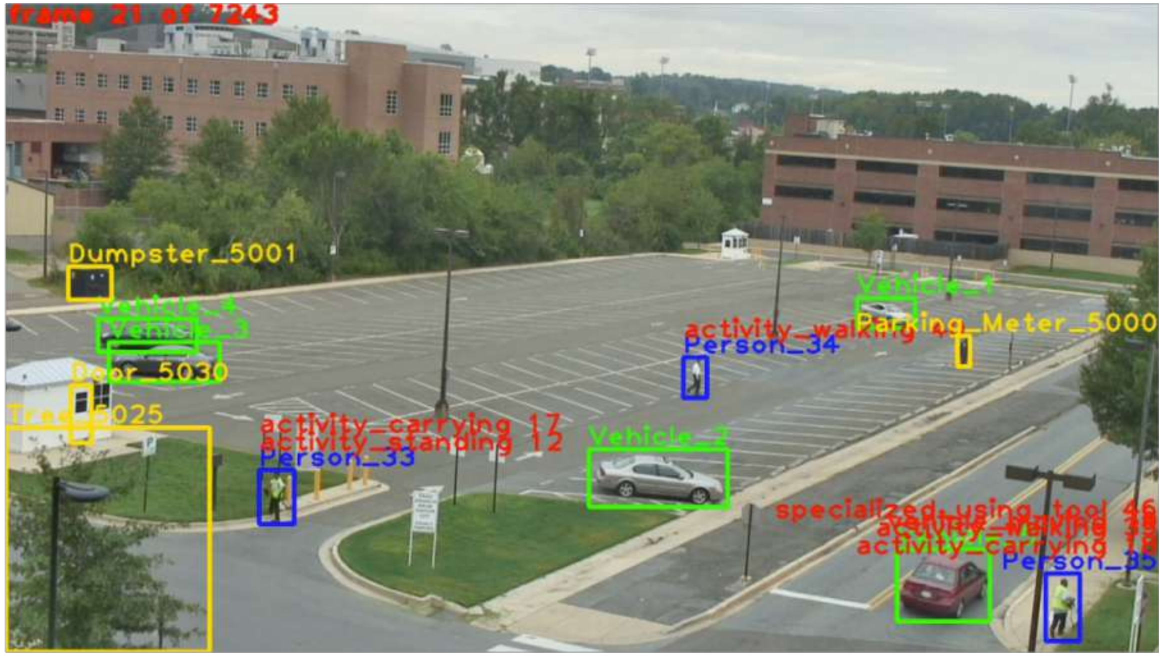
Detecting Activities with Less

Cees Snoek

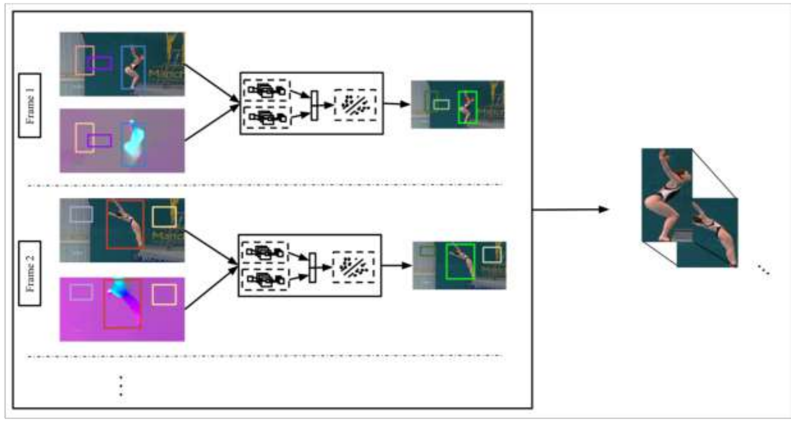
UNIVERSITY OF AMSTERDAM
QUVA Deep Vision Lab
ICAI Innovation Center for Artificial Intelligence

Goal: activity understanding

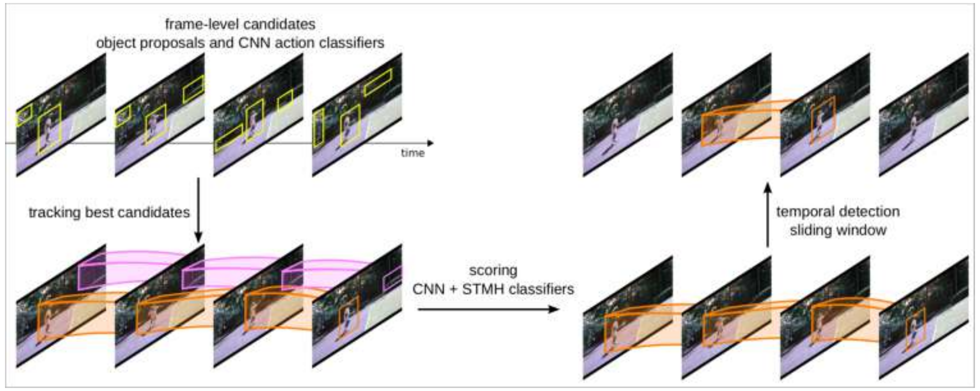


Spatio-temporal localization is key.

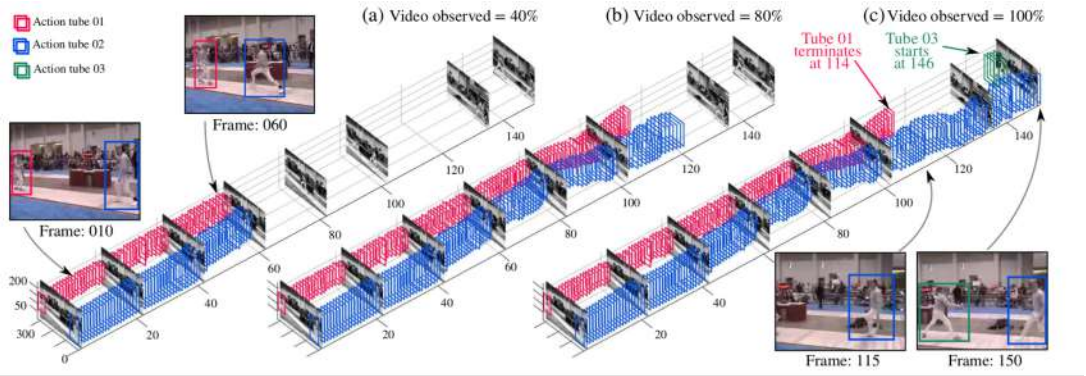
Prior art: box-supervised RGB and flow streams



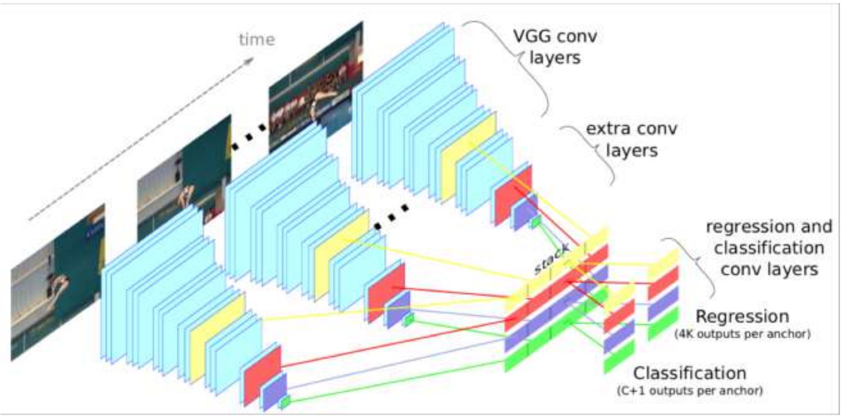
G. Gkioxari and J. Malik, CVPR, 2015.



P. Weinzaepfel et al, ICCV, 2015.



G. Singh et al, ICCV, 2017.



V. Kalogeiton et al, ICCV, 2017.

This talk

i. Detecting activities with less supervision

ii. Detecting activities with less streams

I.

Less supervision

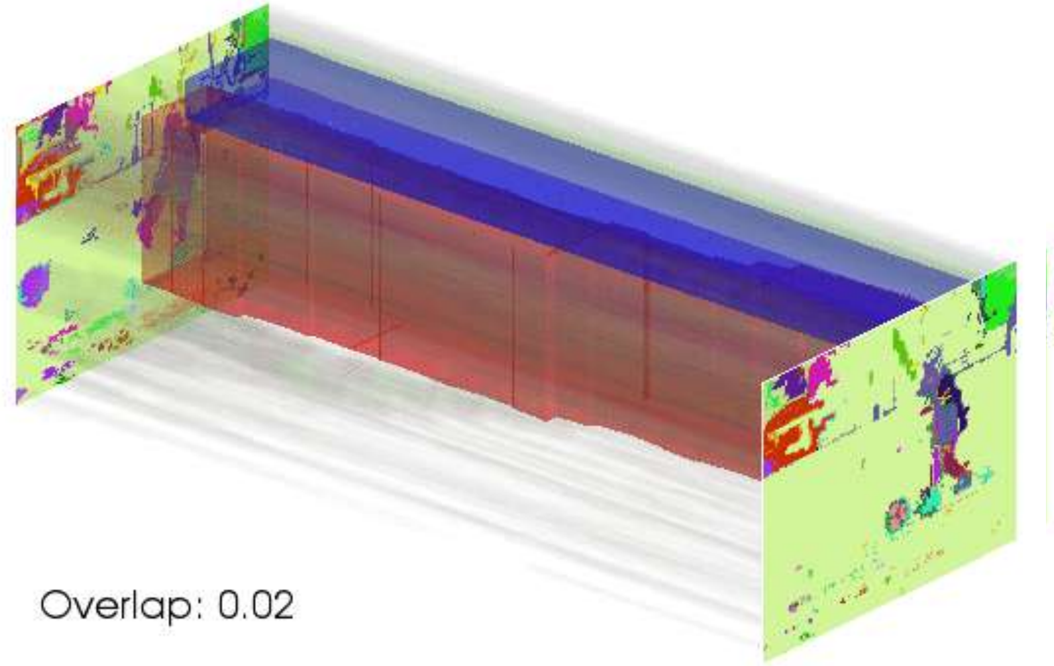
Pointly-Supervised Action Localization
Pascal Mettes and Cees Snoek. IJCV 2019.

Related work: unsupervised action proposals

Analyze space and time jointly to obtain action proposals

Action-class agnostic, covers variable aspect ratios and temporal lengths

High recall with few proposals



Jain *et al.*, CVPR 2014 / IJCV 2017

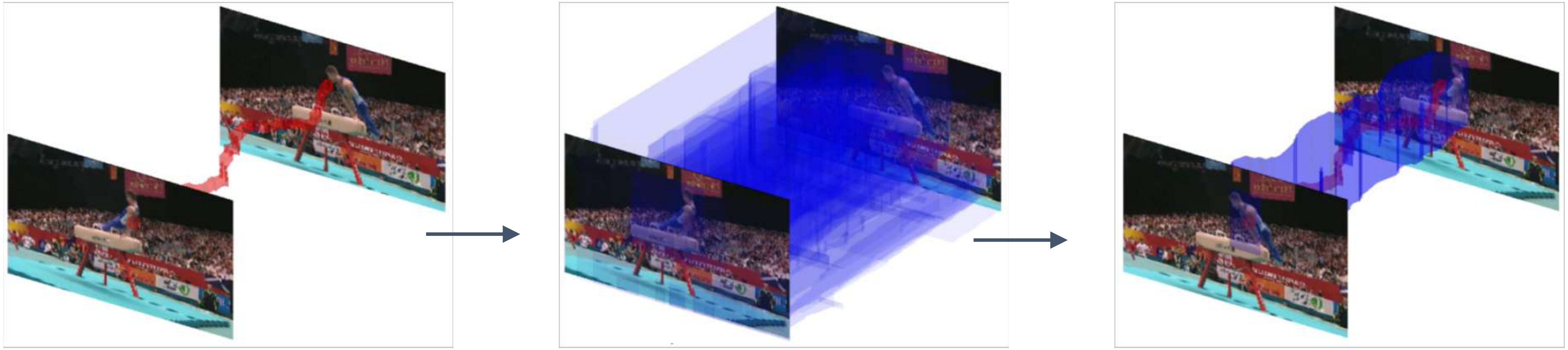
Oneata *et al.*, ECCV 2014

Gemert *et al.*, BMVC 2015

Idea: exploit proposals during training

Training on bounding boxes not required.

Training on action proposals with point annotations is as effective.

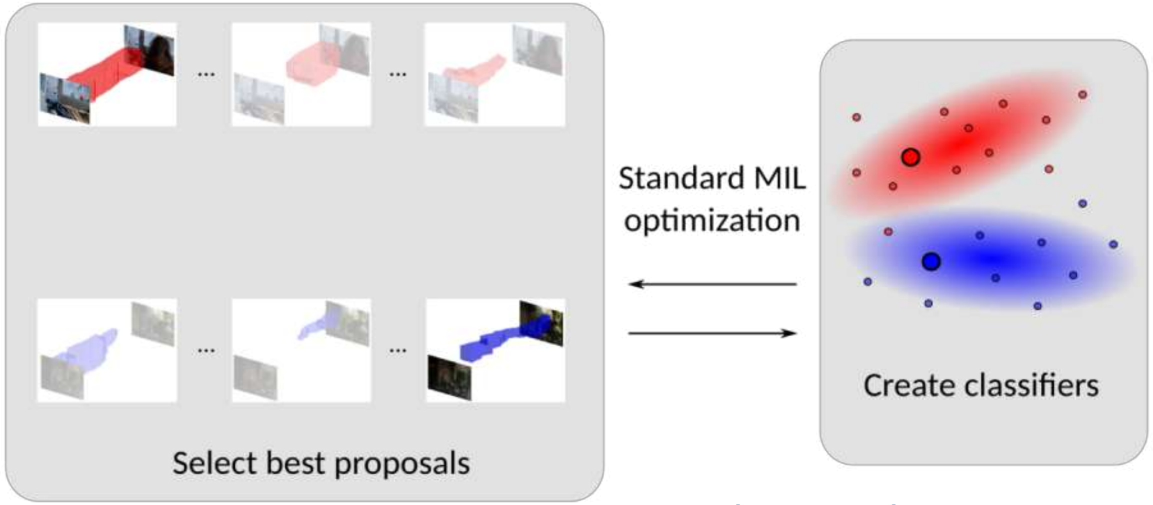


Human point supervision **Compute proposal affinity** **Mine best proposal**

Mining the best proposals

Train action classifiers using best proposals.

Cast as a Multiple Instance Learning problem.

