Revisiting Video Modelling

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

Who Am I?

egavves@uva.nl

@egavves

Efstratios Gavves

- Assistant Professor at the University of Amsterdam
  - Scientific Manager at the QUVA Lab
  - QUVA Lab is a joint Academic-Industry Lab between UVA and Qualcomm
  - Also, teaching Deep Learning (Slides, code available at uvadlc.github.io)
- Co-founder of Ellogon.AI
  - Machine Learning for Clinical Trials and Pharmaceutical Design
  - Partnering up with the Dutch National Cancer Institute against oncology
    - One of the biggest research centers worldwide with huge data
  - If interest, please come find me

UNIVERSITY OF AMSTERDAM

QUVA Deep Vision Lab

ELLOGON.AI

Video Modelling Today: Short

- Spatiotemporal Encoders: convolve up to a few dozen frames
- Action Classification: process up to few seconds
- Efficient Video Models: not really exists
- Self-supervised Learning: predicting immediate spatio-temporal context

VIDEO MODELLING TOMORROW: LONG

- Spatiotemporal Encoders: thousands of frames
- Sequence Learning of Complex Actions: dozens of minutes or hours long
- Efficient Video Models: scaling up cannot be done without contemplating efficiency
- Self-supervised Learning: from spatio-temporal context to temporal properties

Video Temporal Modelling of tomorrow about encoding transitions over long term and dynamics …
… instead of encoding short spatio-temporal (static) patterns

Video Dynamics Learning

- When it comes to long or streaming videos the important questions are:

Is there a difference between a video sequence and other types of sequences?
What are the meaningful dynamics of the video content and how to capture them?
How to encode the meaningful dynamics in a “non-catastrophic forgetting” manner?
How to encode multiple temporal complexities of dynamics?

Can we design video specialized models and architectures for dynamics?
Not models that extend our favorite 2D convnet

