

# Video Prediction

Xiaolong Wang

# This Class

- Video Prediction Background
- Interaction Network for Physical Prediction
- Prediction Space and Time

# Video Prediction Background

# Visual Prediction

- Given a (sequence of) past observations, predict future observations
- “Observations” can be many different things and used for different applications

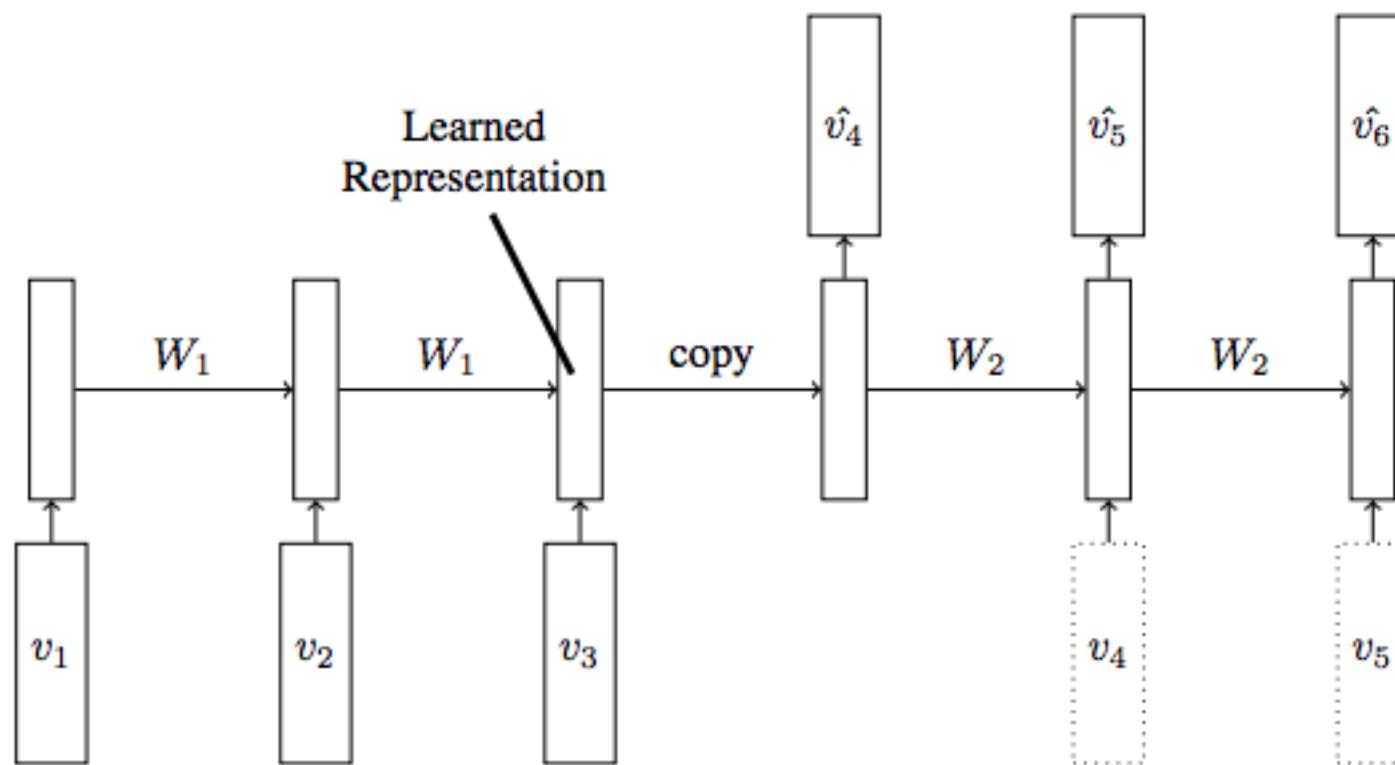
# Why Prediction?

*If an organism carries a model of external reality and its own possible actions within its head, it is able to react in much fuller, safer and more competent manner to emergencies which face it.*

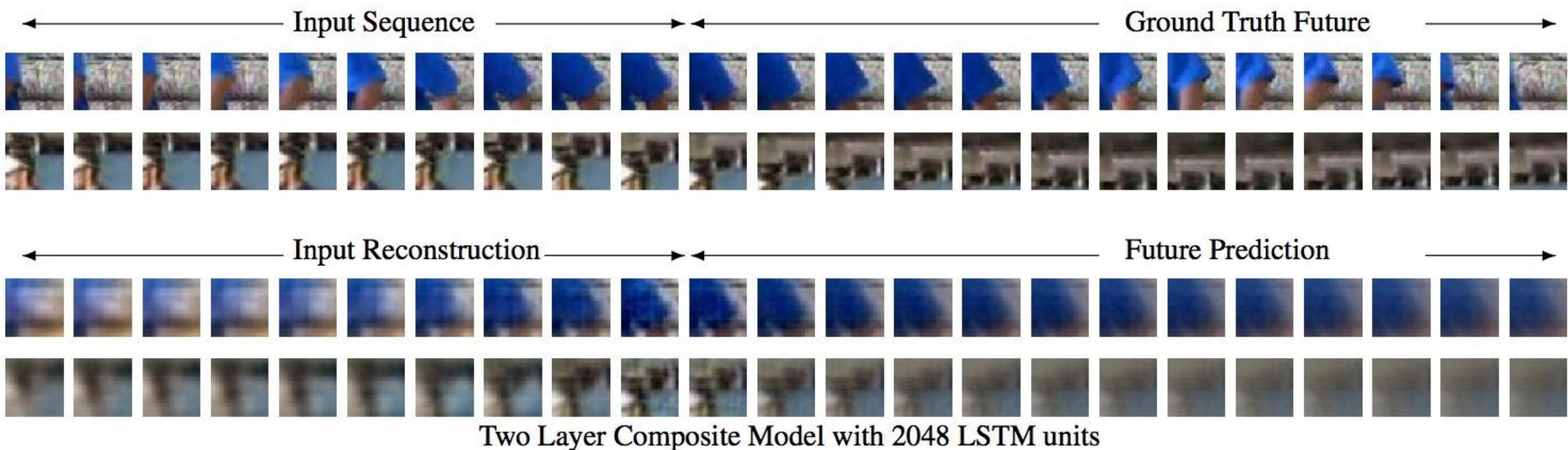
-- Kenneth Craik, in ``The nature of explanation''

- Model-based Planning.
- Learning a deep network provides a differentiable way to adjust the inputs.
- Representation Learning

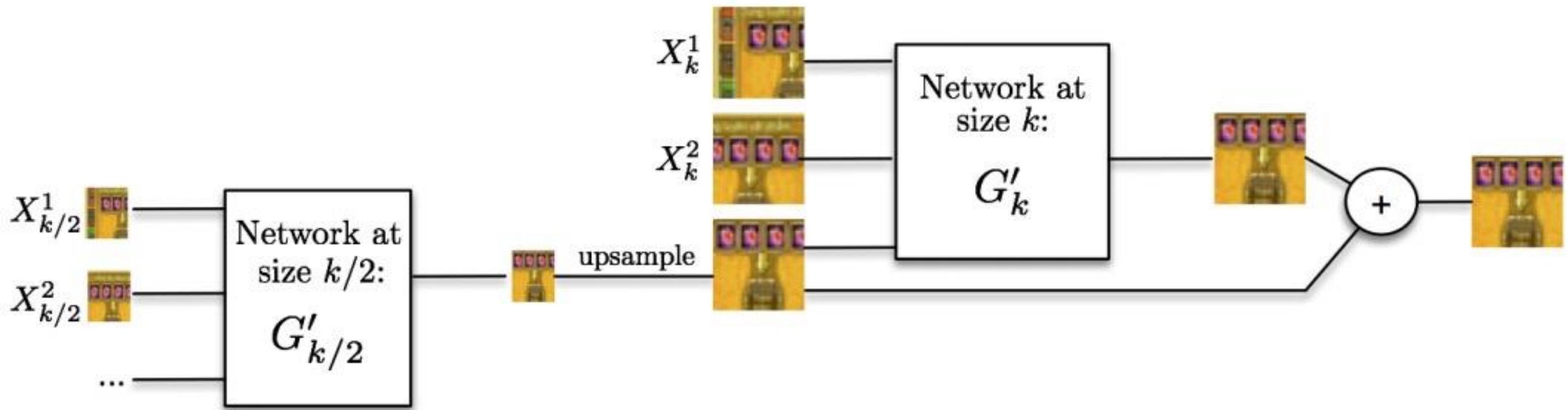
# Visual Prediction in Time



# Visual Prediction in Time



# Visual Prediction in Time



# Visual Prediction in Time

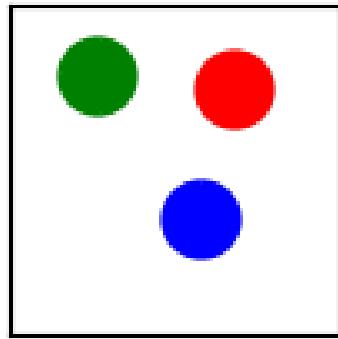


# Visual Prediction in Time

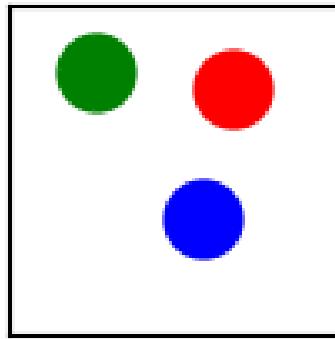
- Not a well-defined problem
- Pixel output space is too large
- Future has a large uncertainty

# Interaction Network for Physical Prediction

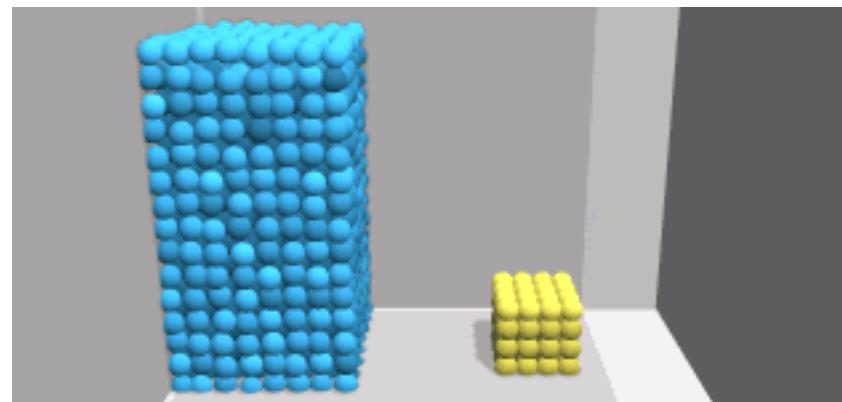
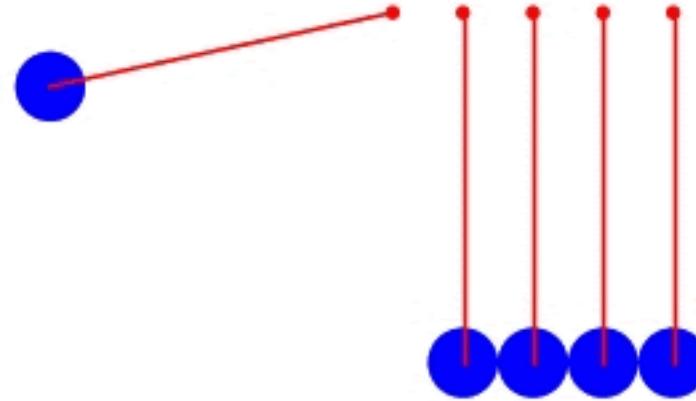
# Object Centric Prediction in a Physical World



