OXFORD BROOKES UNIVERSITY

SARAS
SMART AUTONOMOUS ROBOTIC ASSISTANT SURGEON

European Commission | Horizon 2020 European Union funding for Research & Innovation

ECCV 2018

Predicting Action Tubes

Gurkirt Singh

Saha Suman

Fabio Cuzzolin

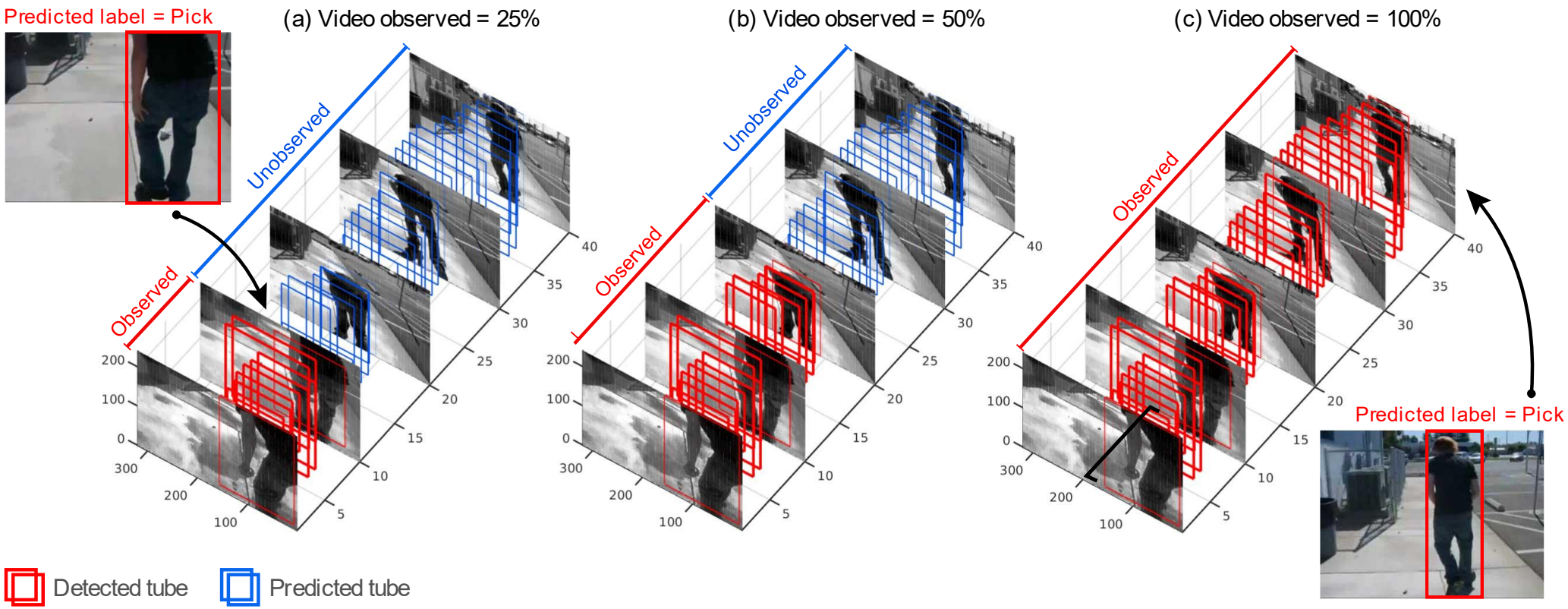
Visual Artificial Intelligence Laboratory (VAIL)
Oxford Brookes University

Outline

- Problem Statement
- Motivation
- Related Work
- Action detection
- Tube Prediction
- Results

Problem statement

The task is to determine/predict what action is occurring in a video, as early as possible using the observed part of the video, localise it (in red), and predict its future locations (in blue)

