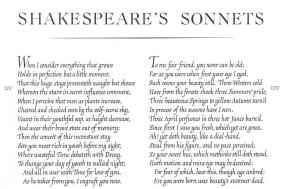
Improved Generative Models of Video

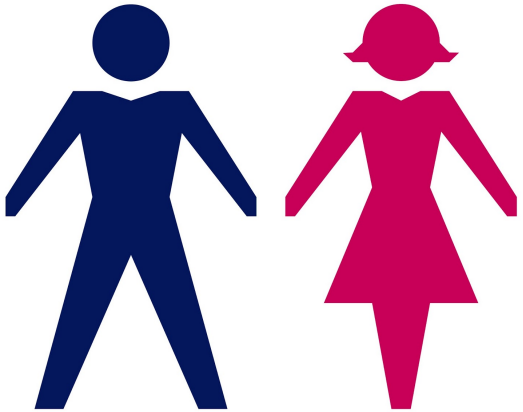
Lluis E. Castrejon Subira

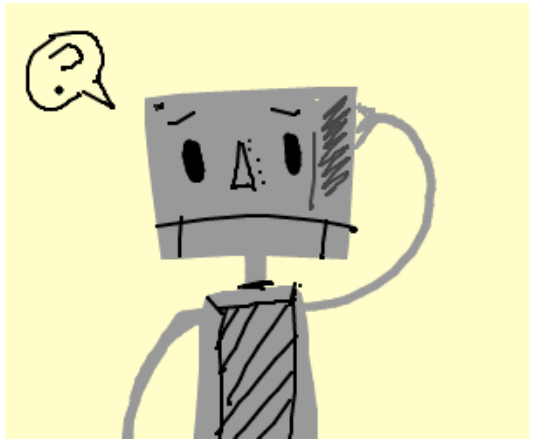
Supervisor: Prof. Aaron Courville


SHAKESPEARE'S SONNETS

When I consider everything that grows
Holds in perfection but a little moment.
That this huge stage presenteth naught but shows
Whereon the stars in secret influence comment.
When I perceive that men as plants increase,
Cheered and checked even by the self-same sky,
Vaunt in their youthful sap, at height decrease,
And wear their brave state out of memory.
Then the conceit of this inconstant stay
Sets you most rich in youth before my sight,
Where wasteful Time debateth with Decay,
To change your day of youth to sullied night;
And all in war with Time for love of you,
As he takes from you, I engraft you new.

To me fair friend, you never can be old,
For as you were when first your eye I eyed,
Such seems your beauty still. Three Winters cold
Have from the forests shook three Summers' pride,
Three beauteous Springs to yellow Autumn turned
In process of the seasons have I seen,
Three April perfumes in three hot Junes burned,
Since first I saw you fresh, which yet are green.
Ah yet doth beauty, like a dial-hand,
Steal from his figure, and no pace perceived;
So your sweet hue, which methinks still doth stand,
Hath motion and mine eye may be deceived:
For fear of which, hear this, thou age unbred:
Ere you were born was beauty's summer dead.





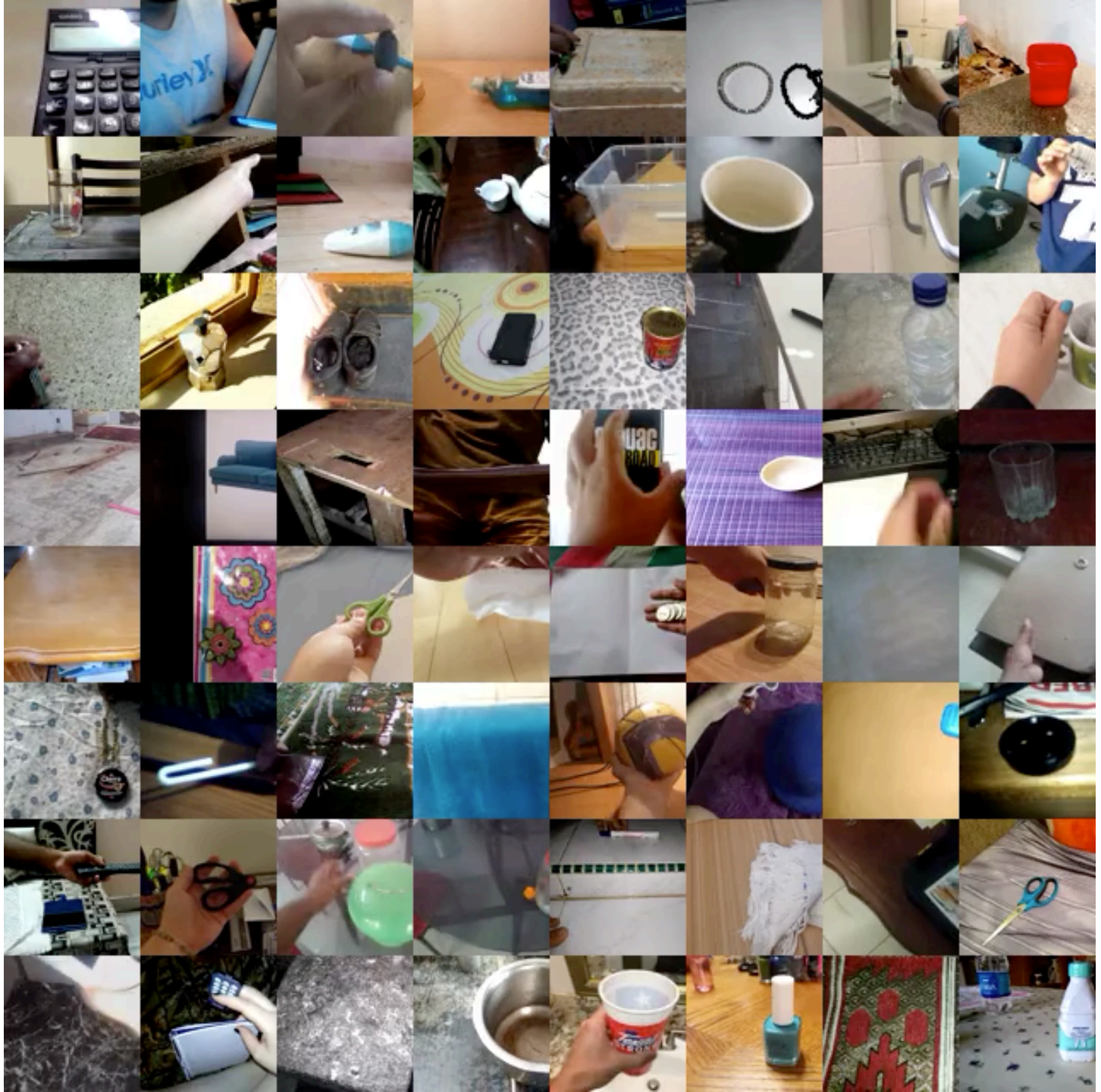


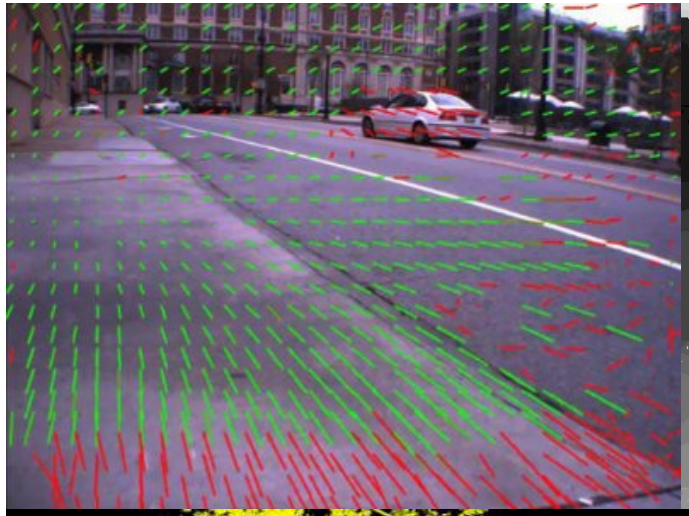
Figure from Something-Something dataset

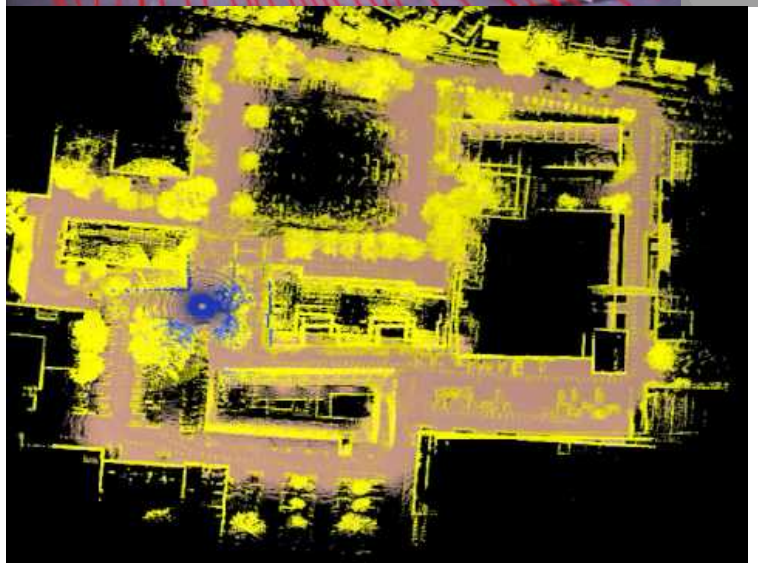
Outline

1. Motivation
2. Background
3. Related Work
4. Improved Conditional VRNNs for Video Prediction
5. GANs for Video Prediction (ongoing work)
6. Future Work and Conclusions


