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

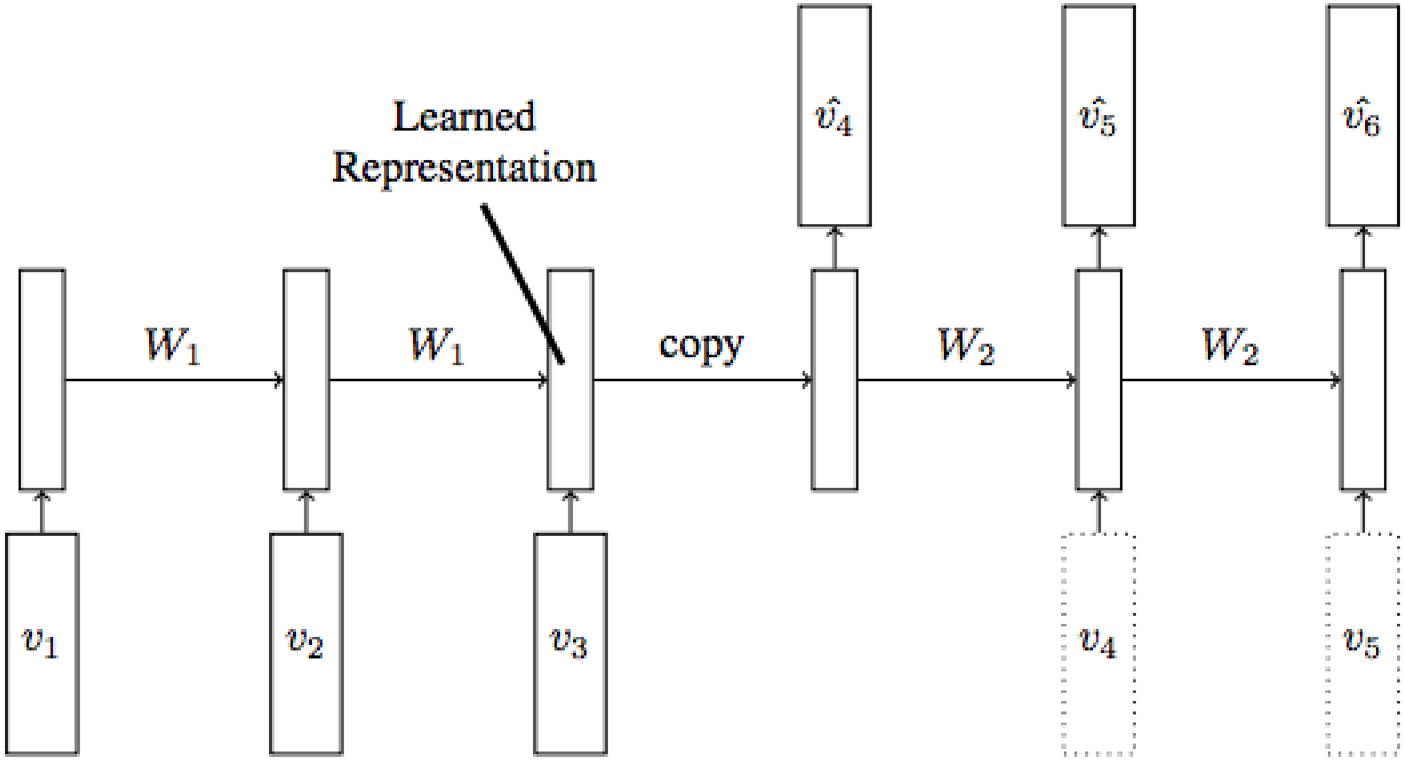
Why Prediction?

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”

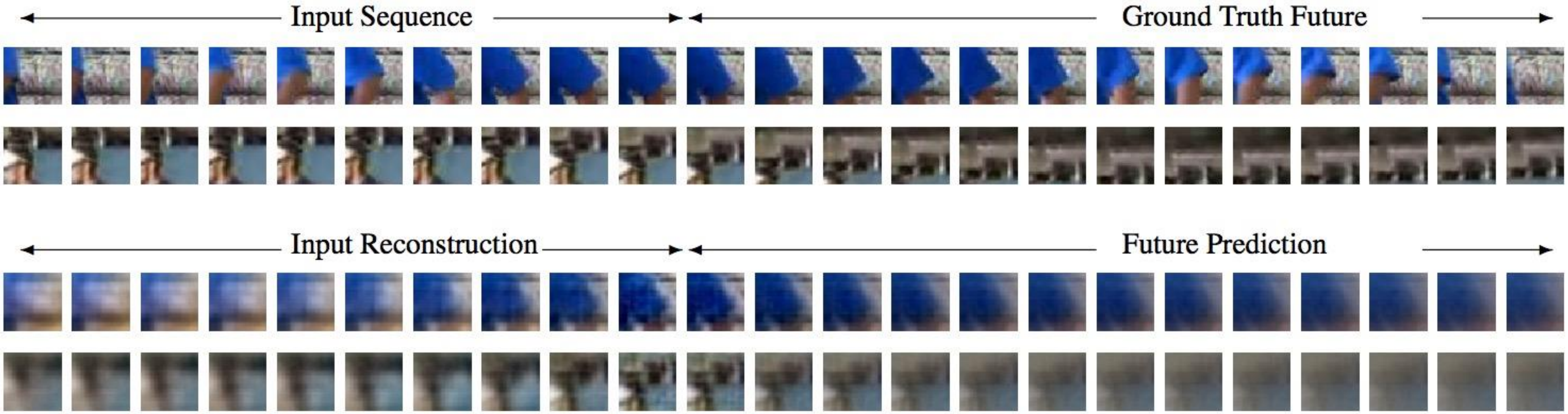
- Model-based Planning.
- Learning a deep network provides a differentiable way to adjust the inputs.
- Representation Learning

Visual Prediction in Time



Srivastava et al., 2015

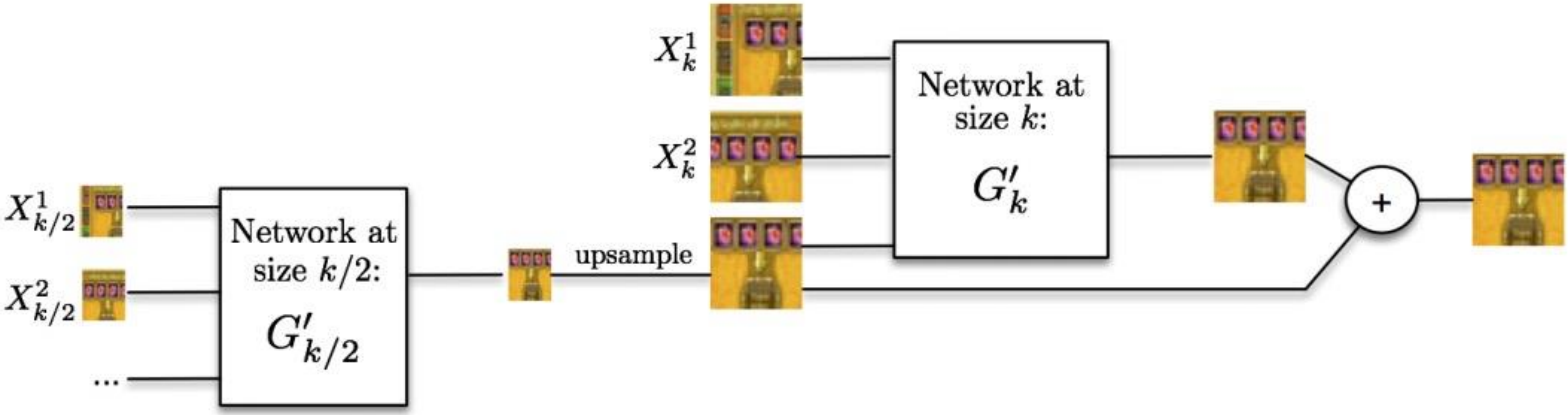
Visual Prediction in Time



Two Layer Composite Model with 2048 LSTM units

Srivastava et al., 2015

Visual Prediction in Time

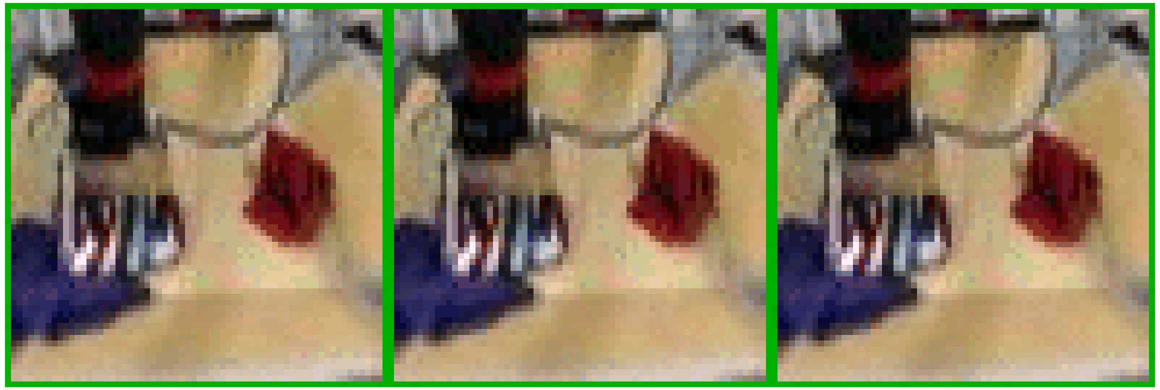


Mathieu et al., 2015

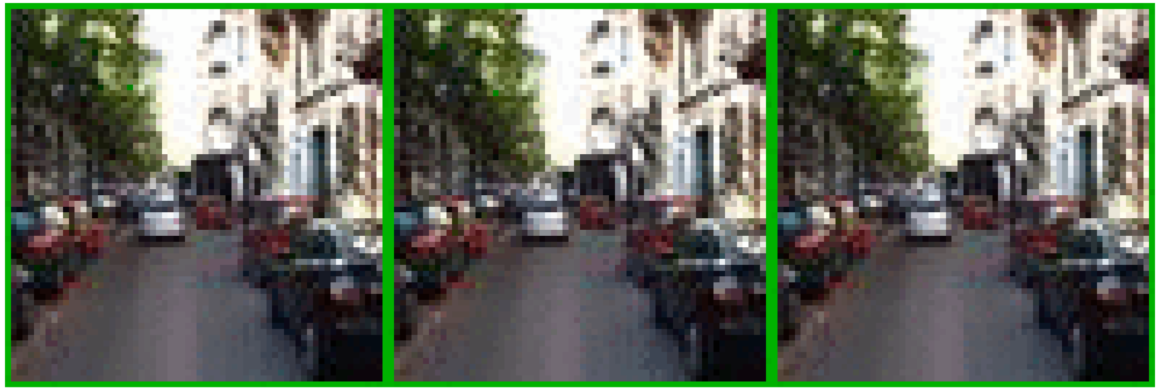
Visual Prediction in Time

Mathieu et al., 2015

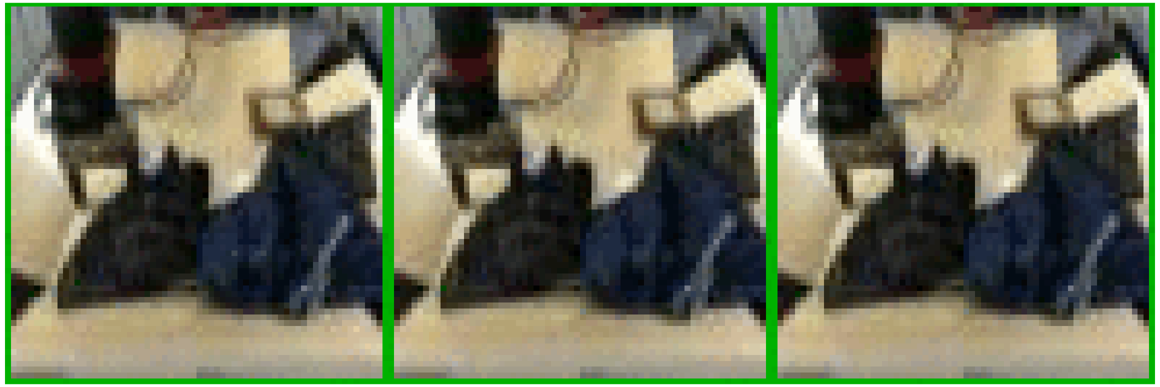
From Pixels to Pixels



Predictions Groundtruth



Predictions Groundtruth



Predictions Groundtruth

Predictions Groundtruth

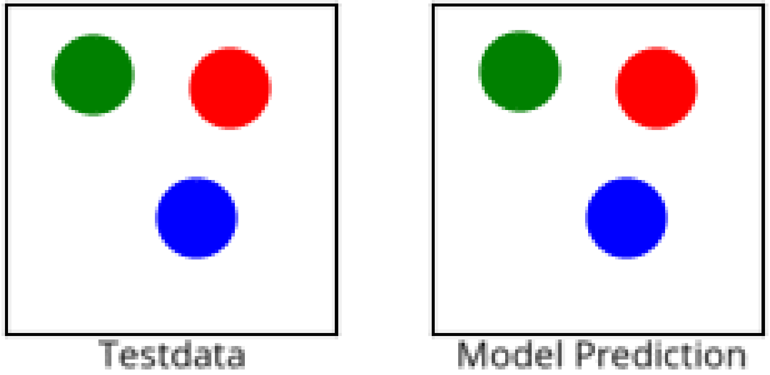
Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V. Le, Honglak Lee. High Fidelity Video Prediction with Large Stochastic Recurrent Neural Networks. NIPS 2019

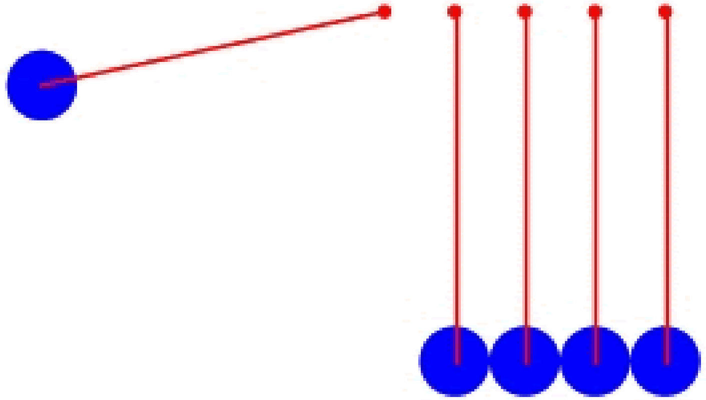
Visual Prediction in Time

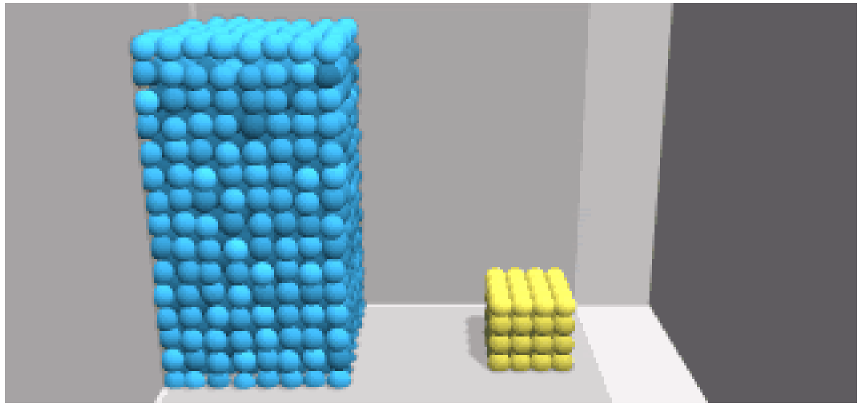
- Not a well-defined problem
- Pixel output space is too large
- Future has a large uncertainty

Interaction Network for Physical Prediction

Object Centric Prediction in a Physical World







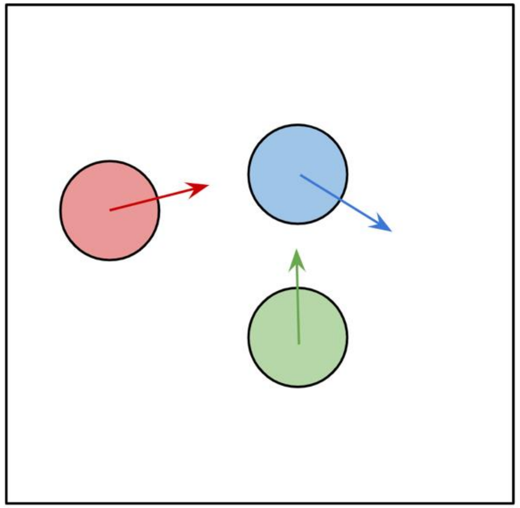
Predicting the physical dynamics

- Given the states of n objects at time t
- We want to predict their states at time t+1

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

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, Koray Kavukcuoglu. "Interaction Networks for Learning about Objects, Relations and Physics" (NIPS 2016)
Michael B. Chang, Tomer D. Ullman, Antonio Torralba, Joshua B. Tenenbaum. "A Compositional Object-Based Approach to Learning Physical Dynamics" (ICLR 2017)

Interaction Module



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

- self-dynamics:

