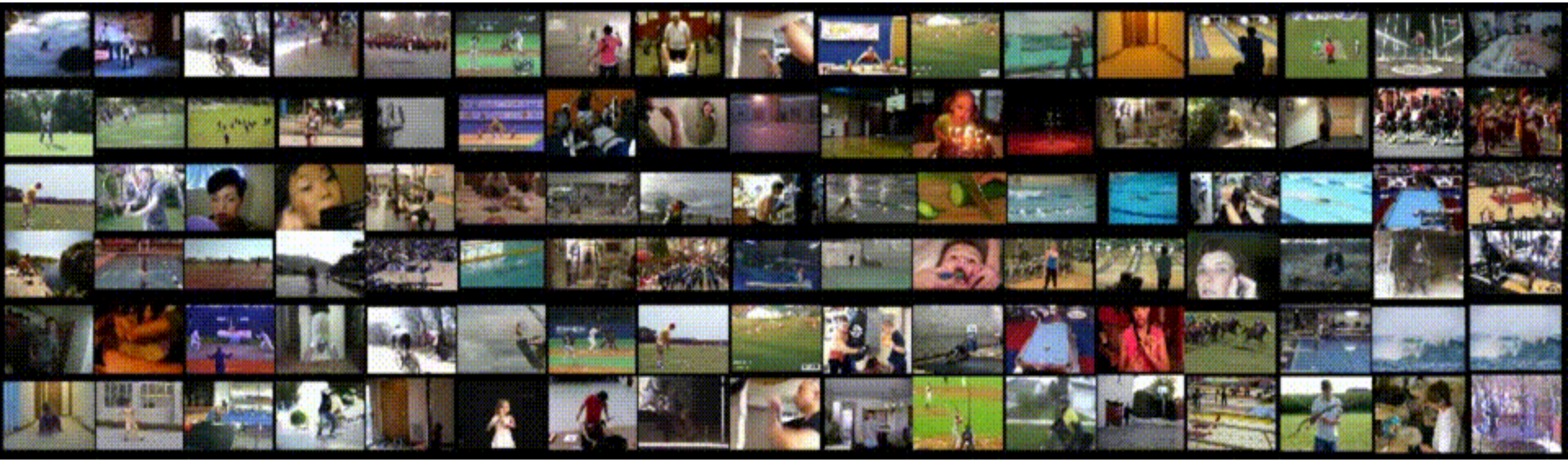
Towards efficient end-to-end architectures for action recognition and detection in videos

Limin Wang

Computer Vision Laboratory, ETH Zurich

Workshop on frontiers of video technology -- 2017

Action recognition in videos



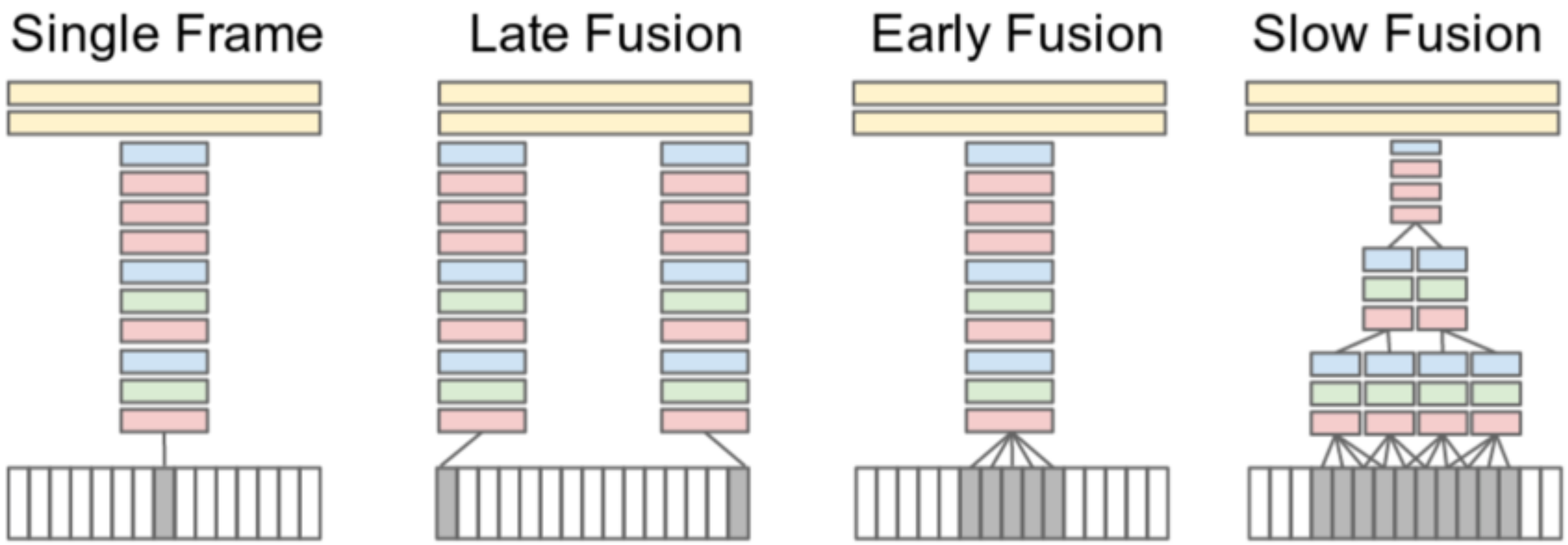
- 1. Action recognition "in the lab": KTH, Weizmann etc.
- 2. Action recognition "in TV, Movies": UCF Sports, Holloywood etc.
- 3. Action recognition "in Web Videos": HMDB, UCF101, THUMOS, ActivityNet etc.

Haroon Idrees et al. **The THUMOS Challenge on Action Recognition for Videos "in the Wild"**, in Computer Vision and Image Understanding (CVIU), 2017.

Action Understanding Tasks

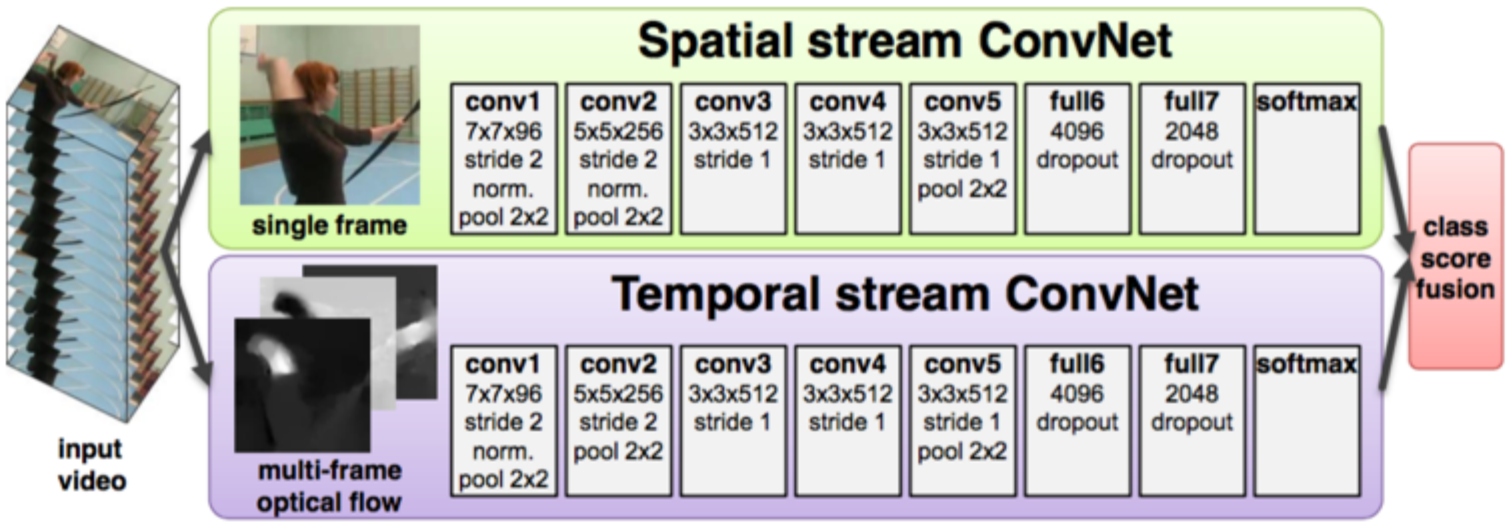
- **Action Recognition:** classify the short clip or untrimmed video into pre-defined class.
- **Action Temporal Localization:** detect starting and ending times of action instances in untrimmed video.
- **Action Spatial Detection:** detect the bounding boxes of actors in trimmed videos.
- **Action Spatial-Temporal Detection:** combine temporal and spatial localization in untrimmed videos.

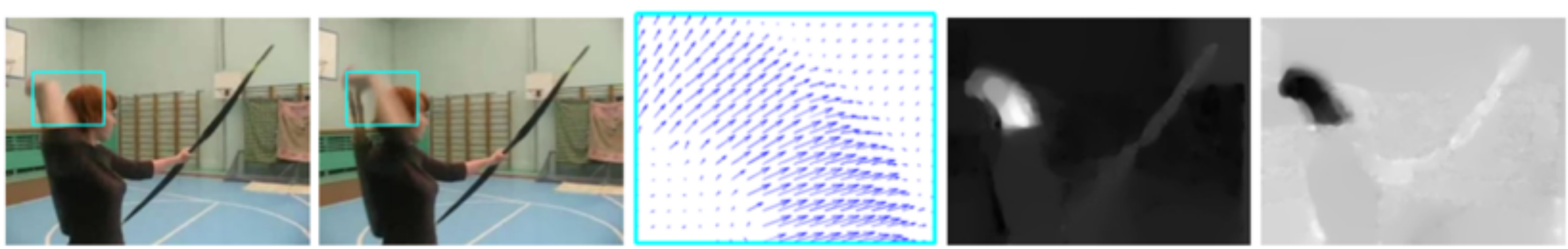
Action recognition -- deep networks (2014)



Andrej Karpathy et al., *Large-scale Video Classification with Convolutional Neural Networks*, in CVPR, 2014.

Action recognition – two stream CNN (2014)

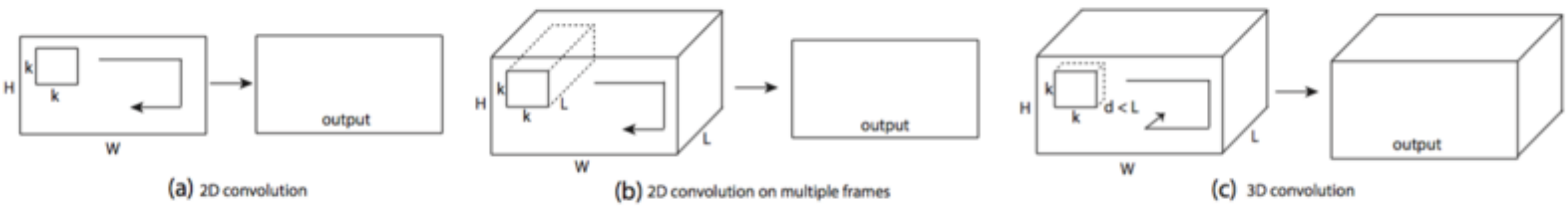




(a) (b) (c) (d) (e)

Karen Simonyan and Andrew Zisserman, *Two-Stream Convolutional Networks for Action Recognition in Videos*, in NIPS, 2014.

Action recognition – 3D CNN (2015)



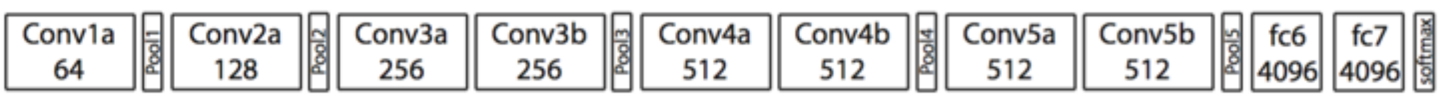


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

Du Tran et al. *Learning Spatiotemporal Features with 3D Convolutional Networks*, in ICCV, 2015.

