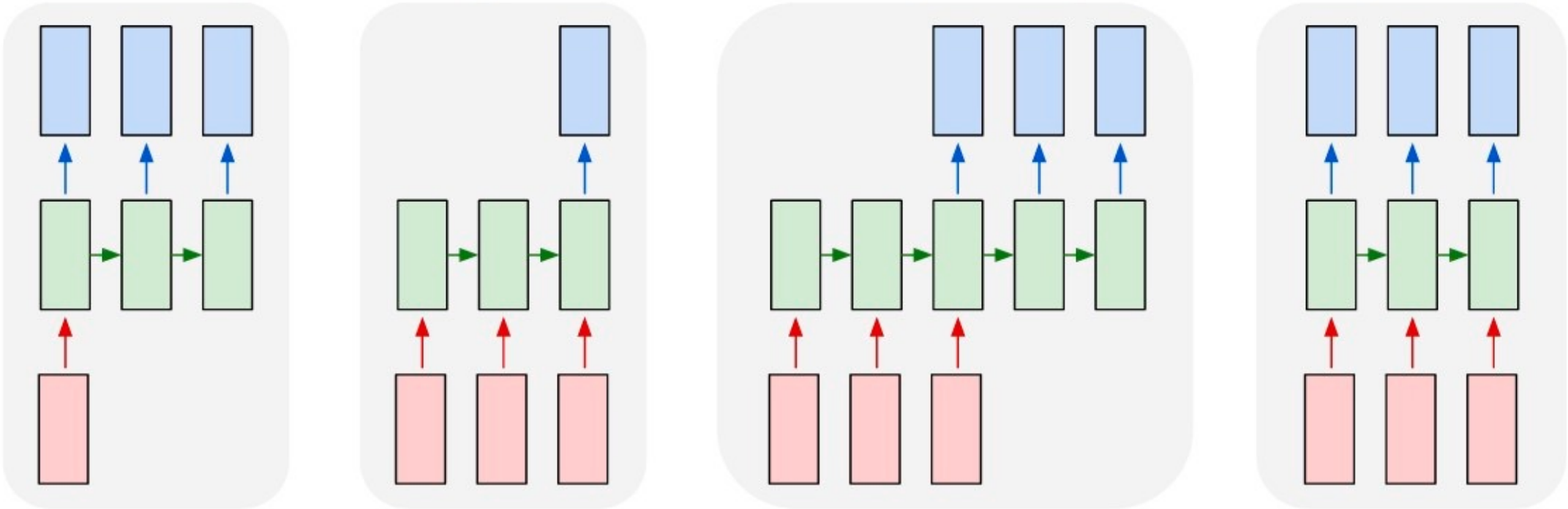
# Video Understanding

Xiaolong Wang

# Previous classes

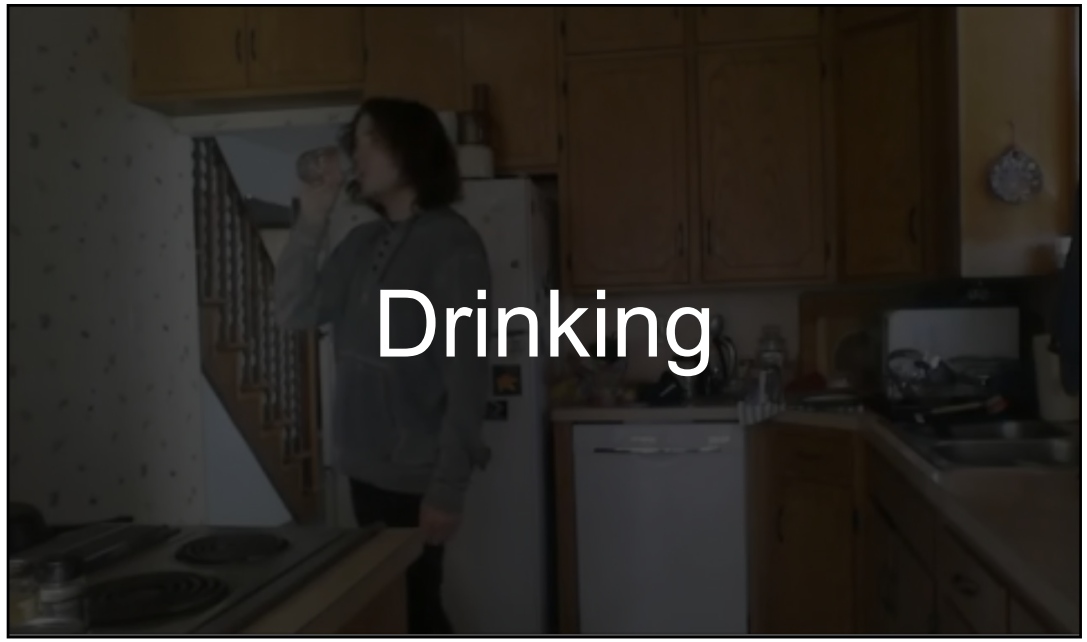


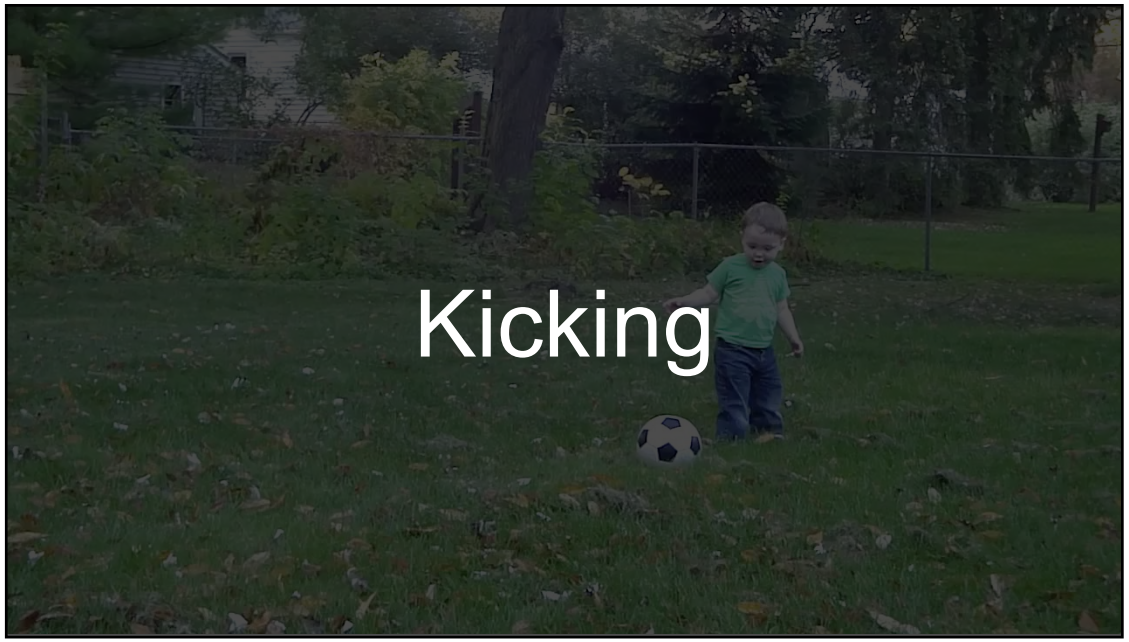
# This Class

- 2-Stream Networks for Action Recognition
- Temporal Convolution and 3D Convolution
- Video Prediction and Interaction Network

# 2-Stream Networks for Action Recognition

# Task: Action Recognition

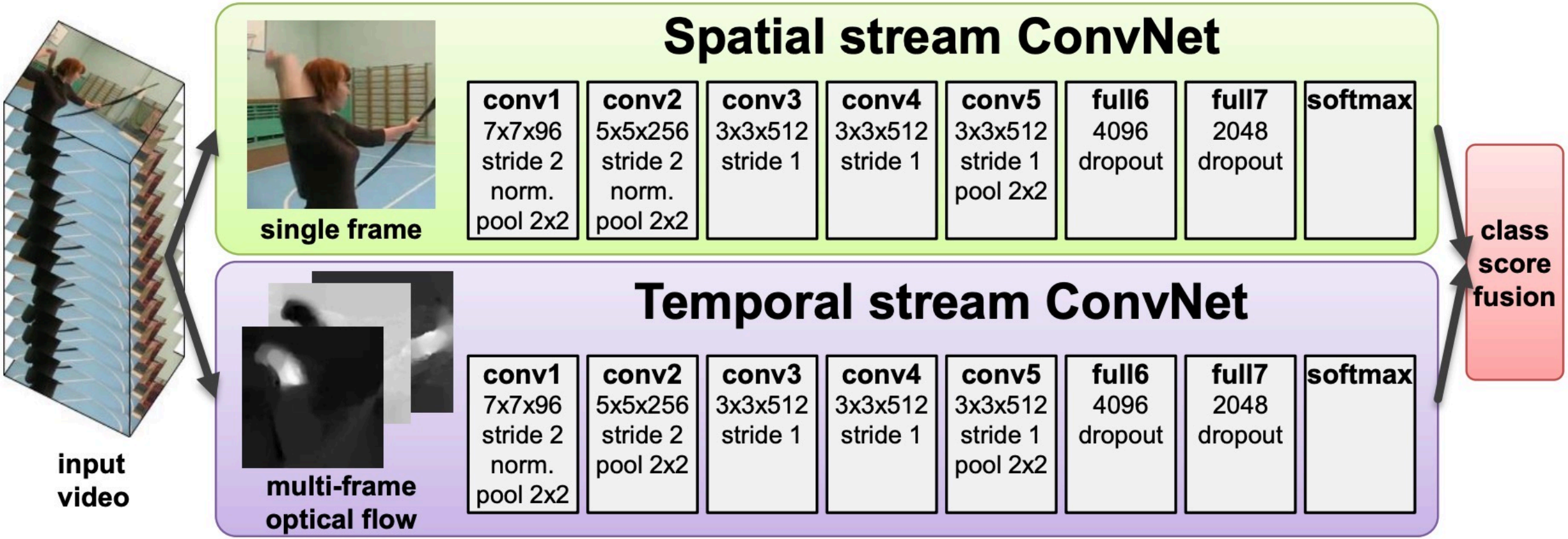


# Task: Action Recognition

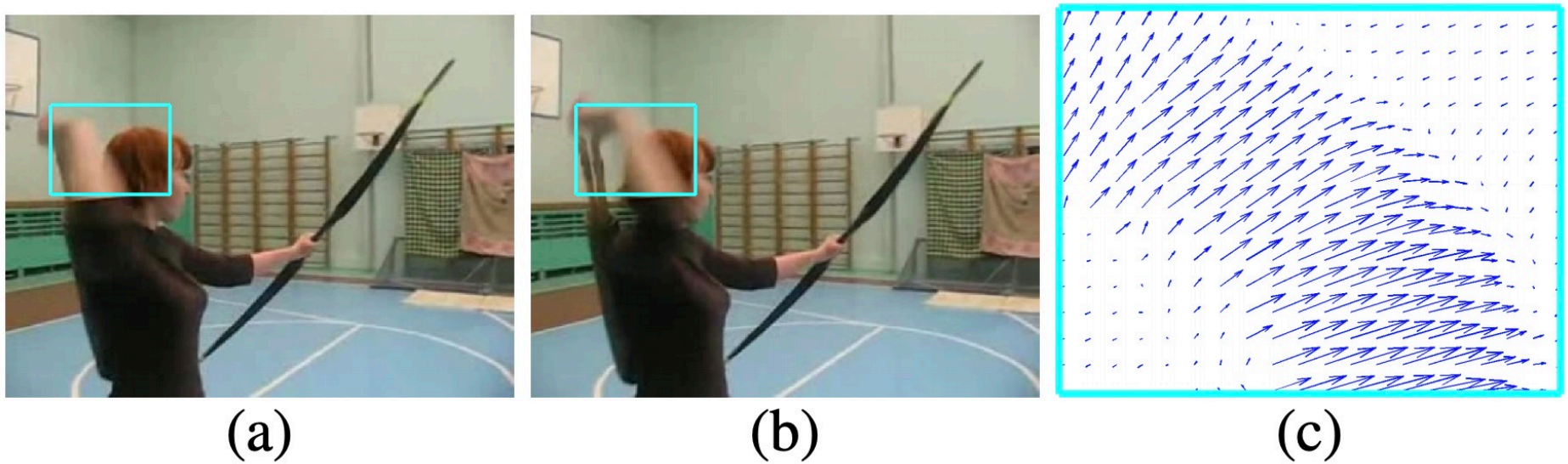
- UCF-101 dataset

# 2-Stream CNNs




Simonyan et al., 2014

# 2-Stream CNNs



# 2-Stream CNNs

How to sample frames in test time

- Given a video, sample 10 frames with equal distance between every two frames

- For example, given a video with 200 frames, we sample frame 1, 21, 41, … , 200 frame as inputs and forward 10 times

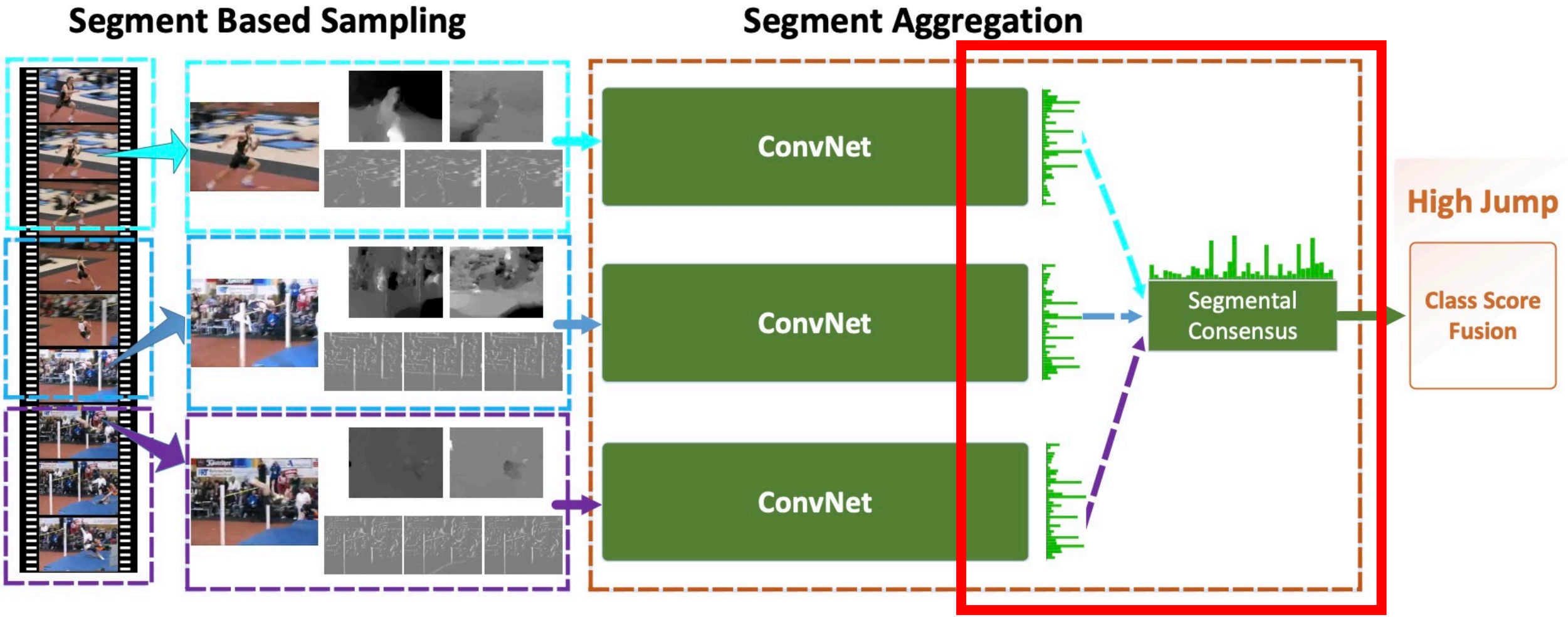
# 2-Stream CNNs


| | |
|---|---|
| Spatial stream ConvNet | 73.0% |
| Temporal stream ConvNet | 83.7% |
| Two-stream model (fusion by averaging) | 86.9% |
| Two-stream model (fusion by SVM) | **88.0%** |

# Temporal Segment Networks (TSN)

- In the previous work, we train each frame individually
- Can we train multiple frames at the same time?