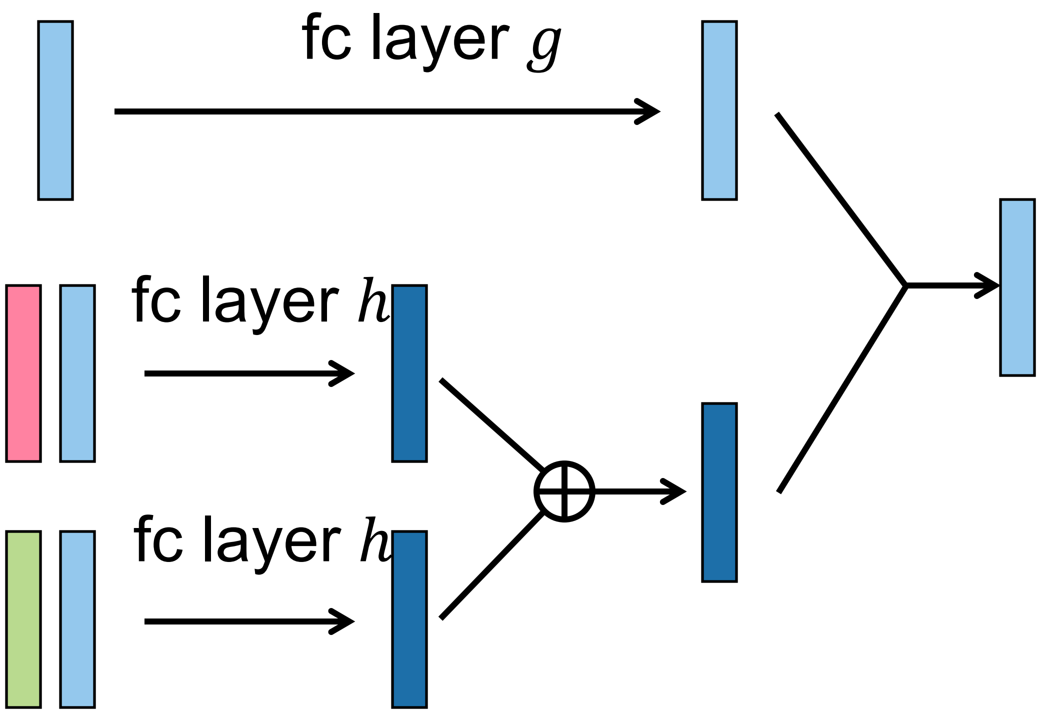
$$g(x_i^t)$$

- relation-dynamics:

$$\sum_{j\neq i} h(x_i^t, x_j^t)$$

- Aggregate the above:

$$F(x_i^t) = f\left(g(x_i^t), \sum_{j\neq i} h(x_i^t, x_j^t) \right)$$



Prediction

Aggregate the unary and binary terms:

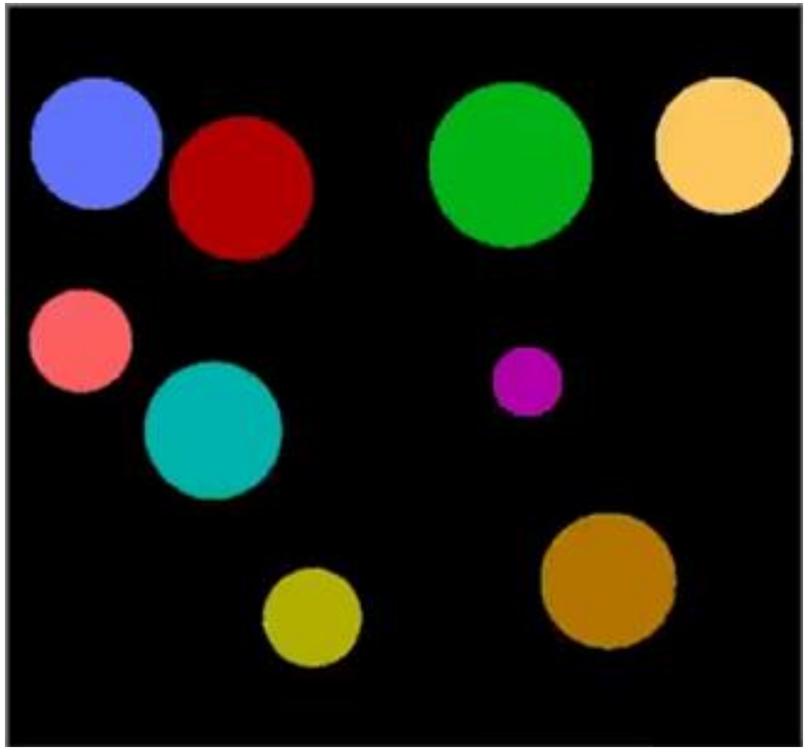
$$x_i^{t+1} = F(x_i^t) = f\big(g(x_i^t), \textstyle\sum_{j \neq i} h(x_i^t, x_j^t) \big)$$

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

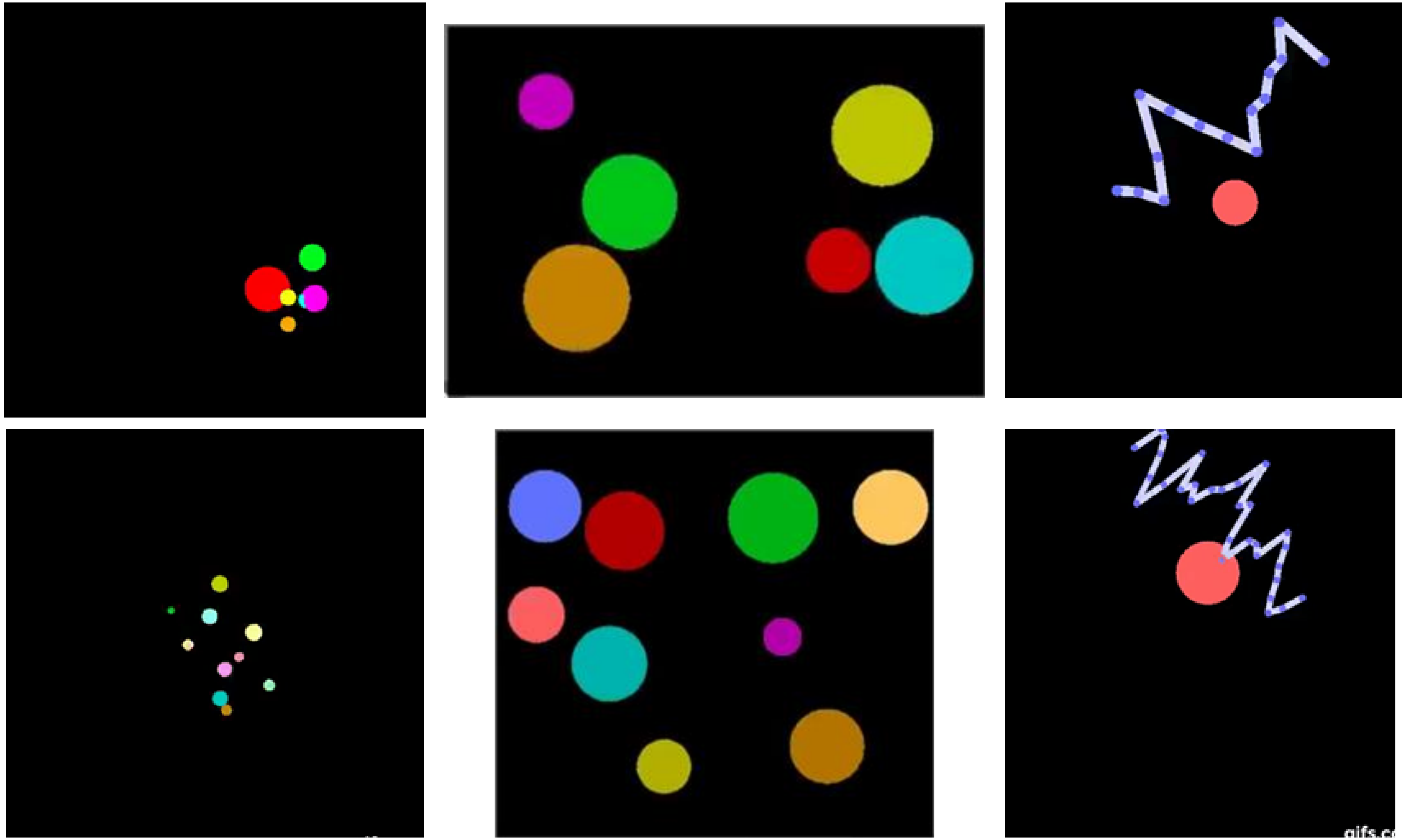
Training loss function: $L_p = \displaystyle\sum_{t=1}^{T} \sum_{i=1}^{n} \left\| \hat{p}_i^{t+1} - p_i^{t+1} \right\|_2^2$

Interaction Network

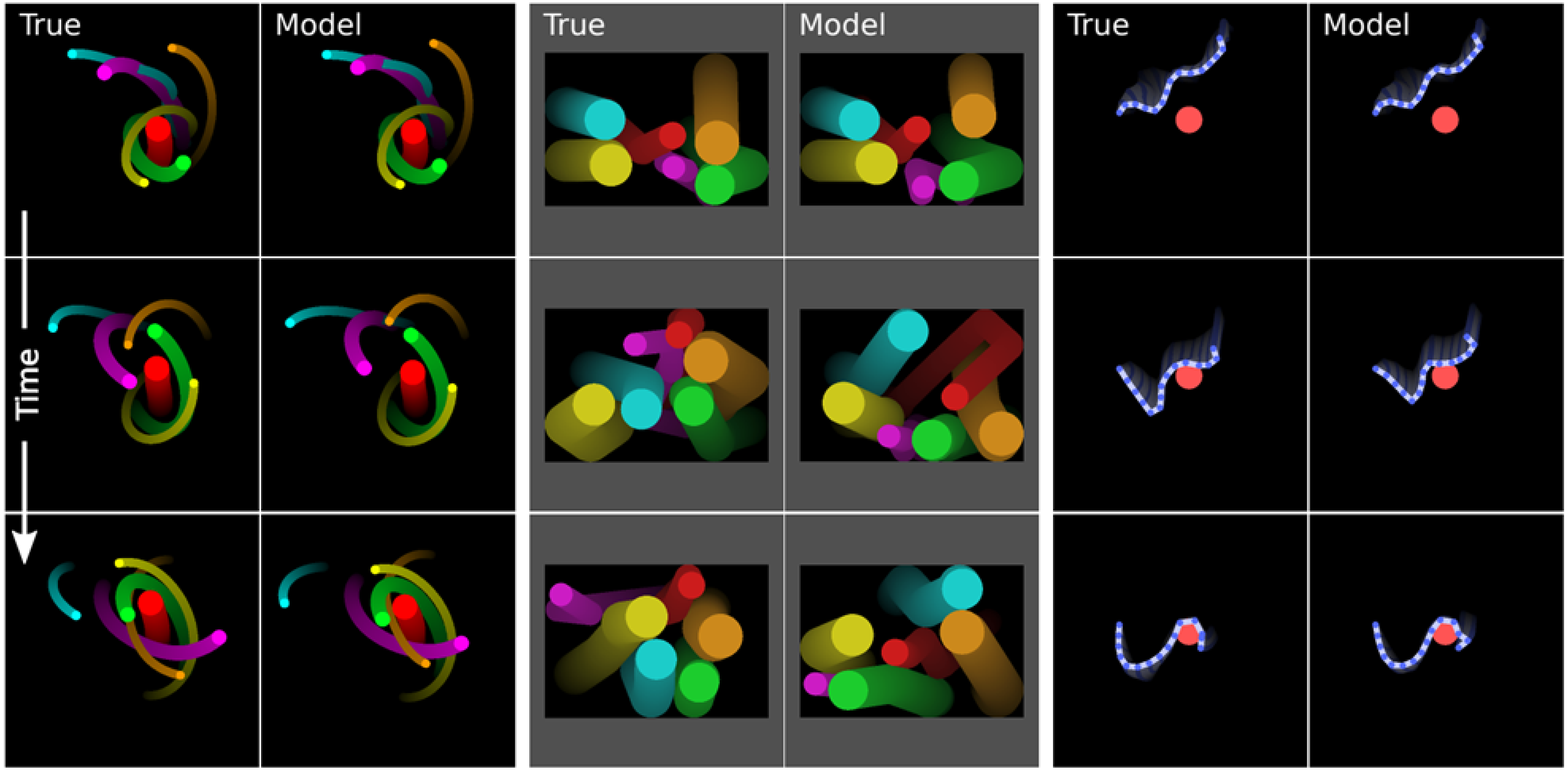
- Object Representation
  - Use ground-truth state as input
  - Rigid Object: mass point (radius, mass, center, velocity)
  - Deformable Object: collection of mass points



