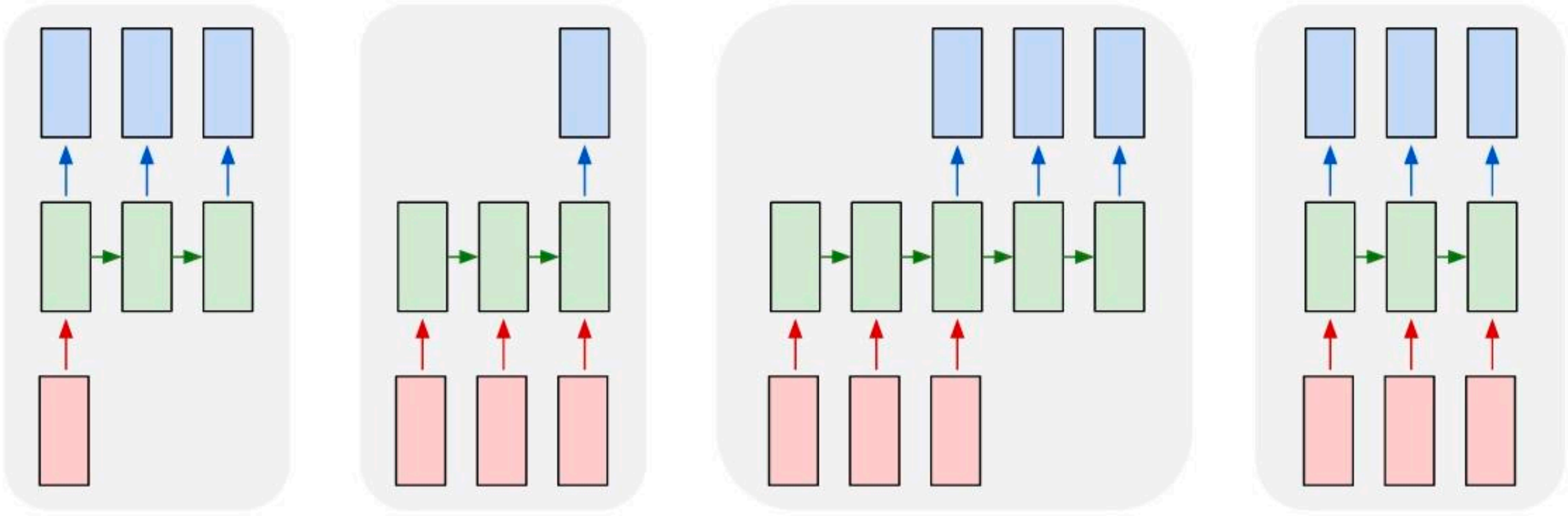
Video Recognition

Xiaolong Wang

Previous classes



This Class

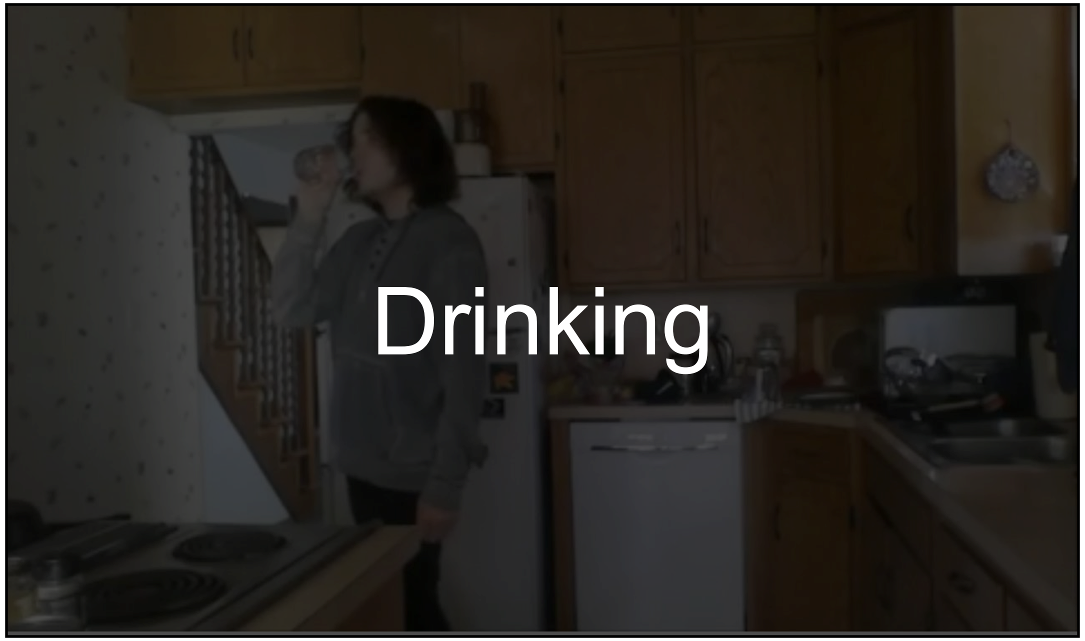
- 2-Stream Networks for Action Recognition

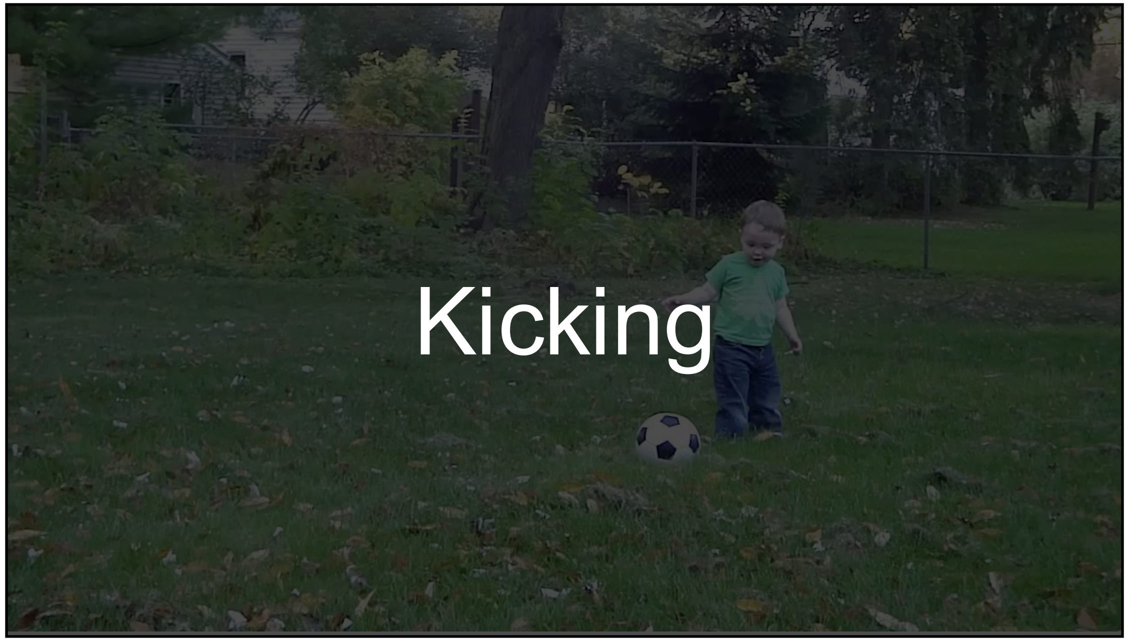
- Temporal Convolution and 3D Convolution