# Temporal Segment Networks (TSN)

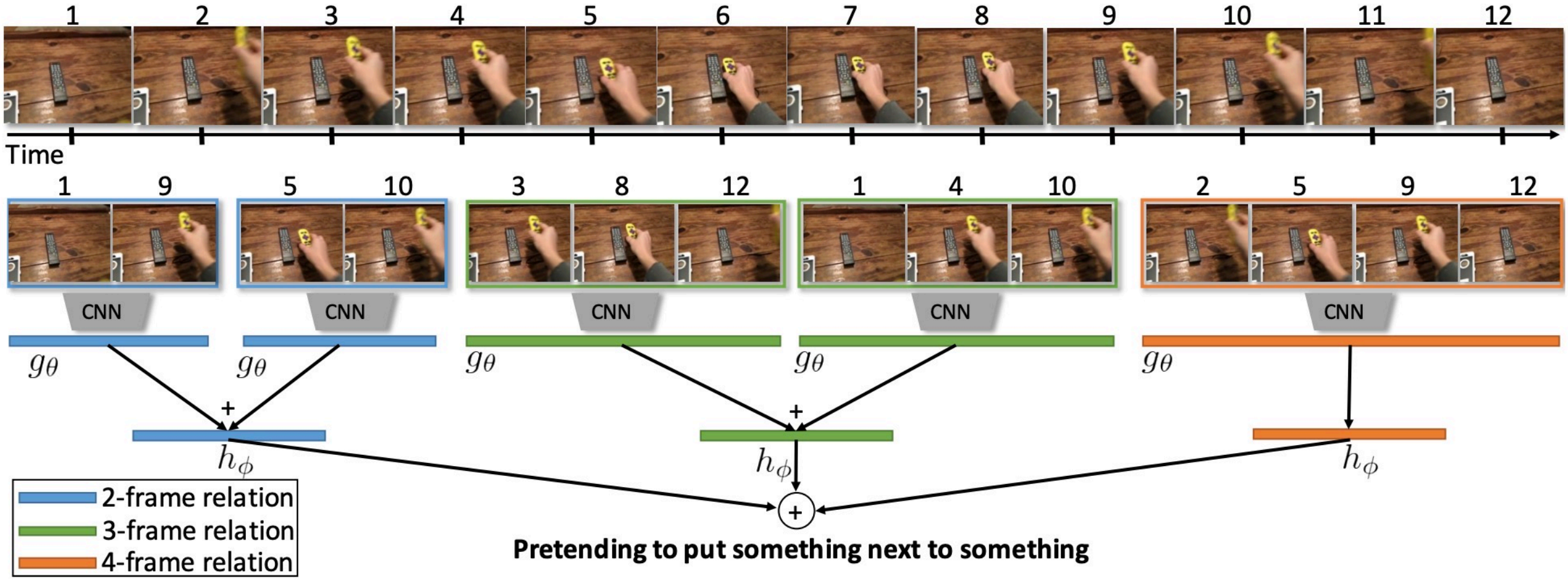



Wang et al., 2016

# Temporal Segment Networks (TSN)

| Modalities | TSN | Accuracy | Speed (FPS) |
|---|---|---|---|
| RGB+Flow | No | 92.4% | 14 |
| RGB+Flow | Yes | 94.9% | 14 |

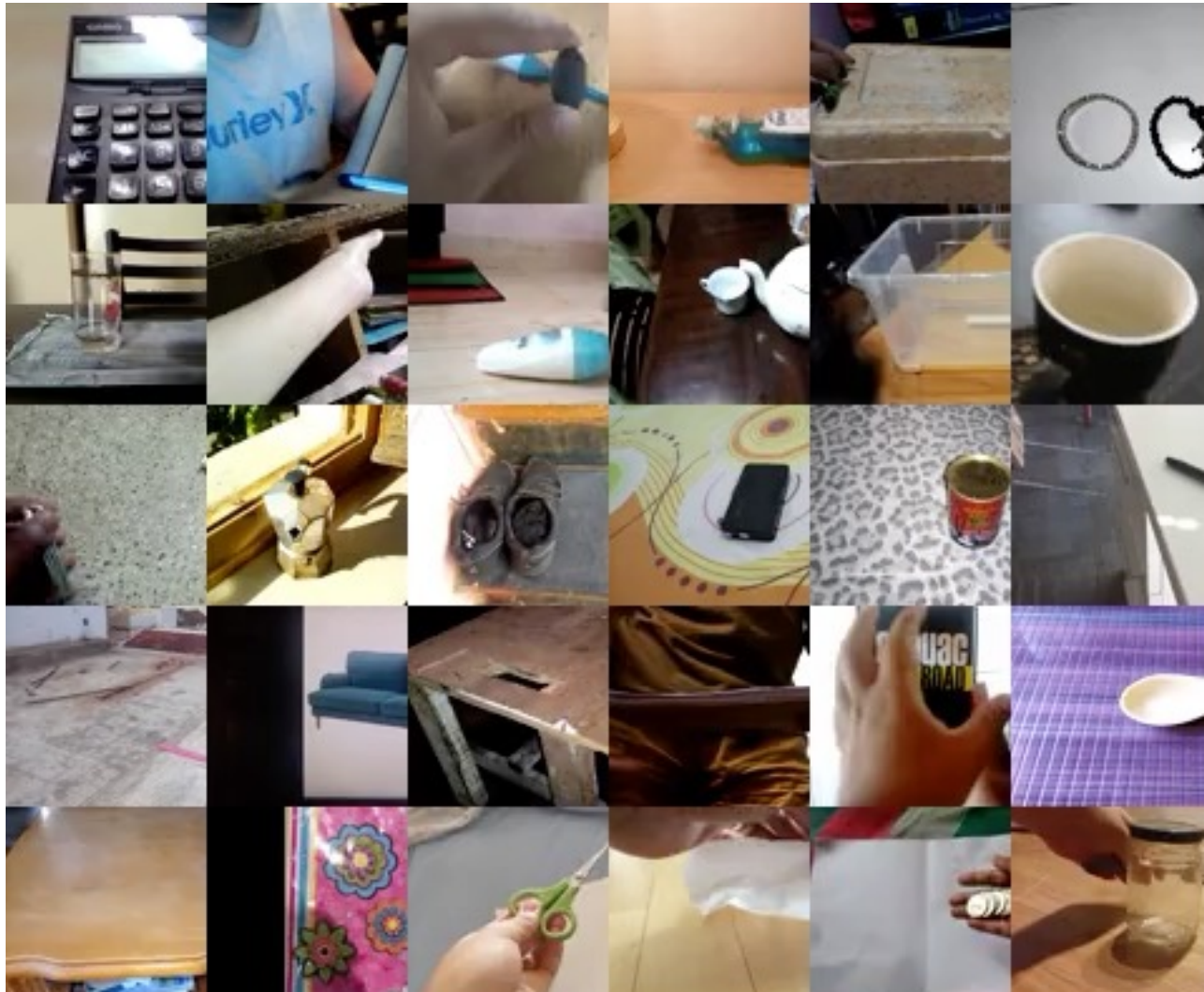
# Temporal Relation Network (TRN)




Zhou et al., 2018

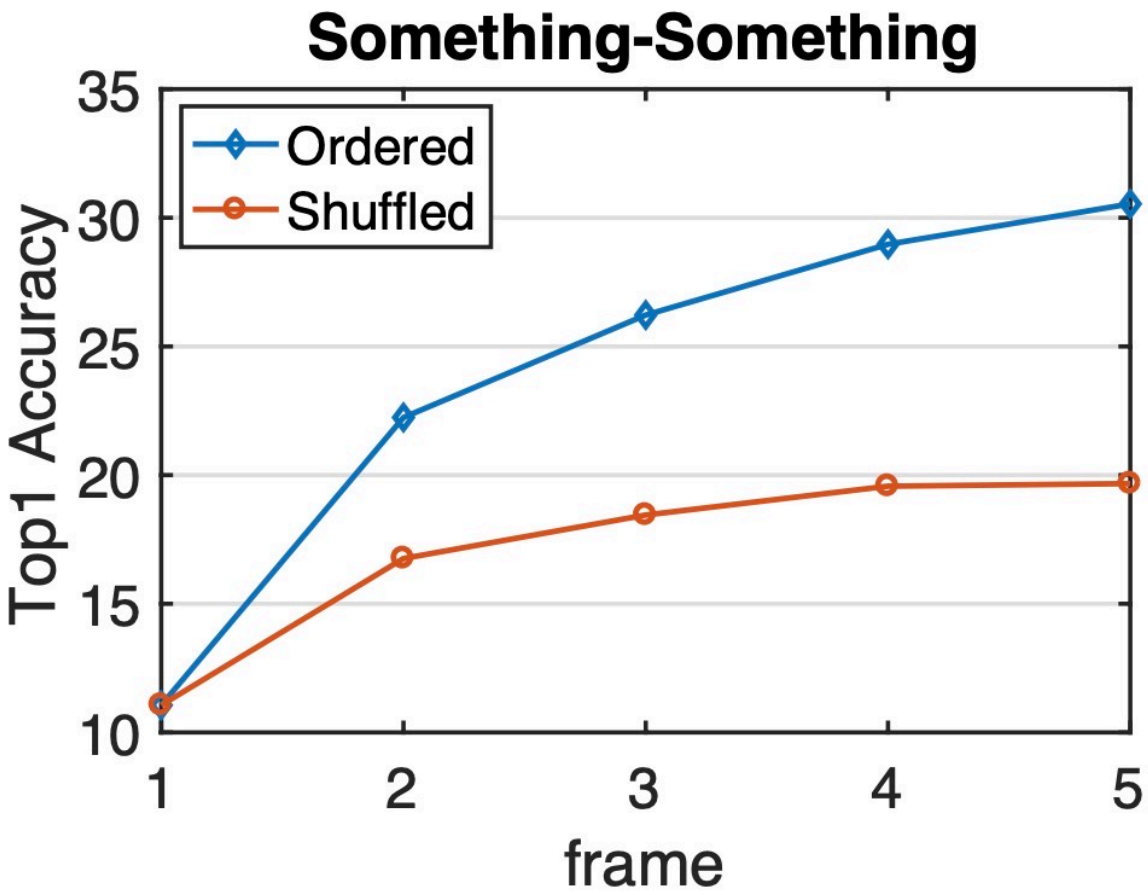
# Something-Something Dataset

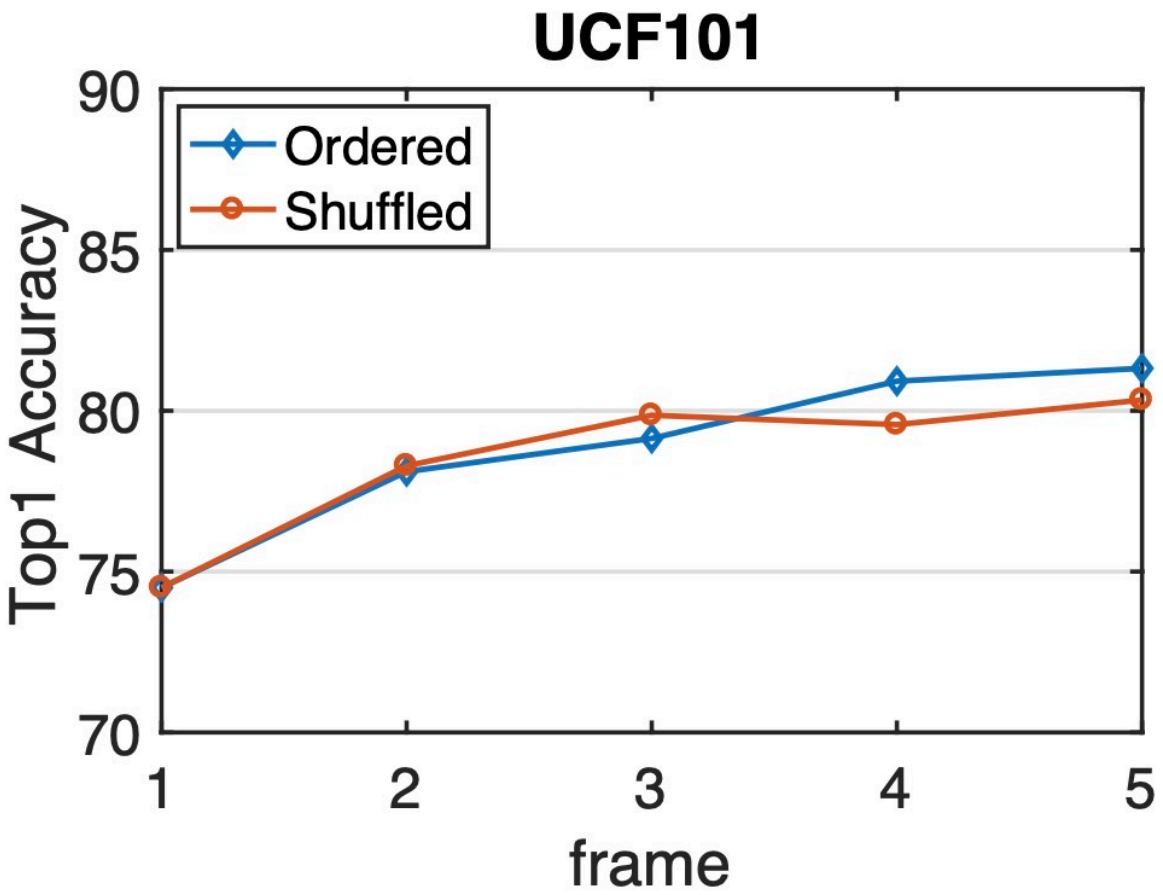
| Classes | |
|---|---|
| Putting something on a surface | 4,081 |
| Moving something up | 3,750 |
| Covering something with something | 3,530 |
| Pushing something from left to right | 3,442 |
| Moving something down | 3,242 |
| Pushing something from right to left | 3,195 |
| Uncovering something | 3,004 |
| Taking one of many similar things on the table | 2,969 |
| Turning something upside down | 2,943 |
| Tearing something into two pieces | 2,849 |
| Putting something into something | 2,783 |
| Squeezing something | 2,631 |



# The problem of Action Recognition

# Temporal Relation Network (TRN)



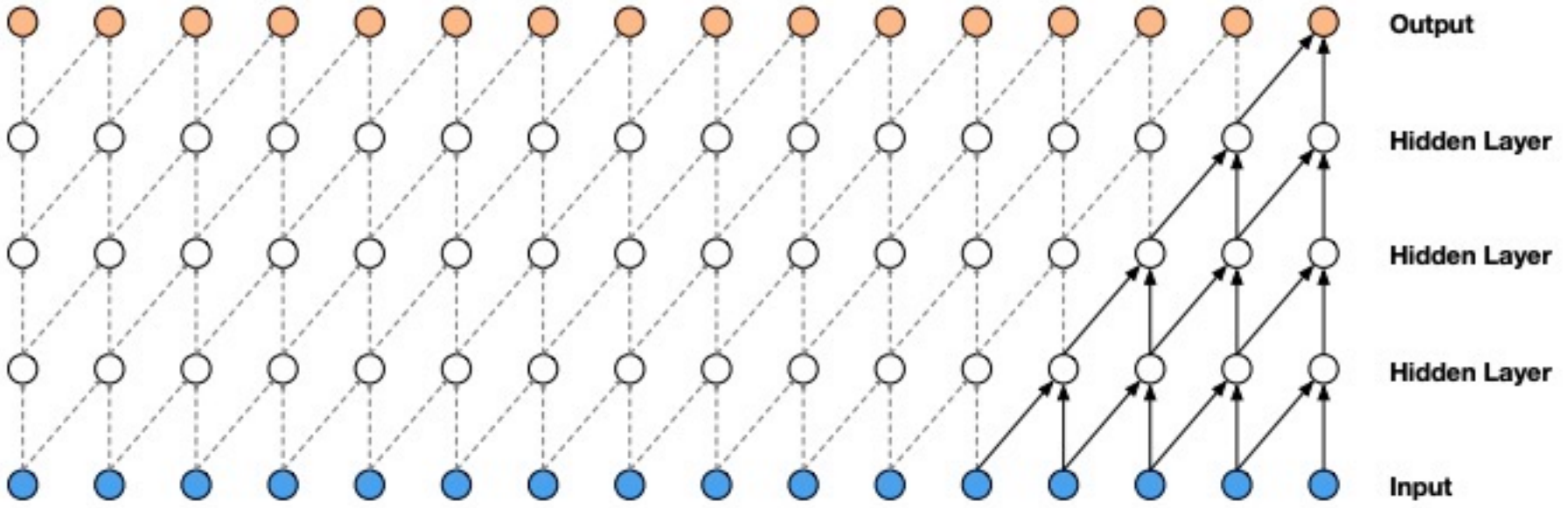
# Short summary

- Basic 2-Stream, train on each frame individually → temporal order does not matter
- TSN, use average pooling to aggregate video frames during training → temporal order does not matter
- TRN, use concatenation and FC to aggregate video frames during training → temporal order matters

