Video Processing

ViViT: A Video Vision Transformer, ICCV21

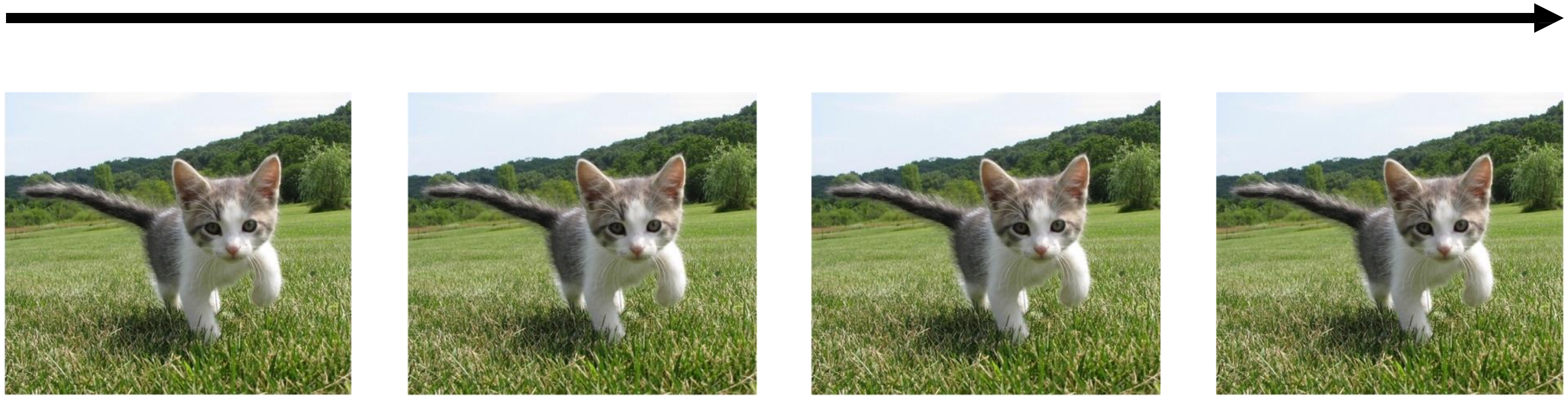
VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, NeurIPS22

Today: Video = 2D + Time

A video is a **sequence** of images

4D tensor: T x 3 x H x W

(or 3 x T x H x W)



Example task: Video Classification

Input video:
T x 3 x H x W

Swimming
Running
Jumping
Eating
Standing

Running video is in the public domain

Example task: Video Classification

Images: Recognize **objects**

Dog
Cat
Fish
Truck

Videos: Recognize **actions**

Swimming
Running
Jumping
Eating
Standing

Running video is in the public domain

Problem: Videos are big!

Input video:
T x 3 x H x W

Videos are ~30 frames per second (fps)

Size of uncompressed video
(3 bytes per pixel):

SD (640 x 480): **~1.5 GB per minute**
HD (1920 x 1080): ~**10 GB per minute**

Problem: Videos are big!

Input video:
T x 3 x H x W

Videos are ~30 frames per second (fps)

Size of uncompressed video
(3 bytes per pixel):

SD (640 x 480): **~1.5 GB per minute**
HD (1920 x 1080): ~**10 GB per minute**

Solution: Train on short **clips:** low fps and low spatial resolution
e.g. T = 16, H=W=112
(3.2 seconds at 5 fps, 588 KB)

Training on Clips

Raw video: Long, high FPS

Training on Clips

Raw video: Long, high FPS

Training: Train model to classify short **clips** with low FPS

Training on Clips

Raw video: Long, high FPS

Training: Train model to classify short **clips** with low FPS

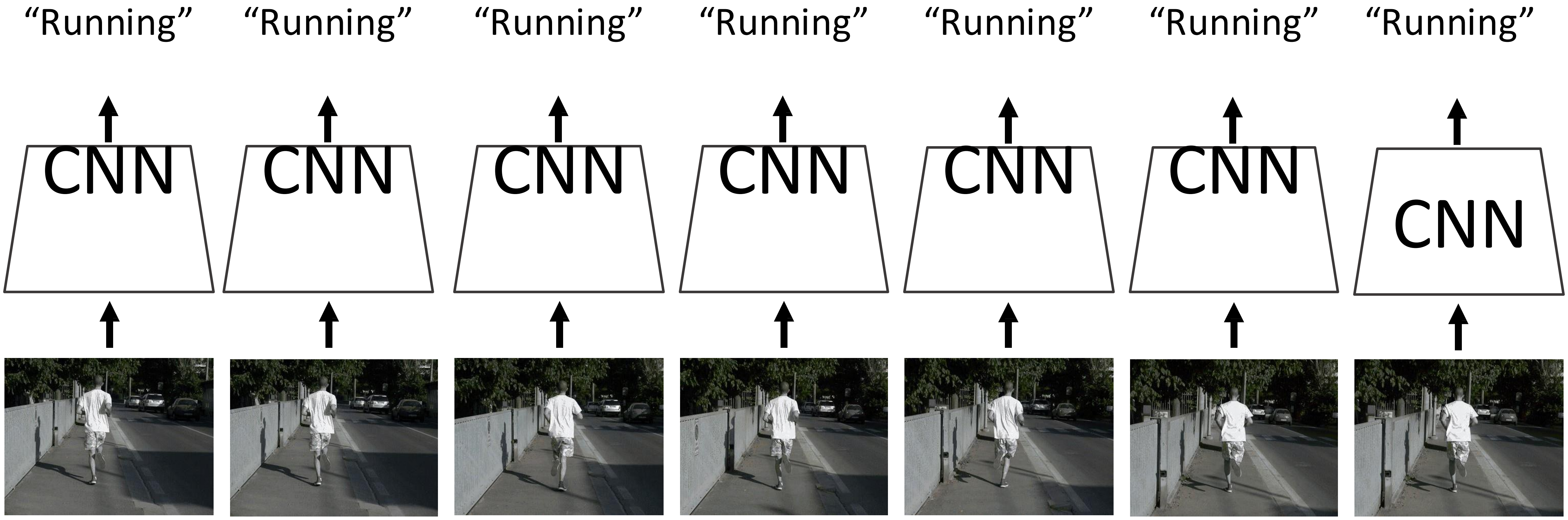
Testing: Run model on different clips, average predictions

- What could be a very simple video classifier?

Video Classification: Single-Frame CNN

Simple idea: train normal 2D CNN to classify video frames independently!

Often a **very** strong baseline for video classification

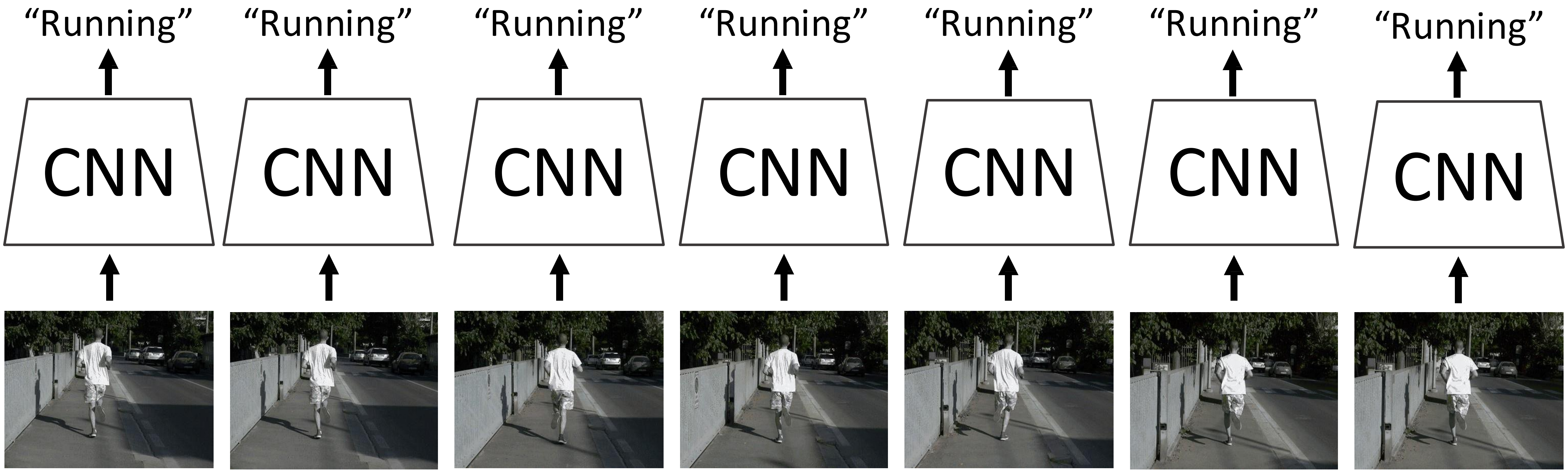


Video Classification: Single-Frame CNN

Simple idea: train normal 2D CNN to classify video frames independently!

(Average predicted probs at test-time)

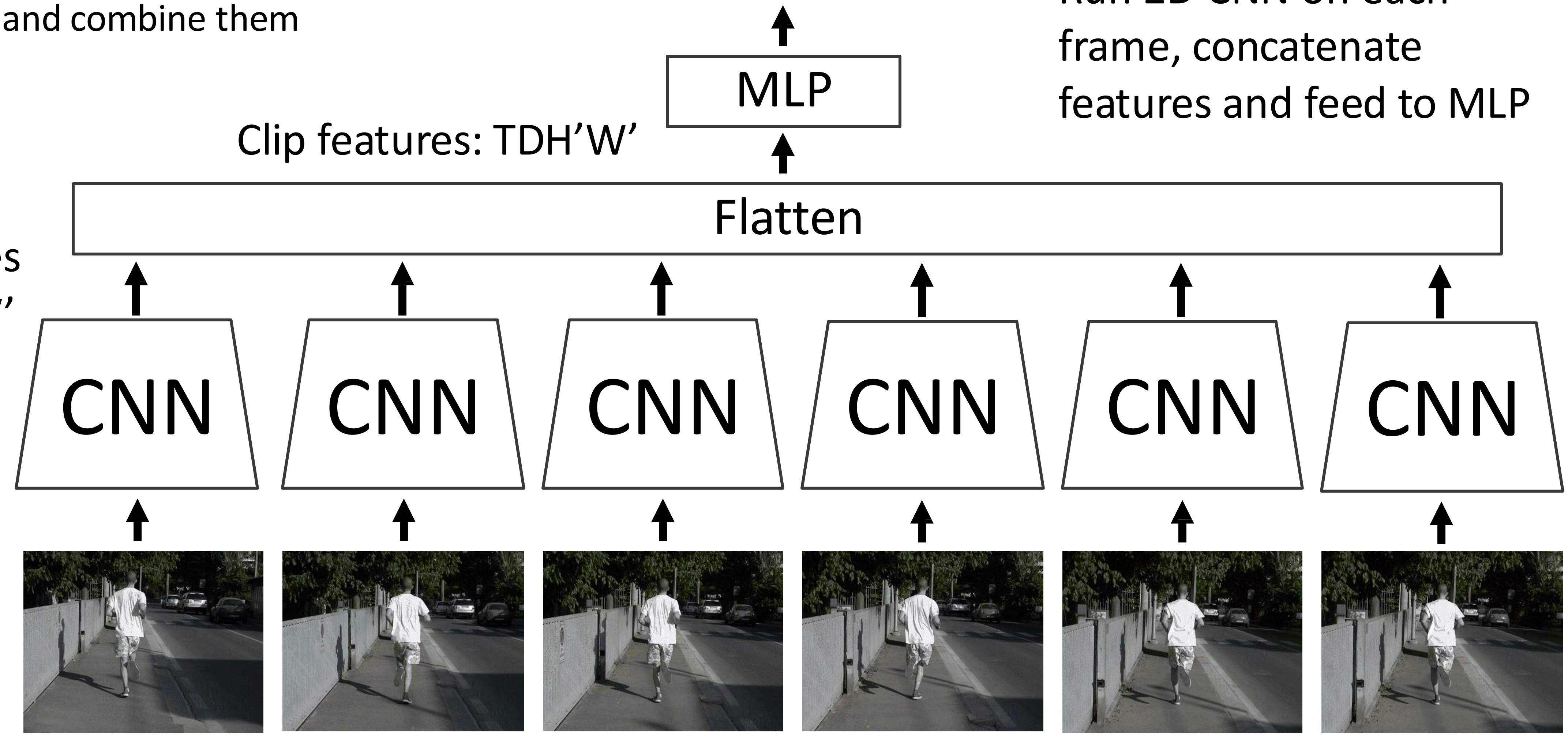
Often a **very** strong baseline for video classification



Video Classification: Late Fusion (with FC layers)

Intuition: Get high-level appearance of each frame, and combine them

Run 2D CNN on each frame, concatenate features and feed to MLP

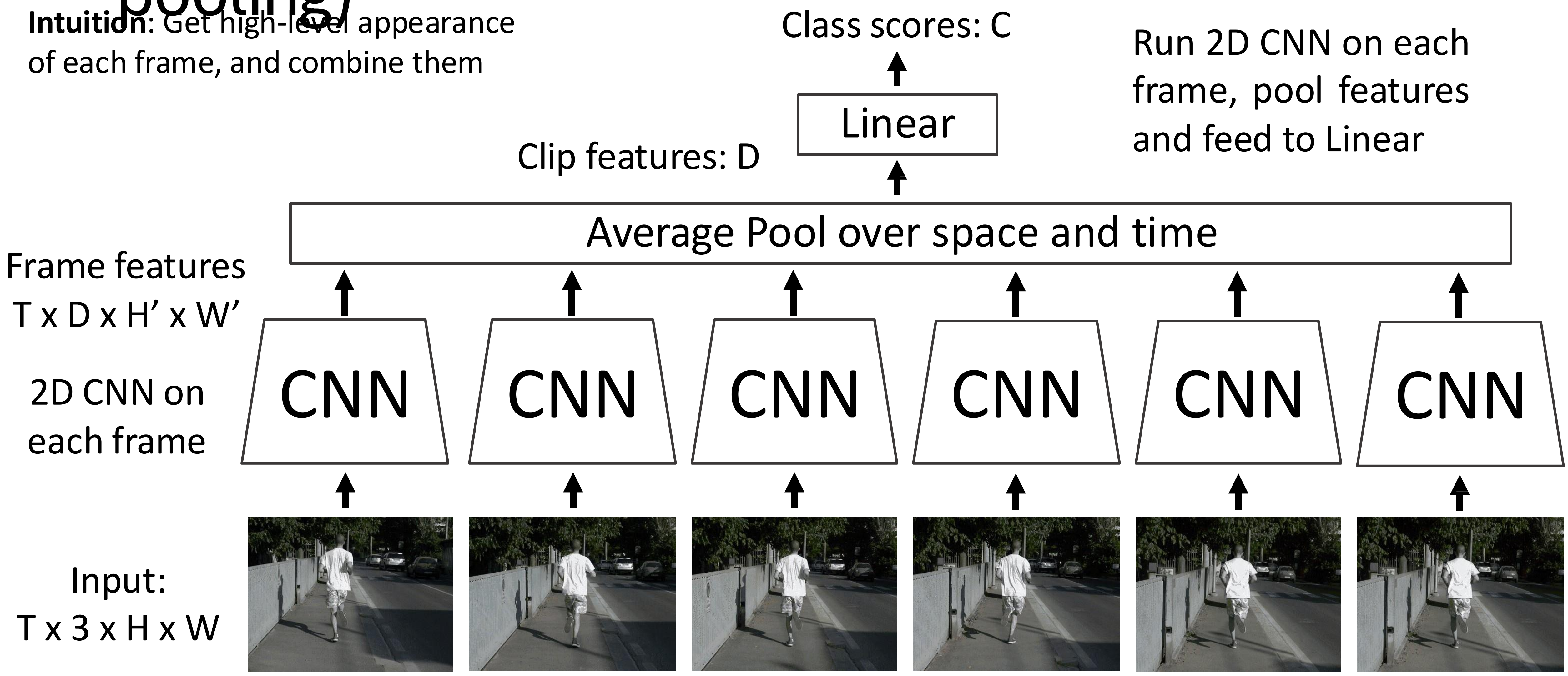


Karpathy et al, “Large-scale Video Classification with Convolutional Neural Networks”, CVPR 2014

Video Classification: Late Fusion (with pooling)

Intuition: Get high-level appearance of each frame, and combine them

Run 2D CNN on each frame, pool features and feed to Linear

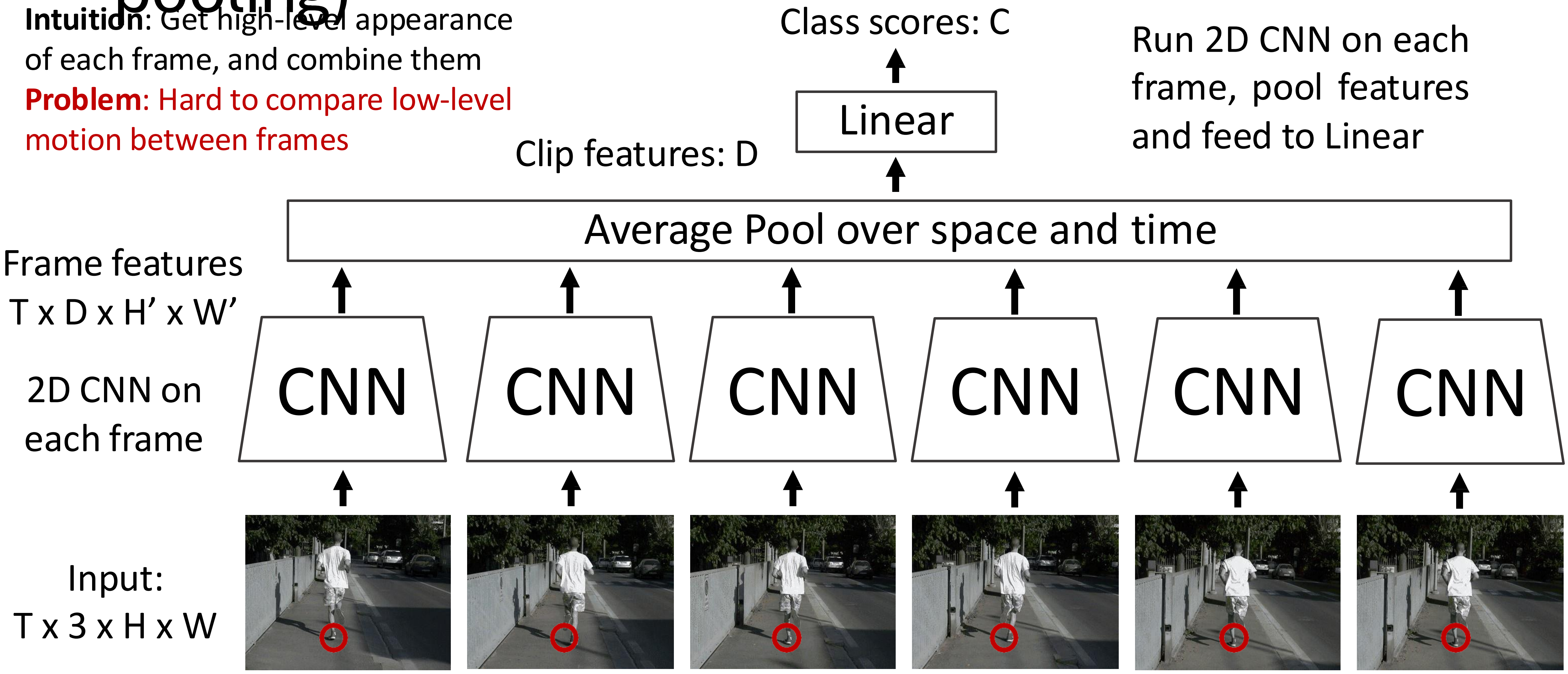


Video Classification: Late Fusion (with pooling)

Intuition: Get high-level appearance of each frame, and combine them

Problem: Hard to compare low-level motion between frames

Run 2D CNN on each frame, pool features and feed to Linear



Video Classification: Early Fusion

Intuition: Compare frames with very first conv layer, after that normal 2D CNN

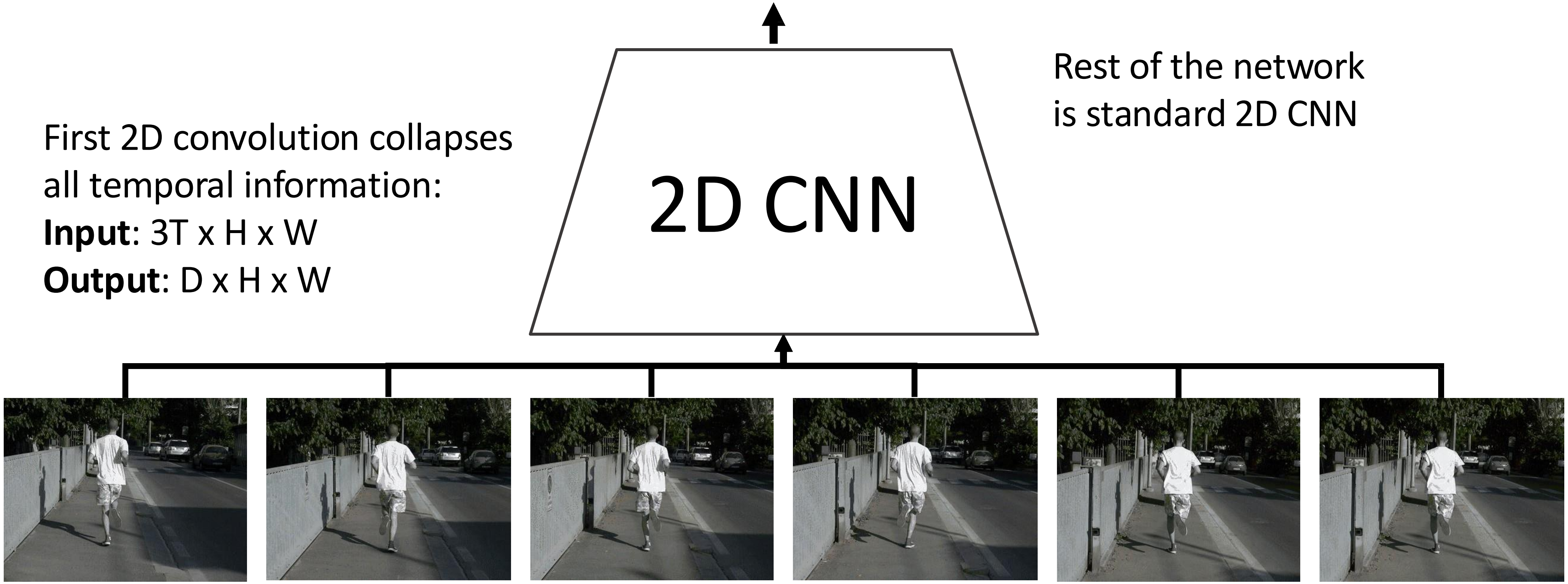
Class scores: C

Rest of the network is standard 2D CNN

First 2D convolution collapses all temporal information:
Input: 3T x H x W
Output: D x H x W

Reshape:
3T x H x W

Input:
T x 3 x H x W



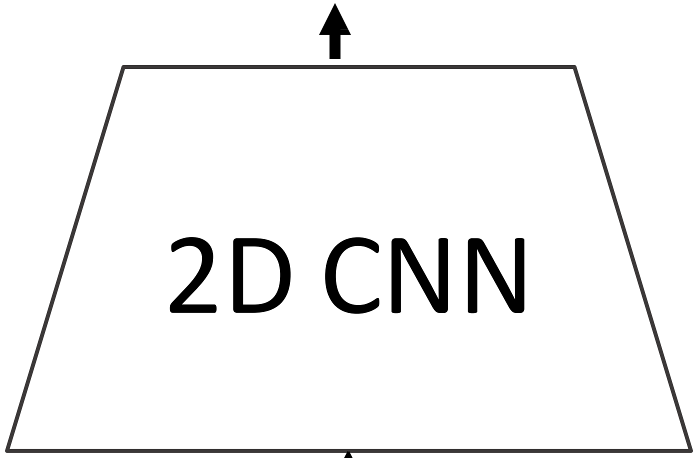
Karpathy et al, "Large-scale Video Classification with Convolutional Neural Networks", CVPR 2014

Video Classification: Early Fusion

Intuition: Compare frames with very first conv layer, after that normal 2D CNN

Problem: One layer of temporal processing may not be enough!

