Computer Vision by Learning

Cees Snoek, UvA
Efstratios Gavves, UvA
Laurens van der Maaten Facebook

Advanced School for Computing and Imaging
University of Amsterdam
ICAI Innovation Center for Artificial Intelligence

Tomorrow

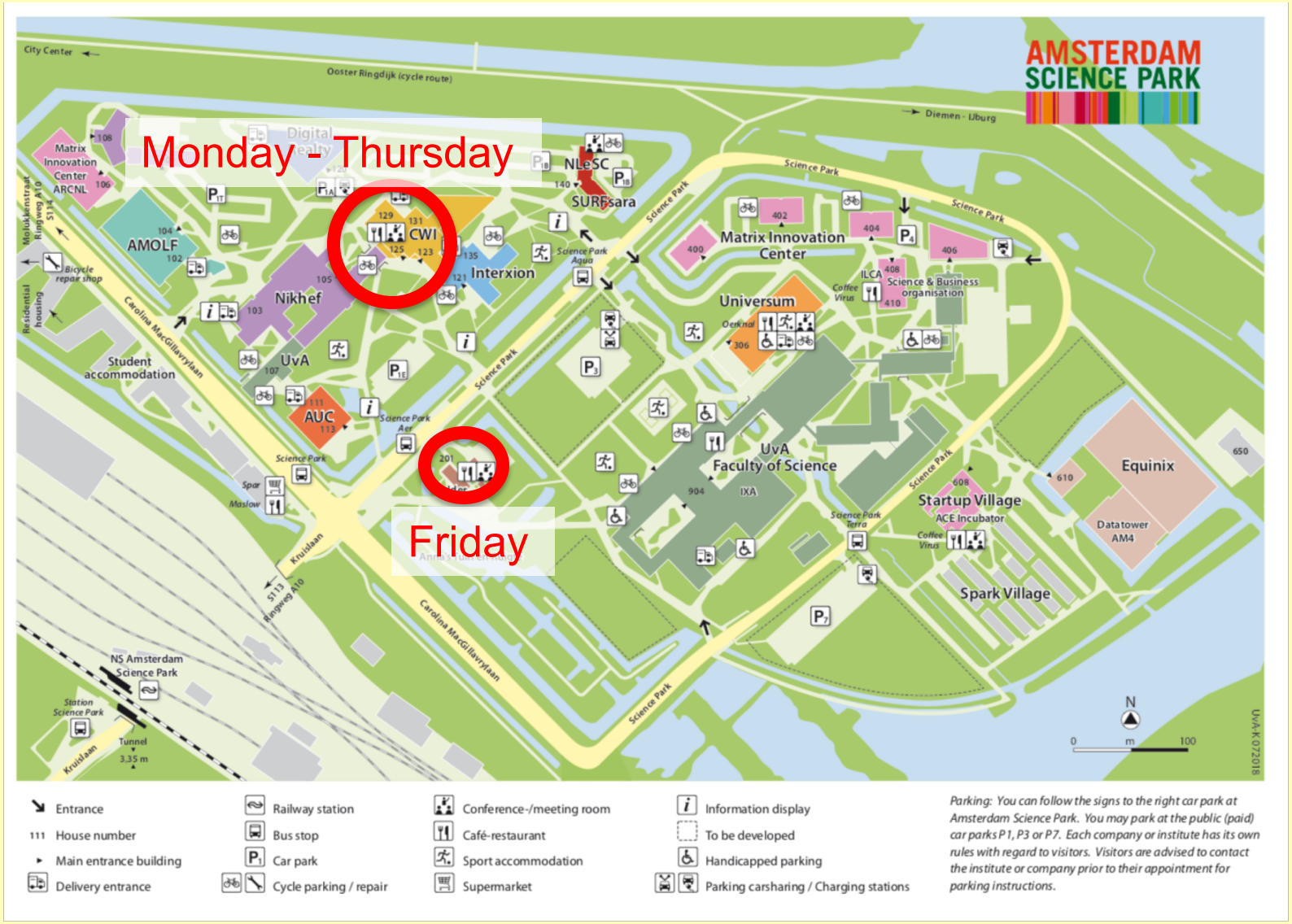
Invited tutorial by **Laurens van der Maaten**

- Efficient convolutional networks
- From visual recognition to visual reasoning

Note change of location

- **Polderzaal, next to Café-Restaurant Polder**

Note: different location tomorrow



Lab

Lab Monday	Vision by multi-layer perceptron
Lab Tuesday	Vision by convolutional neural network
Lab Wednesday	Vision by recurrent neural network
Lab Thursday	Vision by generative adversarial network

Each student hands in the Python notebooks per assignment completed with code and answers.

Deadline: **April 30th, 2019**

Overview Day 4

Computer video by learning

1. Introduction, activities, data, paradox, tasks
2. Video representations, appearance, motion, space and time
3. VideoLSTM, convolutions, attention and localization for free
4. Video time, properties, encoders and evaluation
5. Video and language for tracking and action segmentation
6. Weakly-supervised video recognition

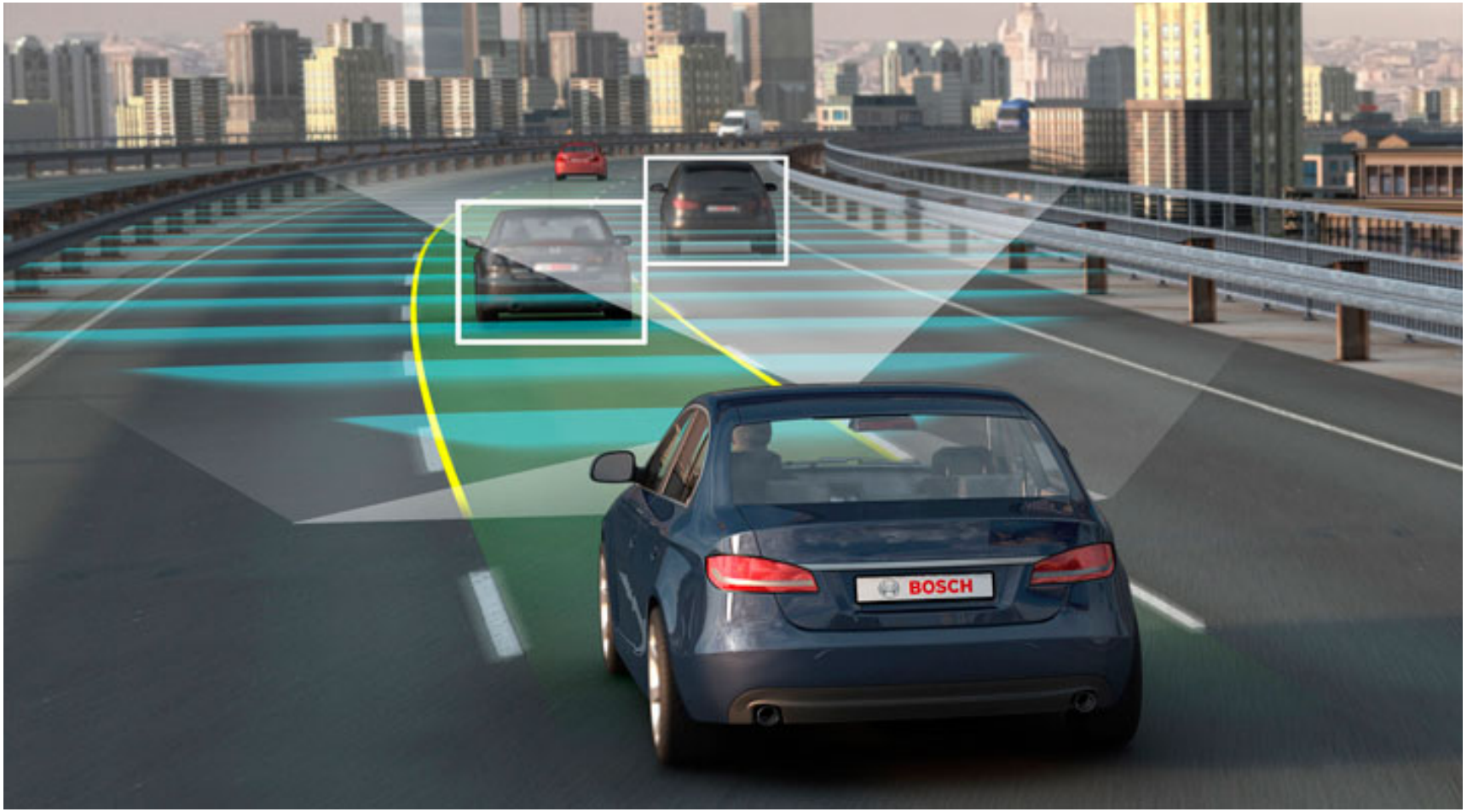
1. Introduction

By 2022 there will be 45 billion cameras in the world, many of them tiny, connected and live streaming 24/7. Self-driving cars, drones and service robots are just three manifestations. For all these applications it will be of critical importance to understand what is happening where and when in the video streams. In this chapter we cover activities, the data paradox, and challenges.

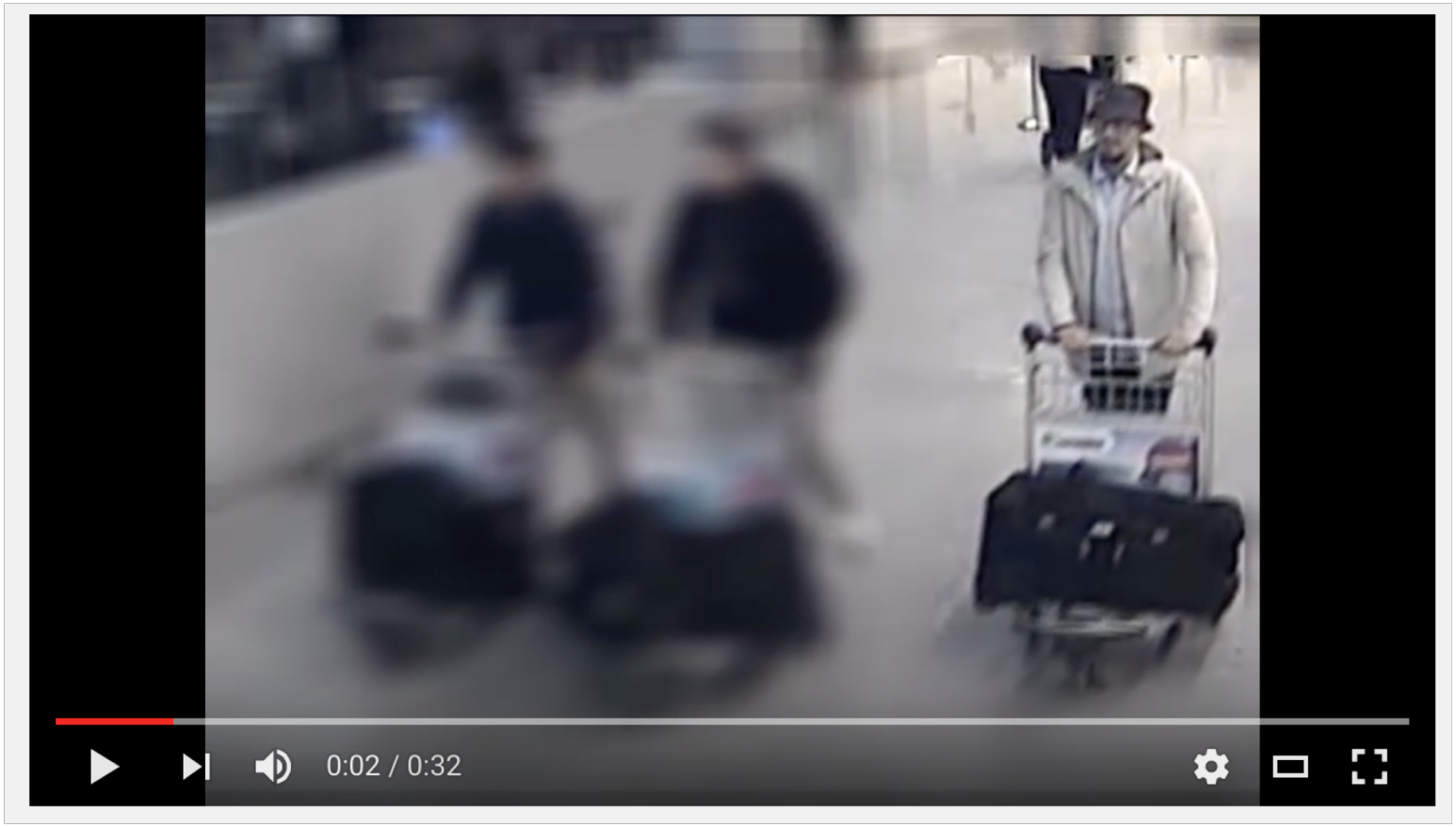
Motivation: Internet of things that video

45 billion cameras by 2022… [LDV Capital]

Technology: self-driving cars



Forensics: Analyzing terrorist behavior



Well-being: elderly monitoring

Figure 1. Examples of interaction patterns in a nursing home

Chen et al. MM 2004

Safety: preventive monitoring

parool.nl

Het Parool HOME AMSTERDAM STADSGIDS OPINIE

Studentencomplexen krijgen meer veiligheidscamera's

Risicoplek aan de Wenckebachweg. © Mats van Soolingen

Studentencomplexen zijn niet onveiliger dan andere woonwijken. Wel moeten in verband met incidenten, waarbij studentes zijn aangevallen, veiligheidsmaatregelen getroffen worden.

GERELATEERD

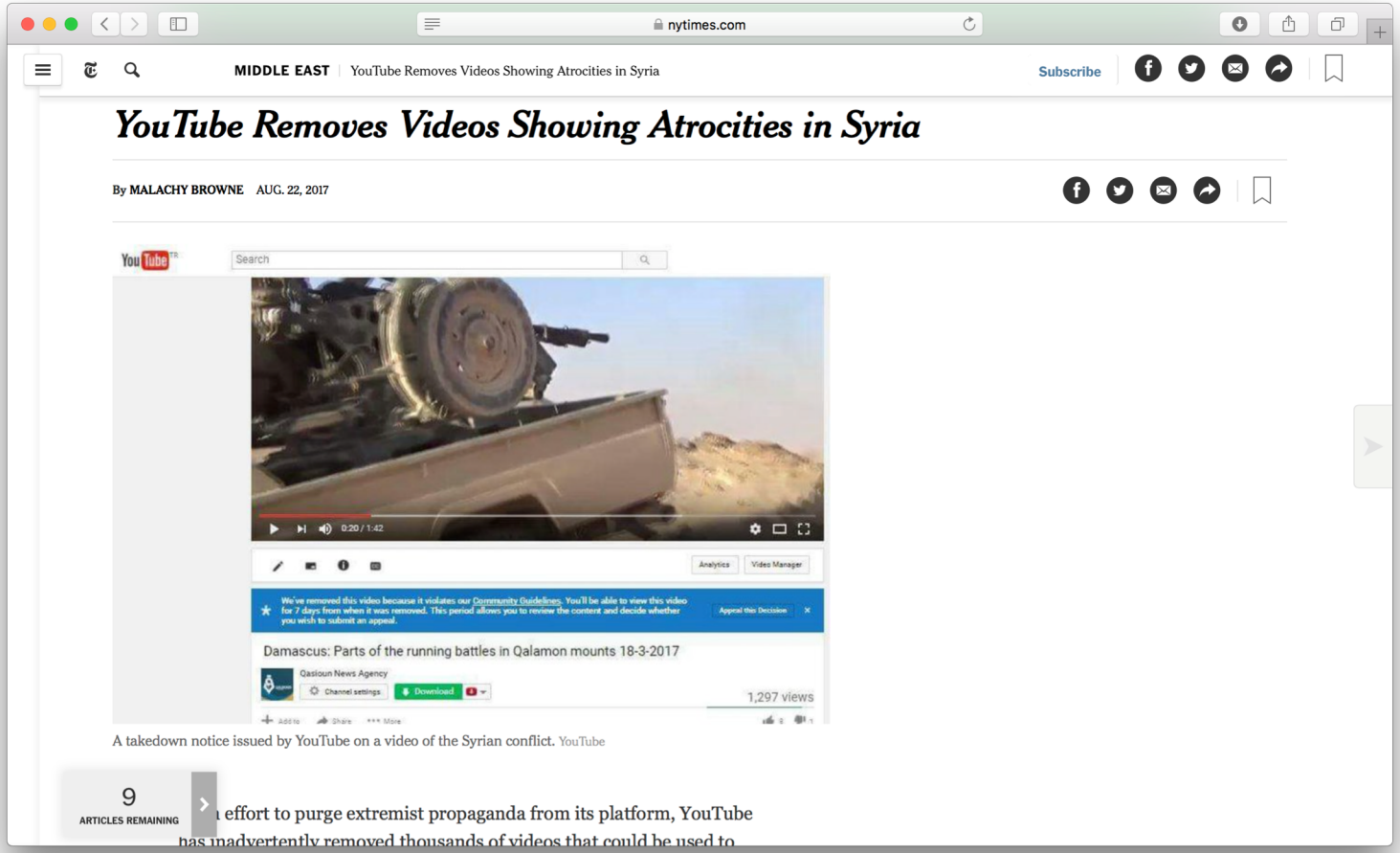
Dit is wat we nu weten over de moord op Djordy Latumahina
30 MAART 2017

Bewoners K-buurt krijgen hun zin (en een politiek slagveld)
30 MAART 2017

Bijzonder bevrijdingsvliegtuig teruggevonden
30 MAART 2017

Street surveillance

Social: media monitoring



Retail: cashier-less shopping

What is an activity?

No clear definition in the literature, also known as action or event.

Typically involves person interacting with an object

Repairing an appliance

Working on sewing project

Grooming an animal

Birthday party

Goal of activity recognition

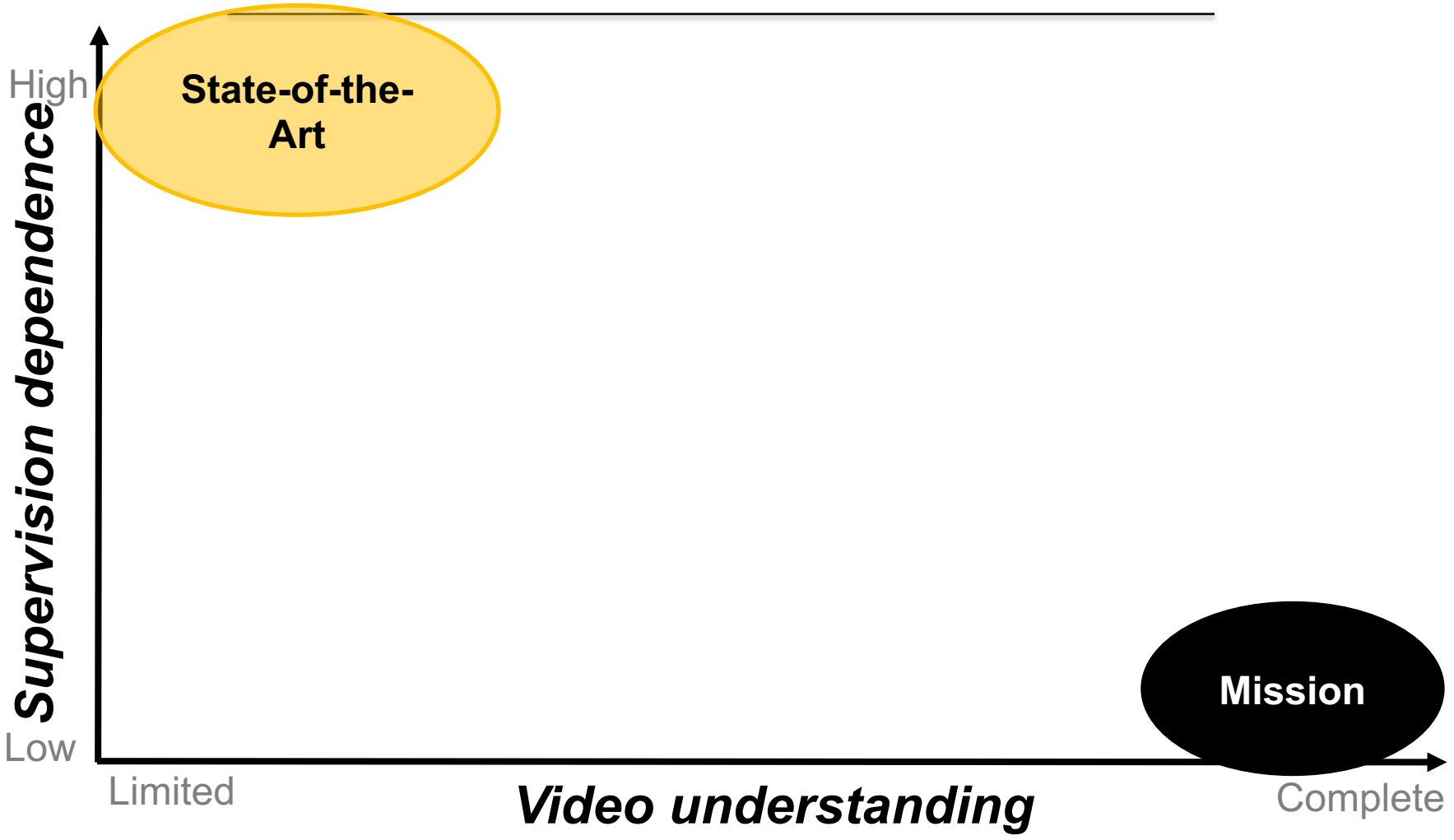
Understand **what** is happening where, **when** and **why**

Kissing

Paradox

As activities become more and more specific, it is unrealistic to assume that ample examples to learn from will be commonly available.

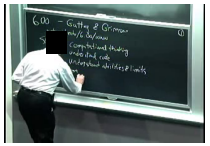
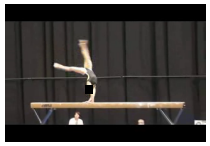




State-of-the-Art
Mission
High
Low
Supervision dependence
Limited
Complete
Video understanding

ImageNet equivalent for videos?

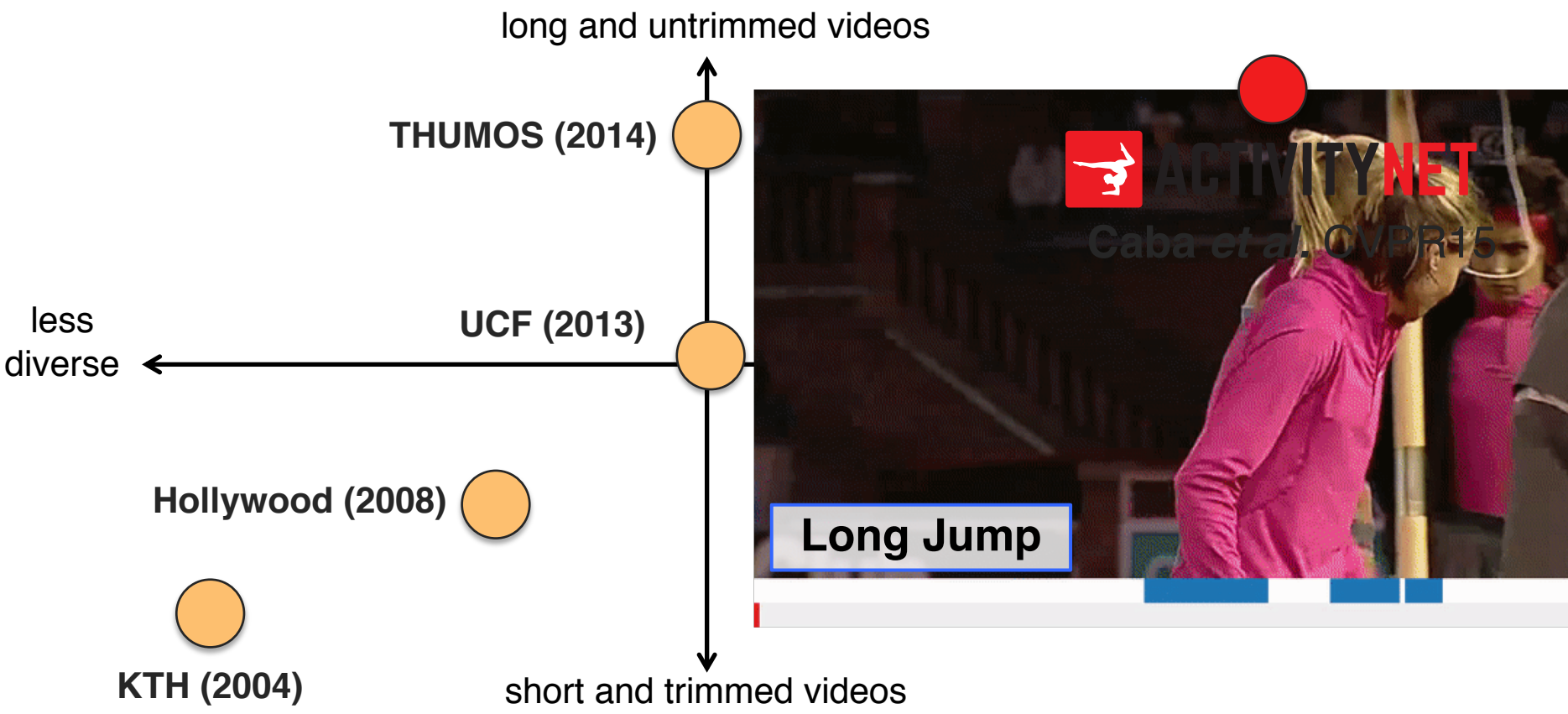
- 1000 classes
- 1.2M images
- “Cleanly” labeled

Activity datasets are small scale

UCF101	 	101 classes / 13,320 clips / web video
THUMOS14	 	101 classes / 15,915 clips / web video
Hollywood2	 	12 classes / 1,707 clips / movies
HMDB51	 	51 classes / 6,766 clips / diverse video
UCF Sports	 	10 classes / 150 clips / sports broadcasts
KTH	 	6 classes by 25 actors

Recently, many new datasets proposed

Activity Understanding Datasets (2015)



ACTIVITYNET

Peeling Potatoes

Playing Badminton

Polishing Shoes

Shoveling Snow

Horse Riding

Vacuuming Floor

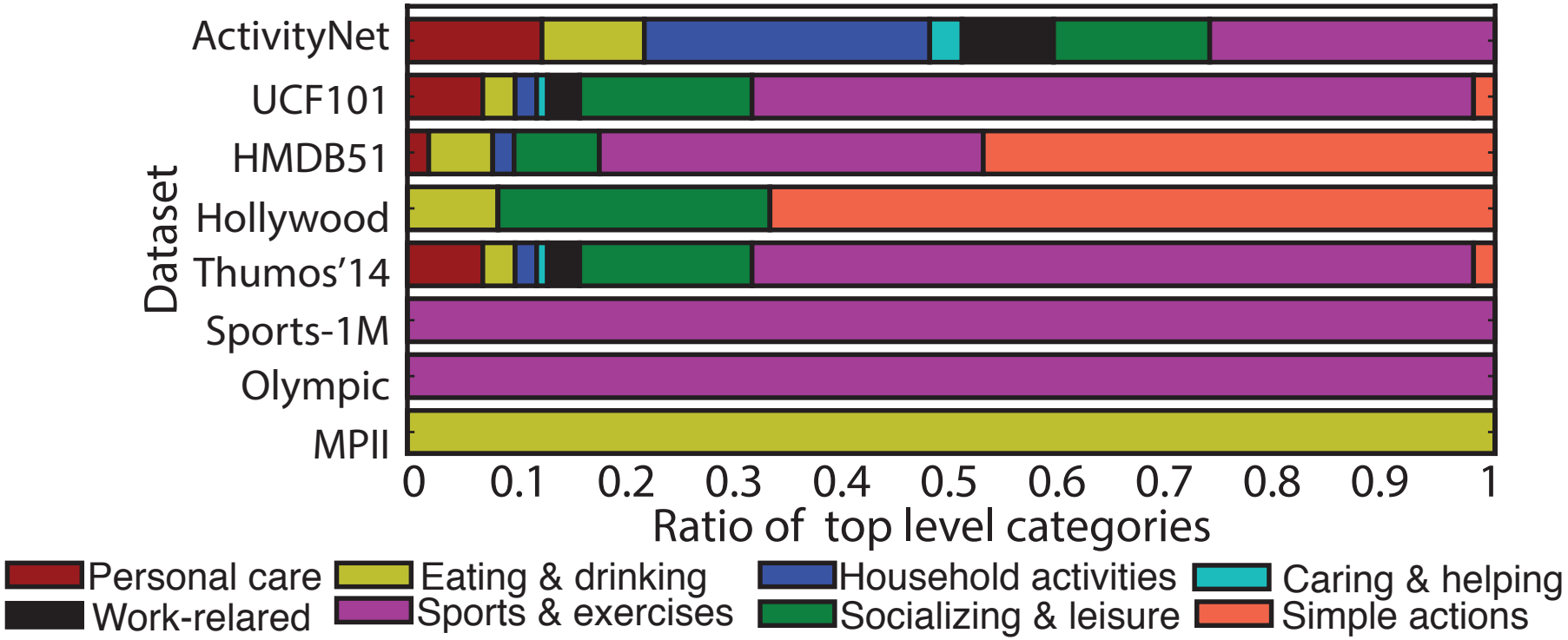
Caba *et al.* CVPR15

ImageNet equivalent for videos?

