What have we learned from deep representations for action recognition?

work with

Axel Pinz
Graz University of Technology

Richard P. Wildes
York University, Toronto

Andrew Zisserman
University of Oxford

Outline

- Two-Stream Architectures for Action Recognition

What have we learned in:

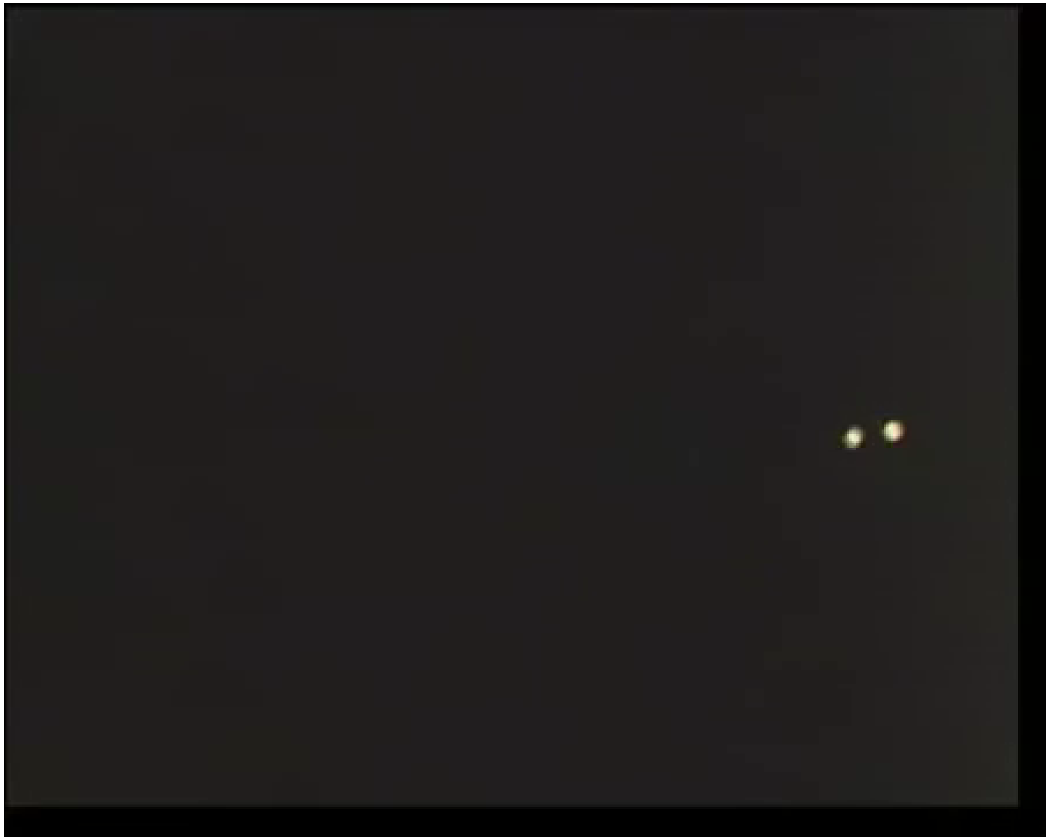
  - Fusion of appearance and motion streams
  - Long-term feature aggregation

- Visualization of Two-Stream representations

Intuitions for why:

  - Explicit motion models perform better
  - Fusion leads to good feature abstractions

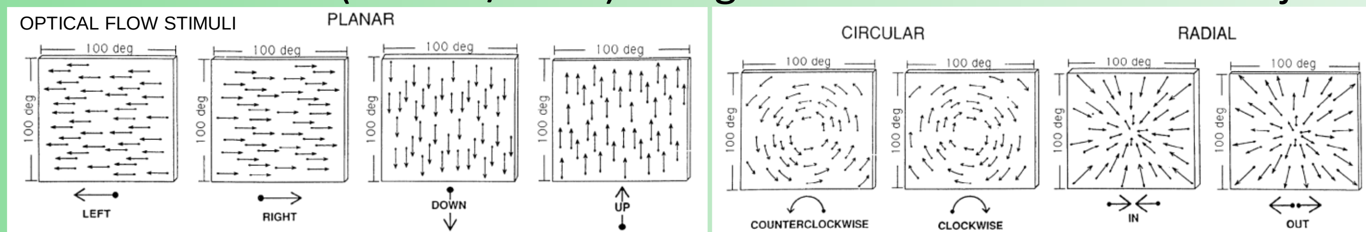
➔ Amazing what the brain can do without appearance information

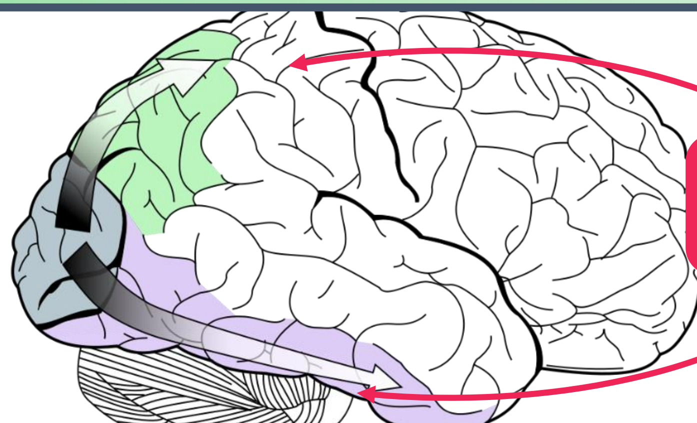


Sources: Johansson, G. “Visual perception of biological motion and a model for its analysis.” Perception & Psychophysics. 14(2):201-211. 1973.

Motivation: Separate visual pathways in nature

➔ Dorsal stream ('where/how') recognizes motion and locates objects





➔ "Interconnection" e.g. in STS area

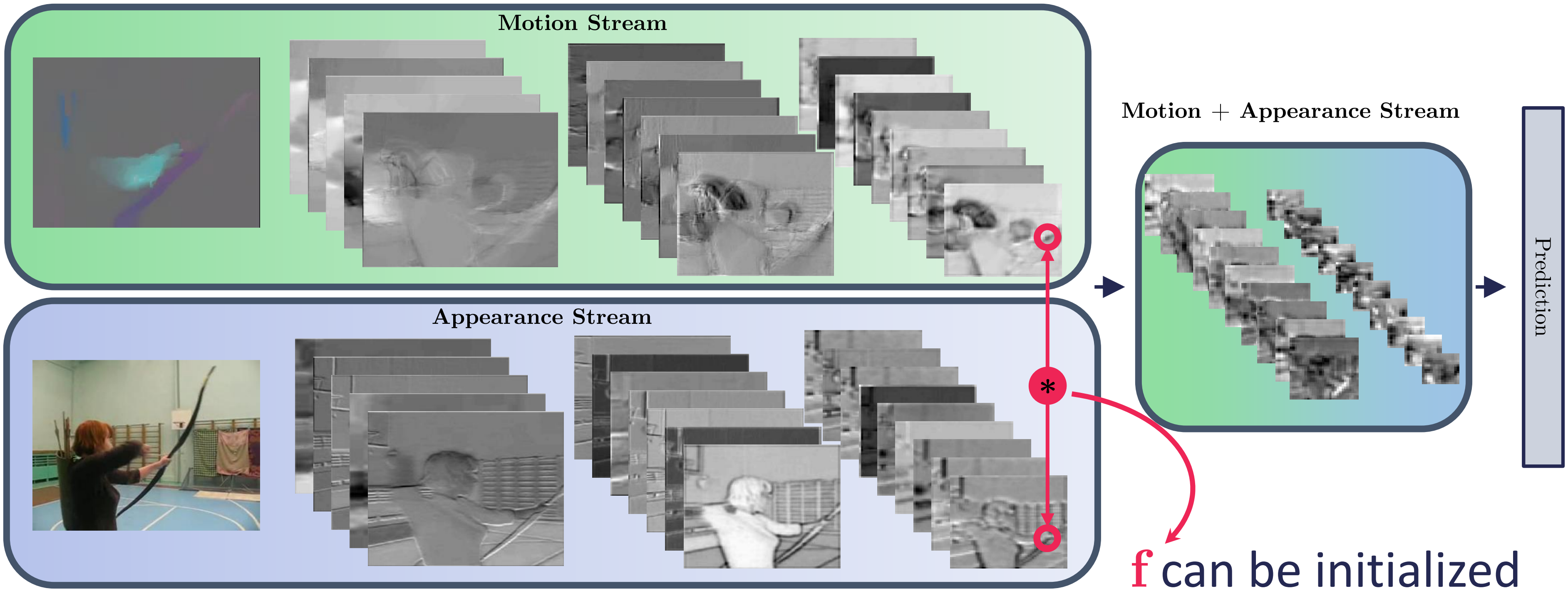
➔ Ventral ('what') stream performs object recognition

Sources: *"Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli." Journal of neurophysiology 65.6 (1991).*
"A cortical representation of the local visual environment", Nature. 392 (6676): 598–601, 2009
https://en.wikipedia.org/wiki/Two-streams_hypothesis

Convolutional Two-Stream Network Fusion

- We study a number of ways of fusing two-stream ConvNets

[Simonyan & Zisserman, NIPS'14]

