Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks

Tianfan Xue*[1], Jiajun Wu*[1], Katherine L. Bouman[1], and William T. Freeman[1,2]

[1]MIT CSAIL [2]Google Research

http://visualdynamics.csail.mit.edu/

* Indicates equal contribution

NIPS 2016

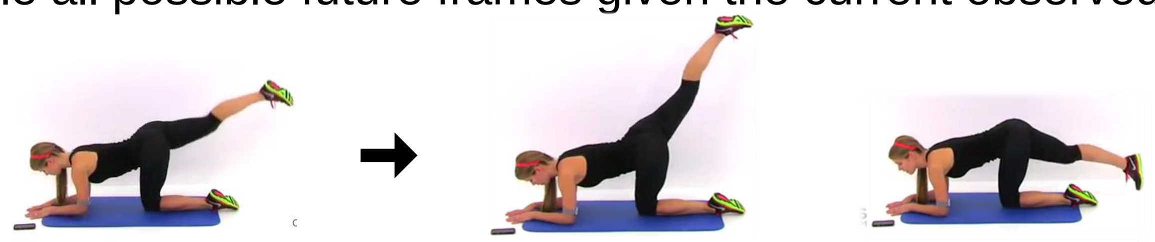
Visual Dynamics

Future frame prediction:

- Predict future frame from current observation
- **Ambiguity**: one observed frame corresponds multiple possible future frames

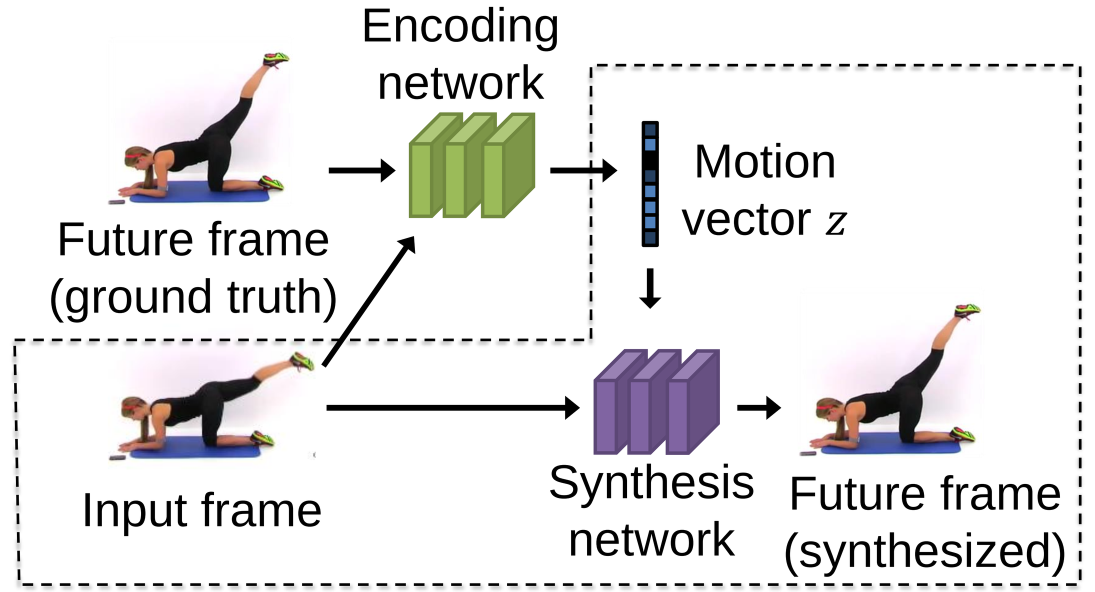
Problem definition: probabilistic future frame synthesis

