Towards a New Level of Action Understanding

Dahua Lin
The Chinese University of Hong Kong

Action Recognition

Classify human actions in short videos

Cutting watermelon

Presenting weather forecast

Climbing a rope

Image Source : https://github.com/open-mmlab/mmaction2

Temporal Action Localization

Temporally localize and classify human actions in long videos.

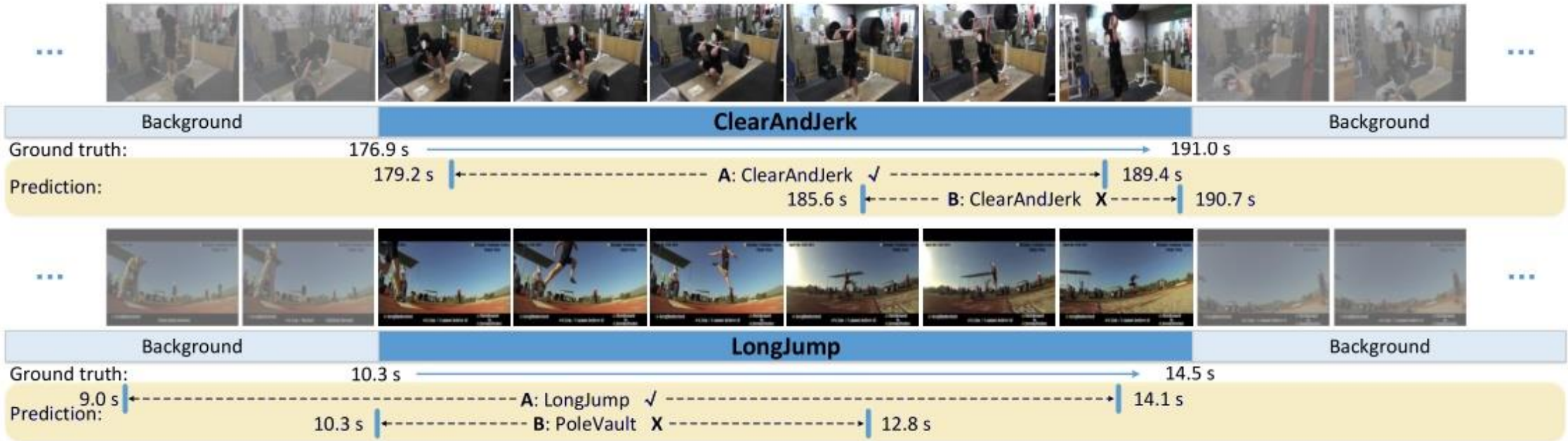


Image Source : https://www.ee.columbia.edu/ln/dvmm/researchProjects/cdc/scnn.html

Spatio-Temporal Action Detection

Spatially detection plus temporally localization of human actions in long videos.

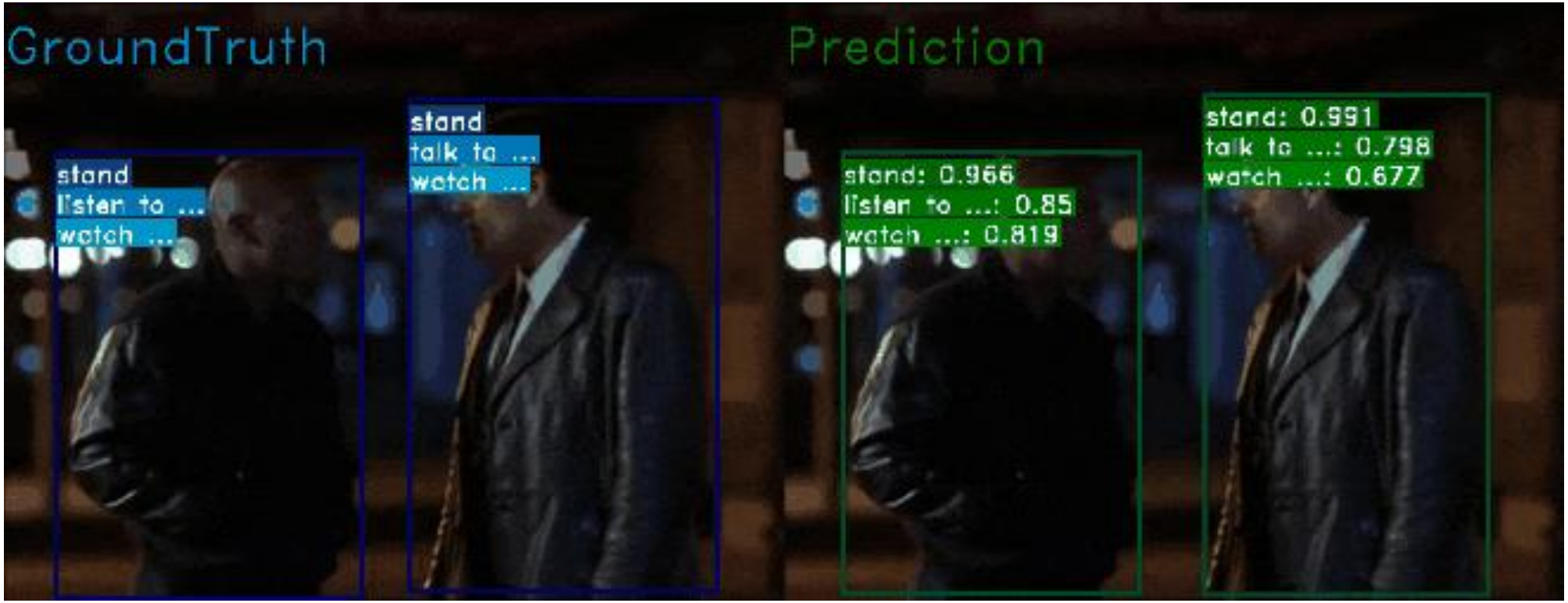


Image Source : https://github.com/open-mmlab/mmaction2

Why Action Understanding is Difficult

Understanding action require analysis of both **appearance** and **motion**.

=

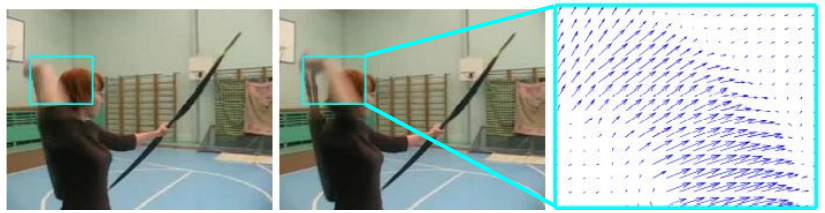
Frames along time

...