Prediction Rollouts

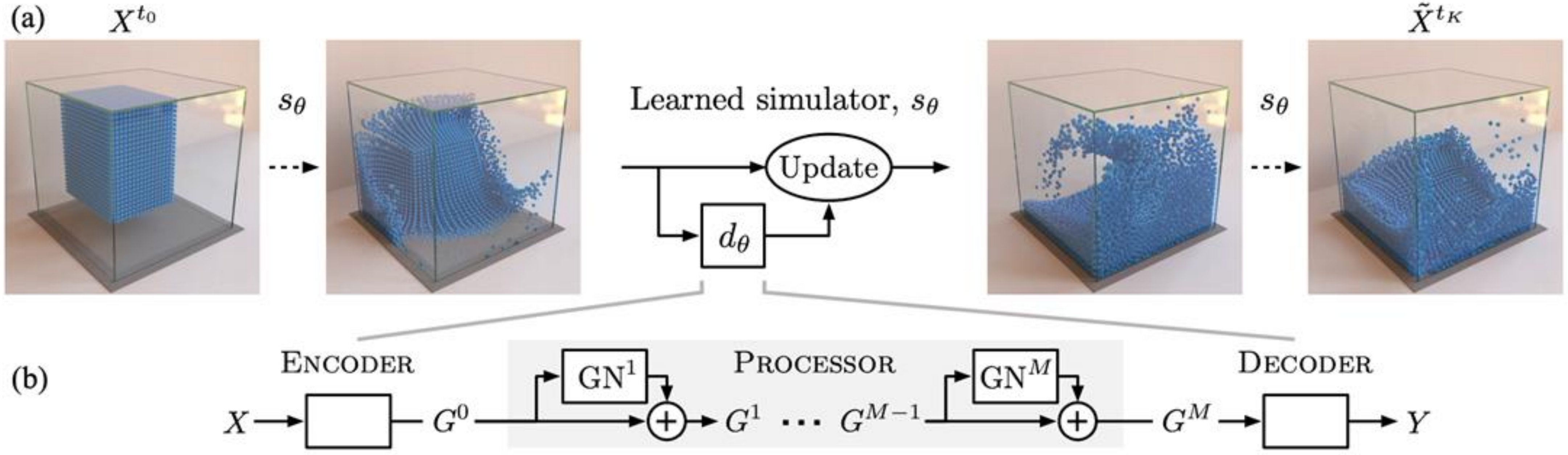


Prediction Results

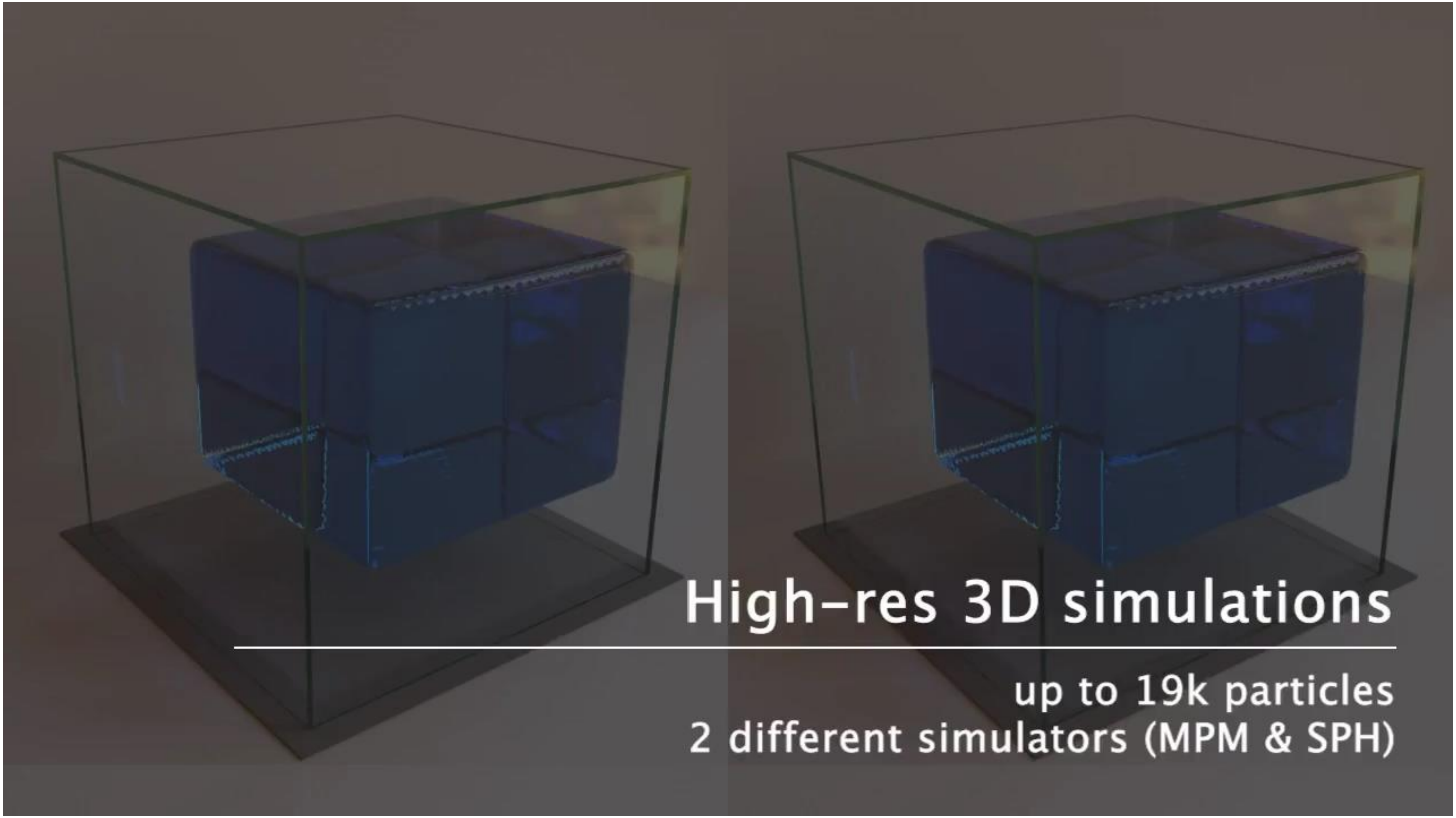


Learning to simulate more complex dynamics

- Propagation Interactions
- Compute Interaction locally

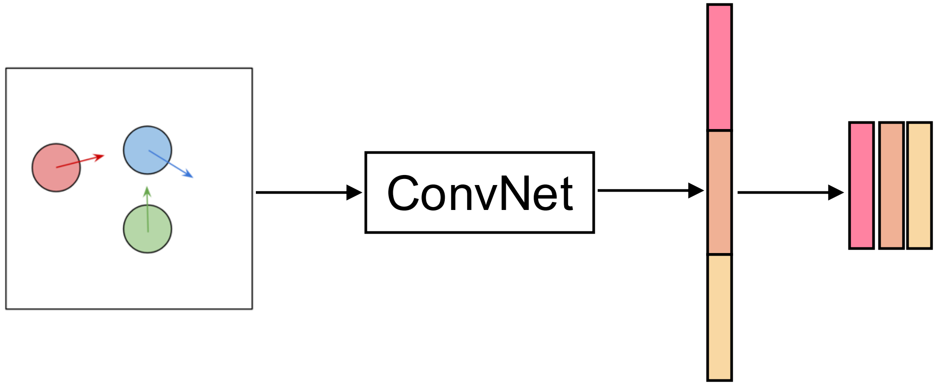


Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, Peter W. Battaglia. "Learning to Simulate Complex Physics with Graph Networks" (arxiv 2020)


High-res 3D simulations
up to 19k particles
2 different simulators (MPM & SPH)

Visual Interaction Network

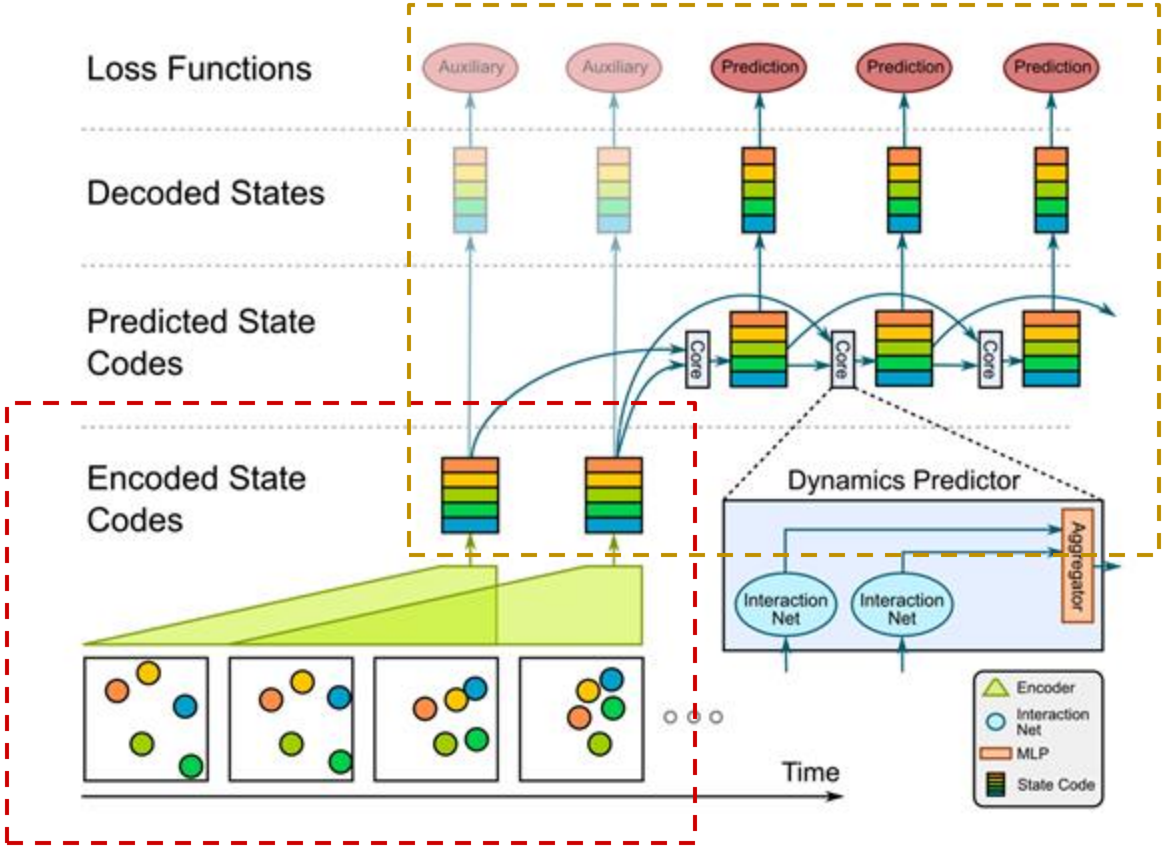
- Visual Interaction Network [1]: Use ConvNet to extract (#obj x 128) feature channels from multiple images.
  - Not very intuitive and cannot generalize to multiple objects
  - Input order is fixed so cannot generalize to multiple appearance



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

Visual Interaction Network

- Estimate the object states from multiple images



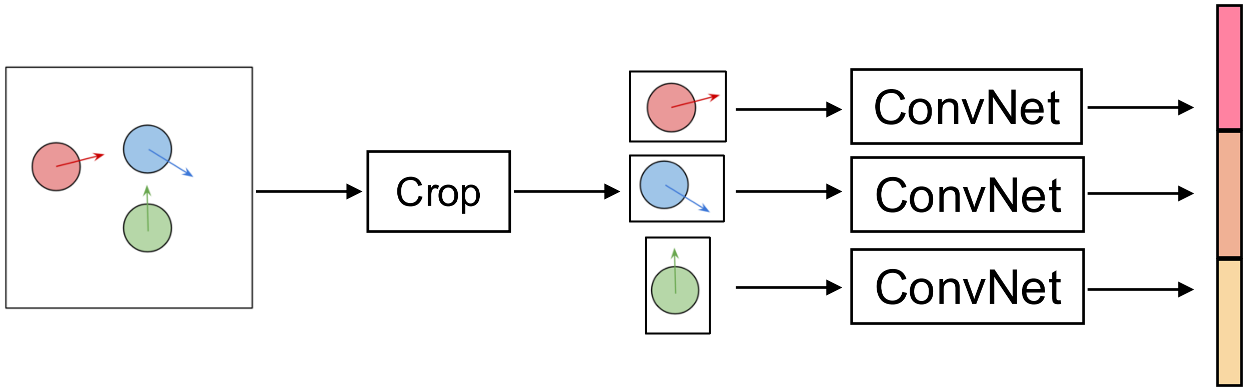
Interaction Network

Visual Encoder

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

Visual Interaction Network

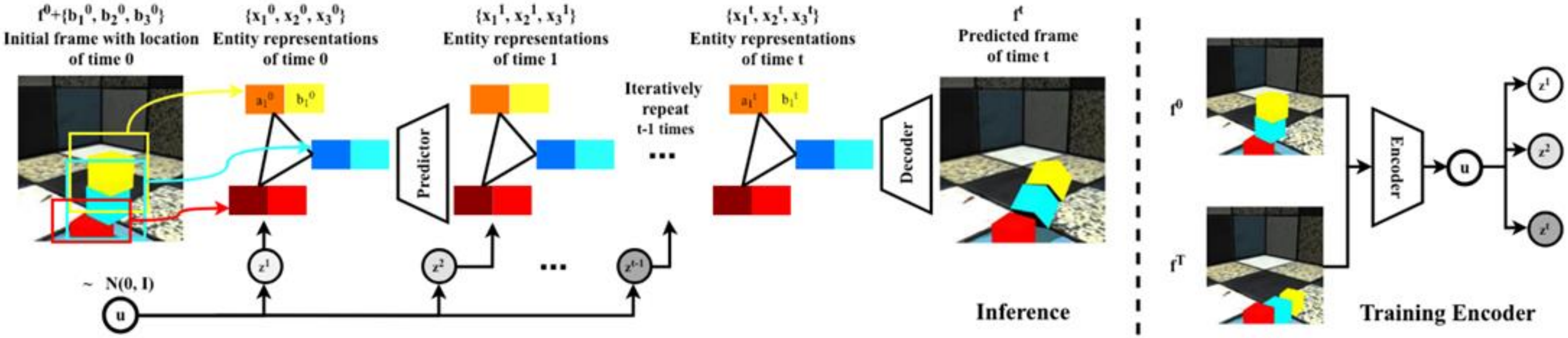
- Visual Interaction Network [1]: Use ConvNet to extract (#obj x 128) features from multiple images.
- Compositional Video Prediction [2,3]: Crop image by RoI and then pass through a ConvNet to get features.



[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

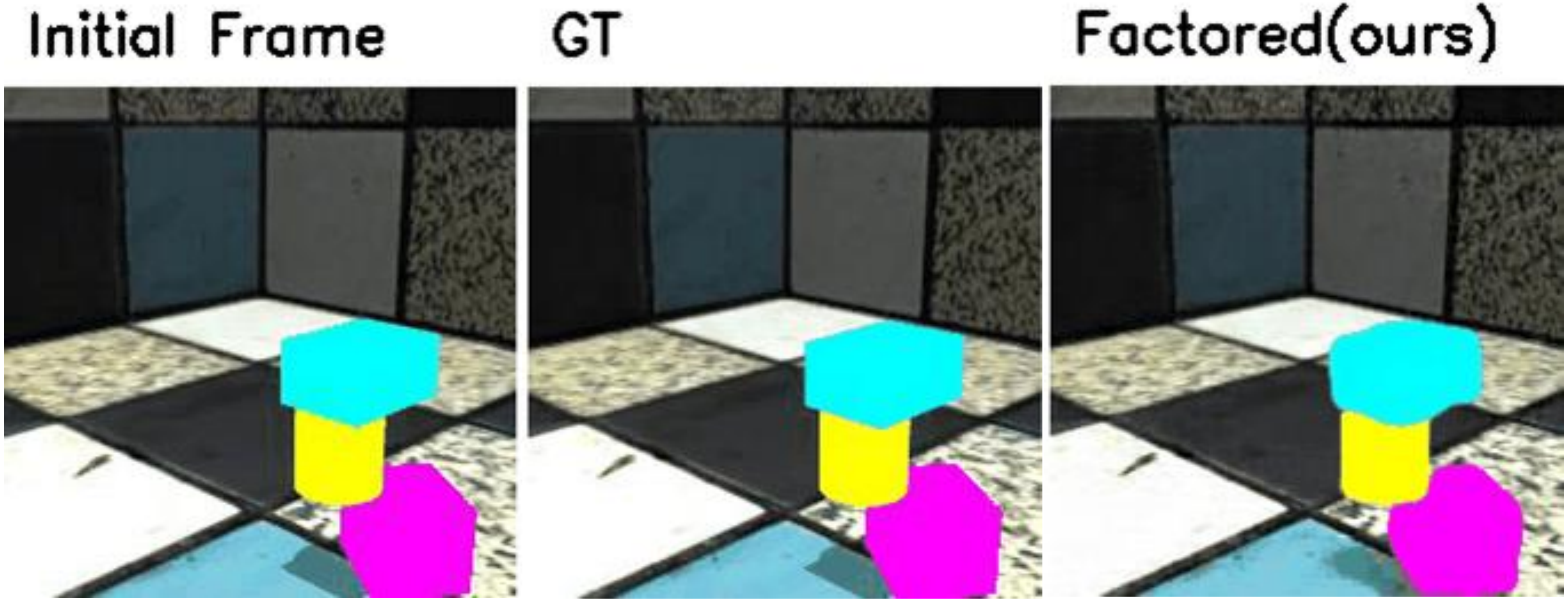
Compositional Video Prediction

- Extract features from cropped object



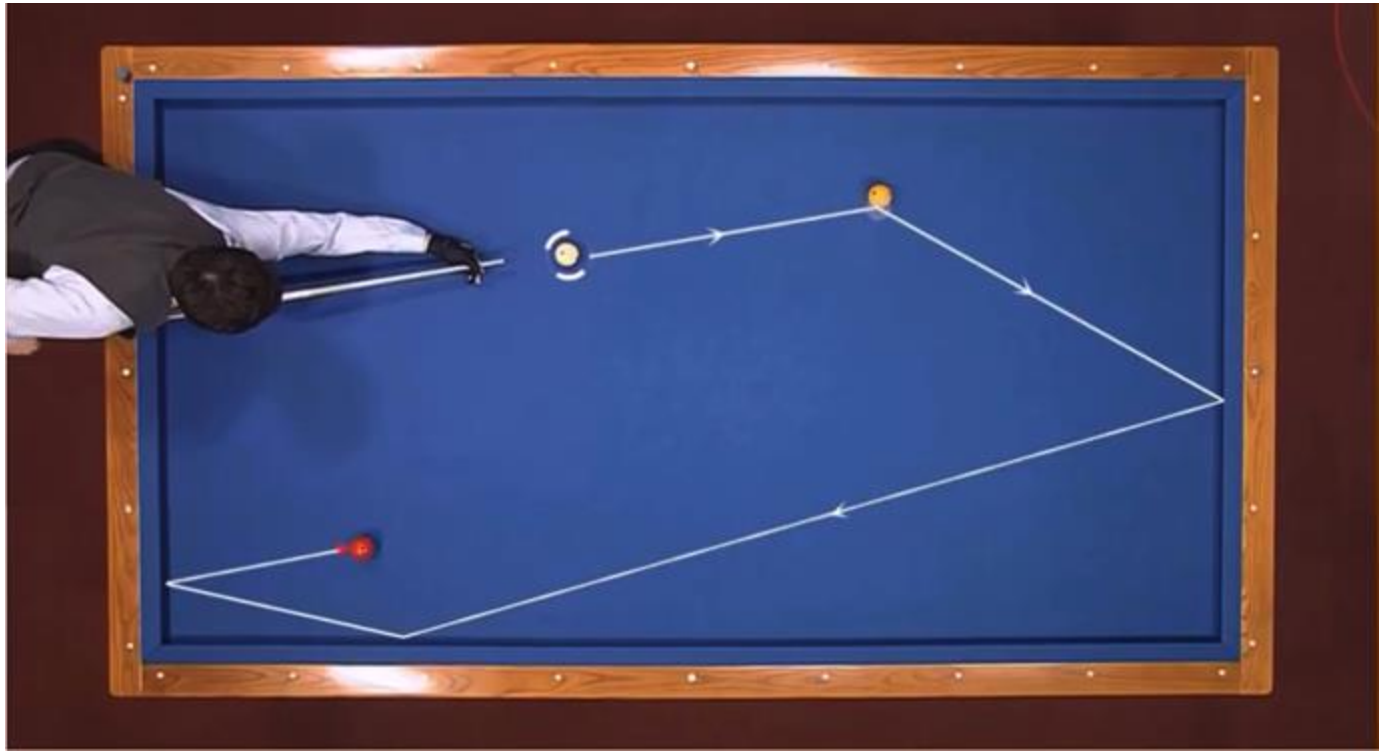
Y. Ye, M. Singh, A. Gupta, S. Tulsiani. “Compositional Video Prediction”. ICCV 2019