Self-Supervised With Odd-One-Out

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

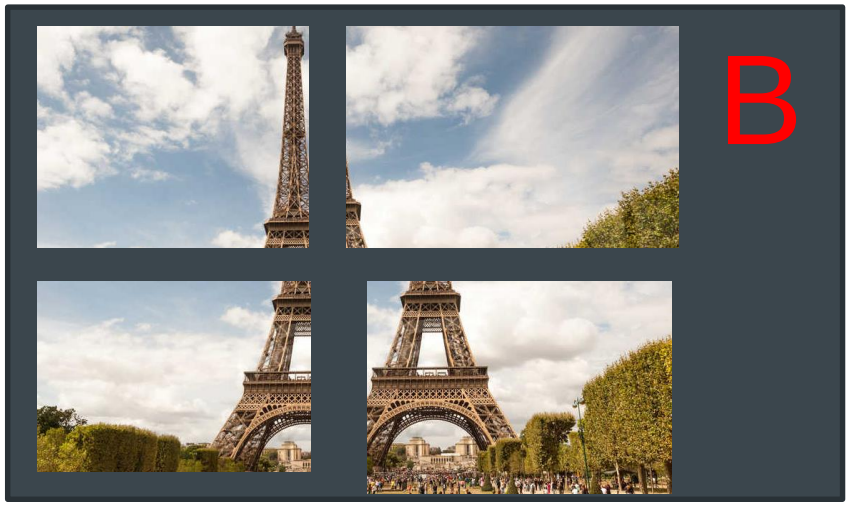
Basura Fernando

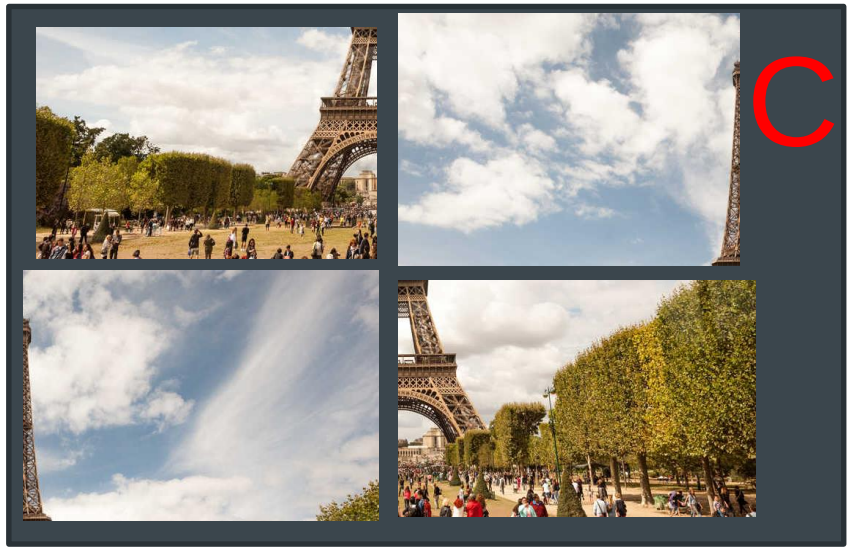
