Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

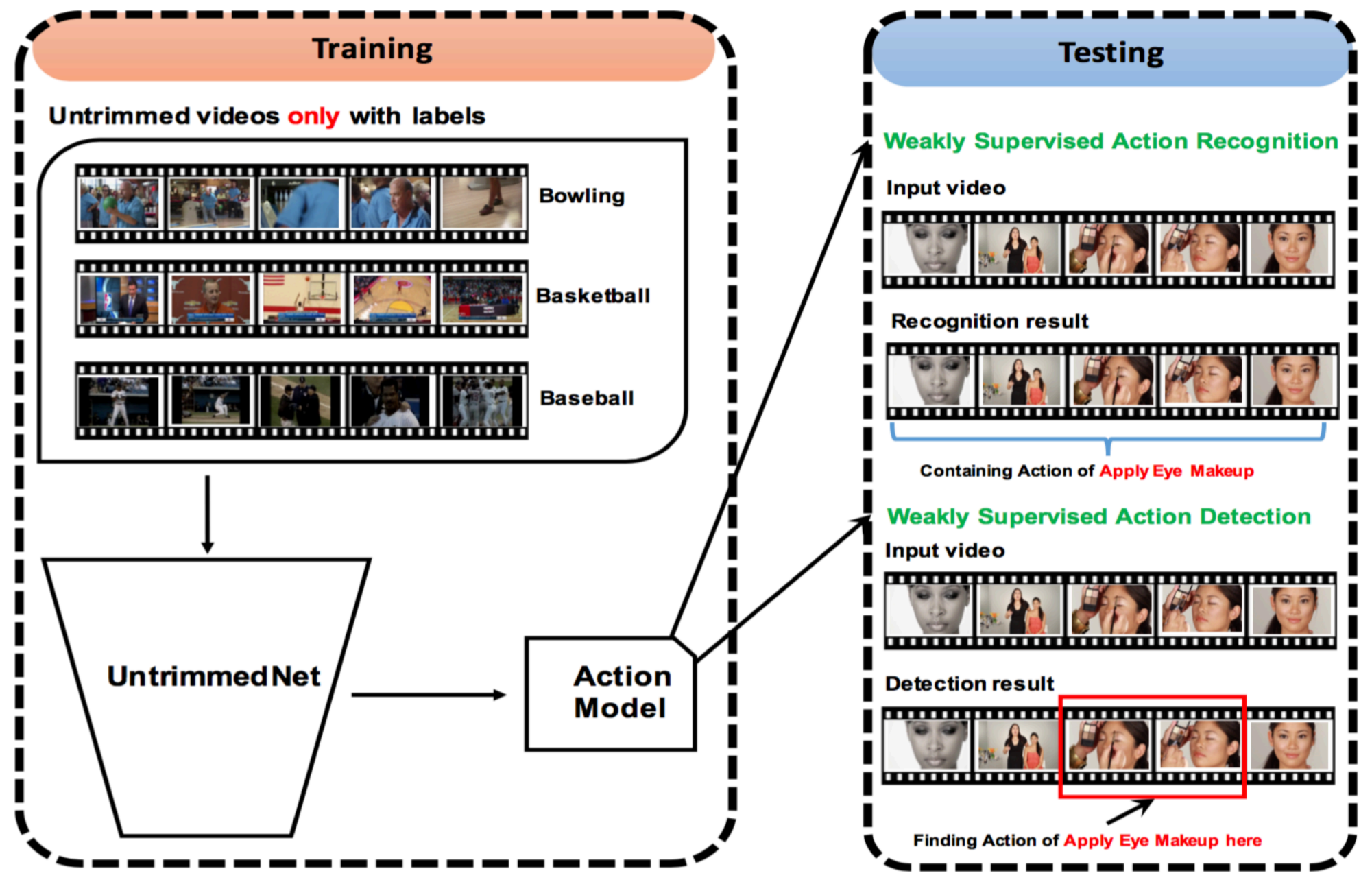
UntrimmedNets for Weakly Supervised Action Recognition and Detection

