Hierarchical Long-term Video Prediction without Supervision ICML 2018

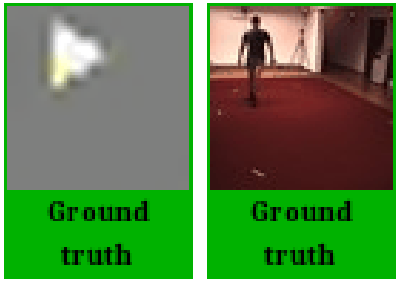
Nevan Wichers, Ruben Villegas, Dumitru Erhan, Honglak Lee

Seminar: Computer Vision SS19

Stefan Möhrle

Introduction

- Given: C context frames
- Goal: predict future frames



Wichers et al. 2018

Motivation

- Self-supervision
  - Autonomously collect virtually unlimited experience
  - Make huge amounts of data usable
- Video prediction
  - Learn representations of motion/actions
  - Future prediction for interacting agents

⇒ Unsupervised feature learning

⇒ Accurate model of the world for reinforcement learning

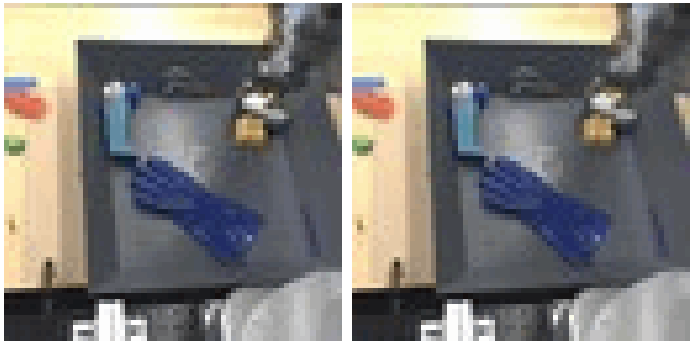
Related Work

Frame-level

- Predicts pixel movement
- Observes own predictions

Finn et al (2016)

- Predict the motions of different objects / image segments
- Merge predictions via masking
- Convolutional dynamic neural advection (CDNA)
  - 10 normalized transformation kernels
  - 11-channel composing mask



Finn et al. 2016

Denton and Fergus (2018)

- Deterministic frame predictor
- Time-dependent stochastic latent variables z_t
  - Modeled by a separate convolutional LSTM
  - High variance in periods of uncertainty

Denton & Fergus 2018

Related Work

Frame-level

- Predicts pixel movement
- Observes own predictions

Related Work

Frame-level

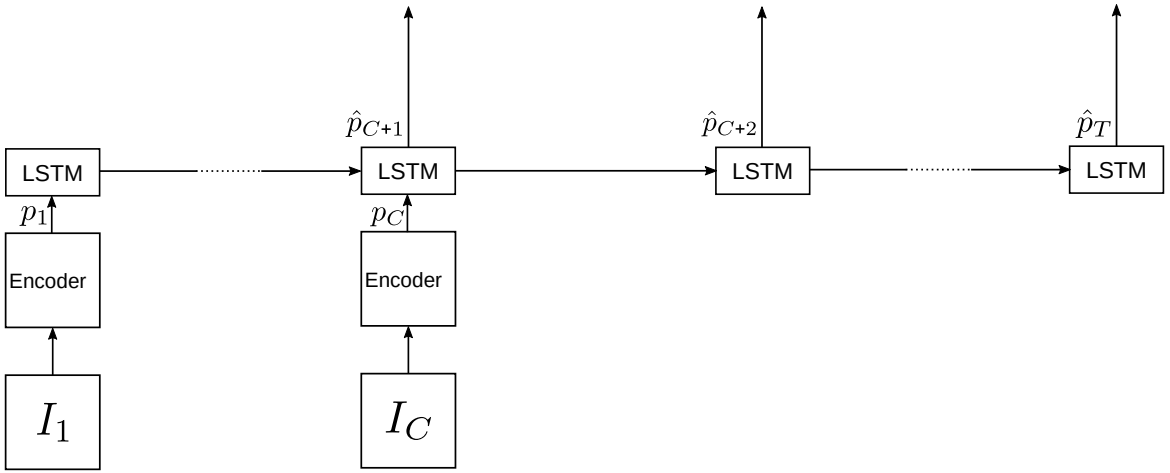
- Predicts pixel movement
- Observes own predictions

Hierarchical

- Predicts high-level features
- Separation between prediction and image generation

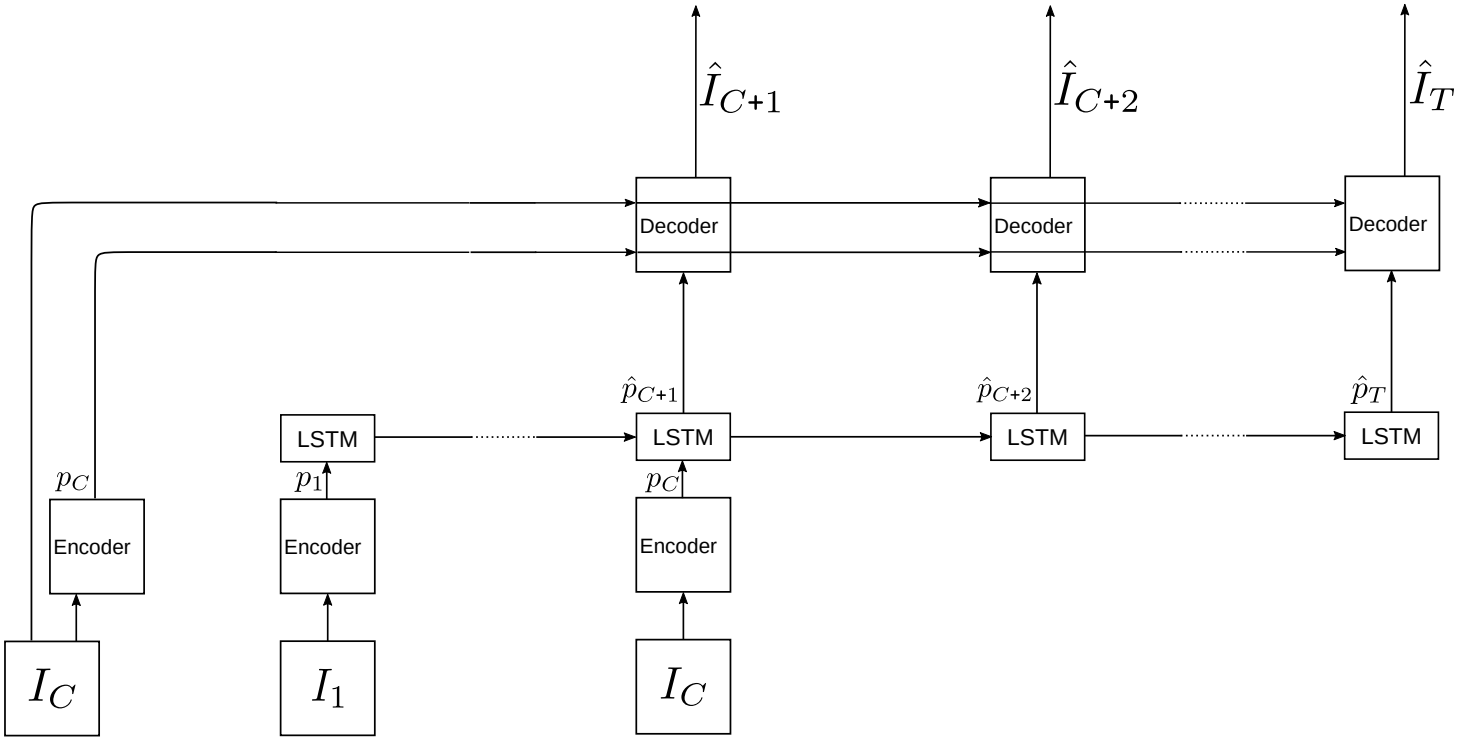
Villegas et al (2017)