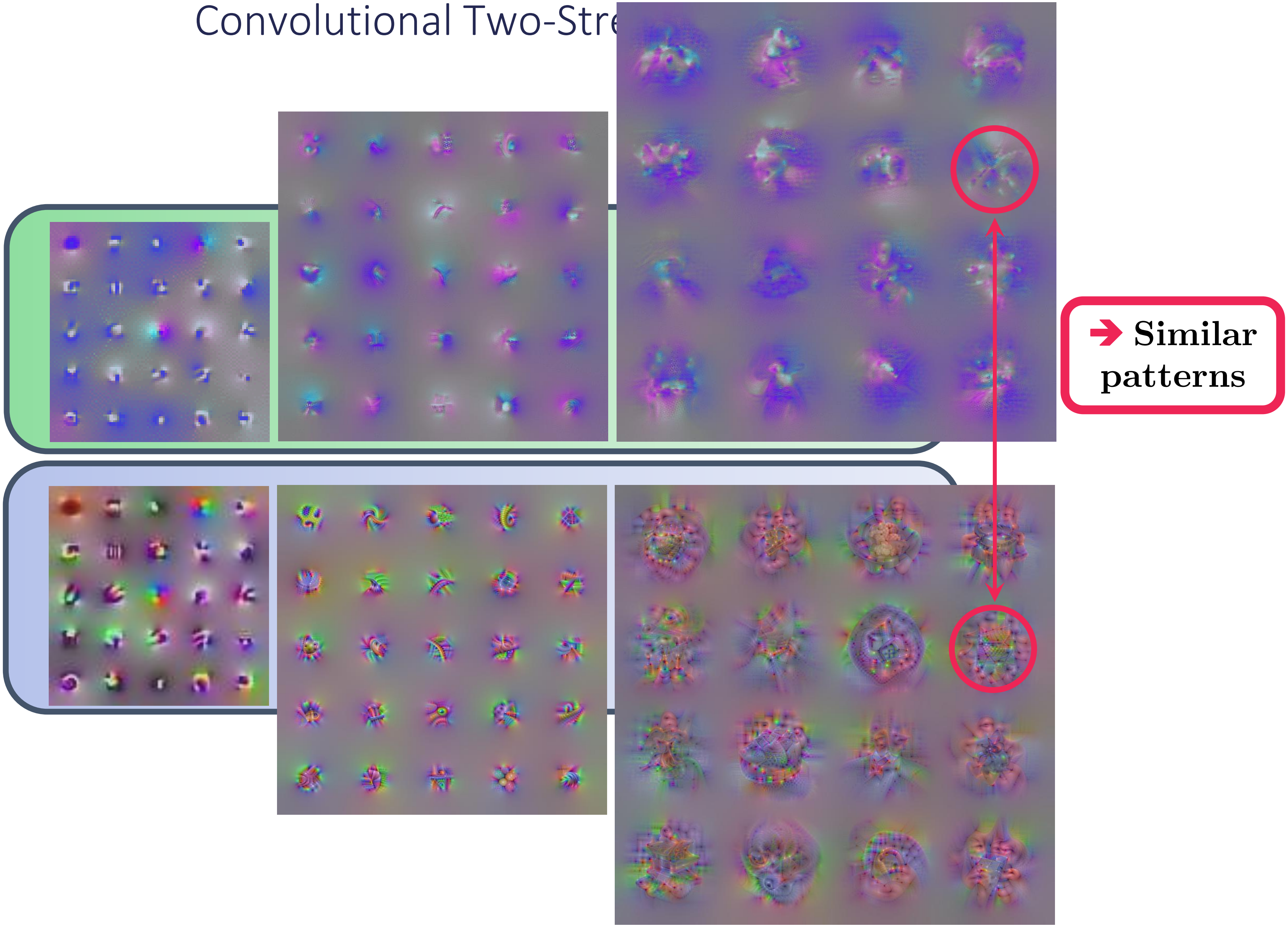


f can be initialized as a sum kernel + feature identity mapping

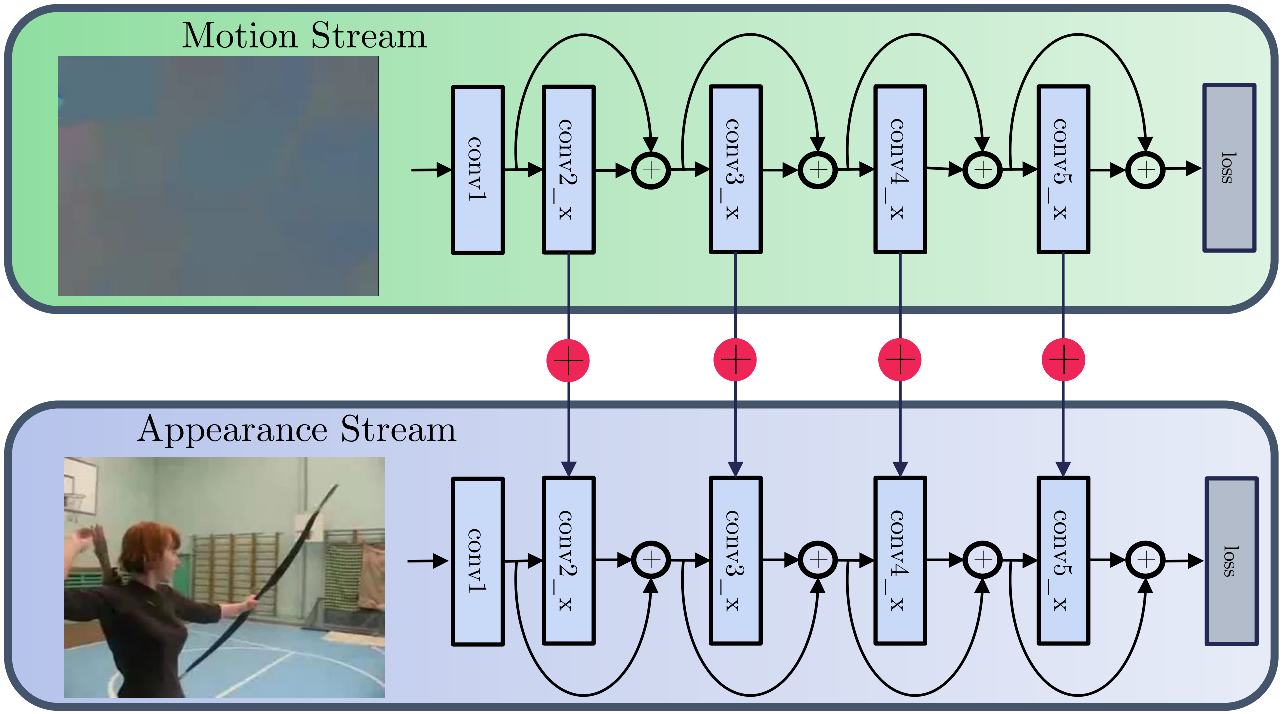
- Sum fusion works surprisingly well

Feichtenhofer, Pinz, Zisserman, CVPR 2016

Convolutional Two-Stre



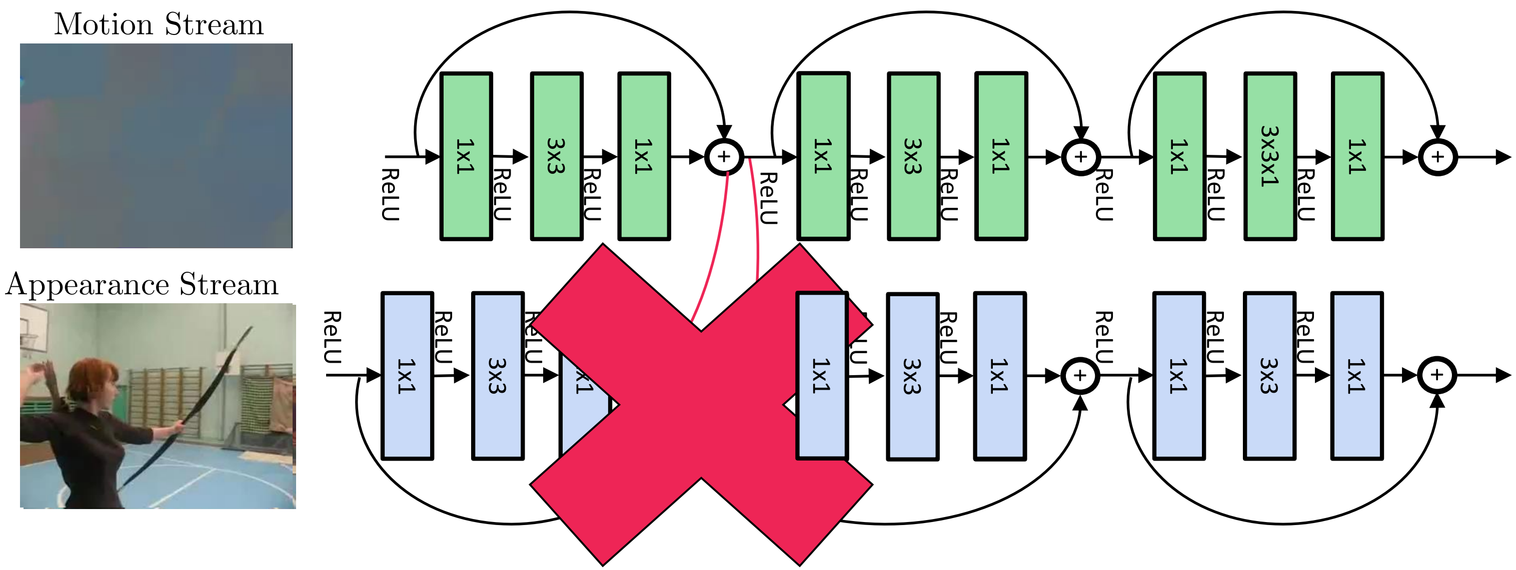
Spatiotemporal Residual Networks



- ST-ResNet allows the hierarchical learning of spacetime features by connecting the appearance and motion channels of a two-stream architecture.
- Though, naive fusion does not work.

Feichtenhofer, Pinz, Wildes, NIPS'16 & CVPR'17

Fusing Two-Stream ResNets & Injecting Temporal Filters

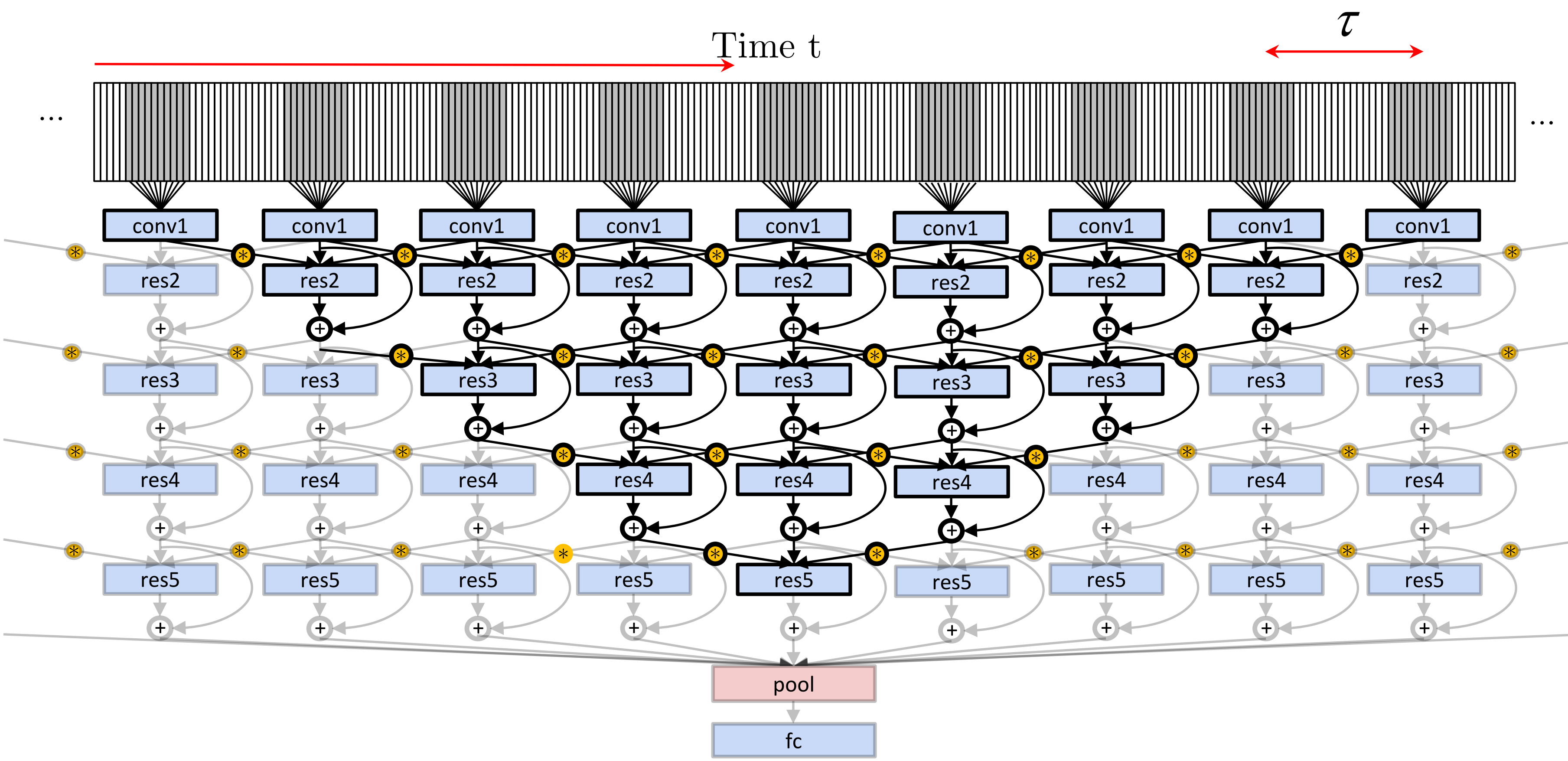


- ResNets for the spatiotemporal domain by introducing residual connections in two ways
  1. Residuals between the motion and appearance pathways to allow spatiotemporal interaction between the streams
  2. Transformation * of pretrained image ConvNets by filters initialized as residuals in time

- Our most recent work (@CVPR'17) reconsiders the combination these approaches more thoroughly to increase our understanding of how these techniques interact.

