CVPR SEATTLE WASHINGTON JUNE 16-18 2020

https://cvpr20-video.mxnet.io/

Tutorial on Video Modeling

A Chronological Review of Recent SoTA and Beyond

Yi Zhu

06/14/2020

aws

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

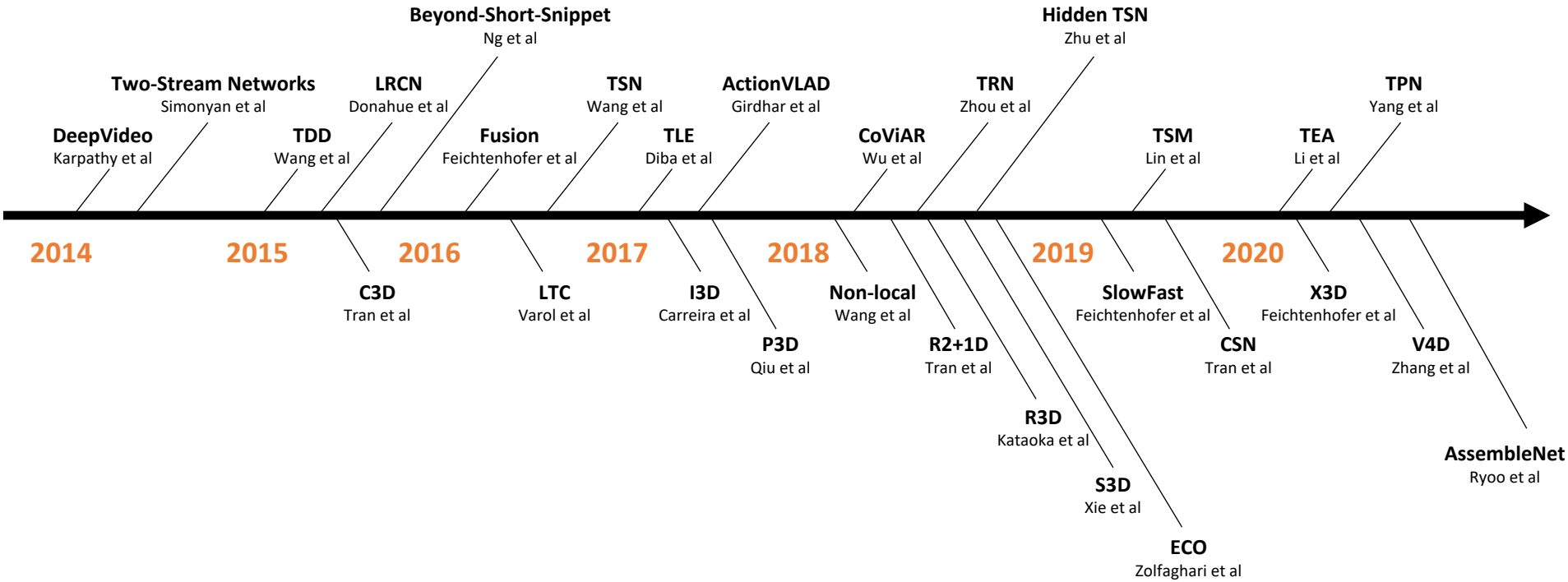
- **Chronological review**
  - Single-stream network
  - Two-stream networks
  - 3D CNNs
  - Other motion modeling methods
- **GluonCV video toolkit**
  - Comprehensive and reproducible model zoo
  - Flexible and customized usage
  - Detailed documentation and tutorials

aws

A Chronological Review of Recent SoTA

aws

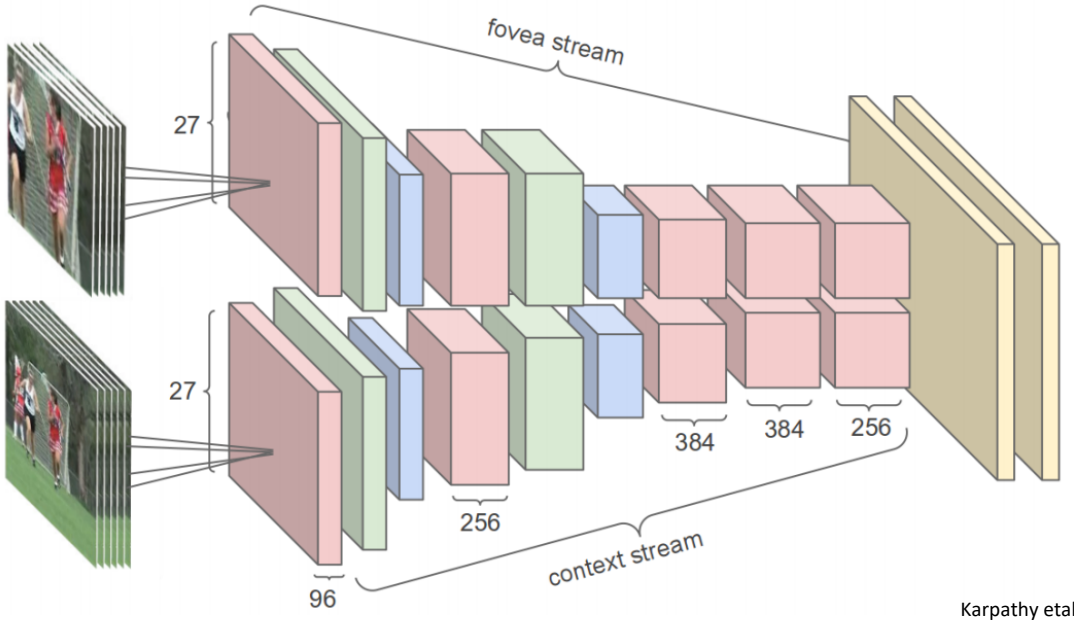
CVPR SEATTLE WASHINGTON JUNE 16-18 2020



aws

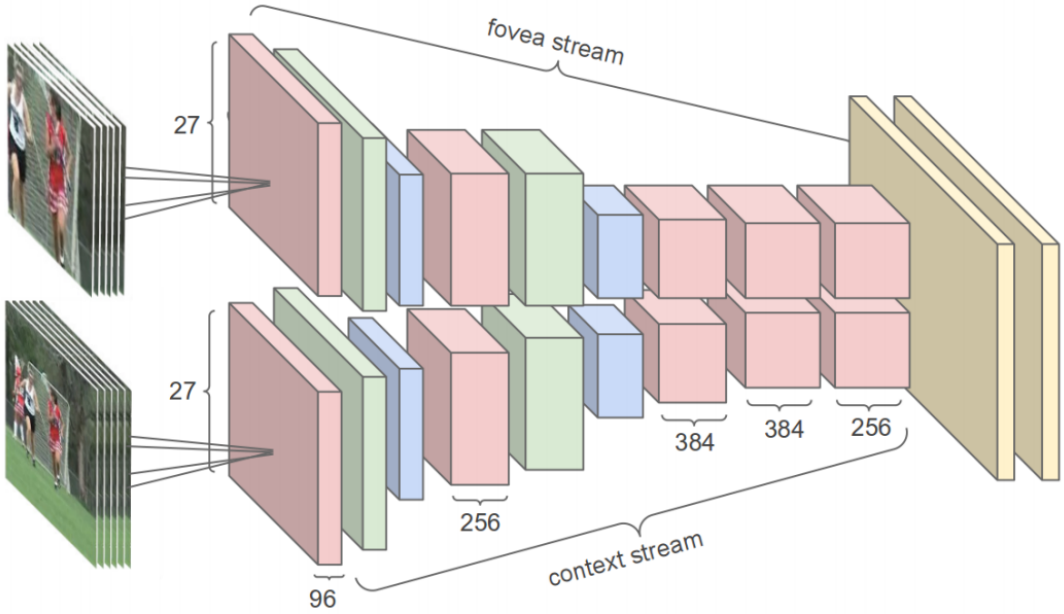
Single-Stream Network

Single Stream Network



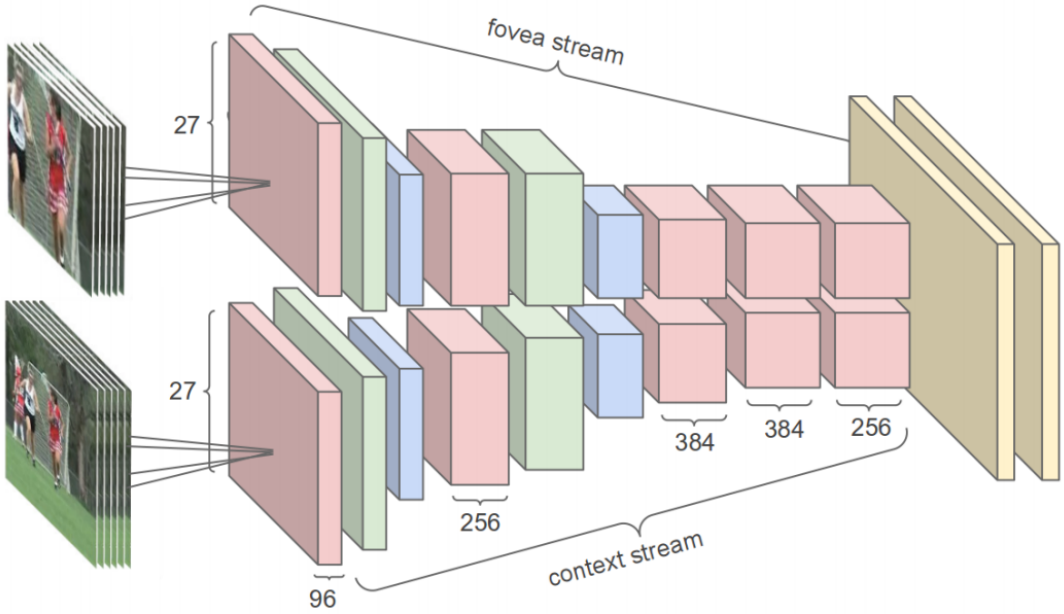
Karpathy etal, Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