- p_t: human pose landmarks, C: #Contextframes



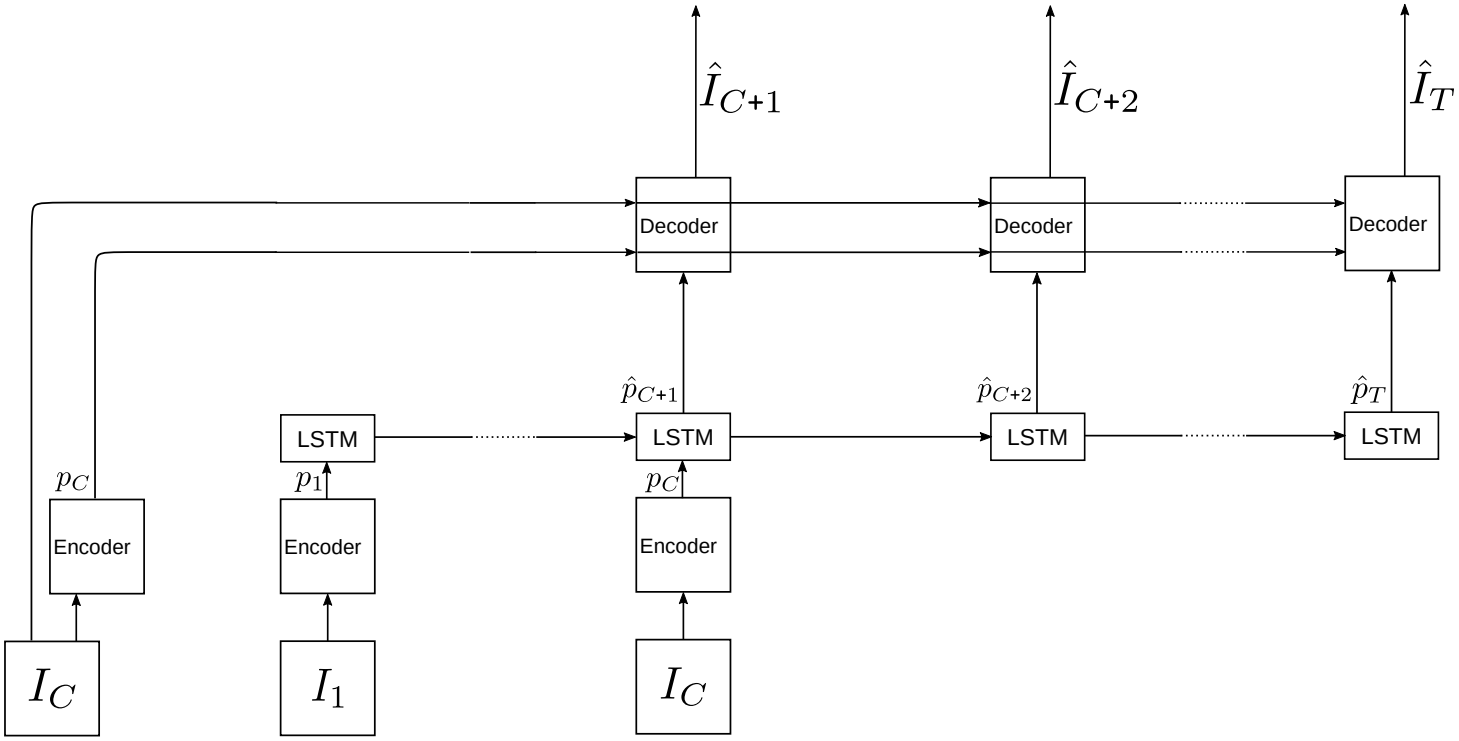
Villegas et al (2017)