Dynamics are simple

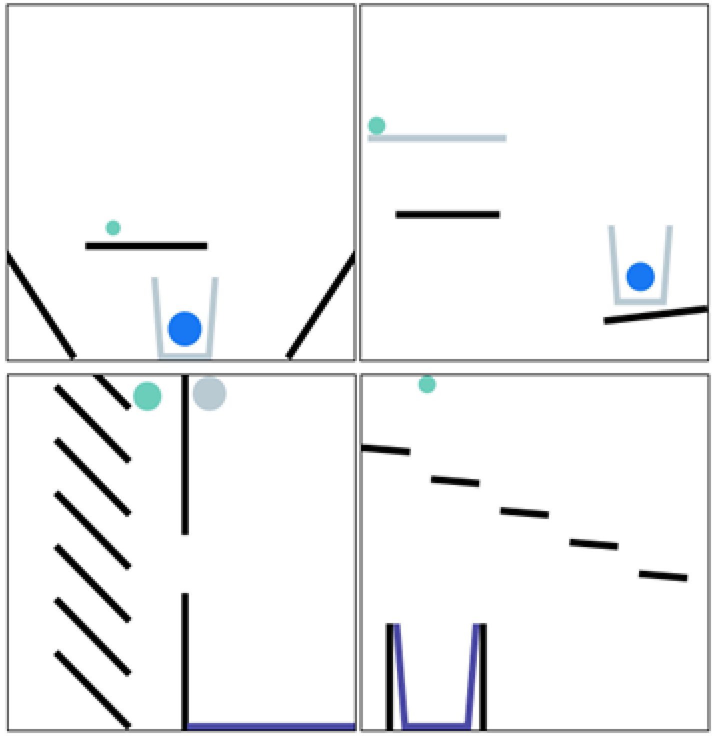


Y. Ye, M. Singh, A. Gupta, S. Tulsiani. “Compositional Video Prediction”. ICCV 2019

More complex / Real World dynamics prediction

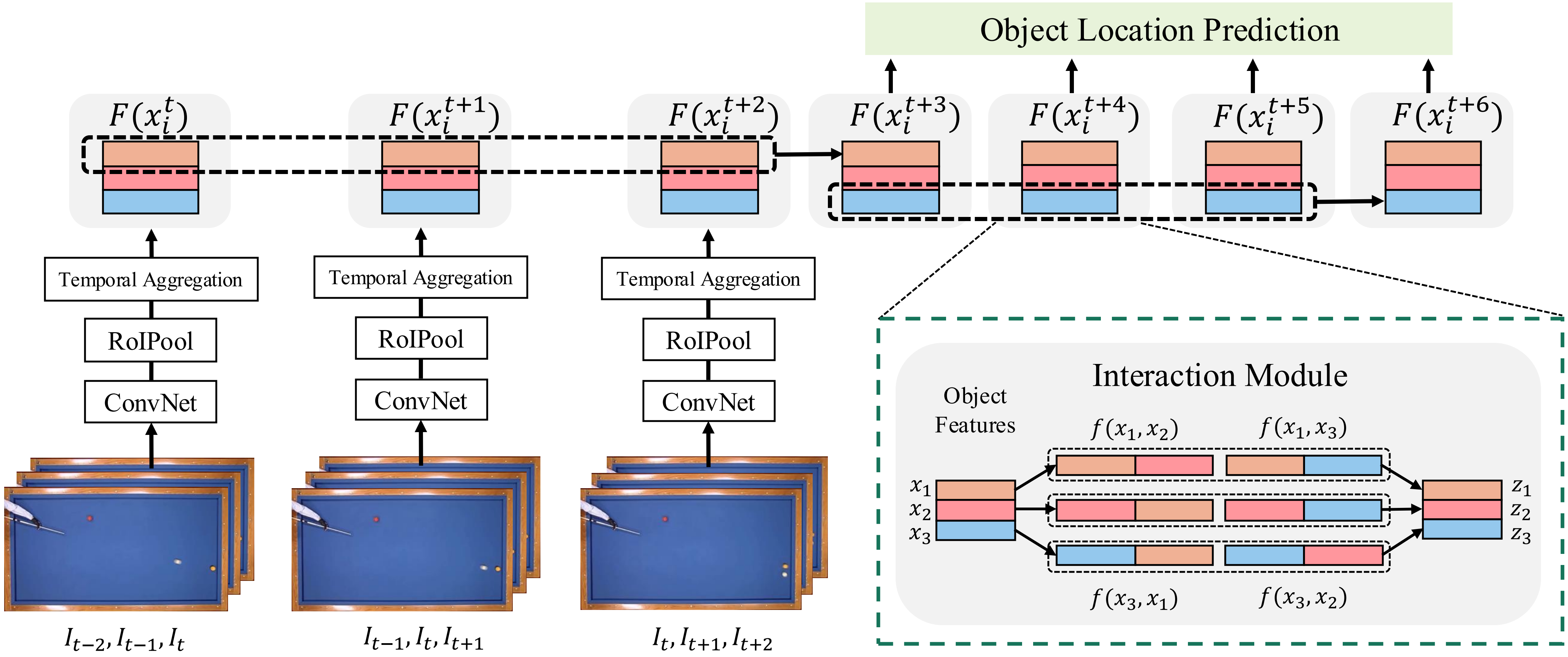


Goal: 1) hit the white ball so that it hits the other object balls. 2) Before hitting the last object ball, the white ball need to hit the cushions at least three times.



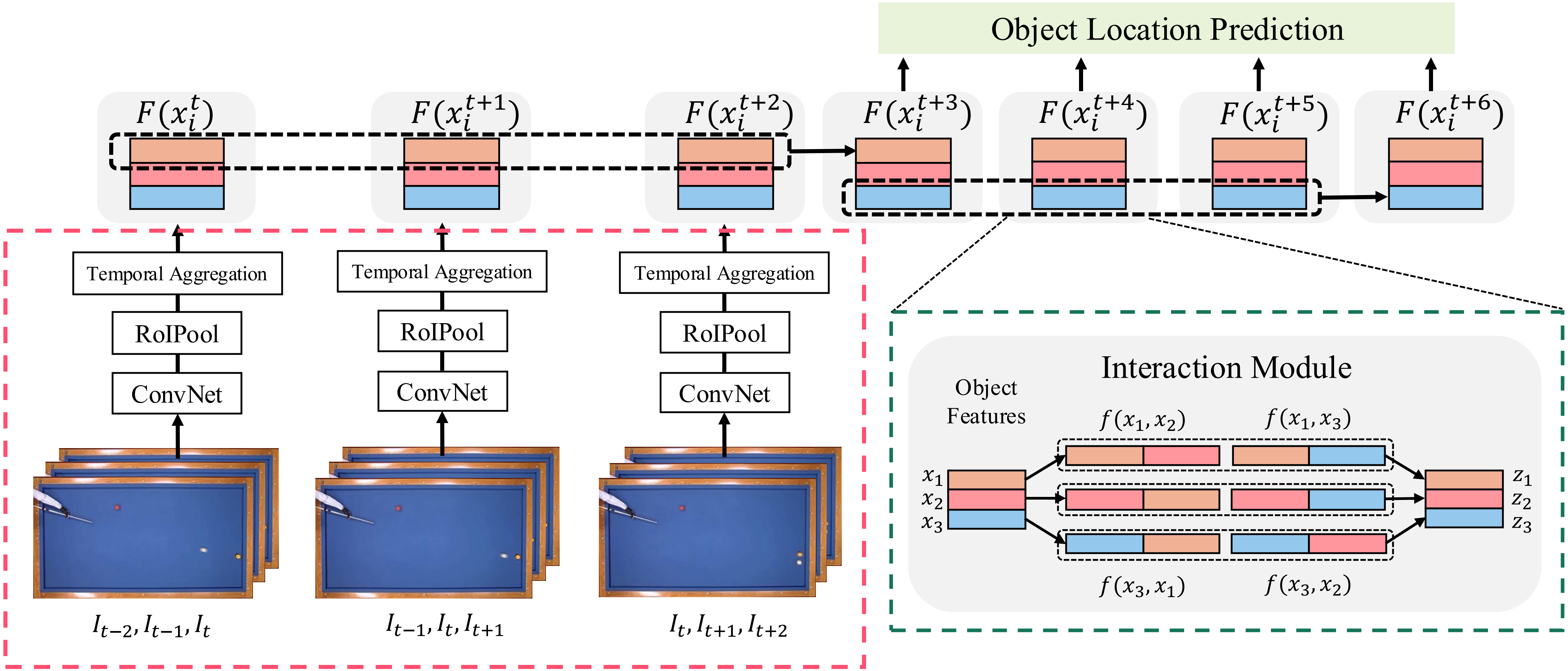
Goal: make the green ball touch the blue/purple object by adding a red ball

Region Proposal Interaction Networks



Qi et al., 2021

Visual Encoder



Visual Encoder

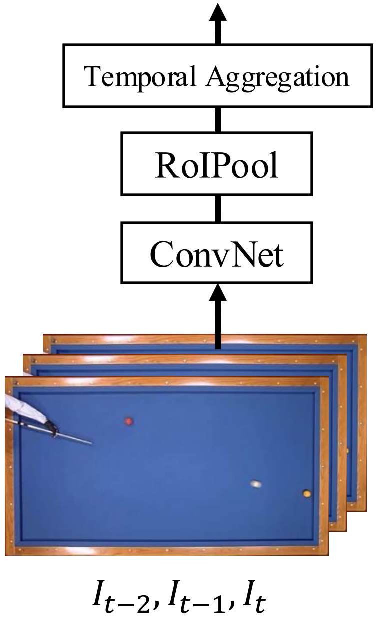
- Object Centric Representation for Prediction

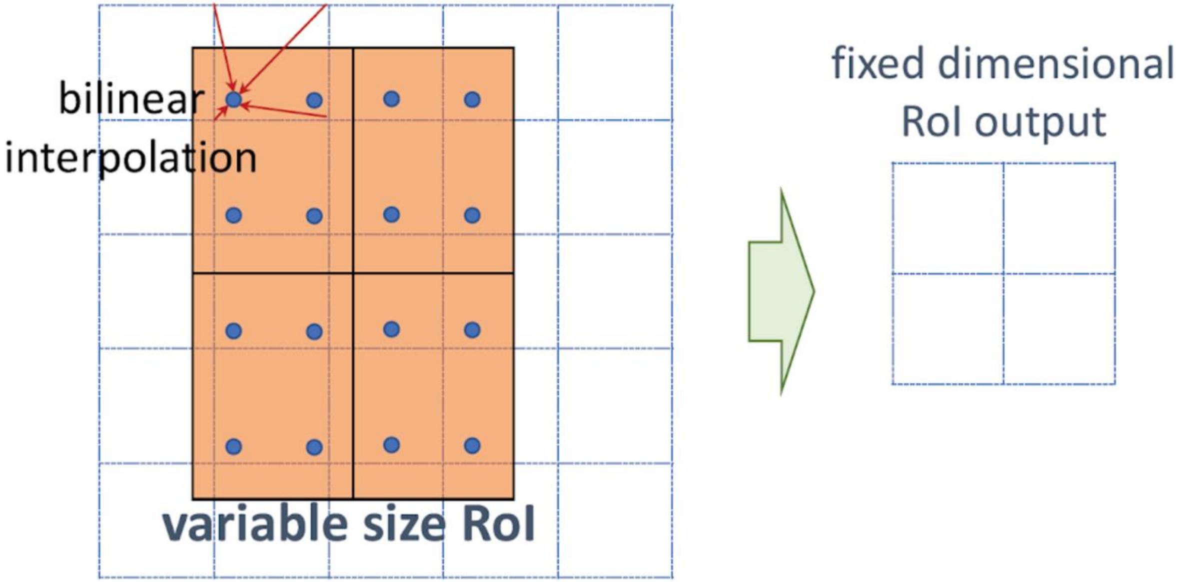
- We extract the state feature representations of n objects in time t, and predict their representations in time $t + 1$.

$$\{x_1^t, x_2^t, \ldots, x_n^t\} \longrightarrow \{x_1^{t+1}, x_2^{t+1}, \ldots, x_n^{t+1}\}$$

