HIER			2milder	Contraction of the second						
Ours w/ Hier			and a							
E						and a				
SVG [5]										
OURS W/O HIER										-
OURS W/ HIER										
E e										
SVG [5]										
Ours w/o Hier					 					
OURS W/ HIER										MAG



Figure C.1: Random Test Samples for Cityscapes.

Context	I				Predicted	Frames				
t=1 $t=2$	t = 3	t = 4	t = 6	t = 8	t = 10	t = 12	t = 15	t = 18	t = 20	t = 25
5 of of	oteo	R		or he	oth	or Ca	or C	or t	or t	or the
SAVP [22]	00000			of the		and in	and in	o in	al.	our en
SVG [5]	otho			orte	ote	oth	ore	of the	d. i	of it
OURS W/O HIER	or in p	A.	AC	de			de.	of the	and the	and the second
Ours w/ Hier	otio		A.C.	of R		or the	or the	o C	of the	or the
E C			-	X	-		- A		X	
SAVP [22]				- CA		-0-	- ele			N.
SVG [5]						- to I		- Col	- Cont	and a start
Ours w/o Hier			-			*	X			
Ours w/ Hier		-		- Co	-	- A	-		-	- the
E				F. 197		100				
SAVP [22]								100	100	
SVG [5]	- 1		1 and	100	100	A series	A Series	1.000	1000	A sea
Ours w/o Hier										
OURS W/ HIER					C and	100	Acres .	1000	1	100

Context	I				Predicted	l Frames				
t=1 $t=2$	t = 3	t = 4	t = 6	t = 8	t = 10	t = 12	t = 15	t = 18	t = 20	t = 25
E										
SAVP [22]				T					The second secon	
SVG [5]		MA	RA		10.10	10.10	The Real		He He	
Ours w/o Hier			Ac		Pecs.	Terry	Per-	1se	-se	-96
Ours w/ Hier										
5	N	1					-			
SAVP [22]	N.	100								
SVG [5]	1	-	A A	-	ar the	ar the	10 m	10 m	-	
Ours w/o Hier			-	-	-	-	*			-
Ours w/ Hier		1						the second		
E F										
SAVP [22]									- Con	
SVG [5]										
Ours w/o Hier	1									
OURS W/ HIER					1.78					

Context				Predicted	l Frames				
t=1 $t=2$	t = 3	t = 4 $t =$	6  t = 8	t = 10	t = 12	t = 15	t = 18	t = 20	t = 25
5	5 - D	1.0?			8 F.	8.32	8.32	872	N FR
SAVP [22]	5 : L B			s : L	5 . F.	87	874	87	a sh
SVG [5]	5 - 1 B			and a	87 L	8.24	8:2	8:2	8 . a
Ours w/o Hier	5 - 5 B	the st		A. A.	10 M	87	81	8 Pr	81
Ours w/ Hier	5 - 5 D					8.4		8.44	84
E									
SAVP [22]									33
SVG [5]				23					
Ours w/o Hier									
OURS W/ HIER									
5			8 3 8	1		3010	20	30 PS	Sales.
SAVP [22]			838	3	3		p-2		12
SVG [5]	1		818	18	3	312	3.18	318	3
OURS W/O HIER				ALC ALC	17 18 18 19 19 19 19 19 19 19 19 19 19 19 19 19		30 2	30 - C	Bold .
OURS W/ HIER	and the second		S H-S	and a	and the second s	30 6	and a	30 10	3-1-2



Figure C.2: Random Test Samples for pushbair.