- Separate training of LSTM and decoder



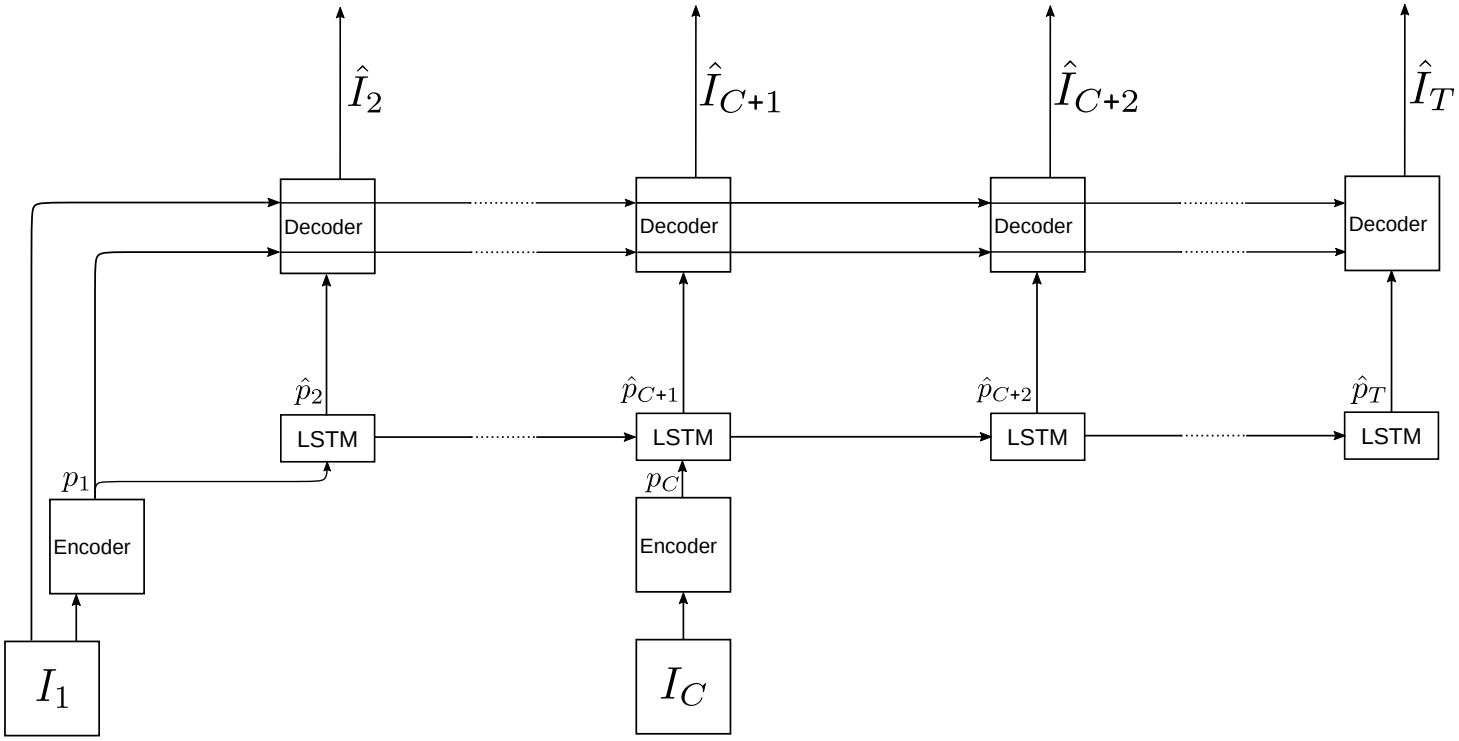
Approach

- Based on network architecture of Villegas et al.
- No ground truth annotations required
- Joint training



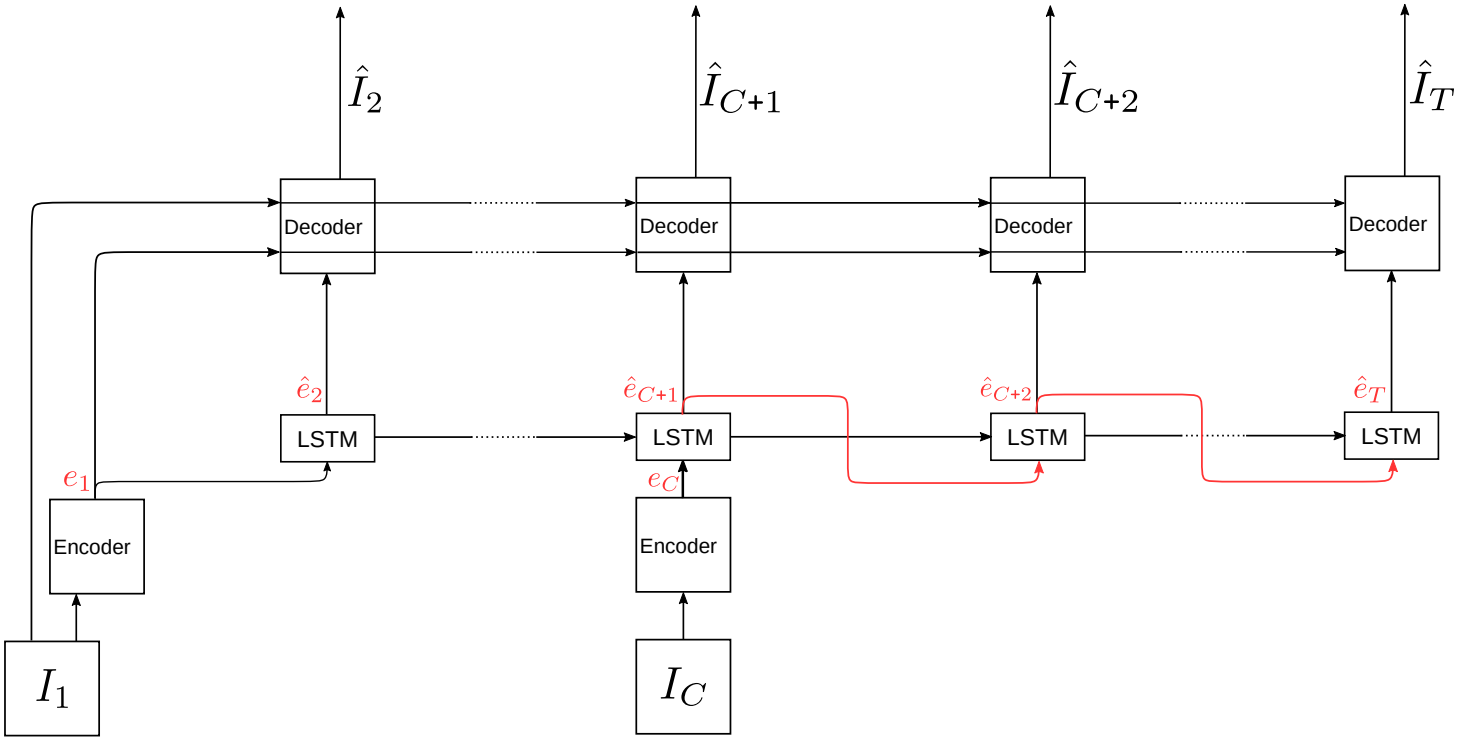
Approach

- Use I_1 instead of I_C for the decoder



Approach

- Use general feature vector with auto-regressive connection



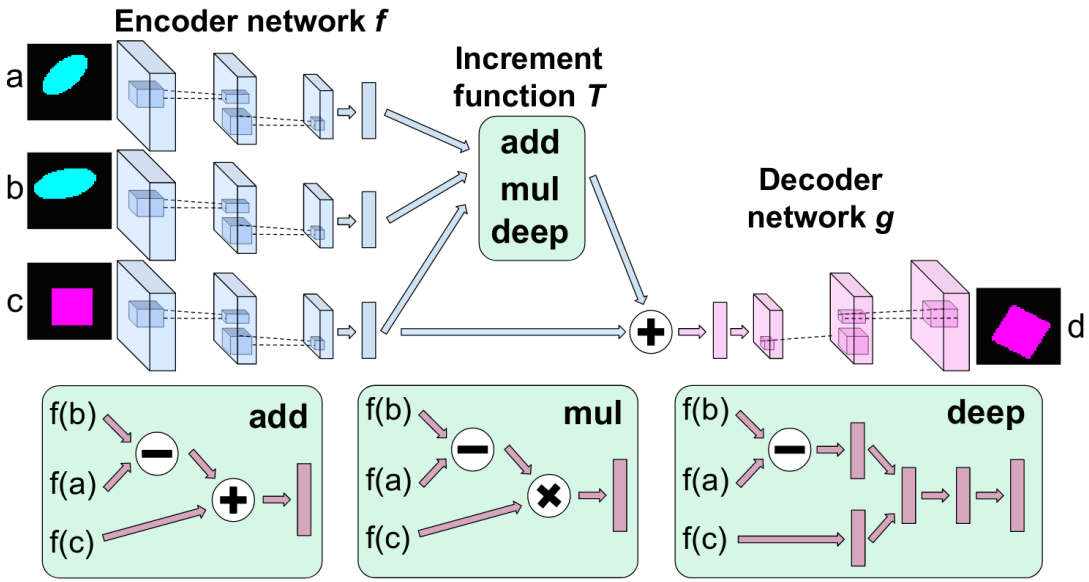
Decoder

- Deep version visual analogy network (VAN, Reed et al., 2015)
  - $\bar{I}_t, M_t = \mathrm{VAN}(e_1, \hat{e}_t, I_1)$
  - $\hat{I}_t = \bar{I}_t \odot M_t + I_1 \odot (1 - M_t)$