Actions in video

Appearance

Motion



Optical flow as description of inter-frames motion

Evolution of Techniques

2014

Two Stream ConvNets

Breakthroughs in Action Understanding

2017

3D ConvNets

Boost performance with the help of large-scaled

2020

Omni-source

Make use of large-scaled web data

FineGym

Fine-grained video dataset

2021

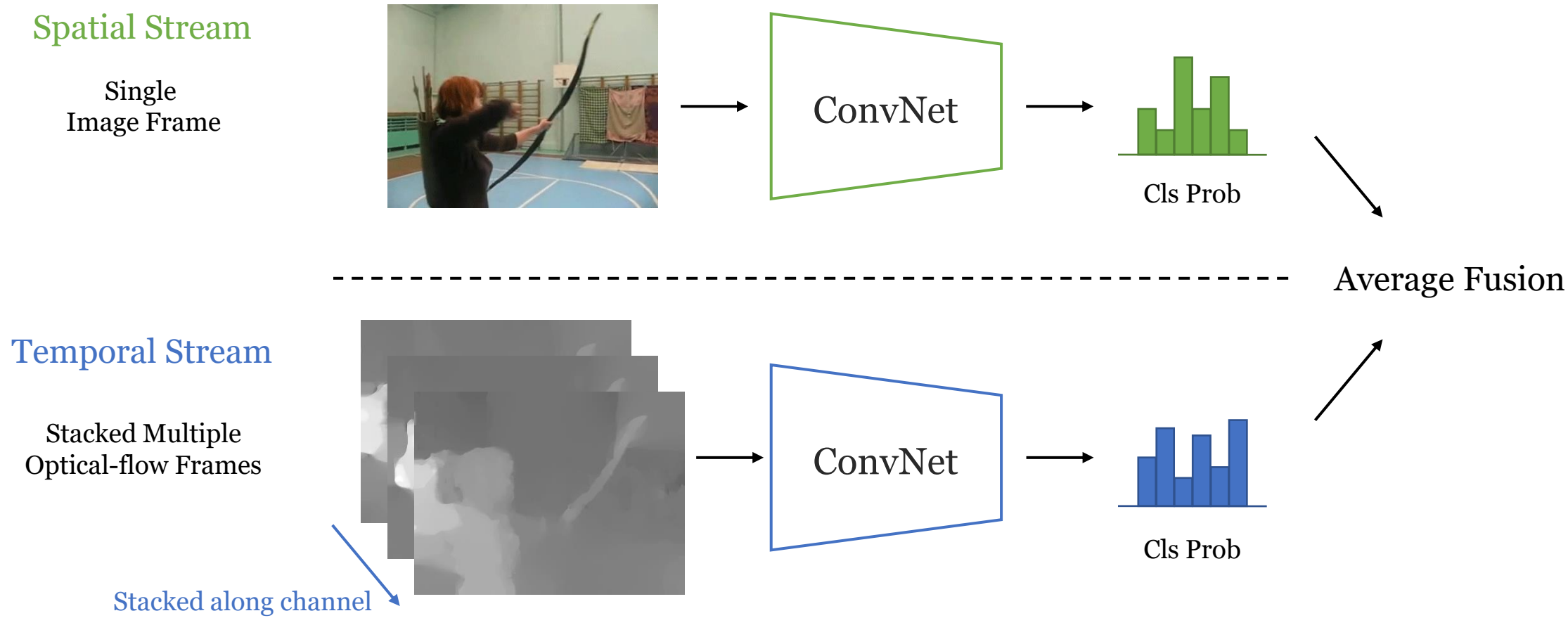
Transformer Based Models

New Trends

Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Two Stream Convolutional Networks

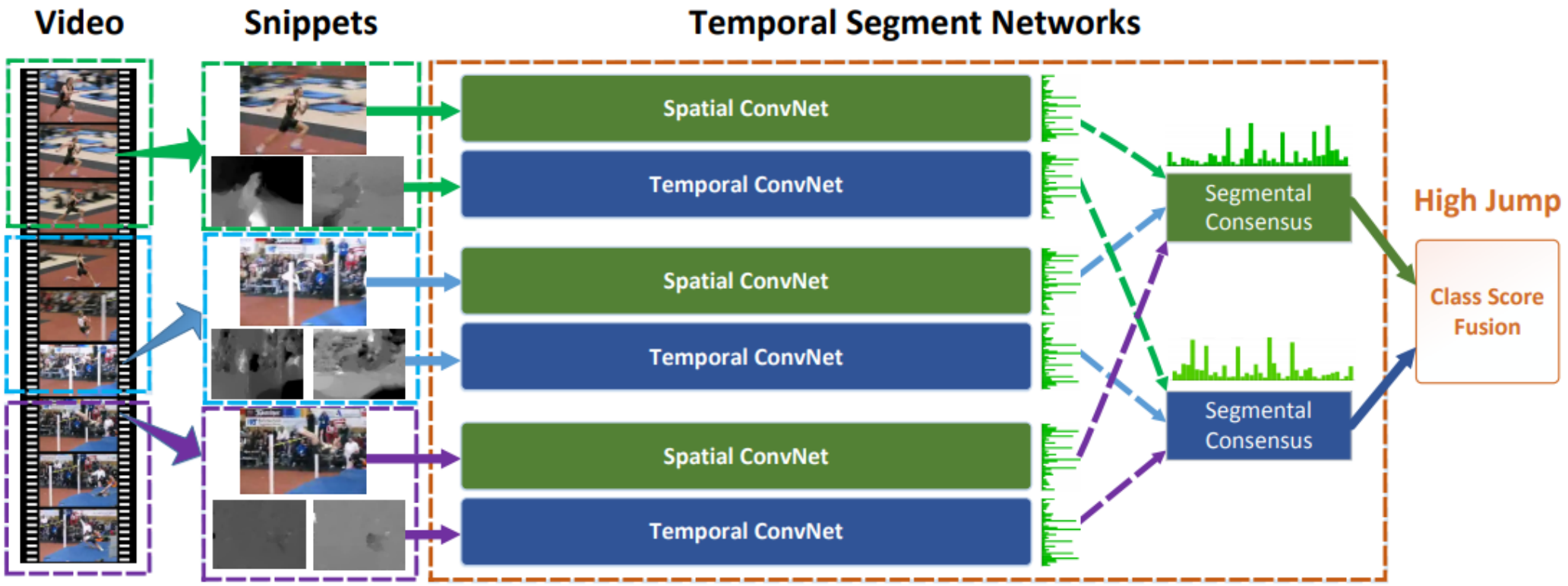
Utilize both appearance and motion information to predict actions in video



Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." European conference on computer vision. Springer, Cham, 2016.

