Self-supervised & Multi-modal Video Learning

Efstratios Gavves
Assistant Professor at University of Amsterdam
Co-founder of Ellogon.AI

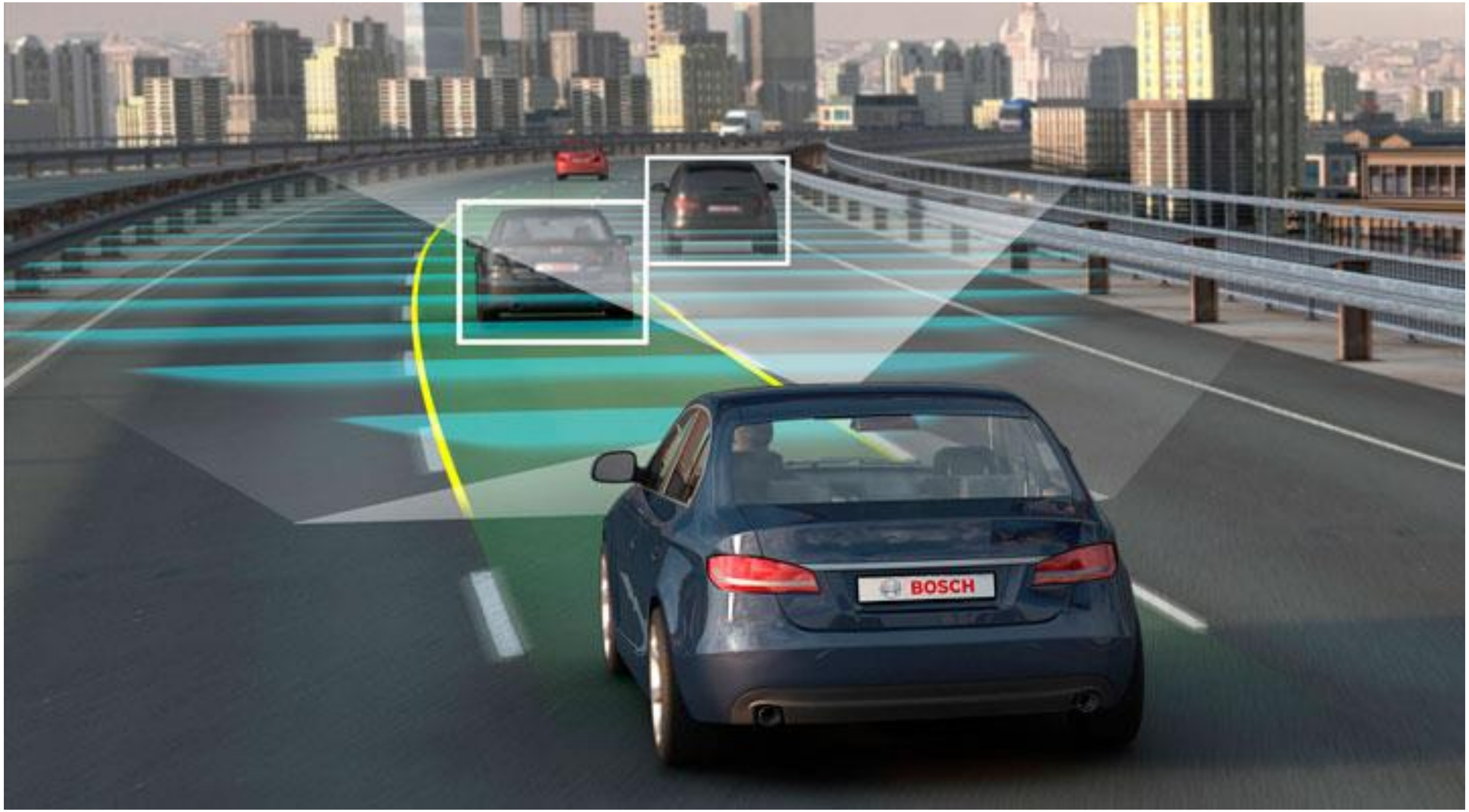
Courtesy of C. Snoek

THE INTERNET OF THINGS THAT VIDEO

45 billion cameras by 2022… [LDV Capital]

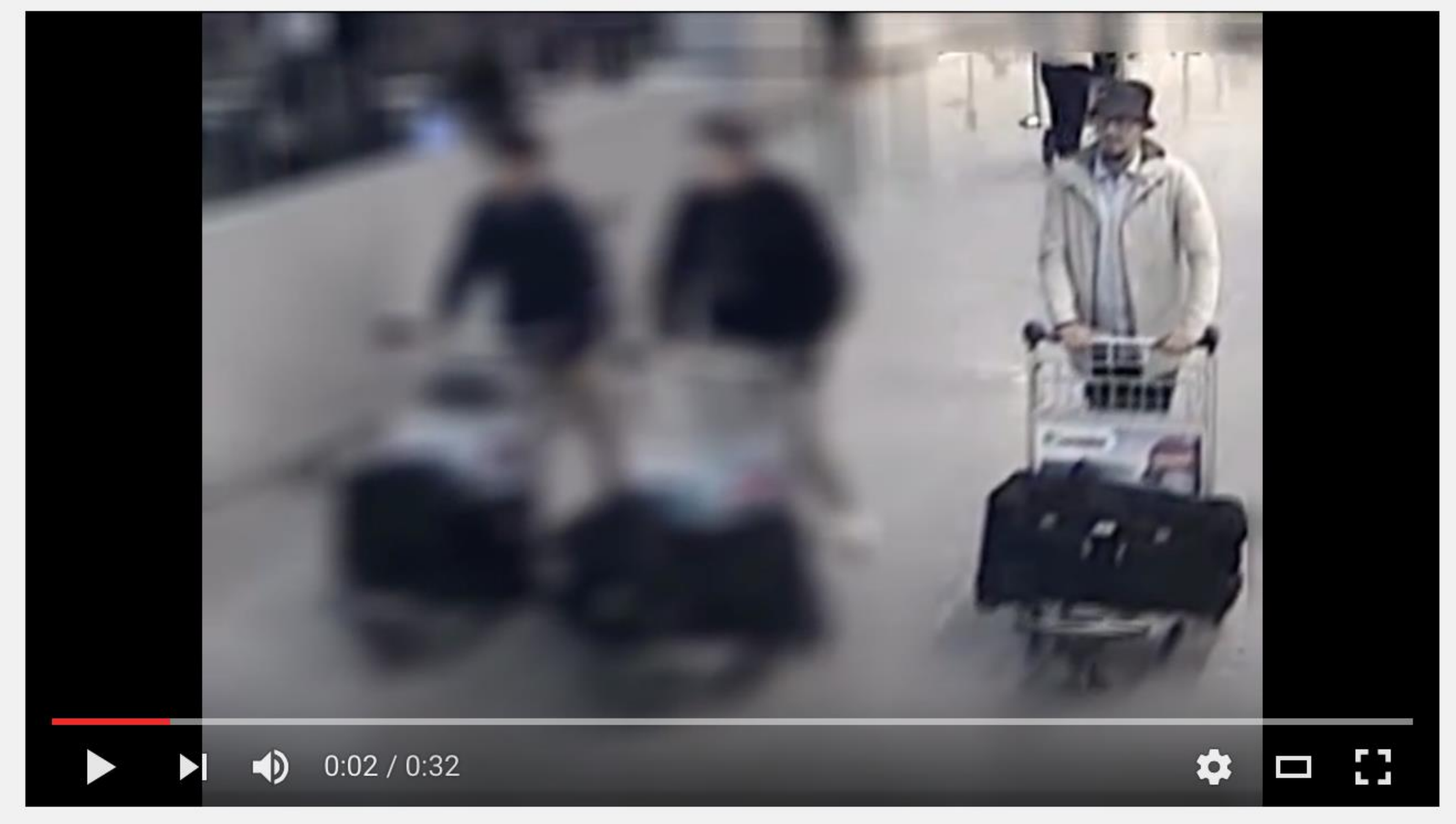
Courtesy of C. Snoek

TECHNOLOGY: SELF-DRIVING CARS



Courtesy of C. Snoek

Forensics: analyzing terrorist behavior



Courtesy of C. Snoek

TRAFFIC SURVEILLANCE

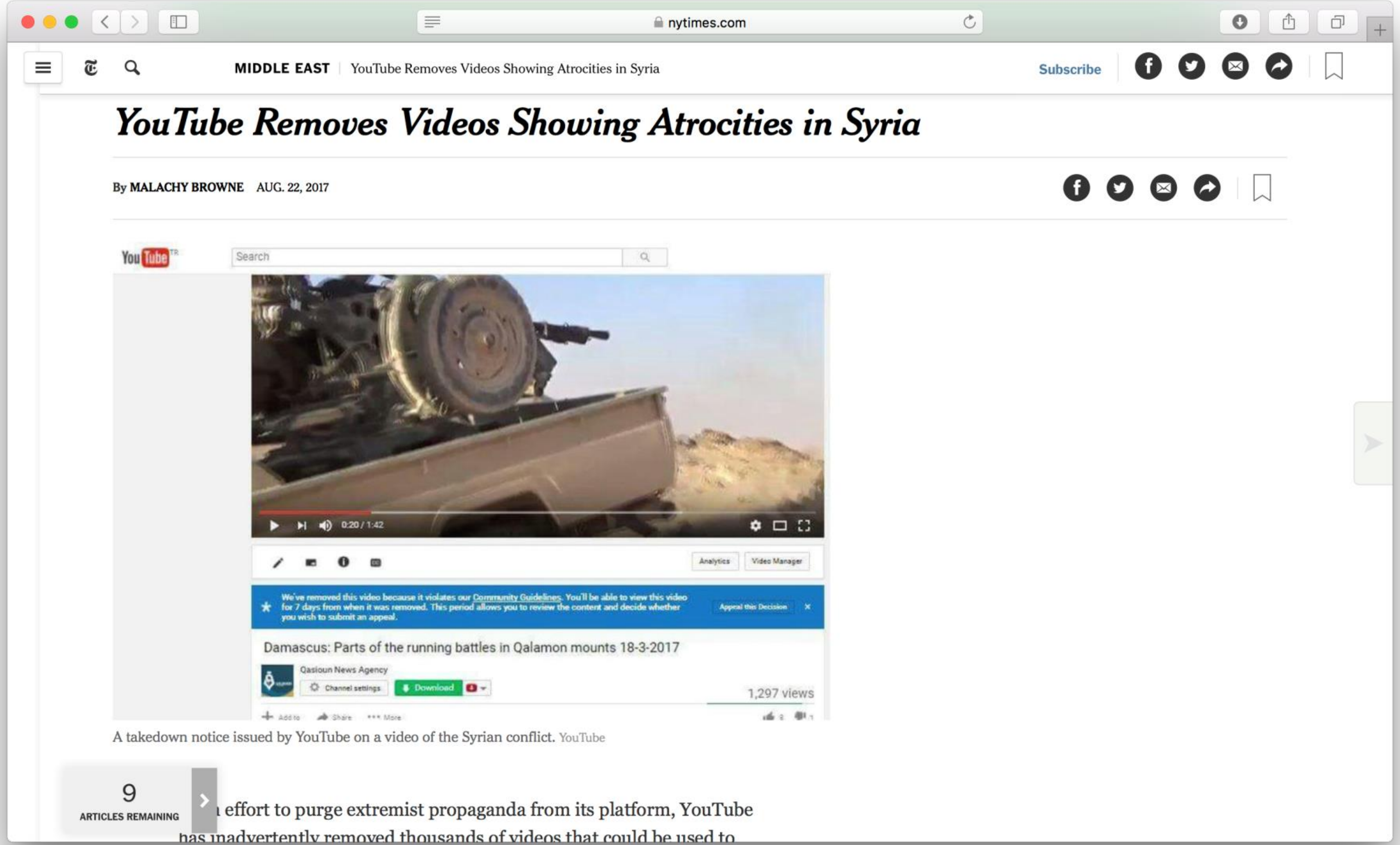
Courtesy of [C. Snoek]