Class scores: C



Rest of the network is standard 2D CNN

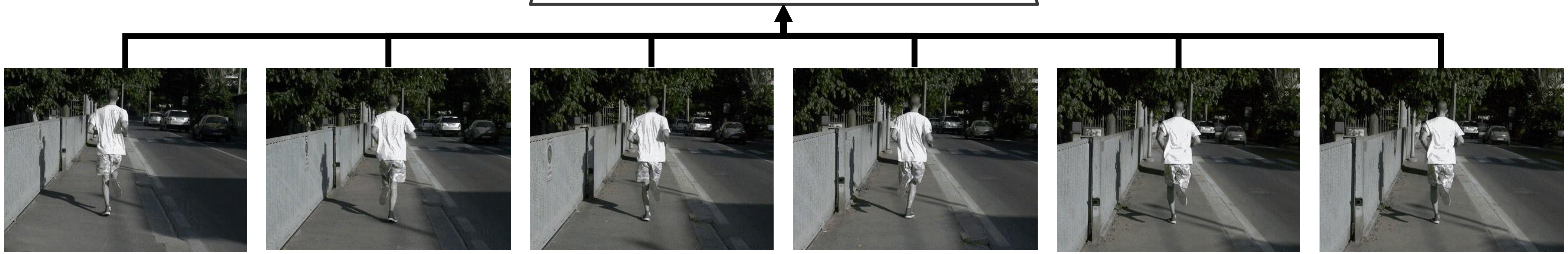
First 2D convolution collapses all temporal information:

Input: 3T x H x W

Output: D x H x W

Reshape:
3T x H x W

Input:
T x 3 x H x W

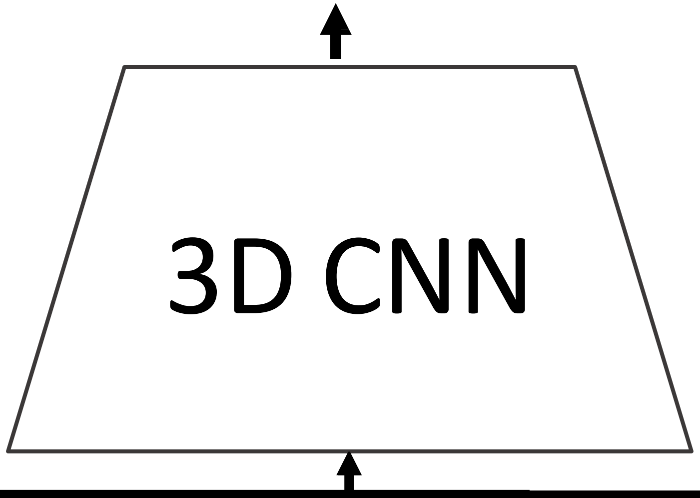


Karpathy et al, "Large-scale Video Classification with Convolutional Neural Networks", CVPR 2014

Video Classification: 3D CNN

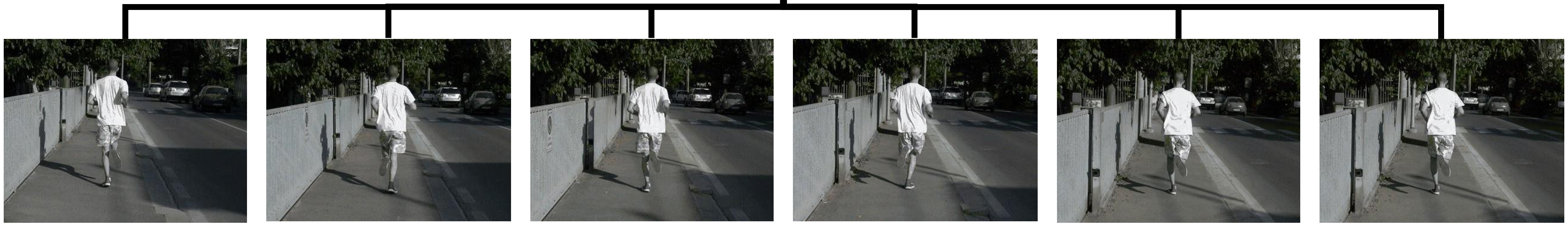
Intuition: Use 3D versions of convolution and pooling to slowly fuse temporal information over the course of the network

Class scores: C



Each layer in the network is a 4D tensor: D x T x H x W
Use 3D conv and 3D pooling operations

Input:
3 x T x H x W



Ji et al, “3D Convolutional Neural Networks for Human Action Recognition”, TPAMI 2010 ; Karpathy et al, “Large-scale Video Classification with Convolutional Neural Networks”, CVPR 2014

Early Fusion vs Late Fusion vs 3D CNN

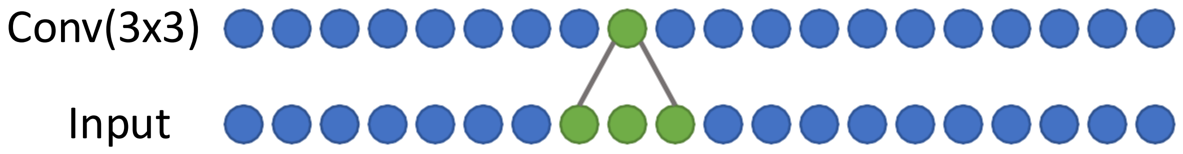
Late Fusion

Layer	Size (C x T x H x W)	Receptive Field (T x H x W)
Input	3 x 20 x 64 x 64	
Conv2D(3x3, 3->12)	12 x 20 x 64 x 64	1 x 3 x 3

Early Fusion vs Late Fusion vs 3D CNN

Late Fusion

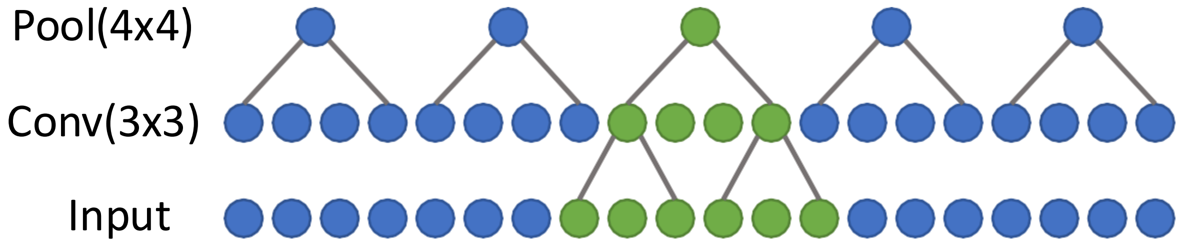
Layer	Size (C x T x H x W)	Receptive Field (T x H x W)
Input	3 x 20 x 64 x 64	
Conv2D(3x3, 3->12)	12 x 20 x 64 x 64	1 x 3 x 3



Early Fusion vs Late Fusion vs 3D CNN

Late Fusion

Layer	Size (C x T x H x W)	Receptive Field (T x H x W)
Input	3 x 20 x 64 x 64	
Conv2D(3x3, 3->12)	12 x 20 x 64 x 64	1 x 3 x 3
Pool2D(4x4)	12 x 20 x 16 x 16	1 x 6 x 6

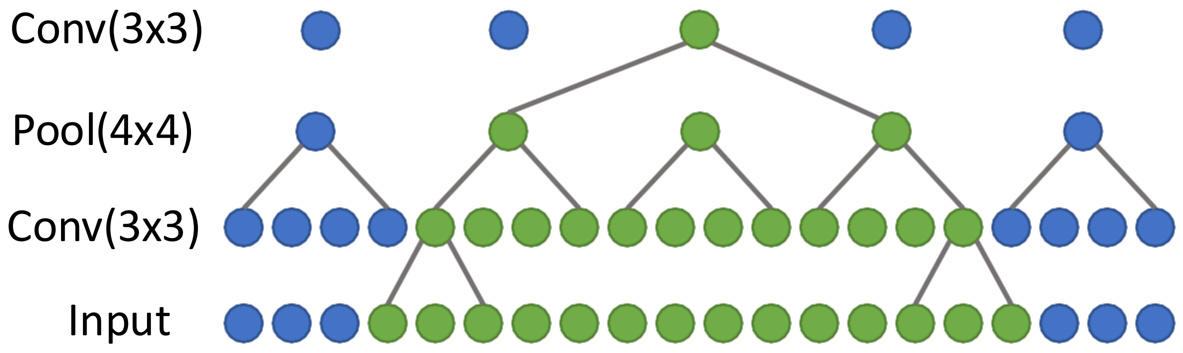


Early Fusion vs Late Fusion vs 3D CNN

Late Fusion

Layer	Size (C x T x H x W)	Receptive Field (T x H x W)
Input	3 x 20 x 64 x 64	
Conv2D(3x3, 3->12)	12 x 20 x 64 x 64	1 x 3 x 3
Pool2D(4x4)	12 x 20 x 16 x 16	1 x 6 x 6
Conv2D(3x3, 12->24)	24 x 20 x 16 x 16	1 x 14 x 14

Build slowly in space

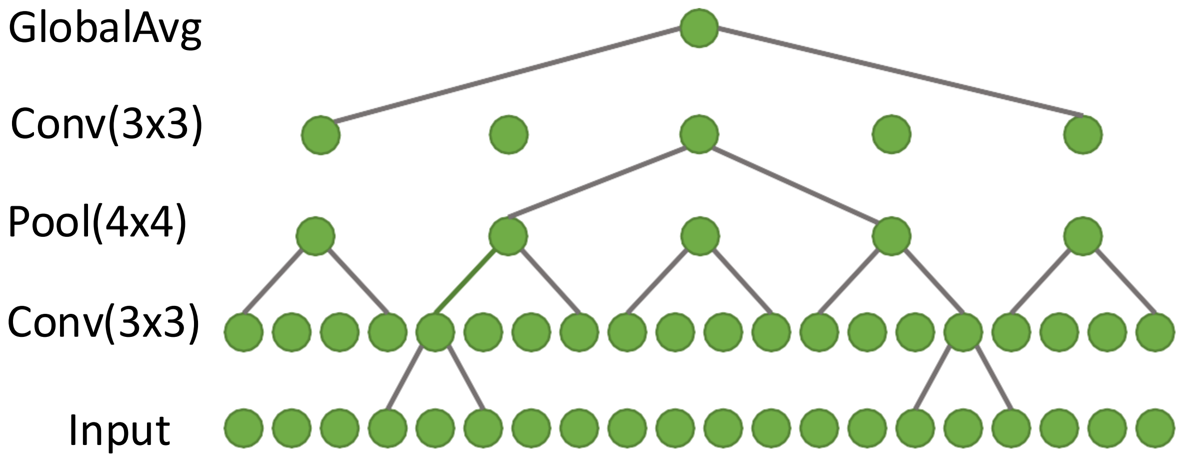


Early Fusion vs Late Fusion vs 3D CNN

Late Fusion

Layer	Size (C x T x H x W)	Receptive Field (T x H x W)
Input	3 x 20 x 64 x 64	
Conv2D(3x3, 3->12)	12 x 20 x 64 x 64	1 x 3 x 3
Pool2D(4x4)	12 x 20 x 16 x 16	1 x 6 x 6
Conv2D(3x3, 12->24)	24 x 20 x 16 x 16	1 x 14 x 14
GlobalAvgPool	24 x 1 x 1 x 1	20 x 64 x 64

Build slowly in space,
All-at-once in time at end



(Small example

Early Fusion vs Late Fusion vs 3D CNN

	Layer	Size (C x T x H x W)	Receptive Field (T x H x W)
Late Fusion	Input	3 x 20 x 64 x 64	
	Conv2D(3x3, 3->12)	12 x 20 x 64 x 64	1 x 3 x 3
	Pool2D(4x4)	12 x 20 x 16 x 16	1 x 6 x 6
	Conv2D(3x3, 12->24)	24 x 20 x 16 x 16	1 x 14 x 14
	GlobalAvgPool	24 x 1 x 1 x 1	20 x 64 x 64
Early Fusion	Input	3 x 20 x 64 x 64	
	Conv2D(3x3, 3*20->12)	12 x 64 x 64	20 x 3 x 3
	Pool2D(4x4)	12 x 16 x 16	20 x 6 x 6
	Conv2D(3x3, 12->24)	24 x 16 x 16	20 x 14 x 14
	GlobalAvgPool	24 x 1 x 1	20 x 64 x 64

Late Fusion: Build slowly in space,
All-at-once in time at end

Early Fusion: Build slowly in space,
All-at-once in time at start

Early Fusion vs Late Fusion vs 3D CNN

	Layer	Size (C x T x H x W)	Receptive Field (T x H x W)	
Late Fusion	Input	3 x 20 x 64 x 64		Build slowly in space, All-at-once in time at end
	Conv2D(3x3, 3->12)	12 x 20 x 64 x 64	1 x 3 x 3	
	Pool2D(4x4)	12 x 20 x 16 x 16	1 x 6 x 6	
	Conv2D(3x3, 12->24)	24 x 20 x 16 x 16	1 x 14 x 14	
	GlobalAvgPool	24 x 1 x 1 x 1	20 x 64 x 64	
Early Fusion	Input	3 x 20 x 64 x 64		Build slowly in space, All-at-once in time at start
	Conv2D(3x3, 3*10->12)	12 x 64 x 64	20 x 3 x 3	
	Pool2D(4x4)	12 x 16 x 16	20 x 6 x 6	
	Conv2D(3x3, 12->24)	24 x 16 x 16	20 x 14 x 14	
	GlobalAvgPool	24 x 1 x 1	20 x 64 x 64	
3D CNN	Input	3 x 20 x 64 x 64		Build slowly in space, Build slowly in time ”Slow Fusion”
	Conv3D(3x3x3, 3->12)	12 x 20 x 64 x 64	3 x 3 x 3	
	Pool3D(4x4x4)	12 x 5 x 16 x 16	6 x 6 x 6	
	Conv3D(3x3x3, 12->24)	24 x 5 x 16 x 16	14 x 14 x 14	
	GlobalAvgPool	24 x 1 x 1	20 x 64 x 64	

(Small example architectures, in practice much bigger)

Early Fusion vs Late Fusion vs 3D CNN

What is the difference?

	Layer	Size (C x T x H x W)	Receptive Field (T x H x W)
Late Fusion	Input	3 x 20 x 64 x 64	
	Conv2D(3x3, 3->12)	12 x 20 x 64 x 64	1 x 3 x 3
	Pool2D(4x4)	12 x 20 x 16 x 16	1 x 6 x 6
	Conv2D(3x3, 12->24)	24 x 20 x 16 x 16	1 x 14 x 14
	GlobalAvgPool	24 x 1 x 1 x 1	20 x 64 x 64
Early Fusion	Input	3 x 20 x 64 x 64	
	Conv2D(3x3, 3*10->12)	12 x 64 x 64	20 x 3 x 3
	Pool2D(4x4)	12 x 16 x 16	20 x 6 x 6
	Conv2D(3x3, 12->24)	24 x 16 x 16	20 x 14 x 14
	GlobalAvgPool	24 x 1 x 1	20 x 64 x 64
3D CNN	Input	3 x 20 x 64 x 64	
	Conv3D(3x3x3, 3->12)	12 x 20 x 64 x 64	3 x 3 x 3
	Pool3D(4x4x4)	12 x 5 x 16 x 16	6 x 6 x 6
	Conv3D(3x3x3, 12->24)	24 x 5 x 16 x 16	14 x 14 x 14
	GlobalAvgPool	24 x 1 x 1	20 x 64 x 64

Build slowly in space,
All-at-once in time at end

Build slowly in space,
All-at-once in time at start

Build slowly in space,
Build slowly in time
"Slow Fusion"

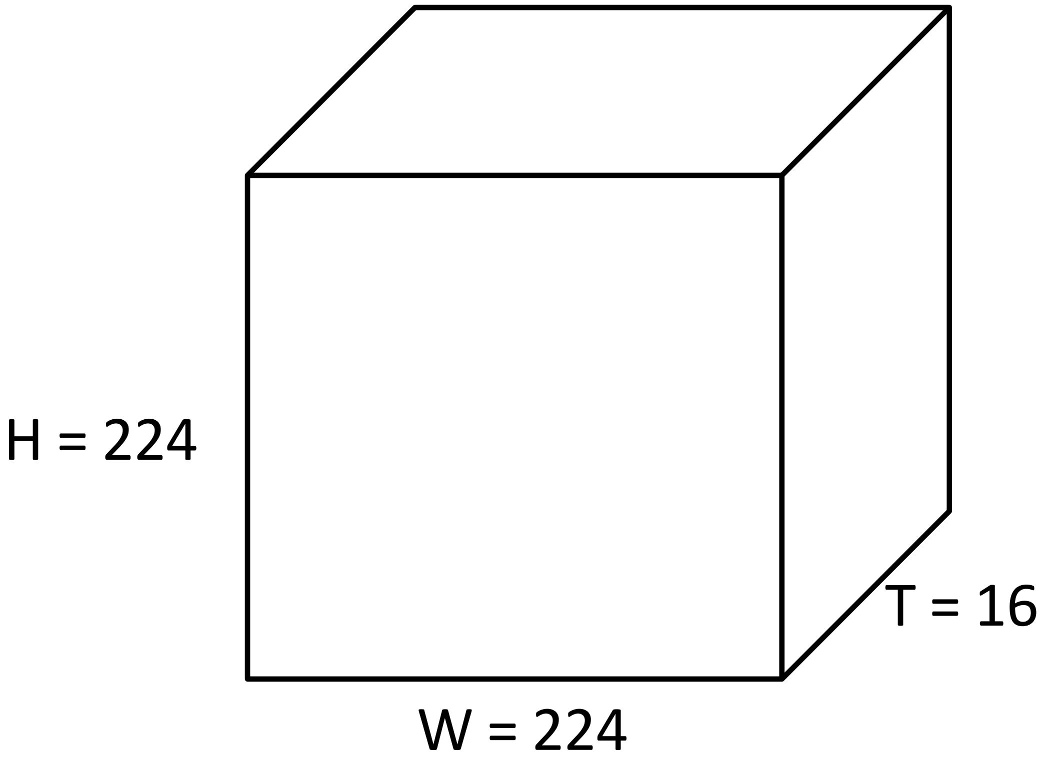
(Small example architectures, in practice much bigger)

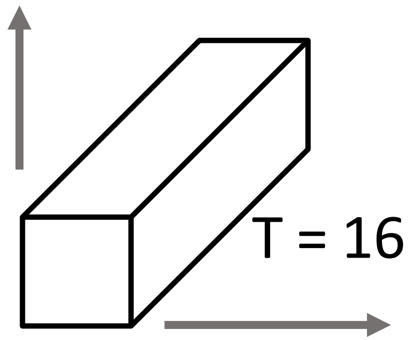
2D Conv (Early Fusion) vs 3D Conv (3D CNN)

Input: C_{in} x T x H x W
(3D grid with C_{in}-dim feat at each point)

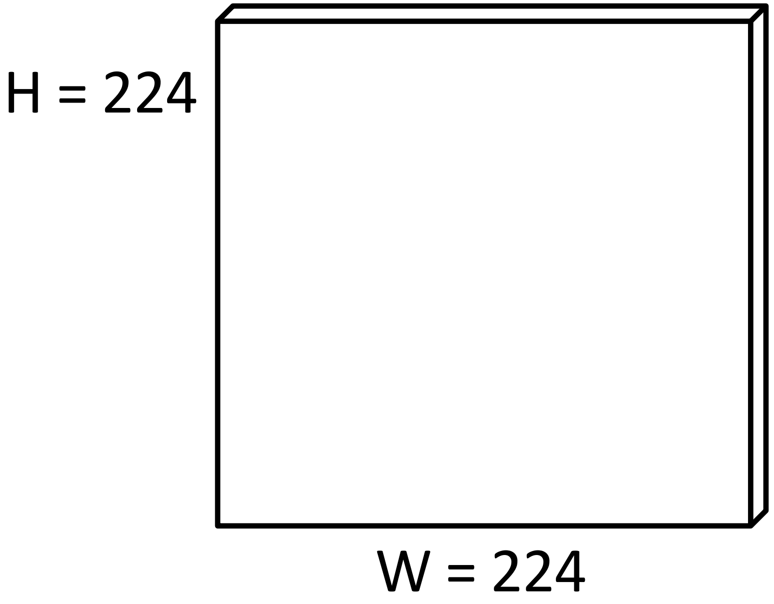
Weight:
C_{out} x C_{in} x T x 3 x 3
Slide over x and y

