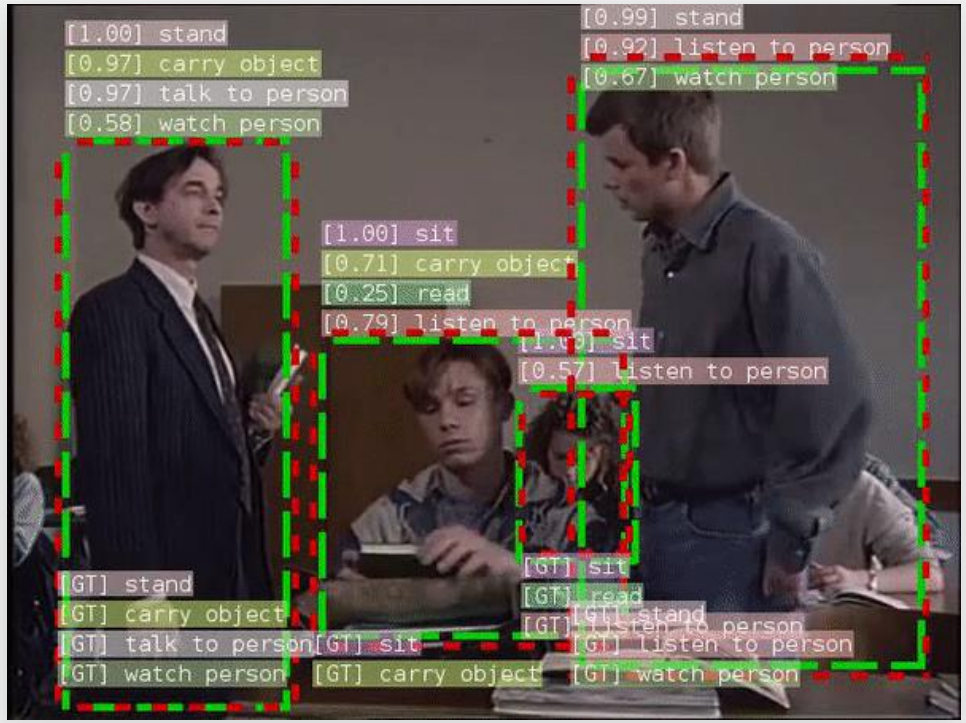
SlowFast Networks for Video Recognition

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He
Facebook AI Research (FAIR)
ICCV 2019

review by Jihun Kim

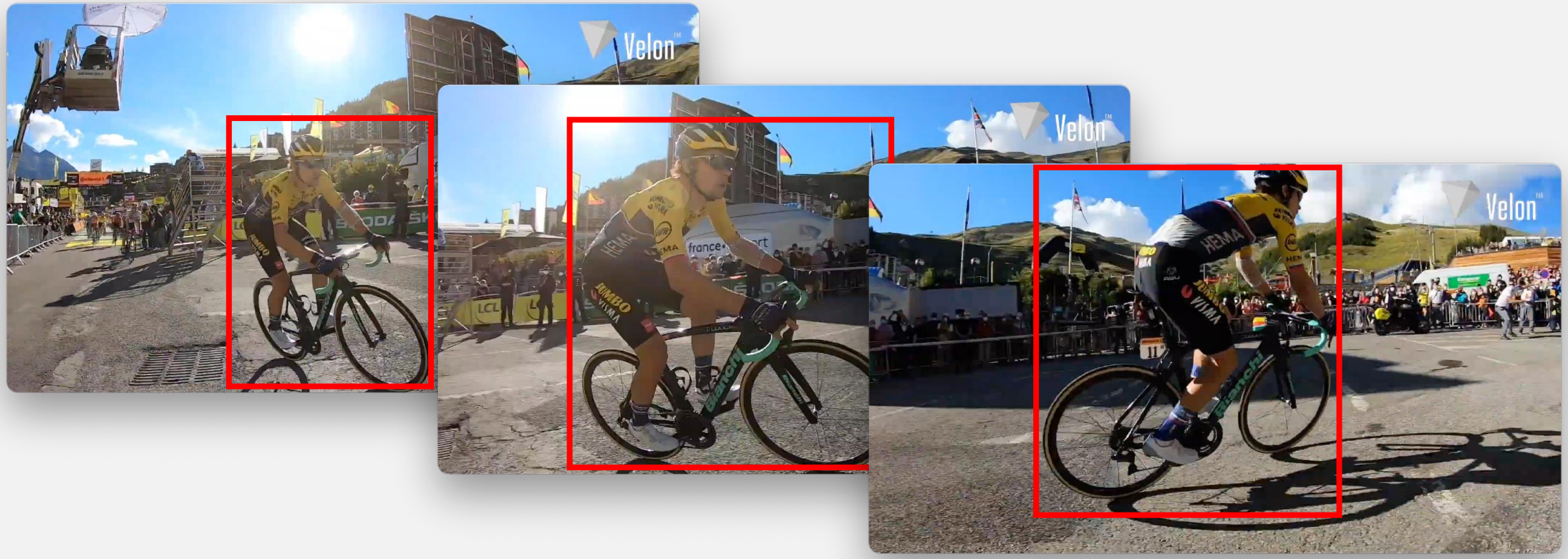
Introduction

Demo: Action Classification



Demo: Action Detection

Introduction



The categorical spatial semantics of the visual content often evolve slowly.
On the other hand, the motion being performed can evolve much faster.