Limin Wang, Yuanjun Xiong, Dahua Lin, Luc Van Gool

ETH Zurich The Chinese University of Hong Kong

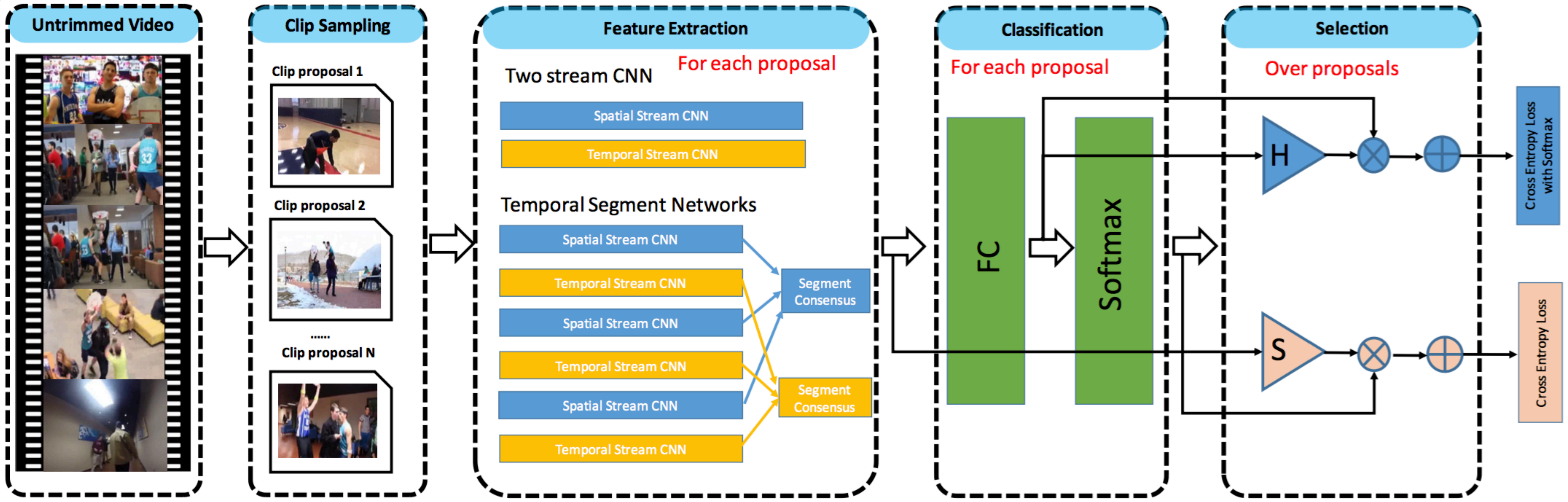
Motivation: weakly supervised action recognition and detection

- Action recognition: training on trimmed videos
- Temporal annotation: expensive and subjective
- Large numbers of videos are untrimmed in nature.
- **Learning directly from untrimmed video without temporal annotations**.


Training
Untrimmed videos only with labels
Bowling
Basketball
Baseball
UntrimmedNet
Action Model
Testing
Weakly Supervised Action Recognition
Input video
Recognition result
Containing Action of Apply Eye Makeup
Weakly Supervised Action Detection
Input video
Detection result
Finding Action of Apply Eye Makeup here

Code available

UntrimmedNet: learning from untrimmed videos


Untrimmed Video

Clip Sampling

Clip proposal 1

Clip proposal 2

......