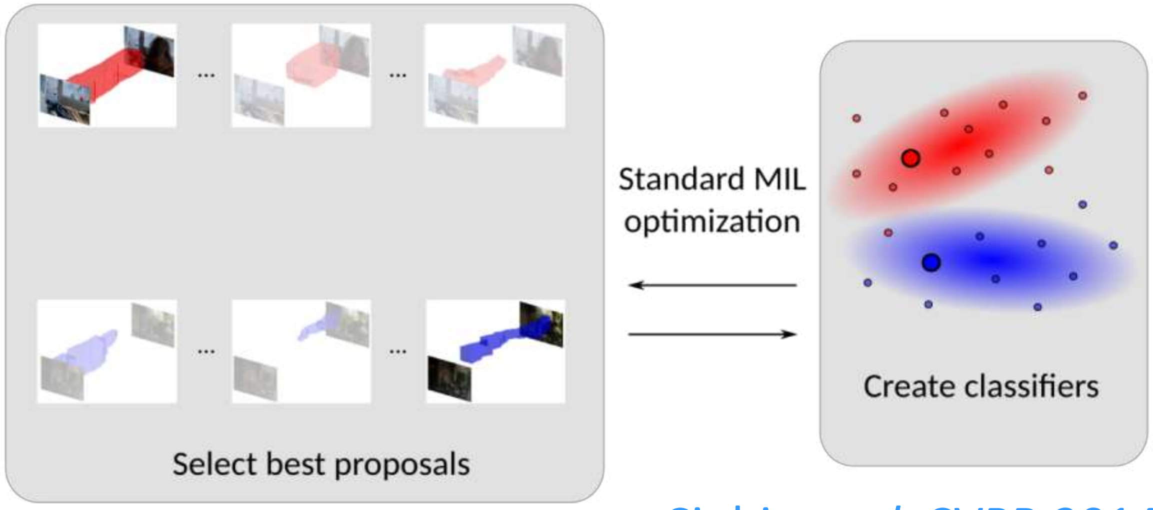


Cinbis et al. CVPR 2014

Mining the best proposals

Train action classifiers using best proposals.

Cast as a Multiple Instance Learning problem.

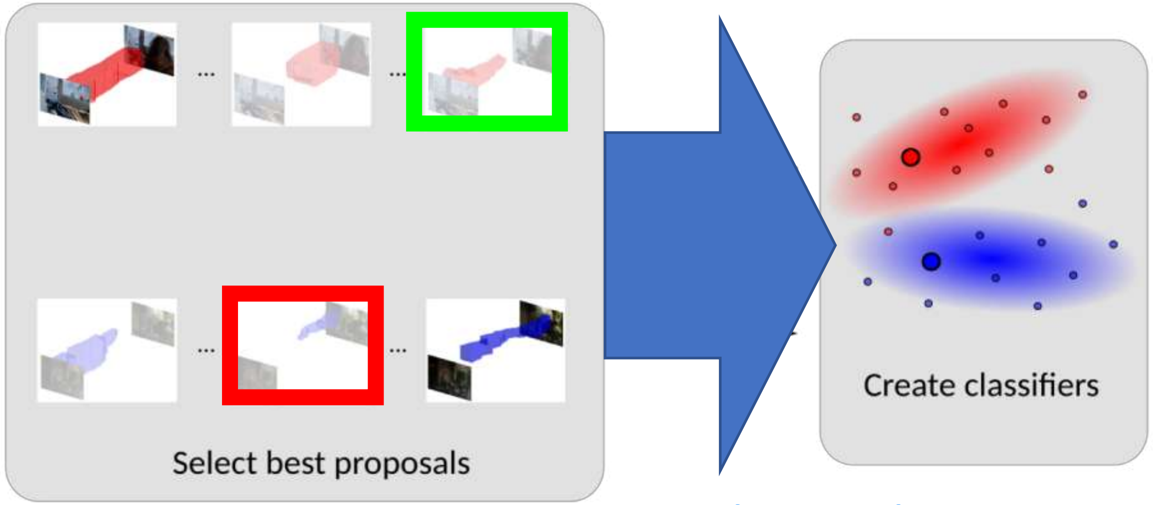


Cinbis *et al.* CVPR 2014

Mining the best proposals

Train action classifiers using best proposals.

Cast as a Multiple Instance Learning problem.

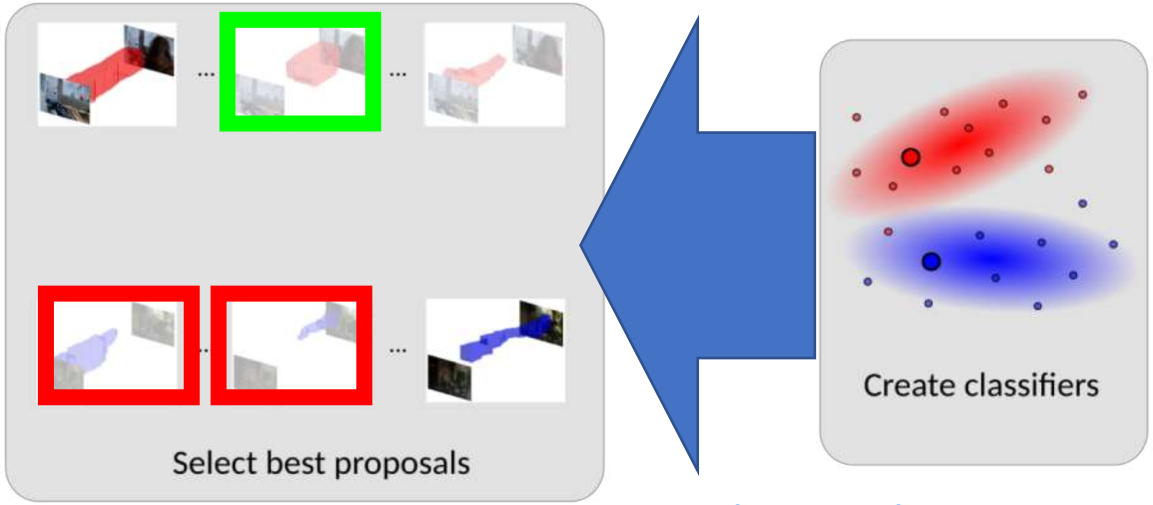


Cinbis *et al.* CVPR 2014

Mining the best proposals

Train action classifiers using best proposals.

Cast as a Multiple Instance Learning problem.

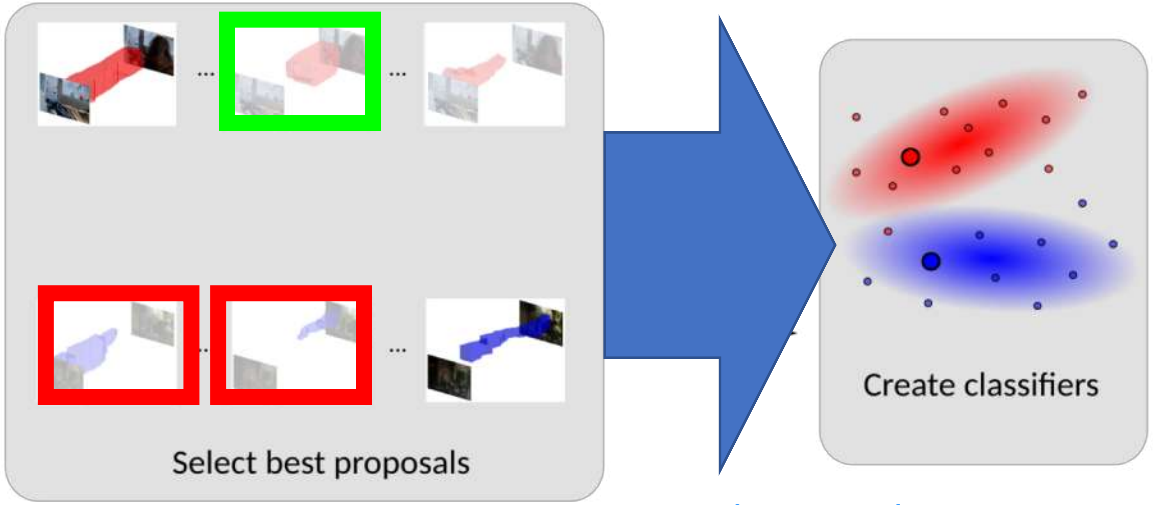


Cinbis *et al.* CVPR 2014

Mining the best proposals

Train action classifiers using best proposals.

Cast as a Multiple Instance Learning problem.

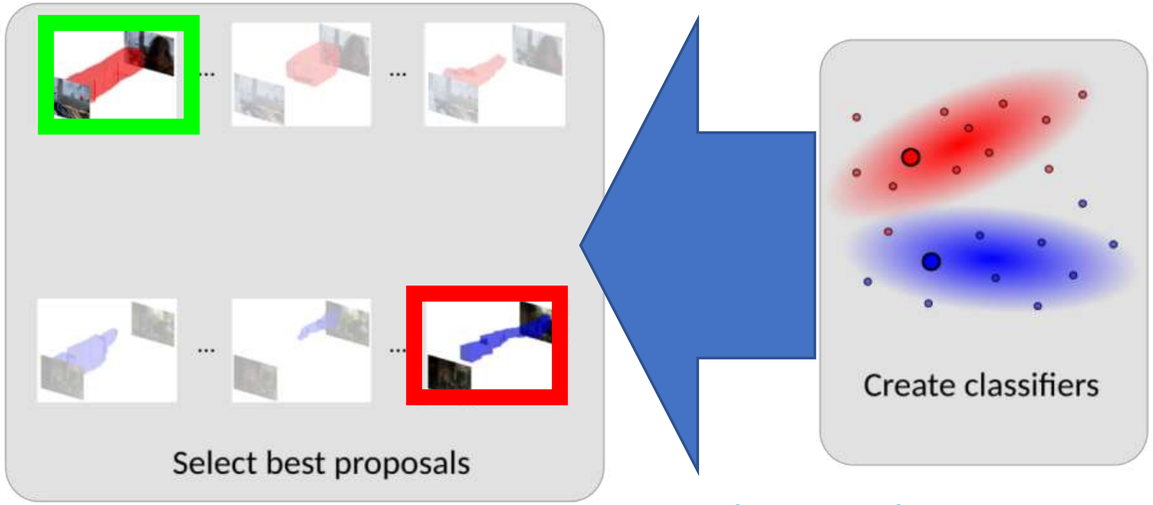


Cinbis *et al.* CVPR 2014

Mining the best proposals

Train action classifiers using best proposals.

Cast as a Multiple Instance Learning problem.

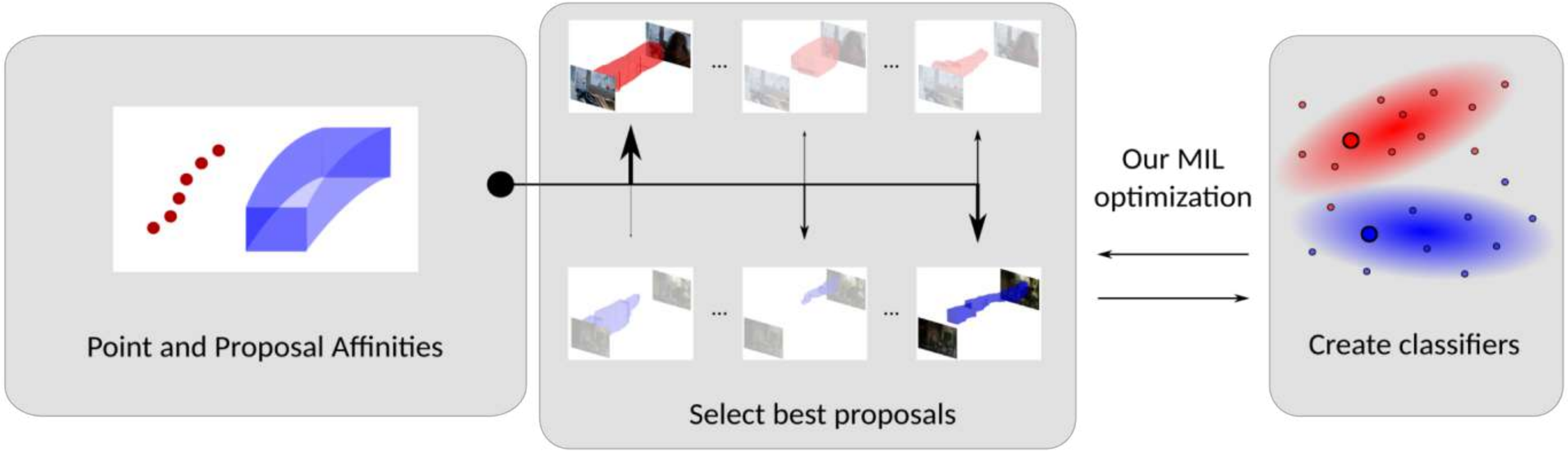


Cinbis *et al.* CVPR 2014

Idea: guide selection by point-supervision

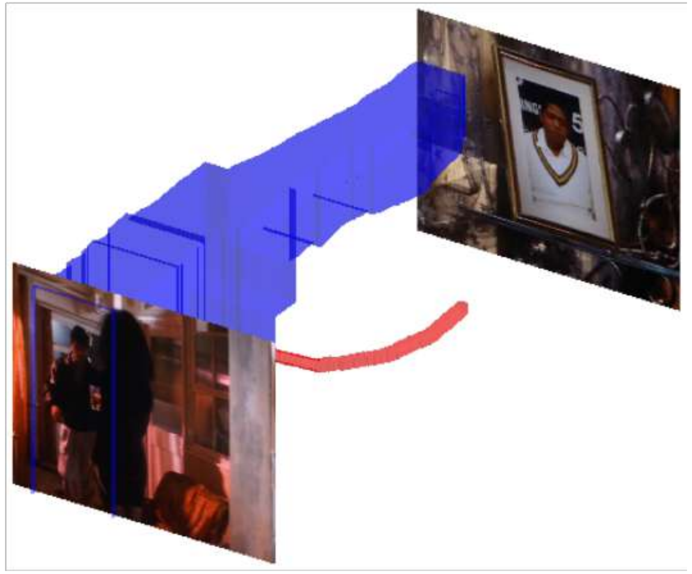
Train action classifiers using best proposals.

Cast as a Multiple Instance Learning problem.