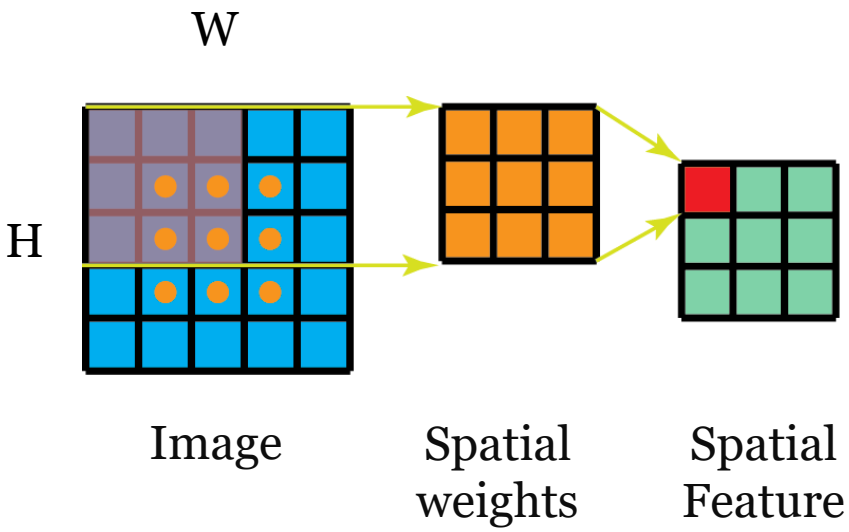
Temporal Segment Network

Long-range modeling using multiple video segments



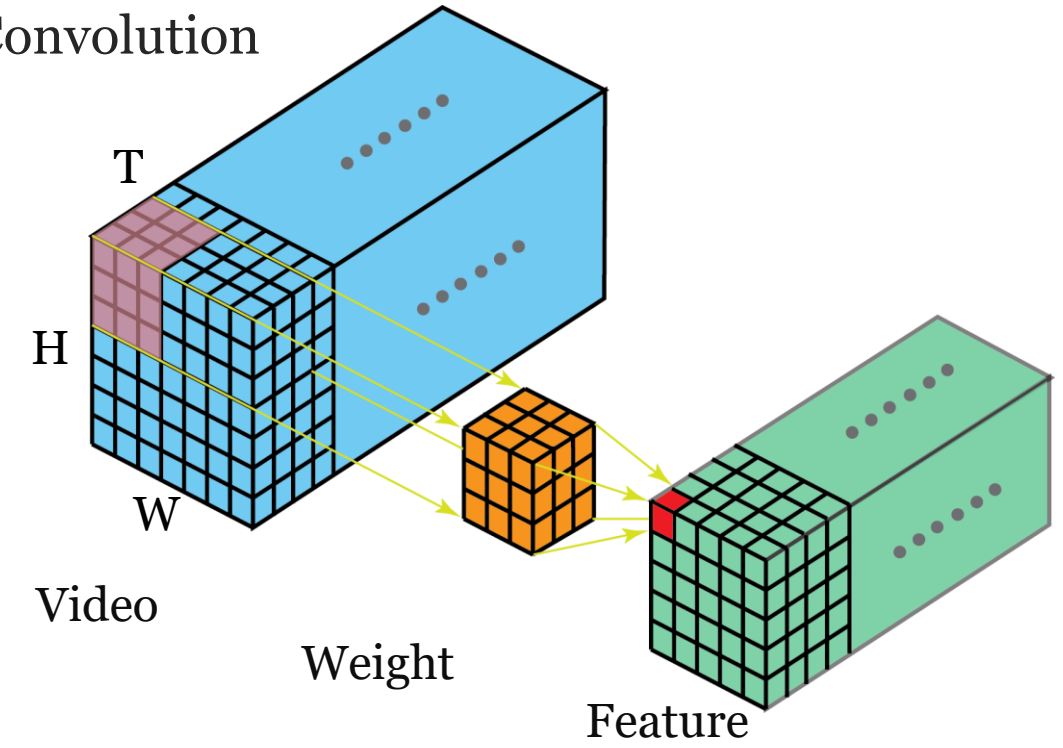
From 2D to 3D Convolution

2D Convolution



Spatial feature from spatially neighboring pixels

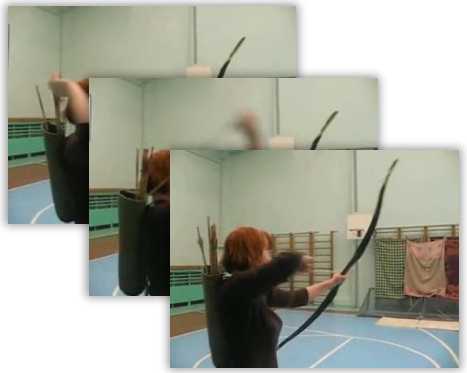
3D Convolution



Spatiotemporal feature from both spatially and temporally neighboring pixels

Image Sources: https://towardsdatascience.com/a-comprehensive-introduction-to-different-types-of-convolutions-in-deep-learning-669281e58215

From Two Stream to 3D ConvNets



Contiguous image frames

Estimate Optical Flow

Stack along channel

2D ConvNet

Motion Feature

Direct forward to network

3D ConvNet

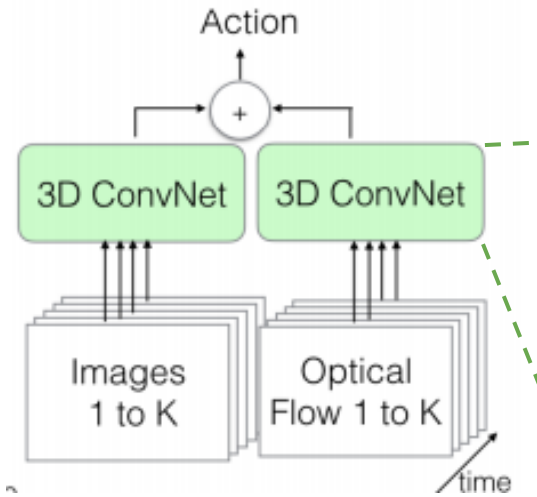
Appearance & Motion Feature

C3D, I3D, R(2+1)D, CSN, SlowFast, ...

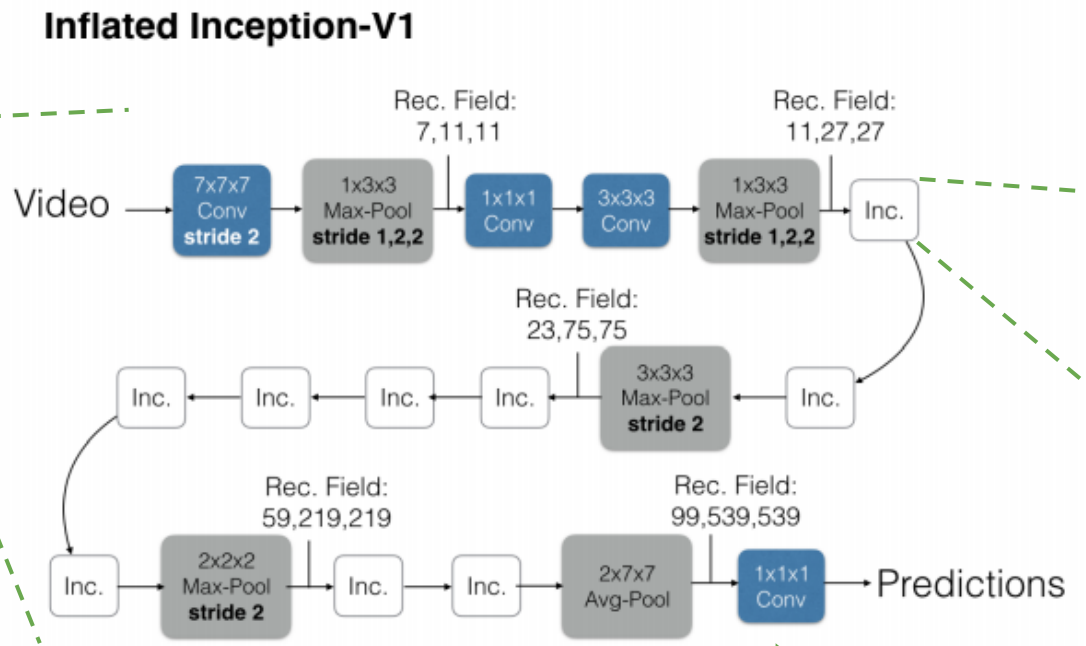
Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

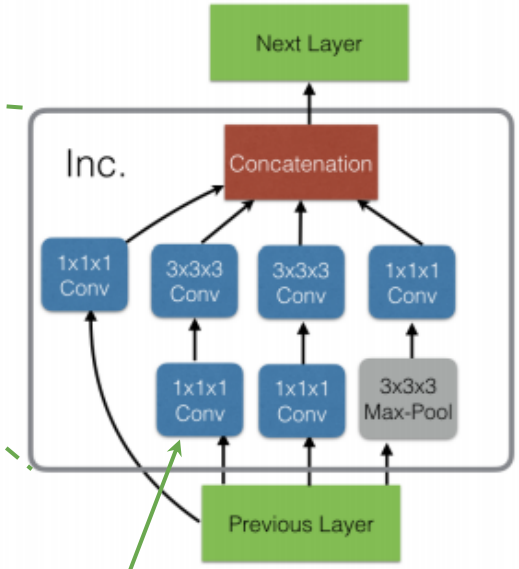
I3D (Inflated 3D ConvNets)