P
F AA 176

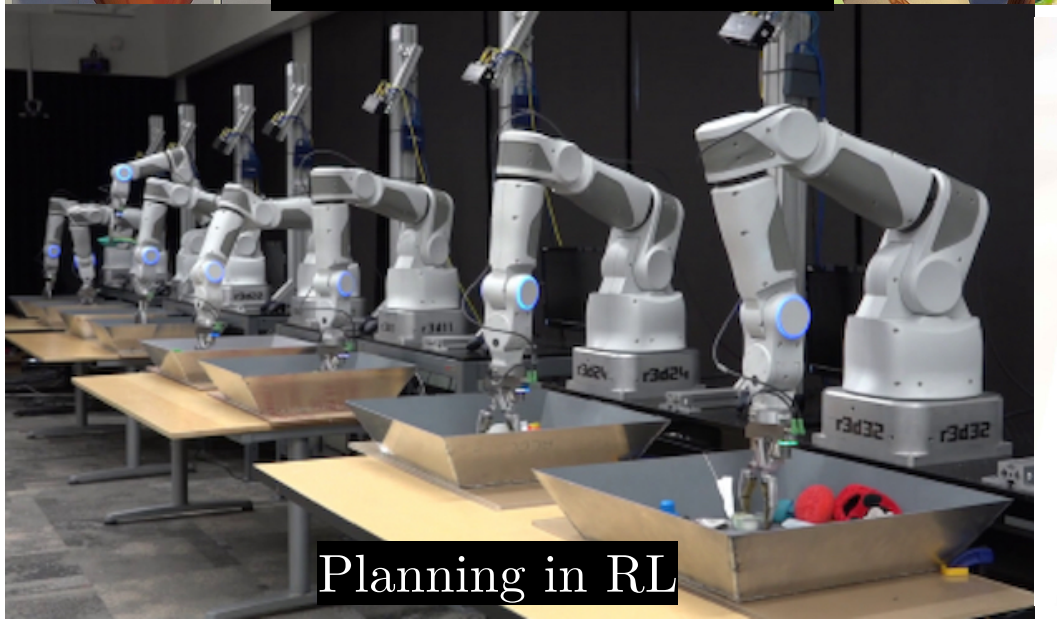

P
F AA 176






Controllable Generation


Video Compression

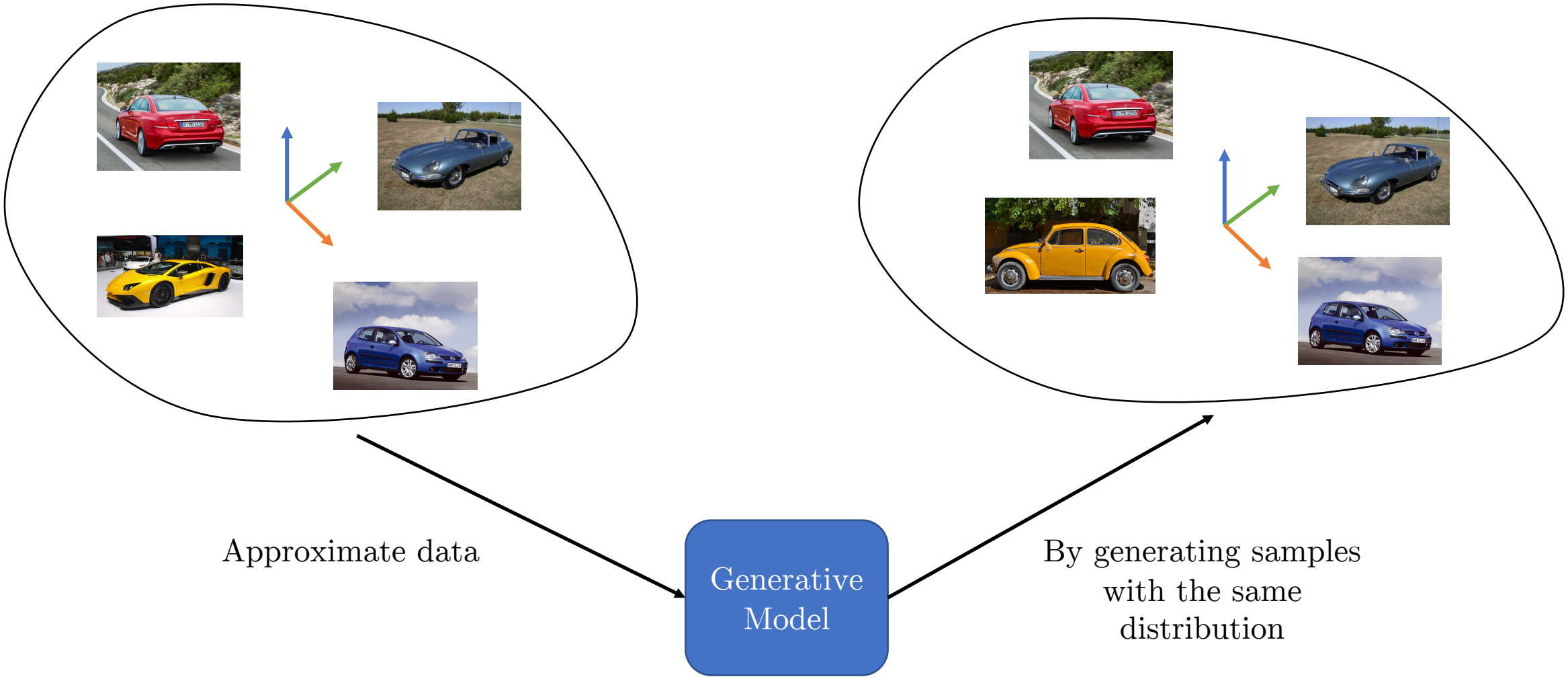

Planning in RL


Representation Learning

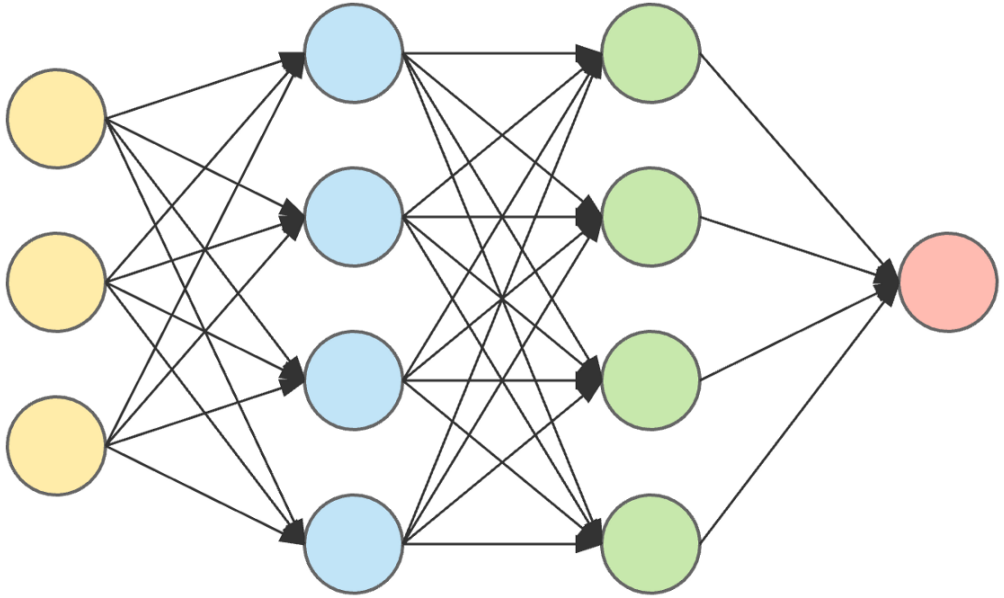
Outline

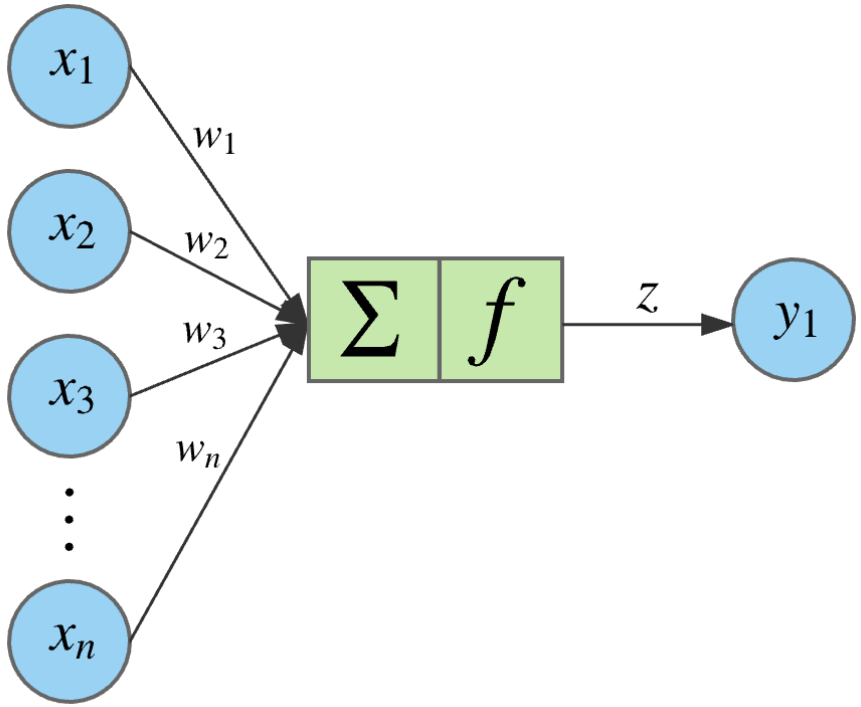
1. Motivation
2. **Background**
3. Related Work
4. Improved Conditional VRNNs for Video Prediction
5. GANs for Video Prediction (ongoing work)
6. Future Work and Conclusions

Generative Models



Neural Networks





$$y = f_n(f_{n-1}(\ldots(f_1(x))))$$

$$z = f(b + x \cdot w) = f\left(b + \sum_{i=1}^{n} x_i w_i\right)$$

Figures from Arden Dertat

CNNs (Fukushima 1980, LeCun 1989)

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n)$$

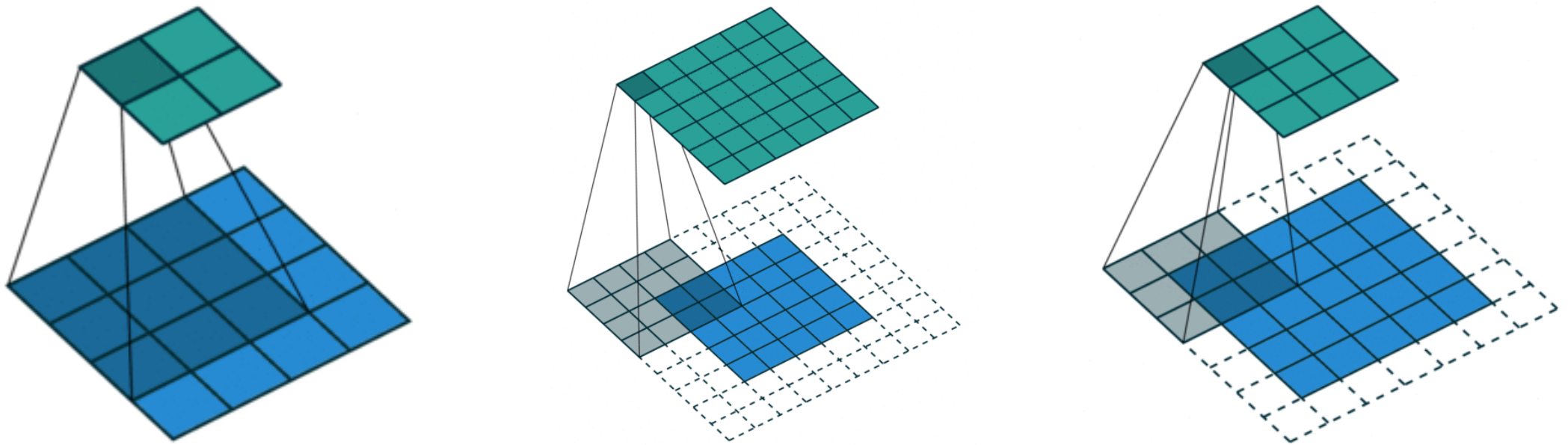


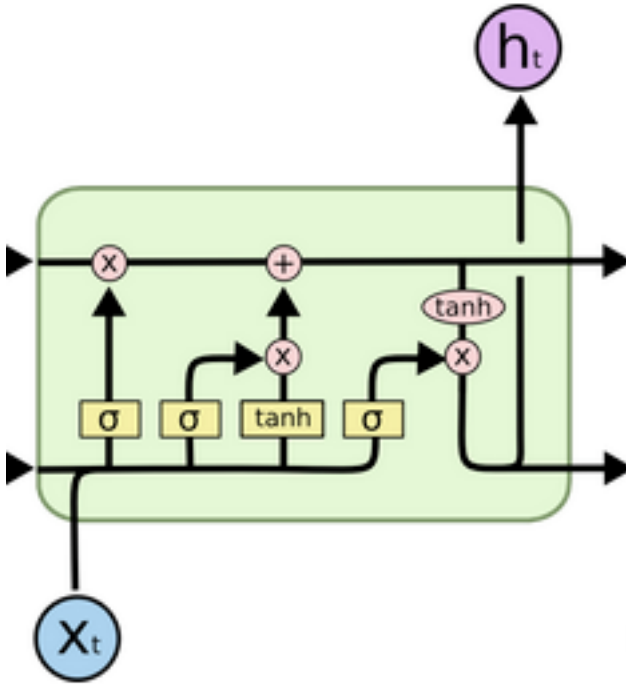
Figure from Dumoulin and Visin, 2016

RNNs/LSTMs (Hochreichter 1997)