Decoder

- $\mathrm{VAN}(e_1, \hat{e}_t, I_1) =$

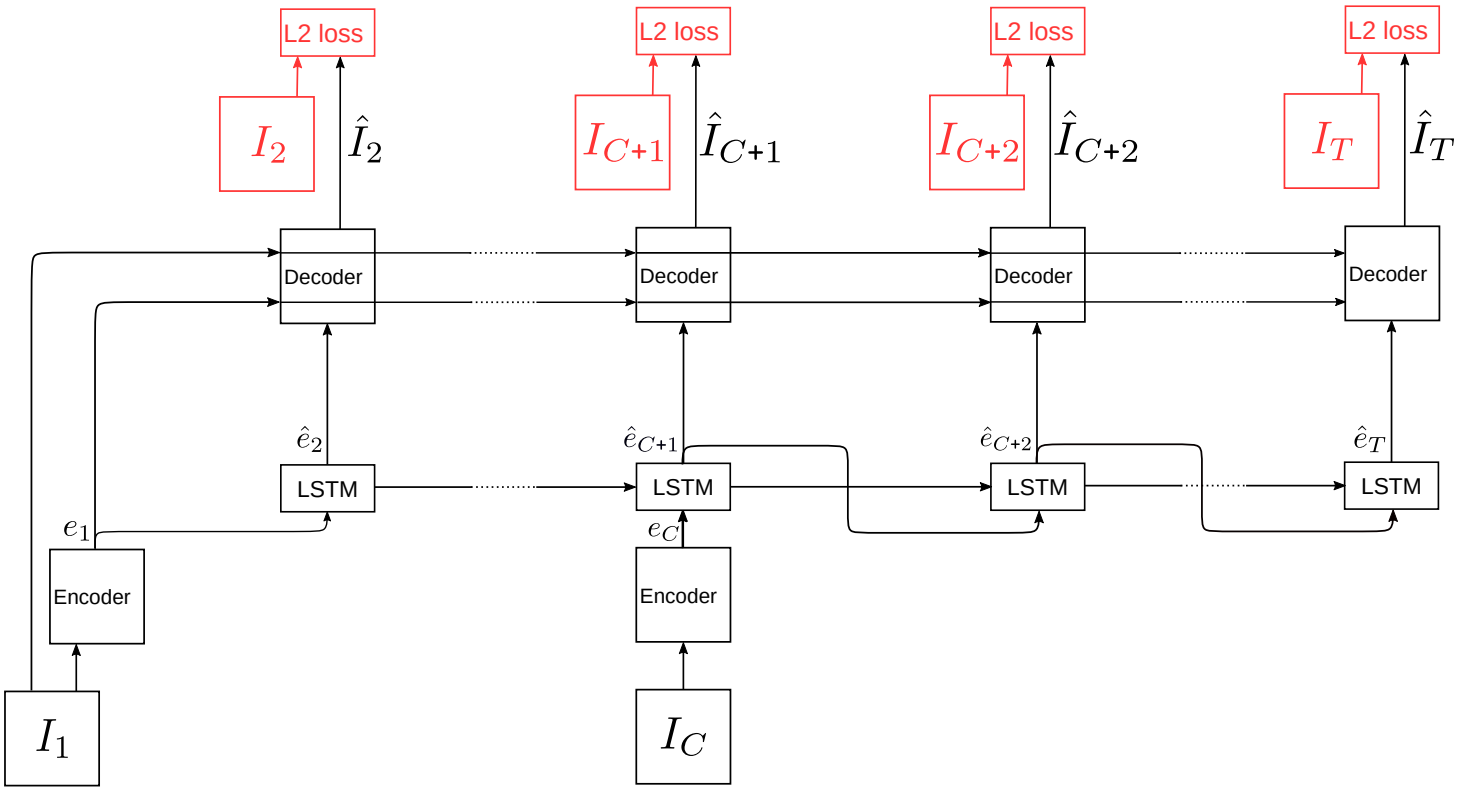
$$g_{\mathrm{dec}}(f_{\mathrm{enc}}(\hat{e}_t) + T(f_{\mathrm{img}}(I_1), f_{\mathrm{enc}}(e_1), f_{\mathrm{enc}}(\hat{e}_t)))$$