WELL-BEING: ELDERLY MONITORING

Figure 1. Examples of interaction patterns in a nursing home

Courtesy of [C. Snoek]

SOCIAL: MEDIA MONITORING



nytimes.com

MIDDLE EAST | YouTube Removes Videos Showing Atrocities in Syria

Subscribe

YouTube Removes Videos Showing Atrocities in Syria

By MALACHY BROWNE AUG. 22, 2017

A takedown notice issued by YouTube on a video of the Syrian conflict. YouTube

9
ARTICLES REMAINING

…n effort to purge extremist propaganda from its platform, YouTube has inadvertently removed thousands of videos that could be used to

Courtesy of [C. Snoek]

RETAIL: CASHIER-LESS SHOPPING

SELF-SUPERVISED WITH ODD-ONE-OUT

- Self-Supervised Video Representation Learning With Odd-One-Out Networks, CVPR 2017

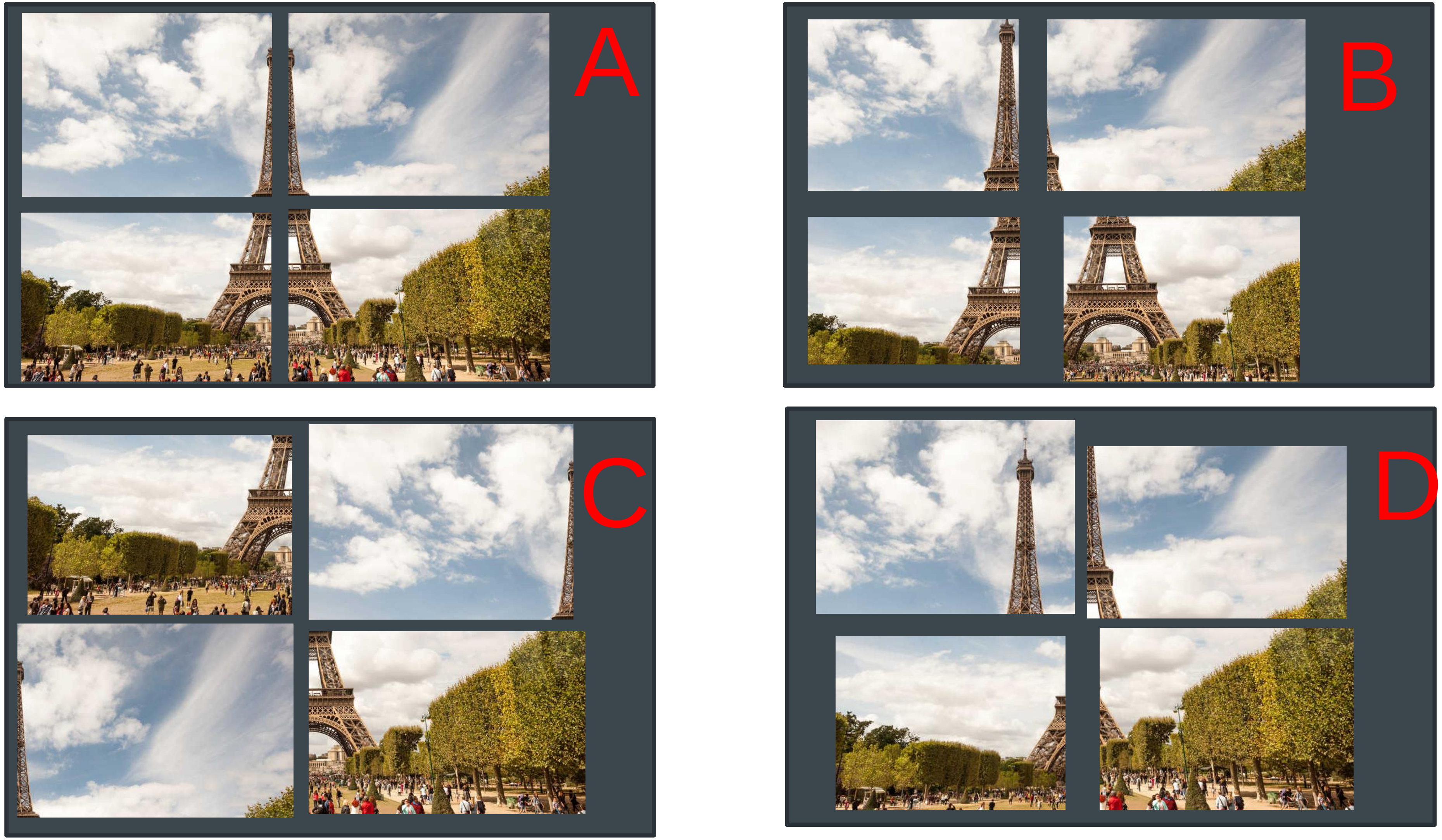
Basura Fernando

Hakan Bilen

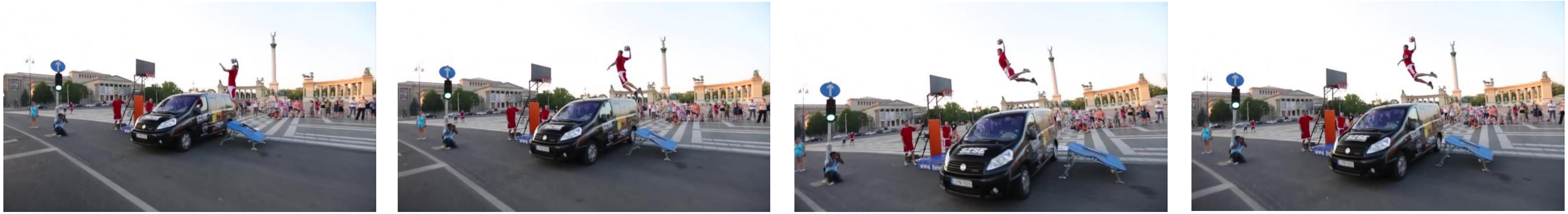
Efstratios Gavves

Stephen Gould

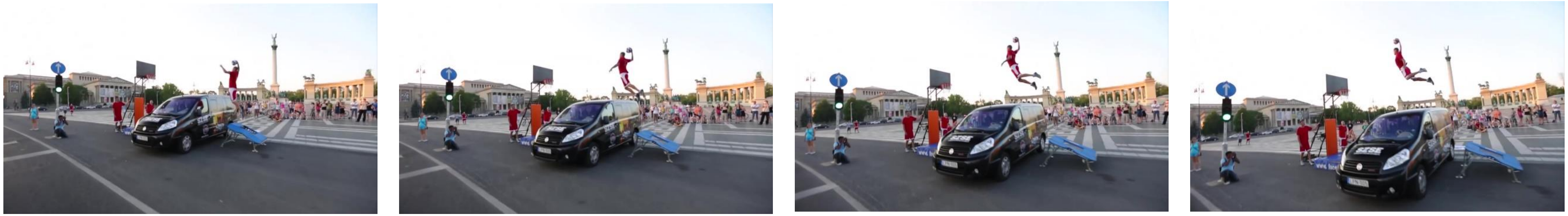
Find the Wrong Input



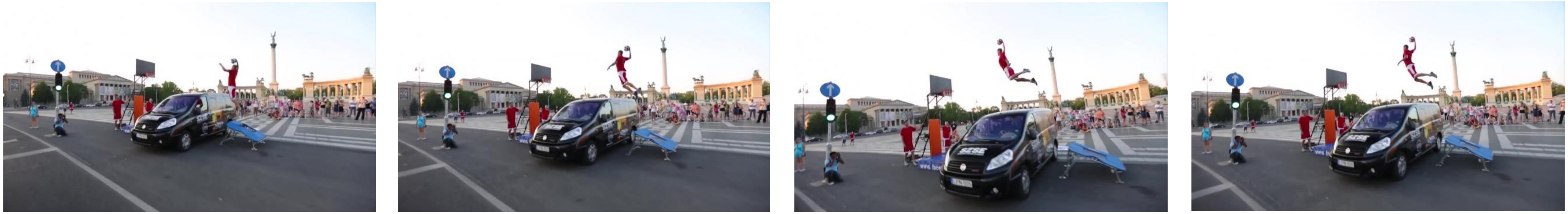
And Temporally



or

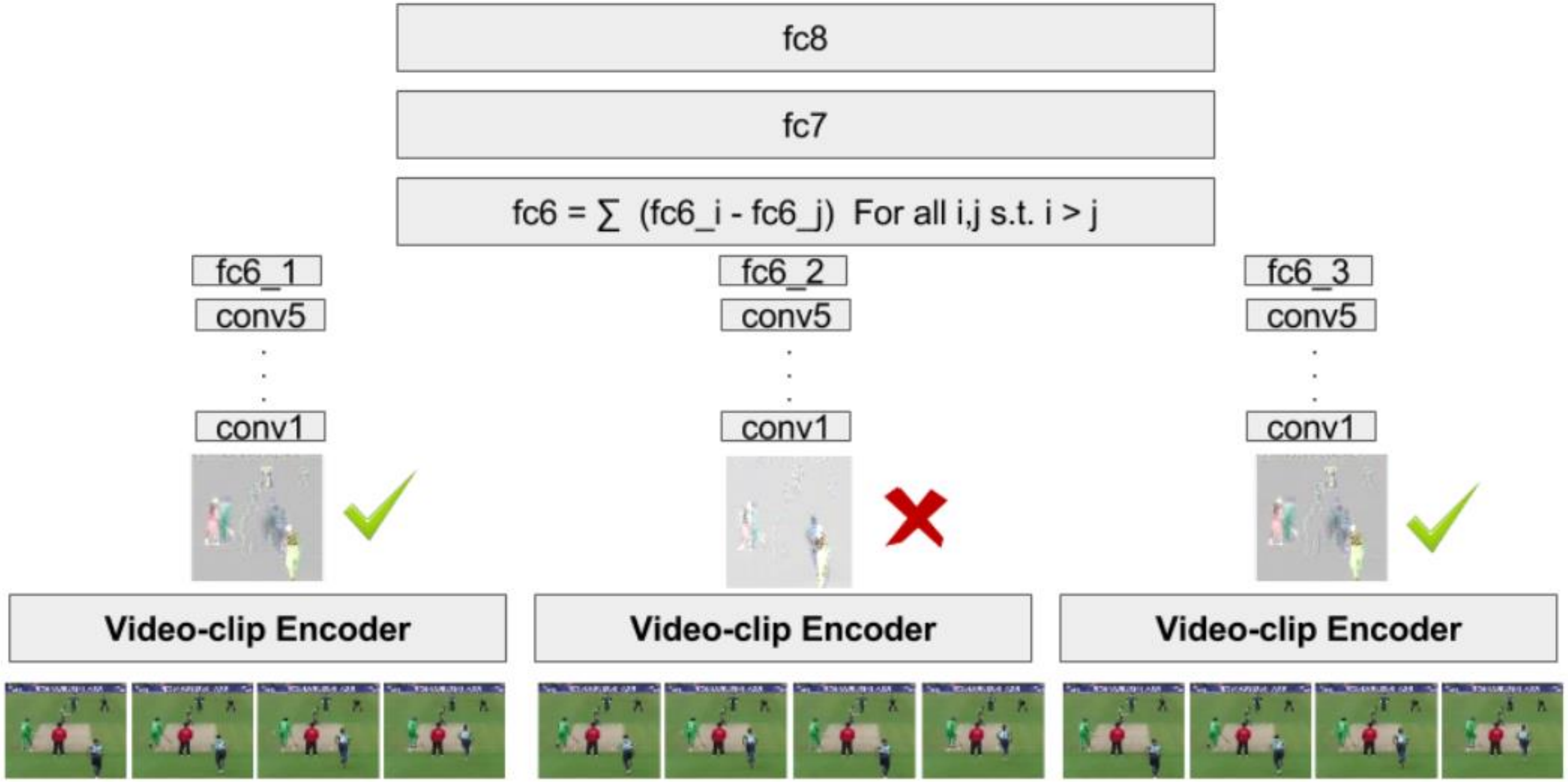


AND TEMPORALLY



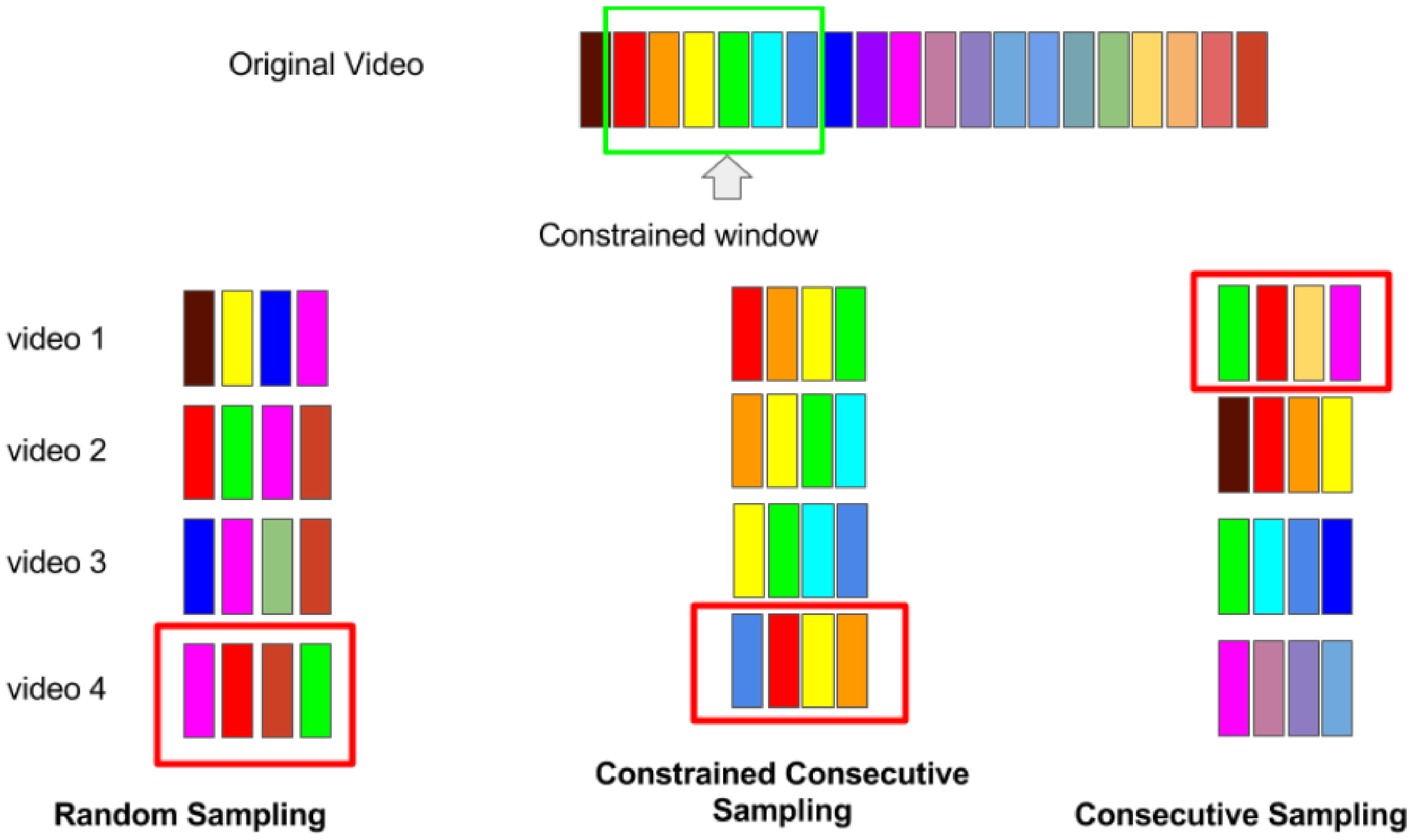
or

Odd-one-out Learning Model

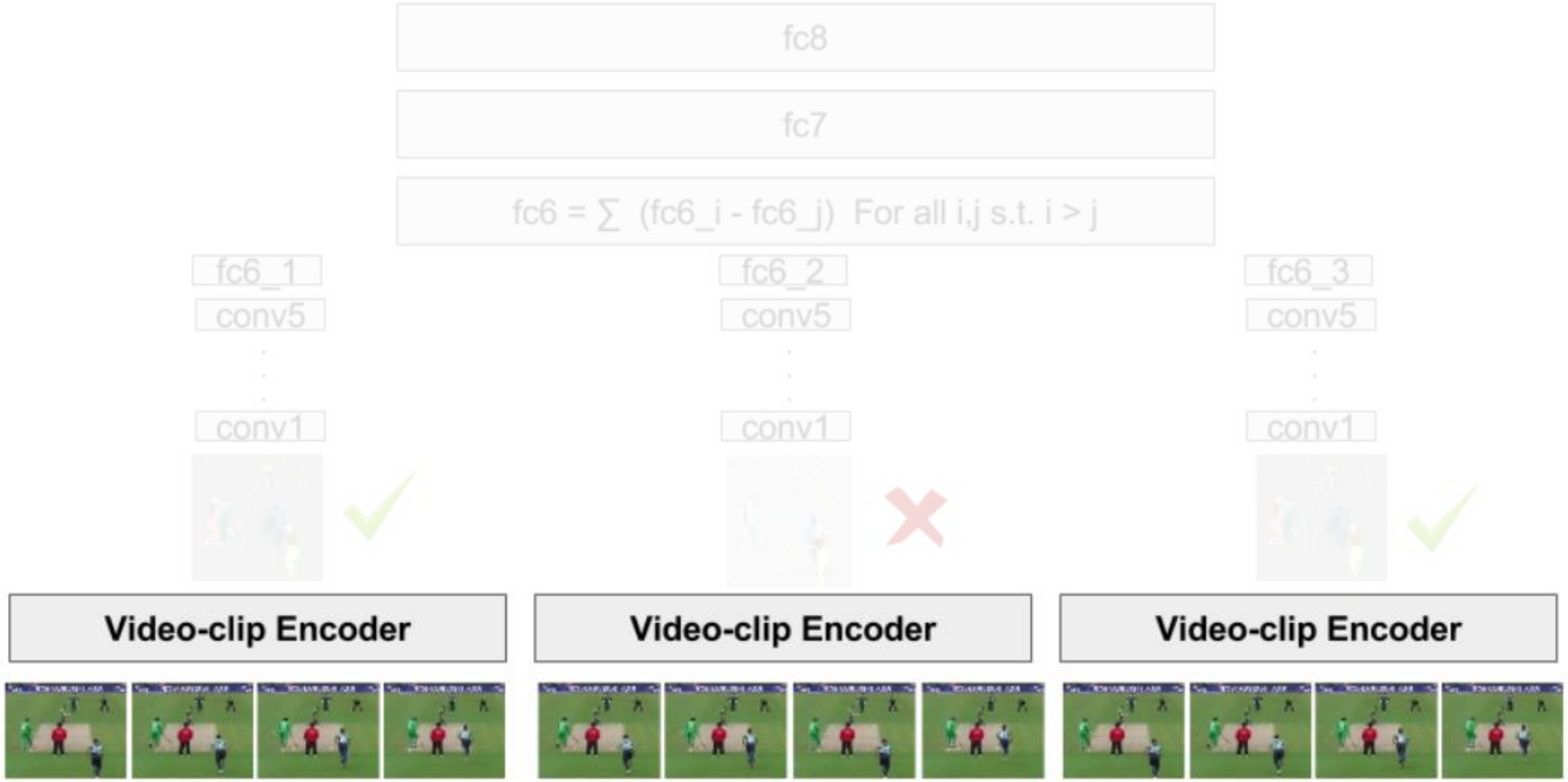


How to Sample Frames?

How to Sample Frames?