- 1000 classes
- 1.2M images
- “Cleanly” labeled

- 200 “fine-grained” classes
- 20K videos
- Untrimmed

Diversity comparison



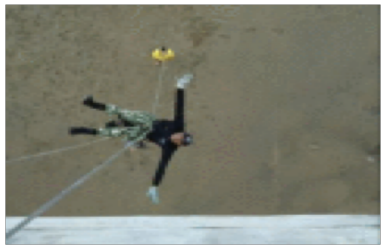
ImageNet equivalent for videos?

- 1000 classes
- 1.2M images
- “Cleanly” labeled

- 200 “fine-grained” classes
- 20K videos
- Untrimmed

- ~~400~~ 600 “fine-grained” classes
- ~~300K~~ 500K videos
- Trimmed

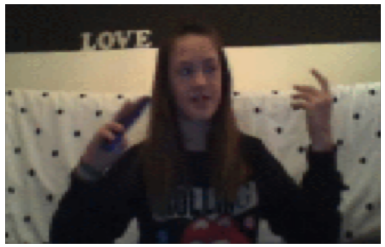
Kinetics Dataset



Abseiling



Air drumming



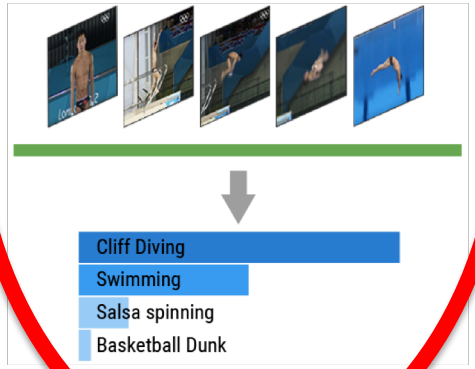
Answering Questions

Carreira & Zisserman CVPR17

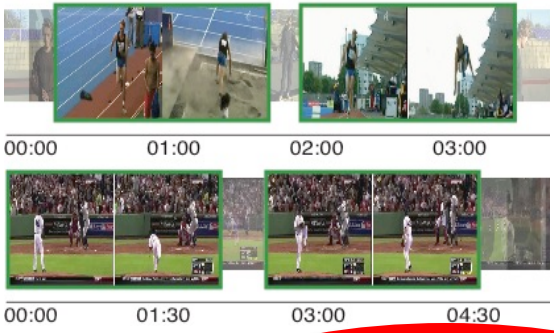
https://deepmind.com/research/open-source/open-source-datasets/kinetics/

Activity Understanding Tasks

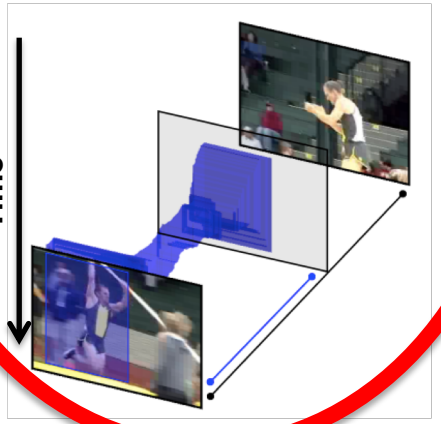
Activity Classification



Temporal Activity Localization



Spatiotemporal Activity Localization



Other tasks: Video Captioning

Actor-action Segmentation

Activity Prediction

Video Summarization

2. Video representations

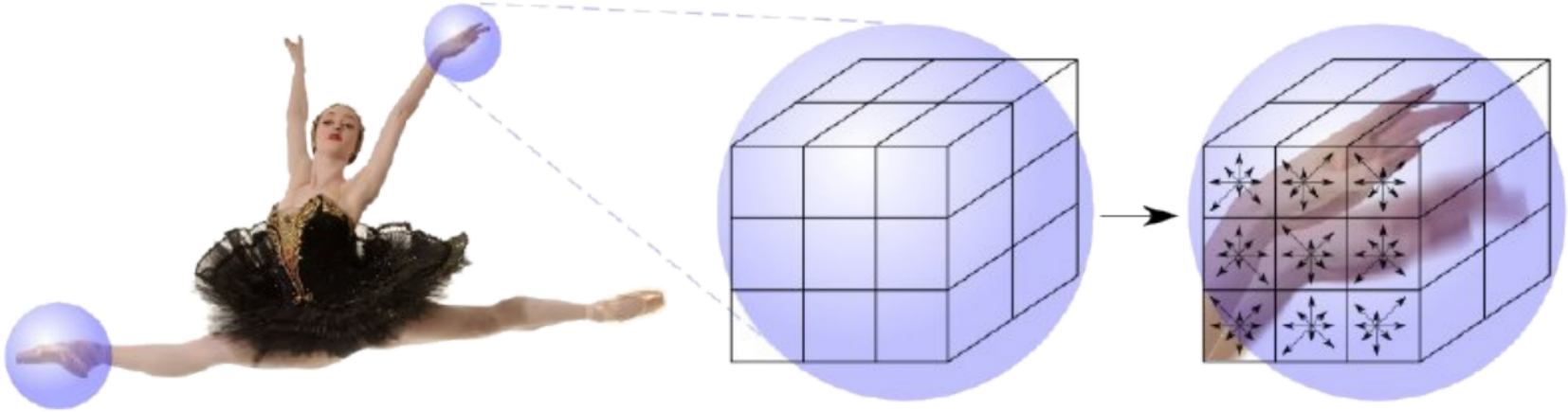
In this chapter we consider video representation learning for action recognition. We analyze existing algorithms for their ability to capture intrinsic and extrinsic action properties, including appearance, motion, spatial and temporal extent.

Shallow action recognition

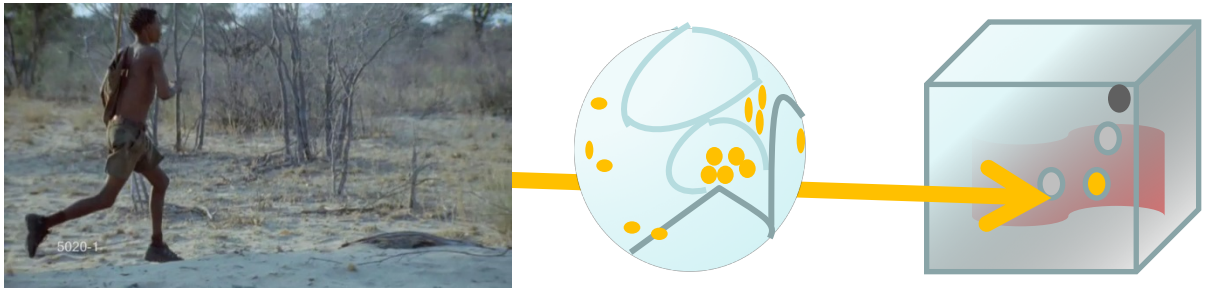
Spatio-temporal Interest point detection

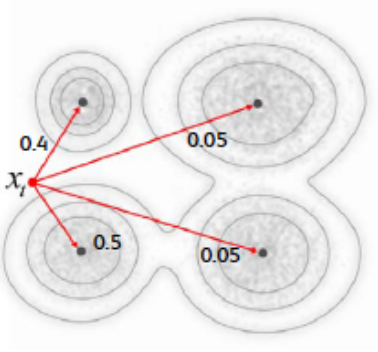
Space-time patch/trajectory

Space-time descriptor



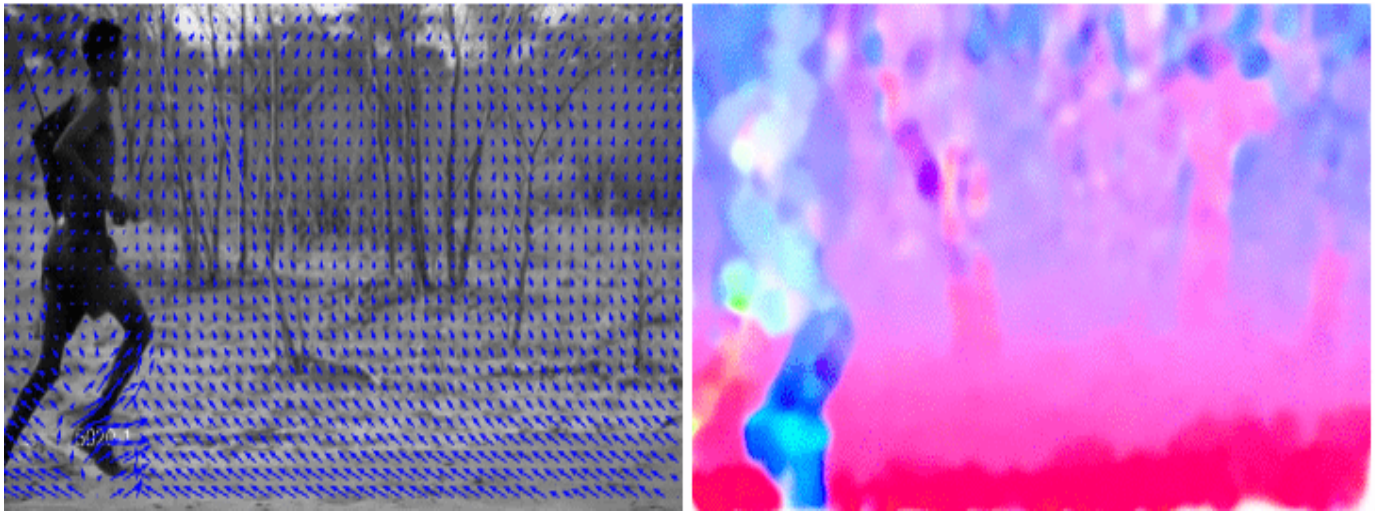
Followed by Bag-of-Words/Fisher vector and SVM



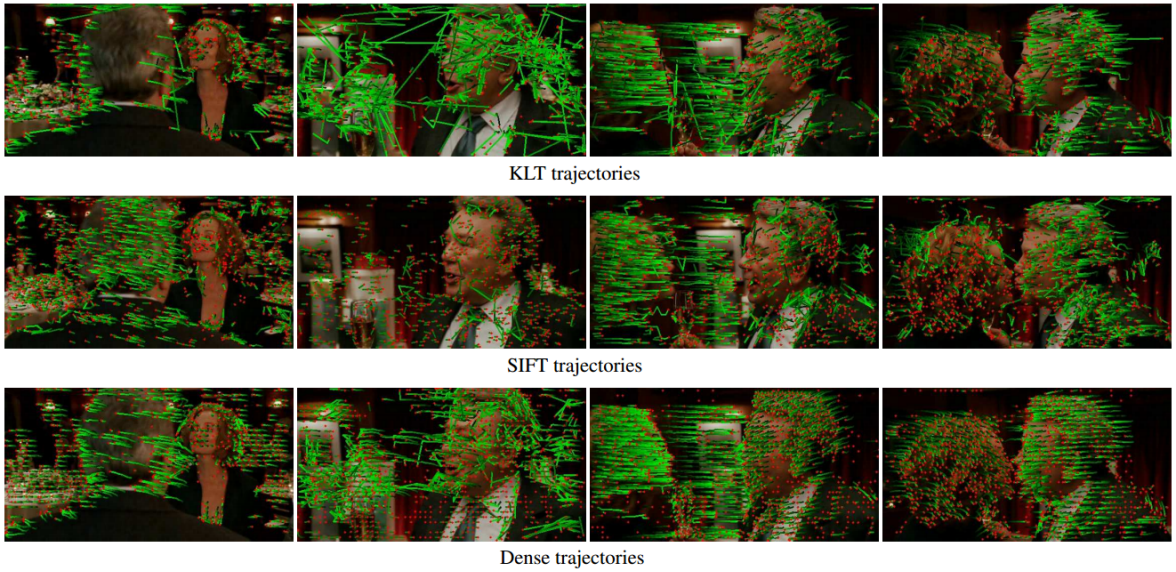


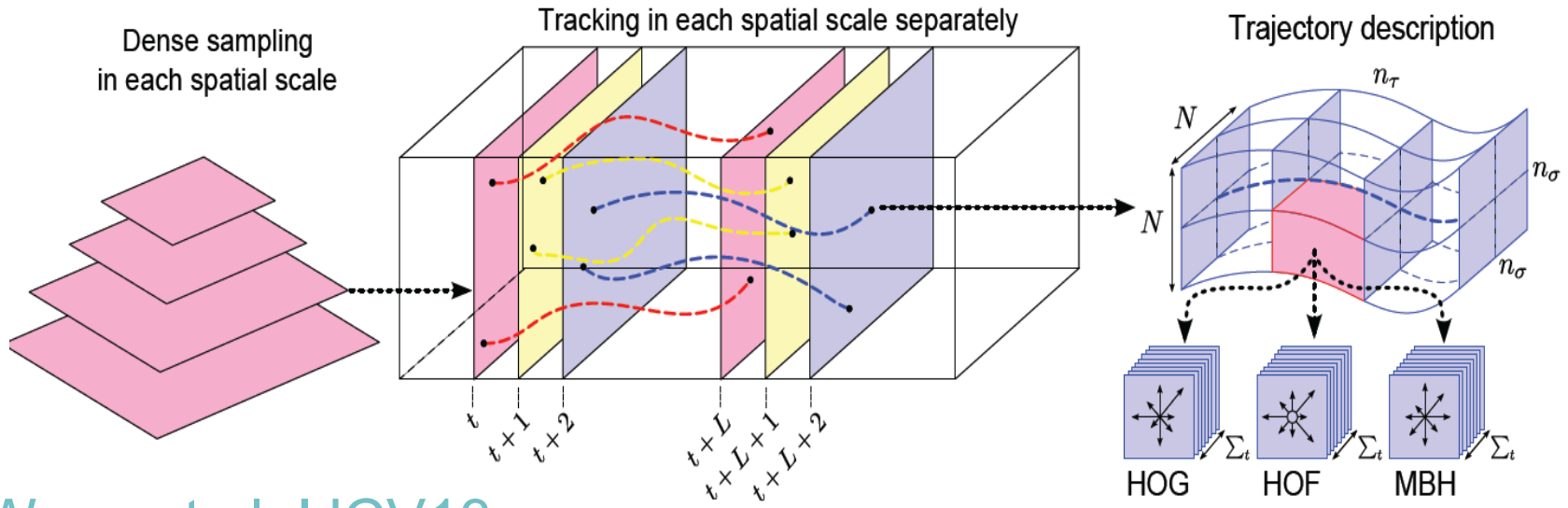
Motion is salient

Motion offers crucial clue where to attend in video



Flow trajectory descriptors





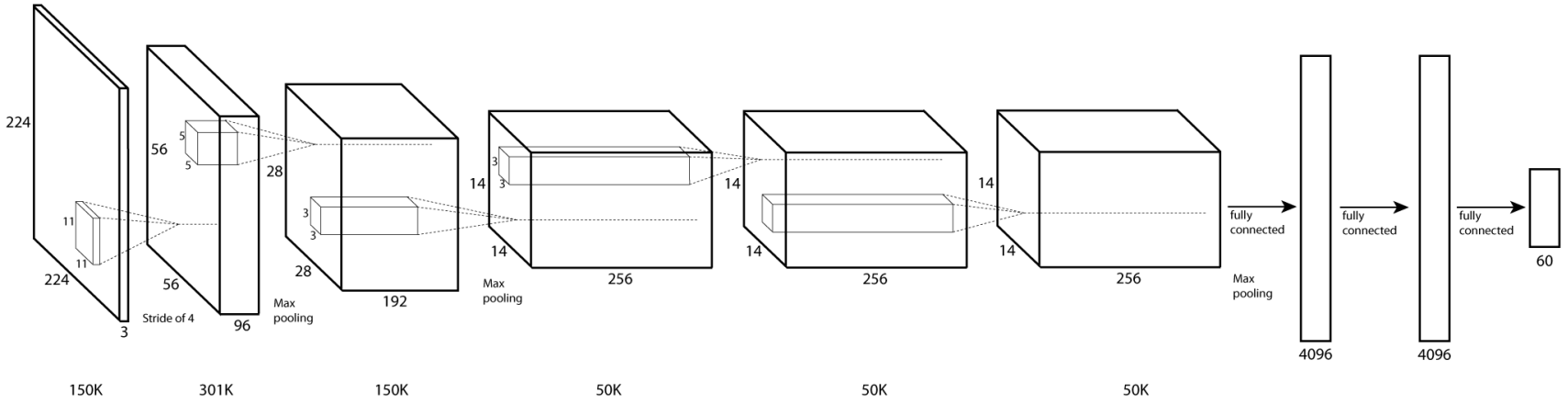
Wang et al. IJCV13

2D Appearance only

Pre-train on ImageNet, fine-tune on video concepts

Average pooling over multiple frames per video shot

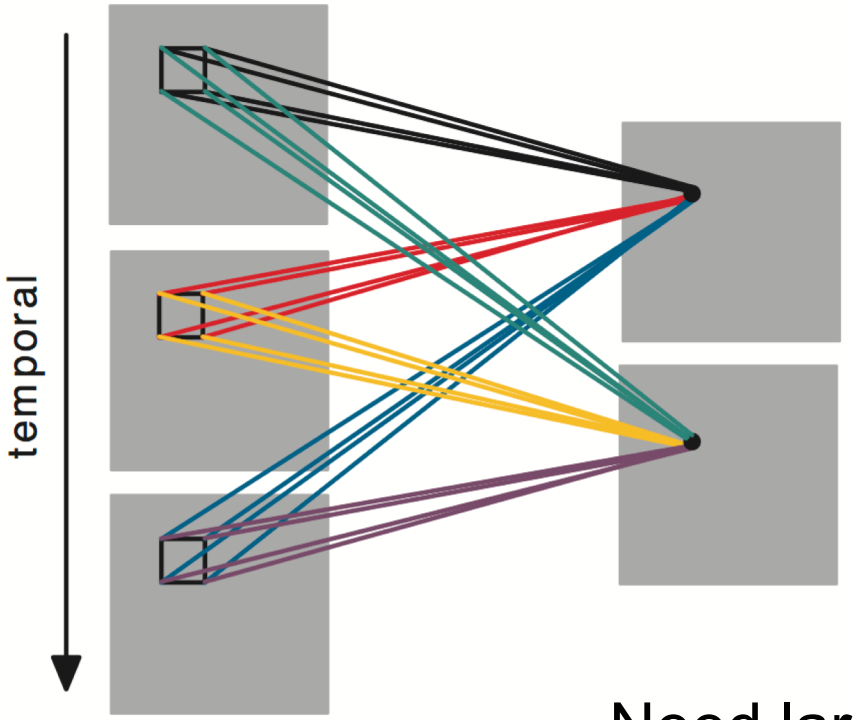
Good for objects and scenes, so and so for actions



Snoek et al. TRECVID13

3D Appearance only

Extracts multiple features from both spatial and temporal dimensions by performing 3D convolutions

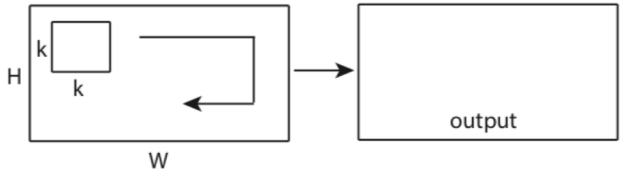


Need large amounts of data to learn filters

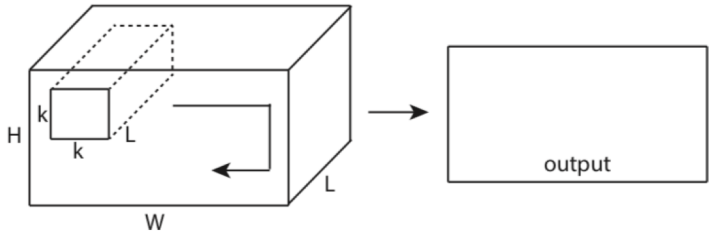
Ji et al. ICML10

2D vs 3D convolution

A) Applying 2D convolution on an image results in an image.



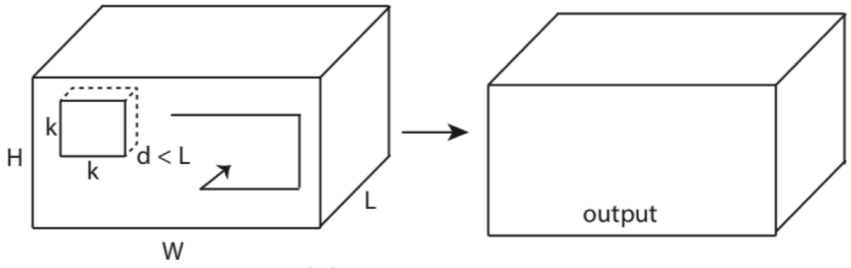
(a) 2D convolution



(b) 2D convolution on multiple frames

B) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image

C) Applying 3D convolution on a video volume results in another volume, preserving **short term** temporal information of the input signal



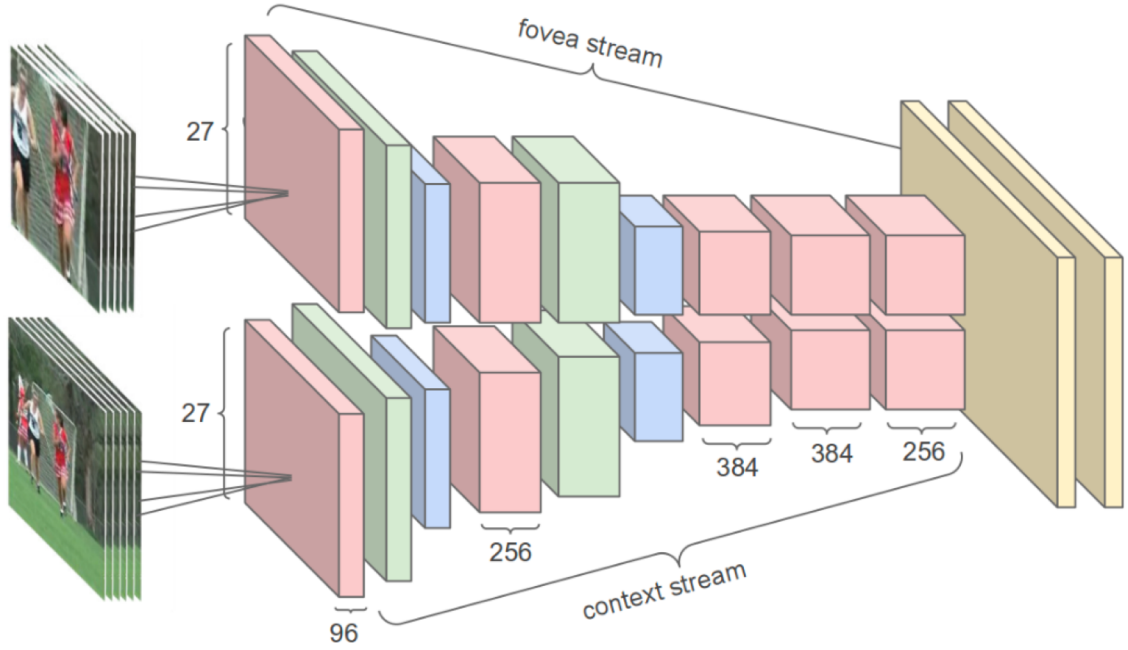
(c) 3D convolution

Tran et al. ICCV 15

2D Appearance and spatial

Models spatial extent of action mildly by separating stream for center crop and entire frame

Considers various temporal pooling schemes

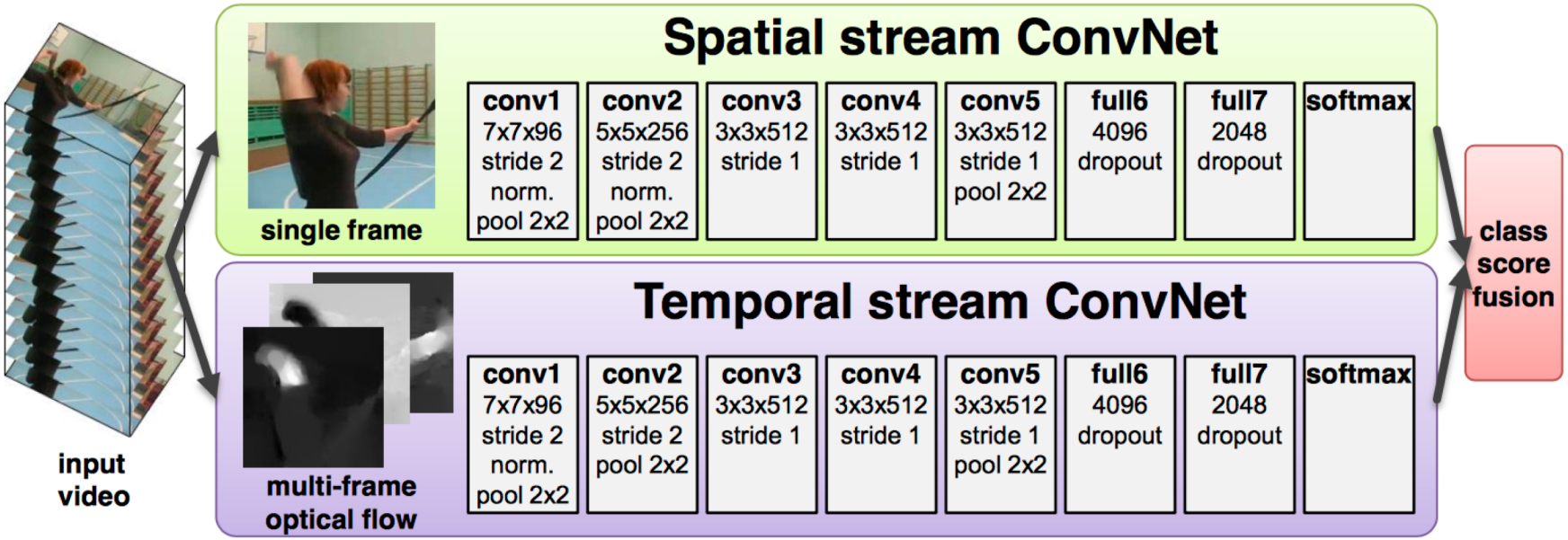


Introduces Sports1M dataset

Karpathy & Fei-Fei. CVPR14

2D Appearance and motion

Learn spatial and temporal filters separately
- Pre-train on ImageNet, fine-tune for actions
- Fusion by averaging or SVM