Output:
C_{out} x H x W
2D grid with C_{out} –dim feat at each point





C_{out} different filters

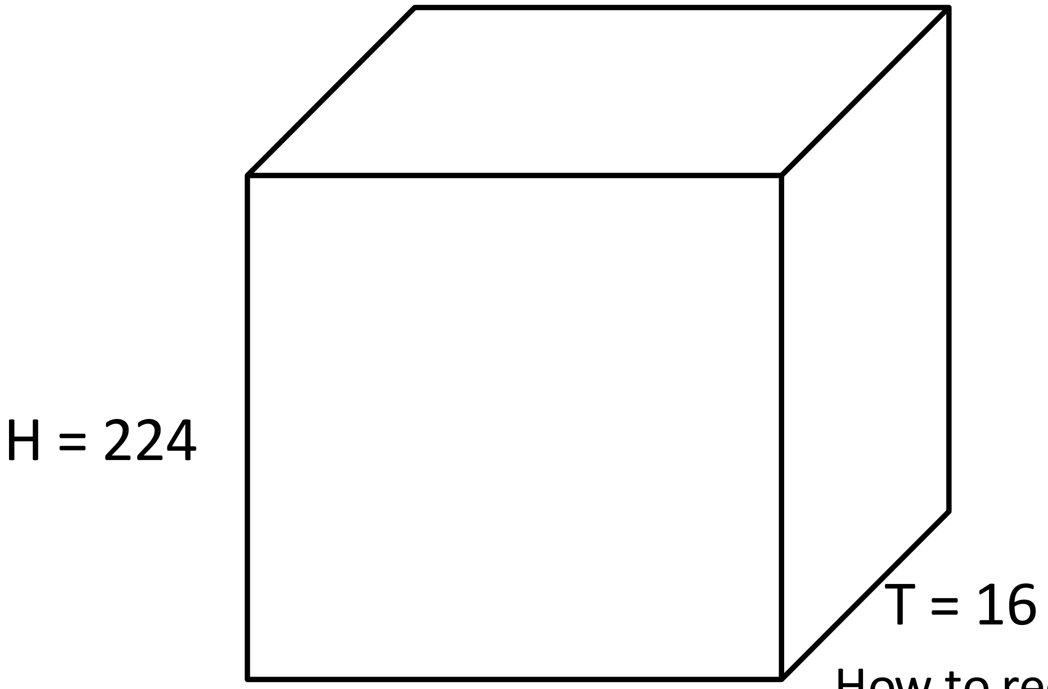


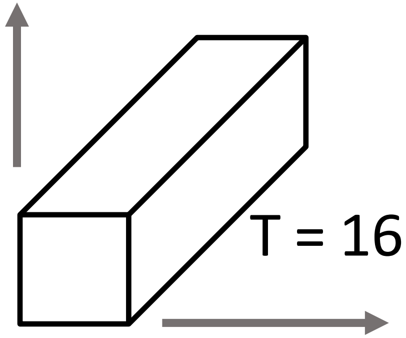
2D Conv (Early Fusion) vs 3D Conv (3D CNN)

Input: C_{in} x T x H x W
(3D grid with C_{in}-dim feat at each point)

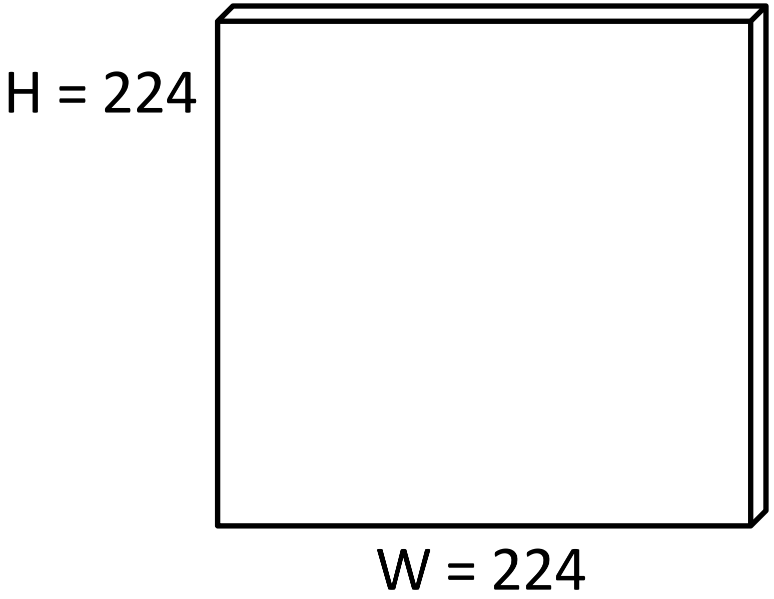
Weight:
C_{out} x C_{in} x T x 3 x 3
Slide over x and y

Output:
C_{out} x H x W
2D grid with C_{out} –dim feat at each point





C_{out} different filters



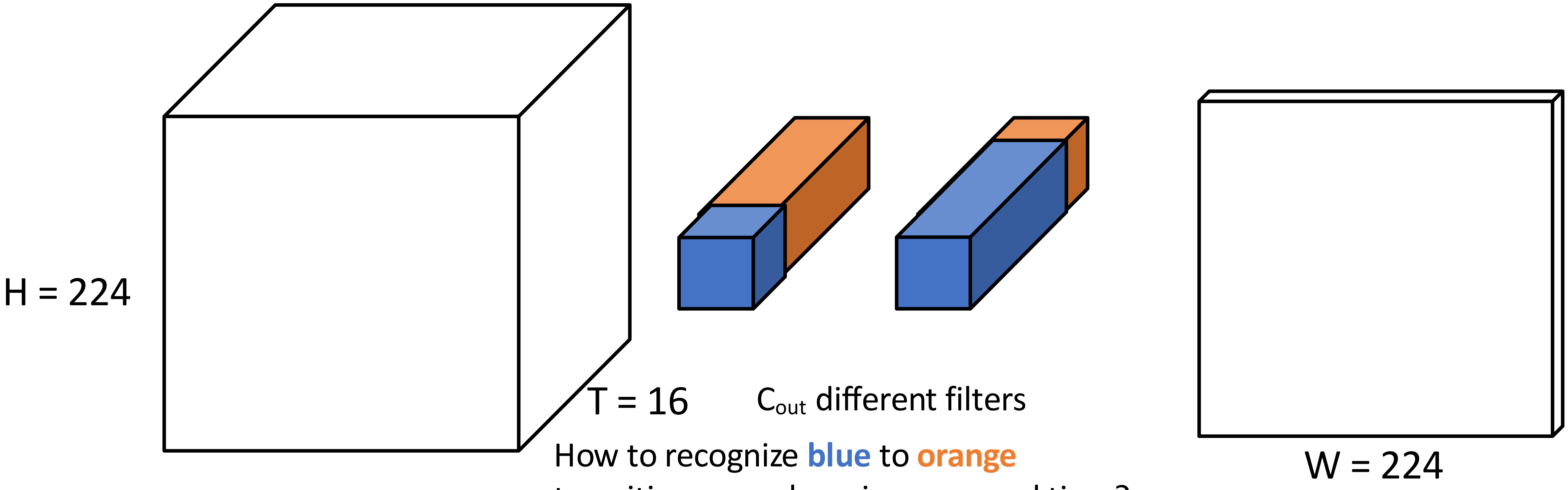
How to recognize **blue** to **orange** transitions anywhere in space and time?

2D Conv (Early Fusion) vs 3D Conv (3D CNN)

Input: C_{in} x T x H x W
(3D grid with C_{in}-dim feat at each point)

Weight:
C_{out} x C_{in} x T x 3 x 3
Slide over x and y

Output:
C_{out} x H x W
2D grid with C_{out} –dim



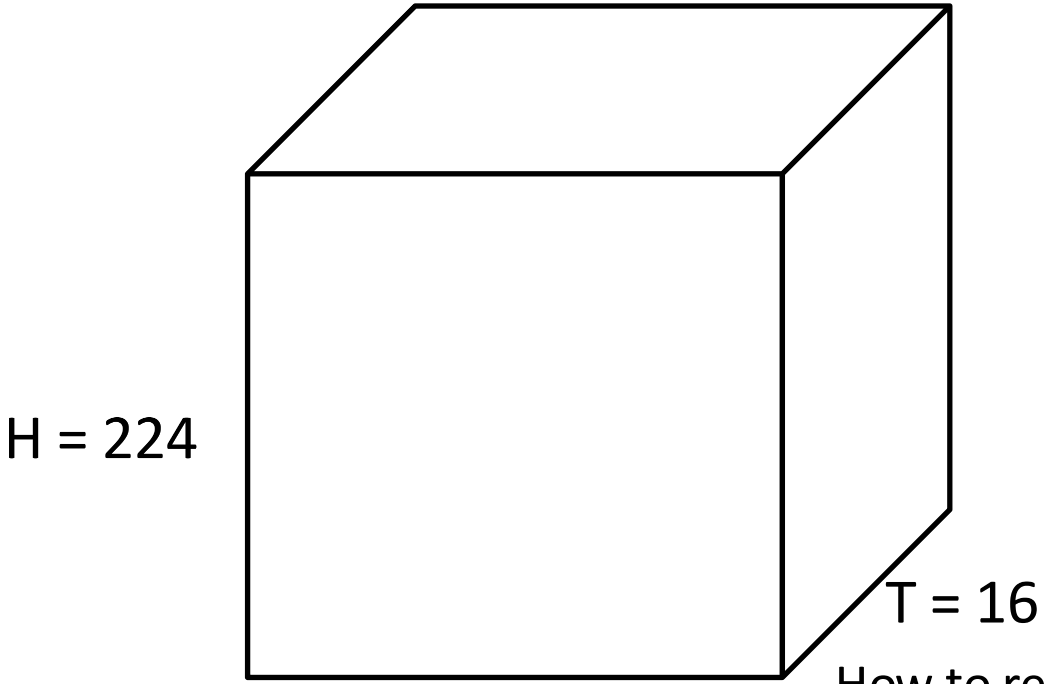
How to recognize **blue** to **orange** transitions anywhere in space and time?

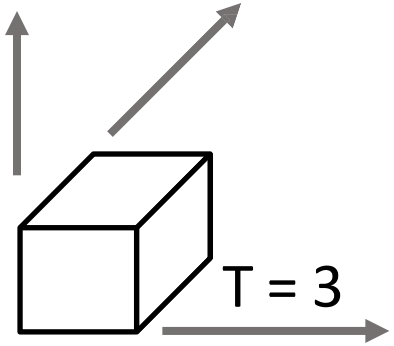
2D Conv (Early Fusion) vs 3D Conv (3D CNN)

Input: C_{in} x T x H x W
(3D grid with C_{in}-dim
feat at each point)

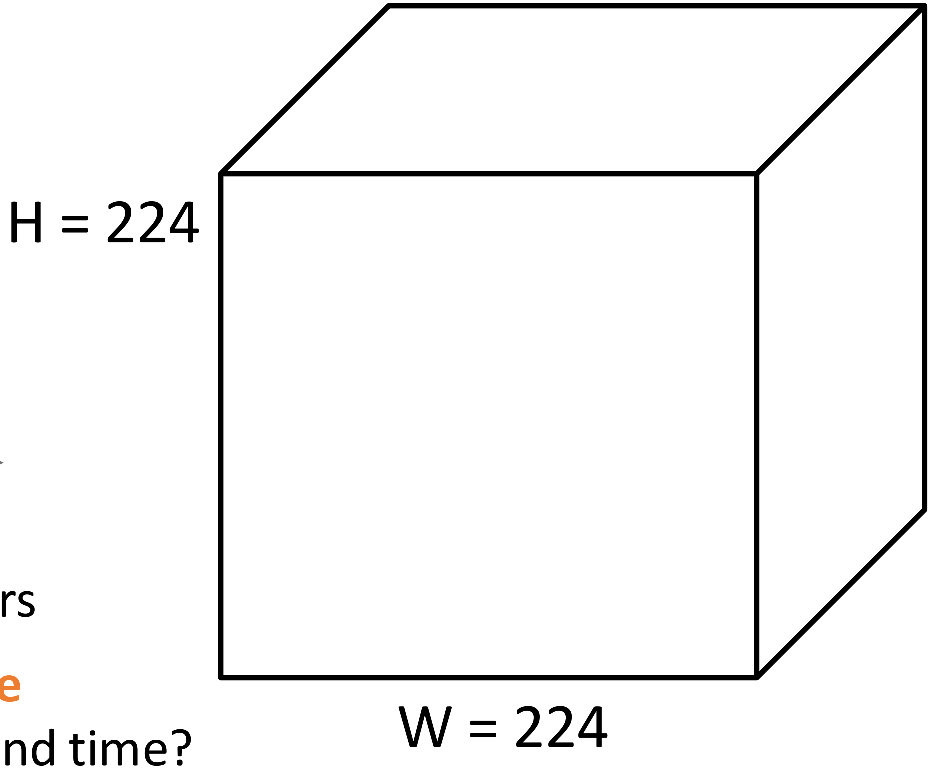
Weight:
C_{out} x C_{in} x 3 x 3 x 3
Slide over x and y

Output:
C_{out} x T x H x W
3D grid with C_{out}–dim
feat at each point





C_{out} different filters



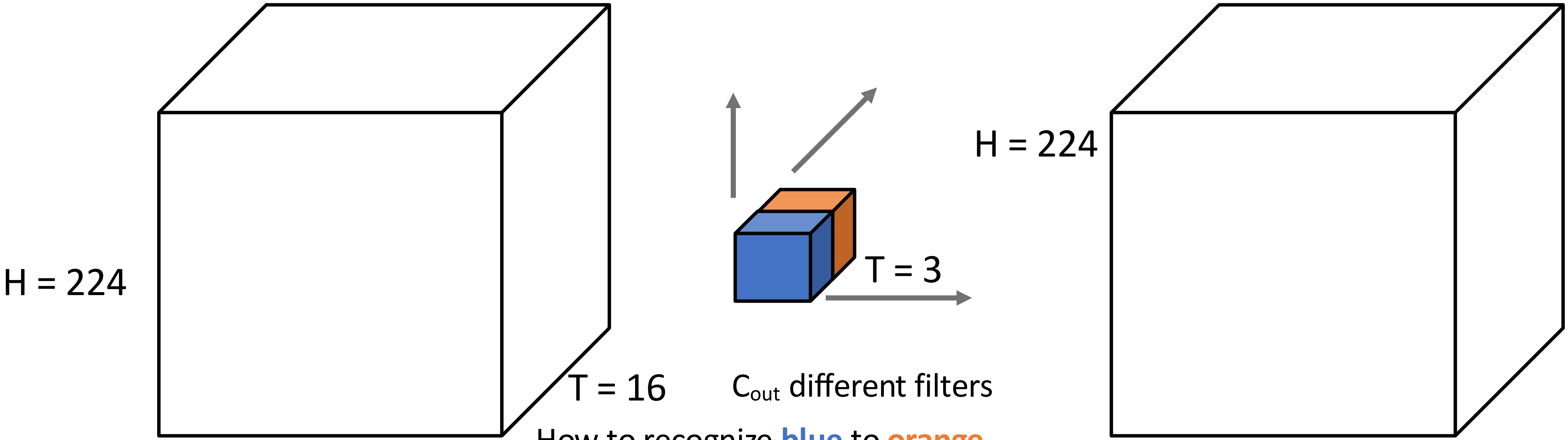
How to recognize **blue** to **orange**
transitions anywhere in space and time?

2D Conv (Early Fusion) vs 3D Conv (3D CNN)

Input: C_{in} x T x H x W
(3D grid with C_{in}-dim feat at each point)

Weight:
C_{out} x C_{in} x 3 x 3 x 3
Slide over x and y

Output:
C_{out} x T x H x W
3D grid with C_{out}–dim feat at each point



C_{out} different filters

How to recognize **blue** to **orange** transitions anywhere in space and time?

Example Video Dataset: Sports-1M

1 million YouTube videos annotated with labels for 487 different types of sports

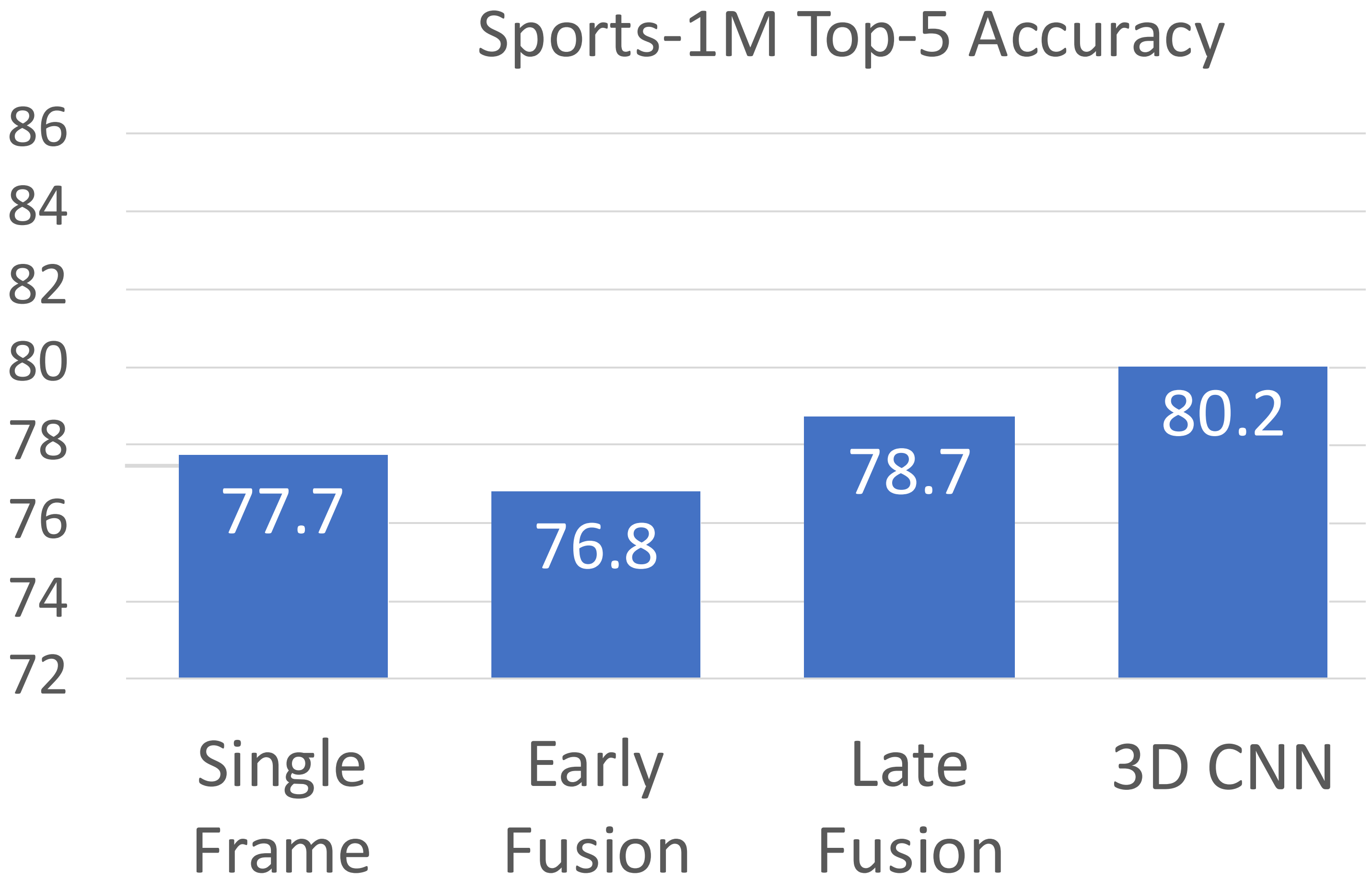
Ground Truth

Correct prediction

Incorrect prediction

Karpathy et al, "Large-scale Video Classification with Convolutional Neural Networks", CVPR 2014

Early Fusion vs Late Fusion vs 3D CNN



Single Frame model works well – always try this first!

3D CNNs have improved a lot since 2014!