Testdata

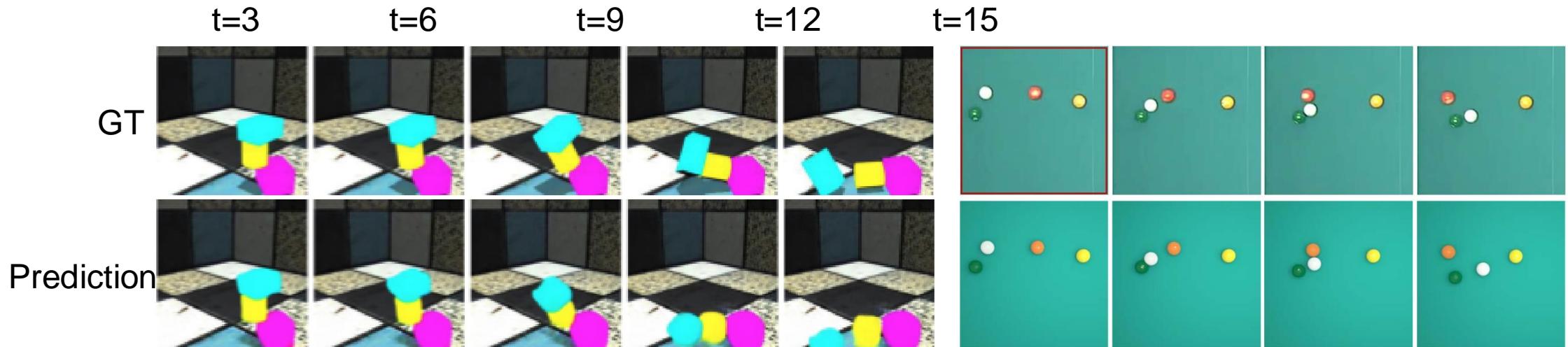


Model Prediction



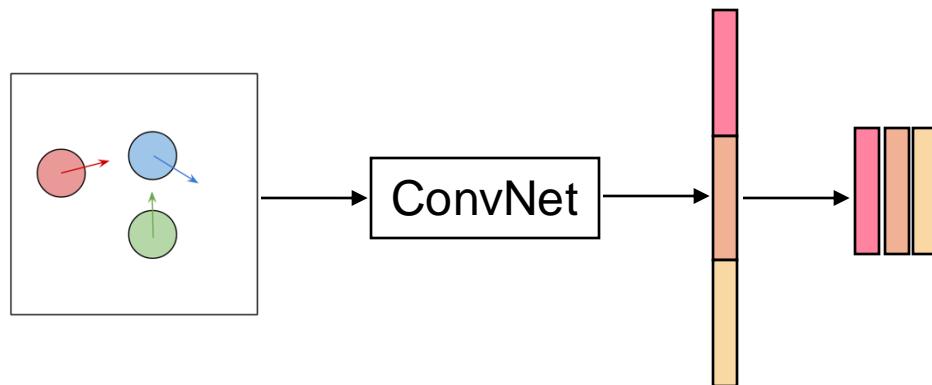
# Current Status of object dynamics prediction

- Prediction from pixels:
  - Not effectively encode object and context features
  - Not aim to do long term prediction (MPC approach)
- We hypothesis the bottleneck is the features of objects.



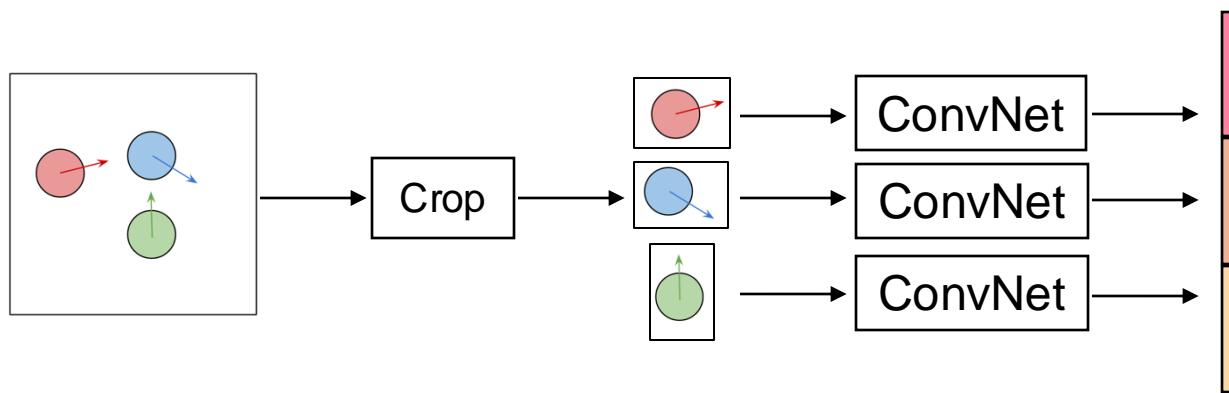
# Interaction Network

- Visual Interaction Network [1]: Use ConvNet to extract (#obj x 128) feature channels from multiple images.
  - Not very intuitive and cannot generalize to multiple objects
  - Input order is fixed so cannot generalize to multiple appearance



# Interaction Network

- Visual Interaction Network [1]: Use ConvNet to extract (#obj x 128) features from multiple images.
- Compositional Video Prediction [2,3]: Crop image by RoI and then pass through a ConvNet to get features.



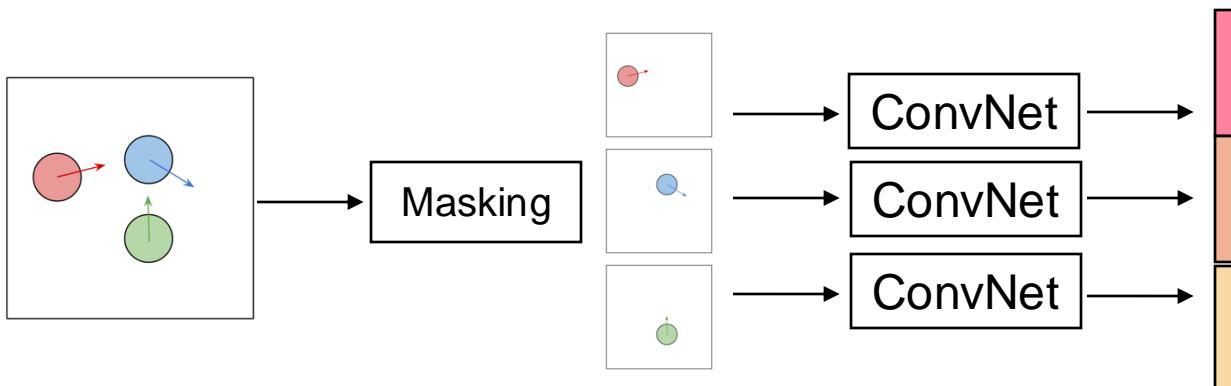
[1] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, A. Tacchetti. "Visual Interaction Networks". NIPS 2017

[2] Y. Ye, M. Singh, A. Gupta, S. Tulsiani. "Compositional Video Prediction". ICCV 2019

[3] Y. Ye, D. Gandhi, A. Gupta, S. Tulsiani. "Object-centric Forward Modeling for Model Predictive Control". CoRL 2019

# Interaction Network

- Visual Interaction Network [1]: Use ConvNet to extract (#obj x 128) features from multiple images.
- Compositional Video Prediction [2,3]: Crop image by RoI and then pass through a ConvNet to get features.
- Masking Based Approach [4,5]



[1] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, A. Tacchetti. "Visual Interaction Networks". NIPS 2017

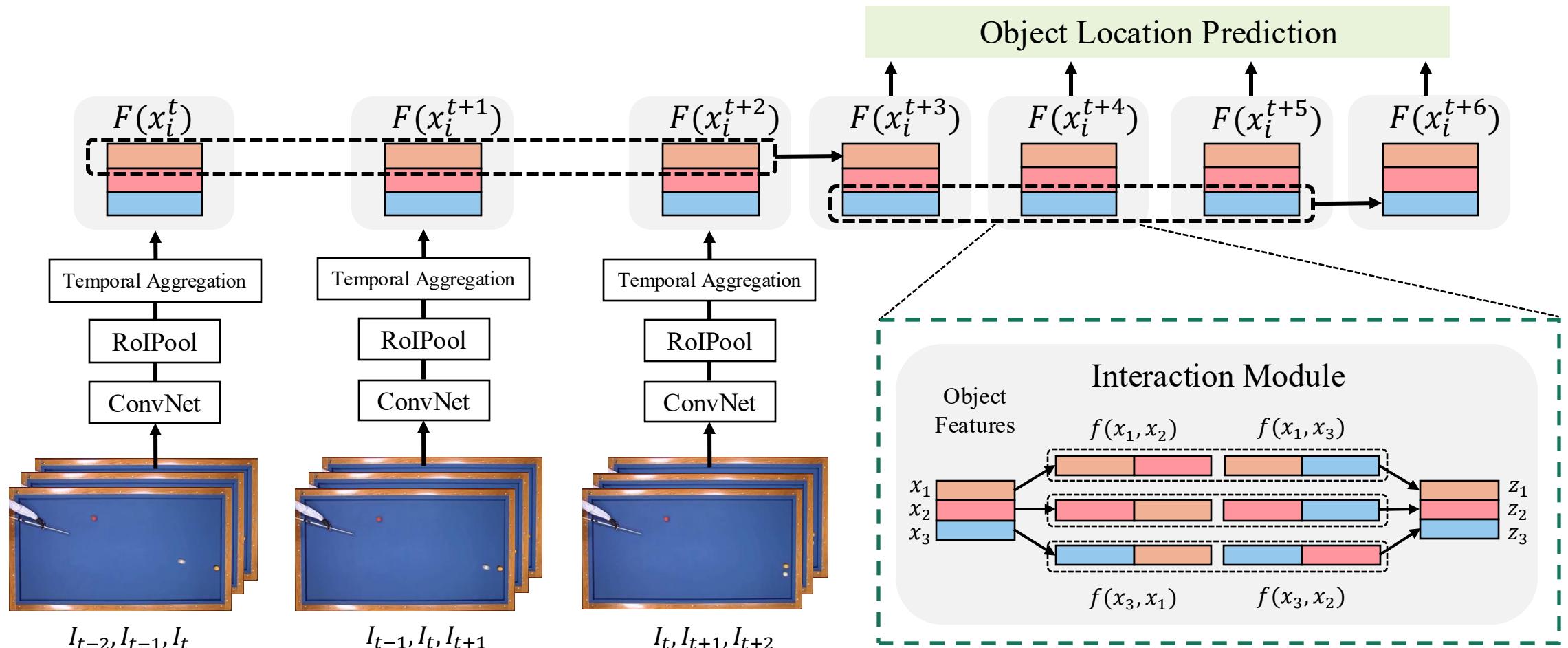
[2] Y. Ye, M. Singh, A. Gupta, S. Tulsiani. "Compositional Video Prediction". ICCV 2019

[3] Y. Ye, D. Gandhi, A. Gupta, S. Tulsiani. "Object-centric Forward Modeling for Model Predictive Control". CoRL 2019

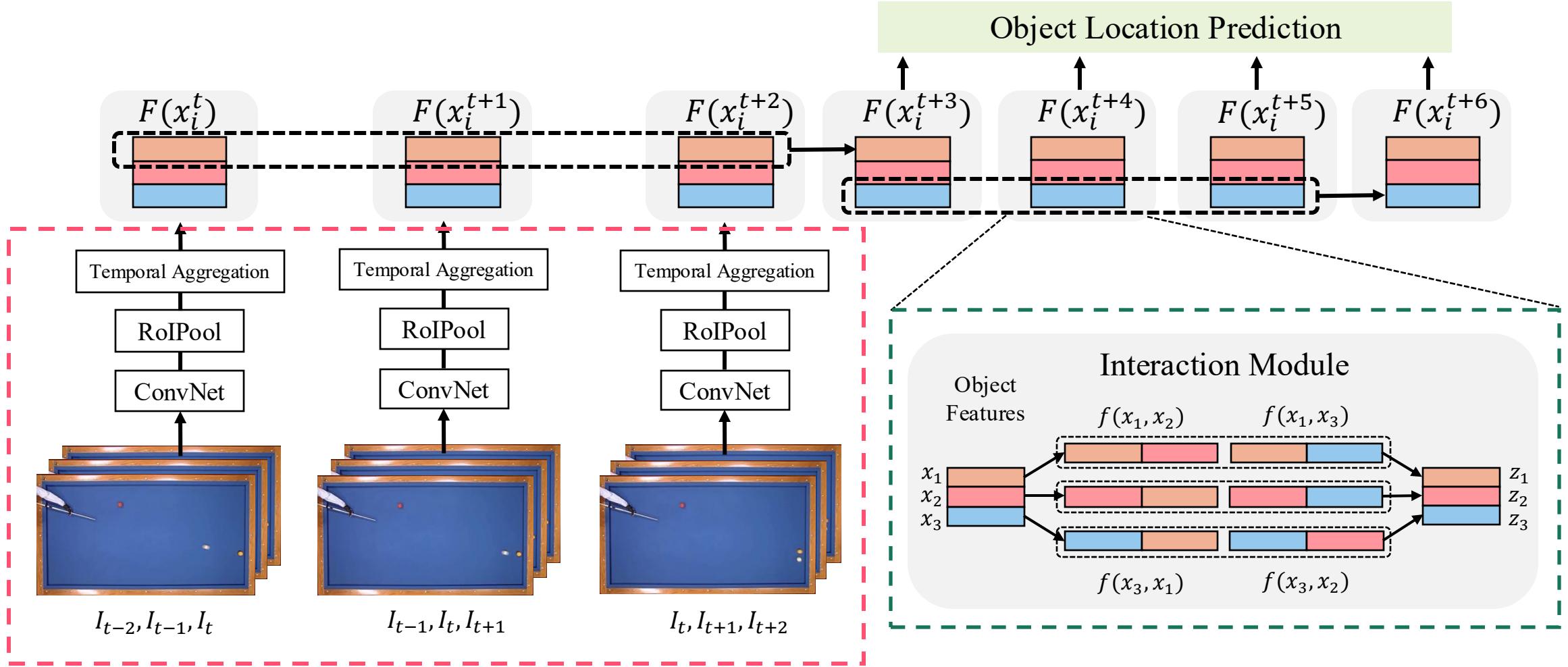
[4] Jiajun Wu, Erika Lu, Pushmeet Kohli, William T. Freeman, Joshua B. Tenenbaum. Learning to See Physics via Visual De-animation. In NIPS 2017