Opportunities and Challenges

- **Opportunities**
  - Videos provide huge and rich data for visual learning
  - Action is important in motion perception and has many applications
- **Challenges**
  - Temporal models and representations
  - High computational and memory cost
  - Noisy and weakly labels

Overview

Action Recognition (AR)
- Temporal Segment Net (TSN)

Action Detection (AD)
- Structured Segment Net (SSN)

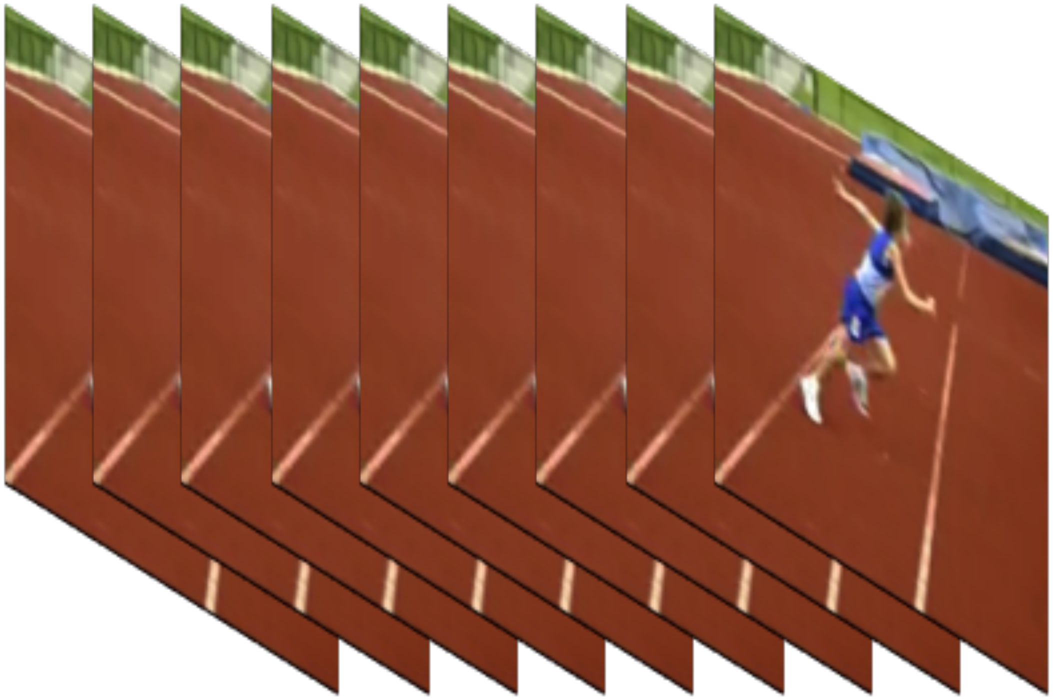
Weakly Supervised AR & AD
- UntrimmedNet

- [1] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition*, in ECCV, 2016.
- [2] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, *UntrimmedNets for Weakly Supervised Action Recognition and Detection*, in CVPR 2017.
- [3] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, D. Lin, and X. Tang, *Temporal Action Detection with Structured Segment Networks*, in ICCV 2017.

Motivation of TSN

- **Towards end-to-end and video-level architecture.**
- Modeling issue: mainstream CNN frameworks focus on appearance and short-term motion.

Modeling Long-Range Structure



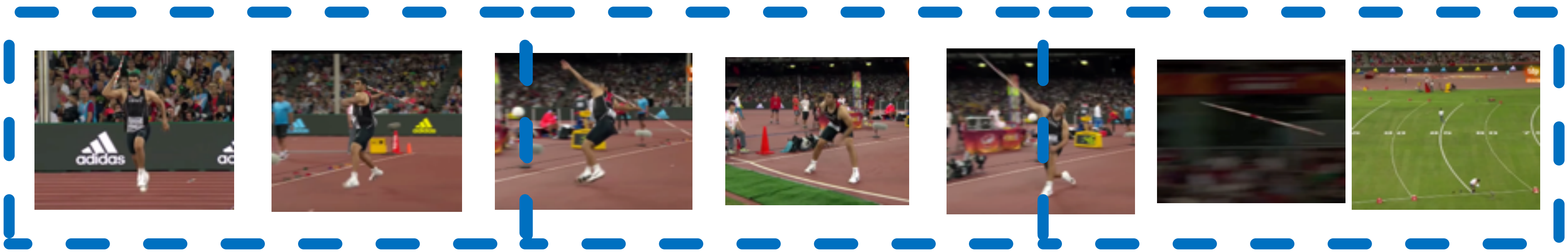
Stacking multiple frames: **dense and local**

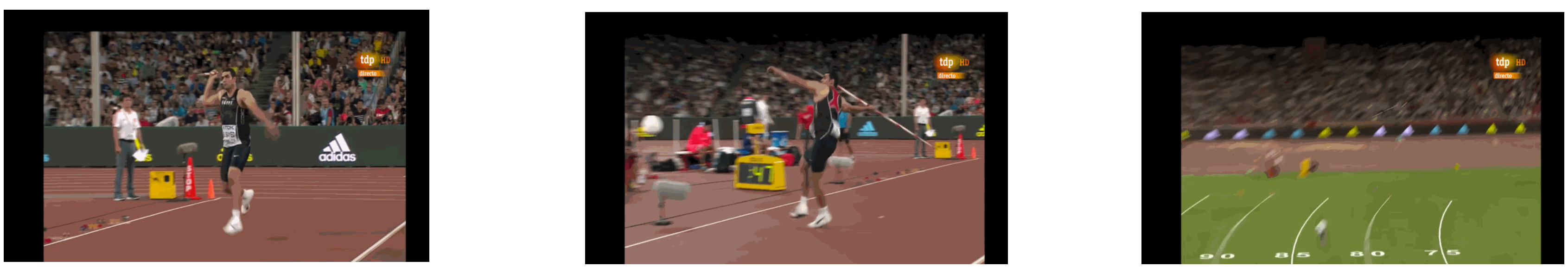
[1] Joe Yue-Hei Ng et al. Beyond Short Snippets: Deep Networks for video classification, in CVPR 2015.
[2] C. Feichtenhofer et al. Convolutional Two-Stream Network Fusion for Video Action Recognition, in CVPR 2016.
[3] Gul Varol et al., Long-term Temporal Convolutions for Action Recognition in PAMI 2017.

Segment Based Sampling

- There are high data **redundancy** in video.
- High-level semantics vary slowly (**slowness**).
- Our segment sampling share two properties:
  - **Sparse:** processing efficiency
  - **Global:** duration invariant and modeling the entire video content.

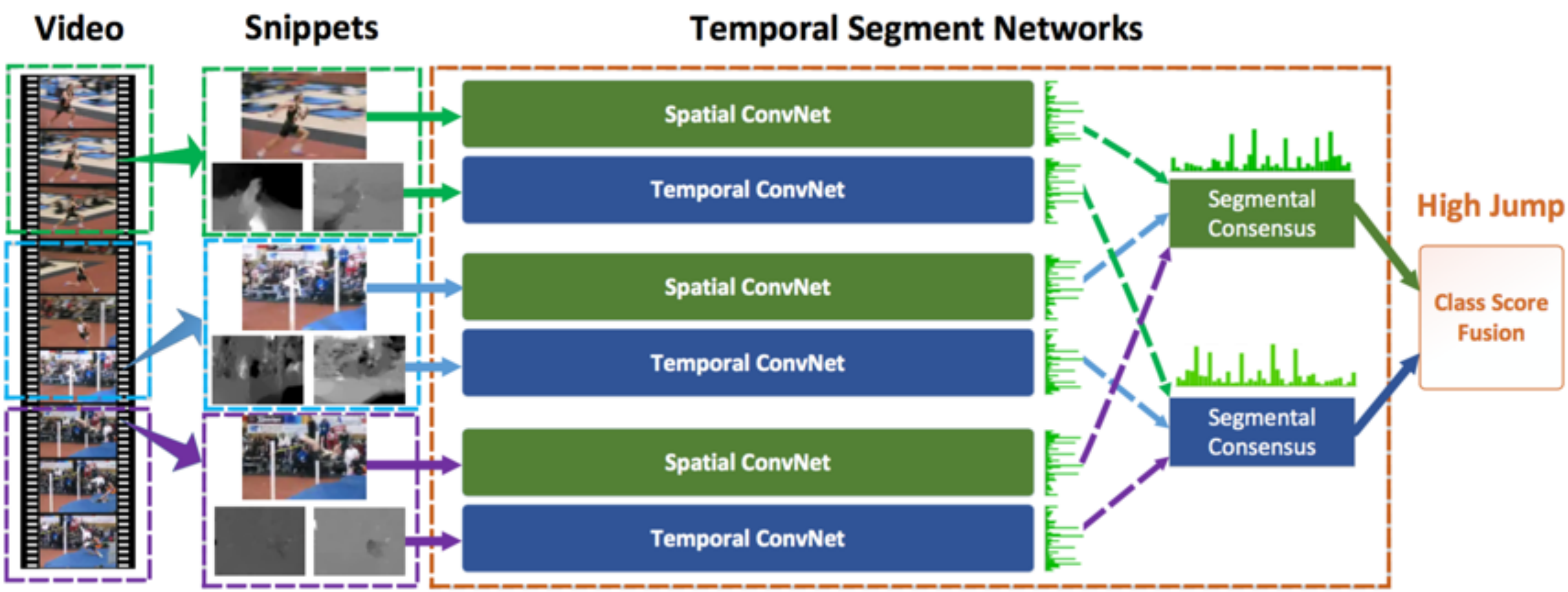
Modeling Long-Range Structure





Segment based sampling: **sparse and global**

Overview of TSN

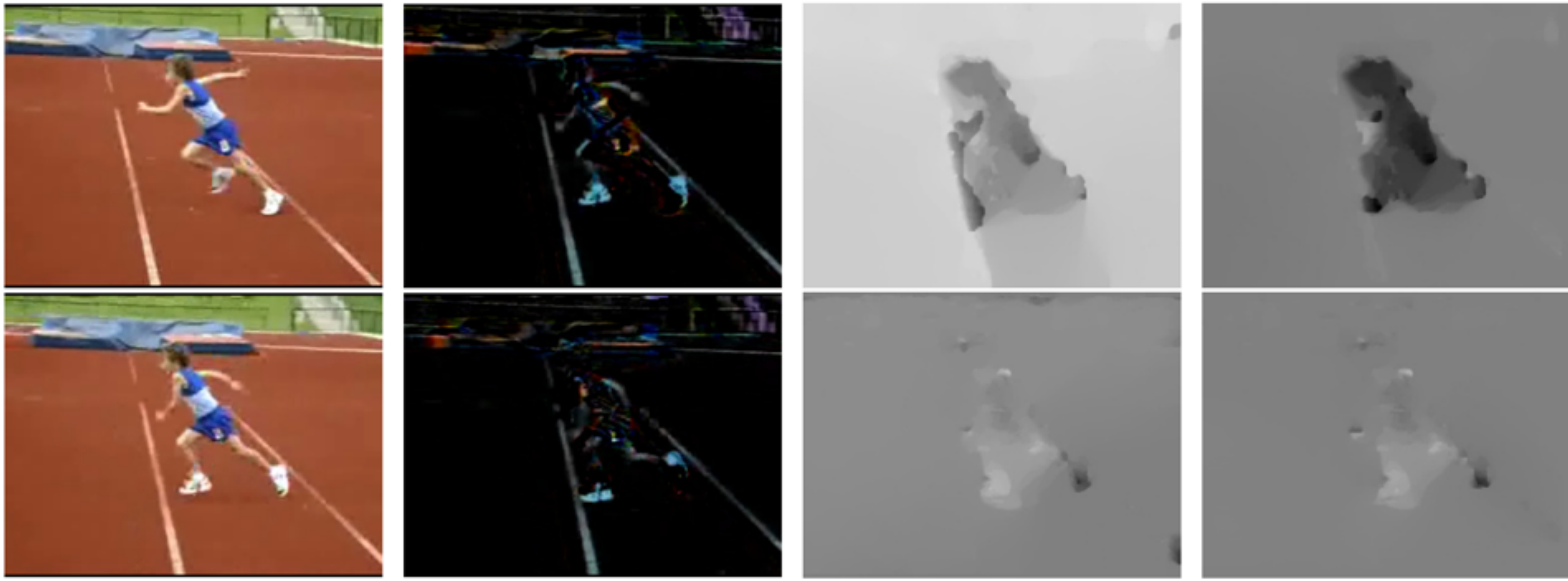


TSN is a **video-level** framework based on simple strategies of **segment sampling** and **consensus aggregation**.