Karpathy et al, "Large-scale Video Classification with Convolutional Neural Networks", CVPR 2014

C3D: The VGG of 3D CNNs

3D CNN that uses all 3x3x3 conv and 2x2x2 pooling
(except Pool1 which is 1x2x2)

Released model pretrained on Sports-1M: Many people used this as a video feature extractor

Layer	Size
Input	3 x 16 x 112 x 112
Conv1 (3x3x3)	64 x 16 x 112 x 112
Pool1 (1x2x2)	64 x 16 x 56 x 56
Conv2 (3x3x3)	128 x 16 x 56 x 56
Pool2 (2x2x2)	128 x 8 x 28 x 28
Conv3a (3x3x3)	256 x 8 x 28 x 28
Conv3b (3x3x3)	256 x 8 x 28 x 28
Pool3 (2x2x2)	256 x 4 x 14 x 14
Conv4a (3x3x3)	512 x 4 x 14 x 14
Conv4b (3x3x3)	512 x 4 x 14 x 14
Pool4 (2x2x2)	512 x 2 x 7 x 7
Conv5a (3x3x3)	512 x 2 x 7 x 7
Conv5b (3x3x3)	512 x 2 x 7 x 7
Pool5	512 x 1 x 3 x 3
FC6	4096
FC7	4096
FC8	C

Tran et al, "Learning Spatiotemporal Features with 3D Convolutional Networks", ICCV 2015

C3D: The VGG of 3D CNNs

3D CNN that uses all 3x3x3 conv and 2x2x2 pooling
(except Pool1 which is 1x2x2)

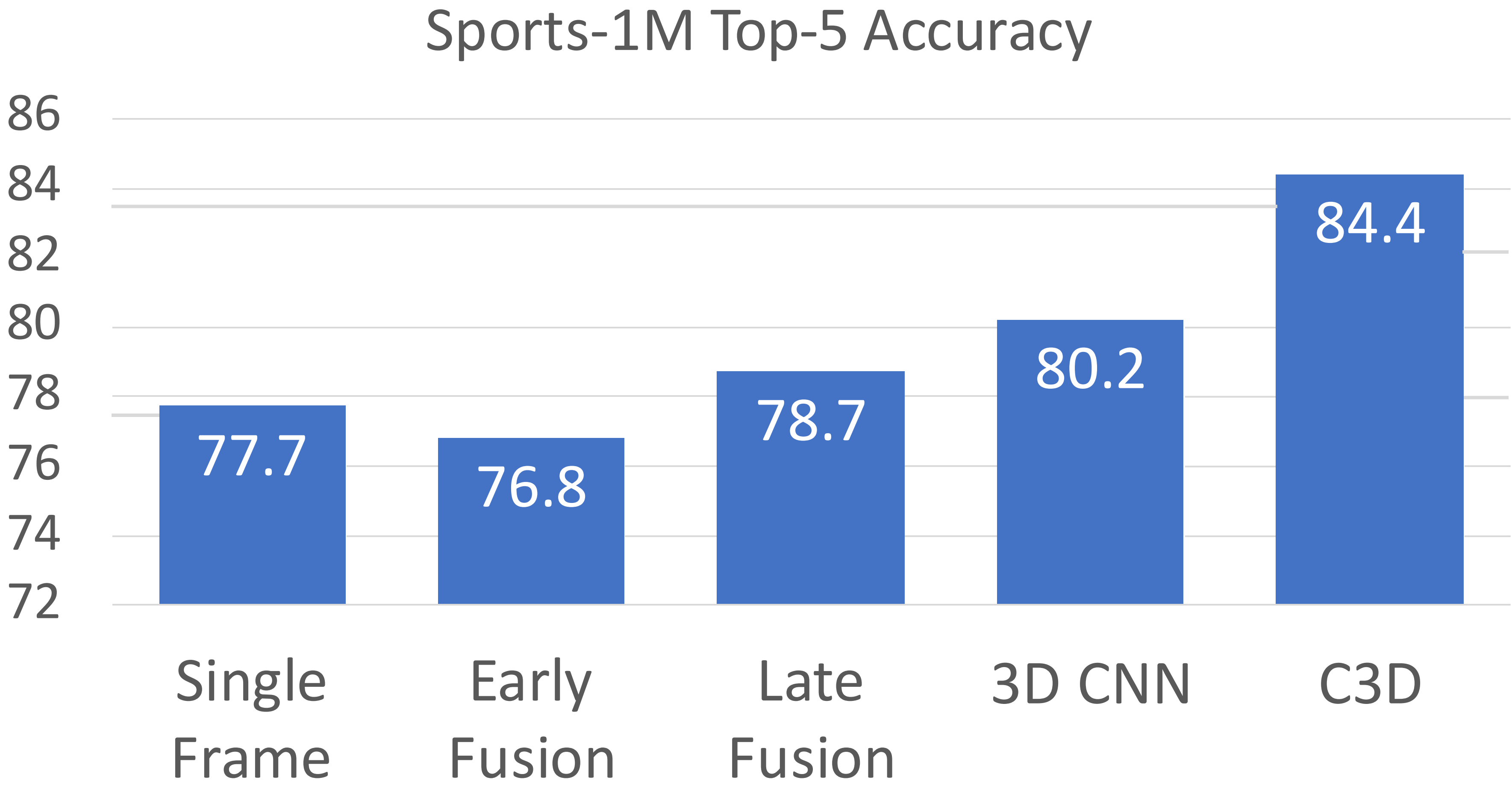
Released model pretrained on Sports-1M: Many people used this as a video feature extractor

Problem: 3x3x3 conv is very expensive!
AlexNet: 0.7 GFLOP
VGG-16: 13.6 GFLOP
C3D: **39.5 GFLOP (2.9x VGG!)**

Layer	Size	MFLOPs
Input	3 x 16 x 112 x 112	
Conv1 (3x3x3)	64 x 16 x 112 x 112	1.04
Pool1 (1x2x2)	64 x 16 x 56 x 56	
Conv2 (3x3x3)	128 x 16 x 56 x 56	11.10
Pool2 (2x2x2)	128 x 8 x 28 x 28	
Conv3a (3x3x3)	256 x 8 x 28 x 28	5.55
Conv3b (3x3x3)	256 x 8 x 28 x 28	11.10
Pool3 (2x2x2)	256 x 4 x 14 x 14	
Conv4a (3x3x3)	512 x 4 x 14 x 14	2.77
Conv4b (3x3x3)	512 x 4 x 14 x 14	5.55
Pool4 (2x2x2)	512 x 2 x 7 x 7	
Conv5a (3x3x3)	512 x 2 x 7 x 7	0.69
Conv5b (3x3x3)	512 x 2 x 7 x 7	0.69
Pool5	512 x 1 x 3 x 3	
FC6	4096	0.51
FC7	4096	0.45
FC8	C	0.05

Tran et al, "Learning Spatiotemporal Features with 3D Convolutional Networks", ICCV 2015

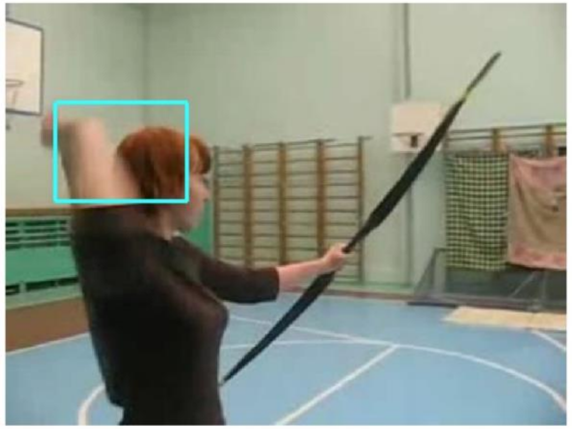
Early Fusion vs Late Fusion vs 3D CNN



Karpathy et al, "Large-scale Video Classification with Convolutional Neural Networks", CVPR 2014
Tran et al, "Learning Spatiotemporal Features with 3D Convolutional Networks", ICCV 2015

Measuring Motion: Optical Flow

Image at frame t



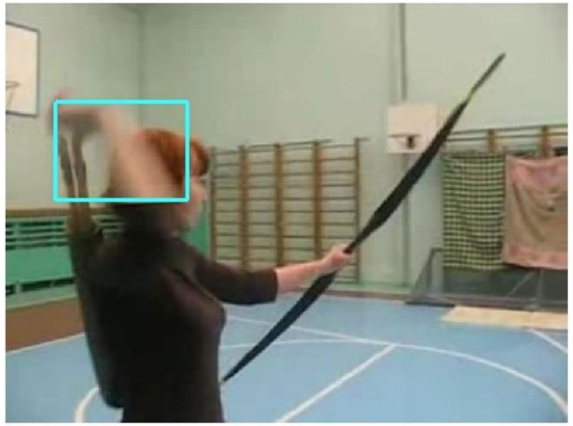
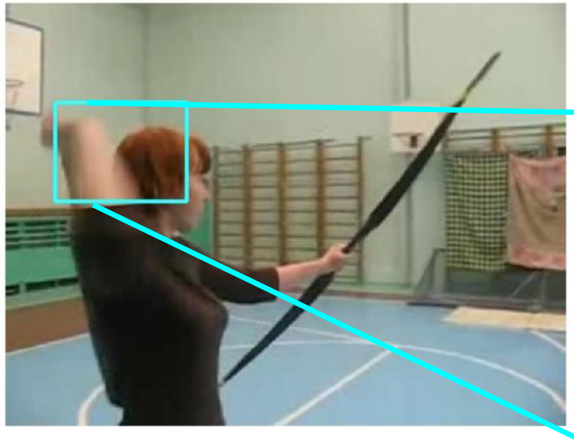


Image at frame t+1

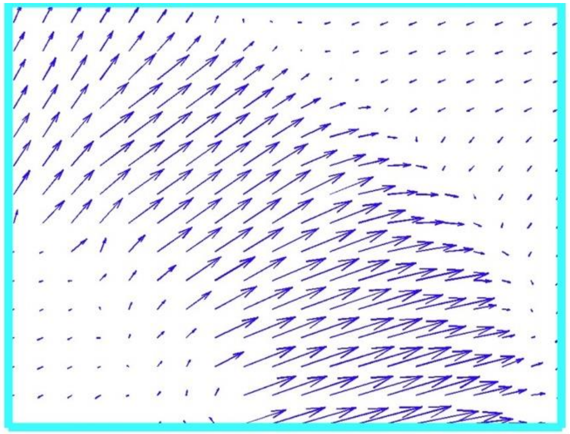
Simonyan and Zisserman, "Two-stream convolutional networks for action recognition in videos", NeurIPS 2014

Measuring Motion: Optical Flow

Image at frame t



Optical flow gives a displacement field F between images I_t and I_{t+1}



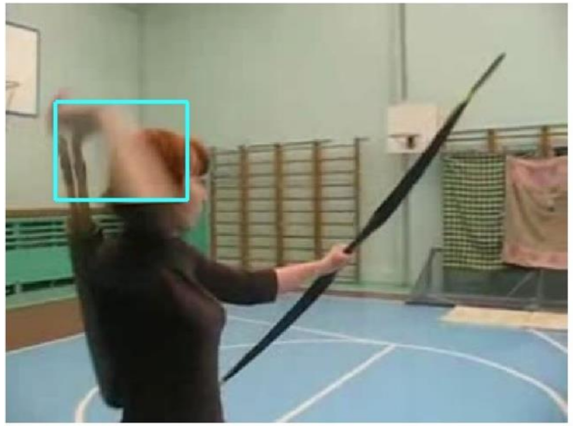


Image at frame t+1

Tells where each pixel will move in the next frame:

$F(x, y) = (dx, dy)$

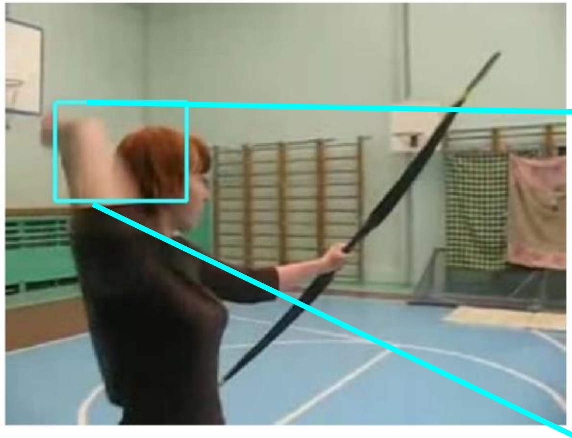
$I_{t+1}(x+dx, y+dy) = I_t(x, y)$

Simonyan and Zisserman, "Two-stream convolutional networks for action recognition in videos", NeurIPS 2014

Measuring Motion: Optical Flow

Optical Flow highlights **local motion**

Image at frame t



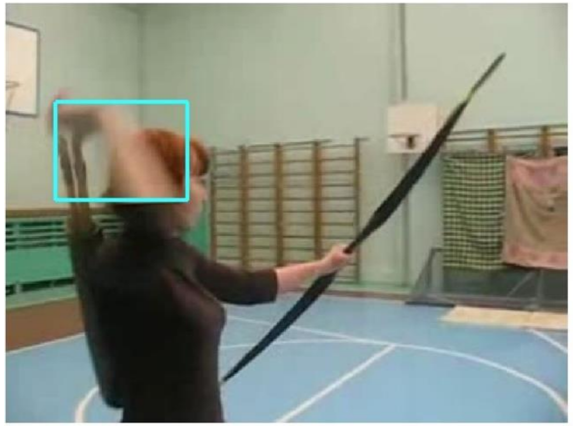
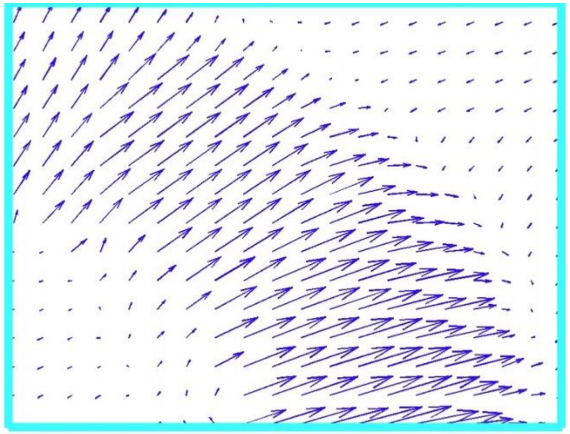


Image at frame t+1

Optical flow gives a displacement field F between images I_t and I_{t+1}

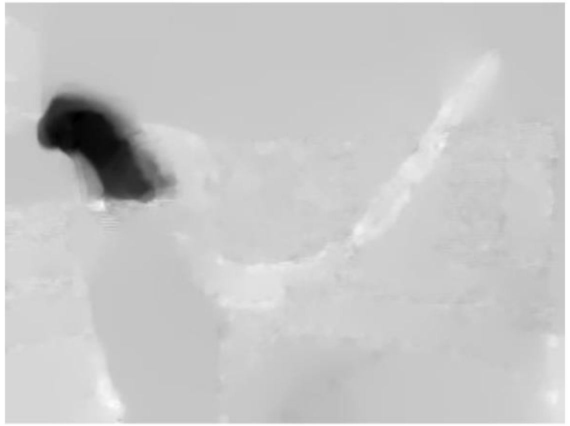


Tells where each pixel will move in the next frame:

$F(x, y) = (dx, dy)$

$I_{t+1}(x+dx, y+dy) = I_t(x, y)$

Horizontal flow dx

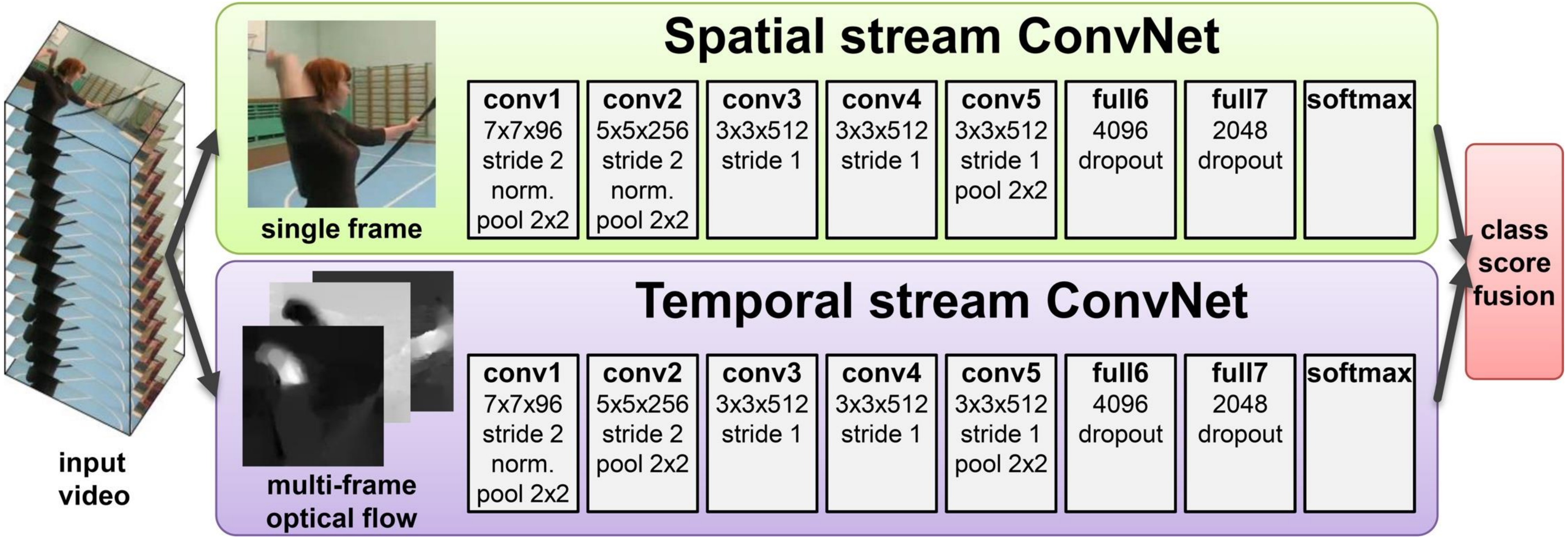


Vertical Flow dy

Simonyan and Zisserman, "Two-stream convolutional networks for action recognition in videos", NeurIPS 2014

Separating Motion and Appearance: Two-Stream Networks

Input: Single Image
3 x H x W

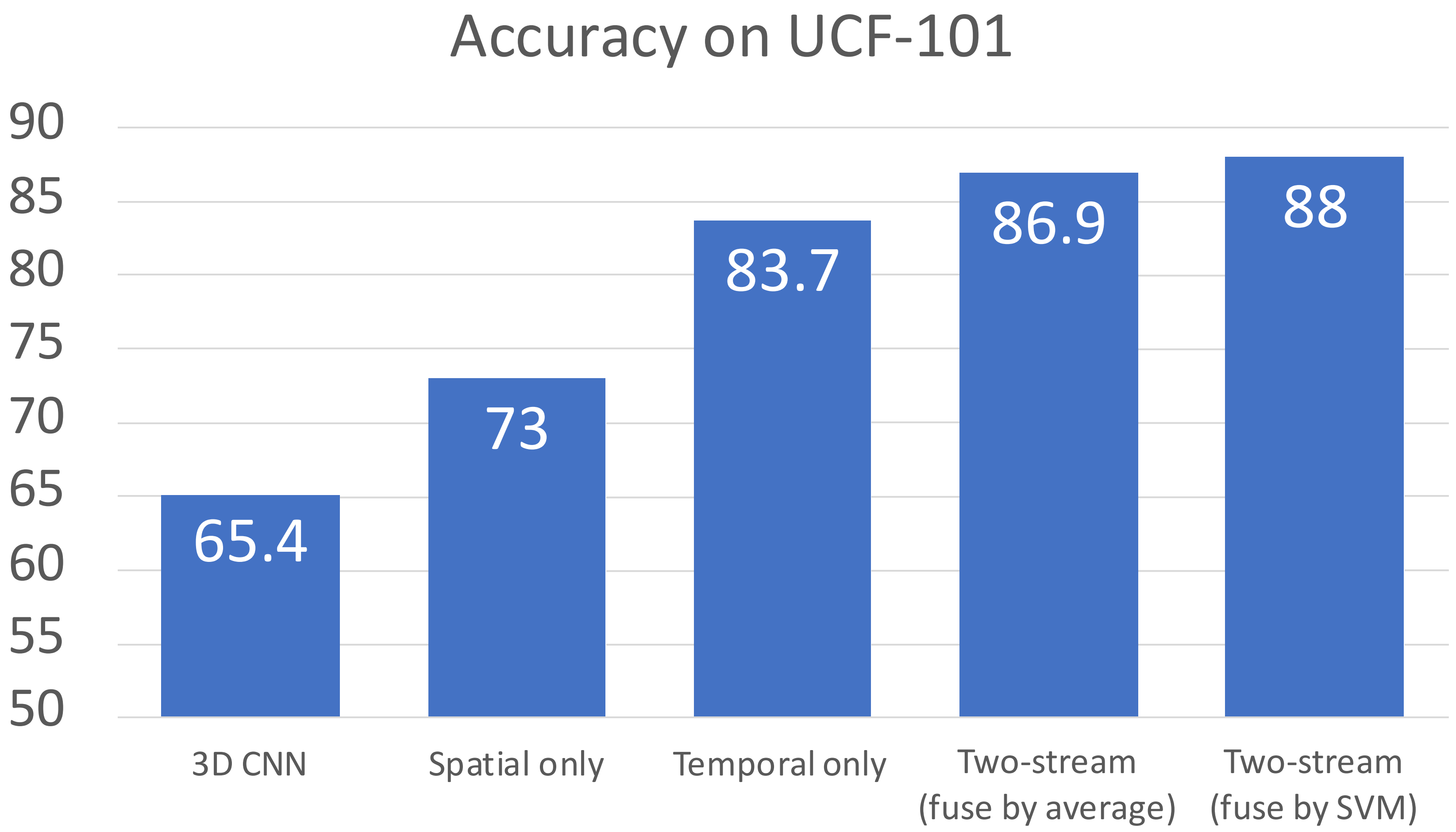


Input: Stack of optical flow:
[2*(T-1)] x H x W

Early fusion: First 2D conv processes all flow images

Simonyan and Zisserman, "Two-stream convolutional networks for action recognition in videos", NeurIPS 2014

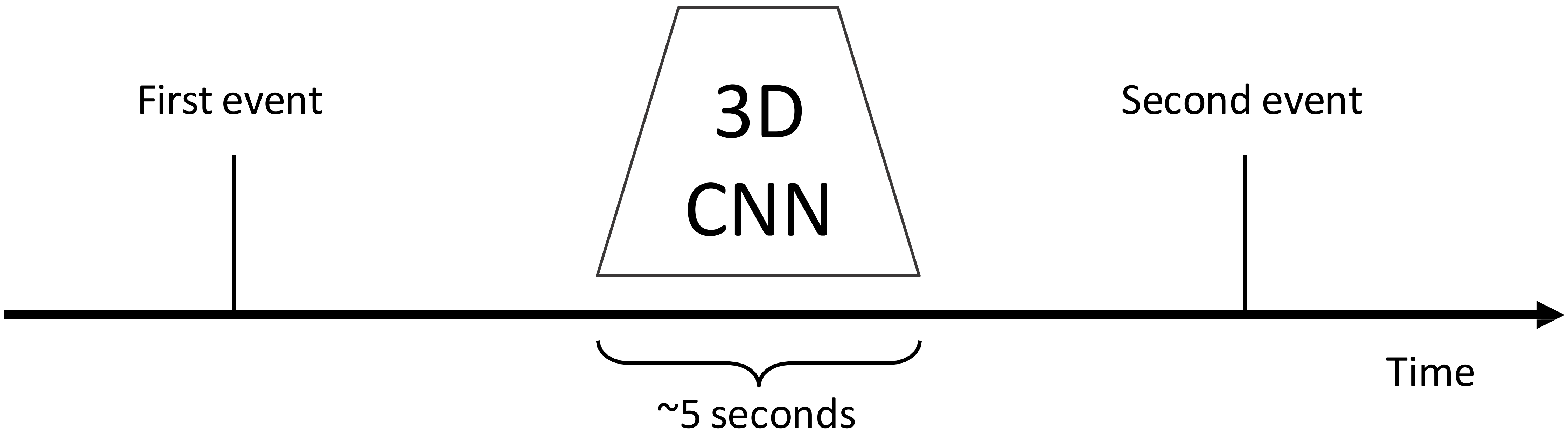
Separating Motion and Appearance: Two-Stream Networks



Simonyan and Zisserman, “Two-stream convolutional networks for action recognition in videos”, NeurIPS 2014

Modeling long-term temporal structure

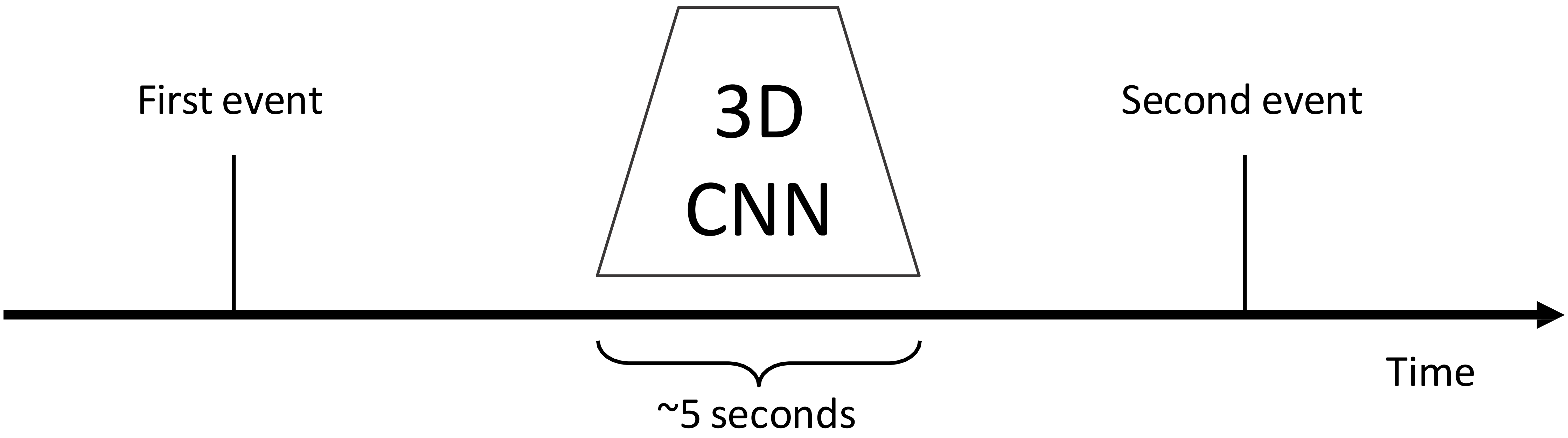
So far all our temporal CNNs only model local motion between frames in very short clips of ~2-5 seconds. What about long-term structure?