# Temporal Convolution and 3D Convolution

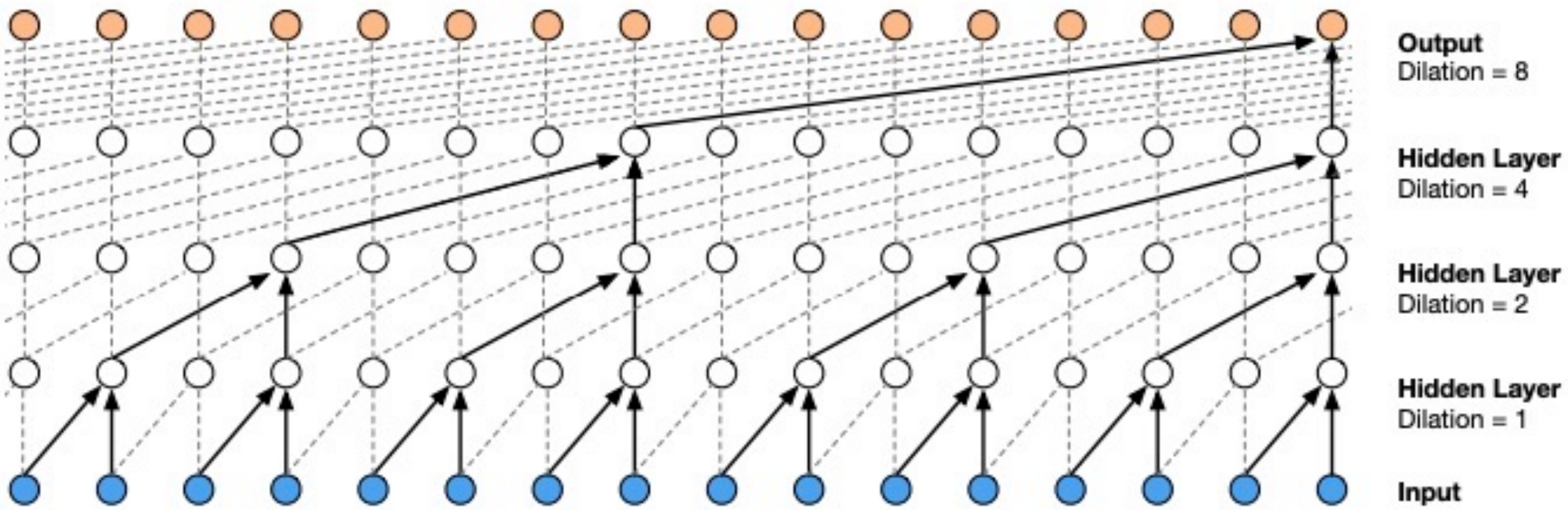
# Temporal Convolution

Figure 1: A second of generated speech.