Aggregation Function

- Aggregation function aims to summarize the predictions of different snippet to yield the video-level prediction.
- Simple aggregation functions:
  - Mean pooling, max pooling, weighted average
- Advanced aggregation functions:
  - Top-k pooling, attention weighting

Input modalities



Stacking RGB difference

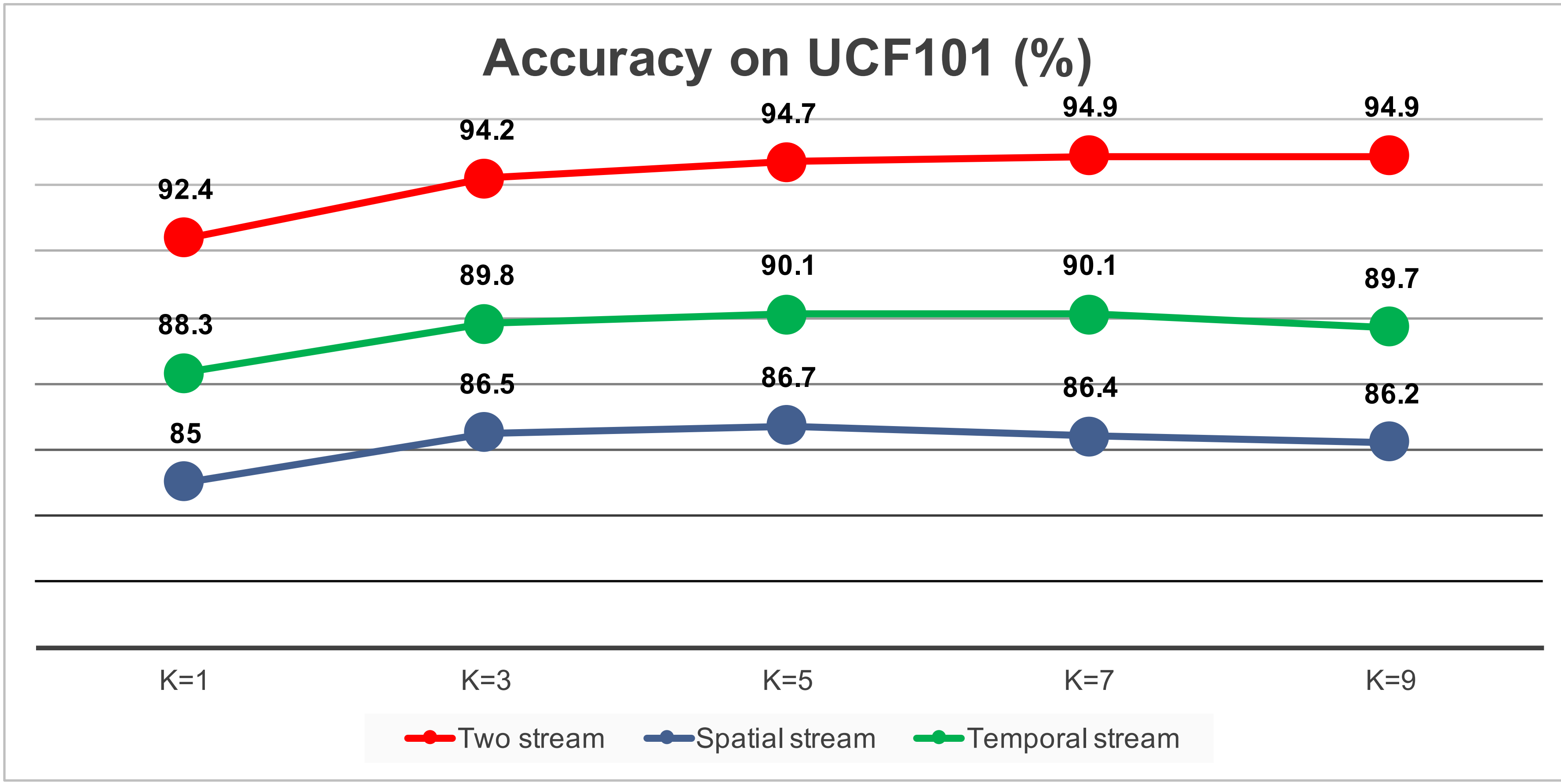
Stacking warped optical field

Experiment result -- input modality

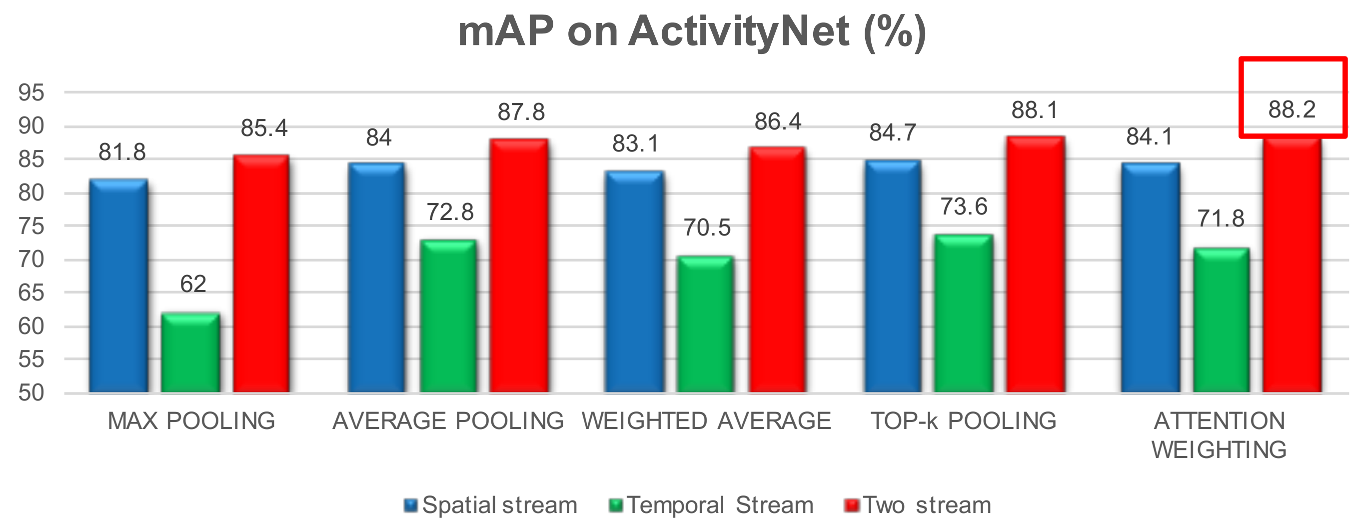
Modalities	TSN	Accuracy	Speed (FPS)
RGB+Flow	No	92.4%	14
RGB+Flow	Yes	94.9%	14
RGB+Flow+Warp	Yes	95.0%	5
Enhanced MV [17]	-	86.4%	390
Two-Stream 3DNet [65]	-	90.2%	246
RGB+RGB Diff.	No	86.8%	340
RGB+RGB Diff.	Yes	91.0%	340

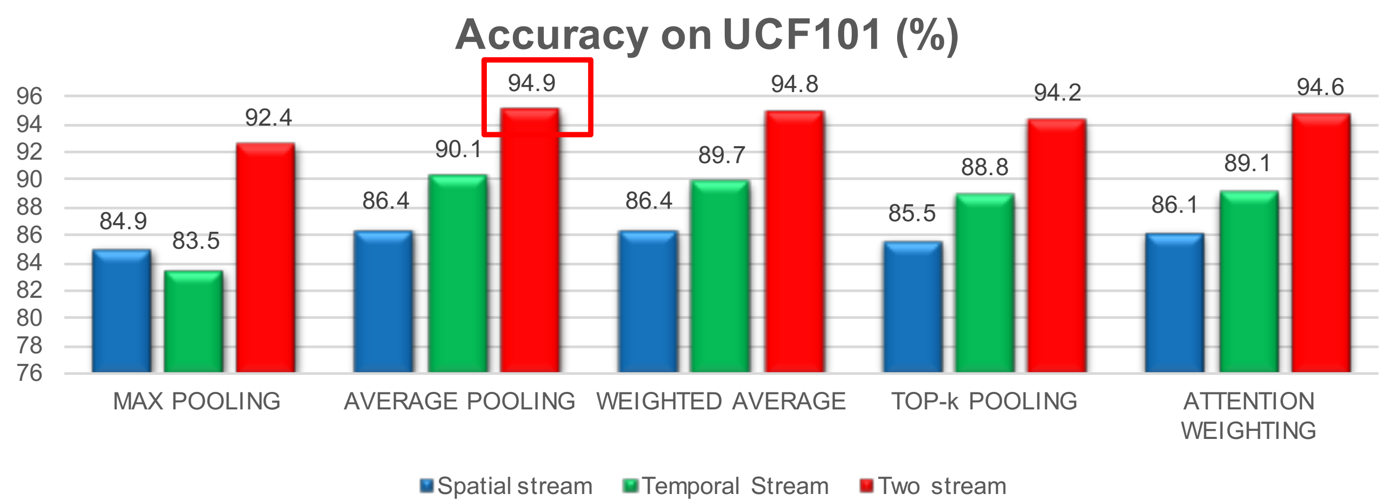
S. Ioffe et al., Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in ICML 2015.

