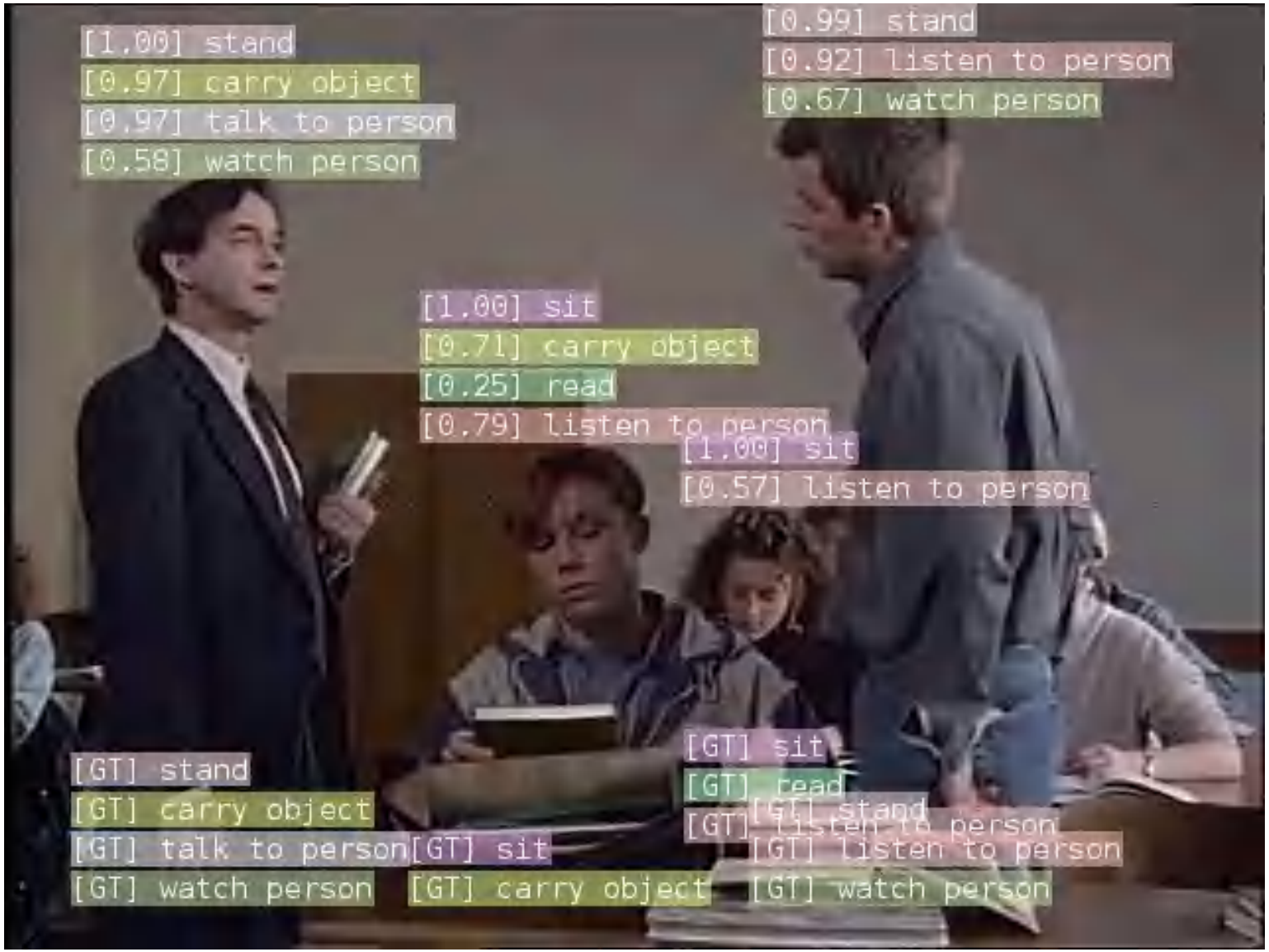
Video Action Classification and Detection Architectures

CVPR 2019 Tutorial

Christoph Feichtenhofer

Facebook AI Research (FAIR)

Task: Human action classification & detection



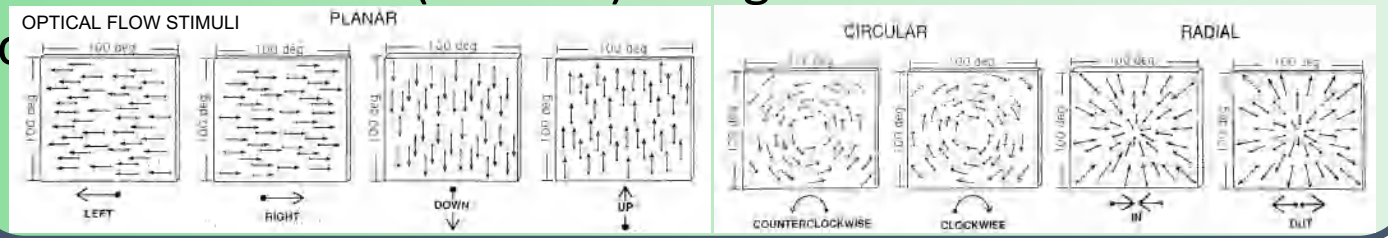
C. Feichtenhofer, H. Fan, J. Malik, K. He SlowFast Networks for Video Recognition. Tech. Report, arXiv, 2018

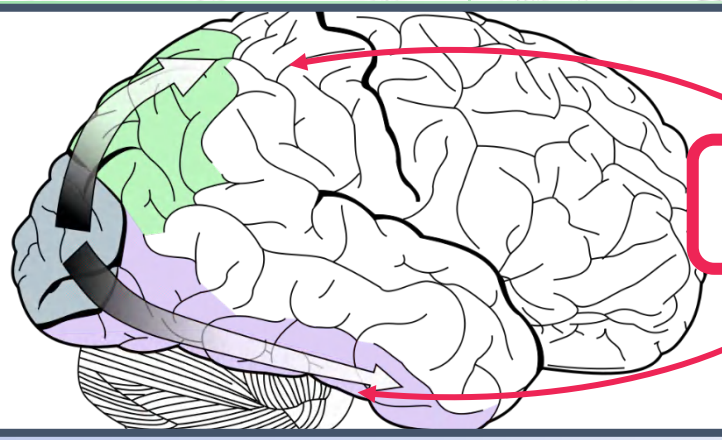
Johansson: Perception of Biological Motion

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’) recognizes motion and locates





➔ “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

Background: Two-Stream Convolutional Networks

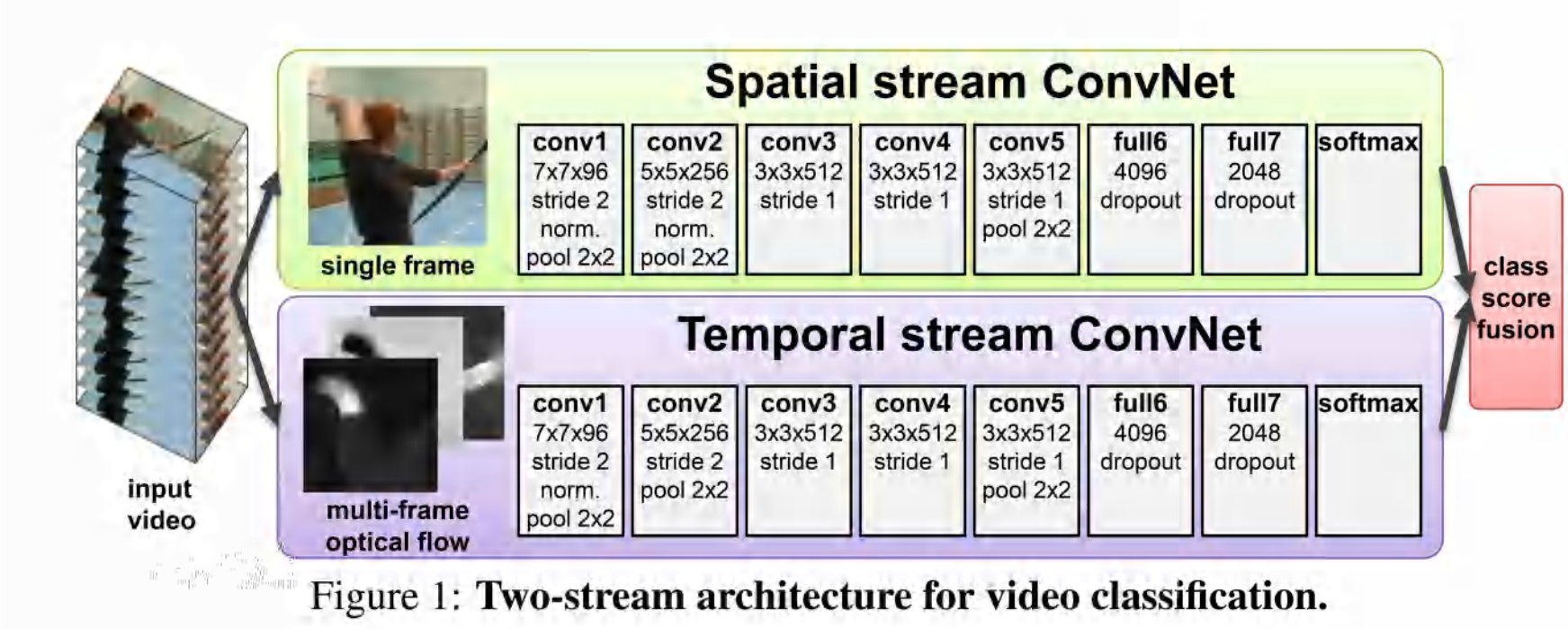


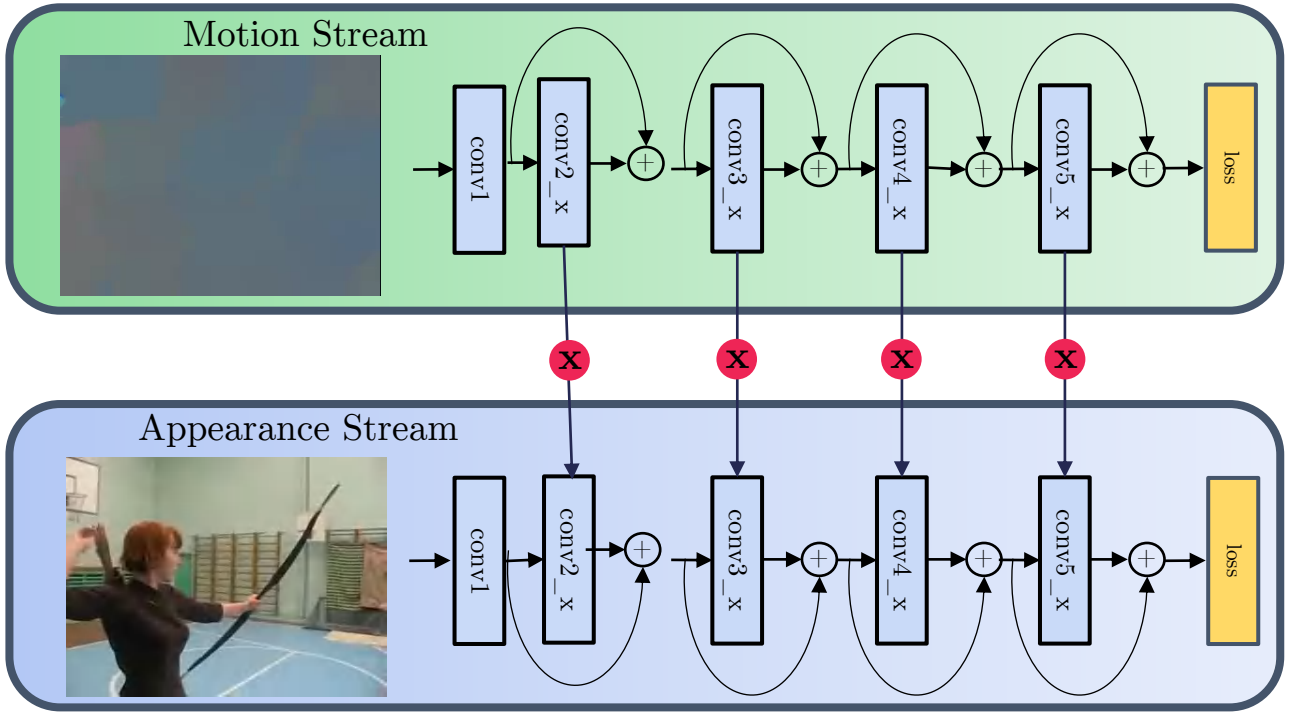
Figure 1: **Two-stream architecture for video classification.**

Individual processing of spatial and temporal information

- Using a separate **2D (x,y)** ConvNet recognition stream for each
- Late fusion via softmax score averaging

K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014

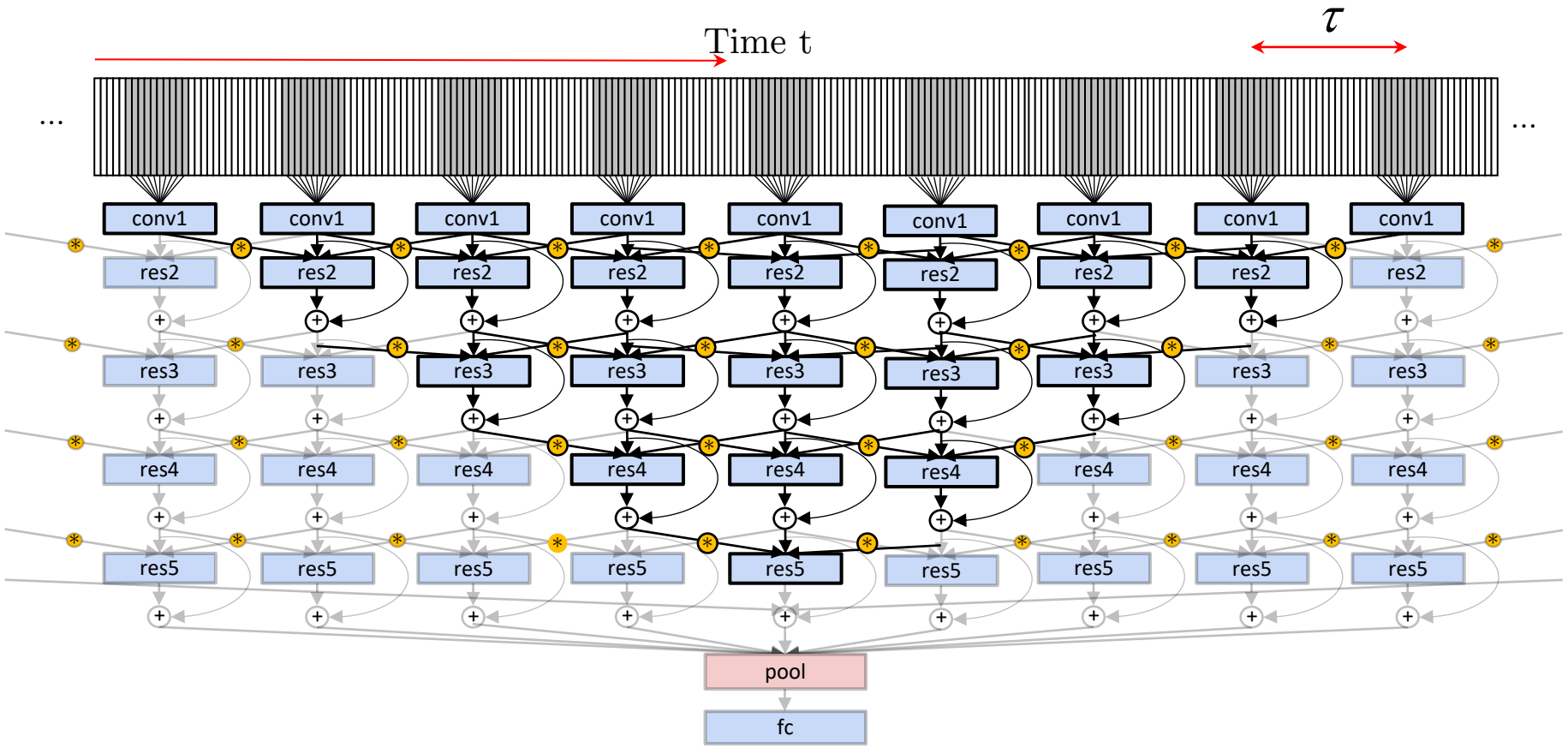
Background: Two-Stream Network Fusion and 2D -> 3D Transformation/Inflation



- ST-ResNet allows the hierarchical learning of spacetime features by connecting the appearance and motion channels of a two-stream architecture.