Task: sample all possible future frames given the current observed snapshot



Observed snapshot Two possible future frames

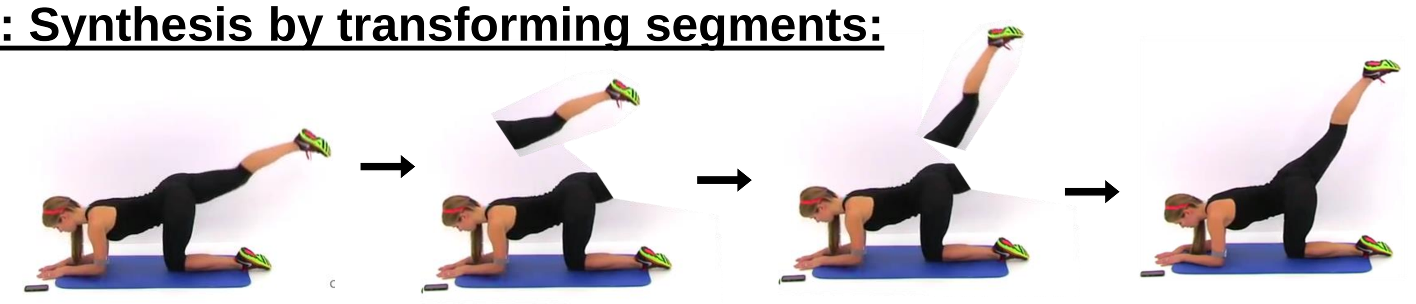
Idea 1: Probabilistic synthesis via conditional variational autoencoder:



Training: use the whole network

Testing: only use the dashed part and randomly sampled z

Idea 2: Synthesis by transforming segments:

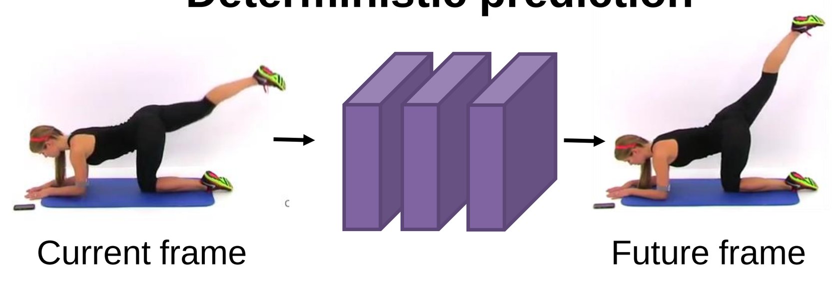


Input frame Segments Transformed segments Sampled future frame

Discussion

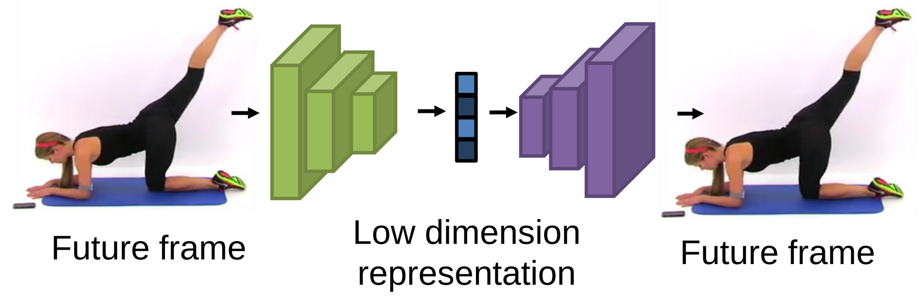
Two naïve baselines:

Deterministic prediction



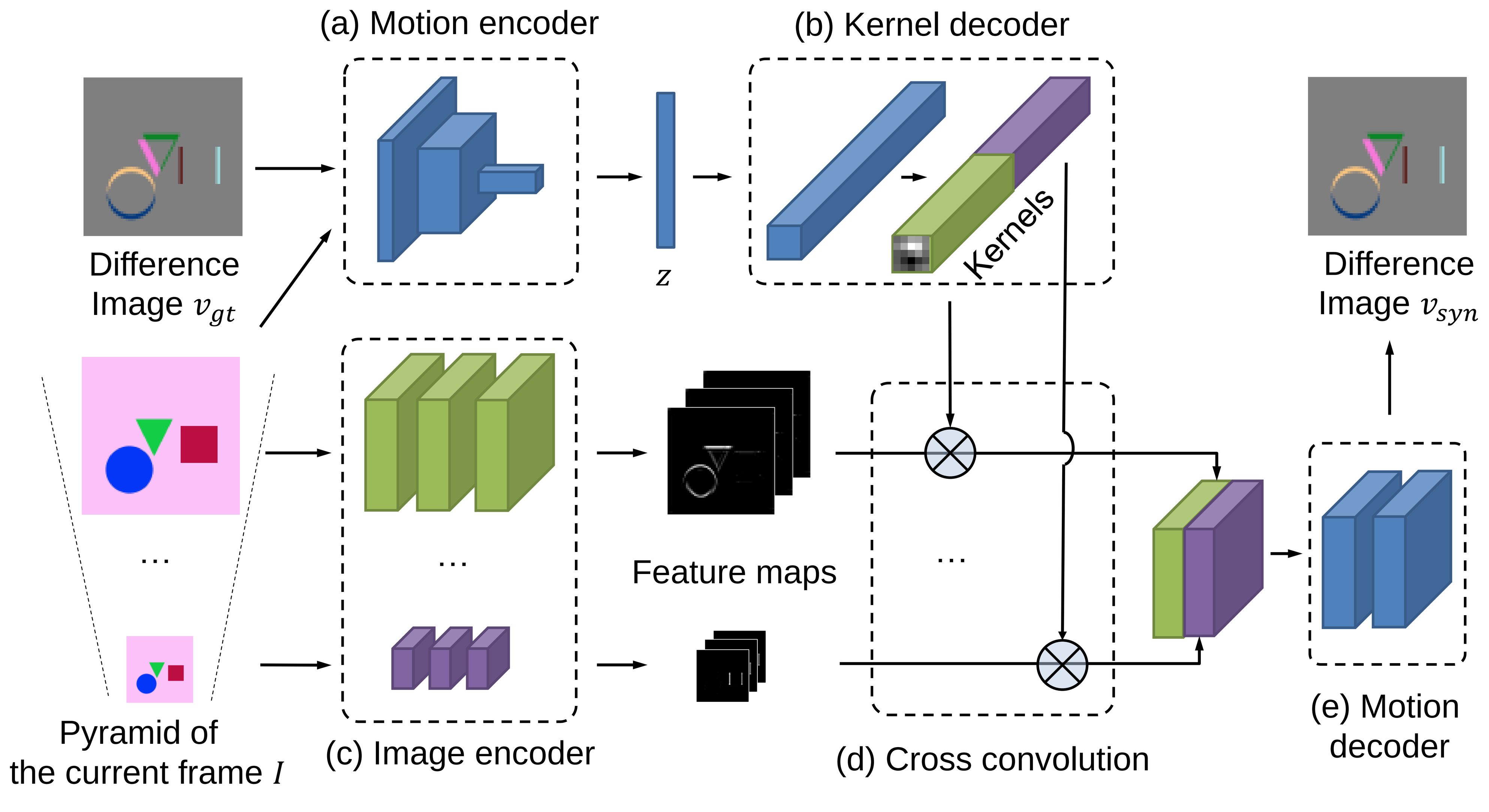
Fails to model the uncertainty of future

Autoencoder



Only learns a prior distribution of the image

Network Structure



Training Objective: $D_{KL}(q_\phi(z|v_{syn}, I)||N(\mathbf{0}, \mathbf{I})) + \lambda \cdot \|v_{syn} - v_{gt}\|$

KL-divergence loss Reconstruction loss

Encoding network $q_\phi(z|v, I)$:

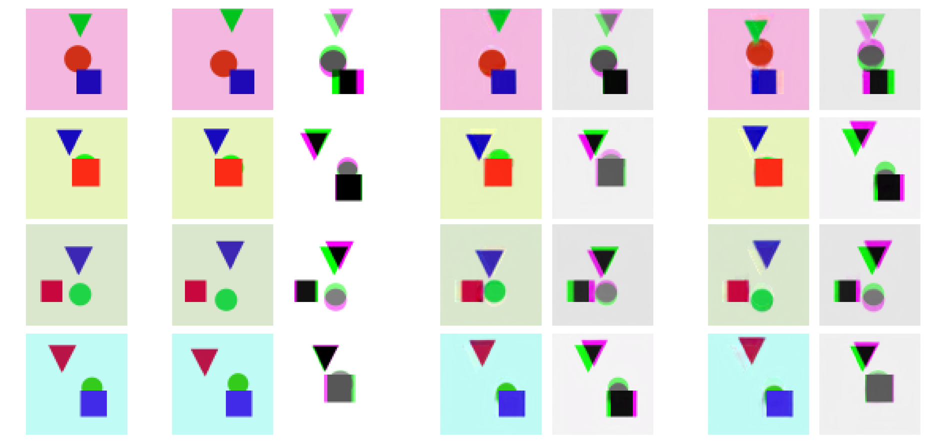
Consists of (a) Motion encoder, which predicts the motion information z from two frames.