Reed et al. 2015

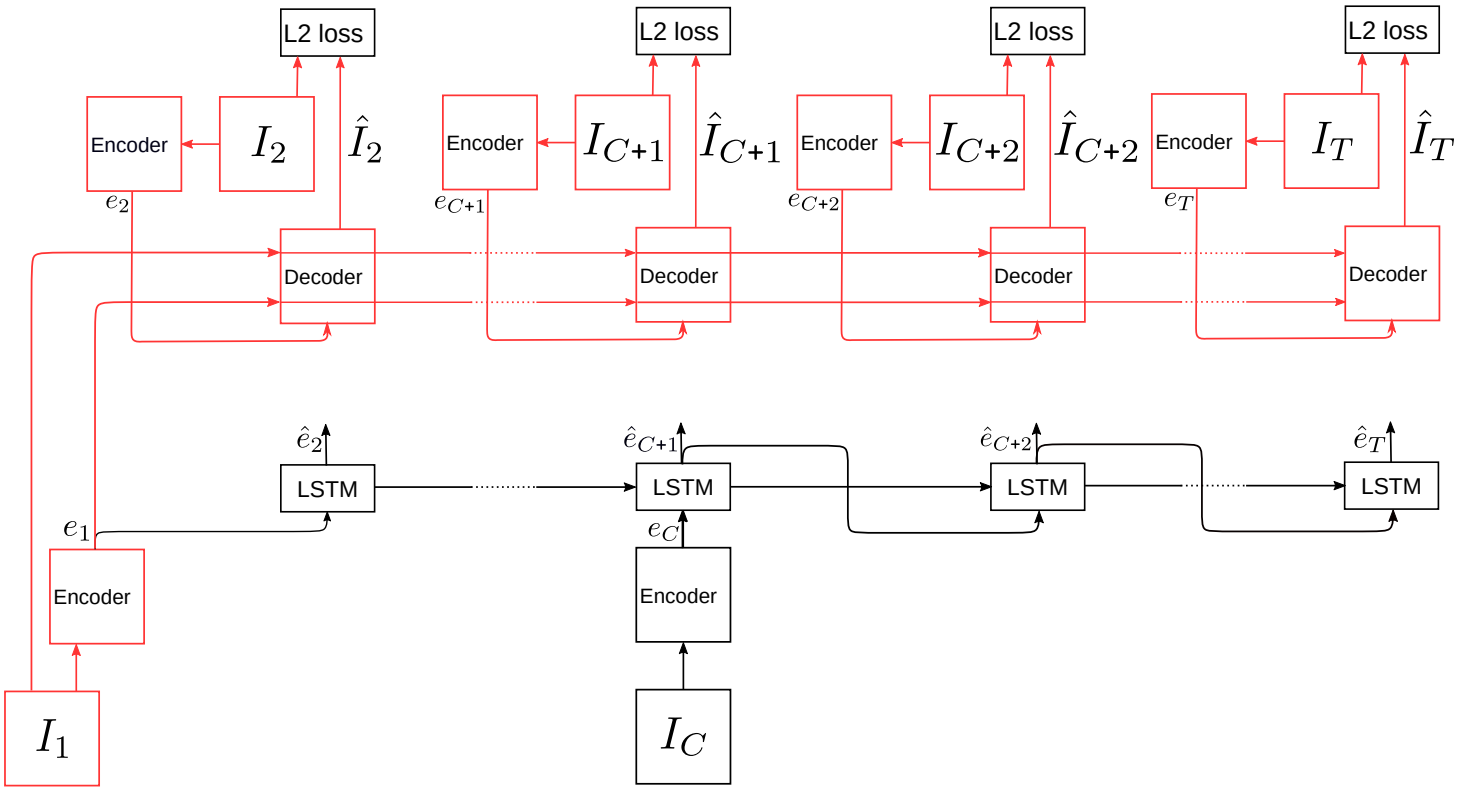
End to End

- Train end-to-end with L2 loss
- Produces blurry predictions



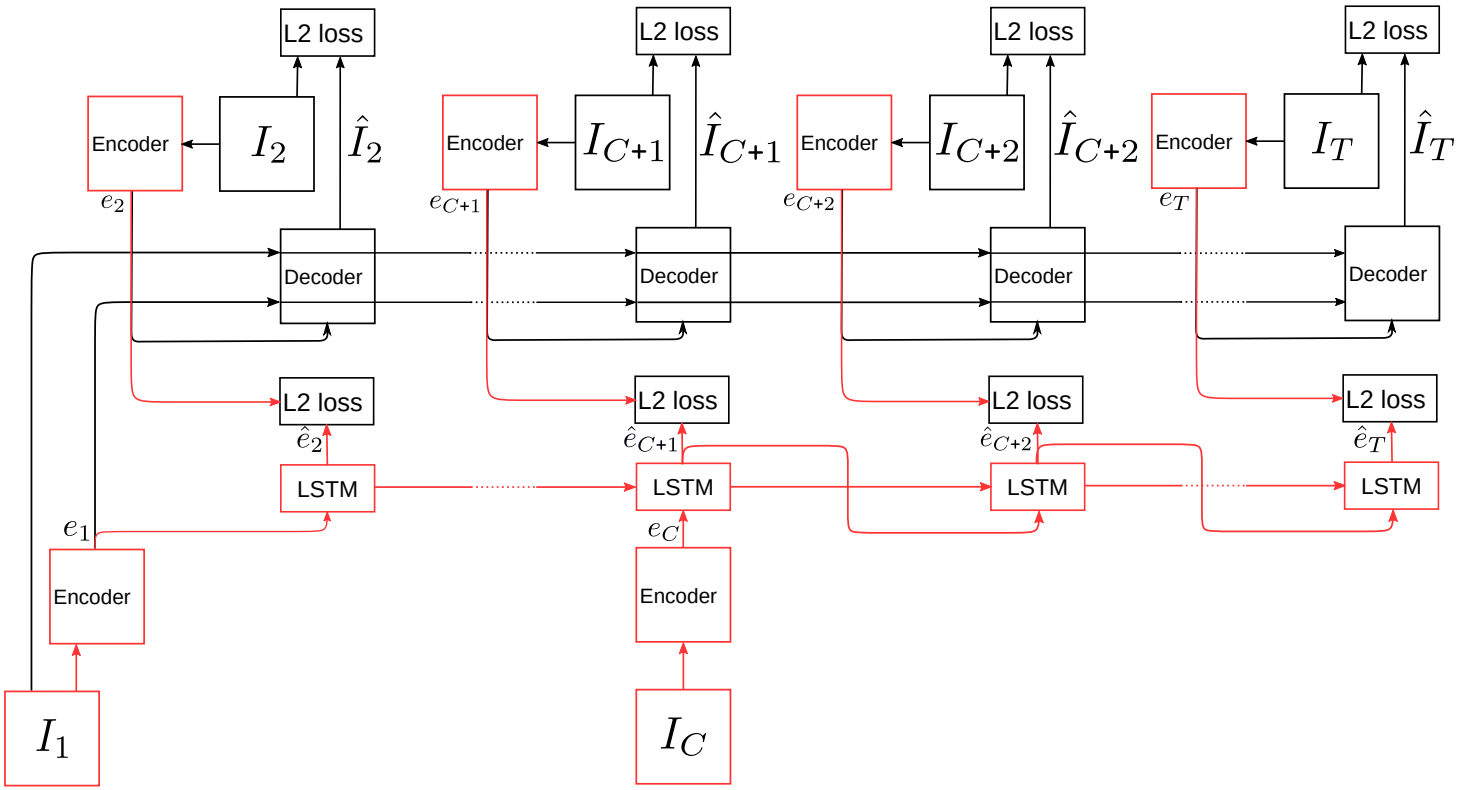
Encoder Predictor with Visual Analogy (EPVA)