- Temporal Detection and Segmentation

2-Stream Networks for Action Recognition

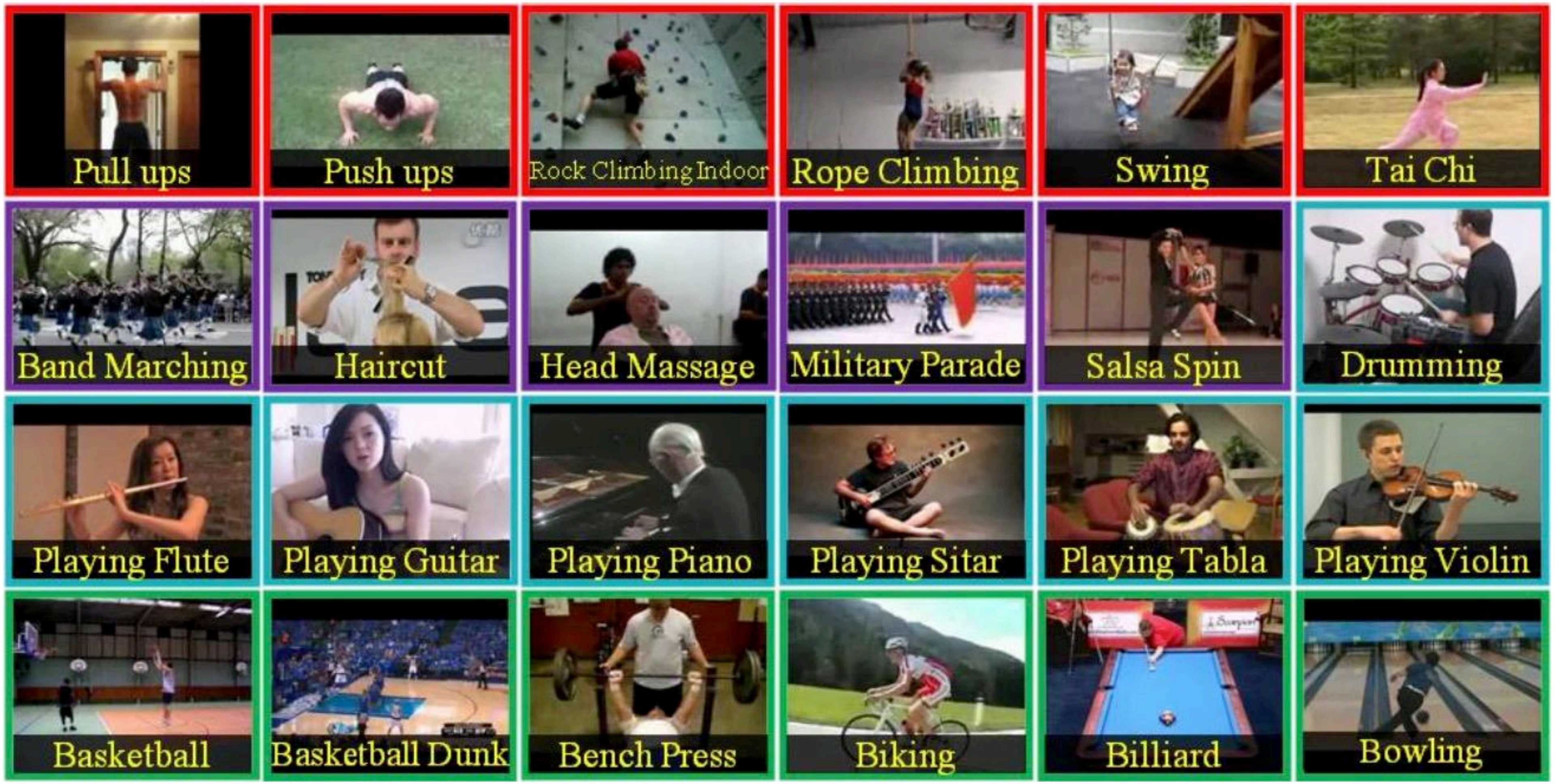
Task: Action Recognition



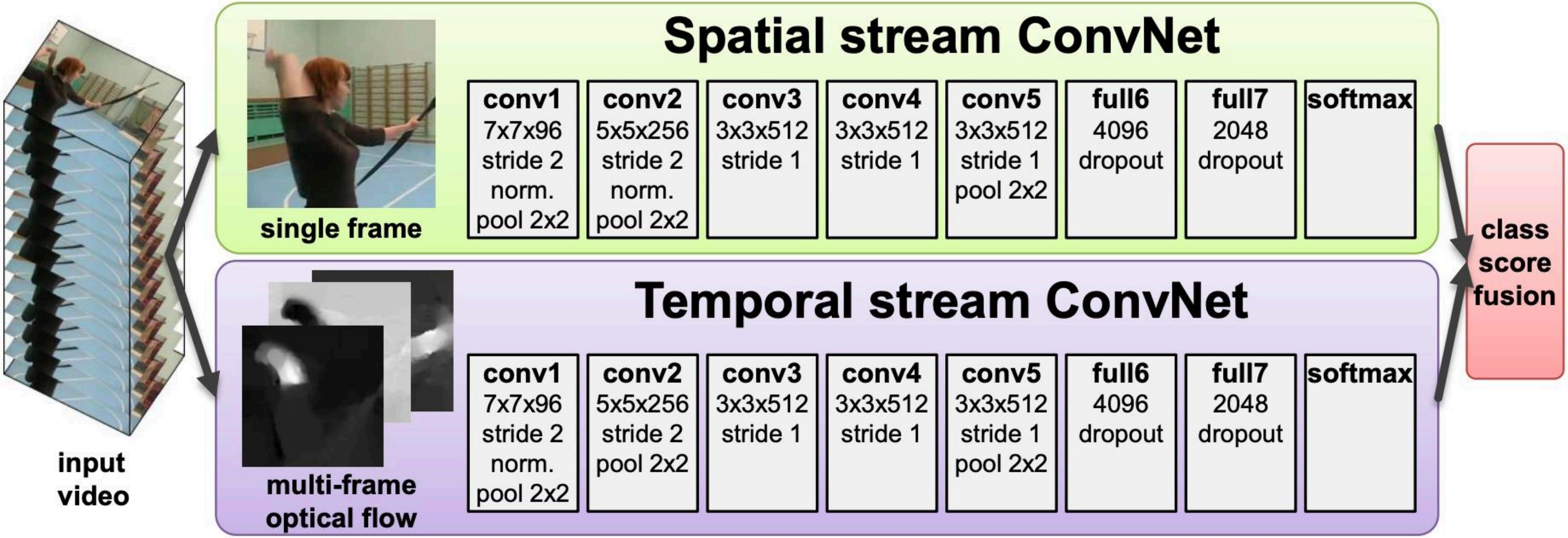


Task: Action Recognition

- UCF-101 dataset

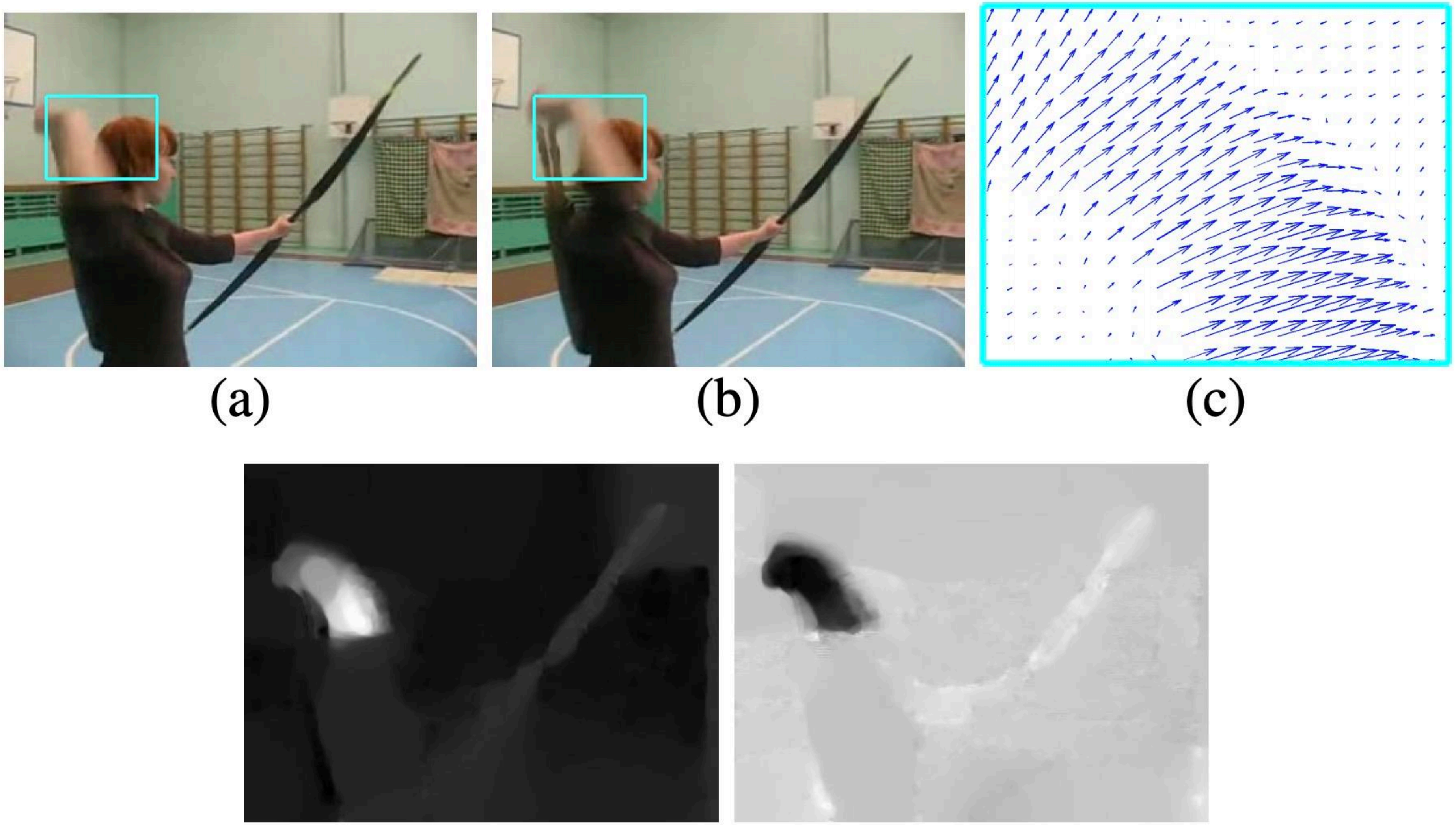


2-Stream CNNs



Simonyan et al., 2014

2-Stream CNNs



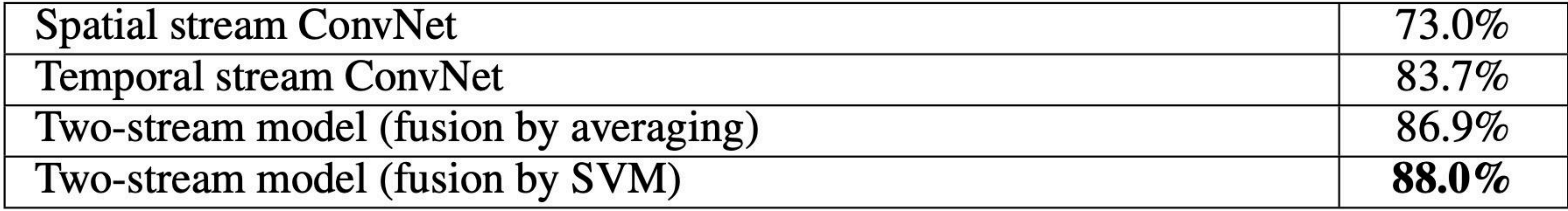
2-Stream CNNs

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

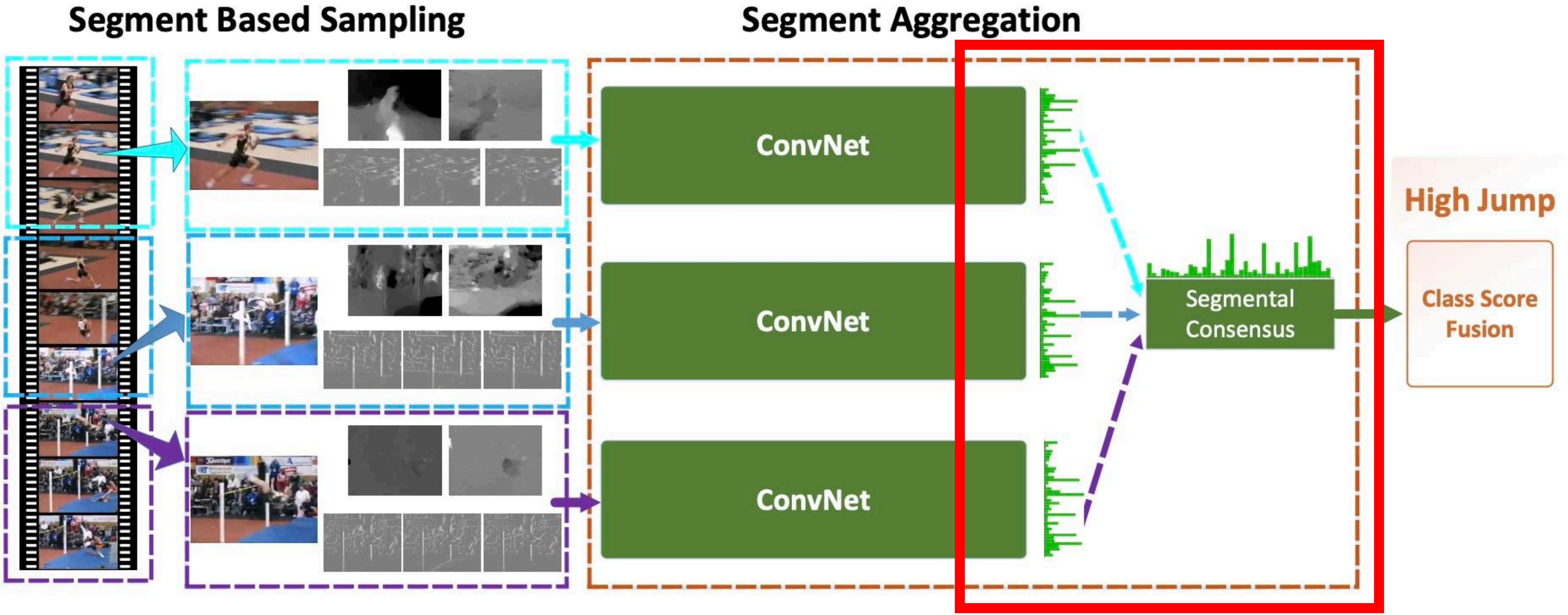


Spatial stream ConvNet	73.0%
Temporal stream ConvNet	83.7%
Two-stream model (fusion by averaging)	86.9%
Two-stream model (fusion by SVM)	**88.0%**

Temporal Segment Networks (TSN)

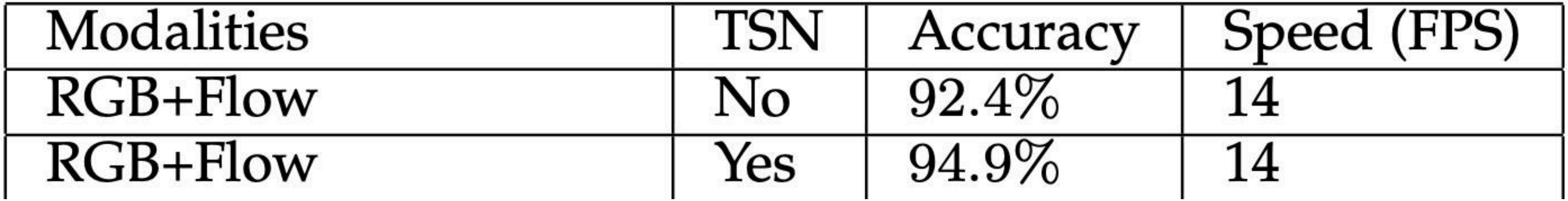
- In the previous work, we train each frame individually
- Can we train multiple frames at the same time?

Temporal Segment Networks (TSN)



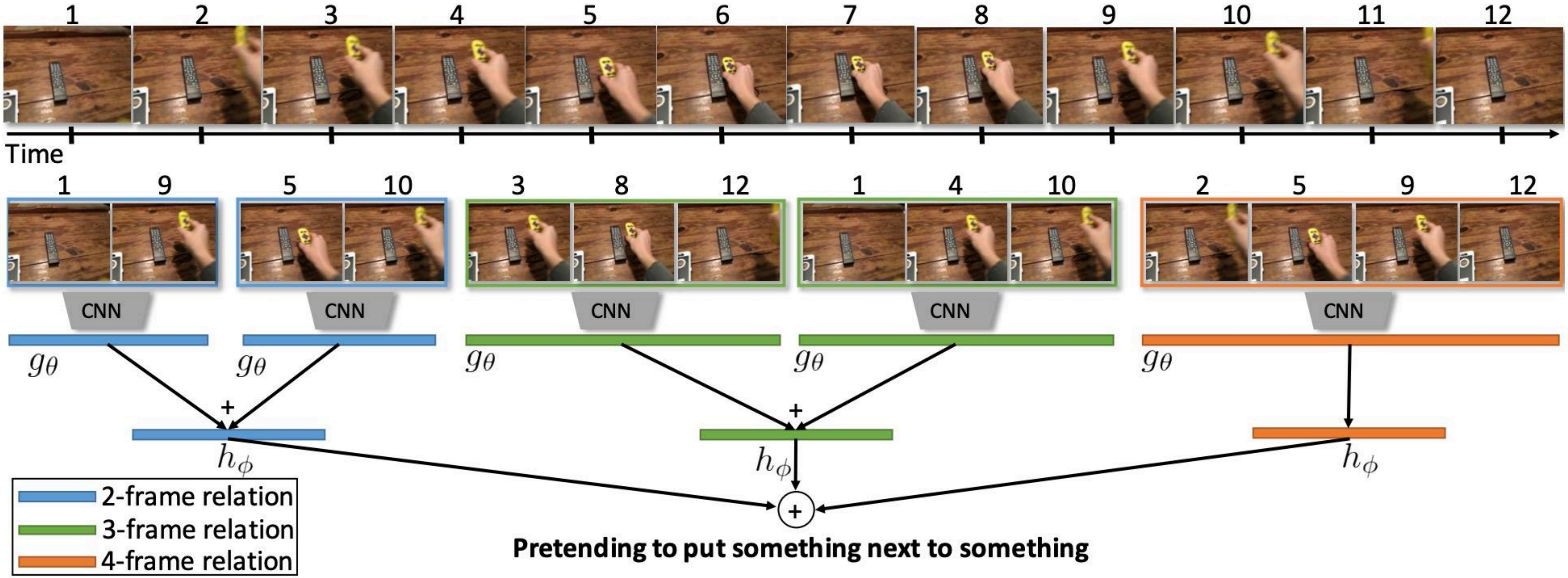
Wang et al., 2016

Temporal Segment Networks (TSN)