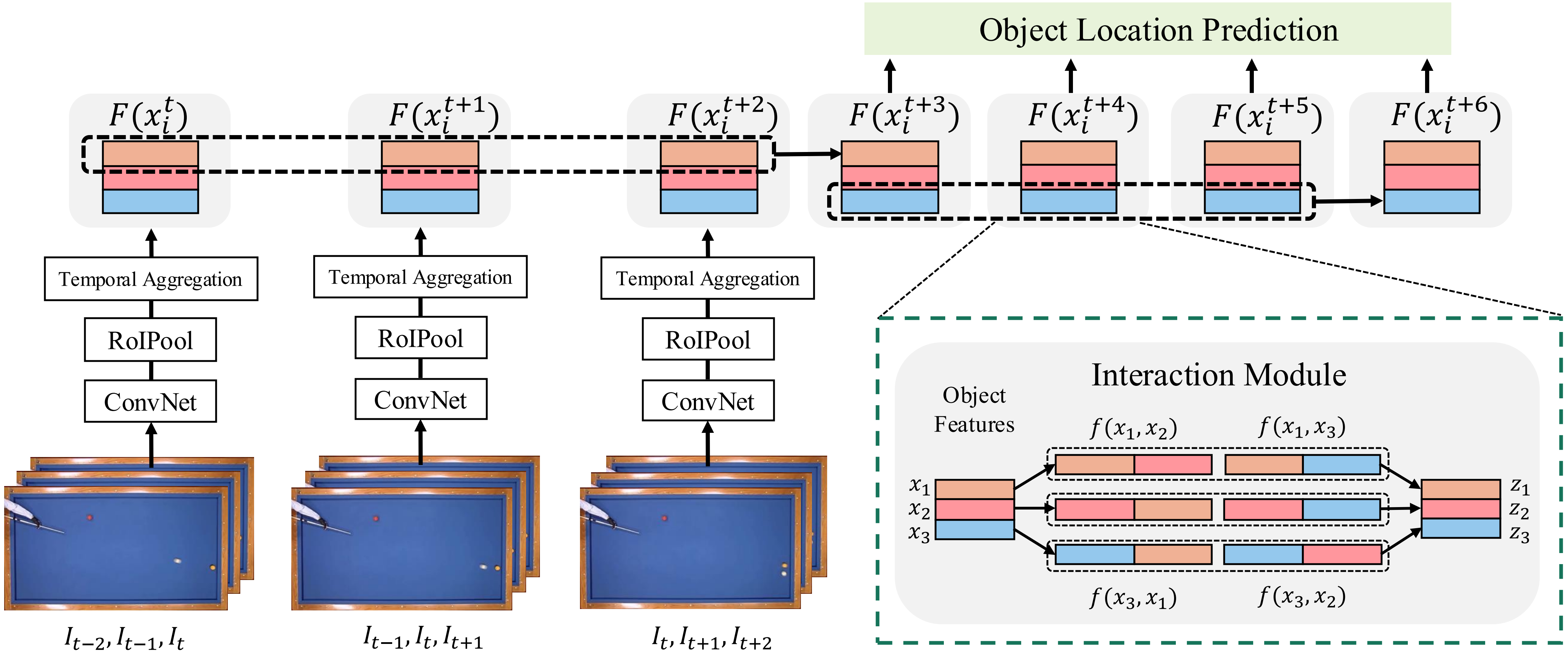
Visual Encoder

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

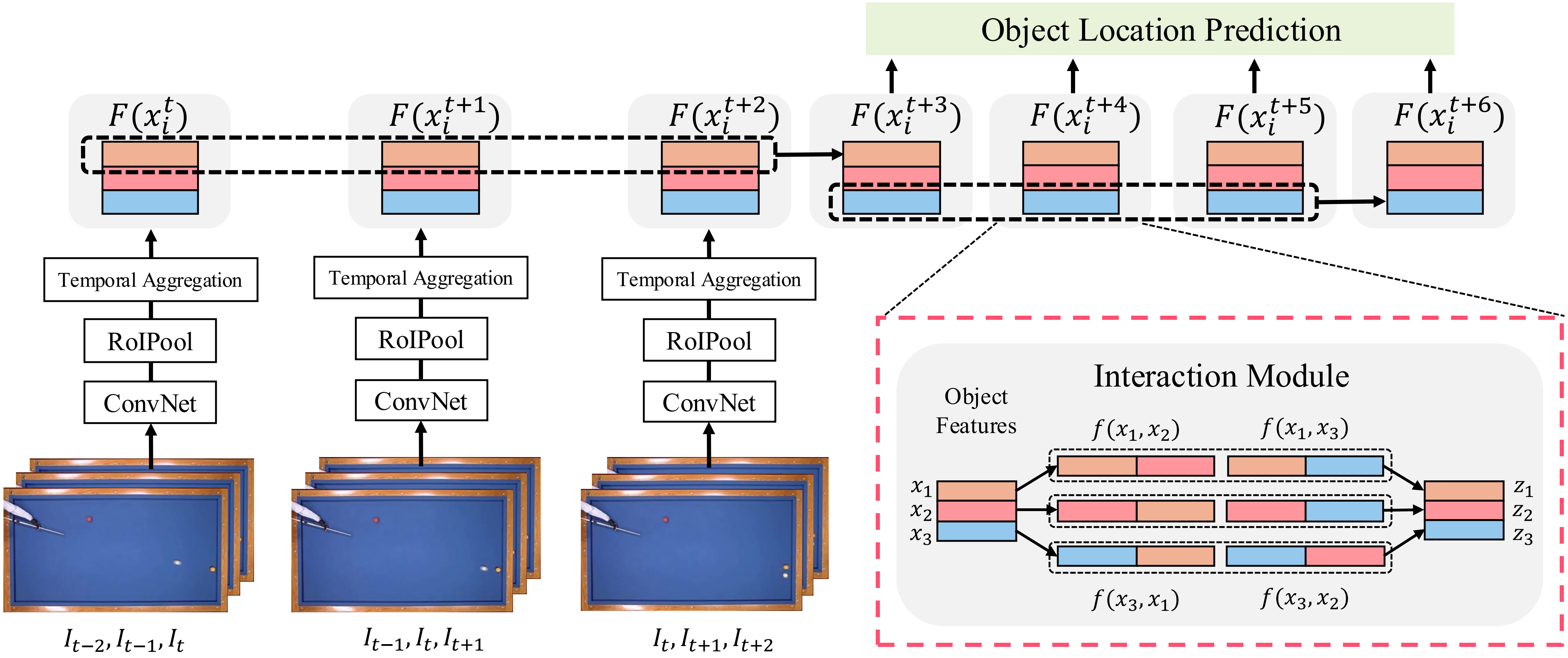




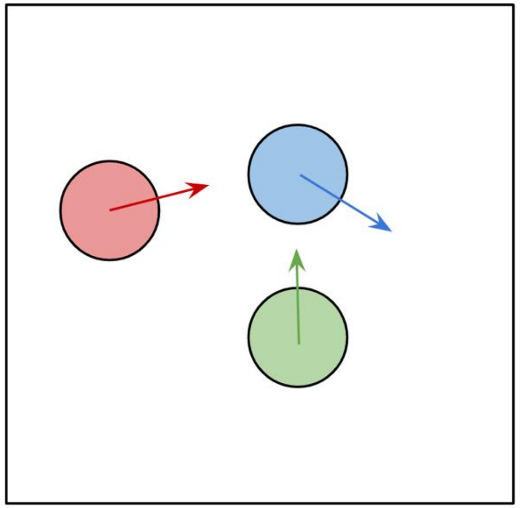
Interaction Module in feature space



Interaction Module in feature space

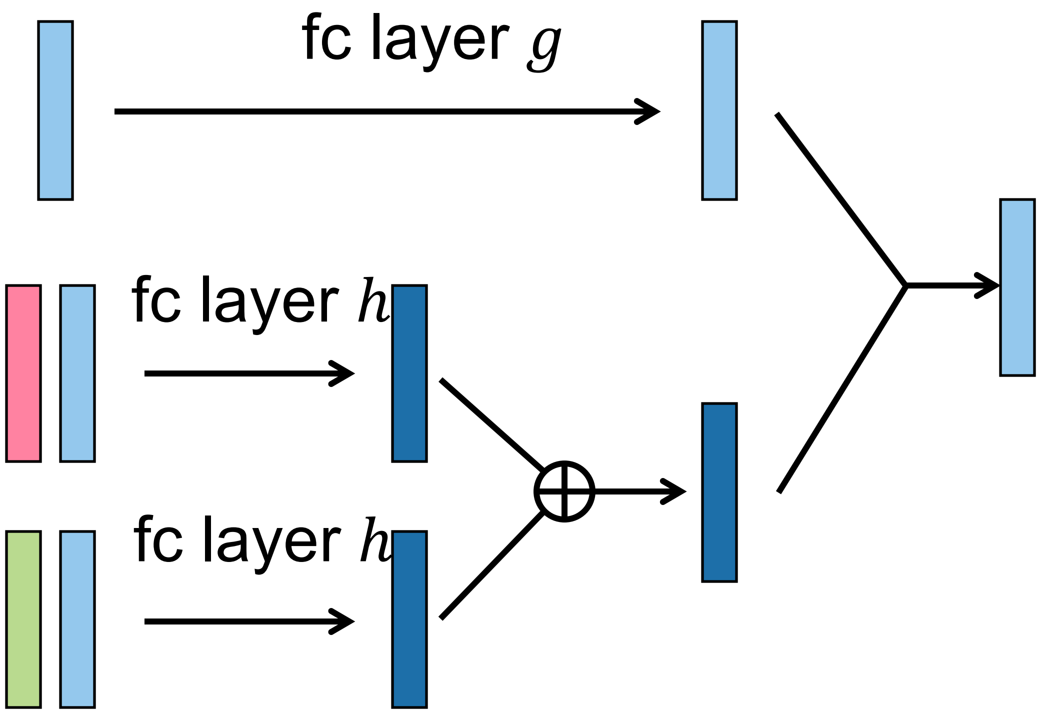


Interaction Module

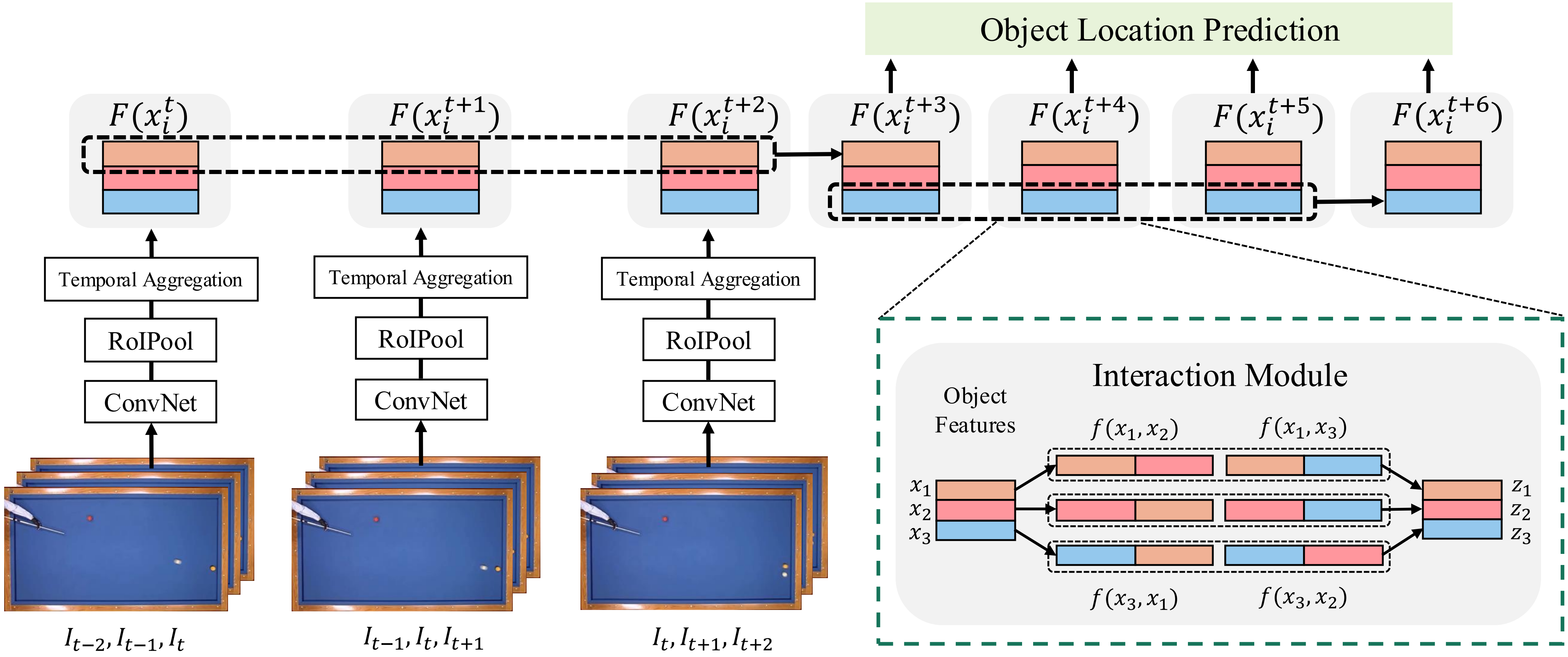


If we want to predict the future movement of the blue billiard

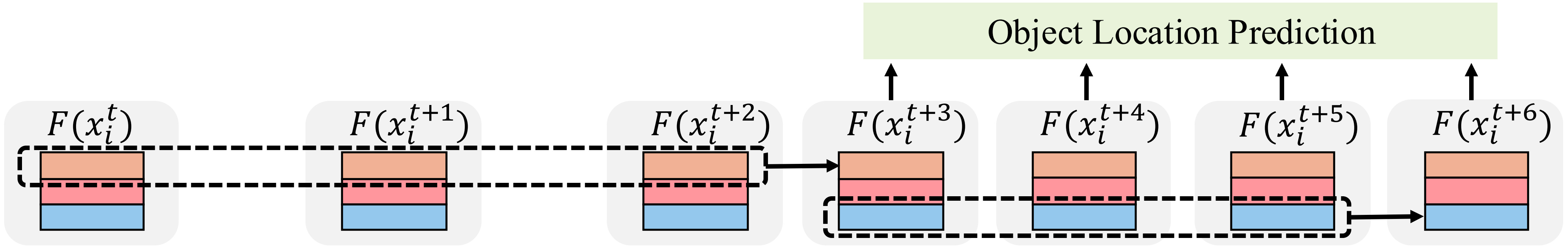
- self-dynamics: (Newton's first law)
  $$g(x_i^t)$$
- relation-dynamics: (Newton's second law)
  $$\sum_{j\neq i} h(x_i^t, x_j^t)$$
- Aggregate the above:
  $$F(x_i^t) = f\left(g(x_i^t), \sum_{j\neq i} h(x_i^t, x_j^t) \right)$$



Prediction


Object Location Prediction
$F(x_i^t)$
$F(x_i^{t+1})$
$F(x_i^{t+2})$
$F(x_i^{t+3})$
$F(x_i^{t+4})$
$F(x_i^{t+5})$
$F(x_i^{t+6})$
Temporal Aggregation
RoIPool
ConvNet
I_{t-2}, I_{t-1}, I_t
I_{t-1}, I_t, I_{t+1}
I_t, I_{t+1}, I_{t+2}
Interaction Module
Object Features
x_1
x_2
x_3
$f(x_1, x_2)$
$f(x_1, x_3)$
$f(x_3, x_1)$
$f(x_3, x_2)$
z_1
z_2
z_3

Prediction

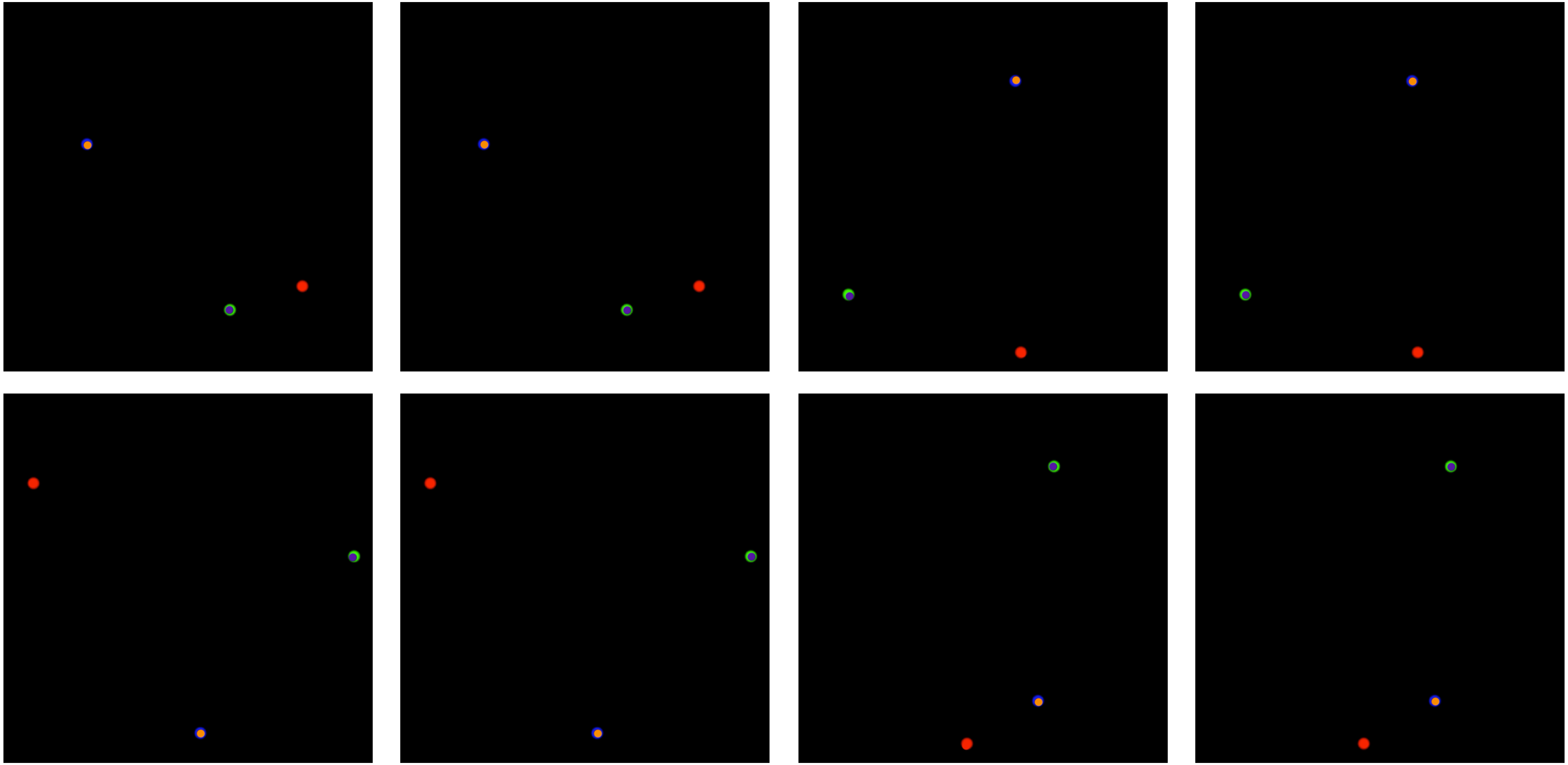


Future feature prediction: $x_i^{t+1} = W_d\left[F\left(x_i^t\right), F\left(x_i^{t-1}\right), \ldots, F\left(x_i^{t-k}\right)\right]$

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

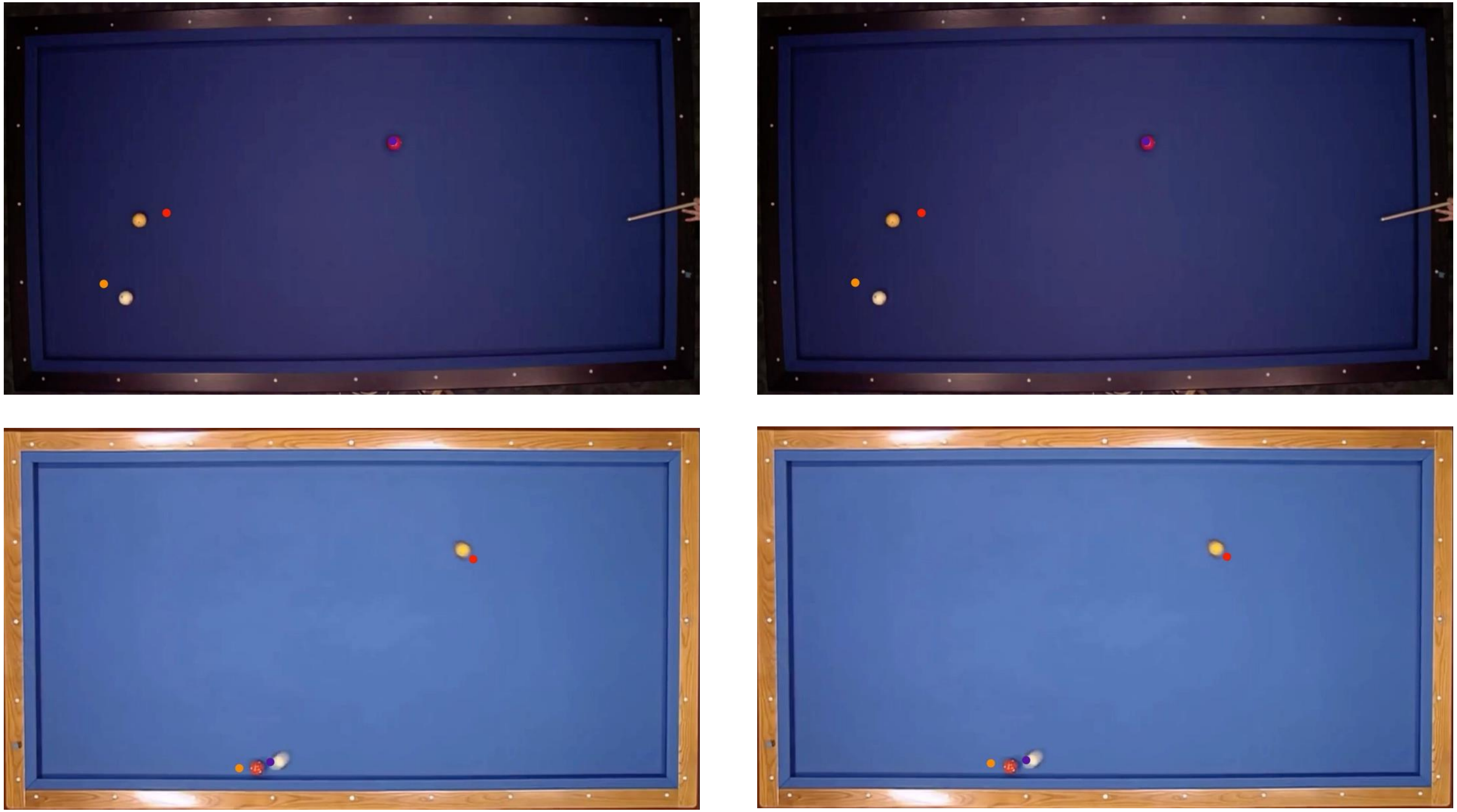
Training loss function: $L_p = \sum_{t=1}^{T} \sum_{i=1}^{n} \left\| \hat{p}_i^{t+1} - p_i^{t+1} \right\|_2^2$