Hakan Bilen

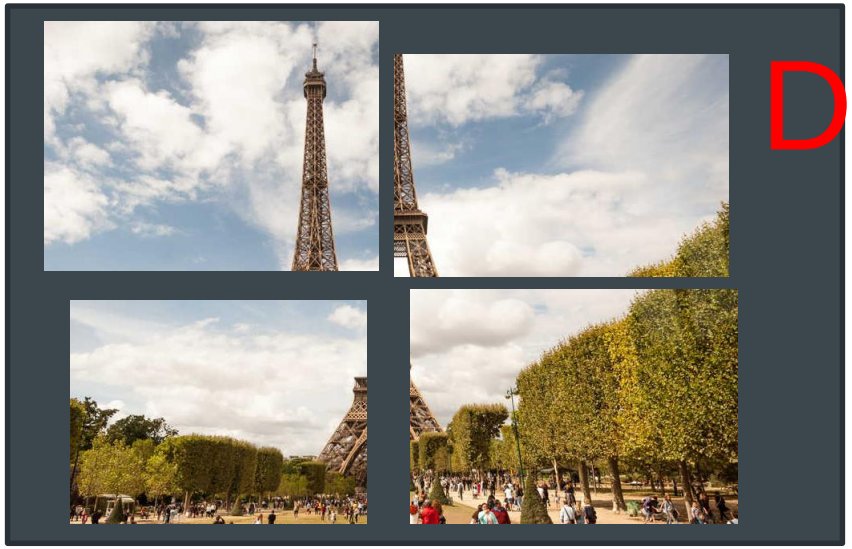
Efstratios Gavves

Stephen Gould

Find the Wrong Input







And Temporally




or



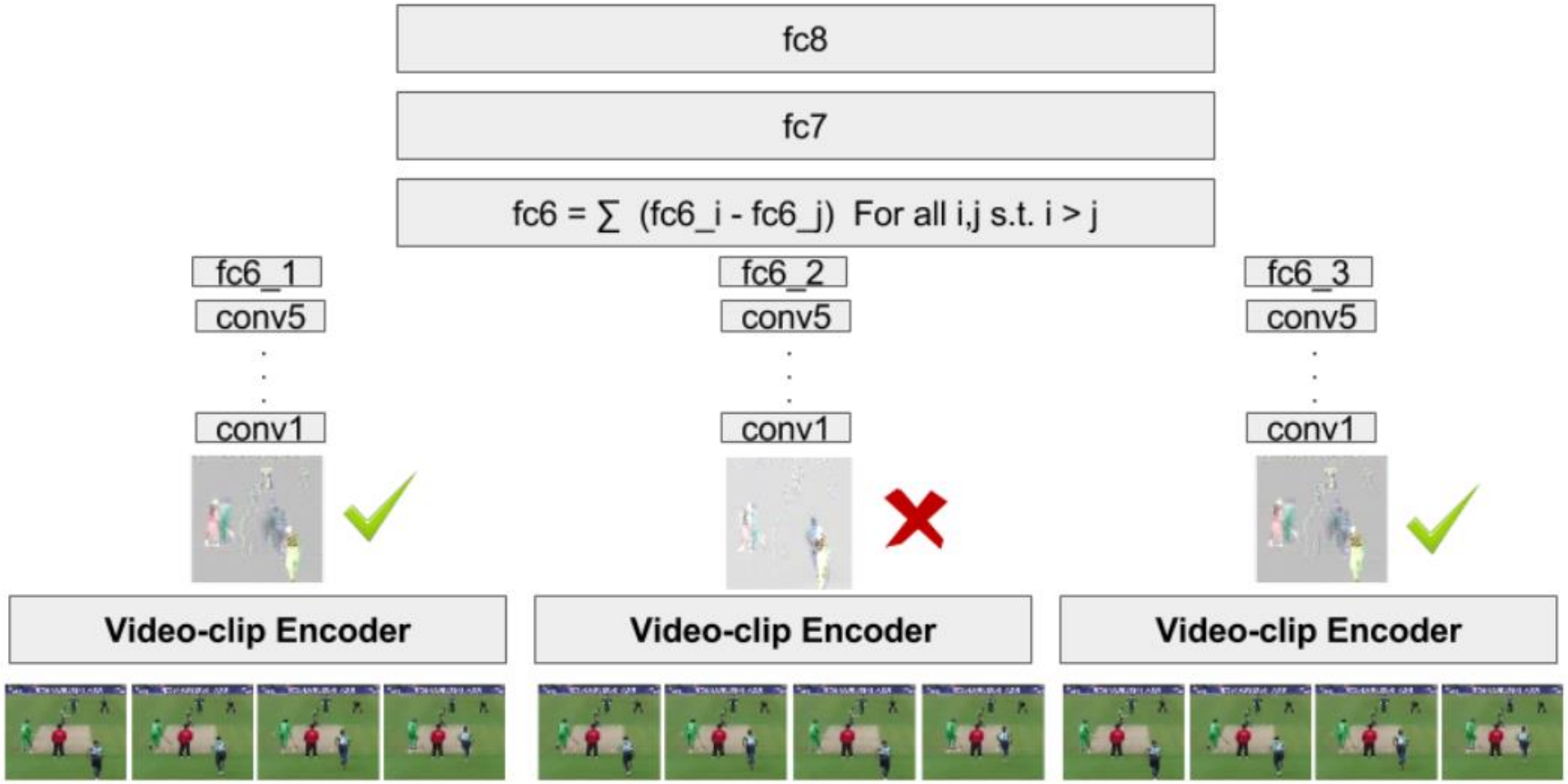