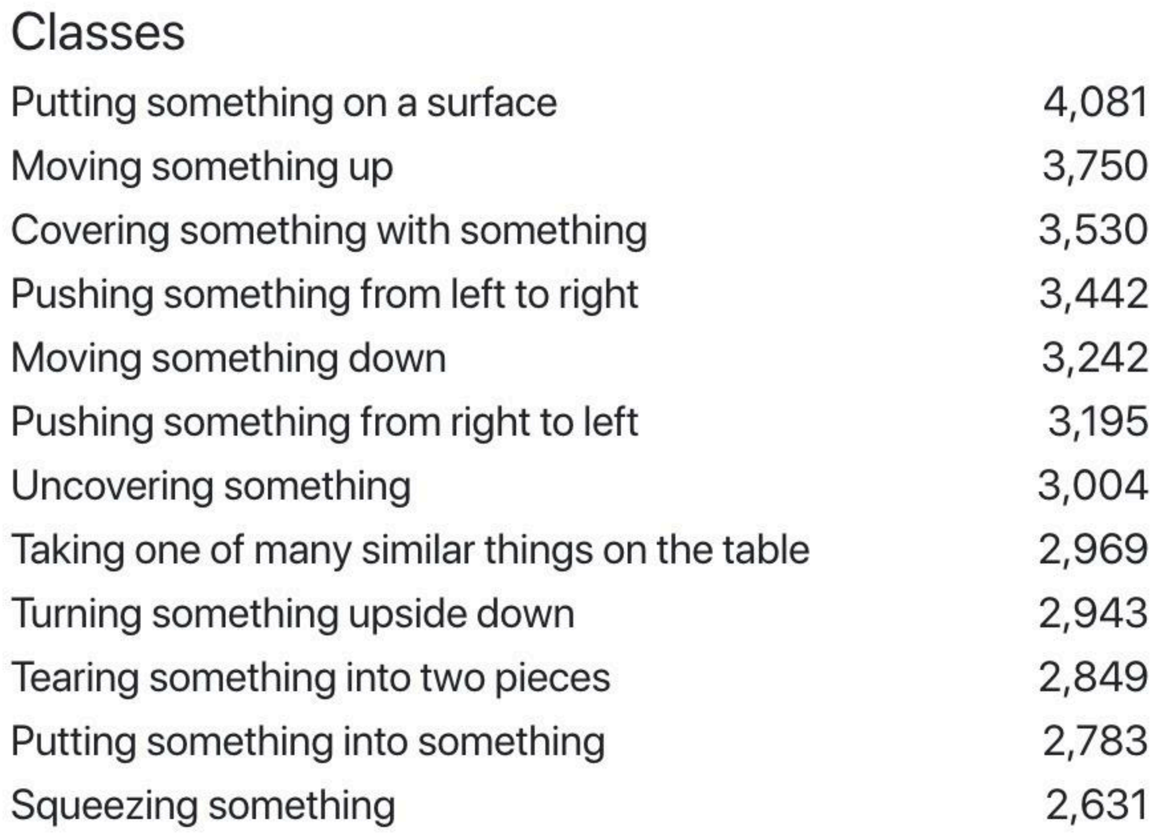
Modalities	TSN	Accuracy	Speed (FPS)
RGB+Flow	No	92.4%	14
RGB+Flow	Yes	94.9%	14

Temporal Relation Network (TRN)

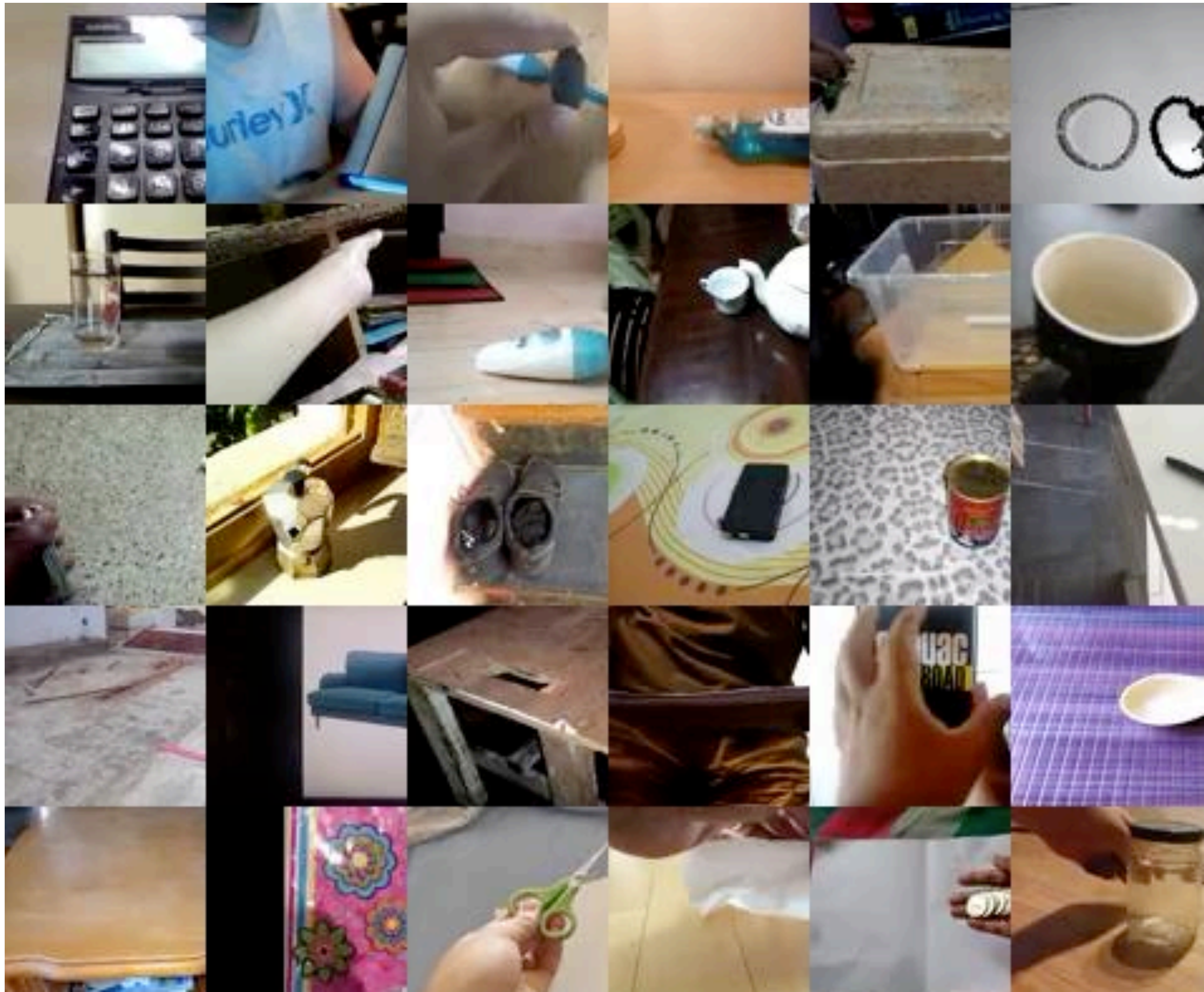


Zhou et al., 2018

Something-Something Dataset

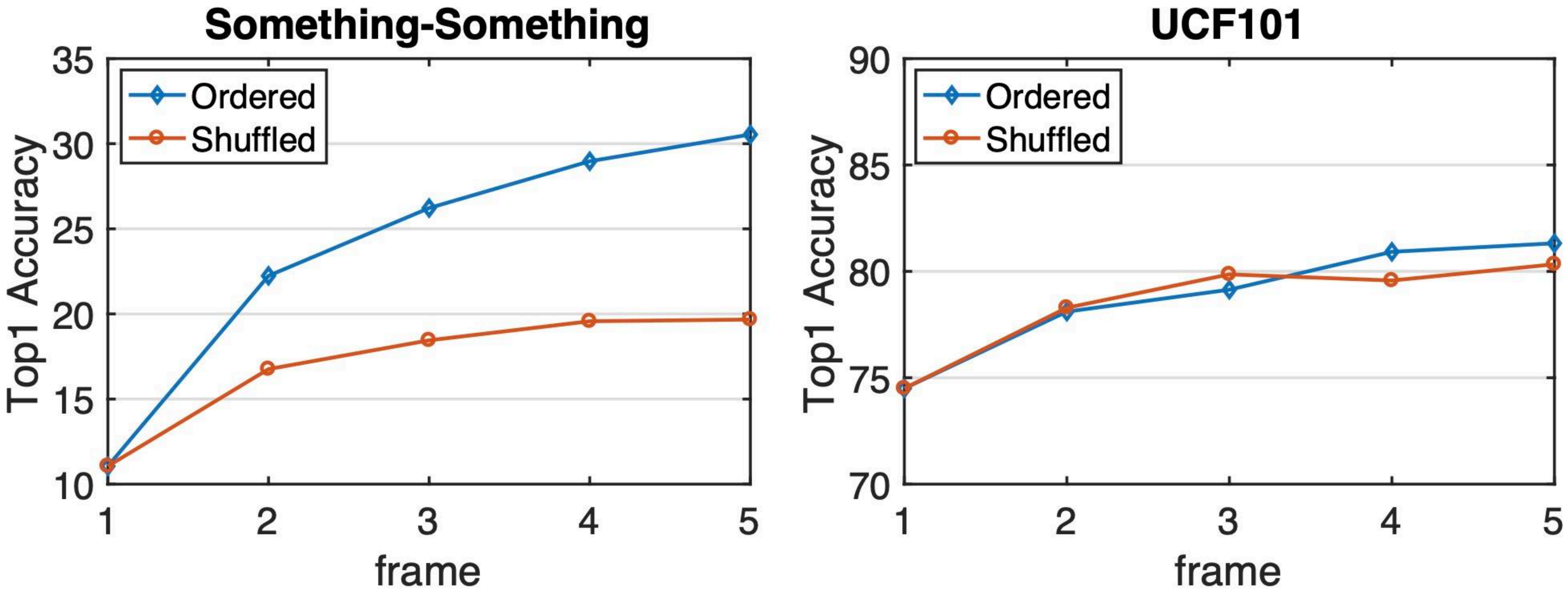


Classes	
Putting something on a surface	4,081
Moving something up	3,750
Covering something with something	3,530
Pushing something from left to right	3,442
Moving something down	3,242
Pushing something from right to left	3,195
Uncovering something	3,004
Taking one of many similar things on the table	2,969
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

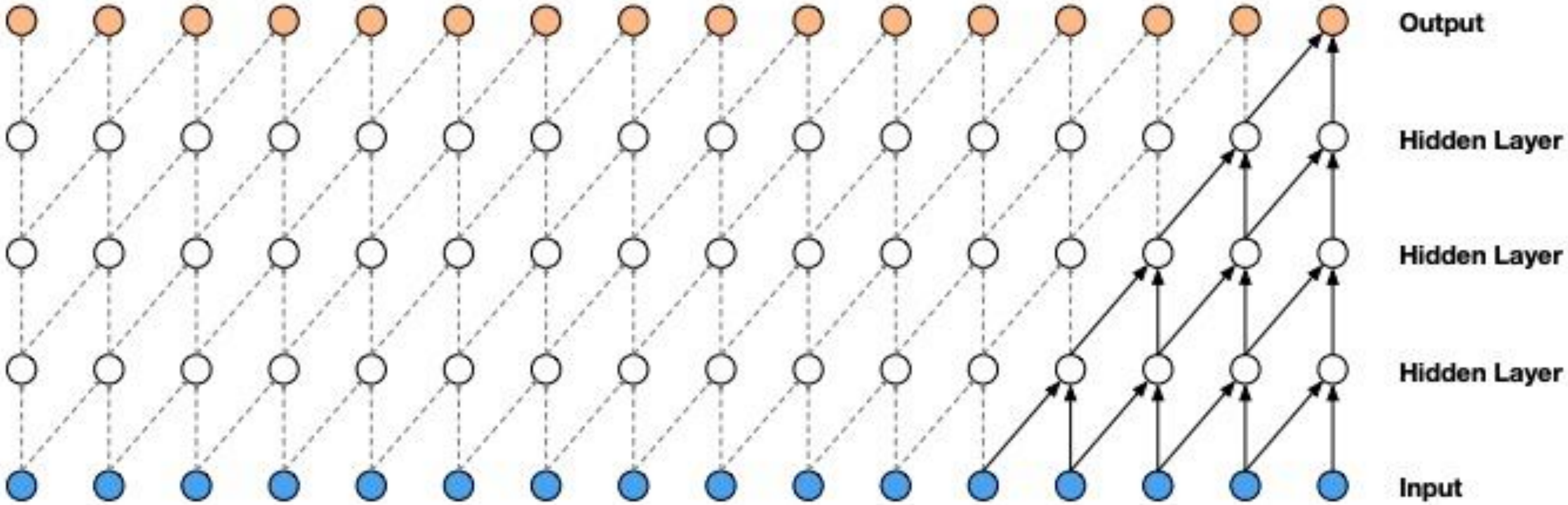
- TSN, use average pooling to aggregate video frames during training → temporal order does not matter

- TRN, use concatenation and FC to aggregate video frames during training → temporal order matters

Temporal Convolution and 3D Convolution

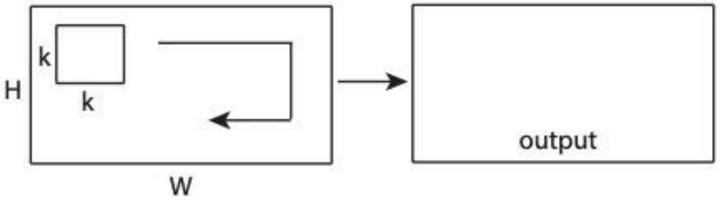
Temporal Convolution

Figure 1: A second of generated speech.