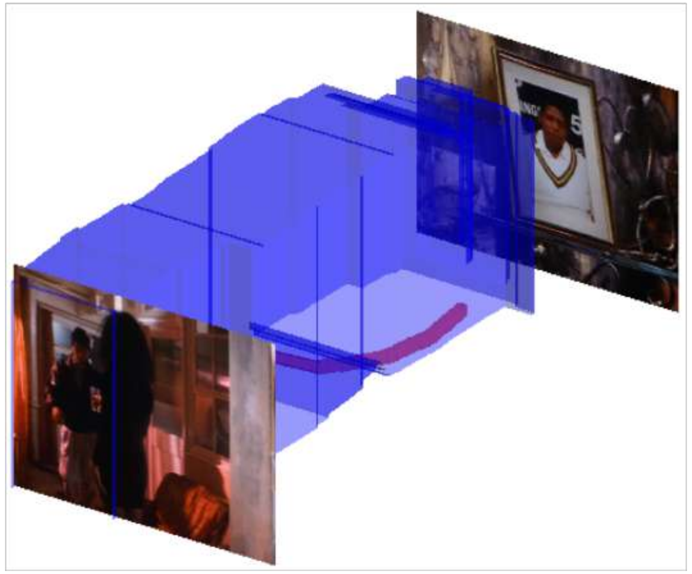


Proposal affinity

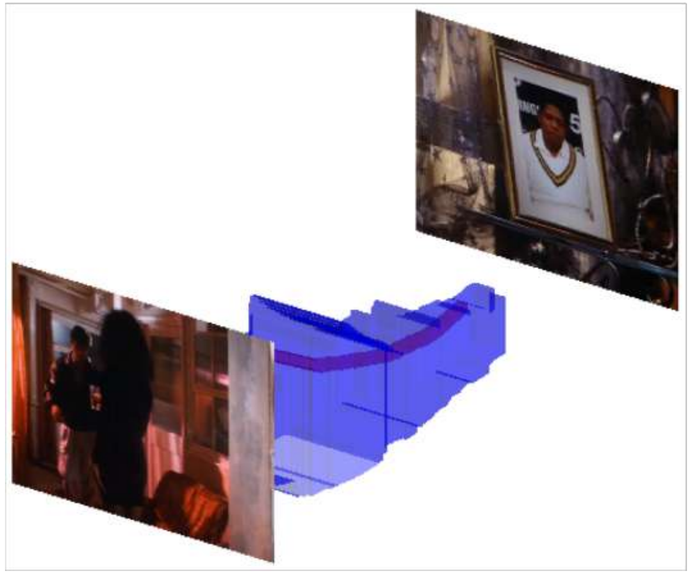
Novel overlap measure between point annotations and proposals.



No overlap



Small overlap

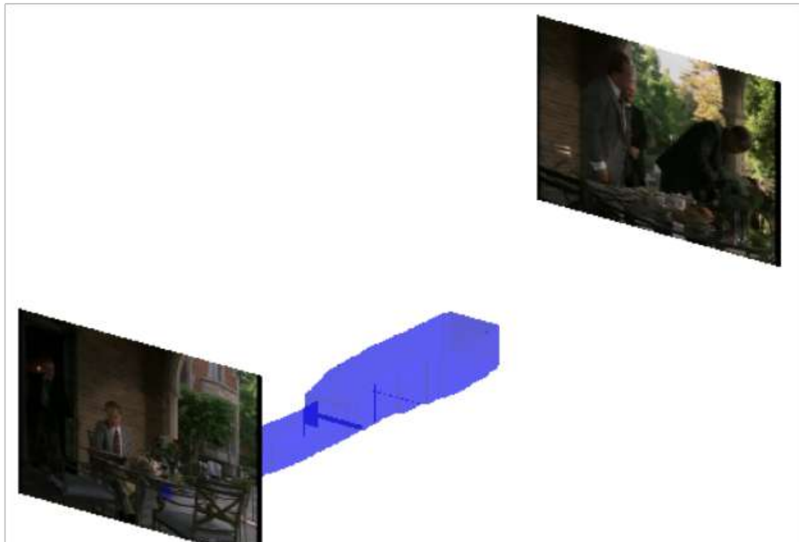


High overlap

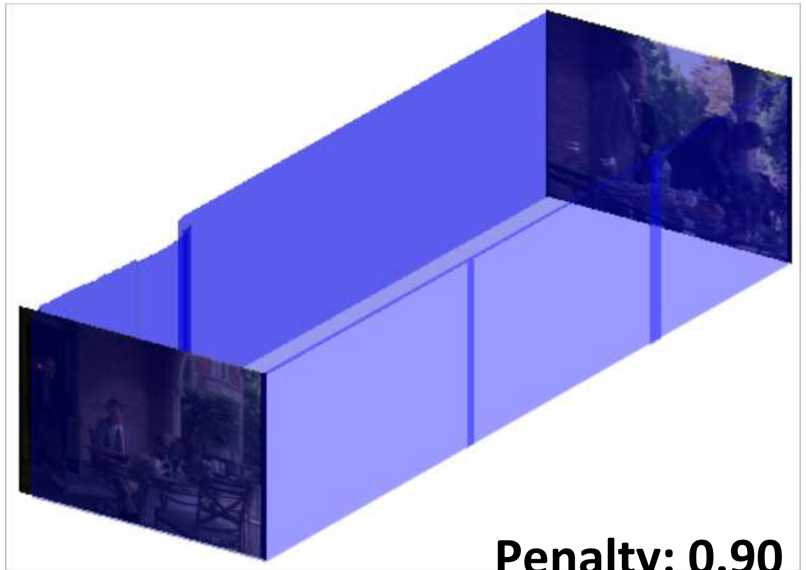
Mind the centre bias

Subtract the size of the proposal from the match.

To alleviate center bias of large proposals.



Penalty: 0.05



Penalty: 0.90

Action localization optimization

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_i \xi_i,$$

$$\text{s.t.} \quad \forall_i : Y_i \cdot (\mathbf{w} \cdot \arg\max_{\mathbf{z} \in X_i} S(\mathbf{z}|\mathbf{w}, b, P)) \geq 1 - \xi_i, \quad \forall_i : \xi_i \geq 0$$

Action localization optimization

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_i \xi_i,$$

Max-margin objective to separate top proposals of positive examples from negative examples.

$$\text{s.t.} \quad \forall_i : Y_i \cdot (\mathbf{w} \cdot \arg\max_{\mathbf{z} \in X_i} S(\mathbf{z}|\mathbf{w}, b, P)) \geq 1 - \xi_i, \quad \forall_i : \xi_i \geq 0$$

Action localization optimization

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_i \xi_i,$$

$$\text{s.t.} \quad \forall_i : Y_i \cdot (\mathbf{w} \cdot \arg\max_{\mathbf{z} \in X_i} \boxed{S(\mathbf{z}|\mathbf{w}, b, P)}) \geq 1 - \xi_i, \quad \forall_i : \xi_i \geq 0$$

Max-margin objective to separate top proposals of positive examples from negative examples.

Select top proposal per video using **likelihood** from current classifier and **prior** from point annotation overlaps.

Experiments

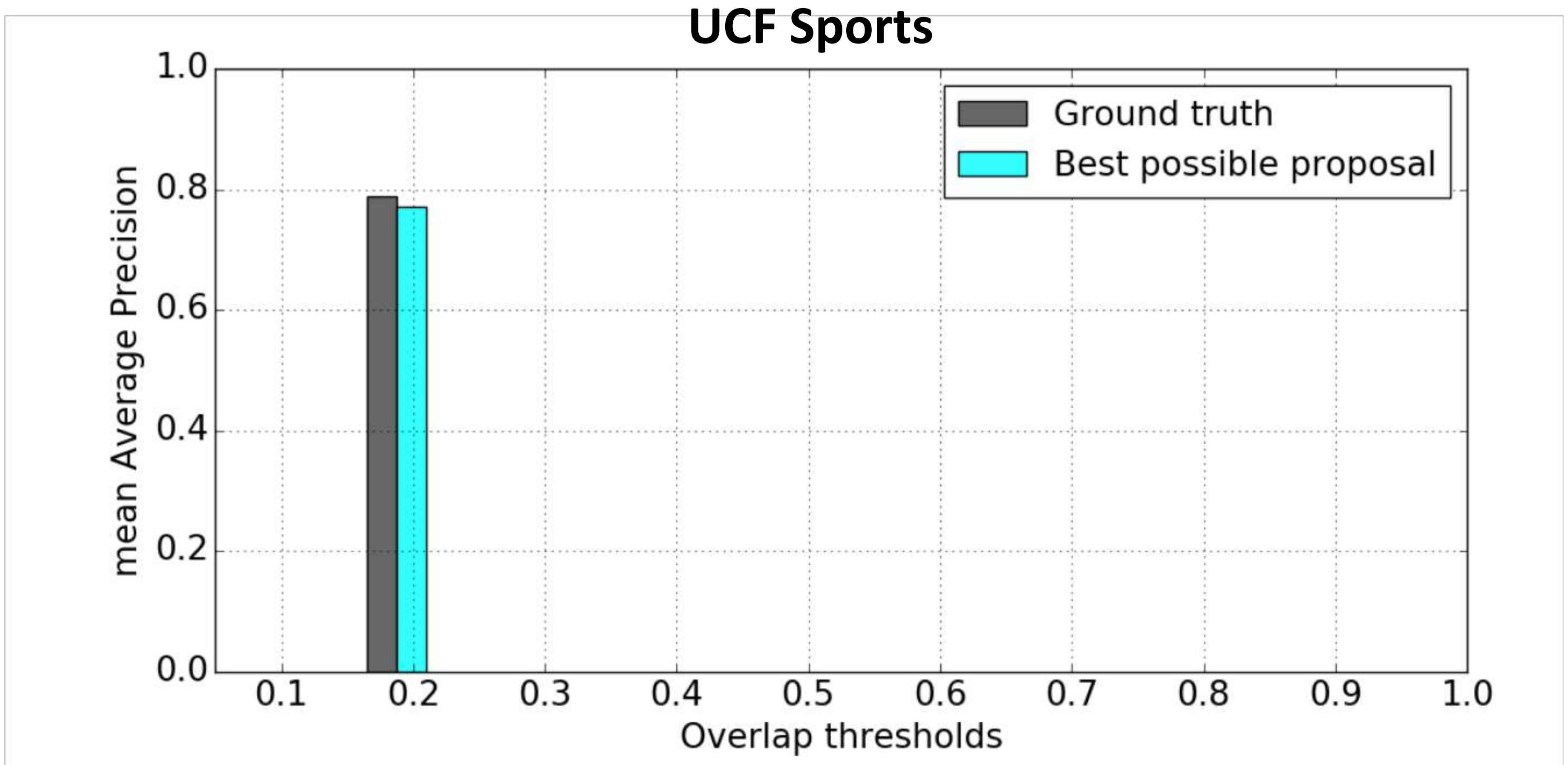
UCF Sports

THUMOS13

Unsupervised proposals from clustered trajectory features.
Evaluated with Fisher Vectors and SVMs.