How to Encode Frames

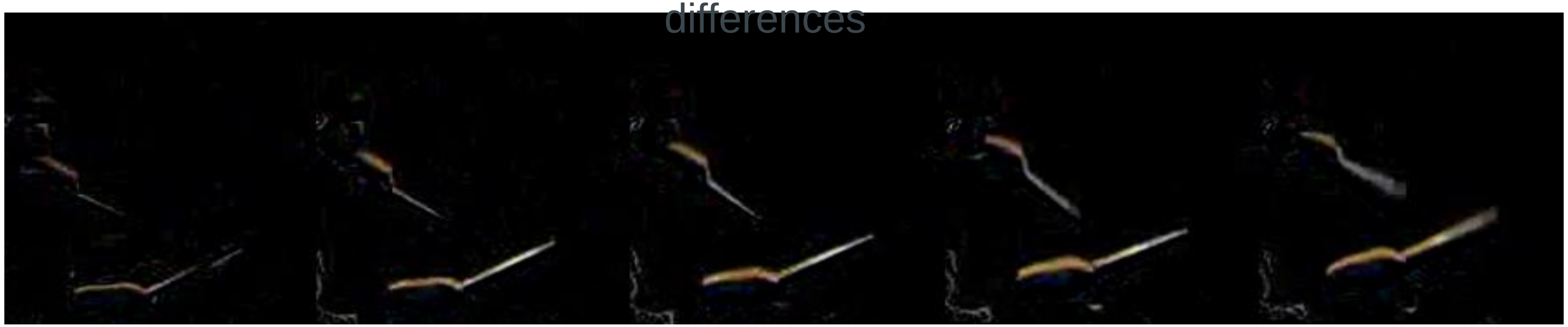


How to Encode Frames

Dynamic images

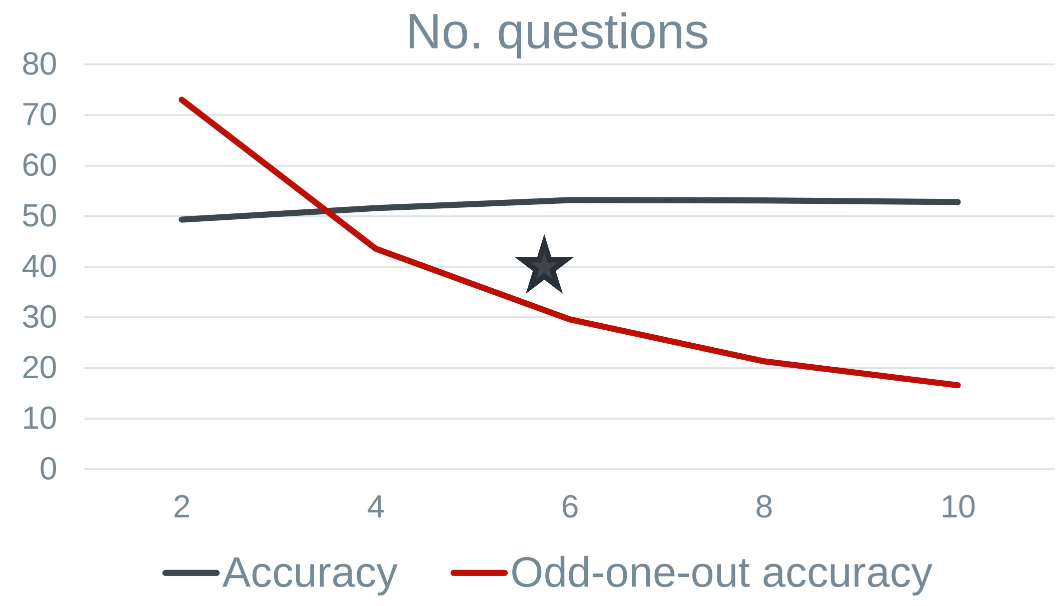
Sum of differences

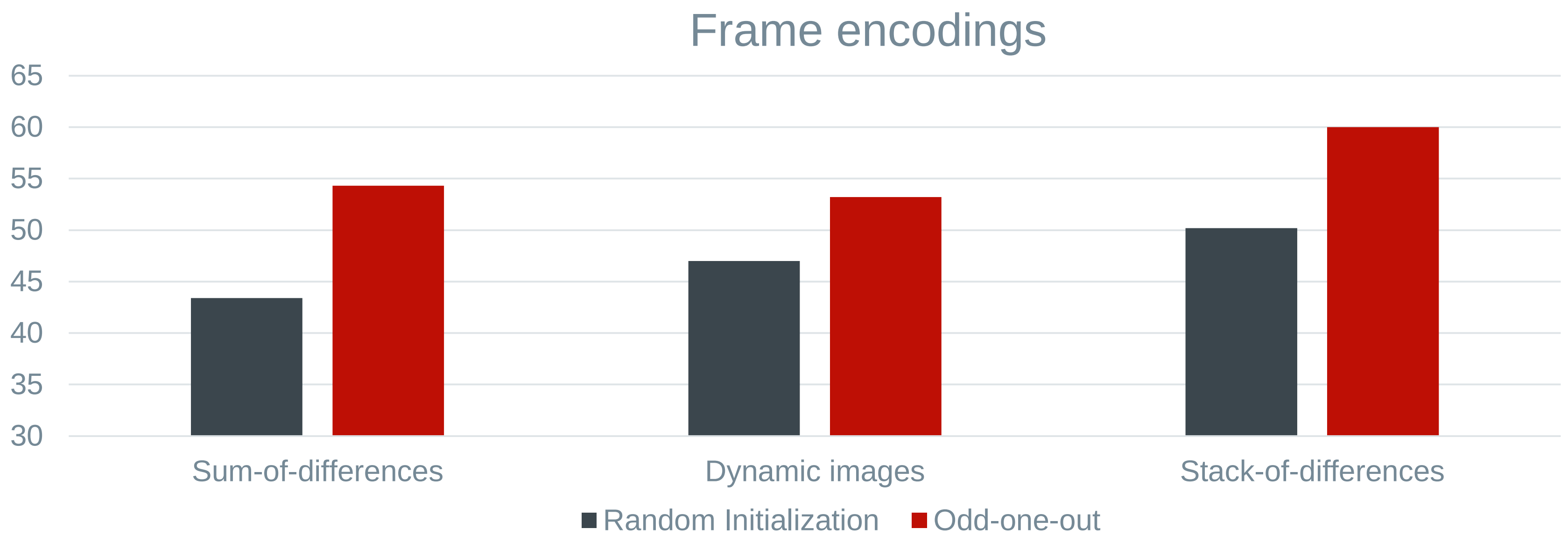
Stack of differences



EXPERIMENTS

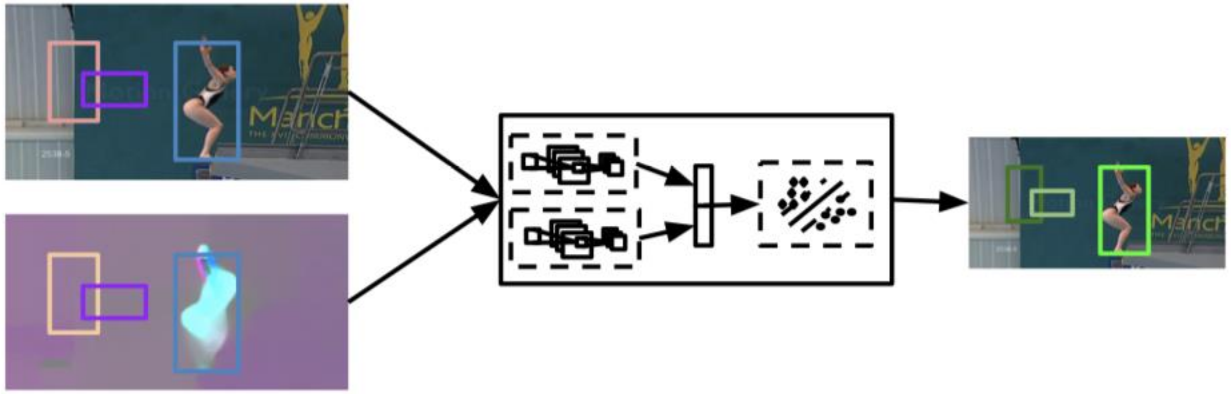
Sampling	Accuracy	Odd-one-out acc.
Consecutive	50.6	27.4
Constrained consecutive	52.4	29.0
Random	53.2	29.6





Two-stream

- Default strategy for action detection and classification.
  - RGB-stream: appearance only
  - Flow-stream: motion only

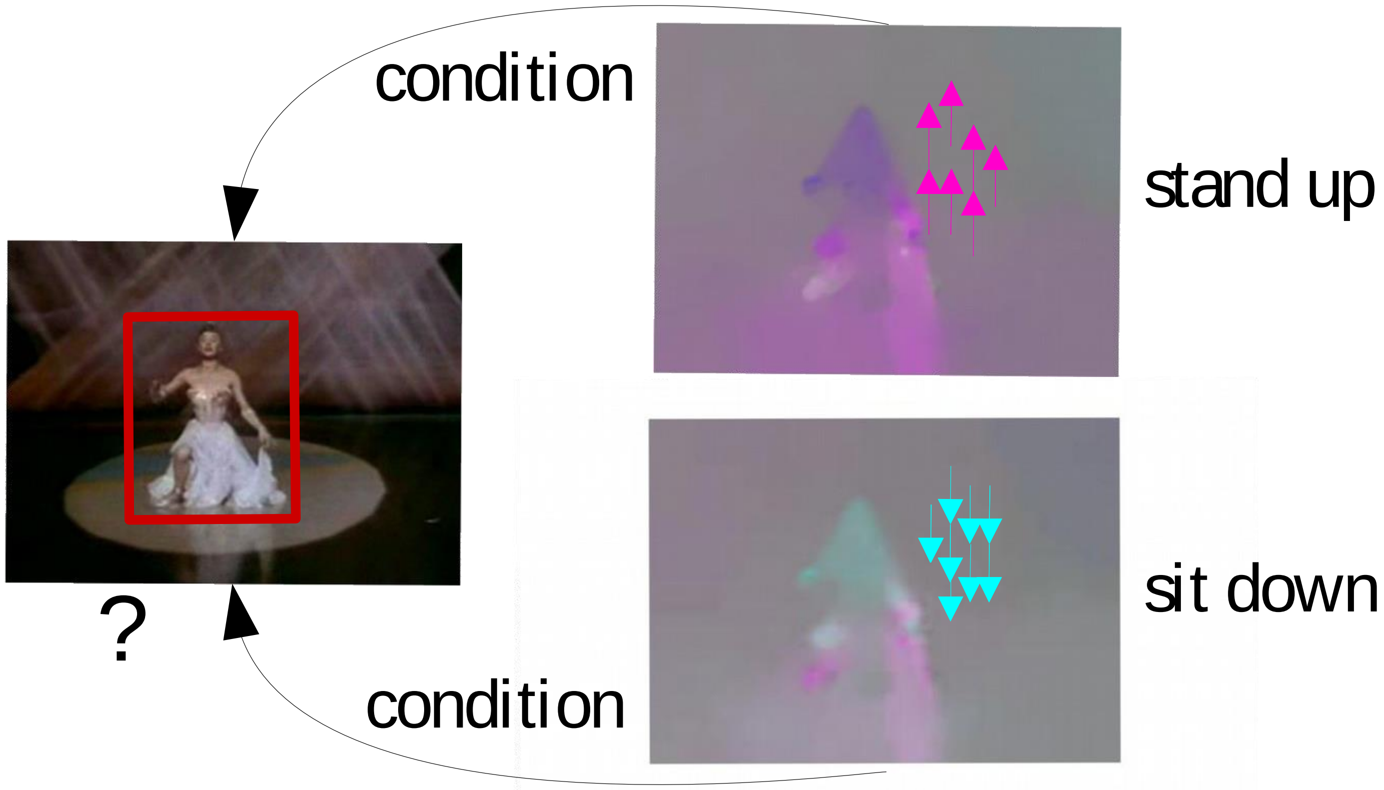


- Doubles computation and parameters for modest accuracy gain.

Simonyan & Zisserman NeurIPS14

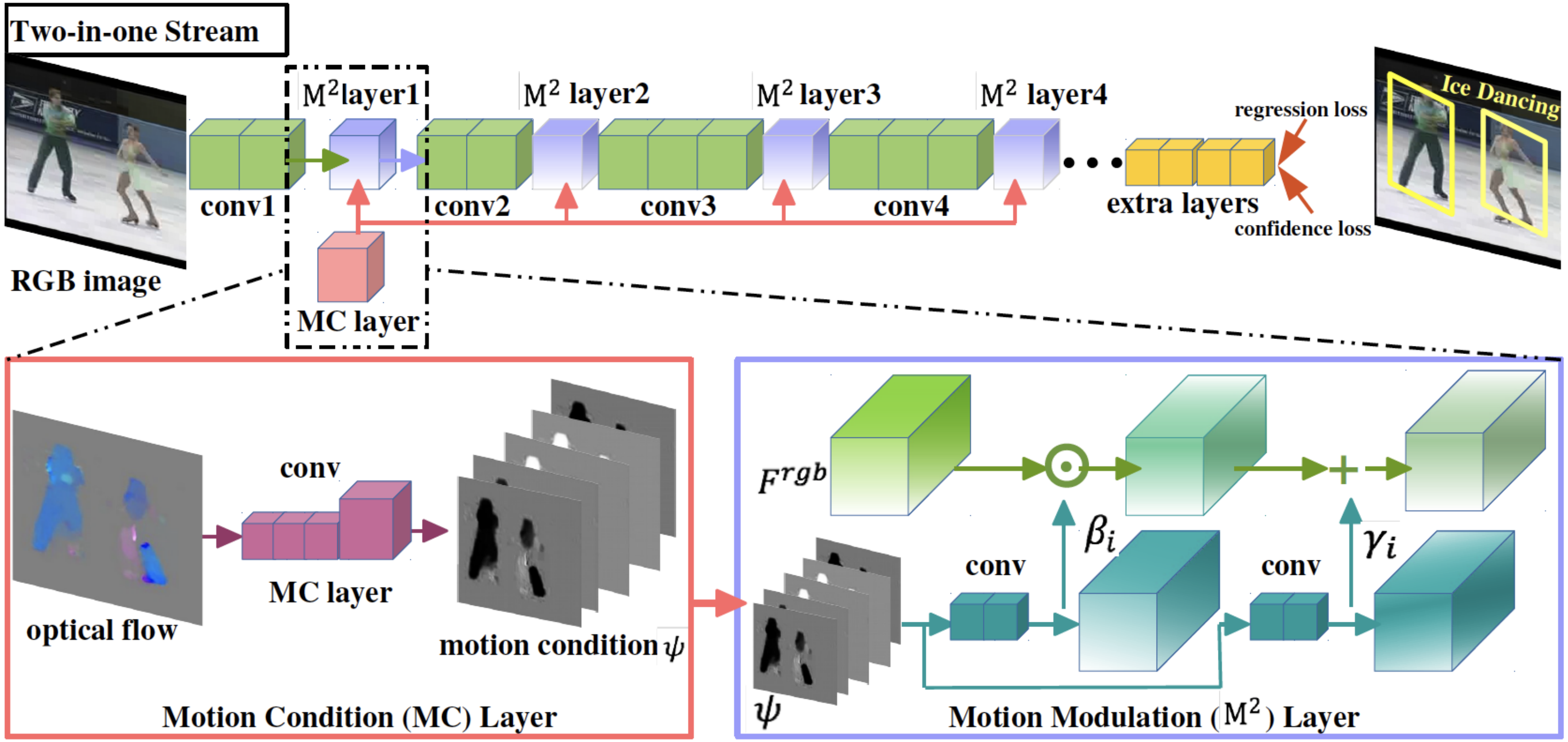
KEY IDEA

Use motion as condition when training a single RGB-stream.