C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In Proc. CVPR, 2016
C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In NIPS, 2016.
C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal multiplier networks for video action recognition. In CVPR, 2017.

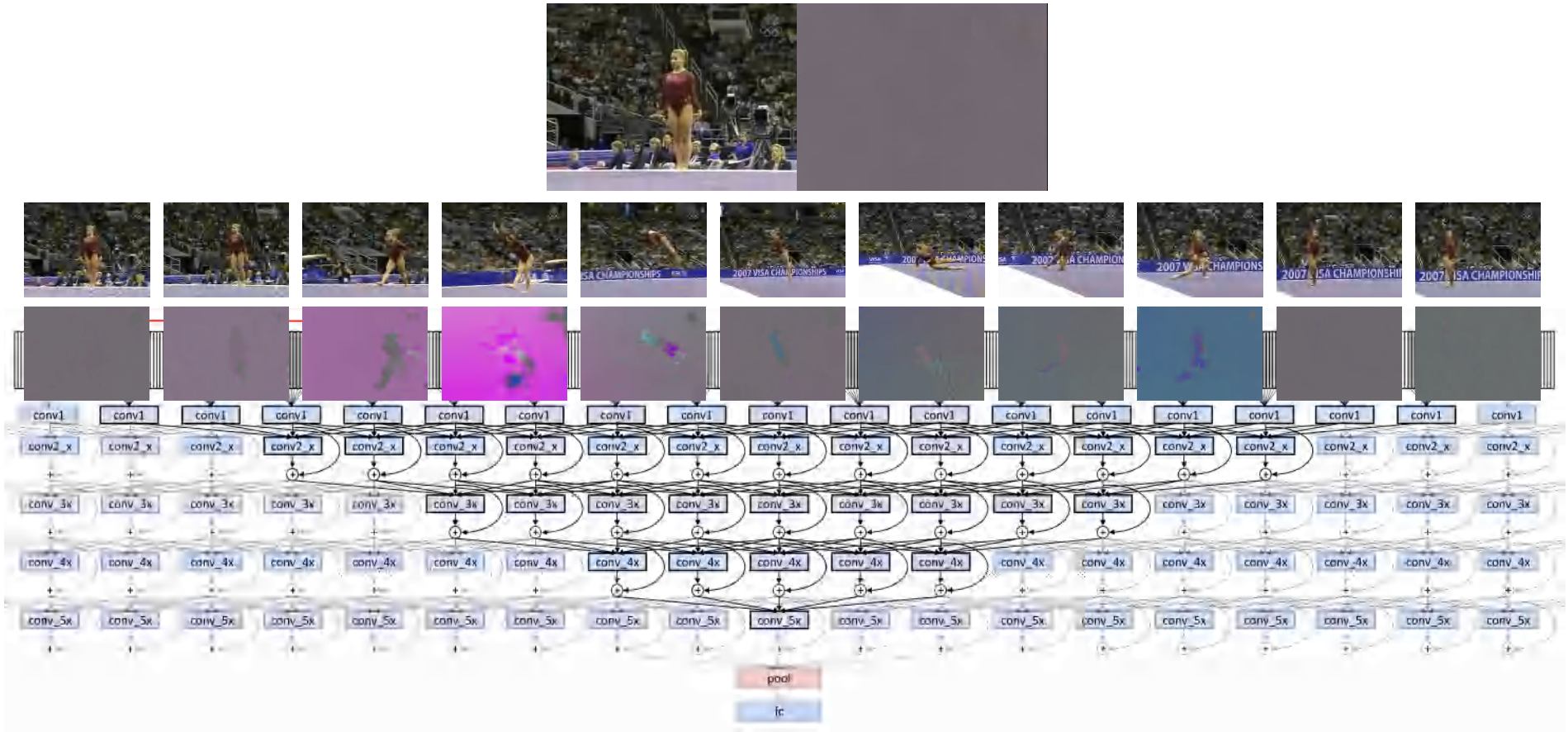
Background: Transforming spatial networks into temporal ones by Inflation



- Inflation allows to transform spatial filters to spatiotemporal ones (3D or 2D spatial +1D temporal)

C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In NIPS, 2016.
J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In Proc. CVPR, 2017.
D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In CVPR, 2018.

Background: Transforming 2D networks into 3D by Inflation

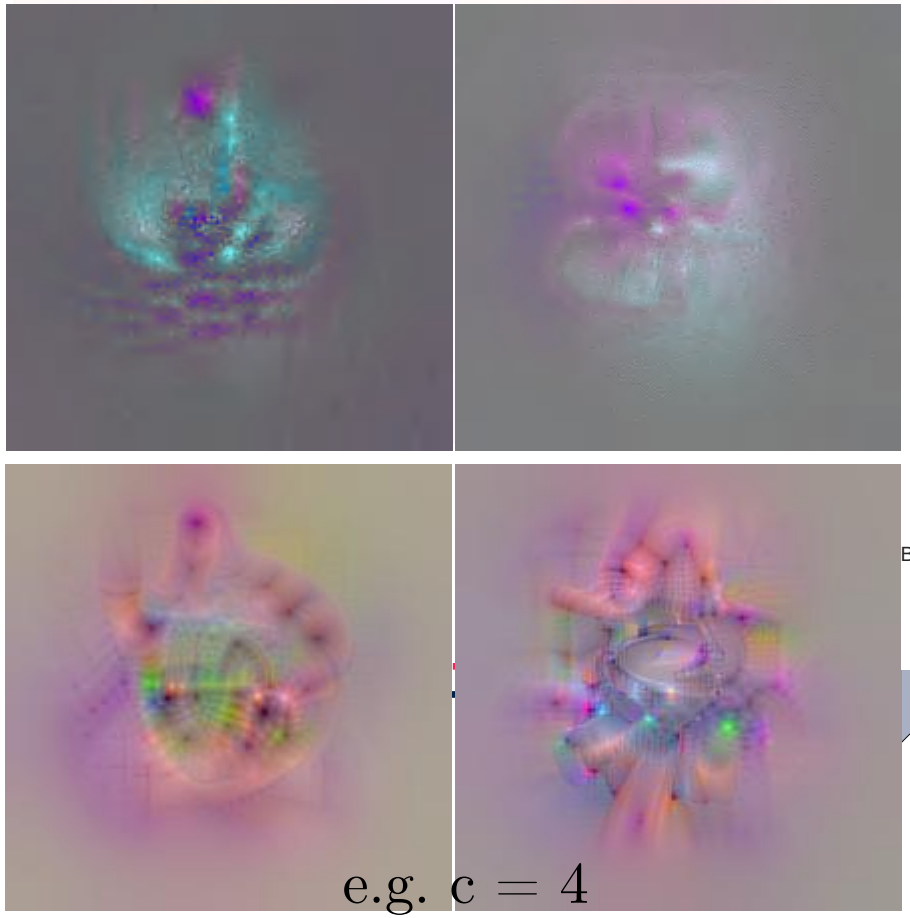


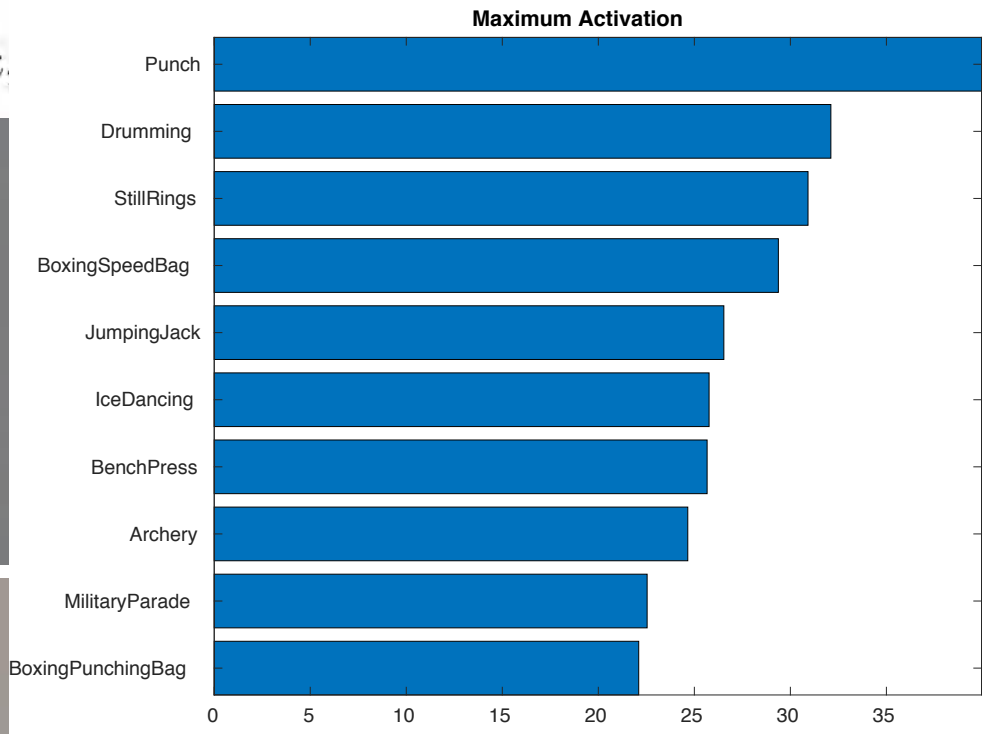
C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In NIPS, 2016.
J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In Proc. CVPR, 2017.

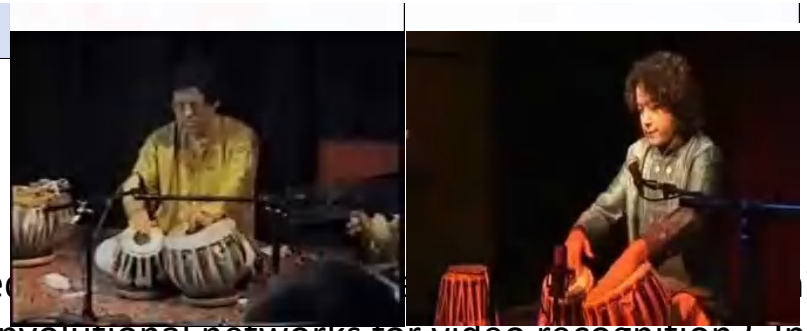
Visualizing the learned representation

Slow motion
(high temporal reg.)

Fast motion
(low temporal reg.)







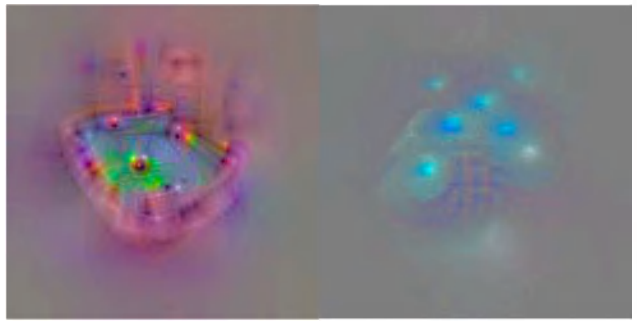
C. Feichtenhofer, A. Pinz, and R. Wildes, A. Zisserman. What have we learne ition?. In CVPR, 2018.
C. Feichtenhofer, A. Pinz, and R. Wildes, A. Zisserman. Deep insights into convolutional networks for video recognition?. In IJCV, 2019.