And Temporally




or

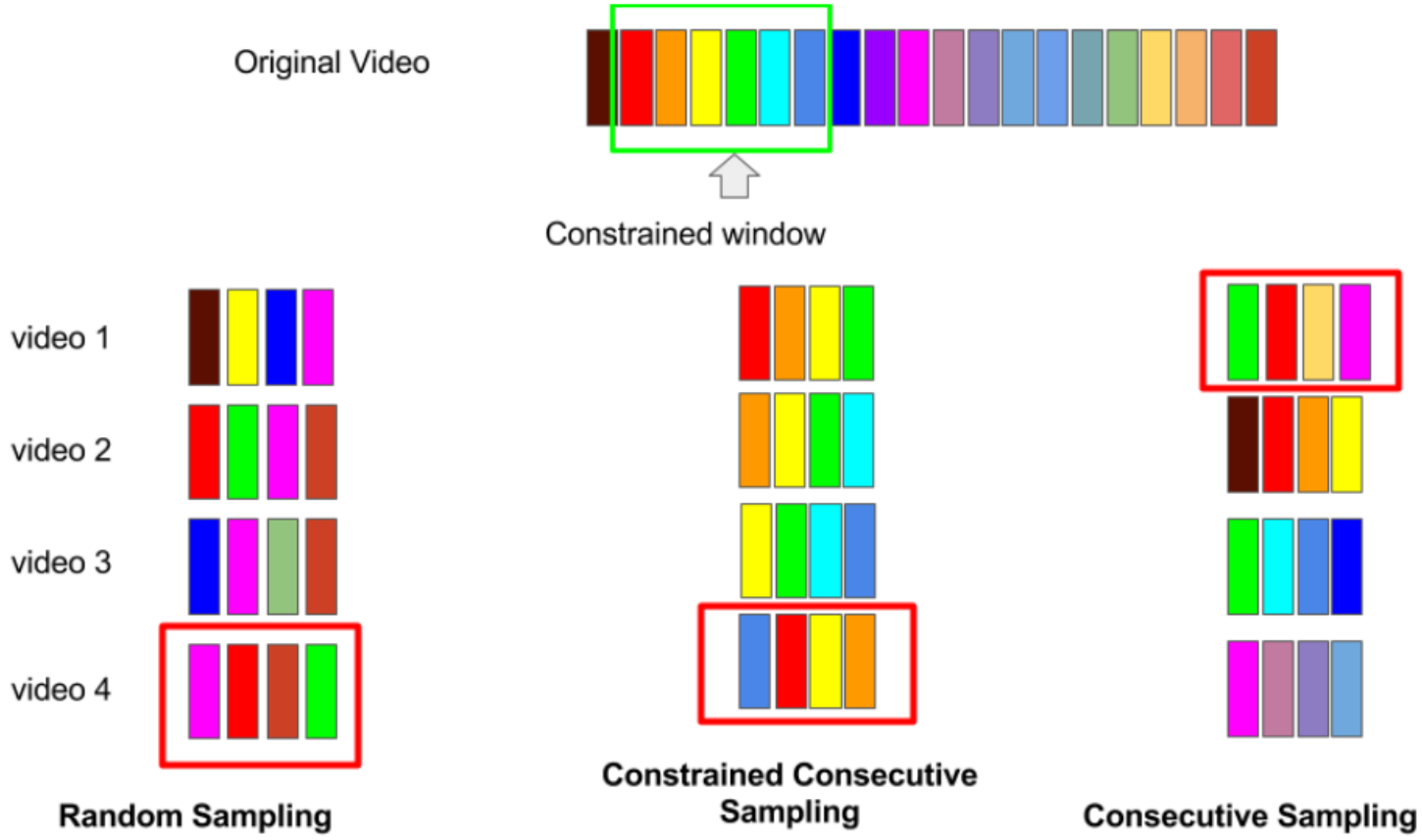



Odd-one-out Learning Model

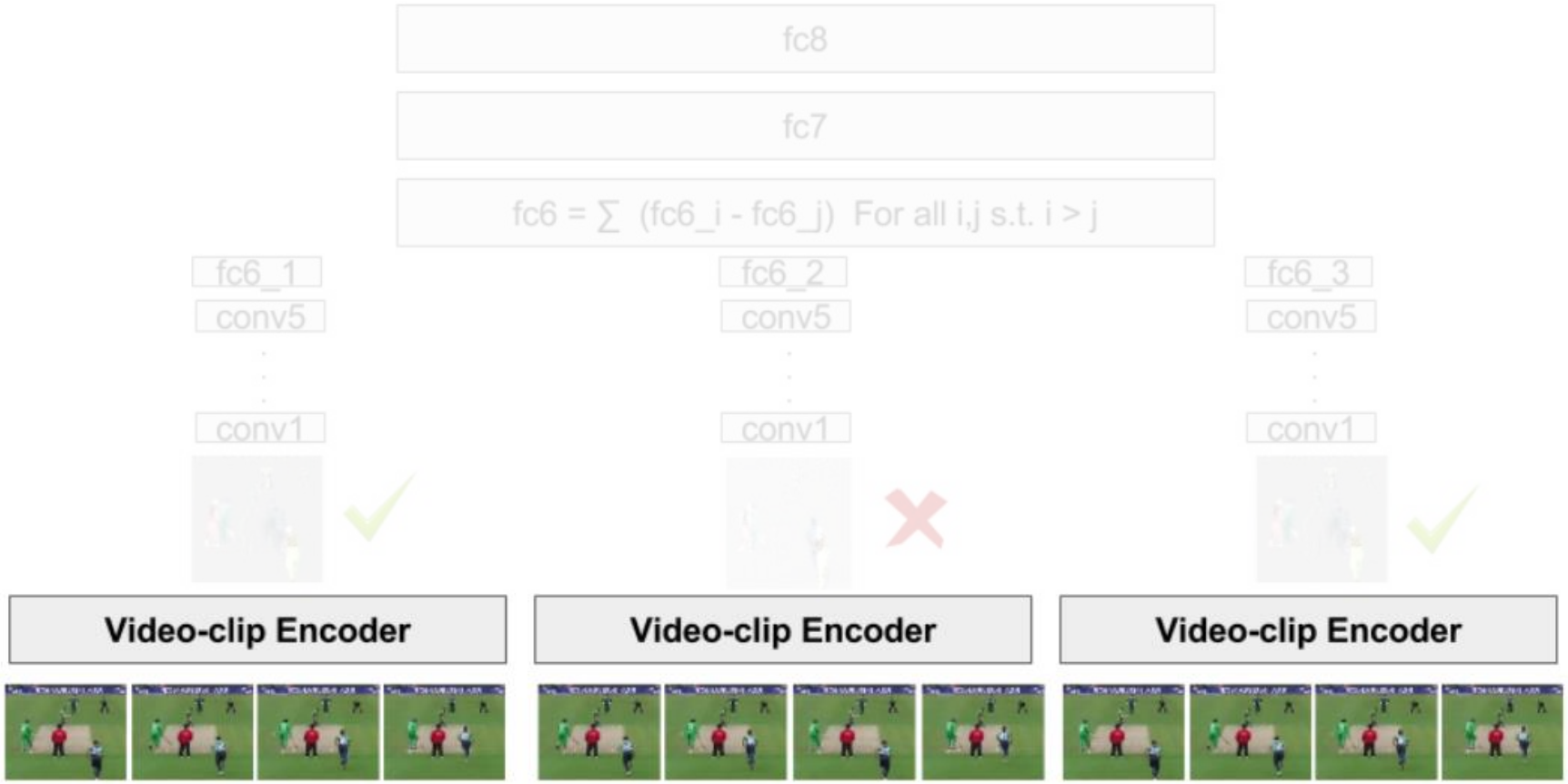


How to Sample Frames?

How to Sample Frames?