SlowFast Networks

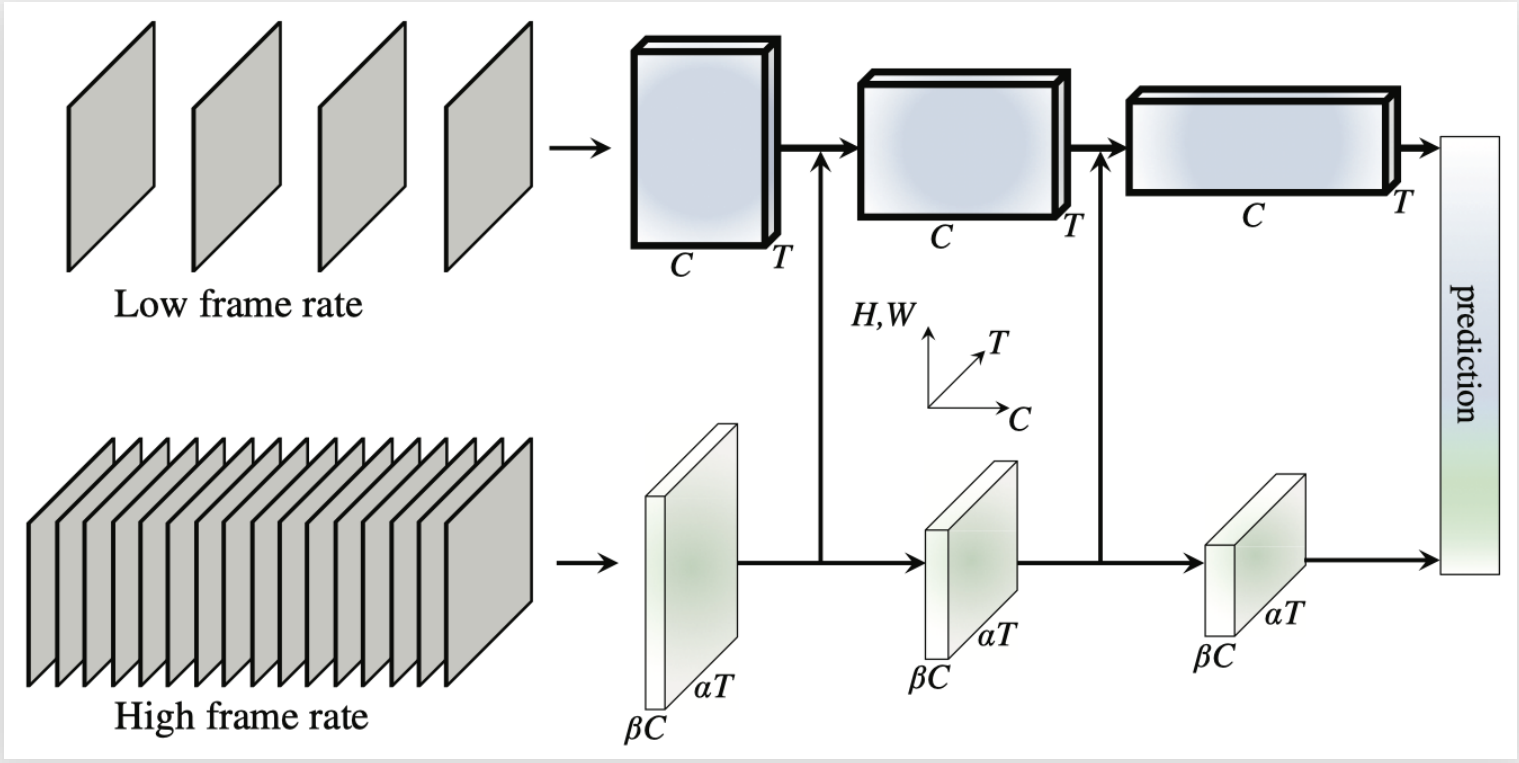
Architecture Overview

Slow Pathway

Captures semantic information

Fast Pathway

Captures rapidly changing motion

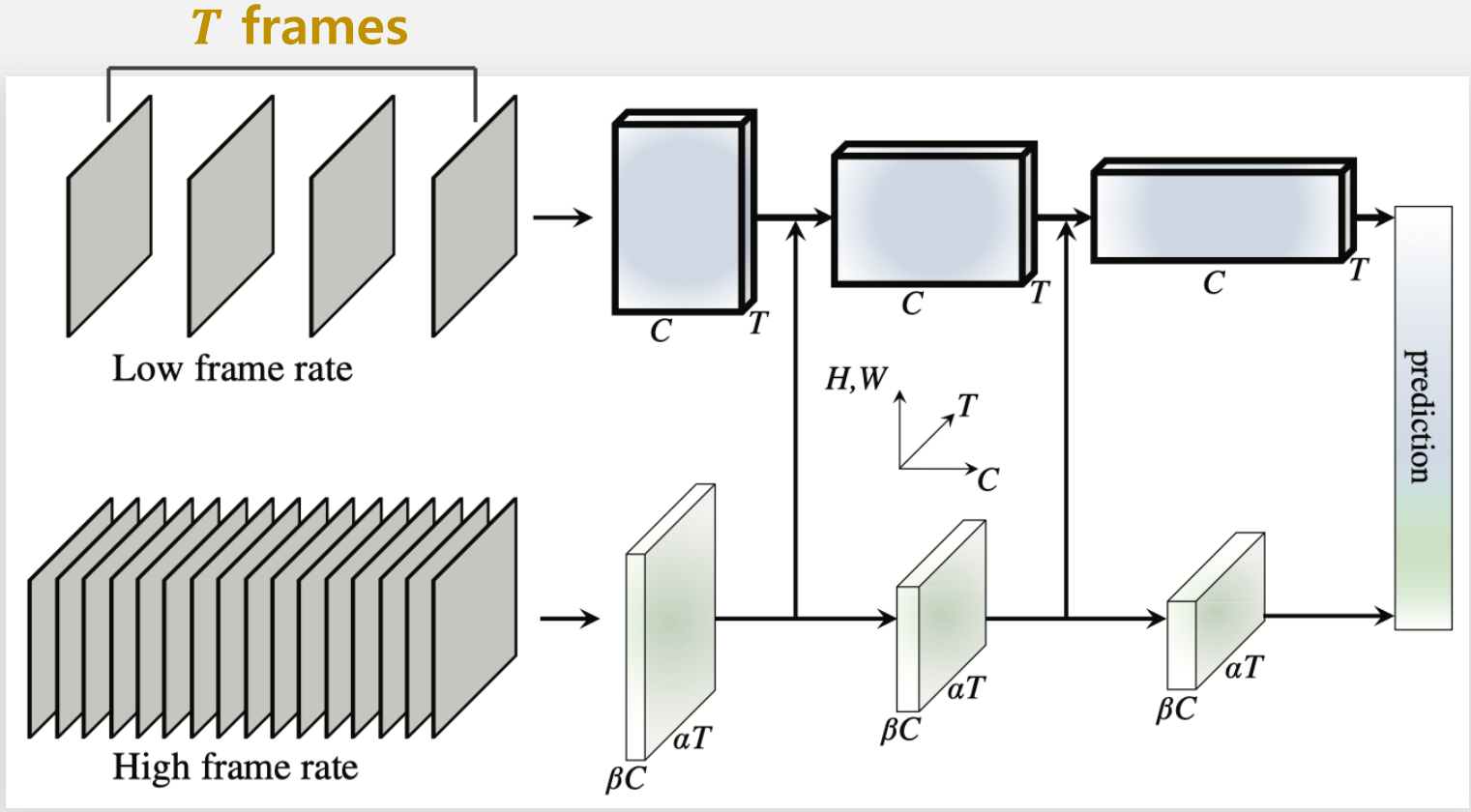


SlowFast Networks

Slow Pathway

Slow Pathway
Captures semantic information

Fast Pathway
Captures rapidly changing motion

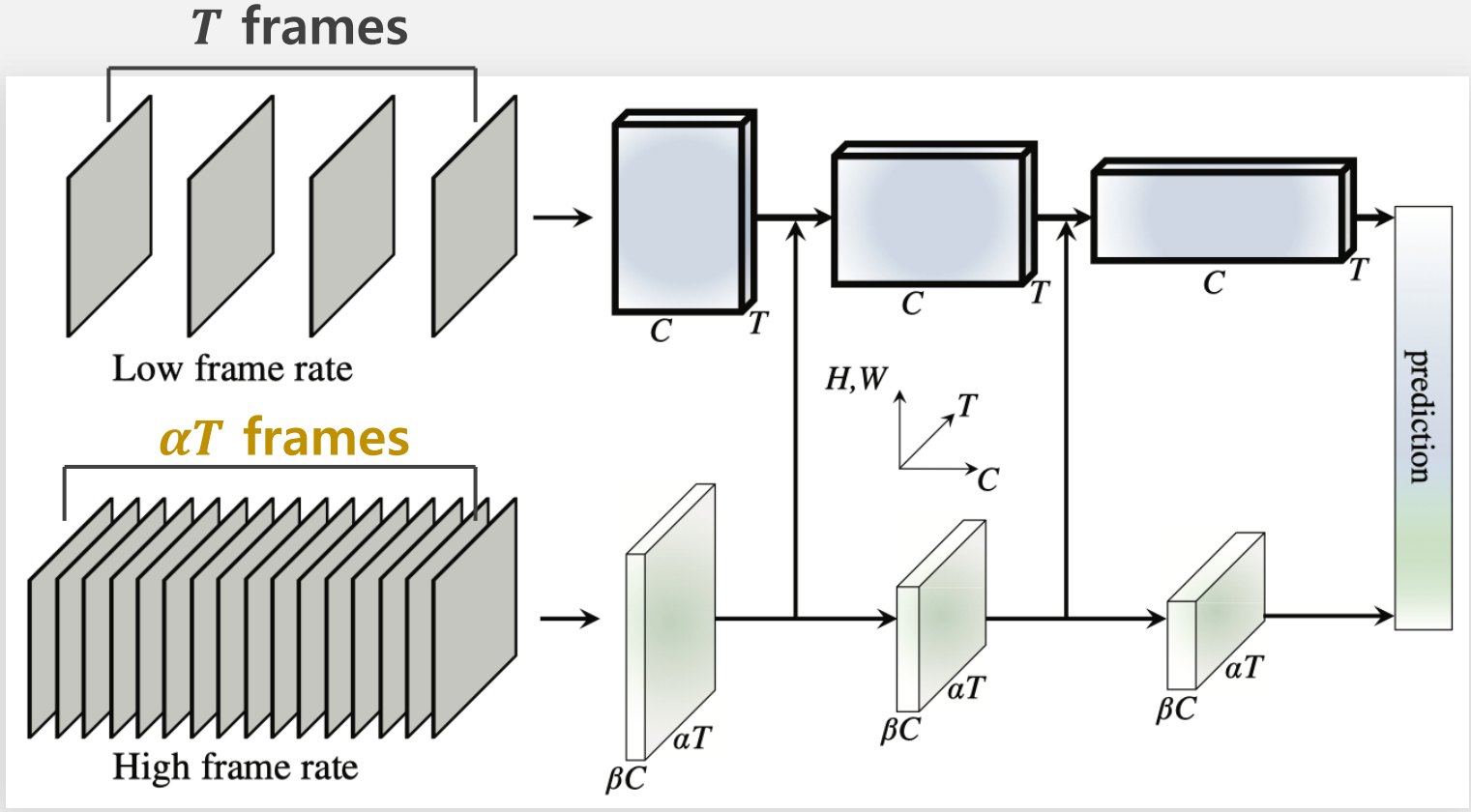


SlowFast Networks

Fast Pathway: High Frame Rate

Slow Pathway
Captures semantic information

Fast Pathway
Captures rapidly changing motion

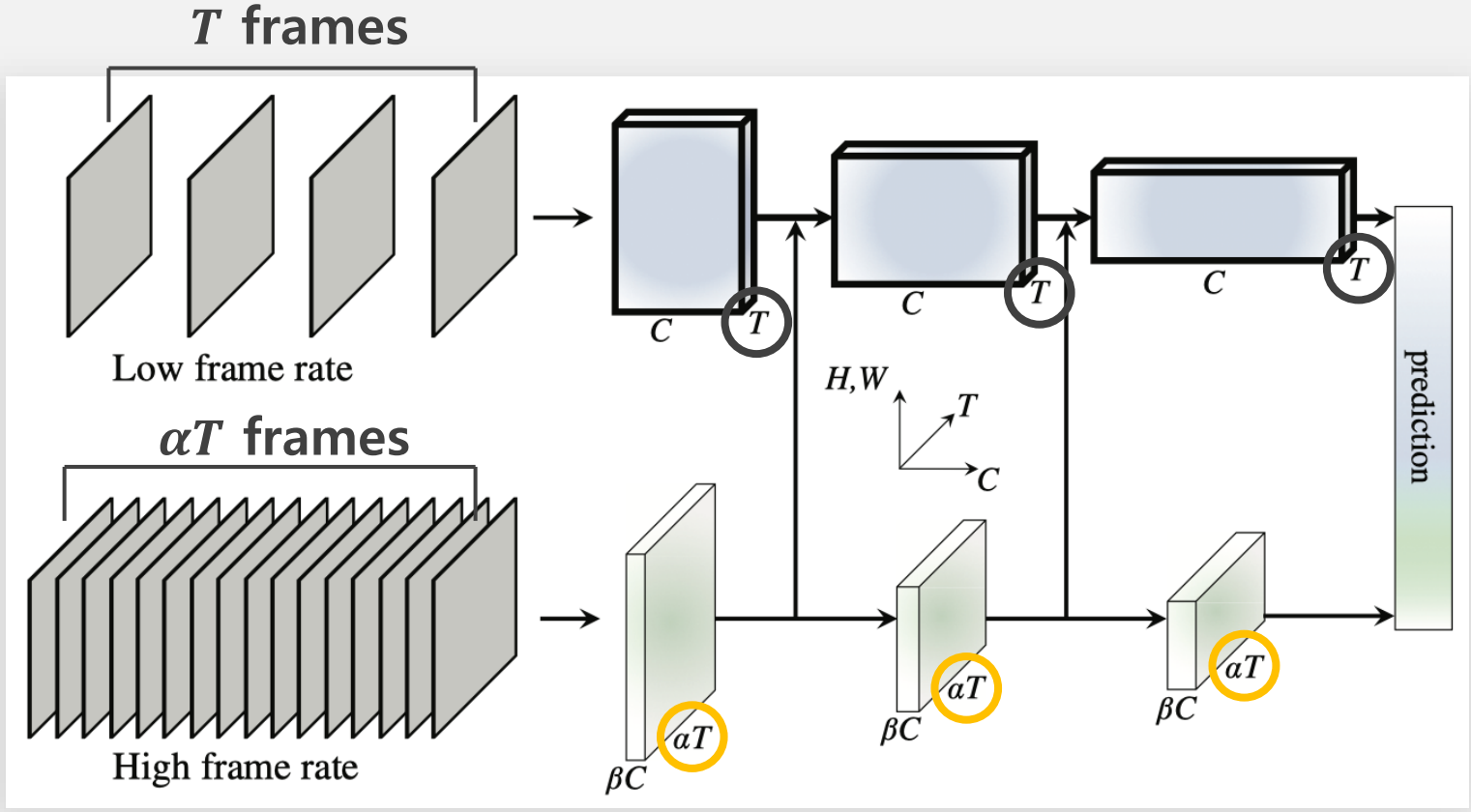


SlowFast Networks

Fast Pathway: High Temporal Resolution Features

Slow Pathway
Captures semantic information

Fast Pathway
Captures rapidly changing motion

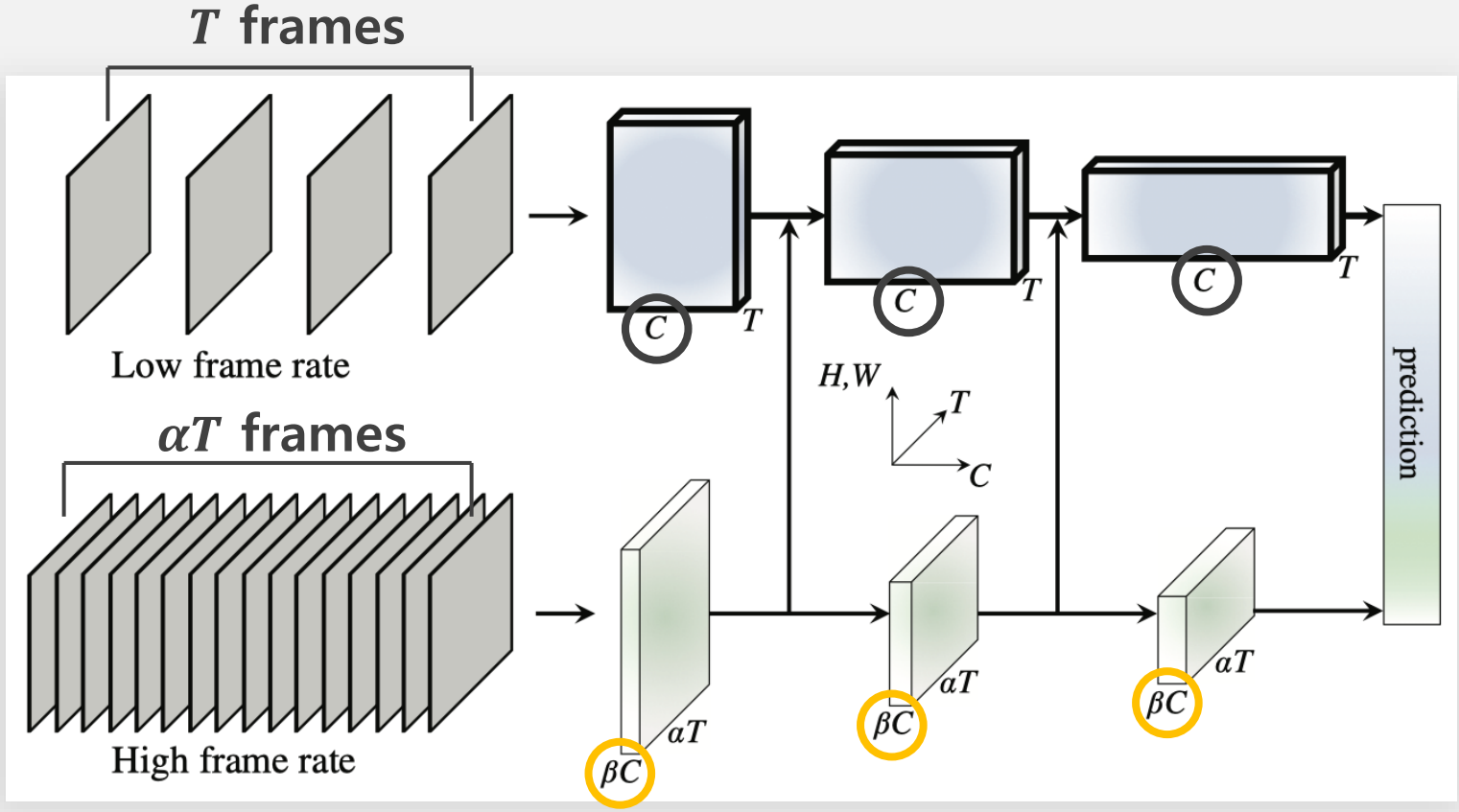


SlowFast Networks

Fast Pathway: Low Channel Capacity

Slow Pathway
Captures semantic information