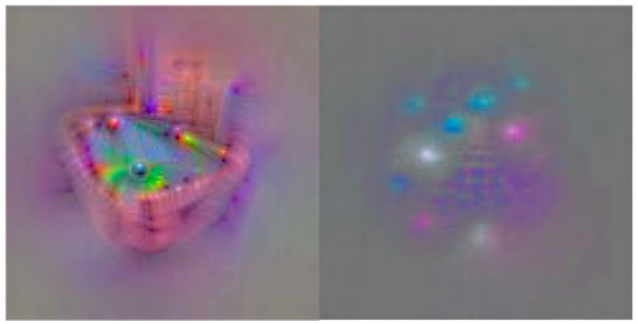
Filter #251 at conv5 fusion – the strongest local Billiards unit

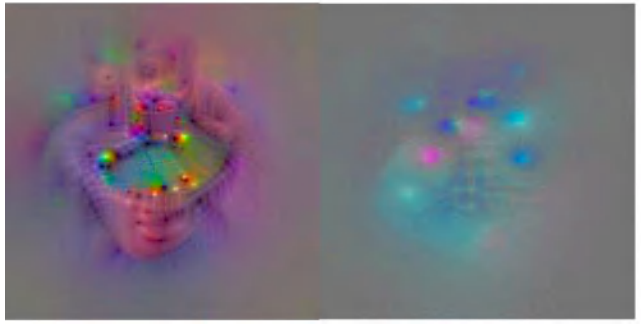
slow

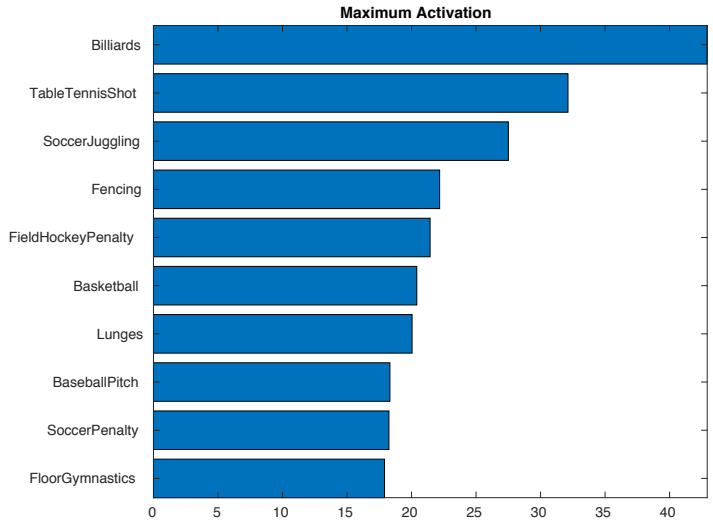
medium

fast



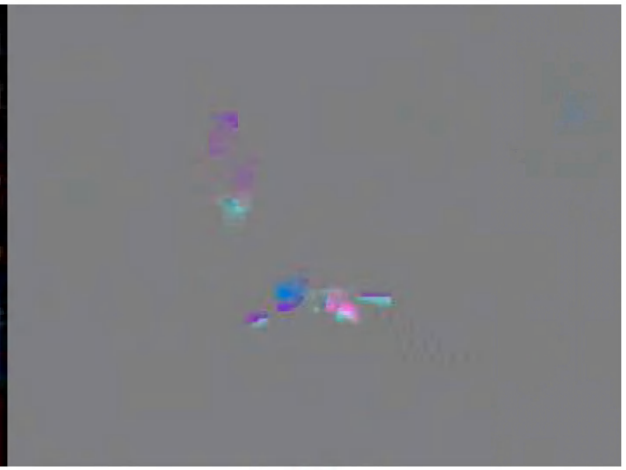






(c) test set activity

(d)



(e)

Last layer

➔ “Billiards”

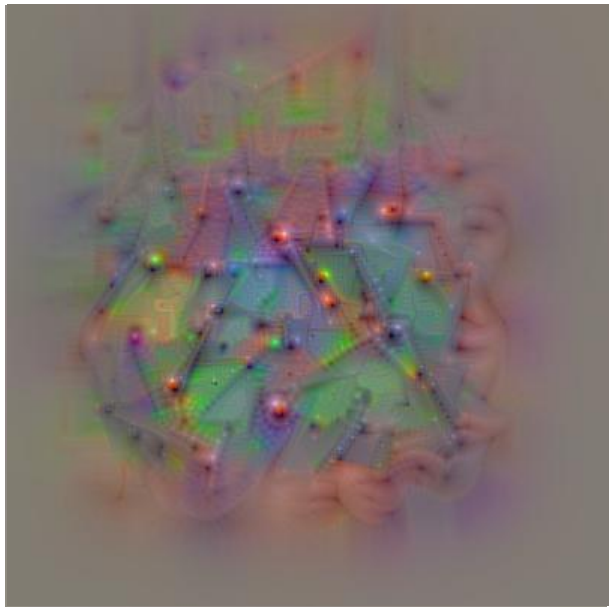
Appearance

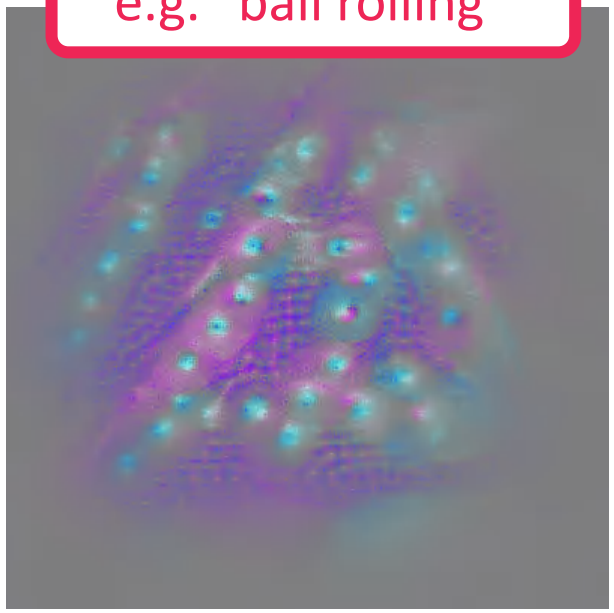
Slow motion

e.g. “ball rolling”

Fast motion

e.g. “player moving”





Last layer

→ “CleanAndJerk”

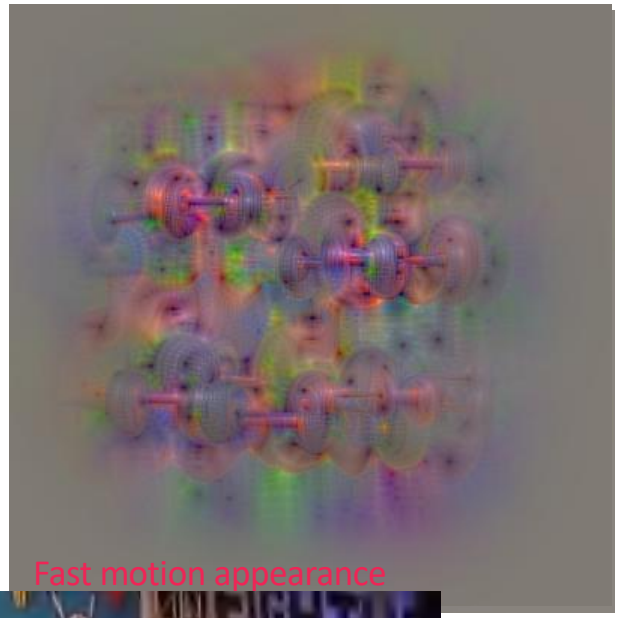
Appearance

Slow motion

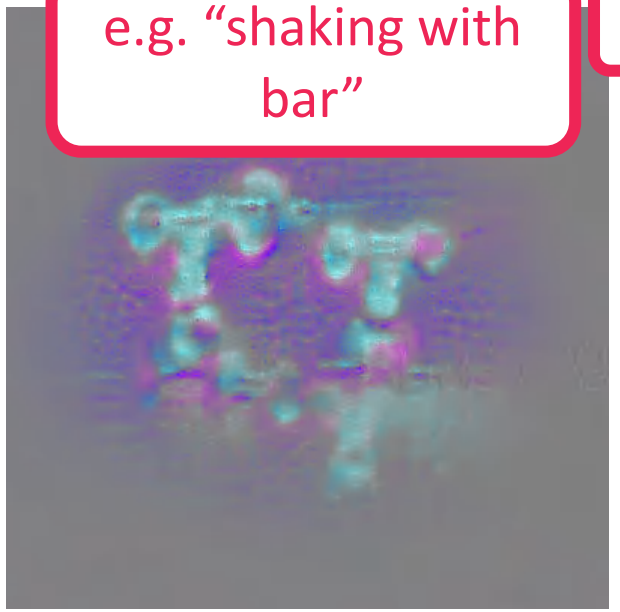
Fast motion

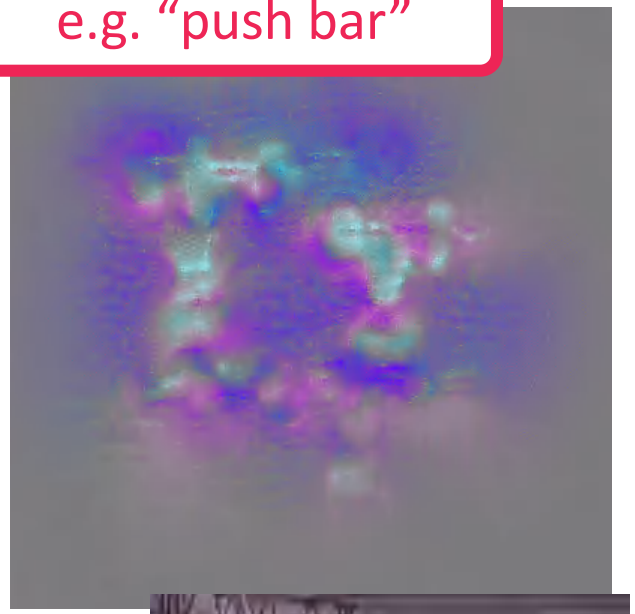
e.g. “shaking with bar”

e.g. “push bar”

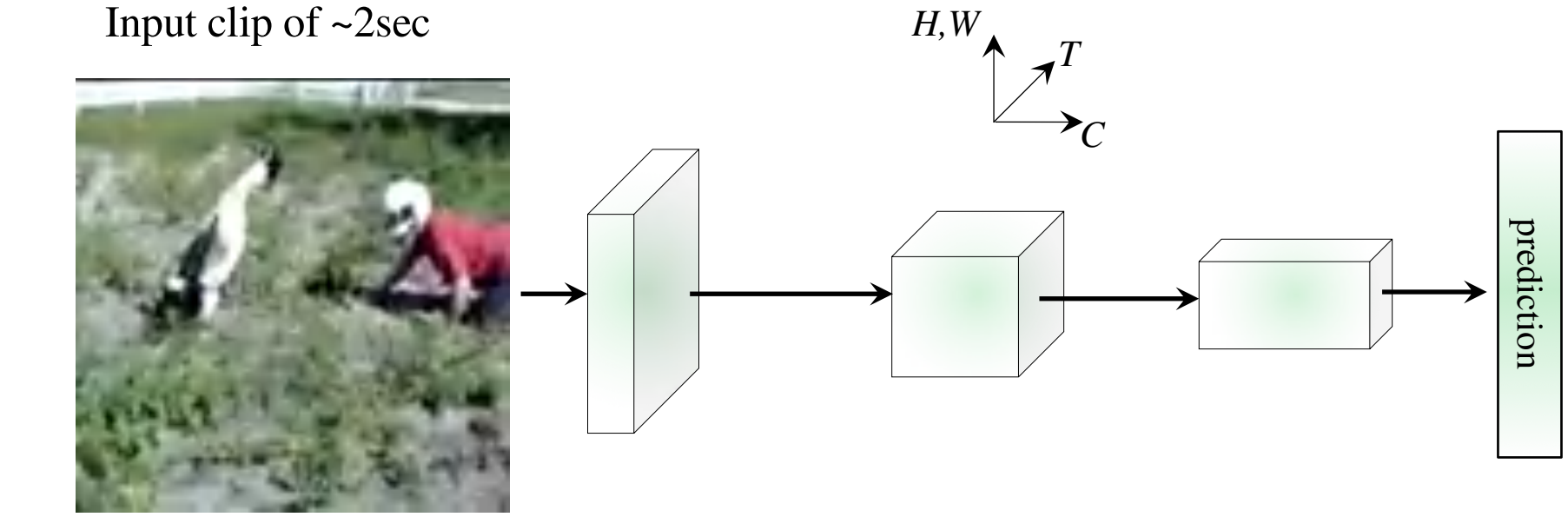


Fast motion appearance



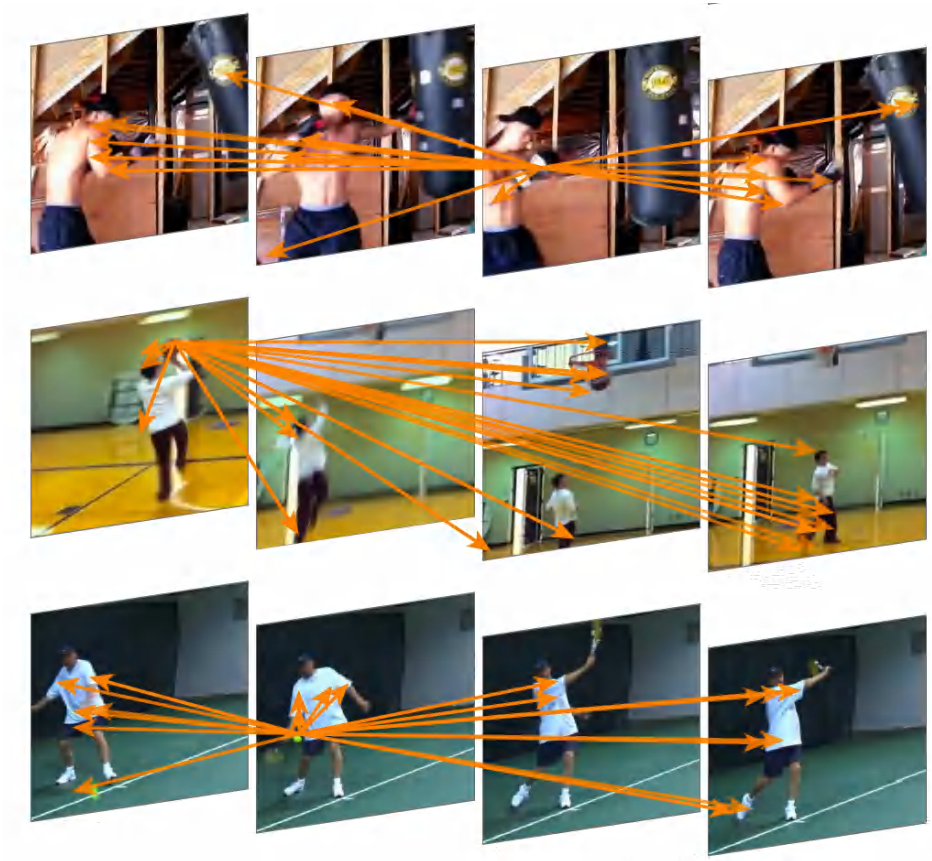


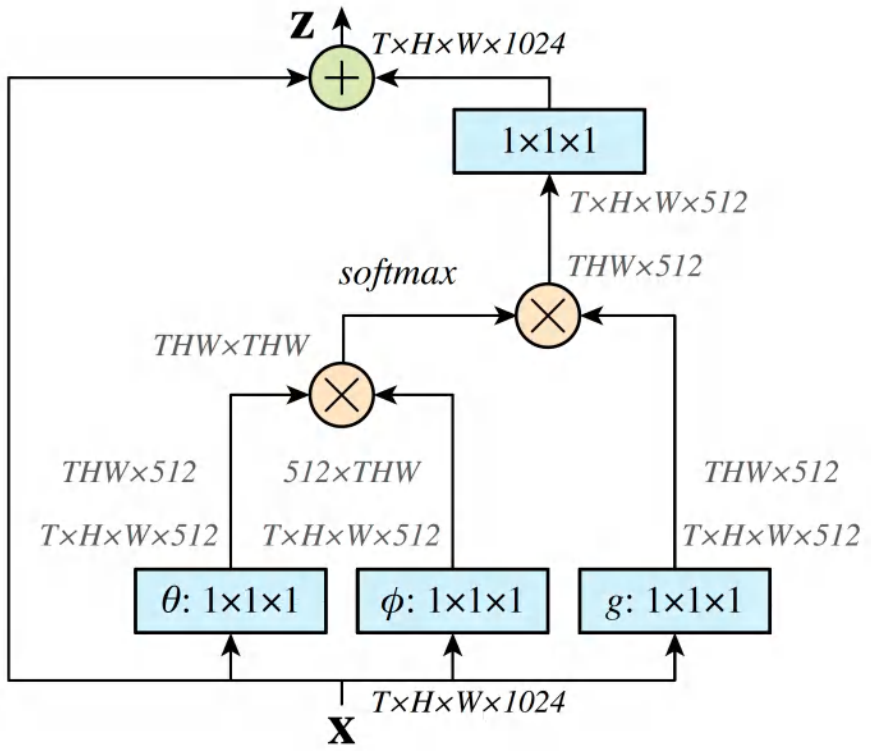
Background: 3D Convolutional Networks



G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In Proc. ECCV, 2010.
D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In Proc. ICCV, 2015.
J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In Proc. CVPR, 2017.