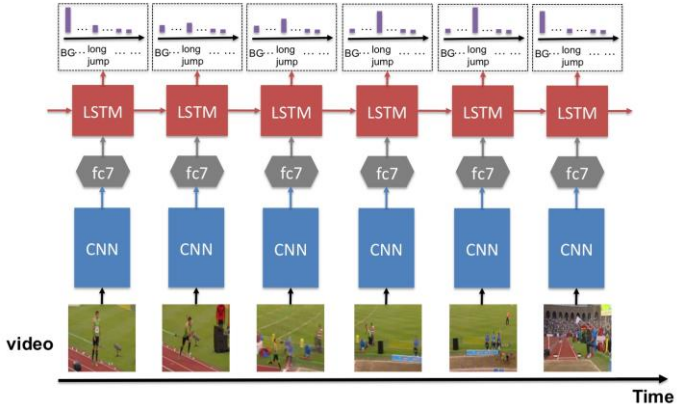


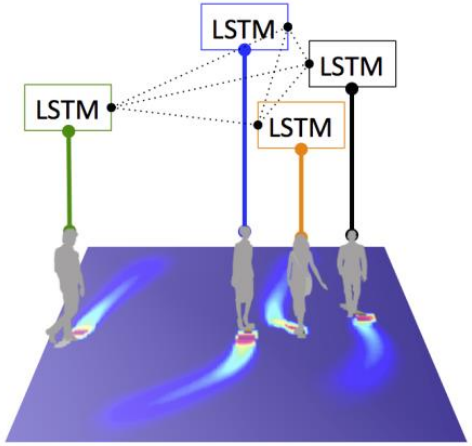
Why?

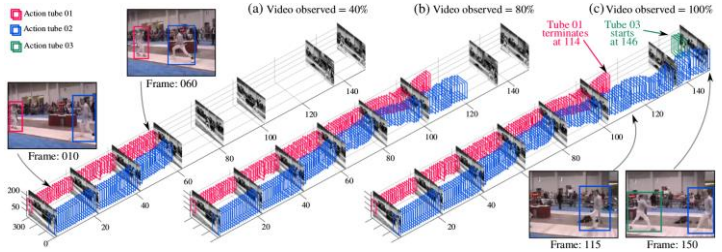
Surveillance, Human-robot interaction, Autonomous driving, Robotic surgery

Related work

Action Prediction + Trajectory Prediction + Online Action Detection







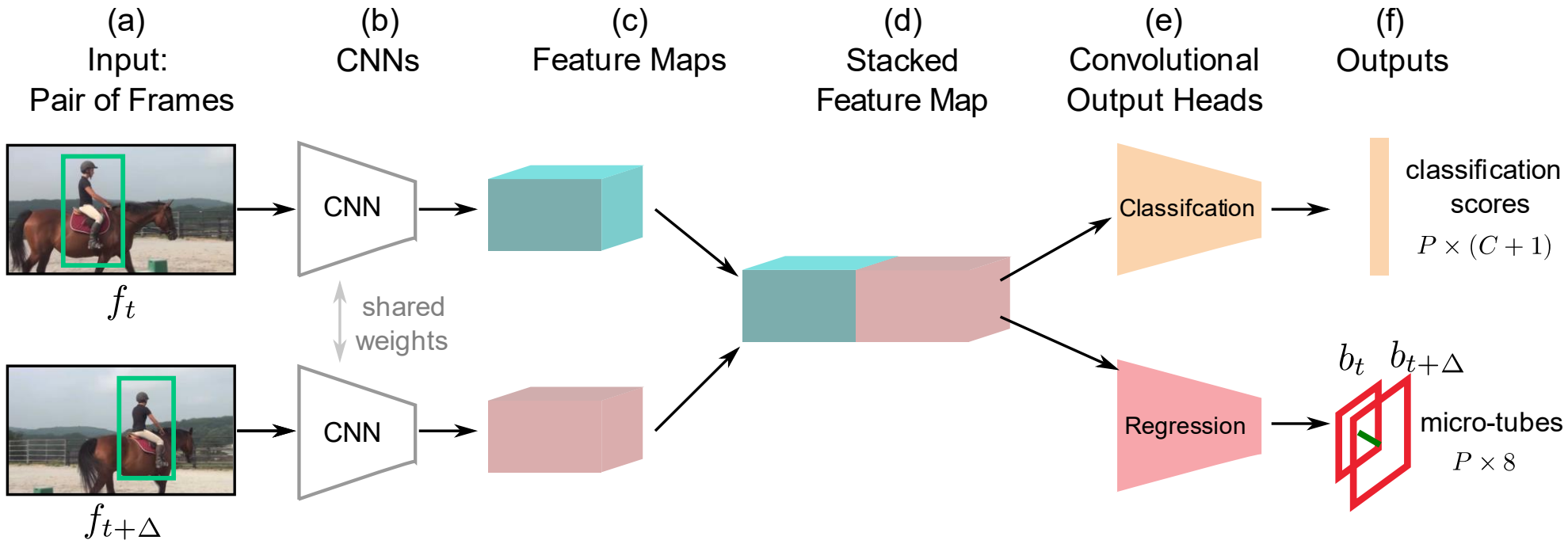
S. Ma et al., CVPR 2016
Y. Kong et al., CVPR 2016
G. Singh et al. ICCV 2017
K. Soomro et al. CVPR 2016