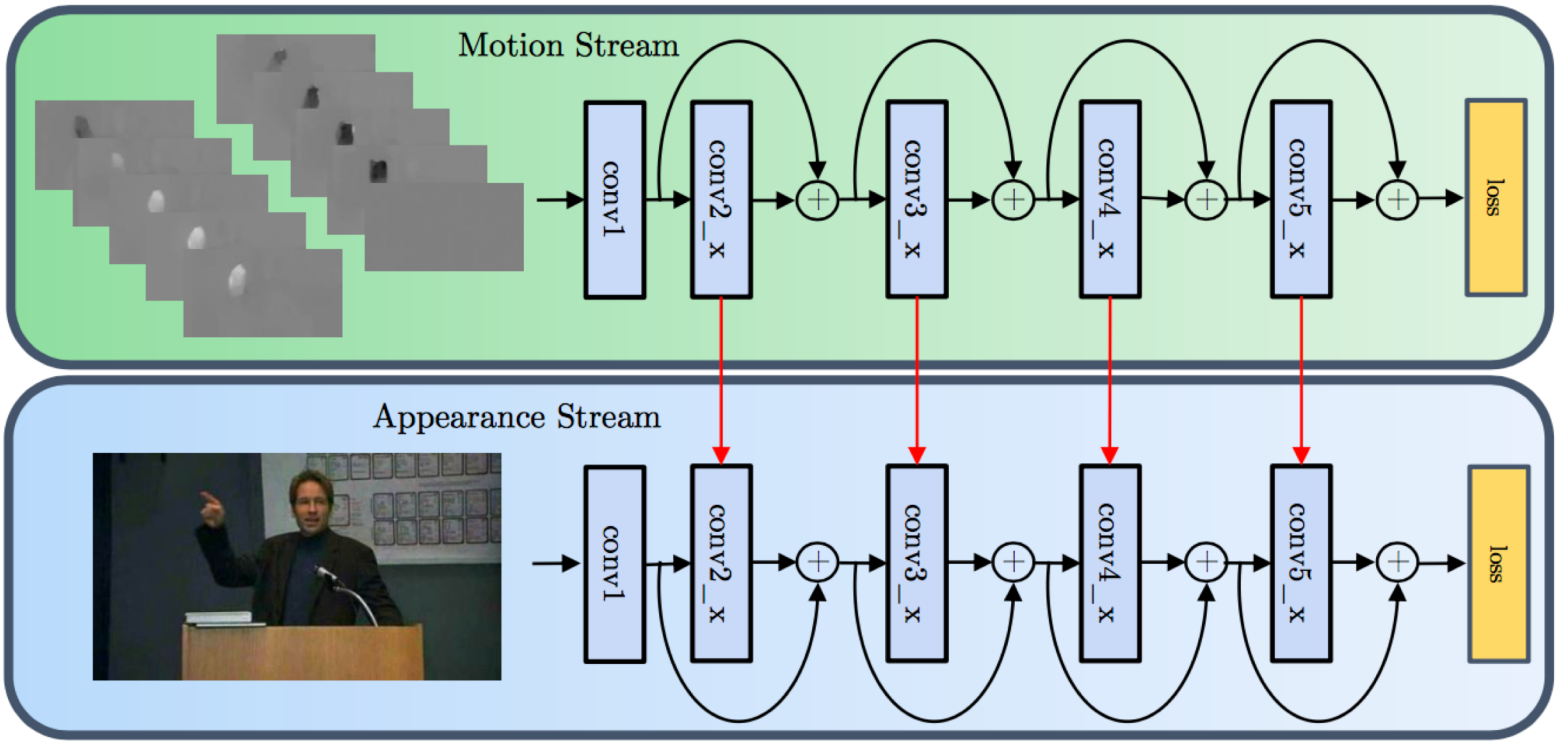


Simonyan & Zisserman, NIPS14

2D Appearance, motion and spatial

Generalize ResNet for video

Remove fully-connected layers



Feichtenhofer et al, NIPS16

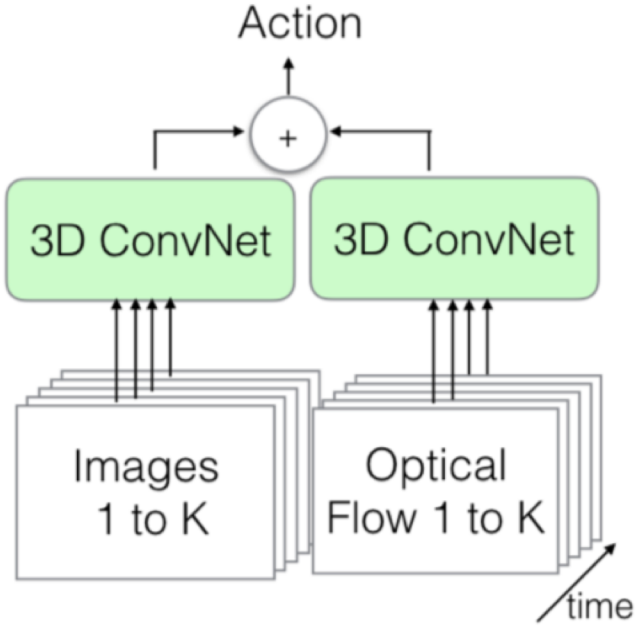
Inflated 3D (appearance and motion)

Simply convert 2D classification models into 3D ConvNets.

From 2D architecture, inflate all the filters and pooling kernels – effectively adding a temporal dimension

Square *N x N* filters become cubic: *N x N x N*

Carreira & Zisserman CVPR17



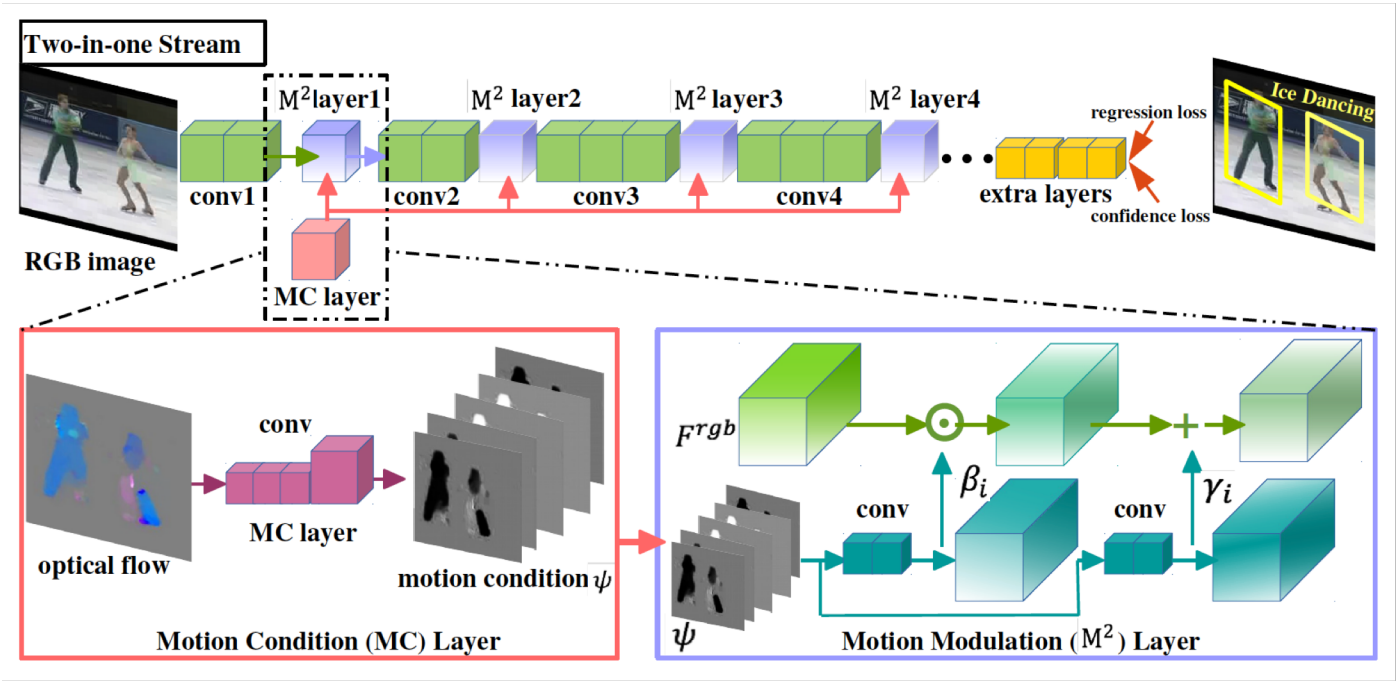
Two-in-one Stream

Appearance, motion and spatial into a single stream

Exploits feature modulation

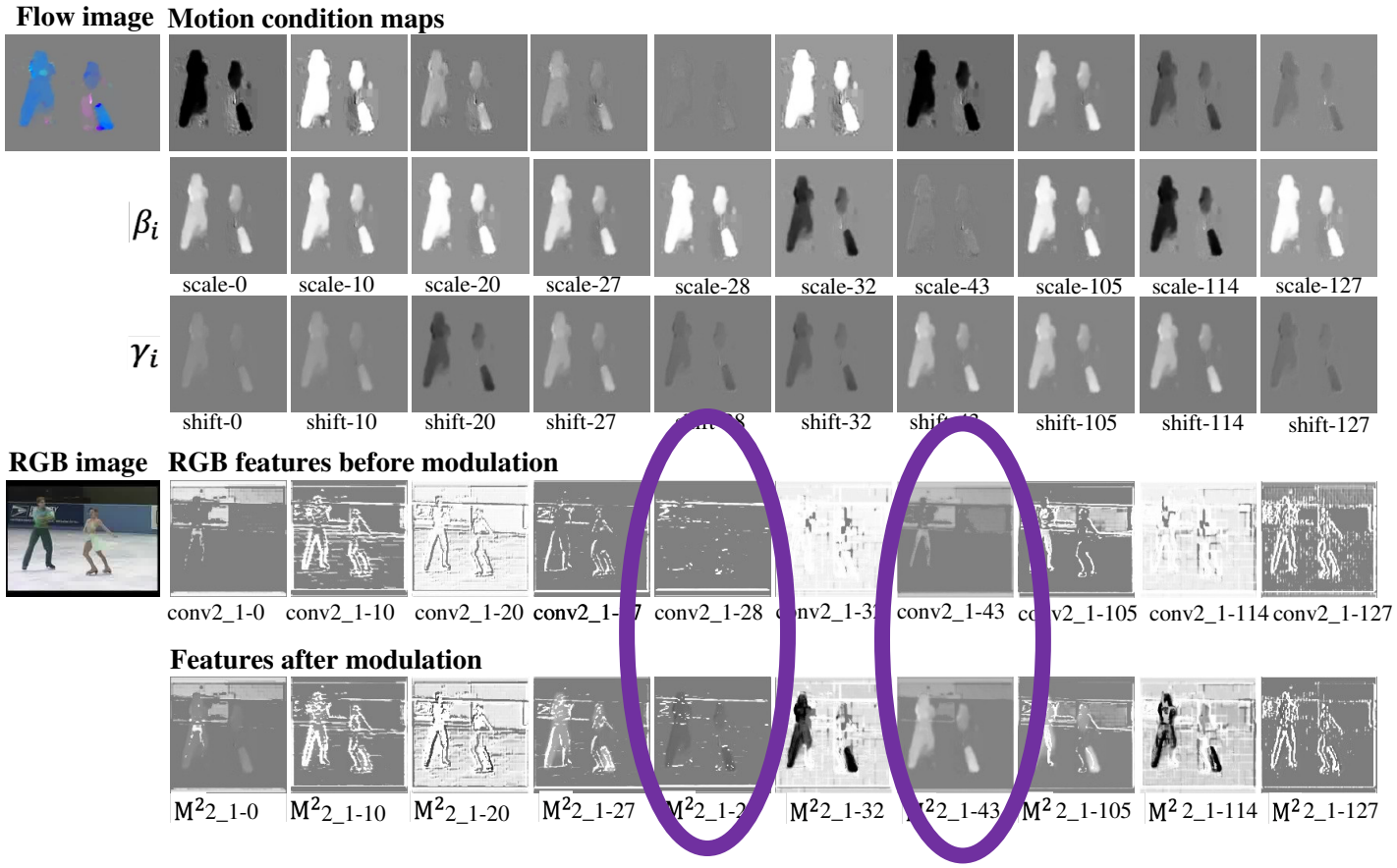
Better accuracy, with half the parameters

Particularly well suited for spatio-temporal action detection



Zhao & Snoek, CVPR19

Two-in-one Stream visualization

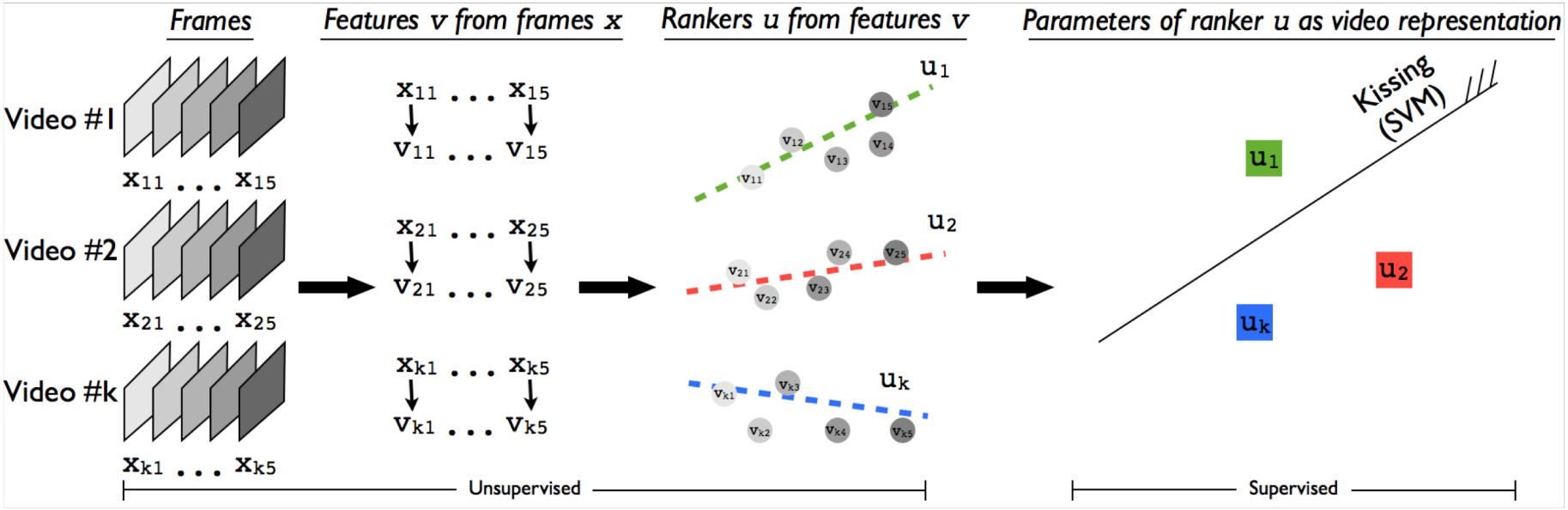


Modulated features focus more on moving actors.

2D Appearance, motion and temporal

Key insight: exploit temporal order as **soft label**

Actions vary in appearance but order is preserved



Video-specific ranker parameters as representation

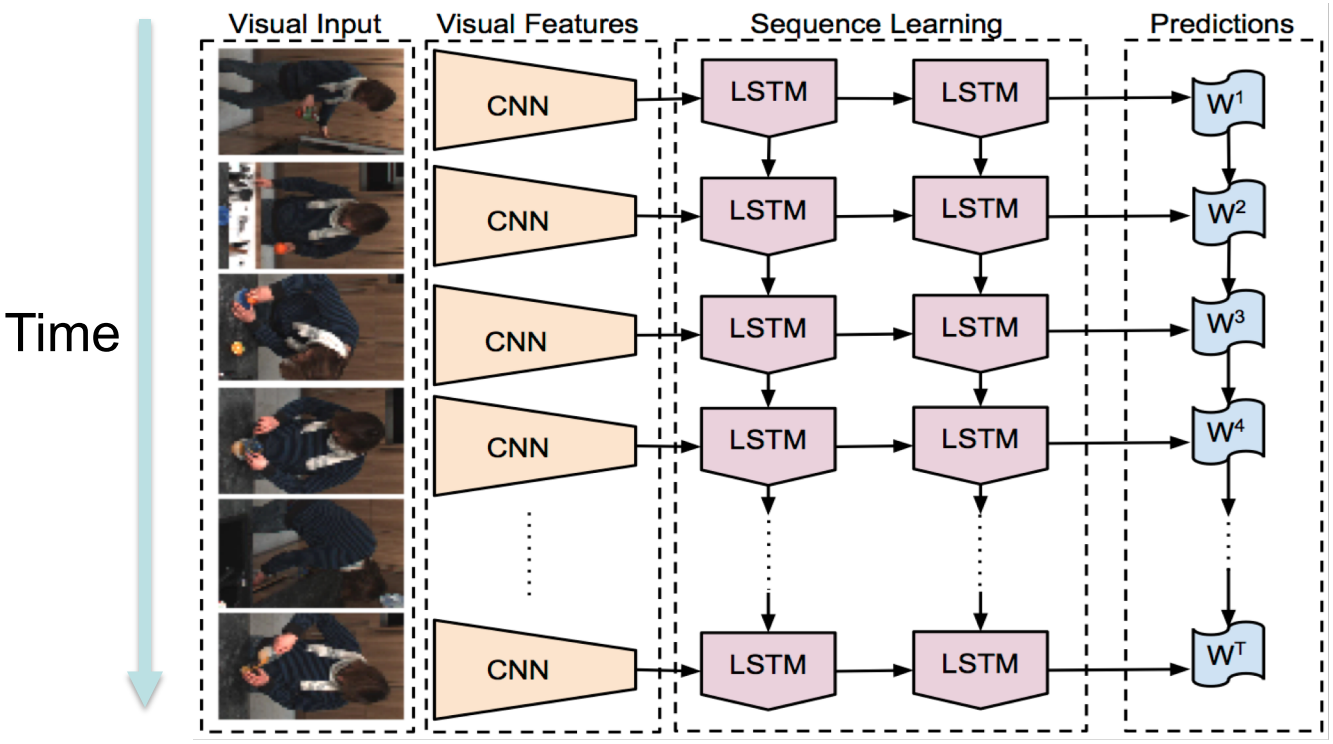
Similar actions → similar ranking parameters

Fernando et al. CVPR15

2D Appearance and temporal

LSTM models sequential memories in the long and short term

ConvNet-fc vectors as input, no spatial information encoded



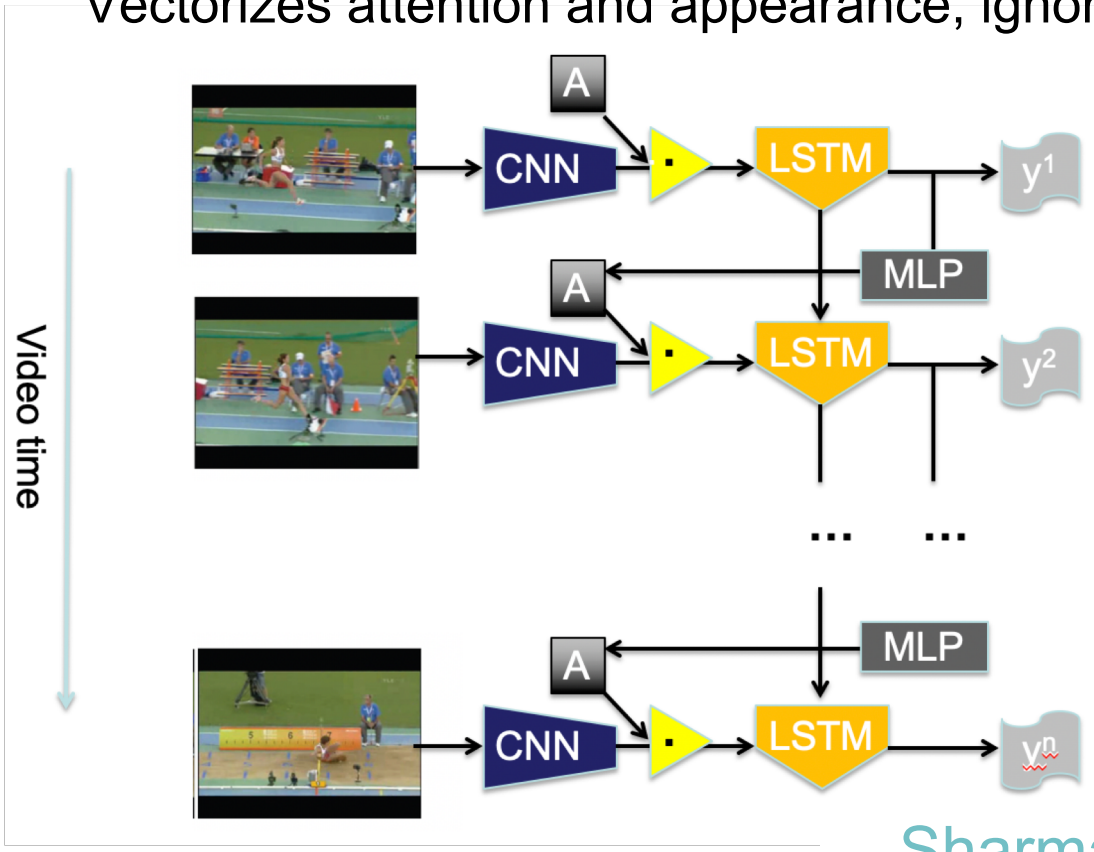
Baccouche et al. ICANN10 / Donahue et al. CVPR15 / Ng et al. CVPR15

2D Appearance, temporal, spatial

Look for best locations leading to correct action classification

Stays close to soft-Attention for image captioning [Xu et al. ICML15],

Vectorizes attention and appearance, ignores the motion inside a video.



Sharma et al. NIPS15 / ICLR16

3. VideoLSTM

This chapter presents VideoLSTM. An LSTM able to model spatiotemporal dynamics of videos by

- preserving 2D spatial structure of the frames over time
- adding motion-based attention
- enabling action localization from action class labels only

VideoLSTM convolves, attends and flows for action recognition.
Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees Snoek.
CVIU 2018

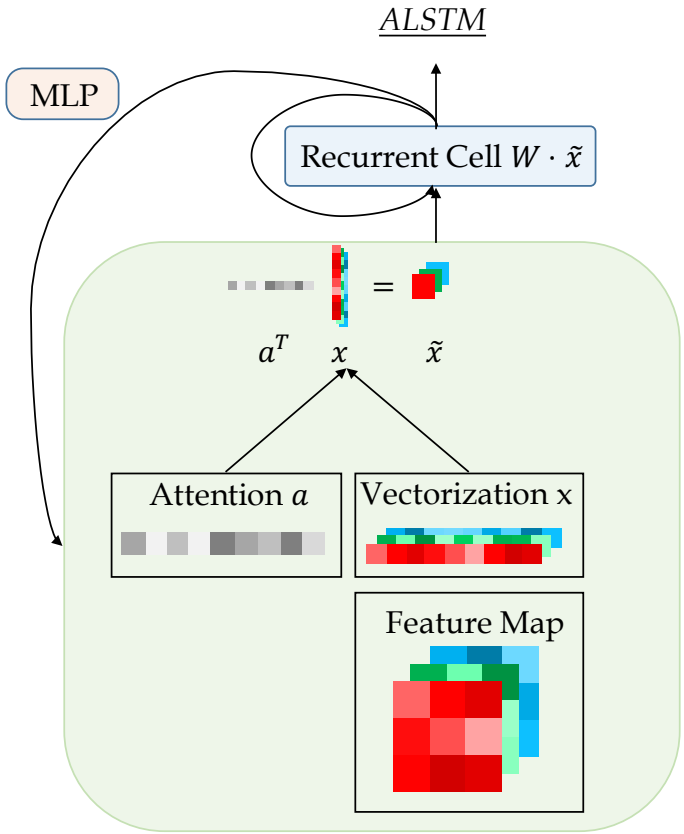
Convolutional (A)LSTM

Replace the fully connected multiplicative operations in an LSTM unit with convolutional operations

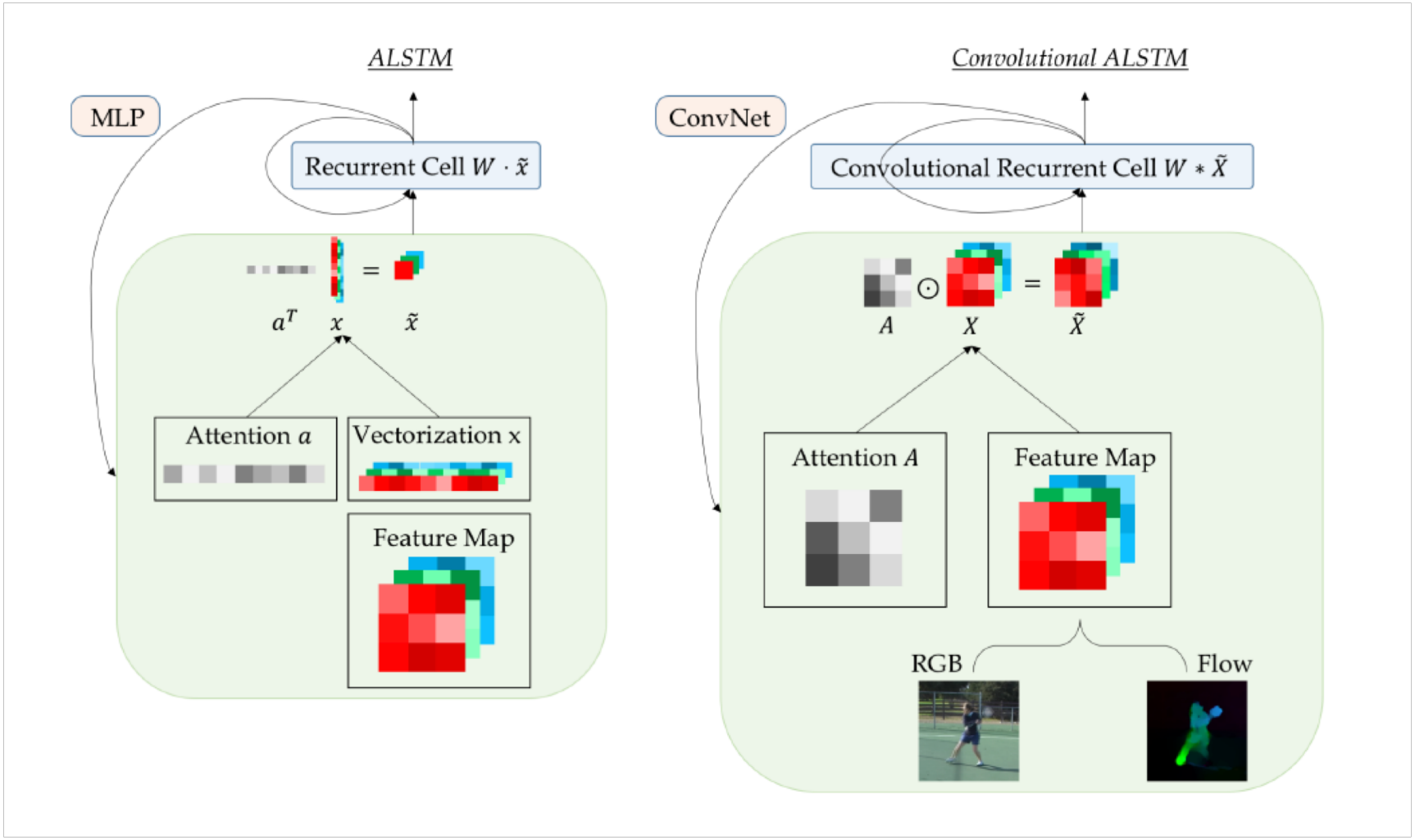
$$
\begin{aligned}
I_t &= \sigma(W_{xi} * \widetilde{X}_t + W_{hi} * H_{t-1} + b_i) \\
F_t &= \sigma(W_{xf} * \widetilde{X}_t + W_{hf} * H_{t-1} + b_f) \\
O_t &= \sigma(W_{xo} * \widetilde{X}_t + W_{ho} * H_{t-1} + b_o) \\
G_t &= \tanh(W_{xc} * \widetilde{X}_t + W_{hc} * H_{t-1} + b_c) \\
C_t &= F_t \odot C_{t-1} + I_t \odot G_t \\
H_t &= O_t \odot \tanh(C_t),
\end{aligned}
$$