Simulation Billiards



prediction ground-truth prediction ground-truth

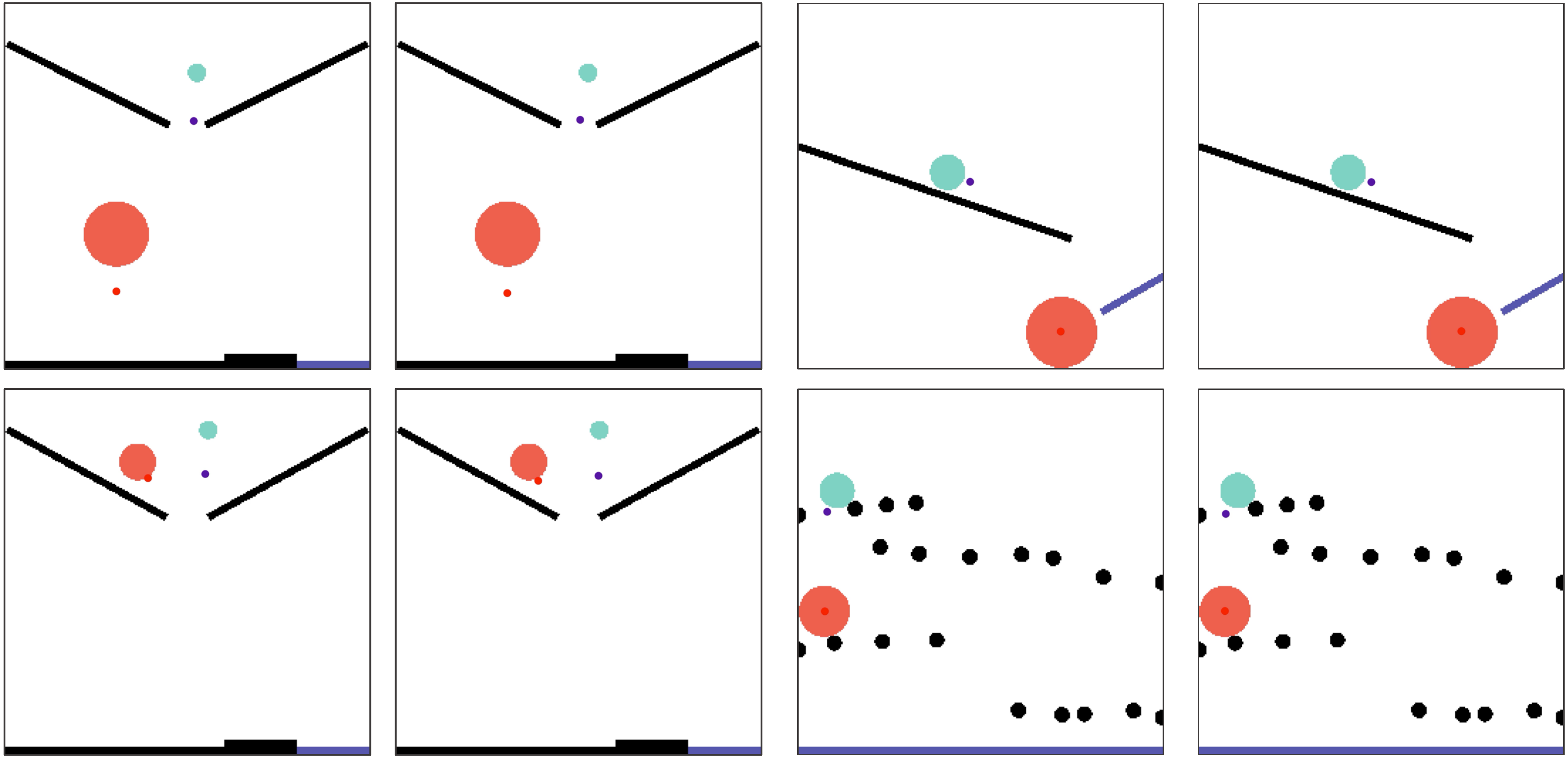
Real Billiards



prediction

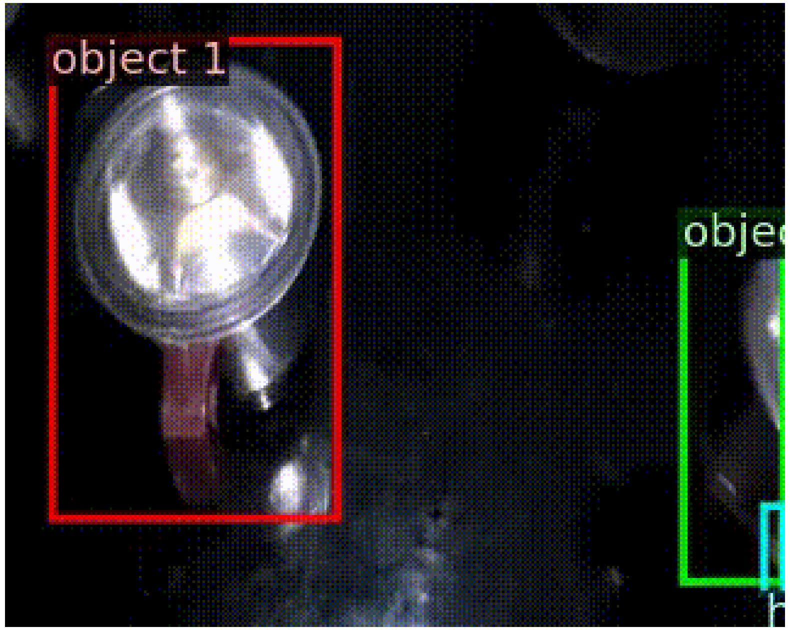
ground-truth

PHYRE

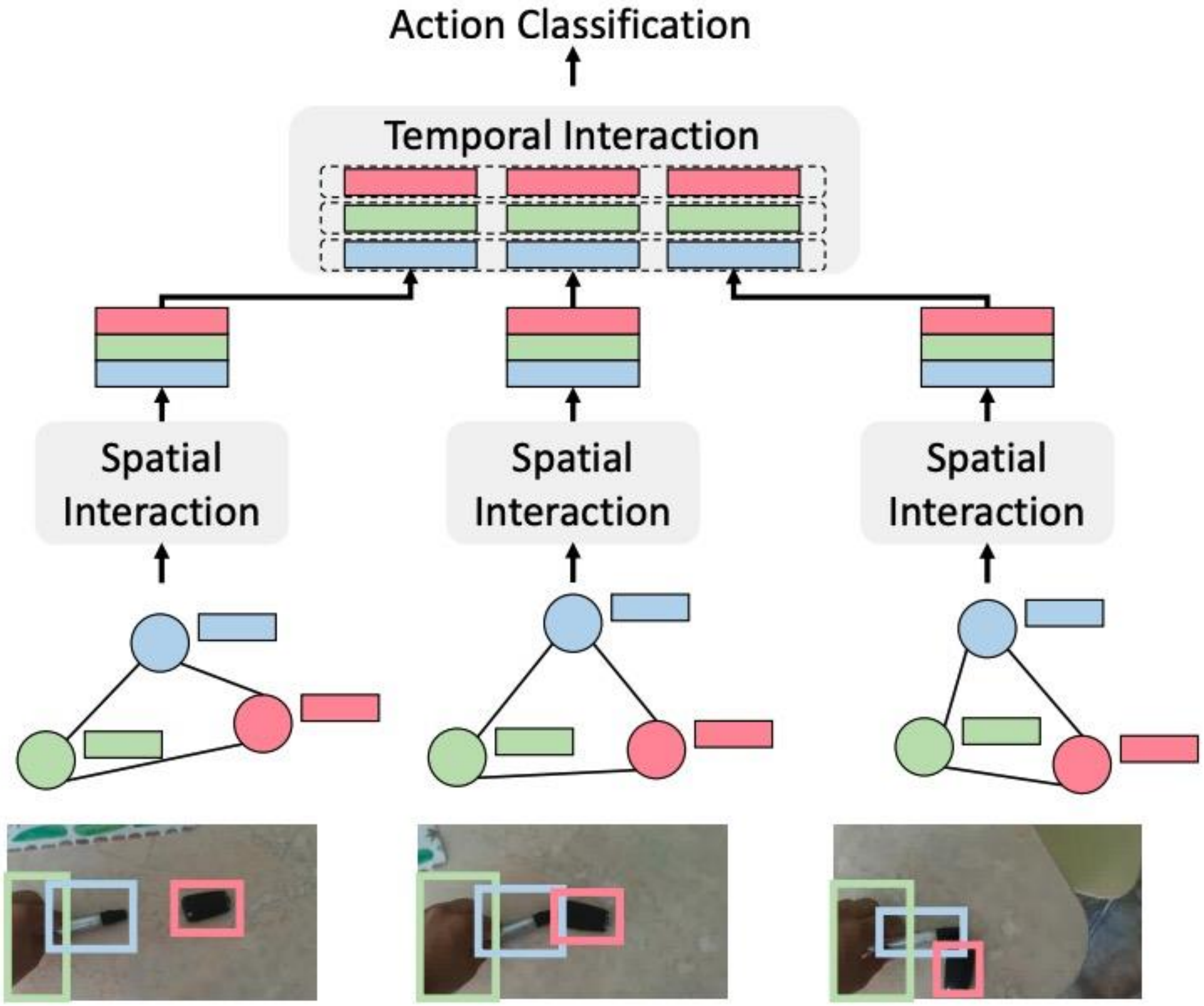


prediction ground-truth prediction ground-truth

Apply to Action Recognition



Materzynska et al., 2020



What Space to Predict

What Space to Predict

Predict Optical Flow:

An Uncertain Future: Forecasting from Static Images using Variational Autoencoders

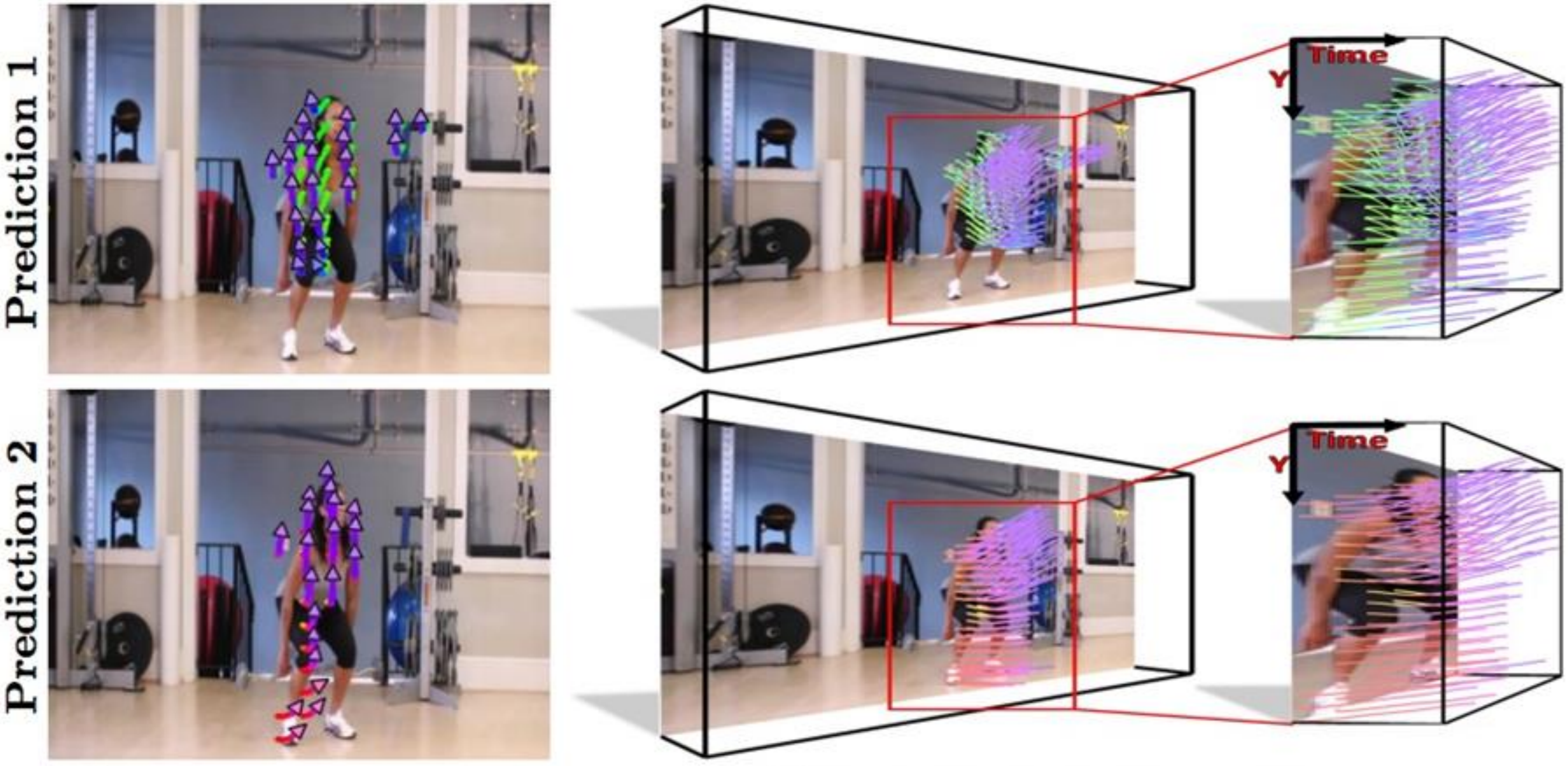
Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert

Predict Skeleton:

Learning to Generate Long-term Future via Hierarchical Prediction

Ruben Villegas[1,*] **Jimei Yang**[2] **Yuliang Zou**[1] **Sungryull Sohn**[1] **Xunyu Lin**[3] **Honglak Lee**[1,4]

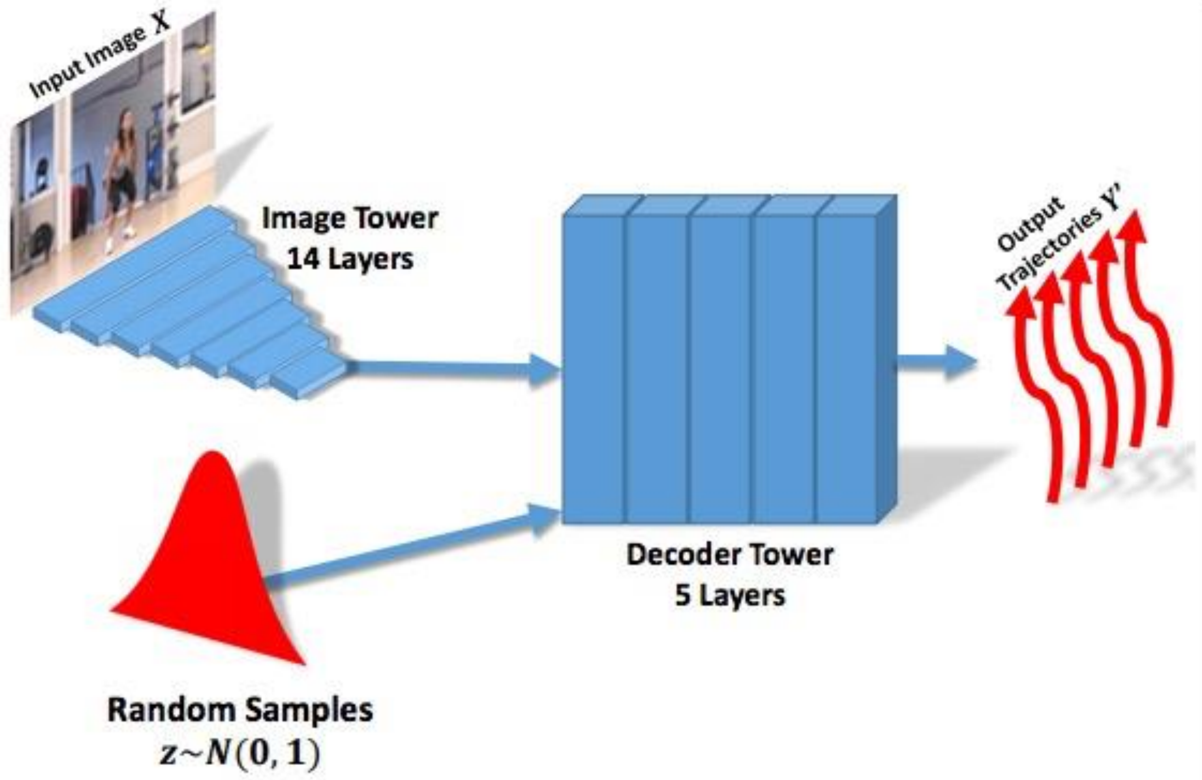
Predict Future Optical Flow from A Single Image