A. Alahi et al. CVPR 2016
K. Kitani et al. ECCV 2012

G. Singh et al. ICCV 2017
K. Soomro et al. CVPR 2016

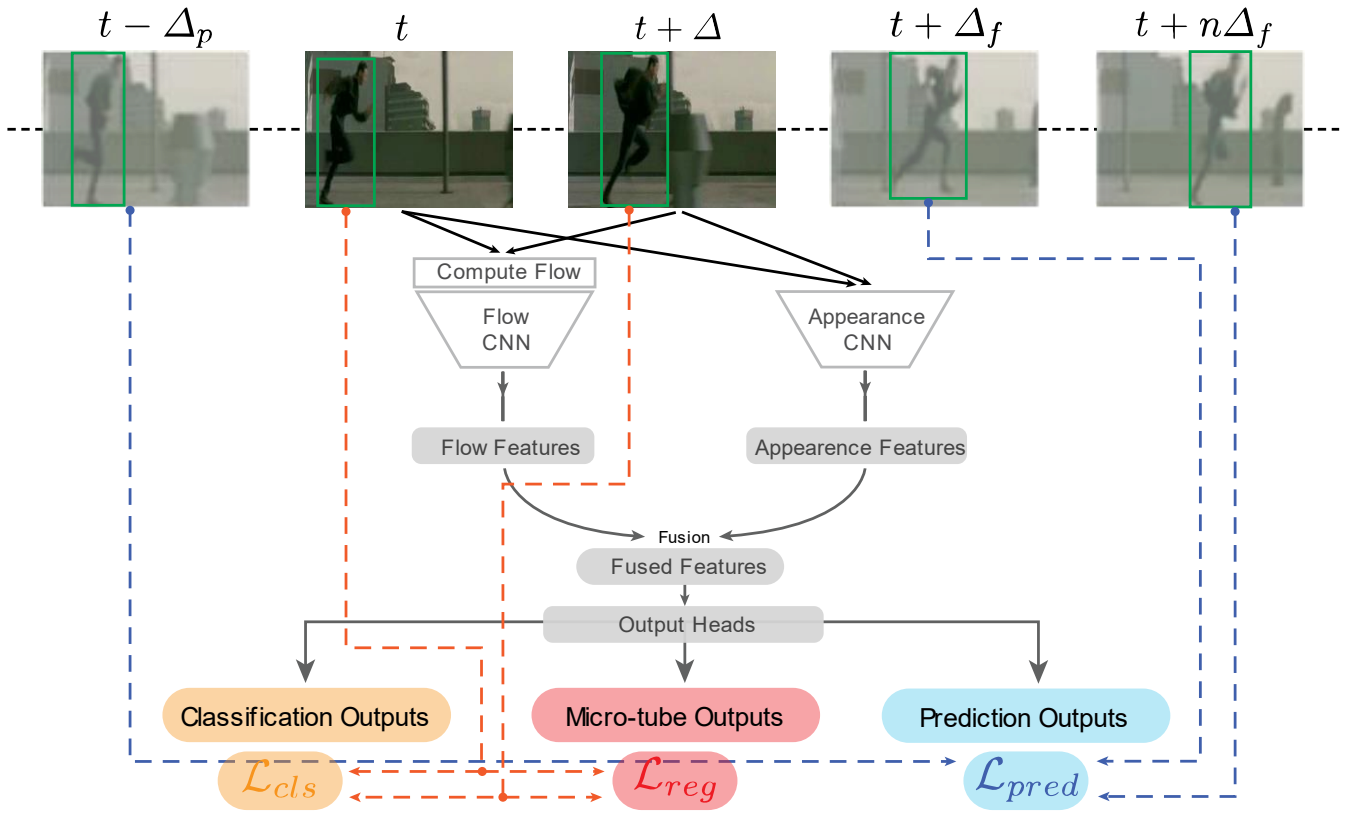
More works in paper