- Train decoder without uncertainty of predictor



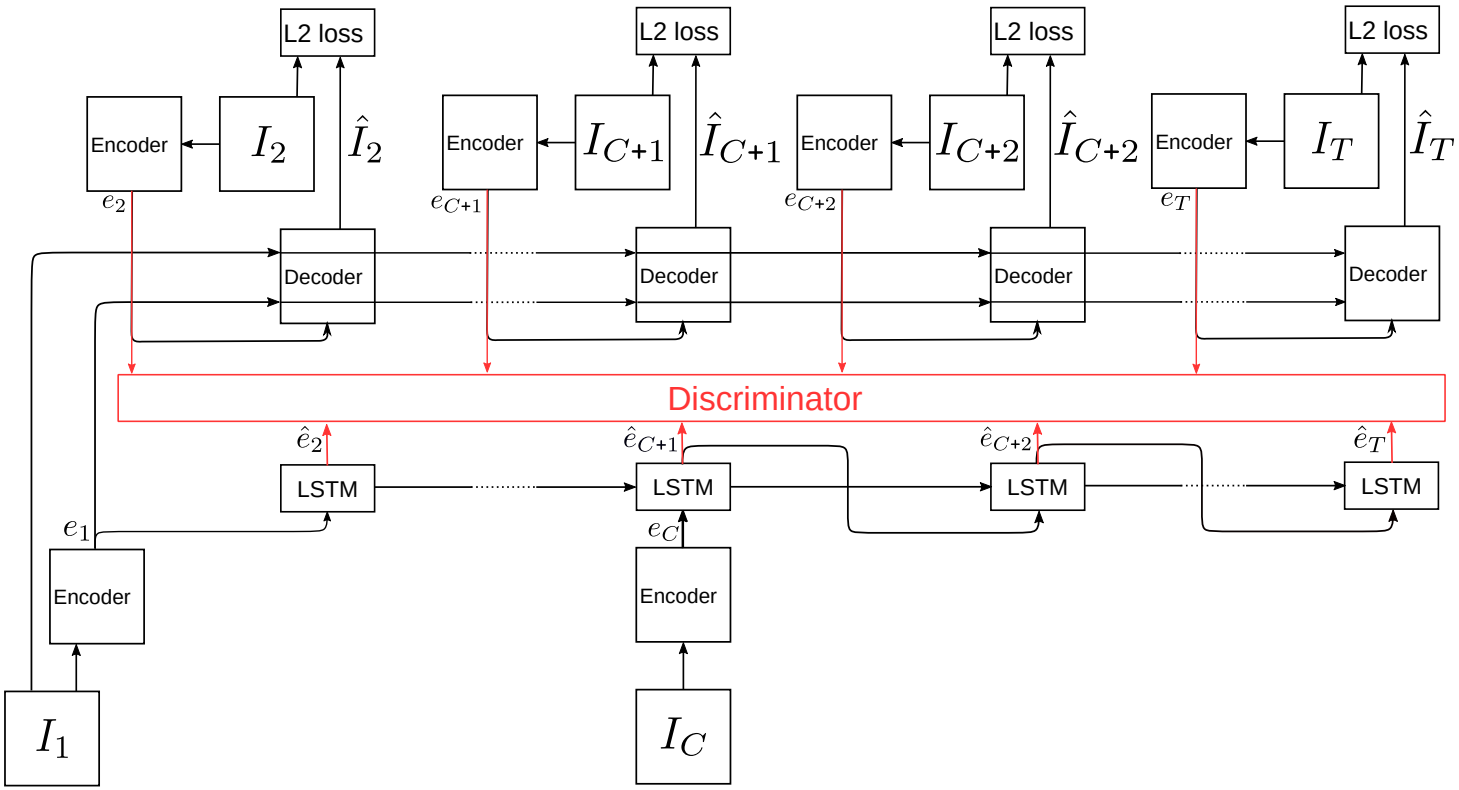
Encoder Predictor with Visual Analogy (EPVA)

- Train predictor without pixel-level noise



EPVA Adversarial Loss

- Reduce blurry predictions with discriminator network

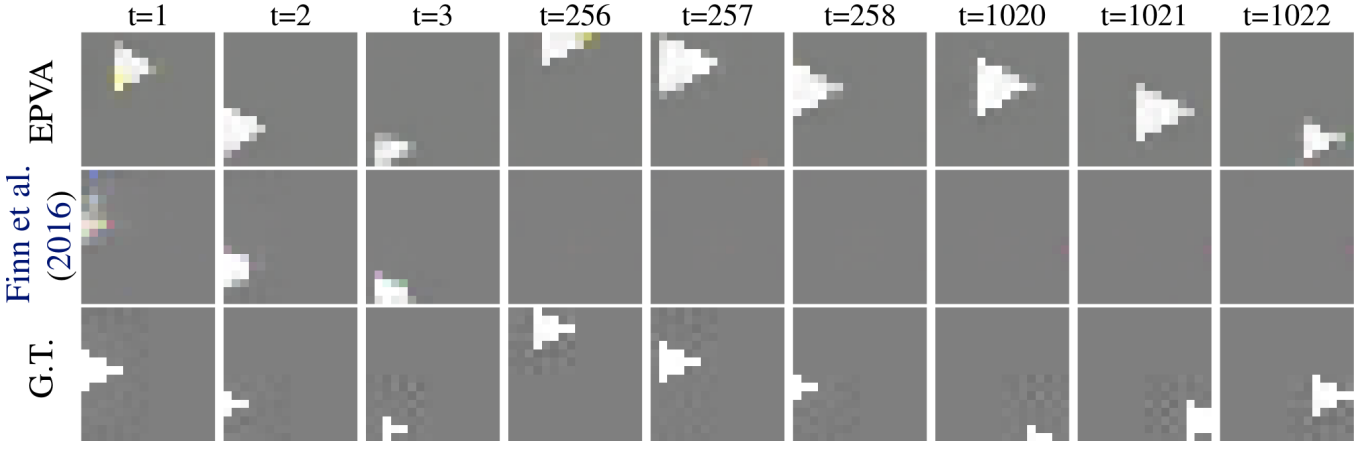


Evaluation: Toy dataset

- Bouncing shape with changing size
- Context: 3 frames
- Train: 16 frames
- Evaluate: ~1000 frames
- Test: shape color and presents (frame 1012 to 1022)

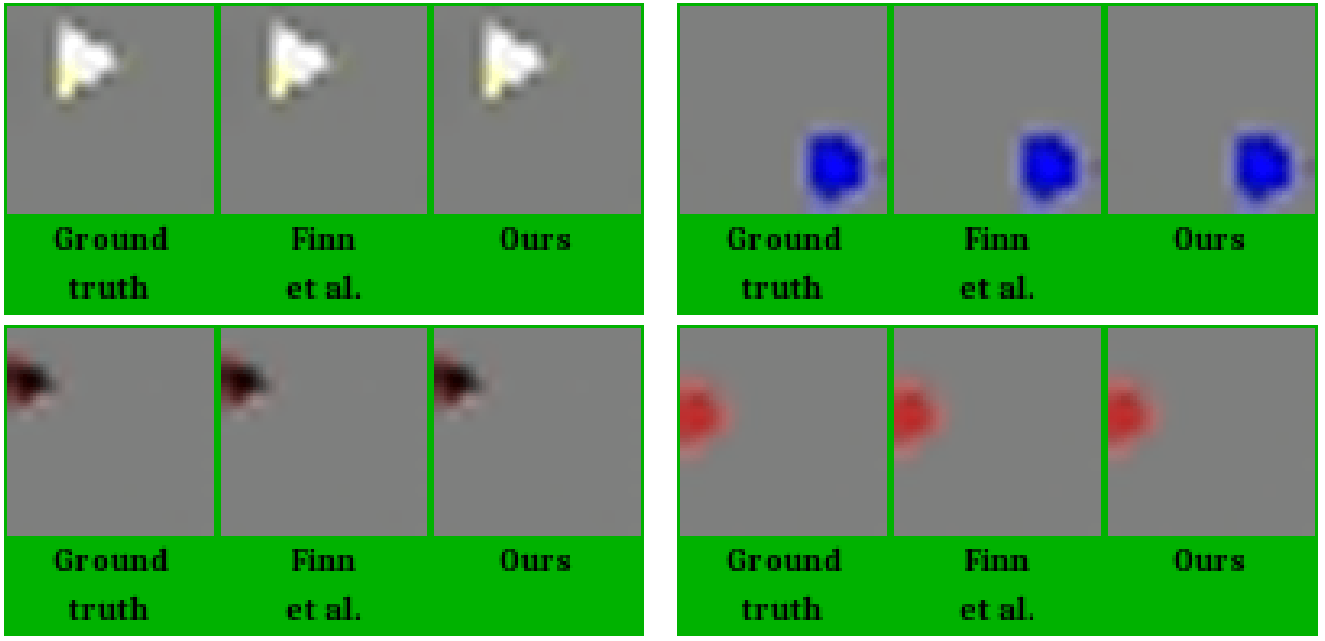
Evaluation: Toy dataset

- Bouncing shape with changing size
- Context: 3 frames
- Train: 16 frames
- Evaluate: ~1000 frames
- Test: shape color and presents (frame 1012 to 1022)