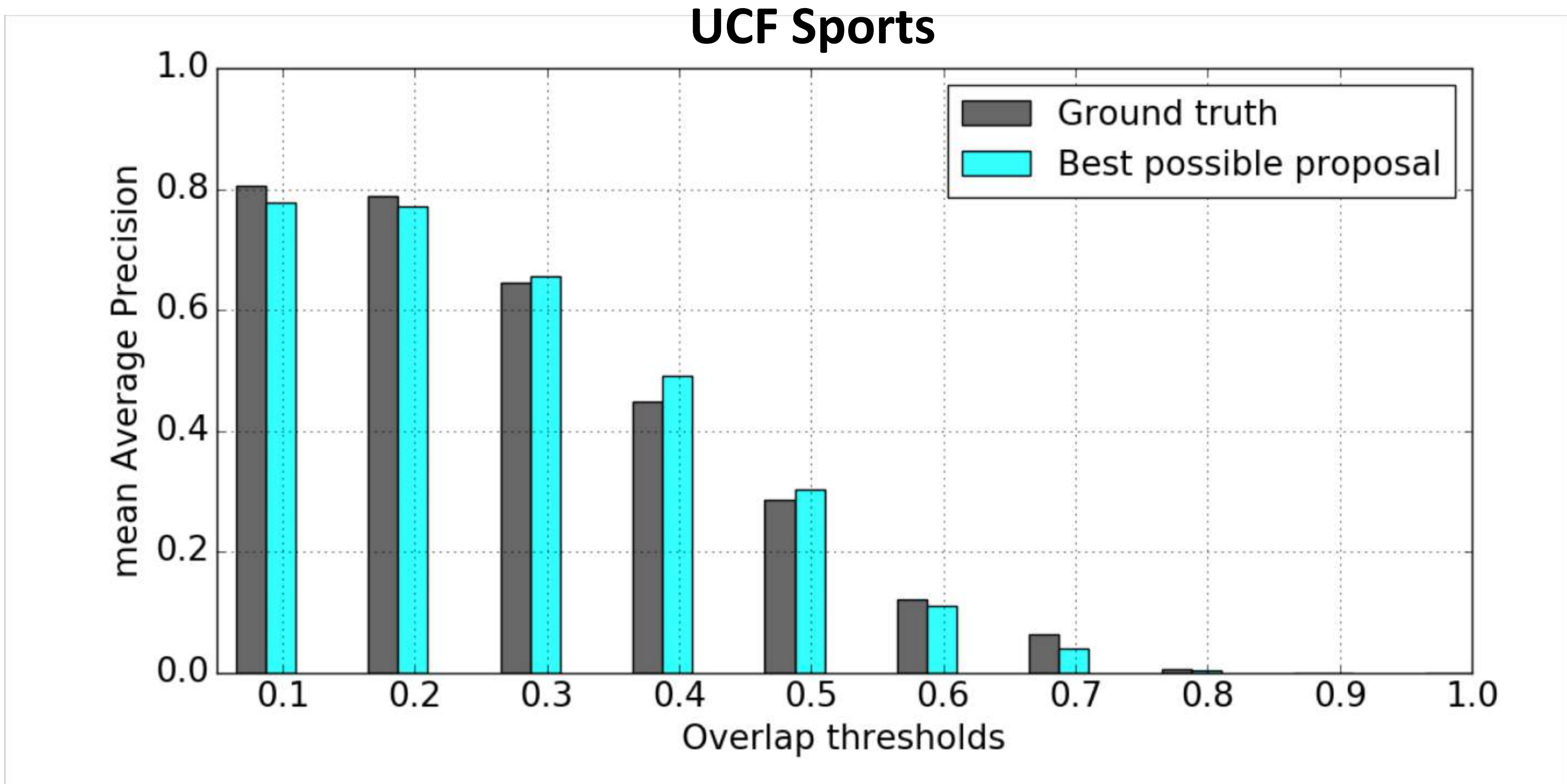
van Gemert *et al.* BMVC 2015

Training without ground truth boxes



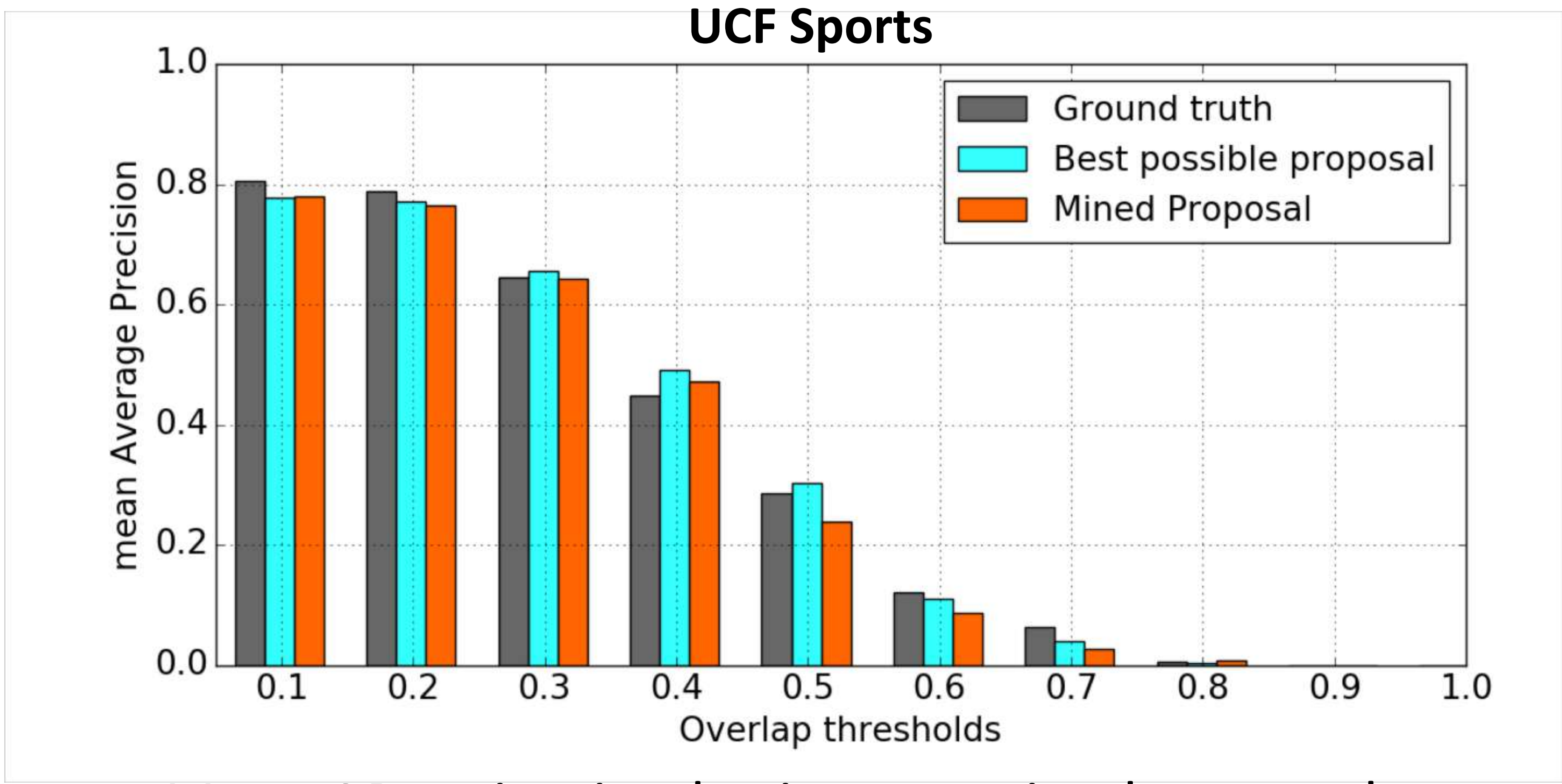
Best possible proposal performs as good as ground truth tube.

Training without ground truth boxes



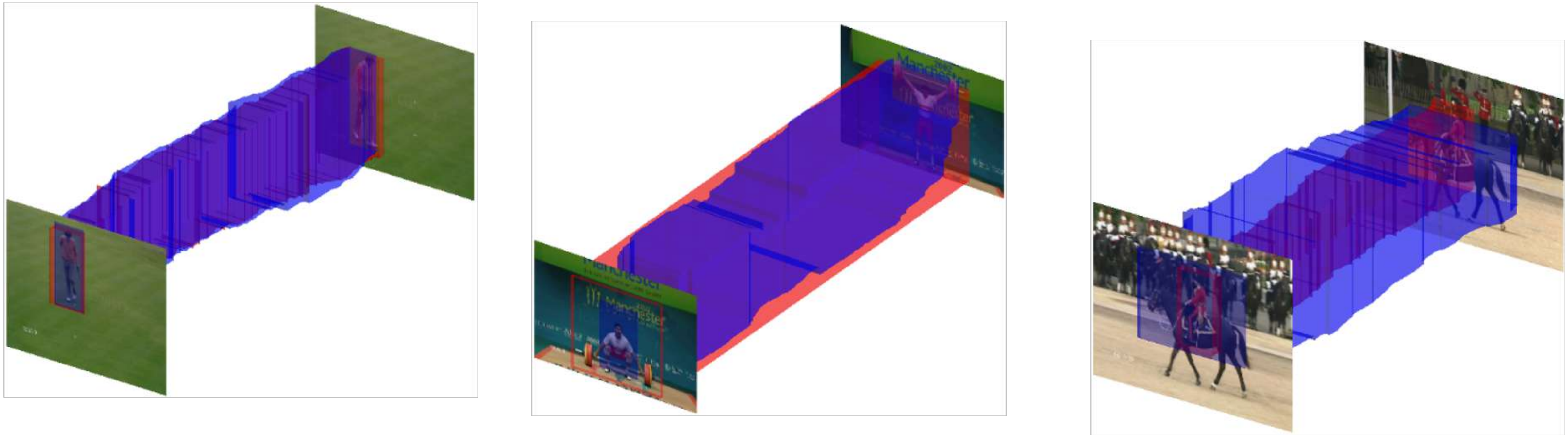
Best possible proposal performs as good as ground truth tube.

Training without ground truth boxes



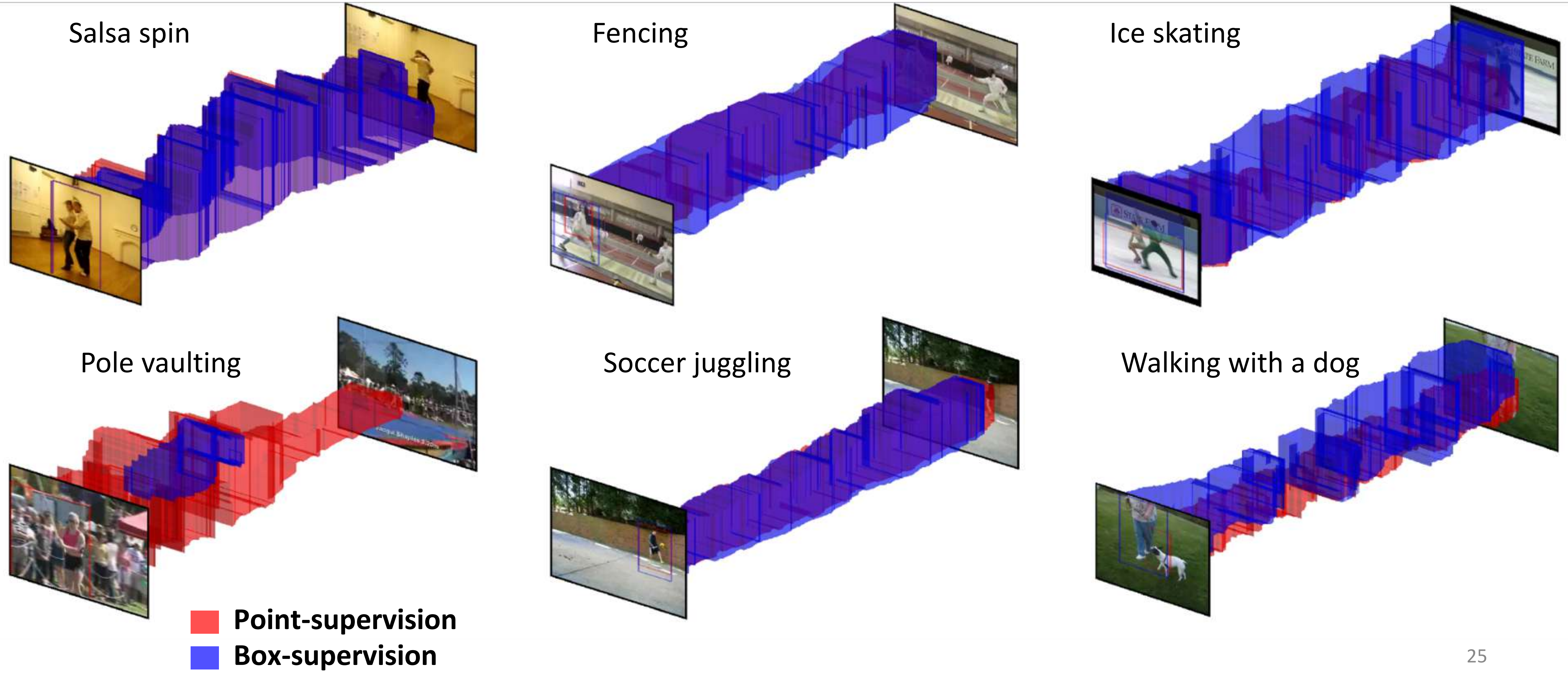
Mean AP maintained using our mined proposals.

Qualitative results

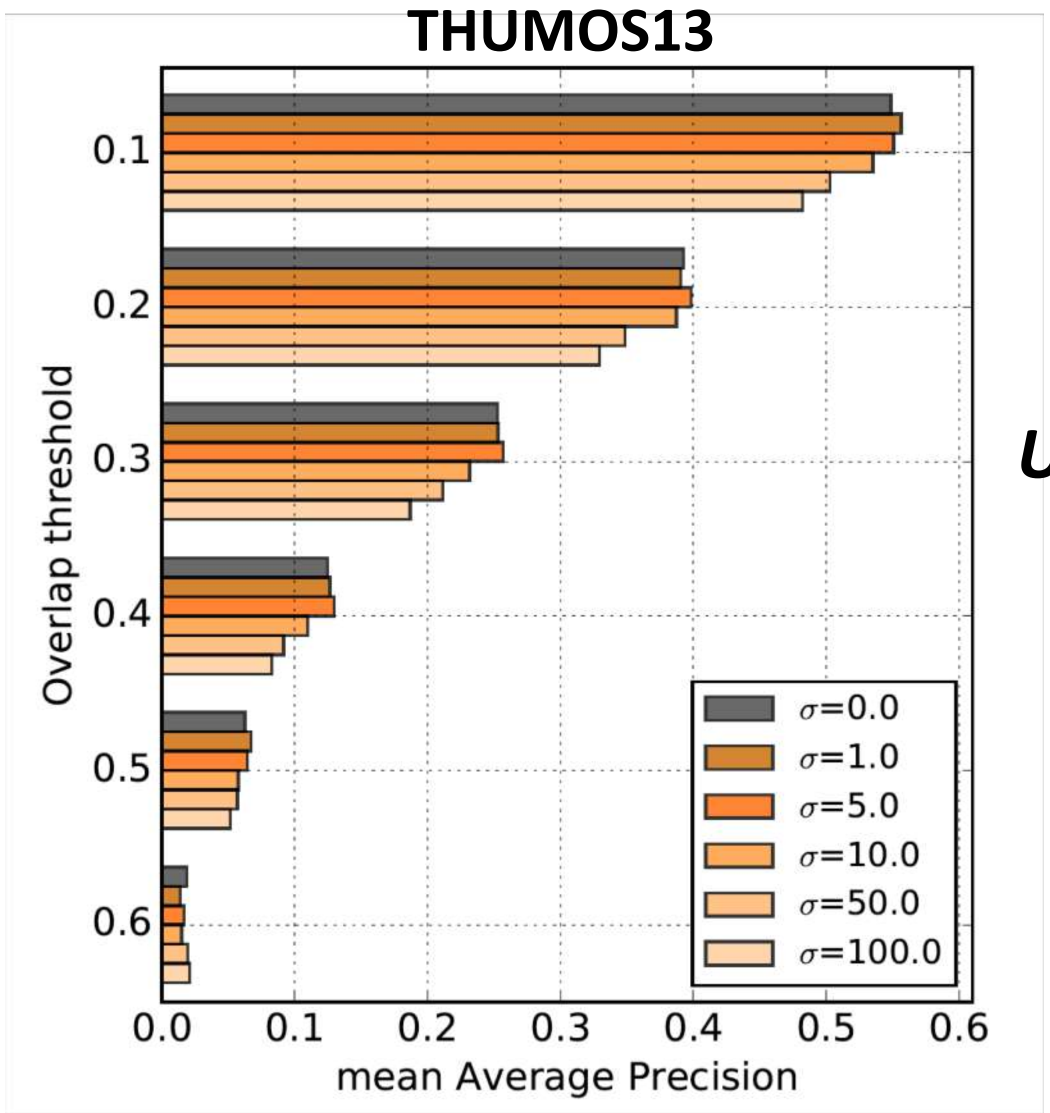


Ground truth boxes
Our mined proposal

Points vs boxes



How precise do we need to point?



Up to 10 pixels from action center good enough.

How much faster?

	Box supervision	Point supervision Annotation stride						
		1	2	5	10	20	50	100
mAP@0.2	0.399	0.393	0.404	0.389	0.384	0.395	0.379	0.371
mAP@0.5	0.074	0.063	0.060	0.068	0.064	0.061	0.064	0.053
Annotation speed-up	**1.0**	**9.8**	**19.3**	**46.0**	**85.0**	**147.6**	**264.6**	**359.6**

Points on par with boxes, with 50-fold speed-up.