Action micro-tubes for action detection



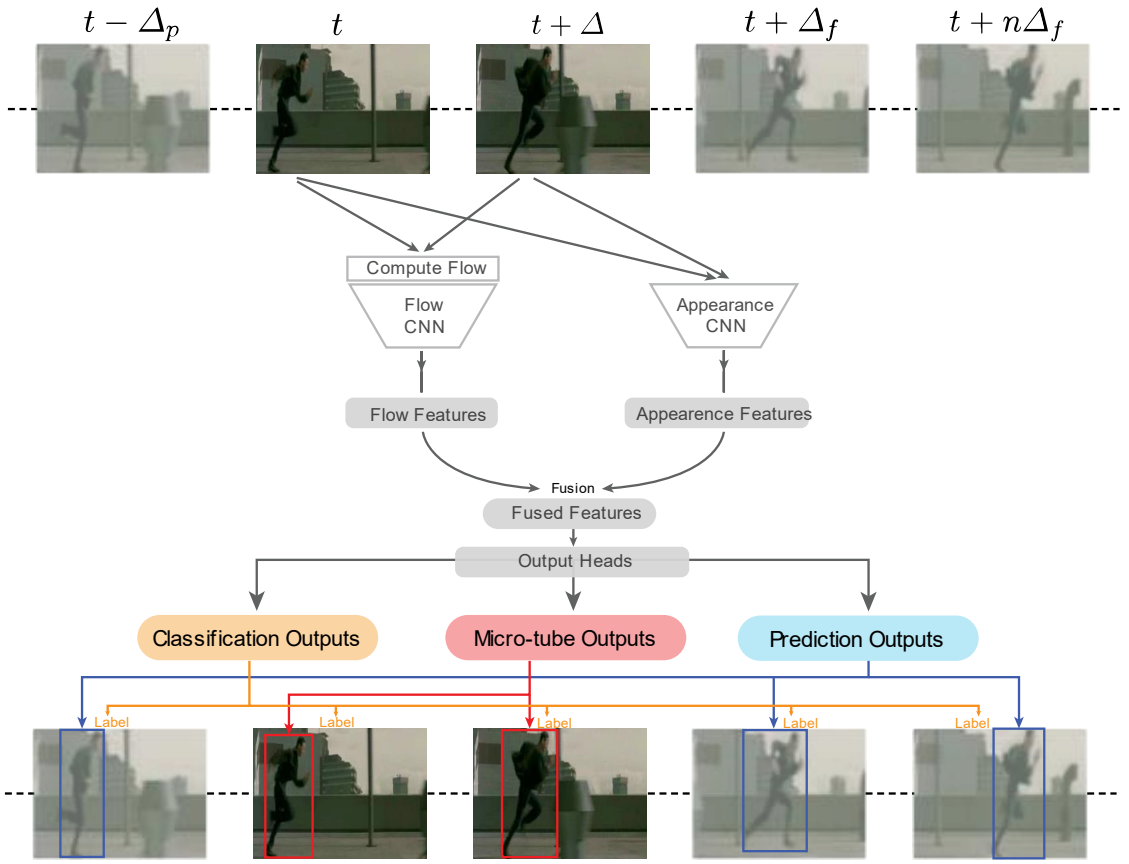
AMTnet Saha et al. ICCV 2017 + Online action tube construction by Singh et al. ICCV 2017