Van den Oord et al., 2016

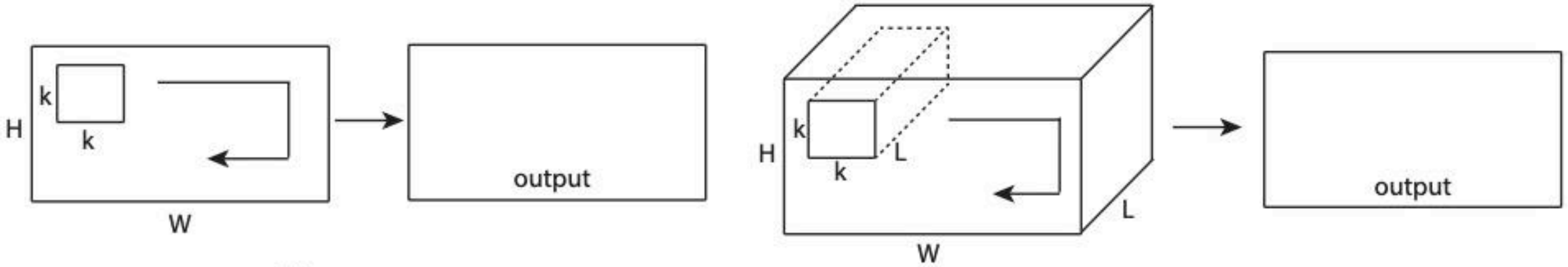
3D Convolution



(a) 2D convolution

Tran et al., 2015

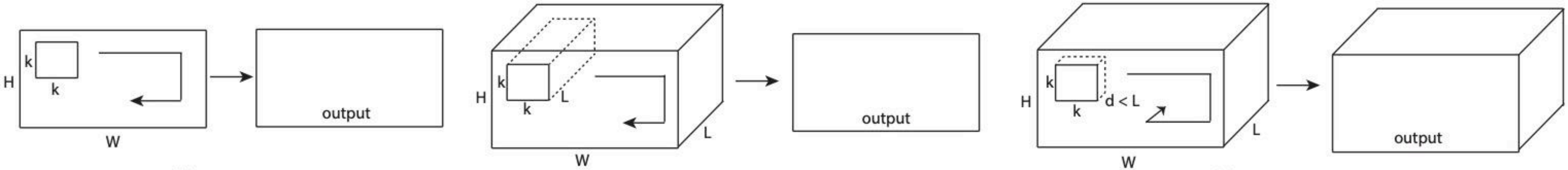
3D Convolution



(a) 2D convolution

(b) 2D convolution on multiple frames

3D Convolution

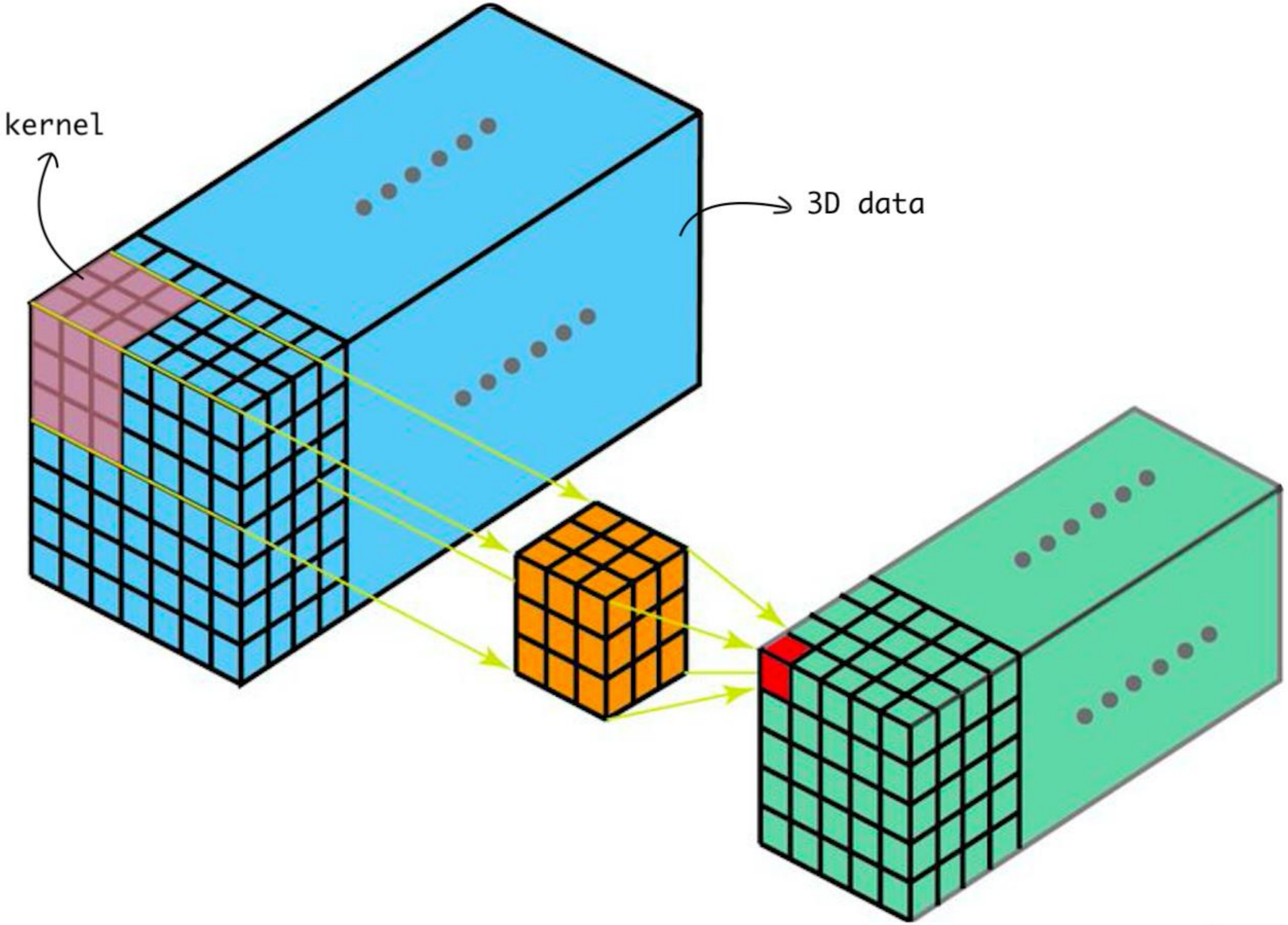


(a) 2D convolution

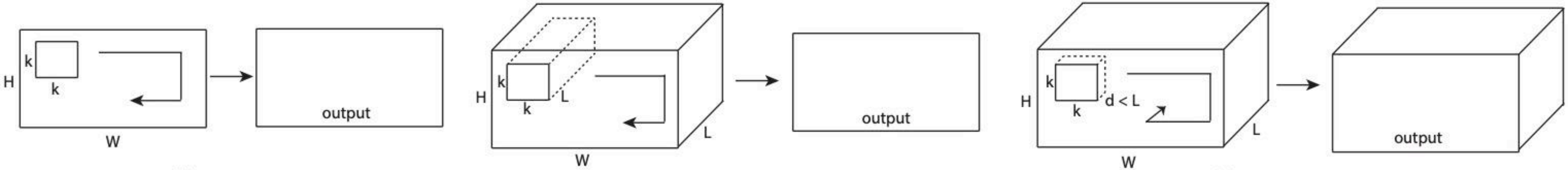
(b) 2D convolution on multiple frames

(c) 3D convolution

3D Convolution



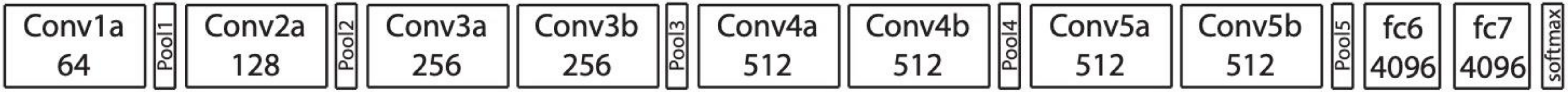
3D Convolution



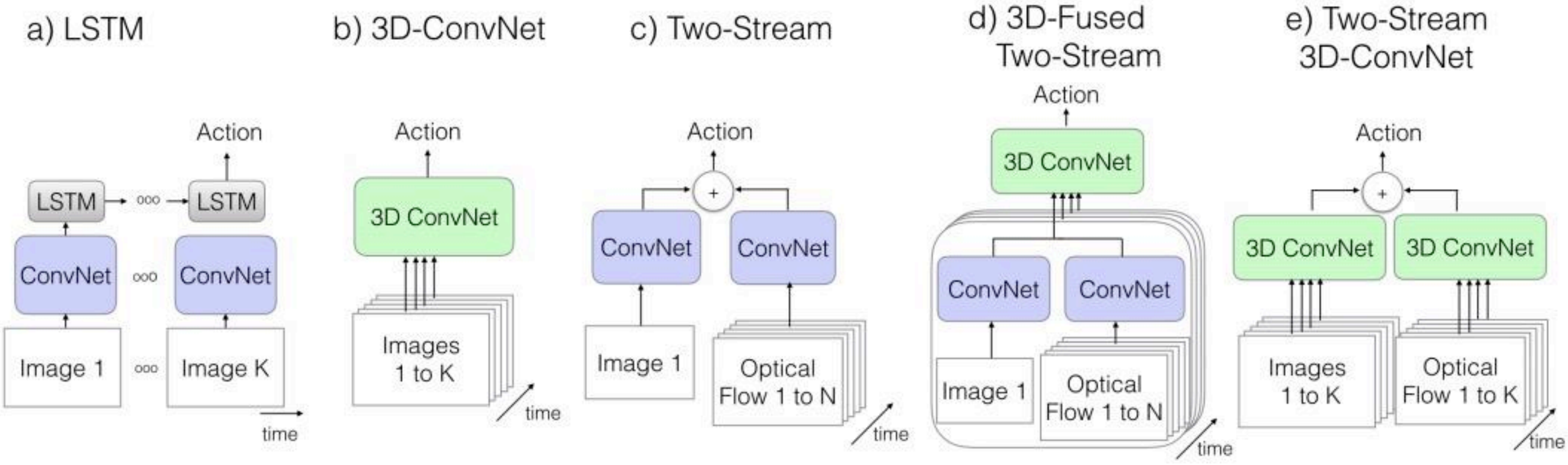
(a) 2D convolution

(b) 2D convolution on multiple frames

(c) 3D convolution


Conv1a 64 | Pool1 | Conv2a 128 | Pool2 | Conv3a 256 | Conv3b 256 | Pool3 | Conv4a 512 | Conv4b 512 | Pool4 | Conv5a 512 | Conv5b 512 | Pool5 | fc6 4096 | fc7 4096 | softmax