Feichtenhofer, Pinz, Wildes, NIPS'16 & CVPR'17

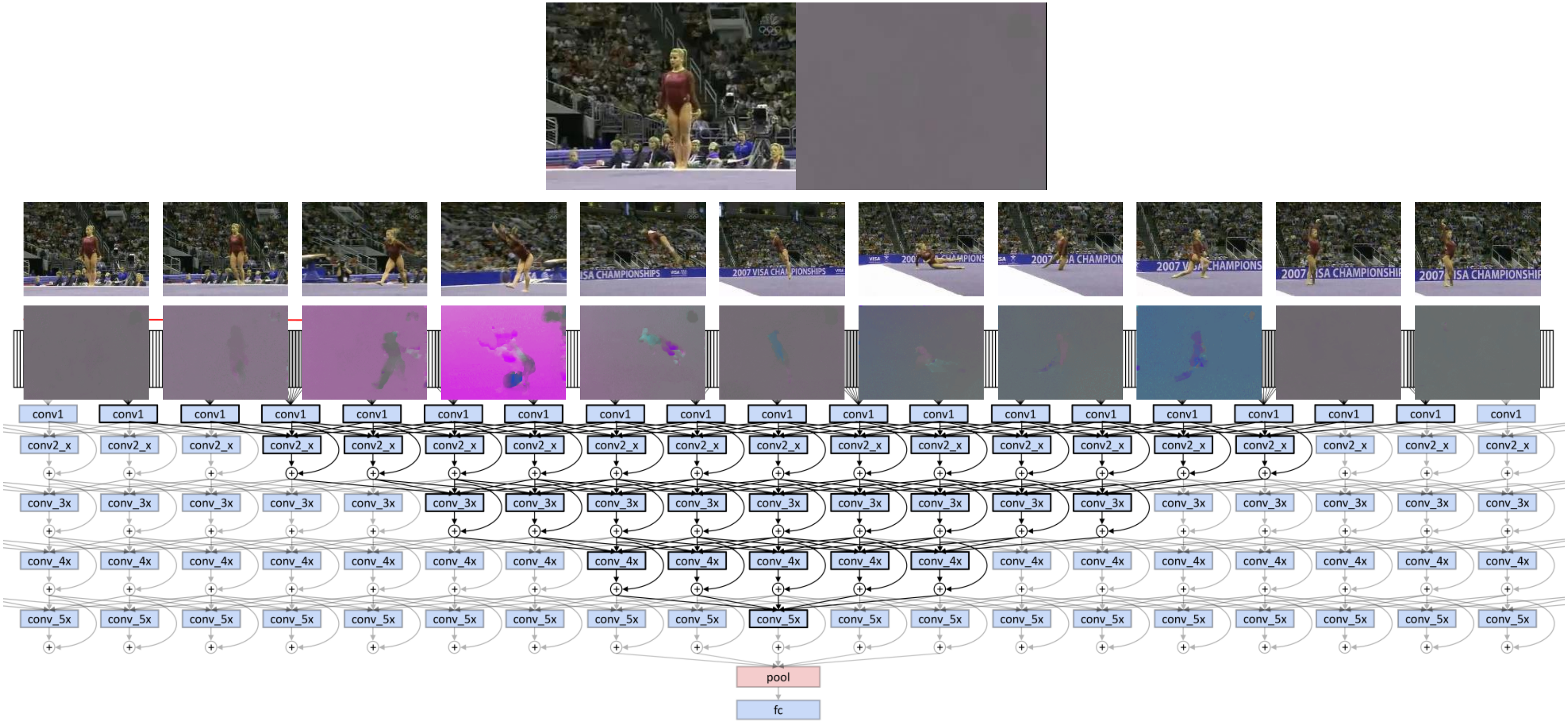
Increasing the temporal receptive field of ResNets



○ The temporal receptive field is modulated by the temporal filters * and input stride τ

Feichtenhofer, Pinz, Wildes, NIPS'16

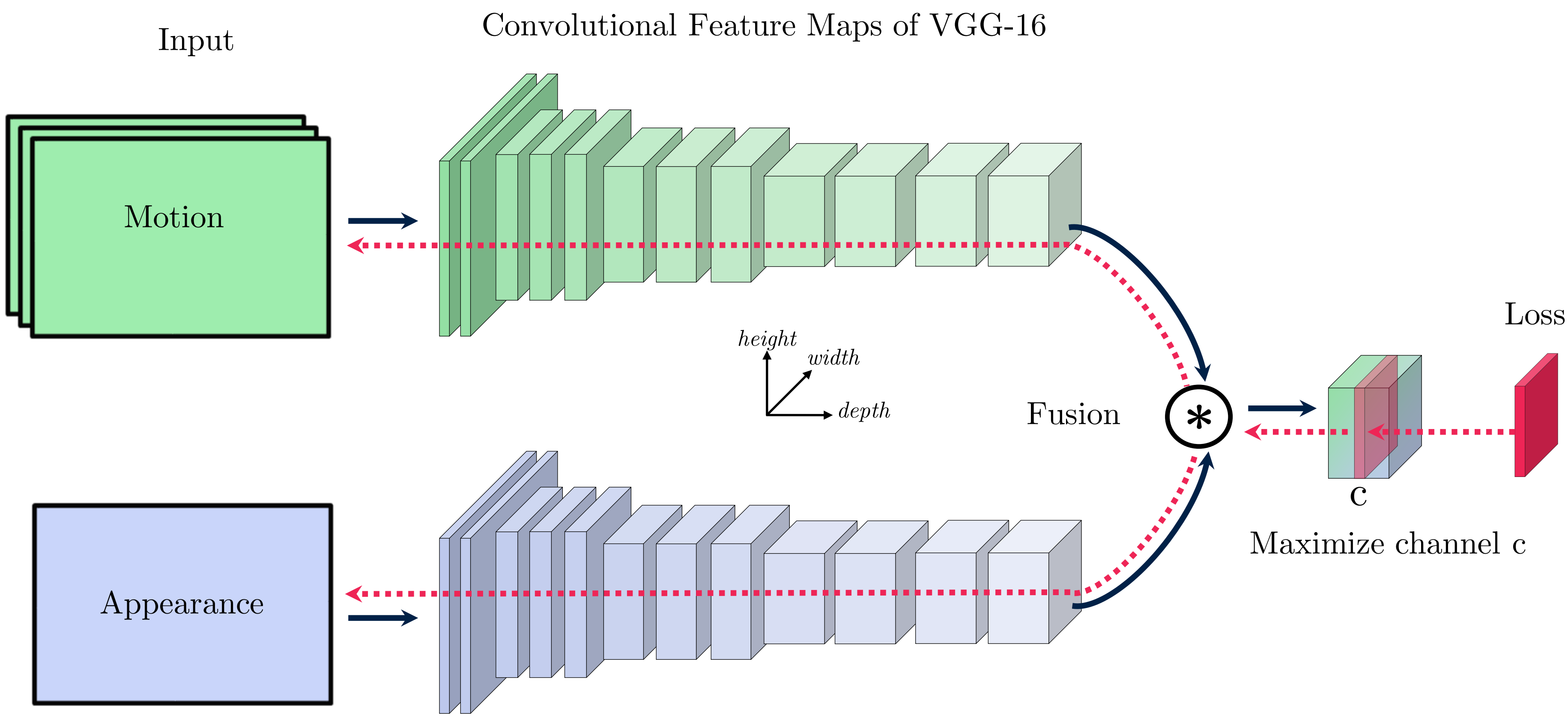
Transforming spatial filters to spatiotemporal ones



- Chaining temporal filters supports *hierarchical* learning of long-term correspondences between features of the appearance and motion stream.
- For example, if the stride is set to $\tau = 15$ frames and we transform 8 filters, a unit at conv5_3 sees a window of $17 \times 15 = 255$ frames.

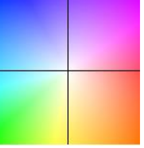
Feichtenhofer, Pinz, Wildes, NIPS'16

Visualizing the learned representation



[Erhan et al. 2009]
[Simonyan et al. 2013]
[Mahendran & Vedaldi 2014]
[Yosinski et al. 2014]