Exploration on TSN



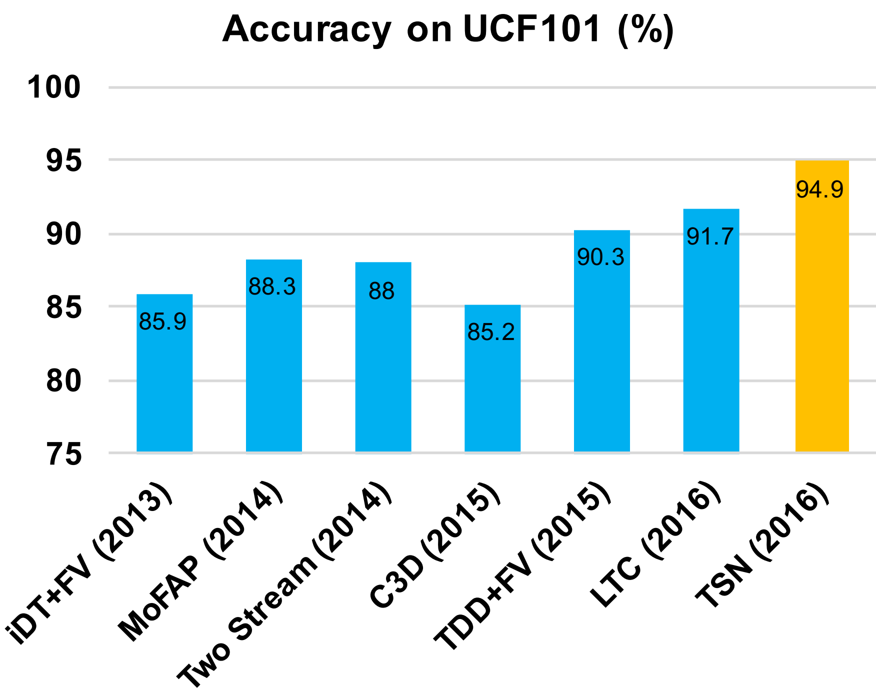
Evaluation on TSN

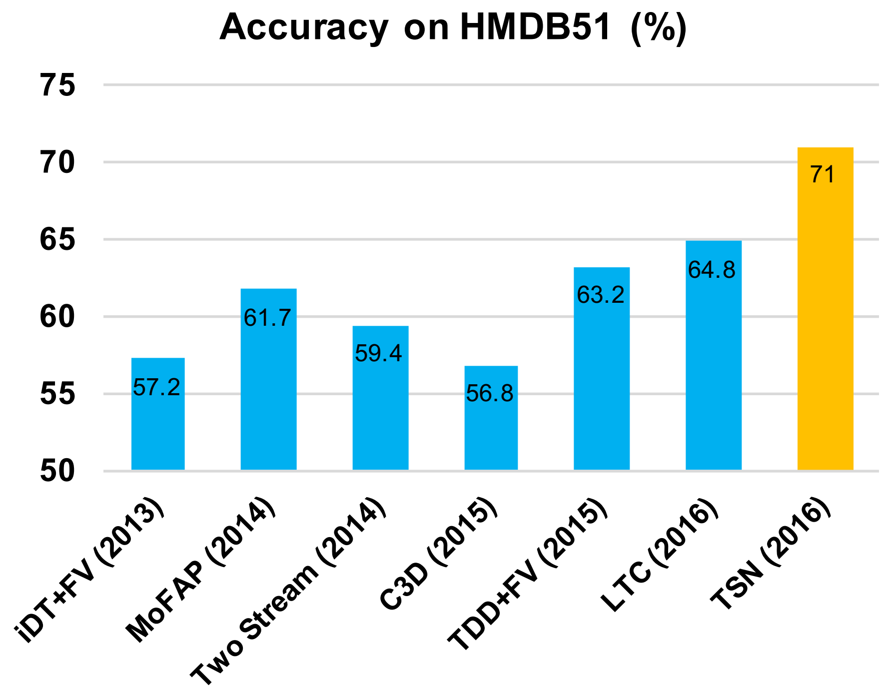




ETH zürich

Comparison on Trimmed Datasets

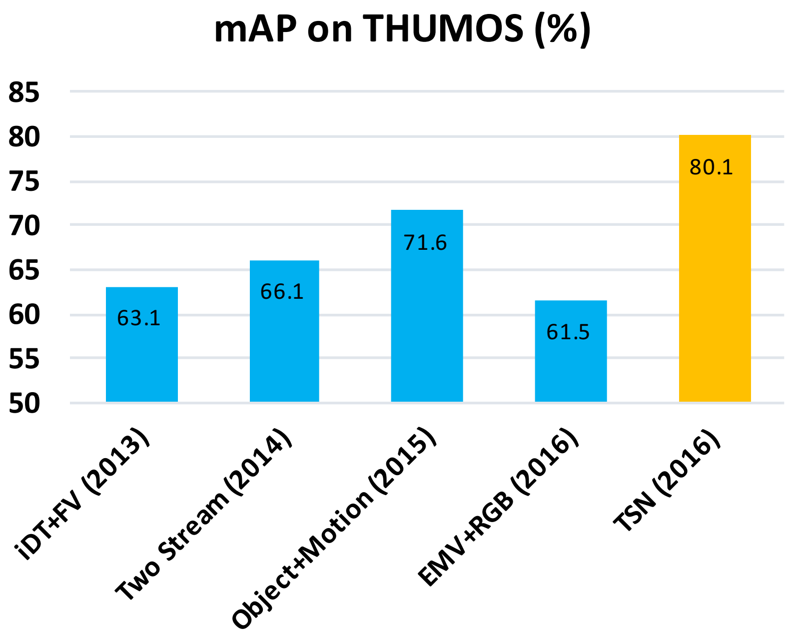


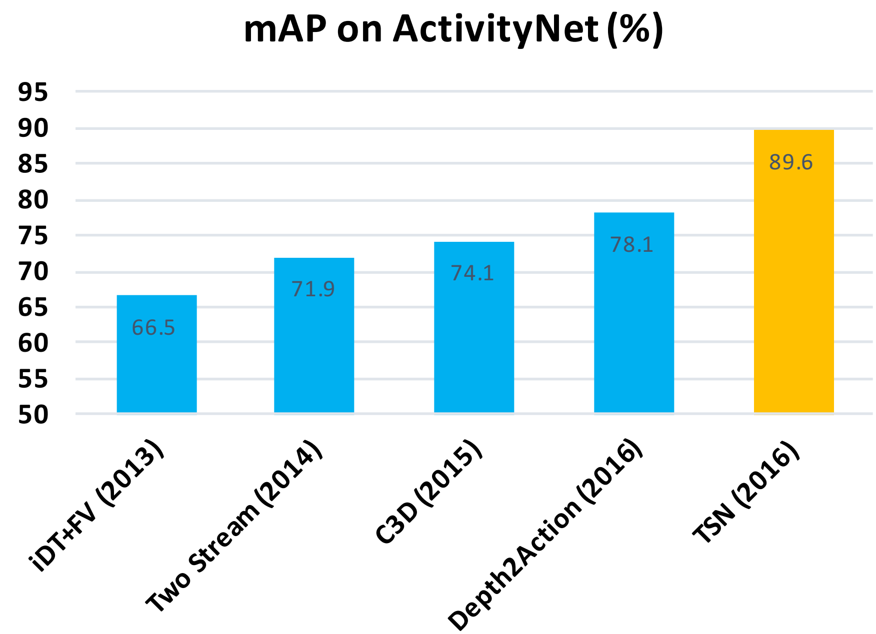


[1] L. Soomro et al., *UCF101: A dataset of 101 human action classes from videos in the wild*, in arXiv 1212.0402, 2012.
[2] H. Kuehne et al., *HMDB: A large video database for human motion recognition*, in ICCV, 2011.

ETH zürich

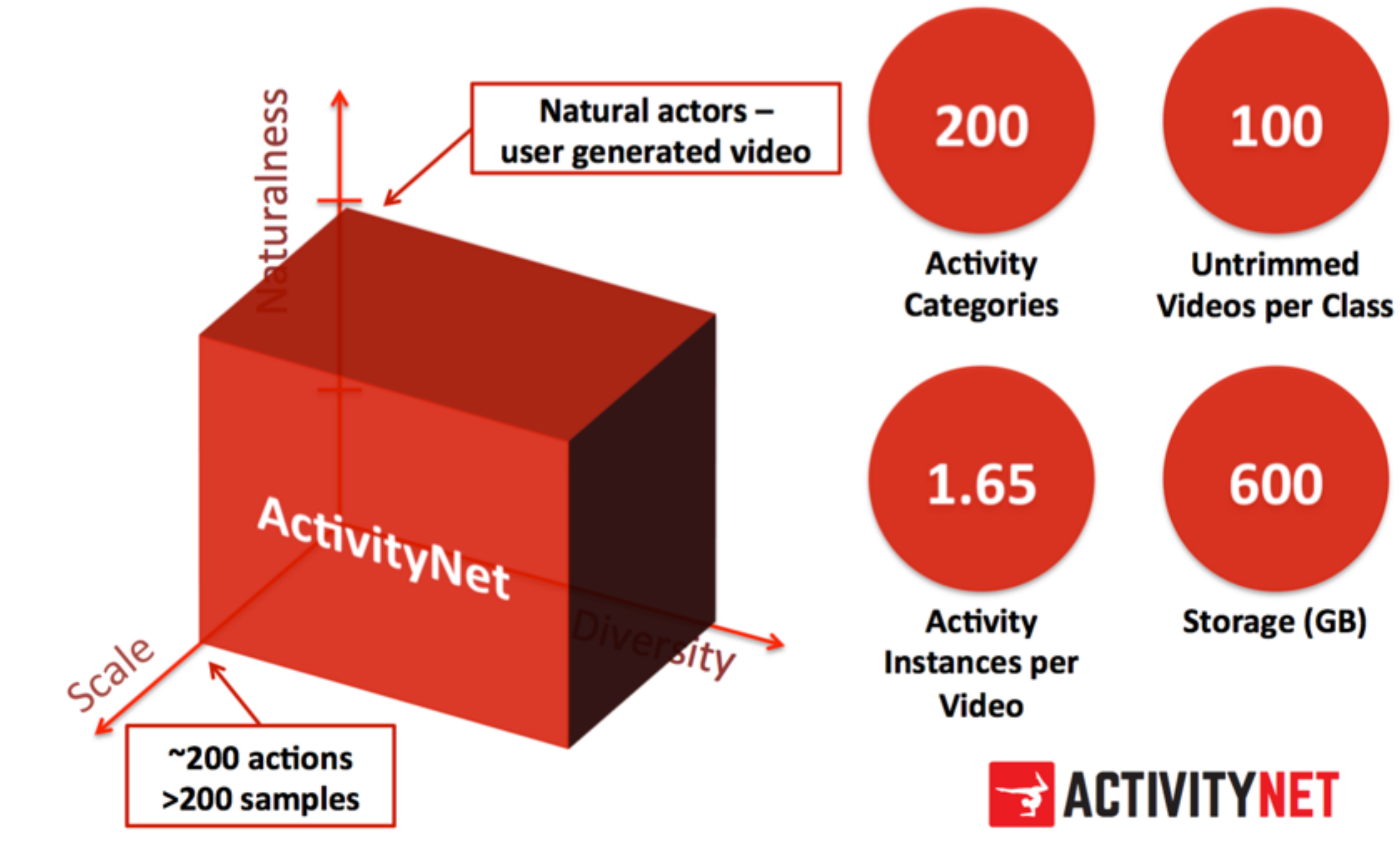
Comparison on Untrimmed Datasets





[1] H. Idrees et al., *The THUMOS Challenge on Action Recognition for Videos "in the Wild"*, in CVIU, 2017.
[2] F. C. Heilbron et al., *ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding*, in CVPR, 2015.

CVPR ActivityNet Challenge -- 2016



CVPR ActivityNet Challenge -- 2016

Settings	mAP on ActivityNet v1.3 Val.		
	Spatial	Temporal	Two Stream
BN-Inception w/o TSN	76.6%	52.7%	78.9%
TSN + BN-Inception	79.7%	63.6%	84.7%
TSN + Inception V3	83.3%	64.4%	87.7%
TSN-Top3 + Inception V3	84.5%	64.0%	88.0%
TSN-Ensemble	85.9%	68.3%	**89.7%**

CVPR ActivityNet Challenge -- 2016

ACTIVITYNET
Large Scale Activity Recognition Challenge

Home People Important Dates Program Guidelines Evaluation Contact Us

CVPR2016

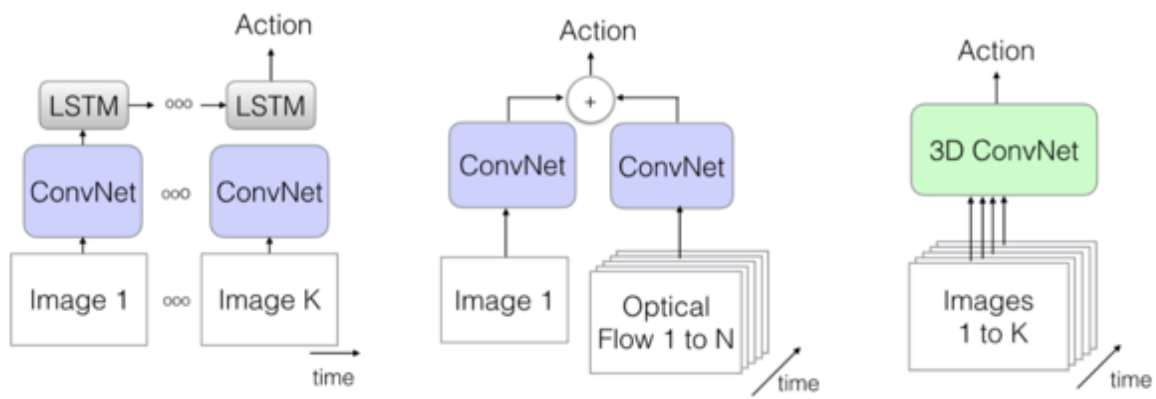
Leaderboard - Untrimmed Video Classification

Ranking	Username	Organization	Upload time	mAP	Top-1	Top-3
1	Limin Wang [Challenge16]	CUHK & ETHZ & SIAT	2016-06-08 14:10:36	0.93233	0.88136	0.96421
2	Ruxin Wang [Challenge16]	QCIS	2016-06-09 06:47:55	0.92413	0.87792	0.97084
3	Ting Yao [Challenge16]	Multimedia Search and Mining Group, MSRA	2016-06-09 07:26:36	0.91937	0.86685	0.95535
4	Linchao Zhu [Challenge16]	UTS	2016-05-08 12:01:45	0.87163	0.849	0.9504
5	Masatoshi Hidaka [Challenge16]	The University of Tokyo	2016-06-09 05:28:49	0.86458	0.80434	0.9262
6	Ke Ning [Challenge16]	Zhejiang University	2016-06-09 05:22:02	0.84104	0.8339	0.93525
7	Cong Guo [Challenge16]	University of Science and Technology of China	2016-05-18 17:18:53	0.84067	0.79654	0.91355
8	Yi Zhu [Challenge16]	UC Merced	2016-06-06 22:05:16	0.831	0.78444	0.91072
9	Cesar Roberto de Souza [Challenge16]	Xerox Research Center Europe	2016-06-08 15:58:55	0.82607	0.78524	0.89584

Kinetics dataset

Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500
ActivityNet-200 [3]	2015	200	avg 141	28,108	19,994
Kinetics	2017	400	min 400	306,245	306,245