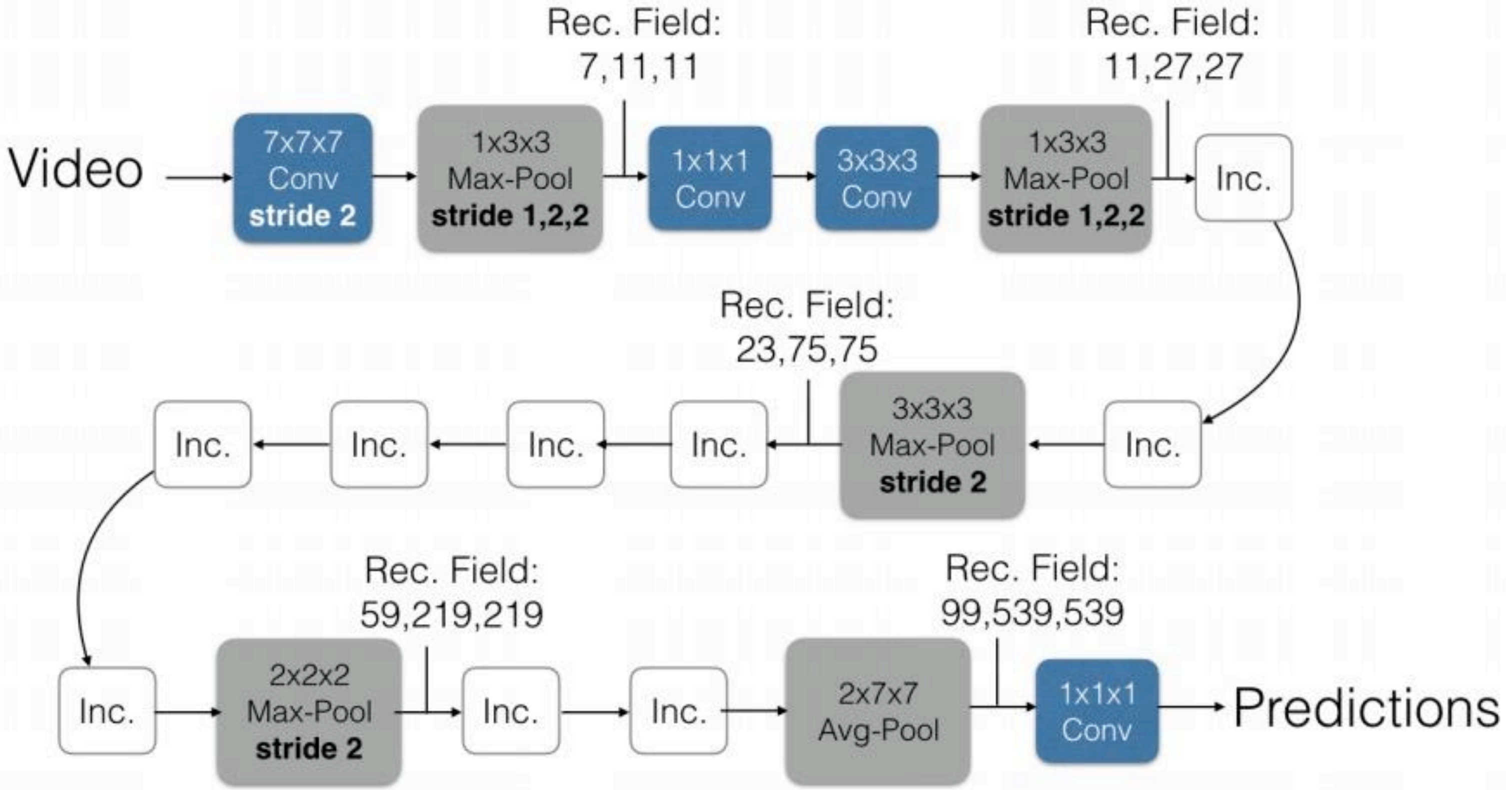
Inflated 3D ConvNets (I3D)



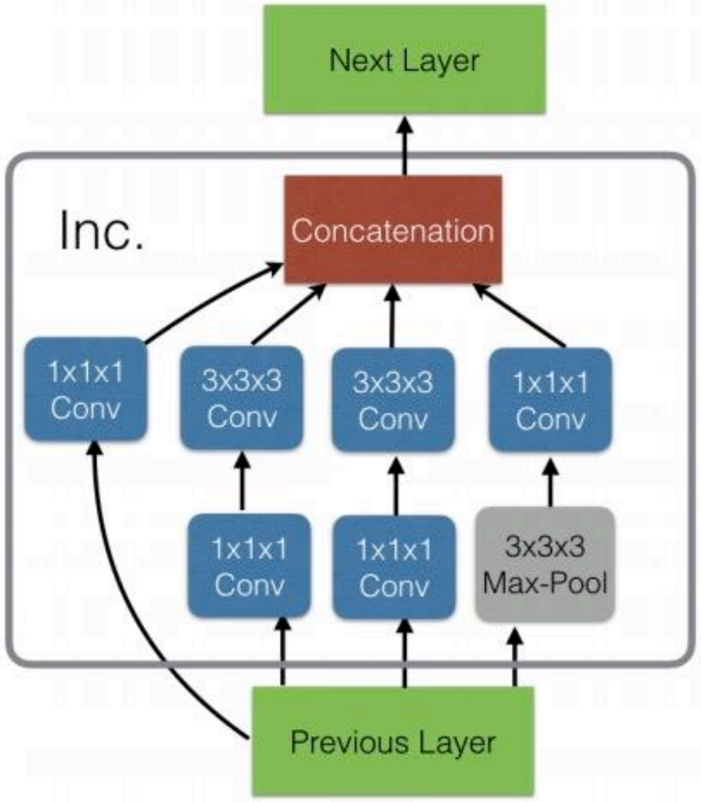
Carreira et al., 2018

Inflated 3D ConvNets (I3D)

Inflated Inception-V1



Inception Module (Inc.)