Background: Non-Local Convolutional Network Blocks

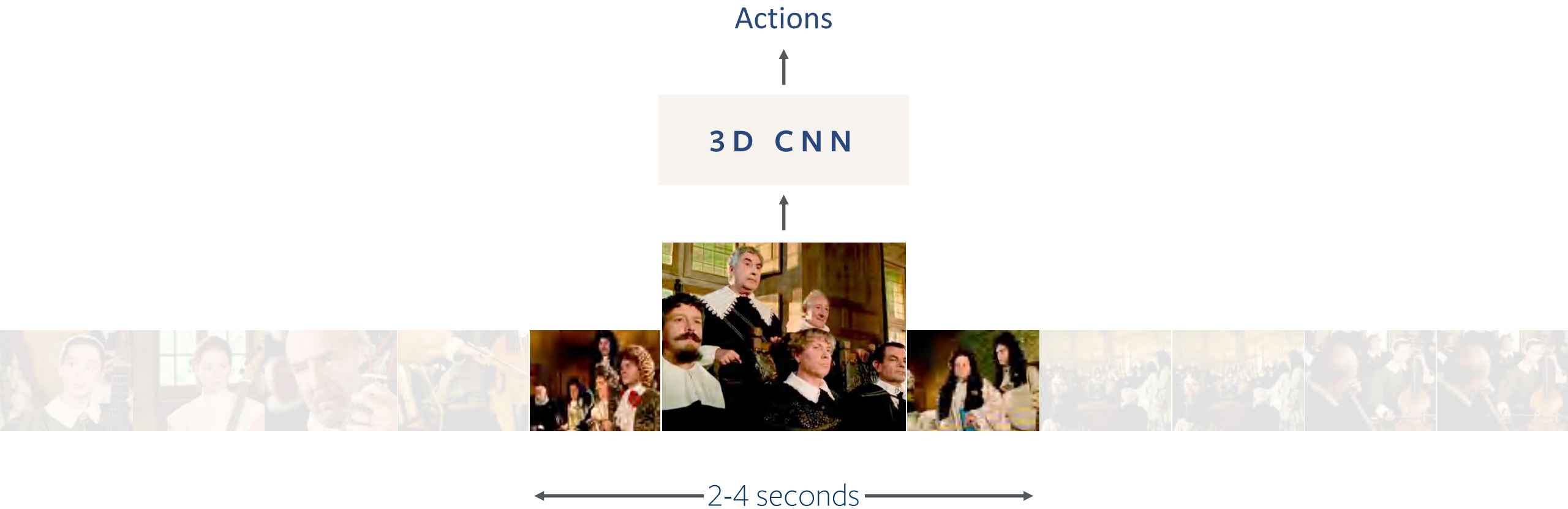




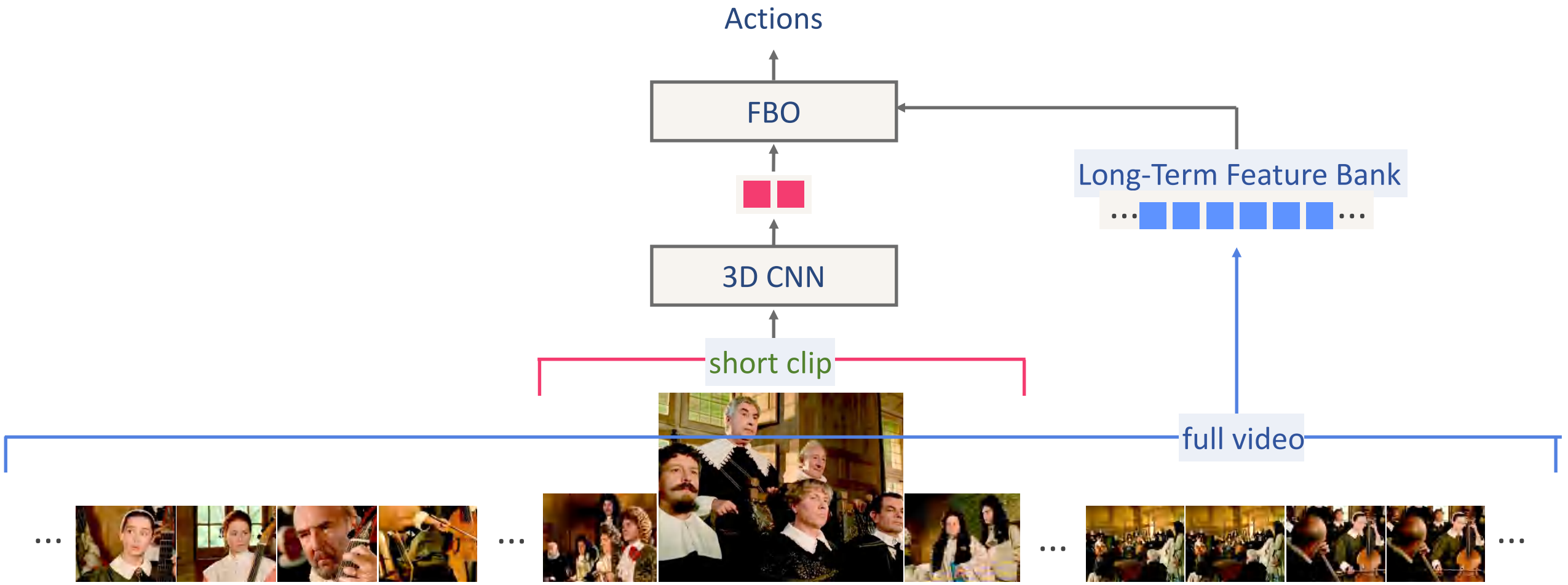
- Self-attention in the spatiotemporal domain allows long-range feature aggregation

X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In Proc. CVPR, 2018.

Background: Limited temporal input length of 3D ConvNets



New work: Long-Term Feature Banks for Video Understanding

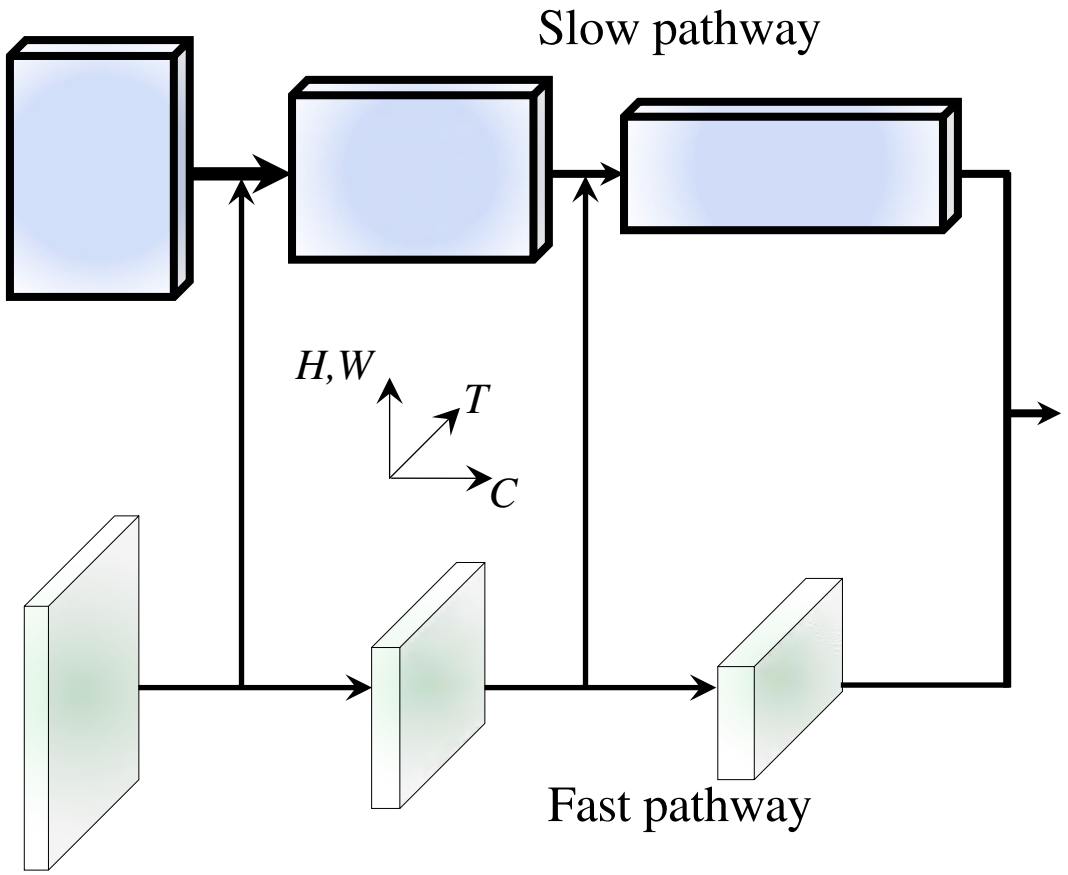


CY Wu, C. Feichtenhofer, H. Fan, K. He, P. Krähenbühl, R. Girshick Long-Term Feature Banks for Detailed Video Understanding. In Proc. CVPR, 2019.

This talk: SlowFast Networks for Video Recognition

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik and Kaiming He

- New backbone network for human action classification & detection

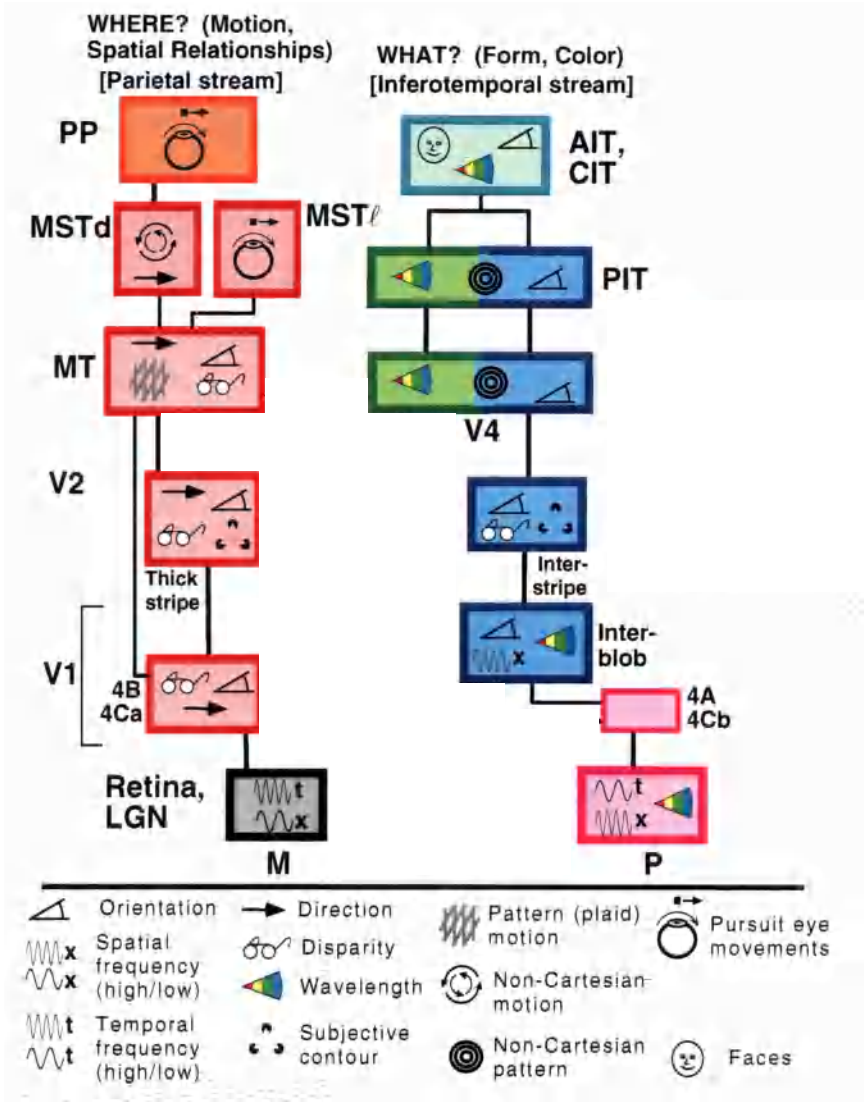


C. Feichtenhofer, H. Fan, J. Malik, K. He SlowFast Networks for Video Recognition. Tech. Report, arXiv, 2018

Motivation: Separate visual pathways for what and where

Magno cell properties[1]

- ➔ Minority of cells in LGN: ~20%
- ➔ Large receptive field
- ➔ High contrast sensitivity
- ➔ Able to differentiate only coarse stimuli
- ➔ Color Blind
- ➔ Processes information about depth & motion
- ➔ Fast conduction rate (more myelin)