How to Encode Frames



How to Encode Frames

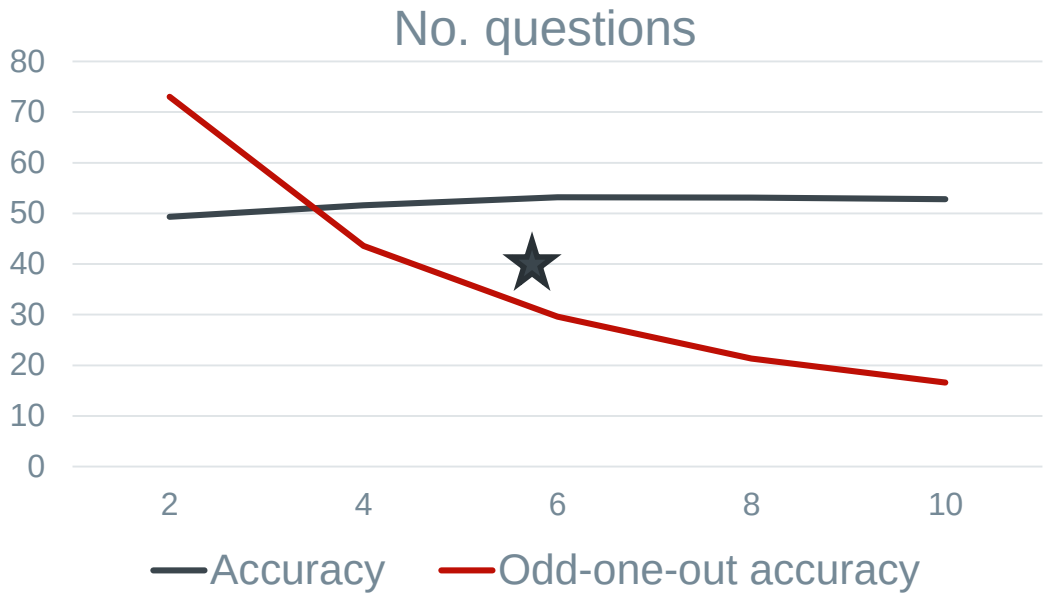
Dynamic images

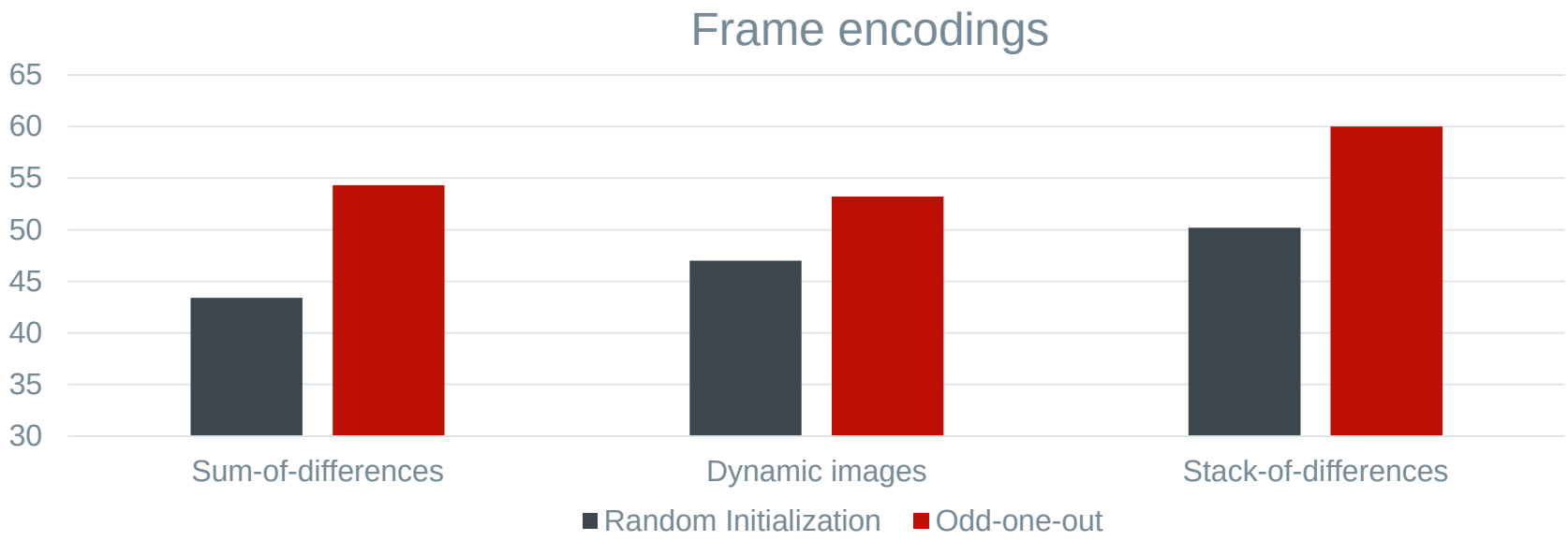
Sum of differences

Stack of differences

EXPERIMENTS

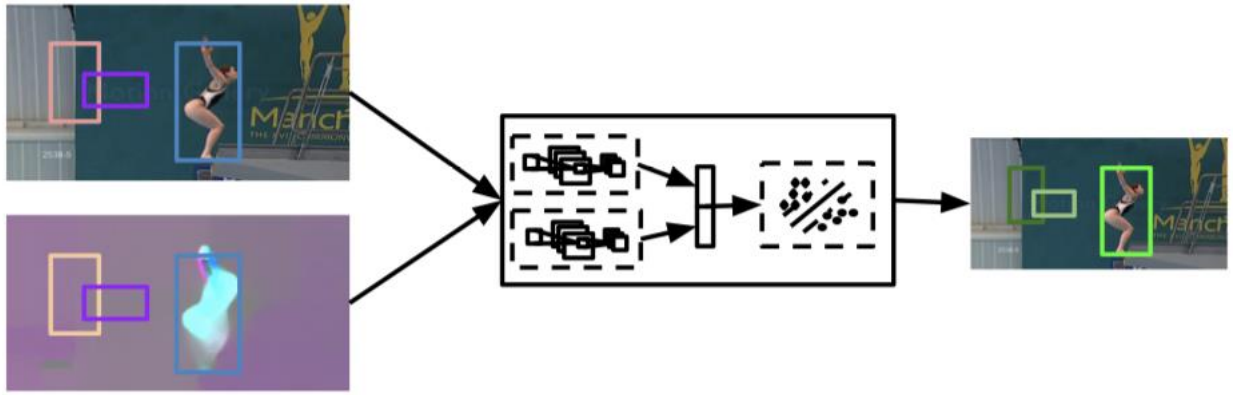
Sampling	Accuracy	Odd-one-out acc.
Consecutive	50.6	27.4
Constrained consecutive	52.4	29.0
Random	53.2	29.6