Generate attention by shallow ConvNet instead of MLP

A(ttention)LSTM

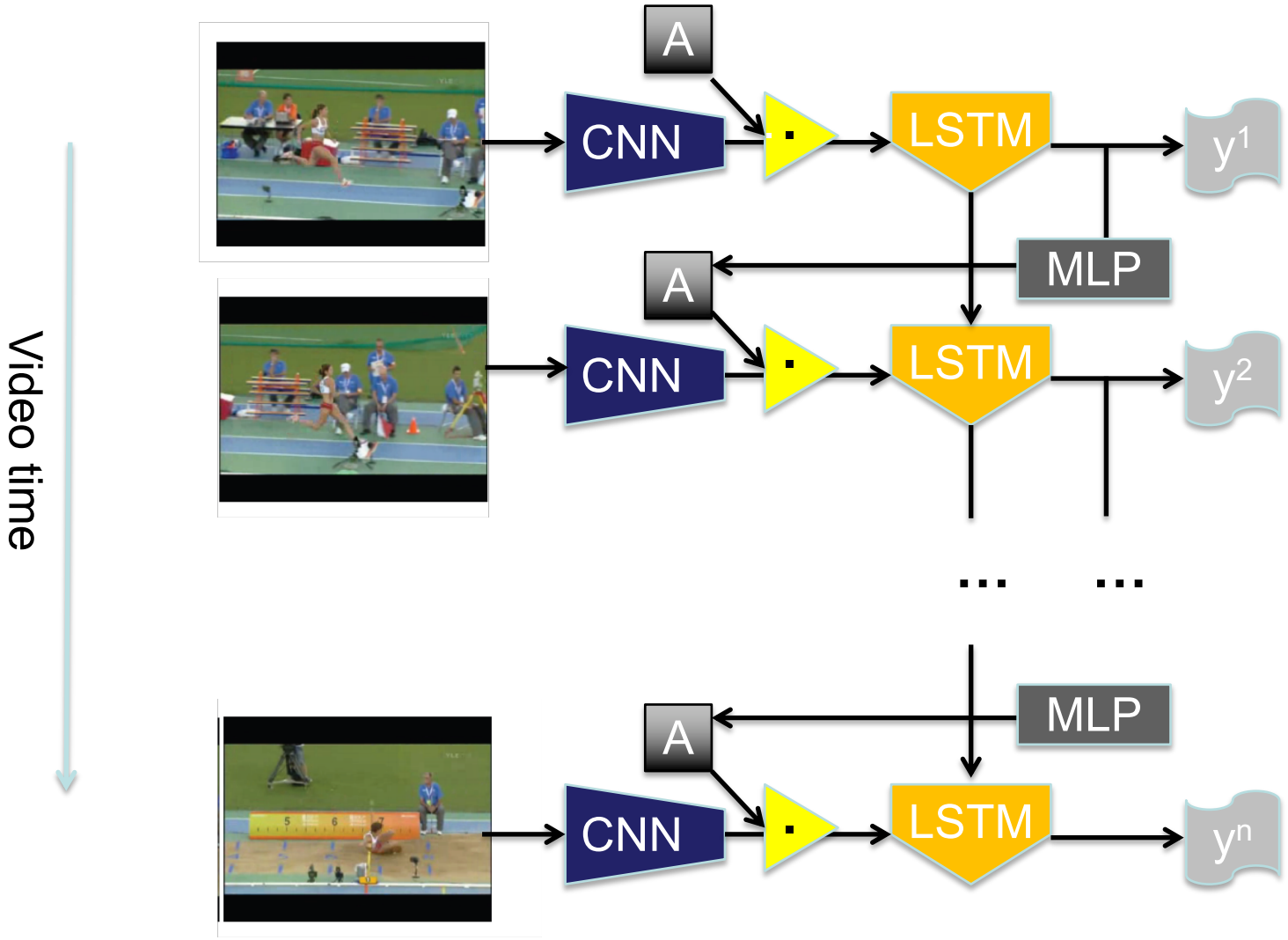


Convolutional ALSTM

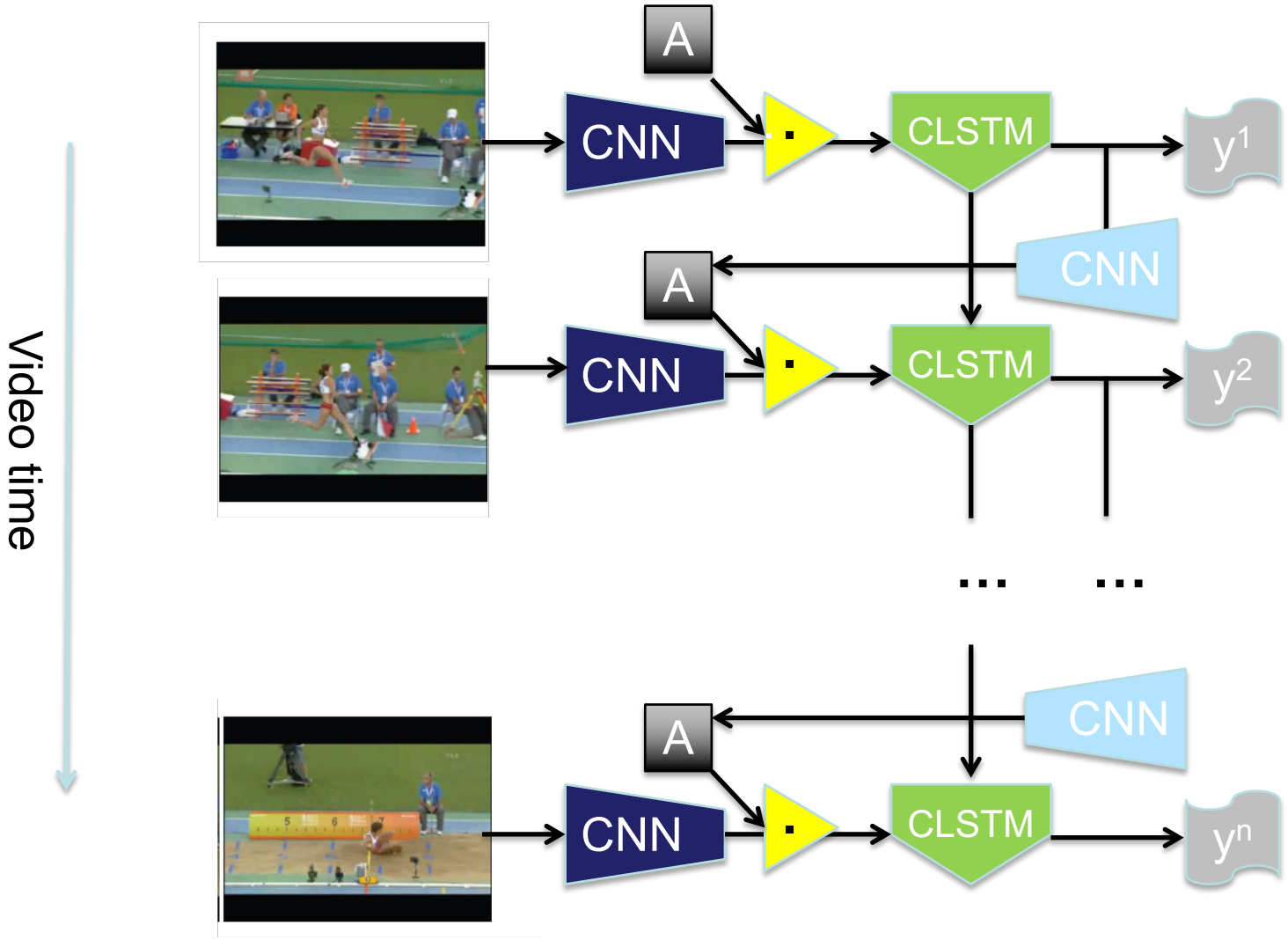


Convolutional ALSTM preserves spatial dimensions over time

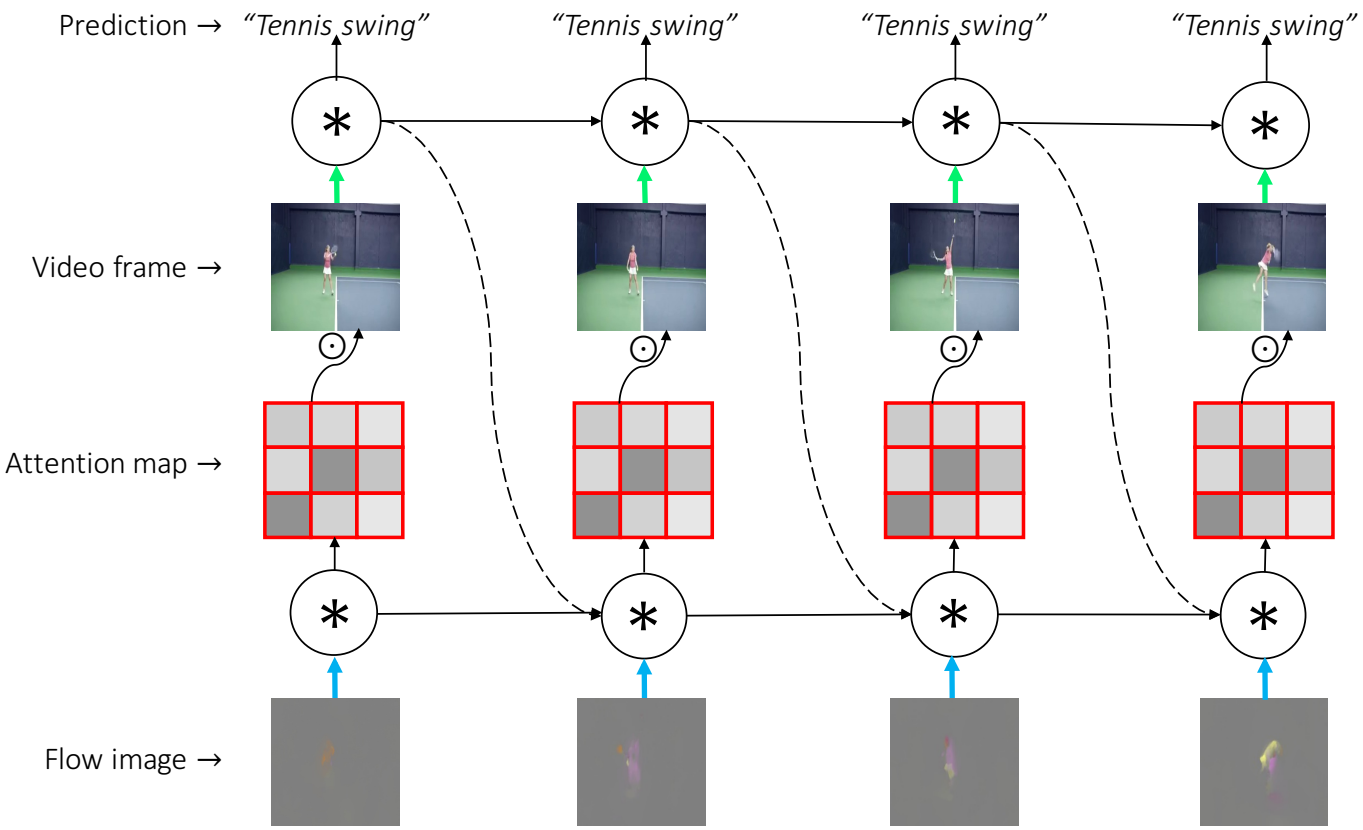
ALSTM



Convolutional ALSTM

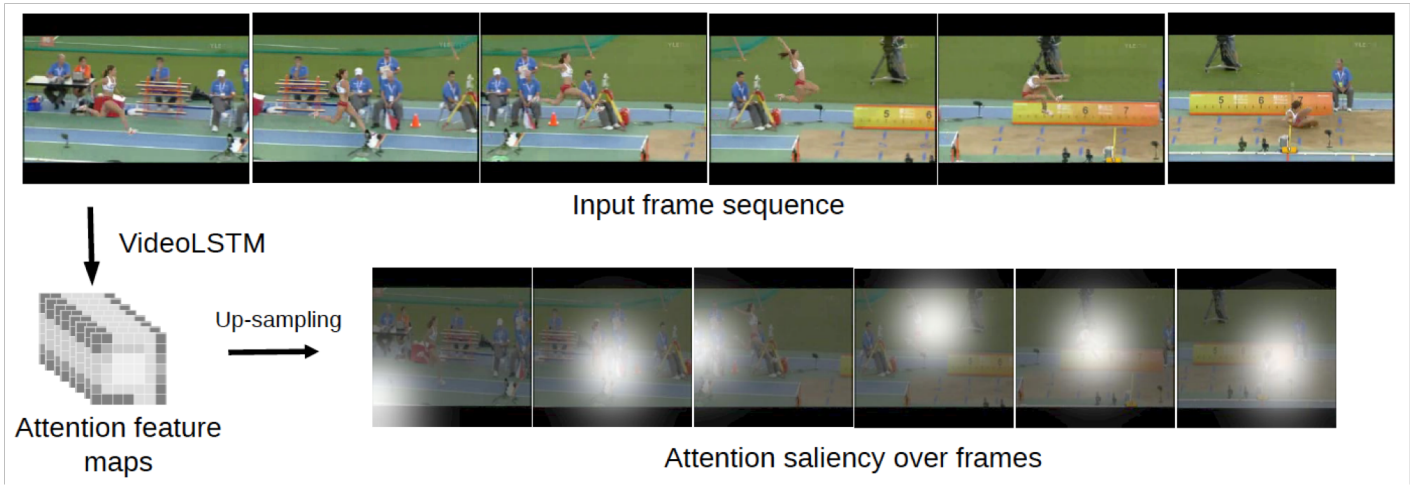


Motion-based attention

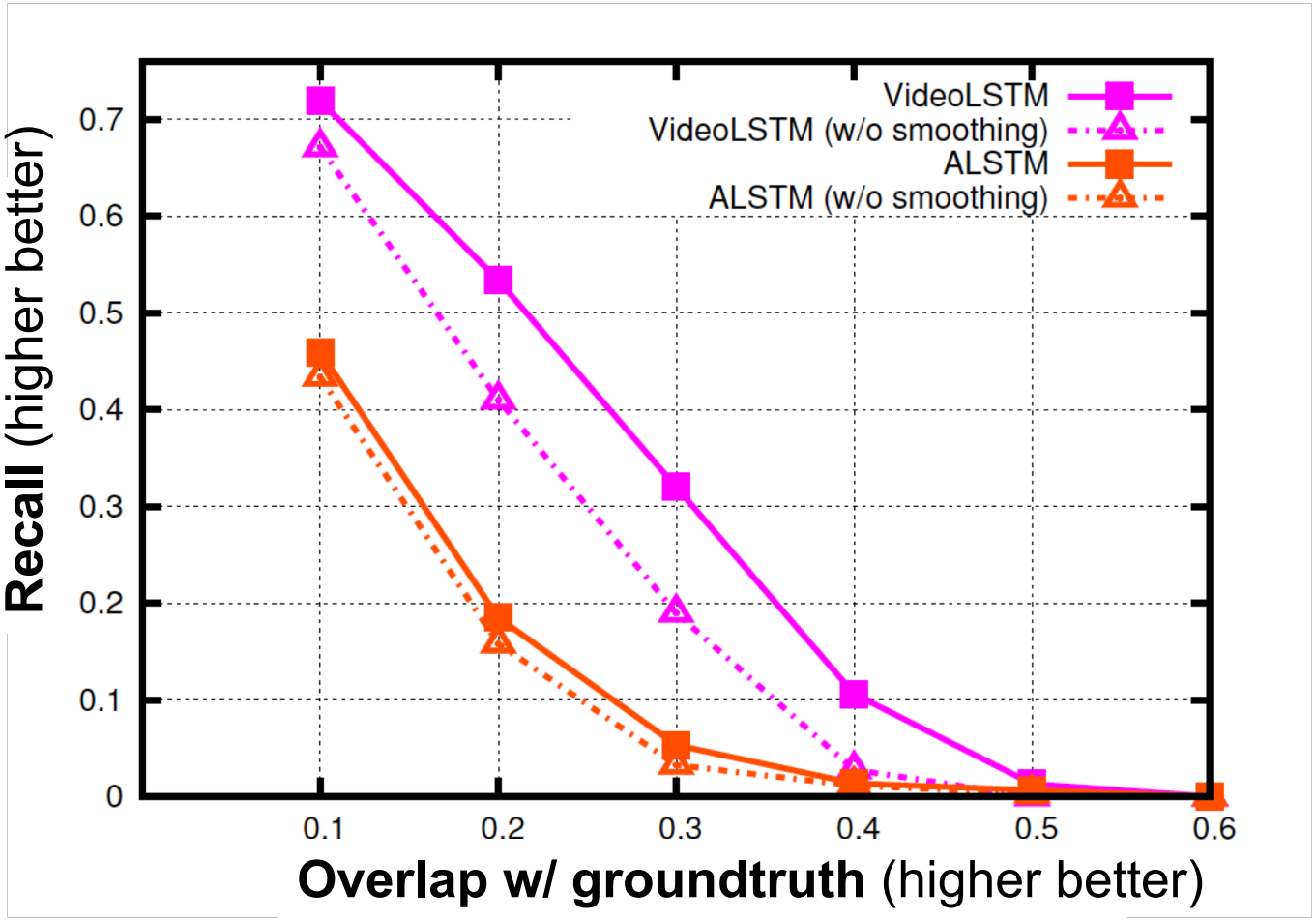


Motion information to infer the attention in each frame

Temporal smoothing

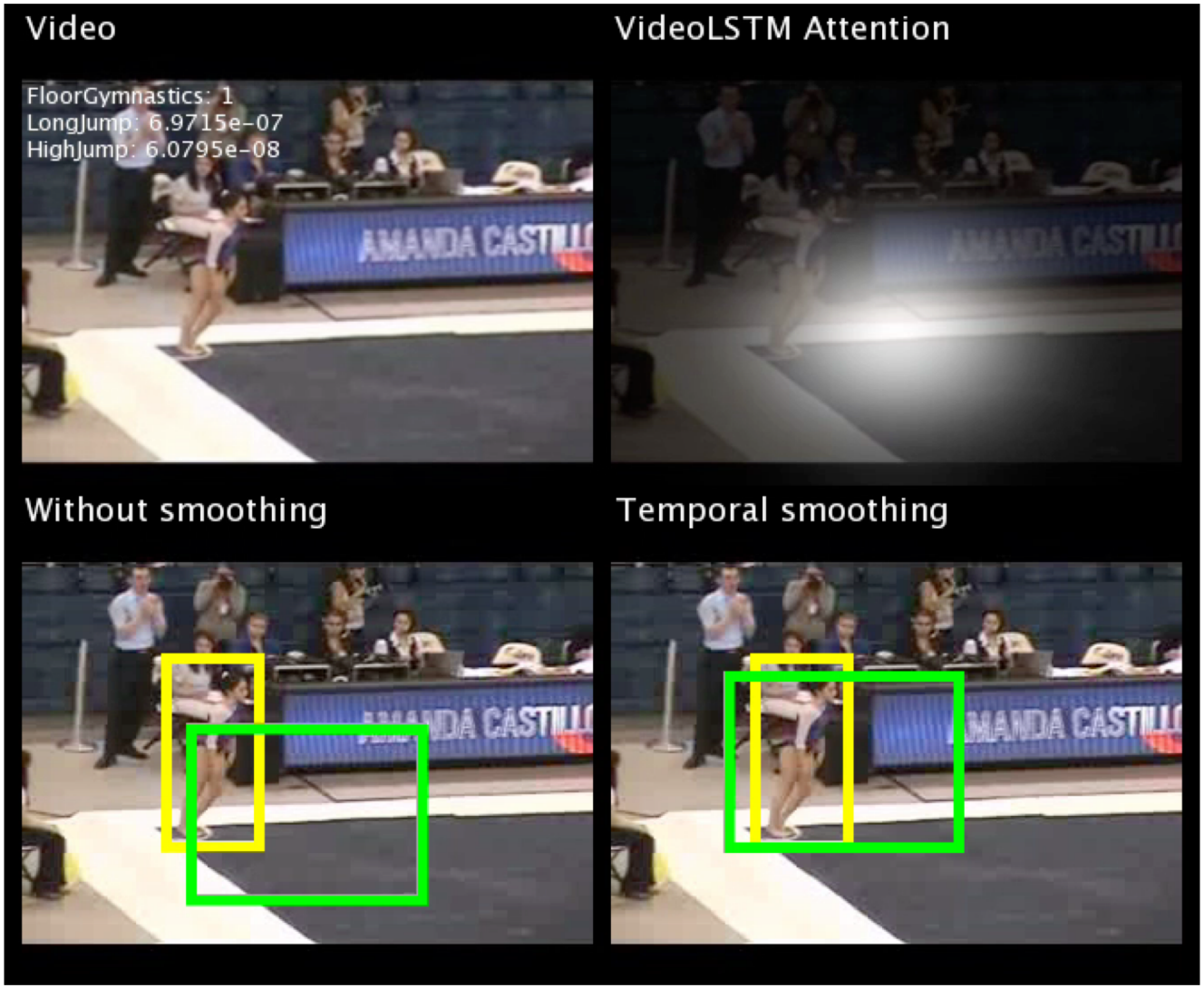


VideoLSTM vs ALSTM

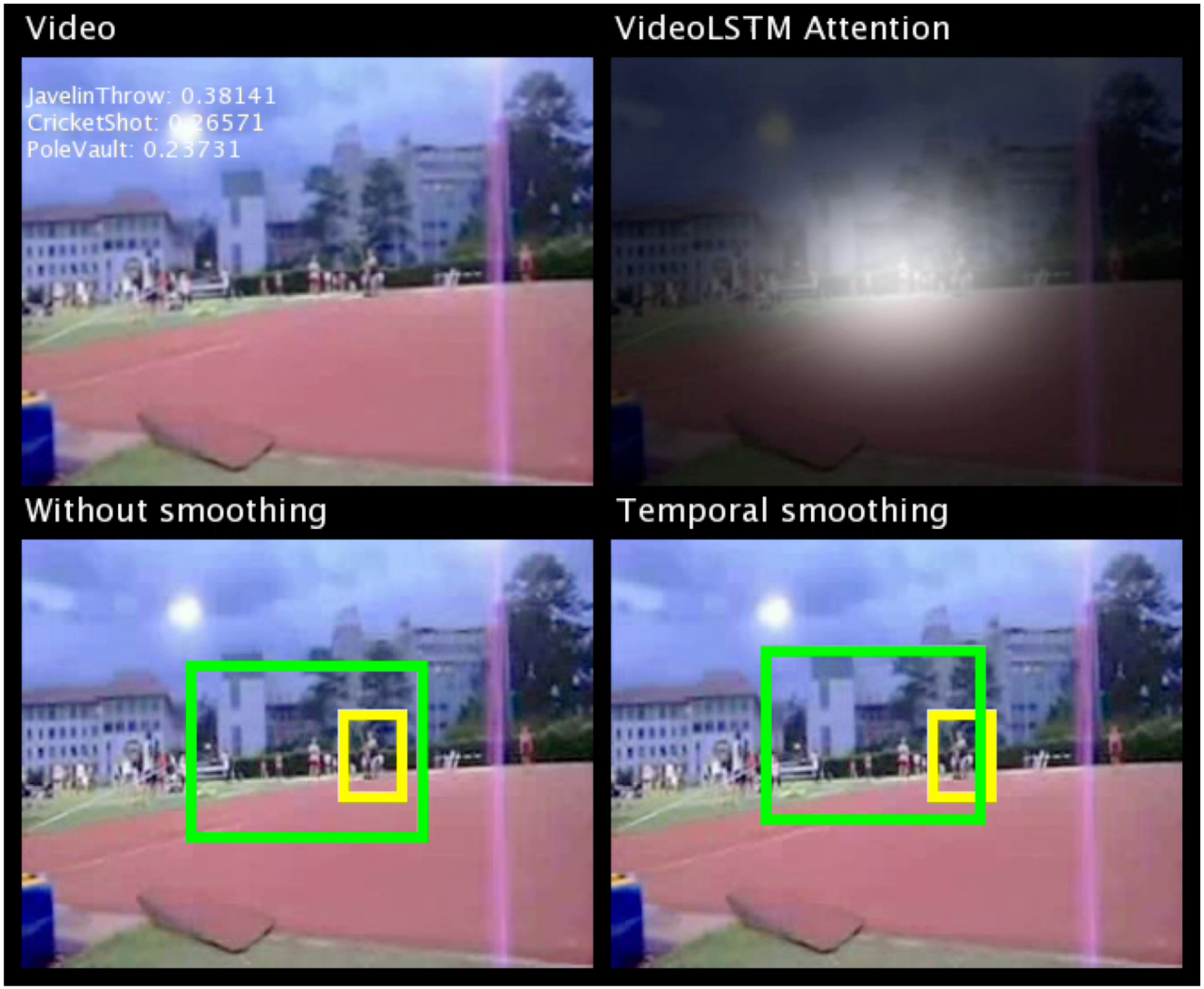


VideoLSTM localizes much better, temporal smoothing helps

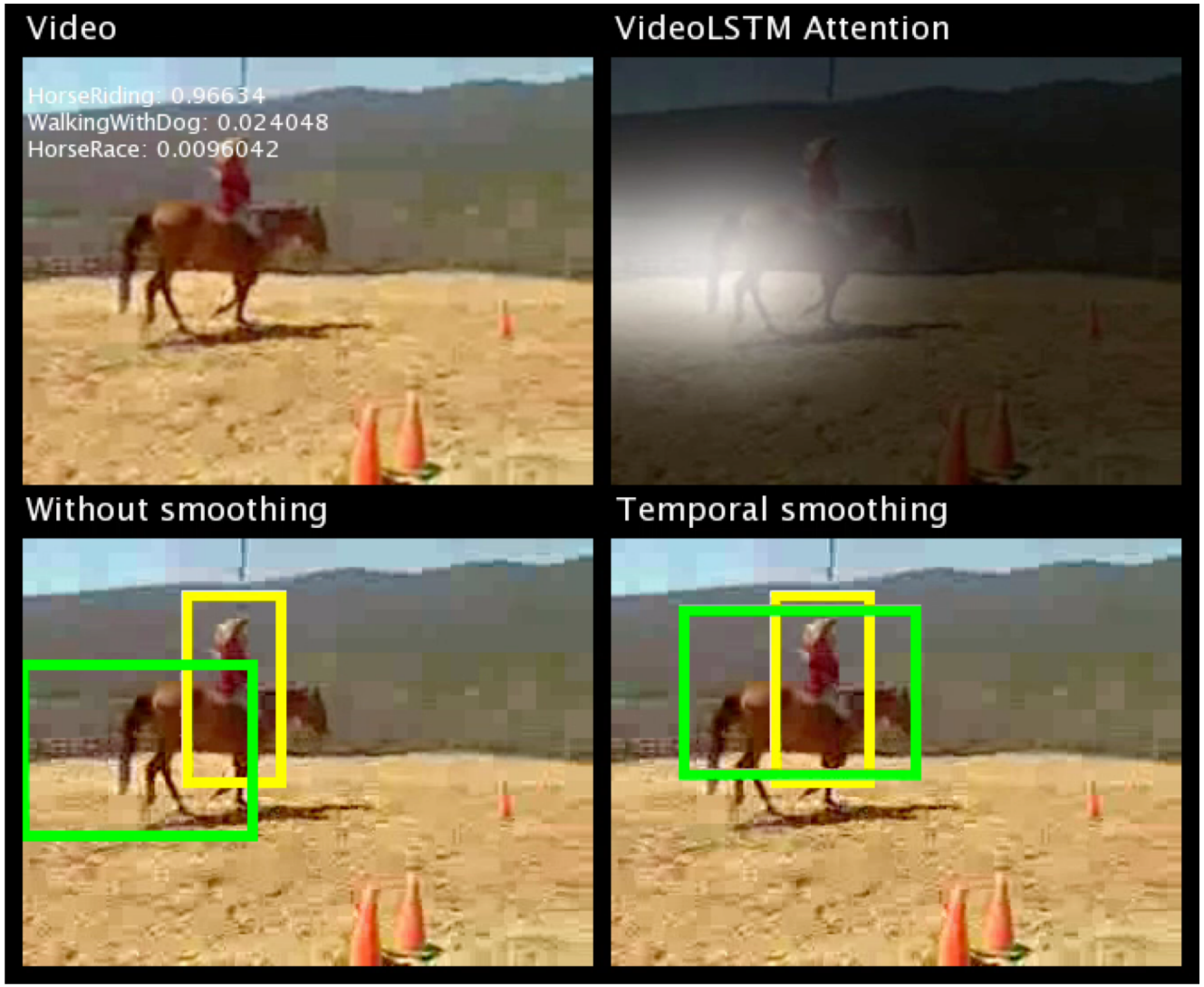
Qualitative results



Qualitative results



Qualitative results



4. Video Time

Time-aware encoding of frame sequences in a video is a fundamental problem in video understanding. While many attempted to model time in videos, an explicit study on quantifying video time is missing. We describe three properties of video time, and formulate tasks able to quantify the associated properties.