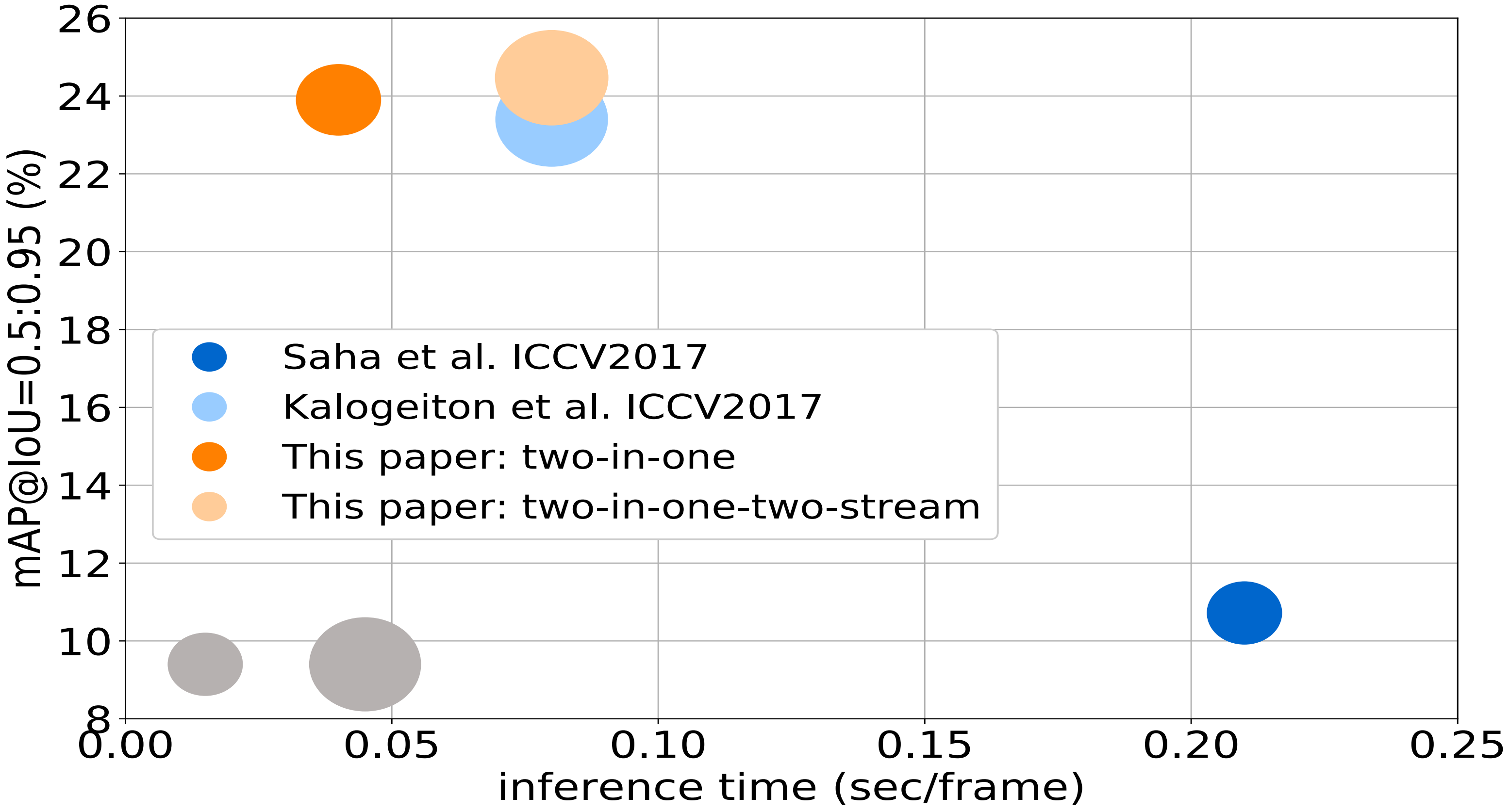
Two-in-one Stream

- Learns a single stream RGB model conditioned on motion information
- Dance With Flow: Two-In-One Stream Action Detection, Zhao and Snoek, CVPR 2019
- To be presented on Thursday at 10.00, Poster 131



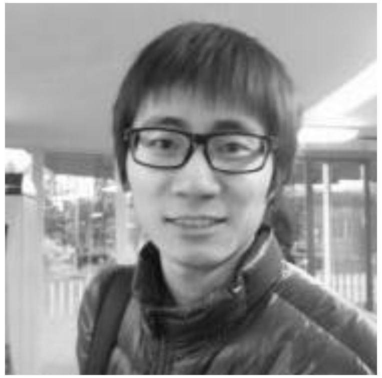
EXPERIMENTS

- Faster, lighter and better accuracy.



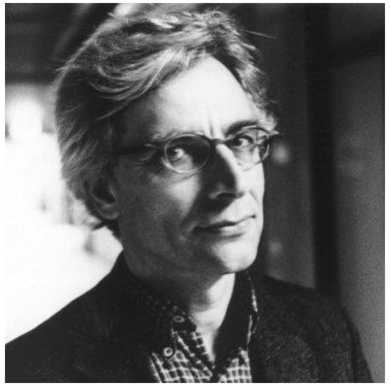
TRACKING A LA SIAMESE

- Siamese Instance Search for Tracking, CVPR 2016



Ran Tao

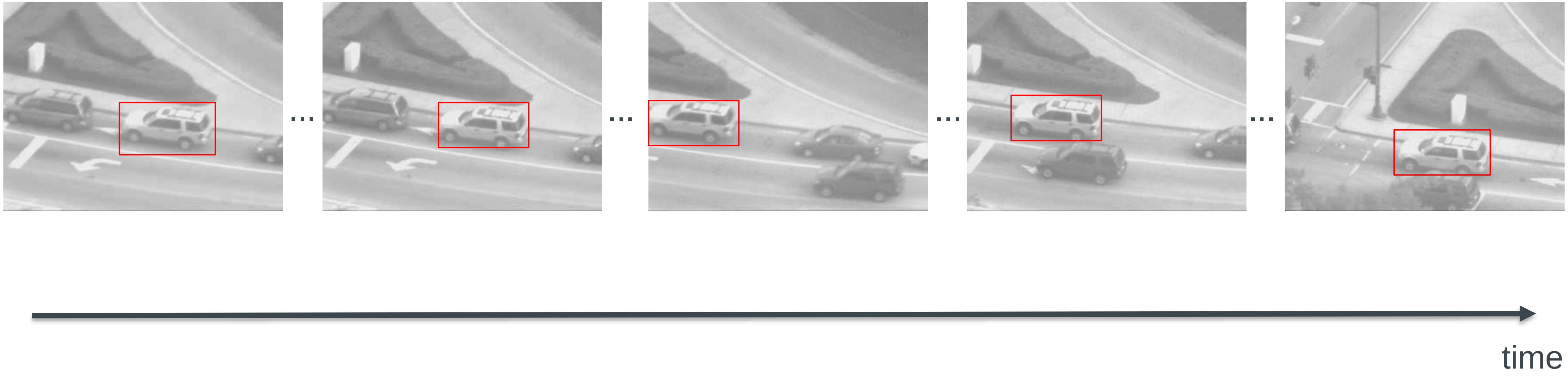
Efstratios Gavves



Arnold W.M. Smeulders

(SINGLE) VISUAL OBJECT TRACKING

Track the target's positions over time in video, given starting box in 1st frame



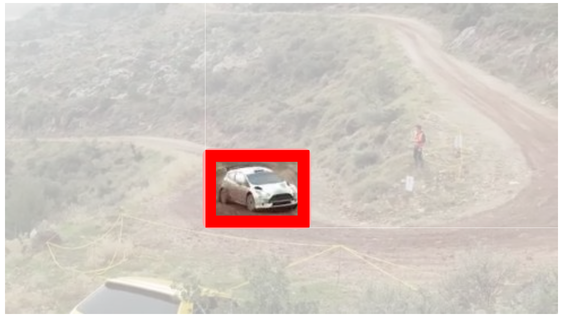
MOTIVATIONS

- Can we learn, *a priori*, invariance which is generically applicable to any object?
  - Online learning: limited, self-inferred data (drifting)
  - Pre-training: rich, reliable data

- Can we solve tracking as an instance search problem?
  - What is tracking: *whether a patch sampled from the frame shows the target?*
    → (relaxation) *which patch in the frame most likely depicts the target?*

Siamese Instance search Tracker (SINT)

candidate patches



Initial patch

$$\operatorname*{argmax}_{j}\ m(initialpatch, patch_j)$$

predicted patch

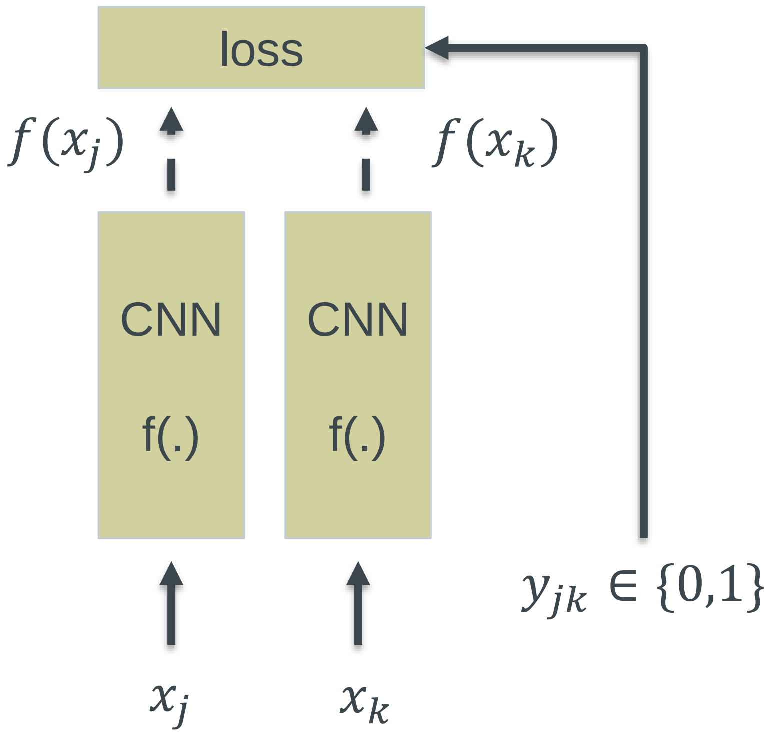
Siamese Instance search Tracker (SINT)

- No online updating
- No occlusion detection
- No geometric matching
- No combination of trackers

Strength is from the matching function $m(\cdot,\cdot)$ learned offline using Siamese network.