Clip proposal N

Feature Extraction

Two stream CNN

For each proposal

Spatial Stream CNN

Temporal Stream CNN

Temporal Segment Networks

Spatial Stream CNN

Temporal Stream CNN

Spatial Stream CNN

Temporal Stream CNN

Spatial Stream CNN

Temporal Stream CNN

Segment Consensus

Segment Consensus

Classification

For each proposal

FC

Softmax

Selection

Over proposals

H

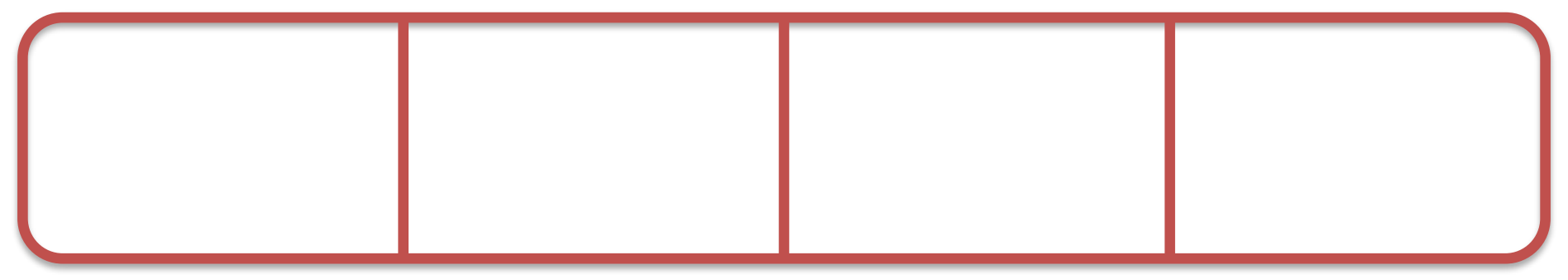
S

Cross Entropy Loss with Softmax

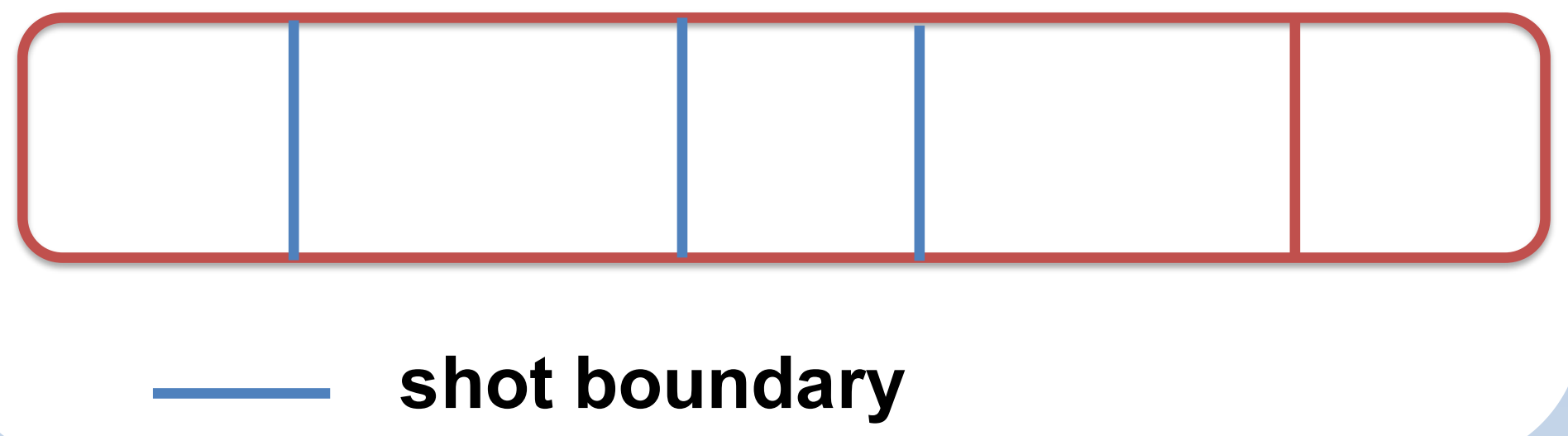
Cross Entropy Loss

Clip sampling

- Uniform sampling



- Shot based sampling


shot boundary

Clip classification