Wichers et al. 2018

Evaluation: Toy dataset



Wichers et al. 2018

Evaluation: Toy dataset

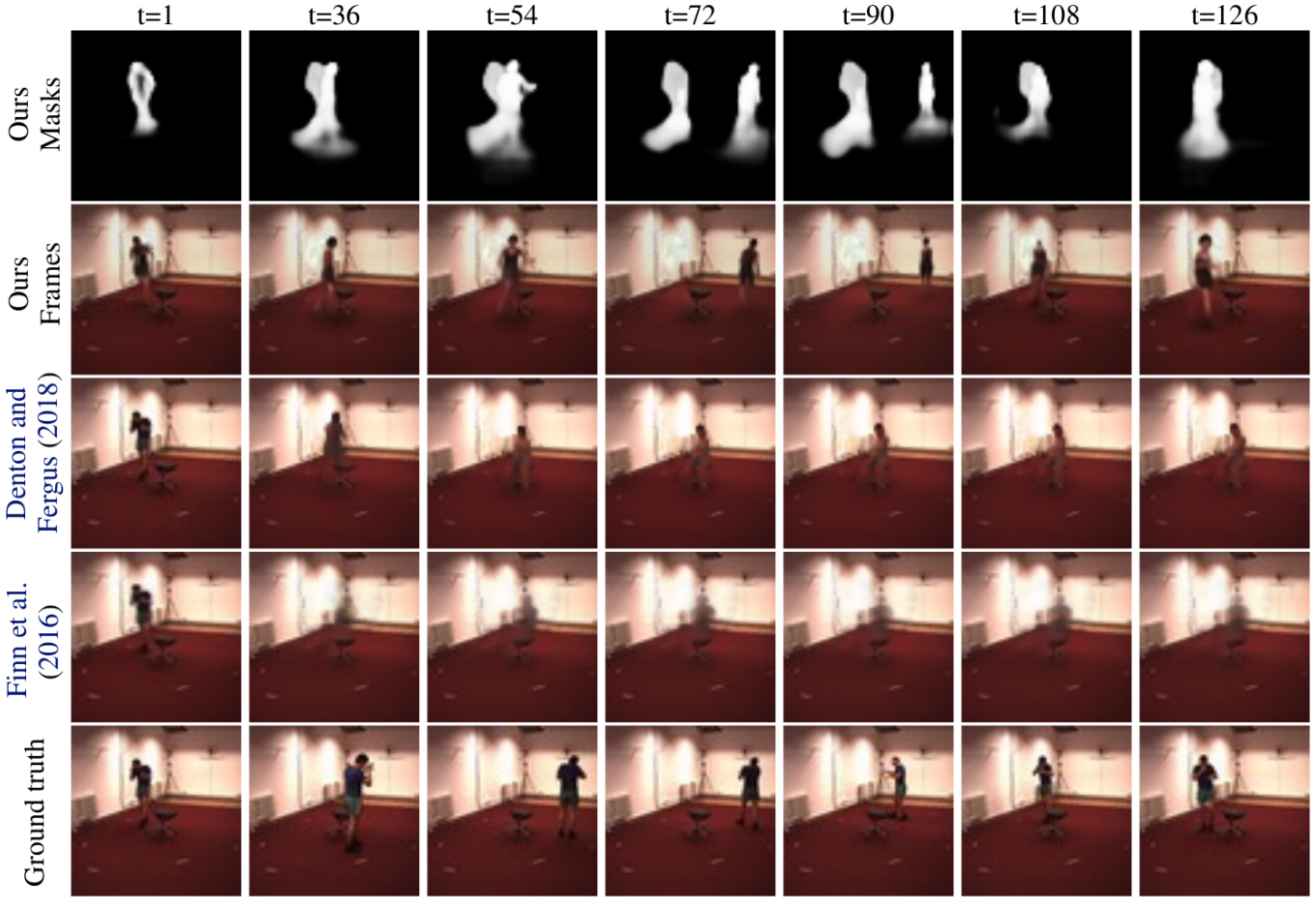
- Bouncing shape with changing size
- Context: 3 frames
- Train: 16 frames
- Evaluate: ~1000 frames
- Test: shape color and presents (frame 1012 to 1022)

Method	Shape has correct color	Shape has wrong color	Shape disappeared
EPVA	96.9%	3.1%	0%
Finn et al. (2016)	24.6%	5.7%	69.7%

Evaluation: Human3.6M

- Subsampled to 6.25 frames per second
- Context: 5 frames (0.8s)
- Train: 32 frames
- Evaluate: 126 frames (~ 20s)
- Test: side by side comparison

Evaluation: Human3.6M



Wichers et al. 2018

Evaluation: Human3.6M

Evaluation: Human3.6M

- Subsampled to 6.25 frames per second
- Context: 5 frames
- Train: 32 frames
- Evaluate: 126 frames
- Test: side by side comparison

Comparison	Frames	Ours is better	Same	Baseline is better
EPVA vs Finn et al. (2016)	1-127	46.4%	40.7%	12.9%
EPVA Adv. vs Finn et al. (2016)	1-127	73.9%	13.2%	12.9%
EPVA Adv. vs Denton & Fergus (2018)	5-127	58.2%	24.0%	17.8%