Basic RNN

$$\boldsymbol{h}^{(t)} = f(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}; \boldsymbol{\theta})$$

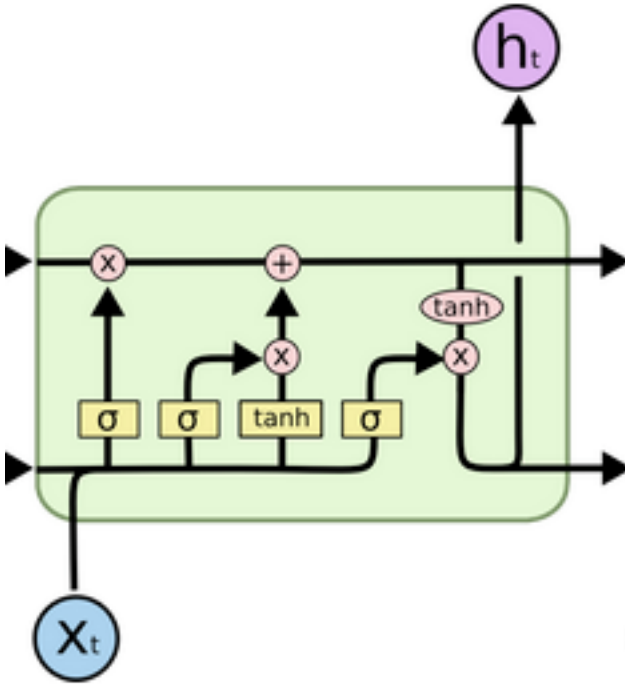
LSTM



LSTM Figure from Colah, 2015

ConvLSTMs (Shi et al. 2015)

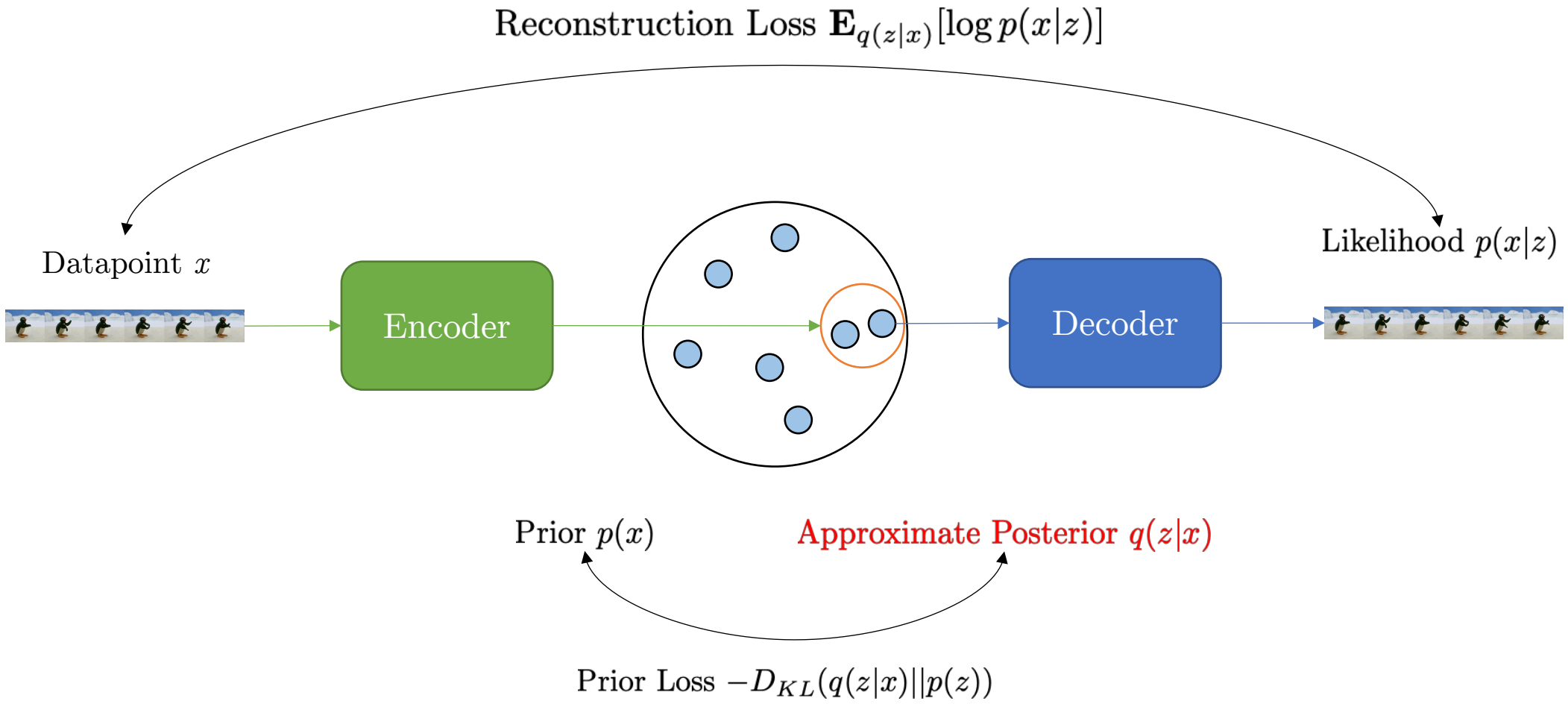
Replace fully-connected affine transformations with convolutions



$$
\begin{aligned}
i_t &= \sigma(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \circ \mathcal{C}_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \circ \mathcal{C}_{t-1} + b_f) \\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \circ \mathcal{C}_t + b_o) \\
\mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t)
\end{aligned}
$$

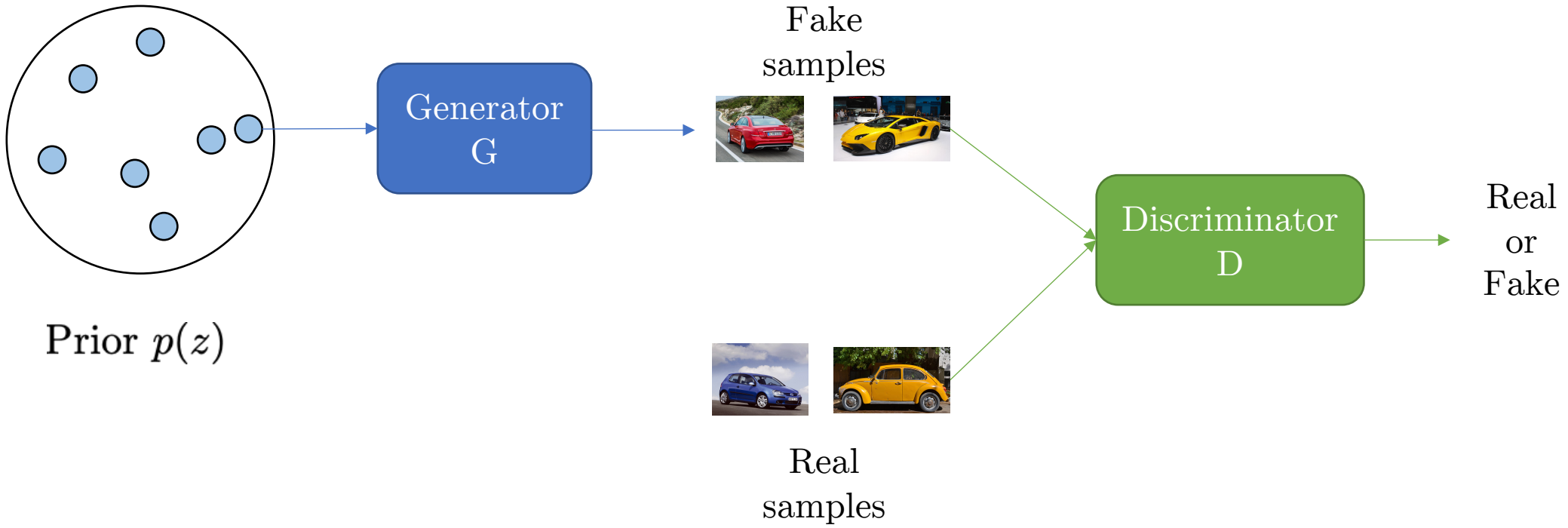
Figure from Colah, 2015

Variational AutoEncoders (VAEs)



First proposed in Kingma et al. 2013, Rezende et al. 2013

Generative Adversarial Networks (GANs)



First proposed in Goodfellow et al. 2014

Outline

1. Motivation
2. Background
3. **Related Work**
4. Improved Conditional VRNNs for Video Prediction
5. GANs for Video Prediction (ongoing work)
6. Future Work and Conclusions

Related Work – First papers

First papers on video generation adapted techniques from the NLP community:

- **Ranzato et al.** Video (language) Modeling: A Baseline for Generative Models of Natural Videos. *ICLR 2015*
- **Srivastava et al.** Unsupervised Learning of Video Representations using LSTMs. *ICML 2015*
- **Oh et al.** Action-Conditional Video Prediction using Deep Networks in Atari Games. *NIPS 2015*

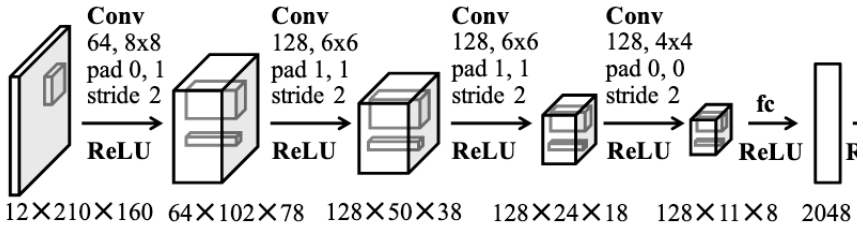


Figure from Oh et al., 2015

Related Work – Disentangled Representations

Disentangle *static* and *dynamic* elements of a scene

- **Denton and Birodkar** Unsupervised Learning of Disentangled Representations from Video. *NIPS 2017*
- **Villegas et al.** Decomposing Motion and Content for Nautral Video Sequence Prediction. *ICLR 2017*
- **Tulyakov et al.** MoCoGAN: Decomposing Motion and Content for Video Generation. *CVPR 2018*
- **Vondrick et al.** Generate Videos with Scene Dynamics. *NIPS 2016*