Single Stream Network



	UCF101
IDT	87.9%
DeepVideo	65.4%

Single Stream Network

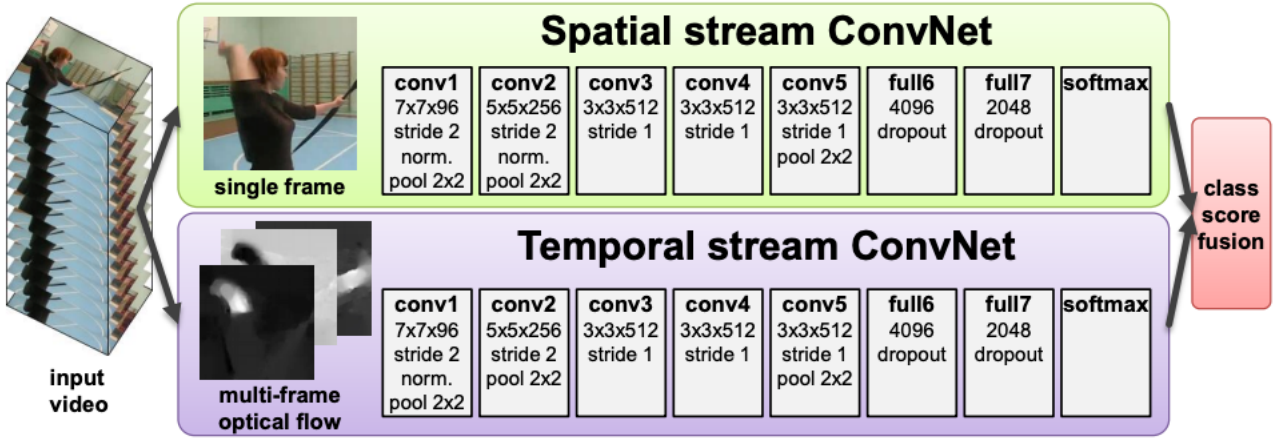


	UCF101
IDT	87.9%
DeepVideo	65.4%

Lack of motion modeling

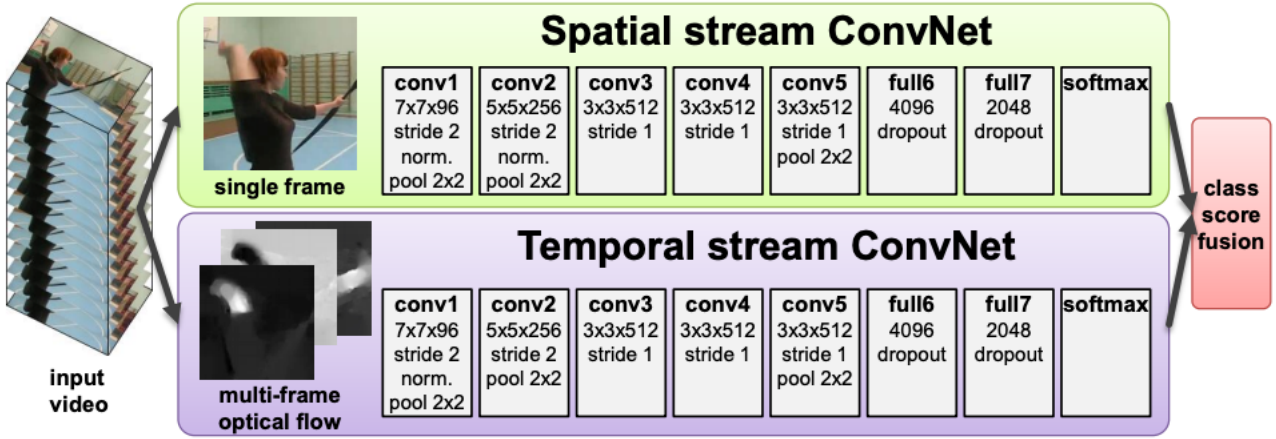
Two-Stream Networks

Two-Stream Networks



Simonyan etal, Two-Stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

Two-Stream Networks

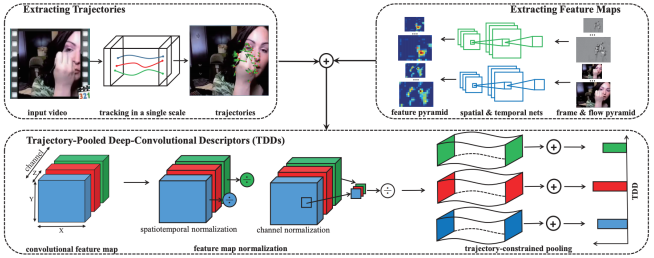


	UCF101
IDT	87.9%
DeepVideo	65.4%
Two-stream	88.0%

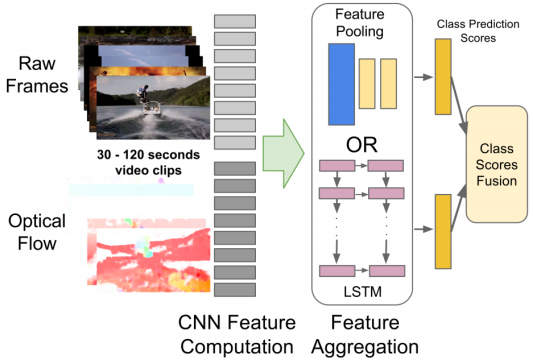
Simonyan etal, Two-Stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