Fast Pathway
Captures rapidly changing motion



SlowFast Networks

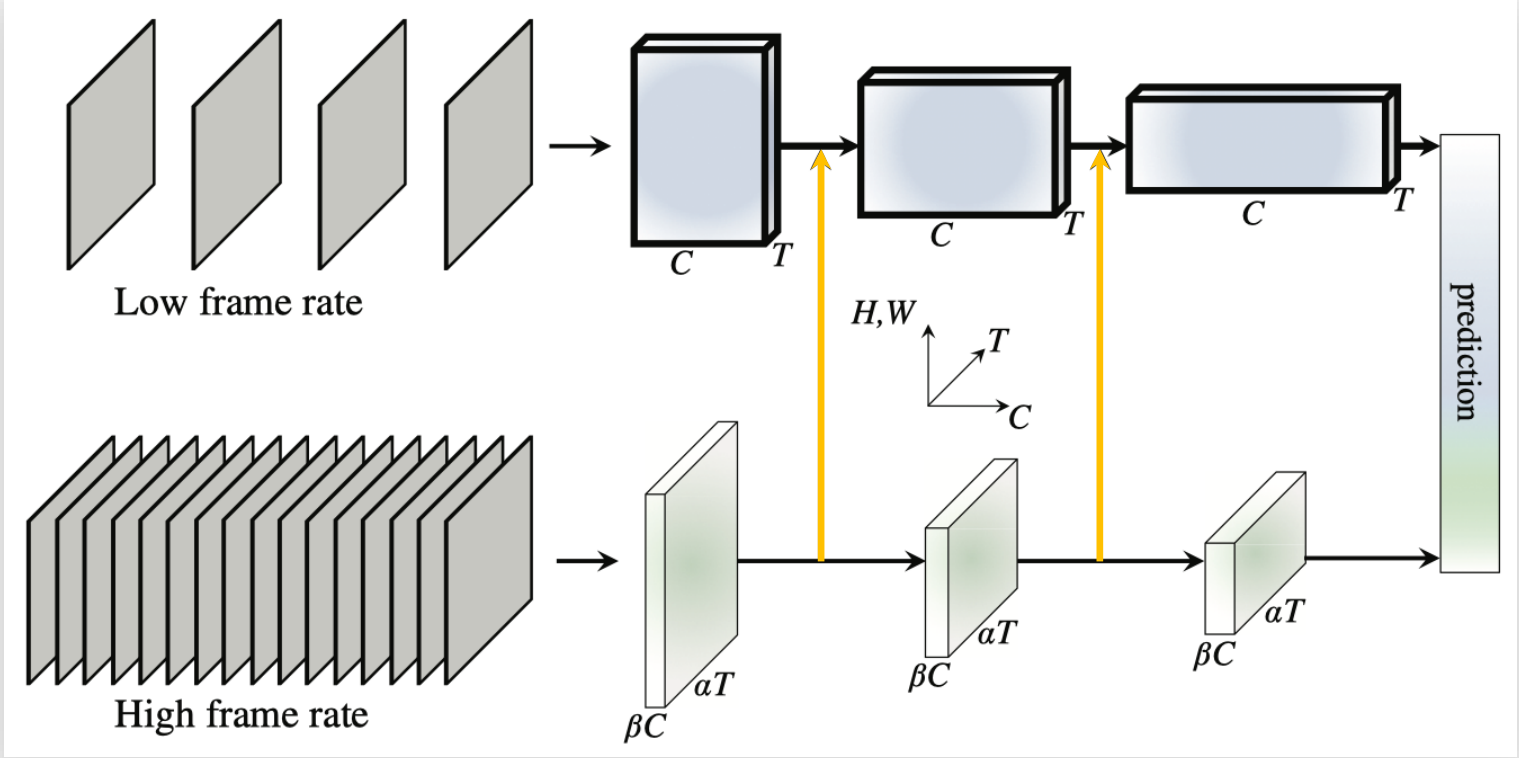
Lateral Connections

Slow Pathway

Captures semantic information

Fast Pathway

Captures rapidly changing motion



SlowFast Networks

Instantiations

Notation: $\{T \times S^2 \times C\}$

Slow Pathway
- 3D ResNet (Temporally Strided)

Fast Pathway
- Much higher temporal resolution (green)
- Much lower channel capacity (orange)

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	5×7^2, 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	$\begin{bmatrix} 1 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 8 \\ 1 \times 3^2, 8 \\ 1 \times 1^2, 32 \end{bmatrix} \times 3$	*Slow* : 4×56^2 *Fast* : 32×56^2
res_3	$\begin{bmatrix} 1 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 1^2, 16 \\ 1 \times 3^2, 16 \\ 1 \times 1^2, 64 \end{bmatrix} \times 4$	*Slow* : 4×28^2 *Fast* : 32×28^2