Ghodrati, Gavves & Snoek BMVC 18

Properties of Video Time

Temporal asymmetry

- There is a clear distinction between the forward and the backward arrow of time

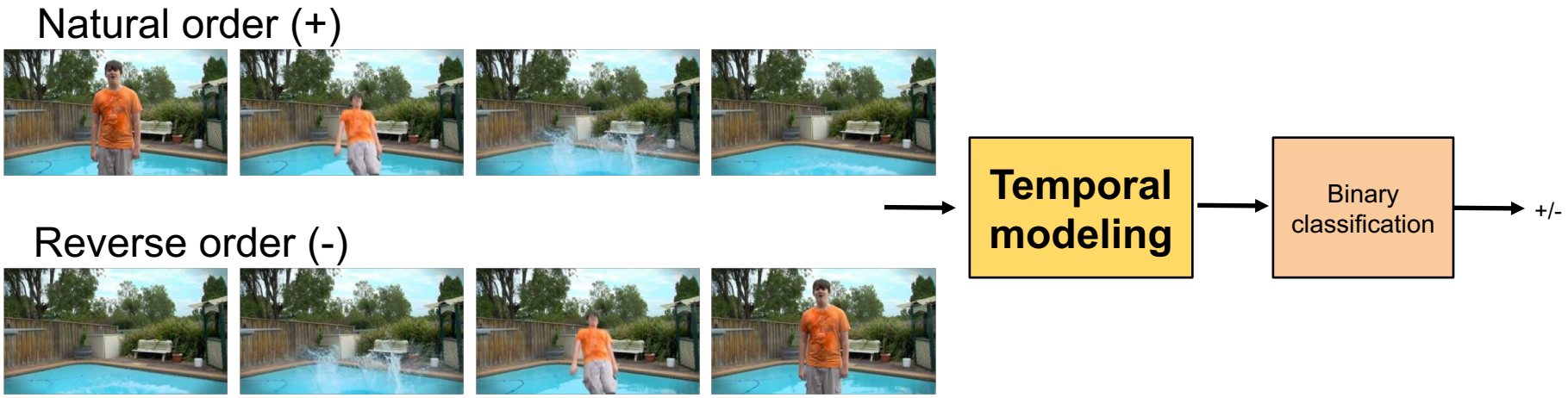
Forward

Backward

Arrow of time prediction

The task of distinguishing natural order and reverse order of frames, a binary classification task.

A conceptual, ultimate temporal task: Output is defined by order of the input, regardless of the input representation.



Properties of Video Time

Temporal asymmetry

- There is a clear distinction between the forward and the backward arrow of time.

Temporal continuity

- Future observations are expected to be a smooth continuation of past observations

Continuous

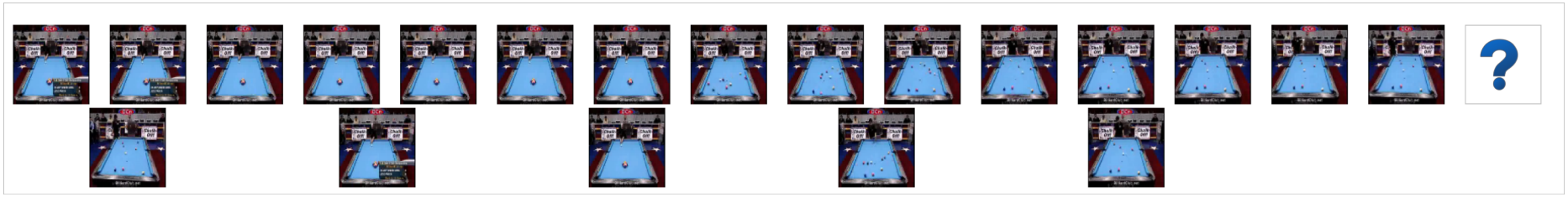
Non-continuous:

swapping two frames

Future frame selection

In this task, we sample *N* frames from a video where the first *N-1* frames are given but the last frame is missing.

The goal is to select the correct future frame, given a list of *K* choices from the same video.



Properties of Video Time

Temporal asymmetry

- There is a clear distinction between the forward and the backward arrow of time.

Temporal continuity

- Future observations are expected to be a smooth continuation of past observations

Temporal causality

- When we observe an event, we observe a chain of causes and then effects

action: Pretending to put something into something

ordered

shuffled

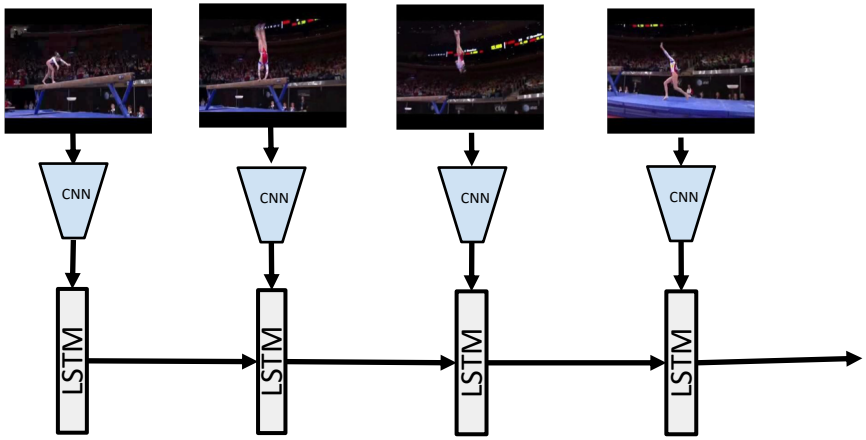
Action template classification

The task of action classification where classes are formulated as template-based textual descriptions e.g. Putting something into something

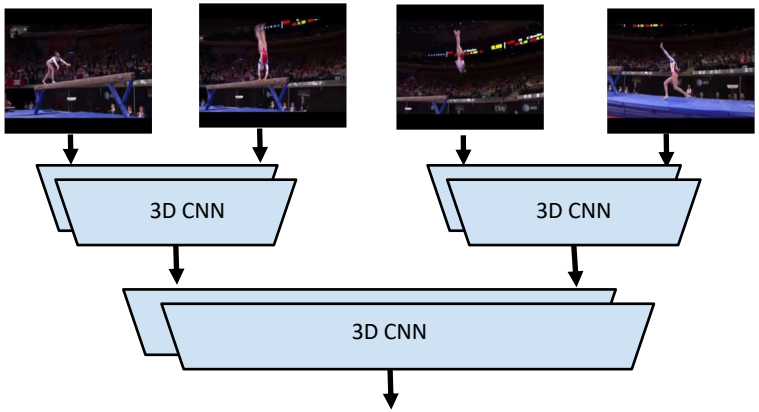
- ☐ **Putting something into something**
- ☑ **Pretending to put something into something**
- ☐ **Holding something behind something**

Modern models

LSTMs learn transitions between subsequent states

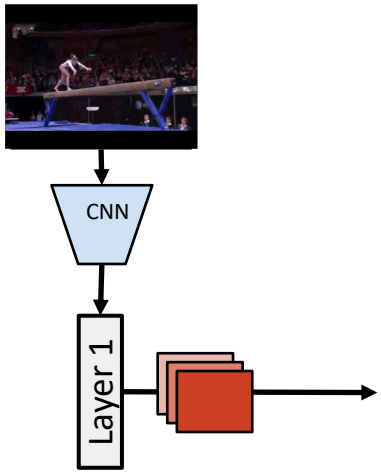


3D convolutions learn spatiotemporal patterns within a video



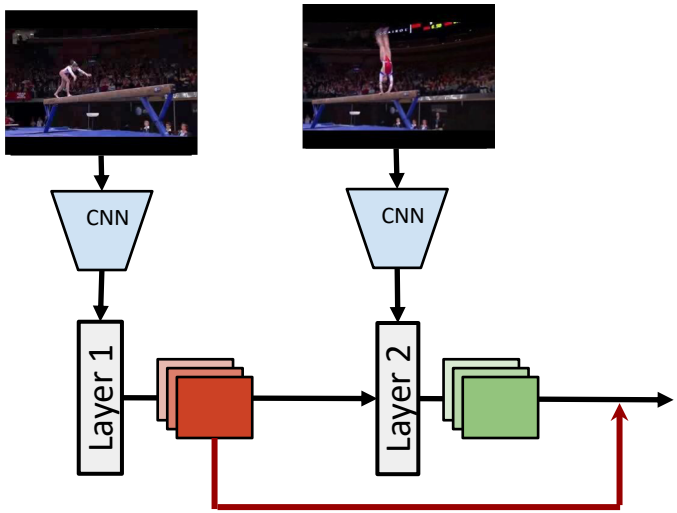
Proposed model

Time-aligned dense net



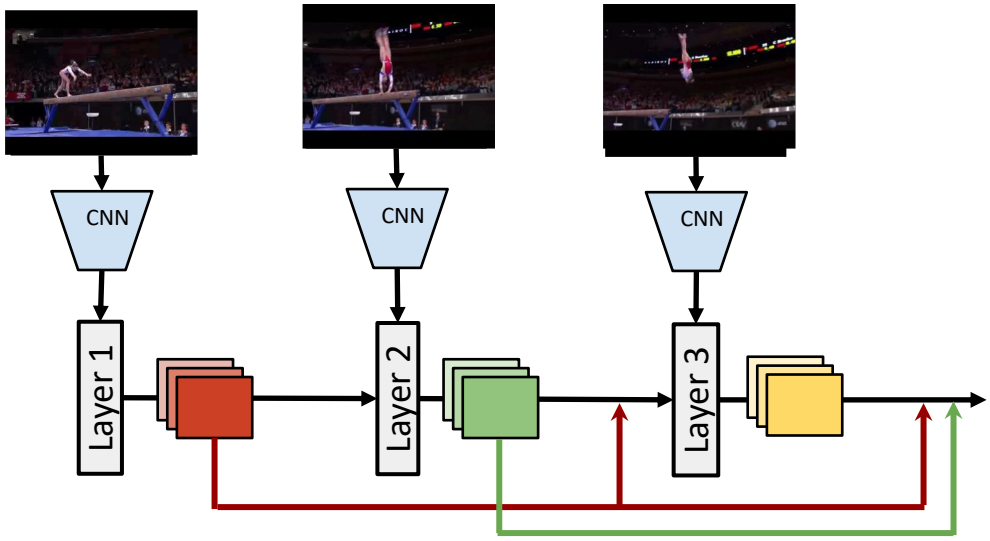
Proposed model

Time-aligned dense net



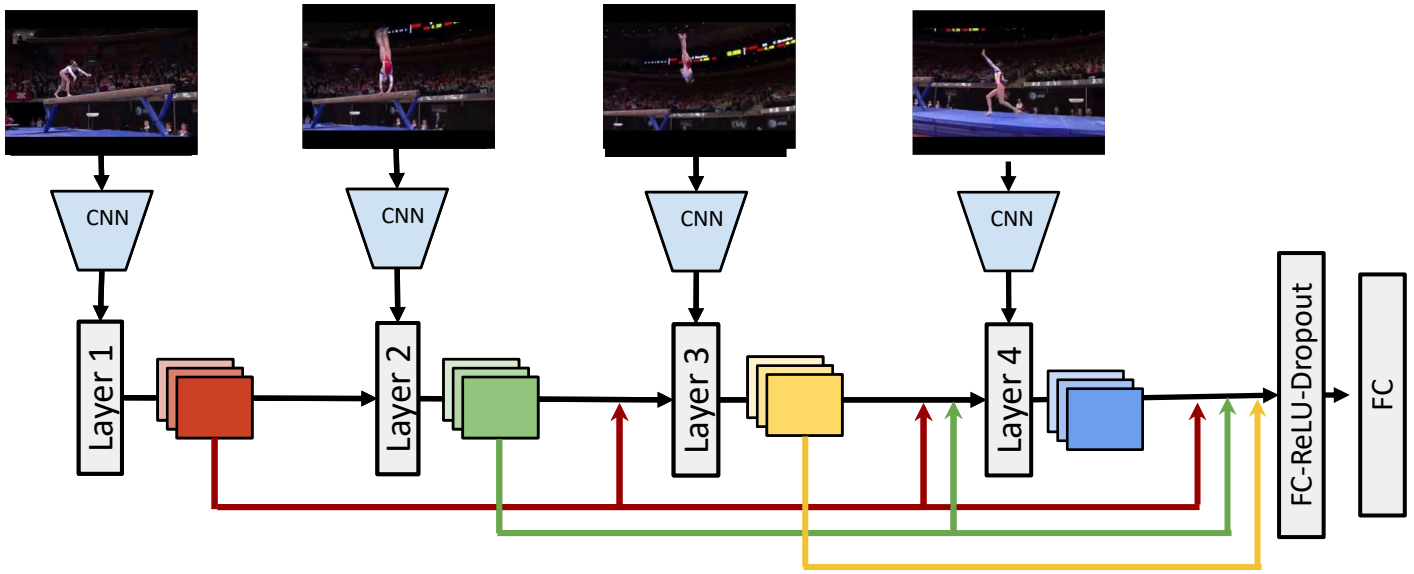
Proposed model

Time-aligned dense net



Proposed model

Time-aligned dense net

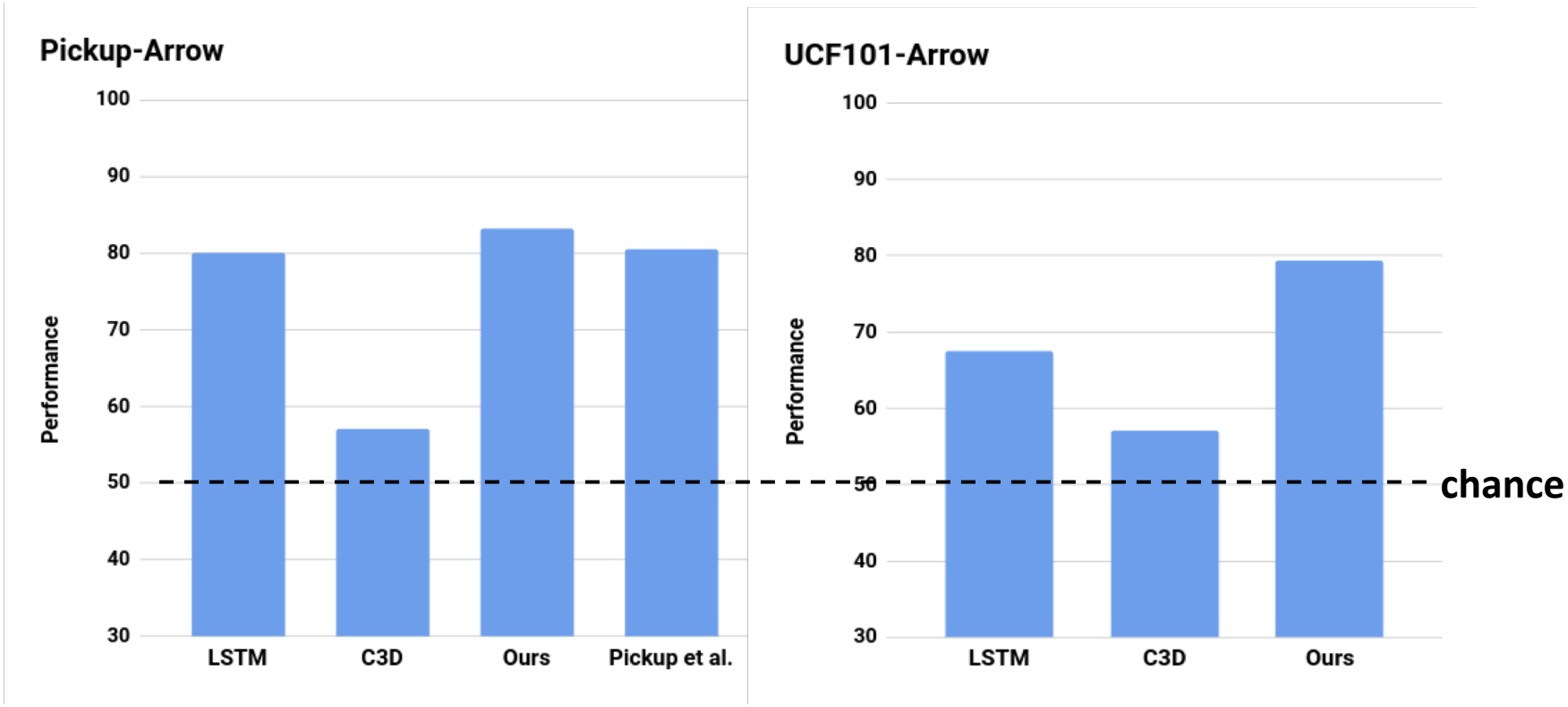


Arrow of time prediction - Datasets

Pickup et al.

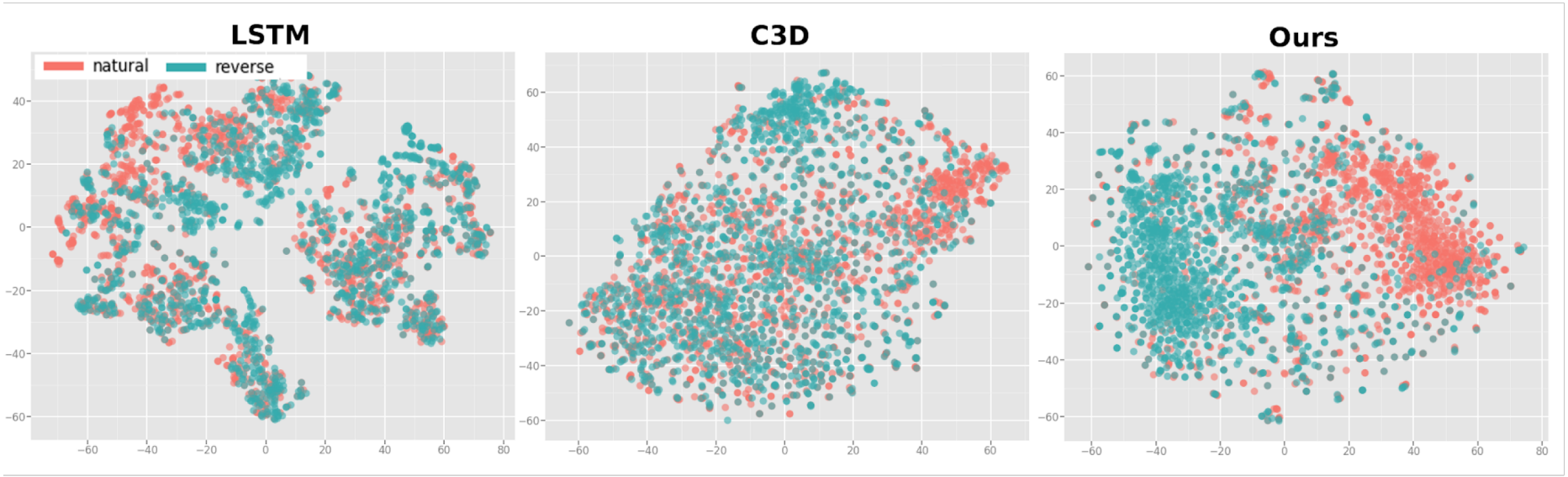
UCF 101

Arrow of time prediction - Results

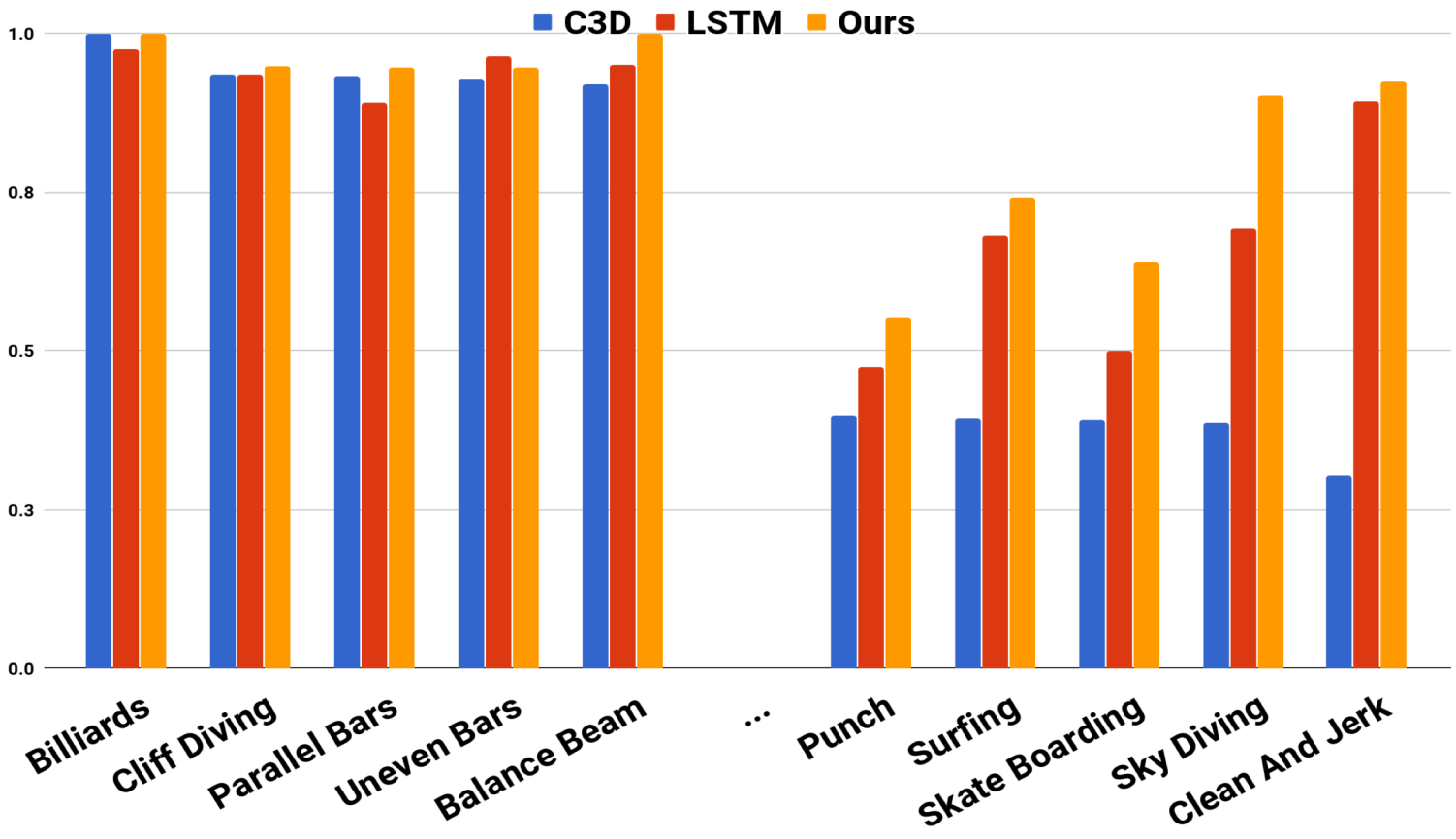


Sequential encoders like LSTM and ours better suited than C3D.

Arrow of time prediction - tSNE



Arrow of time prediction – per class

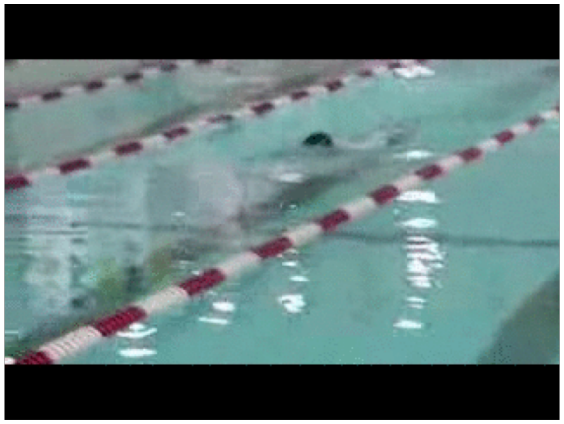
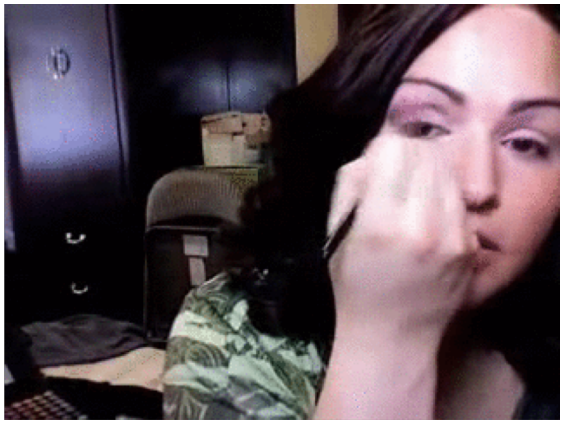


Temporally causal classes easier?