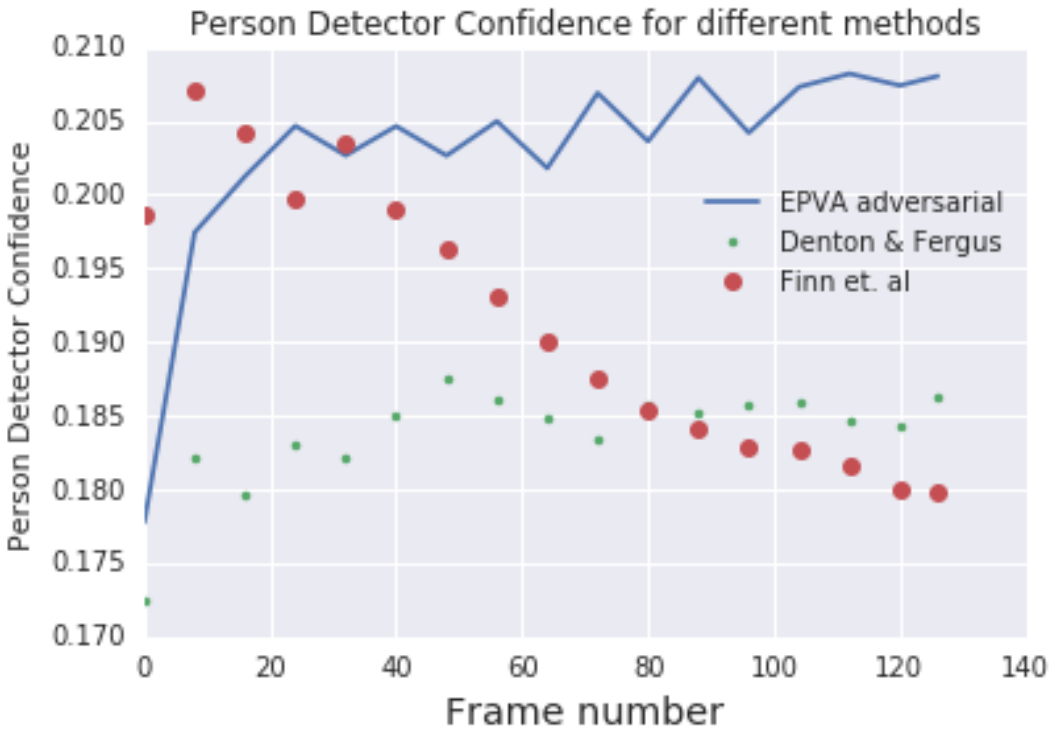
Evaluation: Human3.6M

- Subsampled to 6.25 frames per second
- Context: 5 frames
- Train: 32 frames
- Evaluate: 126 frames
- Test: side by side comparison

Comparison	Frames	Ours is better	Same	Baseline is better
EPVA vs Finn et al. (2016)	1-127	46.4%	40.7%	12.9%
EPVA Adv. vs Finn et al. (2016)	1-127	73.9%	13.2%	12.9%
EPVA Adv. vs Denton & Fergus (2018)	5-127	58.2%	24.0%	17.8%
EPVA Adv. vs Finn et al. (2016)	63-127 1-63	67.2%	17.5%	15.3%

Evaluation: Human3.6M

Person detector



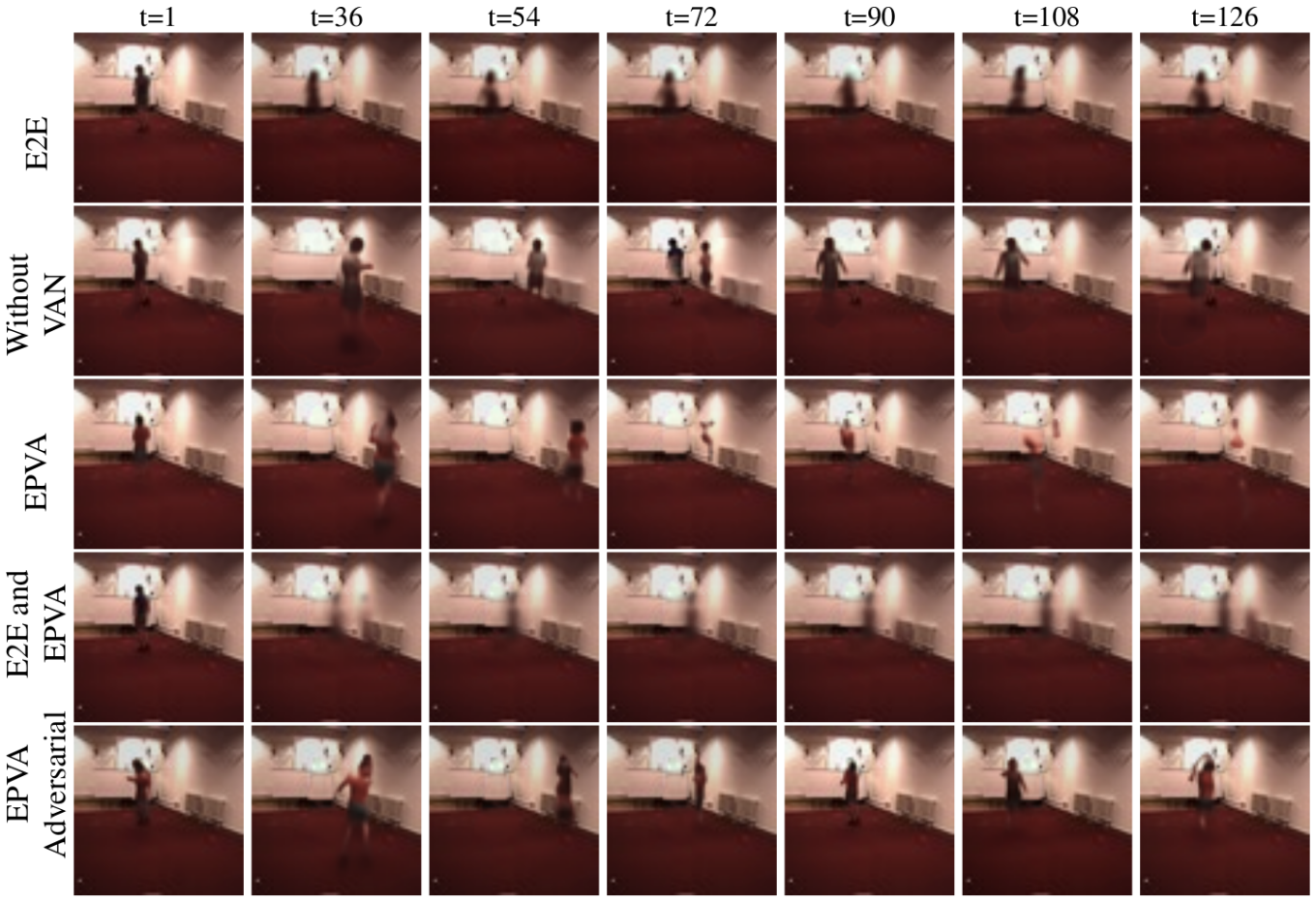
Wichers et al. 2018

Evaluation: Feature transfer

- Learned features + 2-layer MLP
- Baseline: features by VVG-net trained for recognition
- Test: Output human pose landmarks
- 9% relative improvement

Used features	**MSE**
VGG-net	0.0758
Ours	0.0687

Ablation Studies



Wichers et al. 2018

Conclusion

We saw:

- an unsupervised approach for discovering high-level features
- a joint training strategy
- long-term pixel-level video prediction for about 20 seconds

Hierarchical Long-term Video Prediction without Supervision ICML 2018

Nevan Wichers, Ruben Villegas, Dumitru Erhan, Honglak Lee

Seminar: Computer Vision SS19

Stefan Möhrle

References

- E. Denton and R. Fergus. *Stochastic video generation with a learned prior*. In ICML, 2018.
- C. Finn, I. Goodfellow, and S. Levine. *Unsupervised learning for physical interaction through video prediction.* In NIPS, 2016.
- S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee. *Deep visual analogy-making*. In NIPS. 2015.
- R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. *Learning to generate long-term future via hierarchical prediction.* In ICML, 2017.
- N. Wichers, R. Villegas, D. Erhan, H. Lee. *Hierarchical Long-term Video Prediction without Supervision.* In ICML, 2018.