res_4	$\begin{bmatrix} 3 \times 1^2, 256 \\ 1 \times 3^2, 256 \\ 1 \times 1^2, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 1^2, 32 \\ 1 \times 3^2, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 6$	*Slow* : 4×14^2 *Fast* : 32×14^2
res_5	$\begin{bmatrix} 3 \times 1^2, 512 \\ 1 \times 3^2, 512 \\ 1 \times 1^2, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	*Slow* : 4×7^2 *Fast* : 32×7^2
global average pool, concate, fc			# classes

An example instantiations of the SlowFast network.

Experiments

Action Classification

model	flow	pretrain	top-1	top-5	GFLOPs×views
I3D [5]		ImageNet	72.1	90.3	108 × N/A
Two-Stream I3D [5]	✓	ImageNet	75.7	92.0	216 × N/A
S3D-G [61]	✓	ImageNet	77.2	93.0	143 × N/A
Nonlocal R50 [56]		ImageNet	76.5	92.6	282 × 30
Nonlocal R101 [56]		ImageNet	77.7	93.3	359 × 30
R(2+1)D Flow [50]	✓	-	67.5	87.2	152 × 115
STC [9]		-	68.7	88.5	N/A × N/A
ARTNet [54]		-	69.2	88.3	23.5 × 250
S3D [61]		-	69.4	89.1	66.4 × N/A
ECO [63]		-	70.0	89.4	N/A × N/A
I3D [5]	✓	-	71.6	90.0	216 × N/A
R(2+1)D [50]		-	72.0	90.0	152 × 115
R(2+1)D [50]	✓	-	73.9	90.9	304 × 115
SlowFast 4×16, R50		-	75.6	92.1	36.1 × 30
SlowFast 8×8, R50		-	77.0	92.6	65.7 × 30
SlowFast 8×8, R101		-	77.9	93.2	106 × 30
SlowFast 16×8, R101		-	78.9	93.5	213 × 30
SlowFast 16×8, R101+NL		-	**79.8**	**93.9**	234 × 30

Comparison with the SOTA on Kinetics-400.

model	pretrain	top-1	top-5	GFLOPs×views
I3D [3]	-	71.9	90.1	108 × N/A
StNet-IRv2 RGB [21]	ImgNet+Kin400	79.0	N/A	N/A
SlowFast 4×16, R50	-	78.8	94.0	36.1 × 30
SlowFast 8×8, R50	-	79.9	94.5	65.7 ×30
SlowFast 8×8, R101	-	80.4	94.8	106 × 30
SlowFast 16×8, R101	-	81.1	95.1	213 × 30
SlowFast 16×8, R101+NL	-	**81.8**	**95.1**	234 × 30