Future frame selection - datasets

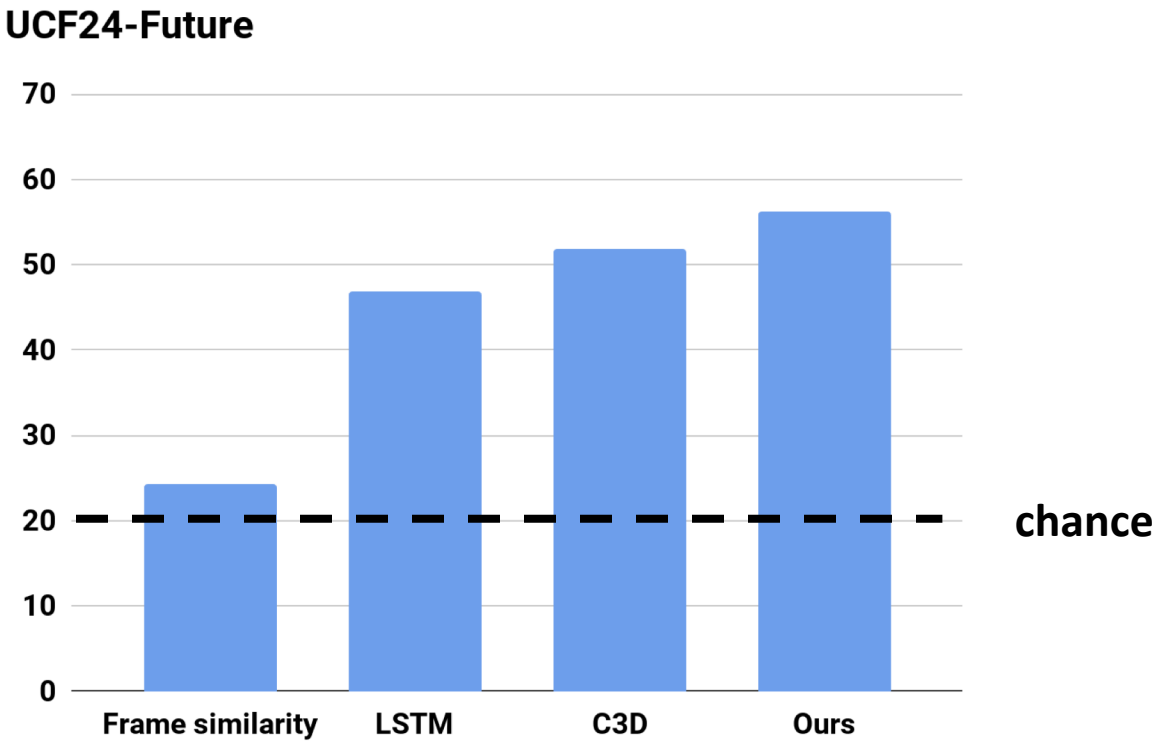
We divide the actions into three categories:

- **Clear arrow** → actions with visible arrow of time e.g. bowling, diving, billiards, … (24 classes)
- **Semi-clear arrow** → actions with semi-visible arrow of time e.g. archery, military march, … (17 classes)
- **unclear arrow** → actions that arrow of time is not visible e.g. haircut, play flute, … (60 classes)

Future frame selection - results

Average over predictions from {0.4, 0.8, 1.3, 1.7, 2.5, 3.3, 4.2, 5.0, 5.8, 6.7, 7.5, 8.3} seconds ahead



Learning both spatial and temporal dependencies are necessary

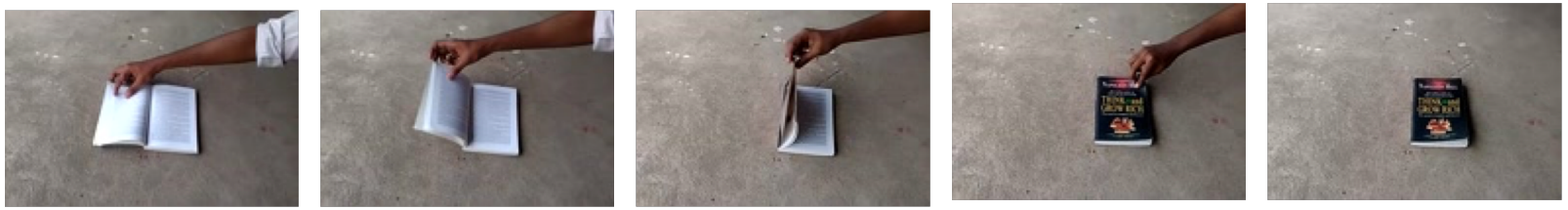
Action Template Classification - Dataset

Something-something dataset

174 template classes

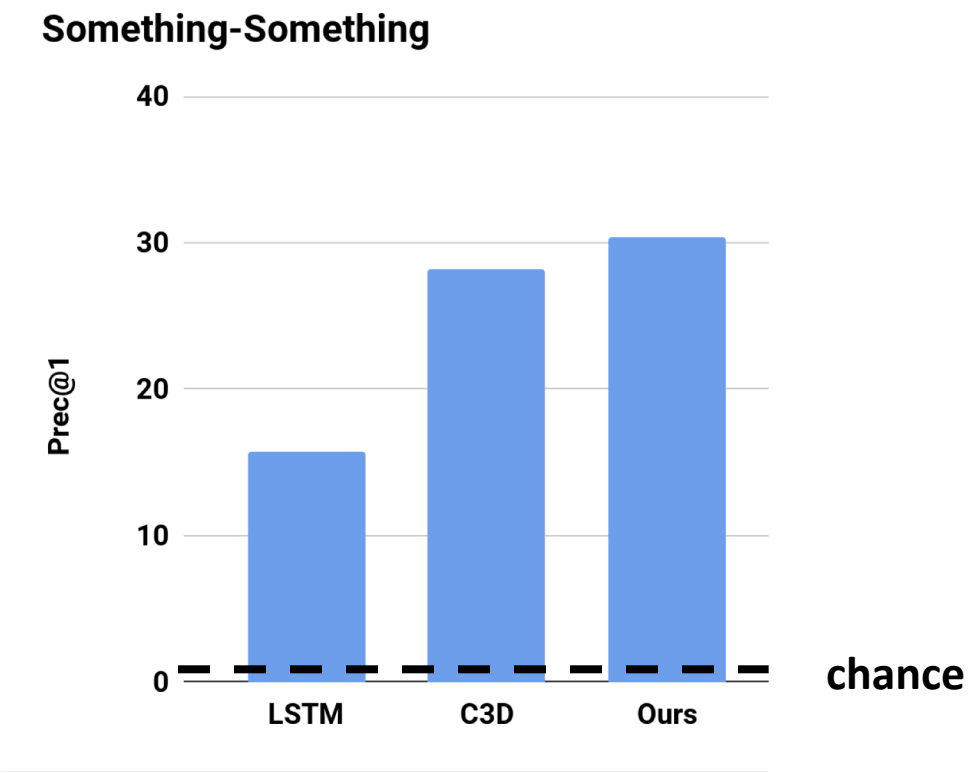
Pretending to close something without closing it

closing something



Goyal et al. ICCV 17

Action Template Classification - Results



Model should be able to parameterize the temporal conditional dependencies per time step freely

Conclusions

Best video representations for actions are learned

- Having sufficient examples available not obvious for actions
- Things become worse for localized actions

VideoLSTM hardwires convolutions inside attention LSTM

- Derives attention from what moves in video
- Localization from a video-level action class label only

Video time representation open challenge

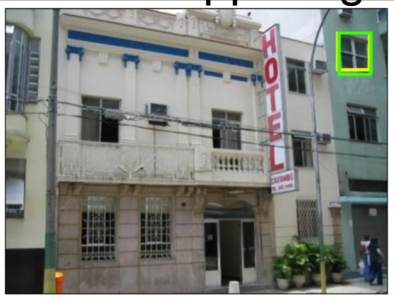
5. Video & Language

The common tactic to spatiotemporal video understanding is to track a human-specified box or to learn a deep classification network from a set of predefined action classes. In this Chapter we will present an alternative approach, that allows for spatiotemporal recognition from a natural language sentence as input, and show its potential for object tracking and action segmentation.

Image understanding from sentence

Find object location in image based on language

“window upper right”



“bottom left window”

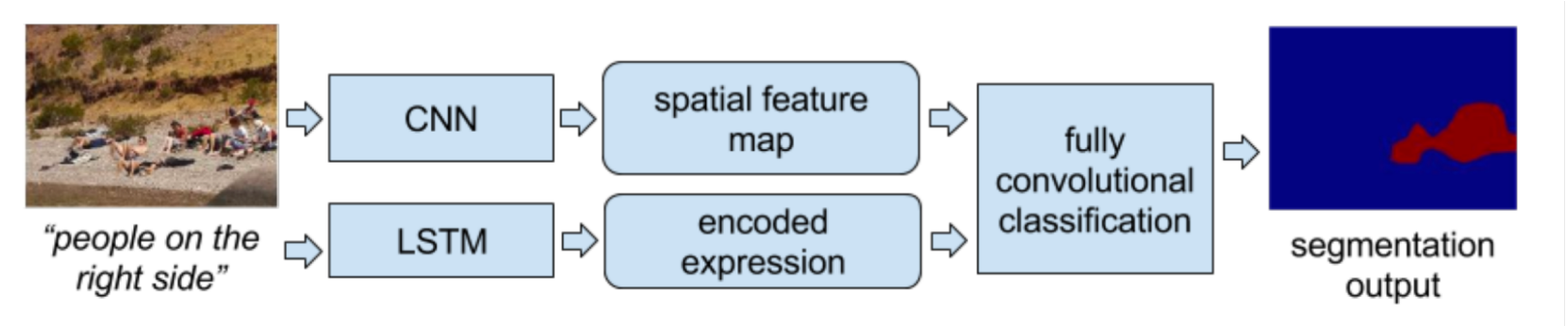
ground truth
prediction

Hu *et al.* CVPR 2016

Object segmentation from a sentence

Image embedding – spatial feature map through CNN

Sentence embedding – final hidden state in LSTM



Fully convolutional classification – match input sentence to every location on the spatial grid and up-sample

Hu *et al.* ECCV 2016

I. Tracking


Zhenyang Li


Ran Tao

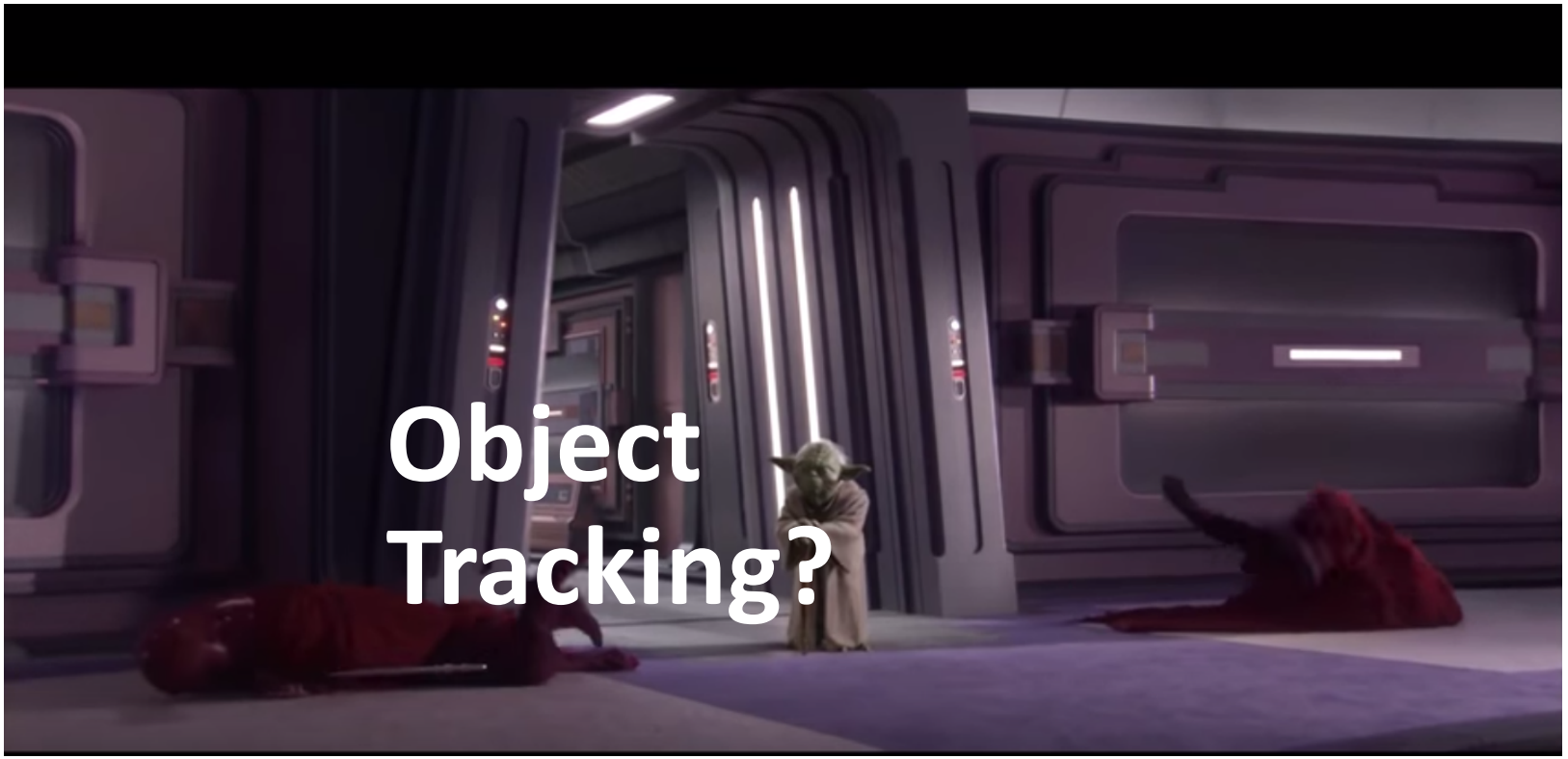

Efstratios Gavves


Cees Snoek

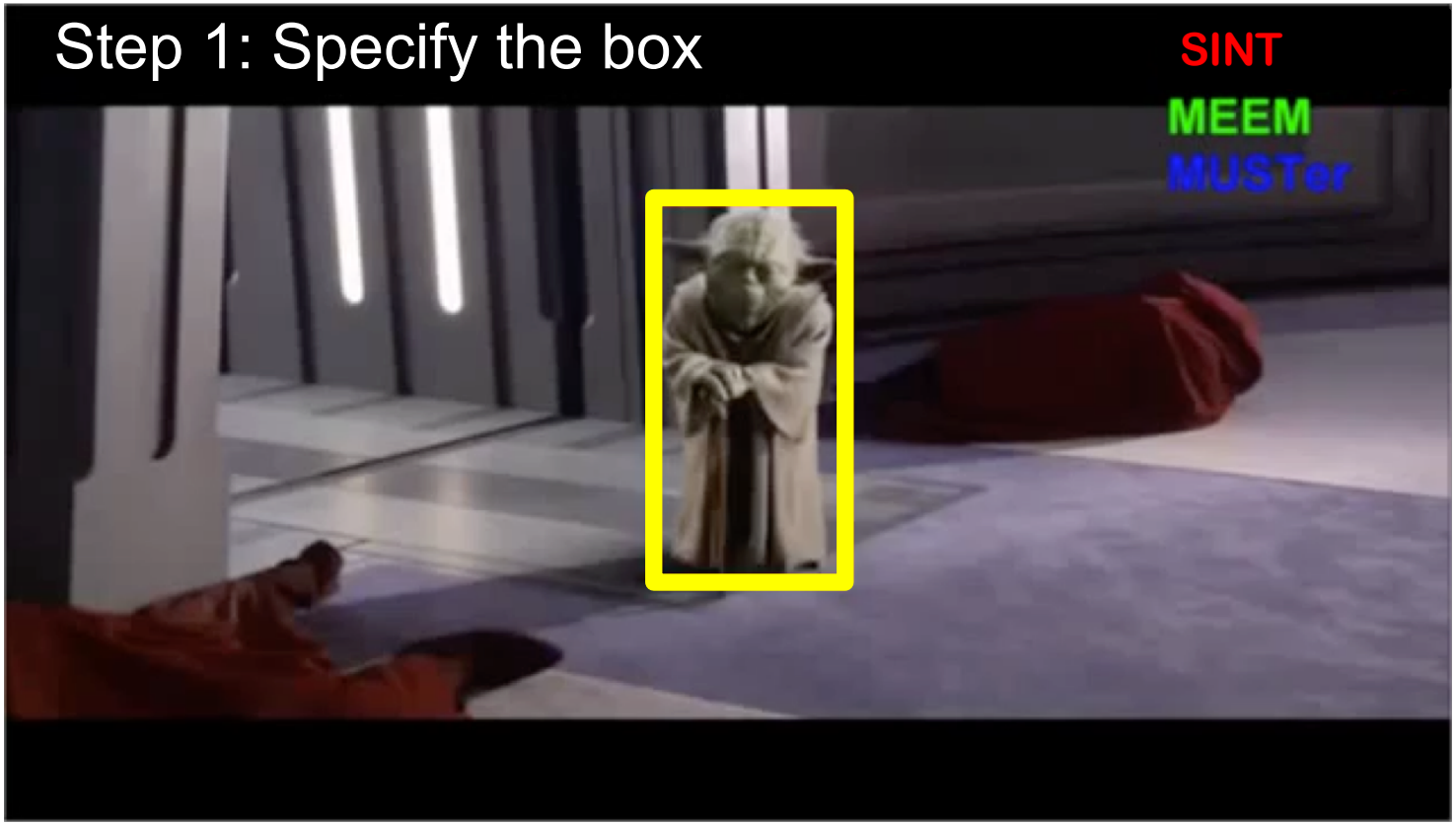

Arnold Smeulders

Tracking by Natural Language Specification. In *CVPR* 2017.

A long standing computer vision challenge



Conventional wisdom




Step 2: Start tracking
SINT
MEEM
MUSTer

Key contribution

Specify the target by language instead of box

“Track the little green person with the pointy ears and the beige robe”

Benefits of language

Tracking objects in multiple videos simultaneously

No ‘first-frame’ requirement, live monitoring across streams

“Man with blue pants”

Challenges

How to obtain a tight box around an object from text?

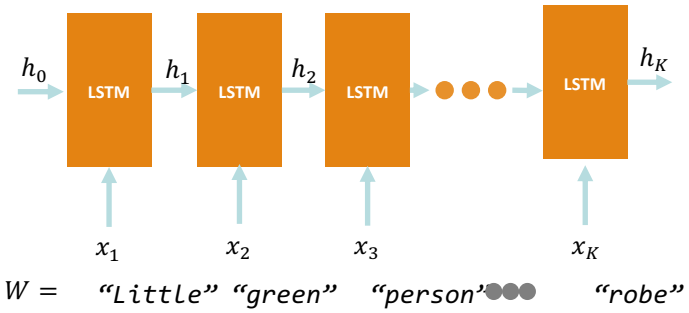
Text ambiguity *vs* object variance *vs* object invariance?

What happens if the description is no longer valid?

Model I: Lingual specification only

LSTM encodes the text query

$$s_t = LSTM(W) = h_K$$



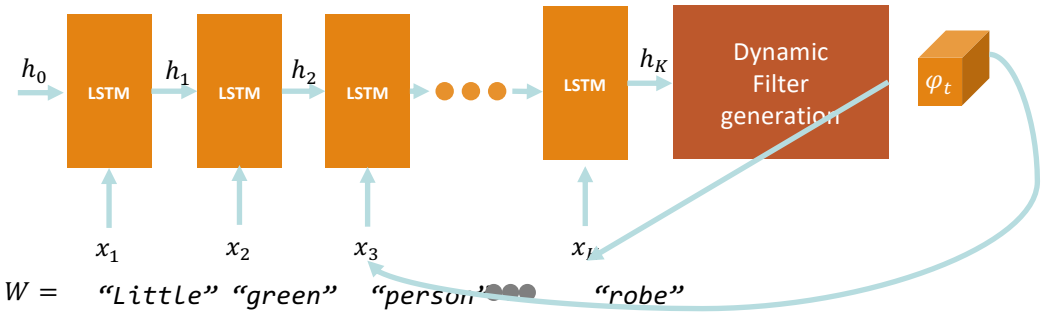
Model I: Lingual specification only

LSTM encodes the text query

$$s_t = LSTM(W) = h_K$$

Dynamically generate filters from LSTM output

$$v_t^{language} = \sigma(W_v s_t + b_v)$$



Model I: Lingual specification only

LSTM encodes the text query

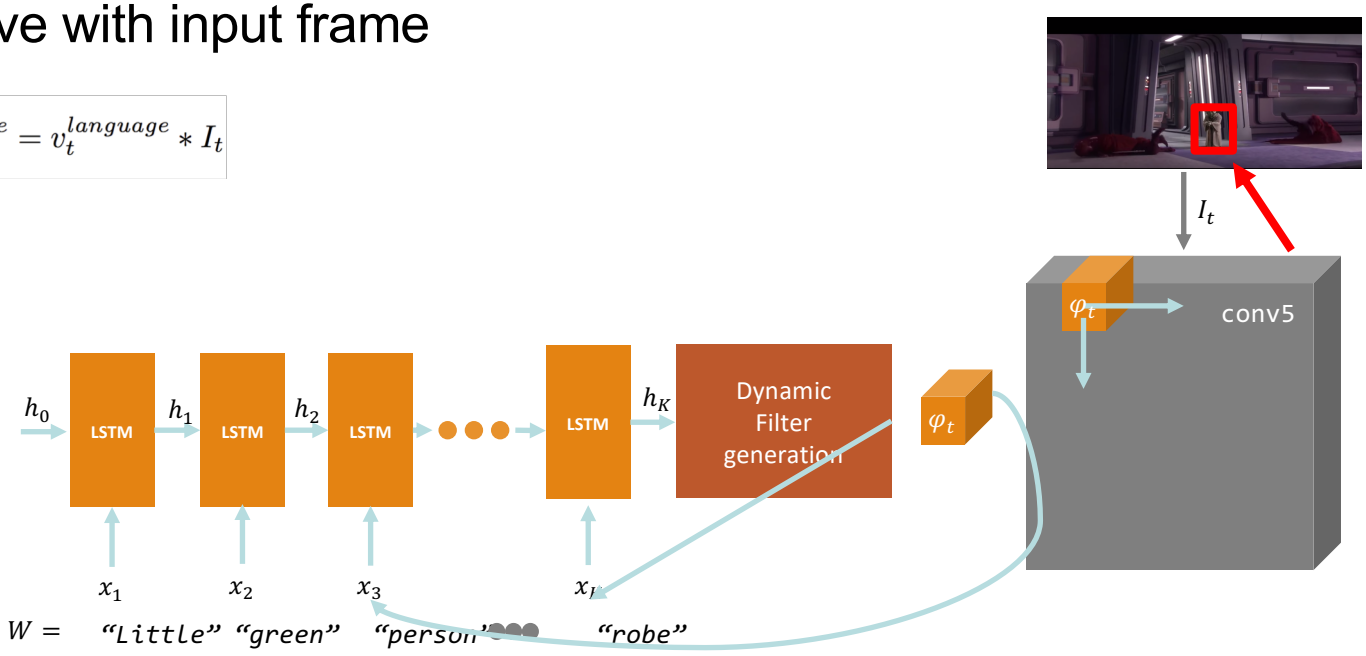
$$s_t = LSTM(W) = h_K$$

Dynamically generate filters from LSTM output

$$v_t^{language} = \sigma(W_v s_t + b_v)$$

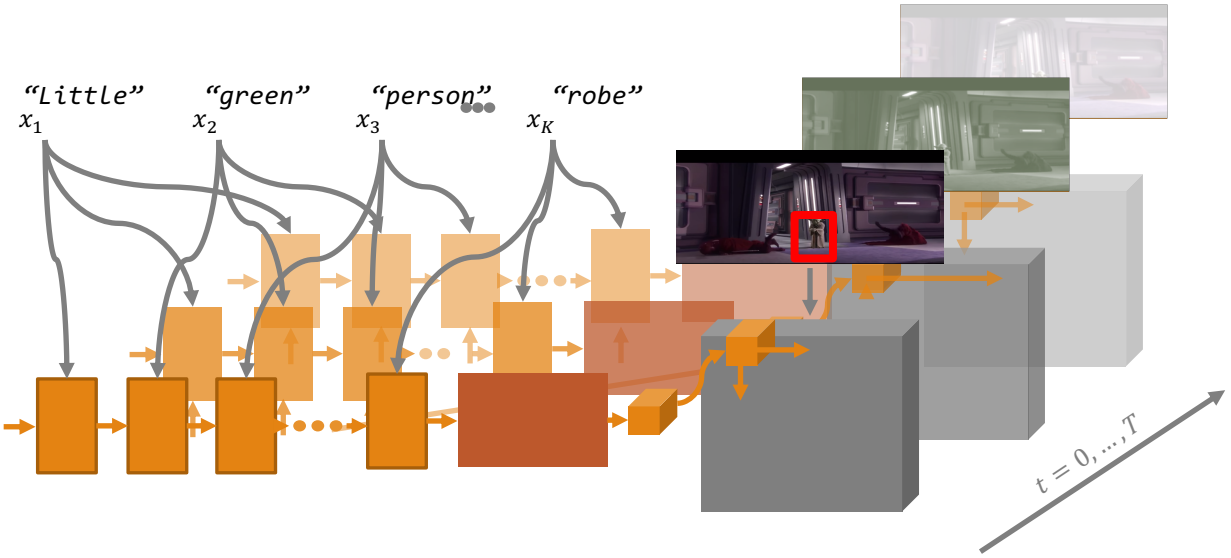
Convolve with input frame

$$A_t^{language} = v_t^{language} * I_t$$



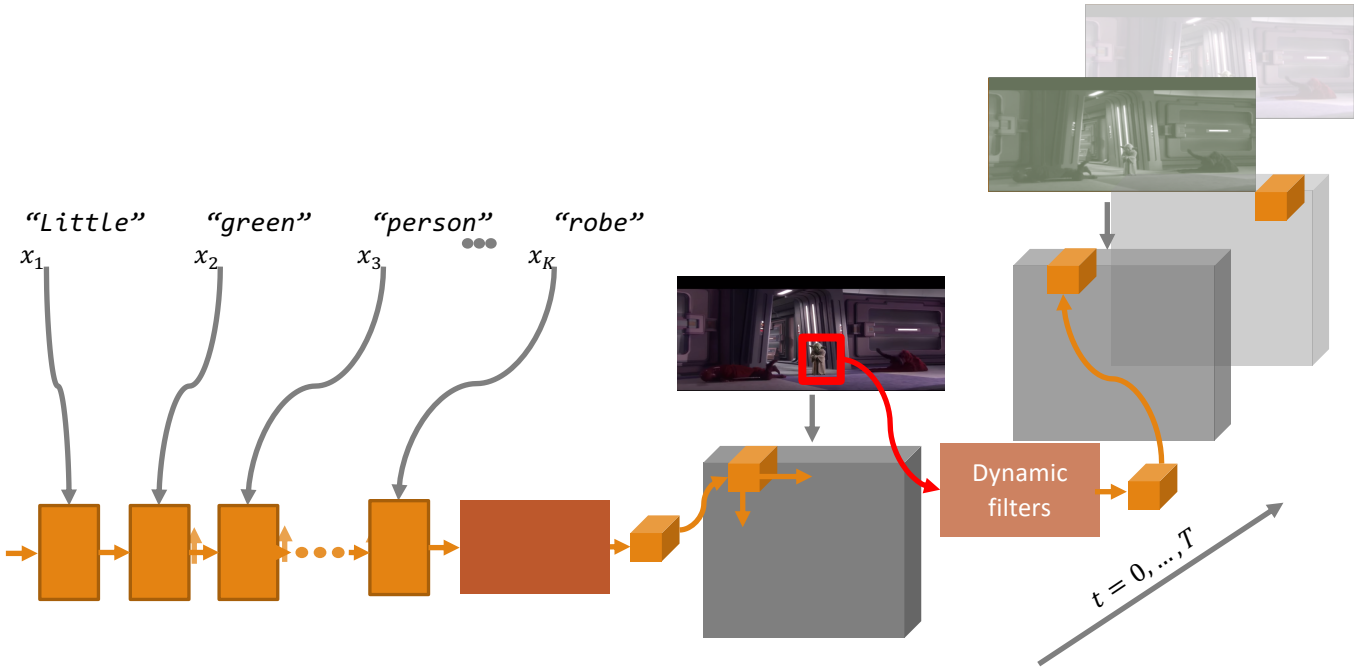
Model I: Lingual specification only

Tracking by repeated ‘detection’