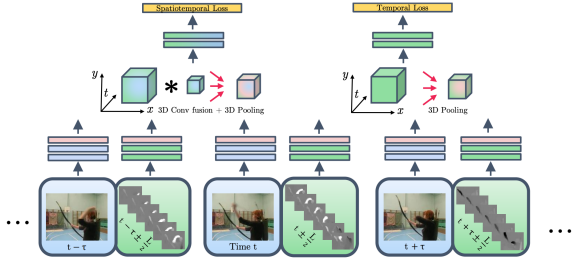
Two-Stream Networks Follow-up



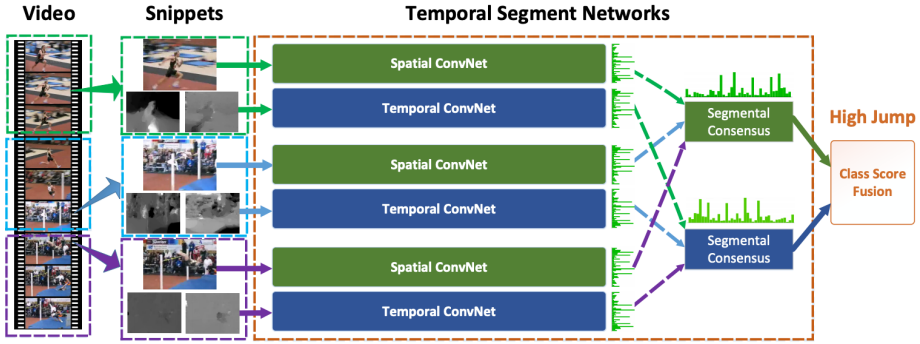
TDD, CVPR 15



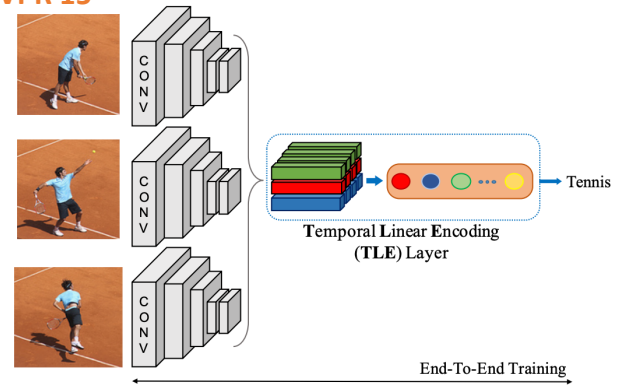
Beyond-Short-Snippet, CVPR 15



Two-stream Fusion, CVPR 16



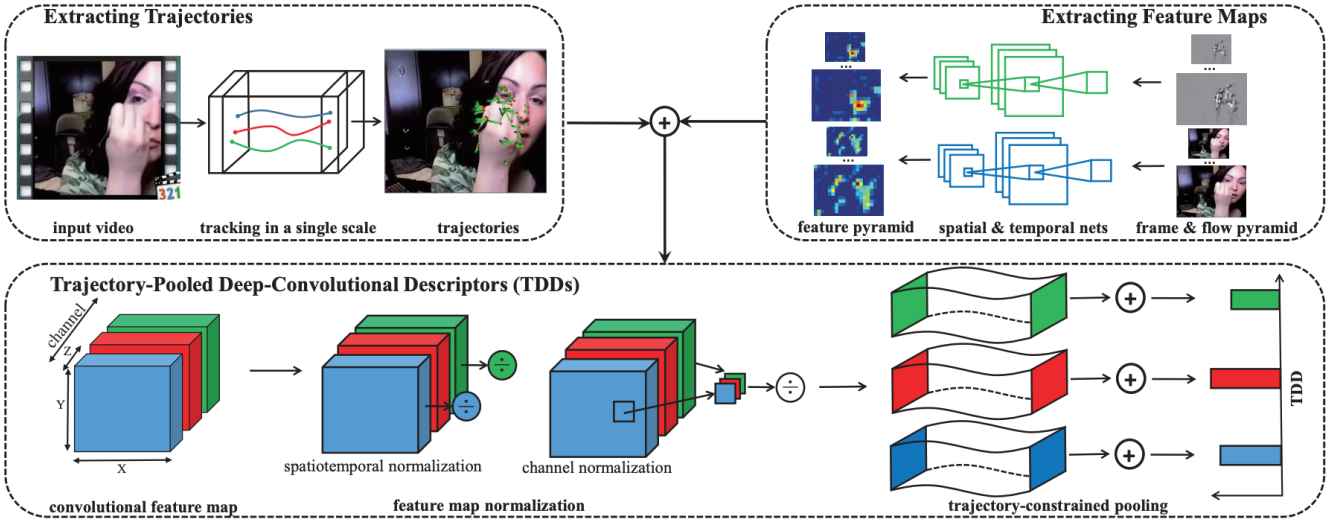
TSN, ECCV 16



TLE, CVPR 17

aws

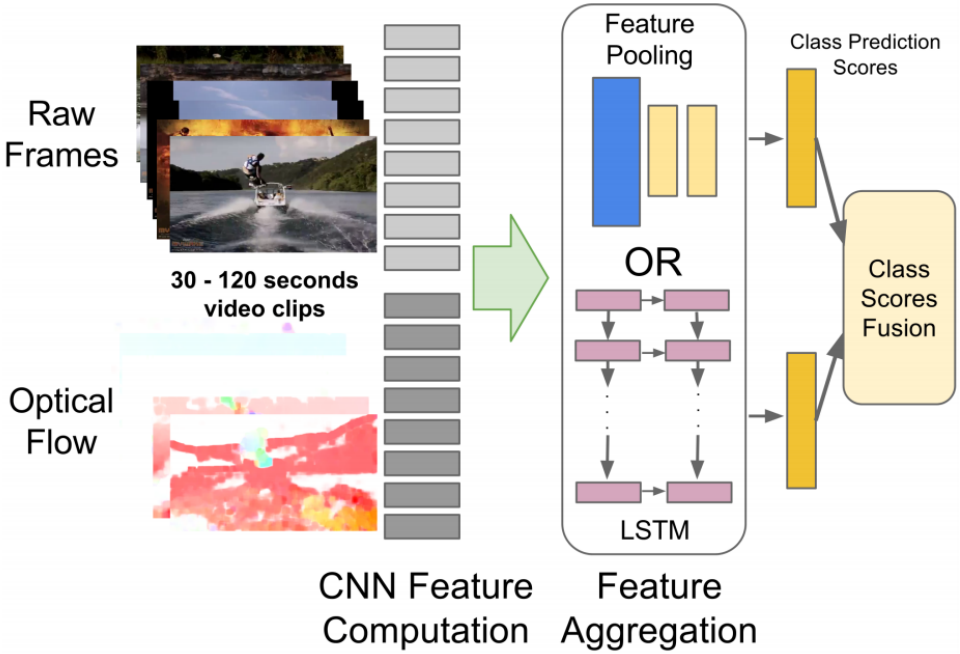
Two-Stream Networks Follow-up



	UCF101
IDT	87.9%
DeepVideo	65.4%
Two-stream	88.0%
TDD	91.5%

Wang etal, Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors, CVPR 2015

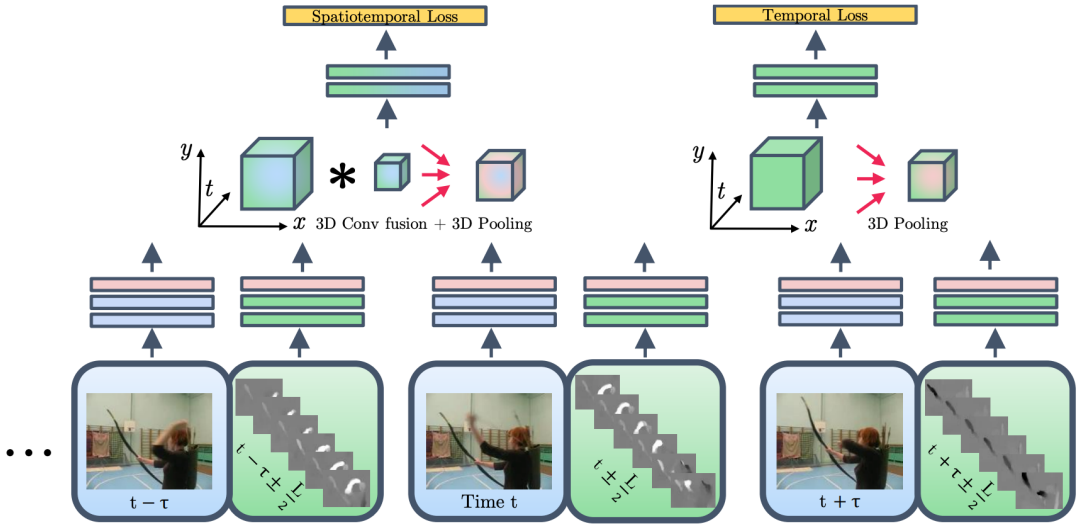
Two-Stream Networks Follow-up



	UCF101
IDT	87.9%
DeepVideo	65.4%
Two-stream	88.0%
TDD	91.5%
Beyond-Short-Snippets	88.6%

Ng etal, Beyond Short Snippets: Deep Networks for Video Classification, CVPR 2015

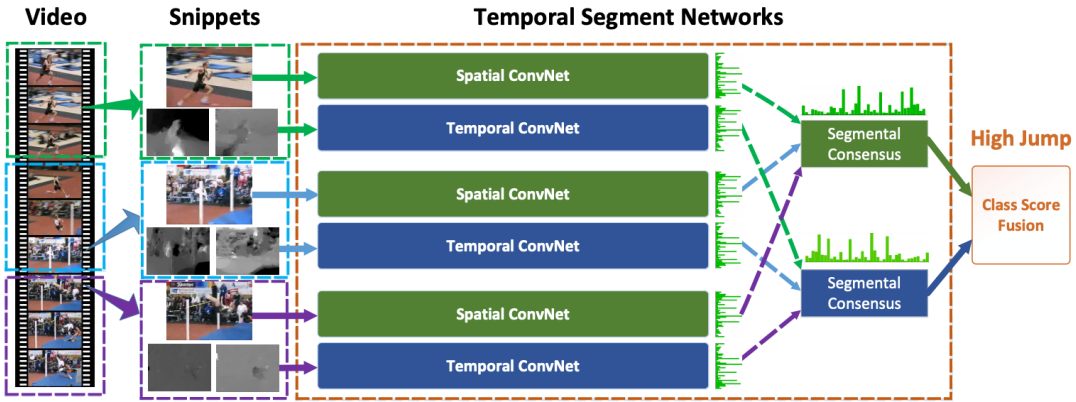
Two-Stream Networks Follow-up



	UCF101
IDT	87.9%
DeepVideo	65.4%
Two-stream	88.0%
TDD	91.5%
Beyond-Short-Snippets	88.6%
Two-Stream Fusion	92.5%

Feichtenhofer etal, Convolutional Two-Stream Network Fusion for Video Action Recognition, CVPR 2016

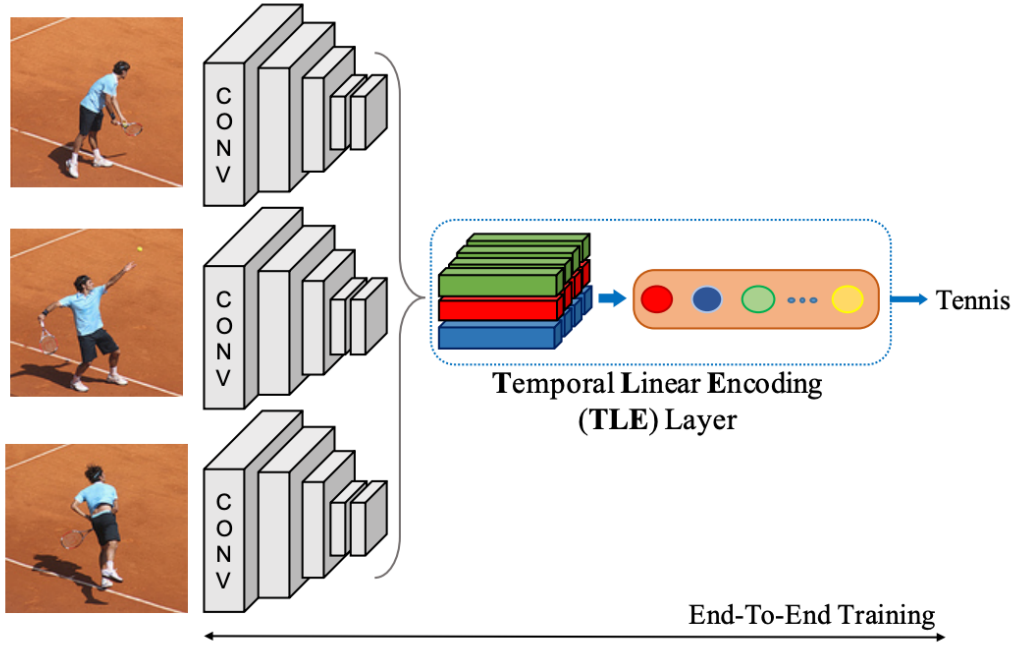
Two-Stream Networks Follow-up



	UCF101
IDT	87.9%
DeepVideo	65.4%
Two-stream	88.0%
TDD	91.5%
Beyond-Short-Snippets	88.6%
Two-Stream Fusion	92.5%
TSN	94.0%

Wang etal, Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, ECCV 2016

Two-Stream Networks Follow-up



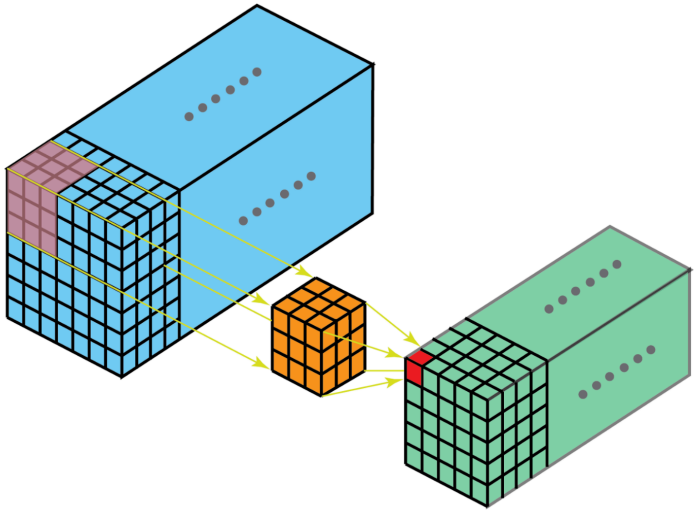
	UCF101
IDT	87.9%
DeepVideo	65.4%
Two-stream	88.0%
TDD	91.5%
Beyond-Short-Snippets	88.6%
Two-Stream Fusion	92.5%
TSN	94.0%
TLE	95.6%