Results on Kinetics Dataset



Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB+Flow	RGB	Flow	RGB+Flow	RGB	Flow	RGB+Flow
(a) ConvNet+LSTM	84.3	–	–	43.9	–	–	57.0 / 79.0	–	–
(b) Two-Stream	84.2	85.9	92.5	51.0	56.9	63.7	56.0 / 77.3	49.5 / 71.9	61.0 / 81.3
(c) 3D-ConvNet	51.6	–	–	24.3	–	–	56.1 / 79.5	–	–

RGB, Pretrain on ImageNet, TSN: top-1 70.28%, 89.13%

RGB, Train from scratch, TSN: top-1 69.55%, 88.68%

Overview

Action Recognition (AR)
- Temporal Segment Net (TSN)

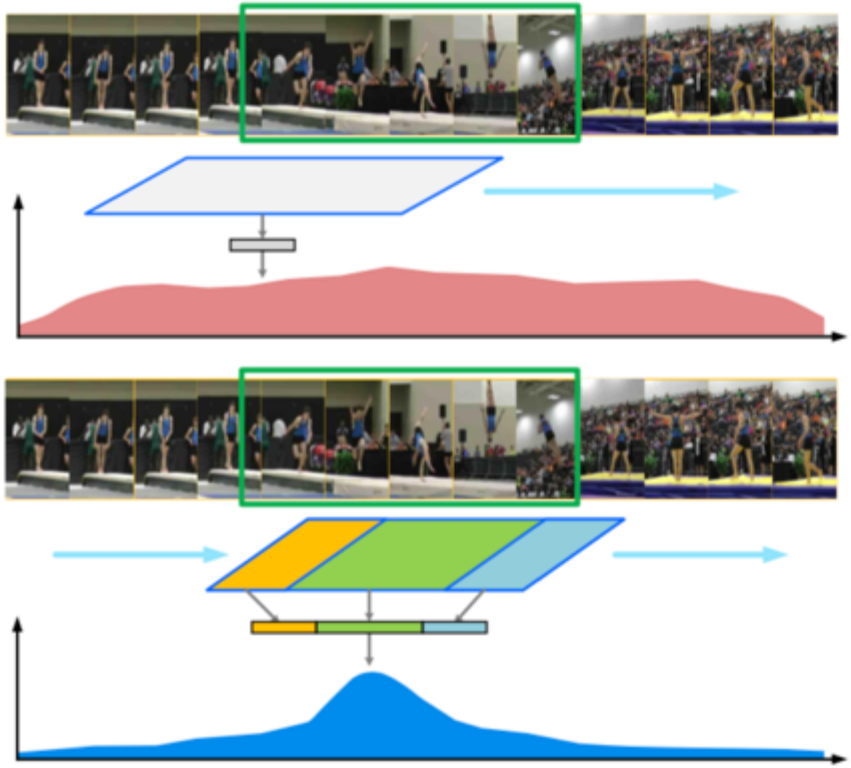
Action Detection (AD)
- **Structured Segment Net (SSN)**

Weakly Supervised AR & AD
- UntrimmedNet

- [1] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition*, in ECCV, 2016.
- [2] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, *UntrimmedNets for Weakly Supervised Action Recognition and Detection*, in CVPR 2017.
- [3] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, D. Lin, and X. Tang, *Temporal Action Detection with Structured Segment Networks*, in ICCV 2017.

Motivation of Structured Segment Network

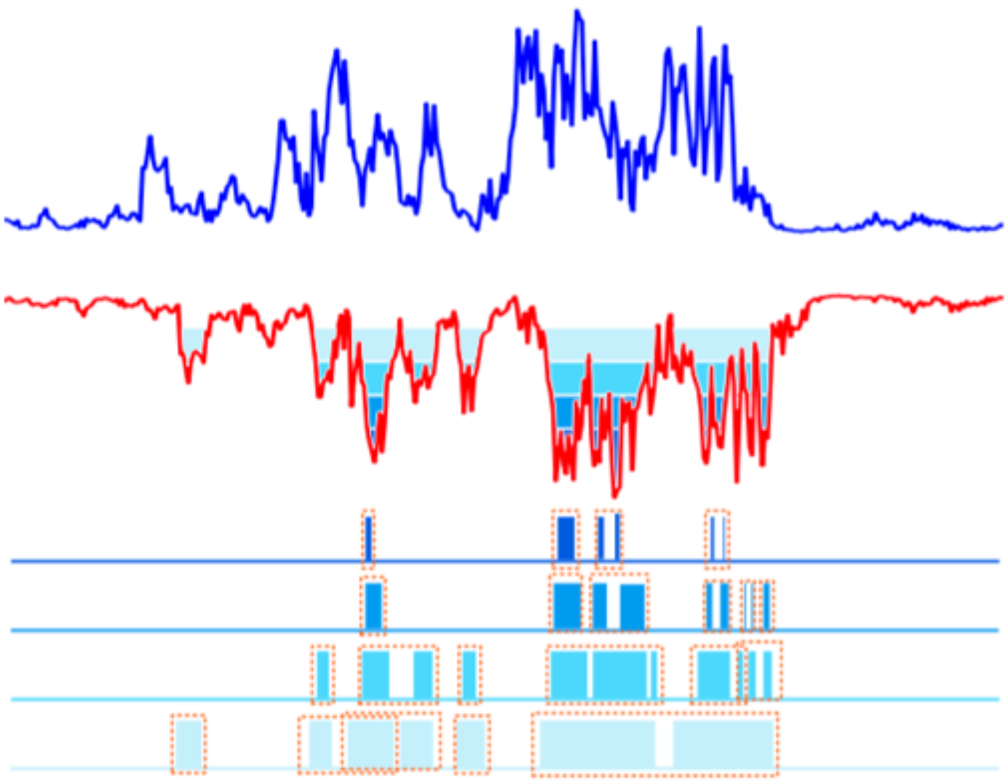
1. Action detection in untrimmed video is an important problem.
2. Snippet-level classifier is difficult to accurately localize the temporal extent of action instance.



Context and Structure Modeling!

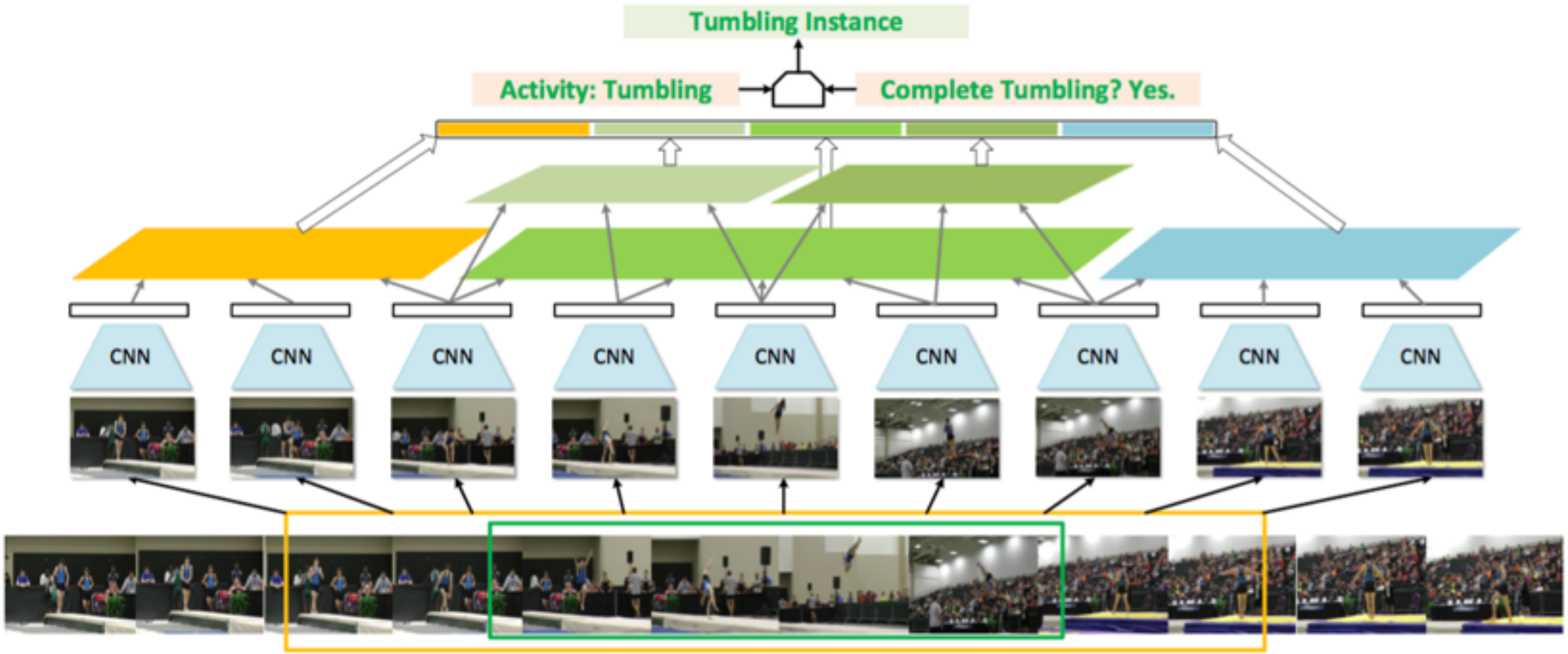
ETH zürich

Temporal Region Proposal



Bottom up proposal generation based on actionness map

Structured Segment Network (SSN)



Two Classifier Design

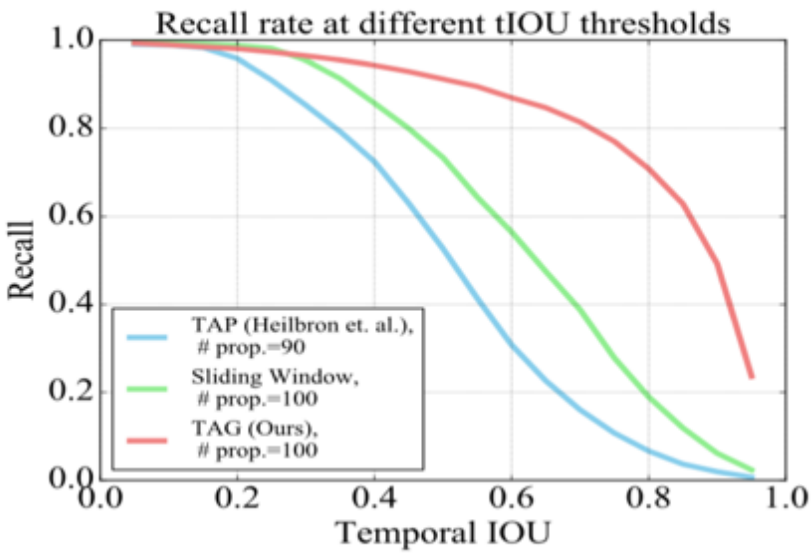
- To model the **class classes** and **completeness of instances**, we design a two classifier loss

P(c,b|p) = P(c|p)P(b|c,p)

- Action class classifier measure the likelihood of action class distribution: P(c|p)
- Completeness classifier measure the likelihood of instance completeness: P(b|c,p)
- A joint loss to optimize these two classifiers:

$$\mathcal{L}_{cls}(c_i, b_i; p_i) = -\log P(c_i|p_i) - 1_{(c_i \geq 1)} P(b_i|c_i, p_i)$$

Experiment result -- Action Proposal



Proposal Method	THUMOS14		ActivityNet v1.2	
	# Prop.	AR	# Prop.	AR
Sliding Windows	204	21.2	100	34.8
SCNN-prop [36]	200	20.0	-	-
TAP [6]	200	23.0	90	14.9
DAP [5]	200	37.0	100	12.1
TAG	200	39.6	100	67.3

[1] V. Escorcia, F. Caba Heilbron, J. C. Niebles, and B. Ghanem. Daps: Deep action proposals for action understanding. In, *ECCV*, pages 768–784, 2016.

[2] Fabian Caba Heilbron et al.. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *CVPR*, 2016.