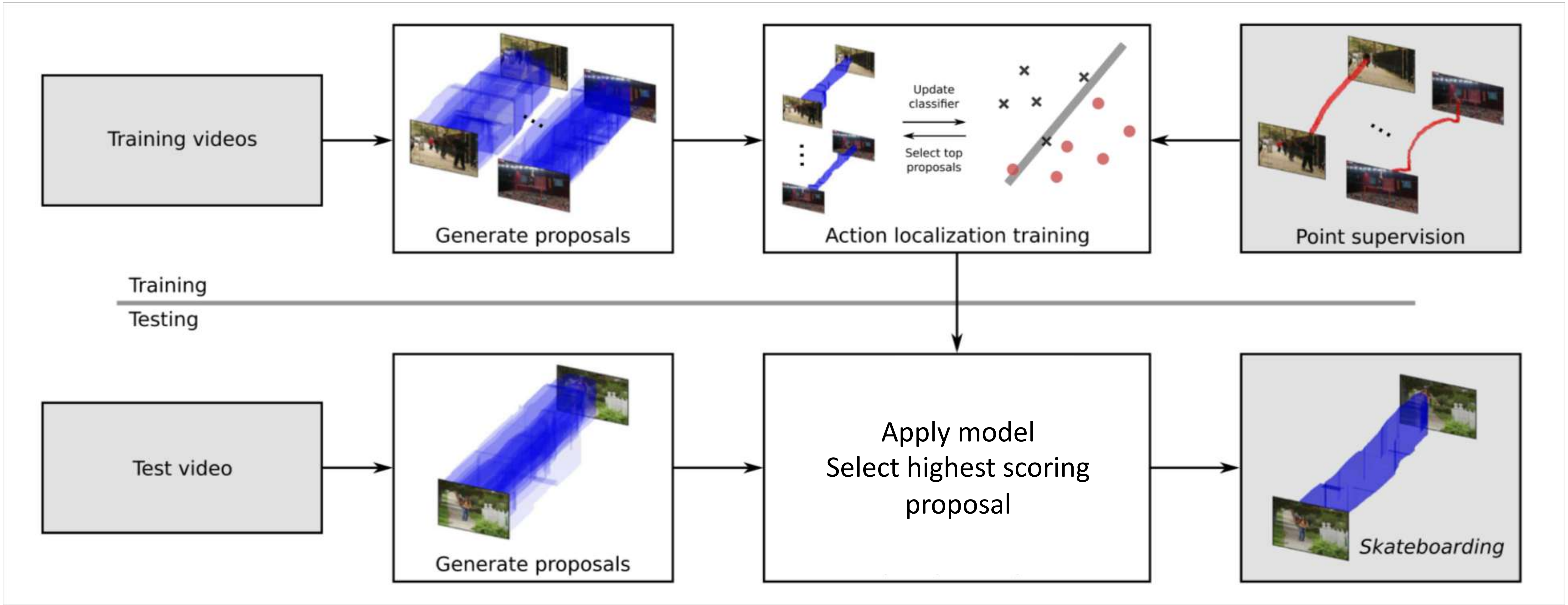
Up to 300-fold speed-up with marginal mAP drop only

Apple-to-apple comparison

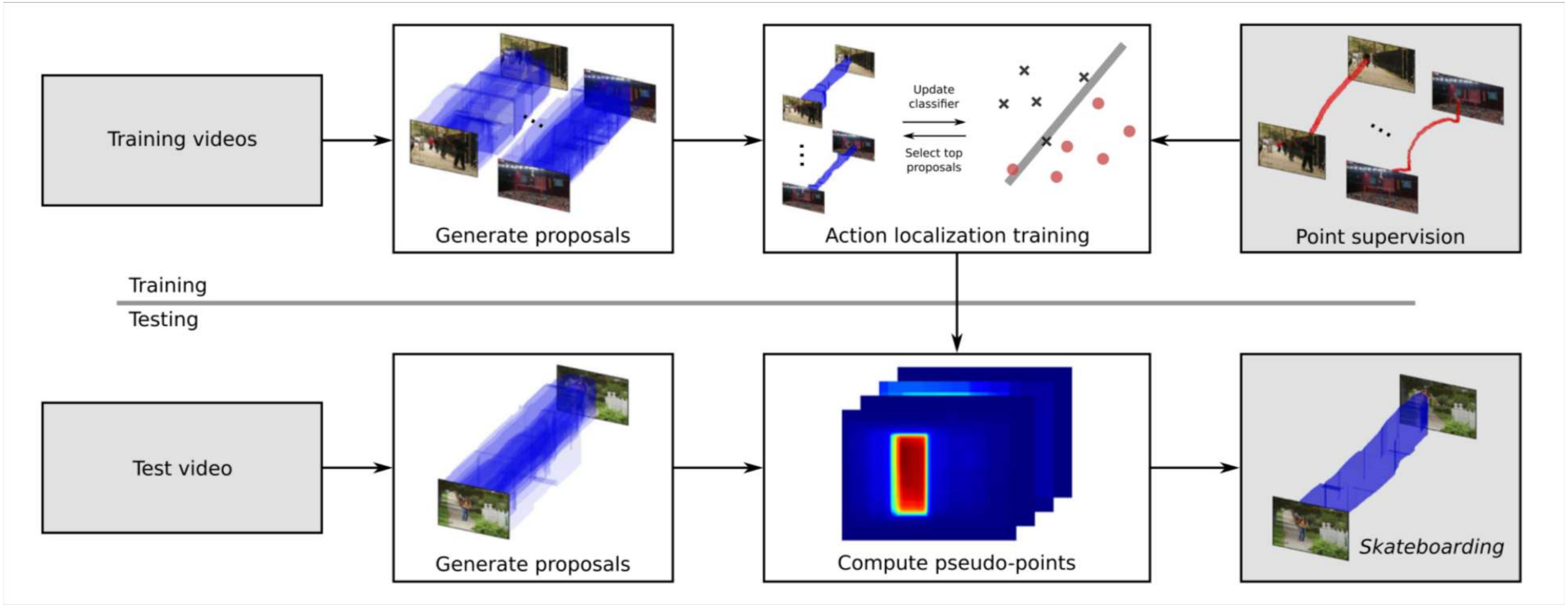
	Action supervision		THUMOS13
	Boxes	Labels	mAP @ 0.2
van Gemert *et al.* BMVC 2015	✓	✓	34.5
Point annotation		✓	34.8

Point annotation good alternative for box annotation.

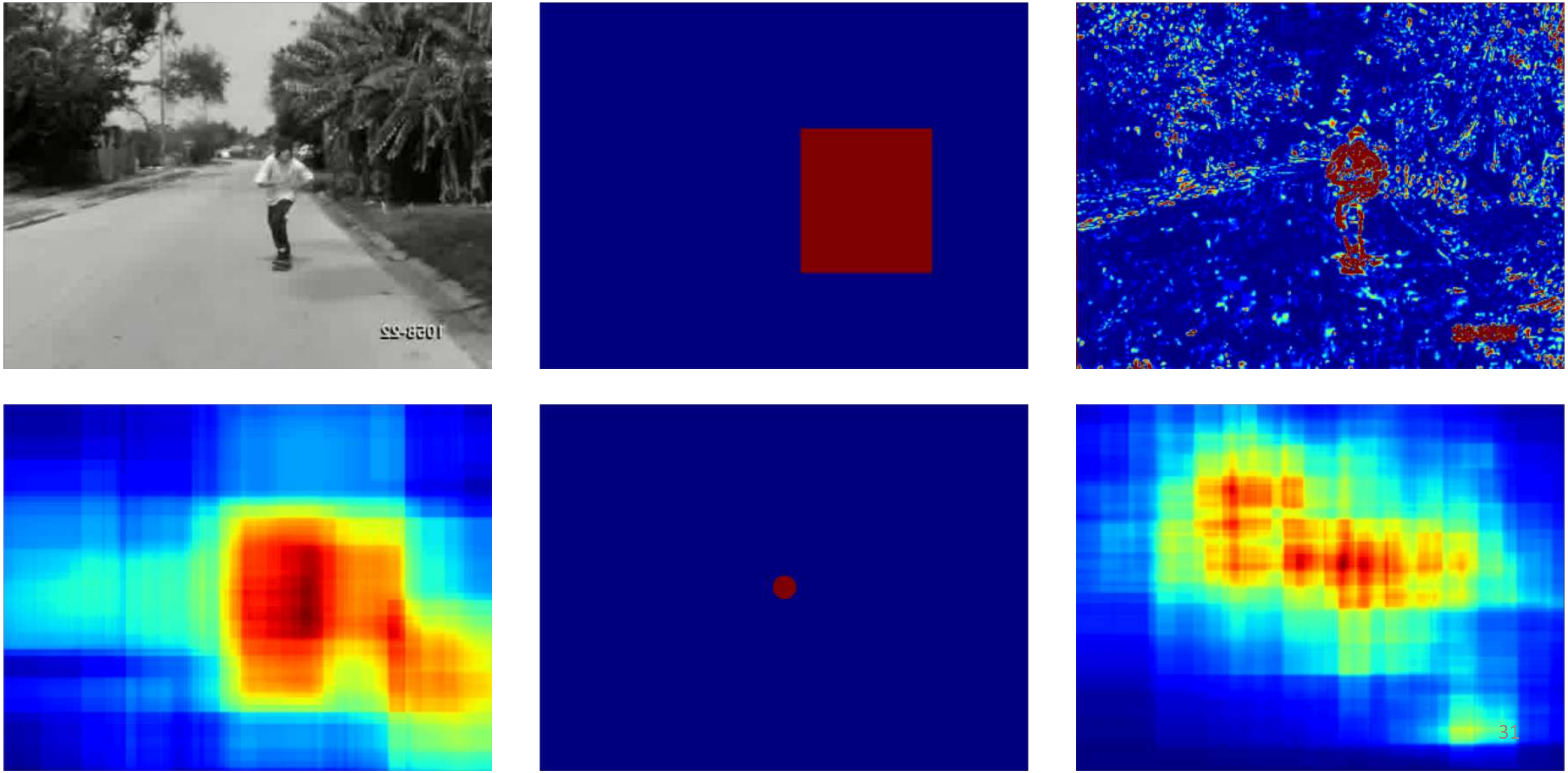
Adding pseudo-points during inference



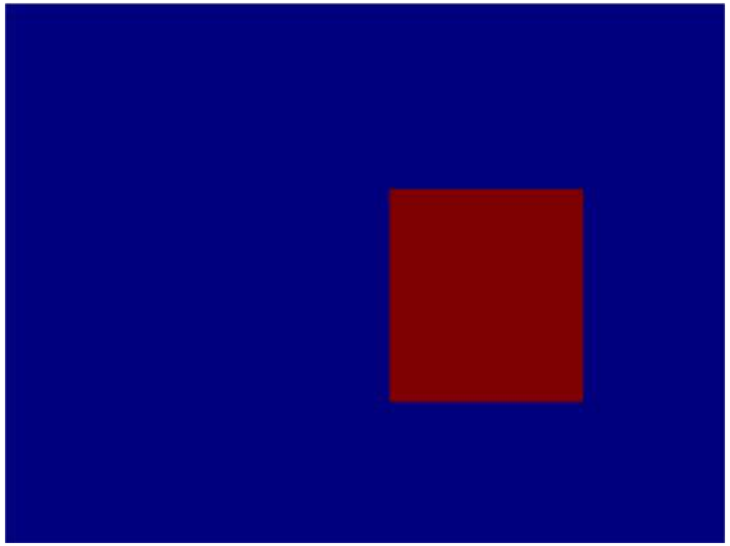
Adding pseudo-points during inference



Pseudo-point examples



Pseudo-pointing with person detector



Person detection

Select box with highest person confidence from pre-trained network.

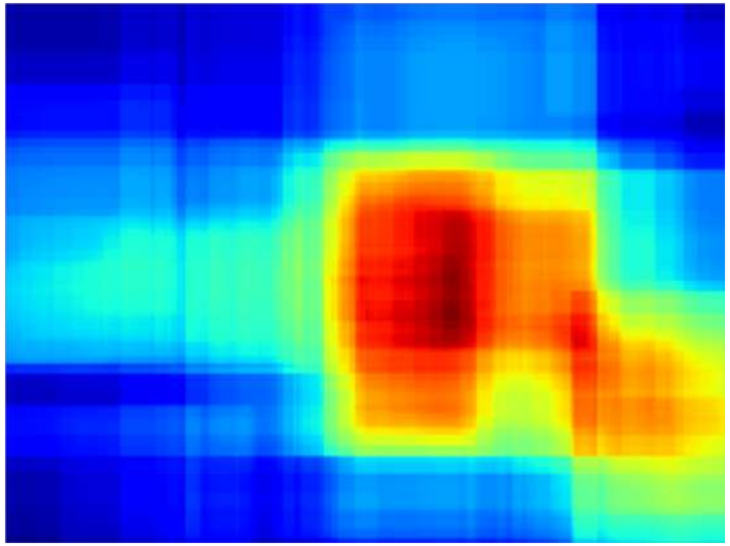
Ren *et al.* NIPS 2015.

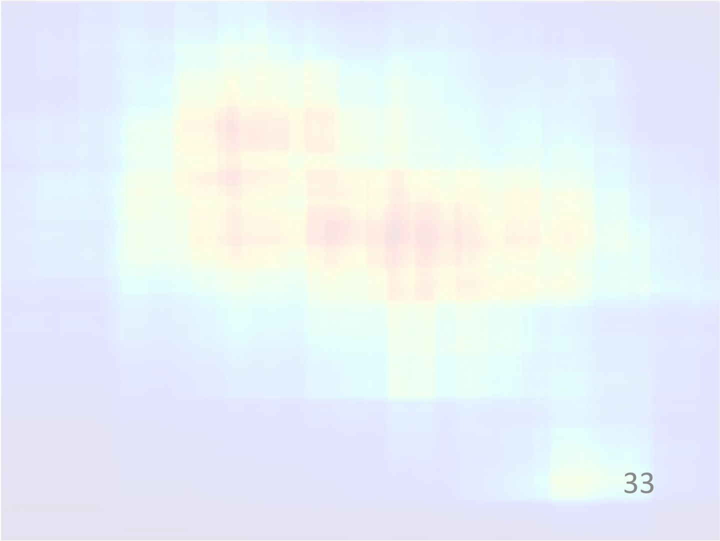
Pseudo-pointing with action proposals

Action proposals

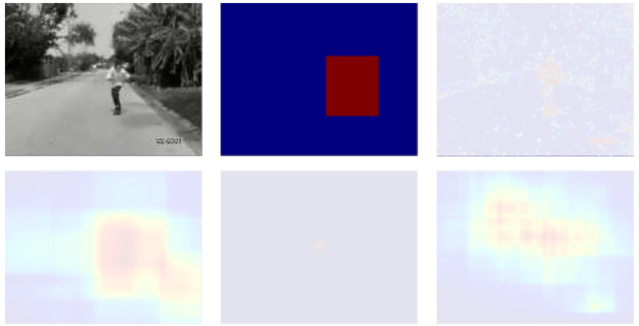
Centre of mass of the
per-pixel action proposal count.

van Gemert *et al.* BMVC 2015.



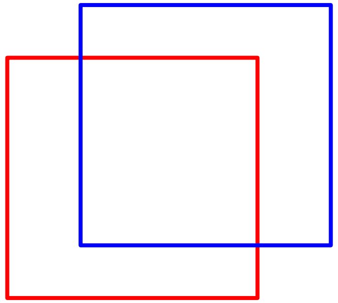


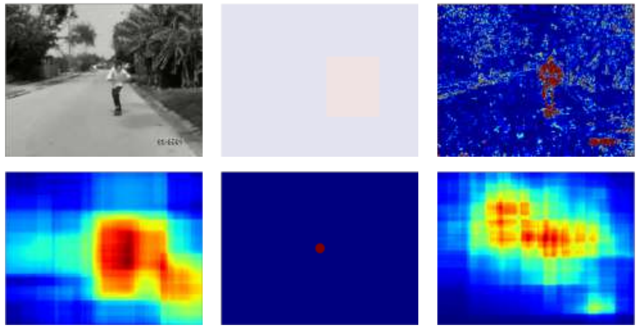
Matching pseudo-points with proposals



Person detection (box)

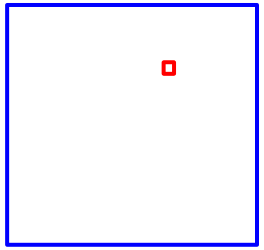
Intersection-over-Union between boxes





Other pseudo-annotations (point)

Match point with box centre



Weighted overlap regularizes proposal selection.