MATCHING FUNCTION LEARNING (INVARIANCE LEARNING)

- Operate on pairs. Take two image patches as input and produce the similarity
- Learn once on a rich video dataset with box annotations following an object.
- Once learned, it is applied as is, to videos of previously unseen targets.



Marginal Contrastive Loss:

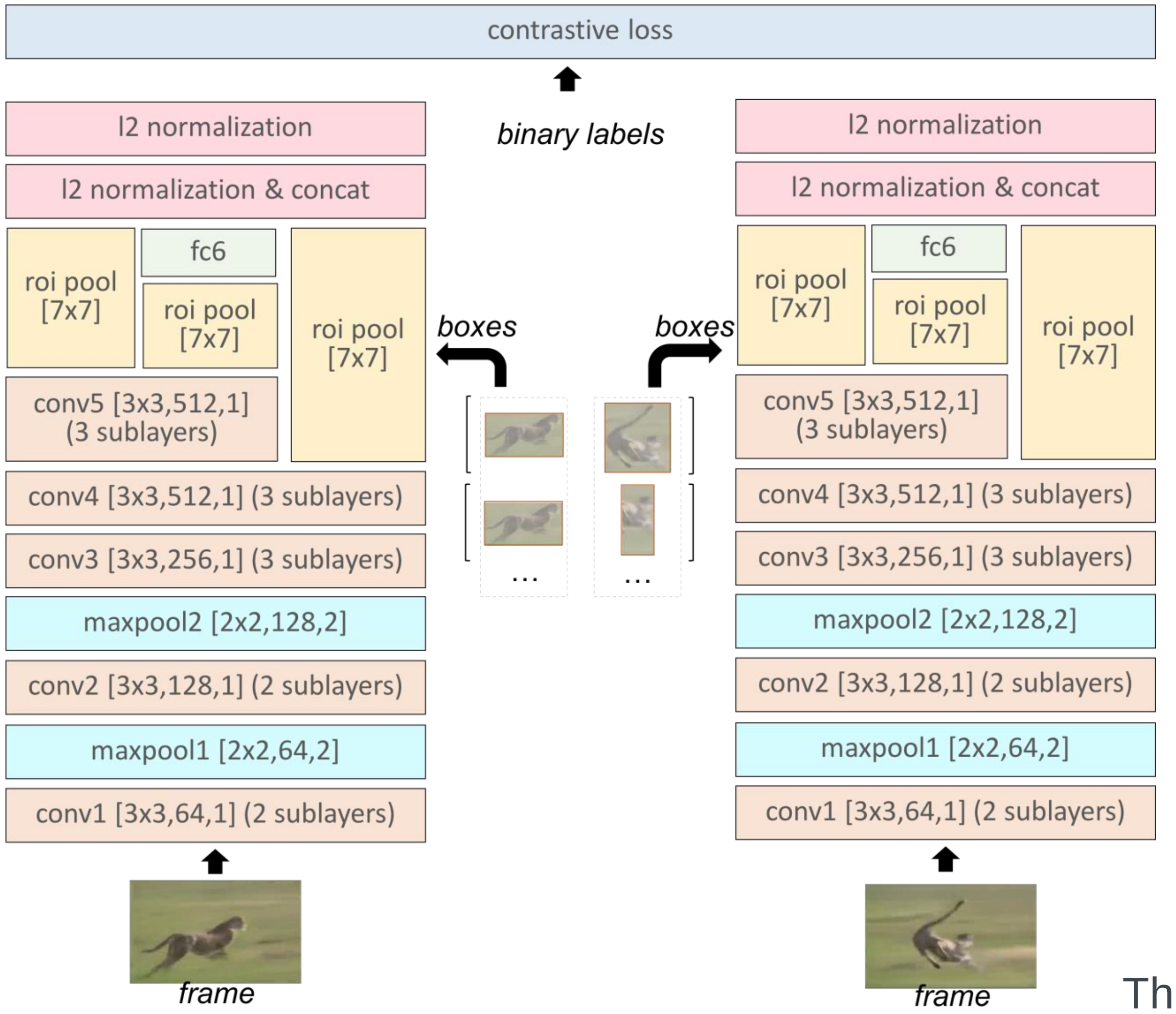
$$L(x_j, x_k, y_{jk}) = \frac{1}{2} y_{jk} D^2 + \frac{1}{2}(1 - y_{jk})\max(0, \sigma - D^2)$$

$$D = \|f(x_j) - f(x_k)\|_2$$

Matching function (after learning):

$$m(x_j, x_k) = f(x_j) \cdot f(x_k)$$

Network Architecture

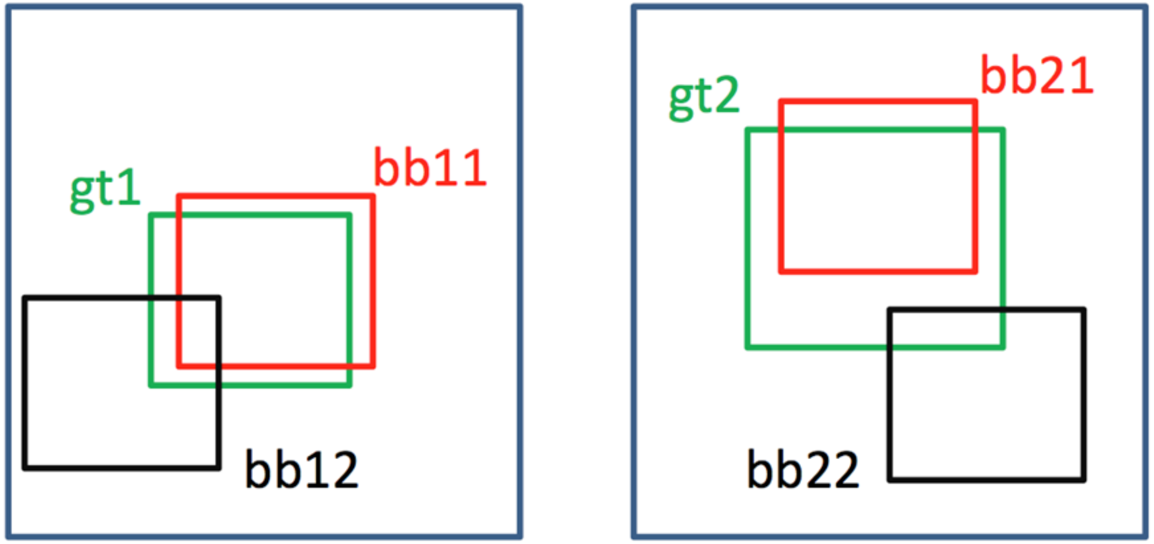


- Very few max pooling → improve localization accuracy
- Region-of-interest (ROI) pooling → process all boxes in a frame in one single pass through the network
- Use outputs of multiple layers (conv4_3, conv5_3, fc6) → to be robust in various situations

The two branches share the parameters.

TRAINING PAIRS

Data: videos of objects with BBox annotation (ALOV)



frame 1 frame 2

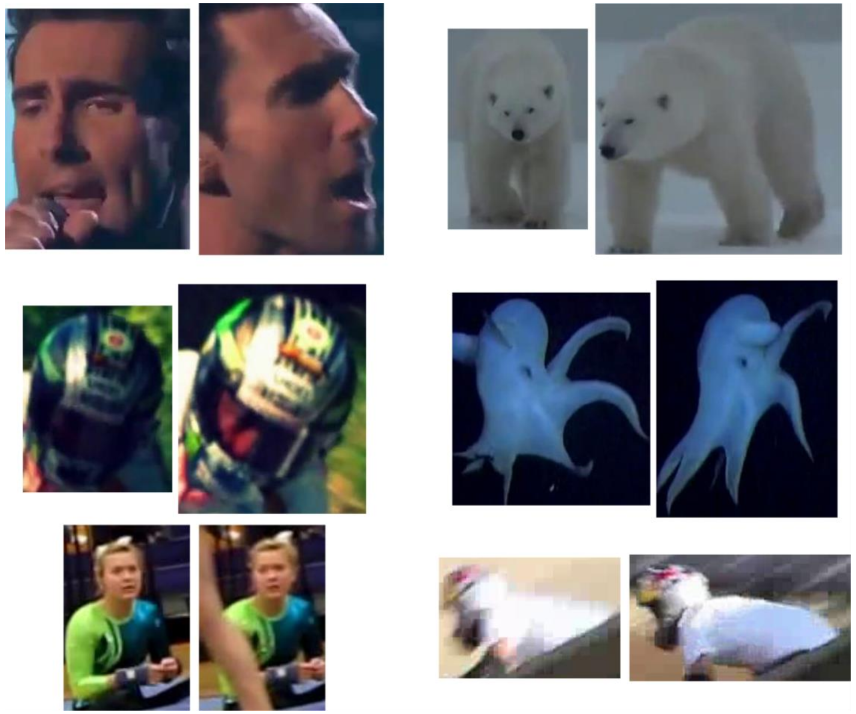
(gt1, gt2, 1)
(gt1, bb21, 1)
(gt1, bb22, 0)
(gt2, bb11, 1)
(gt2, bb12, 0)
...

>0.7, 1
<0.5, 0

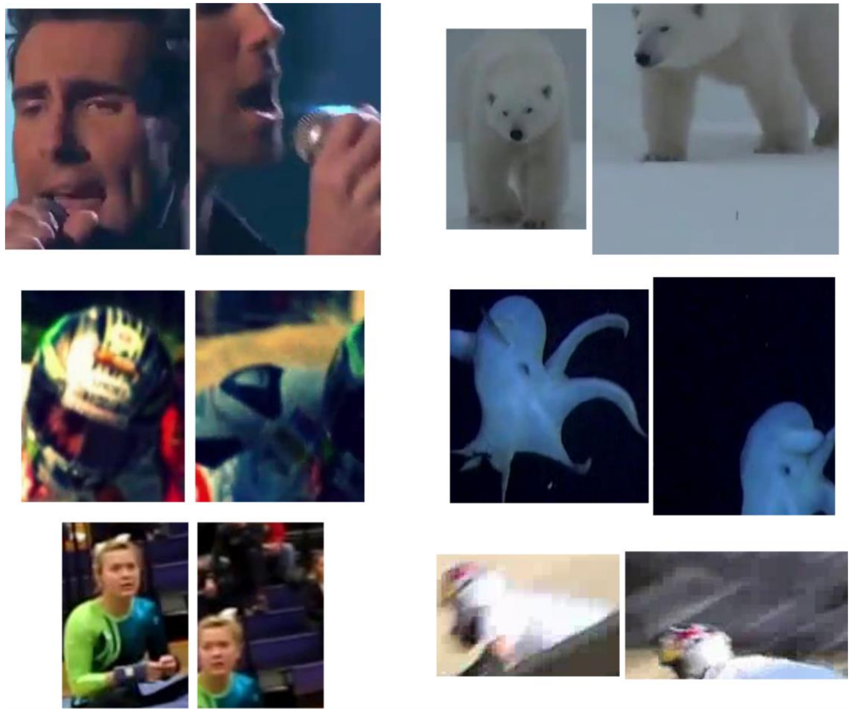
Smeulders *et al*, Visual tracking: An experimental survey, *TPAMI, 2014.*