Diba etal, Deep Temporal Linear Encoding Networks, CVPR 2017

3D CNNs

3D CNNs

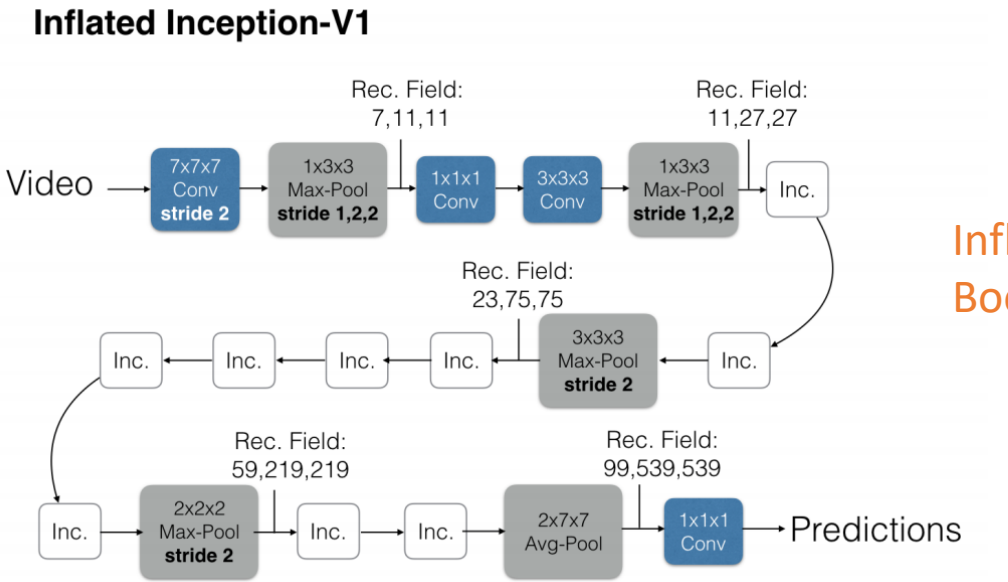
Conv1a 64	Pool1	Conv2a 128	Pool2	Conv3a 256	Conv3b 256	Pool3	Conv4a 512	Conv4b 512	Pool4	Conv5a 512	Conv5b 512	Pool5	fc6 4096	fc7 4096	softmax



	UCF101
IDT	87.9%
DeepVideo	65.4%
Two-stream	88.0%
C3D	82.3%

Tran etal, Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

3D CNNs

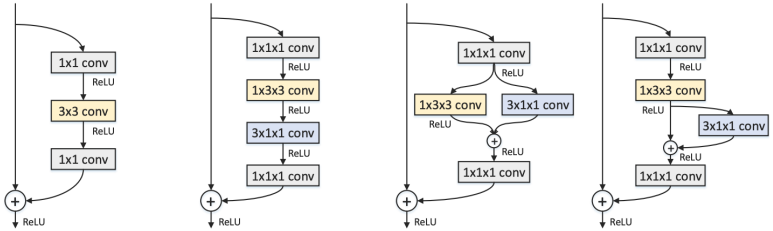


Inflating
Bootstrapping

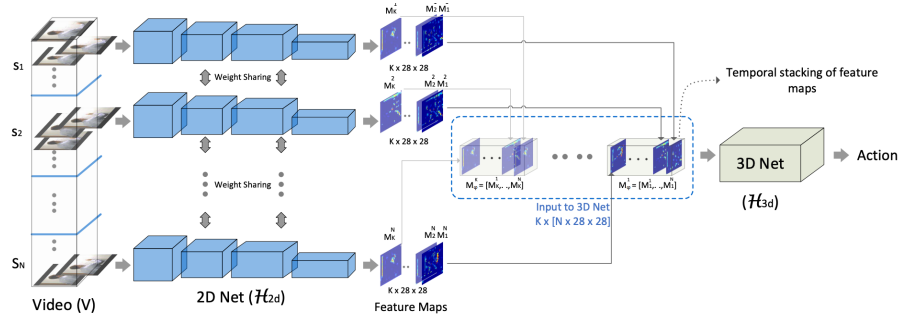
	UCF101
IDT	87.9%
DeepVideo	65.4%
Two-stream	88.0%
C3D	82.3%
I3D	95.6%

Carreira etal, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, CVPR 2017

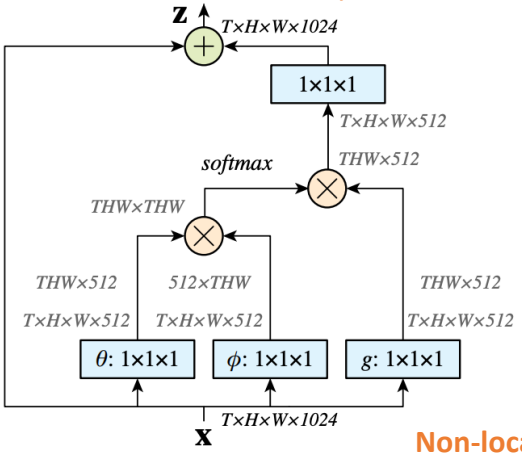
3D CNNs Modifications



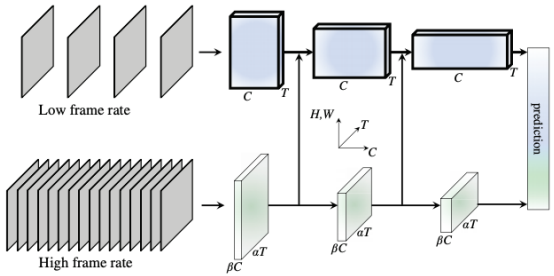
P3D, ICCV 17



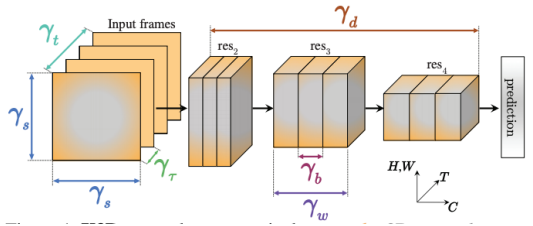
ECO, ECCV 18



Non-local, CVPR 18

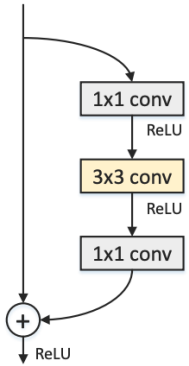


SlowFast, ICCV 19

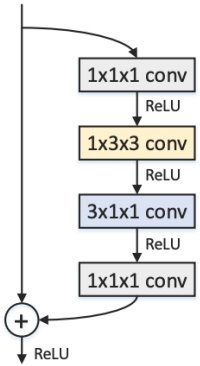


X3D, CVPR 20

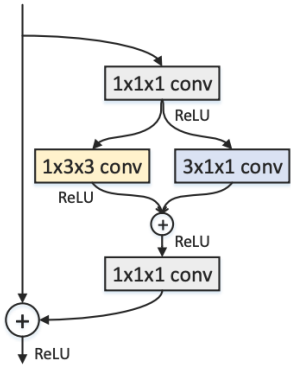
3D CNNs Modifications



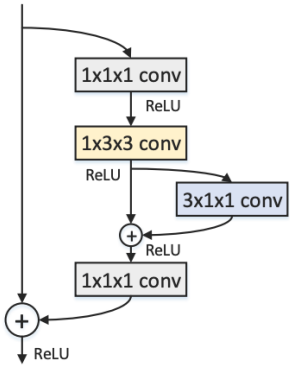
(a) Residual Unit [7]



(b) P3D-A



(c) P3D-B

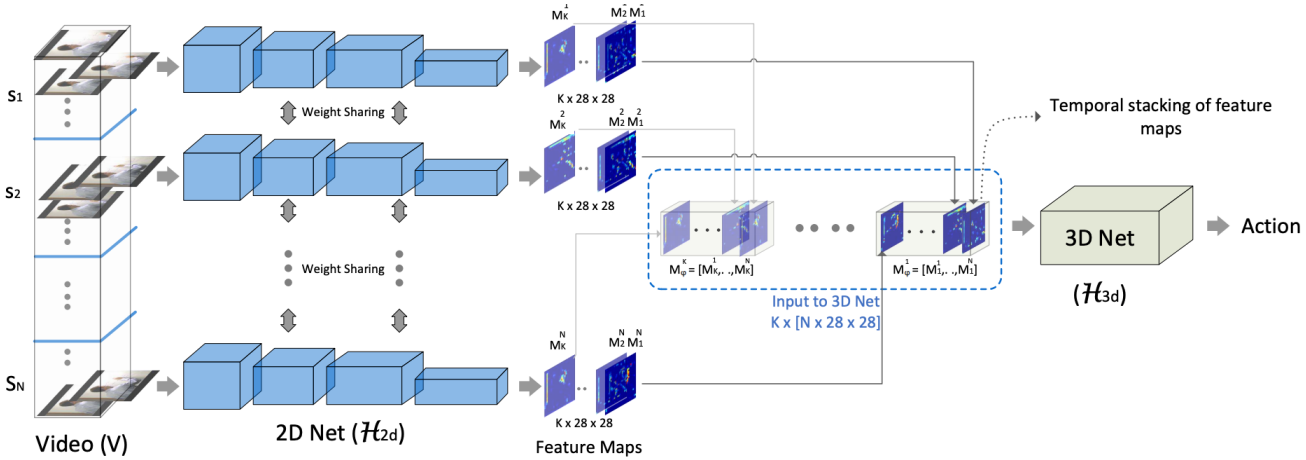


(d) P3D-C

	Kinetics400
C3D	59.5%
I3D	71.1%
P3D	72.6%

Qiu etal, Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, ICCV 2017

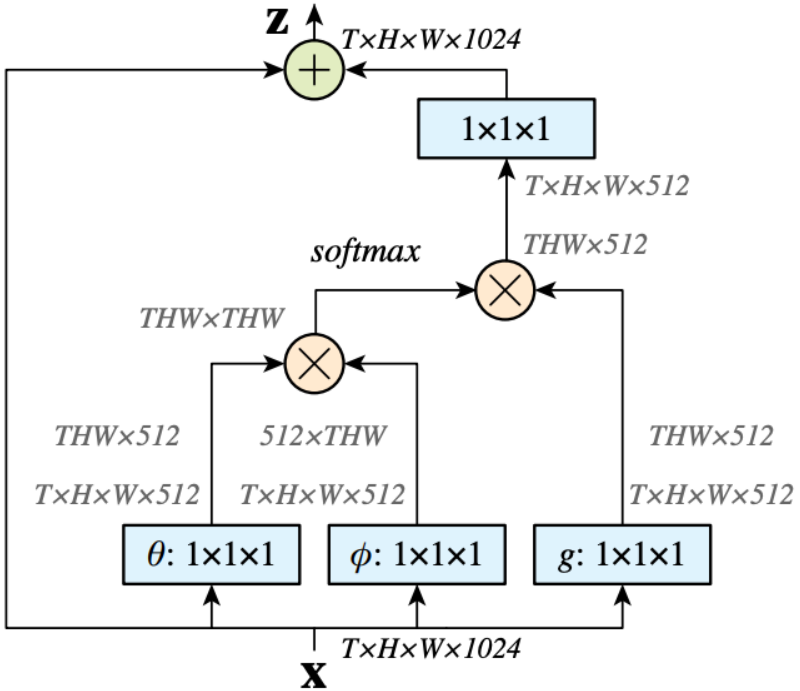
3D CNNs Modifications



	Kinetics400
C3D	59.5%
I3D	71.1%
P3D	72.6%
ECO	70.0%

Zolfaghari etal, ECO: Efficient Convolutional Network for Online Video Understanding, ECCV 2018