Related Work – Disentangled Representations

Disentangle *static* and *dynamic* elements of a scene

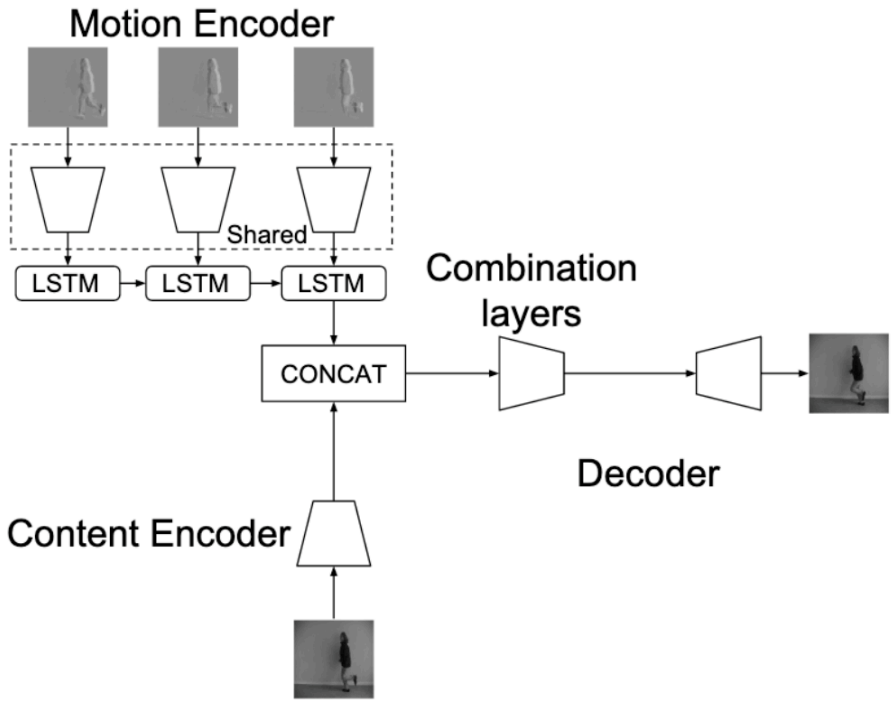


Figure from Villegas et al., 2017

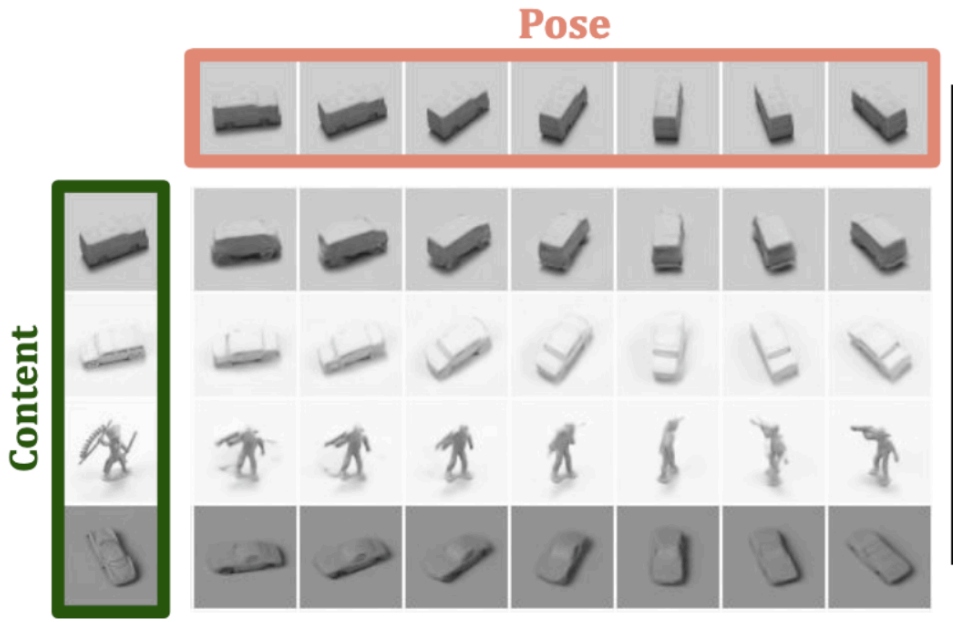


Figure from Denton and Birodkar, 2017

Related Work – Disentangled Representations

Model dynamics on (instance) segmentation space

- **Luc et al.** Predicting Deeper into the Future of Semantic Segmentation *ICCV 2017*
- **Luc et al.** Predicting Future Instance Segmentation by Forecasting Convolutional Features *ECCV 2018*

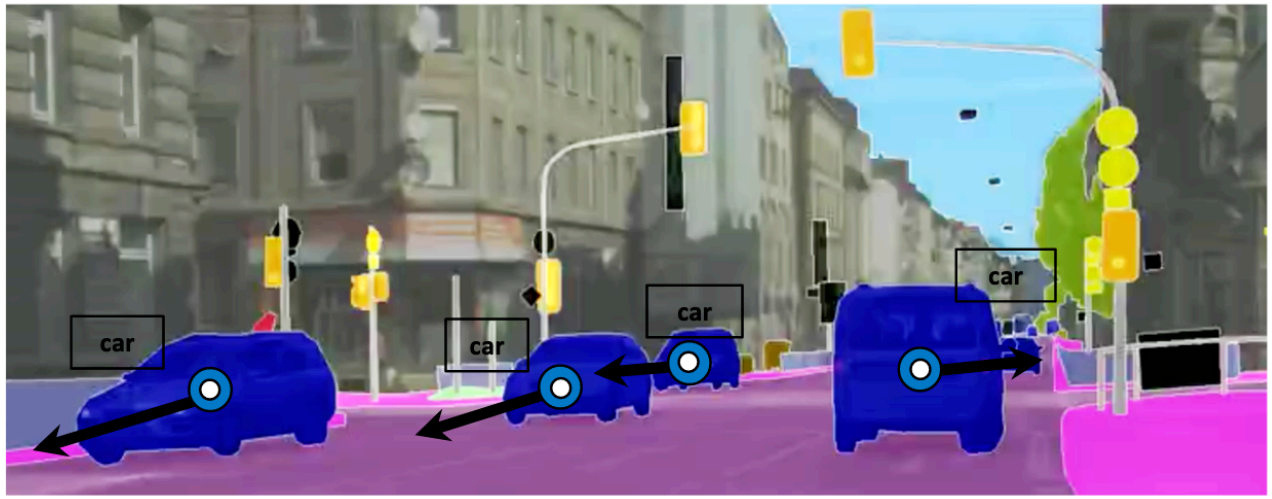


Figure from Luc et al., 2017

Related Work – Stochastic Models

Models capable of producing multiple outputs

- Disentangled GANs
- **Babaizadeh et al.** Stochastic Variational Video Prediction *ICLR 2018*
- **Lee et al.** Stochastic Adversarial Video Prediction *Arxiv 2018*
- **Denton et al.** Stochastic Video Generation with a Learned Prior. *ICML 2018*

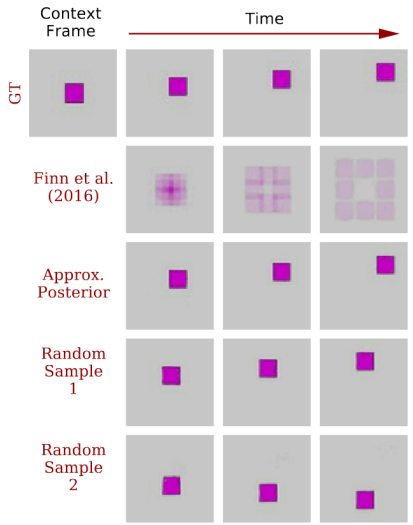


Figure from Babaizadeh et al., 2018

SVG-LP (Denton and Fergus, ICML 2018)

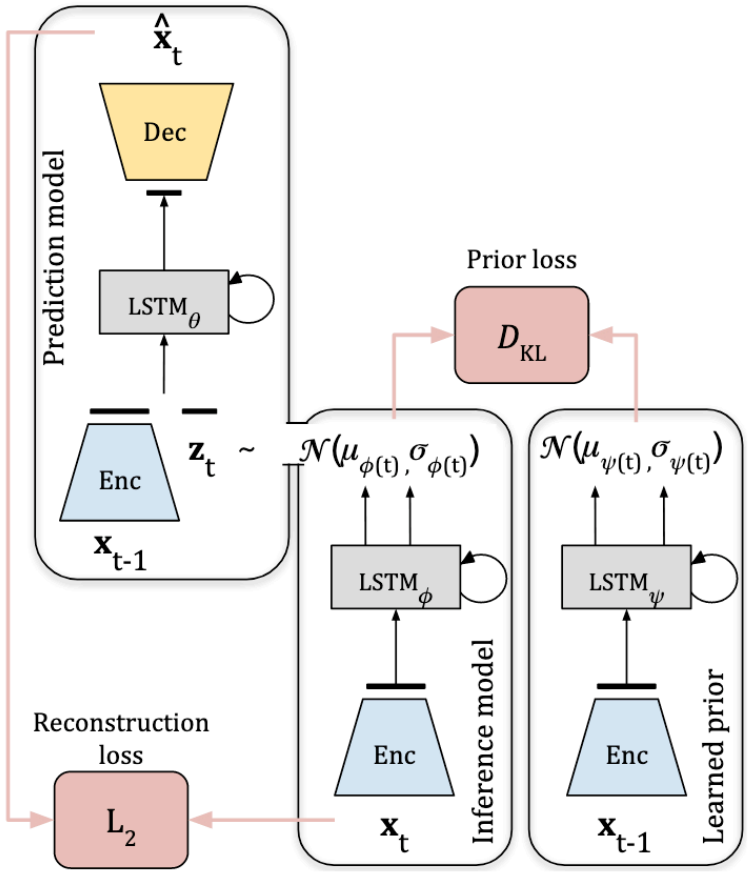
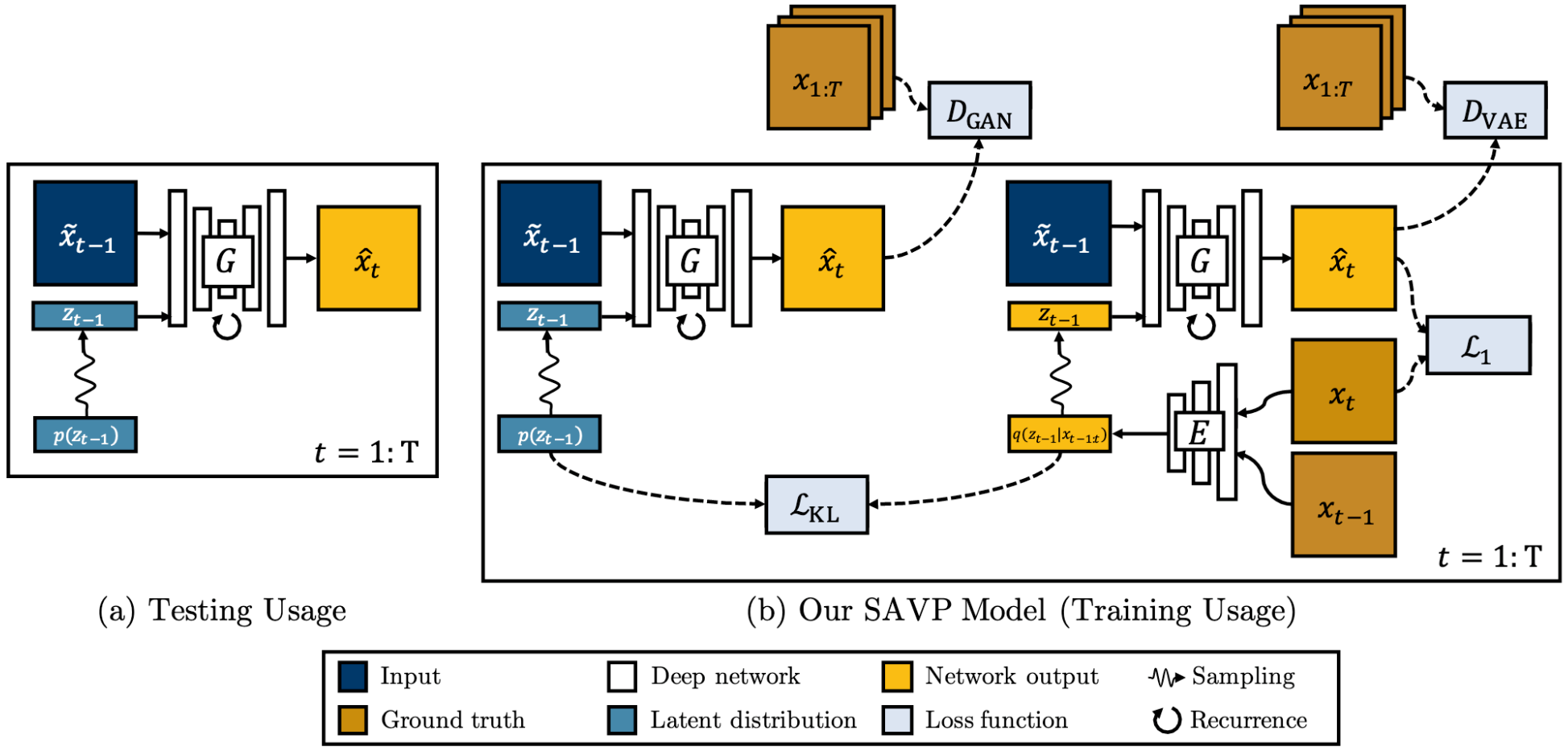