Inflate 2D networks to 3D networks



Two-Stream 3D-ConvNet



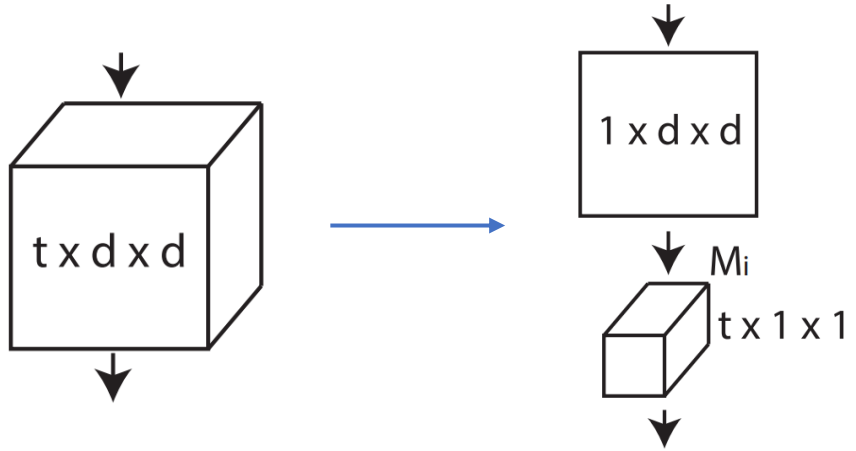


3D convolutional kernels are **Inflated** from succesfully trained image classification networks

A Closer Look at Spatiotemporal Convolutions for Action Recognition
S3D: Single Shot multi-Span Detector via Fully 3D Convolutional Networks
Video Classification with Channel-Separated Convolutional Networks

Parameter-Efficient 3D Networks

Decompose 3D kernels



2D Spatial kernel

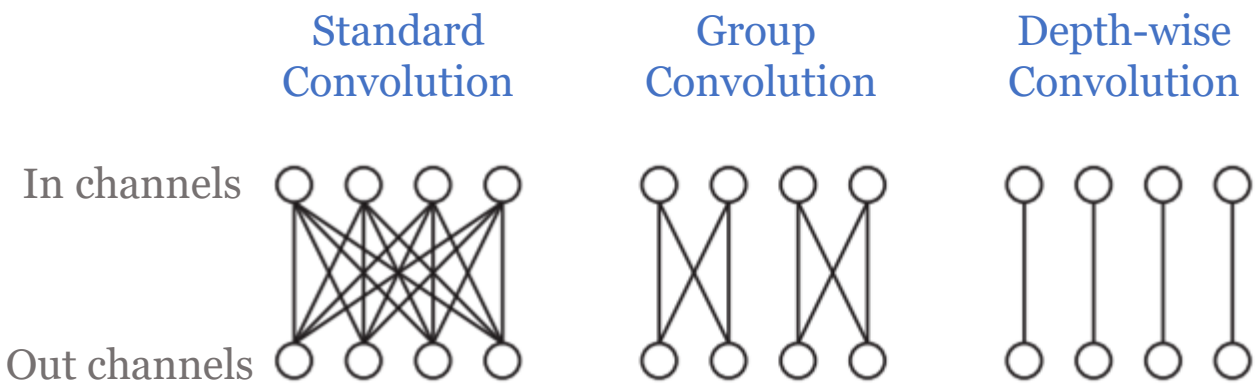
1D Temporal kernel

R(2+1)D

S3D

CSN

Sparse channel connection



◯ = feature maps of single channel

—— = weight parameters

Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

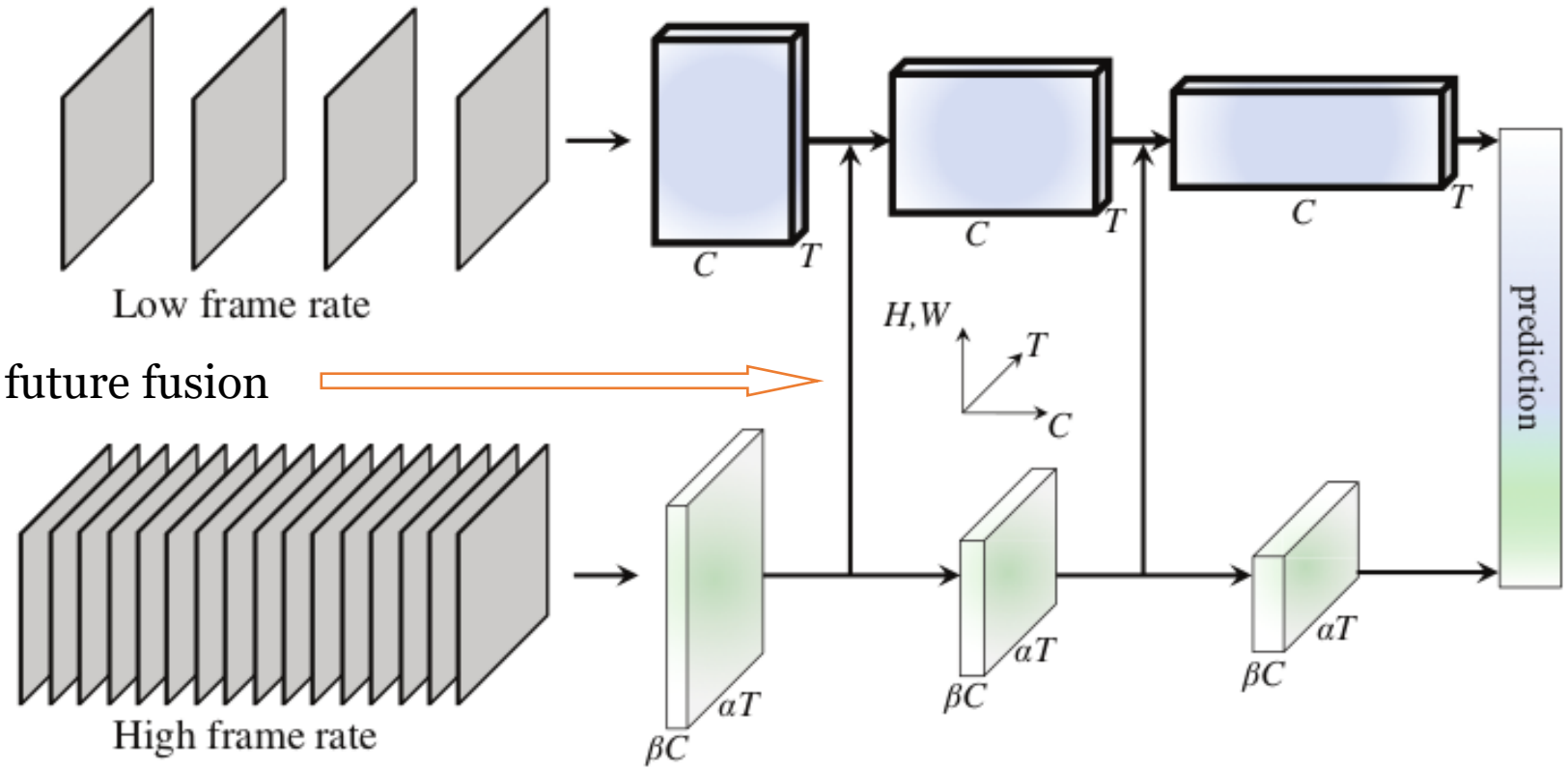
SlowFast Networks

Appearance and motion evolves at different rate, thus can be handled efficiently in different manner.