Feichtenhofer, PhD Thesis 2017

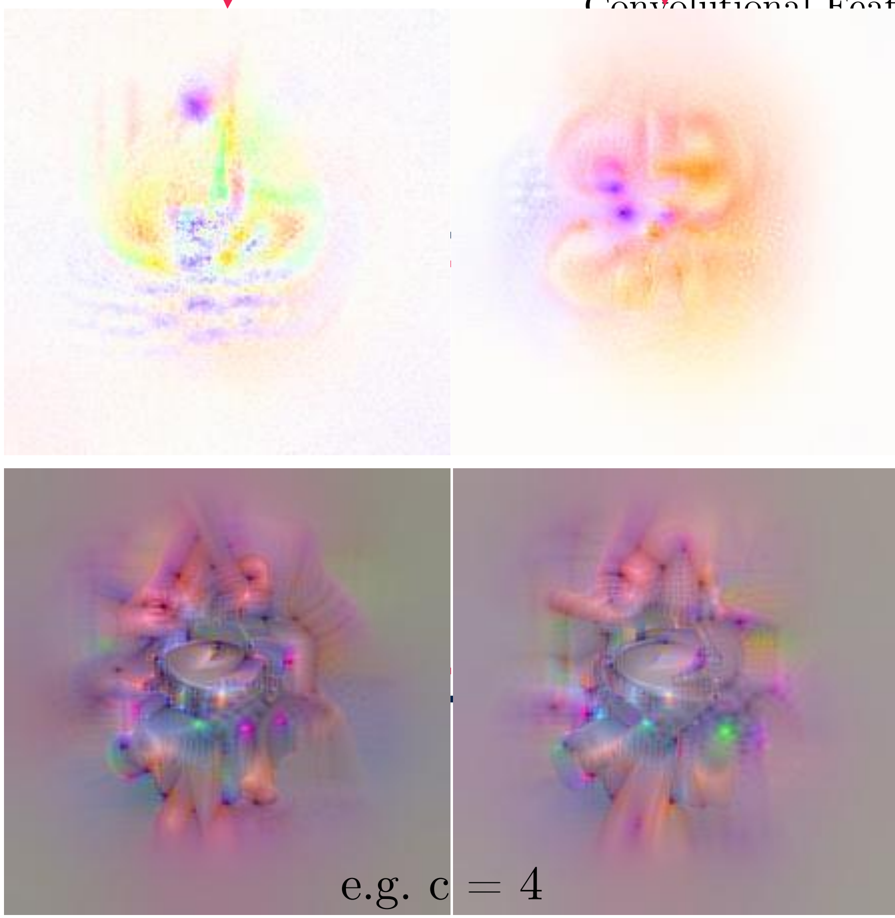


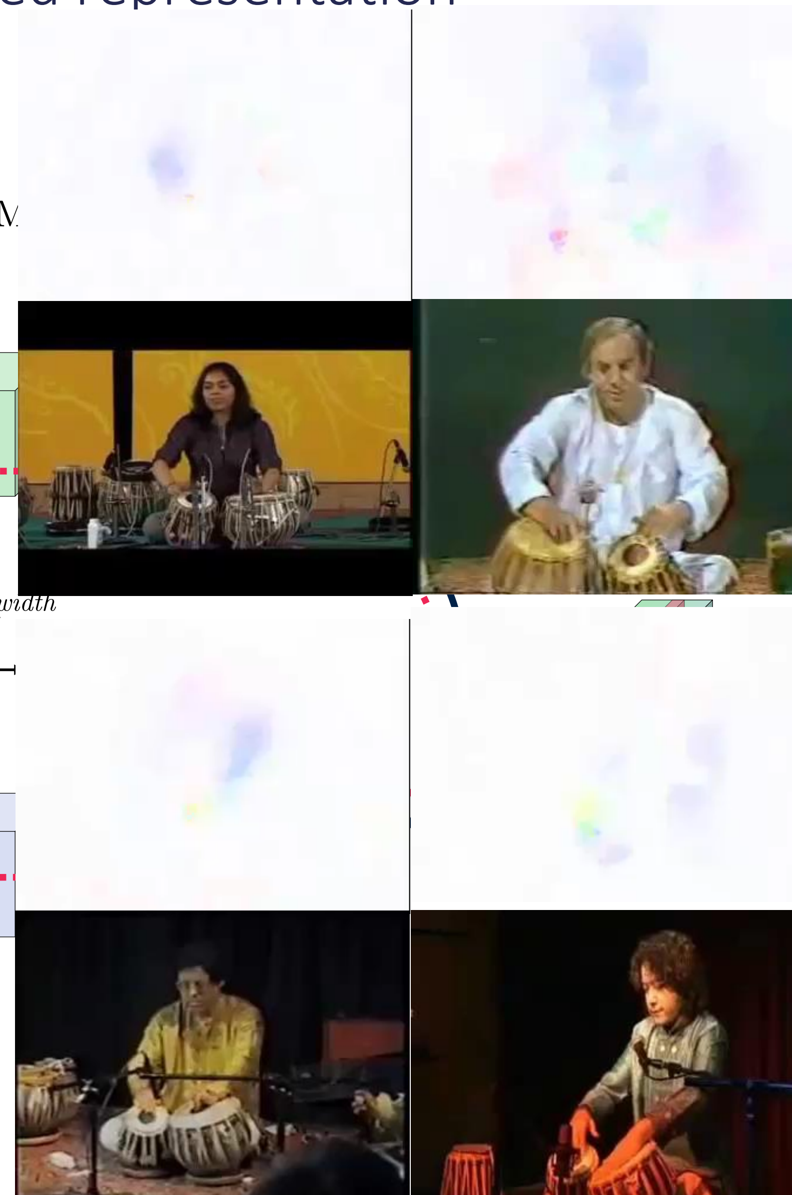
Visualizing the learned representation

Slow motion

Fast motion

Convolutional Feature M





eight

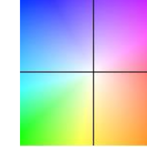
width

Loss

c

2014]

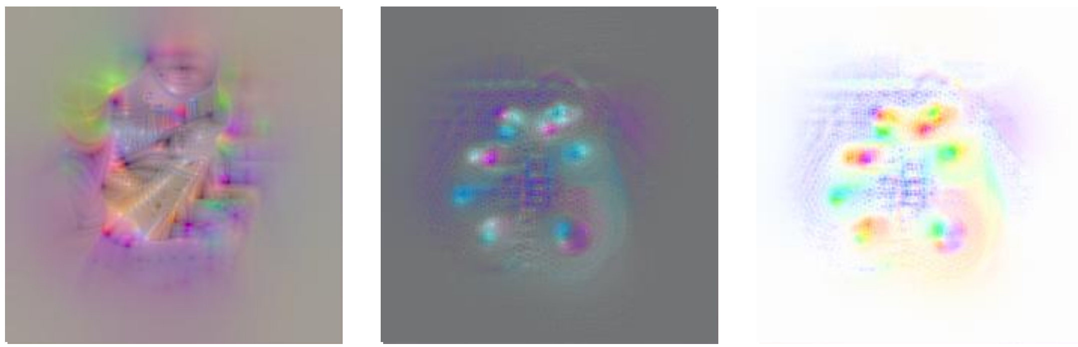
Feichtenhofer, PhD Thesis 2017



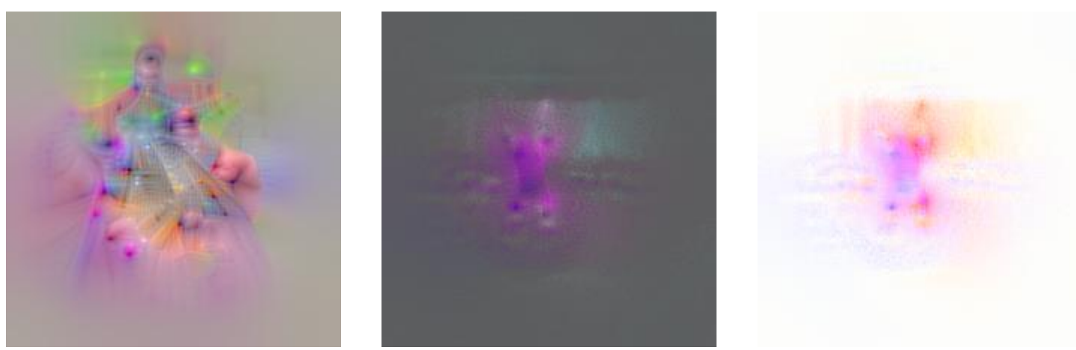
Filter #21 at conv5 fusion – a local Billiard neuron ?

➔ (mostly) linear motion

Slow motion



Fast motion

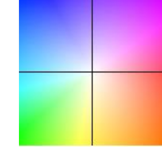


slower

w (direction)

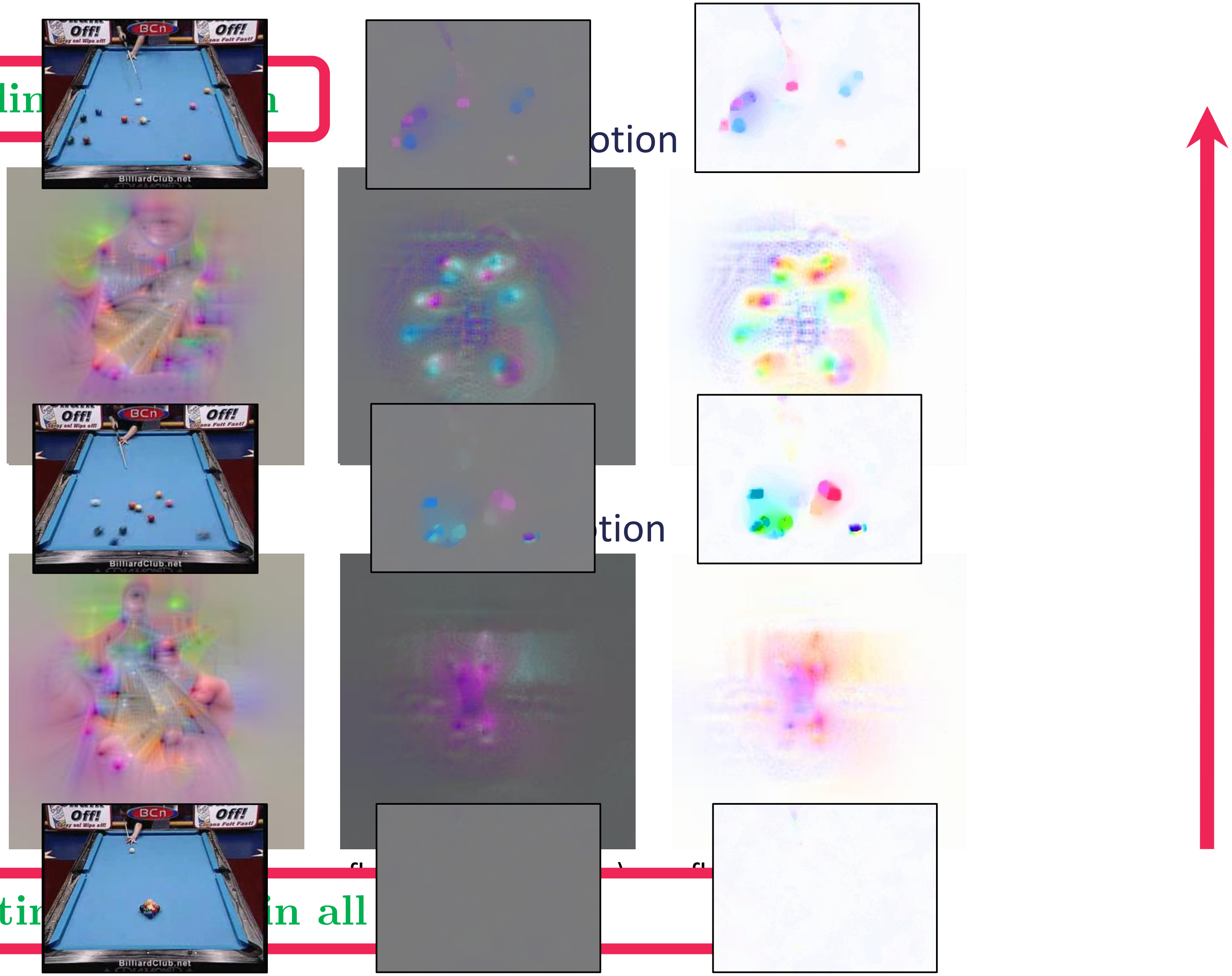
➔ accelerating motion in all directions

Feichtenhofer, PhD Thesis 2017



Filter #21 at conv5 fusion – a local Billiard neuron ?