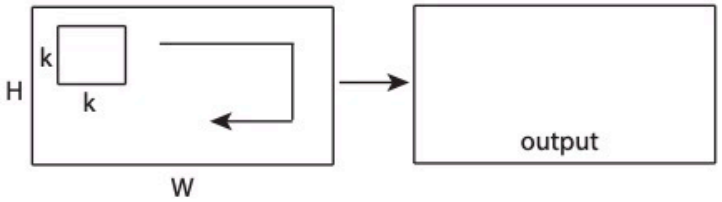



Van den Oord et al., 2016

# Temporal Convolution



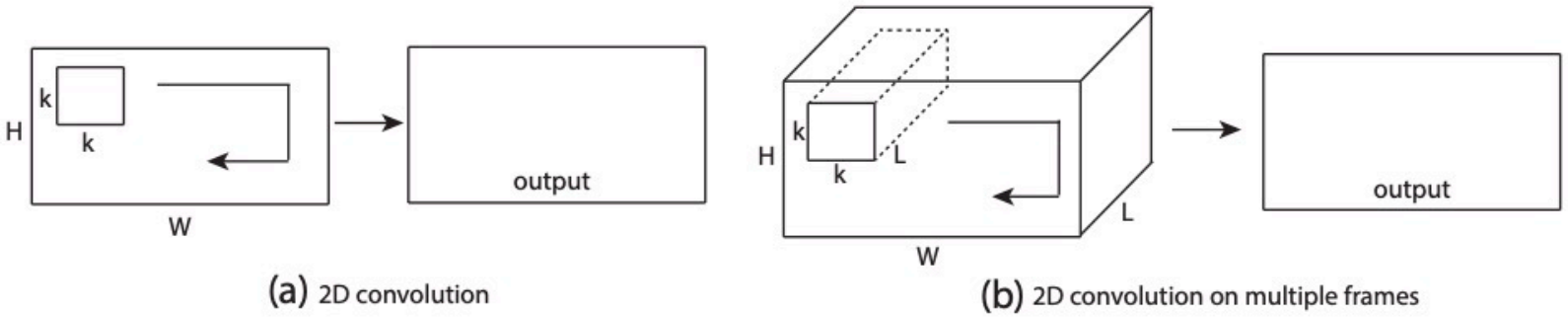
# 3D Convolution



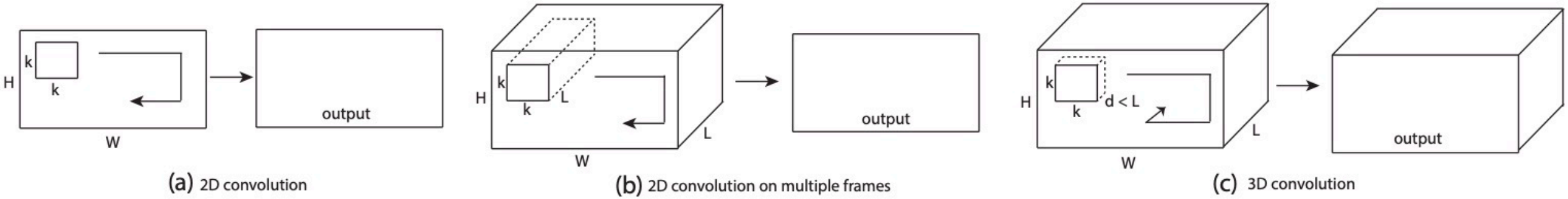

(a) 2D convolution

Train et al., 2015

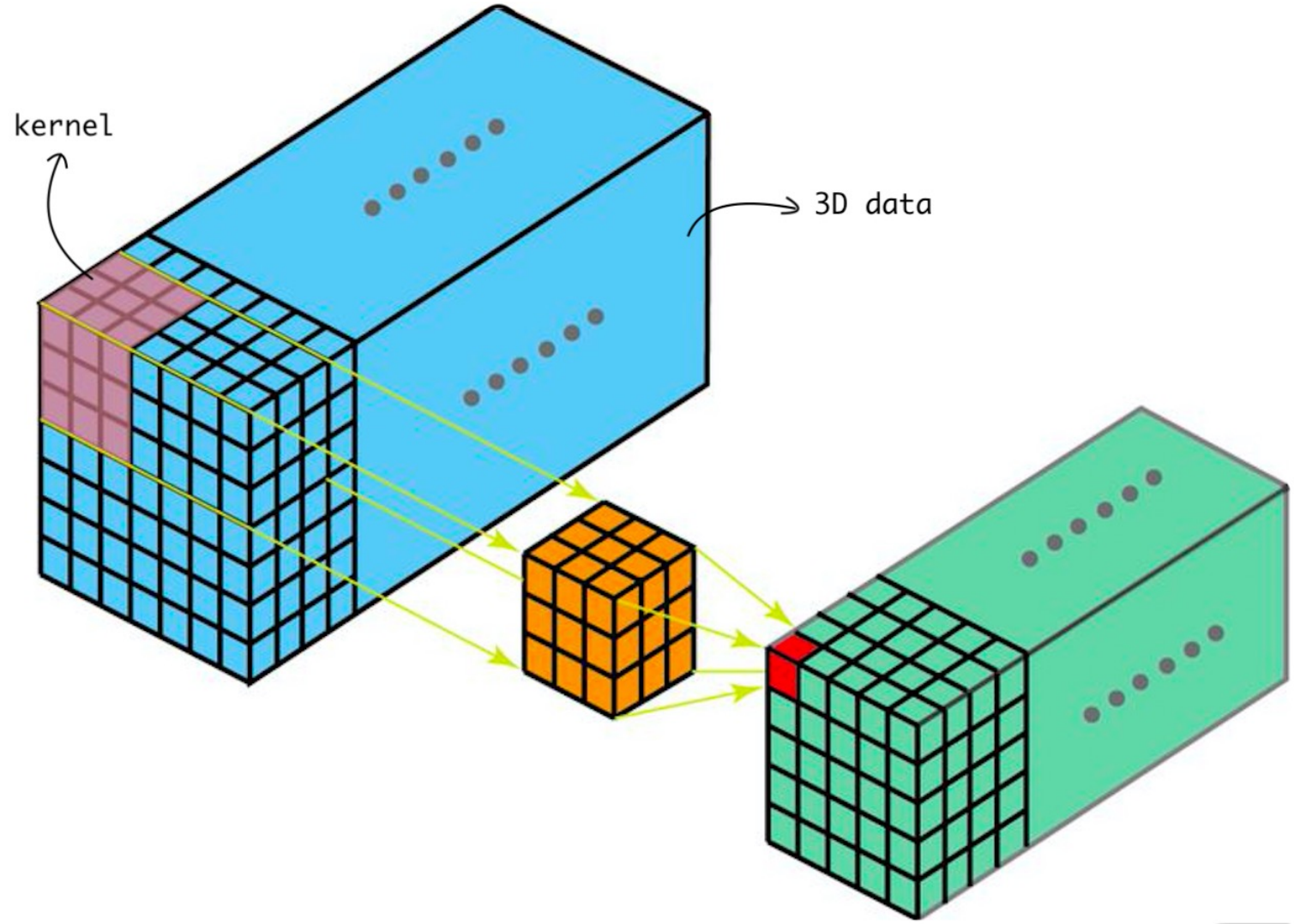
# 3D Convolution



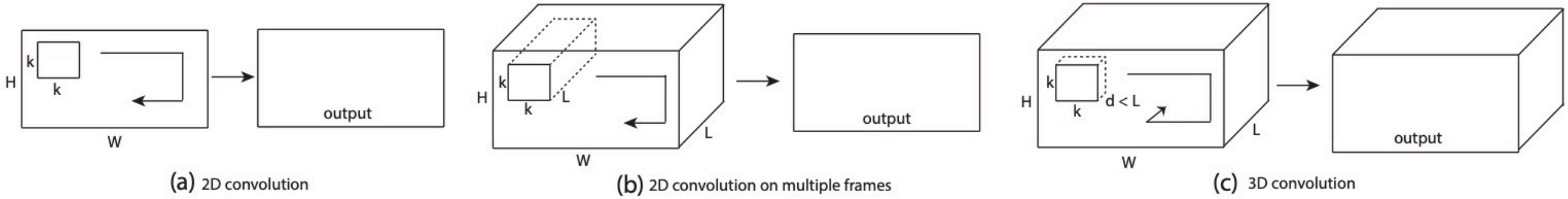
# 3D Convolution

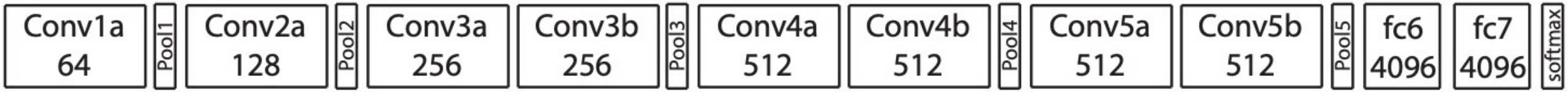


# 3D Convolution

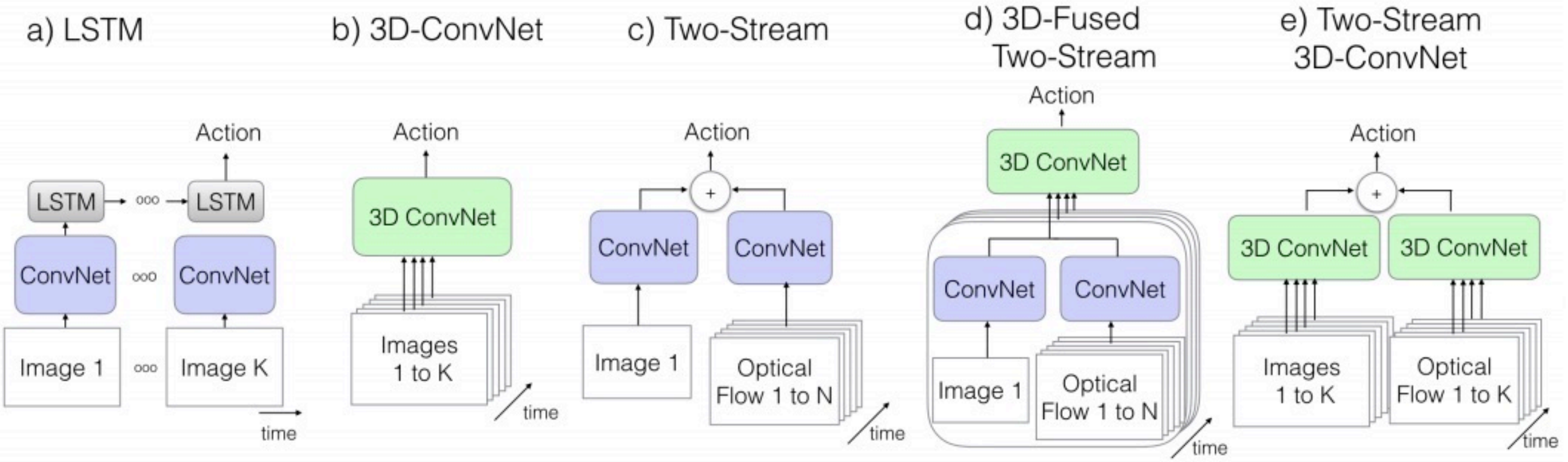


# 3D Convolution



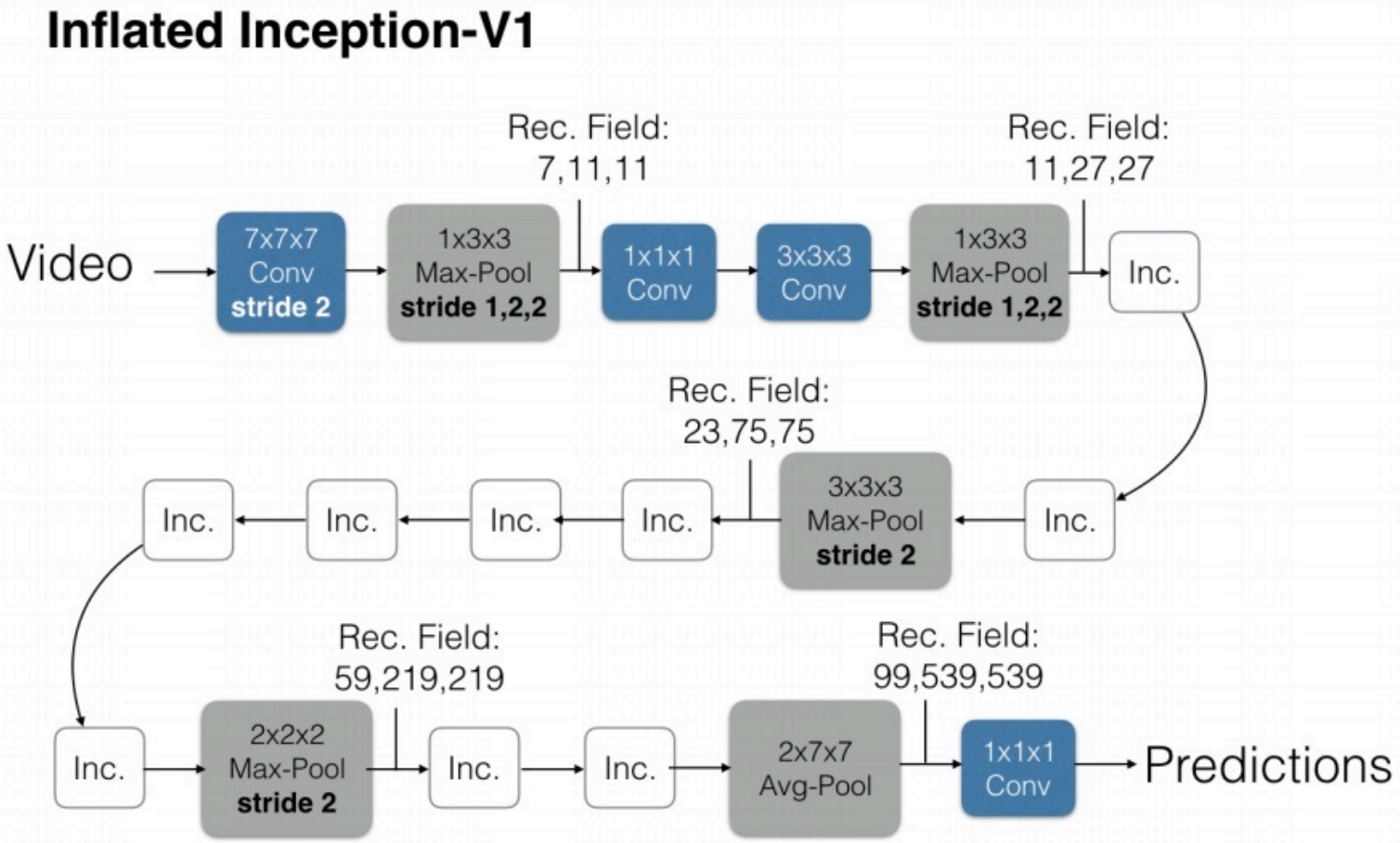



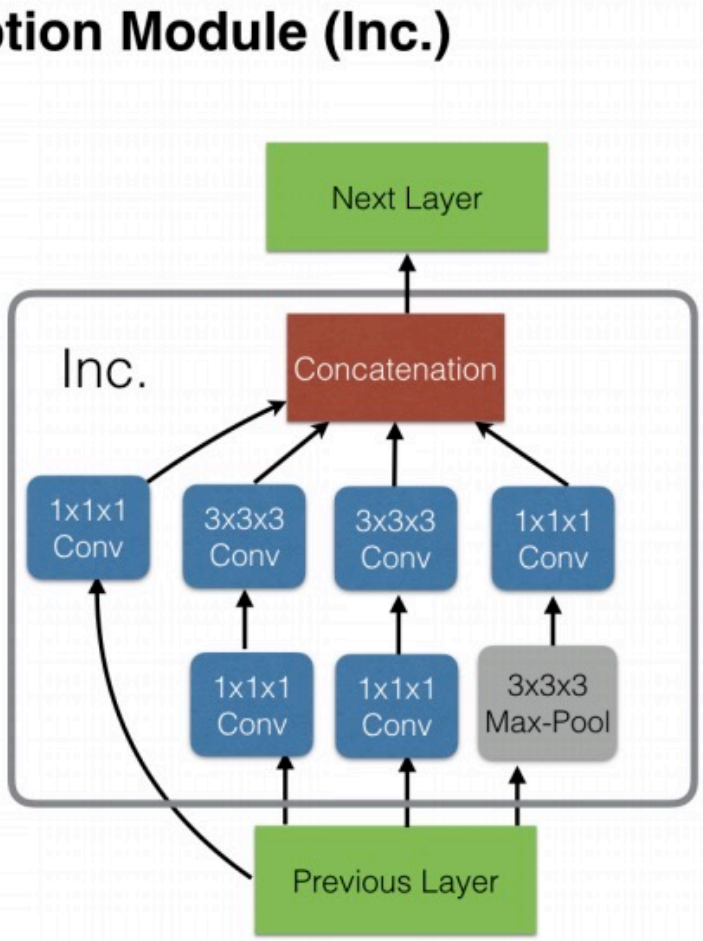
# Inflated 3D ConvNets (I3D)




Carreira et al., 2018

# Inflated 3D ConvNets (I3D)



# Kinetics Dataset

(a) headbanging

(c) shaking hands

# Kinetics Dataset

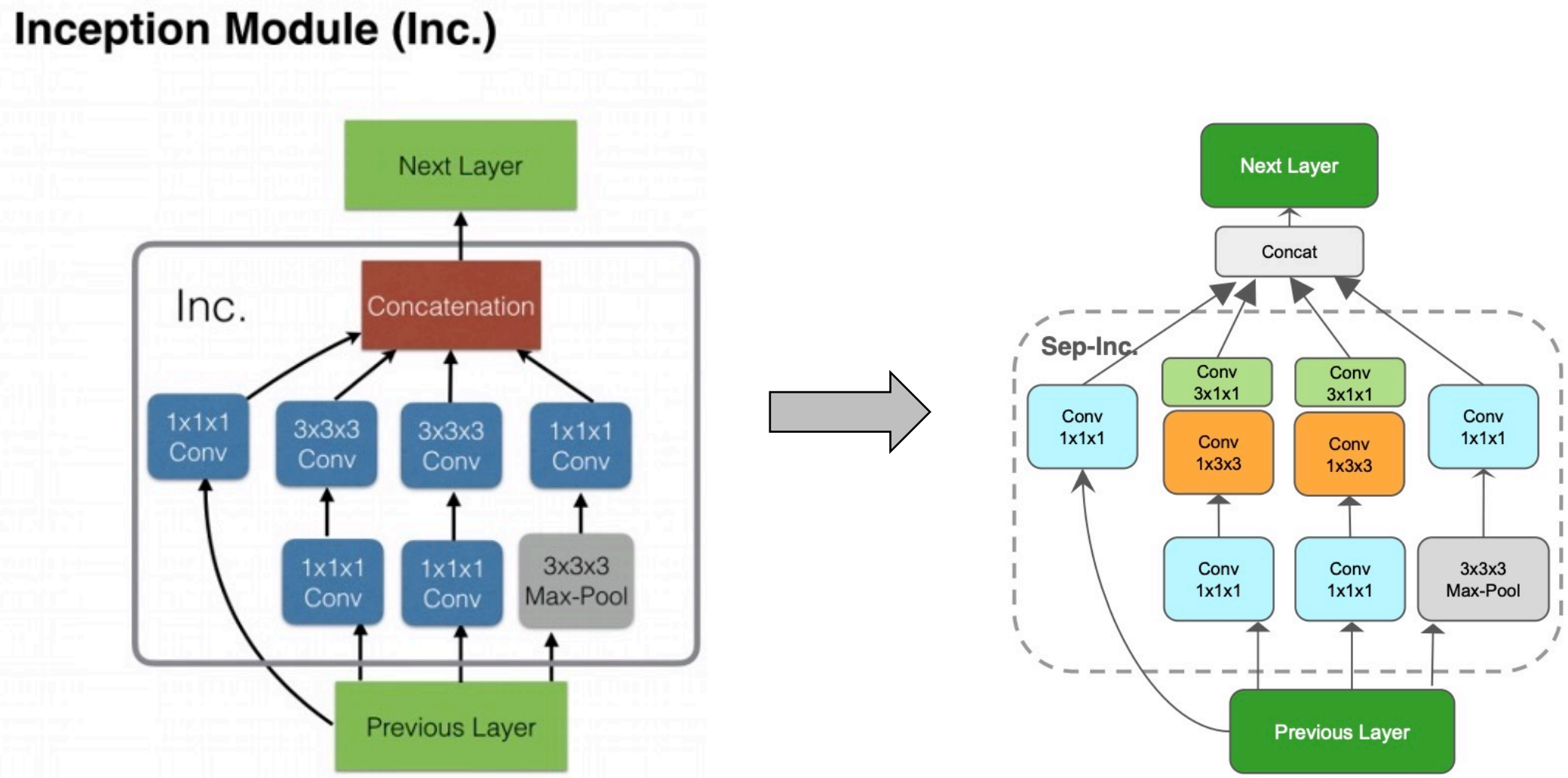
(b) stretching leg

(d) tickling

# Inflated 3D ConvNets (I3D)

| Architecture | UCF-101 | | | HMDB-51 | | | Kinetics | | |
|---|---|---|---|---|---|---|---|---|---|
| | RGB | Flow | RGB + Flow | RGB | Flow | RGB + Flow | RGB | Flow | RGB + Flow |
| (a) LSTM | 81.0 | – | – | 36.0 | – | – | 63.3 | – | – |
| (b) 3D-ConvNet | 51.6 | – | – | 24.3 | – | – | 56.1 | – | – |
| (c) Two-Stream | 83.6 | 85.6 | 91.2 | 43.2 | 56.3 | 58.3 | 62.2 | 52.4 | 65.6 |
| (d) 3D-Fused | 83.2 | 85.8 | 89.3 | 49.2 | 55.5 | 56.8 | – | – | 67.2 |
| (e) Two-Stream I3D | **84.5** | **90.6** | **93.4** | **49.8** | **61.9** | **66.4** | **71.1** | **63.4** | **74.2** |

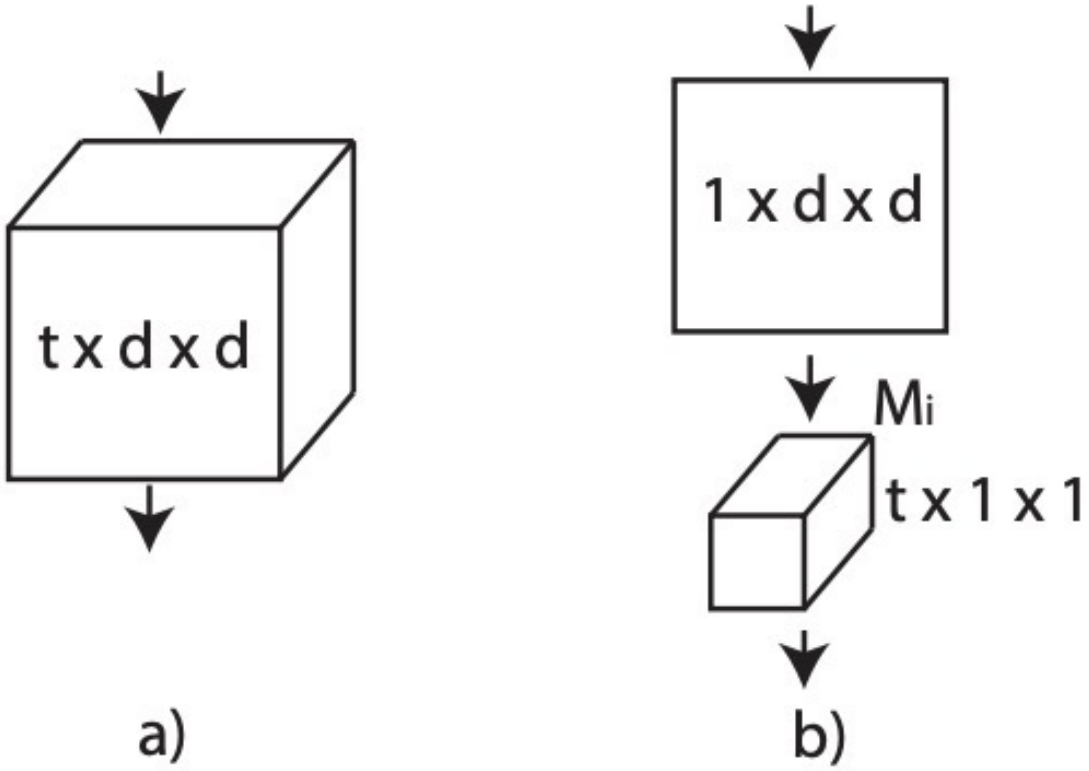
# Separable 3D CNN (S3D)



# Separable 3D CNN (S3D)

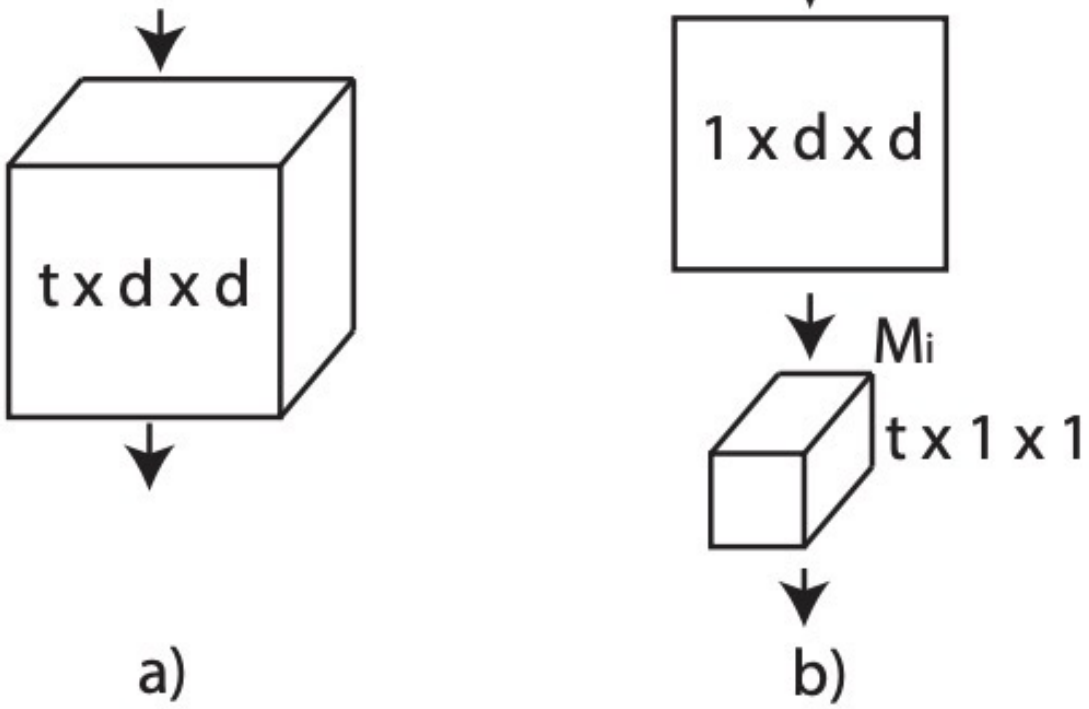
| Model | Top-1 (%) | Top-5 (%) | Params (M) | FLOPS (G) |
|---|---|---|---|---|
| I3D | 71.1 | 89.3 | 12.06 | 107.89 |
| S3D | 72.2 | 90.6 | 8.77 | 66.38 |
| S3D-G | **74.7** | **93.4** | 11.56 | 71.38 |