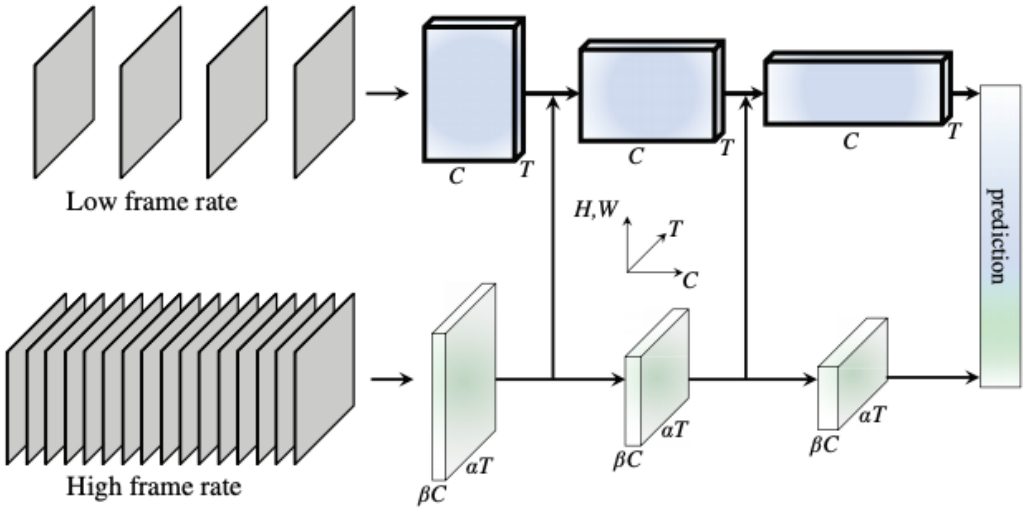
3D CNNs Modifications



	Kinetics400
C3D	59.5%
I3D	71.1%
P3D	72.6%
ECO	70.0%
Non-local	77.7%

Wang etal, Non-local Neural Networks, CVPR 2018

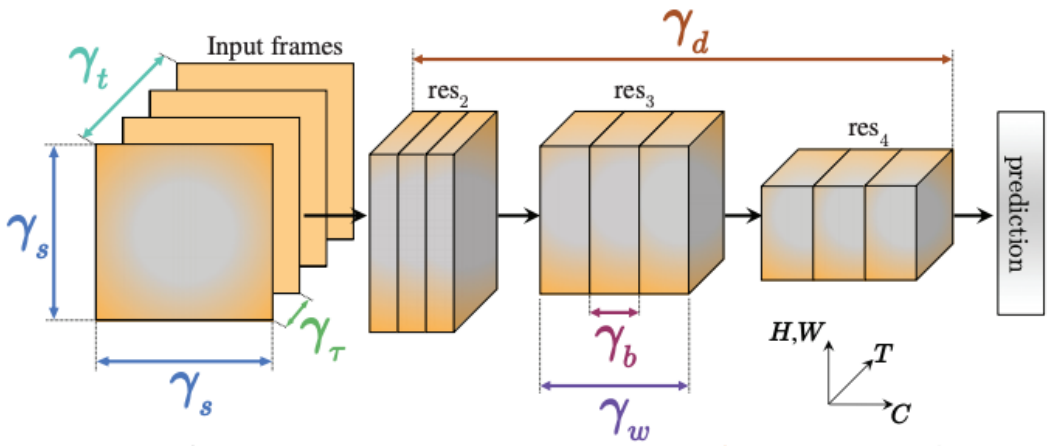
3D CNNs Modifications



	Kinetics400
C3D	59.5%
I3D	71.1%
P3D	72.6%
ECO	70.0%
Non-local	77.7%
SlowFast	78.0%

Feichtenhofer etal, SlowFast Networks for Video Recognition, ICCV 2019

3D CNNs Modifications

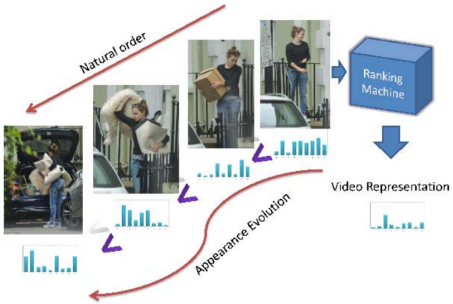


	Kinetics400
C3D	59.5%
I3D	71.1%
P3D	72.6%
ECO	70.0%
Non-local	77.7%
SlowFast	77.9%
X3D	80.4%

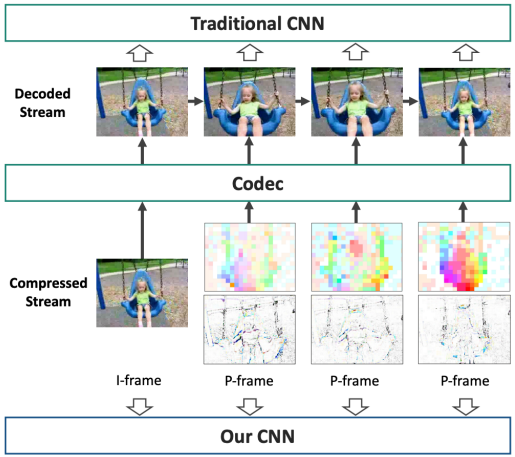
Feichtenhofer, X3D: Expanding Architectures for Efficient Video Recognition, CVPR 2020

Other Motion Modeling

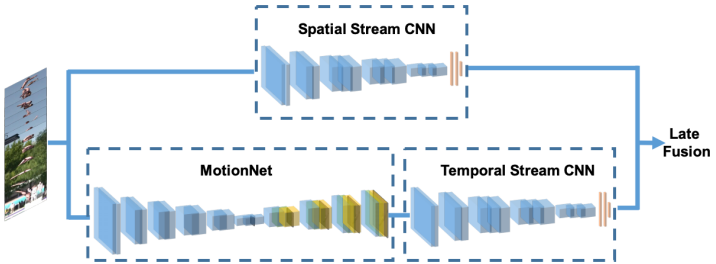
CVPR SEATTLE WASHINGTON JUNE 16-18 2020



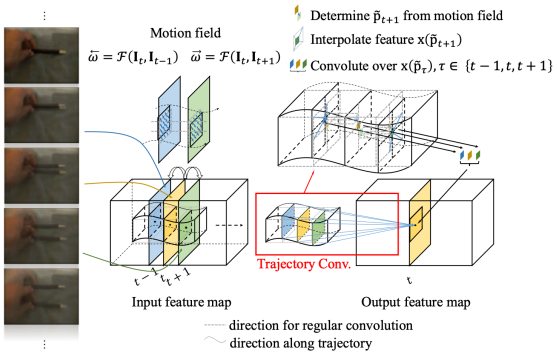
Rank Pooling, CVPR 15/16



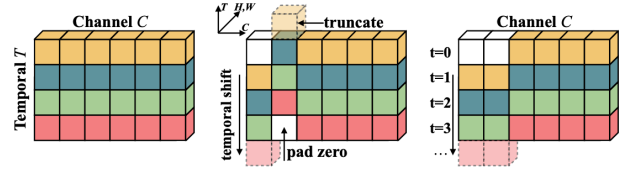
Compressed videos, CVPR 2018



Hidden TSN, ACCV 2018

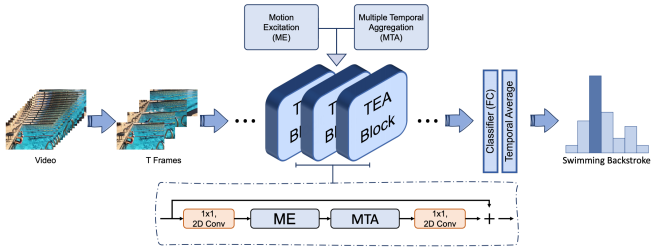


TrajectoryConv, NeurIPS 2018



(a) The original tensor without shift. **(b)** Offline temporal shift (bi-direction). **(c)** Online temporal shift (uni-direction).