→(mostly) lin n



→ Test set example snippets

→ accelerati in all

Feichtenhofer, PhD Thesis 2017

Last layer

→ “Billiards”

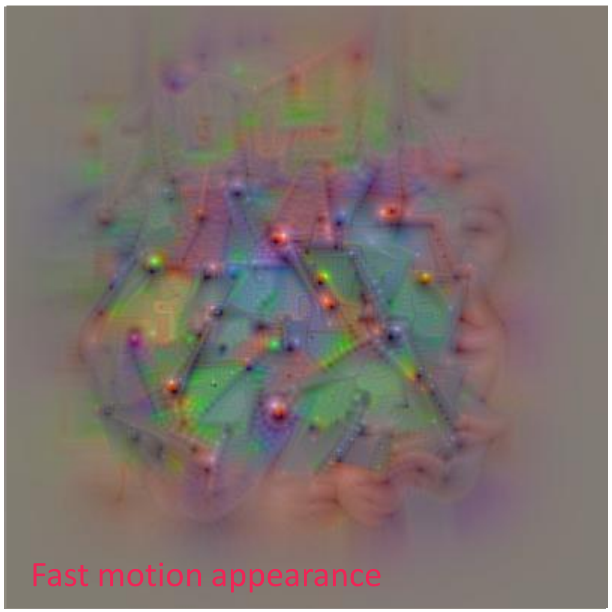
Appearance

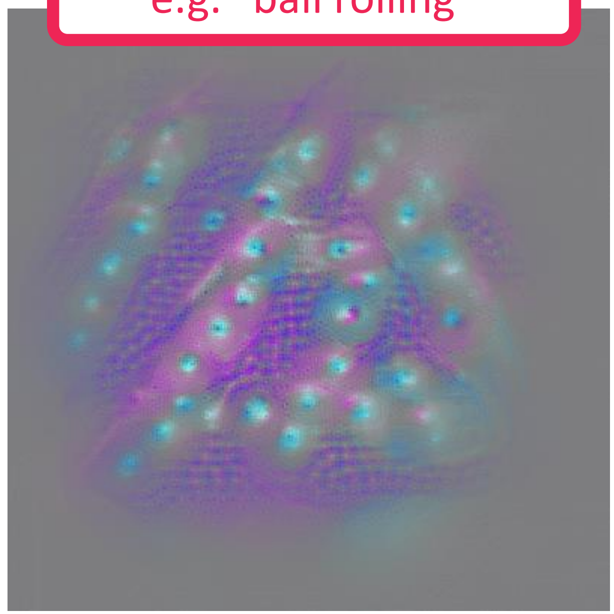
Slow motion

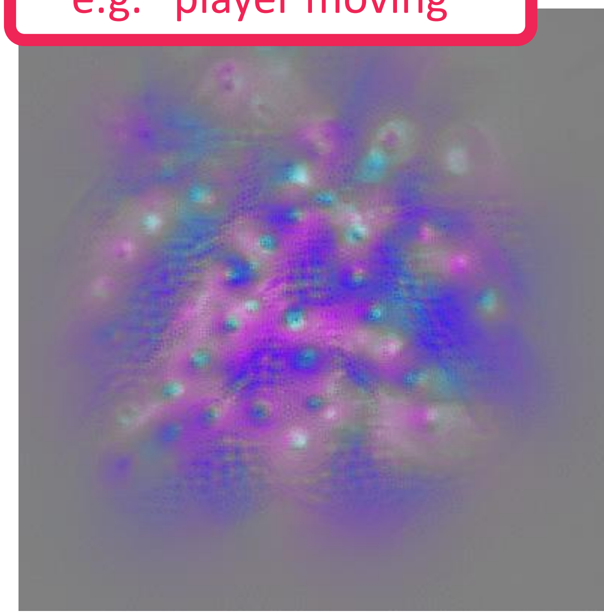
Fast motion

e.g. “ball rolling”

e.g. “player moving”







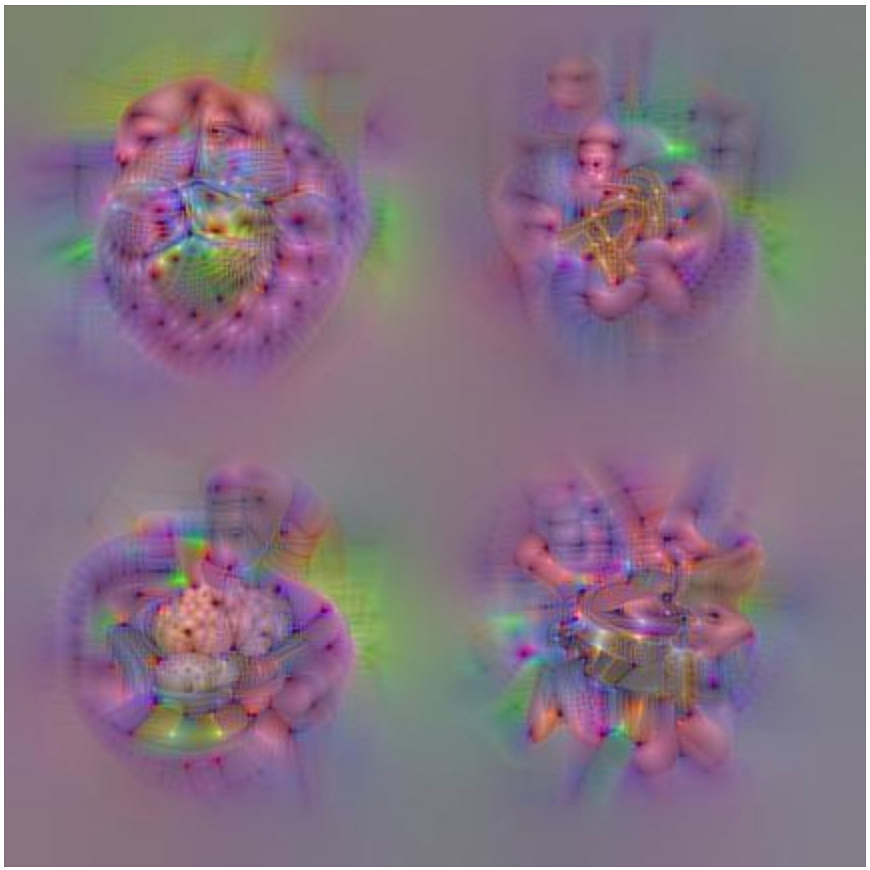
sis 2017

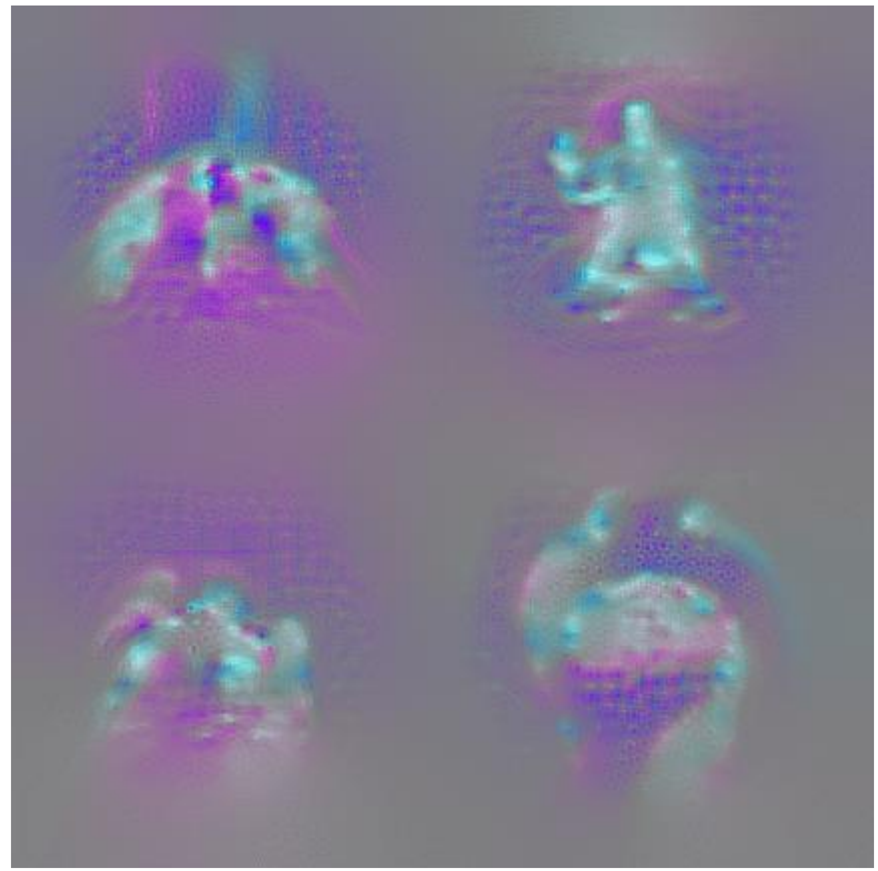
Going through the conv layers of VGG-16
(first four filters of each layer are shown)