(a) Trajectories on Image (b) Trajectories in Space-Time

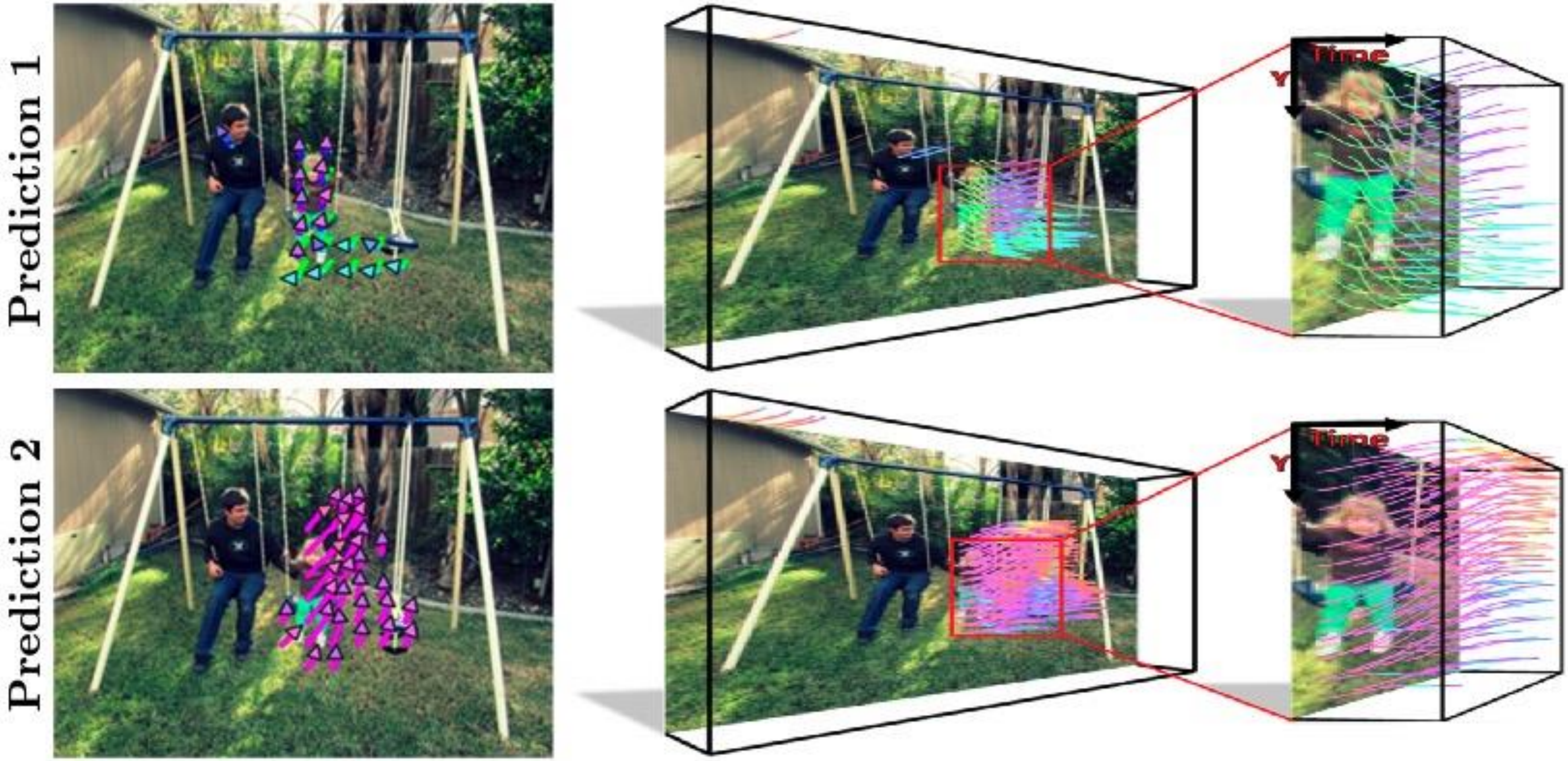
Walker et al. An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders. 2016.