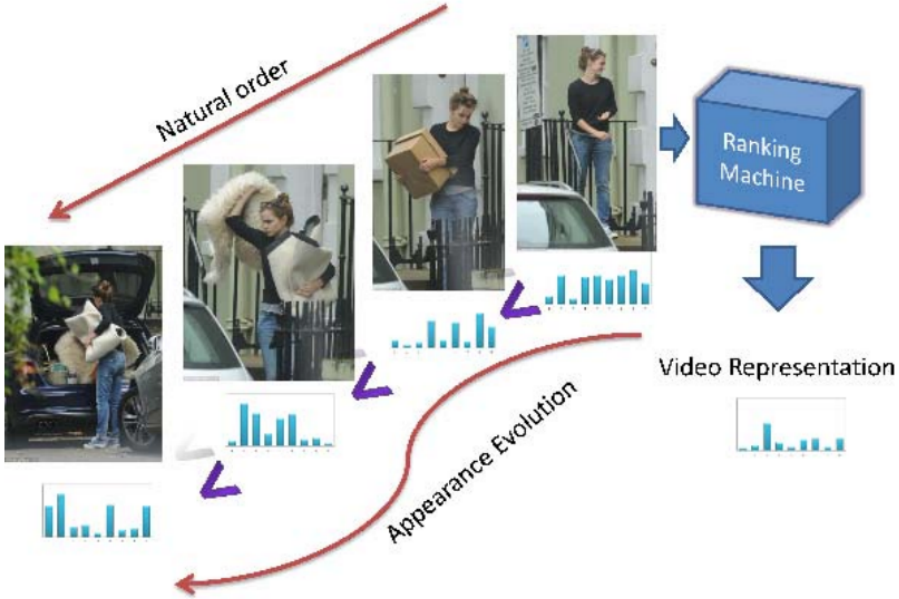
TSM, ICCV 2019



TEA, CVPR 2020

aws

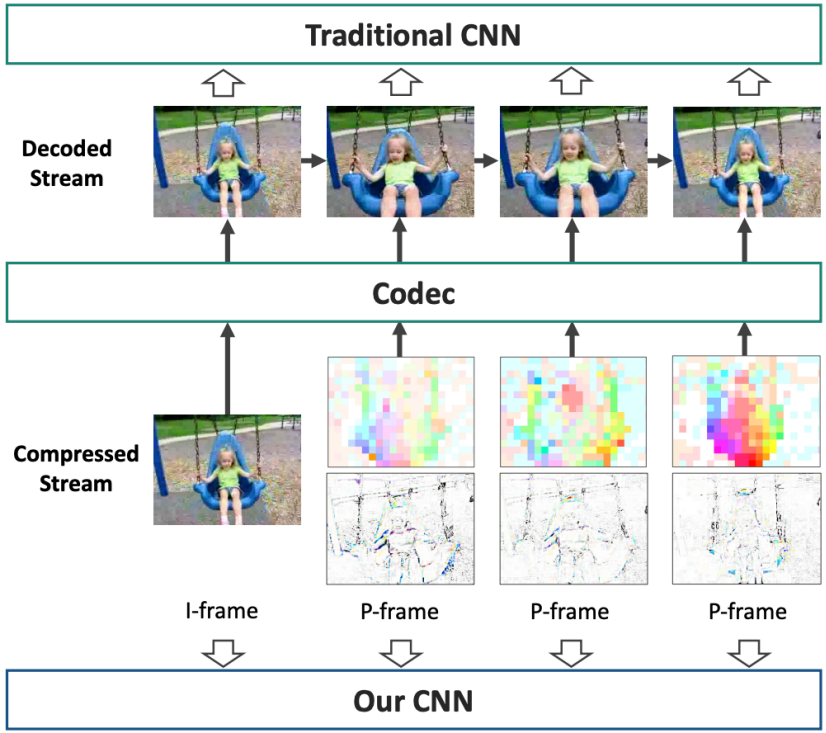
Rank Pooling



Temporal order matters

Fernando etal, Modeling Video Evolution For Action Recognition, CVPR 2015

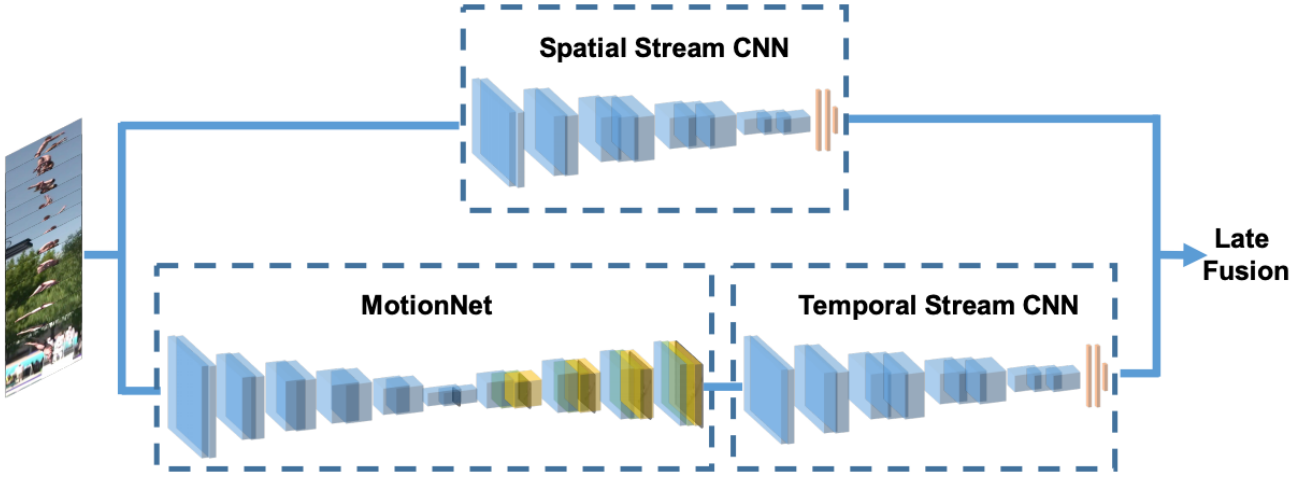
Compressed Videos



Motion vectors replaces optical flow

Wu etal, Compressed Video Action Recognition, CVPR 2018

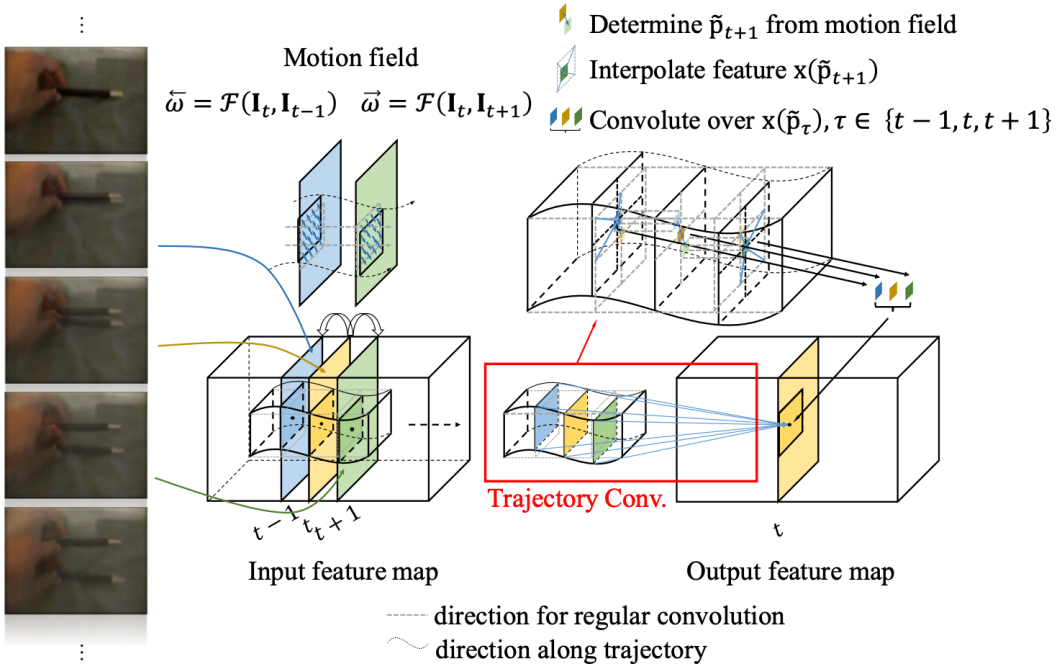
Flow-mimic Approaches



End-to-end learning of motion information by image reconstruction

Zhu etal, Hidden Two-Stream Convolutional Networks for Action Recognition, ACCV 2018

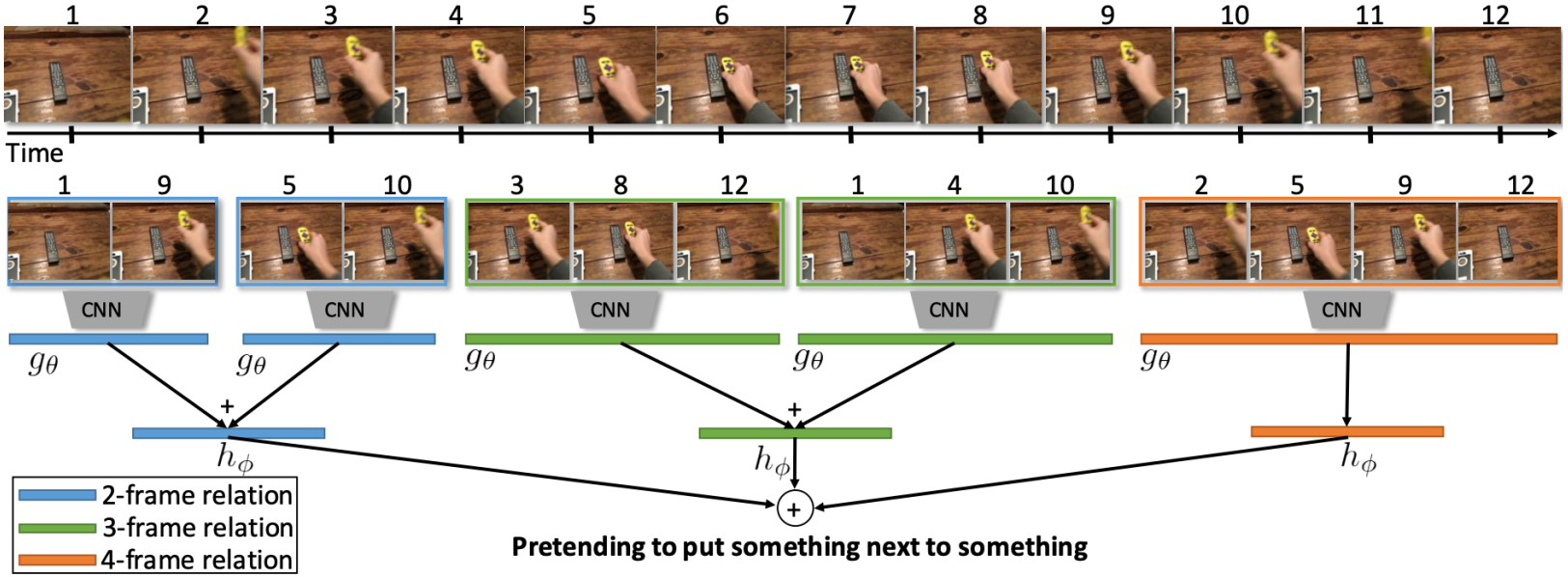
Trajectory-based



Replace temporal convolution by trajectory convolution.

Zhao etal, Trajectory Convolution for Action Recognition, NeurIPS 2018

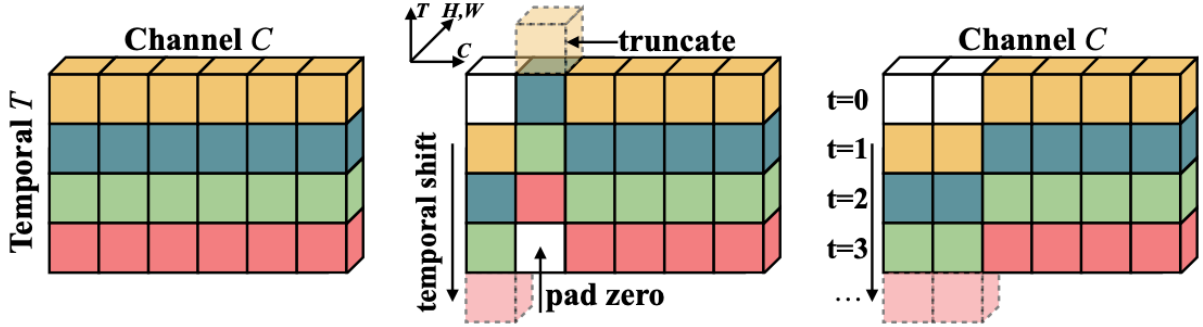
Relationship-based



Relationship reasoning

Zhou etal, Temporal Relational Reasoning in Videos, ECCV 2018

Shifting



(a) The original tensor without shift.