# R(2+1)D




Tran et al., 2018

# R(2+1)D




Tran et al., 2018

# How about using a 3D Network with only 2D Conv?

| layer | | output size |
|---|---|---|
| $\text{conv}_1$ | 7×7, 64, stride 2, 2, 2 | 16×112×112 |
| $\text{pool}_1$ | 3×3×3 max, stride 2, 2, 2 | 8×56×56 |
| $\text{res}_2$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times3$ | 8×56×56 |
| $\text{pool}_2$ | 3×1×1 max, stride 2, 1, 1 | 4×56×56 |
| $\text{res}_3$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times4$ | 4×28×28 |
| $\text{res}_4$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times6$ | 4×14×14 |
| $\text{res}_5$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times3$ | 4×7×7 |
| global average pool, fc | | 1×1×1 |

Wang et al., 2018

# How much does temporal convolution matters?

Same network, remove all temporal conv

| model, R101 | params | FLOPs | top-1 | top-5 |
|---|---|---|---|---|
| C2D baseline | 1× | 1× | 73.1 | 91.0 |
| $I3D_{3\times3\times3}$ | 1.5× | 1.8× | 74.1 | 91.2 |
| $I3D_{3\times1\times1}$ | **1.2**× | 1.5× | 74.4 | 91.1 |

Wang et al., 2018

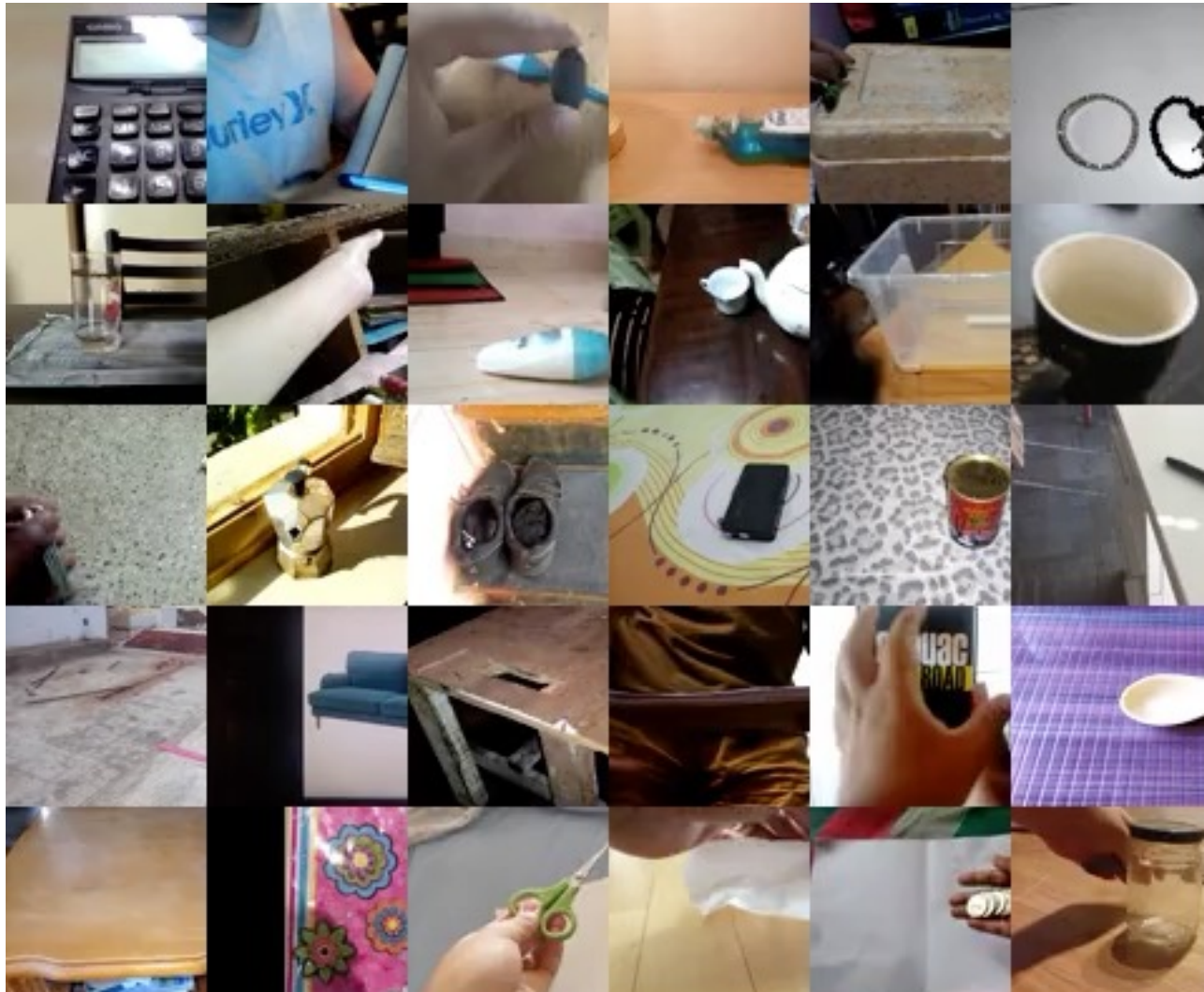
# The Problem is the Dataset

(a) headbanging

(c) shaking hands

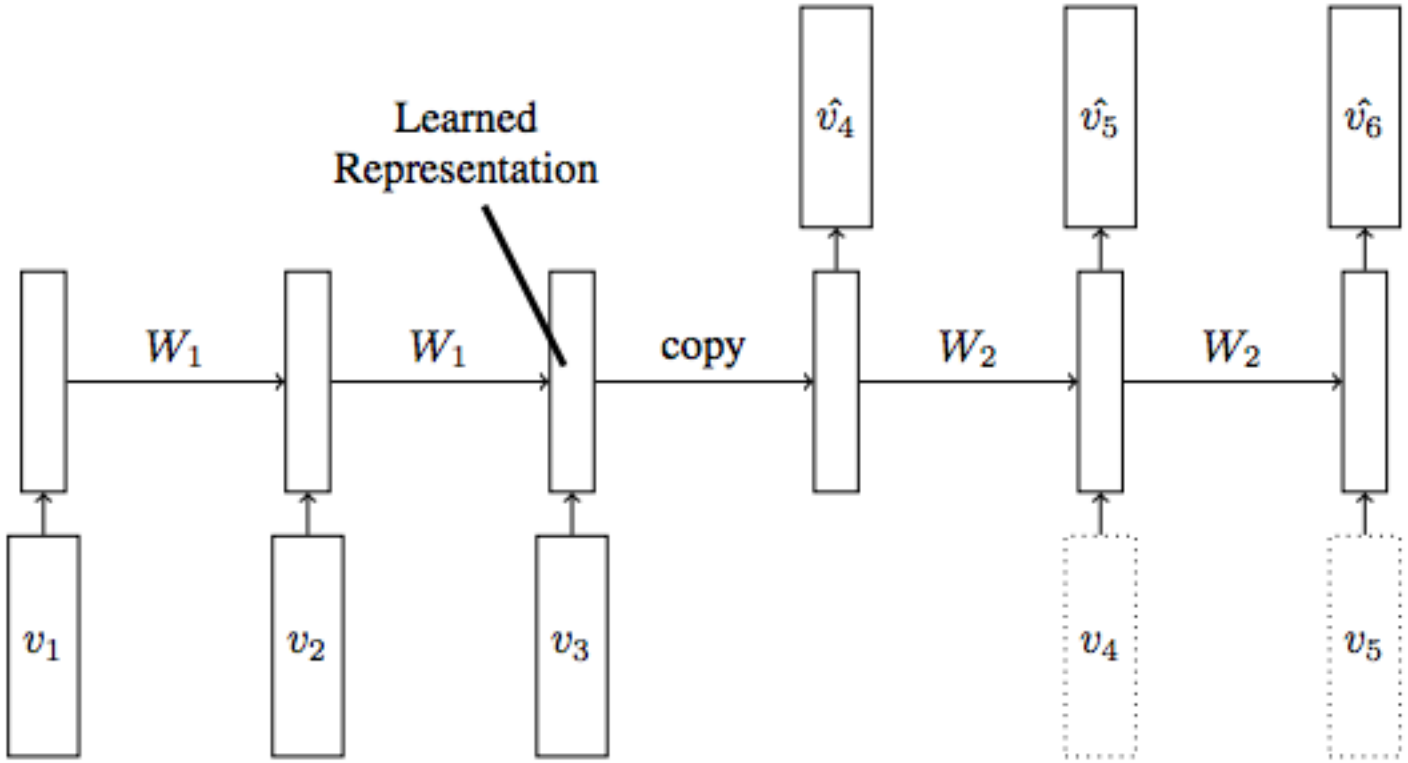
# Something-Something Dataset

| Classes | |
|---|---|
| Putting something on a surface | 4,081 |
| Moving something up | 3,750 |
| Covering something with something | 3,530 |
| Pushing something from left to right | 3,442 |
| Moving something down | 3,242 |
| Pushing something from right to left | 3,195 |
| Uncovering something | 3,004 |
| Taking one of many similar things on the table | 2,969 |
| Turning something upside down | 2,943 |
| Tearing something into two pieces | 2,849 |
| Putting something into something | 2,783 |
| Squeezing something | 2,631 |



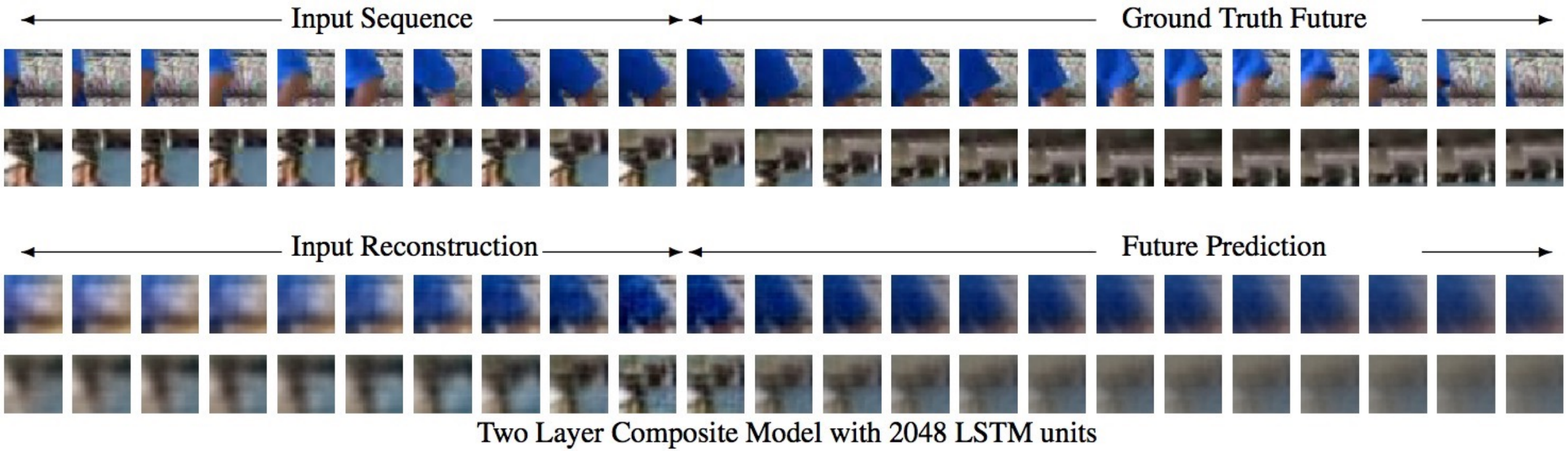
# Video Prediction and Interaction Network