[5] M. Janner, S. Levine, W. Freeman, J. Tenenbaum, C. Finn, J. Wu. "Reasoning About Physical Interactions with object-oriented prediction and planning", ICLR 2019

# Region Proposal Interaction Networks



# Visual Encoder



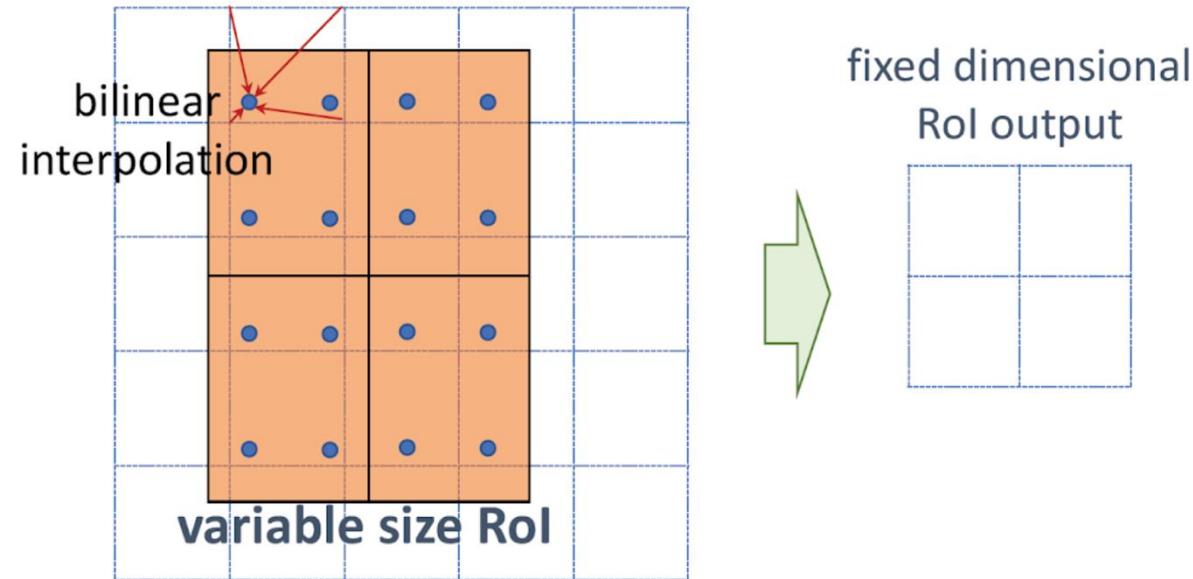
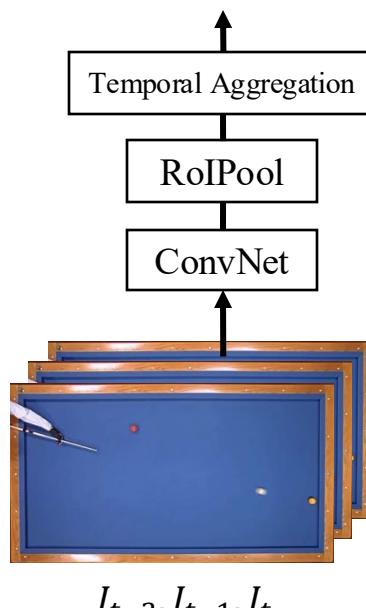
# Visual Encoder

- Object Centric Representation for Prediction
- We extract the state feature representations of  $n$  objects in time  $t$ , and predict their representations in time  $t + 1$ .

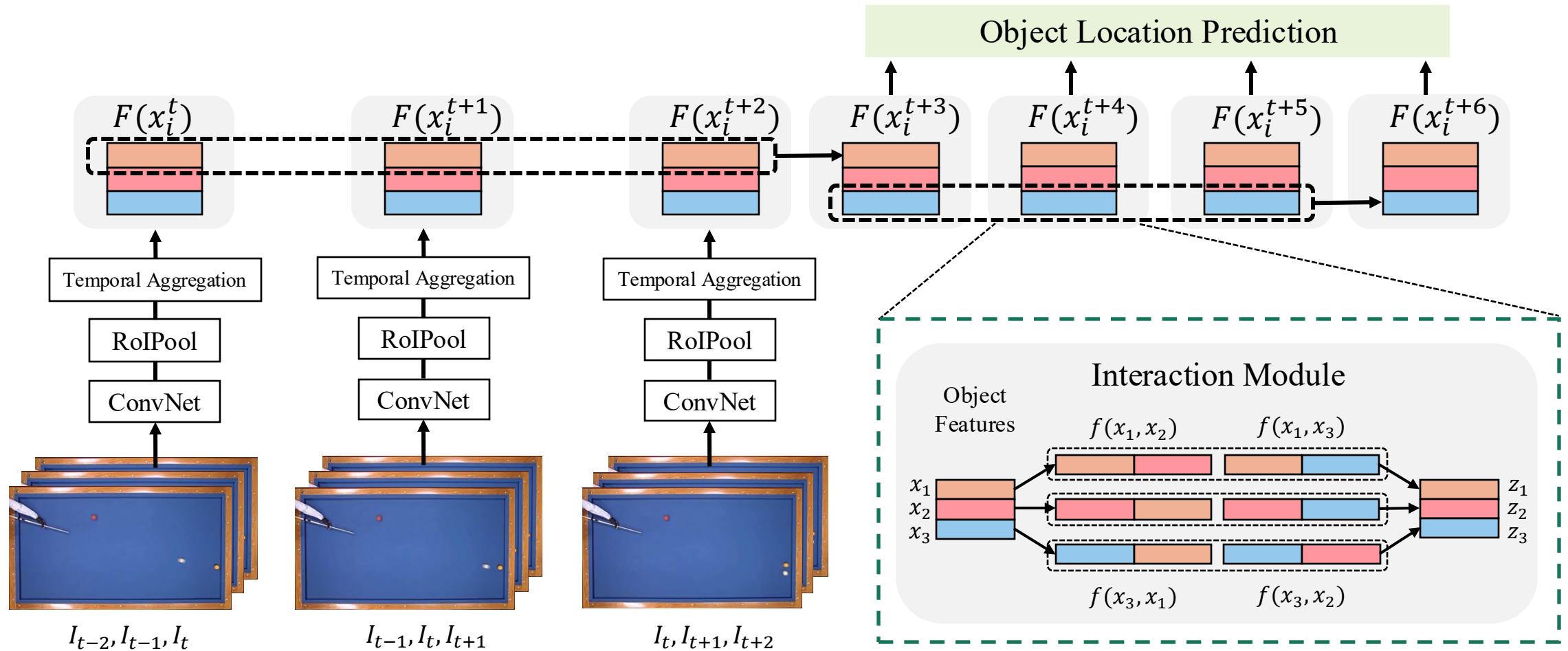
$$\{x_1^t, x_2^t, \dots, x_n^t\} \rightarrow \{x_1^{t+1}, x_2^{t+1}, \dots, x_n^{t+1}\}$$

# Visual Encoder

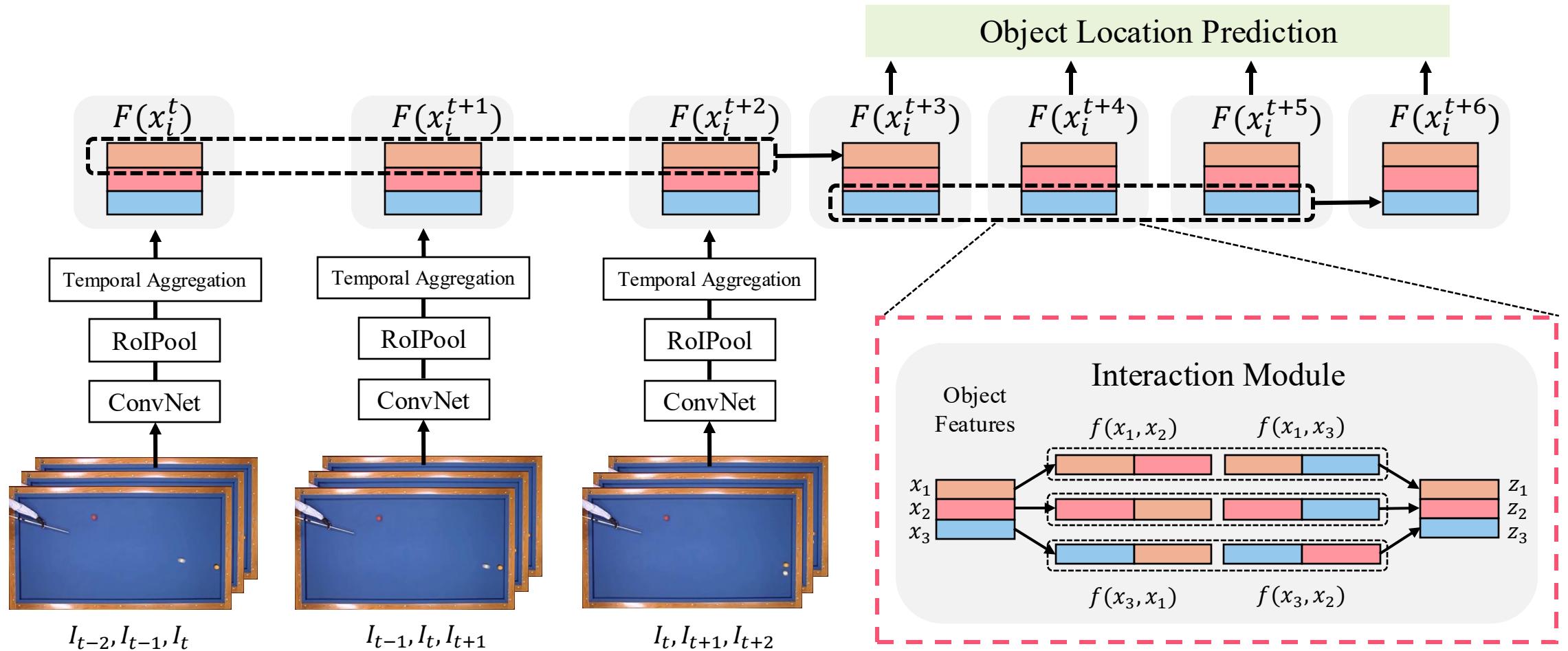
- Use hourglass network to extract image features
- Use aligned RoI Pooling to extract region features