Training Pairs

- 60,000 pairs of frames for training, 2,000 pairs for validation
- 128 pairs of boxes per pair of frames



positive



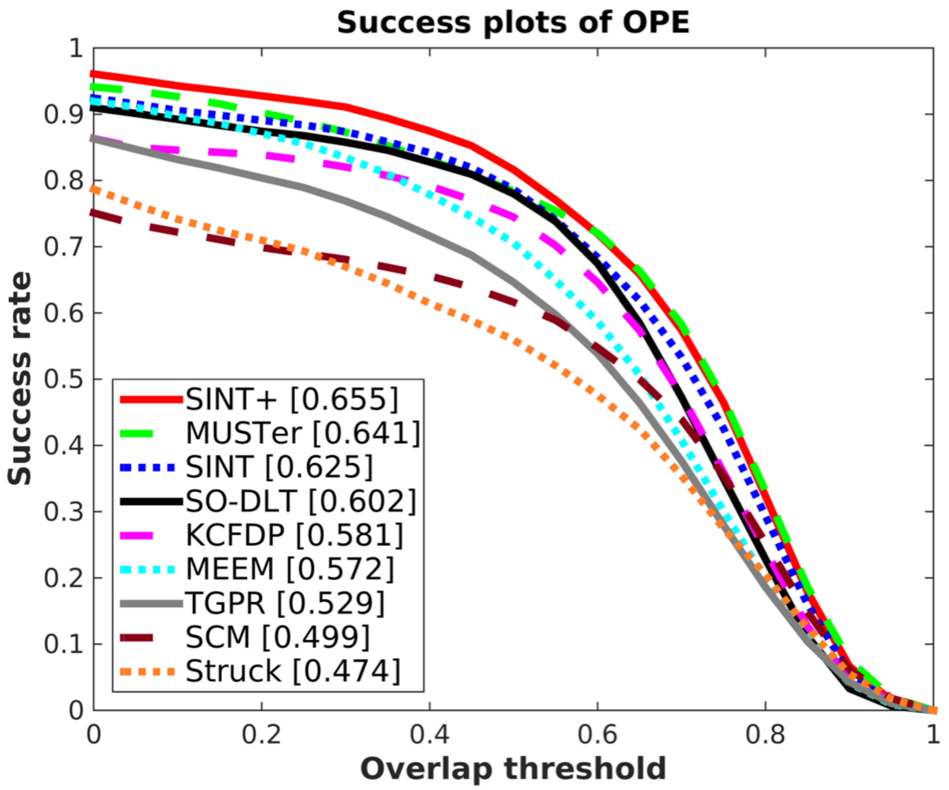
negative

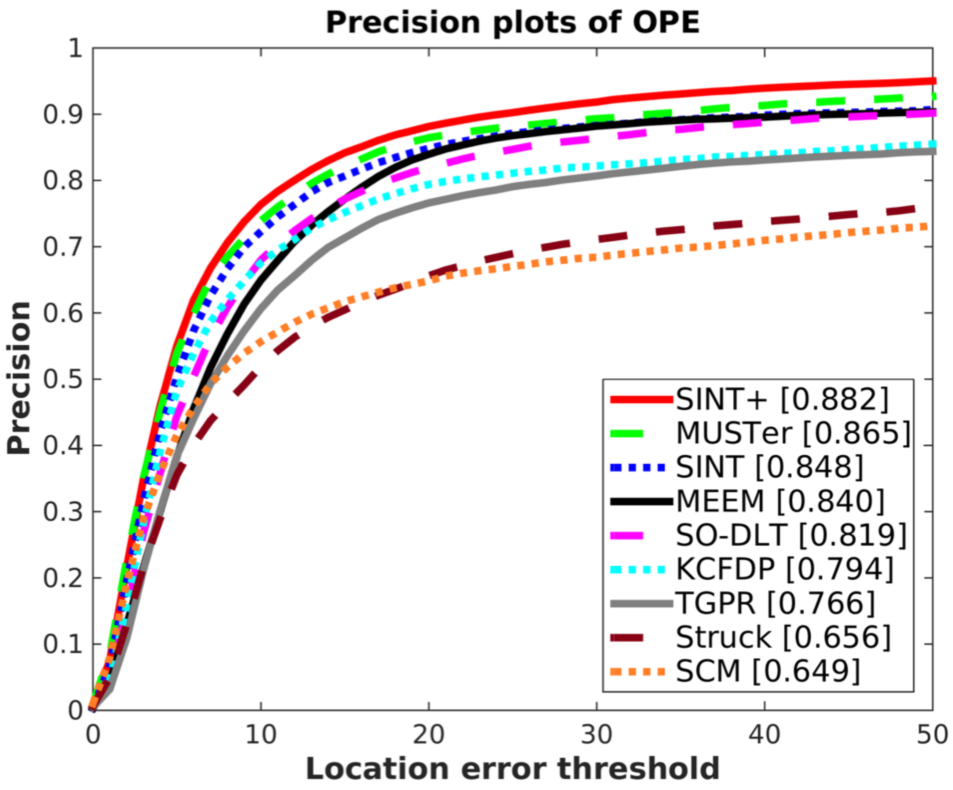
EVALUATION

- Evaluation sets
  - Online tracking benchmark (OTB) [Wu et al, CVPR13]: 51 sequences
  - 6 additional challenging sequences from YouTube

- Evaluation metrics [Wu *et al*, CVPR, 2013]
  - AUC score (box overlap)
  - Precision@20 (center location error)

Results on OTB





SINT+: adaptive sampling range [Want et al, ICCV15] & optical flow to remove motion inconsistent samples

Large potential to improve SINT by integrating advanced online components

Results on 6 Additional Sequences

	MEEM [56]	MUSTer [18]	SINT
Fishing	4.3	11.2	53.7
Rally	20.4	27.5	53.4
BirdAttack	40.7	50.2	66.7
Soccer	36.9	48.0	72.5
GD	13.8	34.9	35.8
Dancing	60.3	54.7	66.8
mean	29.4	37.8	58.1

AUC score

https://youtu.be/K-70sLC6gRU
https://youtu.be/QiCDDQTGcn4
https://youtu.be/r3SgEuuUhDY
https://youtu.be/1GYzl79iXtk
https://youtu.be/gWWHmSCgSno
https://youtu.be/oMG1pJZSno0

FAILURE CASES

similar confusing object

large occlusion

Tracking by Language

- Li et al. Tracking by Natural Language Specification. In CVPR 2017
- Code: https://github.com/QUVA-Lab/lang-tracker
- 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

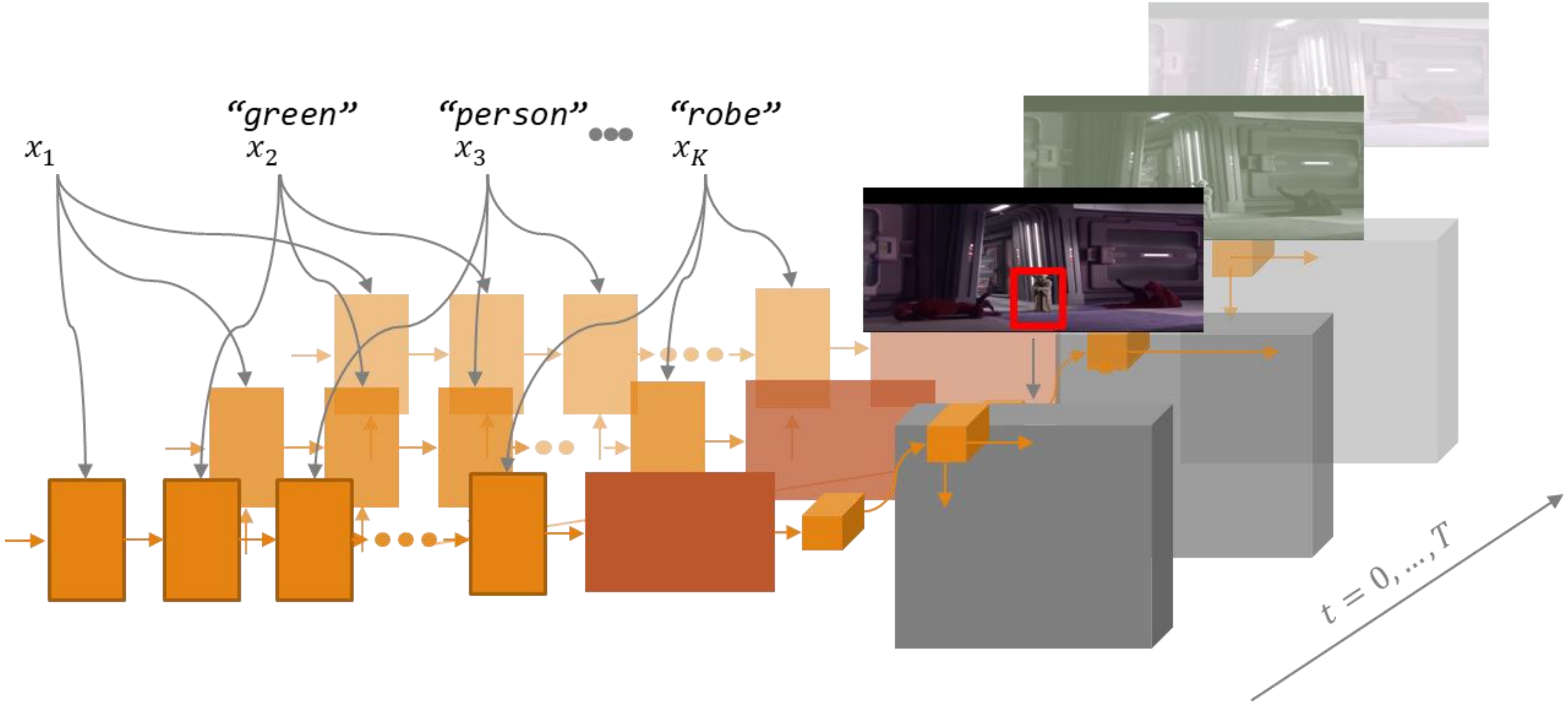
“Man with blue pants”

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?