Experiment result -- Component Analysis

	Stage-Wise				End-to-End	
STPP		✓	✓	✓	✓	✓
Act. + Comp.			✓	✓	✓	✓
Loc. Reg.				✓		✓
SW	0.558	2.26	16.4	18.1	-	-
TAG	4.82	9.55	23.7	24.2	23.8	24.5

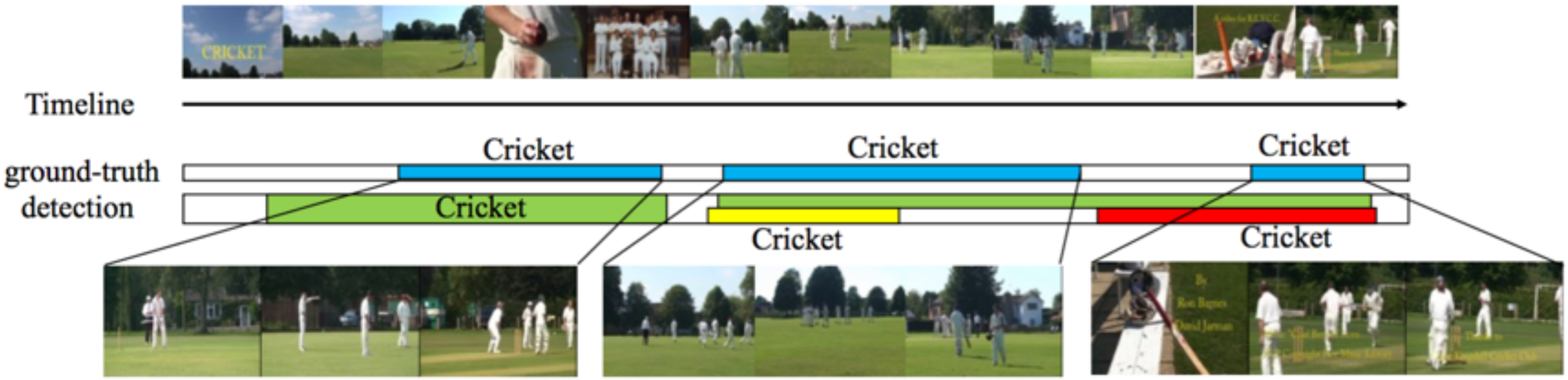
Experiment result -- Comparison

THUMOS14, mAP@α					
Method	0.1	0.2	0.3	0.4	0.5
Wang *et. al.* [46]	18.2	17.0	14.0	11.7	8.3
Oneata *et. al.* [30]	36.6	33.6	27.0	20.8	14.4
Richard *et. al.* [34]	39.7	35.7	30.0	23.2	15.2
S-CNN [36]	47.7	43.5	36.3	28.7	19.0
Yeung *et. al.* [52]	48.9	44.0	36.0	26.4	17.1
Yuan *et. al.* [53]	51.4	42.6	33.6	26.1	18.8
SSN	**64.1**	**57.7**	**48.7**	**39.8**	**28.2**

ActivityNet v1.3 (testing), mAP@α				
Method	0.5	0.75	0.95	Average
Wang *et. al.* [50]	42.478	2.88	0.06	14.62
Singh *et. al.* [39]	28.667	17.78	2.88	17.68
Singh *et. al.* [40]	36.398	11.05	0.14	17.83
SSN	43.261	28.70	5.63	**28.28**

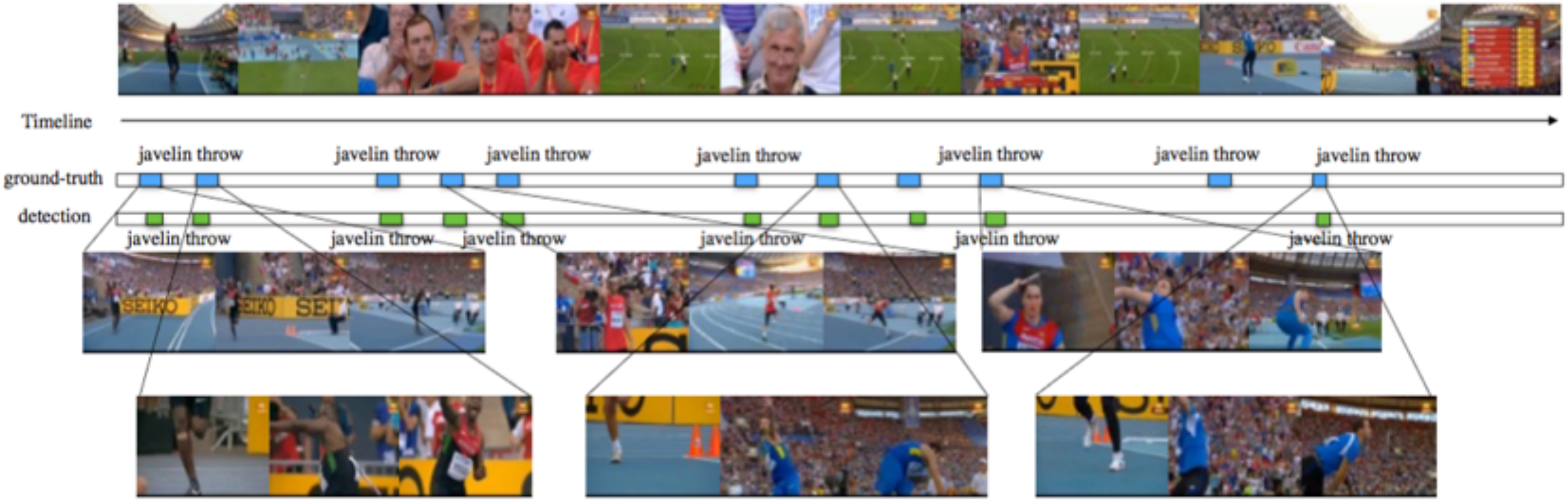
[1] H. Idrees et al., *The THUMOS Challenge on Action Recognition for Videos "in the Wild",* in CVIU, 2017.
[2] F. C. Heilbron et al., *ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding,* in CVPR, 2015.

Detection example (1)



Green: correct detection
Red: bad localization
Yellow: multiple detections

Detection example (2)



Green: correct detection
Red: bad localization
Yellow: multiple detections

ETHzürich

Overview of temporal modeling

Action Recognition (AR)
- Temporal Segment Net (TSN)

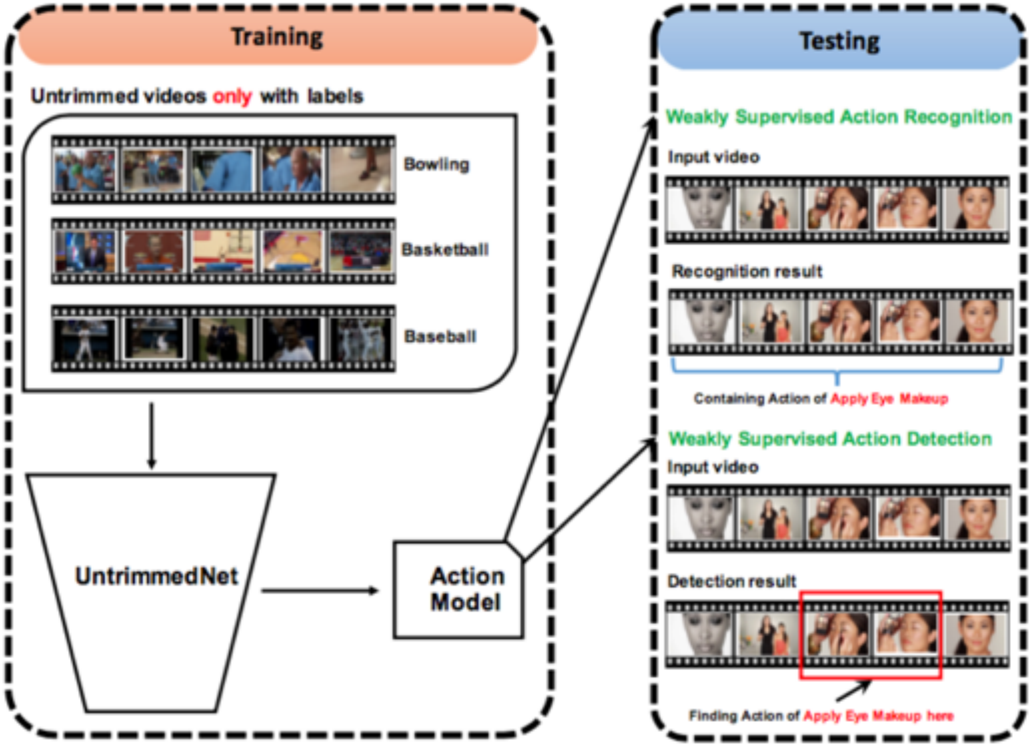
Action Detection (AD)
- Structured Segment Net (SSN)

Weakly Supervised AR & AD
- **UntrimmedNet**

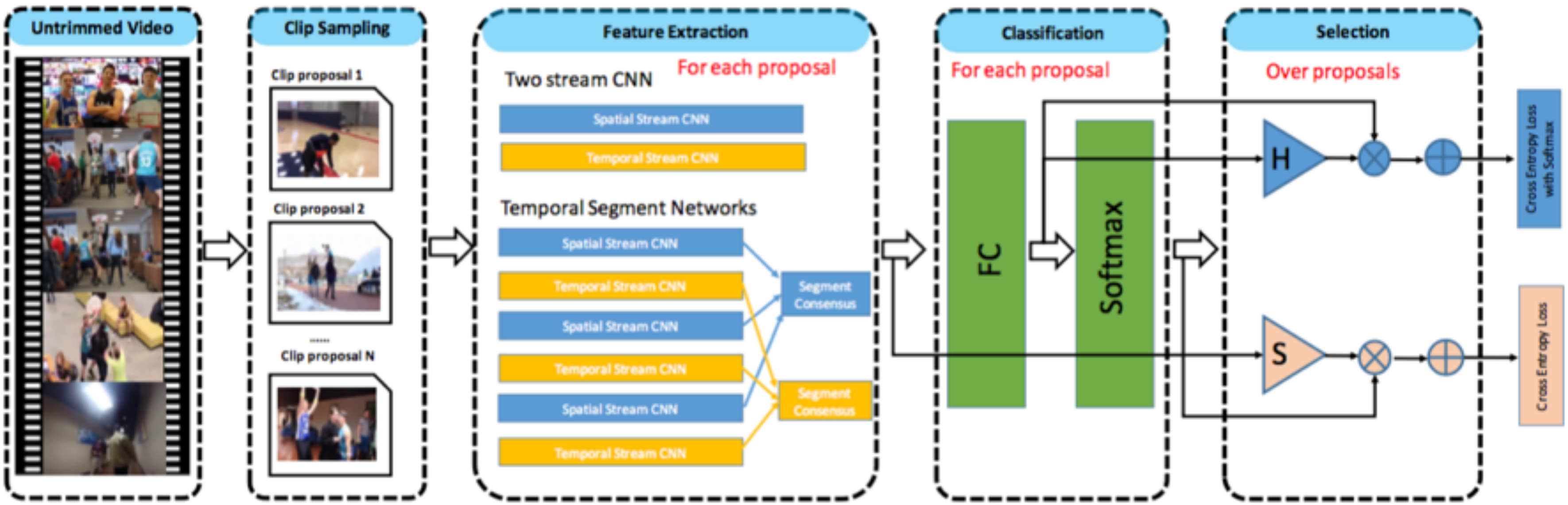
- [1] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition*, in ECCV, 2016.
- [2] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, *UntrimmedNets for Weakly Supervised Action Recognition and Detection*, in CVPR 2017.
- [3] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, D. Lin, and X. Tang, *Temporal Action Detection with Structured Segment Networks*, in ICCV 2017.

Motivation of UntrimmedNet

1. Labeling untrimmed video is expensive and time consuming
2. Temporal annotation is subjective and not consistent across persons and datasets



Overview of UntrimmedNet



Clip Proposal

- **Uniform Sampling**
  - Uniform sampling of fixed duration
- **Shot based Sampling**
  - First shot detection based HOG difference
  - For each shot, perform uniform sampling.