Parvo cell properties[1]

- ➔ Majority of cells in LGN: ~80%
- ➔ Small receptive field
- ➔ Able to differentiate detailed stimuli
- ➔ Color Sensitive
- ➔ Processes information about color & detail
- ➔ Slow conduction rate (less myelin)

[1] https://www.ucalgary.ca/pip369/mod2/visualpathways/magnoparvo

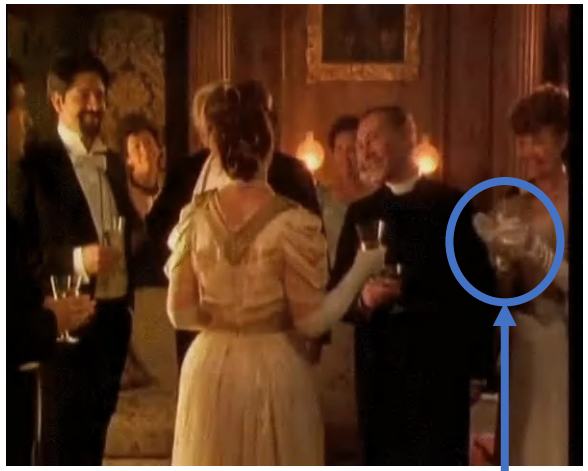
David C. Van Essen, Jack L. Gallant, Neural mechanisms of form and motion processing in the primate visual system, Neuron, Volume 13, Issue 1, July 1994, Pages 1-10, ISSN 0896-6273

The basic idea of SlowFast

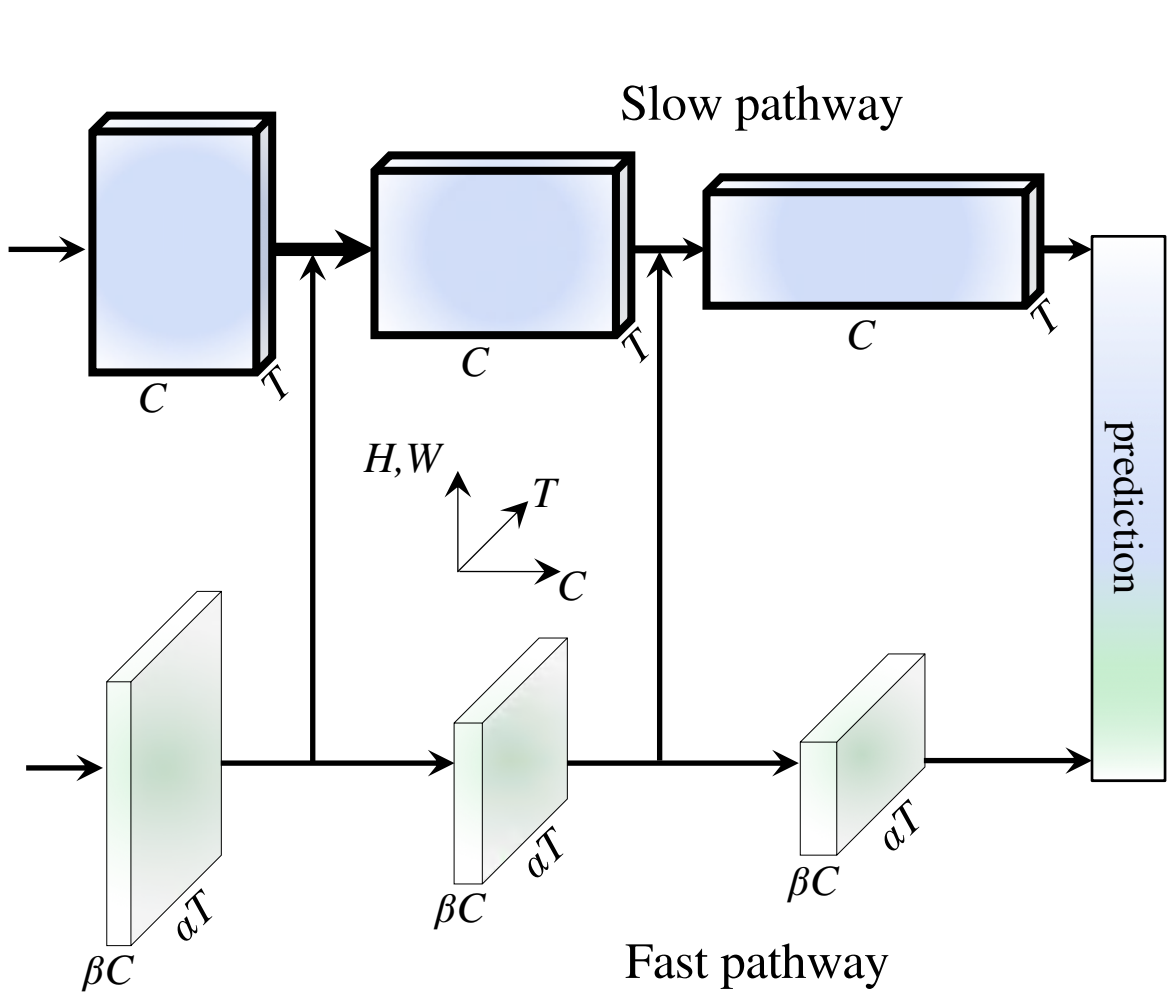
- The network consists of two pathways:
  - (i) a Slow pathway, operating at low frame rate, to capture spatial semantics
  - (ii) a Fast pathway, operating at high frame rate, to capture motion at fine temporal resolution

The basic idea of SlowFast networks

Slow



Fast



"Hand-clap"
(AVA detection annotation)

$e.g.\ \alpha = 8$
$\beta = 1/8$

Example instantiation of a SlowFast network

- Dimensions are $\{T \times S^2, C\}$
- Strides are {temporal, spatial^2}
- The backbone is ResNet-50
- Residual blocks are shown by brackets
- Non-degenerate temporal filters are underlined
- Here the speed ratio is $\alpha = 8$ and the channel ratio is $\beta = 1/8$
- Orange numbers mark fewer channels, for the Fast pathway
- Green numbers mark higher temporal resolution of the Fast pathway
- No temporal *pooling* is performed throughout the hierarchy

stage	*Slow* pathway	*Fast* pathway	output sizes $T \times S^2$
raw clip	-	-	64×224^2
data layer	stride 16, 1^2	stride **2**, 1^2	*Slow* : 4×224^2 *Fast* : 32×224^2
conv_1	1×7^2, 64 stride 1, 2^2	$\underline{5 \times 7^2}$, 8 stride 1, 2^2	*Slow* : 4×112^2 *Fast* : 32×112^2
pool_1	1×3^2 max stride 1, 2^2	1×3^2 max stride 1, 2^2	*Slow* : 4×56^2 *Fast* : 32×56^2
res_2	$\begin{bmatrix} 1 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \underline{3 \times 1^2}, 8 \\ \underline{1 \times 3^2}, 8 \\ 1 \times 1^2, 32 \end{bmatrix} \times 3$	*Slow* : 4×56^2 *Fast* : 32×56^2
res_3	$\begin{bmatrix} 1 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \underline{3 \times 1^2}, 16 \\ \underline{1 \times 3^2}, 16 \\ 1 \times 1^2, 64 \end{bmatrix} \times 4$	*Slow* : 4×28^2 *Fast* : 32×28^2
res_4	$\begin{bmatrix} \underline{3 \times 1^2}, 256 \\ \underline{1 \times 3^2}, 256 \\ 1 \times 1^2, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \underline{3 \times 1^2}, 32 \\ \underline{1 \times 3^2}, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 6$	*Slow* : 4×14^2 *Fast* : 32×14^2
res_5	$\begin{bmatrix} \underline{3 \times 1^2}, 512 \\ \underline{1 \times 3^2}, 512 \\ 1 \times 1^2, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \underline{3 \times 1^2}, 64 \\ \underline{1 \times 3^2}, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	*Slow* : 4×7^2 *Fast* : 32×7^2
global average pool, concate, fc			# classes

SlowFast training recipe, Kinetics action classification

- Kinetics has 240k training videos and 20k validation videos in 400 classes
- Our training recipe for training ***without*** ImageNet initialization (***inflation***)
- T = input size, τ = temporal stride

model	pre-train	$T \times \tau$	t-reduce	top-1	top-5	GFLOPs
3D R-50 [1]	ImageNet	32×2	2^3	73.3	90.7	33.1
3D R-50 (our recipe)	-	32×2	2^3	73.0	90.4	33.1
3D R-50 [1]	ImageNet	8×8	2^1	73.4	90.9	28.1
3D R-50 [1], recipe in [1]	-	8×8	2^1	69.4	88.6	28.1
3D R-50 (our recipe)	-	8×8	2^1	73.5	90.8	28.1

(a) **Baselines trained from scratch**: Using the same structure as [1], our training recipe achieves comparable results *without* ImageNet pre-training. "t-reduce" is the temporal downsampling factor in the network.

[1] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proc. CVPR*, 2018.

SlowFast ablations: Individual paths

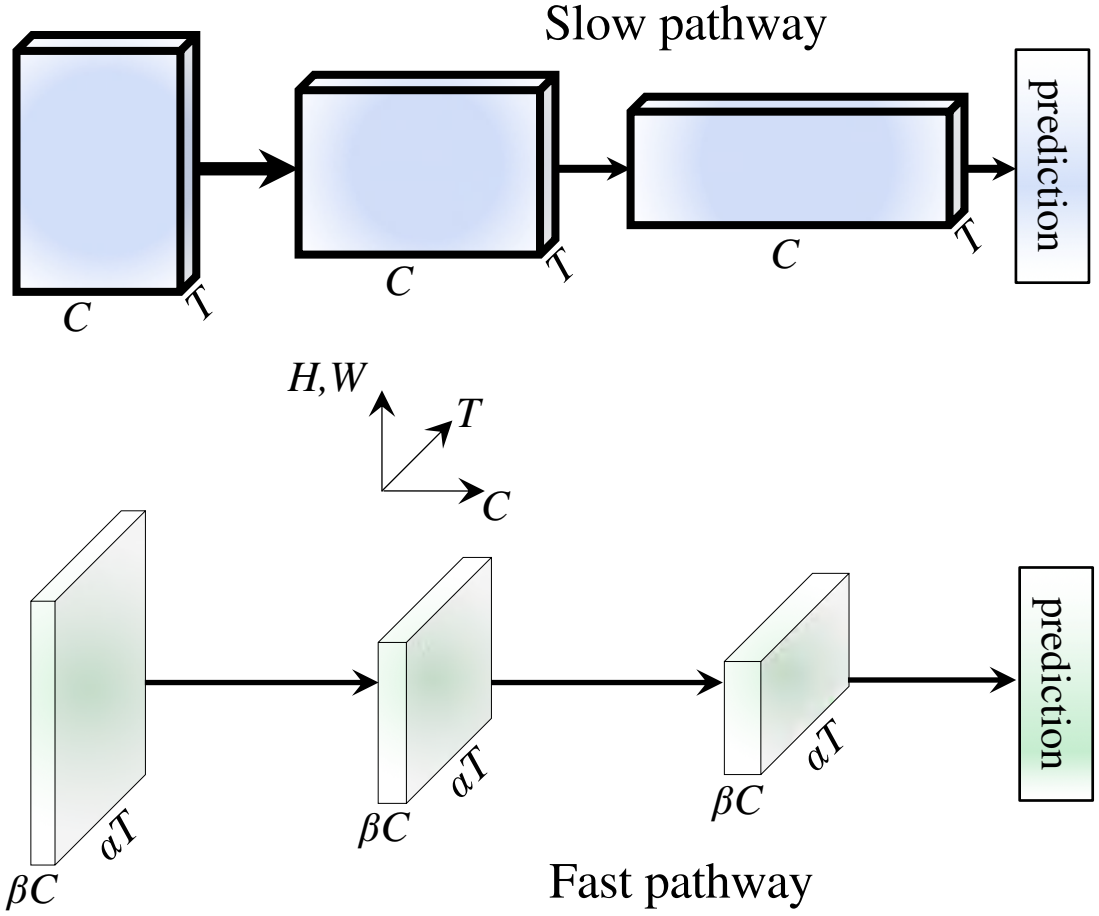
- Kinetics action classification dataset has 240k training videos and 20k validation videos in 400 classes

model	$T \times \tau$	t-reduce	top-1	top-5	GFLOPs
3D R-50	8×8	2^1	73.5	90.8	28.1
3D R-50	8×8	1	**74.6**	**91.5**	44.9
our Slow-only, R-50	4×16	1	72.6	90.3	**20.9**
our Fast-only, R-50	32×2	1	51.7	78.5	**4.9**

(b) **Individual pathways**: Training our Slow-only or Fast-only pathway alone, using the structure specified in Table 1. "t-reduce" is the total temporal downsampling factor within the network.