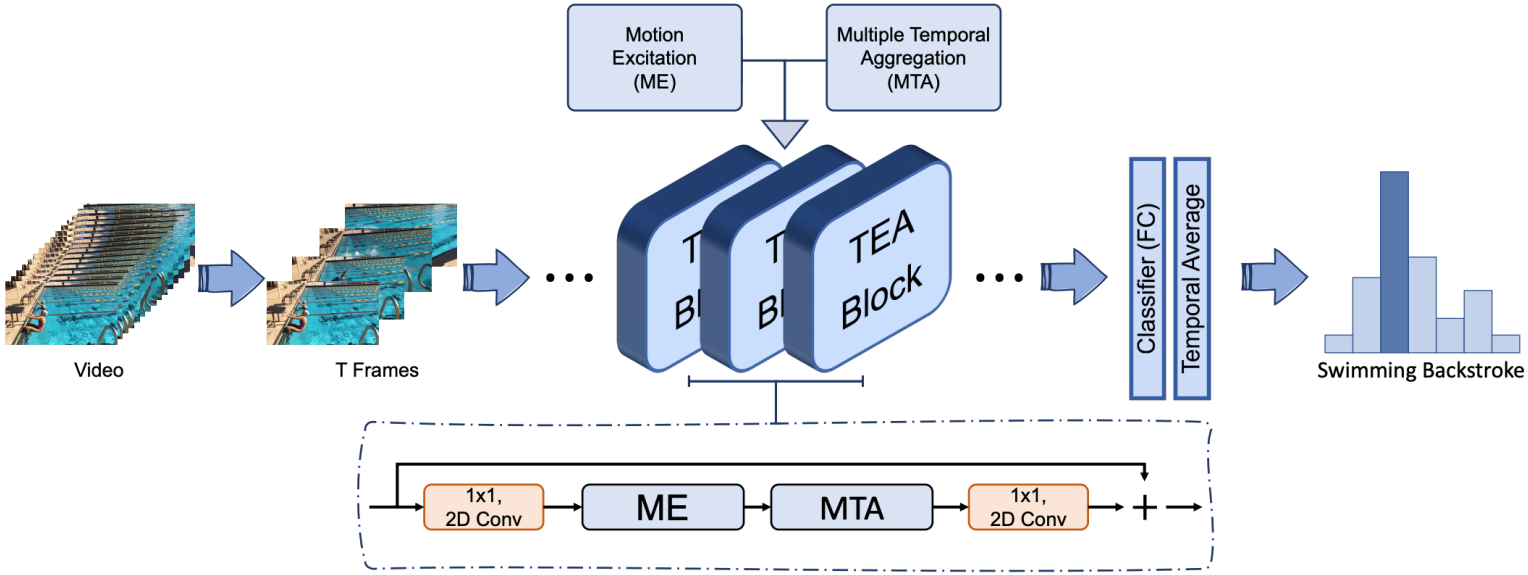
(b) Offline temporal shift (bi-direction).

(c) Online temporal shift (uni-direction).

Moving the feature map along the temporal dimension, enabling 2D Conv to model motion

Lin etal, TSM: Temporal Shift Module for Efficient Video Understanding, ICCV 2019

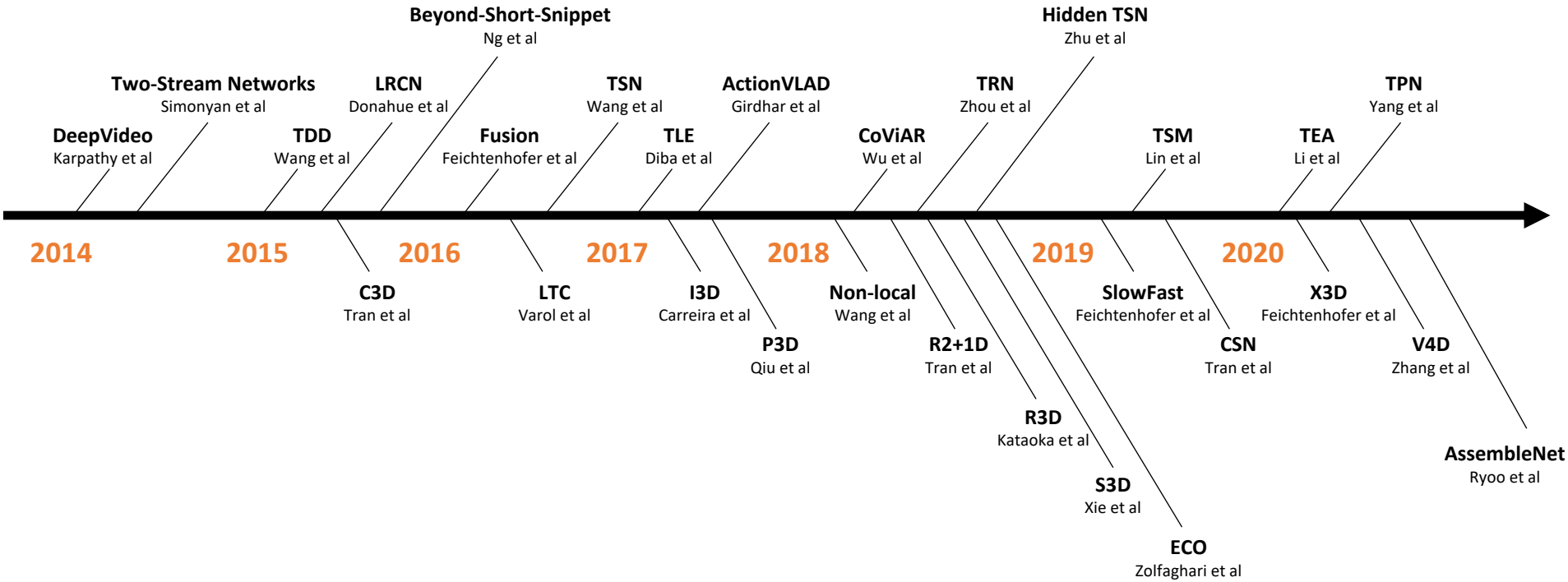
Attention



Using attention for motion modeling for 2D CNNs

Li etal, TEA: Temporal Excitation and Aggregation for Action Recognition, CVPR 2020

CVPR SEATTLE WASHINGTON JUNE 16-18 2020



aws

GluonCV Video Toolkit

aws

https://gluon-cv.mxnet.io/model_zoo/action_recognition.html

Model Zoo

Name	Pretrained	Segments	Length	1	Hashtag	Command	Log
inceptionv1_kinetics400 [3]	ImageNet	7	1	69.1	6dcdafb1	shell script	log
inceptionv3_kinetics400 [3]	ImageNet	7	1	72.5	8a4a6946	shell script	log
resnet18_v1b_kinetics400 [3]	ImageNet	7	1	65.5	46d5a985	shell script	log
resnet34_v1b_kinetics400 [3]	ImageNet	7	1	69.1	8a8d0d8d	shell script	log
resnet50_v1b_kinetics400 [3]	ImageNet	7	1	69.9	cc757e5c	shell script	log
resnet101_v1b_kinetics400 [3]	ImageNet	7	1	71.3	5bb6098e	shell script	log
resnet152_v1b_kinetics400 [3]	ImageNet	7	1	71.5	9bc70c66	shell script	log
c3d_kinetics400 [2]	Scratch	1	16 (32/2)	59.5	a007b5fa	shell script	log
p3d_resnet50_kinetics400 [5]	Scratch	1	16 (32/2)	71.6	671ba81c	shell script	log
p3d_resnet101_kinetics400 [5]	Scratch	1	16 (32/2)	72.6	b30e3a63	shell script	log
r2plus1d_resnet18_kinetics400 [6]	Scratch	1	16 (32/2)	70.8	5a14d1f9	shell script	log
r2plus1d_resnet34_kinetics400 [6]	Scratch	1	16 (32/2)	71.6	de2e592b	shell script	log
r2plus1d_resnet50_kinetics400 [6]	Scratch	1	16 (32/2)	73.9	deaefb14	shell script	log
i3d_inceptionv1_kinetics400 [4]	ImageNet	1	32 (64/2)	71.8	81e0be10	shell script	log
i3d_inceptionv3_kinetics400 [4]	ImageNet	1	32 (64/2)	73.6	f14f8a99	shell script	log
i3d_resnet50_v1_kinetics400 [4]	ImageNet	1	32 (64/2)	74.0	568a722e	shell script	log
i3d_resnet101_v1_kinetics400 [4]	ImageNet	1	32 (64/2)	75.1	6b69f655	shell script	log
i3d_nl5_resnet50_v1_kinetics400 [7]	ImageNet	1	32 (64/2)	75.2	3c0e47ea	shell script	log
i3d_nl10_resnet50_v1_kinetics400 [7]	ImageNet	1	32 (64/2)	75.3	bfb58c41	shell script	log
i3d_nl5_resnet101_v1_kinetics400 [7]	ImageNet	1	32 (64/2)	76.0	fbfc1d30	shell script	log
i3d_nl10_resnet101_v1_kinetics400 [7]	ImageNet	1	32 (64/2)	76.1	59186c31	shell script	log
slowfast_4x16_resnet50_kinetics400 [8]	Scratch	1	36 (64/1)	75.3	9d650f51	shell script	log
slowfast_8x8_resnet50_kinetics400 [8]	Scratch	1	40 (64/1)	76.6	d6b25339	shell script	log
slowfast_8x8_resnet101_kinetics400 [8]	Scratch	1	40 (64/1)	77.2	fbde1a7c	shell script	log