Appearance

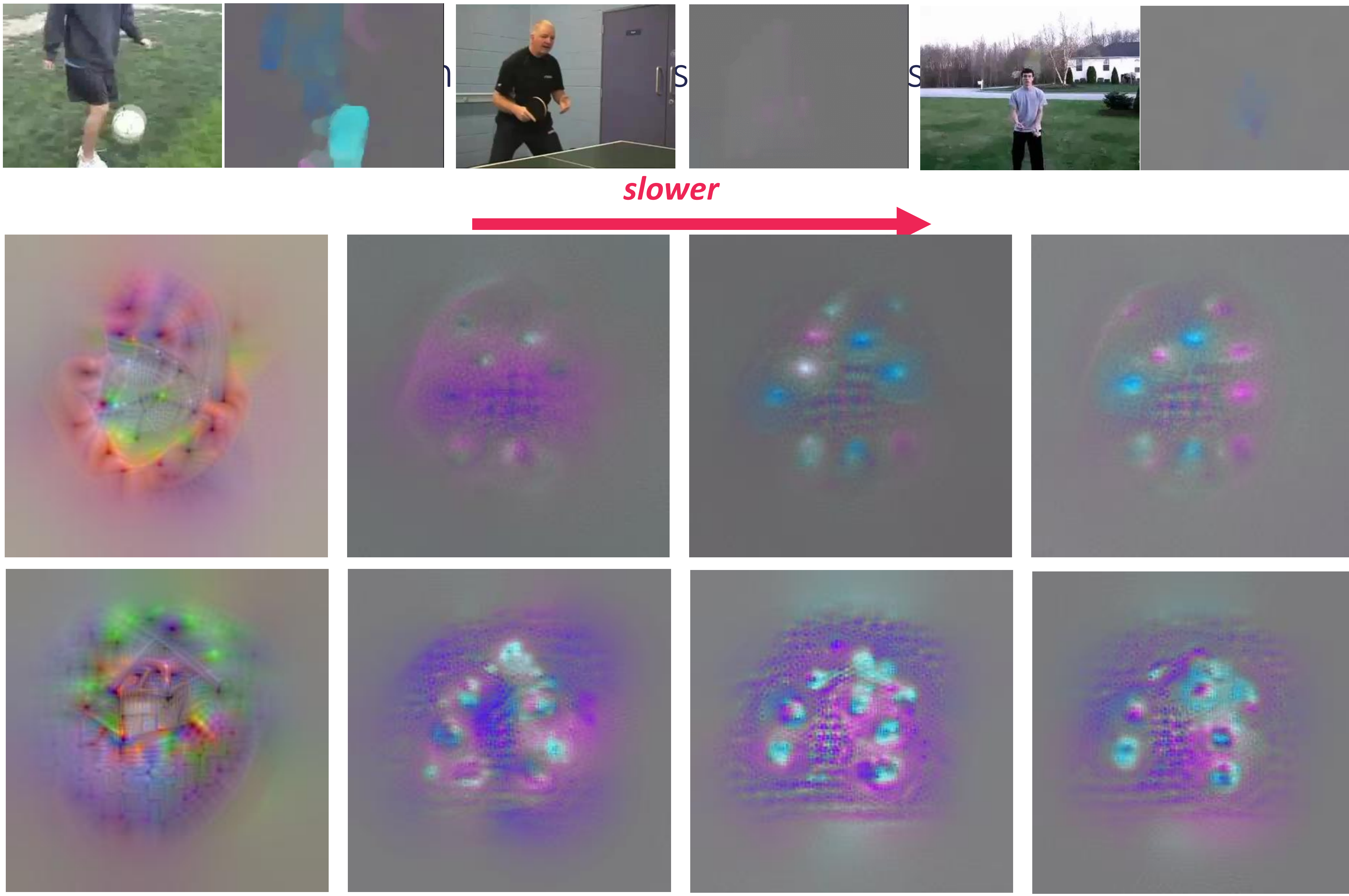
conv5_3 f1-4

Slow motion

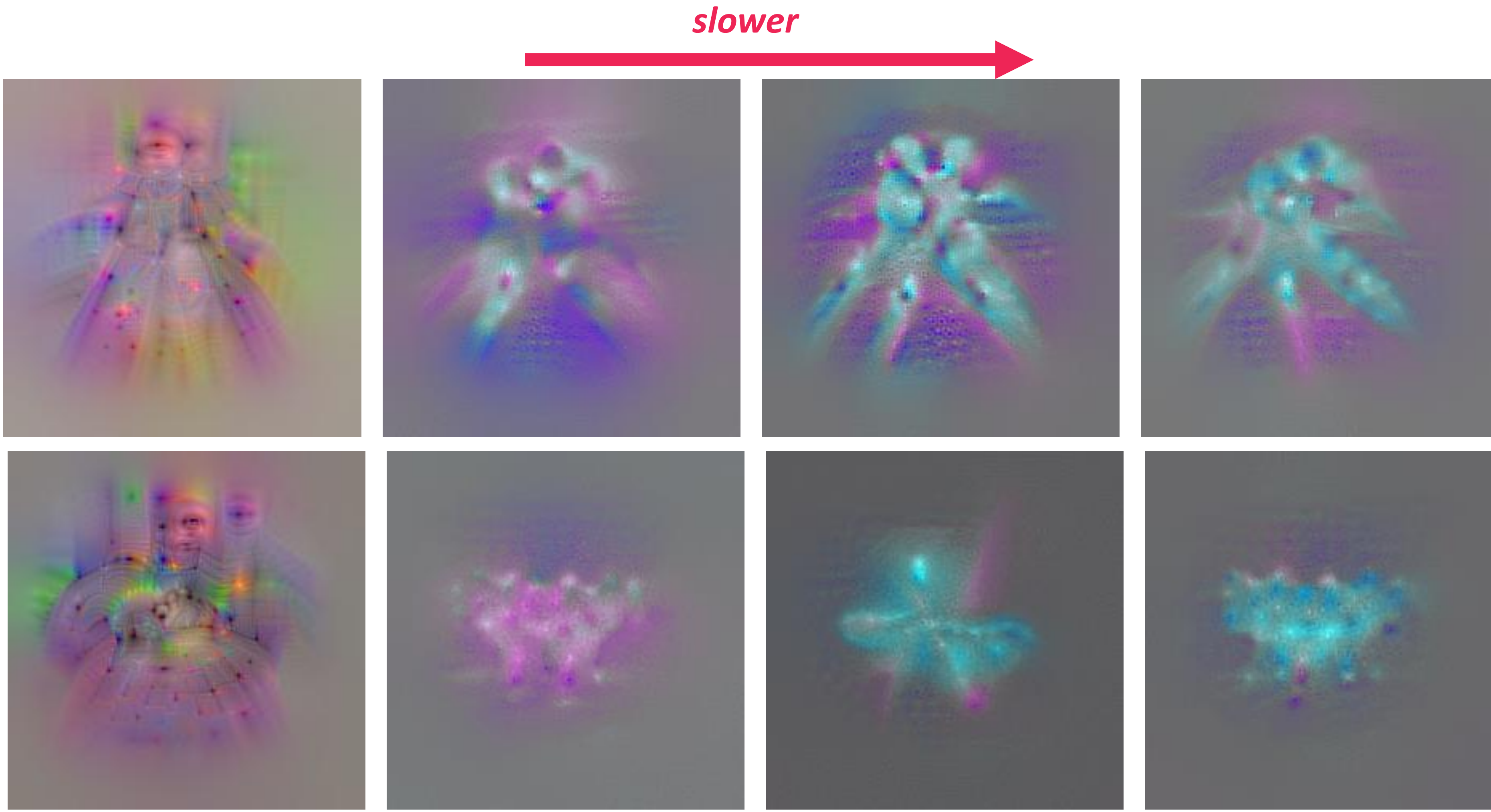




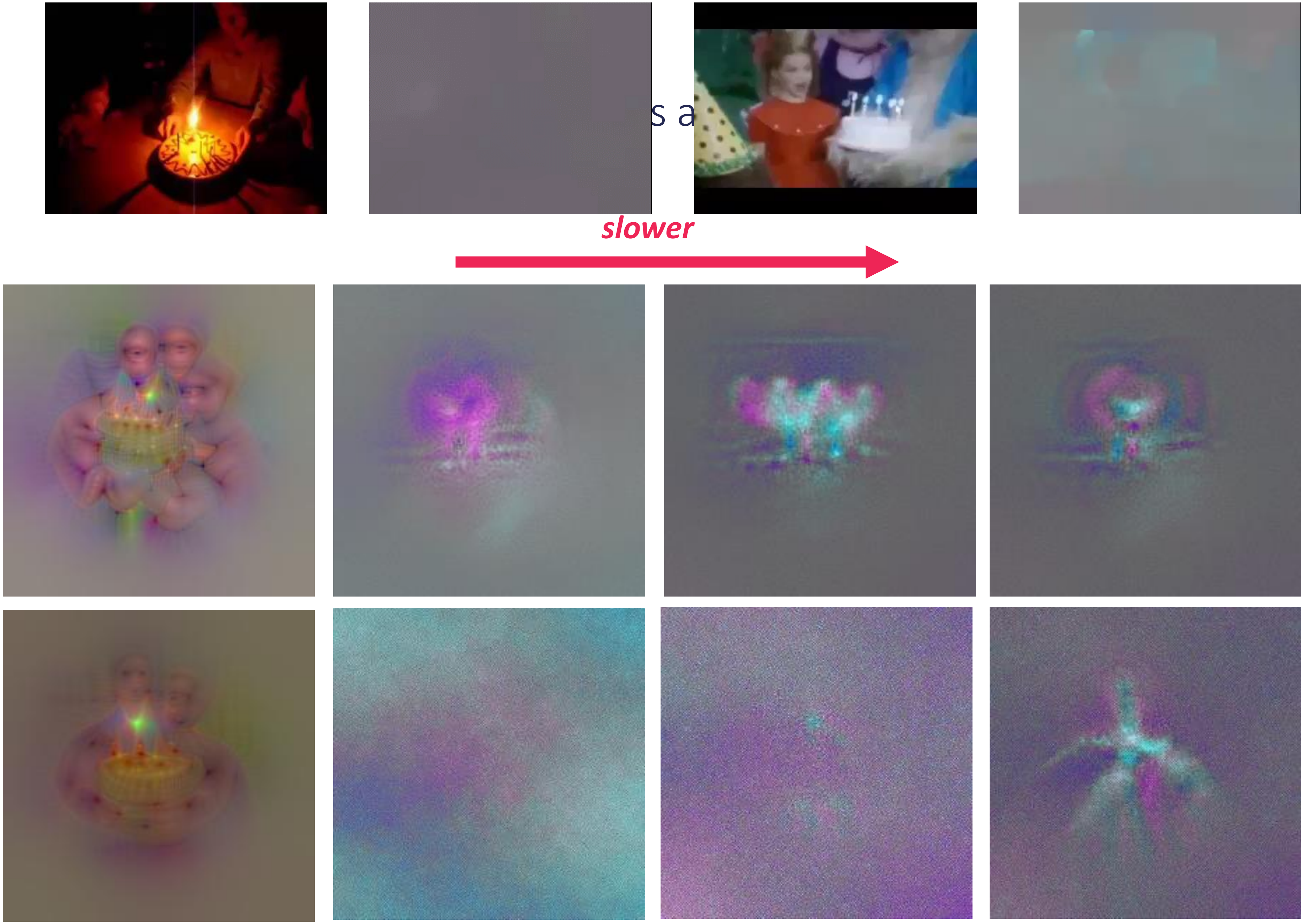
Feichtenhofer, PhD Thesis 2017



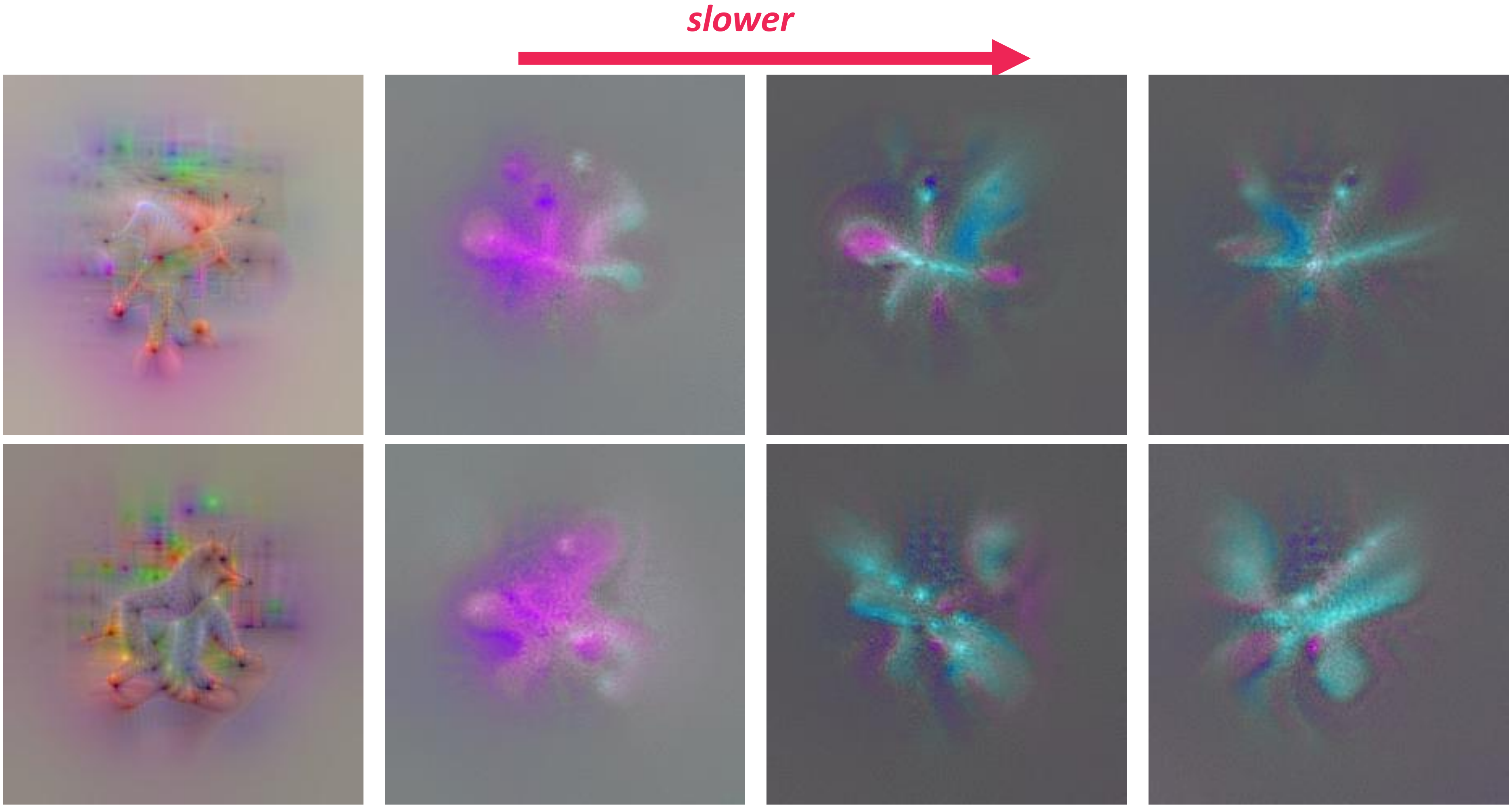
Feichtenhofer, PhD Thesis 2017



Feichtenhofer, PhD Thesis 2017



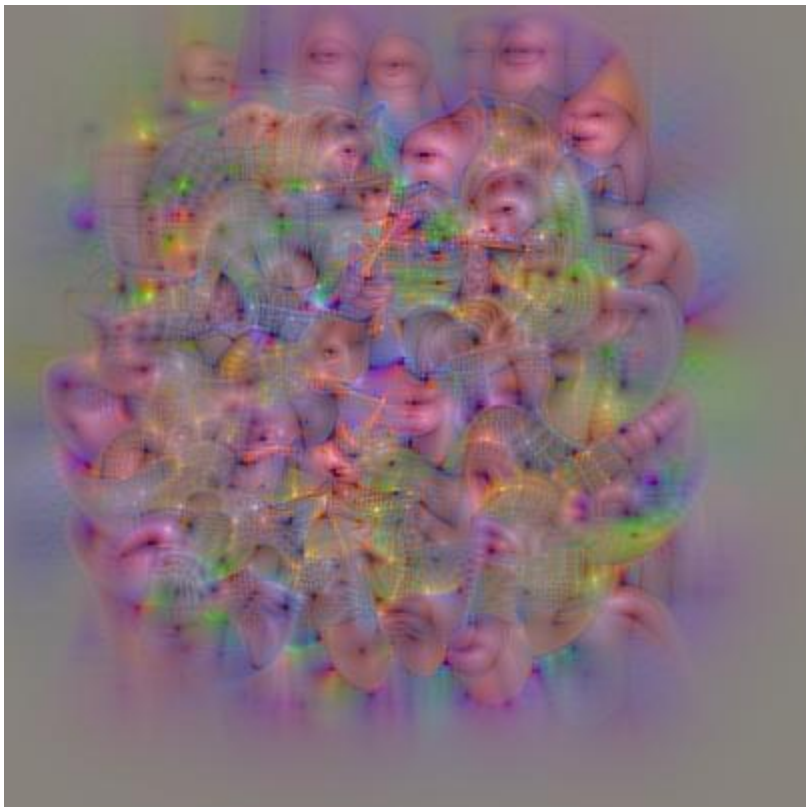