A micro-tube's future and past



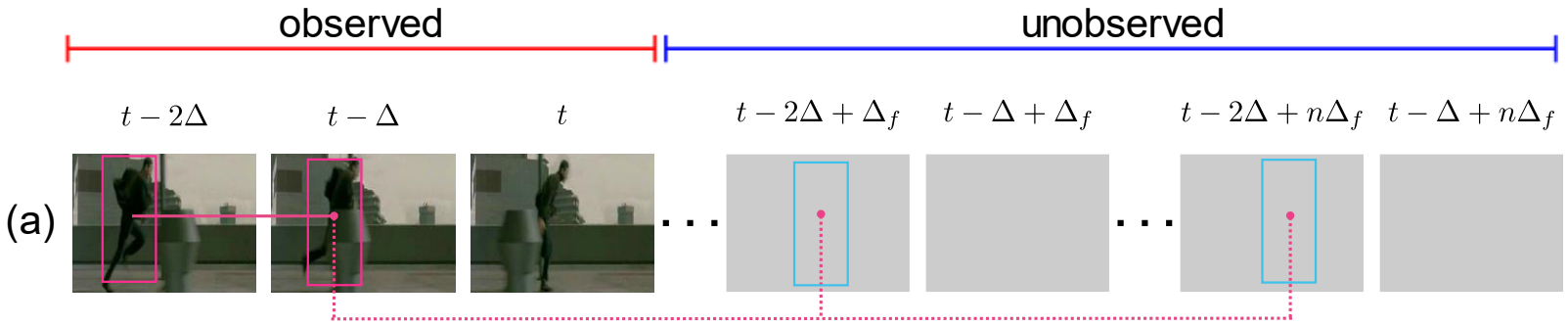
Training process overview

A micro-tube's future and past