$\alpha = 8$

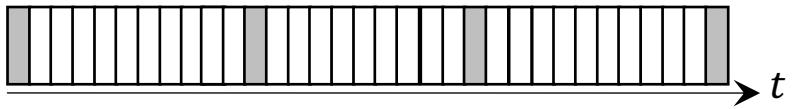
$\beta = 1/8$



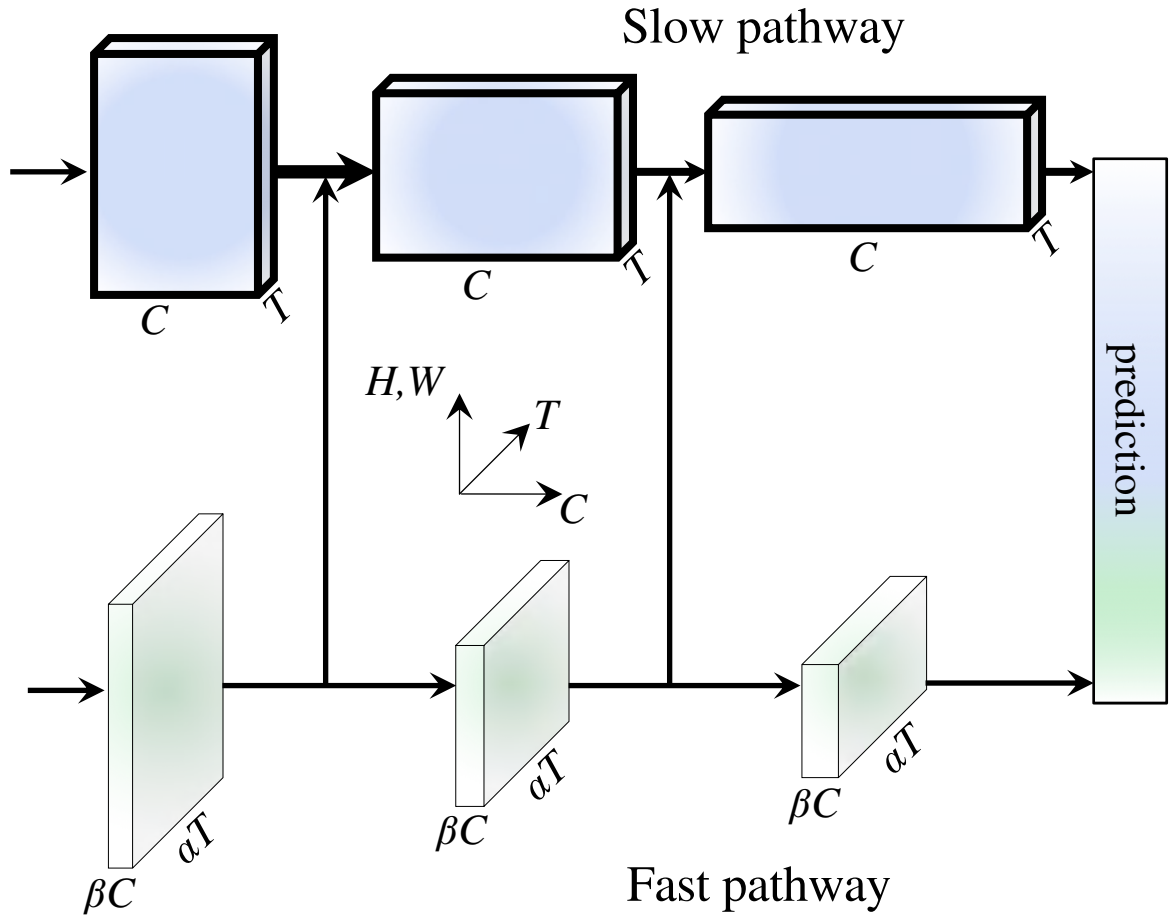
SlowFast ablations: Lateral fusion, Kinetics action classification

- Kinetics dataset has 240k training videos and 20k validation videos in 400 classes

	lateral	top-1	top-5	GFLOPs
Slow-only	-	72.6	90.3	20.9
Fast-only	-	51.7	78.5	**4.9**
SlowFast	-	73.5	90.3	26.2
SlowFast	T-sample	75.4	91.8	26.7
SlowFast	T-conv	**75.6**	**92.1**	27.6



(c) **SlowFast fusion**: Fusing Slow and Fast pathways with various lateral connections is consistently better than the Slow-only baseline. Backbone: R-50.



SlowFast ablations: Learning curves

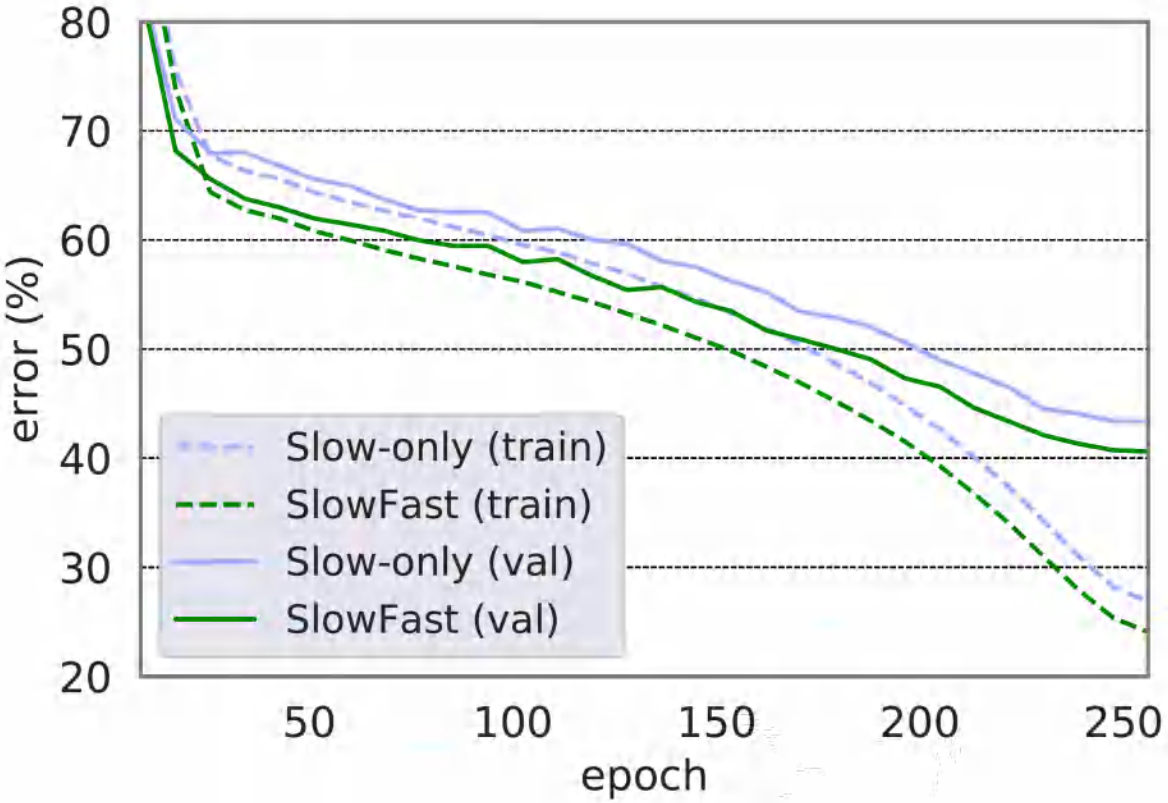


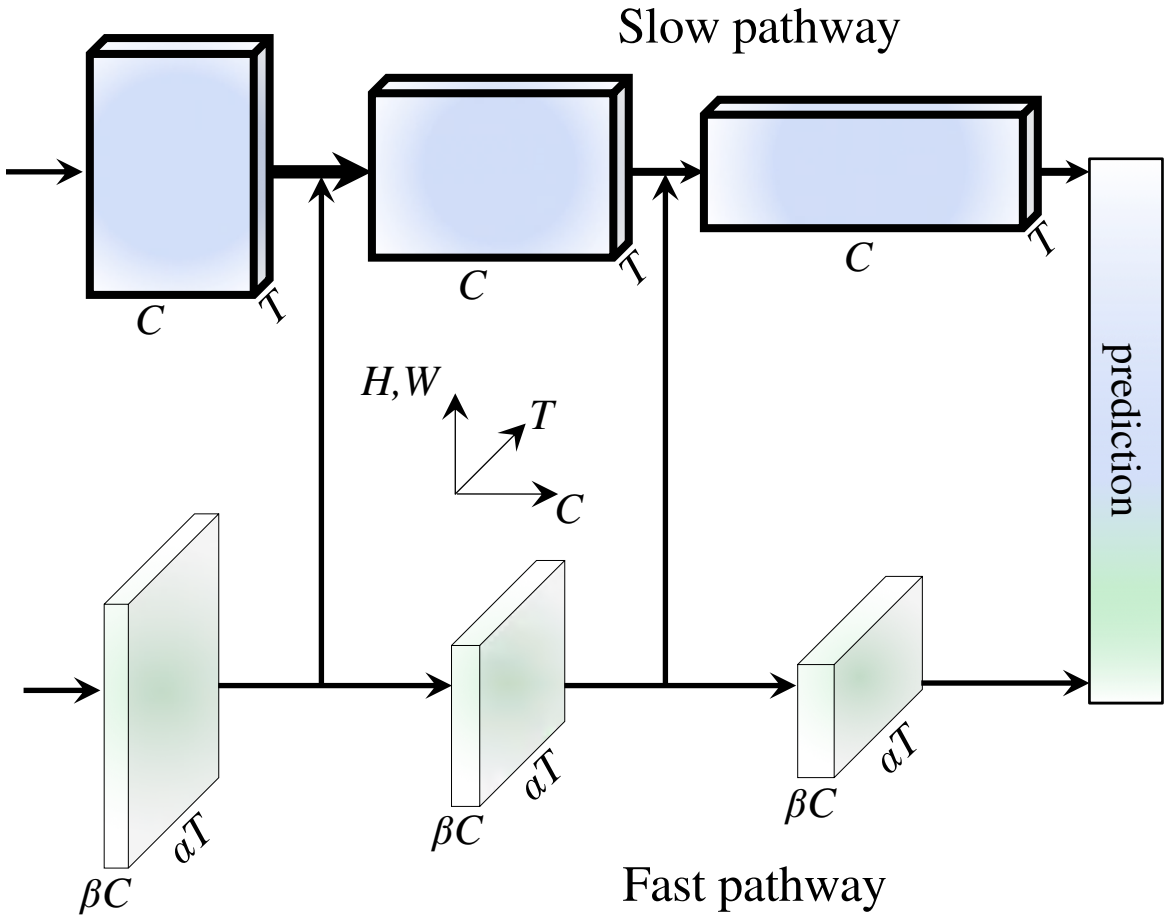
Figure 2. Training procedure on Kinetics for Slow-only (blue) *vs.* SlowFast (green) network. We show the top-1 training error (dash) and validation error (solid). The curves are single-crop *errors*; the video *accuracy* is 72.6% *vs.* 75.6% (see also Table 2c).

SlowFast ablations: Making the Fast path thin in channel dimension

- Kinetics dataset has 240k training videos and 20k validation videos in 400 classes

	top-1	top-5	GFLOPs
Slow-only	72.6	90.3	20.9
$\beta = 1/4$	75.6	91.7	41.7
1/6	**75.8**	92.0	32.0
1/8	75.6	**92.1**	27.6
1/12	75.2	91.8	25.1
1/16	75.1	91.7	23.4
1/32	74.2	91.3	21.9

(d) **Channel capacity ratio**: Varying values of β, the channel capacity ratio of the Fast pathway. Backbone: R-50.

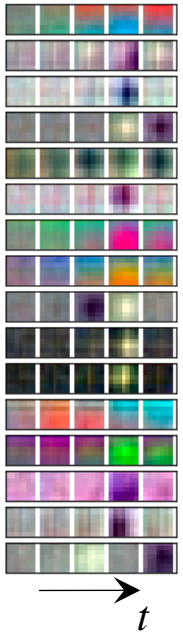


SlowFast ablatios: Weak input

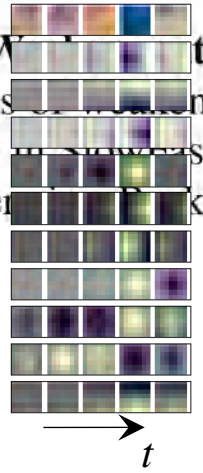
Fast pathway	spatial	top-1	top-5	GFLOPs
RGB	-	**75.6**	**92.1**	27.6
gray-scale	-	**75.5**	**91.9**	**26.1**
time diff	-	74.5	91.6	26.2

(e) **W tial input to Fast pathway**: Various ways ning spatial inputs to the Fast pathway st models. β=1/8 unless specified othe kbone R-50