# Visual Prediction in Time



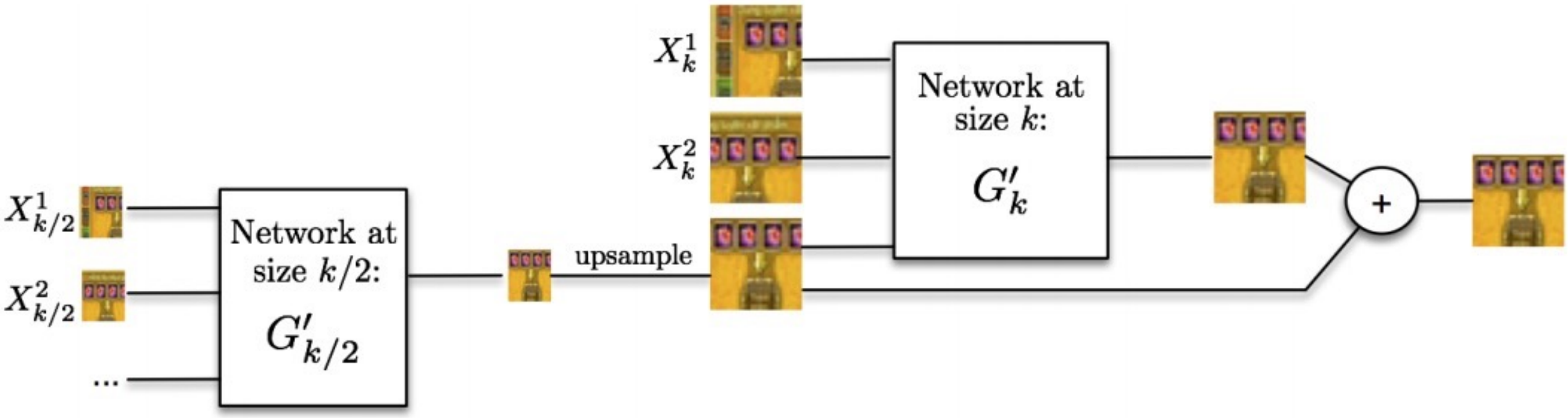

Srivastava et al., 2015

# Visual Prediction in Time




Srivastava et al., 2015

# Visual Prediction in Time

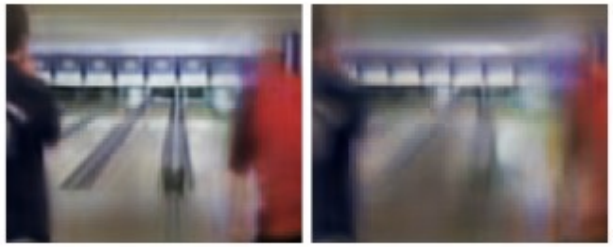



Mathieu et al., 2015

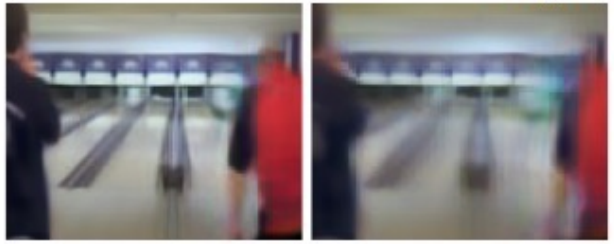
# Visual Prediction in Time

Input frames

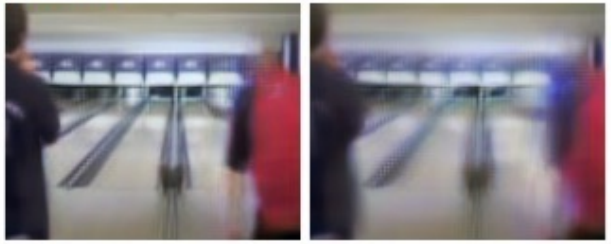
Ground truth



$\ell_2$ result



$\ell_1$ result



GDL $\ell_1$ result

Adversarial result

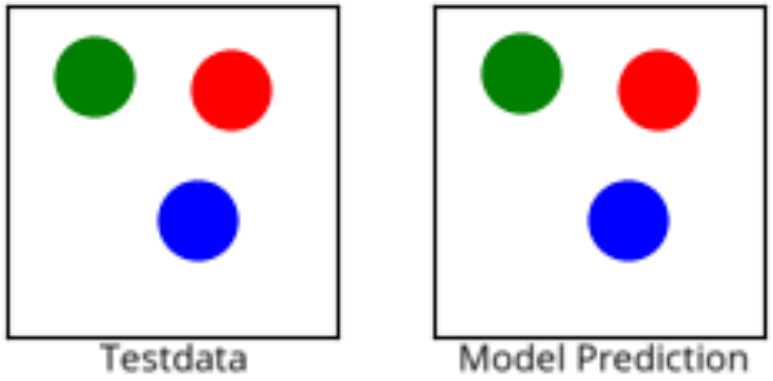
Adversarial+GDL result

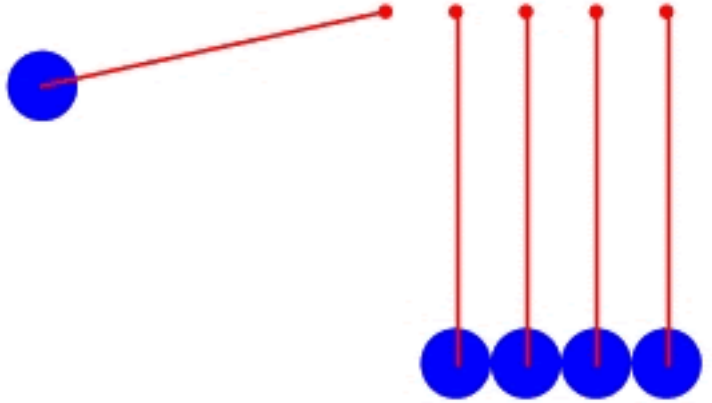
Mathieu et al., 2015

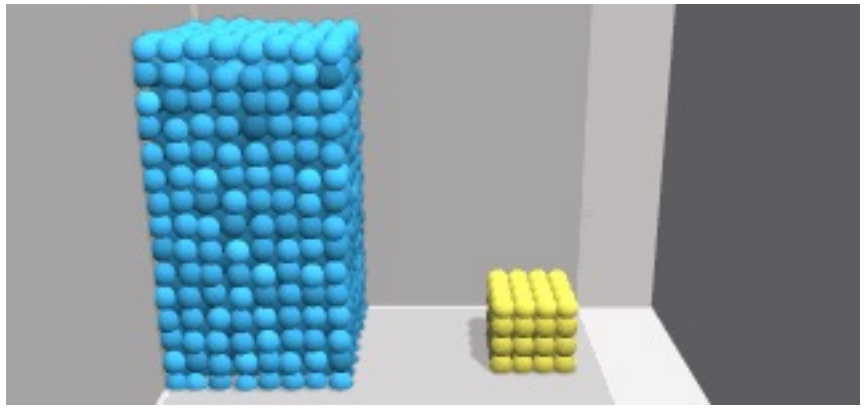
# Visual Prediction in Time

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

# Object Centric Prediction in a Physical World

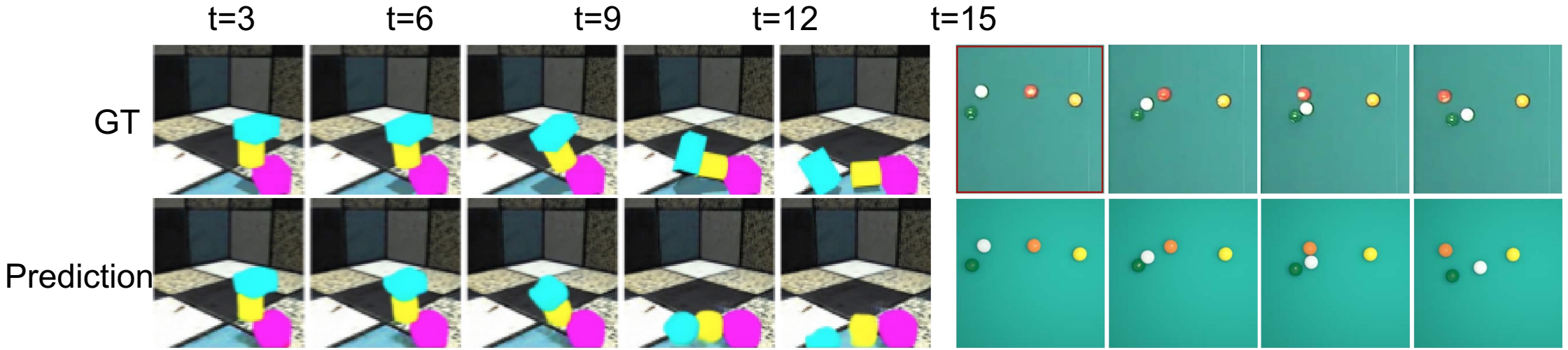




# Current Status of object dynamics prediction

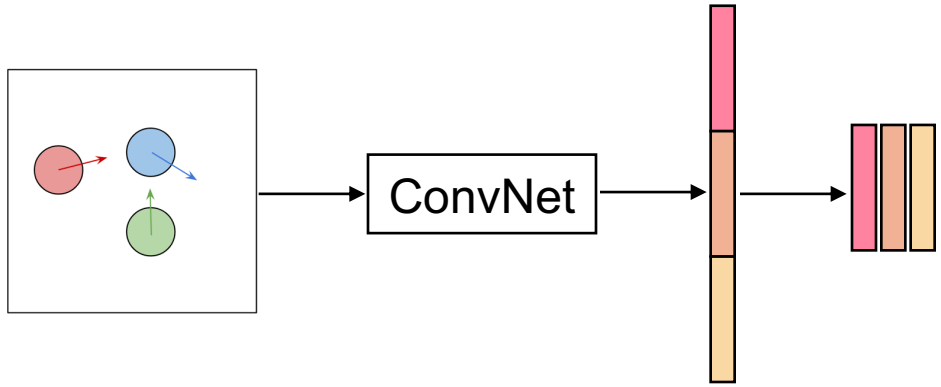
- Prediction from pixels:
  - Not effectively encode object and context features
  - Not aim to do long term prediction (MPC approach)
- We hypothesis the bottleneck is the features of objects.




[1] Y. Ye, M. Singh, A. Gupta, S. Tulsiani. “Compositional Video Prediction”. ICCV 2019
[2] Jiajun Wu, Erika Lu, Pushmeet Kohli, William T. Freeman, Joshua B. Tenenbaum. Learning to See Physics via Visual De-animation. In NIPS 2017

# Interaction Network

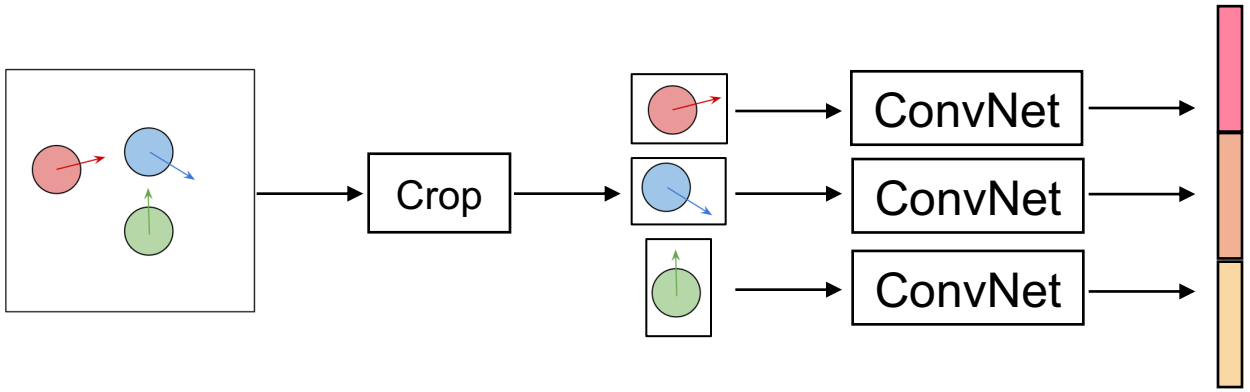
- Visual Interaction Network [1]: Use ConvNet to extract (#obj x 128) feature channels from multiple images.
  - Not very intuitive and cannot generalize to multiple objects
  - Input order is fixed so cannot generalize to multiple appearance




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

# Interaction Network

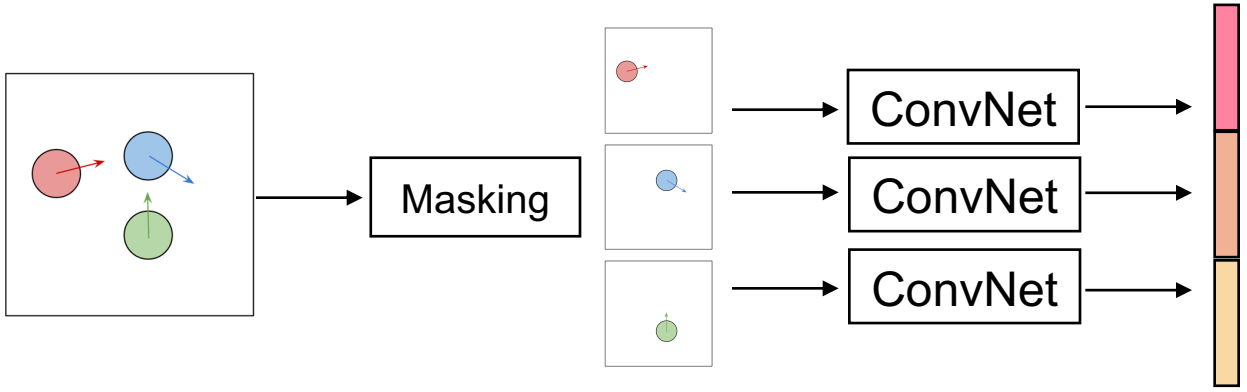
- Visual Interaction Network [1]: Use ConvNet to extract (#obj x 128) features from multiple images.
- Compositional Video Prediction [2,3]: Crop image by RoI and then pass through a ConvNet to get features.




[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

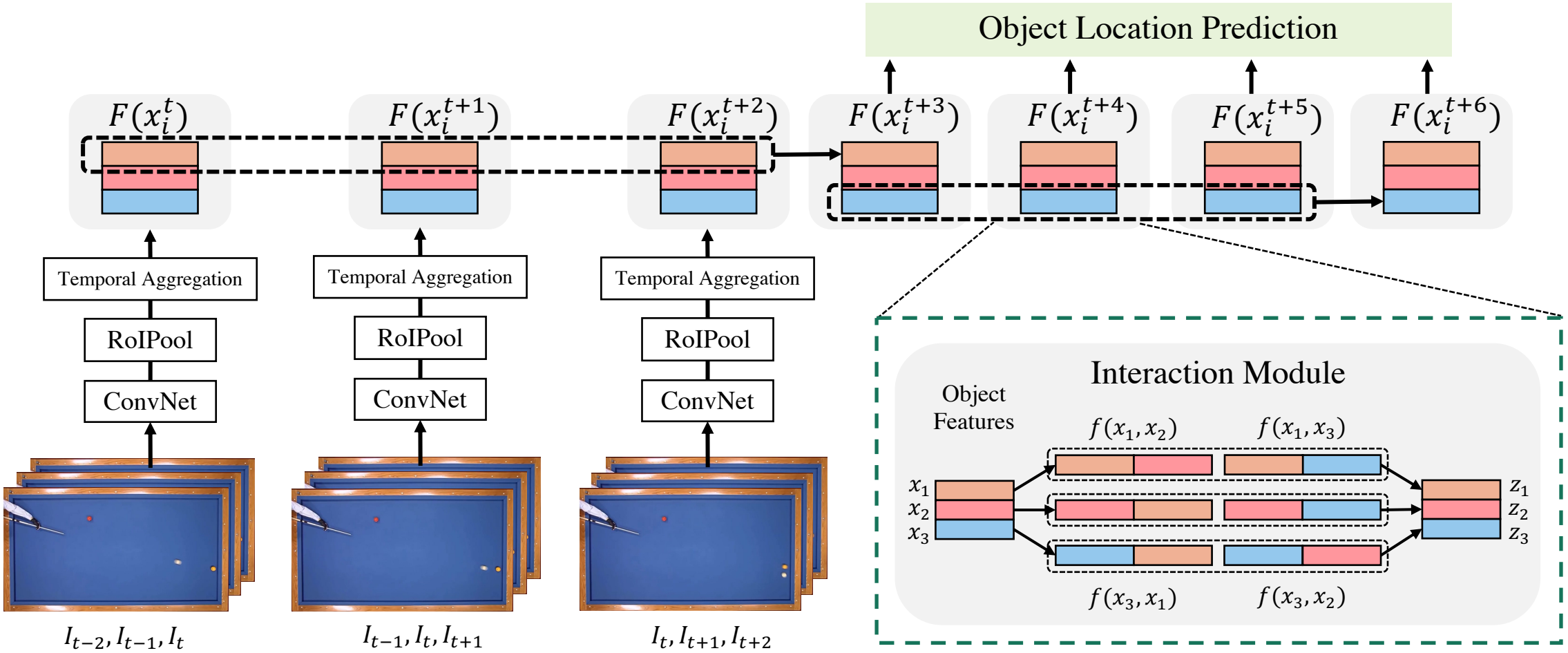
# Interaction Network

- 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]



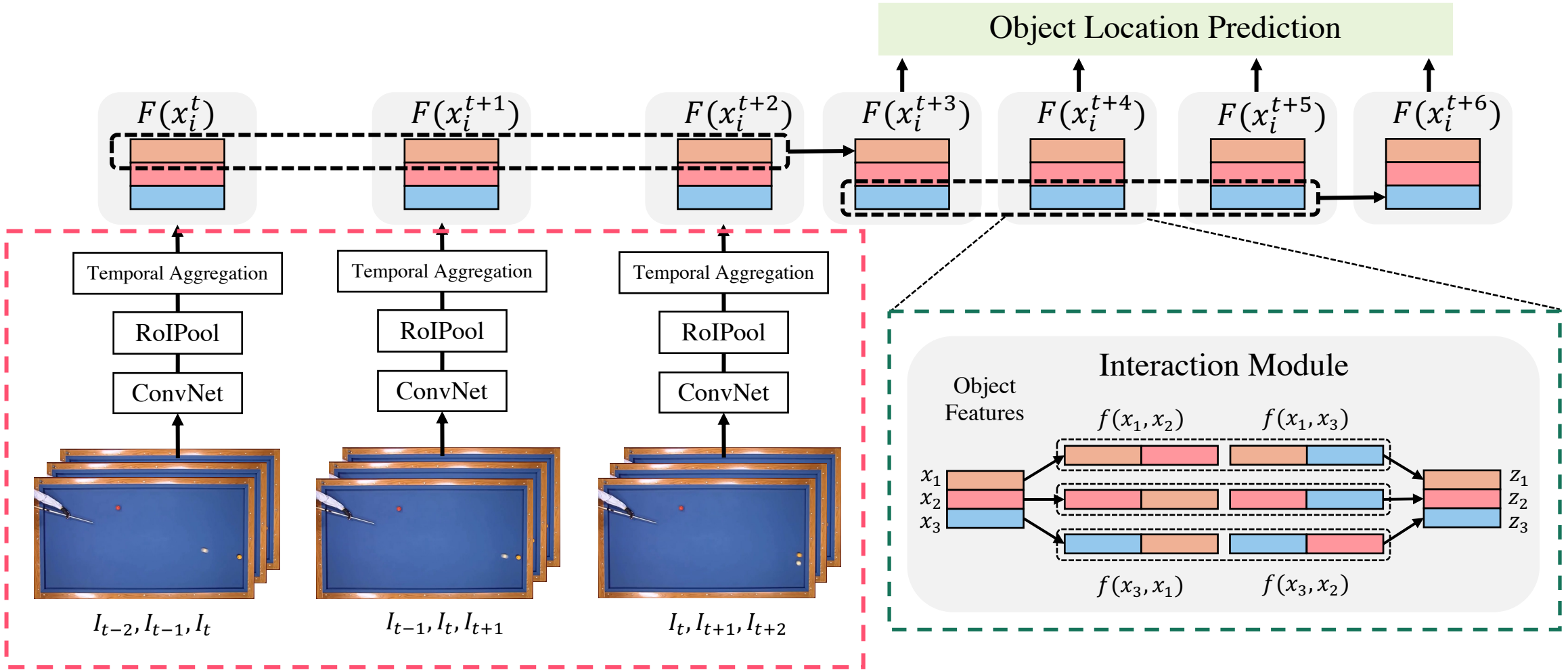

[1] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, A. Tacchetti. “Visual Interaction Networks”. NIPS 2017
[2] Y. Ye, M. Singh, A. Gupta, S. Tulsiani. “Compositional Video Prediction”. ICCV 2019
[3] Y. Ye, D. Gandhi, A.Gupta, S. Tulsiani. “Object-centric Forward Modeling for Model Predictive 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