Apple-to-apple comparison

	Action supervision		THUMOS13
	Boxes	Labels	mAP @ 0.2
van Gemert *et al.* BMVC 2015	✓	✓	34.5
Point annotation		✓	34.8
\w pseudo points at inference		✓	41.8

Points and pseudo-points better than box.

Take-aways

Points provide a fast and viable alternative to box-supervision

Pseudo-points at inference aid action localization accuracy

II.

Less streams

Dance with Flow: Two-in-One Stream Action Detection
Jiaojiao Zhao and Cees Snoek. In *CVPR* 2019.

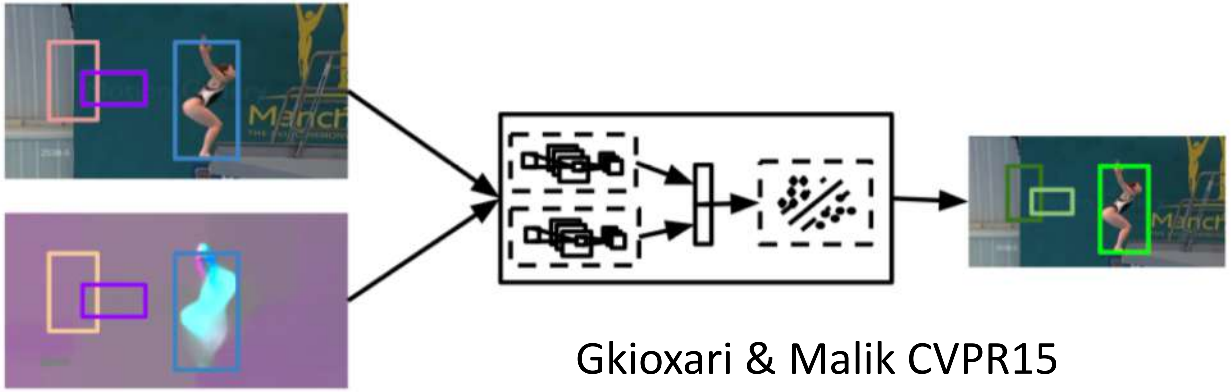
Two-stream

Simonyan & Zisserman NeurIPS14

Default strategy for action detection and classification.

RGB-stream: appearance only

Flow-stream: motion only

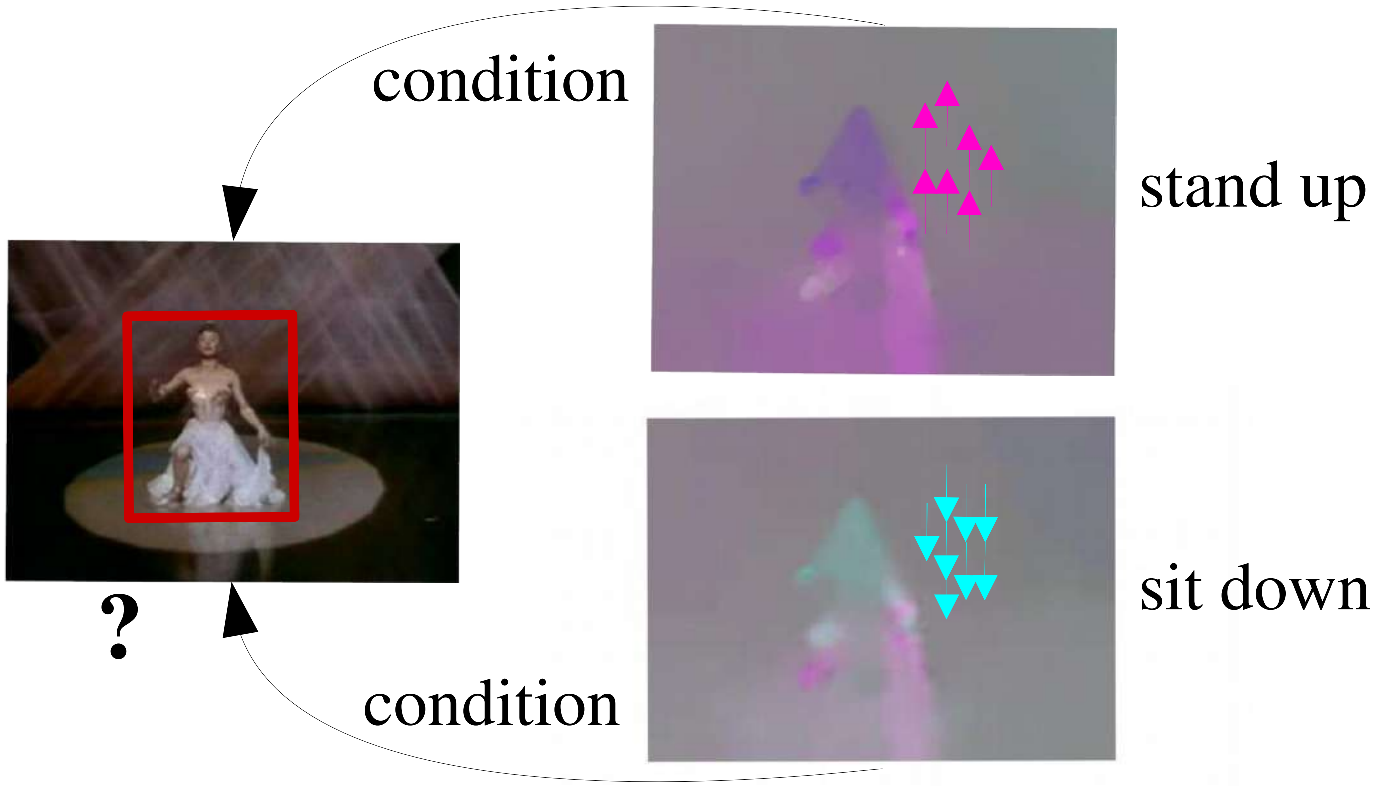


Gkioxari & Malik CVPR15

***Doubles computation and parameters for modest accuracy gain*.**

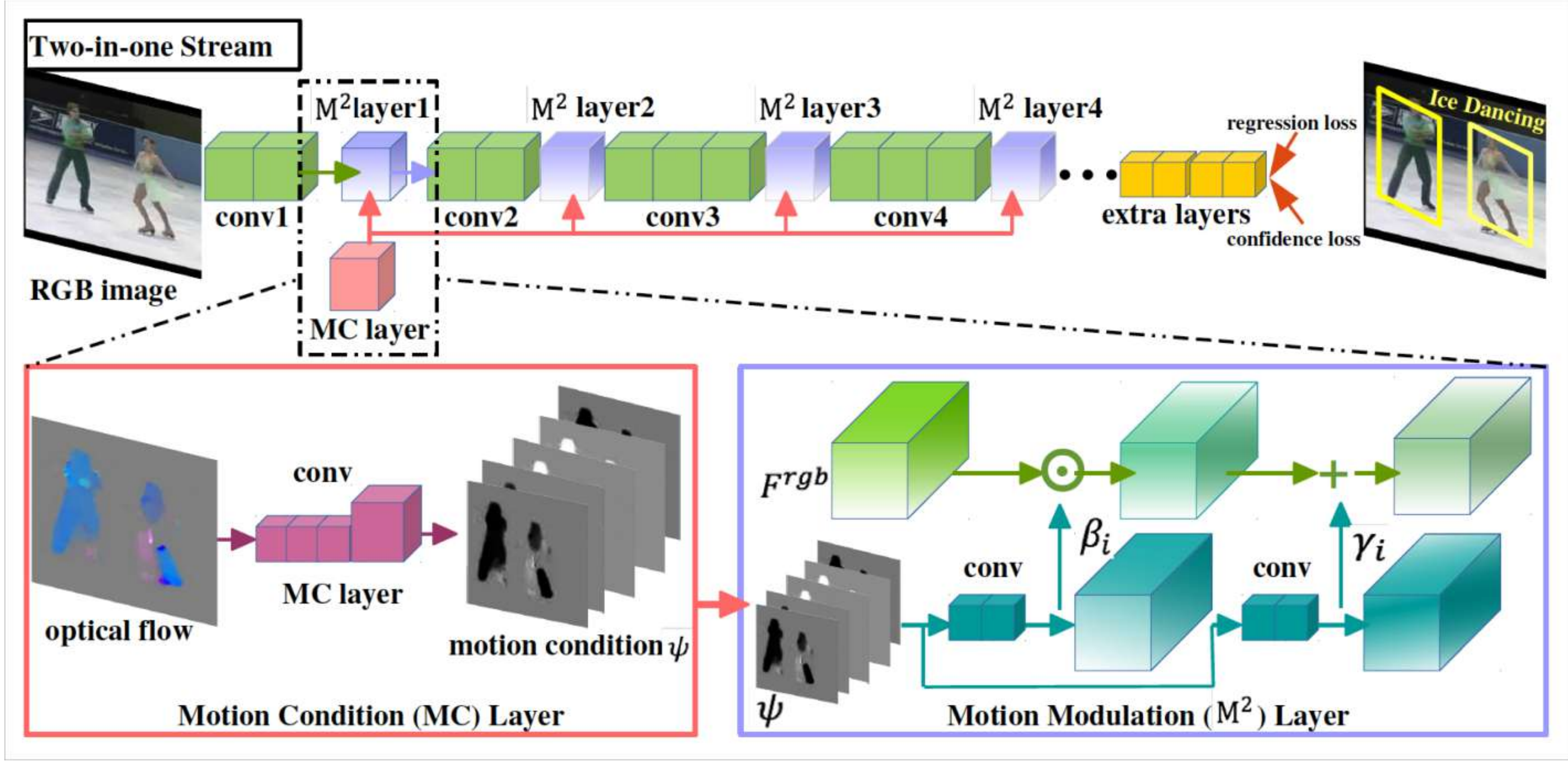
Key idea

Use motion as condition when training **a single RGB-stream**.