Feichtenhofer, PhD Thesis 2017



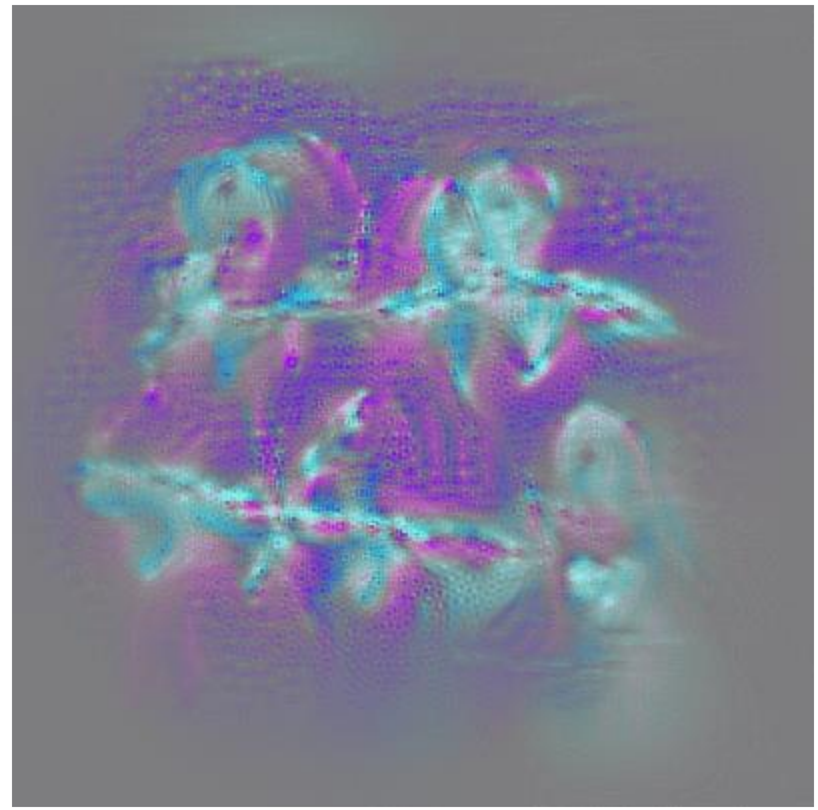
Feichtenhofer, PhD Thesis 2017

FC 6 (4096 features; RF 404x404)

Appearance



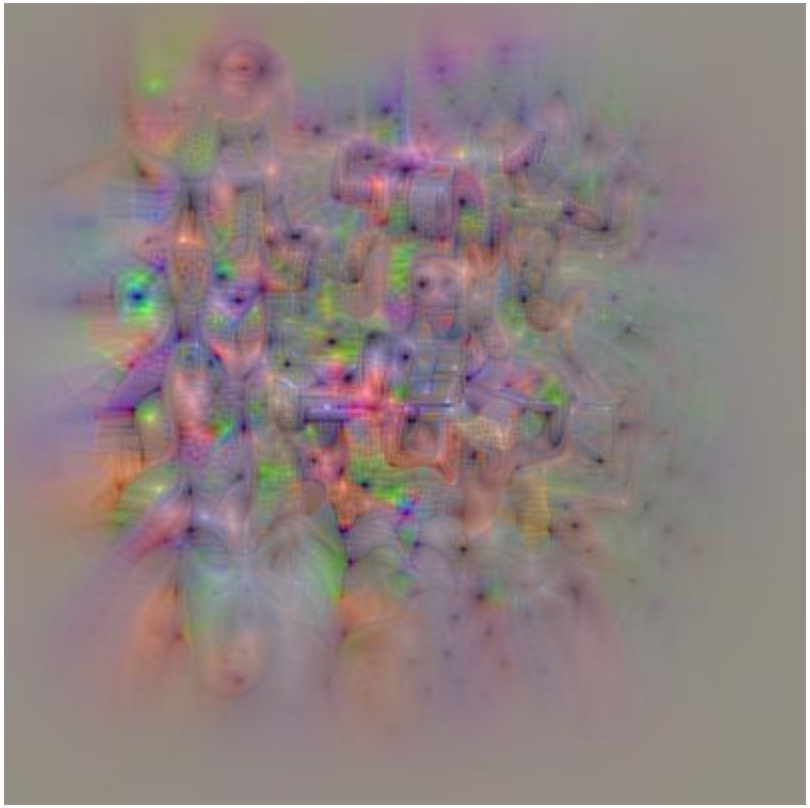
Slow motion



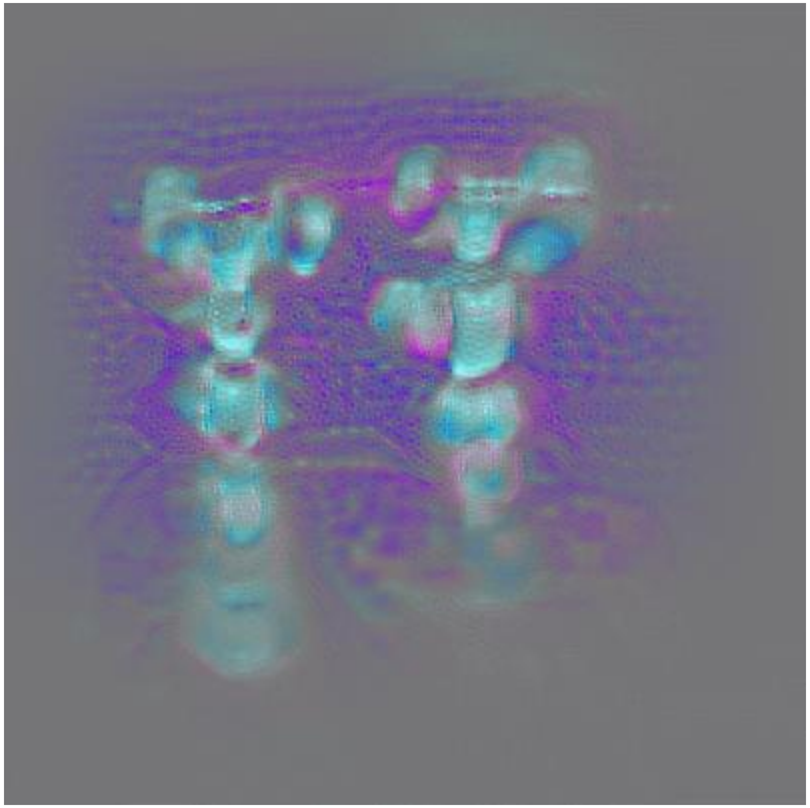
Feichtenhofer, PhD Thesis 2017

FC 7 (4096 features; RF 404x404)