TWO-STREAM

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

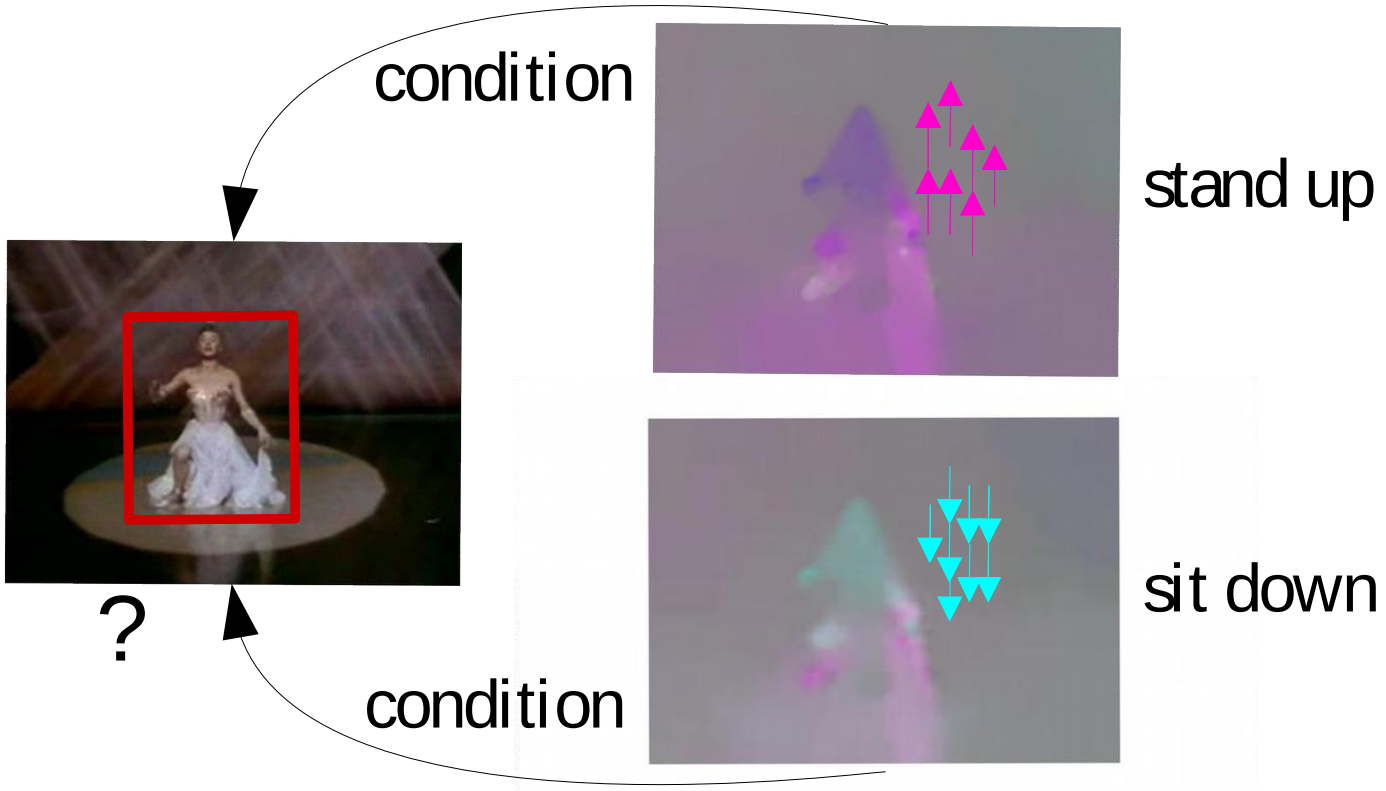


- Doubles computation and parameters for modest accuracy gain.