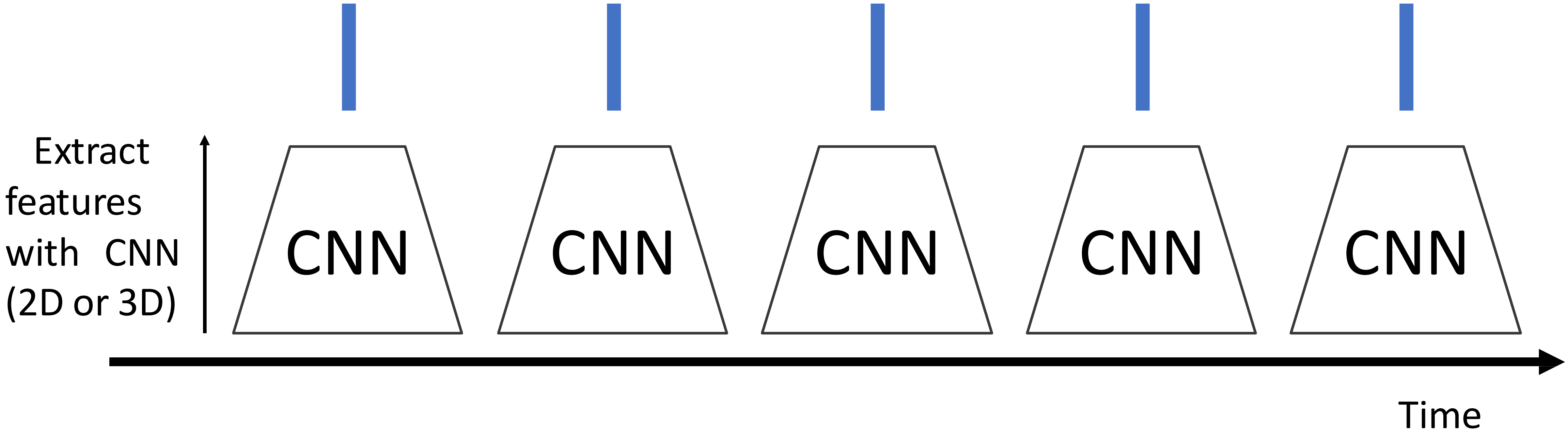
Modeling long-term temporal structure

So far all our temporal CNNs only model local motion between frames in very short clips of ~2-5 seconds. What about long-term structure?

We know how to handle sequences! How about recurrent networks?

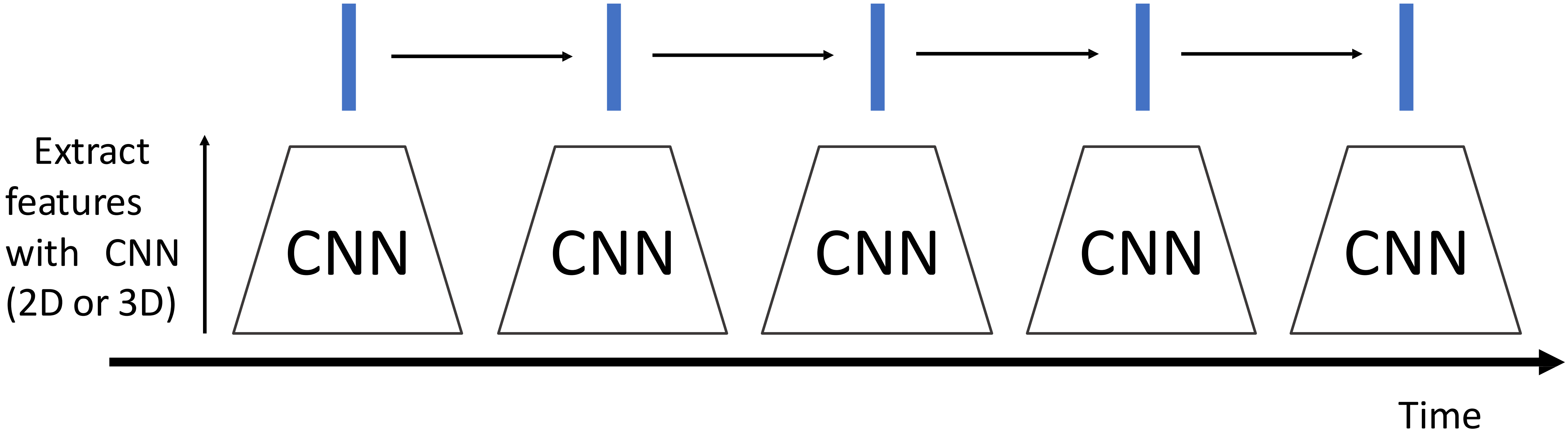


Modeling long-term temporal structure



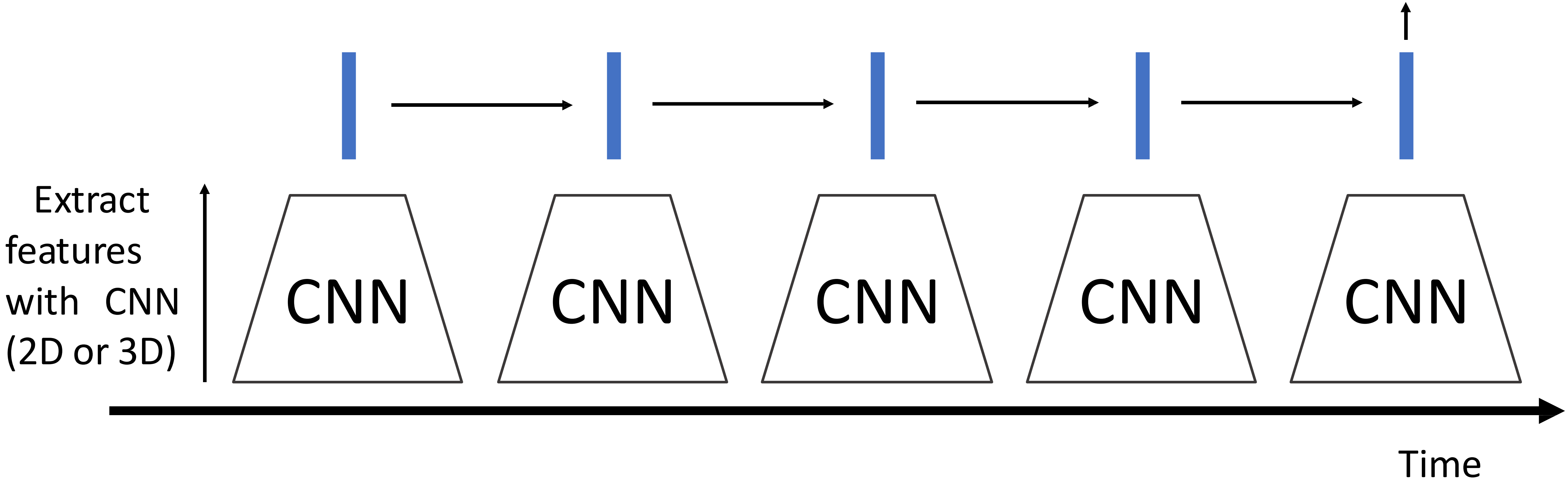
Modeling long-term temporal structure

Process local features using recurrent network (e.g. LSTM)



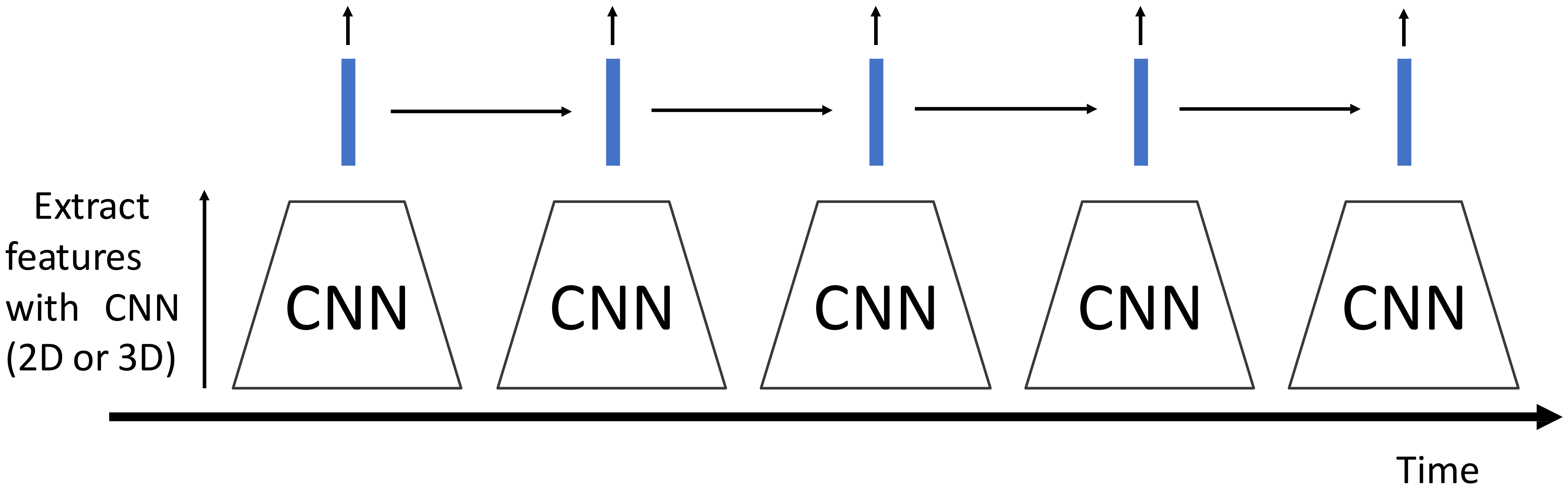
Modeling long-term temporal structure

Process local features using recurrent network (e.g. LSTM)
Many to one: One output at end of video



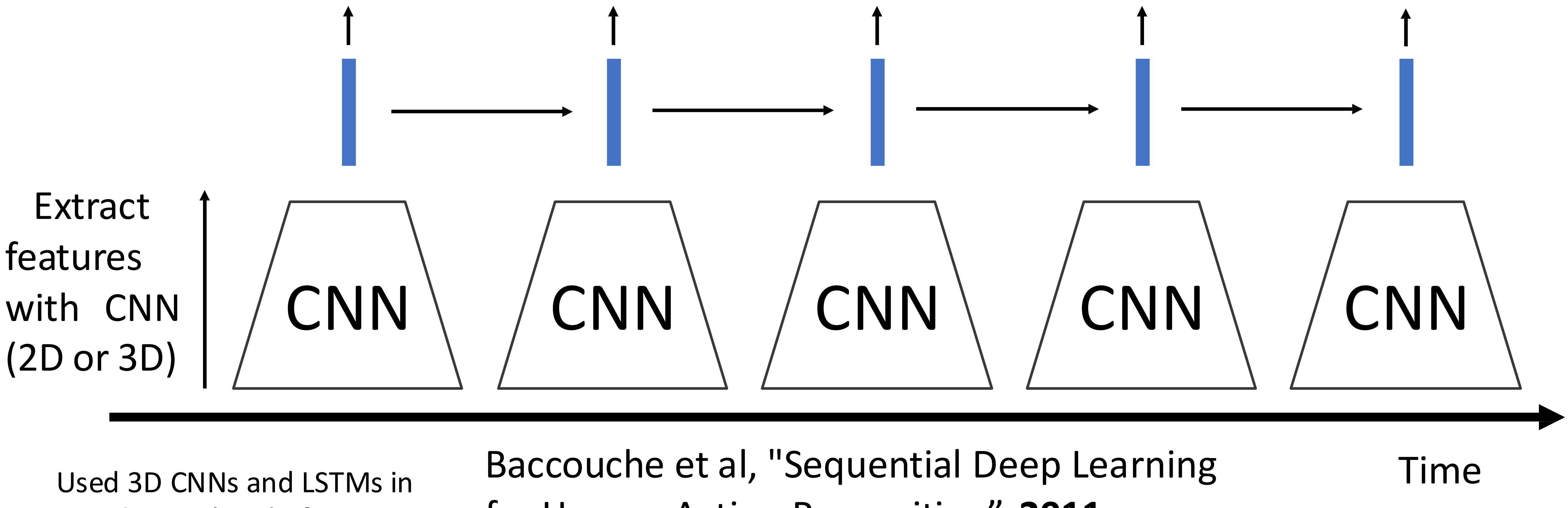
Modeling long-term temporal structure

Process local features using recurrent network (e.g. LSTM)
Many to many: one output per video frame



Modeling long-term temporal structure

Process local features using recurrent network (e.g. LSTM)
Many to many: one output per video frame

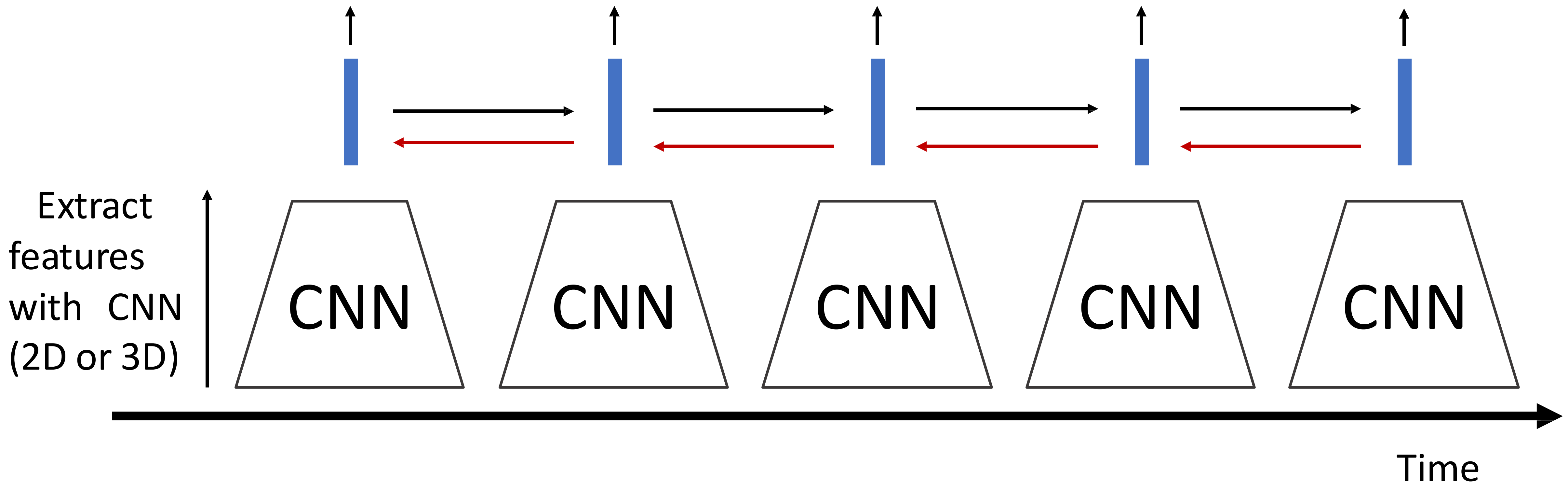


Used 3D CNNs and LSTMs in 2011! Way ahead of its time

Baccouche et al, "Sequential Deep Learning for Human Action Recognition", **2011**

Modeling long-term temporal structure

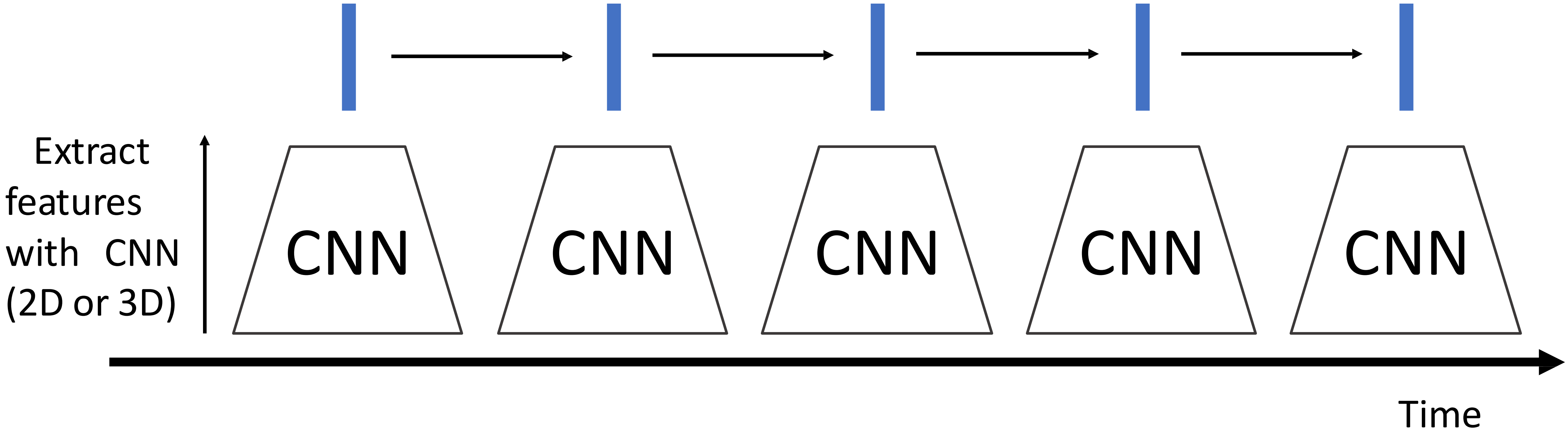
Sometimes don't backprop to CNN to save memory;
pretrain and use it as a feature extractor



Baccouche et al, "Sequential Deep Learning for Human Action Recognition", 2011
Donahue et al, "Long-term recurrent convolutional networks for visual recognition and description", CVPR 2015

Modeling long-term temporal structure

Inside CNN: Each value a function of a fixed temporal window (local temporal structure)
Inside RNN: Each vector is a function of all previous vectors (global temporal structure)
Can we merge both approaches?