Context	I			Pre	dicted Fra	mes			
t = 1 $t = 2$	t = 3	t = 4	t = 6	t = 8	t = 10	t = 12	t = 15	t = 18	t = 20
ह ह	34	34	34	34	NJT	MJT	3	Ŷ	397
SVG [5]	સ્પ	3 <sub>4</sub>	з <sub>ц</sub>	34	зч	ઝપ	34	34	କ୍
Ours w/o Hier	સ્વ	34	34	3 <sub>4</sub>	ngt	nar	3	S.	÷
OURS W/ HIER	34	34	34	34	ЗЦ	PH T	R	5	B
62 62	24	24	26	24	Ð	42	J.	24	24
SVG [5]	26	26	26	26	Å	é	g2	«2	«2
Ours w/o Hier	24	26	26	2%	24	24	N	ap	¢,
Ours w/ Hier	26	26	26	24	26	af.	12	62	62
ज ज	のす	ਅਧ	᠕ᠴ	BJ	ઉ્વ	रू	ゴい	ન ૩	43
SVG [5]	るり	MJ	MJ	35	34	34	Щ.	uļ_	cy_
Ours w/o Hier	ろす	ゔゴ	প্য	শ্রে	R.	き	3	1	tr)
Ours w/ Hier	やす	लग	MJ	N-T	F	સ	3	લ	Rev J
,2 ,2	2	24	24	2	3	7	¥	24	24
SVG [5]	2	2	2	24	24	24	24	24	24
Ours w/o Hier	2	Ş	24	24	24	24	<b>2</b> <sub>4</sub>	2	$\mathbf{Z}_{\boldsymbol{\varphi}}$
	2	2	2	2	2	2	2.	2.4	24

Context	1			Pre	dicted Fra	mes				
$t = 1 \qquad t = 2$	t = 3	t = 4	t = 6	t = 8	t = 10	t = 12	t = 15	t = 18	t = 20	_
<sup>5</sup> 3, 3,	3	3	37	3	37	3	23	3	3	
SVG [5]	37	<u>з</u> 7	ა 7	ح 7	<sup>ර</sup> 7	<sup>ර</sup> 7	<sup>3</sup>	<sup>8</sup> 7	8 2	
Ours w/o Hier	3	3	37	37	-tu	z	\$	<b>(?</b> )	3	
Ours w/ Hier	37	25	3 <b>7</b>	75	37	37	37	3 <b>7</b>	3 <b>7</b>	
57 <sub>9</sub> 7 <sub>9</sub>	79	79	Ŧ	9	Ŋ	97	<sup>9</sup> 7	<sup>9</sup> 7	9 7	
SVG [5]	79	カ	Ŋ	P	P	Þ	74	<sup>7</sup> 7	77	
Ours w/o Hier	79	Τŋ	¥1	9	Ħ	97	<sup>9</sup> 7	<sup>9</sup> 7	9 7	
Ours w/ Hier	79	Τŋ	Ħ	9	Ħ	Ŋ	<sup>9</sup> 7	<sup>9</sup> 7	9 7	
<sup>15</sup> 7 <sup>7</sup> 7 <sup>7</sup>	7	7	7	ア	ア	7	ラ	P	7	
SVG [5]	7	7	7	3	R	7	ろ	フ	フ	
Ours w/o Hier	7	27	2	77	P	ア	7	TP	77	
Ours w/ Hier	7	7	27	77	77	Ŋ	3	77	77	
<sup>2</sup> 3131	3	<b>A1</b>	3	31	31	31	3	21	12-	
SVG [5]	Ð	Ð	47	طأله	لغي	38	36	39	39	
Ours w/o Hier	3	<b>A11</b>	t,	3	31	31	31	X	*	
Ours w/ Hier	3	3	3	31	31	31	31	3	-H-	

Context		Predicted Frames							
t = 1 $t = 2$	t = 3	t = 4	t = 6	t = 8	t = 10	t = 12	t = 15	t = 18	t = 20
69 <b>69</b>	9.9	9 <sub>9</sub>	96	96	96	96	96	96	96
SVG [5]	9.9	200	970	er O	96	<i>q</i> 0	96	00	aO.
Ours w/o Hier	وم	9.9	9.9	96	96	96	99	9	9
OURS W/ HIER	9.9	ويَ	9-9-0	Z	9	ф	99	99	59
5 <b>9</b> 9	94	94	94	97	97	9 4	49	49	¶¶.
SVG [5]	94	94	44	44	4	Ę	¥	ų	4
Ours w/o Hier	94	94	94	94	9	49	49	49	4
Ours w/ Hier	<b>94</b> Figure (	94 3: Rando	94 m Test Sa	H	<b>A</b>		<b>H</b>	भ	49
Figure U.S. Kanuom fest Samples for Stochastic Moving MINIS I.									



Figure D.1: A sample from PredNet on Cityscapes.