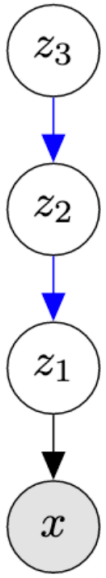
Figure from Denton and Fergus, 2018

SAVP (Lee et al., Arxiv 2018)



Related Work – Hierarchical VAEs

VAEs with a hierarchy of latents



- **Gulrajani et al.** PixelVAE: A Latent Variable Model for Natural Images. *ICLR 2017*
- **Ranganath et al.** Hierarchical Variational Models. *ICML 2016*
- **Maaløe et al.** Auxiliary Deep Generative Models. *ICML 2016*
- **Sønderby et al.** Ladder Variational Autoencoders. *NIPS 2016*
- **Kingma et al.** Improved Variational Inference with Inverse Autoregressive Flows. *NIPS 2016*
- **Maaløe et al.** BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling. *Arxiv, 2019*

Figure from Maaløoe et al., 2019

Outline

1. Motivation
2. Background
3. Related Work
4. **Improved Conditional VRNNs for Video Prediction**
5. GANs for Video Prediction (ongoing work)
6. Future Work and Conclusions

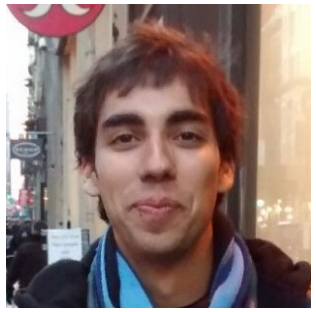
Improved Conditional VRNNs for Video Prediction

Lluis Castrejon
Mila, Université de Montréal

Nicolas Ballas
Facebook AI Research

Aaron Courville
CIFAR Fellow, Mila,
Université de Montréal

International Conference in Computer Vision ICCV 2019
Seoul, South Korea



Summary

We propose to improve upon current VRNNs for video prediction:

- More flexible hierarchical prior and approximate posterior
- Better decoder/likelihood model

State-of-the-art results on Stochastic Moving MNIST, Cityscapes and BAIR Push

VRNNs

Extension of VAEs to sequences with a latent variable per timestep

Random Variables

$c \sim$ context
$z \sim$ latents
$x \sim$ future frames

$$p(\mathbf{x}, \mathbf{z}|\mathbf{c}) = \prod_{t=1}^{T} p(x_t|\mathbf{z}_{\leq \mathbf{t}}, \mathbf{x}_{<\mathbf{t}}, \mathbf{c}) p(z_t|\mathbf{z}_{<\mathbf{t}}, \mathbf{x}_{<\mathbf{t}}, \mathbf{c})$$

Likelihood *Prior*

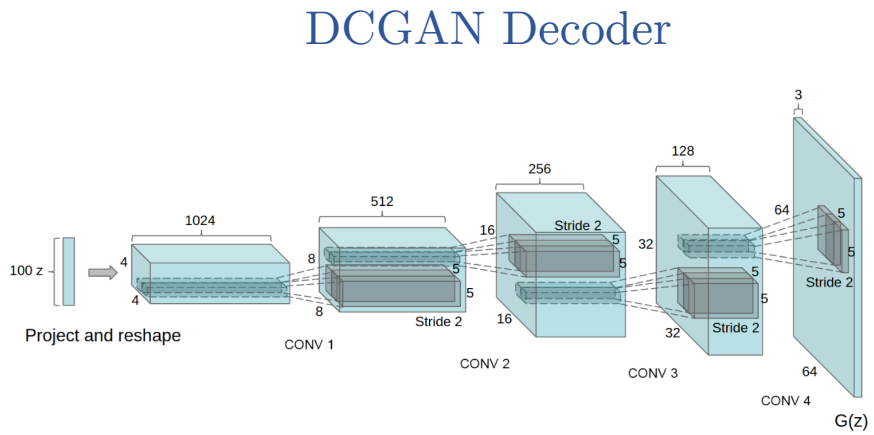
ELBO for training

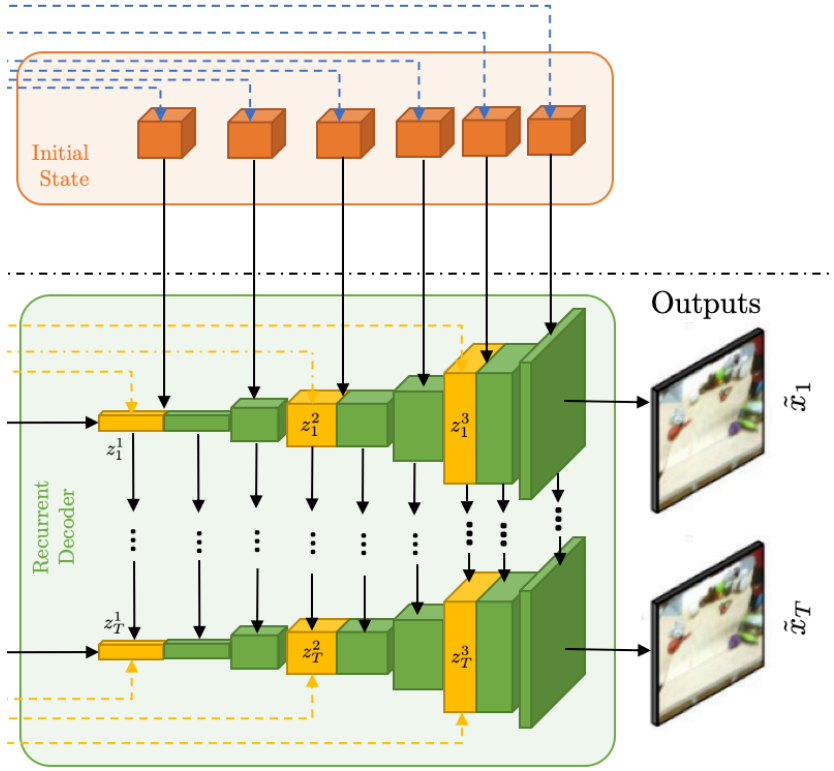
$$\log p(\mathbf{x}|\mathbf{c}) \geq \sum_{t=1}^{T} \mathbb{E}_{q(z_t|\mathbf{z}_{<\mathbf{t}}, \mathbf{x}_{\leq \mathbf{t}}, \mathbf{c})} \log p(x_t|\mathbf{z}_{\leq \mathbf{t}}, \mathbf{x}_{<\mathbf{t}}, \mathbf{c})$$
$$- D_{KL}(q(z_t|\mathbf{z}_{<\mathbf{t}}, \mathbf{x}_{\leq \mathbf{t}}, \mathbf{c})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

Approximate Posterior

Higher Capacity Likelihood

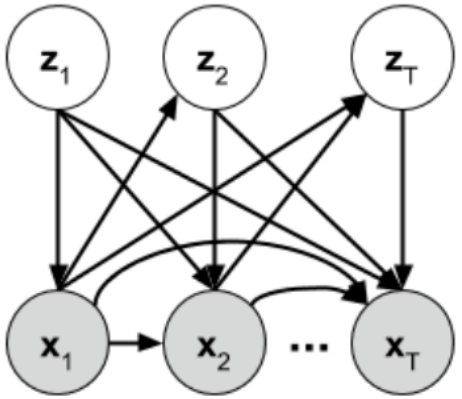
Increase the capacity of the decoder



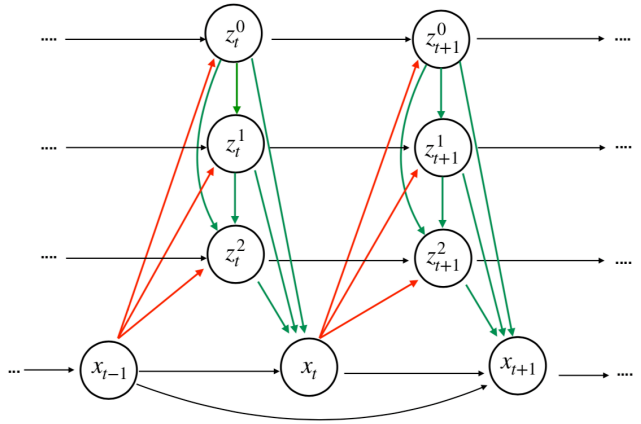


Hierarchical Latent Distribution

Single Latent Level



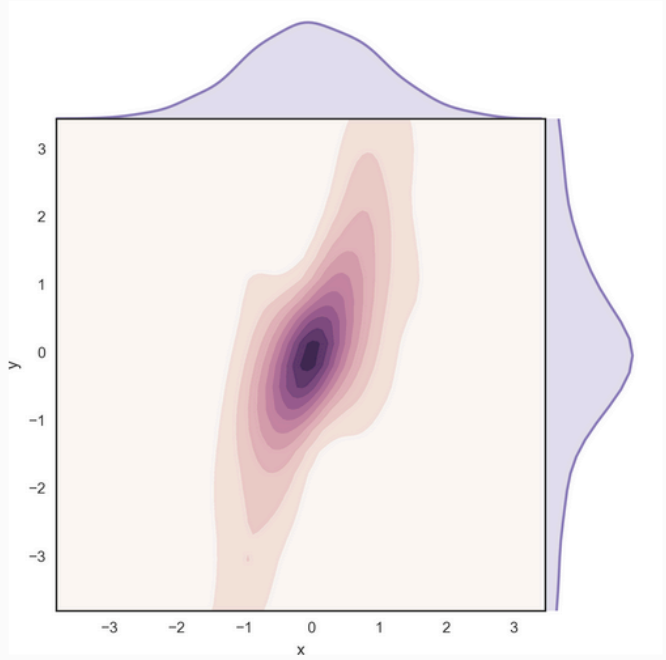
Multiple Latent Levels



Each latent distribution is a diagonal Gaussian conditioned on the previous latents in the hierarchy