Standard branch at low frame rate for appearance

Lateral connection for future fusion

Lightweight branch at high frame for motion



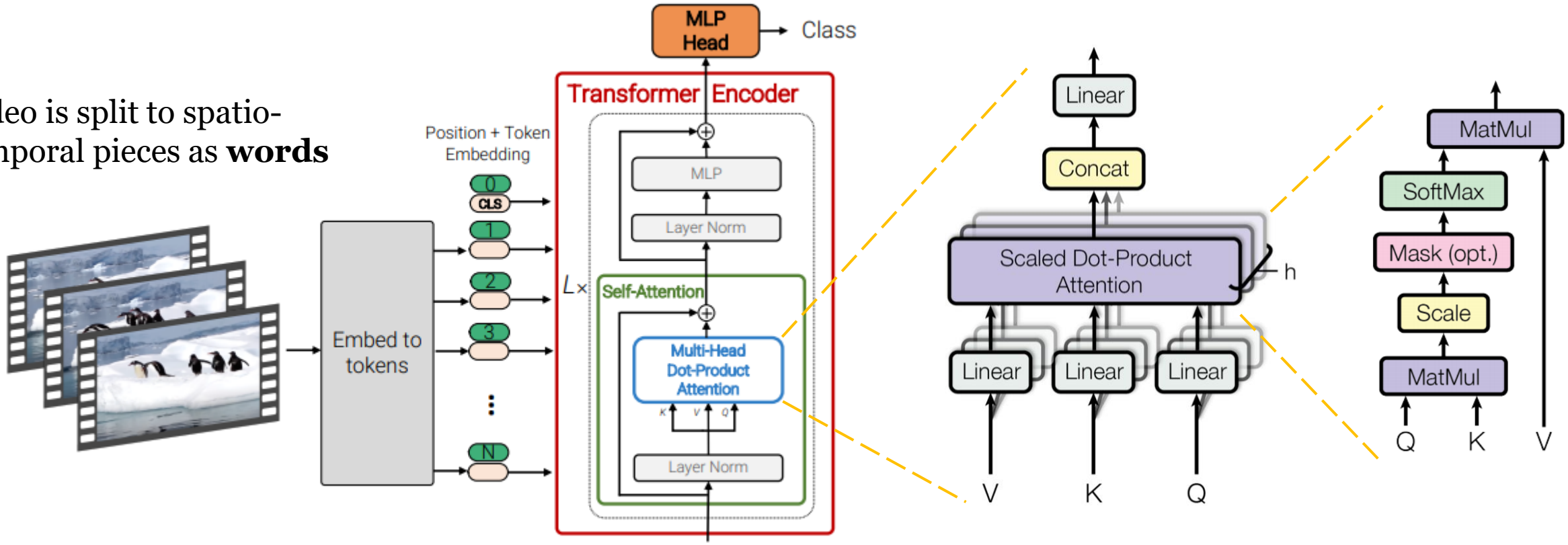
e.g., α=8, β=1/8

Arnab, Anurag, et al. "Vivit: A video vision transformer." arXiv preprint arXiv:2103.15691 (2021).
Vaswani, Ashish, et al. "Attention is all you need." arXiv preprint arXiv:1706.03762 (2017).

Video Transformer

Transformer a successful model in NLP and start to emerge in computer vision, including video understanding.

Video is split to spatio-temporal pieces as **words**



https://github.com/open-mmlab/mmaction2

Unified Framework for Video Understanding

Single Framework

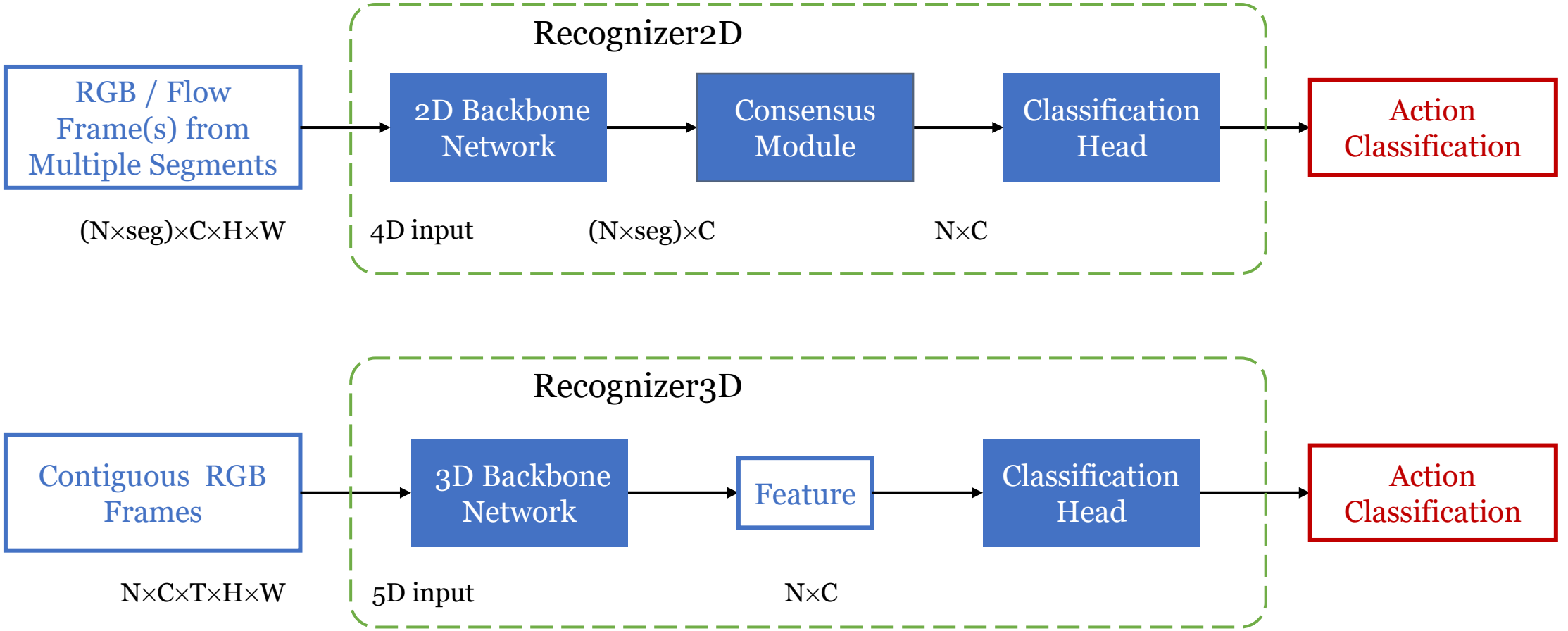
Multiple Tasks

- Action Recognition
- Temporal Action Localization
- Spatio-Temporal Action Detection

Various Models

- Two Stream based Models
- 3D Convnet based Models
- Transformer based Models

Unified Model Design



Unified Model Design

Modular design allows for unified model configuration

Example Config of TSN

```
model = dict(
    type='Recognizer2D',
    backbone=dict(
        type='ResNet',
        pretrained='torchvision://resnet50',
        depth=50,
        norm_eval=False),
    cls_head=dict(
        type='TSNHead',
        num_classes=400,
        in_channels=2048,
        spatial_type='avg',
        consensus=dict(type='AvgConsensus', dim=1),
        dropout_ratio=0.4,
        init_std=0.01),
```

Backbone

Head

Example config of I3D

```
model = dict(
    type='Recognizer3D',
    backbone=dict(
        type='ResNet3d',
        pretrained2d=True,
        pretrained='torchvision://resnet50',
        depth=50,
        # omit some parameters),
    cls_head=dict(
        type='I3DHead',
        num_classes=400,
        in_channels=2048,
        spatial_type='avg',
        dropout_ratio=0.5,
        init_std=0.01),
```

Unified Frame Loader

Single frame sampler for multiple type of frames required by different models.