aws

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Model Zoo

Name	Pretrained	Segments	Length	1	Hashtag	Command	Log
inceptionv1_kinetics400 [3]	ImageNet	7	1	69.1	6dcdafb1	shell script	log
inceptionv3_kinetics400 [3]	ImageNet	7	1	72.5	8a4a6946	shell script	log
resnet18_v1b_kinetics400 [3]	ImageNet	7	1	65.5	46d5a985	shell script	log
resnet34_v1b_kinetics400 [3]	ImageNet	7	1	69.1	8a8d0d8d	shell script	log
resnet50_v1b_kinetics400 [3]	ImageNet	7	1	6			
resnet101_v1b_kinetics400 [3]	ImageNet	7	1	7			
resnet152_v1b_kinetics400 [3]	ImageNet	7	1	7			
c3d_kinetics400 [2]	Scratch	1	16 (32/2)	59.5	a007b5fa	shell script	log
p3d_resnet50_kinetics400 [5]	Scratch	1	16 (32/2)	71.6	671ba81c	shell script	log
p3d_resnet101_kinetics400 [5]	Scratch	1	16 (32/2)	72.6	b30e3a63	shell script	log
r2plus1d_resnet18_kinetics400 [6]	Scratch	1	16 (32/2)	70.8	5a14d1f9	shell script	log
r2plus1d_resnet34_kinetics400 [6]	Scratch	1	16 (32/2)	71.6	de2e592b	shell script	log
r2plus1d_resnet50_kinetics400 [6]	Scratch	1	16 (32/2)	73.9	deaefb14	shell script	log
i3d_inceptionv1_kinetics400 [4]	ImageNet	1	32 (64/2)	71.8	81e0be10	shell script	log
i3d_inceptionv3_kinetics400 [4]	ImageNet	1	32 (64/2)	73.6	f14f8a99	shell script	log
i3d_resnet50_v1_kinetics400 [4]	ImageNet	1	32 (64/2)	74.0	568a722e	shell script	log
i3d_resnet101_v1_kinetics400 [4]	ImageNet	1	32 (64/2)	75.1	6b69f655	shell script	log
			32 (64/2)	75.2	3c0e47ea	shell script	log
			32 (64/2)	75.3	bfb58c41	shell script	log
[7]	ImageNet	1	32 (64/2)	76.0	fbfc1d30	shell script	log
i3d_nl10_resnet101_v1_kinetics400 [7]	ImageNet	1	32 (64/2)	76.1	59186c31	shell script	log
slowfast_4x16_resnet50_kinetics400 [8]	Scratch	1	36 (64/1)	75.3	9d650f51	shell script	log
slowfast_8x8_resnet50_kinetics400 [8]	Scratch	1	40 (64/1)	76.6	d6b25339	shell script	log
slowfast_8x8_resnet101_kinetics400 [8]	Scratch	1	40 (64/1)	77.2	fbde1a7c	shell script	log

TSM, TVN, TPN, TEA, etc. are coming in future release!

aws

Datasets

UCF101
HMDB51
Somethig-Something-v1/v2
Kinetics400

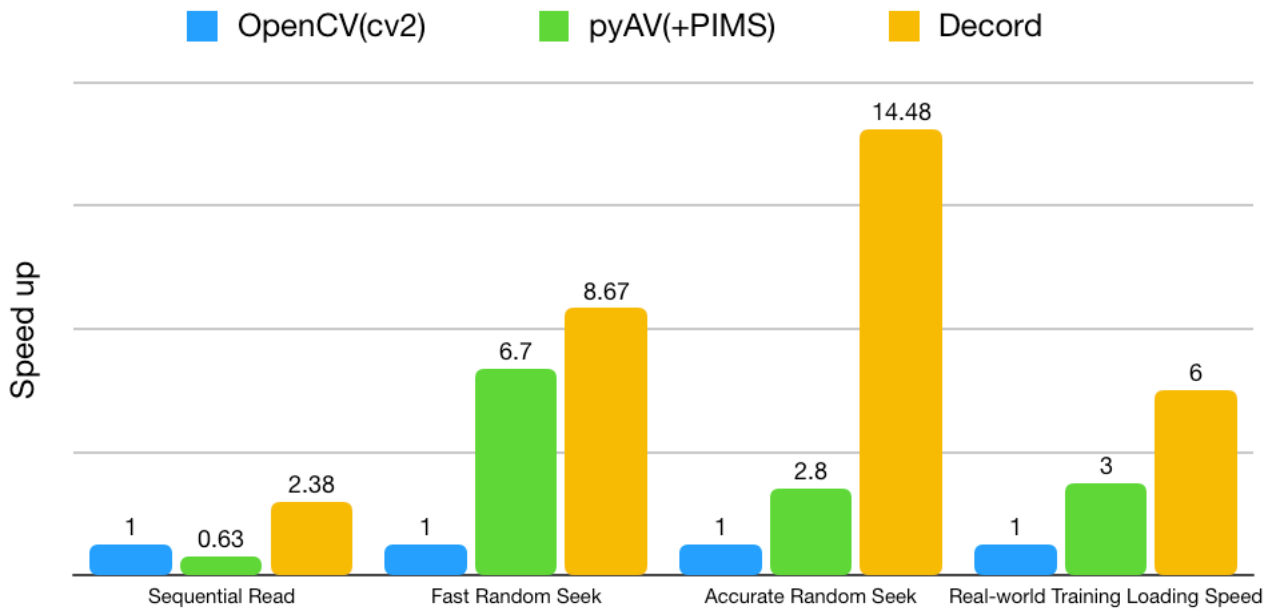
Datasets

UCF101
HMDB51
Somethig-Something-v1/v2
Kinetics400

HACS
Kinetics600/700
Moment in time

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Fast Video Reader: Decord



aws

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Tutorials

Prepare datasets
Inference
Training
Fine-tuning
Feature extraction
Distributed training

Action Recognition

Pre-trained TSN Models on UCF101 Recognize human actions in real-world videos with pre-trained TSN models	**Training TSN models on UCF101** Hands on TSN action recognition model training on UCF101 dataset	**Pre-trained I3D Models on Kinetics400** Recognize human actions in real-world videos with pre-trained I3D models
Training I3D Models on Kinetics400 Hands on I3D action recognition model training on Kinetics400 dataset	**Pre-trained SlowFast Models on Kinetics400** Recognize human actions in real-world videos with pre-trained SlowFast models	**Training SlowFast Models on Kinetics400** Hands on SlowFast action recognition model training on Kinetics400 dataset
Fine-tuning video mdoels on Your Own Dataset Hands on SOTA video models fine-tuning on your own dataset	**Extracting video features from pre-trained models** Extracting video features from pre-trained models on your own videos	**Inference on your own videos using pre-trained models** Inference on your own videos using pre-trained models and save the predictions.	**Distributed training of deep video models: SlowFast** Hands on distributed training of SlowFast models on Kinetics400 dataset.

aws

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Customized Usage

We provide a generic video dataloader, VideoClsCustom, for users to load their data.

```
train_dataset = VideoClsCustom(setting='train.txt',
                               new_length=32,
                               new_step=2,
                               new_height=224,
                               new_width=224)
```

1. Only need a text file, no specific hierarchy required.
2. Support both frame loading and video loading
3. Support all video classification datasets

```
def __init__(self,
             root,
             setting,
             train=True,
             test_mode=False,
             name_pattern='img_%05d.jpg',
             video_ext='mp4',
             is_color=True,
             modality='rgb',
             num_segments=1,
             num_crop=1,
             new_length=1,
             new_step=1,
             new_width=340,
             new_height=256,
             target_width=224,
             target_height=224,
             temporal_jitter=False,
             video_loader=False,
             use_decord=False,
             slowfast=False,
             slow_temporal_stride=16,
             fast_temporal_stride=2,
             transform=None):
```


aws

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Customized Usage

Prepare datasets
Inference
Training
Fine-tuning
Feature extraction
Distributed training

All you need is a text file!

```
video_001.mp4 200 0
video_001.mp4 200 0
video_002.mp4 300 0
video_003.mp4 100 1
video_004.mp4 400 2
......
video_100.mp4 200 10
```

Any number when using video loader

aws

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Customized Usage

We provide several off-the-shelf customized popular model for users to train/fine-tune on their data, e.g., C2D, I3D and SlowFast.

```
net = get_model(name='i3d_resnet50_v1_custom',
                nclass=USERCLASSES)

net = get_model(name='i3d_resnet50_v1_custom',
                use_kinetics_pretrain=True,
                nclass=USERCLASSES)

net = get_model(name='i3d_resnet50_v1_custom',
                freeze_backbone=True,
                nclass=USERCLASSES)
```


aws

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

More or Less Computing Resources

- Support multi-GPU training
- Support multi-machine distributed training
  6x speed up using 8 machines
- Support single-GPU training on large models with gradient accumulation
  can train I3D with 1 GPU as well.

aws

https://gluon-cv.mxnet.io/build/examples_deployment/int8_inference.html

Deployment

Model	Dataset	Batch Size	Speedup (INT8/FP32)	FP32 Accuracy	INT8 Accuracy
vgg16_ucf101	UCF101	64	4.46	81.86	81.41
inceptionv3_ucf101	UCF101	64	5.16	86.92	86.55
resnet18_v1b_kinetics400	Kinetics400	64	5.24	63.29	63.14
resnet50_v1b_kinetics400	Kinetics400	64	6.78	68.08	68.15
inceptionv3_kinetics400	Kinetics400	64	5.29	67.93	67.92

INT8 models: same performance with ~5x speed up

More details about deployment using TVM can be seen in the next talk.

Some observations

Classification head

```
self.fc = nn.Dense(in_units=self.feat_dim, units=nclass, weight_initializer=init.Normal(sigma=self.init_std))
```

	UCF101		Kinetics400	
init_std	0.01	**0.001**	**0.01**	0.001
TSN	84.3	**86.1**	**69.1**	68.7
I3D	94.5	**95.6**	**74.0**	73.4

init_std: important

A small value for small-scale datasets, particularly during fine-tuning

A big value for large-scale datasets with more classes, particularly training from scratch

Some observations

Number of frames used during training and testing

All frames sampled from a 64-frame clip

Not stable!

Kinetics400 (I3D_ResNet50)		Test		
		8	16	32
Train	8	**73.5**	73.3	73.8
	16	72.4	**73.8**	73.8
	32	72.1	72.9	**74.0**

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Some observations

Do we need 2D ImageNet pre-trained weights as model initialization for 3D CNNs?

No, at least for most 3D CNNs, like C3D, P3D, R2+1D, I3D, S3D and SlowFast.

aws

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Conclusion

- Video is the next battle field.
- Try GluonCV video toolkit! Welcome to leave issues, request features, and make contributions. https://gluon-cv.mxnet.io/
- Stay tuned. We have a survey paper coming! (200+ papers reviewed)
- All pre-recorded videos, slides and the survey paper will be uploaded to https://cvpr20-video.mxnet.io/.

aws