Hierarchical Latent Distribution

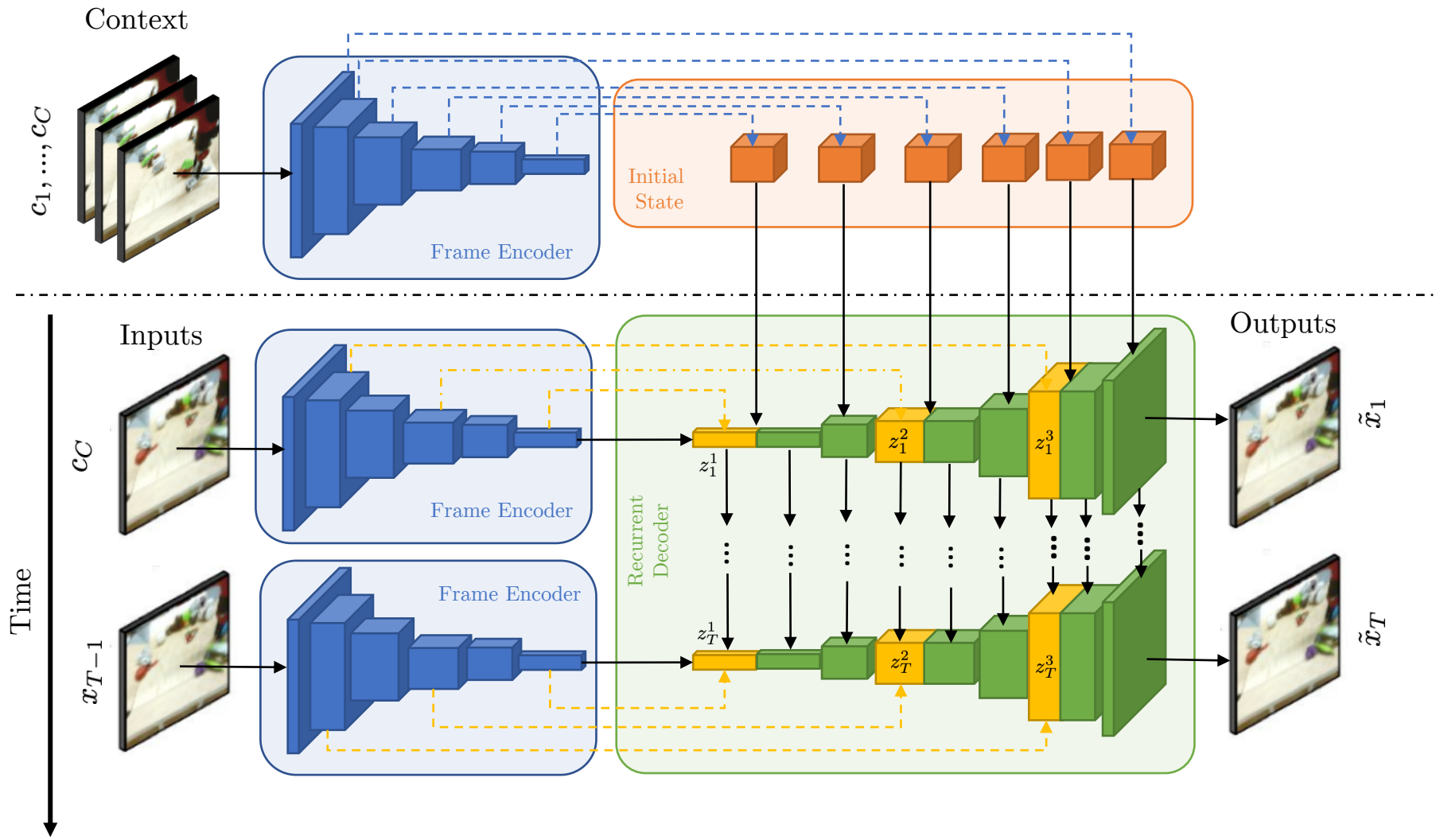
$$q(v) = \mathcal{N}(v \mid 0, I)$$
$$q(u \mid v) = \mathcal{N}(u \mid Fv, Fvv^T F^T + \beta I),$$



The joint distribution of latents whose maginal probability is a Gaussian conditioned on other Gaussian variables does not necessarily follow a Gaussian distribution and can be quite flexible

Image from this URL

Model


Context
$c_1, \ldots, c_C$
Frame Encoder
Initial State
Inputs
c_C
Frame Encoder
Recurrent Decoder
Outputs
z_1^1
z_1^2
z_1^3
$\tilde{x}_1$
Time
x_{T-1}
Frame Encoder
z_T^1
z_T^2
z_T^3
$\tilde{x}_T$

Ablation: Likelihood $p(x_t|\mathbf{z}_{\leq \mathbf{t}}, \mathbf{x}_{<\mathbf{t}}, \mathbf{c})$

We increase the capacity by adding additional ConvLSTM layers at different resolutions

More capacity ↓

Model	Parameters	Train/Test ELBO(↓)
1-ConvLSTM	62.22M	3237/3826
3-ConvLSTM	86.64M	1948/2355
6-ConvLSTM	93.81M	1279.21/1731.31

Performance improves as we increase the capacity of the decoder

Ablation: Hierarchy of Latents

Add multiple levels of latents, each one conditional on the previous one

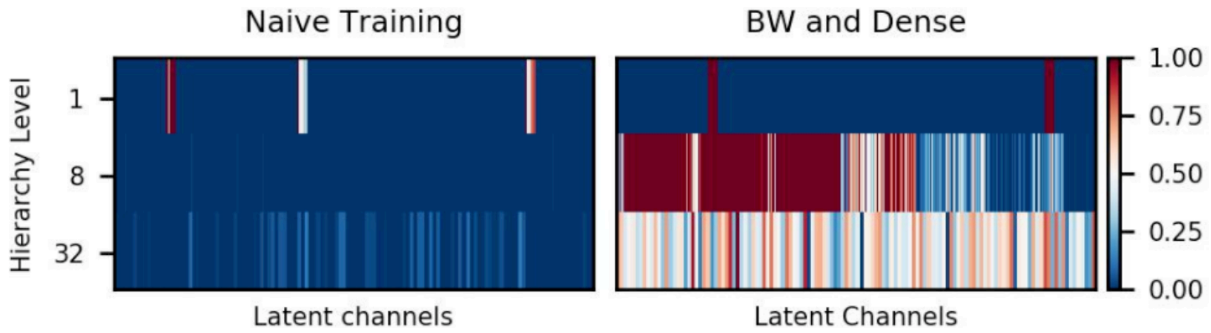
More Levels ↓

Model	Parameters	Train/Test ELBO (↓)
1	166.55M	1141.85/1536.93
1-8	220.60M	989.39/1313.02
1-8-32	230.74M	**883.10/1162.24**
1-8-16-32	245.19M	956.63/1256.22

Only with Beta Warmup (BW) and Dense Connections the hierarchy brings an improvement

Ablation: Hierarchy of Latents

We qualitatively evaluate the effect of BW and Dense Connectivity on the latent variables



Without these techniques most of the latents are not used

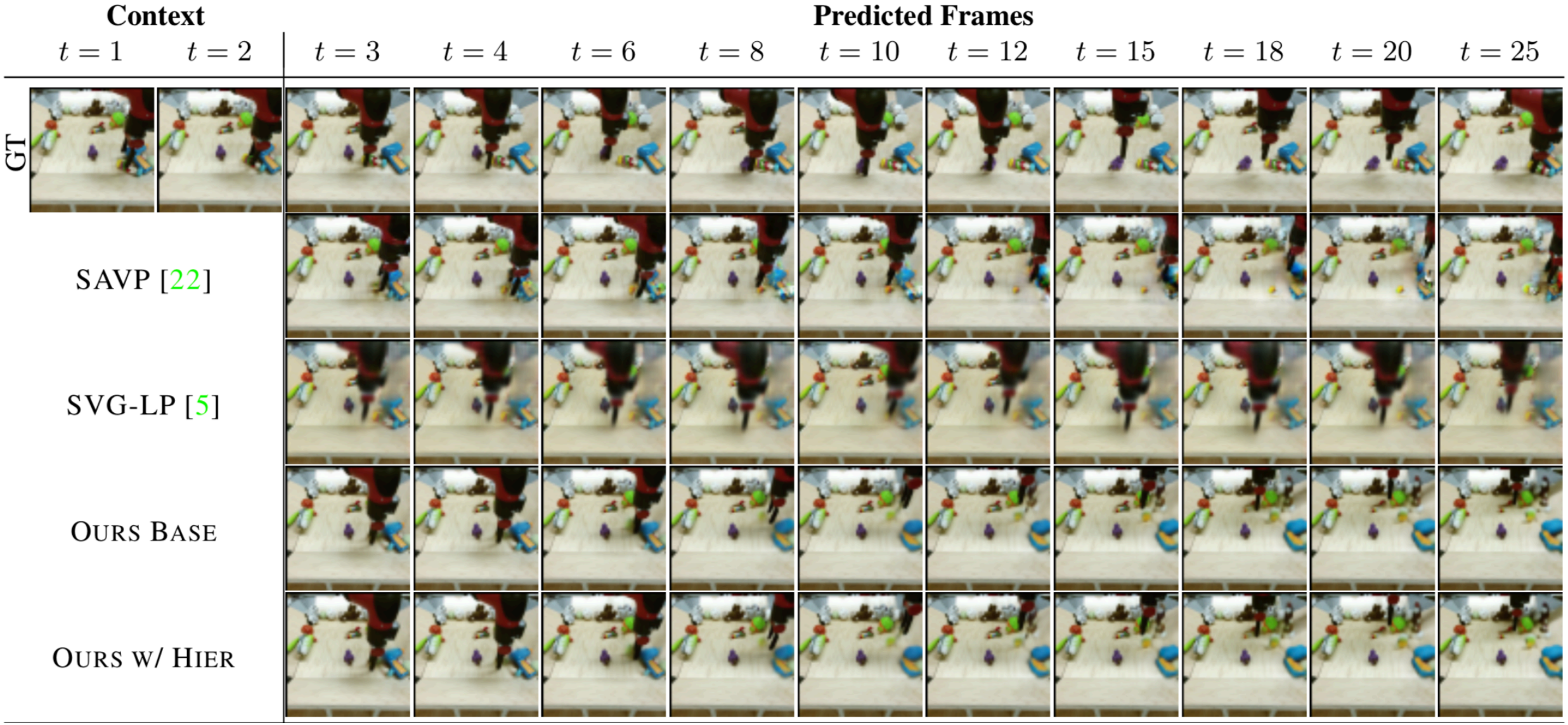
Evaluation Metrics

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

LPIPS $$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$

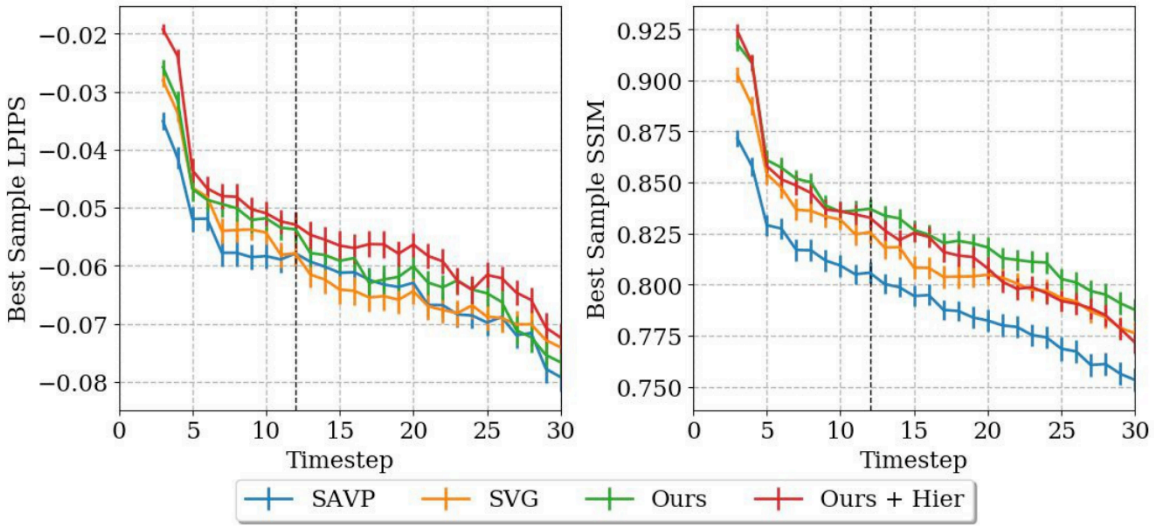
FVD $$|\mu_R - \mu_G|^2 + \text{Tr}\left(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{\frac{1}{2}}\right)$$