Synthesis network $p_\theta(v|z, I)$

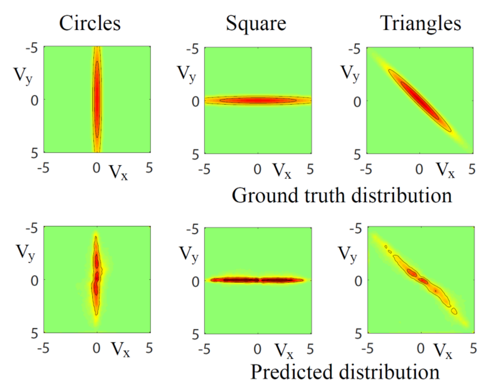
Consists of (b) Kernel decoder, (c) Image encoder, (d) Cross convolution, and (e) Motion decoder:

Experiments

Synthetic dataset:



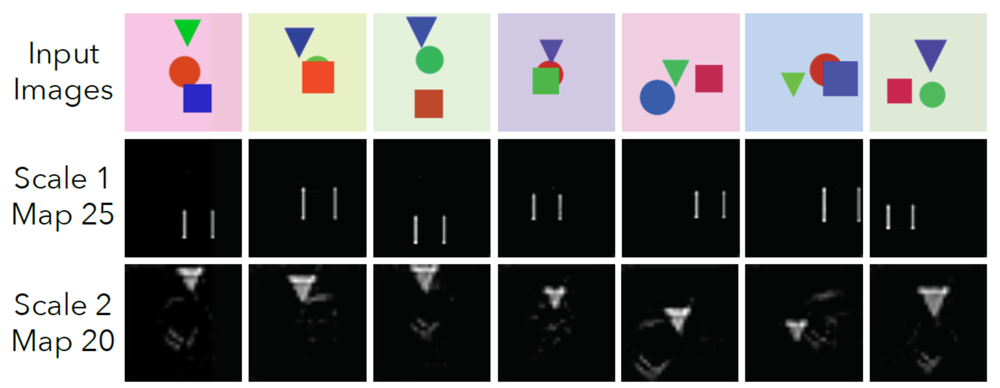
Frame 1 Frame 2 (Ground truth) Frame 2 (Sample 1) Frame 2 (Sample 2)



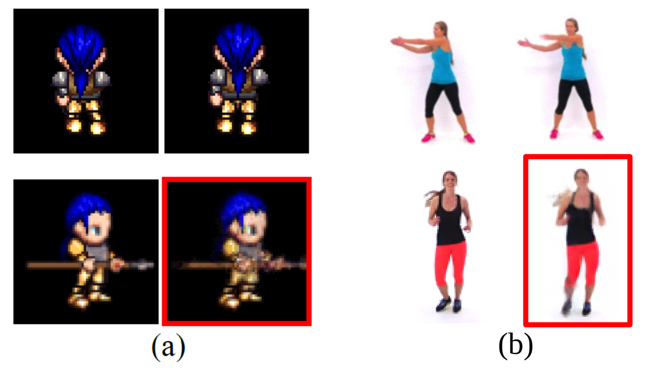
KL divergence ($KL(p_{gt} || p_{pred})$) between predicted and ground truth distributions

Method	Shapes			
	C.	S.	T.	C.-T.
Flow	6.77	7.07	6.07	8.42
AE	8.76	12.37	10.36	10.58
Ours	**1.70**	**2.48**	**1.14**	**2.46**

Visualization of learned feature maps:



Visual analogy:



(a) (b)