Qi et al., 2021

# Visual Encoder



# Visual Encoder

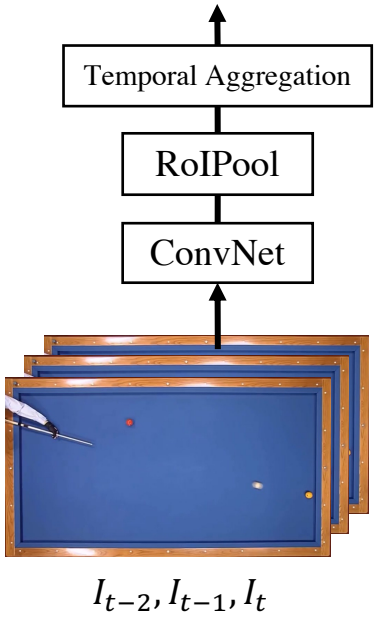
- Object Centric Representation for Prediction

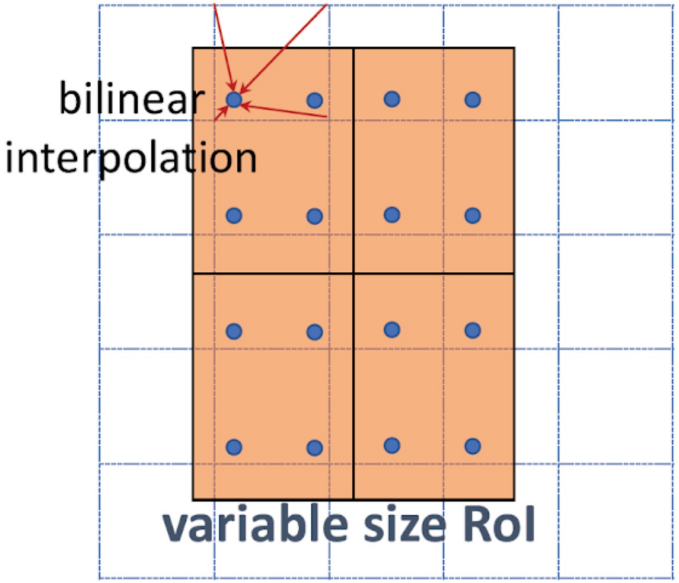
- We extract the state feature representations of $n$ objects in time $t$, and predict their representations in time $t + 1$.

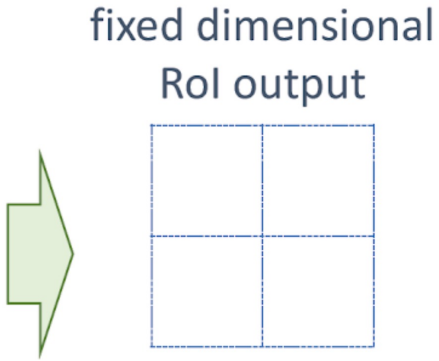
$$\{x_1^t, x_2^t, \ldots, x_n^t\} \longrightarrow \{x_1^{t+1}, x_2^{t+1}, \ldots, x_n^{t+1}\}$$

# Visual Encoder

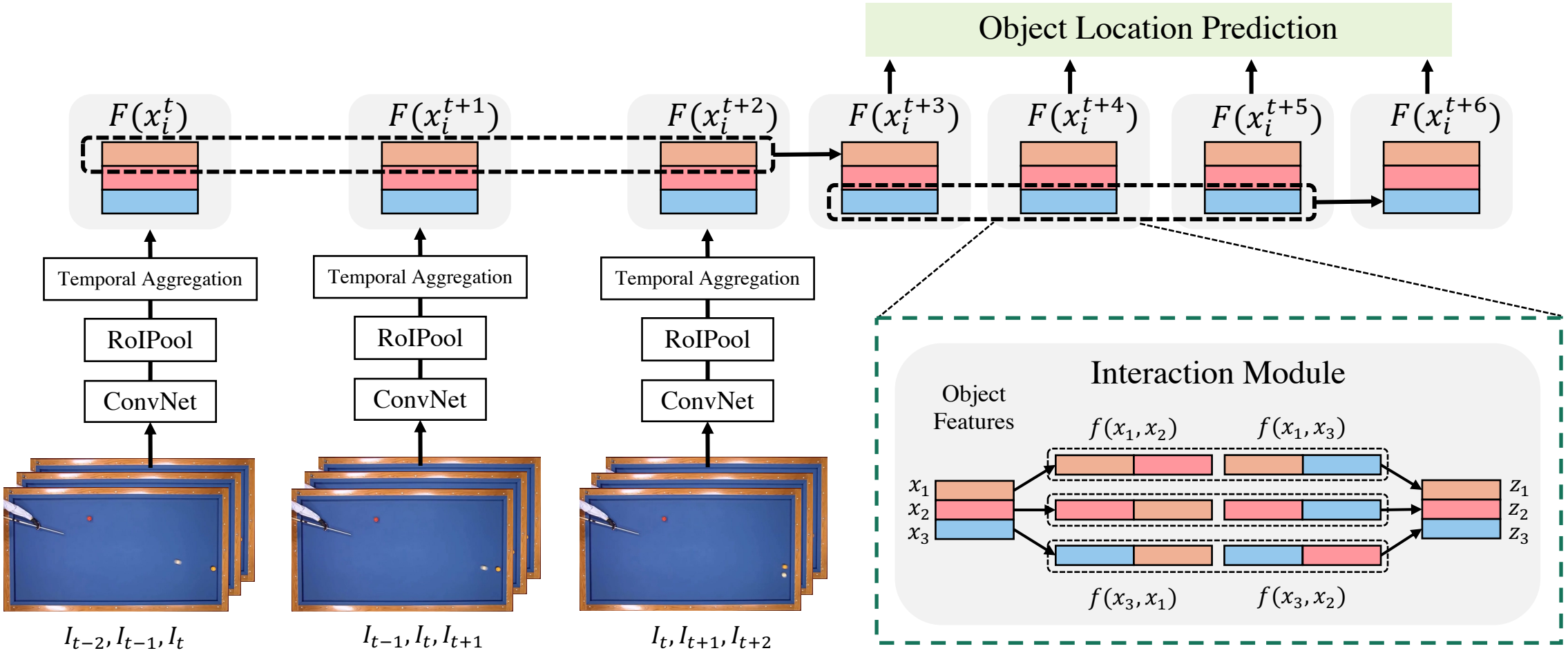
- Use hourglass network to extract image features
- Use aligned RoI Pooling to extract region features



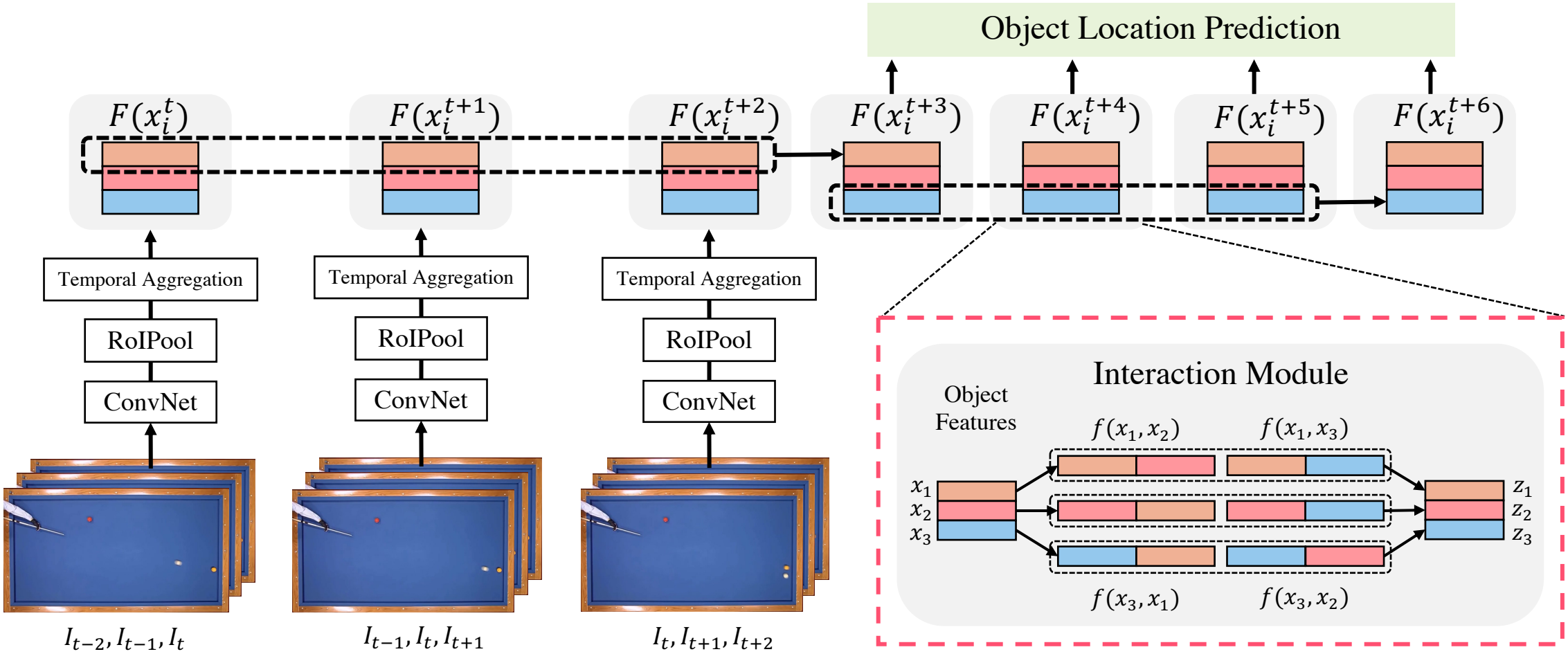



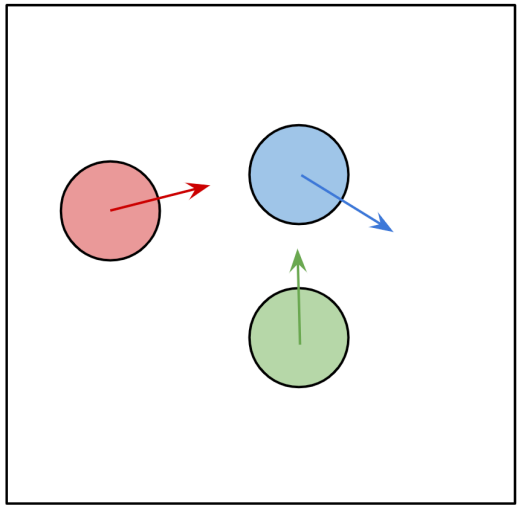
# Interaction Module in feature space



# 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)

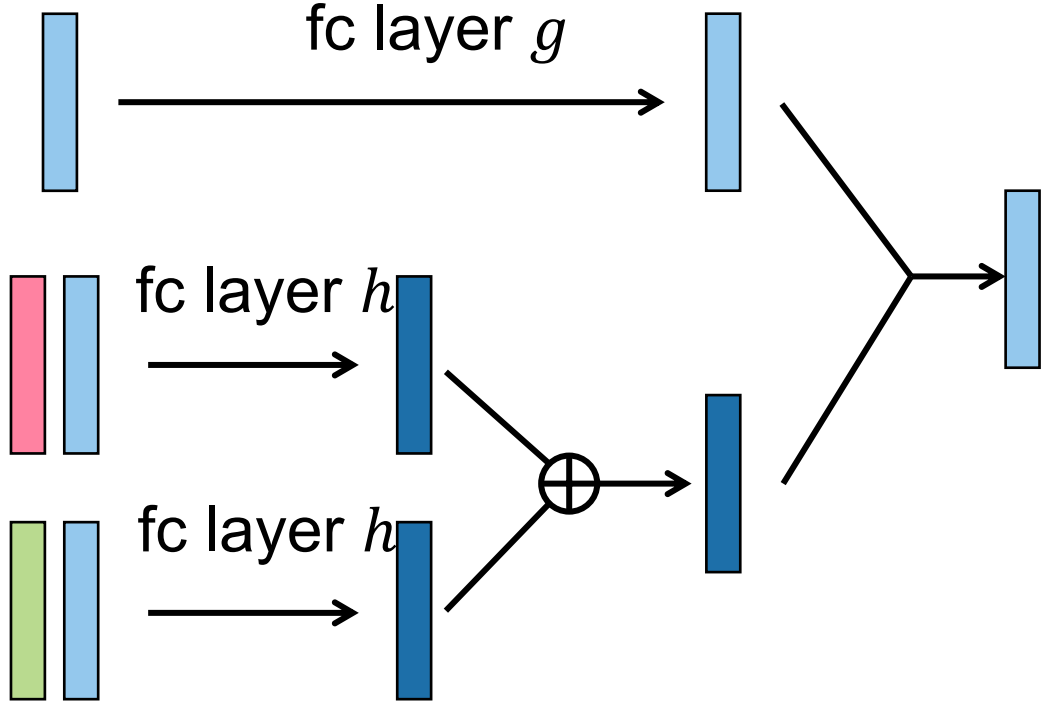
$$g(x_i^t)$$

- relation-dynamics: (Newton's second law)

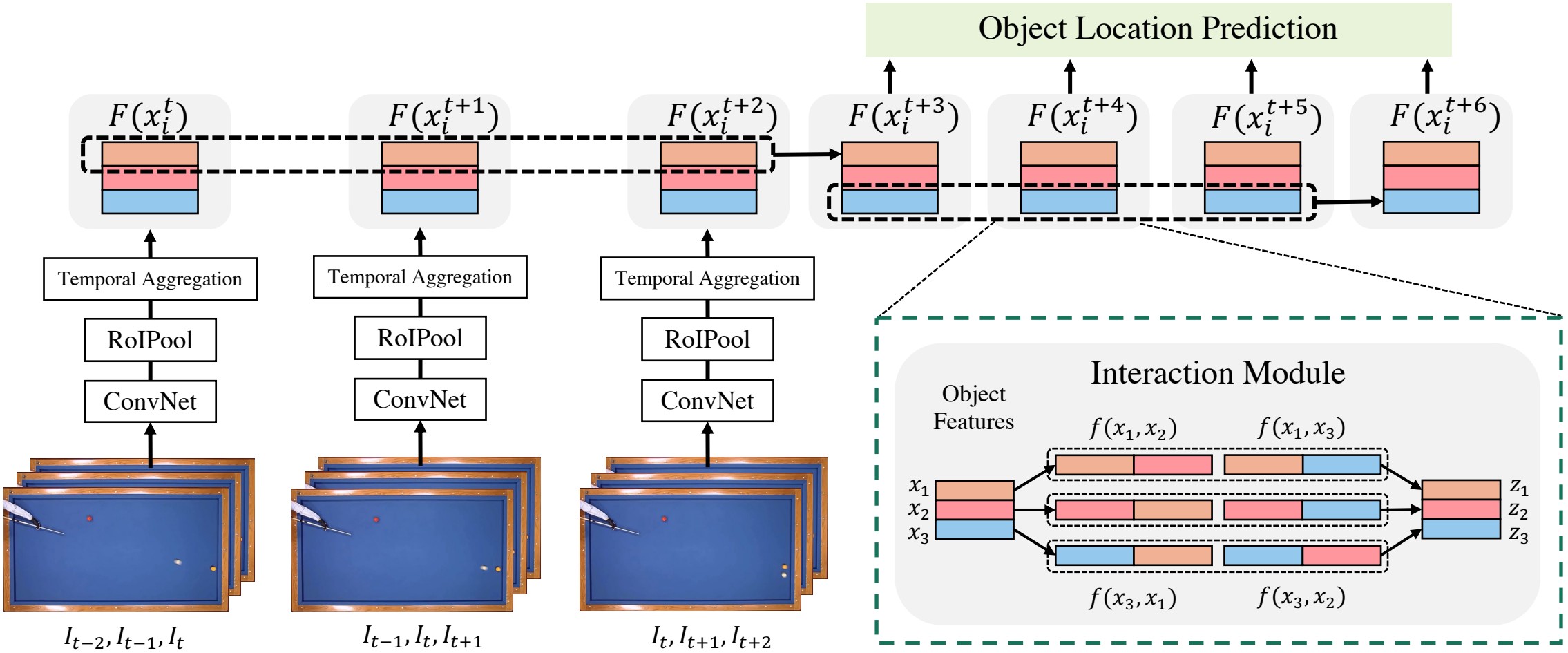
$$\sum_{j \neq i} h(x_i^t, x_j^t)$$

- Aggregate the above:

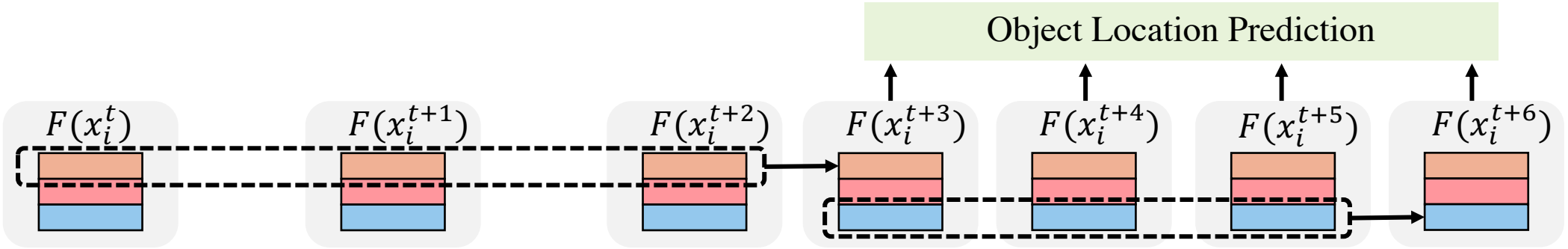
$$F(x_i^t) = f\left( g(x_i^t), \sum_{j \neq i} h(x_i^t, x_j^t) \right)$$



# 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
Temporal Aggregation
Temporal Aggregation
RoIPool
RoIPool
RoIPool
ConvNet
ConvNet
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
$f(x_1, x_2)$
$f(x_1, x_3)$
$x_1$
$x_2$
$x_3$
$z_1$
$z_2$
$z_3$
$f(x_3, x_1)$
$f(x_3, x_2)$

# Prediction

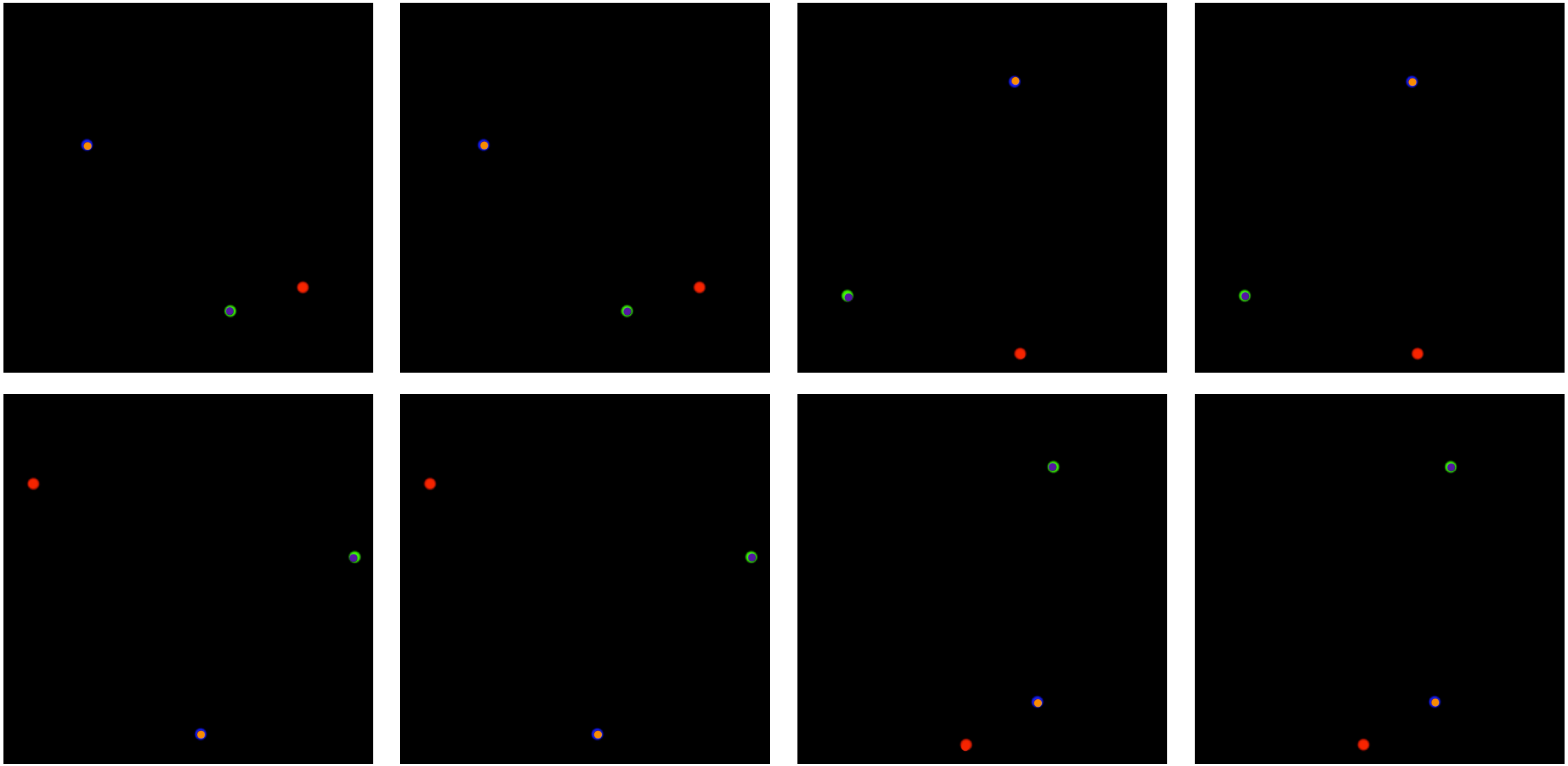



Future feature prediction: $x_i^{t+1} = W_d\left[F\left(x_i^t\right), F\left(x_i^{t-1}\right), \ldots, F\left(x_i^{t-k}\right)\right]$

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

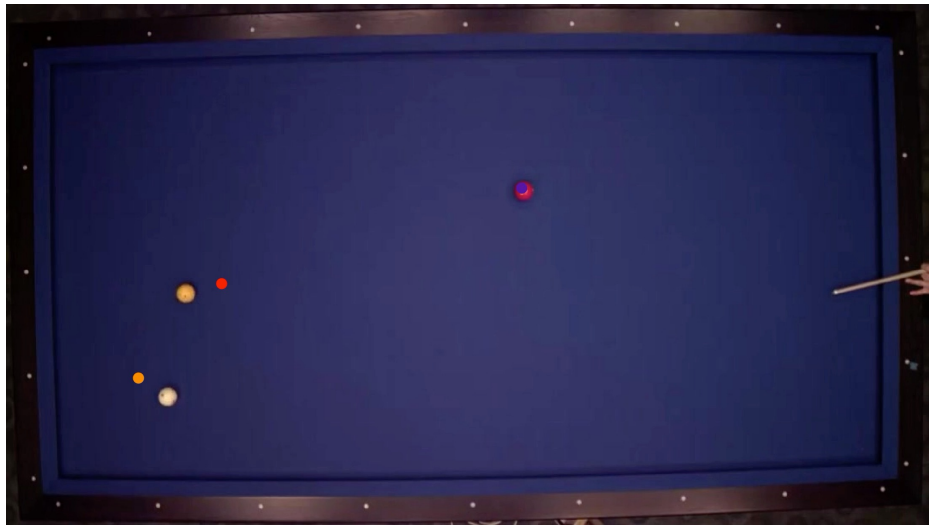
Training loss function: $L_p = \sum_{t=1}^{T}\sum_{i=1}^{n}\left\|\hat{p}_i^{t+1} - p_i^{t+1}\right\|_2^2$

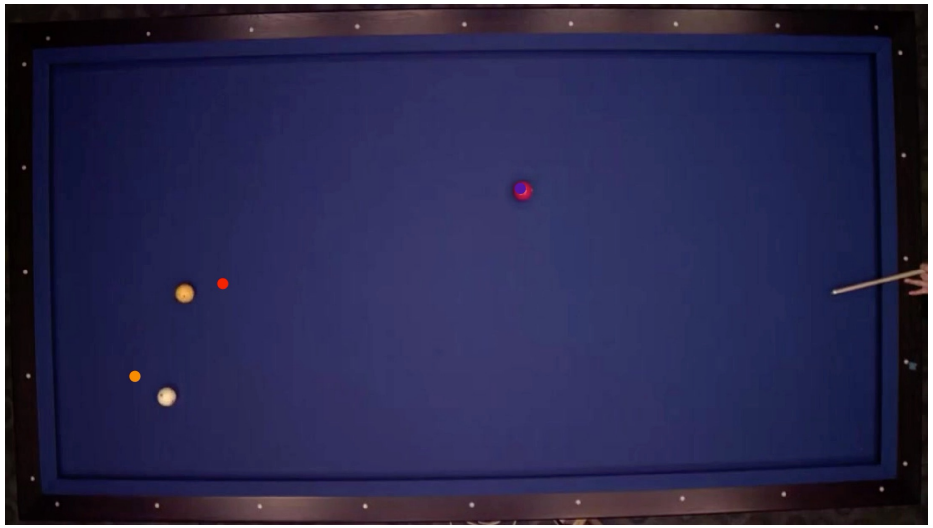
# Simulation Billiards

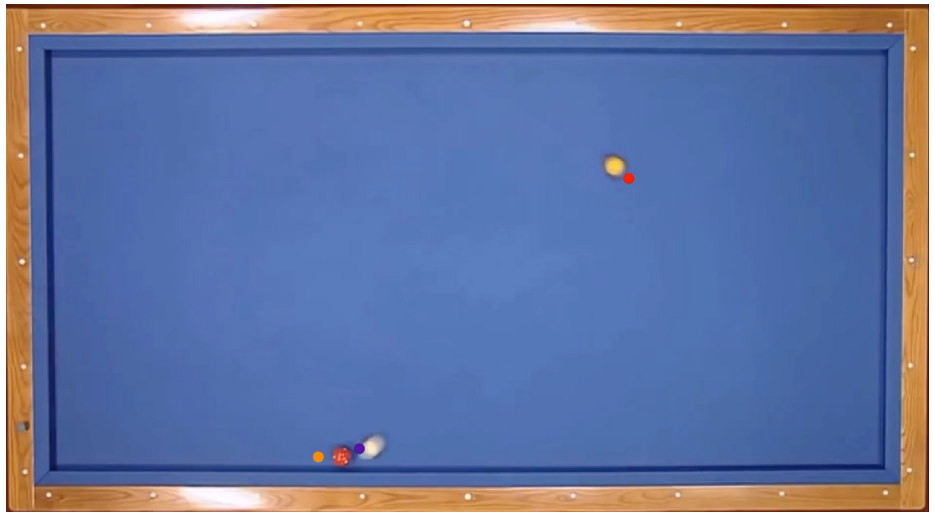


prediction ground-truth prediction ground-truth

# Real Billiards



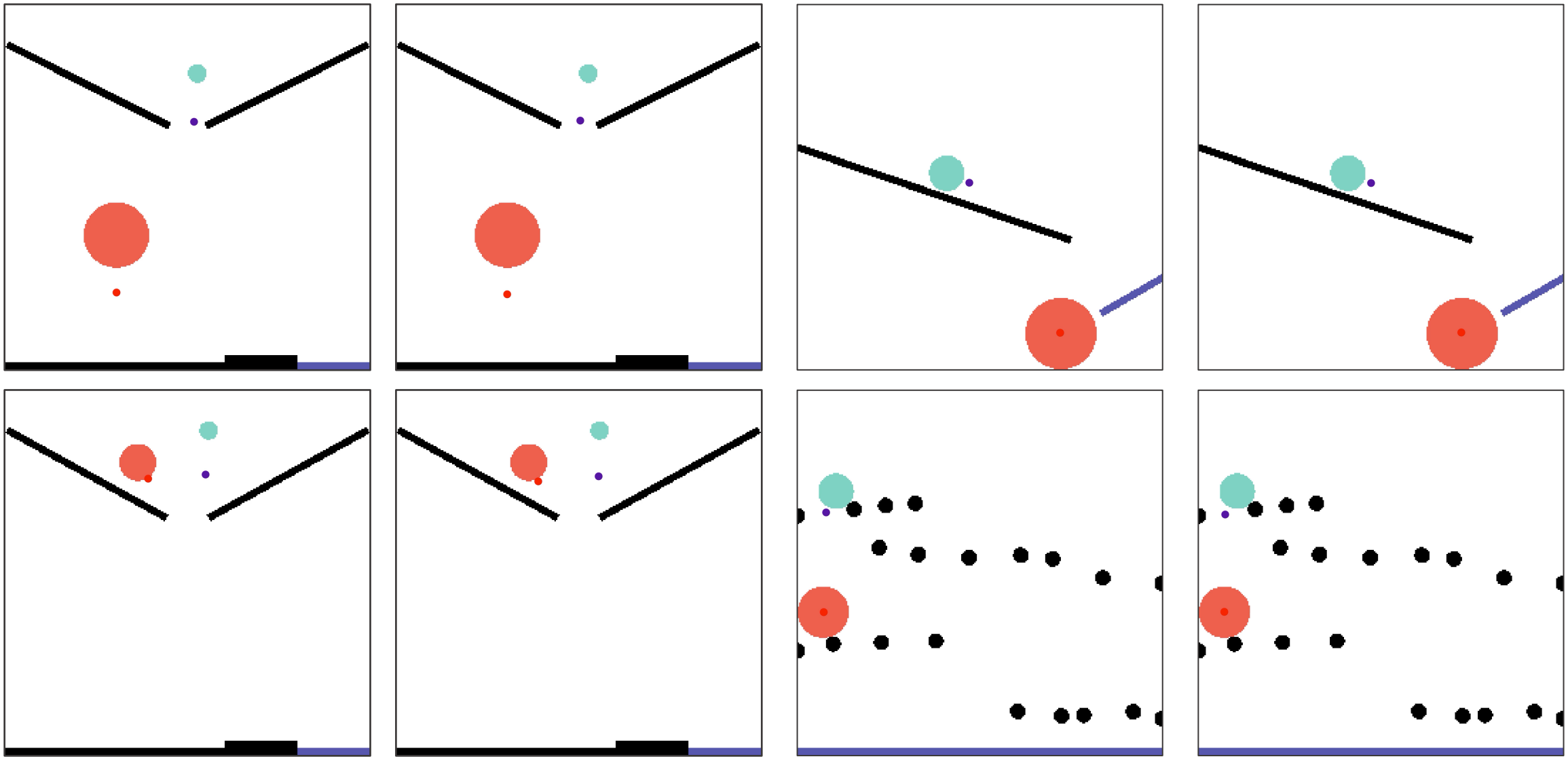




prediction

ground-truth

# PHYRE



prediction ground-truth prediction ground-truth

# Summary

- 2-Stream Networks for Action Recognition
- Temporal Convolution and 3D Convolution
- Video Prediction and Interaction Network

# Next Class

Attention and transformer