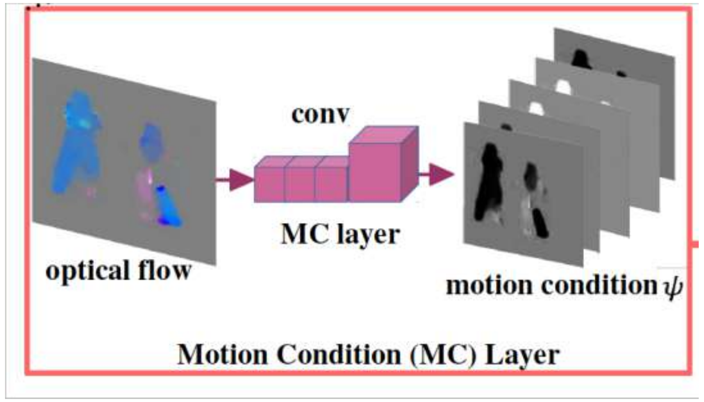


Two-in-one Stream

Learns a single stream RGB model conditioned on motion information



Motion condition layer

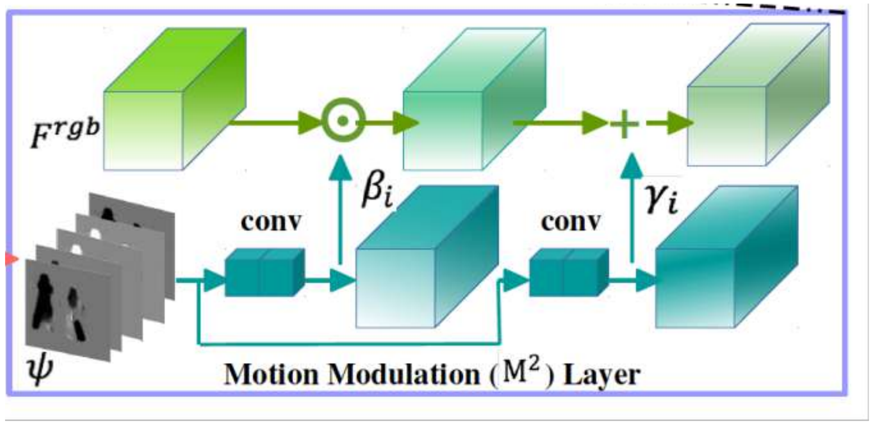


Generates simple features from flow images

Flow images are sparse, simple 1x1 or 3x3 convolution layer sufficient

Flow **Motion condition maps**

Motion modulation layer

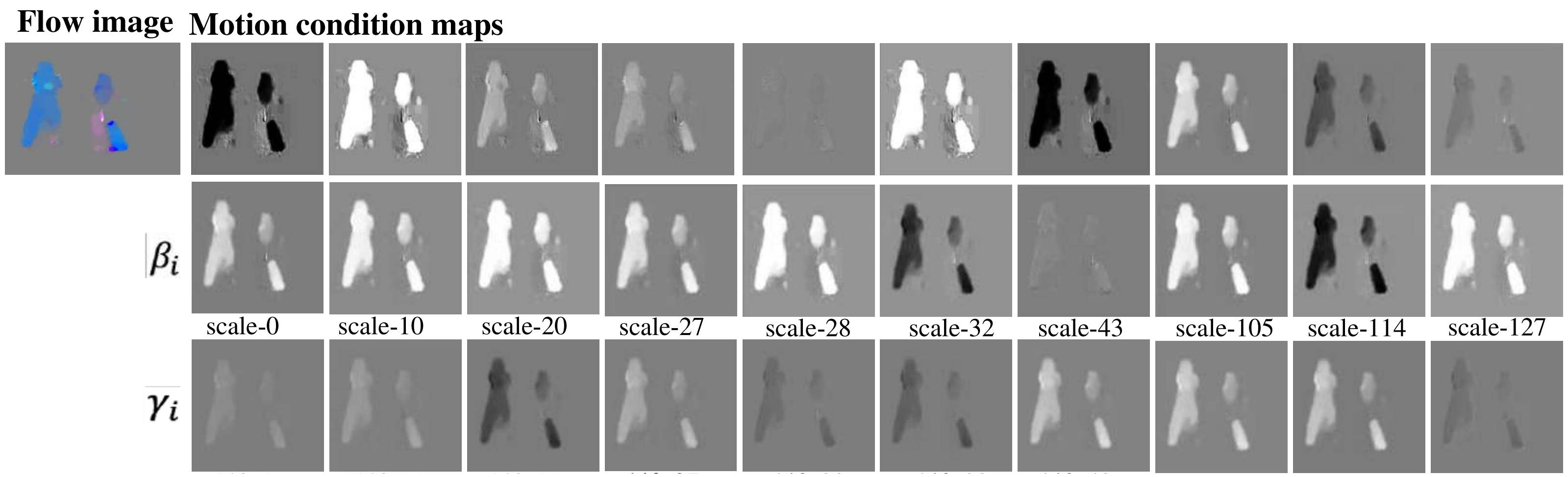


Generates a pair of transformation parameters

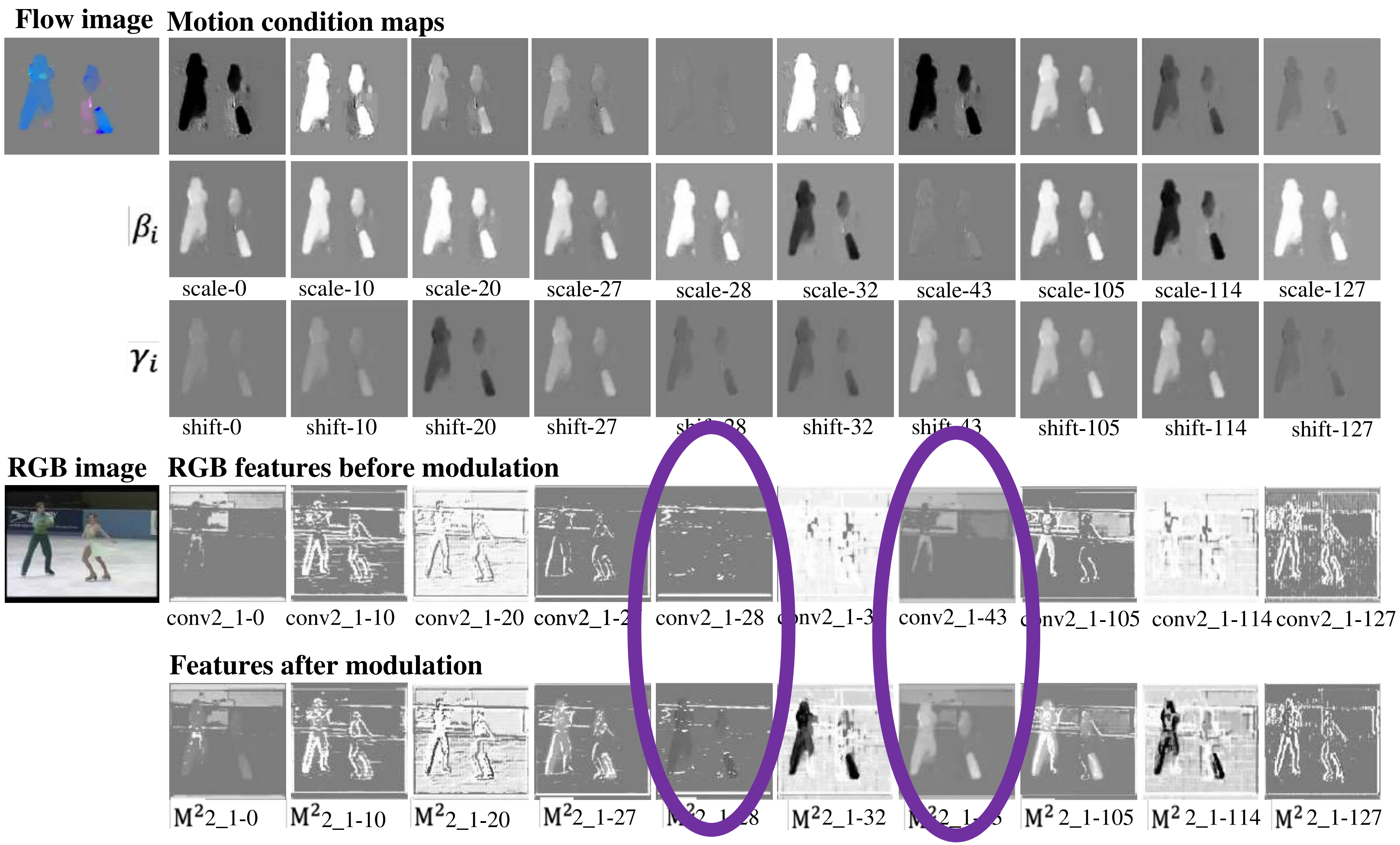
Two groups of 1x1 convolutional layers generate the parameters

RGB features are modulated by element-wise multiplication

Feature visualization



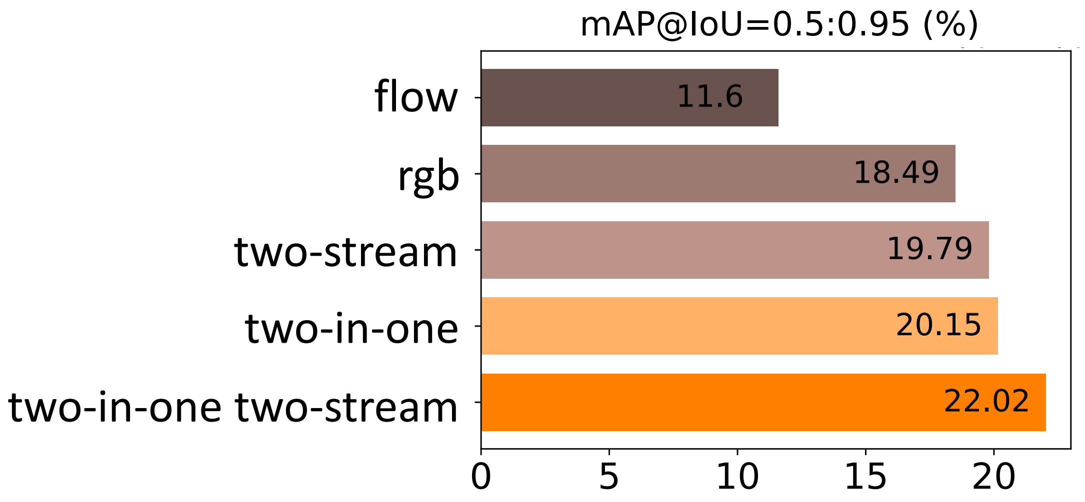
Feature visualization



Modulated features focus more on moving actors.

Ablation: *Two-in-one detection vs. baselines*

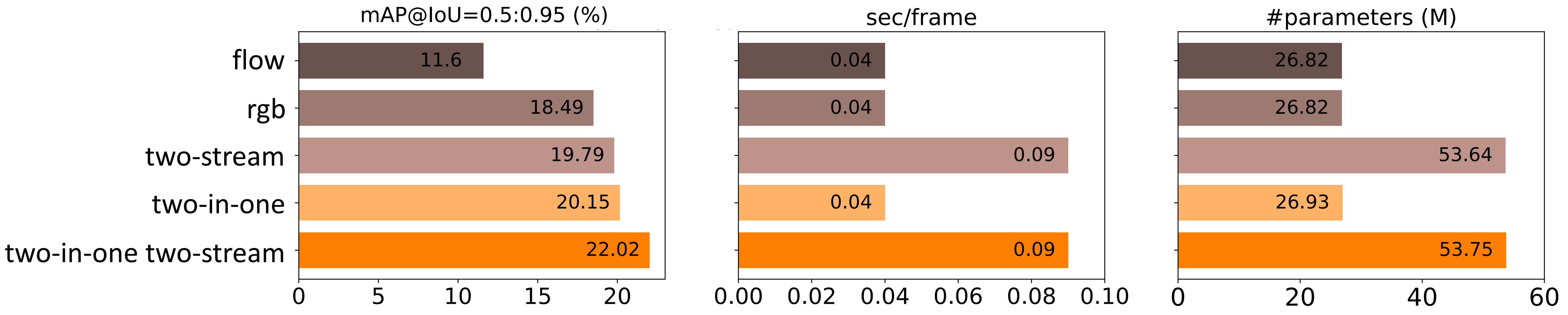
Single-frame SSD by Singh *et al.* ICCV17, on UCF101-24.



Better action detection

Ablation: *Two-in-one detection vs. baselines*

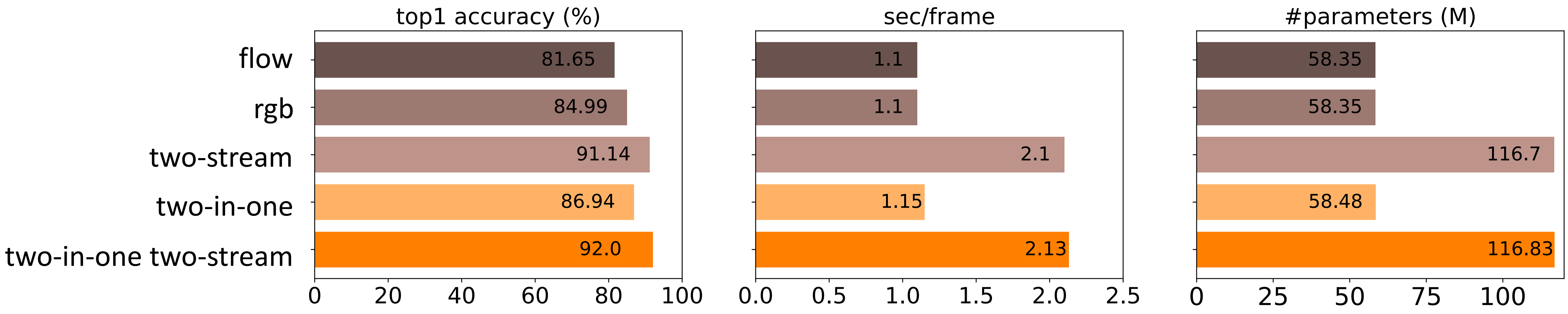
Single-frame SSD by Singh *et al.* ICCV17, on UCF101-24.



Better action detection with only <u>half</u> the computation and parameters.

Ablation: *Two-in-one classification vs. baselines*

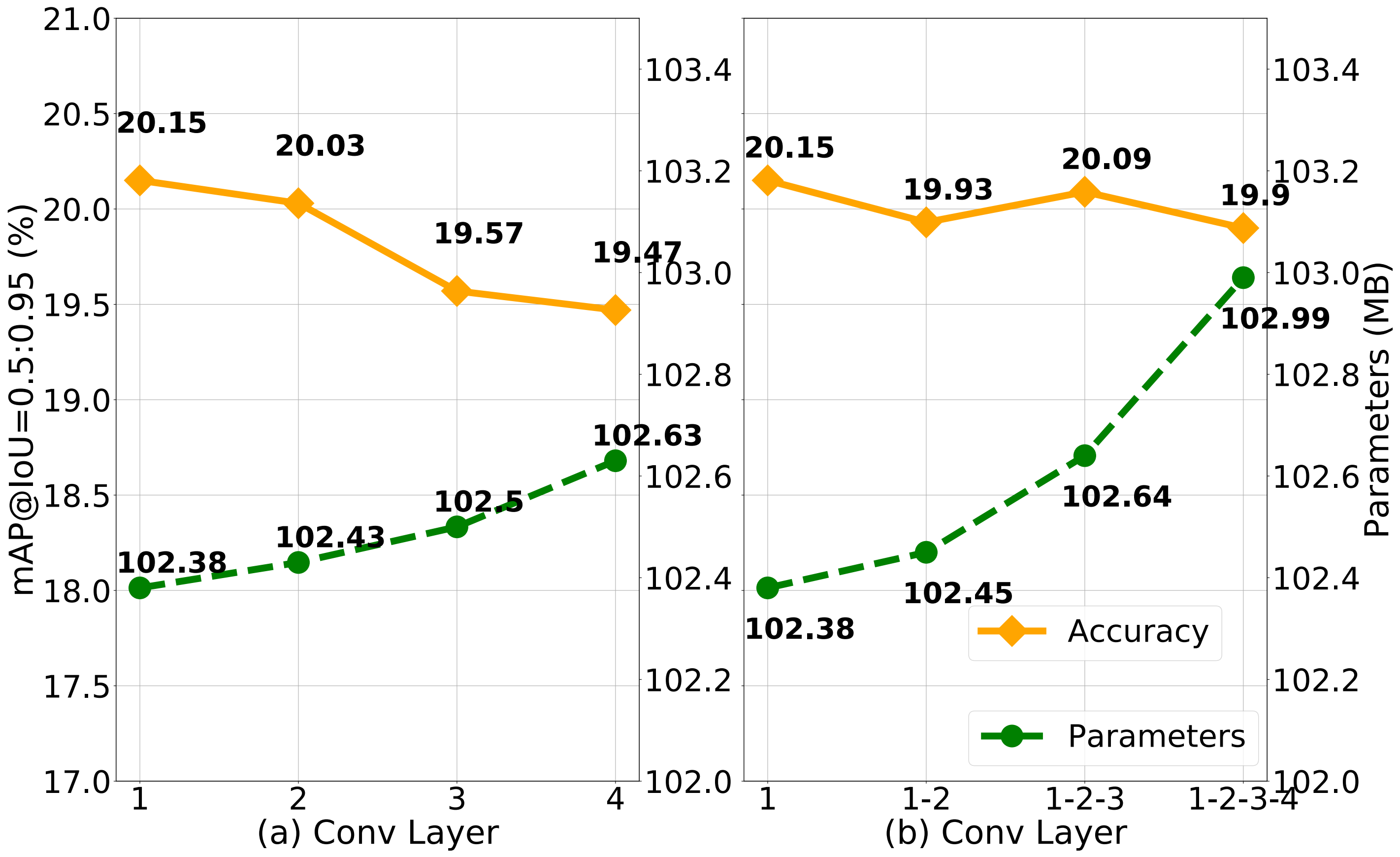
ResNet152 by Wang *et al.* ArXive15, on UCF101.



Action classification profits less, accuracy-wise.

Ablation: *Where to modulate?*

Single-frame SSD by Singh *et al.* ICCV17, on UCF101-24.



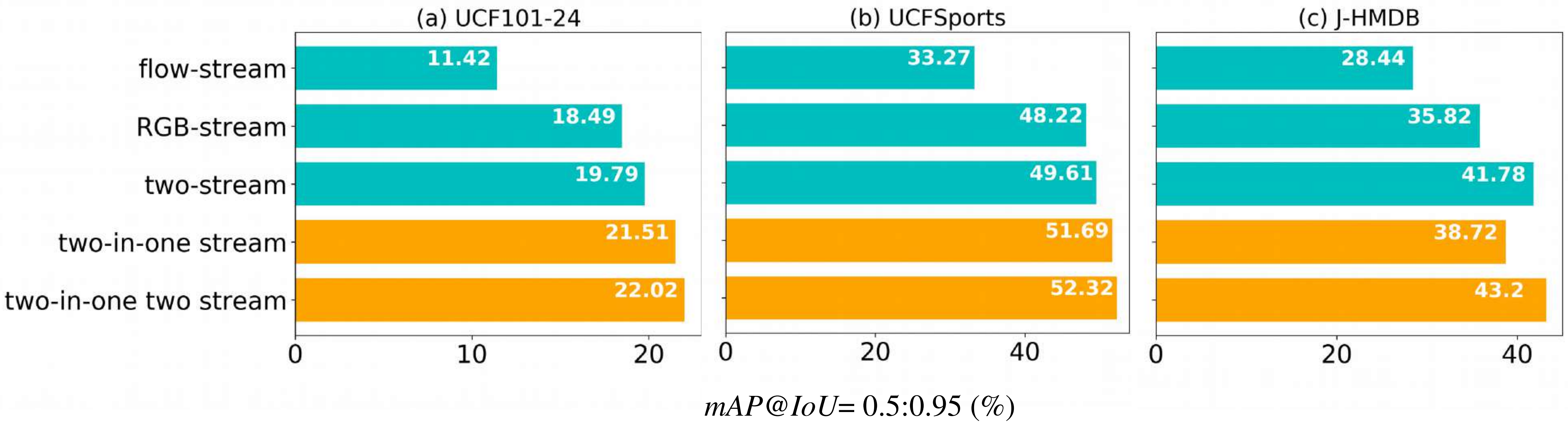
Modulating after Conv 1 gives us the best result with least parameters.