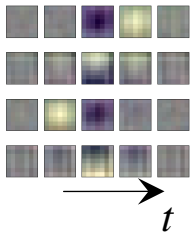
$\beta = 1/4$



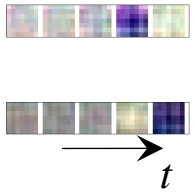
$\beta = 1/6$



$\beta = 1/16$

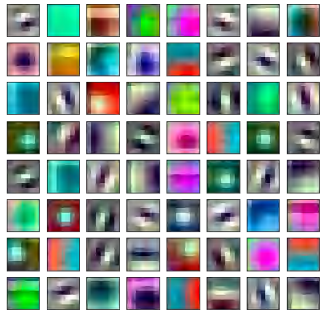


$\beta = 1/32$



Conv1 filters

Slow



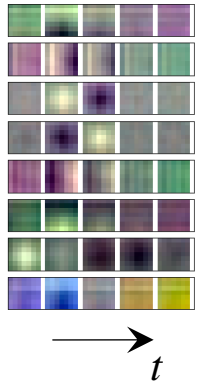
rgb

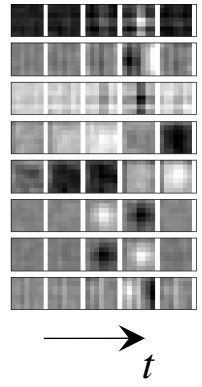
grayscale

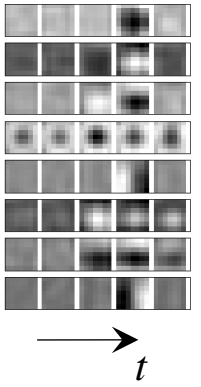
time diff

$\beta = 1/8$

Fast







rgb

grayscale

dt

SlowFast ablations: Temporal sampling rates

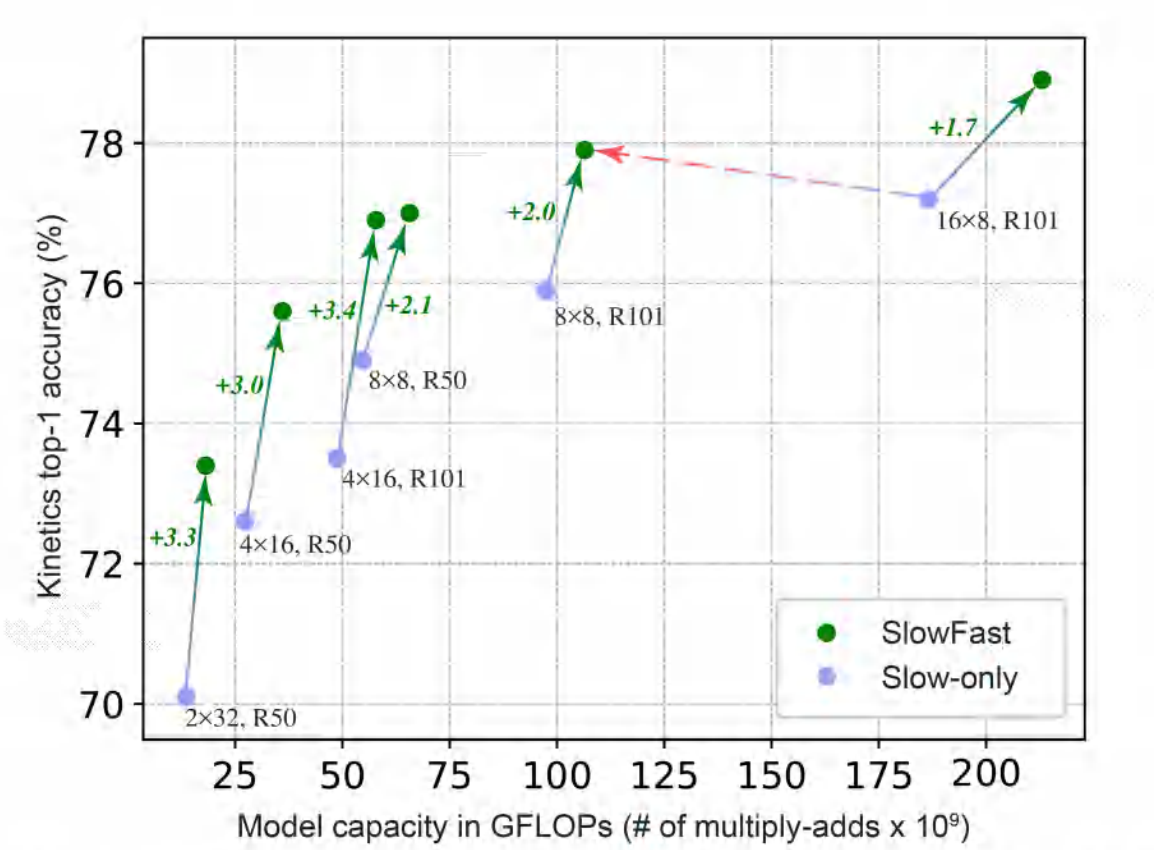
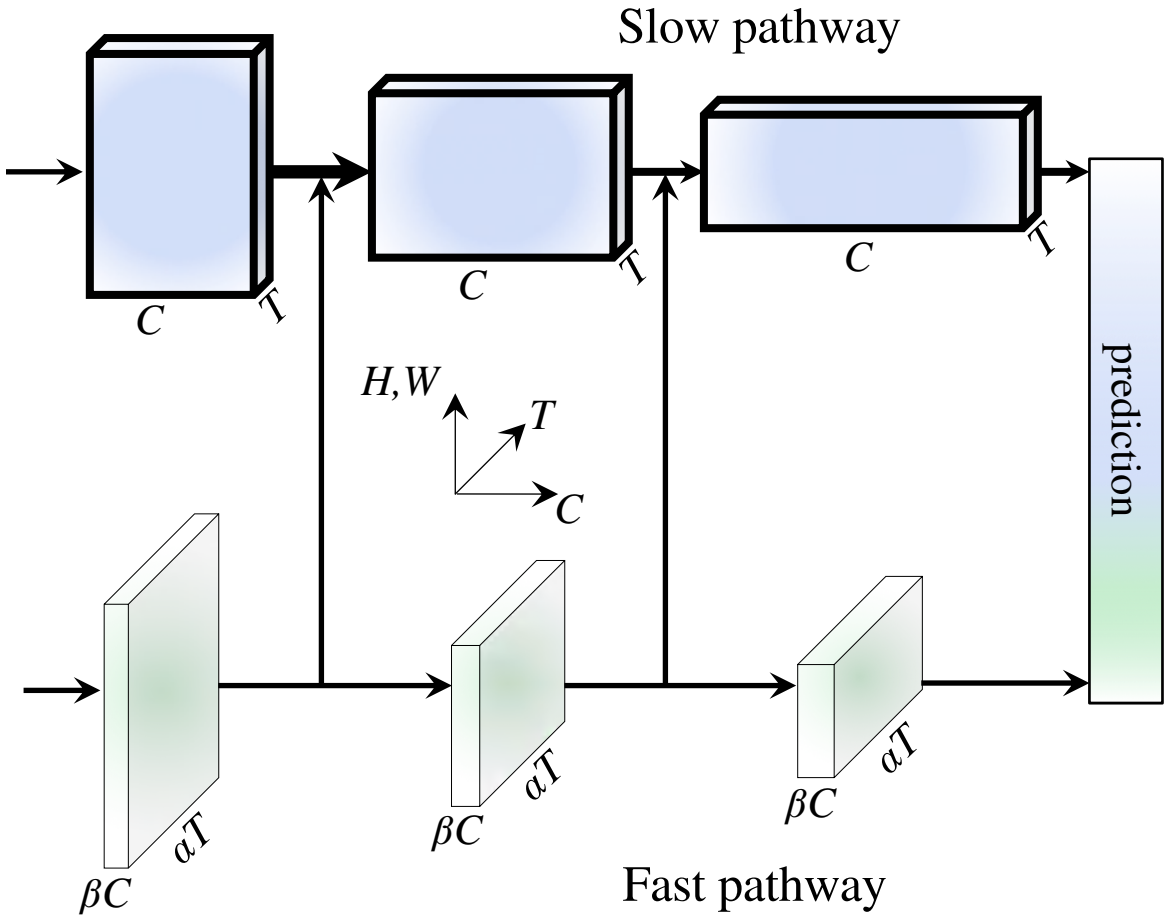


Figure 2. Accuracy (top-1) / complexity (FLOPs) tradeoff for the SlowFast (green) *vs*. Slow-only (blue) architectures on **Kinetics-400**. SlowFast is consistently better than its Slow-only counterpart in all cases (green arrows). SlowFast provides higher accuracy *and* lower cost than temporally heavy Slow-only (*e.g*. red arrow).



SlowFast: State-of-the-art comparison on Kinetics

model	flow	pretrain	top-1	top-5	inference GFLOPs×crops
I3D [1]		ImageNet	72.1	90.3	108 × N/A
Two-Stream I3D [1]	✓	ImageNet	75.7	92.0	216 × N/A
S3D-G [6]	✓	ImageNet	74.7	93.4	142.8 × N/A
Nonlocal R-50 [5]		ImageNet	76.5	92.6	282 × 30
Nonlocal R-101 [5]		ImageNet	77.7	93.3	359 × 30
R(2+1)D Flow [3]	✓	-	67.5	87.2	152 × 115
STC [2]		-	68.7	88.5	N/A × N/A
ARTNet [4]		-	69.2	88.3	23.5 × 250
S3D [6]		-	69.4	89.1	66.4 × N/A
ECO [7]		-	70.0	89.4	N/A × N/A
I3D [1]	✓	-	71.6	90.0	216 × N/A
R(2+1)D [3]		-	72.0	90.0	152 × 115
R(2+1)D [3]	✓	-	73.9	90.9	304 × 115
SlowFast, R50 (4×16)		-	75.6	92.1	36.1 × 30
SlowFast, R50		-	77.0	92.6	65.7 × 30
SlowFast, R50 + NL		-	77.7	93.1	80.8 × 30
SlowFast, R101		-	77.9	93.2	106 × 30
SlowFast, R101 + NL		-	**79.0**	**93.6**	115 × 30

+ 5.1% top-1

at 10% of FLOPs

Table 1. **Comparison with the state-of-the-art on Kinetics-400.** In the column of computational cost, we report the cost of a single spacetime crop and the numbers of such crops used. "N/A" indicates the numbers are not available for us. The SlowFast models are the $T\times\tau = 8\times8$ versions, unless specified.

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. CVPR*, 2017.

[2] A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool. Spatio-temporal channel correlation networks for action classification. In *Proc. ECCV*, 2018.

[3] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proc. CVPR*, 2018.

[4] L. Wang, W. Li, W. Li, and L. Van Gool. Appearance-and-relation networks for video classification. In *Proc. CVPR*, 2018.

[5] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proc. CVPR*, 2018.

[6] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 2017.

[7] M. Zolfaghari, K. Singh, and T. Brox. ECO: efficient convolutional network for online video understanding. In *Proc. ECCV*, 2018.

SlowFast: State-of-the-art comparison Kinetics-600

- Kinetics-600 has 392k training videos and 30k validation videos in 600 classes

model	pretrain	top-1	top-5	inference GFLOPs×crops
I3D [2]	-	71.9	90.1	108 × N/A
StNet-IRv2 RGB [3]	ImgNet+Kinetics400†	79.0	N/A	N/A
SlowFast, R50	-	79.9	94.5	65.7 ×30
SlowFast, R101	-	80.4	94.8	106 × 30
SlowFast, R101 + NL	-	**81.1**	**94.9**	115 × 30

Table 1. **Kinetics-600 results**. SlowFast models are with $T\times\tau = 8\times8$. †: The Kinetics-400 training set partially overlaps with the Kinetics-600 validation set, and "*it is therefore not ideal to evaluate models on Kinetics-600 that were pre-trained on Kinetics-400*" [1].

[1] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about Kinetics-600. *arXiv:1808.01340*, 2018.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. CVPR*, 2017.

[3] D. He, F. Li, Q. Zhao, X. Long, Y. Fu, and S. Wen. Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition. *arXiv preprint arXiv:1806.10319*, 2018.

SlowFast: State-of-the-art comparison Charades[1]

- Charades has 9.8k training videos and 1.8k validation videos in 157 classes
- Multi-label classification setting of longer activities spanning 30 seconds on average

model	pretrain	mAP	inference GFLOPs×views
CoViAR, R-50 [55]	ImageNet	21.9	N/A
Asyn-TF, VGG16 [39]	ImageNet	22.4	N/A
MultiScale TRN [58]	ImageNet	25.2	N/A
Nonlocal, R101 [52]	ImageNet+Kinetics400	37.5	544 × 30
STRG, R101+NL [53]	ImageNet+Kinetics400	39.7	630 × 30
our baseline (Slow-only)	Kinetics-400	39.0	187 × 30
SlowFast	Kinetics-400	41.8	213 × 30
SlowFast, +NL	Kinetics-400	42.5	234 × 30
SlowFast, +NL	Kinetics-600	**45.2**	234 × 30

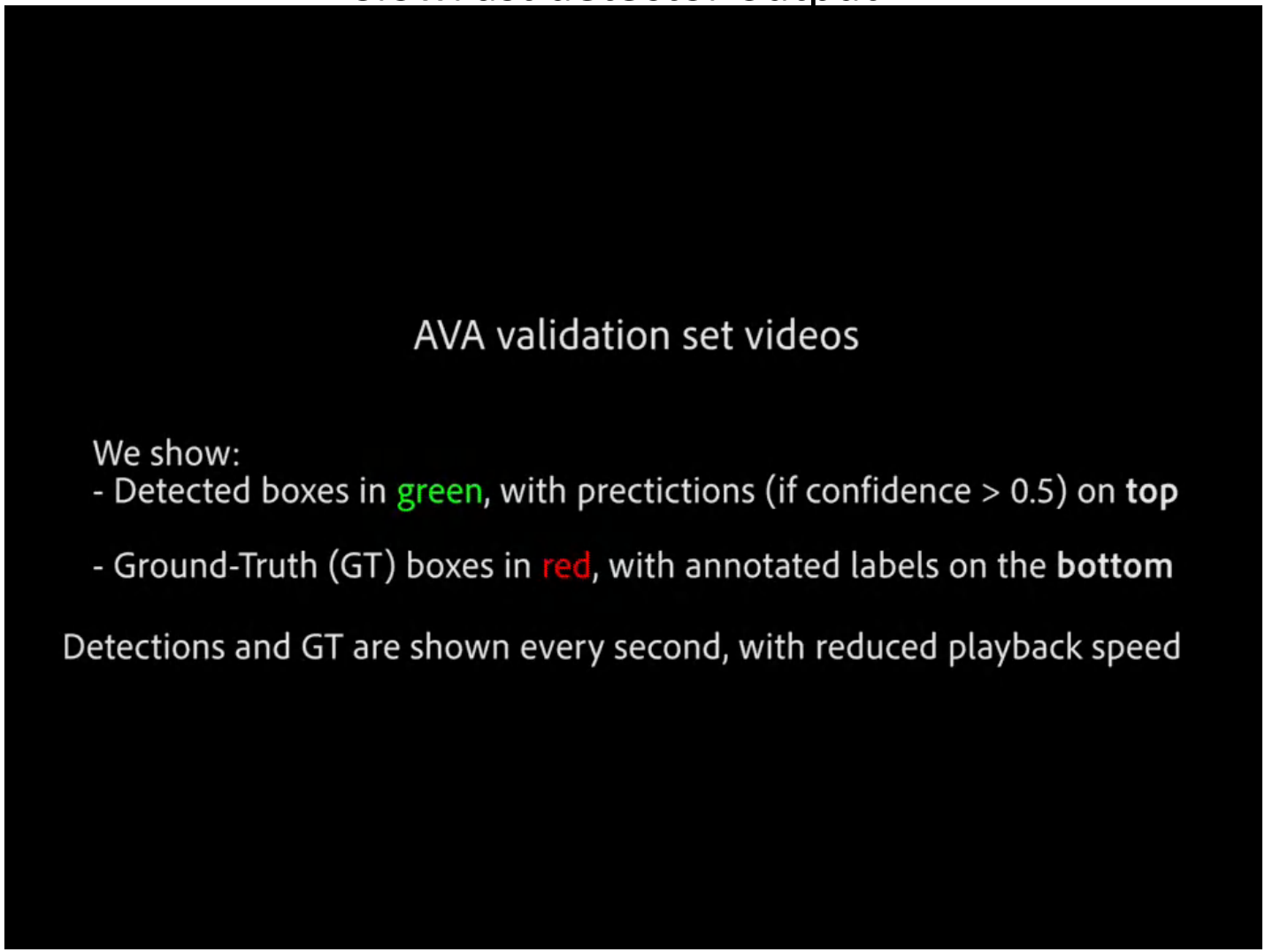
Table 4. **Comparison with the state-of-the-art on Charades**. All our variants are based on $T\times\tau = 16\times8$, R101.