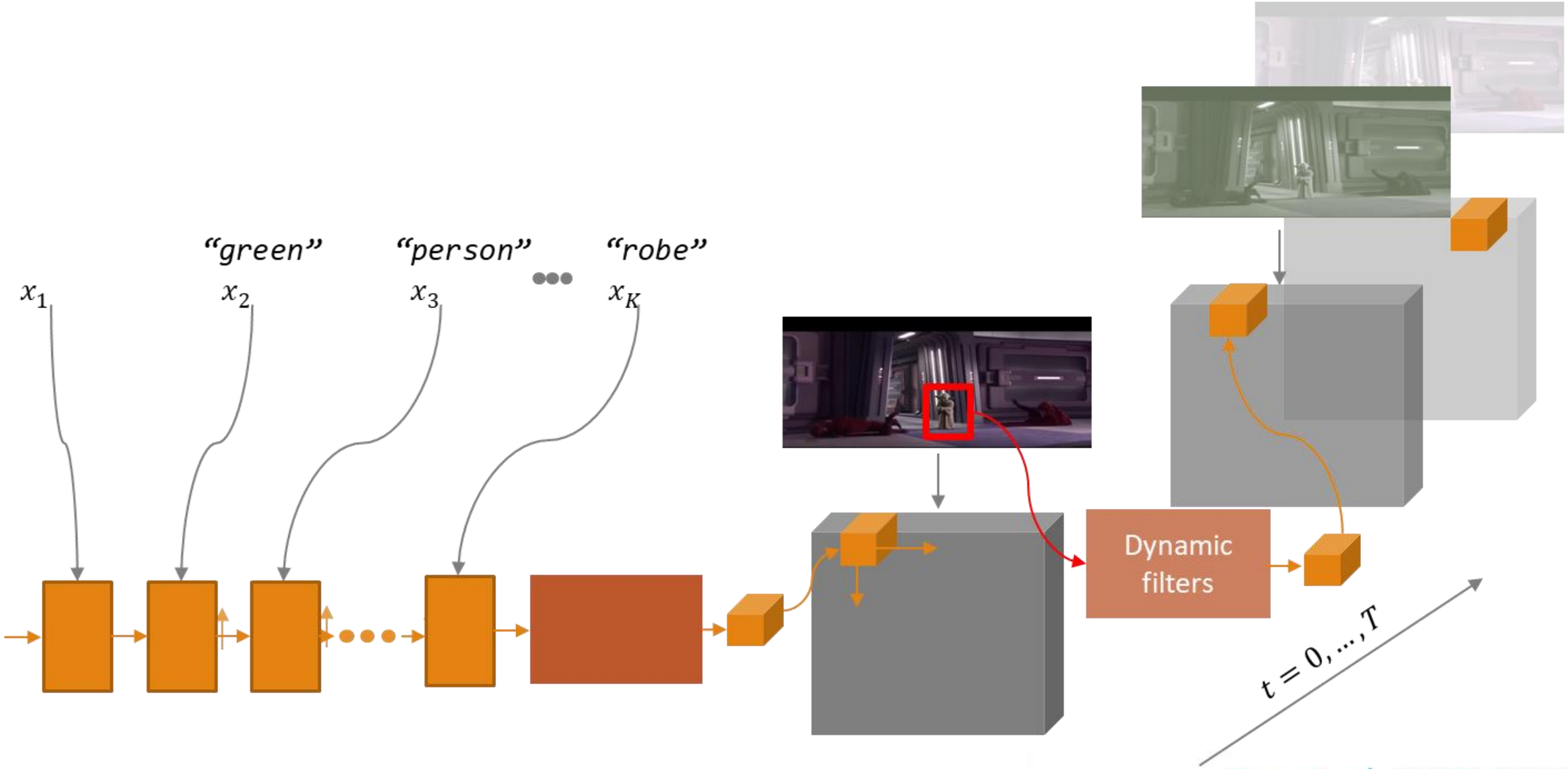
MODEL I: LINGUAL SPECIFICATION ONLY

- Tracking by repeated 'detection'



Model II: Lingual first, then visual

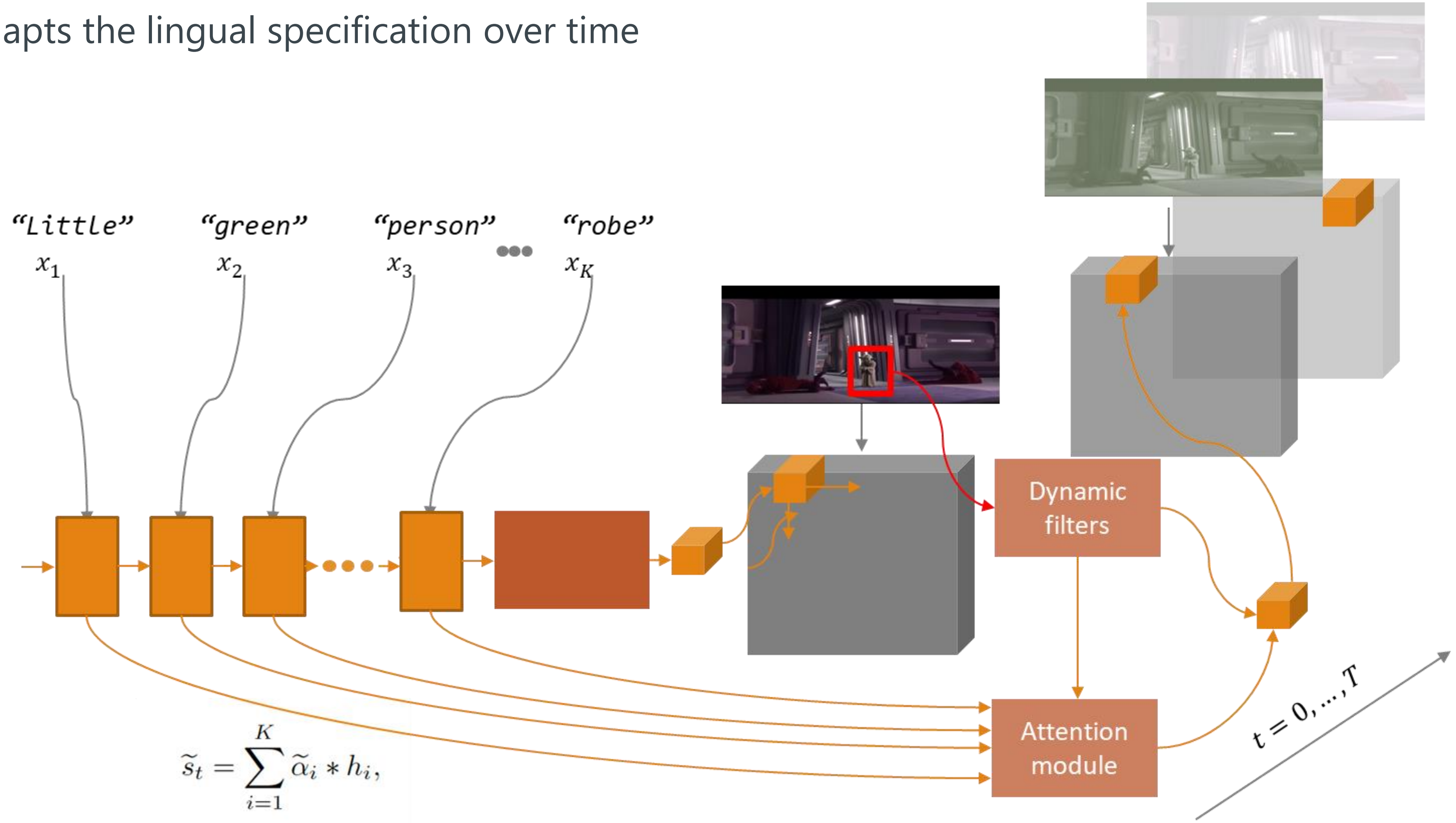
- Use Model I for initialization, then track



Tao *et al.* CVPR 2016

Model III: Lingual & visual

- Adapts the lingual specification over time



“GIRL IN YELLOW SHIRT AND PURPLE PANTS”

Lingual only   Lingual, then visual   Lingual & visual

ACTORS & ACTIONS

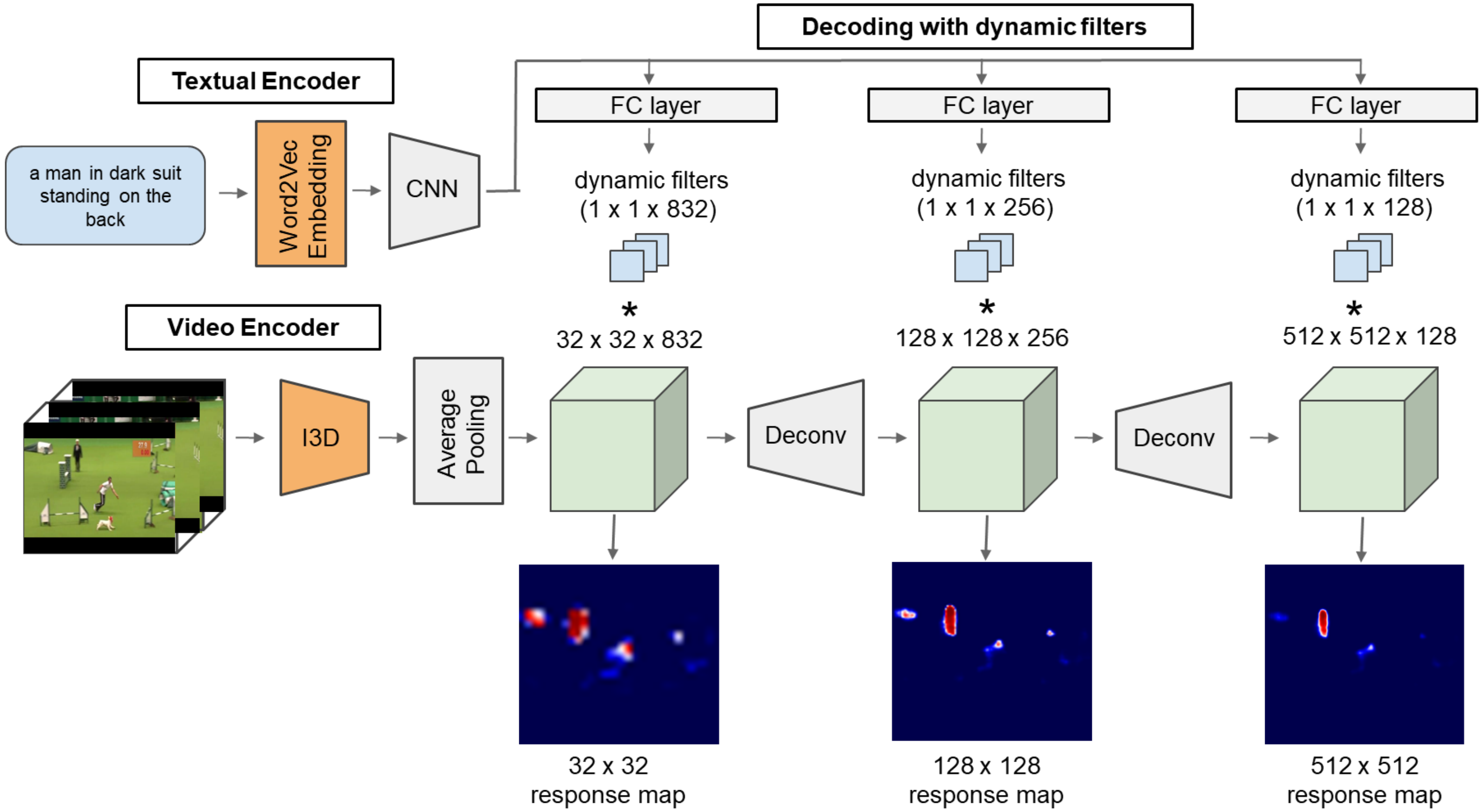
Input video

“woman in purple dress running”

“gray dog running on a leash during dog show”

Action Recognition by Language

- Gavrilyuk et al. Actor and Action Video Segmentation from a Sentence. In CVPR 2018.
- Word2Vec is pre-trained on GoogleNews
- I3D is pre-trained on Kinetics and ImageNet



Conclusions

- Self-supervised spatio-temporal representations still not as good as supervised pretraining
  - But the gap with supervised, pre-trained networks is closing
  - It seems that the temporal domain hides lots of information still

- Better interplay between motion and RGB can help with efficiency and accuracy

- Language and video reinforce each other in multiple way
  - Object tracking, on multiple videos simultaneously and with no first frame requirement
  - Action classification, beyond closed set of predefined labels

THANK YOU!