- Two stream CNNs
- Temporal Segment Networks

$$\mathbf{x}^c(c) = \mathbf{W}^c\phi(c)$$

$$\bar{x}_i^c(c) = \frac{\exp(x_i^c(c))}{\sum_{k=1}^{C}\exp(x_k^c(c))}$$

Clip selection

$$x_i^s(c_j) = \delta(j \in S_i^k)$$

$$x_i^p(V) = \sum_{n=1}^{N} x_i^s(c_n) x_i^c(c_n),$$

$$\bar{x}_i^p(V) = \frac{\exp(x_i^r(V))}{\sum_{k=1}^{C}\exp(x_k^r(V))},$$

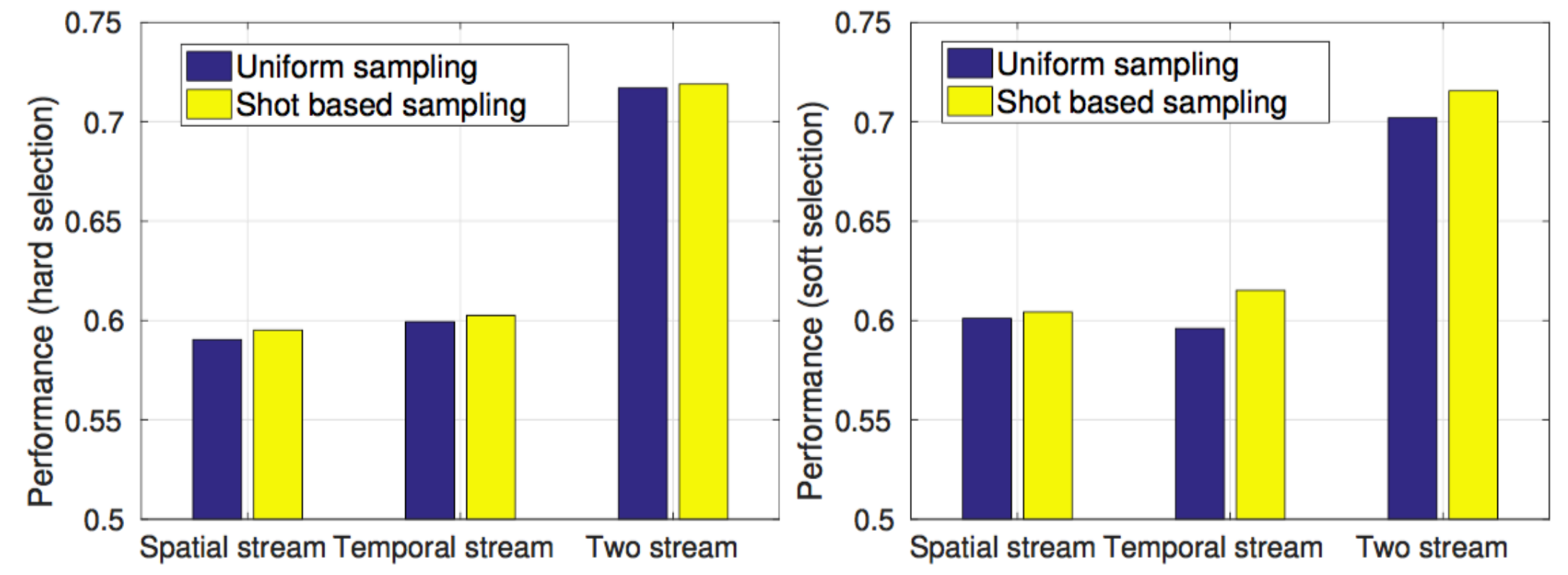
$$x^s(c) = \mathbf{w}^{sT}\phi(c)$$

$$\bar{x}^s(c_i) = \frac{\exp(x^s(c_i))}{\sum_{n=1}^{N}\exp(x^s(c_n))}$$

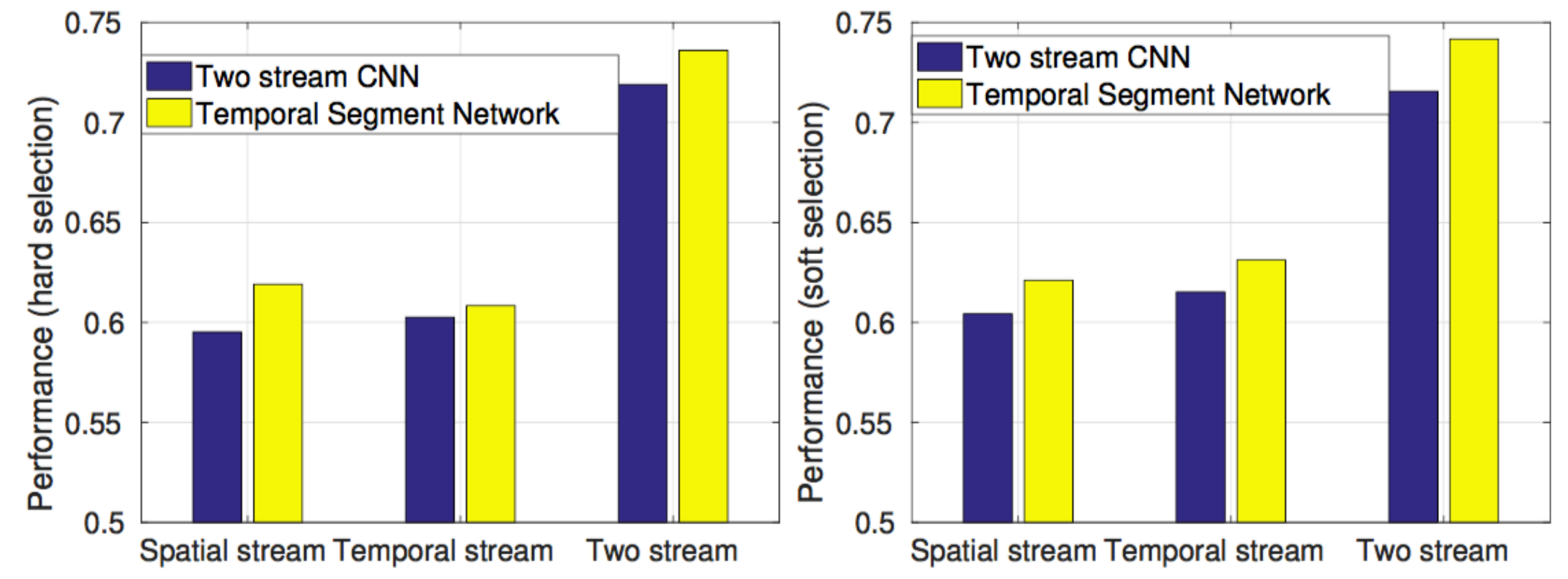
$$\bar{\mathbf{x}}^p(V) = \sum_{n=1}^{N} \bar{x}^s(c_n)\bar{\mathbf{x}}^c(c_n).$$