[1]G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In ECCV, 2016. , CVPR 2016

Experiments: AVA[1] Action Detection

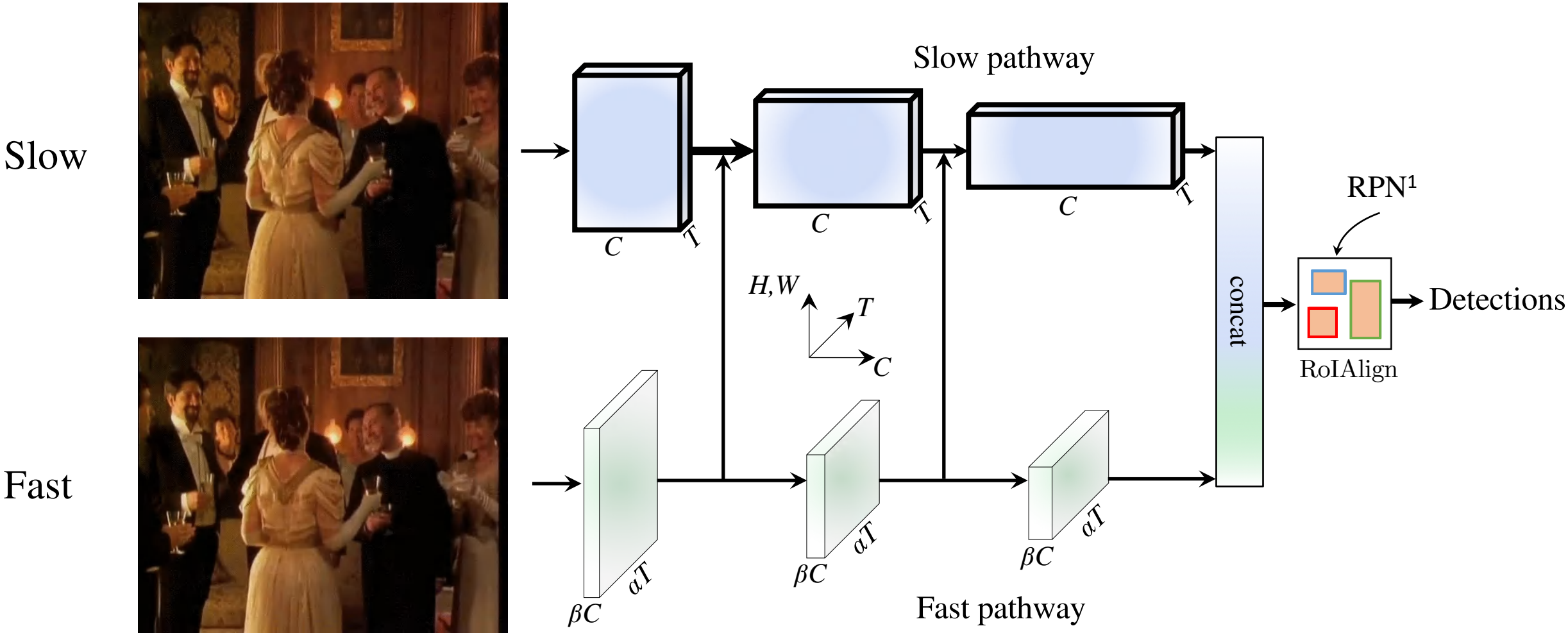
- Fine-scale localization of 80 different physical actions
- Data from 437 different movies and spatiotemporal labels are provided in a 1Hz interval
- 211k training and 57k validation video segments
- We follow the standard protocol of evaluating on 60 most freqent classes
- Every person is annotated with a bounding box and (possibly multiple) actions

SlowFast detector output



[1]Gu et al. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions, CVPR 2018

Experiments: AVA Action Detection



[1]Faster R-CNN with a ResNeXt-101-FPN backbone pretrained on COCO keypoints

SlowFast: state-of-the-art on AVA action detection

model	$T \times \tau$	α	mAP
Slow-only, R-50	4×16	-	19.0
SlowFast, R-50	4×16	8	24.2

+5.2mAP

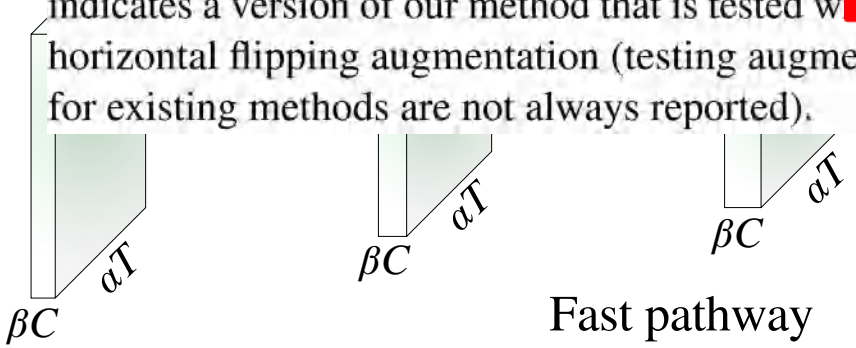
Table 3. **Comparison with the state-of-the-art** on AVA v2.1. "..." indicates a version of our method that is tested with ... horizontal flipping augmentation (testing augmentation strategies for existing methods are not always reported).

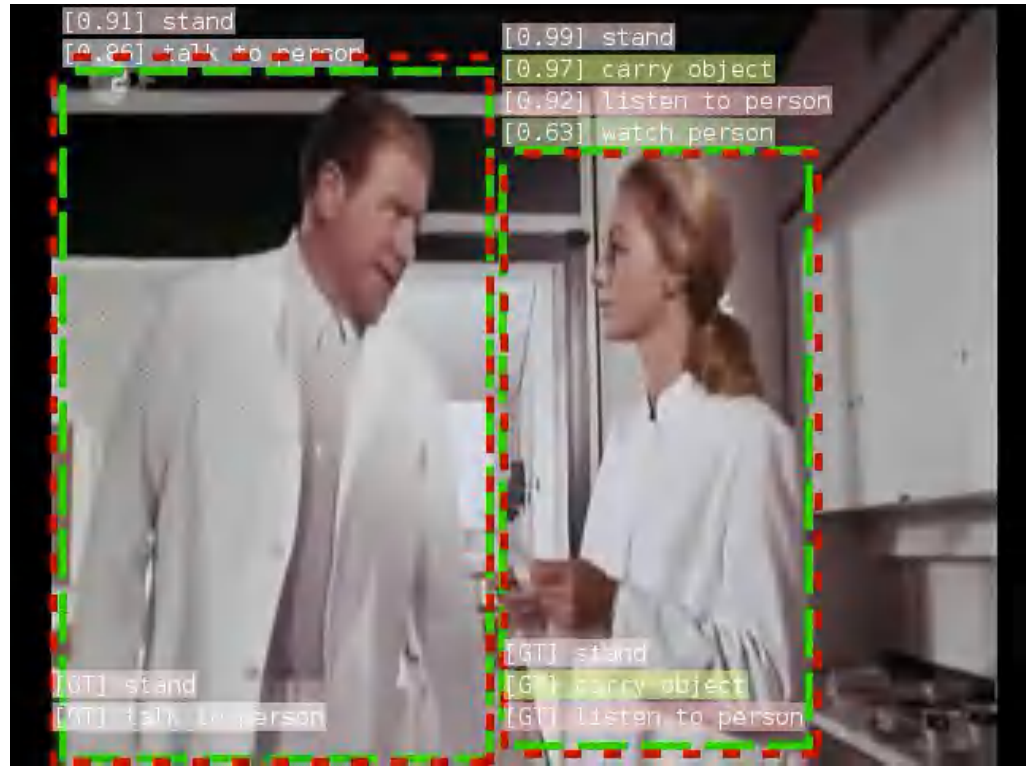
model	flow	video pretrain	val mAP	test mAP
I3D [2]		Kinetics-400	14.5	-
I3D [2]	✓	Kinetics-400	15.6	-
ACRN, S3D [4]	✓	Kinetics-400	17.4	-
ATR, R50 + NL [3]		Kinetics-400	20.0	-
ATR, R50 + NL [3]	✓	Kinetics-400	21.7	-
9-model ensemble [3]	✓	Kinetics-400	25.6	21.1
I3D [1]		Kinetics-600	21.9	21.0
SlowFast, R101		Kinetics-400	26.1	-
SlowFast, R101		Kinetics-600	26.8	**26.6**
SlowFast, R101 + NL		Kinetics-600	27.3	
SlowFast++, R101 + NL		Kinetics-600	**28.3**	

+6.4mAP

+13mAP

34.25mAP





RPN

RoIAlign

ACTIVITYNET
Large Scale Activity Recognition Challenge

SlowFast ablations: AVA class level performance

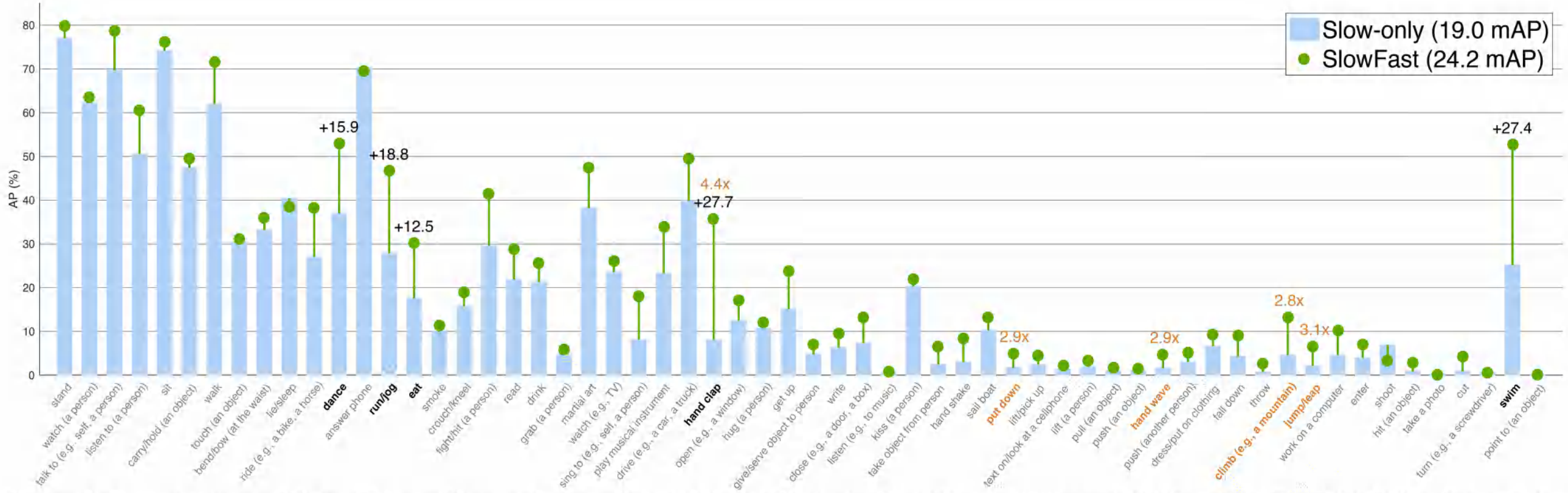


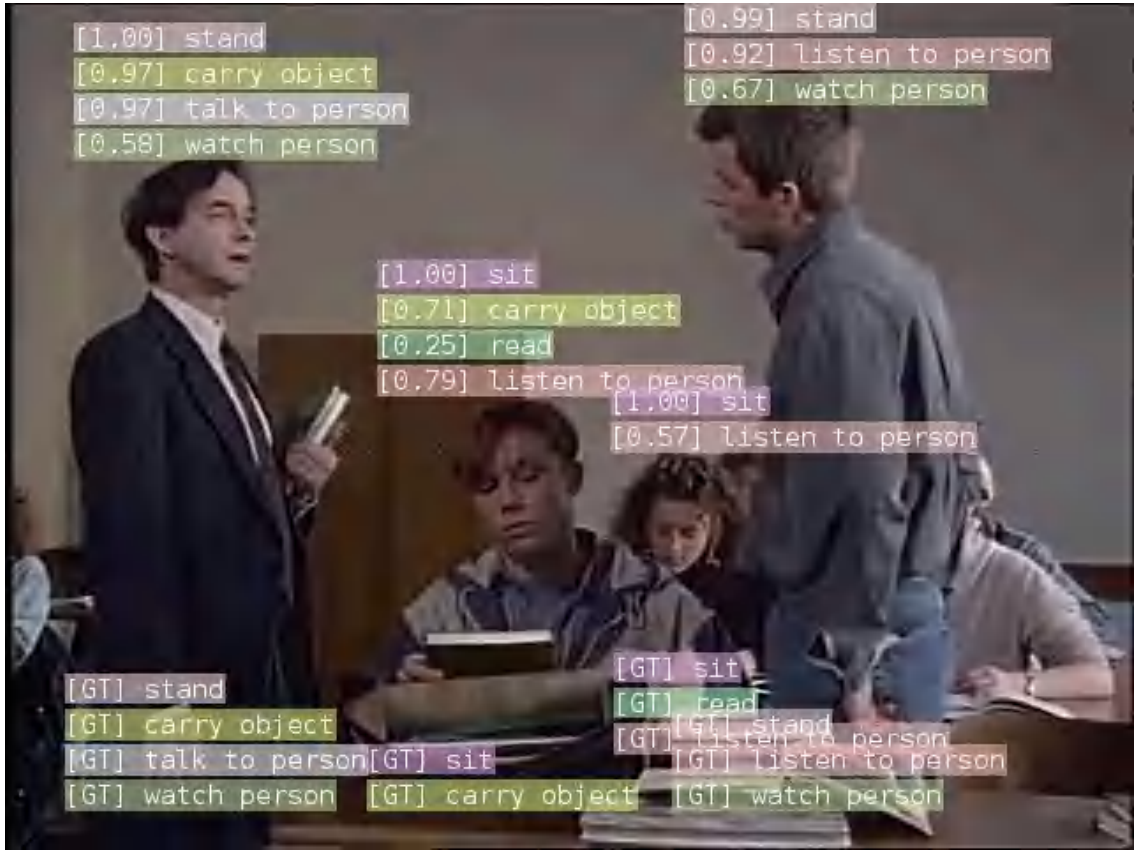
Figure 2. **Per-category AP on AVA**: a Slow-only baseline (19.0 mAP) *vs*. its SlowFast counterpart (24.2 mAP). The highlighted categories are the 5 highest absolute increase (**black**) or 5 highest relative increase with Slow-only AP $>$ 1.0 (**orange**). Categories are sorted by number of examples. Note that the SlowFast instantiation in this ablation is not our best-performing model.

Experiments: AVA Qualitative results

Experiments: AVA Qualitative results

Conclusion

- The time axis is a special dimension of video.
- *3D ConvNets* treat space and time uniformly.
- *SlowFast* and *Two-Stream* networks share motivation from biological studies.
- We investigate an architecture design that focuses on contrasting the speed along the temporal axis.
- The SlowFast architecture achieves state-of-the-art accuracy for video action classification and detection *without* need of any (e.g. ImageNet) pretraining.
- Given the mutual benefits of jointly modeling video with different temporal speeds, we hope that this concept can foster further research in video analysis.



FAIR Research Engineer

Menlo Park, CA
Seattle, WA

ACCELERATE AND SCALE CV RESEARCH

Familiarity with CV and ML
Ability to write high-quality and performance-critical code

wlo@fb.com