Independent RGB frames from multiple segments (e.g. for training TSN)

```
dict(type='SampleFrames', clip_len=1, frame_interval=1, num_clips=3),
```

Contiguous Flow frames from multiple segments (e.g. for training TSN)

```
dict(type='SampleFrames', clip_len=5, frame_interval=1, num_clips=3),
```

Contiguous frames (e.g. for training 3D networks)

```
dict(type='SampleFrames', clip_len=32, frame_interval=2, num_clips=1),
```

FineGym

Towards Finer-grained Action Recognition

Semantic Granularities

Different classes FROM UCF101

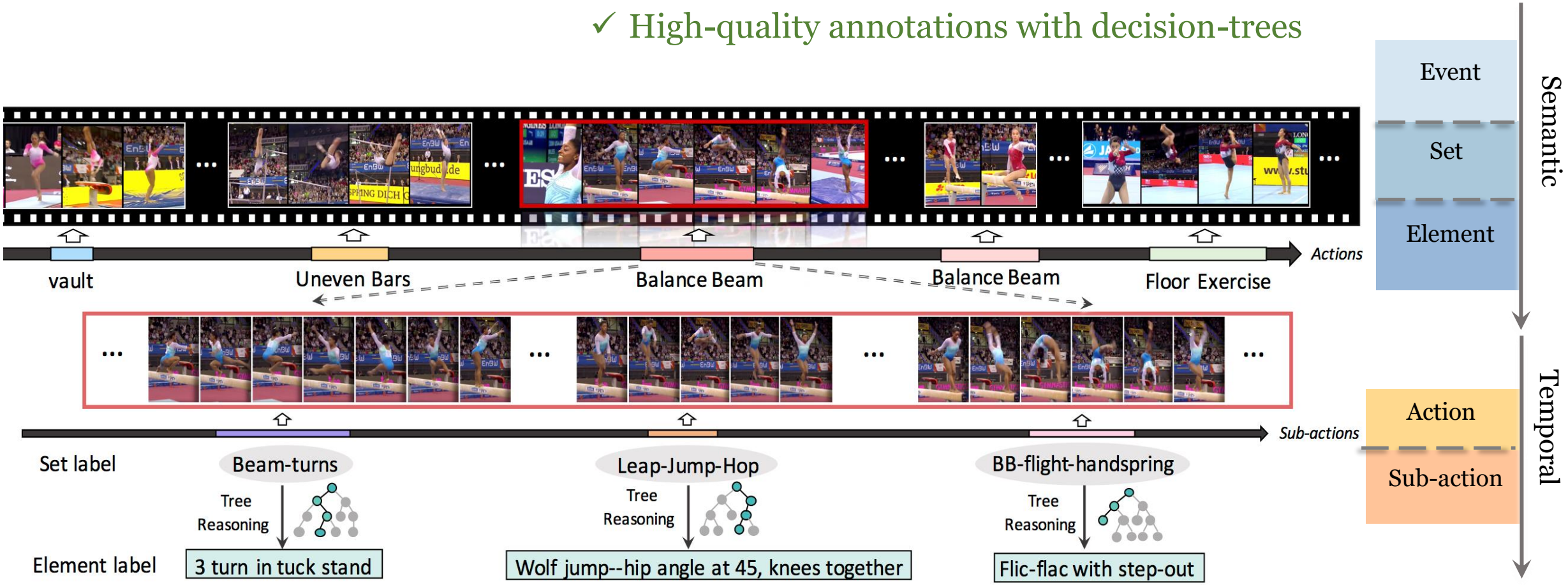
Different classes FROM FineGym

FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding.
D. Shao, Y. Zhao, B. Dai, D. Lin. CVPR 2020.

FineGym

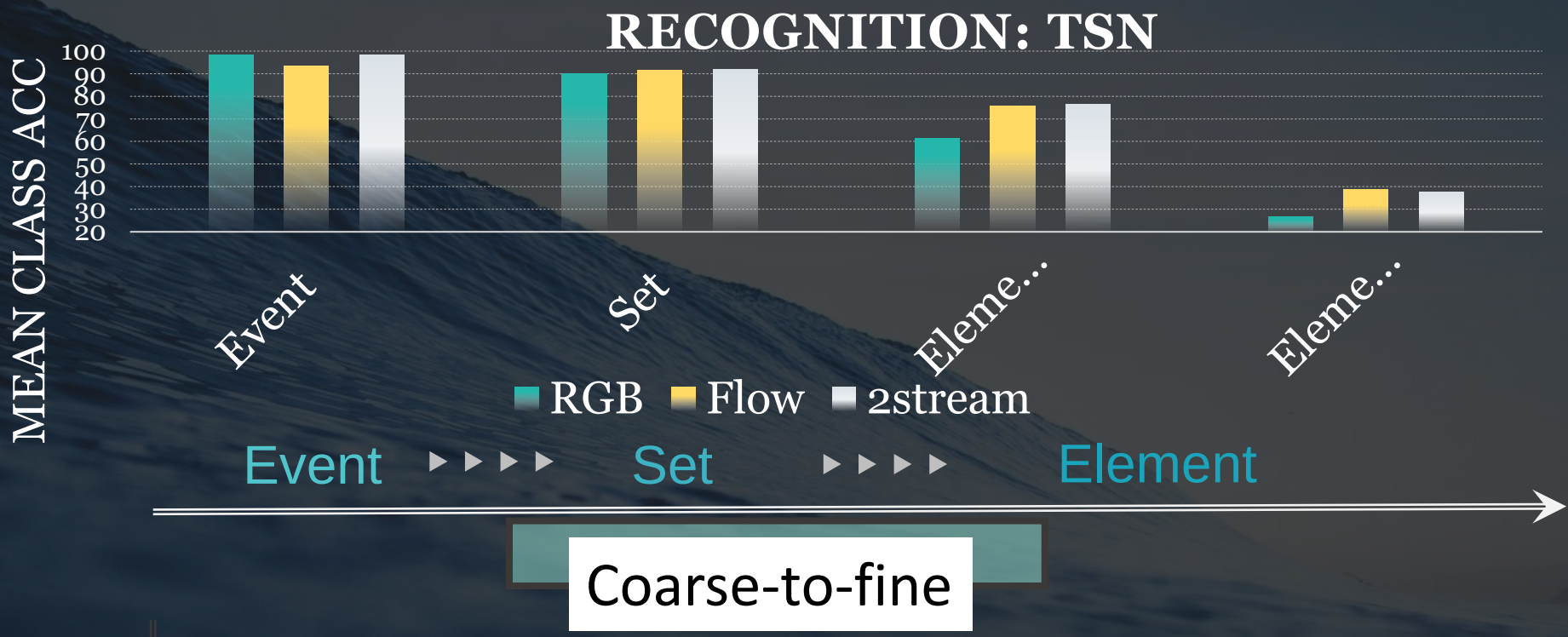
530 well-defined categories

- ✓ Rich semantic & temporal structures
- ✓ Action-centric
- ✓ High-quality annotations with decision-trees