Exploration study

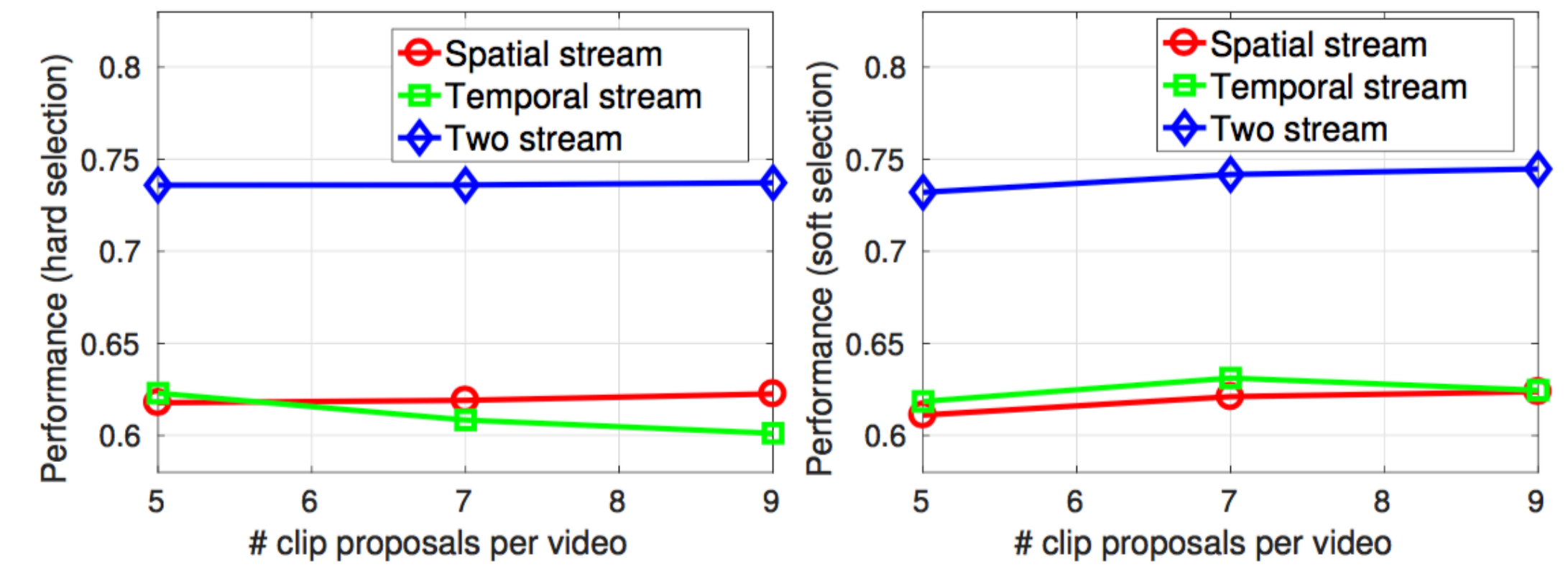
- **Study on sampling method**



- **Study on clip classifier**



- **Study on proposal number**



Comparisons

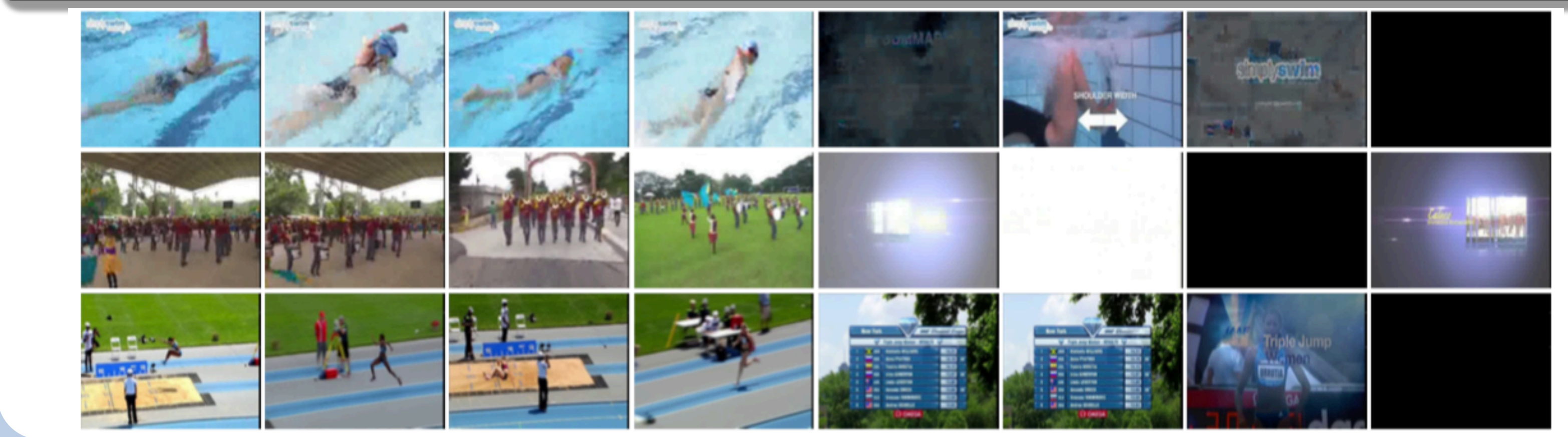
Method	THUMOS14	ActivityNet (a)	ActivityNet (b)
TSN (3 seg) [50]	67.7%	85.0%	88.5%
TSN (21 seg)	68.5%	86.3%	90.5%
UntrimmedNet (hard)	73.6%	**87.7%**	**91.3%**
UntrimmedNet (soft)	**74.2%**	86.9%	90.9%

IoU (α)	$\alpha = 0.5$	$\alpha = 0.4$	$\alpha = 0.3$	$\alpha = 0.2$	$\alpha = 0.1$
Oneata *et al.* [33]*	14.4	20.8	27.0	33.6	36.6
Richard *et al.* [35]*	15.2	23.2	30.0	35.7	39.7
Shou *et al.* [39]*	19.0	28.7	36.3	43.5	47.7
Yeung *et al.* [54]*	17.1	26.4	36.0	44.0	48.9
Yuan *et al.* [55]*	18.8	26.1	33.6	42.6	51.4
UntrimmedNet (soft)	13.7	21.1	28.2	37.7	44.4

THUMOS14		ActivityNet	
iDT+FV [45]	63.1%	iDT+FV [45]	66.5%*
Two Stream [40]	66.1%	Two Stream [40]	71.9%*
EMV+RGB [56]	61.5%	C3D [42]	74.1%*
Objects+Motion [19]	71.6%	Depth2Action [57]	78.1%*
TSN (3 seg) [50]	78.5%	TSN (3 seg) [50]	88.8%*
UntrimmedNet (hard)	81.2%	UntrimmedNet (hard)	**91.3%**
UntrimmedNet (soft)	**82.2%**	UntrimmedNet (soft)	90.9%

Evaluation on datasets of THUMOS14 and ActivityNet 1.2

Examples



References:
[1] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS 2014.
[2] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In ECCV 2016.

Limin Wang, Yuanjun Xiong, Dahua Lin, Luc Van Gool IEEE Conference on Computer Vision and Pattern Recognition 2017.