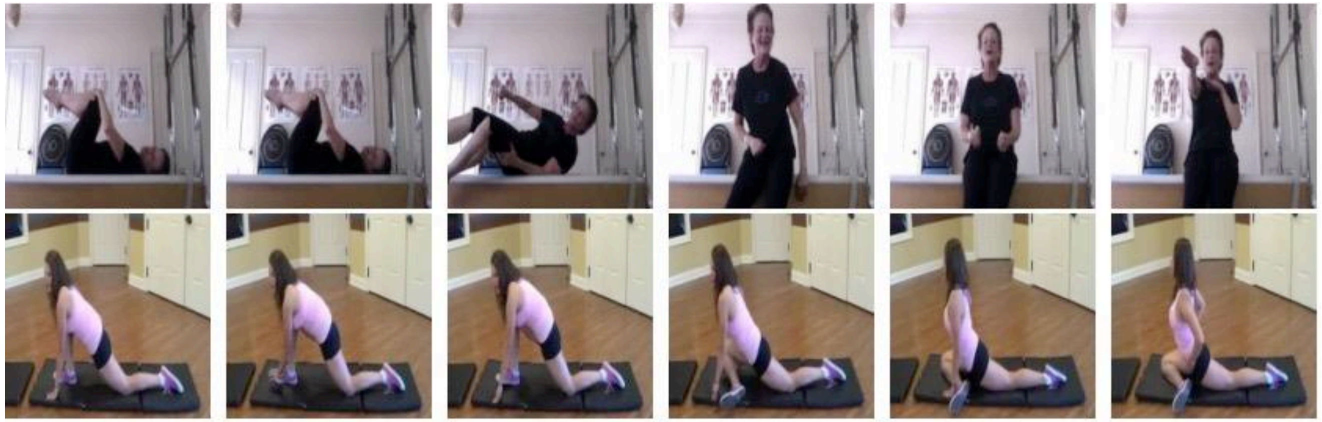


Kinetics Dataset

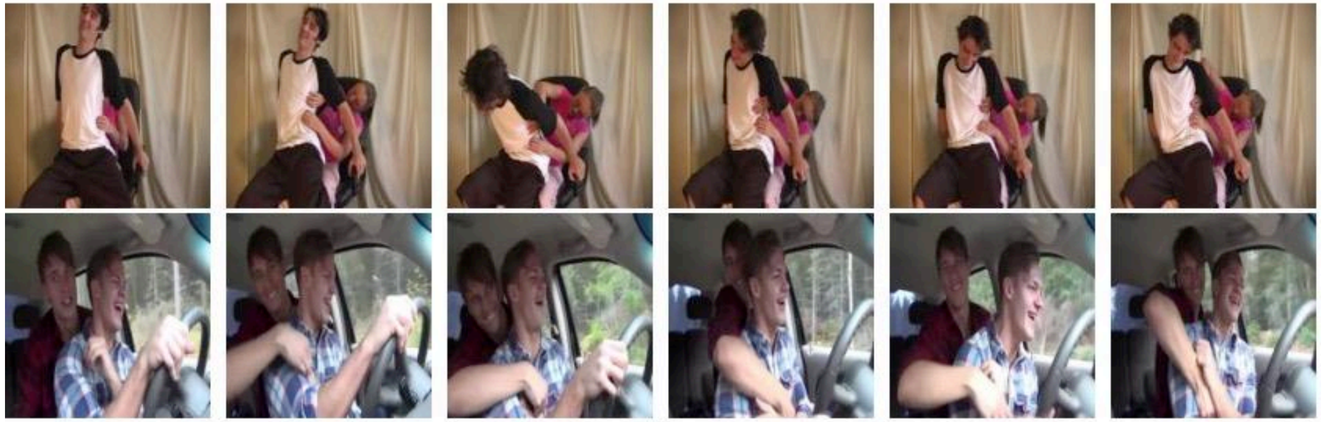
(a) headbanging

(c) shaking hands

Kinetics Dataset

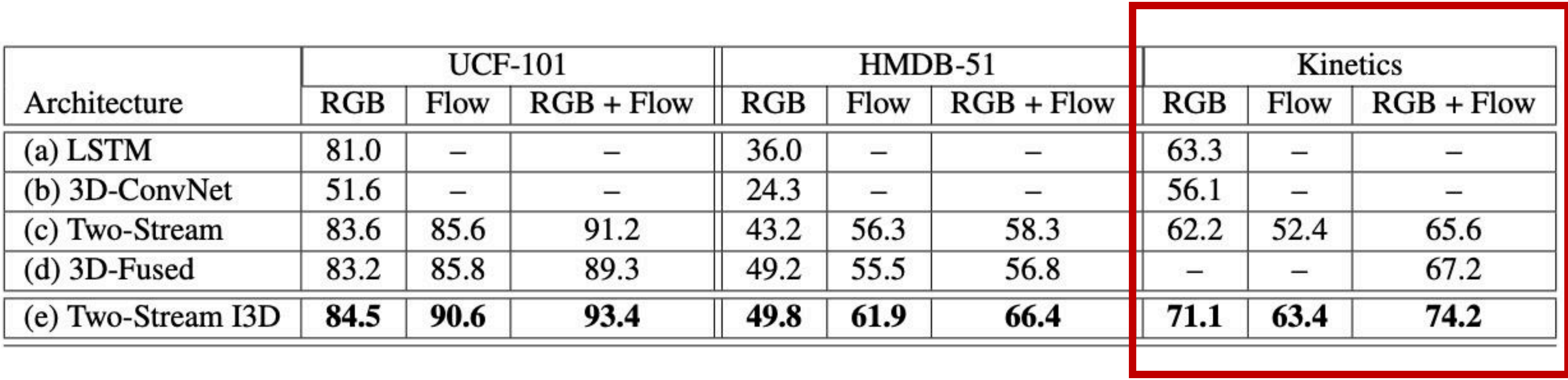


(b) stretching leg



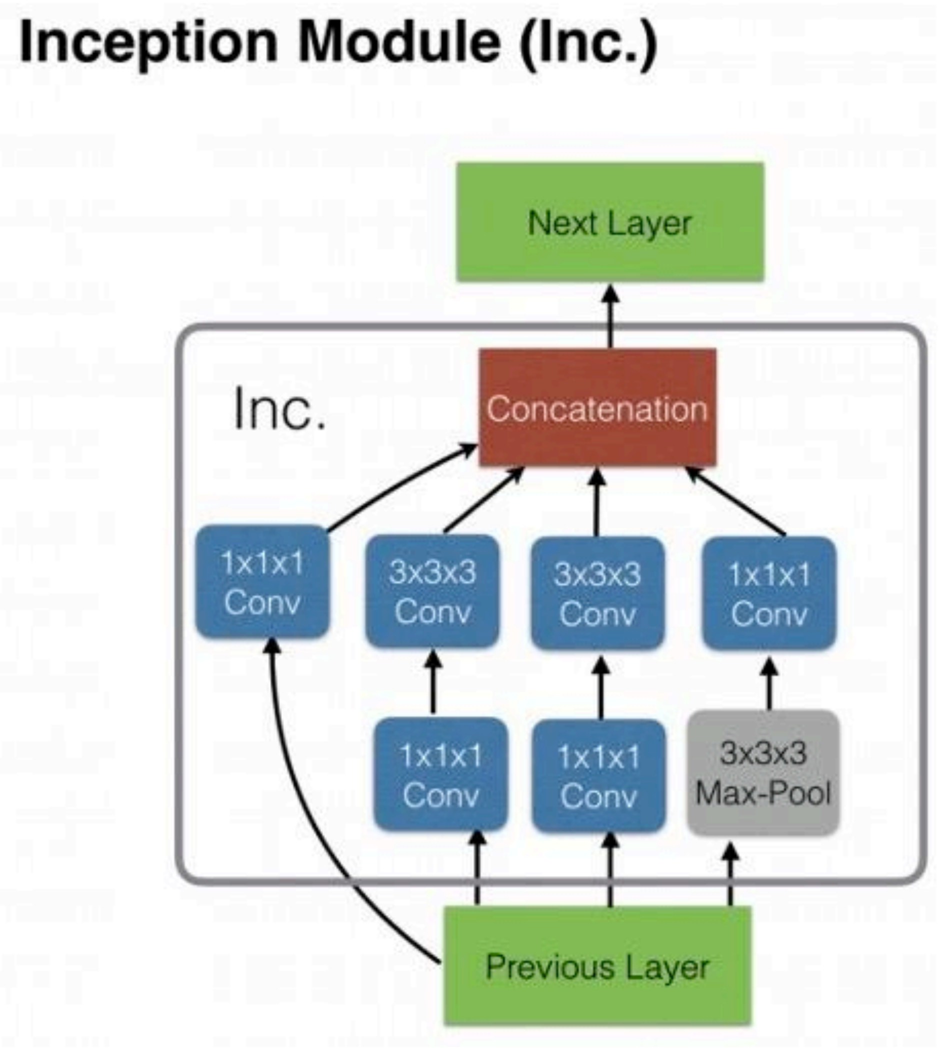
(d) tickling

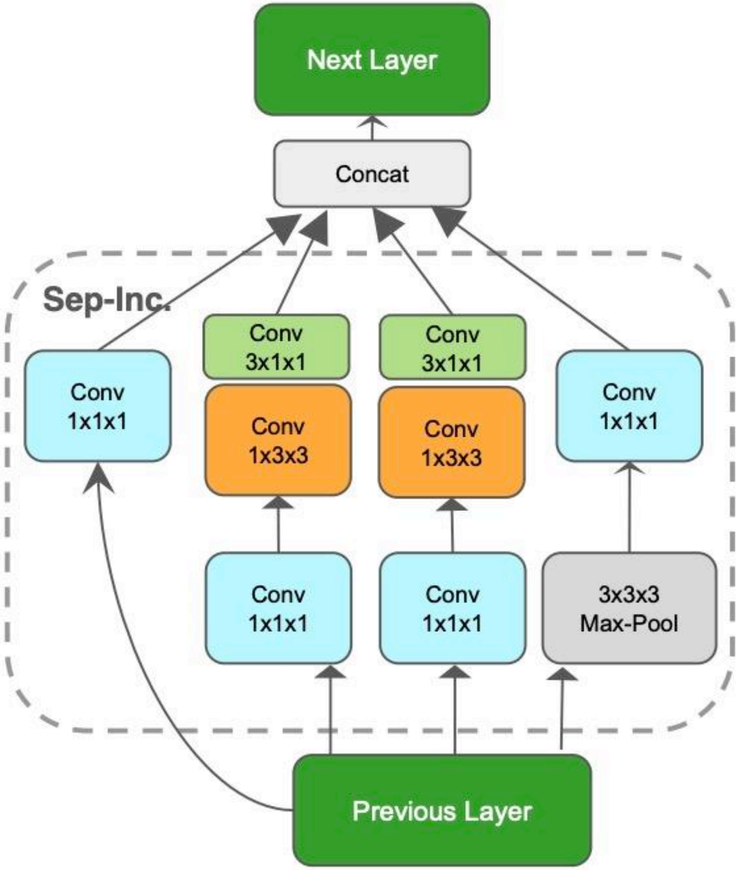
Inflated 3D ConvNets (I3D)



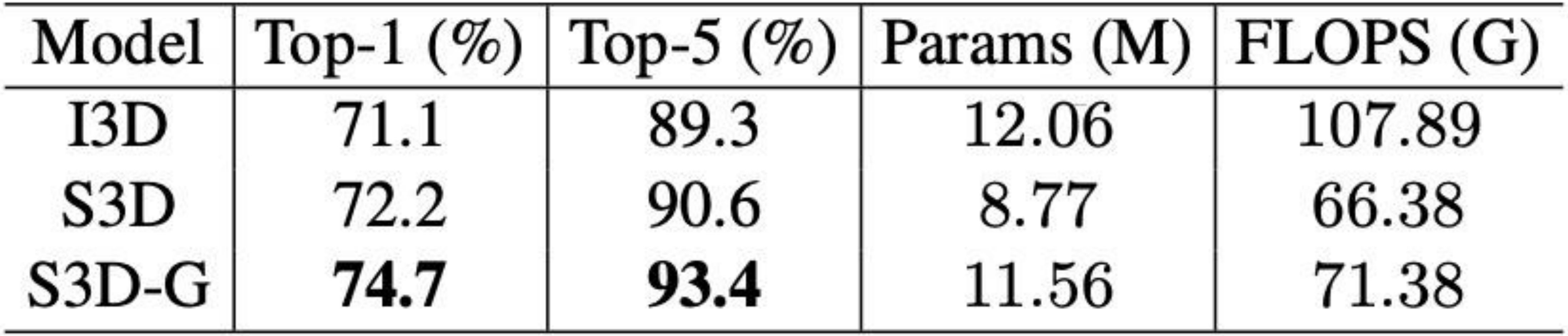
Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	63.3	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	56.1	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	62.2	52.4	65.6
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	–	–	67.2
(e) Two-Stream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

Separable 3D CNN (S3D)


Inception Module (Inc.)

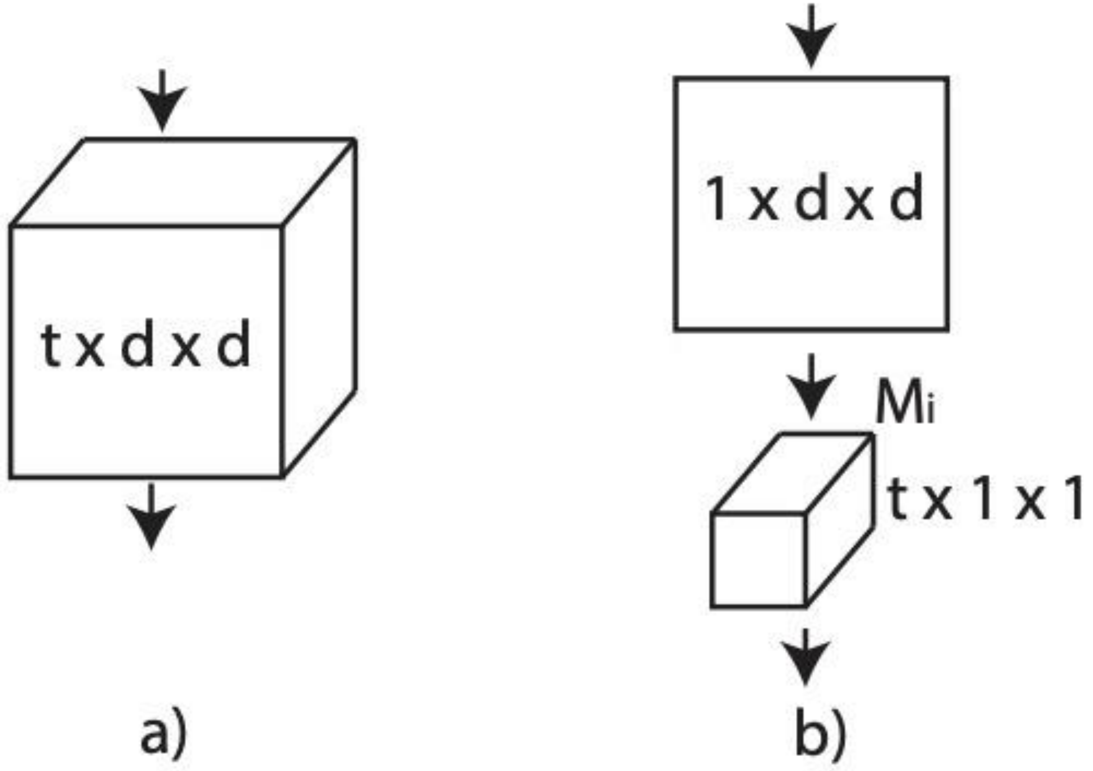


Separable 3D CNN (S3D)



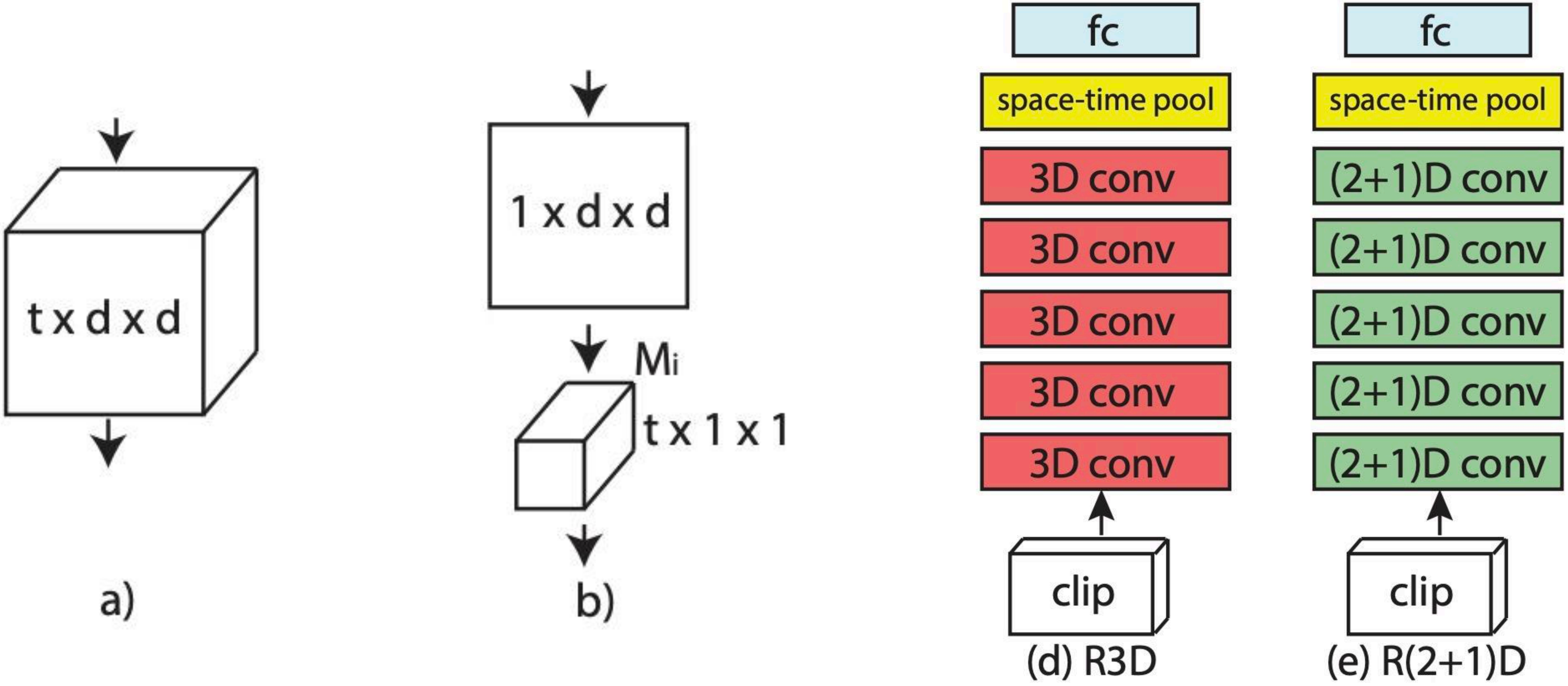
Model	Top-1 (%)	Top-5 (%)	Params (M)	FLOPS (G)
I3D	71.1	89.3	12.06	107.89
S3D	72.2	90.6	8.77	66.38
S3D-G	**74.7**	**93.4**	11.56	71.38