# Interaction Module in feature space



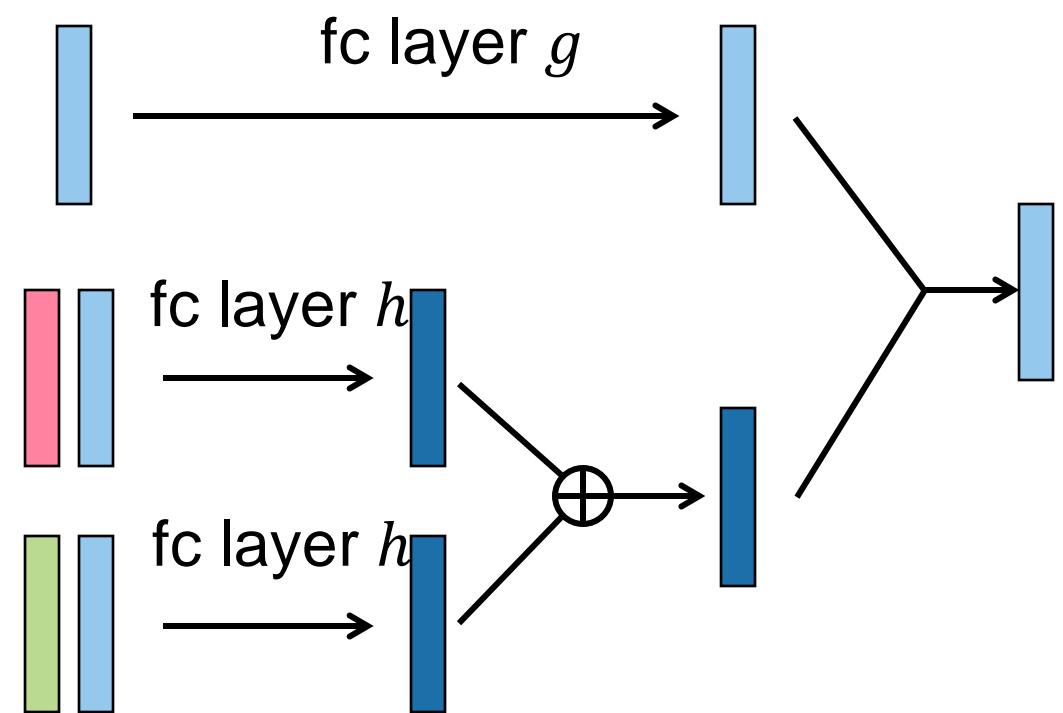
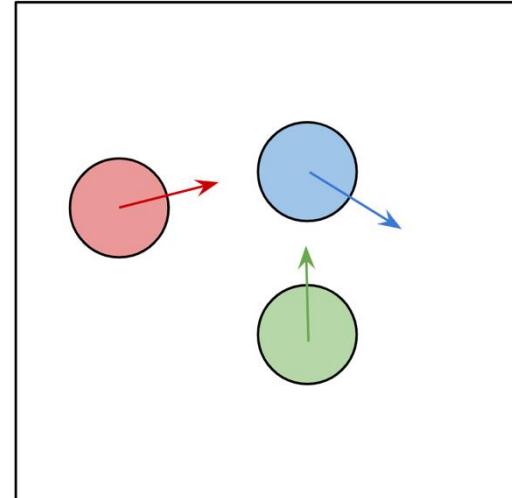
# Interaction Module in feature space



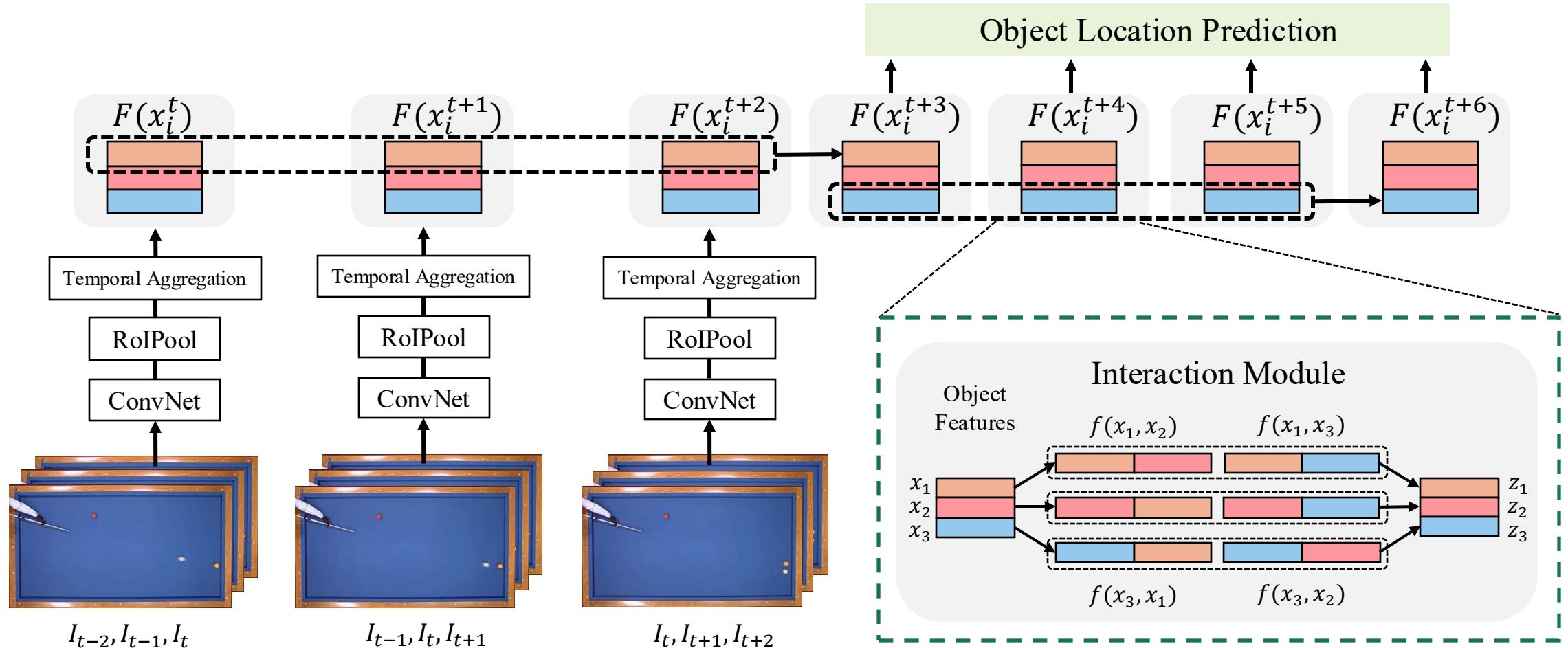
# Interaction Module

If we want to predict the future movement  
of the blue billiard

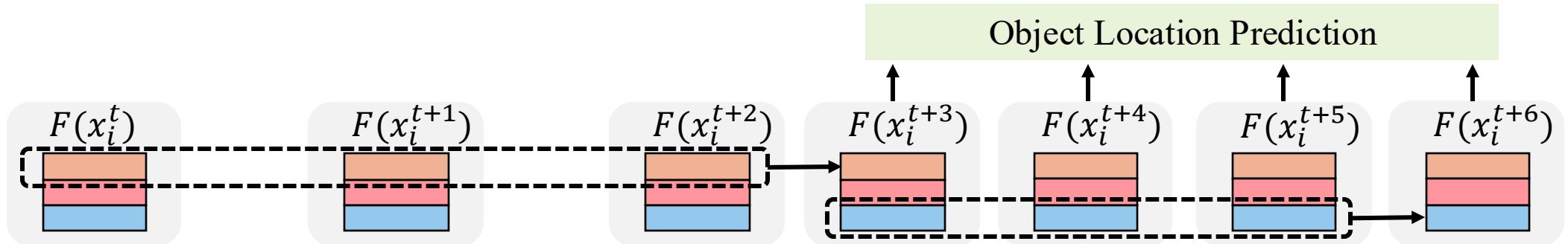
- self-dynamics: (Newton's first law)  
 $g(x_i^t)$
- relation-dynamics: (Newton's second law)  
 $\sum_{j \neq i} h(x_i^t, x_j^t)$
- Aggregate the above:  
 $F(x_i^t) = f(g(x_i^t), \sum_{j \neq i} h(x_i^t, x_j^t))$



# Prediction



# Prediction

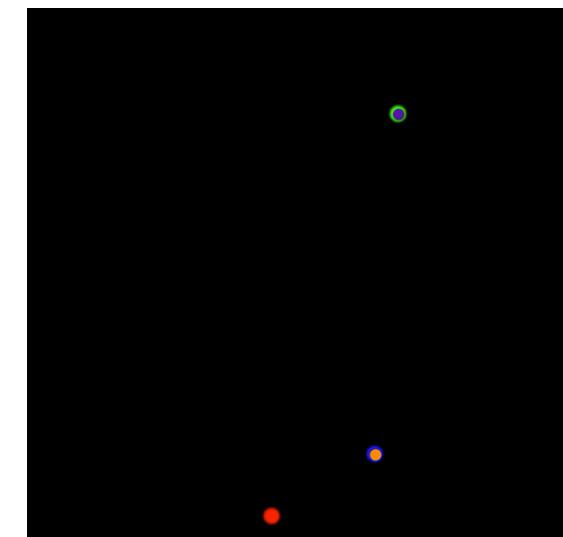
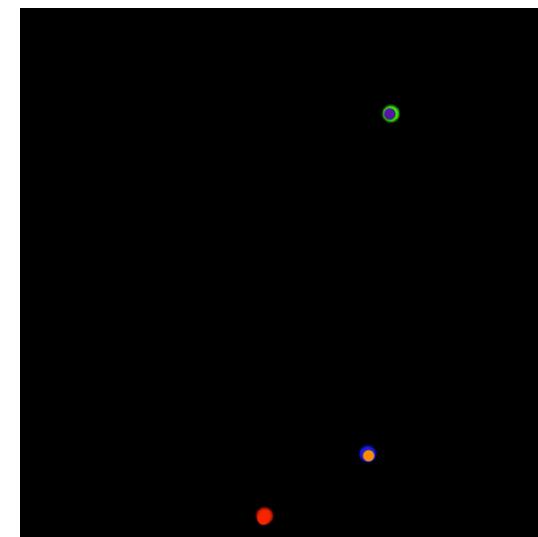
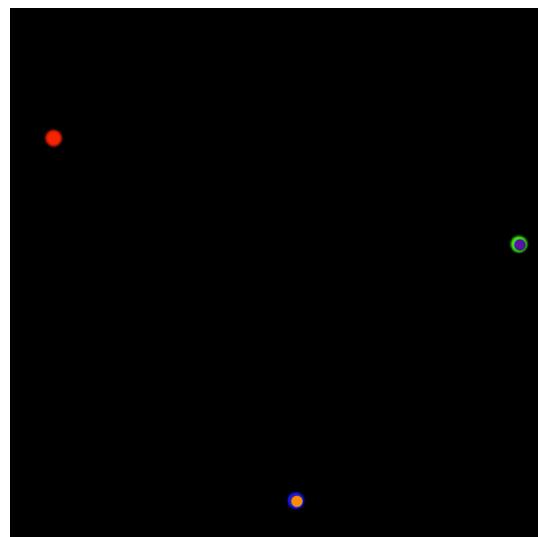
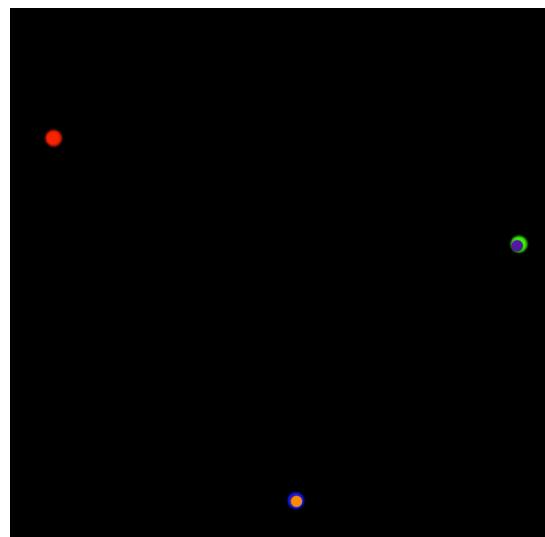
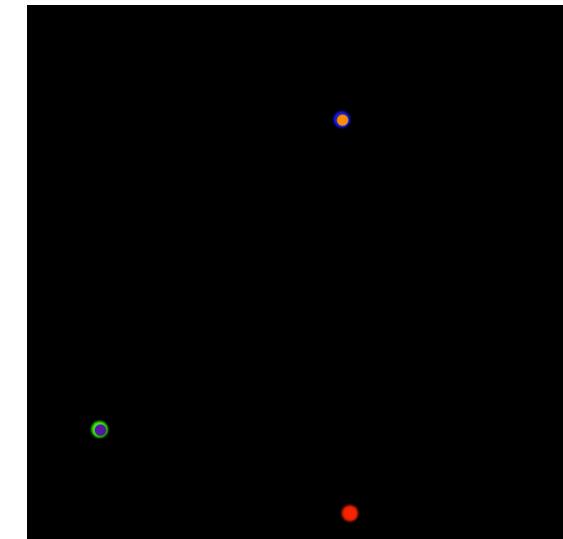
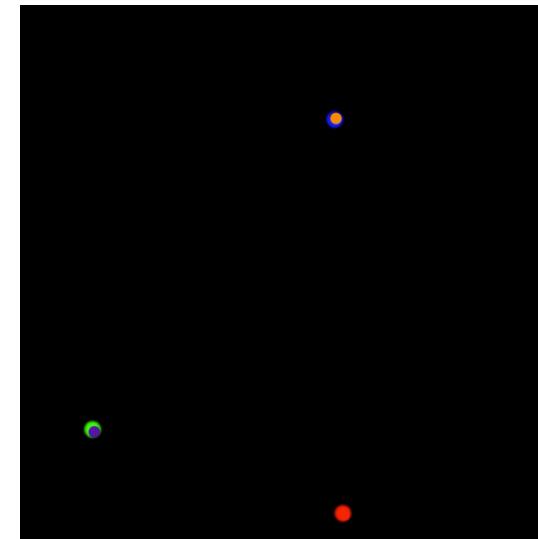
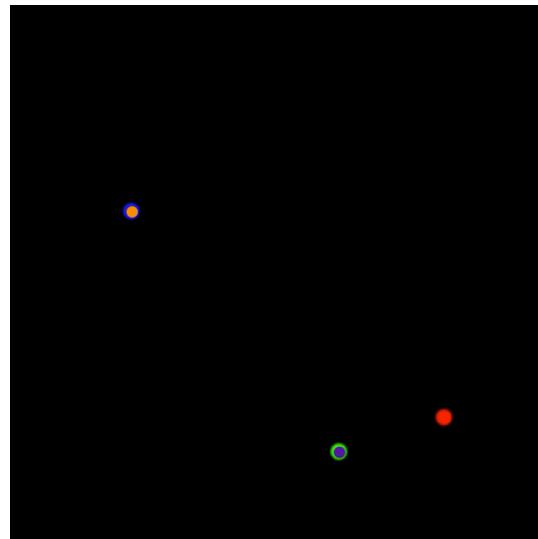
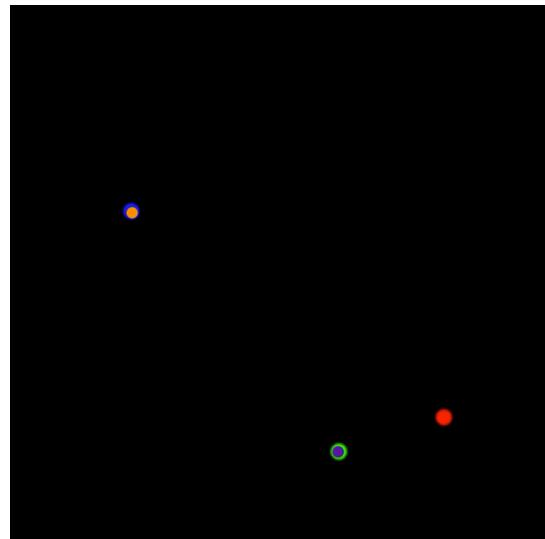