Simonyan & Zisserman NeurIPS14

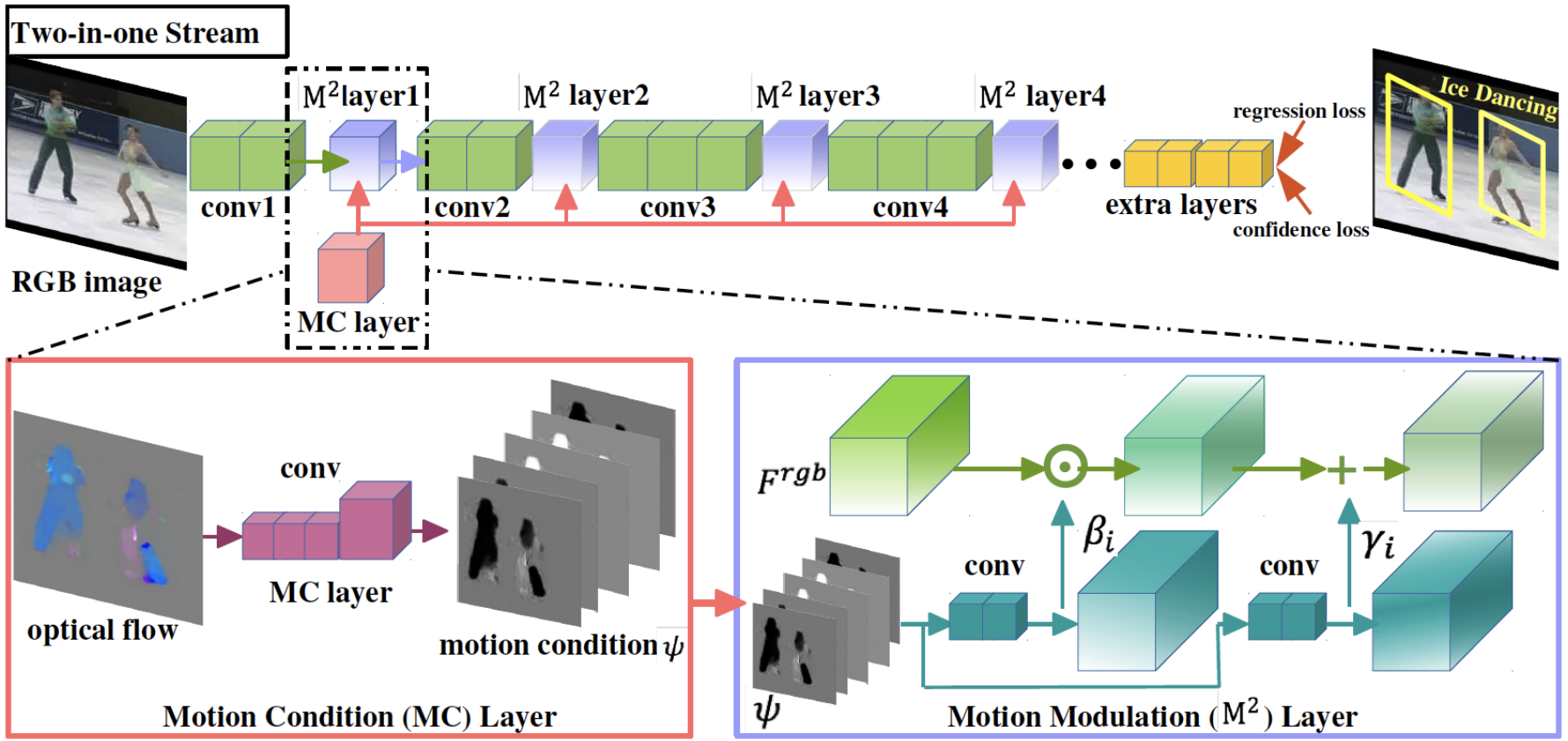
KEY IDEA

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



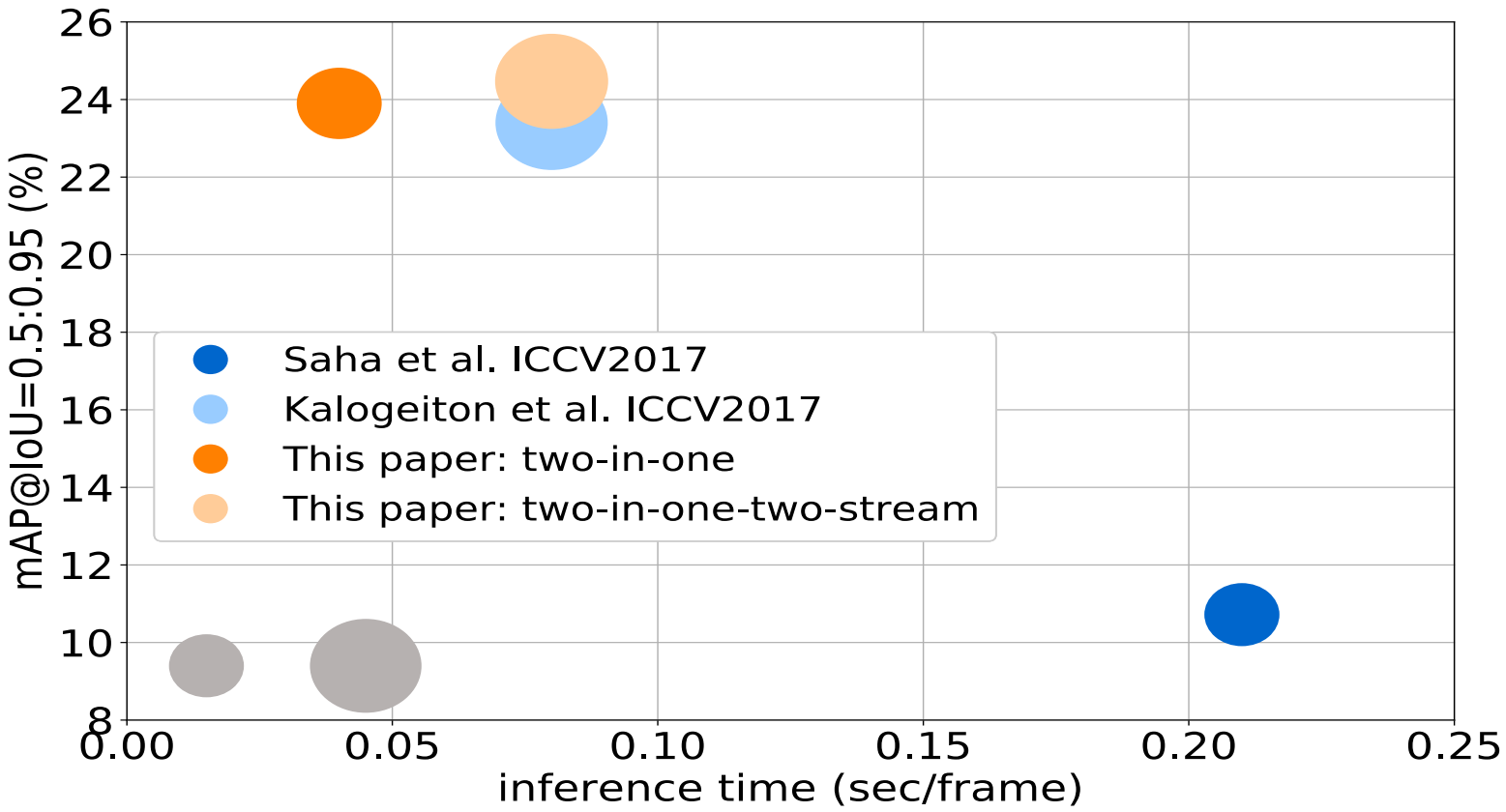
Two-in-one Stream

- 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.



THANK YOU!

Conclusions

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