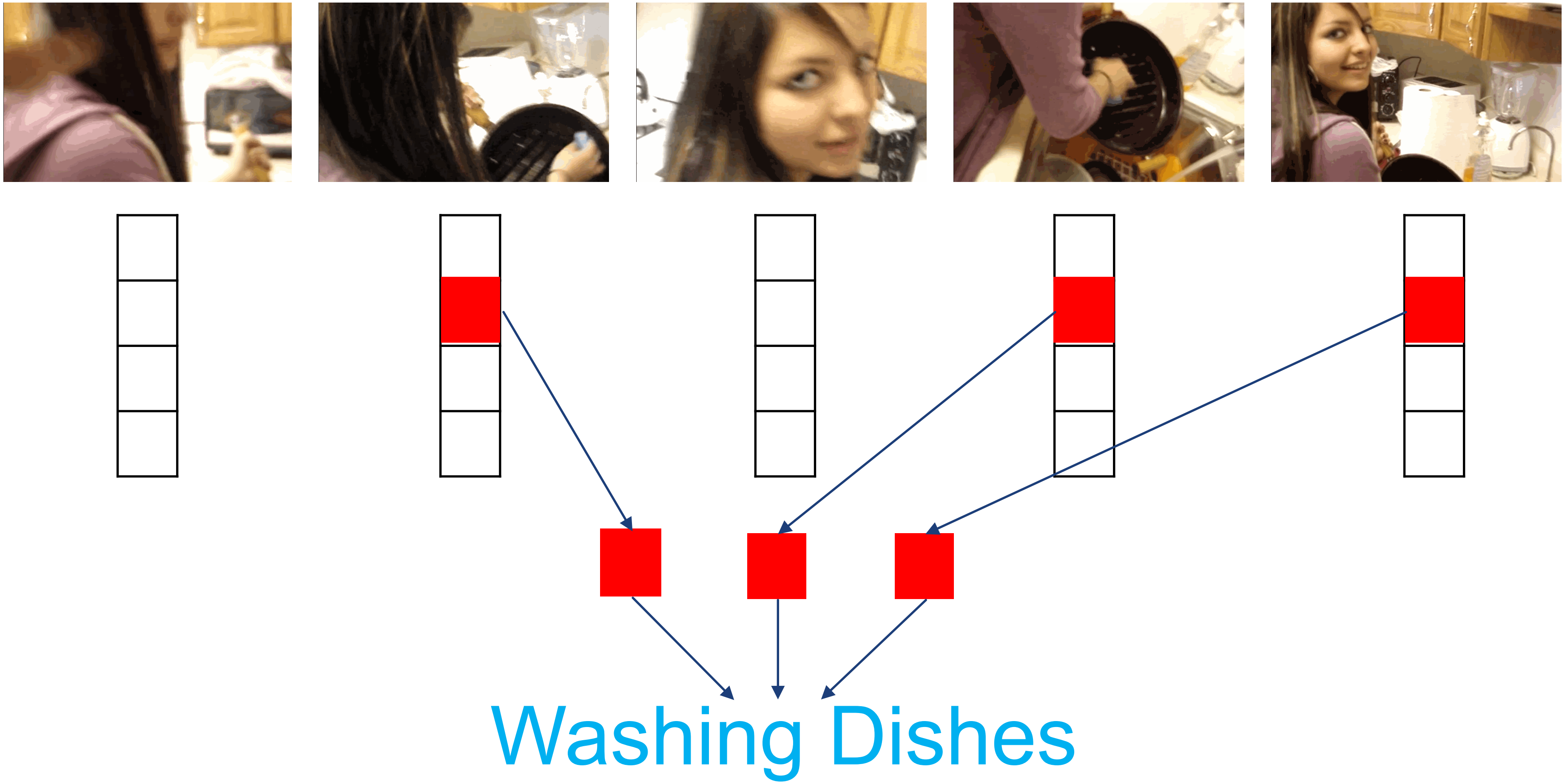
Clip Classification

- Following TSN framework:
  - Sampling a few snippets from each clip.
  - Aggregating snippet-level predictions with average pooling
- In practice, we use two stream input: RGB and Optical Flow

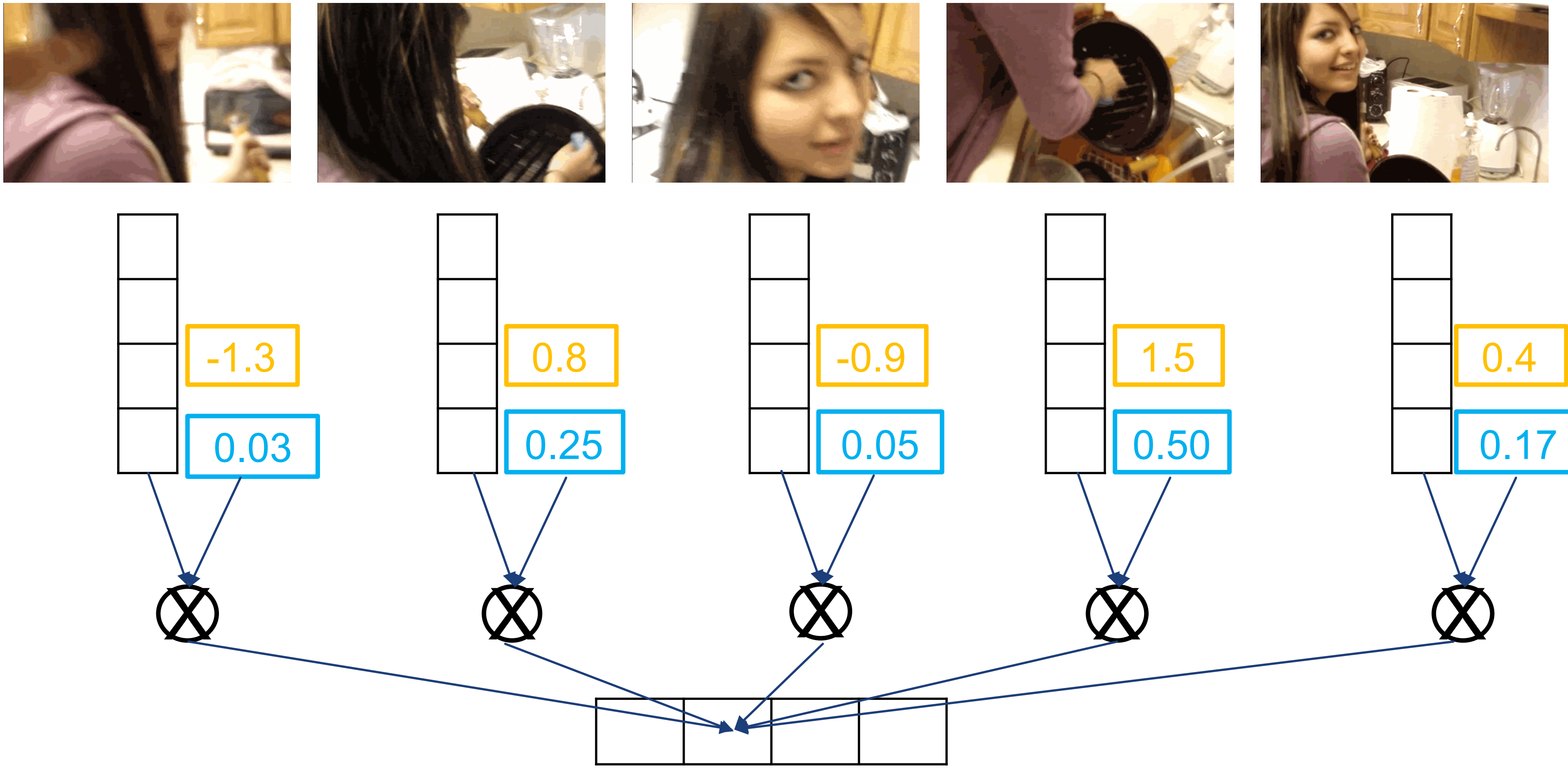
Clip Selection

- Selection aims to select discriminative clips or rank them with attention weights.
- Two selection methods:
  - Hard selection: top-k pooling over clip-level prediction
  - Soft selection: learning attention weights for different clips

Top-k Pooling



Attention weighting



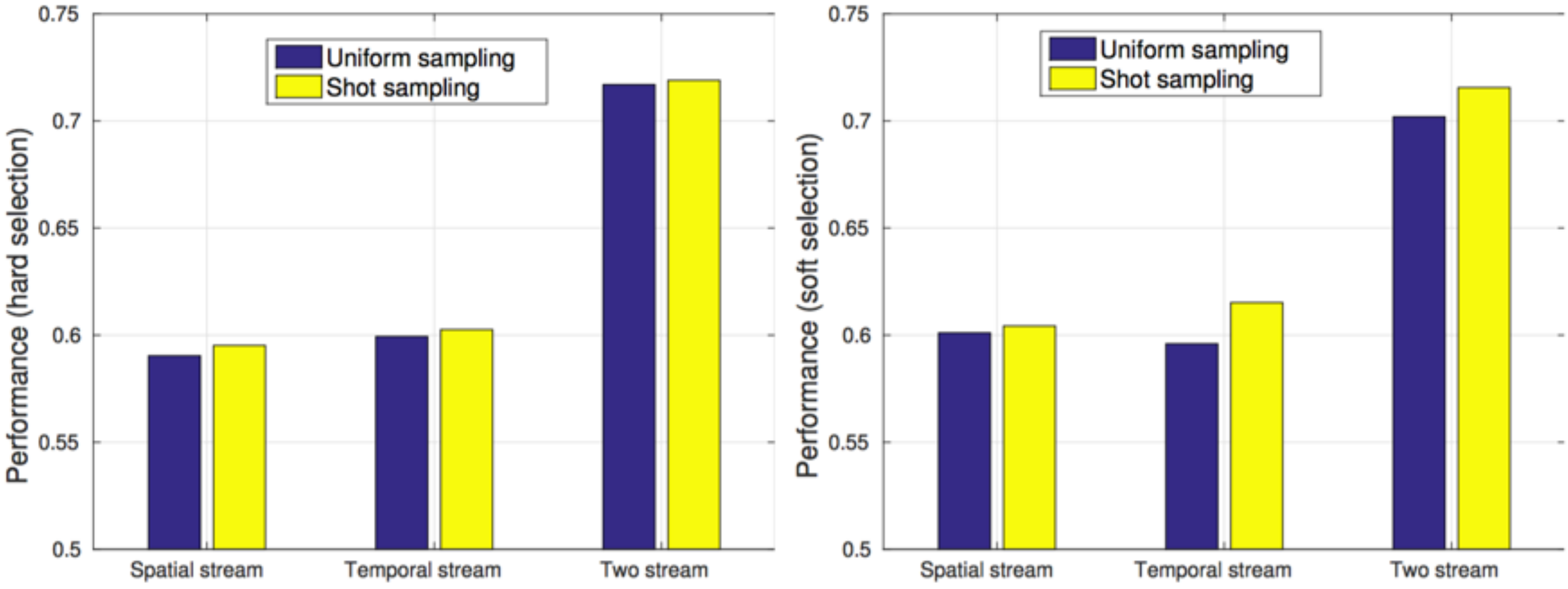
UntrimmedNet

- UntrimmedNet is an end-to-end learning architecture, combing three modules: feature extraction, classification module, selection module.
- Video-level prediction: a bilinear model over classification score and selection weights.
- The whole pipeline could be optimized with standard back propagation algorithm.