Future feature prediction:  $x_i^{t+1} = W_d[F(x_i^t), F(x_i^{t-1}), \dots, F(x_i^{t-k})]$

Location estimation:  $\hat{p}_i^{t+1} = W_p x_i^{t+1}$

Training loss function:  $L_p = \sum_{t=1}^T \sum_{i=1}^n \|\hat{p}_i^{t+1} - p_i^{t+1}\|_2^2$

# Simulation Billiards



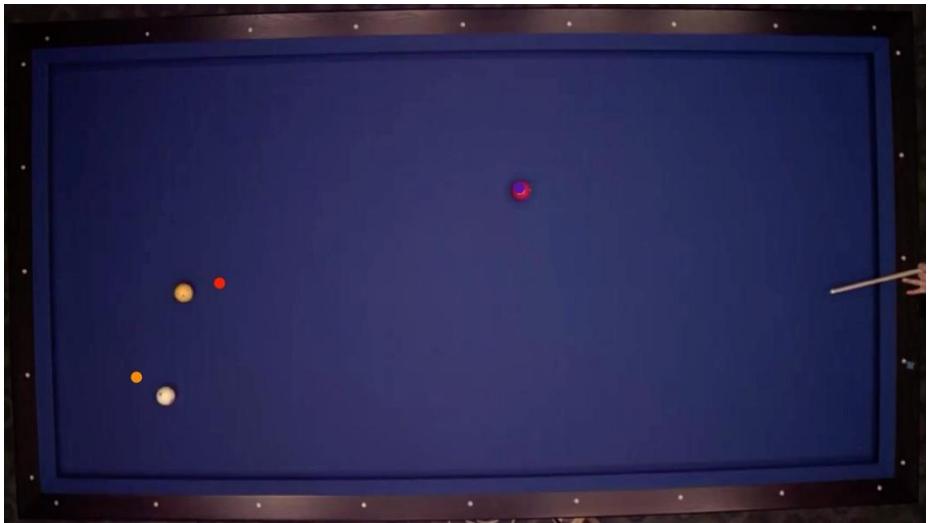
prediction

ground-truth

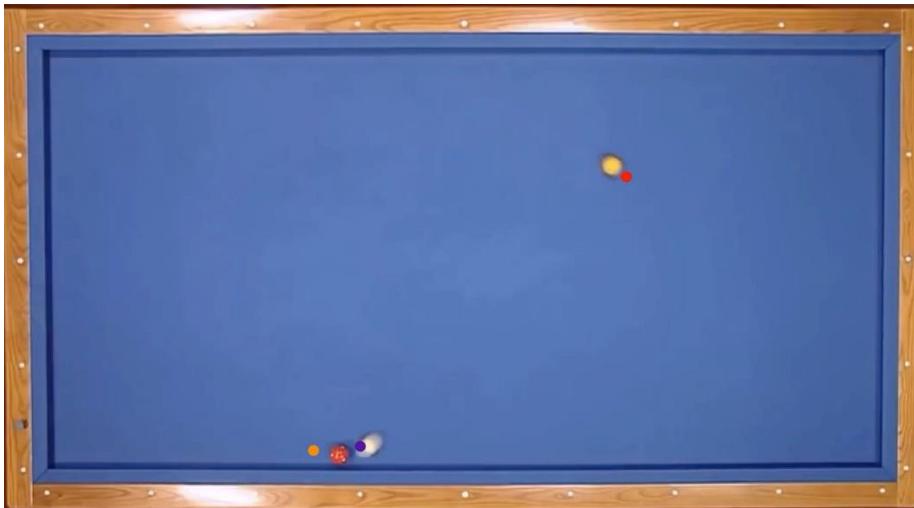
prediction

ground-truth

# Real Billiards

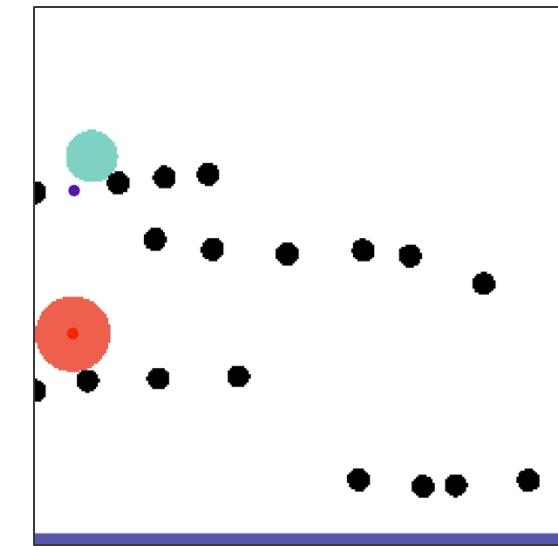
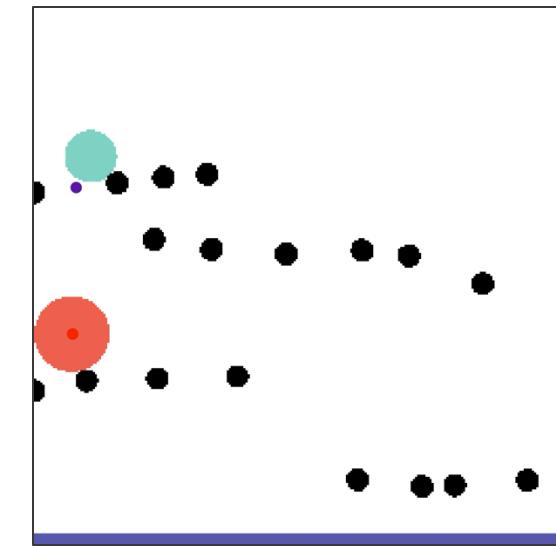
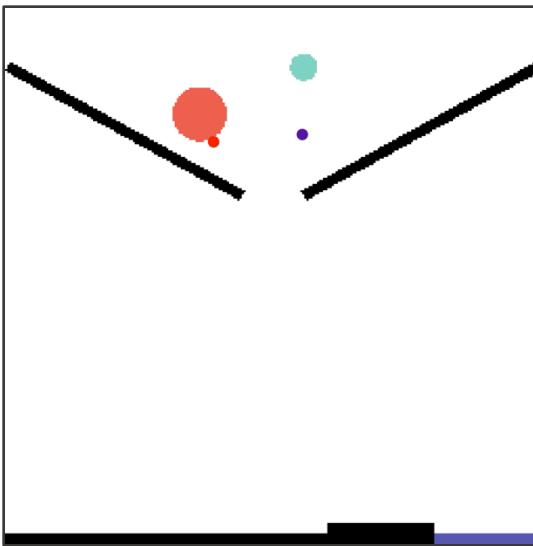
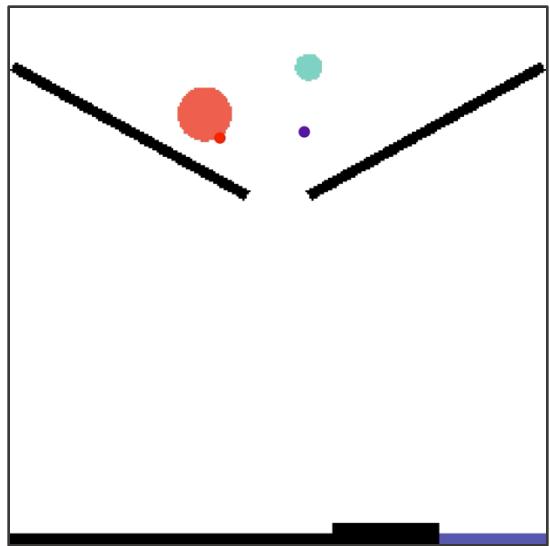
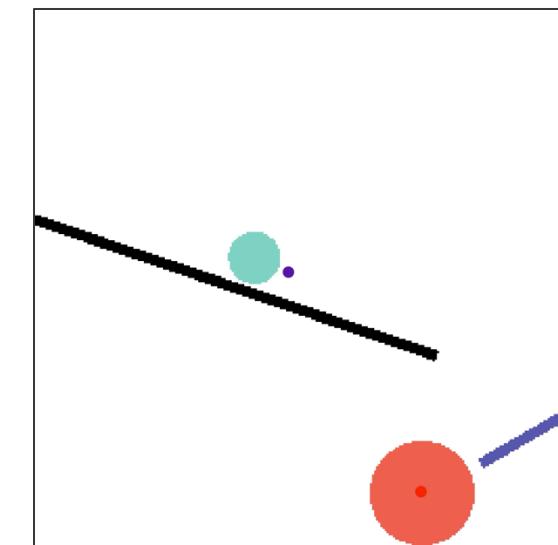
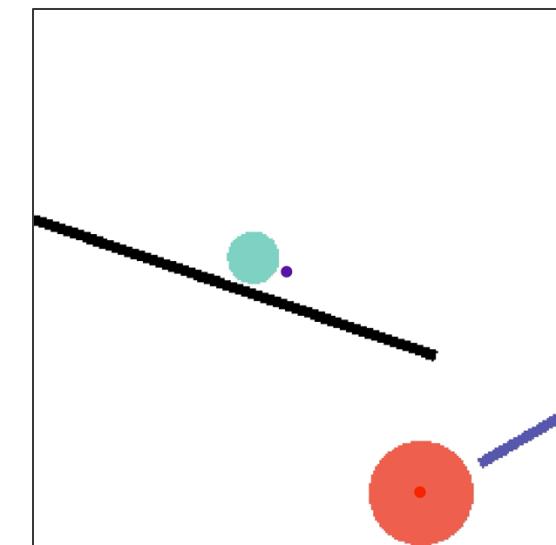
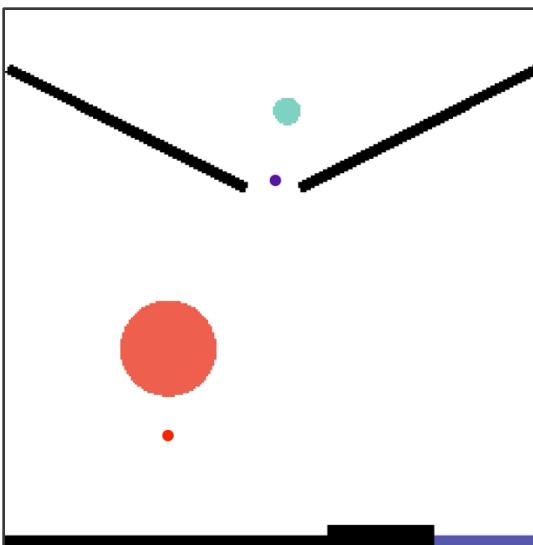
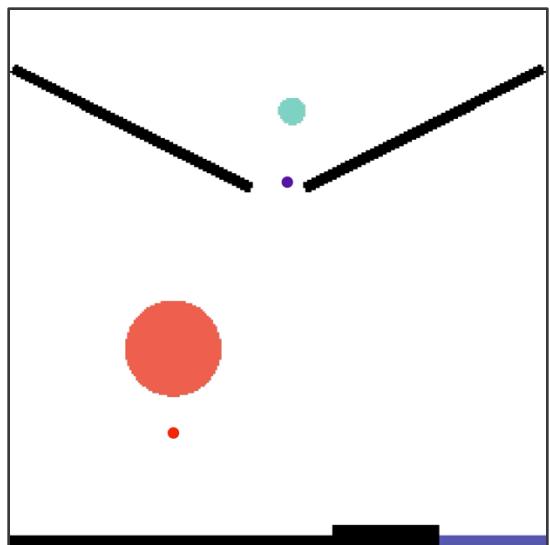


prediction



ground-truth

# PHYRE



prediction

ground-truth

prediction

ground-truth

# What Space to Predict

# What Space to Predict

Predict Optical Flow:

**An Uncertain Future: Forecasting from Static  
Images using Variational Autoencoders**

Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert

Predict Skeleton:

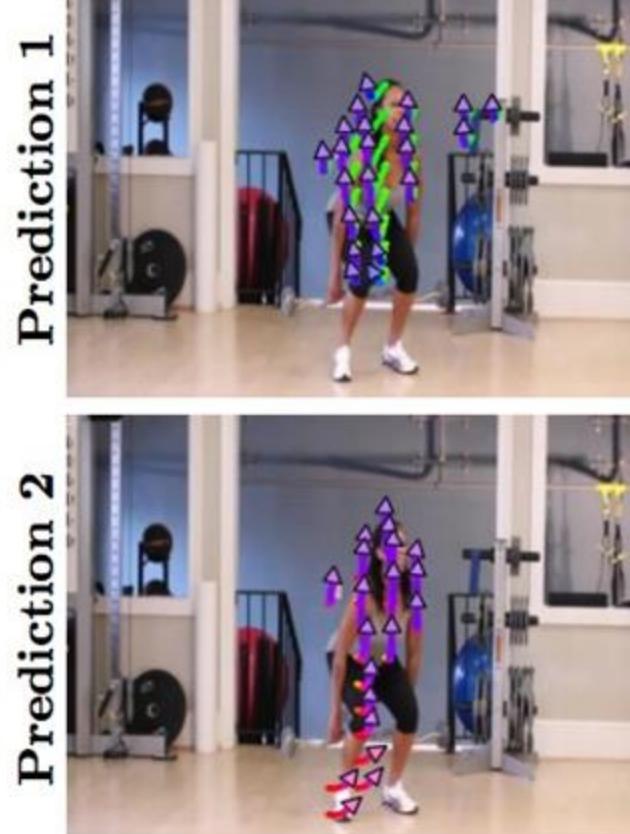
---

**Learning to Generate Long-term Future via Hierarchical Prediction**

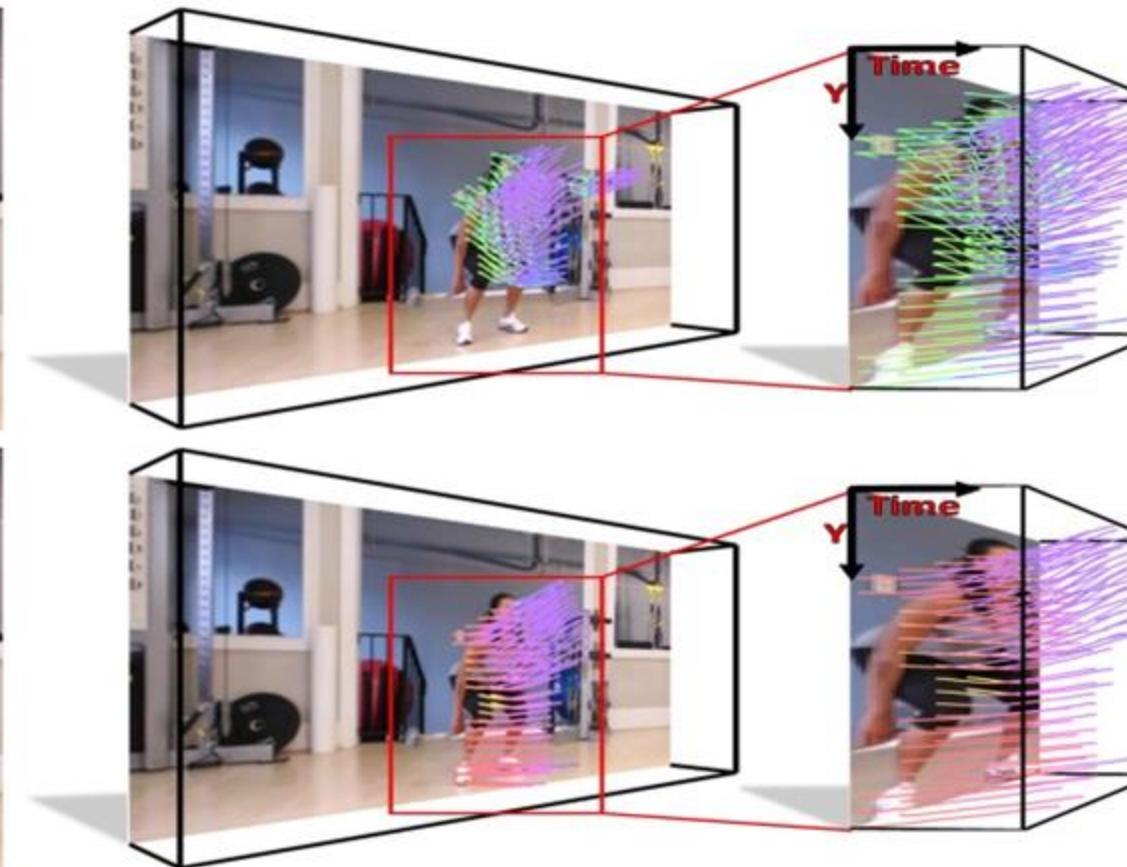
---

**Ruben Villegas<sup>1</sup>\*** **Jimei Yang<sup>2</sup>** **Yuliang Zou<sup>1</sup>** **Sungryull Sohn<sup>1</sup>** **Xunyu Lin<sup>3</sup>** **Honglak Lee<sup>1,4</sup>**