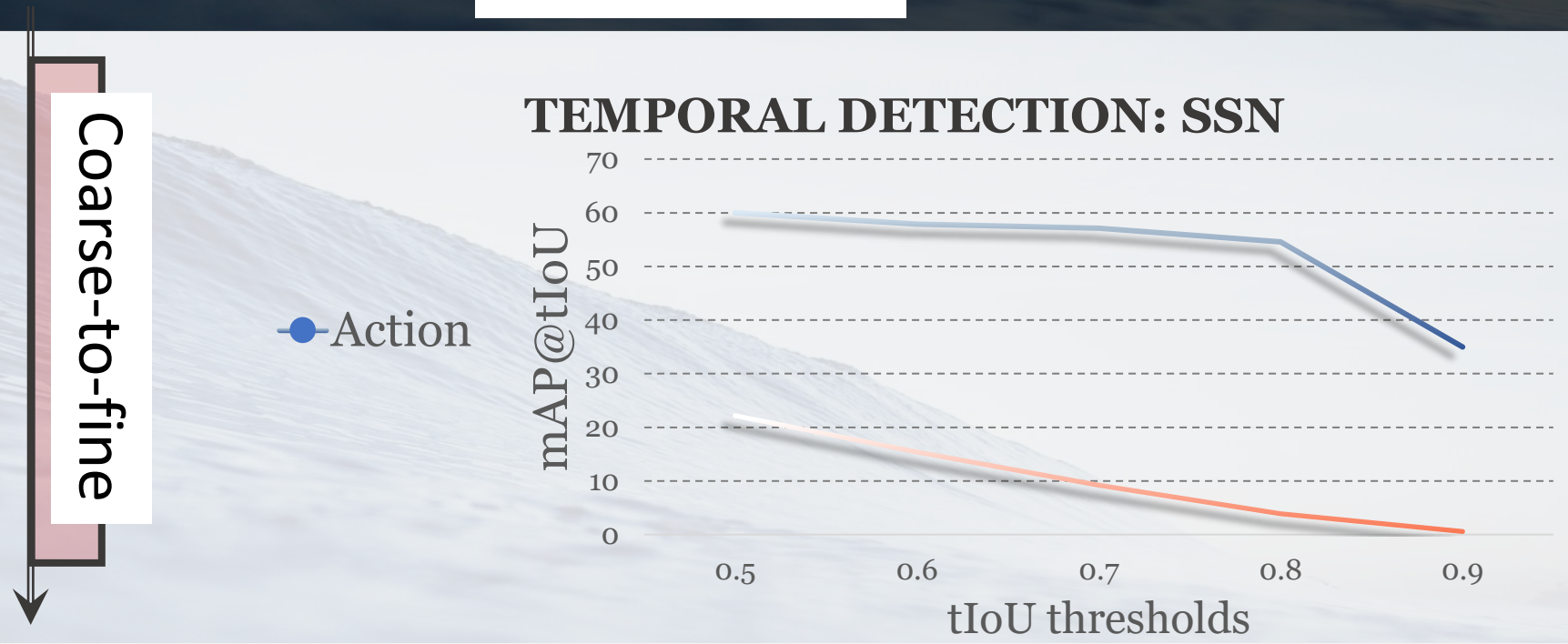
S

Semantic



T

Temporal

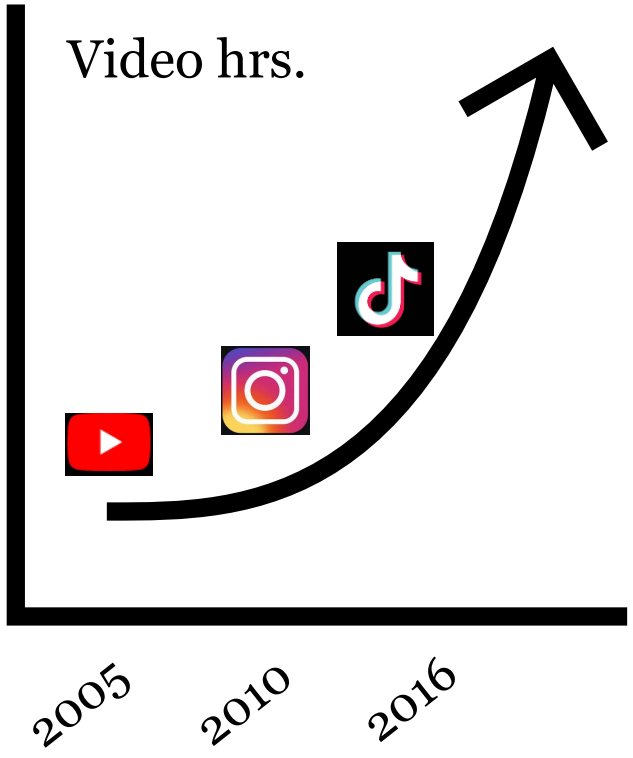


Omni-Source

Towards larger-scaled dataset with less cost

Motivation

Unlabeled Dataset



> 500 Hours

Generated on Youtube **per minute**

65M Videos

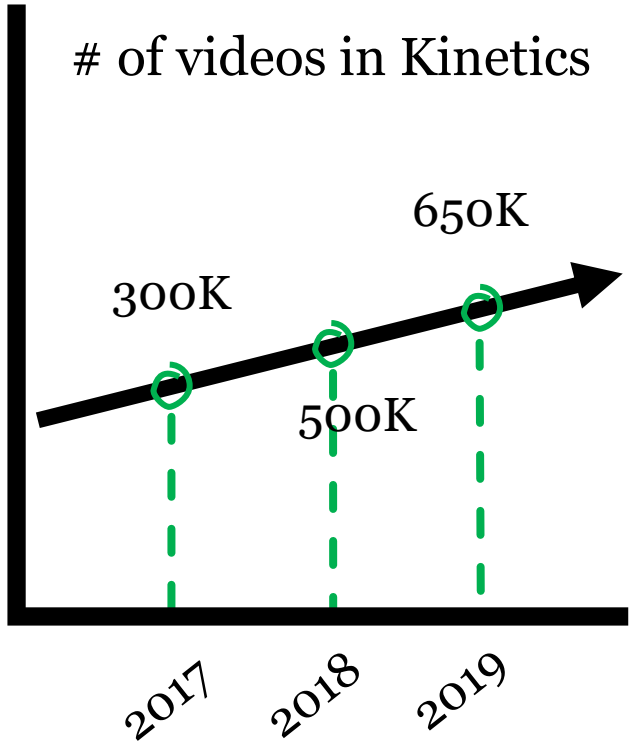
Weakly supervised video recognition dataset collected from Instagram

v.s.

< 2000 Hours

Kinetics-700: one of the largest video recognition dataset

Labeled Dataset



Motivation


Unlabeled Dataset
Video hrs.
2005
2010
2016
Omni Source
Target Dataset
Labeled Dataset
# of videos in Kinetics
300K
500K
650K
2017
2018
2019

OmniSource Pipeline

1. Task-specific Data Collection

Keywords

Crawl with

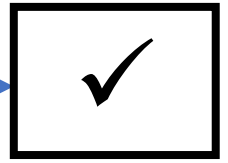
Multiple Sources

From

2. Teacher Filtering

Teacher Net

keep



Abseiling?