Ablation: *What flow?*

	**Brox** 	**Flownet** 	**RealTimeFlow** 
Flow-stream	11.60	7.13	3.58
RGB-stream	18.49	18.49	18.49
Two-stream	19.79	19.75	18.53
Two-in-one stream	**21.51**	**19.97**	**19.16**

Works with any flow, best with Brox.

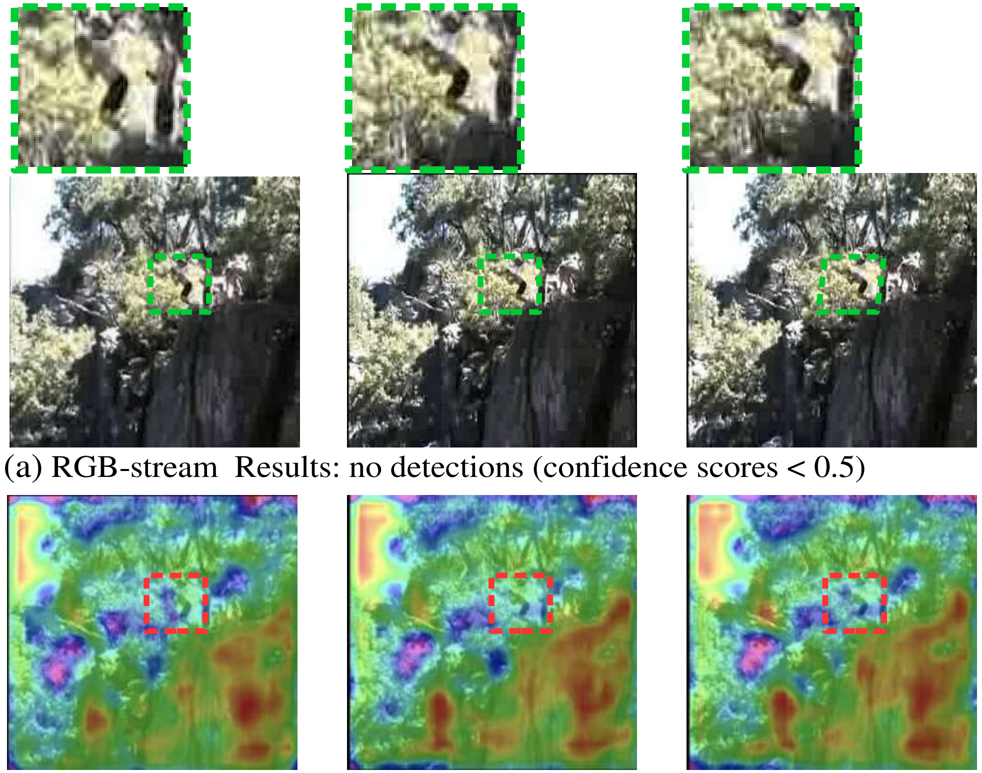
Ablation: *Generalization ability*



Also better than two-stream on UCF-Sports, worse on J-HMDB.

Qualitative analysis

Two-in-one stream has higher activation on actions, resulting in correct detection.

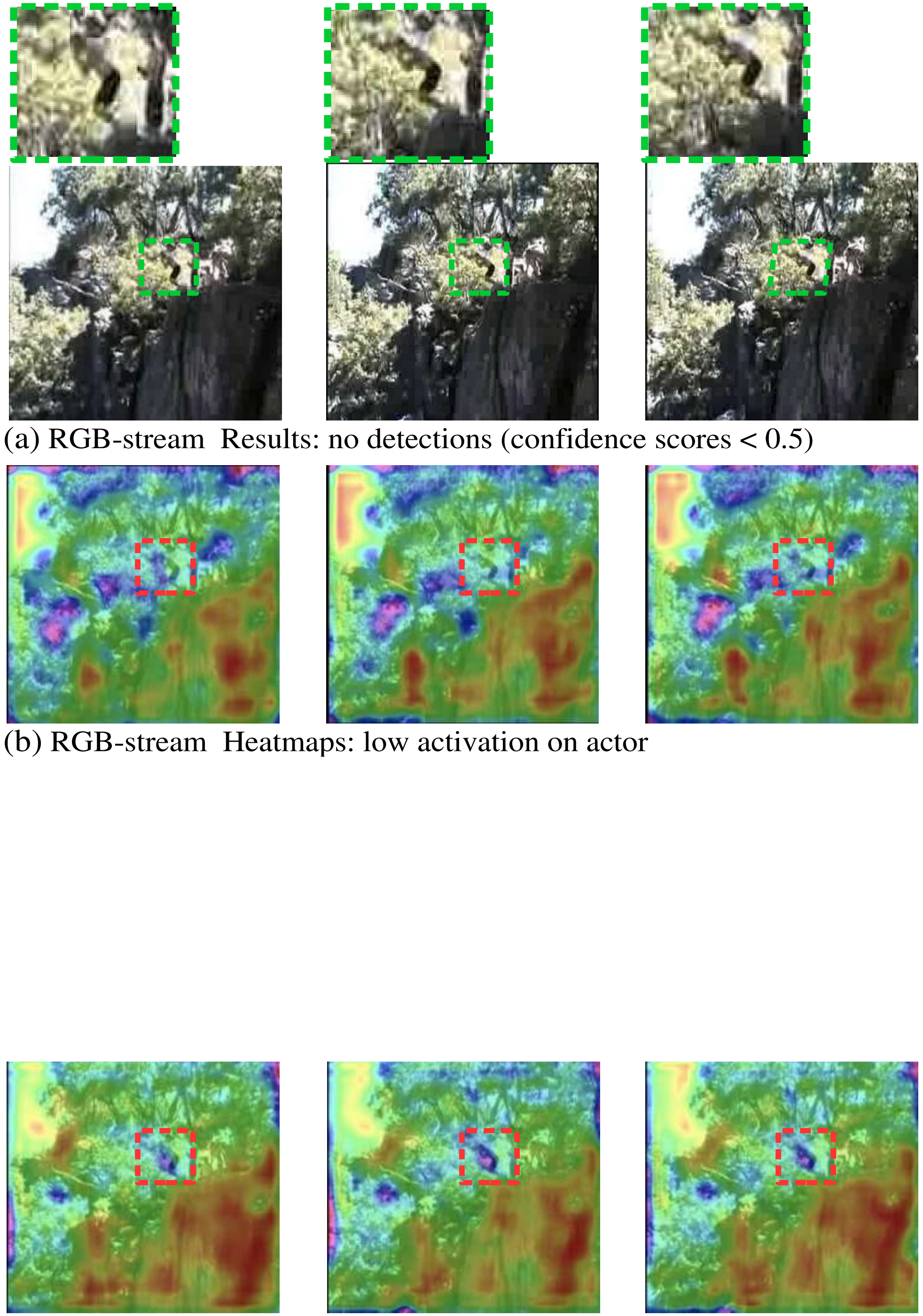


(a) RGB-stream Results: no detections (confidence scores < 0.5)

(b) RGB-stream Heatmaps: low activation on actor

Qualitative analysis

Two-in-one stream has higher activation on actions, resulting in correct detection.

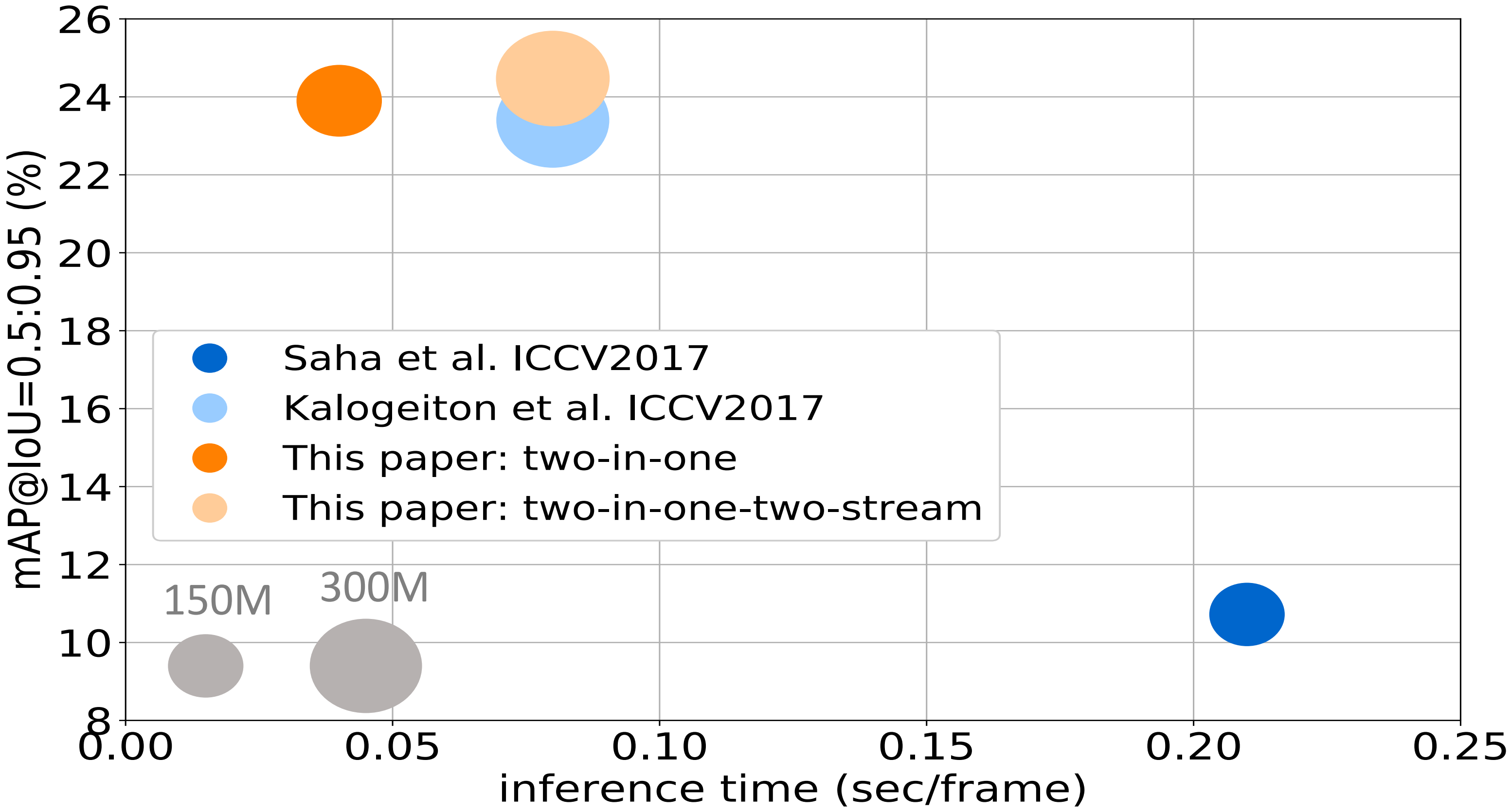


(a) RGB-stream Results: no detections (confidence scores < 0.5)

(b) RGB-stream Heatmaps: low activation on actor

(d) Two-in-one Heatmaps: high activation on actor

Comparison with state-of-the-art



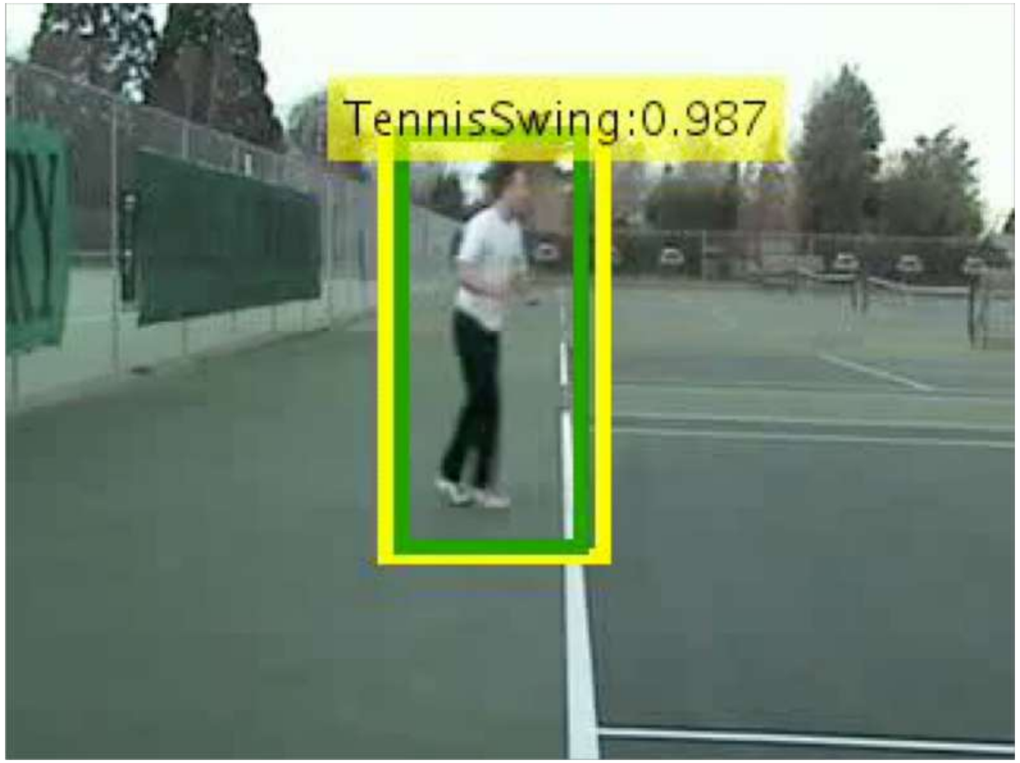
Faster, lighter and better accuracy.

Results: *success*

Ground truth

Our prediction

Results: *failures*



Ground truth

Our prediction

Take-aways

Two-in-one stream is simple, effective and efficient

but we still need to pre-compute optical flow

Modulation may profit from other priors as well

Thank you

www.ceessnoek.info