Appearance



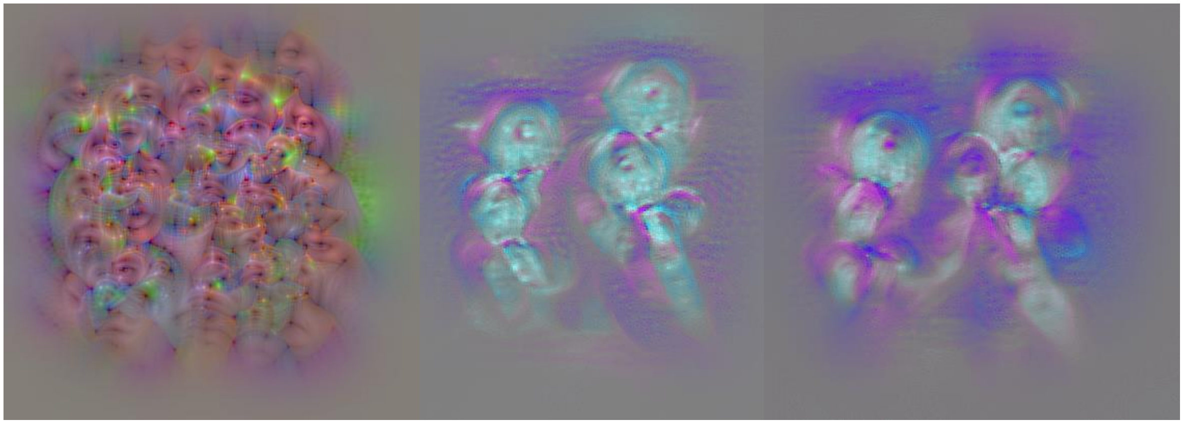
Slow motion



Feichtenhofer, PhD Thesis 2017

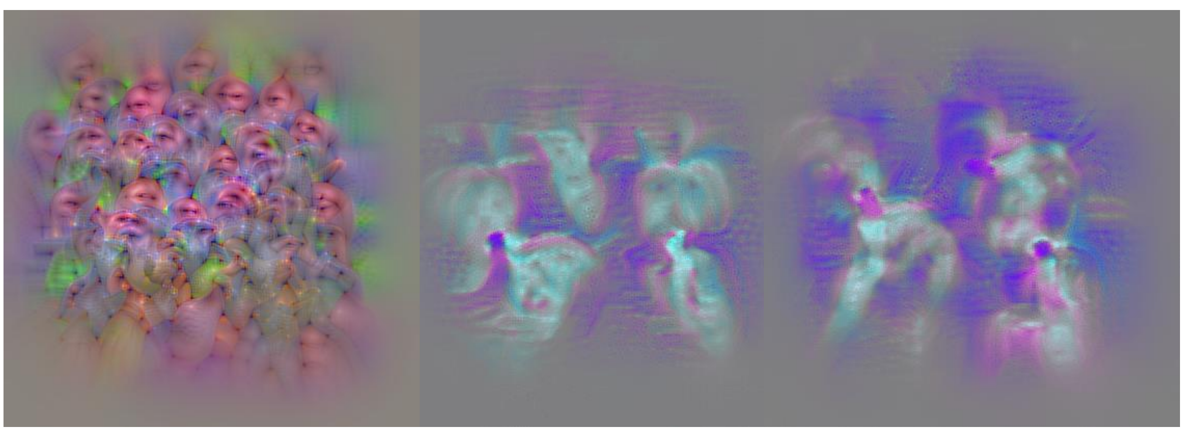
Explaining failure cases:

Appearance | Slow motion | Fast motion



BrushingTeeth 52% accuracy

(33% confused with ShavingBeard)

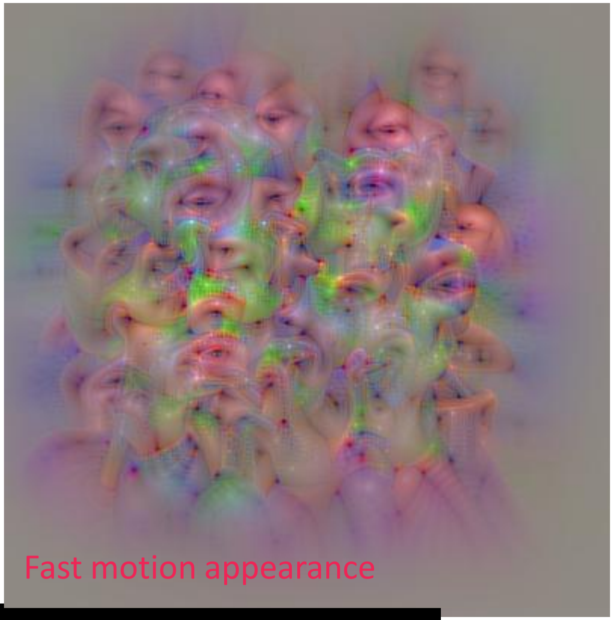


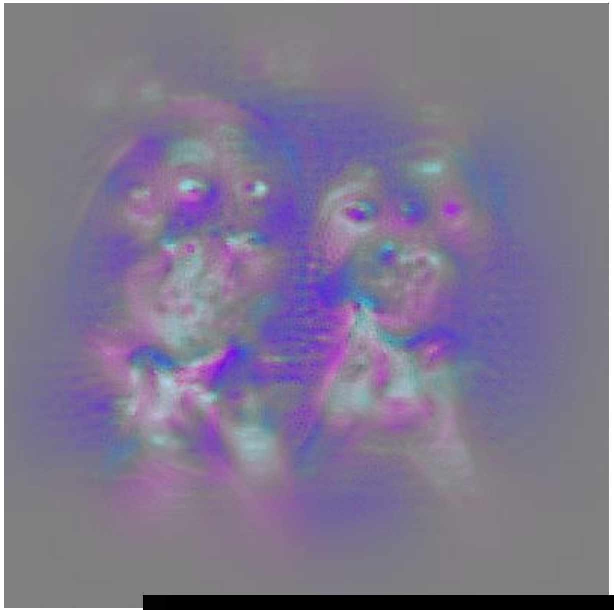
Feichtenhofer, PhD Thesis 2017

Revealing idiosyncracies in data

➔ "ApplyLipstick"

Appearance | Slow motion | Fast motion





sis 2017

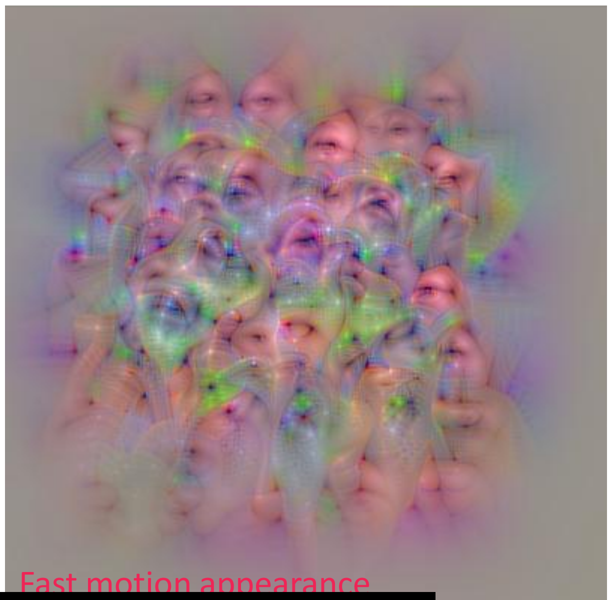
Revealing idiosyncracies in data

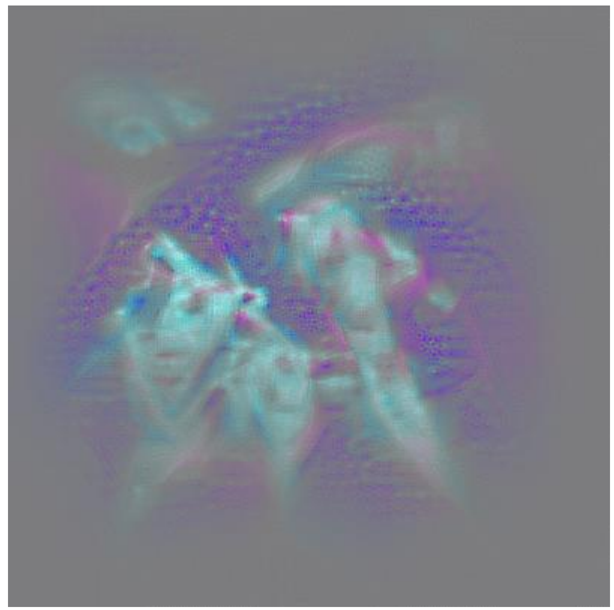
➔ "ApplyEyeMakeup"

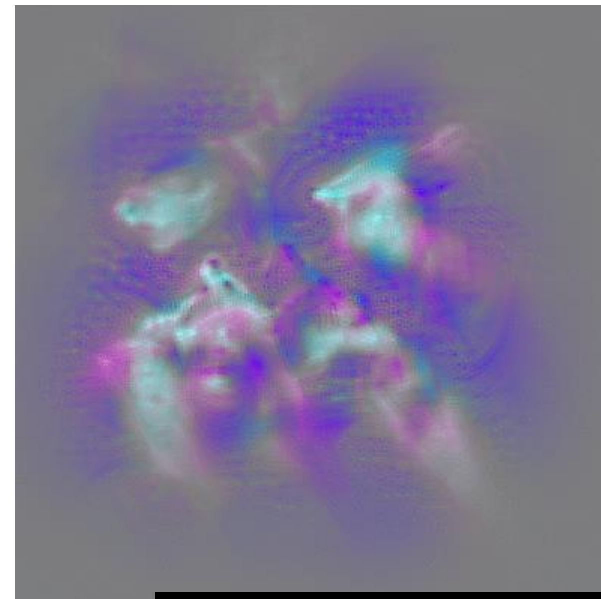
Appearance

Slow motion

Fast motion





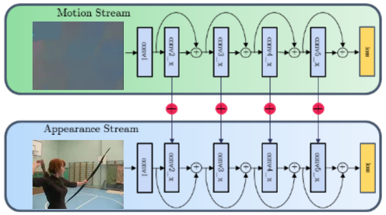


sis 2017

Summary of our insights

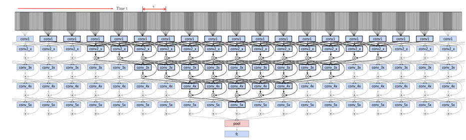
We study ways of connecting appearance and motion ConvNets

spatially



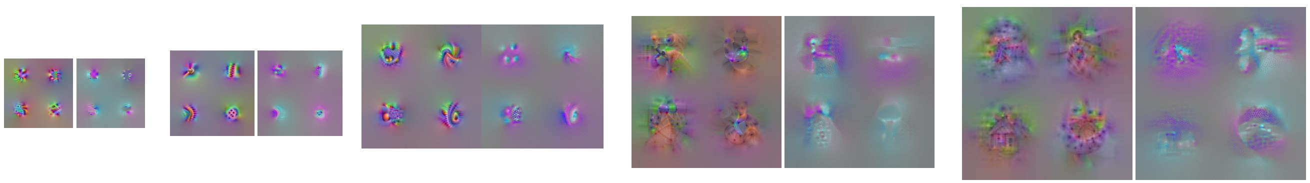
and

temporally

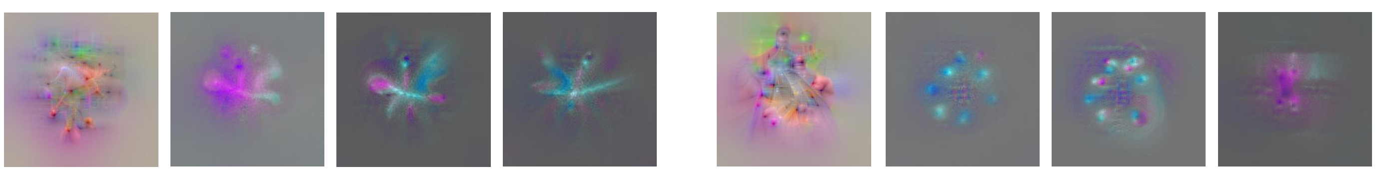


By visualizing the learned representation we find that:

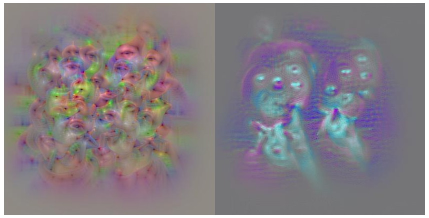
- Early layers show similar spatial structures for appearance and flow



- Higher layer conv-fusion-filters are broadly tuned to multiple speeds and can be specific but also generic across classes



- Class visualizations aid in analyzing system strengths and weaknesses



vs.