R(2+1)D



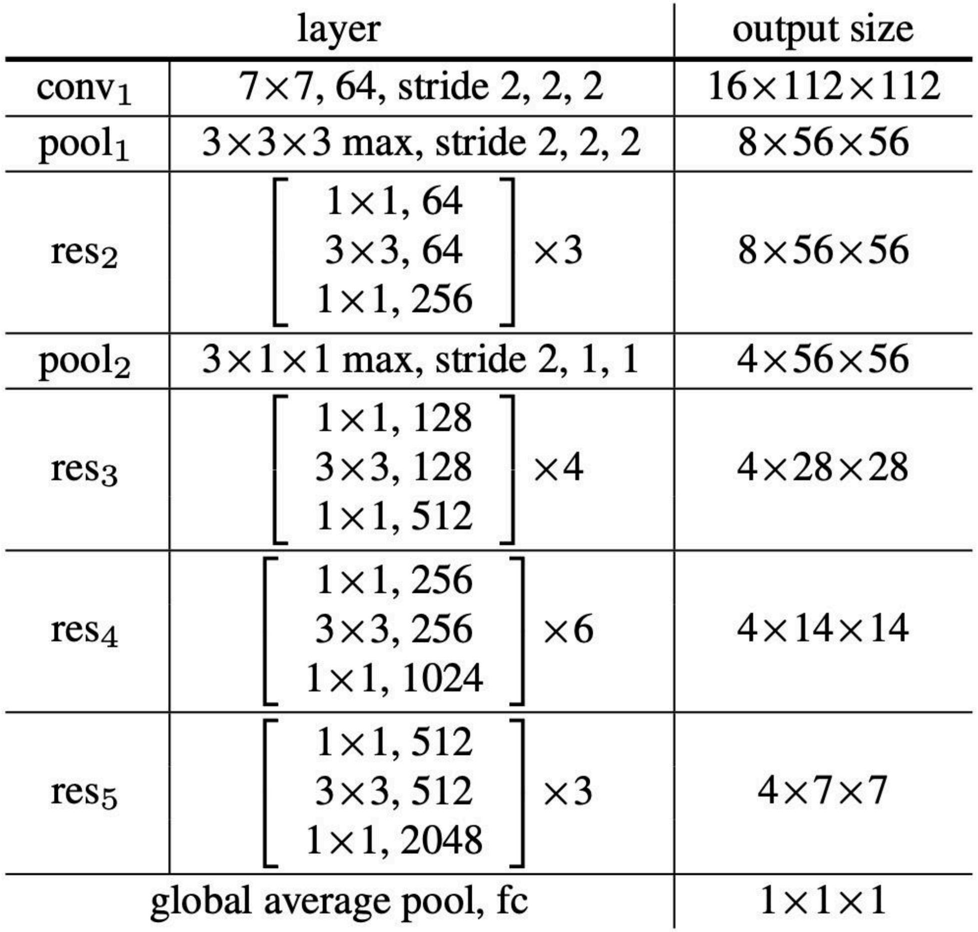
Tran et al., 2018

R(2+1)D



Tran et al., 2018

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

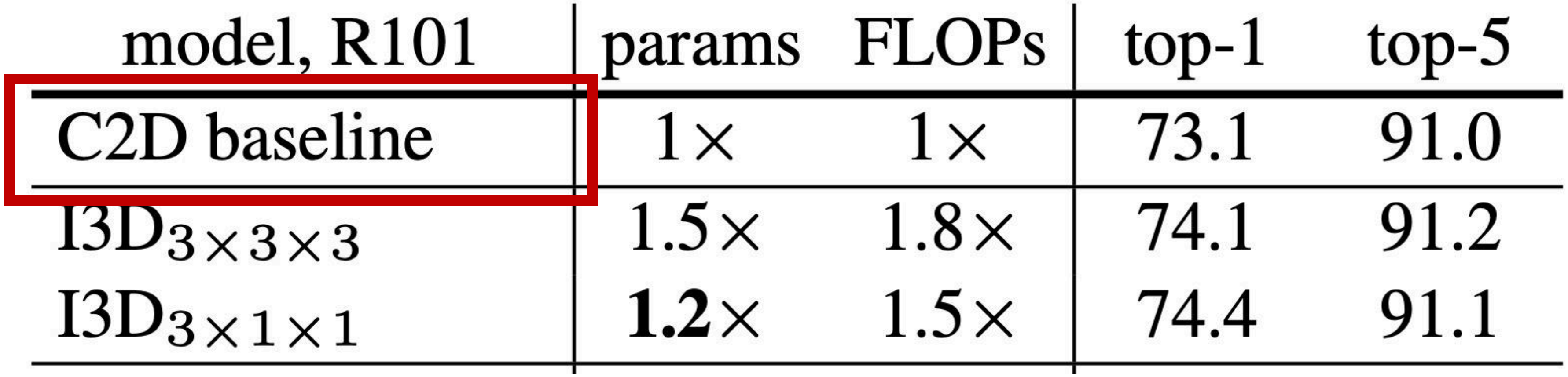


layer		output size
conv$_1$	7×7, 64, stride 2, 2, 2	16×112×112
pool$_1$	3×3×3 max, stride 2, 2, 2	8×56×56
res$_2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	8×56×56
pool$_2$	3×1×1 max, stride 2, 1, 1	4×56×56
res$_3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3,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