Baccouche et al, "Sequential Deep Learning for Human Action Recognition", 2011
Donahue et al, "Long-term recurrent convolutional networks for visual recognition and description", CVPR 2015

Recall: Self-Attention

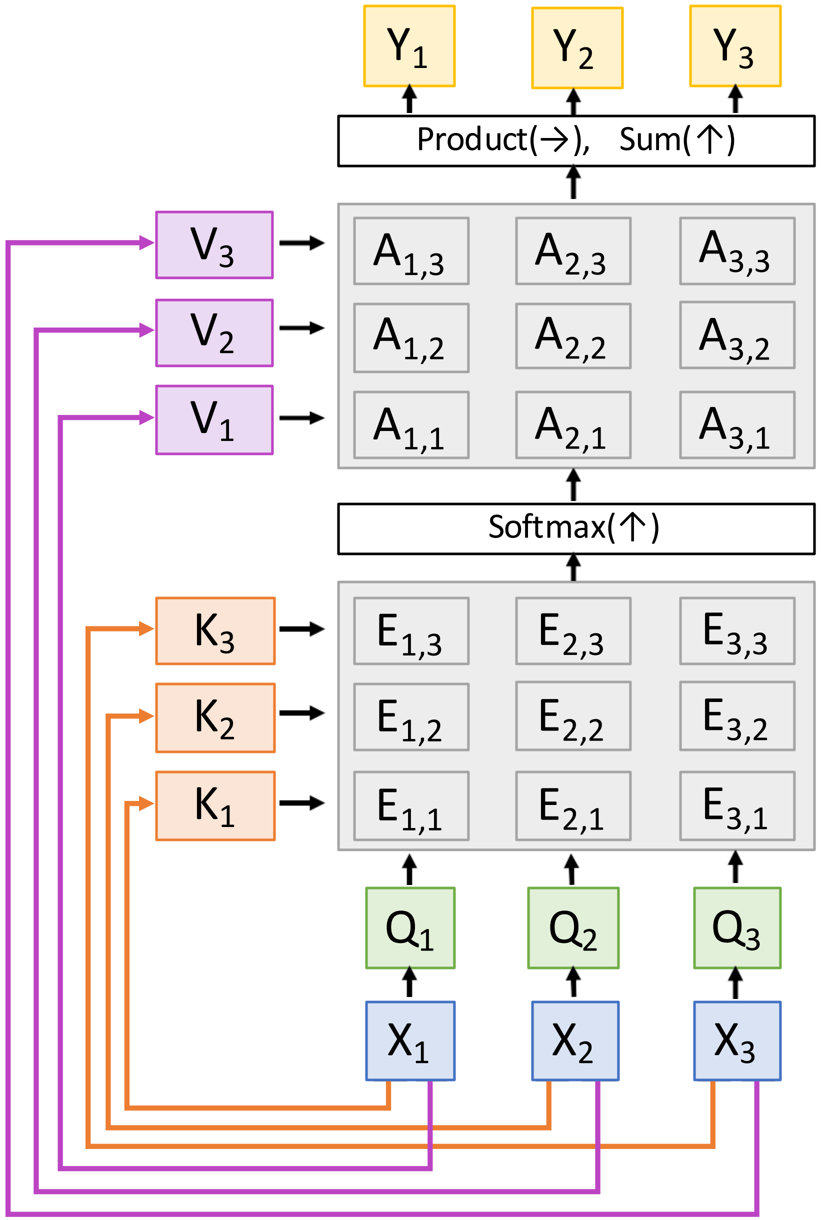
Input: Set of vectors $x_1, \ldots, x_N$

Keys, Queries, Values: Project each x to a key, query, and value using linear layer

Affinity matrix: Compare each pair of x, (using scaled dot-product between keys and values) and normalize using softmax

Output: Weighted sum of values, with weights given by affinity matrix

Features in 3D CNN: C x T x H x W
Interpret as a set of THW vectors of dim C

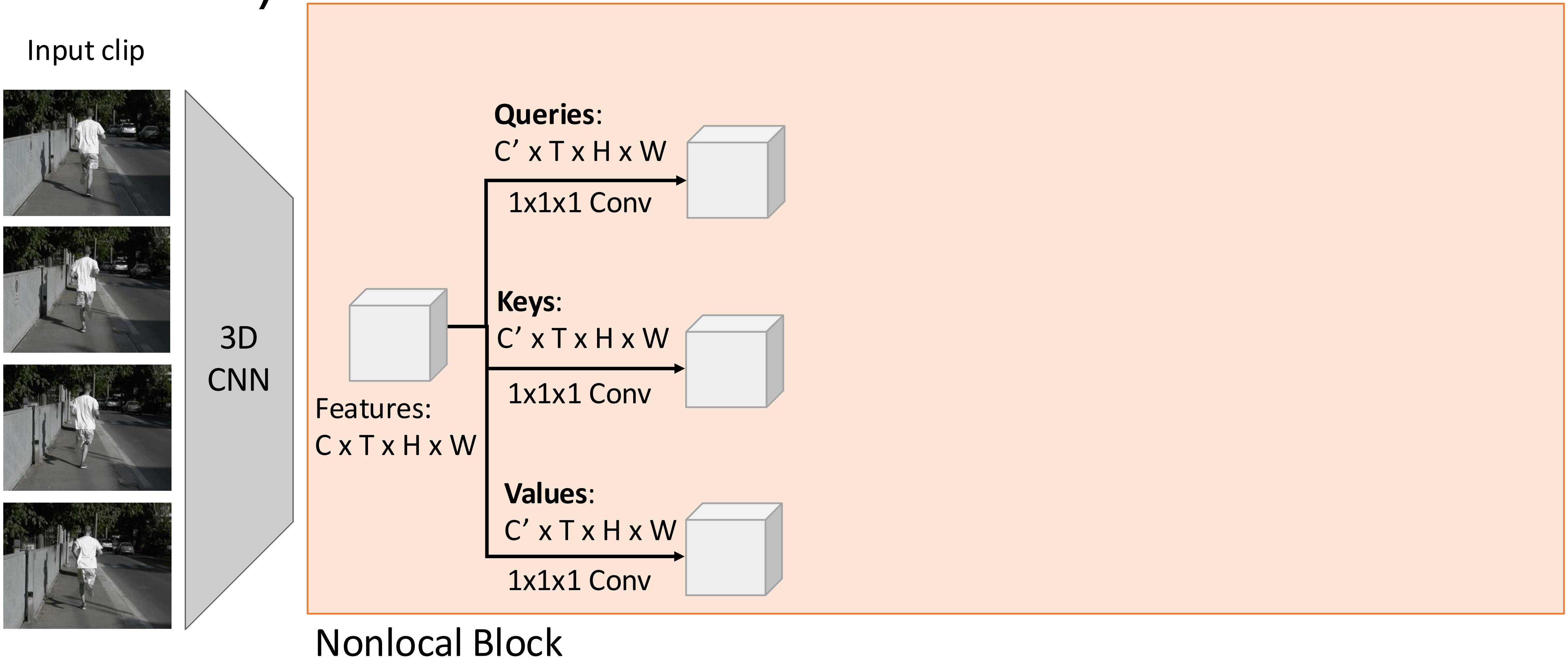


Vaswani et al, "Attention is all you need", NeurIPS 2017

Spatio-Temporal Self-Attention (Nonlocal Block)

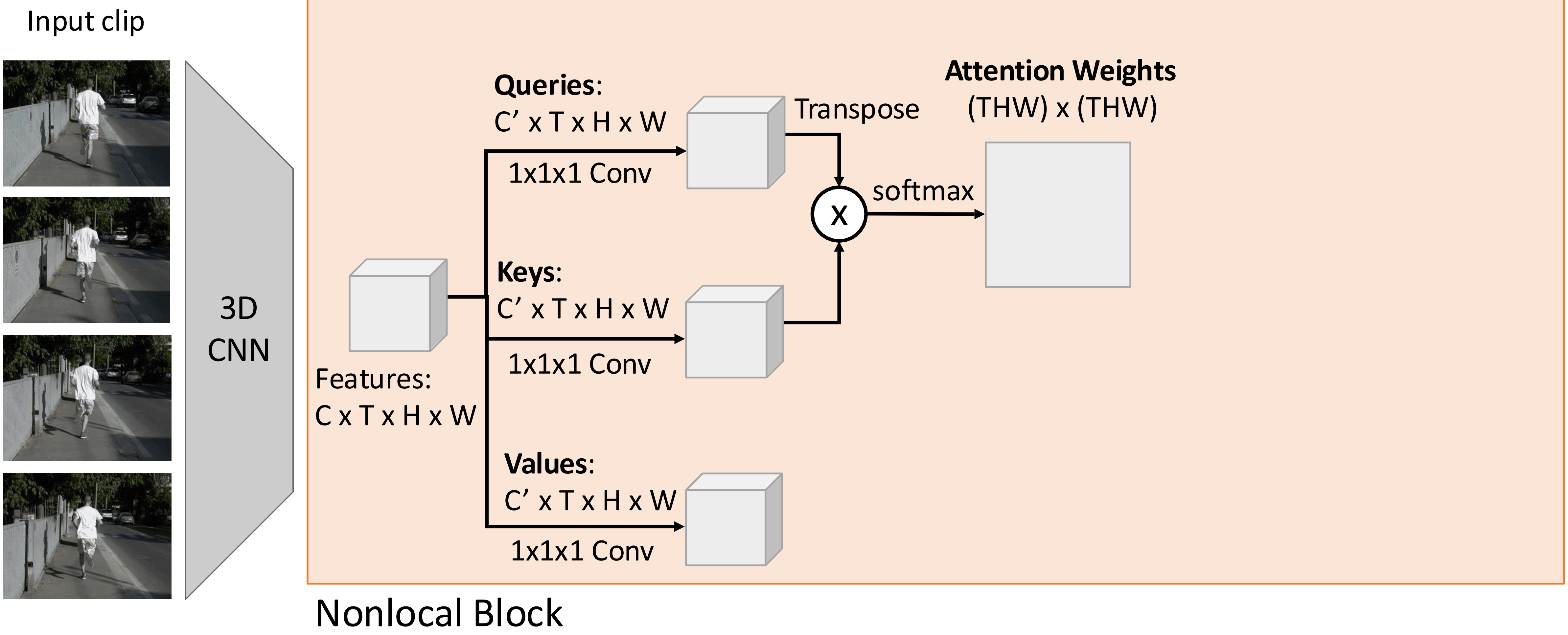
Wang et al, “Non-local neural networks”, CVPR 2018

Spatio-Temporal Self-Attention (Nonlocal Block)



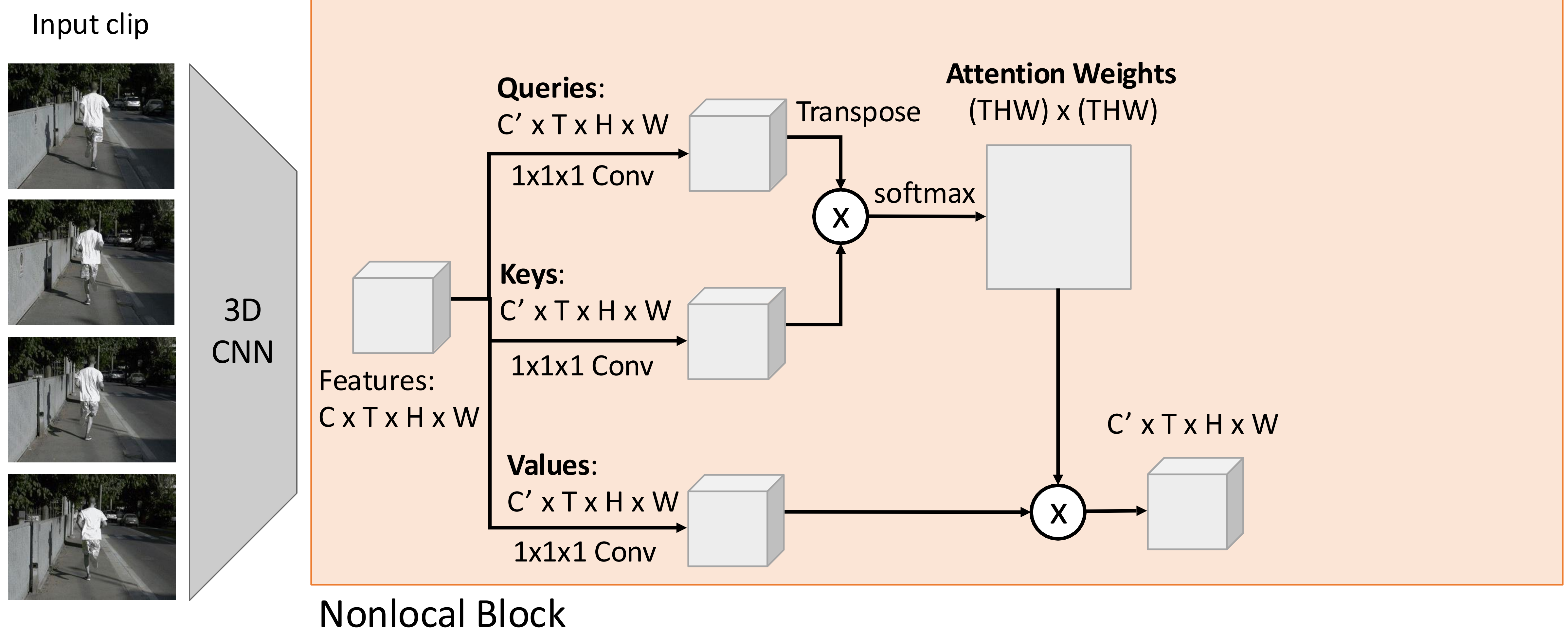
Wang et al, "Non-local neural networks", CVPR 2018

Spatio-Temporal Self-Attention (Nonlocal Block)



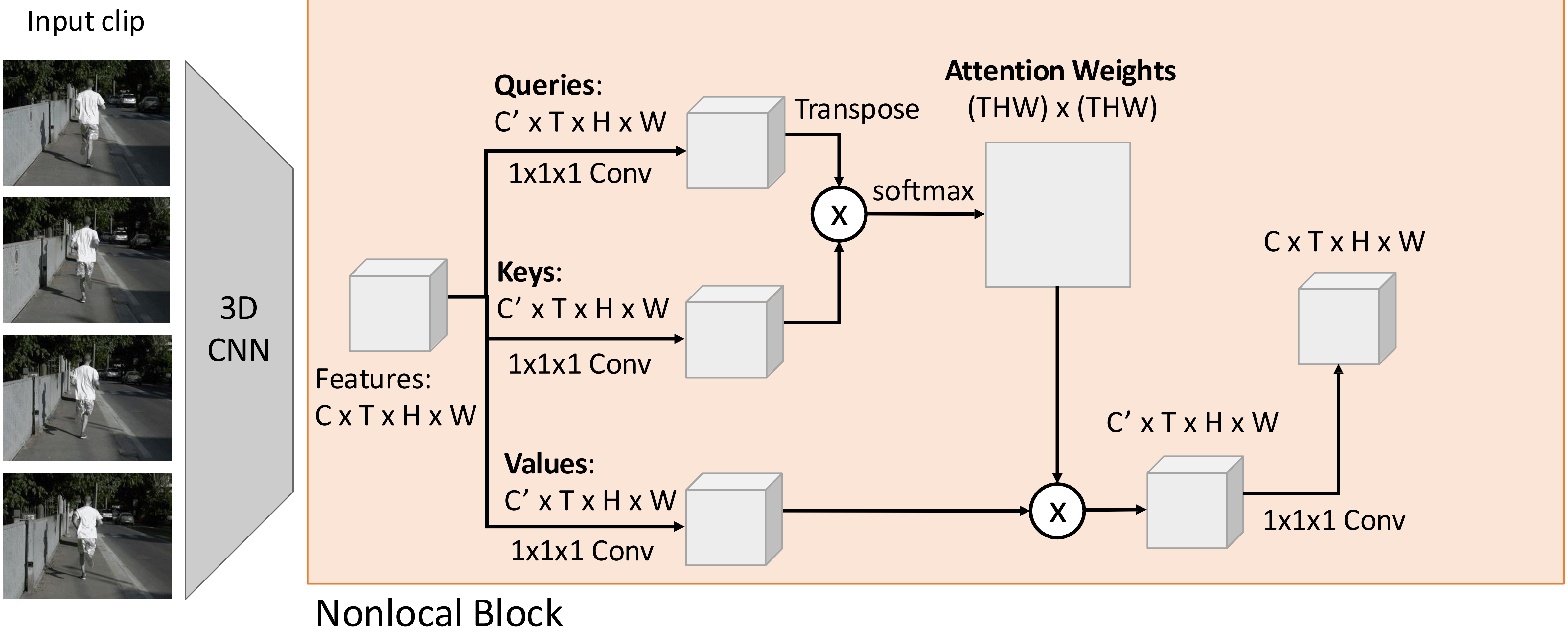
Wang et al, "Non-local neural networks", CVPR 2018

Spatio-Temporal Self-Attention (Nonlocal Block)



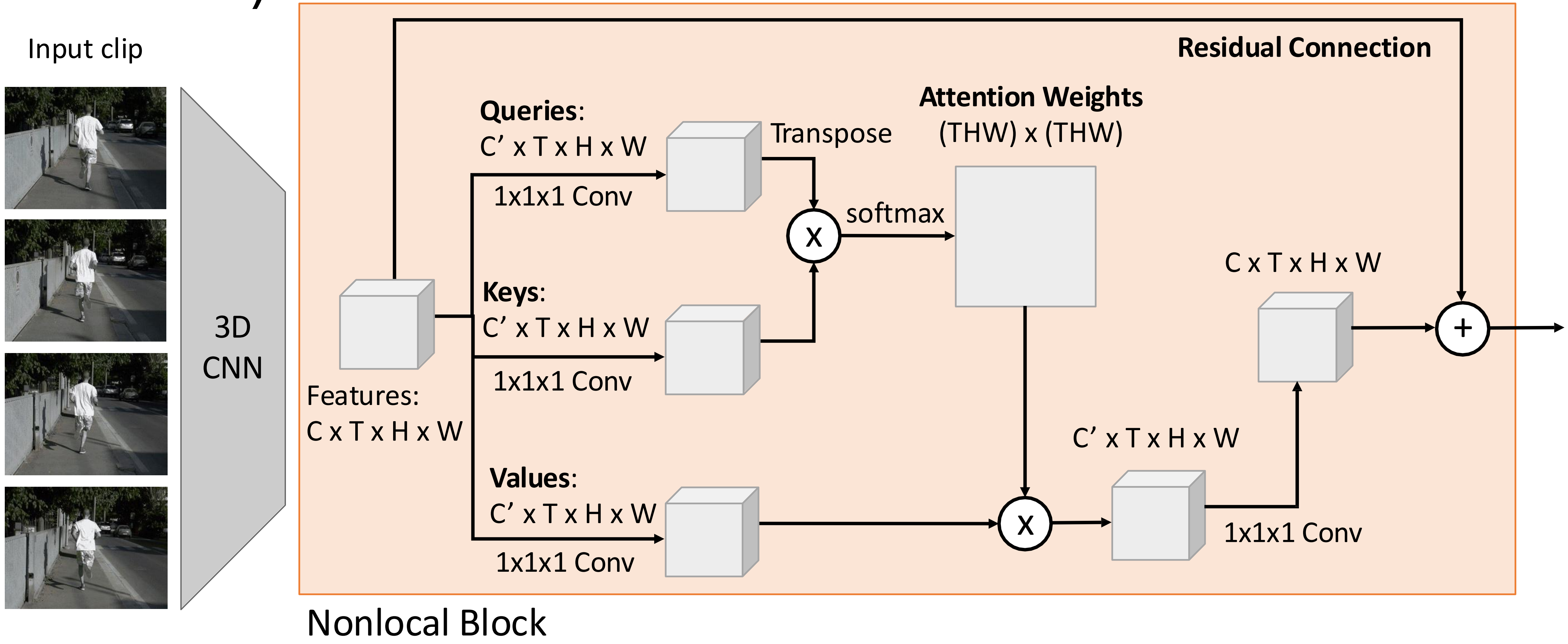
Wang et al, "Non-local neural networks", CVPR 2018

Spatio-Temporal Self-Attention (Nonlocal Block)



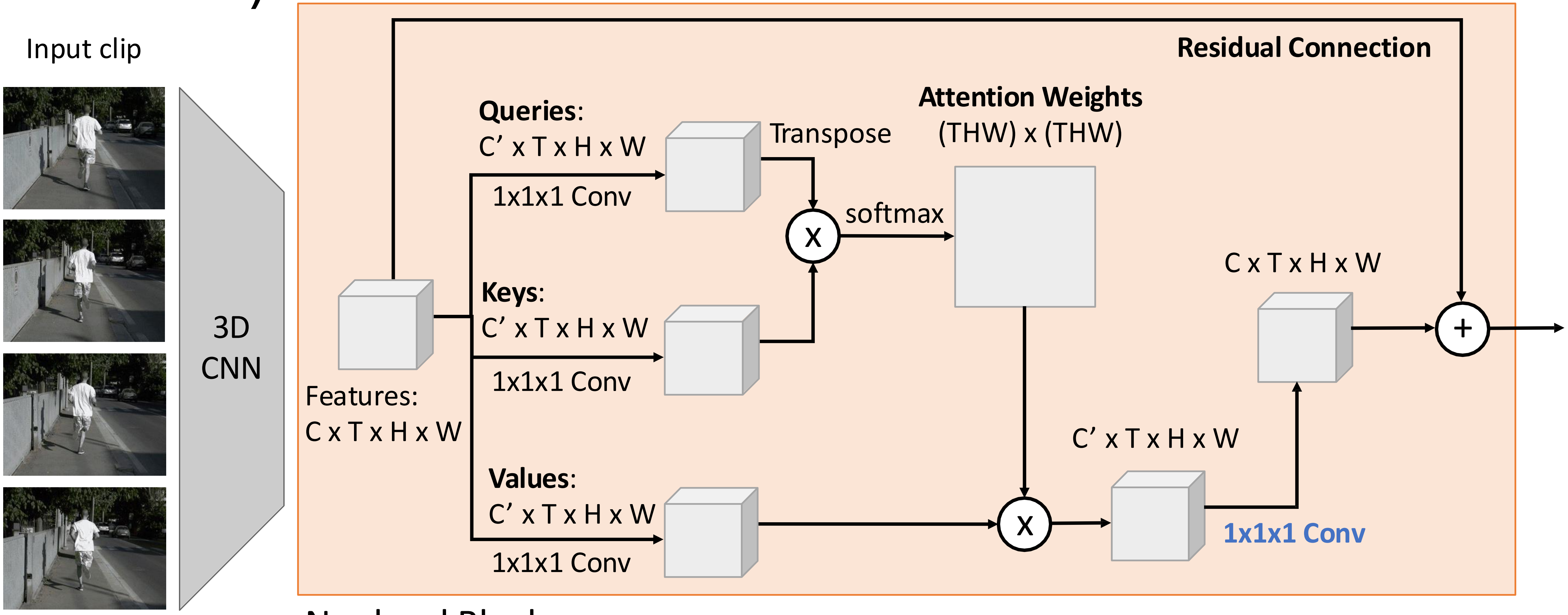
Wang et al, "Non-local neural networks", CVPR 2018

Spatio-Temporal Self-Attention (Nonlocal Block)



Wang et al, "Non-local neural networks", CVPR 2018

Spatio-Temporal Self-Attention (Nonlocal Block)



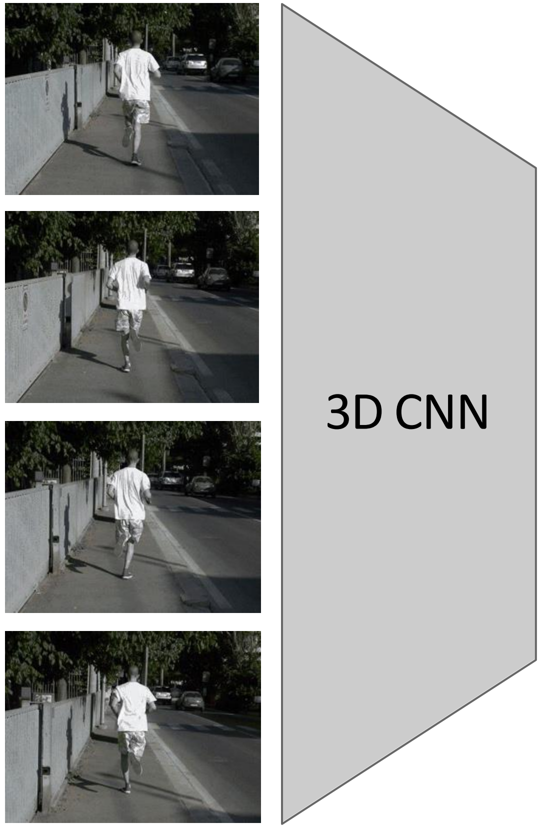
Nonlocal Block

Trick: Initialize last conv to 0, then entire block computes identity. Can insert into existing 3D CNNs

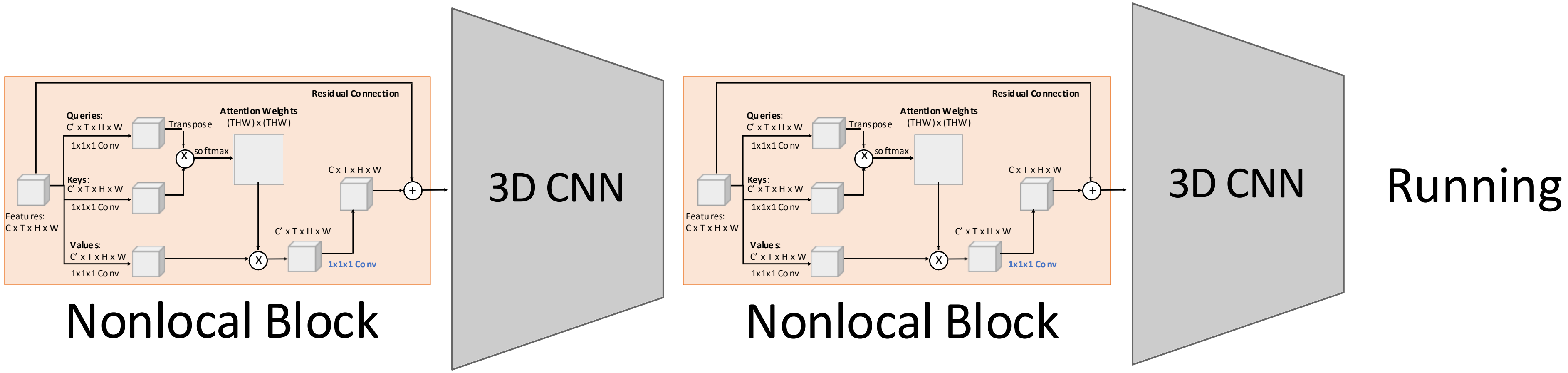
Wang et al, "Non-local neural networks", CVPR 2018

Spatio-Temporal Self-Attention (Nonlocal Block)

Input clip



We can add nonlocal blocks into existing 3D CNN architectures.
But what is the best 3D CNN architecture?