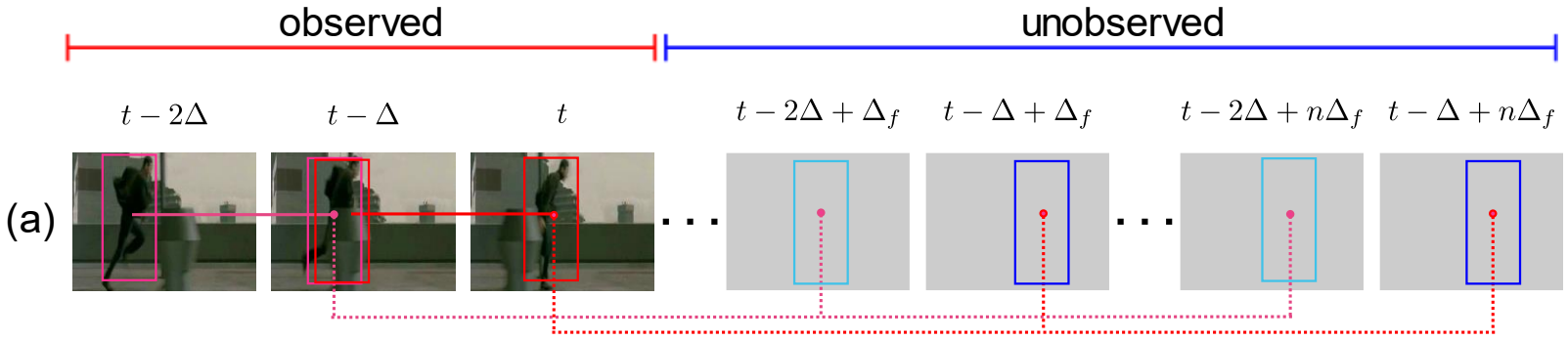


Testing process overview

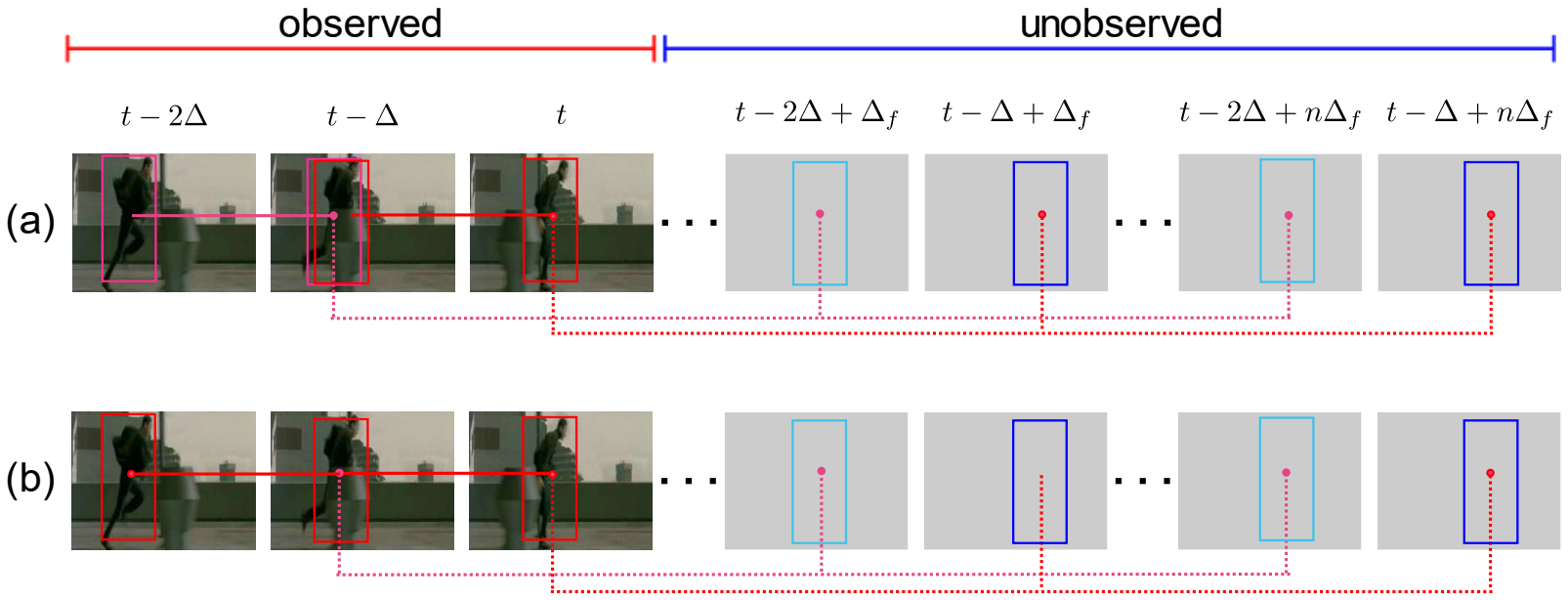
How linking micro-tubes affect the tube's future