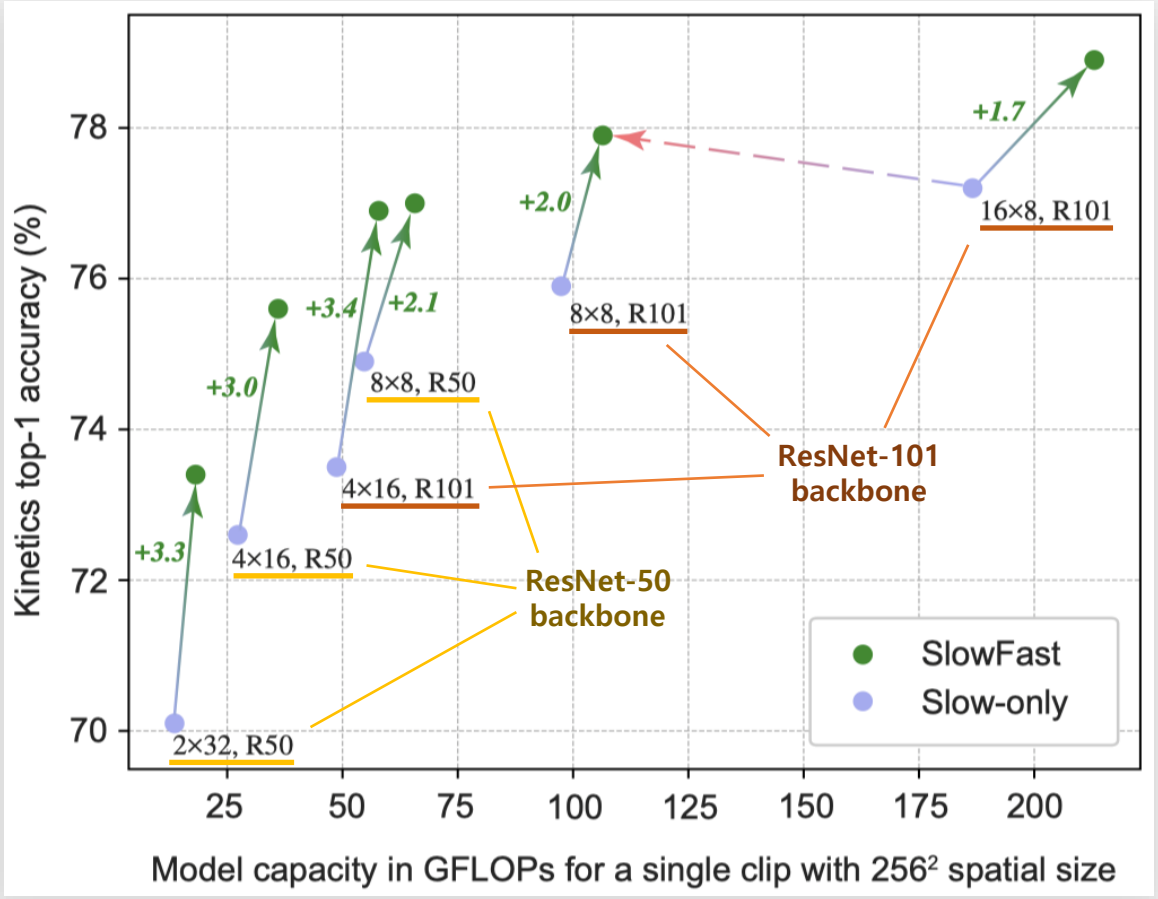
Comparison with the SOTA on Kinetics-600.

model	pretrain	mAP	GFLOPs×views
CoViAR, R-50 [59]	ImageNet	21.9	N/A
Asyn-TF, VGG16 [42]	ImageNet	22.4	N/A
MultiScale TRN [62]	ImageNet	25.2	N/A
Nonlocal, R101 [56]	ImageNet+Kinetics400	37.5	544 × 30
STRG, R101+NL [57]	ImageNet+Kinetics400	39.7	630 × 30
our baseline (Slow-only)	Kinetics-400	39.0	187 × 30
SlowFast	Kinetics-400	42.1	213 × 30
SlowFast, +NL	Kinetics-400	42.5	234 × 30
SlowFast, +NL	Kinetics-600	**45.2**	234 × 30

Comparison with the SOTA on Charades.

Ablation Study

Action Classification

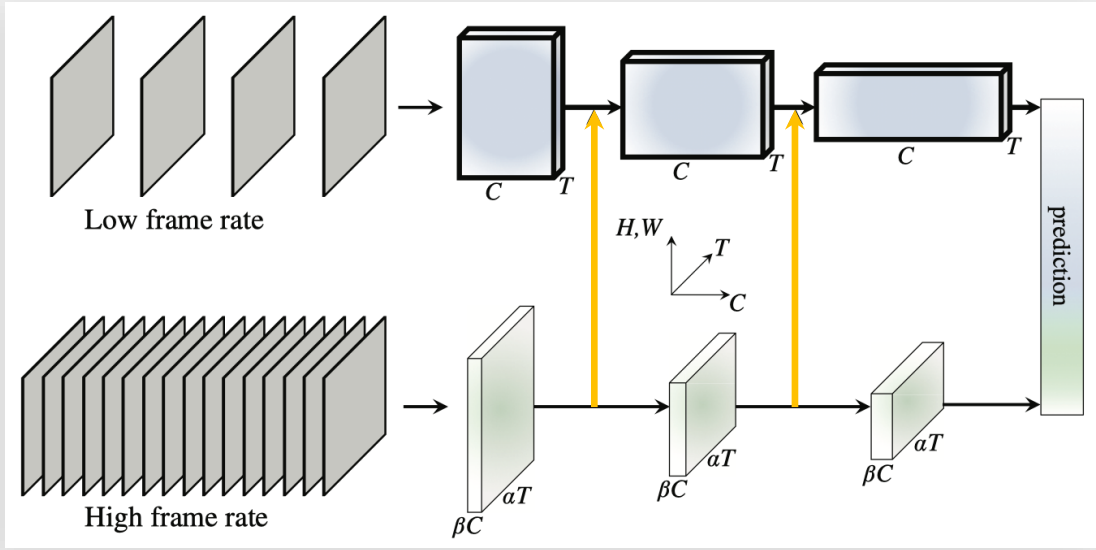


Accuracy/complexity tradeoff on Kinetics-400 for the SlowFast (green) vs. Slow-only (blue) architectures.

Ablation Study

Action Classification

	lateral	top-1	top-5	GFLOPs
Slow-only	-	72.6	90.3	27.3
Fast-only	-	51.7	78.5	**6.4**
SlowFast	-	73.5	90.3	34.2
SlowFast	TtoC, sum	74.5	91.3	34.2
SlowFast	TtoC, concat	74.3	91.0	39.8
SlowFast	T-sample	75.4	91.8	34.9
SlowFast	T-conv	**75.6**	**92.1**	36.1

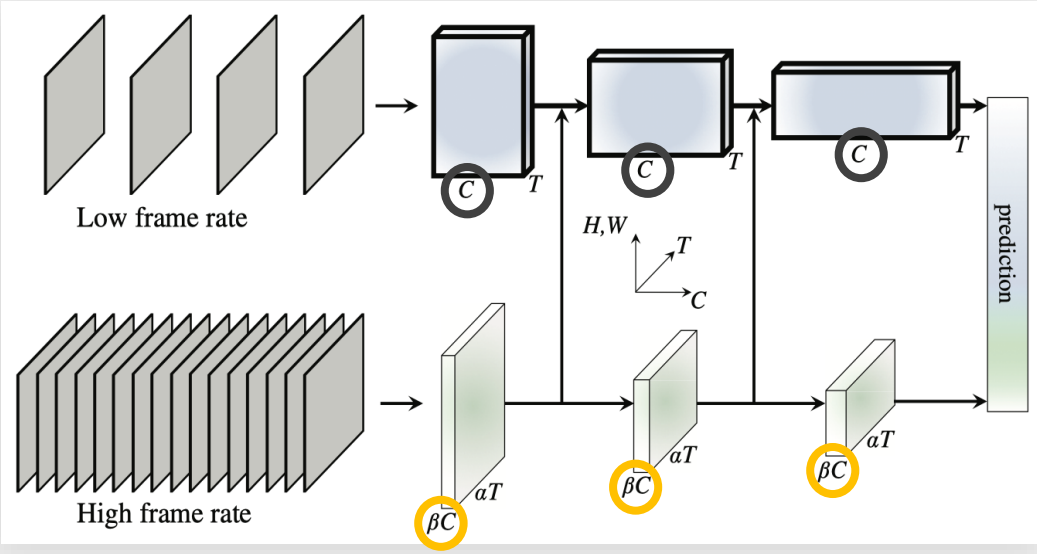


SlowFast Fusion:
Fusing Slow and Fast pathways with various types of lateral connections.