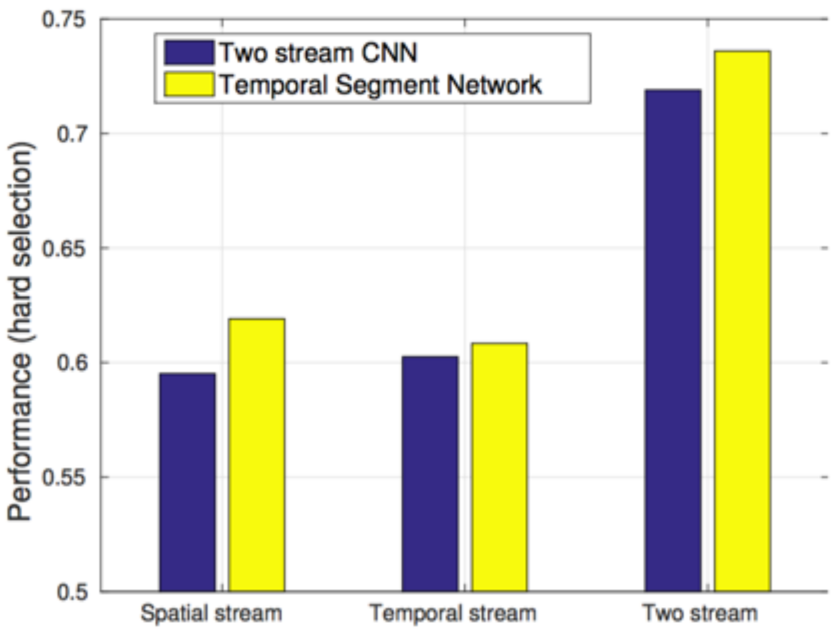
Weakly supervised AR and AD

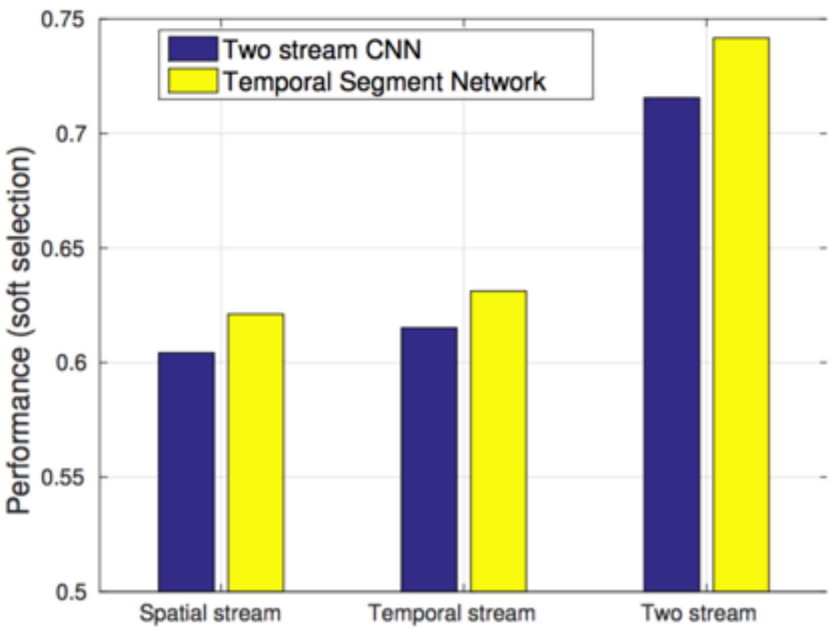
- **Action Recognition:**
  - In practice, we sample a single frame (or 5 frame stacking of optical flow) every 30 frames.
  - The recognition from sampled frames are aggregated with top-k pooling (k set to 20) to yield the final video-level prediction.
- **Action Detection:**
  - we sample frames every 15 frame and for each frame, we get both prediction scores and attention weights.
  - we remove background by thresholding (set to 0.0001) on the attention weights .
  - we produce the final detection results by thresholding (set to 0.5) on the classification scores.

Exploration Study



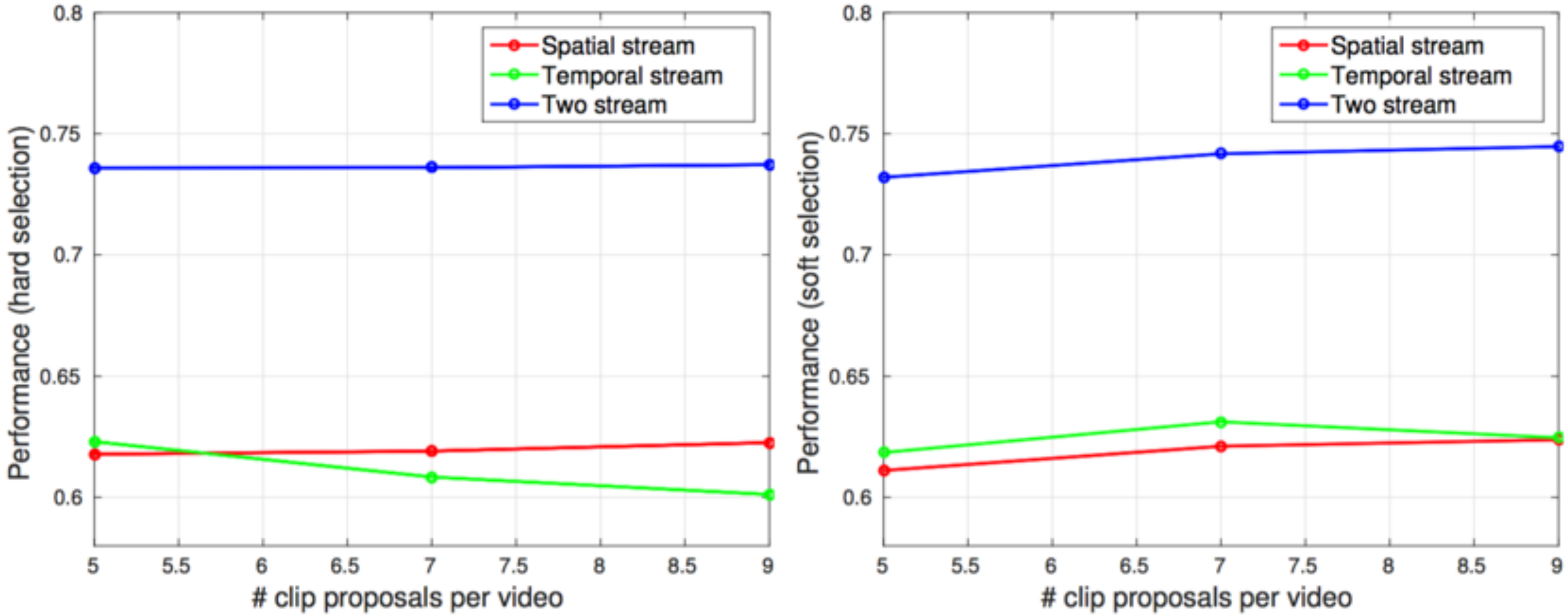
Exploration Study





ETH zürich

Exploration Study



Experiment Results -- Action recognition

Method	THUMOS14	ActivityNet (a)	ActivityNet (b)
TSN (3 seg) [37]	67.7%	85.0%	88.5%
TSN (21 seg)	68.5%	86.3%	90.5%
UntrimmedNet (h)	73.6%	**87.7%**	**91.3%**
UntrimmedNet (s)	**74.2%**	86.9%	90.9%

[1] H. Idrees et al., *The THUMOS Challenge on Action Recognition for Videos "in the Wild"*, in CVIU, 2017.
[2] F. C. Heilbron et al., *ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding*, in CVPR, 2015.

Experiment Results -- Action recognition

THUMOS14		ActivityNet	
iDT+FV [35]	63.1%	iDT+FV [35]	66.5%*
Two Stream [31]	66.1%	Two Stream [31]	71.9%*
EMV+RGB [42]	61.5%	C3D [33]	74.1%*
Objects+Motion [12]	71.6%	Depth2Action [43]	78.1%*
TSN [37]	78.5%	TSN [37]	88.8%*
UntrimmedNet (hard)	81.2%	UntrimmedNet (hard)	**91.3%**
UntrimmedNet (soft)	**82.2%**	UntrimmedNet (soft)	90.9%

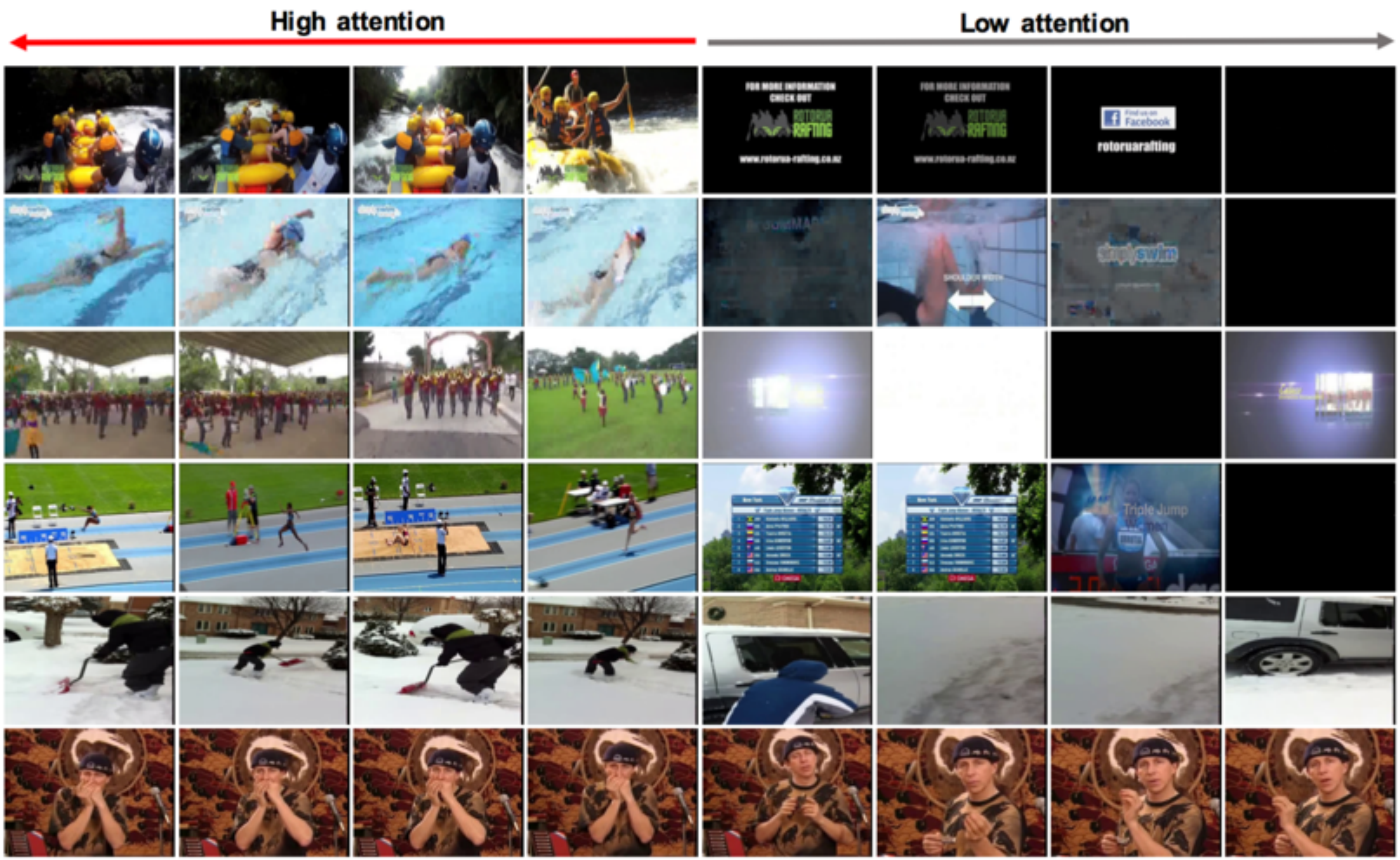
[1] H. Idrees et al., *The THUMOS Challenge on Action Recognition for Videos "in the Wild"*, in CVIU, 2017.
[2] F. C. Heilbron et al., *ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding*, in CVPR, 2015.

Experiment Results -- Action detection

	IoU = 0.5	IoU = 0.4	IoU = 0.3	IoU = 0.2	IoU = 0.1
Wang *et al.* [36]*	8.3	11.7	14.0	17.0	18.2
Oneata *et al.* [25]*	14.4	20.8	27.0	33.6	36.6
Richard *et al.* [26]*	15.2	23.2	30.0	35.7	39.7
Shou *et al.* [30]*	19.0	28.7	36.3	43.5	47.7
Yeung *et al.* [40]*	17.1	26.4	36.0	44.0	48.9
Yuan *et al.* [41]*	18.8	26.1	33.6	42.6	51.4
UntrimmedNet (soft)	13.7	21.1	28.2	37.7	44.4

[1] H. Idrees et al., *The THUMOS Challenge on Action Recognition for Videos "in the Wild"*, in CVIU, 2017.

Examples of Attention



Summary

- Temporal modeling is important for action understanding.
- Segment based sampling shares two properties: **global** and **sparse**.
- **TSN** is a general and flexible framework for action modeling.
- **SSN** extends TSN for action detection with context and structure modeling.
- **UntrimmedNet** extends TSN for weakly supervised setting with attention modeling.

Code and References

- **Temporal segment network:**

https://github.com/yjxiong/temporal-segment-network

- **Structured segment network:**

https://github.com/yjxiong/action-detection

- **UntrimmedNet:**

https://github.com/wanglimin/UntrimmedNet

- **Video Caffe:**

https://github.com/yjxiong/caffe

- [1] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition*, in ECCV, 2016.
- [2] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, *UntrimmedNets for Weakly Supervised Action Recognition and Detection*, in CVPR 2017.
- [3] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, D. Lin, and X. Tang, *Temporal Action Detection with Structured Segment Networks*, in ICCV 2017.

ETH zürich

Collaborators