Model II: Lingual first, then visual

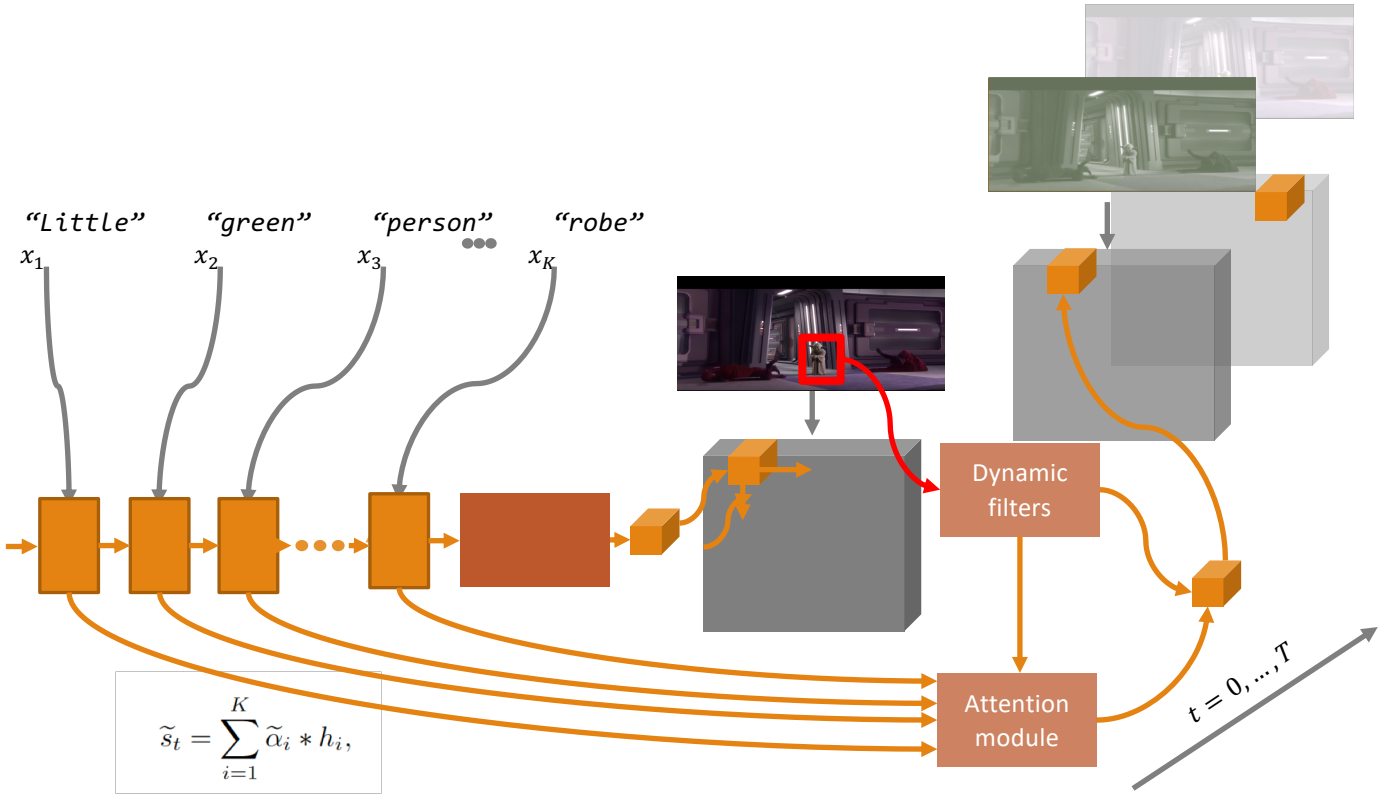
Use Model I for initialization, then track



Tao *et al.* CVPR 2016

Model III: Lingual & visual

Adapts the lingual specification over time



Augment tracking video with sentences

Tracking datasets: OTB100 and ImageNet Videos

Augment tracking video with sentences

Tracking datasets: OTB100 and ImageNet Videos

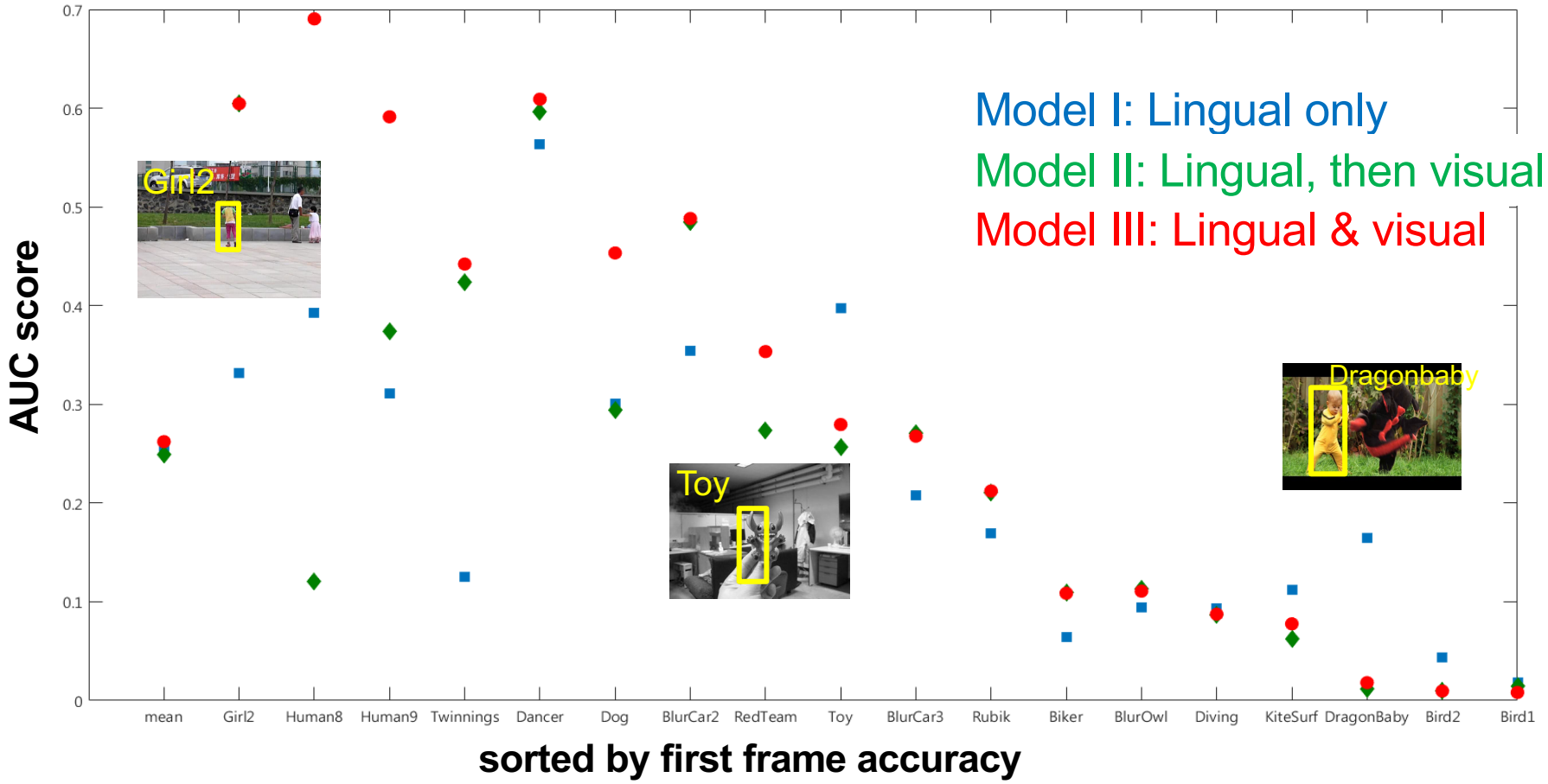
We add language description about the target in first frame

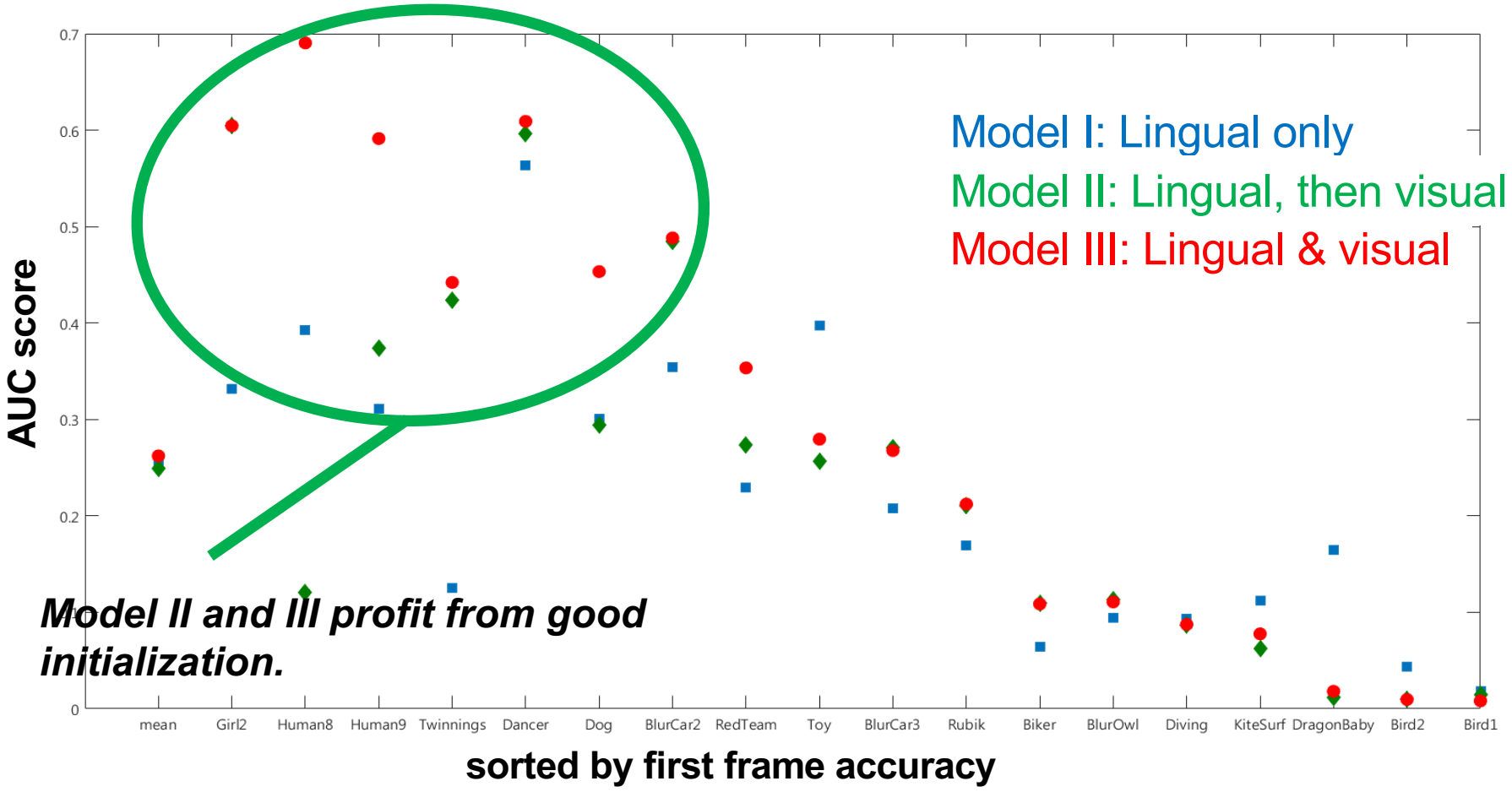
"*left singer in white dress*"

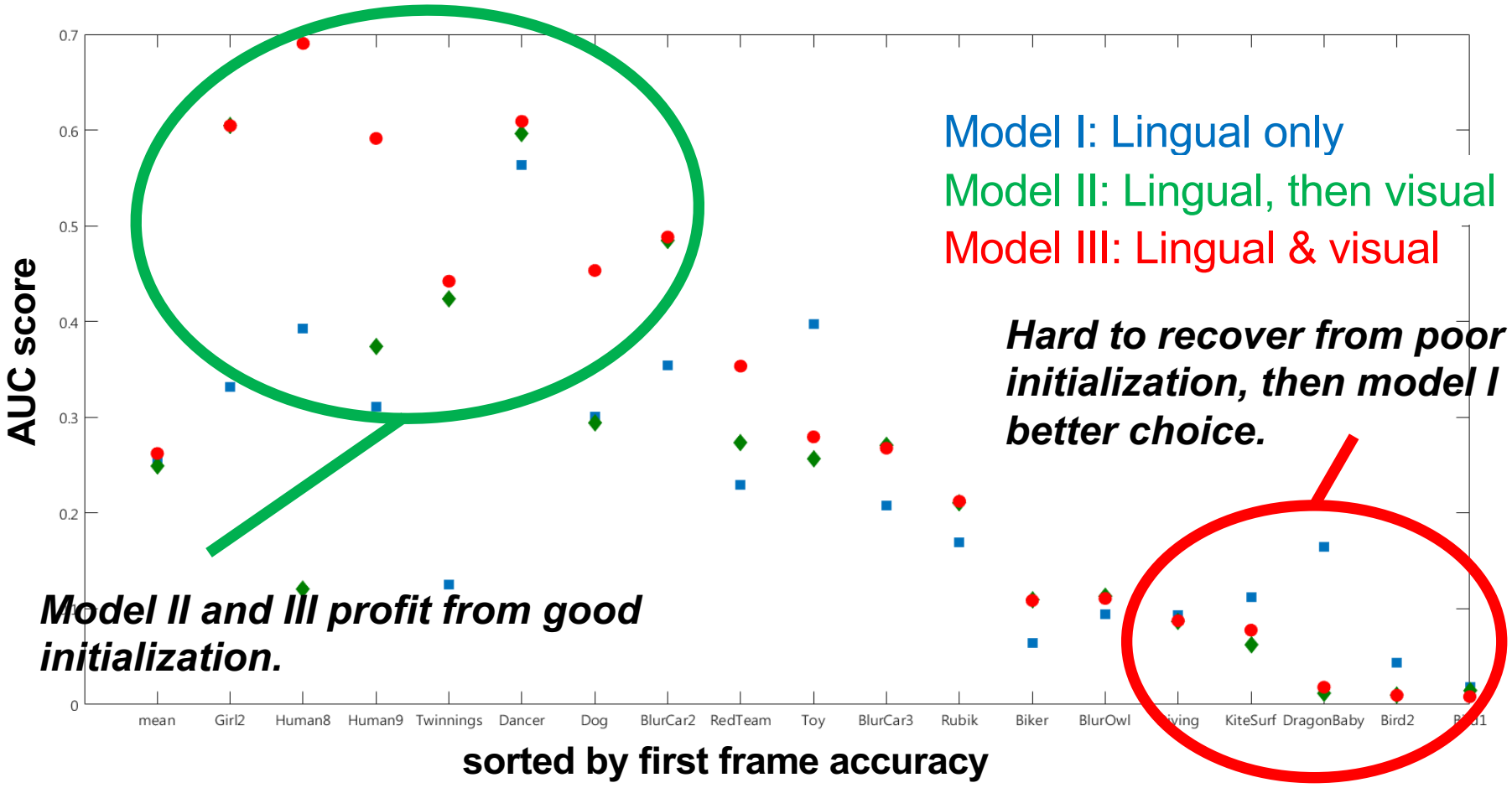
"*white car on the left*"

"*woman in dark pants on left*"

"*woman in white next to the woman in dark pants*"


Model I: Lingual only
Model II: Lingual, then visual
Model III: Lingual & visual
AUC score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Girl2
Toy
Dragonbaby
mean
Girl2
Human8
Human9
Twinnings
Dancer
Dog
BlurCar2
RedTeam
Toy
BlurCar3
Rubik
Biker
BlurOwl
Diving
KiteSurf
DragonBaby
Bird2
Bird1
sorted by first frame accuracy


AUC score
0.7
0.6
0.5
0.4
0.3
0.2
0
mean
Girl2
Human8
Human9
Twinnings
Dancer
Dog
BlurCar2
RedTeam
Toy
BlurCar3
Rubik
Biker
BlurOwl
Diving
KiteSurf
DragonBaby
Bird2
Bird1
sorted by first frame accuracy
Model I: Lingual only
Model II: Lingual, then visual
Model III: Lingual & visual
Model II and III profit from good initialization.


AUC score
0.7
0.6
0.5
0.4
0.3
0.2
0
mean
Girl2
Human8
Human9
Twinnings
Dancer
Dog
BlurCar2
RedTeam
Toy
BlurCar3
Rubik
Biker
BlurOwl
KiteSurf
DragonBaby
Bird2
sorted by first frame accuracy
Model I: Lingual only
Model II: Lingual, then visual
Model III: Lingual & visual
Hard to recover from poor initialization, then model I better choice.
Model II and III profit from good initialization.

“The black and white dog”

Lingual only **Lingual, then visual** **Lingual & visual**

“White car on the left”

Lingual only **Lingual, then visual** **Lingual & visual**

“Girl in yellow shirt and purple pants”

Lingual only Lingual, then visual Lingual & visual

Track by language ***and*** box specification

"female skater in red"

Ground truth **Box specification** **Language and box specification**

Language helps against drift.

Track by language ***and*** box specification

"*people on the right next to a big tree*"

Ground truth **Box specification** **Language and box specification**

Language helps against drift.

II. Actions

Kirill Gavrilyuk

Amir Ghodrati

Zhenyang Li

Cees Snoek

Actor and Action Video Segmentation from a Sentence. In *CVPR* 2018.

Goal

Input video

“woman in purple dress running”

“gray dog running on a leash during dog show”

Benefits of language

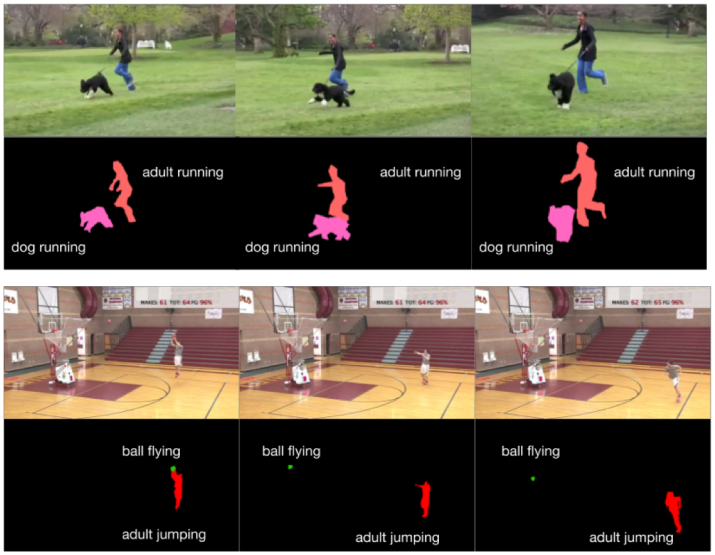
Distinguish fine-grained actors within same super-category

Identify actor and action instances

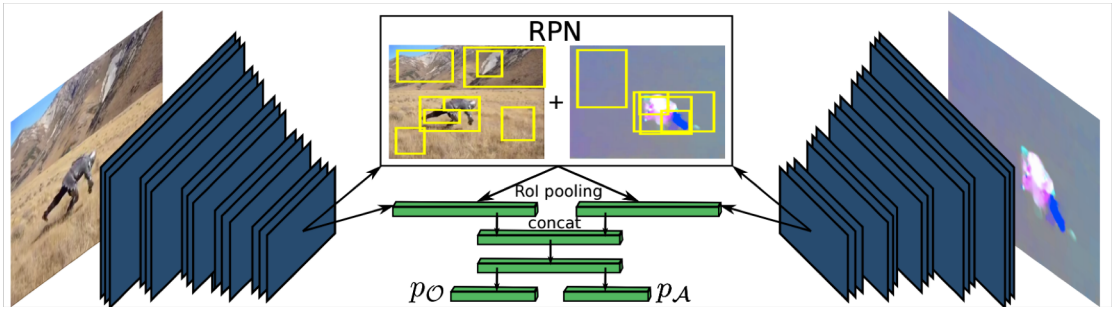
Segment actors & actions outside pre-defined vocabulary

Actor-action video segmentation

Different goal: recognize after learning from provided actor and action label pairs



3782 videos, 7 actors, 8 actions

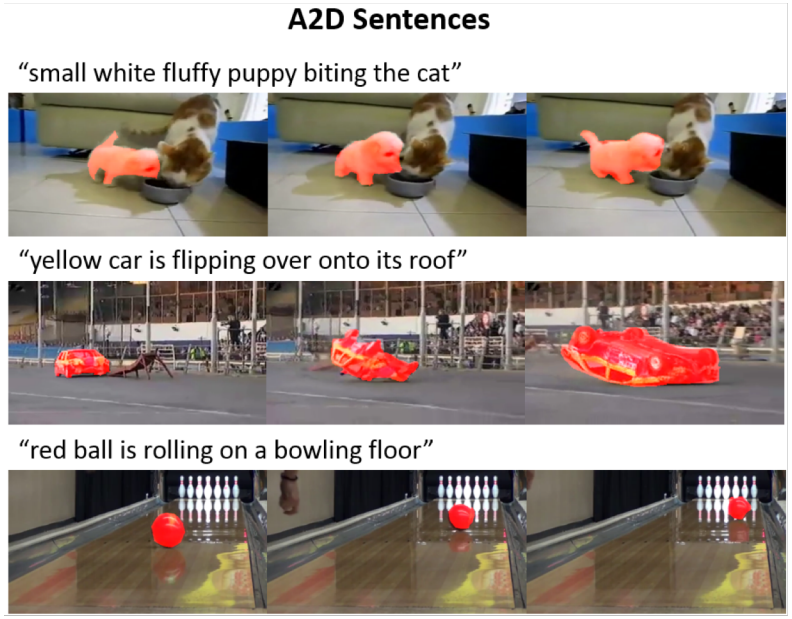


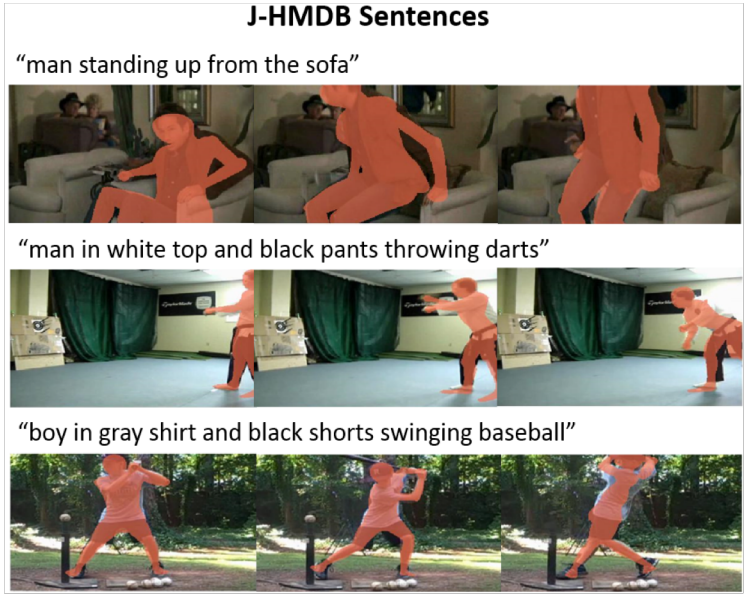
Joint actor-action detection and segmentation in videos using single RGB and Flow frames as input

Xu et al. CVPR15

Kalogeiton et al. ICCV17

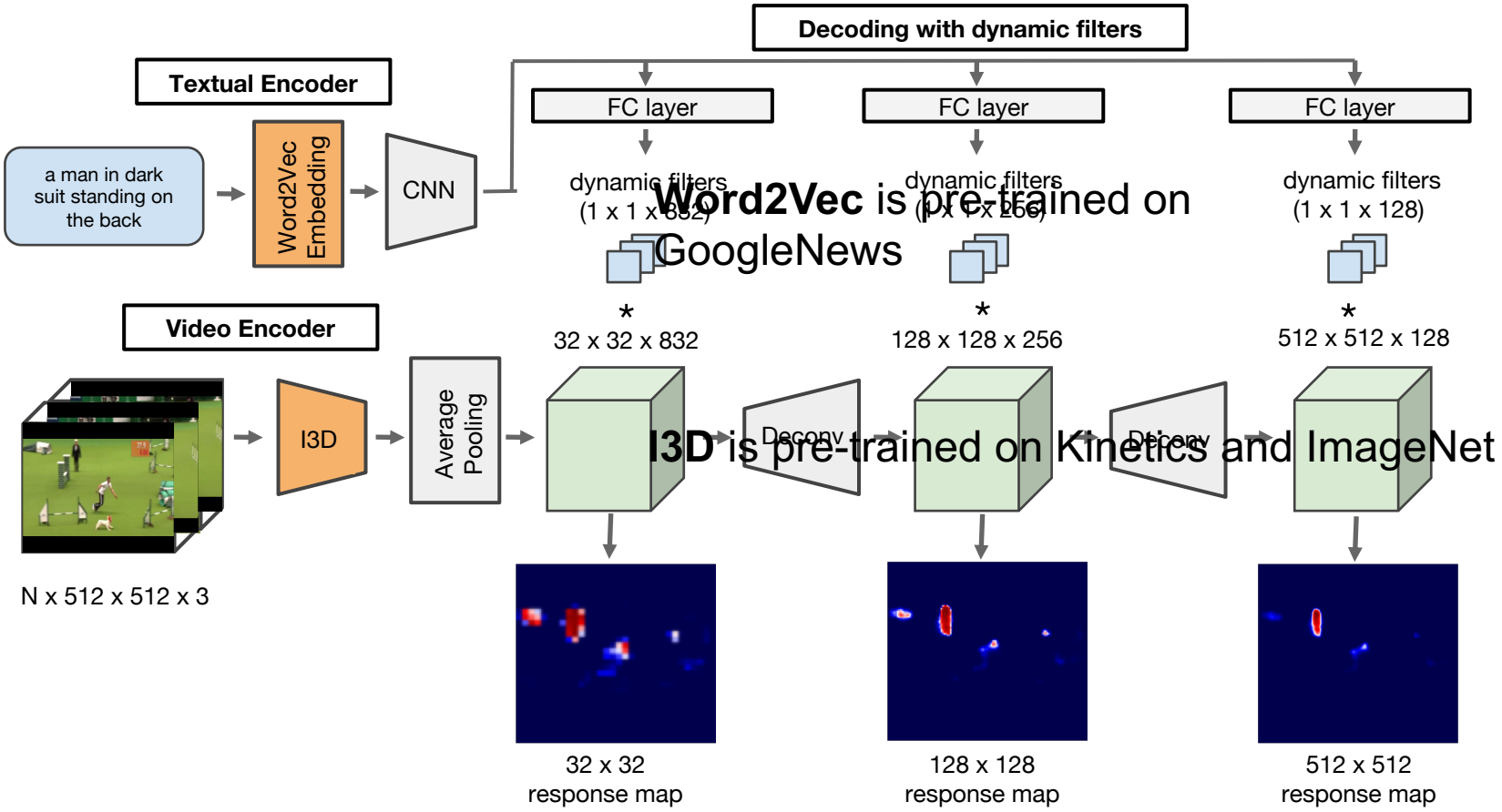
Augment action video with sentences





Two datasets extended with more than 7,500 natural language descriptions

Model


Decoding with dynamic filters
Textual Encoder
a man in dark suit standing on the back
Word2Vec Embedding
CNN
FC layer
FC layer
FC layer
dynamic filters (1 x 1 x 832)
dynamic filters (1 x 1 x 256)
dynamic filters (1 x 1 x 128)
Word2Vec is pre-trained on GoogleNews
Video Encoder
I3D
Average Pooling
32 x 32 x 832
128 x 128 x 256
512 x 512 x 128
Deconv
Deconv
I3D is pre-trained on Kinetics and ImageNet
N x 512 x 512 x 3
32 x 32 response map
128 x 128 response map
512 x 512 response map