Ablation Study

Action Classification

	top-1	top-5	GFLOPs
Slow-only	72.6	90.3	27.3
$\beta = 1/4$	75.6	91.7	54.5
1/6	**75.8**	92.0	41.8
1/8	75.6	**92.1**	36.1
1/12	75.2	91.8	32.8
1/16	75.1	91.7	30.6
1/32	74.2	91.3	28.6

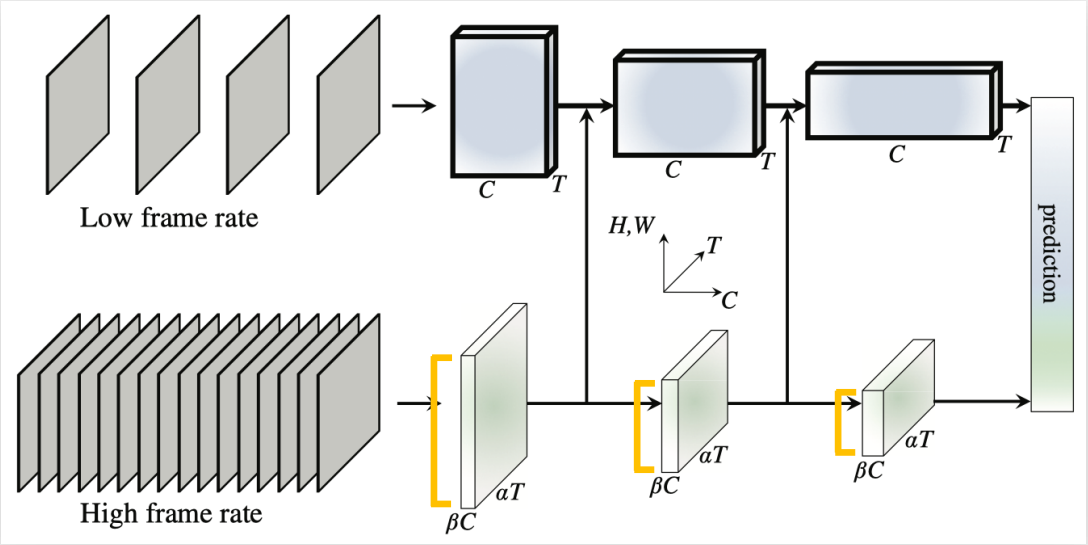


Channel capacity ratio:
Varying the channel capacity ratio of the Fast pathway to make SlowFast lightweight.

Ablation Study

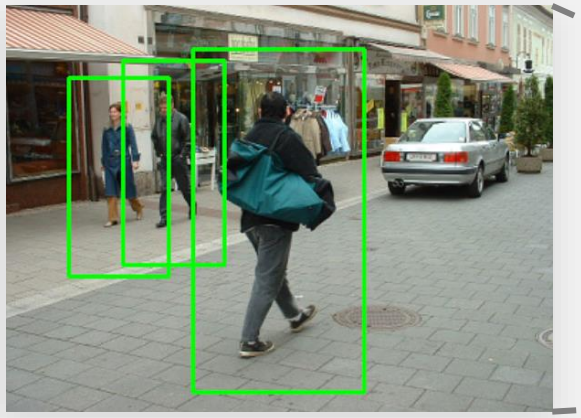
Action Classification

Fast pathway	spatial	top-1	top-5	GFLOPs
RGB	-	**75.6**	**92.1**	36.1
RGB, β=1/4	*half*	74.7	91.8	34.4
gray-scale	-	**75.5**	**91.9**	**34.1**
time diff	-	74.5	91.6	34.2
optical flow	-	73.8	91.3	35.1

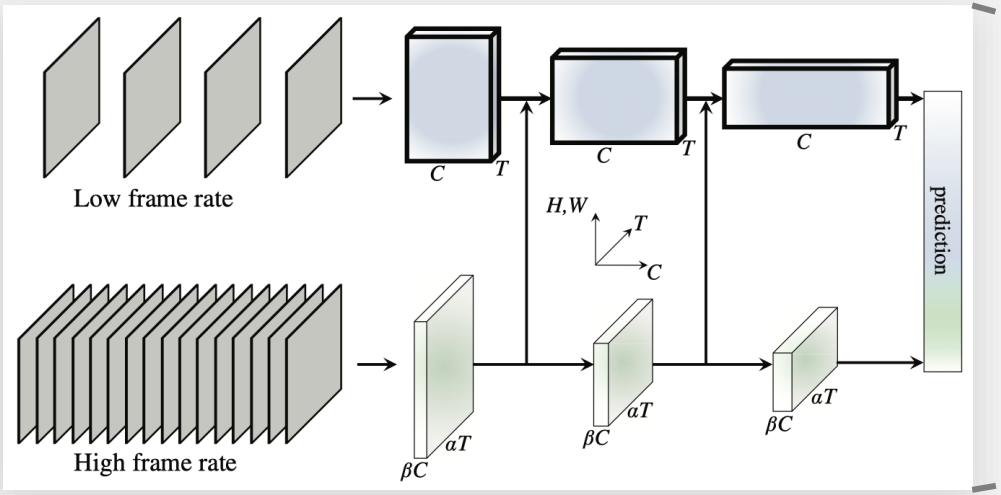


Weaker spatial input to Fast pathway:
Alternative ways of weakening spatial inputs.