Wang et al, "Non-local neural networks", CVPR 2018

Inflating 2D Networks to 3D (I3D)

There has been a lot of work on architectures for images.
Can we reuse image architectures for video?

Idea: take a 2D CNN architecture.

Replace each 2D K_h x K_w conv/pool layer with a 3D K_t x K_h x K_w version

Carreira and Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", CVPR 2017

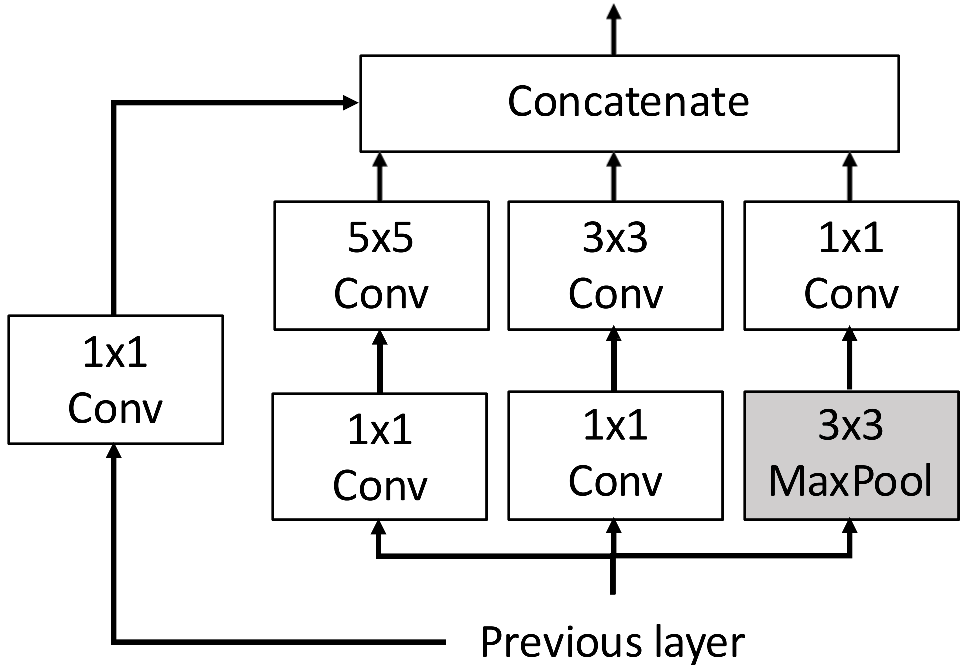
Inflating 2D Networks to 3D (I3D)

There has been a lot of work on architectures for images.
Can we reuse image architectures for video?

Idea: take a 2D CNN architecture.

Replace each 2D K_h x K_w conv/pool layer with a 3D K_t x K_h x K_w version

Inception Block: Original



Carreira and Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", CVPR 2017

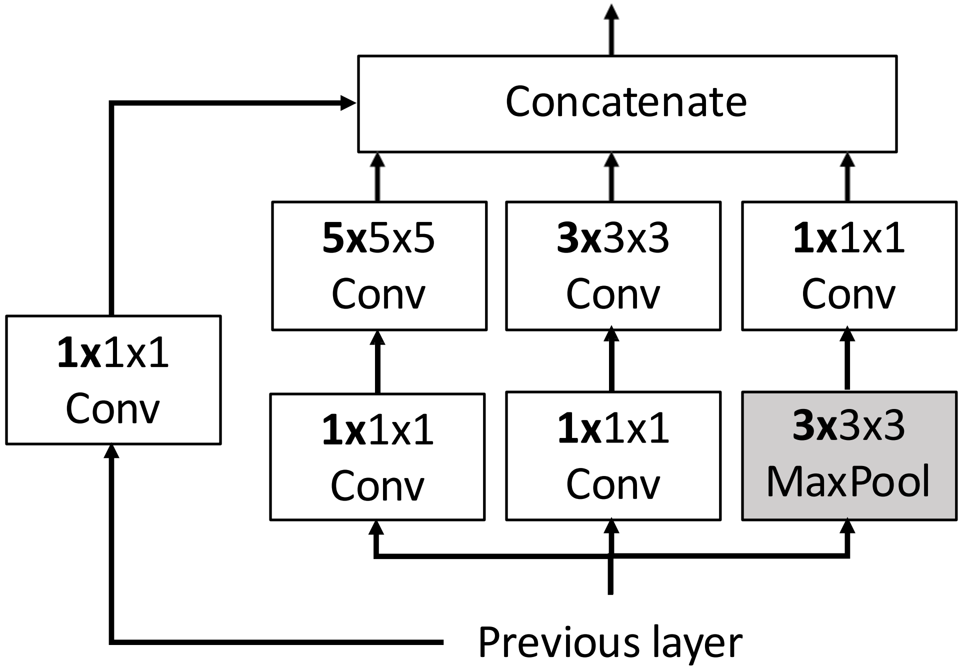
Inflating 2D Networks to 3D (I3D)

There has been a lot of work on architectures for images.
Can we reuse image architectures for video?

Idea: take a 2D CNN architecture.

Replace each 2D K_h x K_w conv/pool layer with a 3D K_t x K_h x K_w version

Inception Block: Inflated



Carreira and Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", CVPR 2017

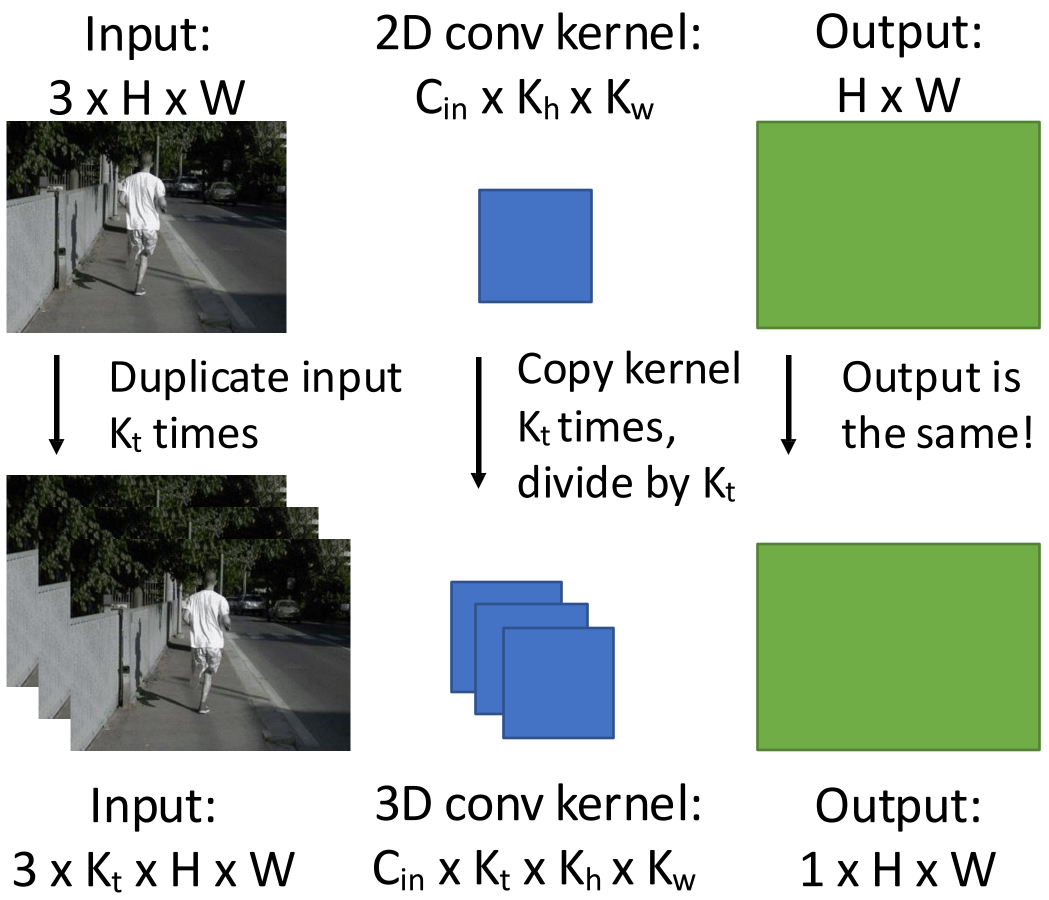
Inflating 2D Networks to 3D (I3D)

There has been a lot of work on architectures for images.
Can we reuse image architectures for video?

Idea: take a 2D CNN architecture.

Replace each 2D K_h x K_w conv/pool layer with a 3D K_t x K_h x K_w version

Can use weights of 2D conv to initialize 3D conv: copy K_t times in space and divide by K_t
This gives the same result as 2D conv given "constant" video input



Carreira and Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", CVPR 2017

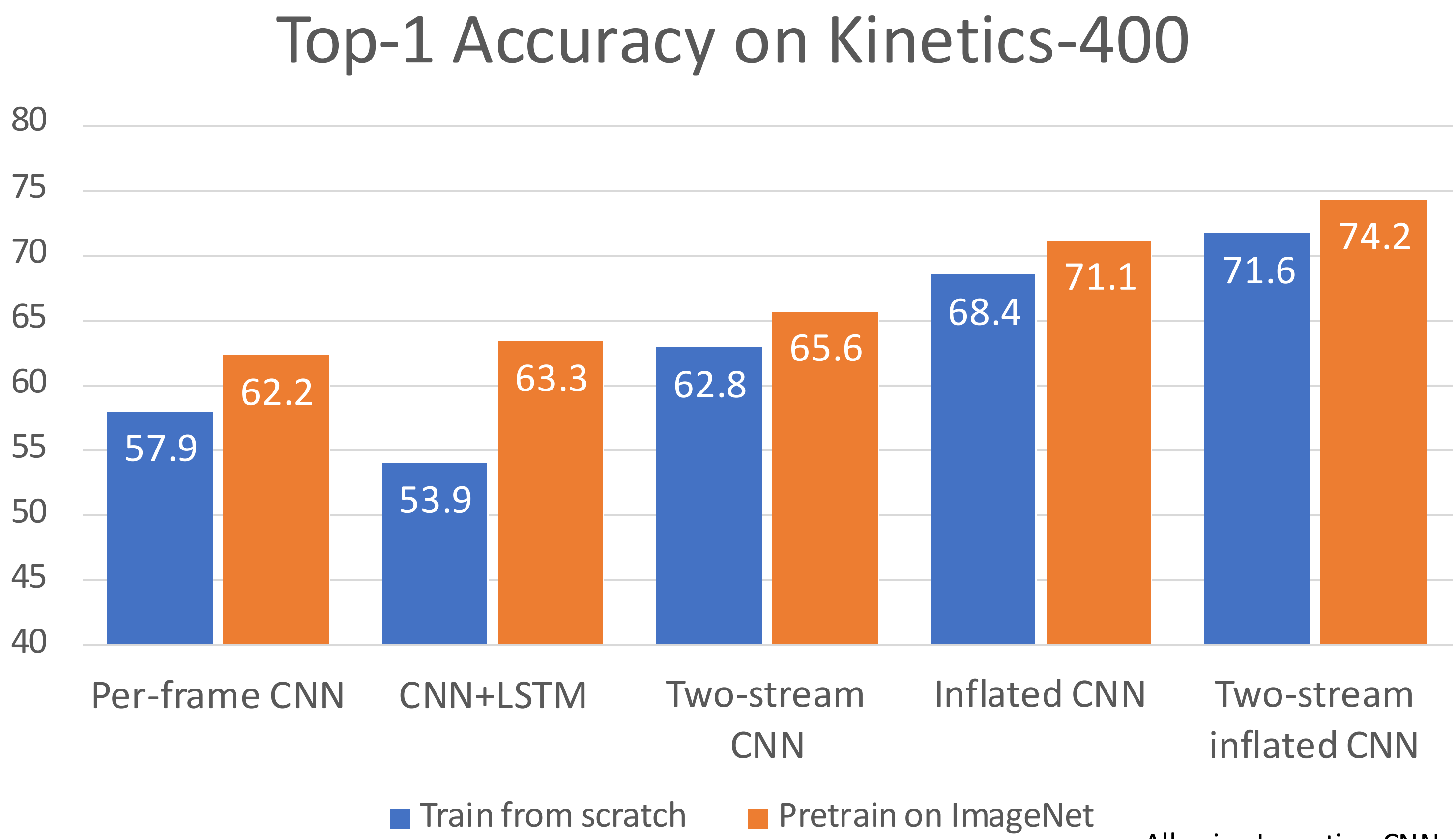
Inflating 2D Networks to 3D (I3D)

There has been a lot of work on architectures for images.
Can we reuse image architectures for video?

Idea: take a 2D CNN architecture.

Replace each 2D K_h x K_w conv/pool layer with a 3D K_t x K_h x K_w version

Can use weights of 2D conv to initialize 3D conv: copy K_t times in space and divide by K_t
This gives the same result as 2D conv given "constant" video input

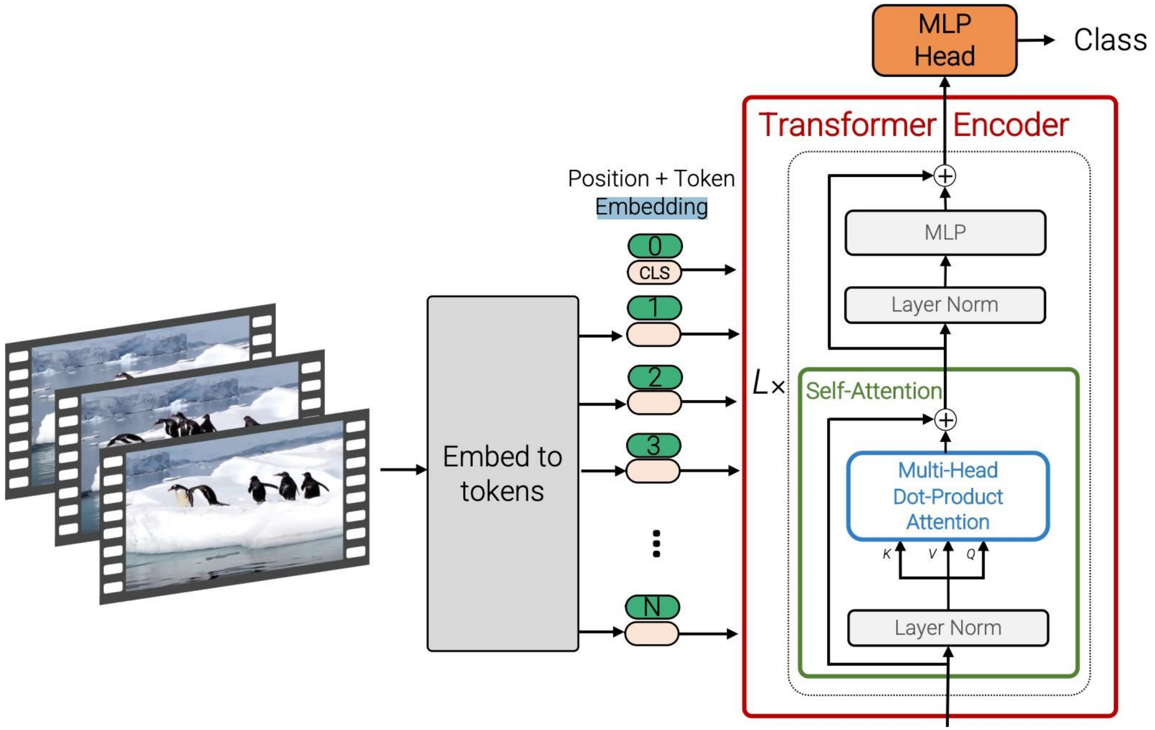


All using Inception CNN

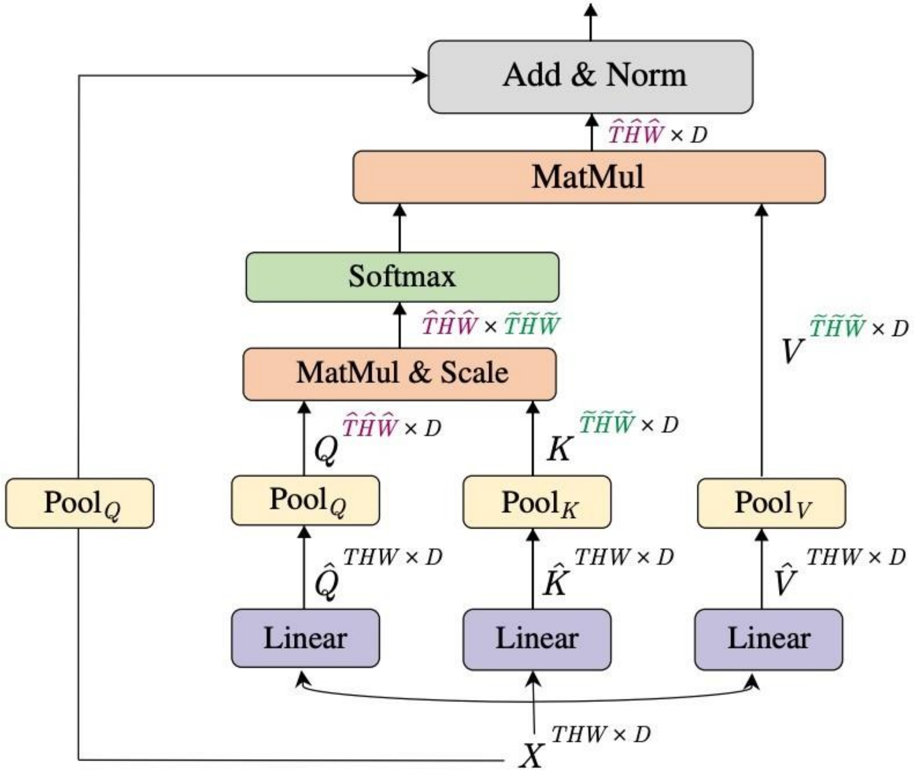
Carreira and Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", CVPR 2017

Vision Transformers for Video

Factorized attention: Attend over space / time



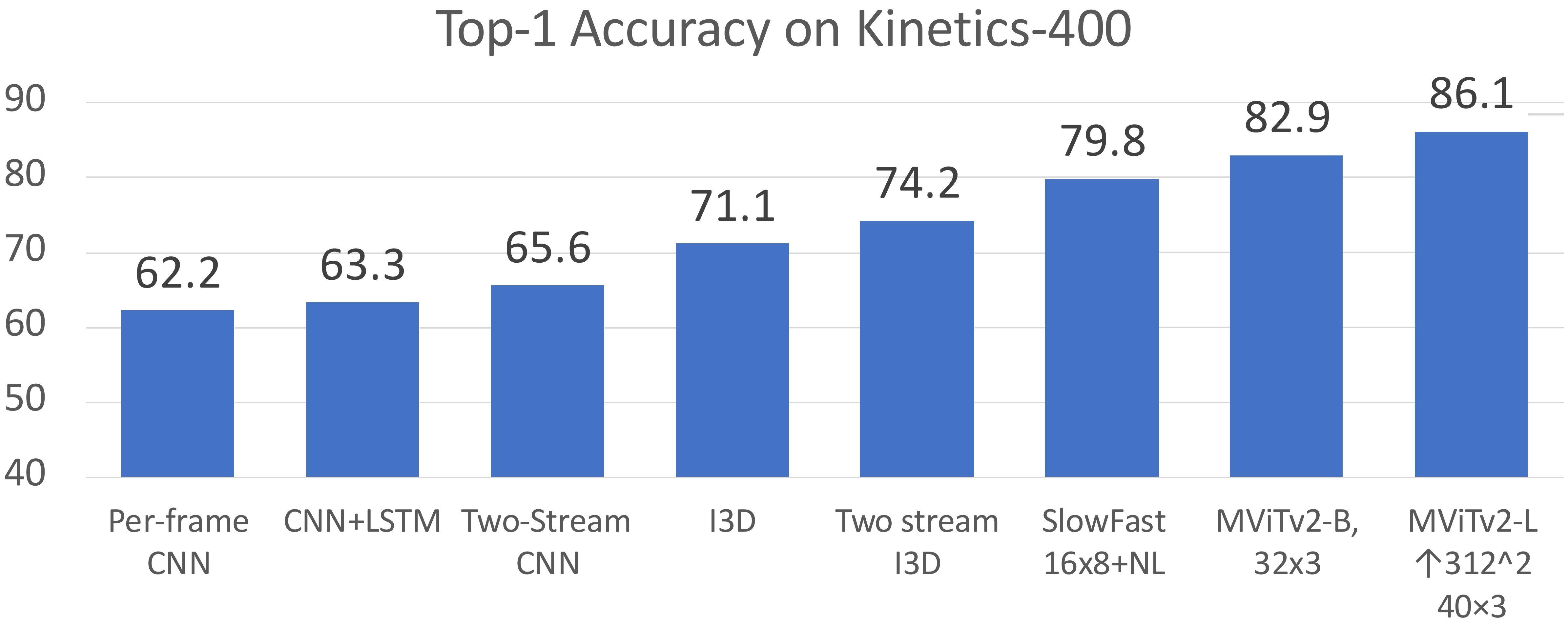
Pooling module: Reduce number of tokens



Bertasius et al, "Is Space-Time Attention All You Need for Video Understanding?", ICML 2021
Arnab et al, "ViViT: A Video Vision Transformer", ICCV 2021