Training

Training sample - video clip, sentence and binary segmentation mask

Loss

Loss per sample takes into account multiple resolutions

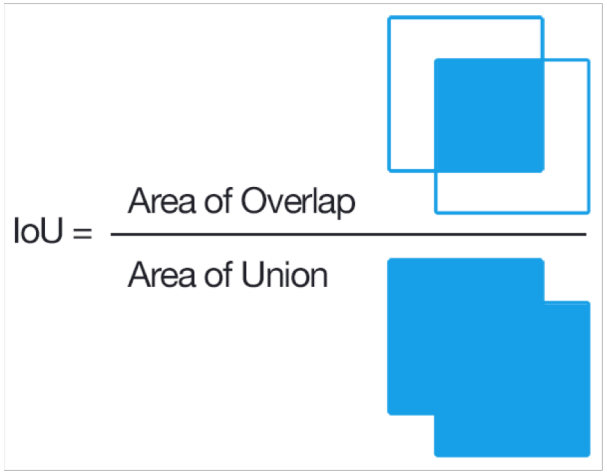
$$\mathcal{L} = \sum_{r \in R} \alpha_r \mathcal{L}^r \qquad \mathcal{L}^r = \frac{1}{r^2} \sum_{i=1}^{r} \sum_{j=1}^{r} \mathcal{L}_{ij}^r$$

Logistic loss for per-pixel classification

$$\mathcal{L}_{ij}^r = log(1 + \exp\left(-S_{ij}^r Y_{ij}^r\right))$$

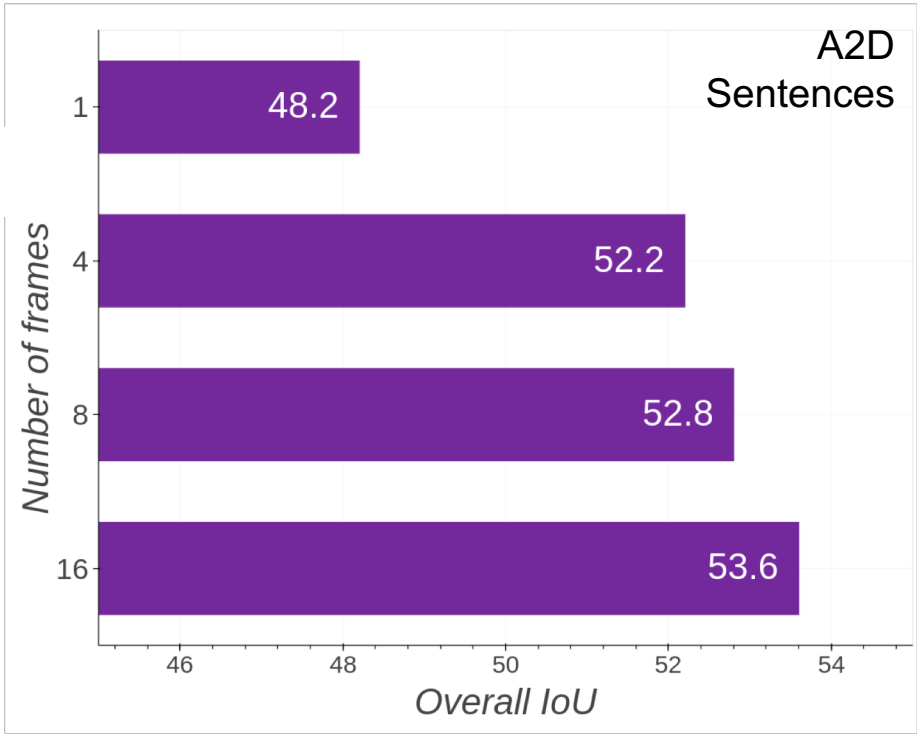
Evaluation metric

Intersection over Union



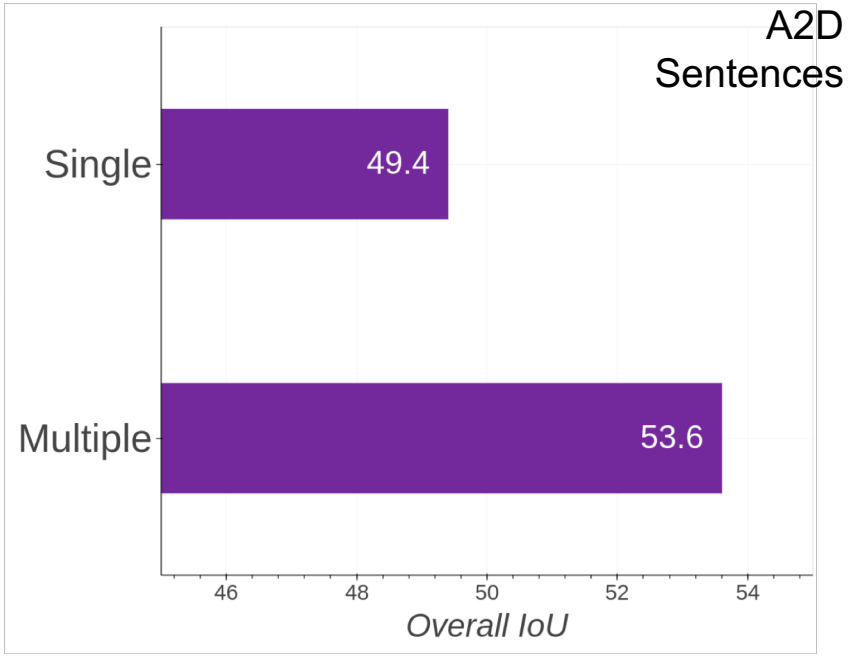
mAP 0.5:0.95: mean over precision for various IoU levels in range of [0.5:0.05:0.95]

Ablation: image or video?

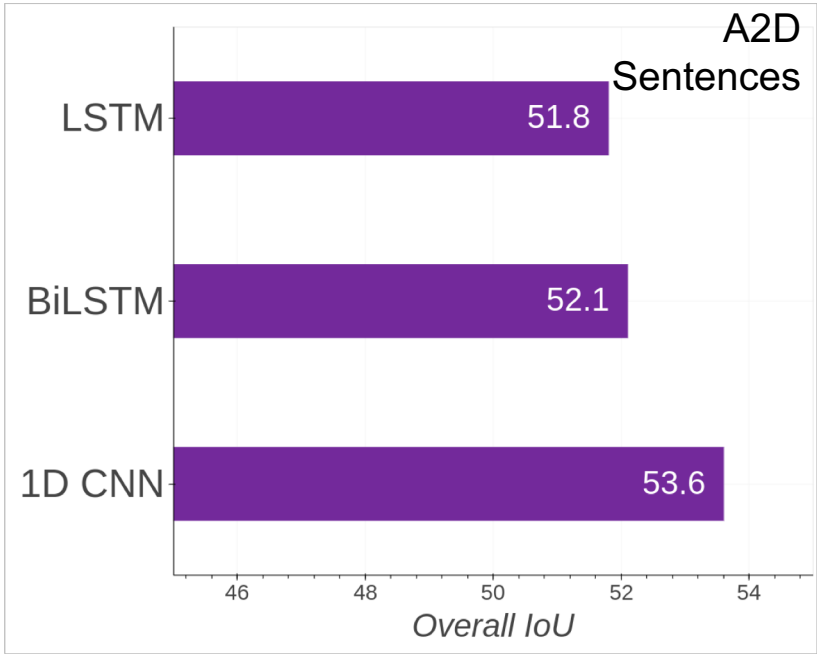


For segmentation, video is more than just a set of independent frames

More ablation

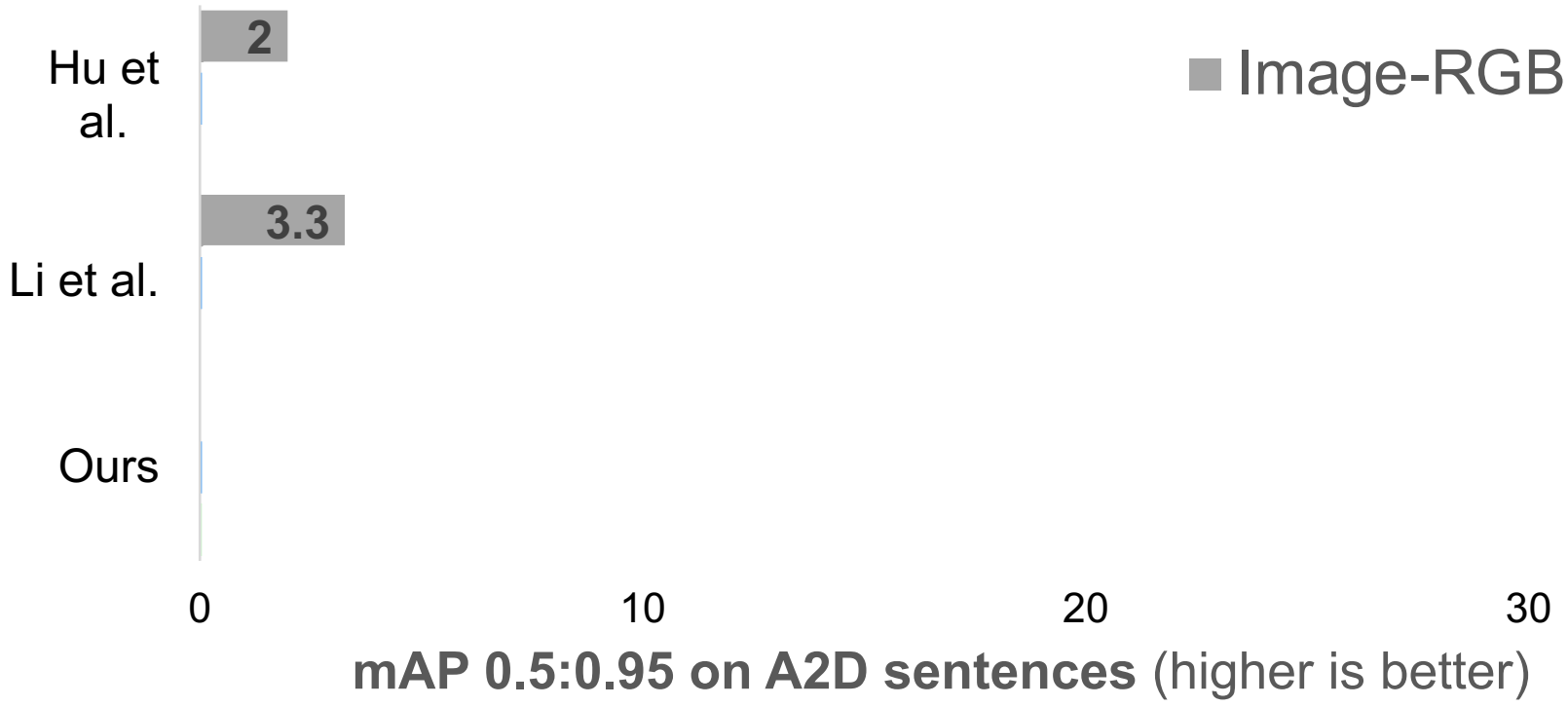


Multi-resolution for better training



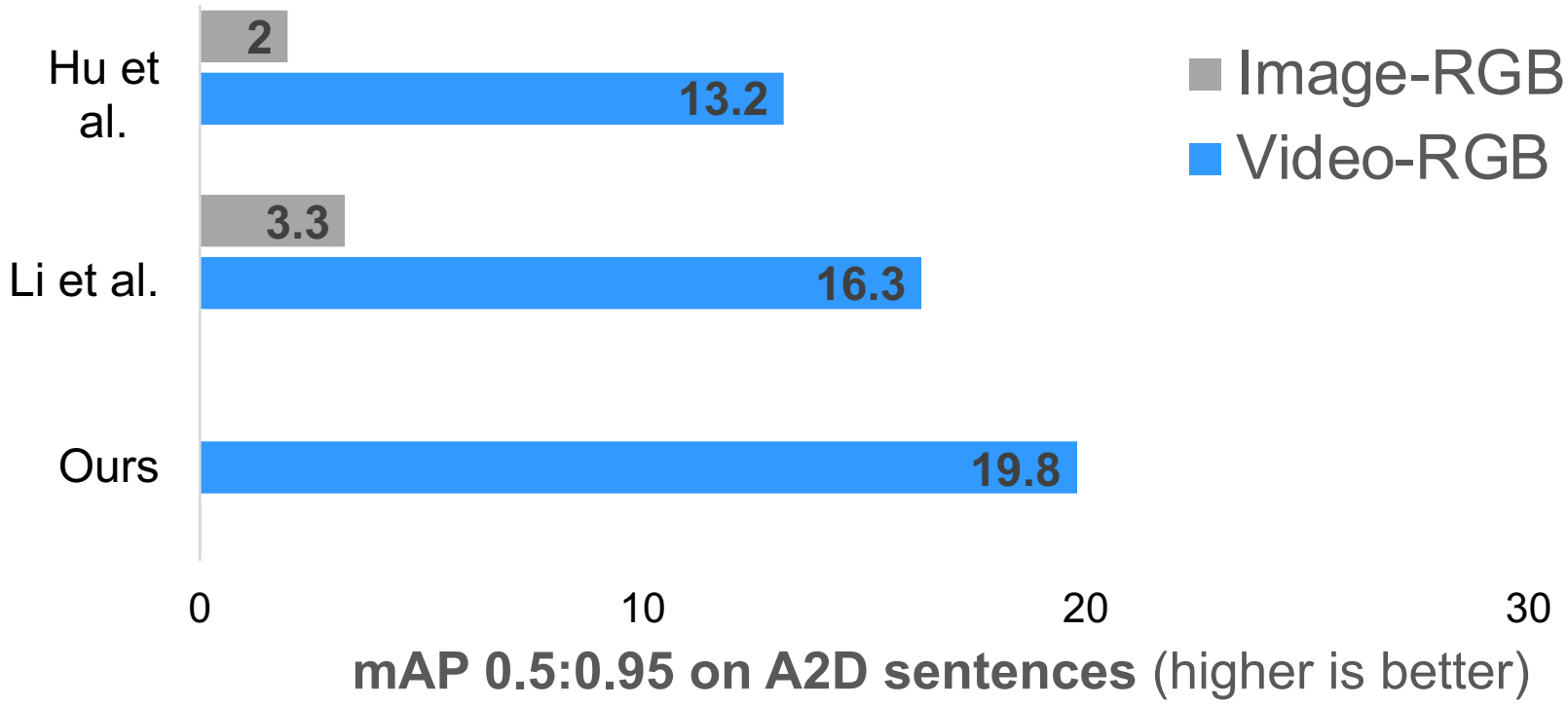
Simple 1D CNN outperforms LSTM

Comparison with image-baselines



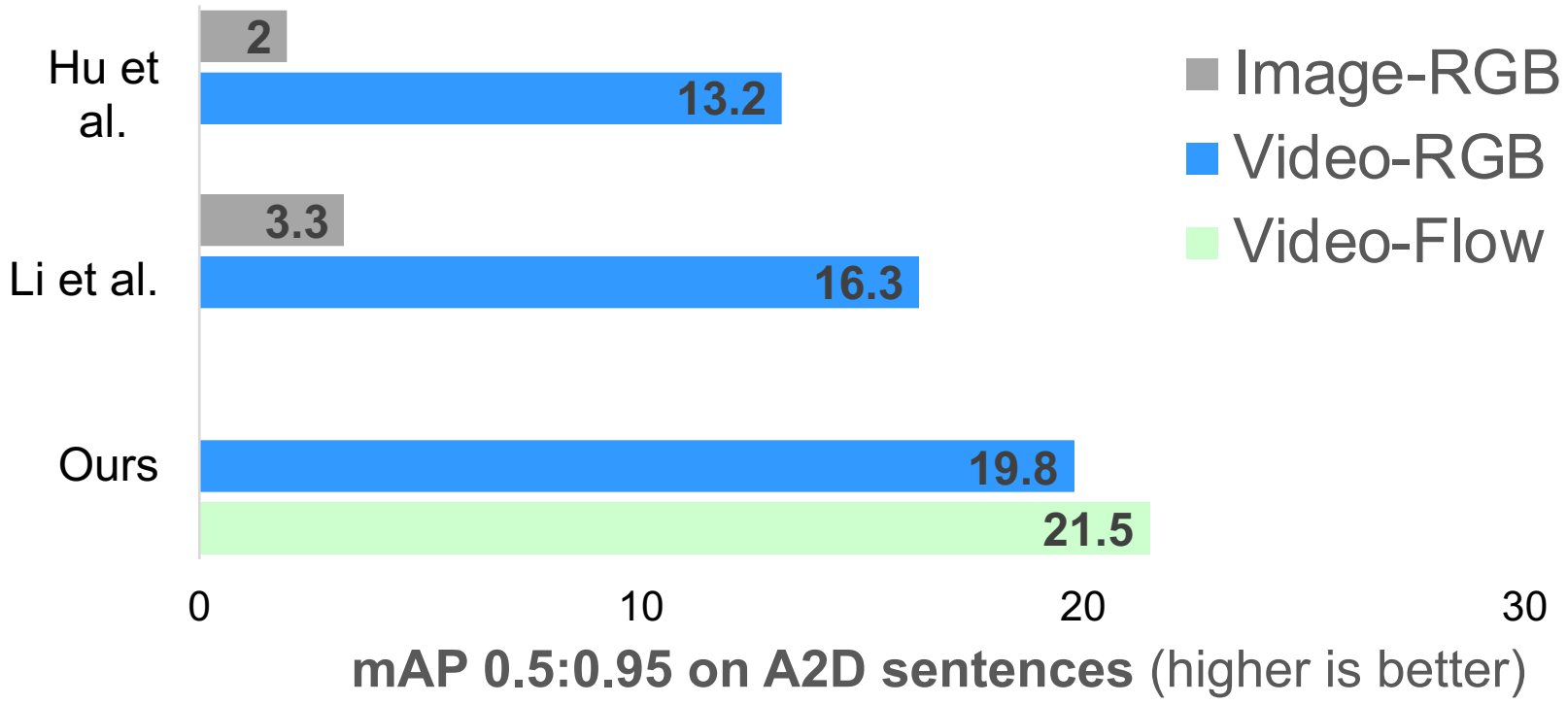
Existing models without pre-training results in modest recognition.

Comparison with image-baselines



Our model outperforms both baselines.

Comparison with image-baselines



Incorporating flow in our video model further improves results.

Qualitative results: simple cases

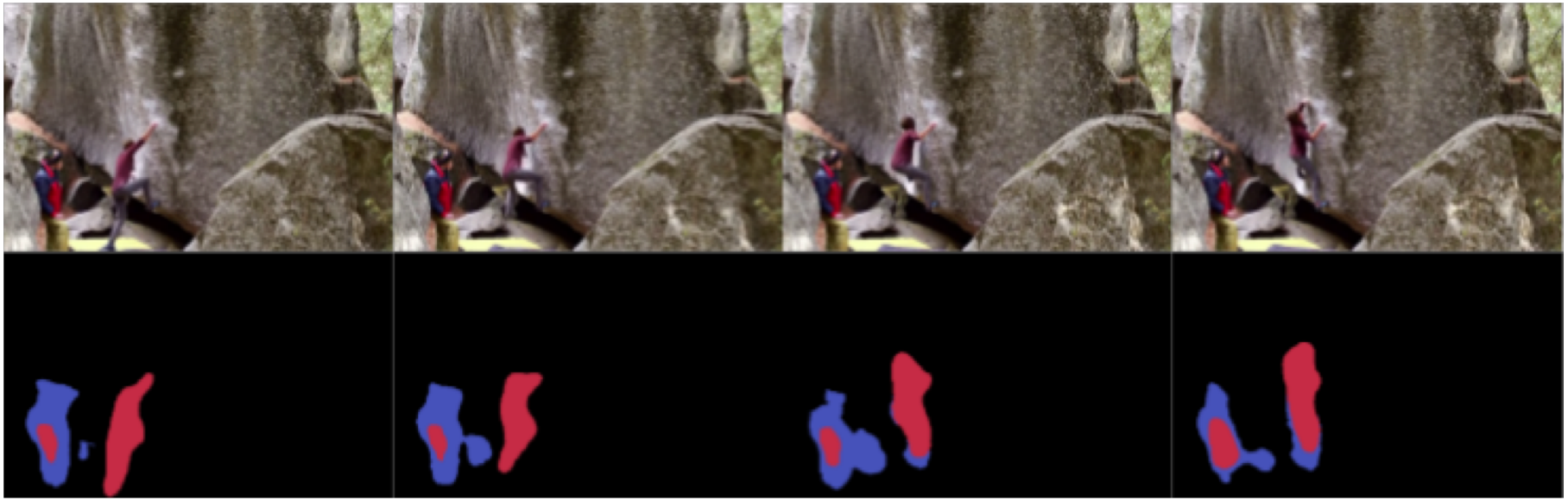
“kid rolling over”

“brown dog crawling on the ground”

Qualitative results: hard cases

“man standing on the left”

“a man is climbing a rock”

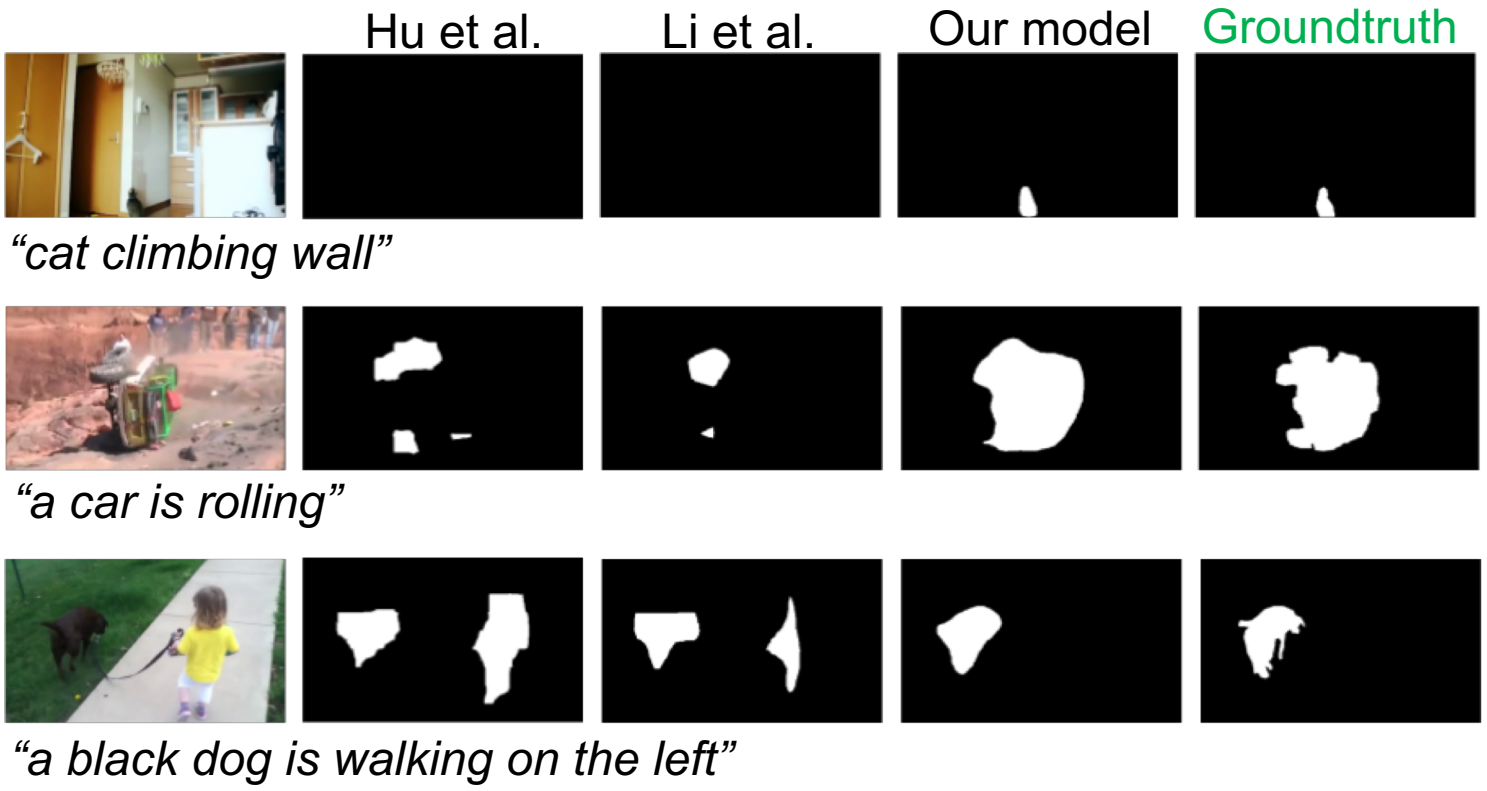


Qualitative results: failure cases

“man walking with a woman on the beach”

“woman walking with a man on the beach”

Qualitative comparison with baselines

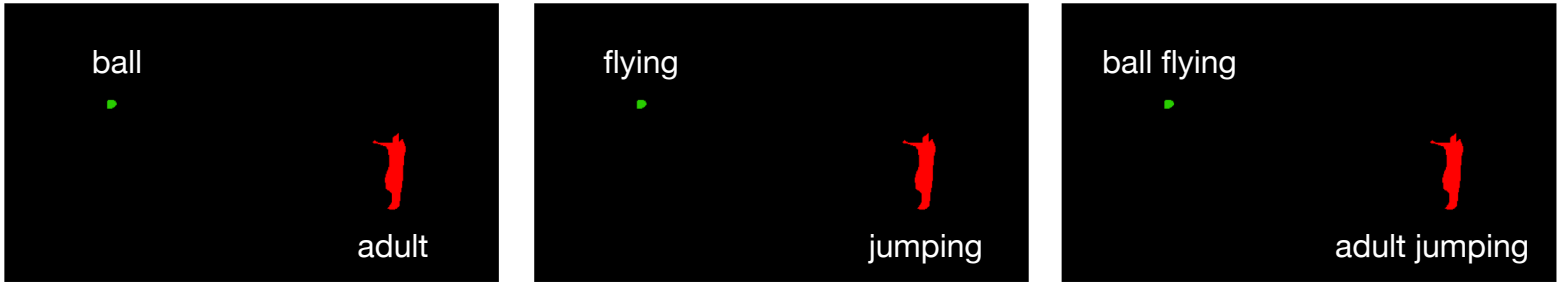


Segmentation from actor-action pairs

We use actor-action pairs from original A2D dataset as input

Segmentation from actor-action pairs

	Actor	
	Class-Average	Mean IoU
Xu *et al.*	45.7	-
Kalogeiton *et al.*	**73.7**	49.5
Our model	71.4	**53.7**



Our method outperforms the state-of-the-art in most cases.

Conclusion

New type of human-machine interaction for video understanding.

More robust tracking and segmentation of actors and actions.

Representations enable novel application scenarios.

Thank you!