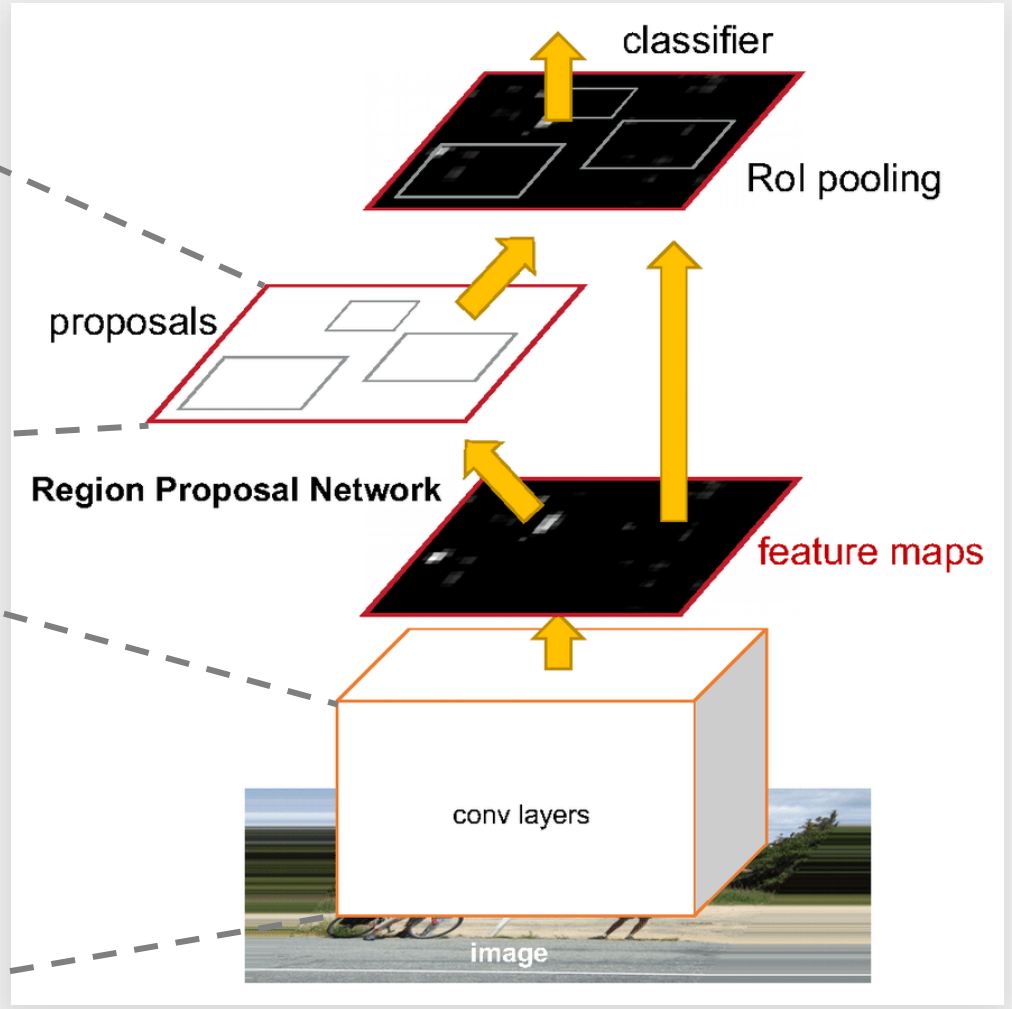
Action Detection



Person Detector



SlowFast Network



Faster R-CNN

Experiments

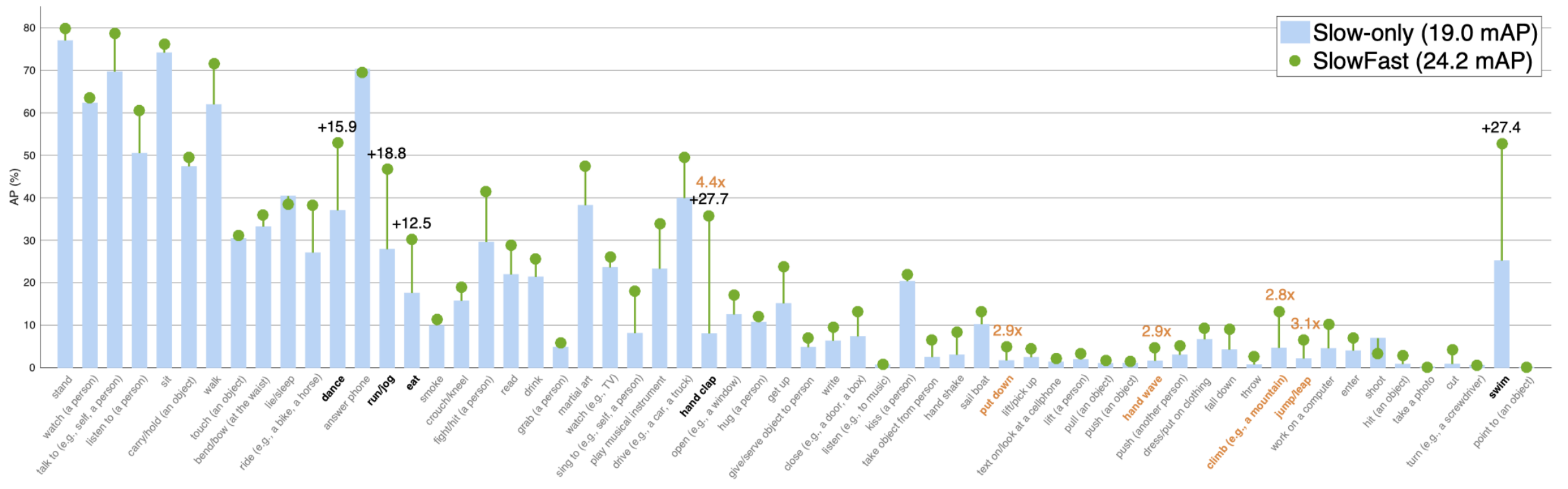
Action Detection

model	flow	video pretrain	val mAP	test mAP
I3D [20]		Kinetics-400	14.5	-
I3D [20]	✓	Kinetics-400	15.6	-
ACRN, S3D [46]	✓	Kinetics-400	17.4	-
ATR, R50+NL [29]		Kinetics-400	20.0	-
ATR, R50+NL [29]	✓	Kinetics-400	21.7	-
9-model ensemble [29]	✓	Kinetics-400	25.6	21.1
I3D [16]		Kinetics-600	21.9	21.0
SlowFast		Kinetics-400	26.3	-
SlowFast		Kinetics-600	26.8	-
SlowFast, +NL		Kinetics-600	27.3	**27.1**
SlowFast*, +NL		Kinetics-600	**28.2**	-

Comparison with the SOTA on AVA 2.1.

Experiments

Action Detection



Per-category AP on AVA:
a Slow-only baseline (19.0 mAP) vs. its SlowFast counterpart (24.2 mAP).

Conclusion

- The time axis is a special dimension.

- This paper has investigated an architecture design that contrasts the speed along this axis.

- It achieves state-of-the-art accuracy for video action classification and detection.

SlowFast Networks for Video Recognition

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He
Facebook AI Research (FAIR)
ICCV 2019

review by Jihun Kim