3. Transformation

Image

Clip

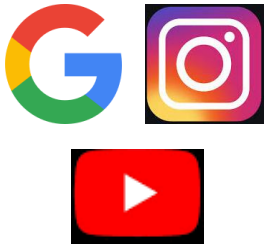
Untrimmed Video

Clip

4. Joint Training

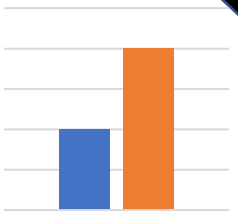
Target Dataset

+



Web Dataset

=

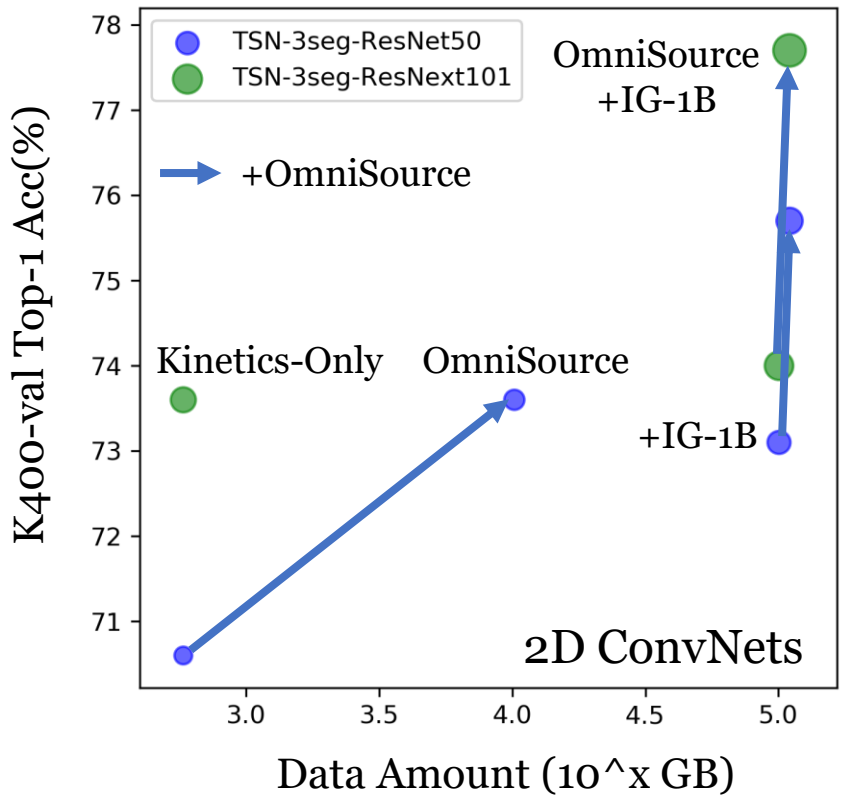


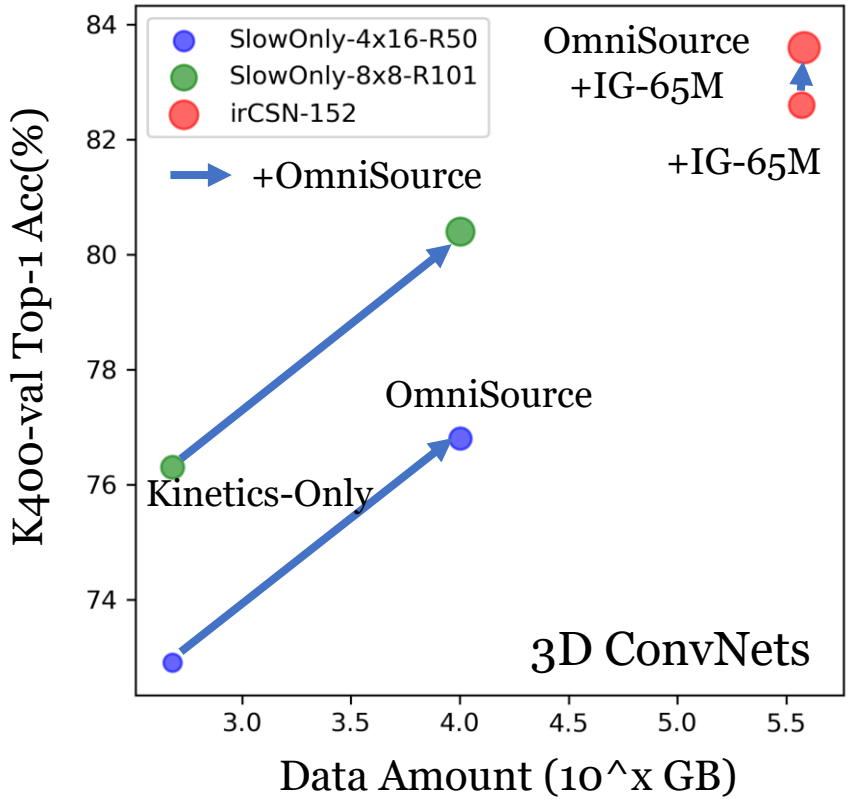
Performance

w/o. Omni

w. Omni

Data Efficiency





Note: Since the resolution of images and videos might be different. For fair comparison, we assume that each image is 100KB and one minute video is 12MB. We assume IG videos last 30 seconds on average.

Improvement on Kinetics & Downstream Tasks

v.s.

Arch	Backbone	Pretrain	w/o. Omni	w/. Omni	Δ
TSN-3seg	ResNet50	ImageNet	70.6 / 89.4	73.6 / 91.0	+3.0 / +1.6
TSN-3seg	ResNet50	IG-1B	73.1 / 90.4	75.7 / 91.9	+2.6 / +1.5
TSN-3seg	Efficient-b4	ImageNet	73.3 / 91.0	75.2 / 92.0	+1.9 / +1.0
SlowOnly-4x16	ResNet50	-	72.9 / 90.9	76.8 / 92.5	+3.9 / +1.6
SlowOnly-4x16	ResNet50	ImageNet	73.8 / 90.9	76.6 / 92.5	+2.8 / +1.6
SlowOnly-8x8	ResNet101	-	76.3 / 92.6	80.4 / 94.4	+4.1 / +1.8
SlowOnly-8x8	ResNet101	ImageNet	76.8 / 92.8	80.5 / 94.4	+3.7 / +1.6
irCSN-32x2	irCSN-152	IG-65M	82.6 / 95.3	83.6 / 96.0	+1.0 / +0.7

Table 1. Recognition performance improvement on Kinetics400.

Table 2. Detailed results of transfer learning.

Top-1 accuracies on the official split-1 are reported.

Architecture	w/. ImageNet-pretrain	w/. OmniSource	UCF101-Top1	HMDB51-Top1	
TSN-3seg ResNet50	✓		91.51	63.53	
	✓	✓	93.29	65.88	v.s.
TSN-3seg Efficient-b4	✓		92.52	66.27	
	✓	✓	93.05	66.54	v.s.
SlowOnly-4x16 ResNet50	✓		94.69	69.35	
	✓	✓	95.98	70.71	v.s.
			94.05	65.82	
		✓	96.01	70.98	v.s.
SlowOnly-8x8 ResNet101	✓		96.40	76.41	
	✓	✓	97.38	78.95	v.s.
			96.61	75.82	
		✓	97.52	79.02	v.s.

Summary

- Action recognition models evolves from two stream, 3D convolution, to the use of transformers
- Modular design allow unified model composition in MMAction2
- Carefully designed datasets and effective use of large scale web data help improving action recognition models