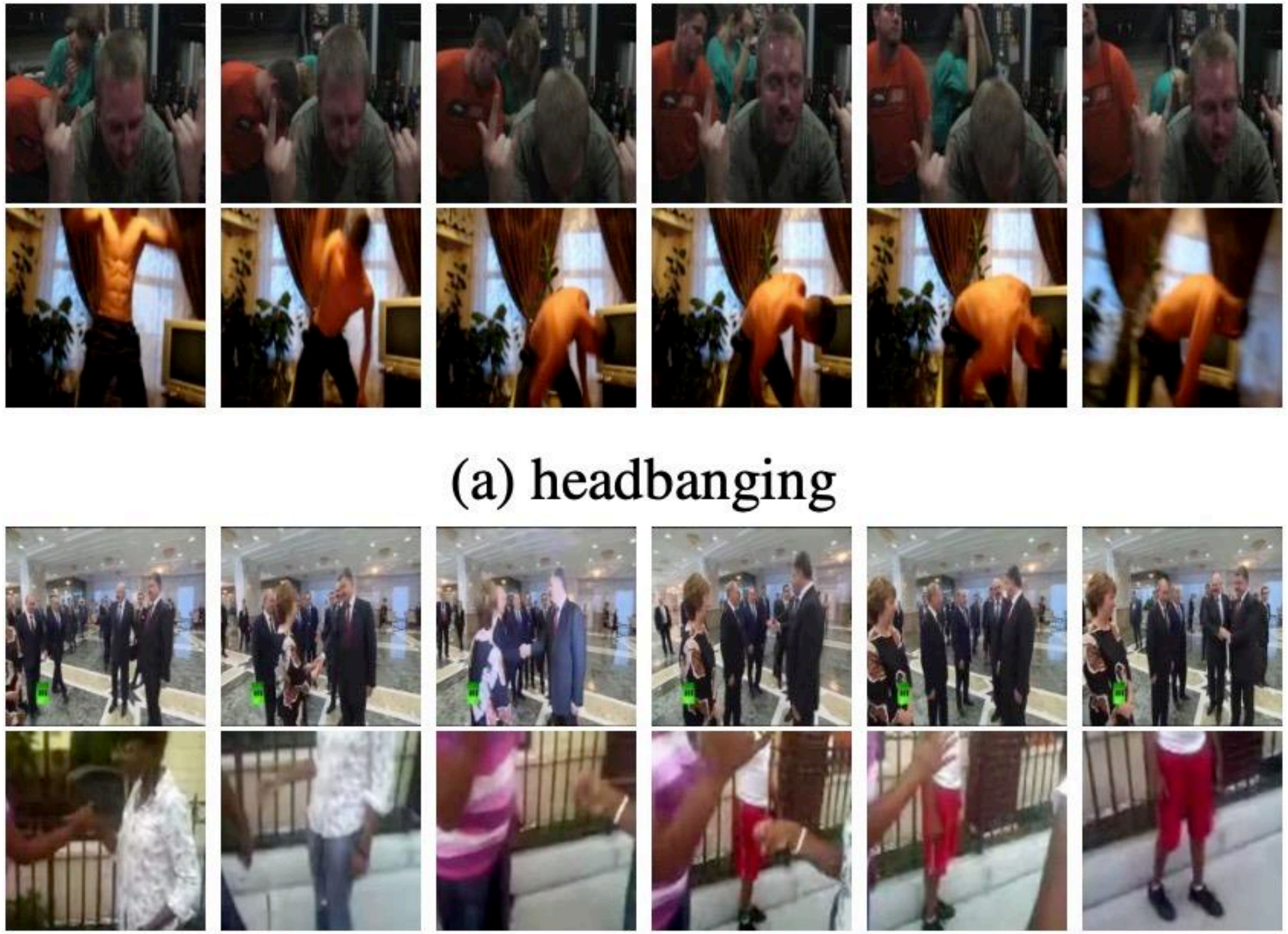
Same network, remove all temporal conv



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

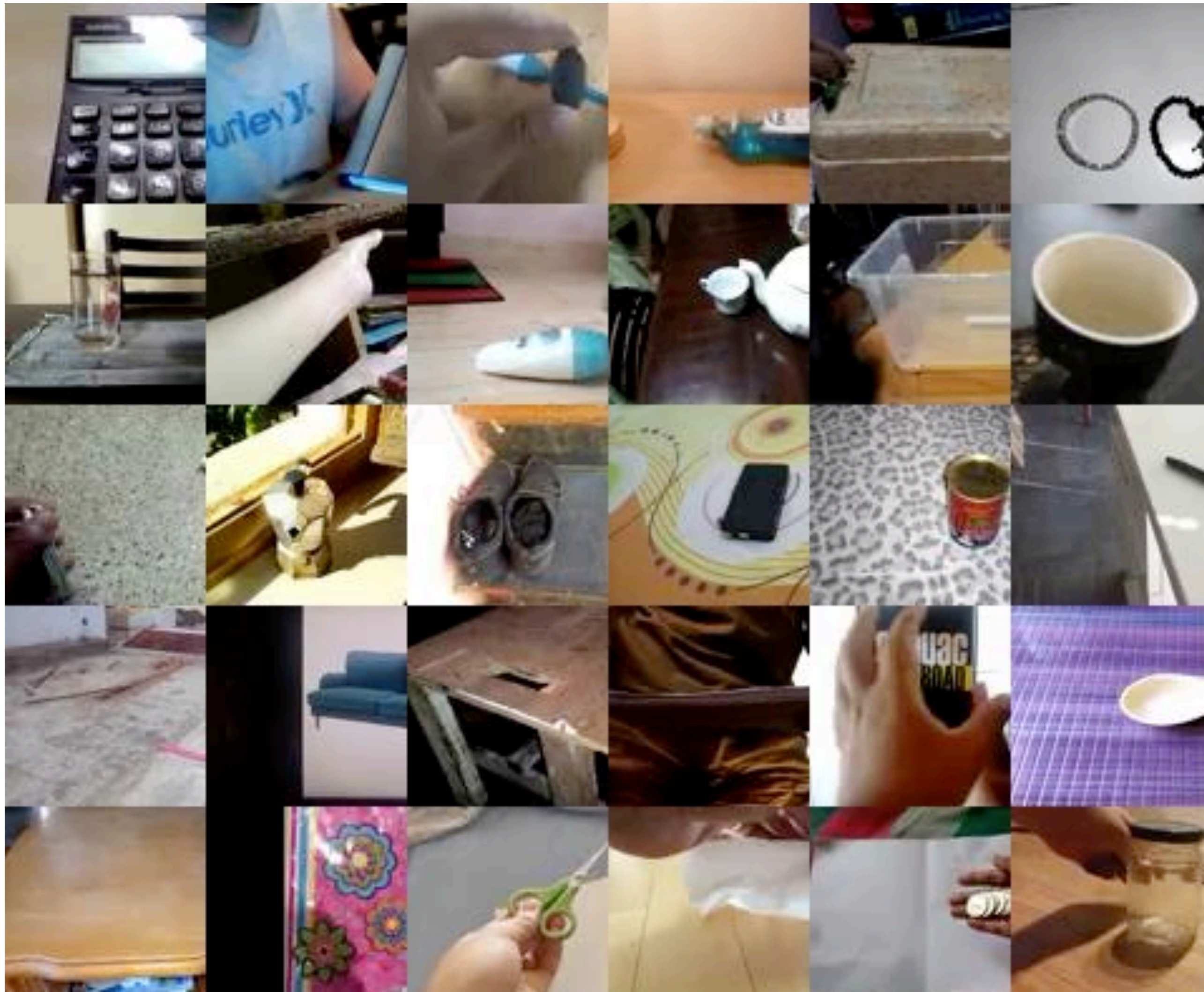


(a) headbanging

(c) shaking hands

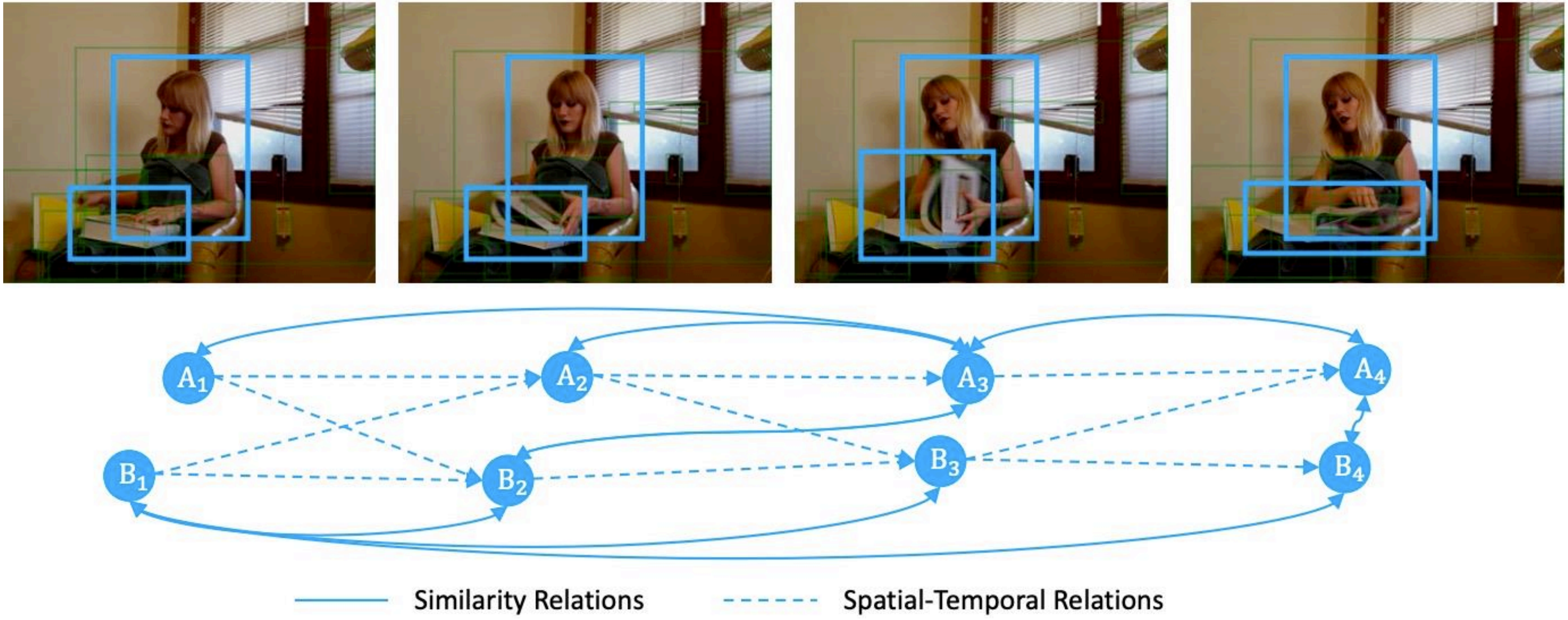
Something-Something Dataset

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



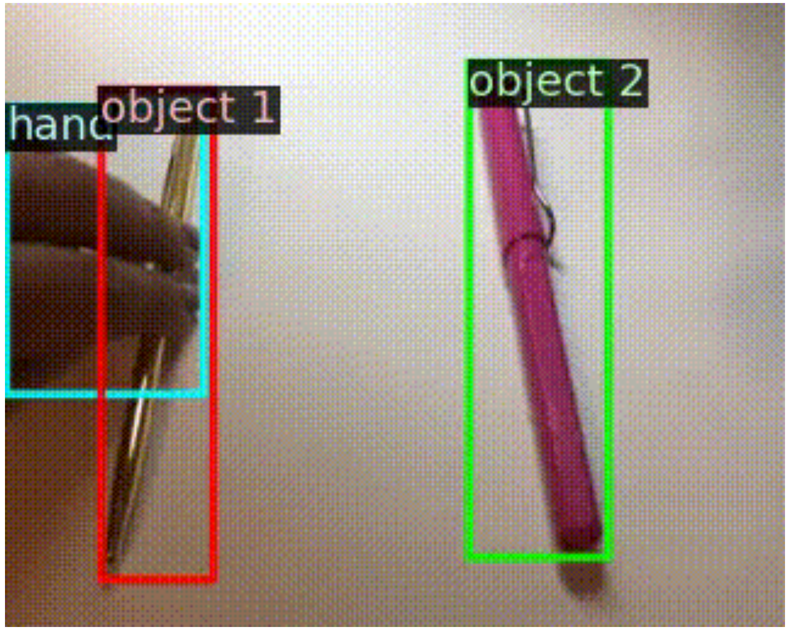
Spatial-Temporal Graph in Videos

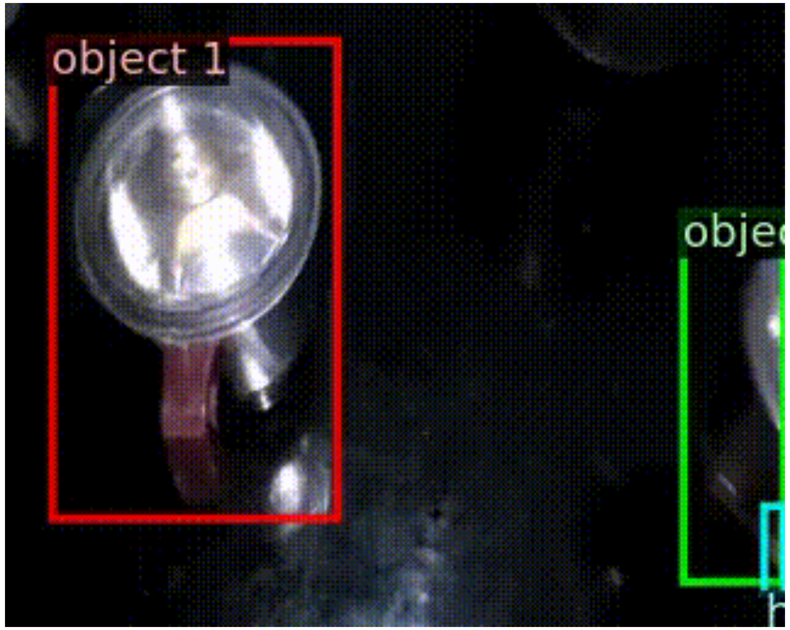
Videos as Space-Time Region Graphs

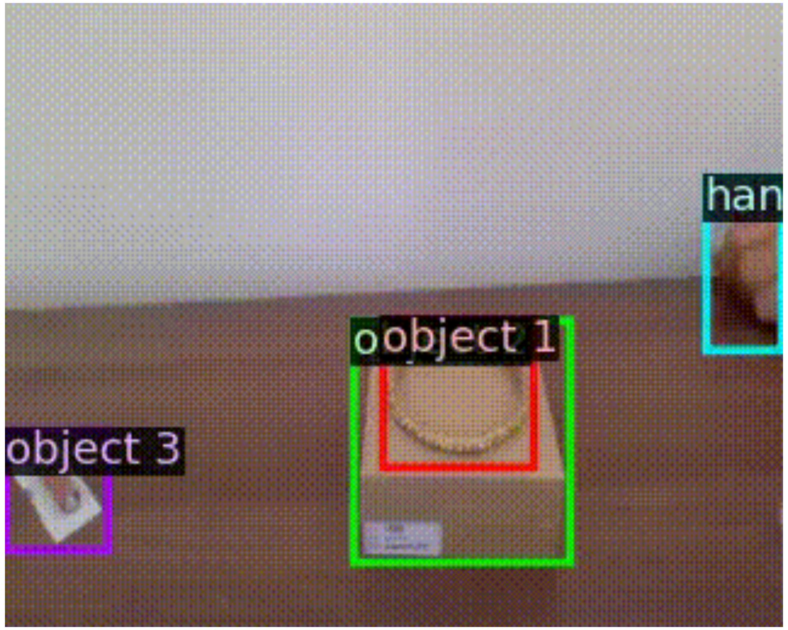


Wang et al., 2018

Space-Time Interactions

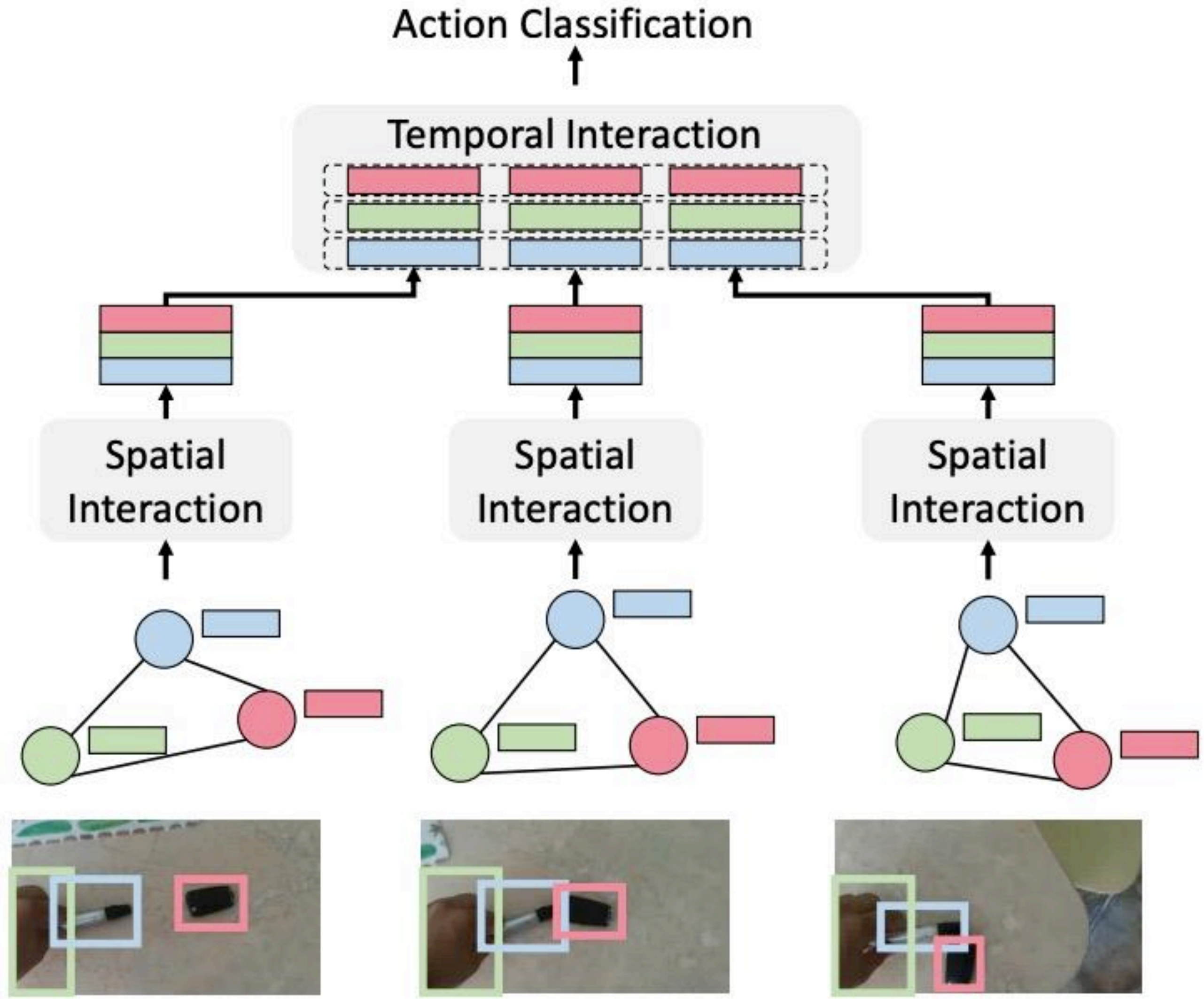




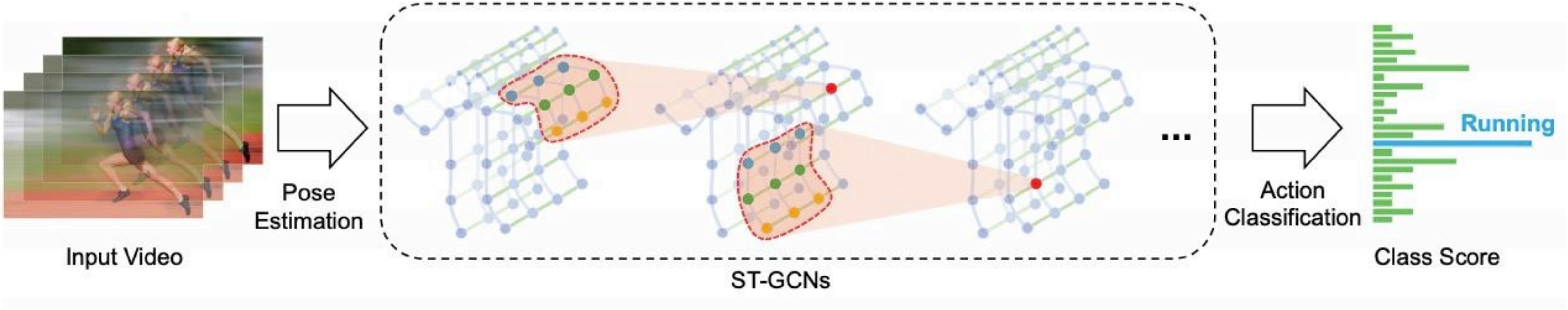


Materzynska et al., 2020

Space-Time Interactions



Skeleton-Based Action Recognition



Yan et al., 2018