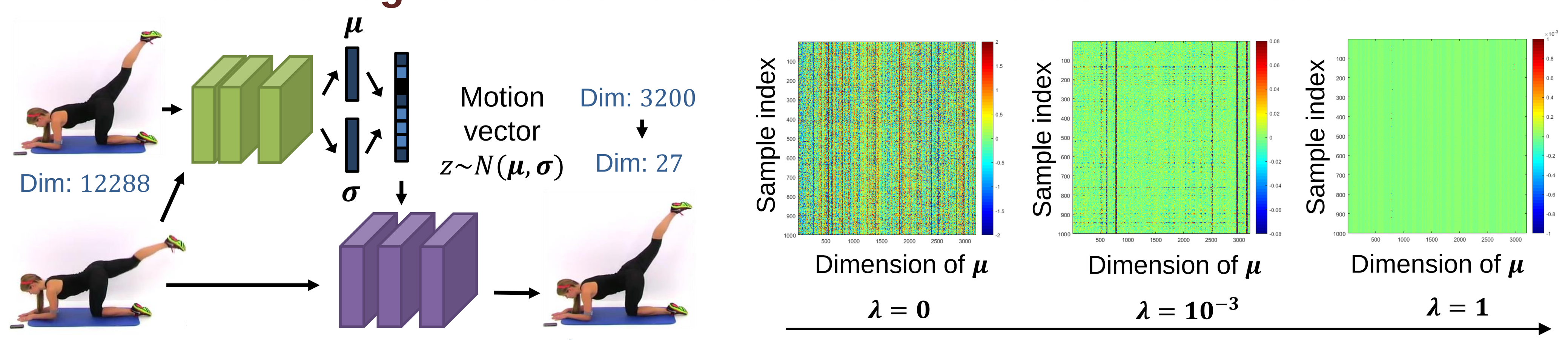
Model	spellcast	thrust	walk	slash	shoot	average
Add [Reed et al., 2015]	41.0	53.8	55.7	52.1	77.6	56.0
Dis [Reed et al., 2015]	40.8	55.8	52.6	53.5	79.8	56.5
Dis + Cls [Reed et al., 2015]	13.3	24.6	17.2	**18.9**	40.8	23.0
Our Model	**9.5**	**11.5**	**11.1**	28.2	**19.0**	**15.9**

(c) Comparison with [Reed et al. 2015]

Video demo & motion vector visualization

Video demo

DL-divergence ensures the motion vector is low dimension



The effective dimension of z decreases as λ increases

- $D_{KL}(N(\boldsymbol{\mu}, \boldsymbol{\sigma})||N(\mathbf{0}, \mathbf{I})) = \sum_j \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2$, D_{KL} is minimized when $\mu_j = 0$ and $\sigma_j = 1$
- Shown in [Hinton and Camp 1993], KL-divergence penalizes the information z carries, so it reduce its effective dimension

Derivation of training objective:

Generative process in testing:

To sample a future frame J from observation I:
1) Sample z from a prior distribution $z \sim p_z(z) = N(\mathbf{0}, I)$;
2) Given z, sample the intensity difference image from $v \sim p_\theta(v|I, z)$.
3) Synthesize the future frame $J = I + v$.

Training:

- Maximize the marginal distribution:
$$\sum_i \log \int_z p_\theta(v^{(i)}|I^{(i)}, z) p_z(z) dz$$
where $(I^{(i)}, v^{(i)})$ are training samples
- Approximate the distribution by the variational upper bound:
$$-D_{KL}(q_\phi(z|v^{(i)}, I^{(i)})||p_z(z)) + \frac{1}{L}\sum_{l=1}^{L}\left[\log p_\theta\left(v^{(i)}|z^{(i,l)}, I^{(i)}\right)\right]$$

Notation:

- $q_\phi(z|v^{(i)}, I^{(i)})$ is the variational distribution of $p(z|v^i, I^i)$, defined by the encoding network.
- p_θ is defined by the synthesis network.

References:

1. G. Hinton and D. Camp. Keeping the neural networks simple by minimizing the description length of the weights, 1993
2. D. Kingma and M. Welling. Auto-encoding variational bayes, ICLR, 2014
3. C. Finn, I. Goodfellow, S. Levine. Unsupervised learning for physical interaction through video prediction, NIPS, 2016
4. J. Walker, C. Doersch, and A. Gupta. An uncertain future: Forecasting from static images using variational autoencoders, ECCV, 2016
5. B. Brabandere, X. Jia, T. Tuytelaars, and L. Gool. Dynamic filter networks, NIPS, 2016