# Predict Future Optical Flow from A Single Image

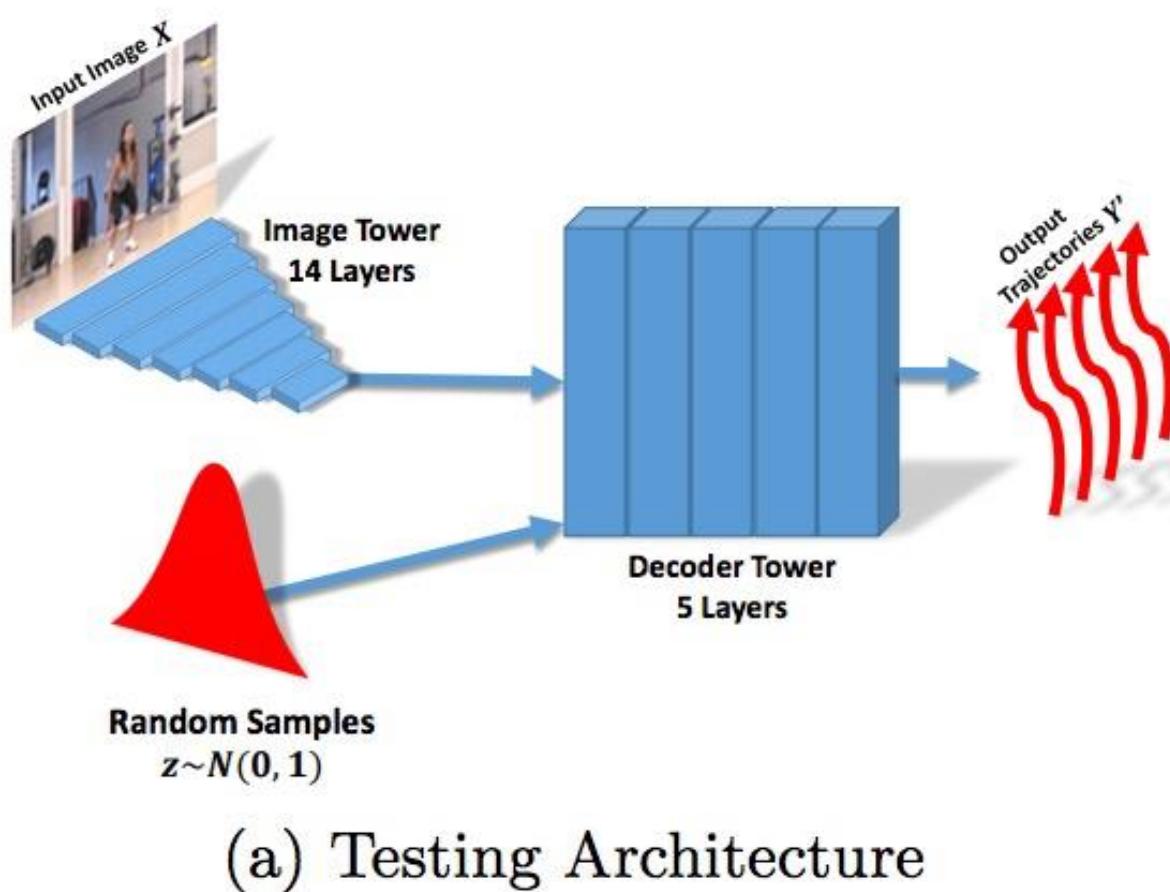


(a) Trajectories on Image



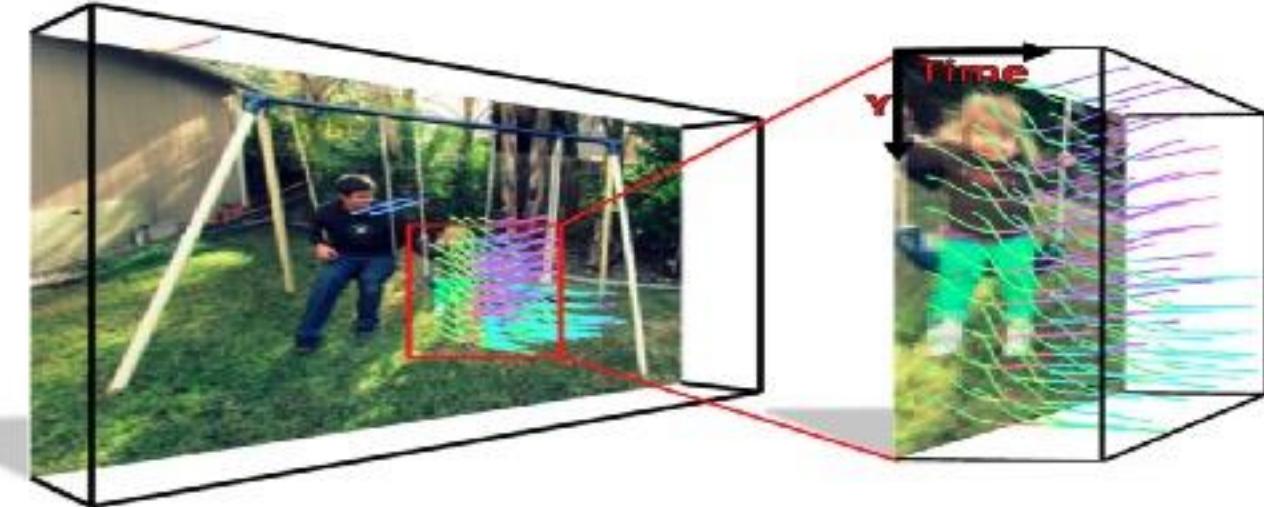
(b) Trajectories in Space-Time

# CVAE for Modeling Uncertainty

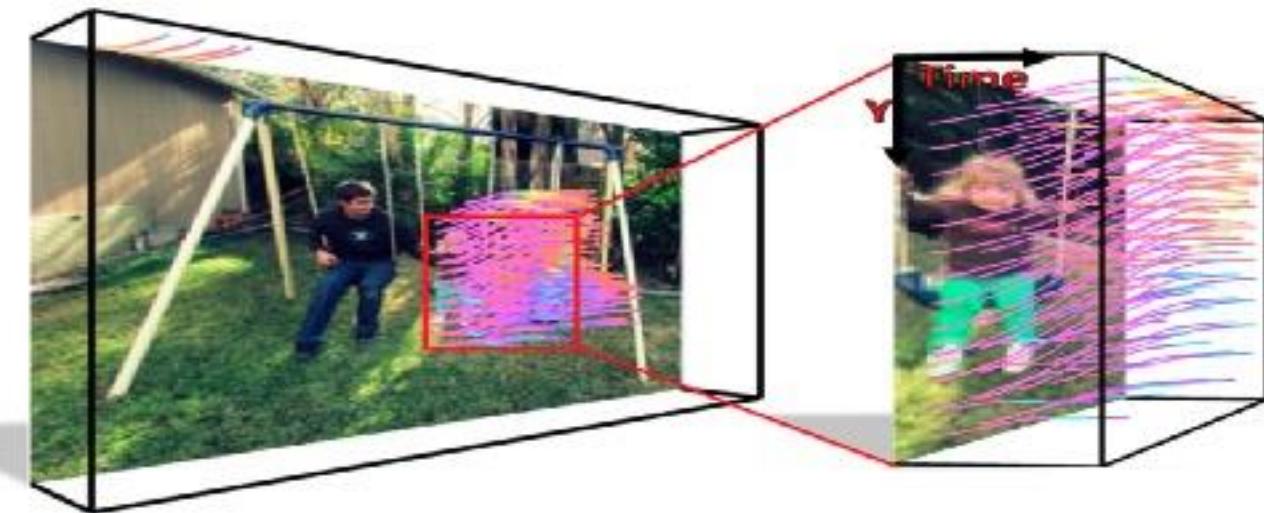


# Results

Prediction 1



Prediction 2

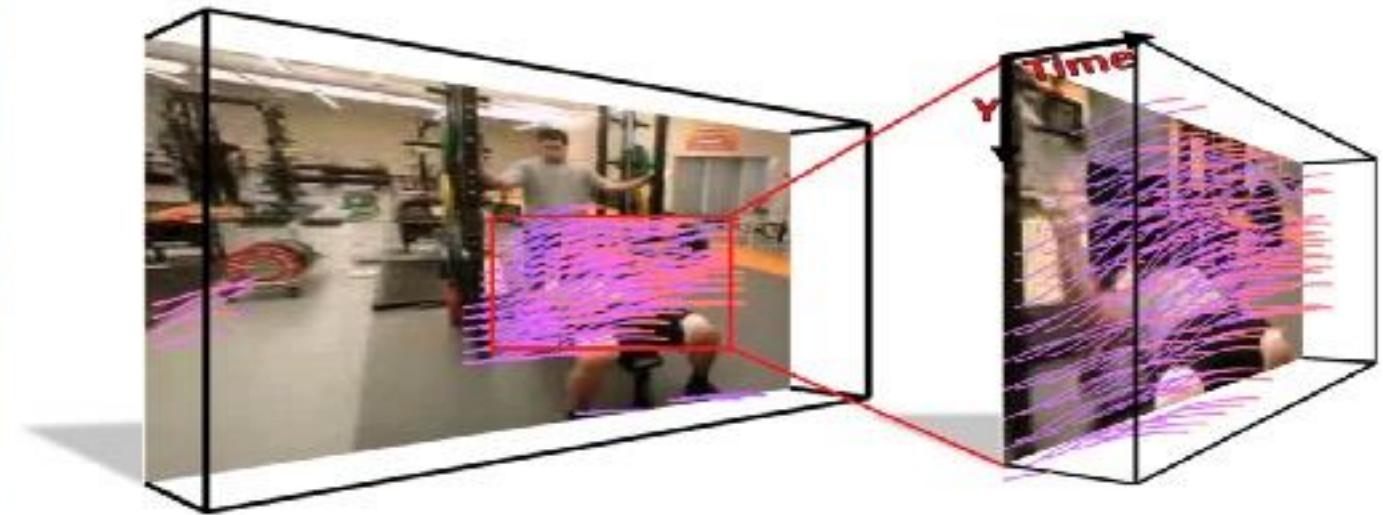
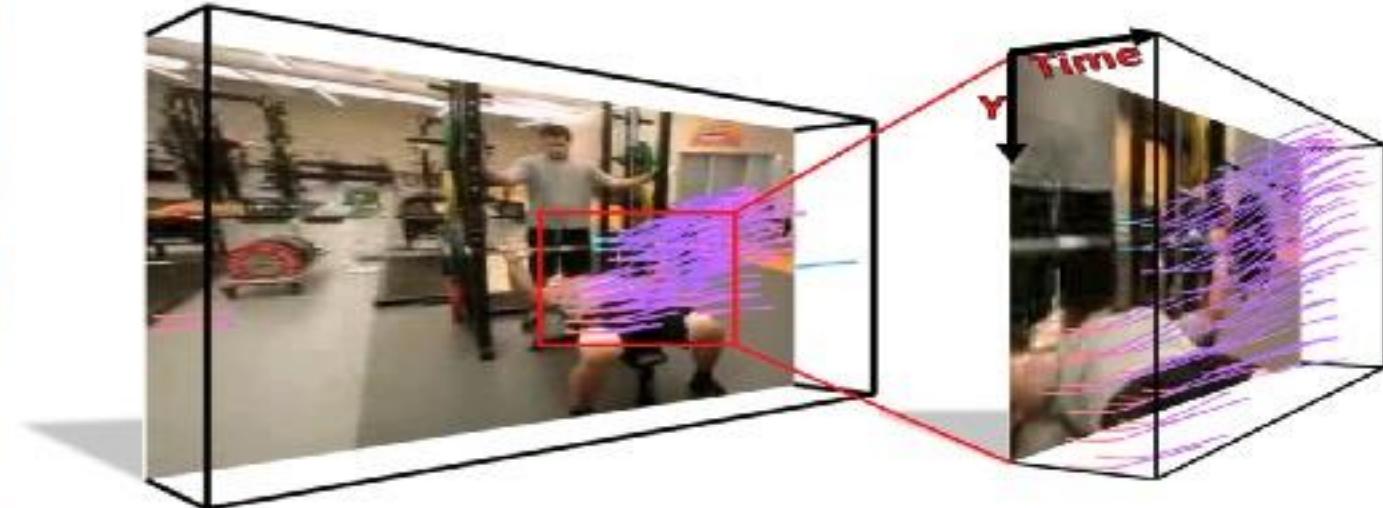


# Results

Prediction 1

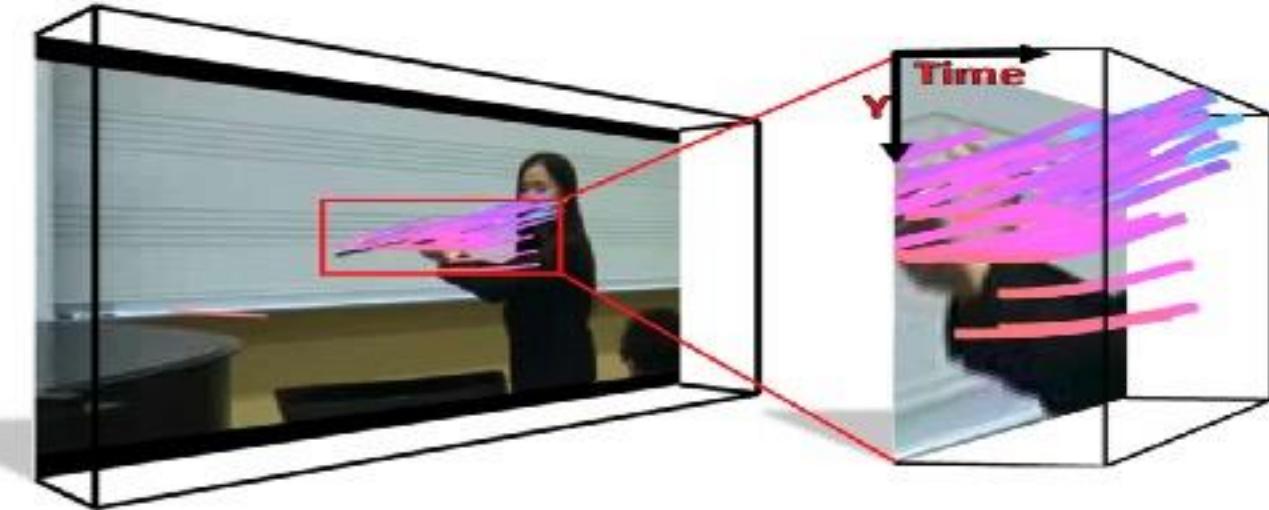
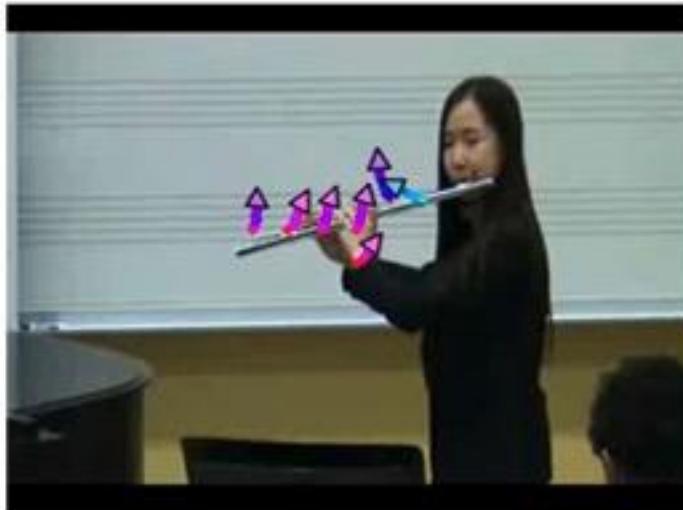


Prediction 2

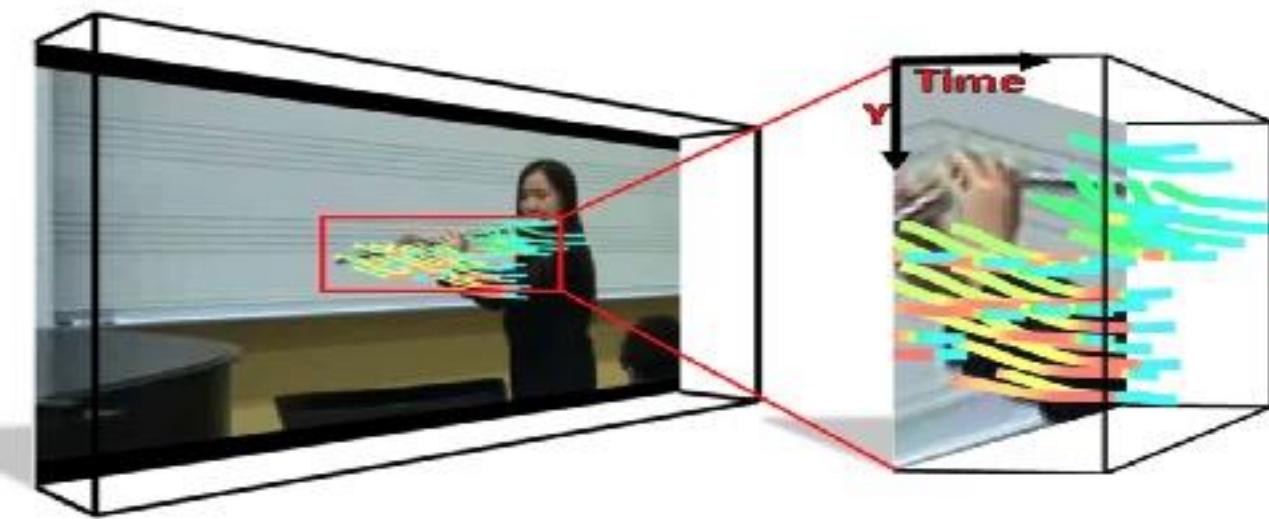
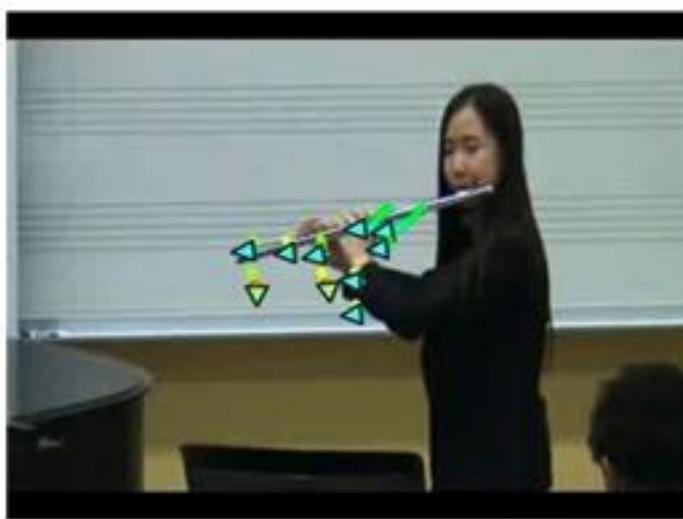