BAIR Push Dataset

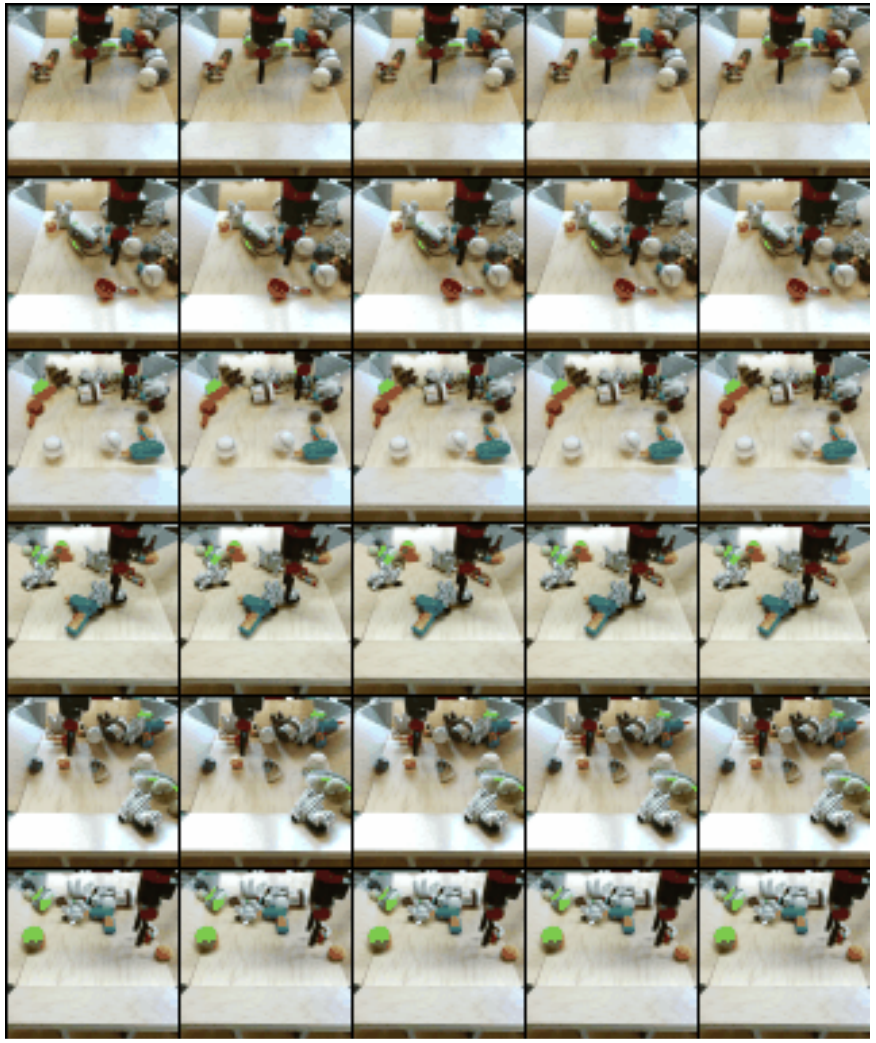


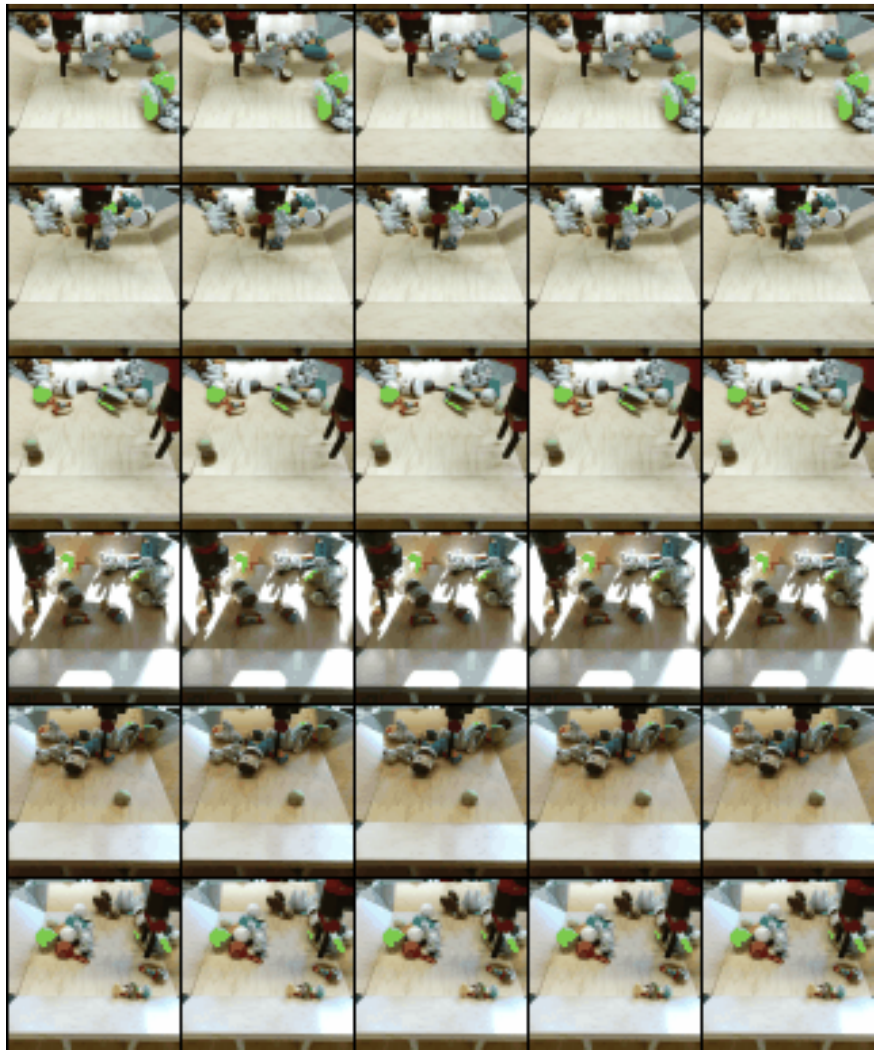
BAIR Push Dataset

Model	FVD (↓)	LPIPS (↓)	SSIM (↑)
SVG-LP [4]	256.62	0.061 ± 0.03	0.816 ± 0.07
SAVP [21]	**143.43**	0.062 ± 0.03	0.795 ± 0.07
Ours w/o Hier	149.22	0.058 ± 0.03	$\mathbf{0.829 \pm 0.06}$
Ours w/ Hier	**143.40**	$\mathbf{0.055 \pm 0.03}$	0.822 ± 0.06

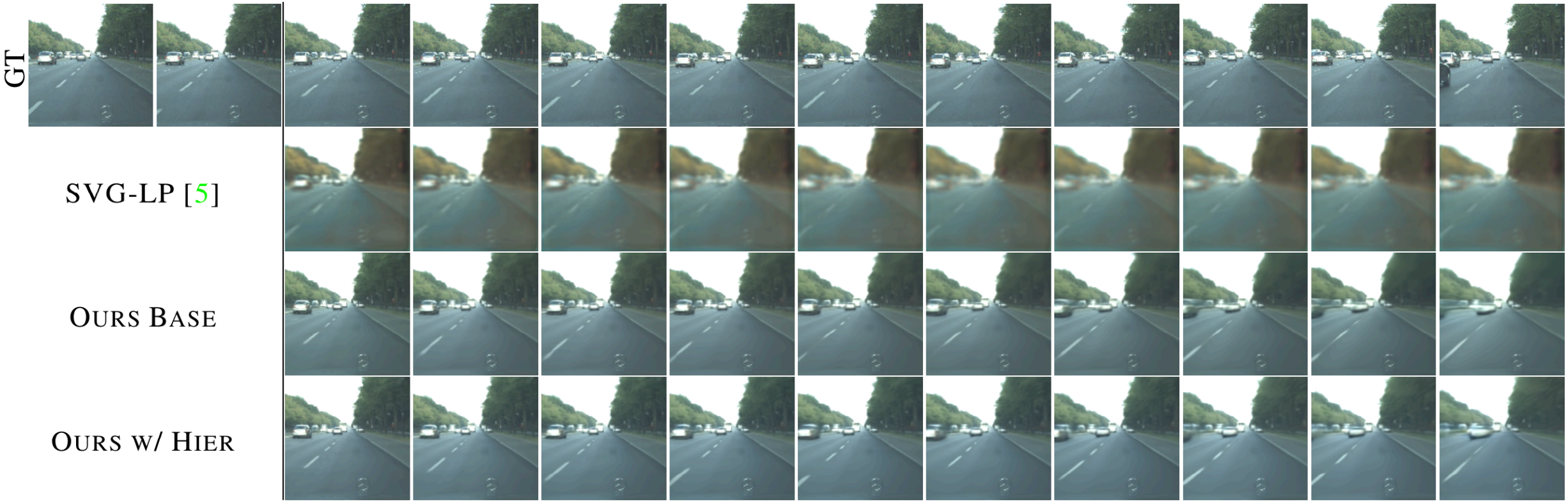


BAIR Push Samples



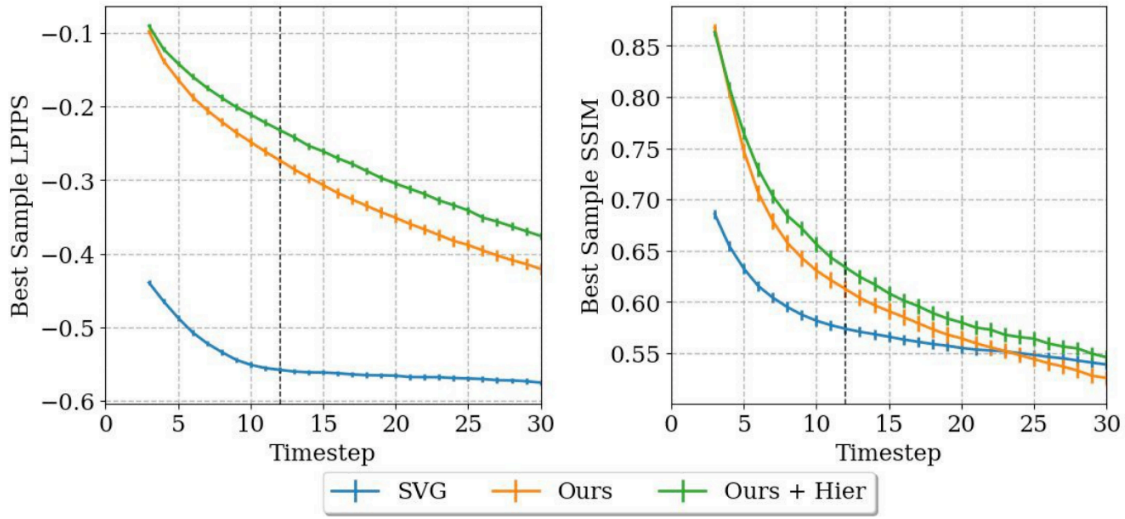


Cityscapes



Cityscapes

Model	FVD (↓)	LPIPS (↓)	SSIM (↑)
SVG-LP [4]	1300.26	0.549 ± 0.06	0.574 ± 0.08
Ours w/o Hier	682.08	0.304 ± 0.10	0.609 ± 0.11
Ours w/ Hier	**567.51**	$\mathbf{0.264 \pm 0.07}$	$\mathbf{0.628 \pm 0.10}$