Fan et al, "Multiscale Vision Transformers", ICCV 2021
Li et al, "MViTv2: Improved Multiscale Vision Transformers

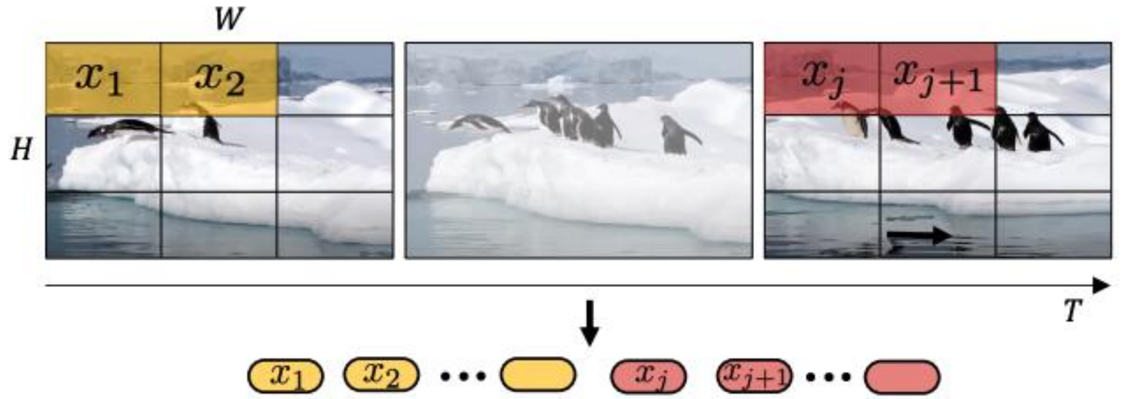
Vision Transformers for Video



Li et al, "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection", CVPR 2022

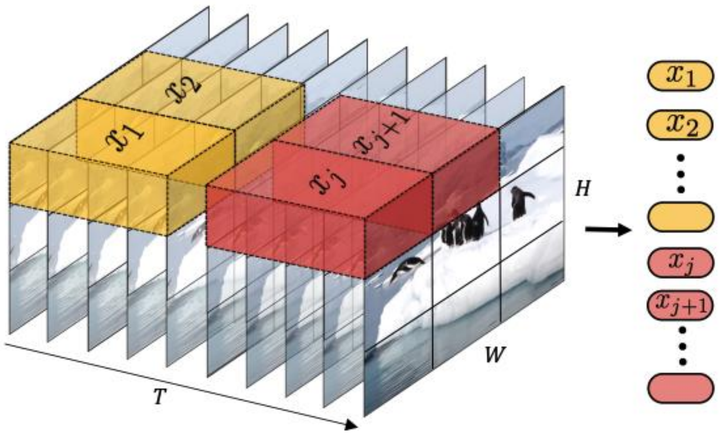
ViViT: A Video Vision Transformer

- Tokenization: Uniform frame sampling
- uniformly sample *nt* frames from the input video clip, embed each 2D frame independently using the same method as ViT
- Concretely, if *nh* x *nw* non-overlapping image patches are extracted from each frame, then a total of *nt* x *nh* x *nw* tokens will be forwarded through the transformer encoder.



Tubelet embedding

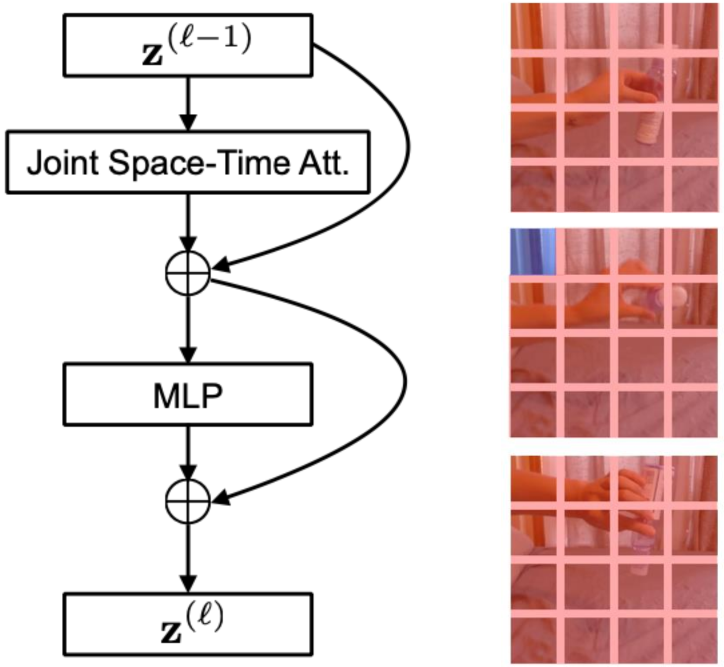
- Extract non-overlapping, spatio-temporal “tubes” from the input volume, and linearly project to d dimensions.
- For a tubelet of dimension t × h × w, nt = floor(T/t), nh = floor(H/h)
- Smaller tubelet dimensions thus result in more tokens which increases the computation.
- Intuitively, this method fuses spatio-temporal information during tokenisation



Model 1: Spatio-temporal attention

- ST

NF + 1 keys



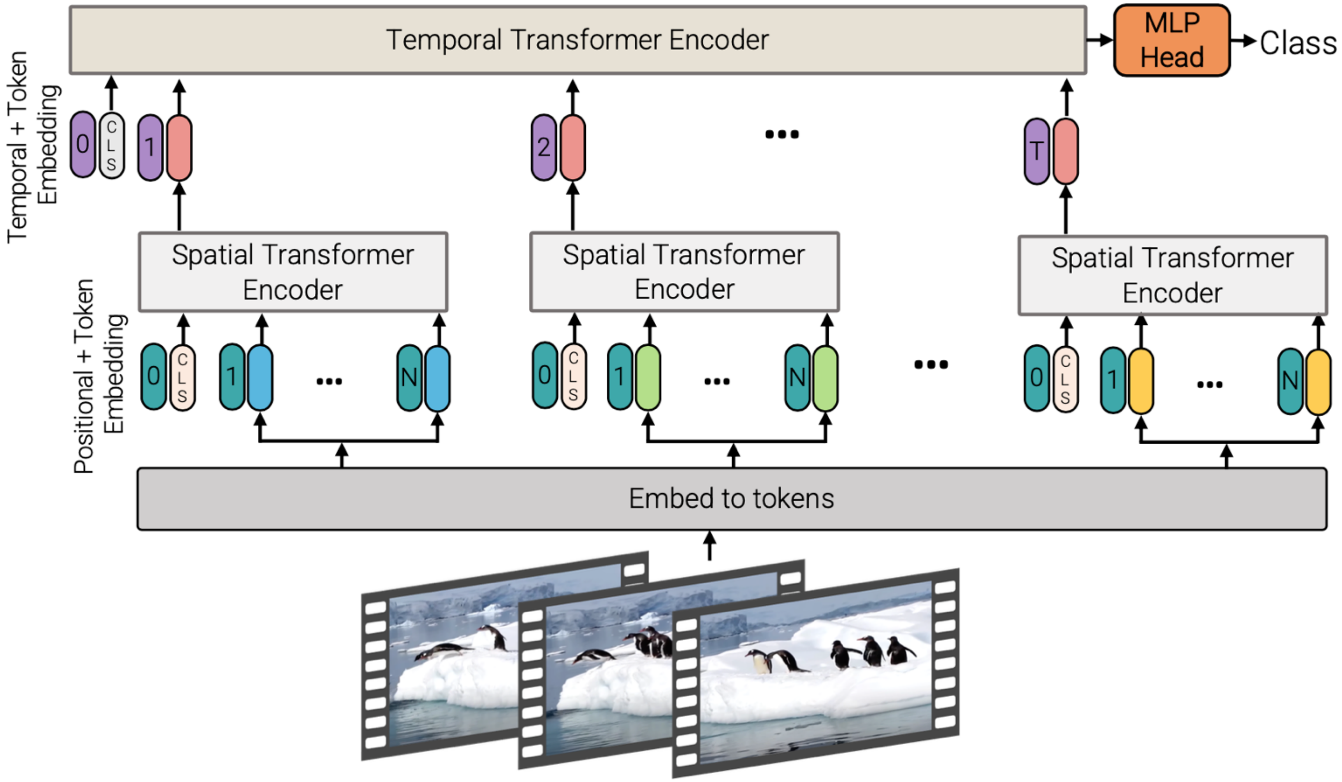
$$\boldsymbol{\alpha}_{(p,t)}^{(\ell,a)} = \mathrm{SM}\left(\frac{\mathbf{q}_{(p,t)}^{(\ell,a)}}{\sqrt{D_h}}^{\top} \cdot \left[\mathbf{k}_{(0,0)}^{(\ell,a)} \left\{\mathbf{k}_{(p',t')}^{(\ell,a)}\right\}_{\substack{p'=1,\ldots,N \\ t'=1,\ldots,F}}\right]\right)$$

$k_{0,0}$ *is the cls token*, $\alpha \in R^{NF+1}$

Gedas et. al. Is Space-Time Attention All You Need for Video Understanding? Arxiv 21

Model 2: Factorised encoder

- Two separate transformer encoders
- The first, spatial encoder, only models interactions between tokens extracted from the same temporal index (same frame)



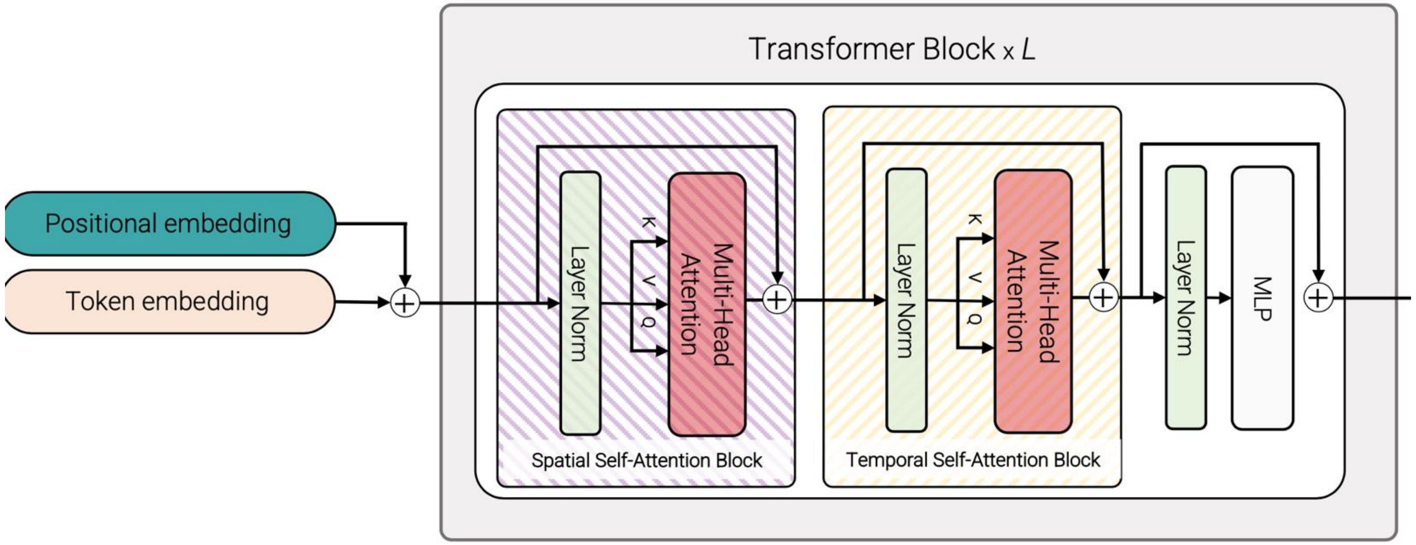
Contd.

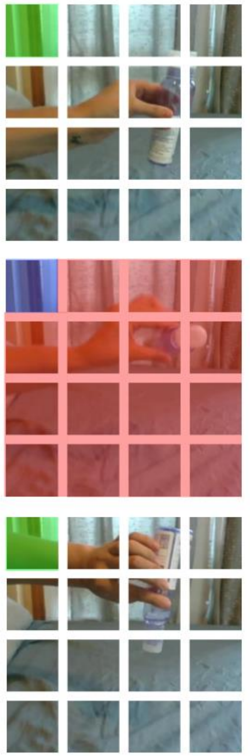
- The frame-level representations, hi, are concatenated into nt ×d , and then forwarded through a temporal encoder consisting of Lt transformer layers to model interactions between tokens from different temporal indices. The output token of this encoder is then finally classified.
- Although this model has more transformer layers than Model 1 (and thus more parameters), it requires fewer floating point operations (FLOPs), as the two separate transformer blocks have a complexity of $O((nh \cdot nw)^2 + nt^2)$ compared to $O((nt \cdot nh \cdot nw)^2)$ of Model 1.

Model 3: Factorised self-attention

Nt is batch for spatial
Nh.nw is batch for temporal

- This operation can be performed efficiently by reshaping the tokens z from nt·nh·nw x d to nt × nh·nw·d to compute spatial self-attention. Similarly, the input to temporal self-attention, is reshaped to nh·nw x nt·d.
- Computing self-attention separately on spatial and temporal dimensions reduc the quadratic complexity typically associated with attention mechanisms from
- $O((nt.nh.nw)^2)$ to $O(nt.(nh.nw)^2) + O(nh.nw.nt^2)$





Divided Space-Time Attention (T+S)

Training

- ViT has been shown to only be effective when trained on large-scale datasets, as transformers lack some of the inductive biases of convolutional networks.
- How- ever, even the largest video datasets such as Kinetics, have several orders of magnitude less labelled examples when compared to their image counterparts.
- As a result, training large models from scratch to high accuracy is extremely challenging.
- To sidestep this issue, and enable more efficient training we initialise our video models from pretrained image models.

Positional Emeddings

- A positional embedding p is added to each input token. However, our video models have nt times more tokens than the pretrained image model.
- As a result, we initialise the positional embeddings by "repeating" them temporally from $nw \cdot nh \times d$ to $nt \cdot nh \cdot nw \times d$.
- Therefore, at initialisation, all tokens with the same spatial index have the same embedding which is then fine-tuned.

Embedding weights

- When using the “tubelet embed- ding” tokenisation method, the embedding filter E is a 3D tensor, compared to the 2D tensor in the pre- trained model, Eimage.
- A common approach for initialising 3D convolutional filters from 2D filters for video classification is to “inflate” them by replicating the filters along the temporal dimension and averaging them
- E = [Eimage,...,Eimage,...,Eimage]/t.

- We consider an additional strategy, which we denote as “central frame initialisation”, where E is initialised with zeroes along all temporal positions, except at the centre $\lfloor t/2 \rfloor$,
- E = [0,...,Eimage,...,0].

Transformer weights for Model 3

- The transformer block in Model 3 (Fig. 5) differs from the pretrained ViT model, in that it contains two multi-headed self attention (MSA) modules.
- In this case, we initialise the spatial MSA module from the pretrained module, and initialise all weights of the temporal MSA with zeroes.

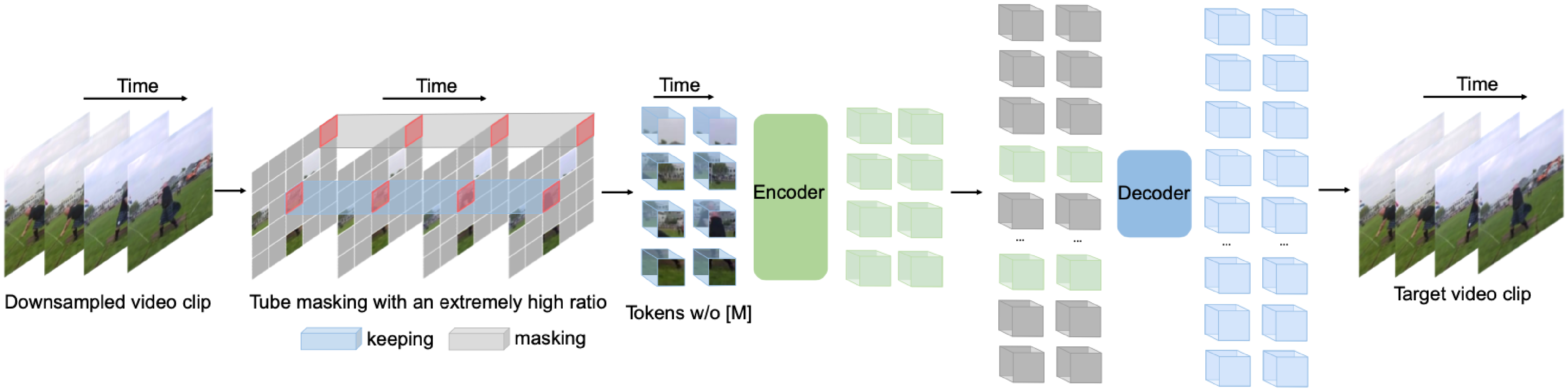
- ViT-Base (ViT-B, L=12, NH =12, d=768),
- L is the number of trans- former layers, each with a self-attention block of NH heads and hidden dimension d
- ViViT-B/16x2 denotes a ViT- Base backbone with a tubelet size of h×w×t = 16×16×2

Table 1: Comparison of input encoding methods using ViViT-B and spatio-temporal attention on Kinetics. Further details in text.

	Top-1 accuracy
Uniform frame sampling	78.5
Tubelet embedding	
Random initialisation [25]	73.2
Filter inflation [8]	77.6
Central frame	79.2

Video MAE – NeurIPS22

- VideoMAE takes the *downsampled frames* as inputs and uses the *cube embedding* to obtain video tokens. Then, we propose a simple design of *tube masking with high ratio* to perform MAE pre-training with an asymmetric encoder- decoder architecture. Our backbone uses the vanilla ViT with *joint space-time attention*.



Contd.

- one video clip consisting of t consecutive frames is first randomly sampled from the original video V .
- We then use temporal sampling to compress the clip to T frames, each of which contains $H \times W \times 3$ pixels.
- In experiments, the stride τ is set to 4 and 2 on Kinetics and Something-Something, respectively.

Cube embedding

- We adopt the joint space-time cube embedding in our VideoMAE, where we treat each cube of size $2 \times 16 \times 16$ as one token embedding.
- Thus, the cube embedding layer obtains $T/2 \times H/16 \times W/16$ 3D tokens and maps each token to the channel dimension D.
- This design 2 x16 x16 can decrease the spatial and temporal dimension of input, which helps to alleviate the spatiotemporal redundancy in videos.
- Mask ratio 90-95%.

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
NL I3D [78]	ResNet101	ImageNet-1K	✓	128	359×10×3	62	77.3	93.3
TANet [41]	ResNet152		✓	16	242×4×3	59	79.3	94.1
TDN_{En} [75]	ResNet101		✓	8+16	198×10×3	88	79.4	94.4
TimeSformer [6]	ViT-L	ImageNet-21K	✓	96	8353×1×3	430	80.7	94.7
ViViT FE [3]	ViT-L		✓	128	3980×1×3	N/A	81.7	93.8
Motionformer [51]	ViT-L		✓	32	1185×10×3	382	80.2	94.8
Video Swin [39]	Swin-L		✓	32	604×4×3	197	83.1	95.9
ViViT FE [3]	ViT-L	JFT-300M	✓	128	3980×1×3	N/A	83.5	94.3
ViViT [3]	ViT-H	JFT-300M	✓	32	3981×4×3	N/A	84.9	95.8
VIMPAC [65]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	77.4	N/A
BEVT [77]	Swin-B	IN-1K+DALLE	✗	32	282×4×3	88	80.6	N/A
MaskFeat↑352 [80]	MViT-L	Kinetics-600	✗	40	3790×4×3	218	87.0	97.4
ip-CSN [69]	ResNet152	*no external data*	✗	32	109×10×3	33	77.8	92.8
SlowFast [23]	R101+NL		✗	16+64	234×10×3	60	79.8	93.9
MViTv1 [22]	MViTv1-B		✗	32	170×5×1	37	80.2	94.4
MaskFeat [80]	MViT-L		✗	16	377×10×1	218	84.3	96.3
VideoMAE	ViT-S	*no external data*	✗	16	57×5×3	22	79.0	93.8
VideoMAE	ViT-B		✗	16	180×5×3	87	81.5	95.1
VideoMAE	ViT-L		✗	16	597×5×3	305	85.2	96.8
VideoMAE	ViT-H		✗	16	1192×5×3	633	**86.6**	**97.1**
VideoMAE↑320	ViT-L	*no external data*	✗	32	3958×4×3	305	86.1	97.3
VideoMAE↑320	ViT-H		✗	32	7397×4×3	633	**87.4**	**97.6**

Comparison with the state-of-the-art methods on Kinetics-400

Further reading

- An Empirical Study of End-to-End Video-Language Transformers with Masked Visual Modeling, CVPR2023
- Masked Video Distillation: Rethinking Masked Feature Modeling for Self-supervised Video Representation Learning, CVPR2023
- VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking, CVPR2023
-