CVAE for Modeling Uncertainty

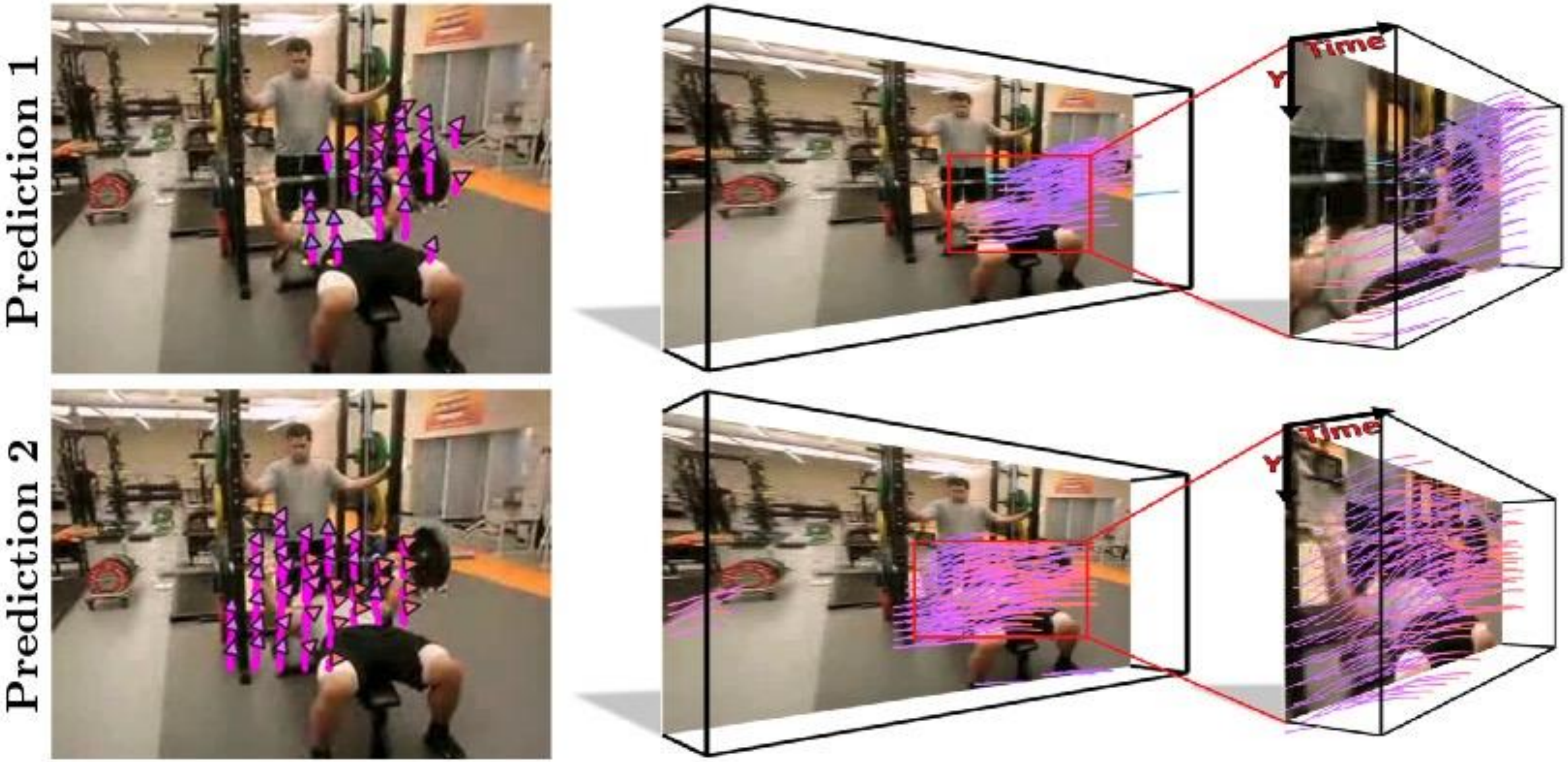


(a) Testing Architecture

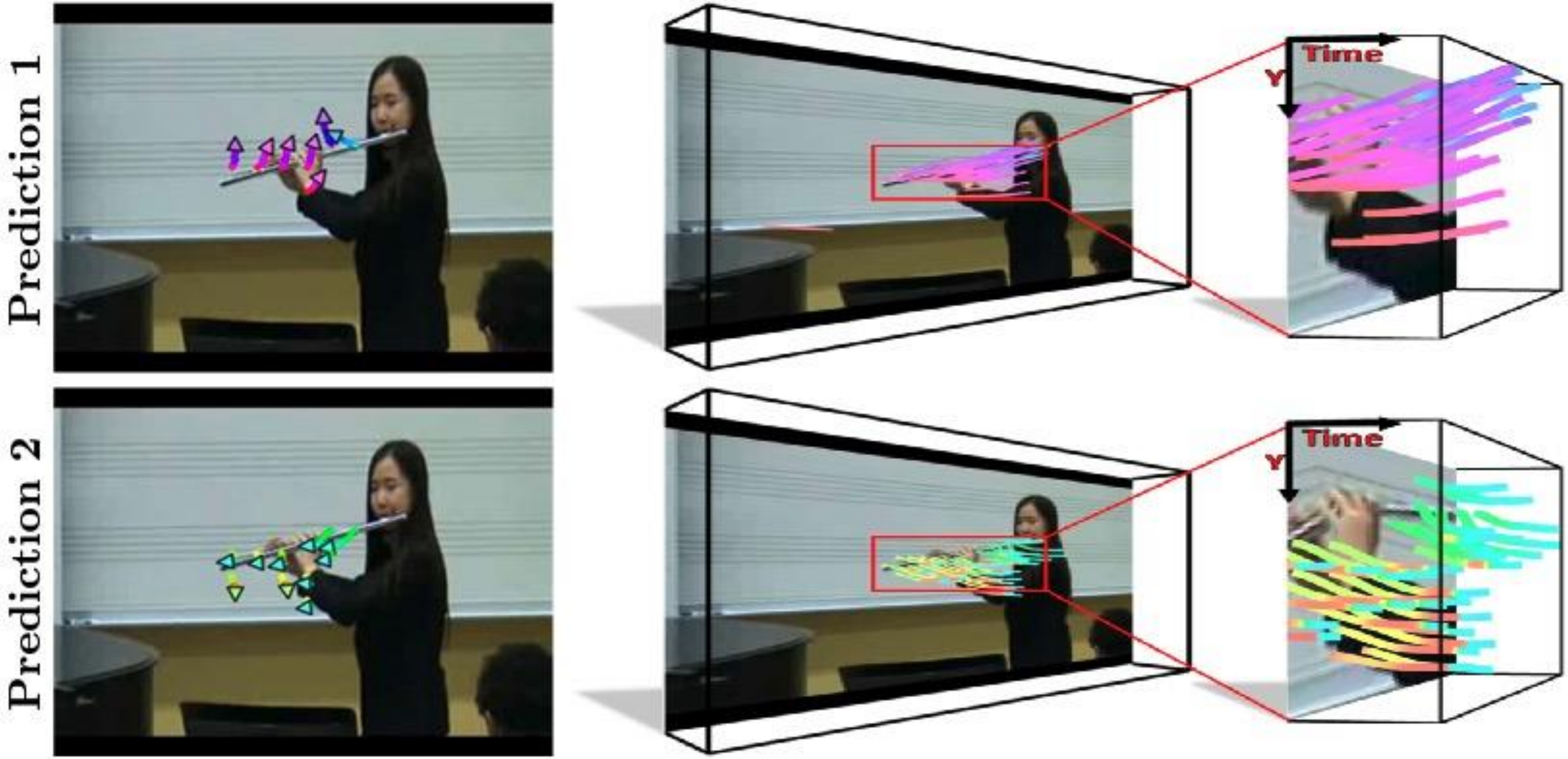
Results



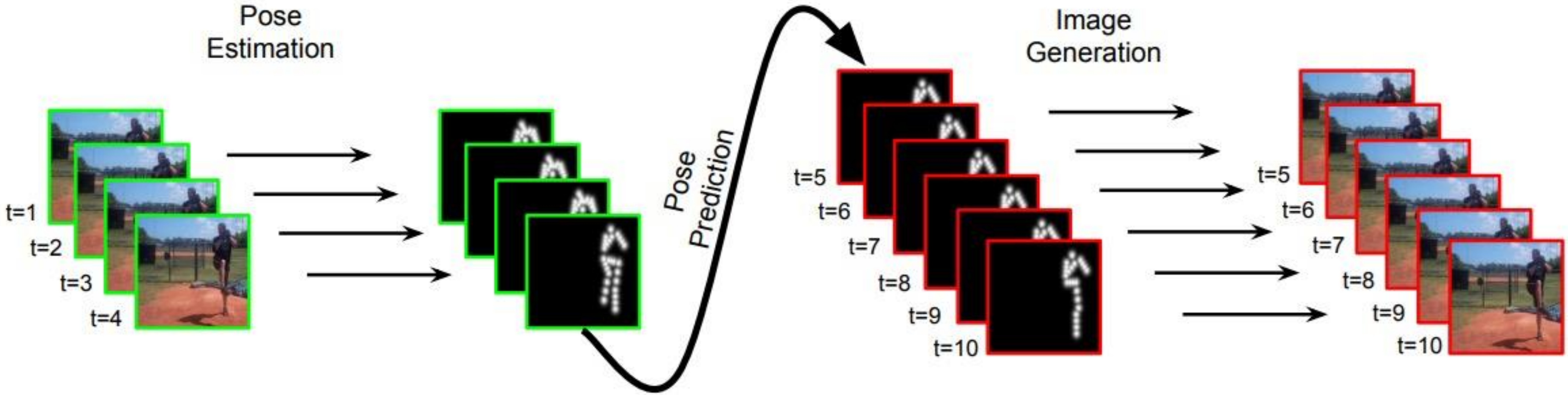
Results



Results

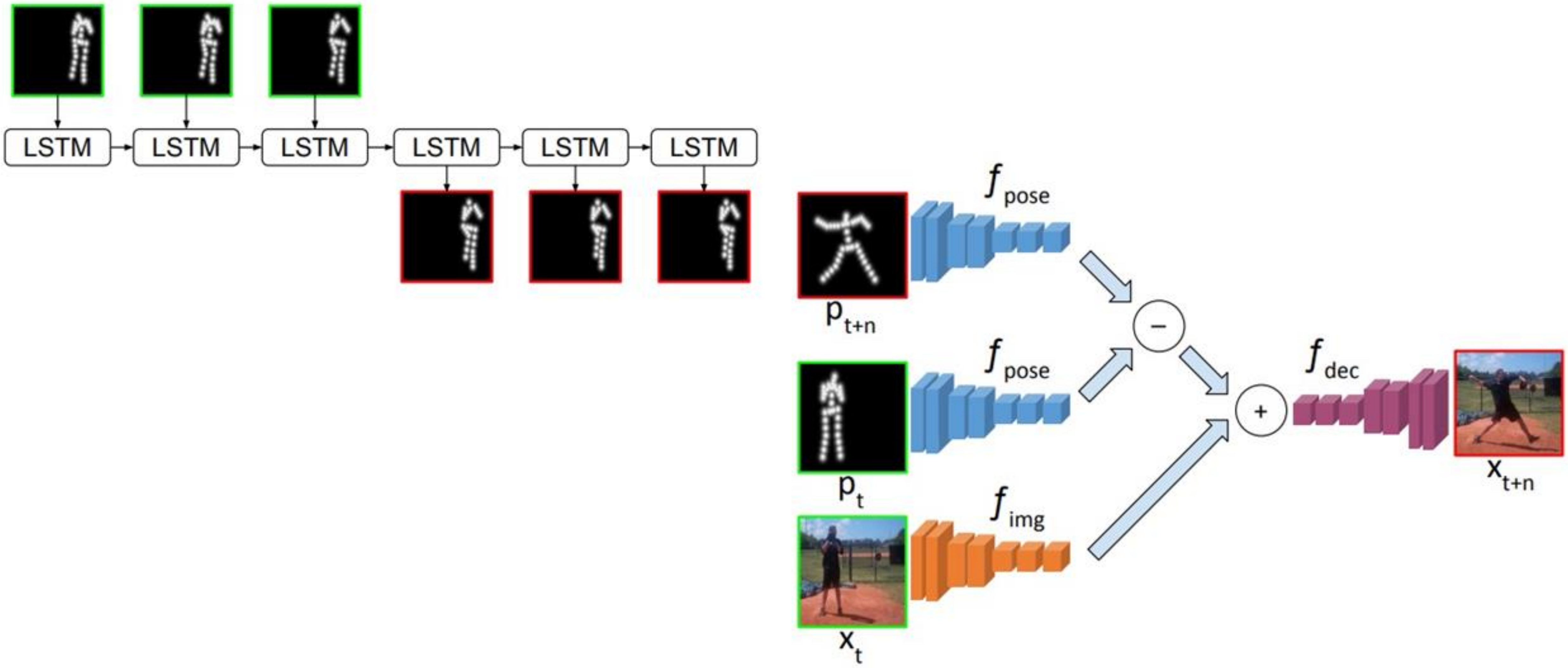


Predict Future Pose



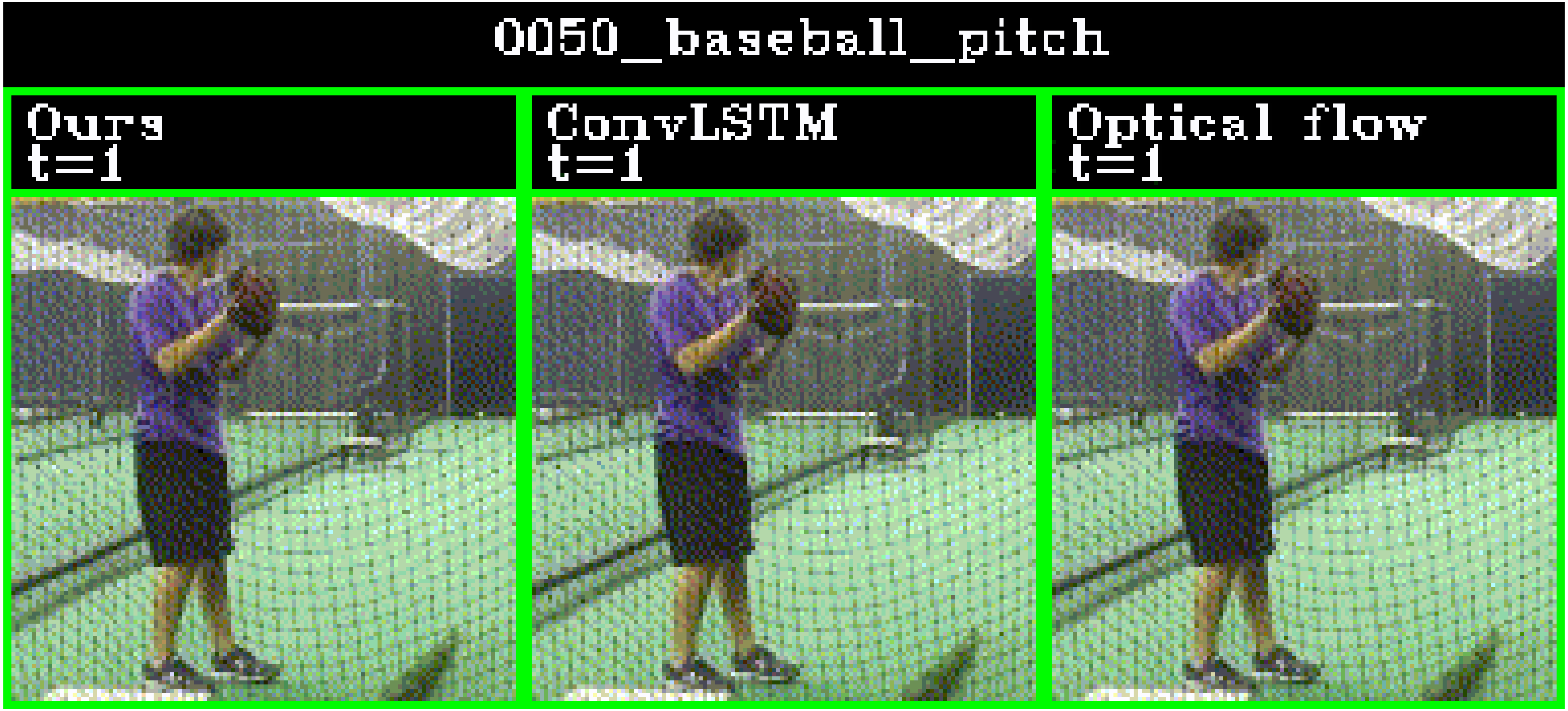
Villegas et al. Learning to Generate Long-term Future via Hierarchical Prediction. 2017.

Method

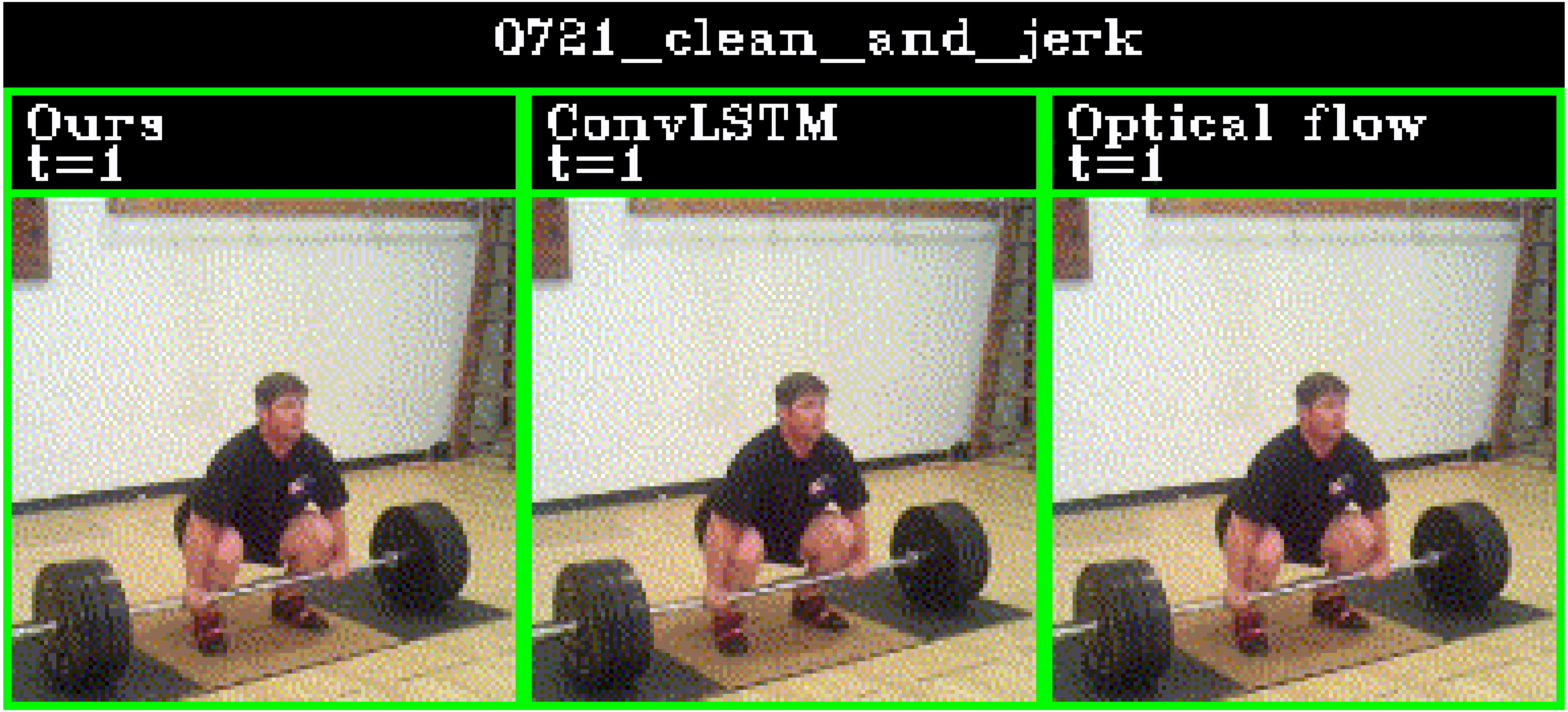


LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
f_{pose}
f_{pose}
f_{img}
f_{dec}
p_{t+n}
p_t
x_t
x_{t+n}
−
+

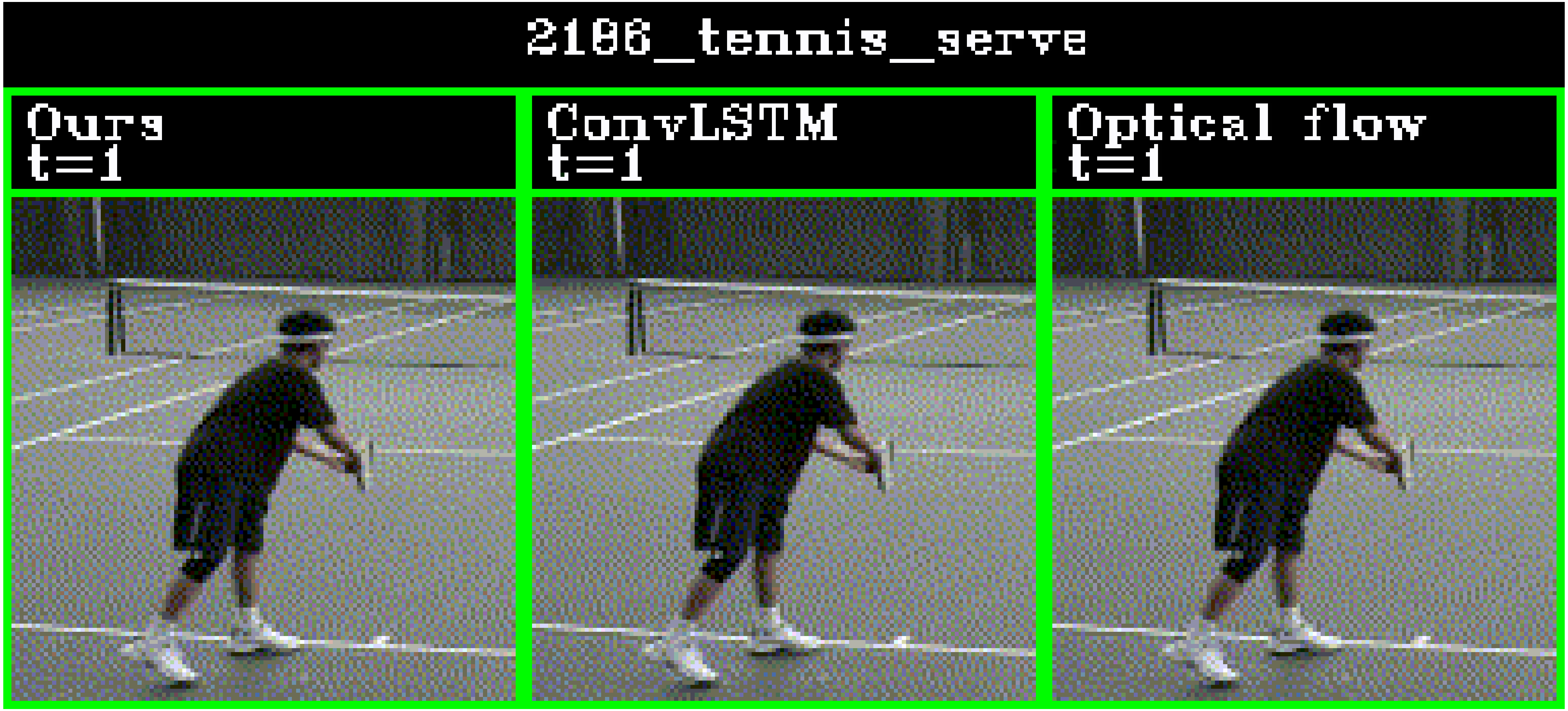
Results



Results

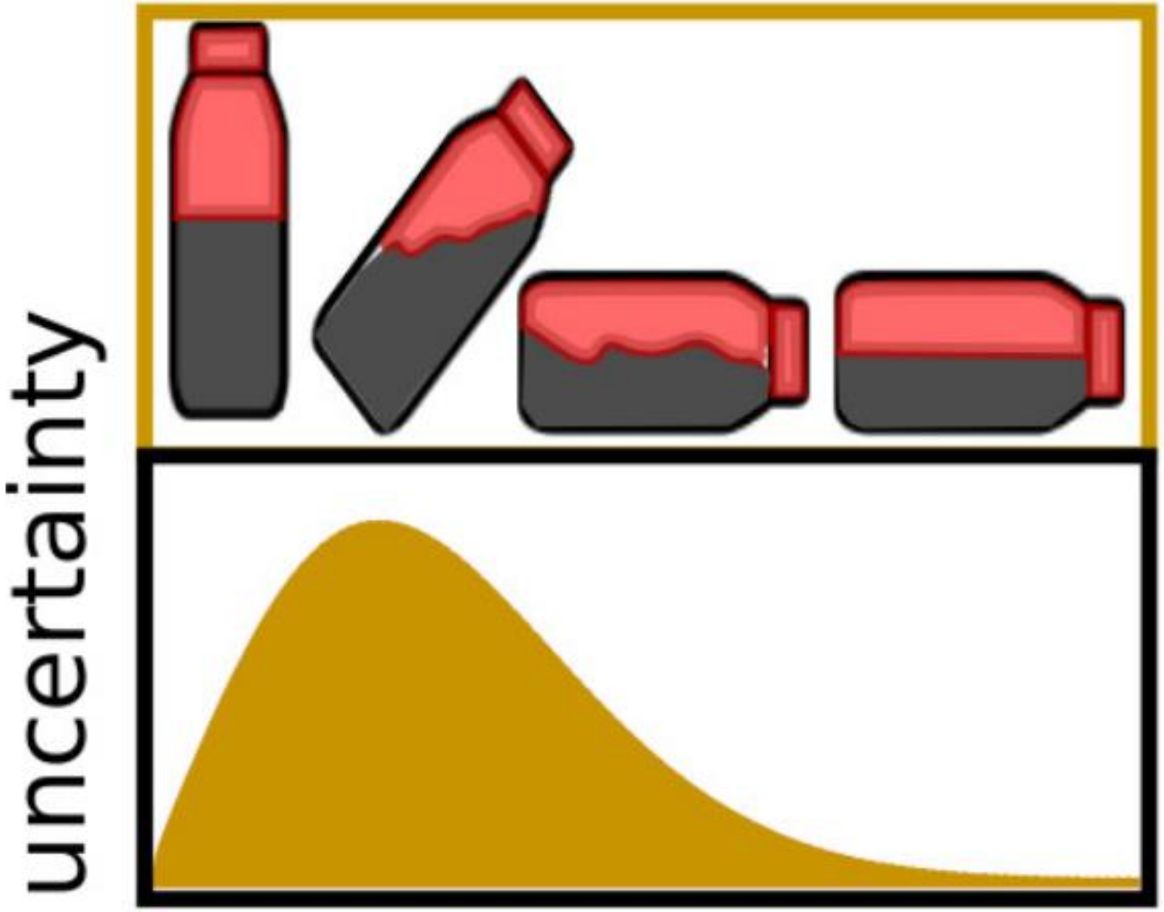


Results



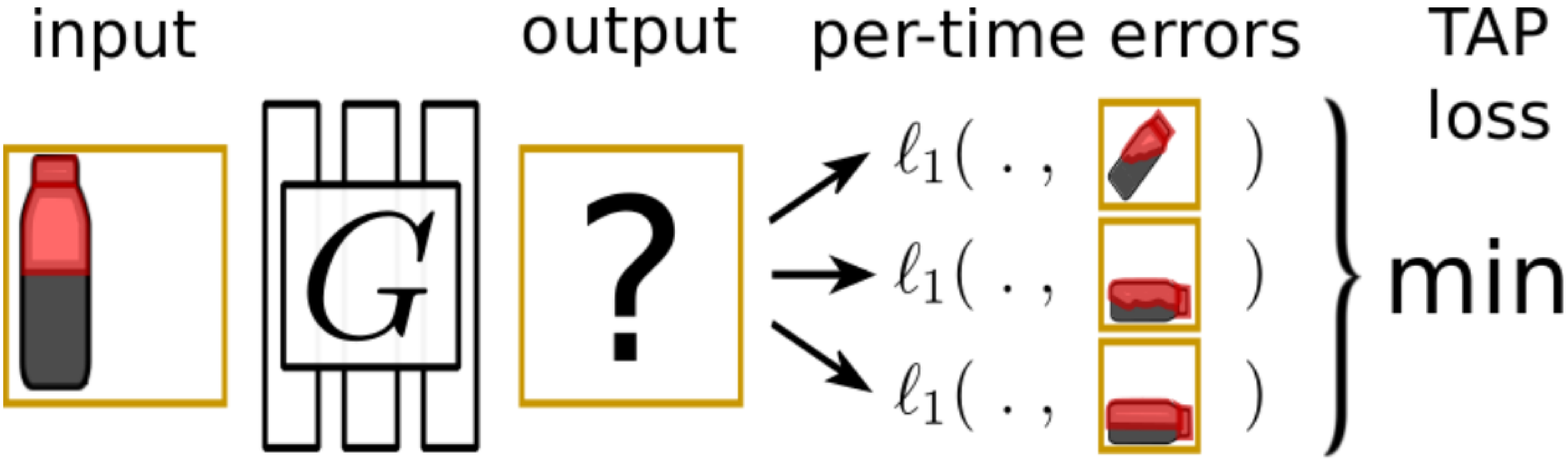
What Time to Predict

Uncertainty in Time



Jayaraman et al. Time-Agnostic Prediction: Predicting Predictable Video Frames, 2019.

Predict the Predictable Future



Predict the Predictable Future

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

Find a state with low uncertainty.

But it is unclear what exactly the T is in testing

Next Class

Self-Attention, GNN and Transformer