# Results

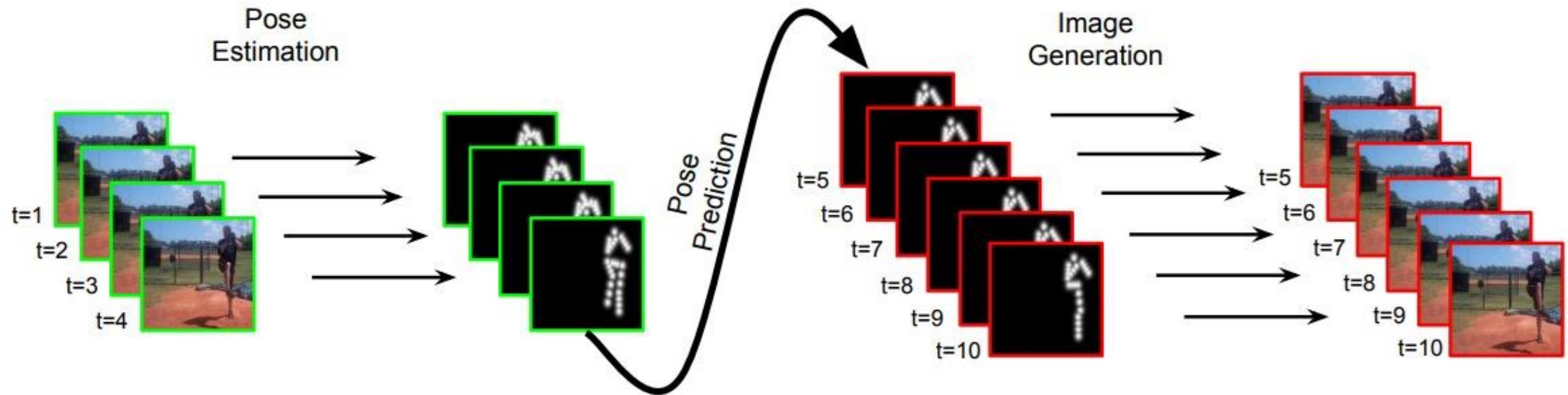
Prediction 1



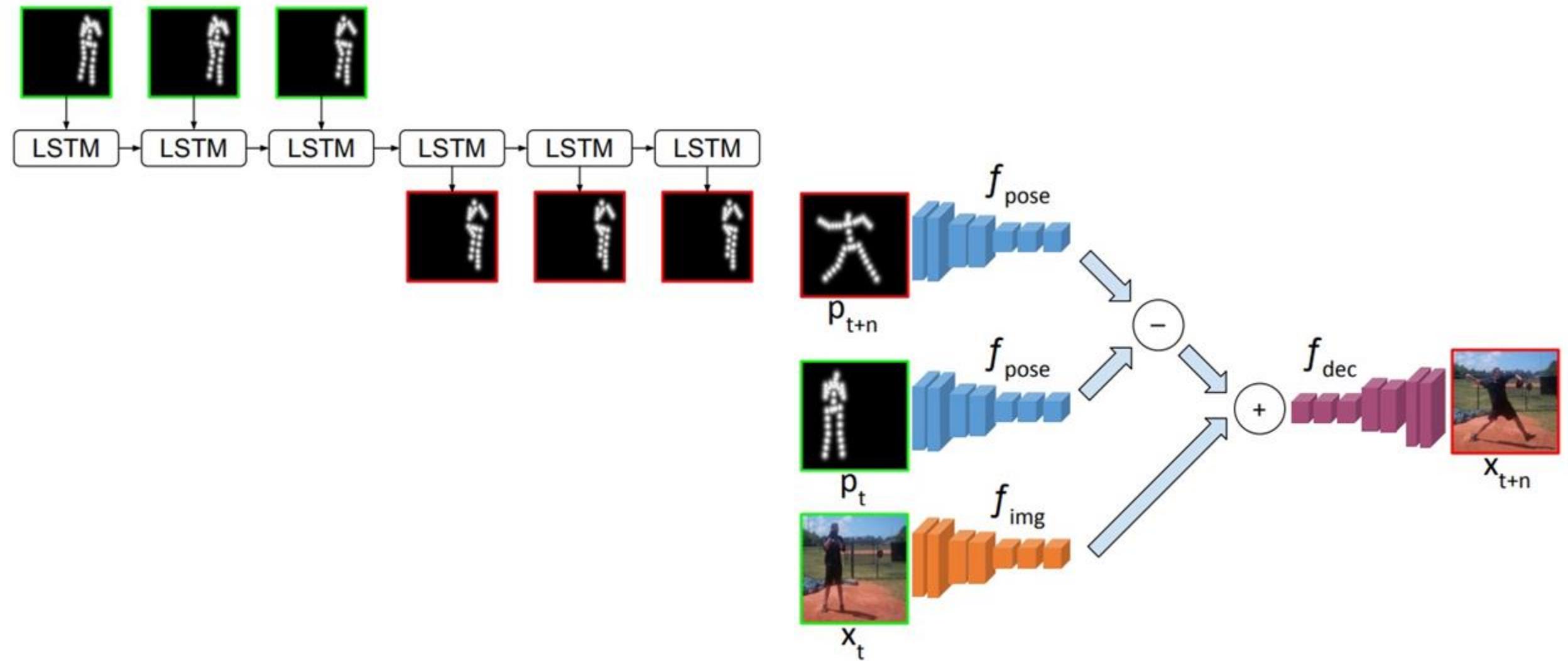
Prediction 2



# Predict Future Pose



# Method



# Results

0050\_baseball\_pitch

Ours  
 $t=1$



ConvLSTM  
 $t=1$



Optical flow  
 $t=1$



# Results

0721\_clean\_and\_jerk

Ours  
 $t=1$



ConvLSTM  
 $t=1$



Optical flow  
 $t=1$



# Results

2196\_tennis\_serve

Ours  
 $t=1$



ConvLSTM  
 $t=1$

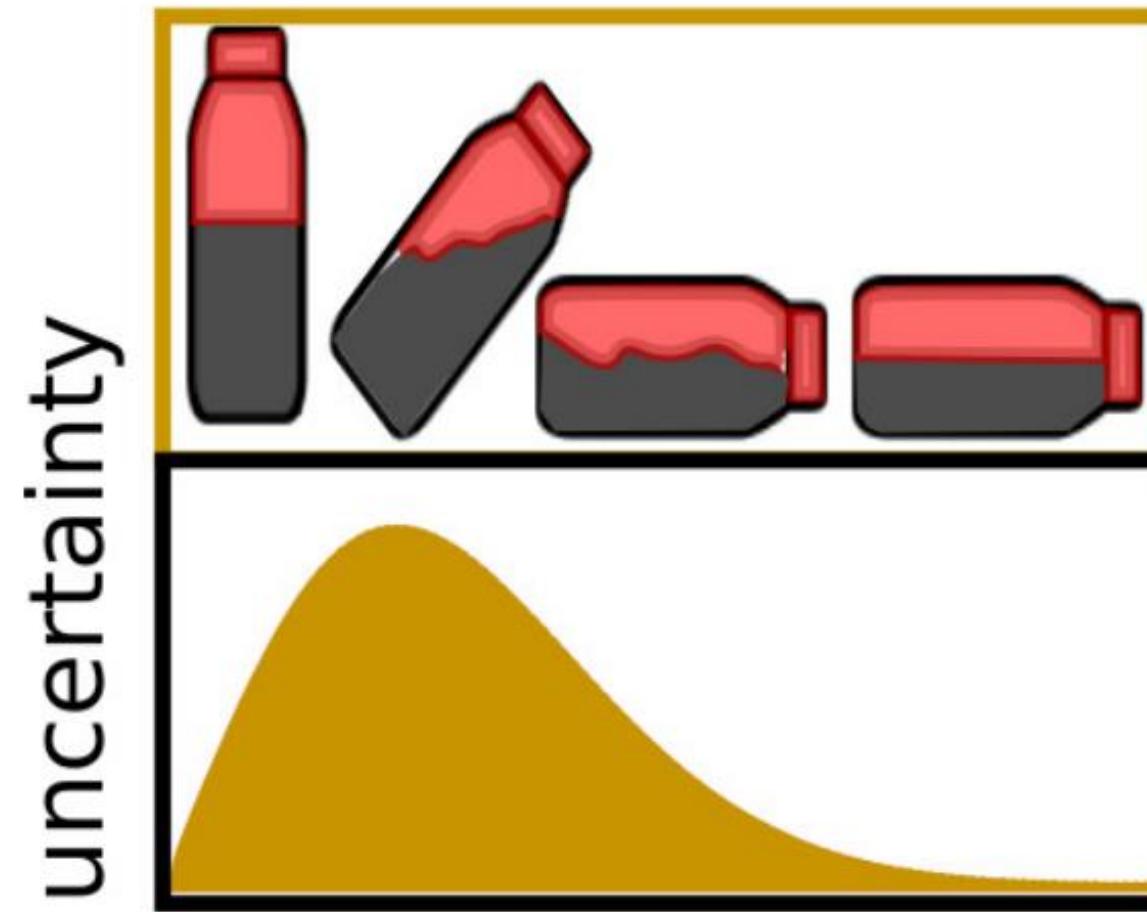


Optical flow  
 $t=1$

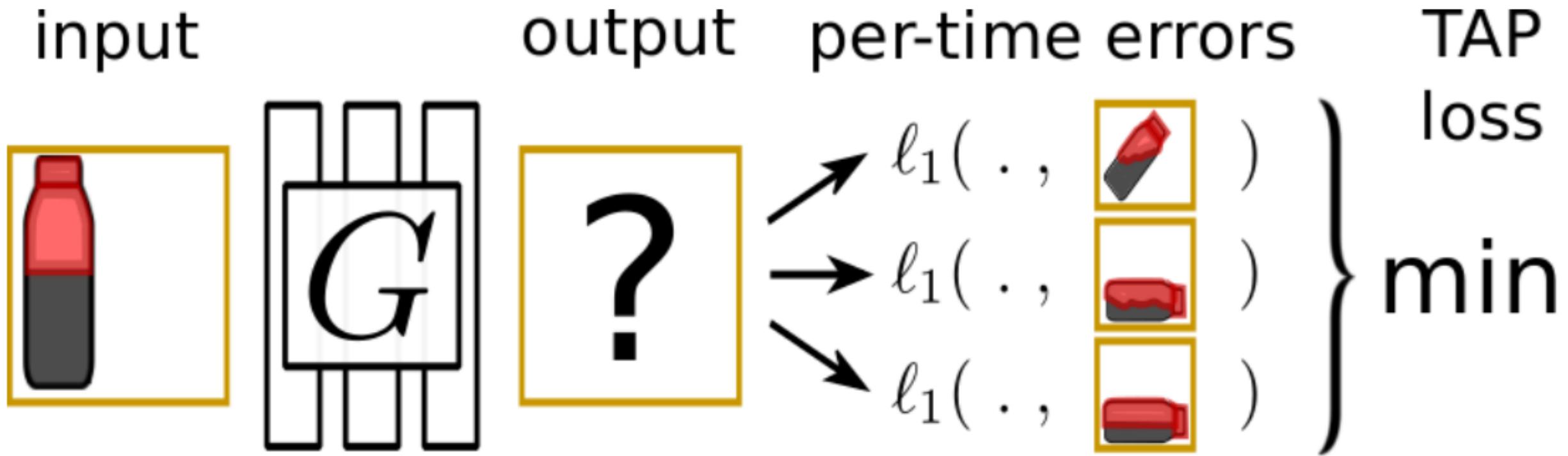


# What Time to Predict

# Uncertainty in Time



# Predict the Predictable Future



# Predict the Predictable Future

$$G^* = \arg \min_G \mathcal{L}(G) = \arg \min_G \min_{t \in T} \mathcal{E}(G(c), x_t)$$

Find a state with low uncertainty.

But it is unclear what exactly the T is in testing

# Next Class

Attention and transformer