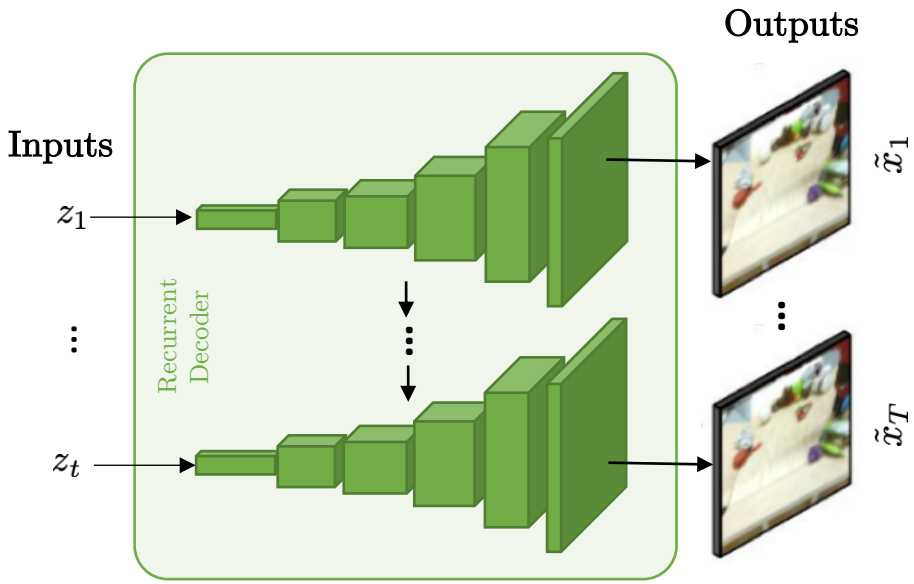
Cityscapes Samples

Outline

1. Motivation
2. Background
3. Related Work
4. Improved Conditional VRNNs for Video Prediction
5. **GANs for Video Prediction (ongoing work)**
6. Future Work and Conclusions

Generator G

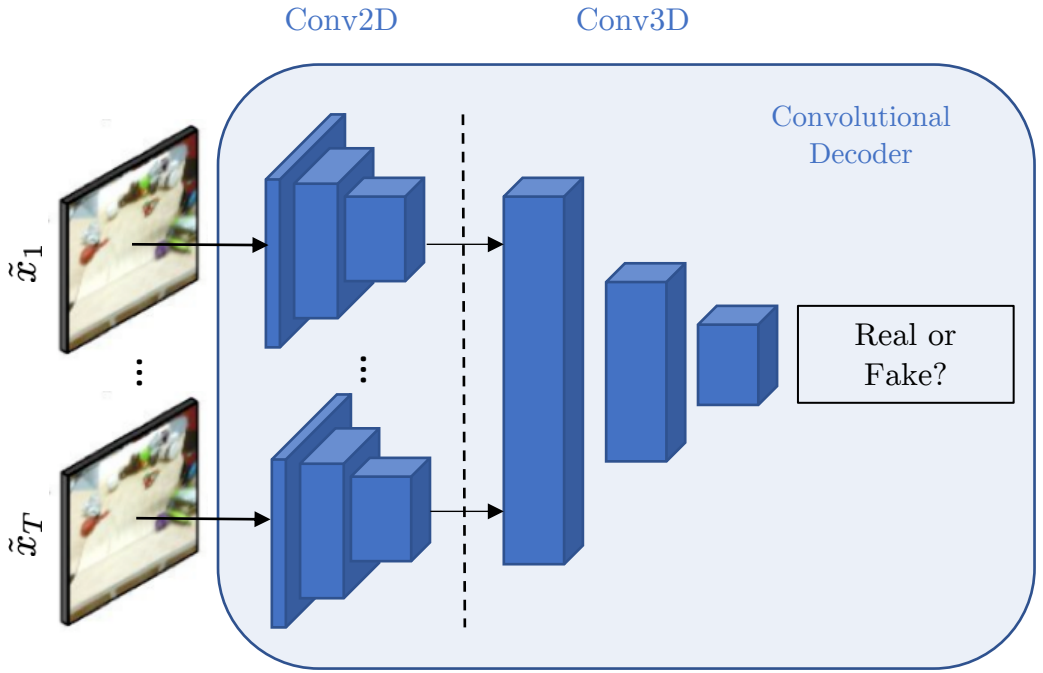
We reuse the Generator G architecture that we used as the decoder for the VRNN



However, there is only a latent per timestep and there is no output-input autoregression

Discriminator D

We use an architecture based on Conv2D to process low level details and Conv3D on features



Full Conv3D usually diverges, this setup is close to a dual discriminator (image + video)

Cost function

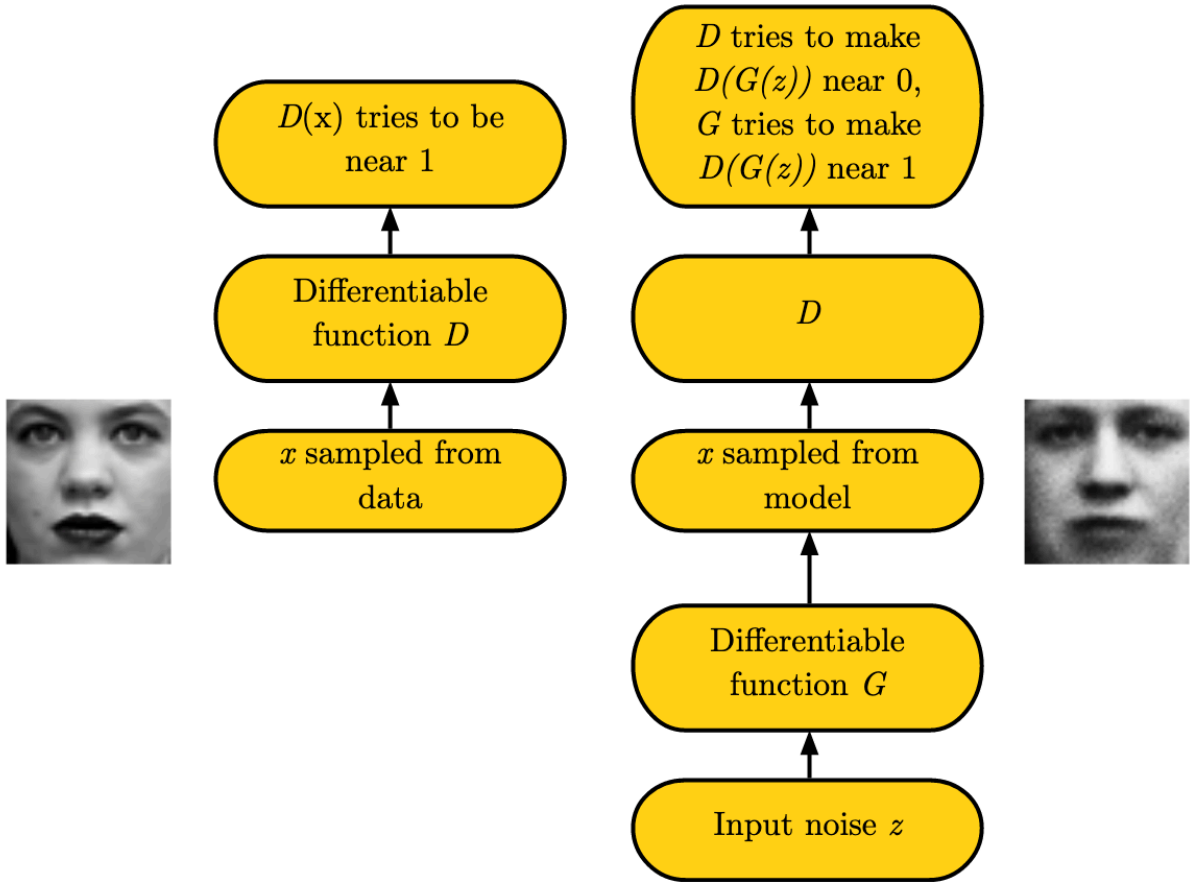


Figure from Goodfellow 2016

Cost function

$$J^{(D)}(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)}) = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log\left(1 - D\left(G(z)\right)\right)$$

$$J^{(G)} = -J^{(D)}$$

$$\boldsymbol{\theta}^{(G)*} = \arg\min_{\boldsymbol{\theta}^{(G)}} \max_{\boldsymbol{\theta}^{(D)}} V\left(\boldsymbol{\theta}^{(D)}, \boldsymbol{\theta}^{(G)}\right)$$

$$J^{(G)} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log D(G(\boldsymbol{z}))$$

BAIR Push Samples

Cityscapes Samples

Cityscapes Samples

SlowTV Dataset

Outline

1. Motivation
2. Background
3. Related Work
4. Improved Conditional VRNNs for Video Prediction
5. GANs for Video Prediction (ongoing work)
6. **Future Work and Conclusions**

Future Work

Current models operate at *low resolution* and predict a *small amount of frames*

This is mostly a matter of computational resources

For both of these issues we can use two-stage models that first generate a smaller-dimension sequence with temporal gaps and then have another model that upscales the prediction and interpolates temporally between the frames.

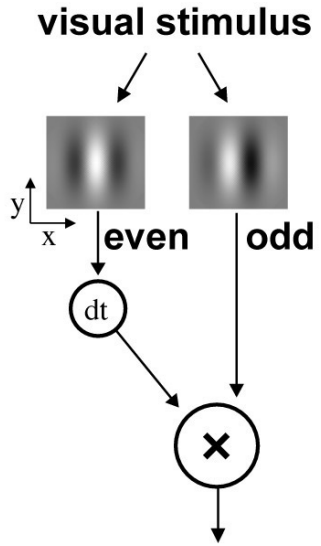
Better Inductive Biases

Neither ConvLSTMs nor Conv3D have been shown to work that well for video tasks.

A future research direction is to find better components to build video models.

Better Inductive Biases

We can take inspiration from Cognitive Sciene



Can be seen as interpreted as computing optical flow on features

Image from this URL

Conclusions

- Motivated the importance of Generative Models for Video
- Overview of the basic concepts and literature
- Proposed an improved VRNN model for Video Prediction
  - Improved Likelihood model
  - More flexible prior and approximate posterior
- Presented some early work on GANs for video
- Described some future research directions

Collaborators

Mila

facebook artificial intelligence research

Thank you for attending

+

Questions