How linking micro-tubes affect the tube's future

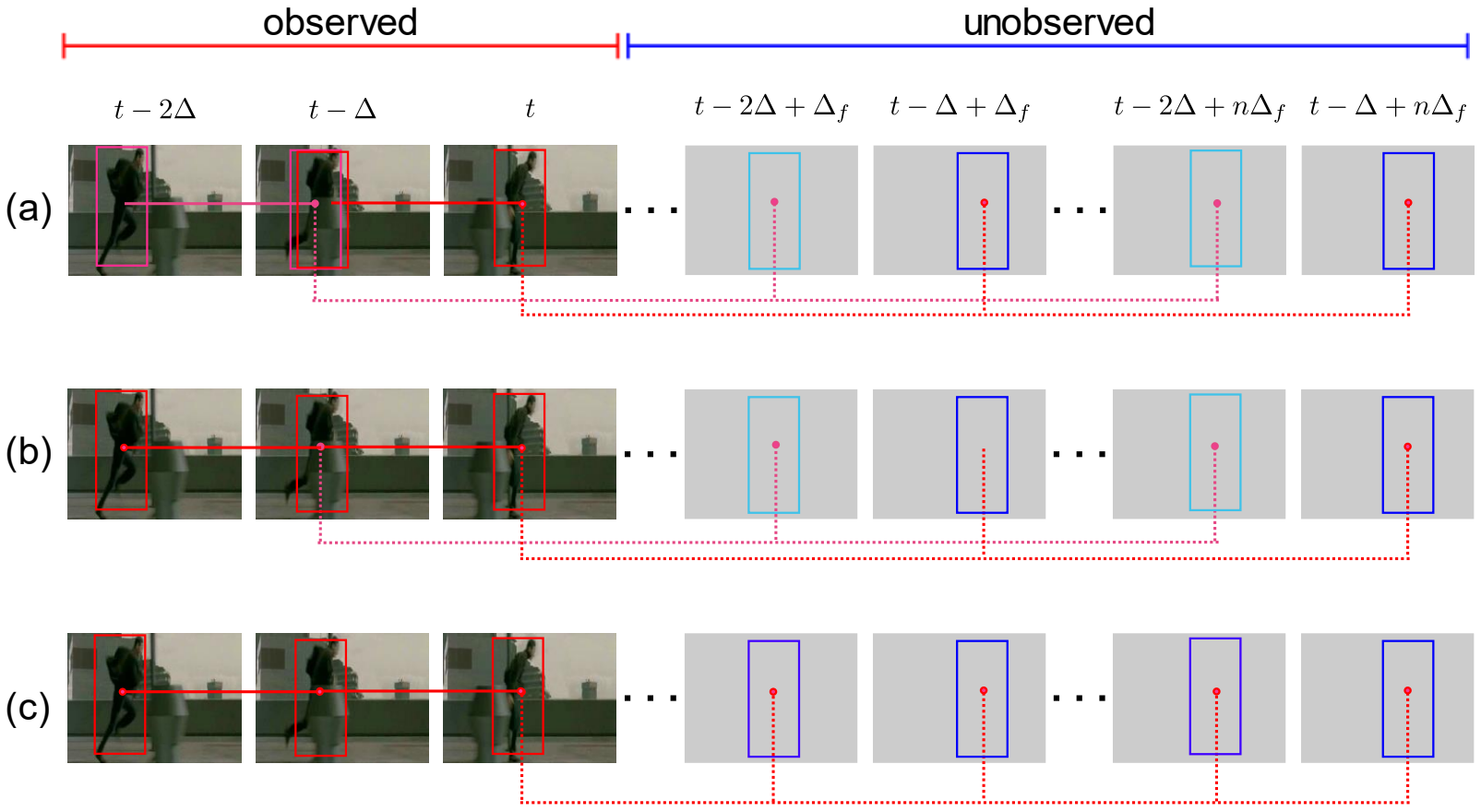


How linking micro-tubes affect the tube's future



Linking is done as in Singh et al. ICCV 2017, overlap and classification score are constraints

How linking micro-tubes affect the tube's future



Linking is done as in Singh et al. ICCV 2017, overlap and classification score are constraints

JHMDB-21 and parameter

- 21 atomic class
- Short videos up to 40 frames, i.e. 1.5 seconds
- Temporally trimmed videos

- 4 parameters to parameterize future and past locations
- Δ is the gap between two frames (both at training and testing) - we set it to 1;
- Δ_p is the gap between the first frame of the microtube and a past frame
- Δ_f is the gap between the first frame and a future frame
- n is number of future steps
- We cross-validated these parameters and showed their effect in our experiments

Action detection on JHMDB-21 (fusion aspect)

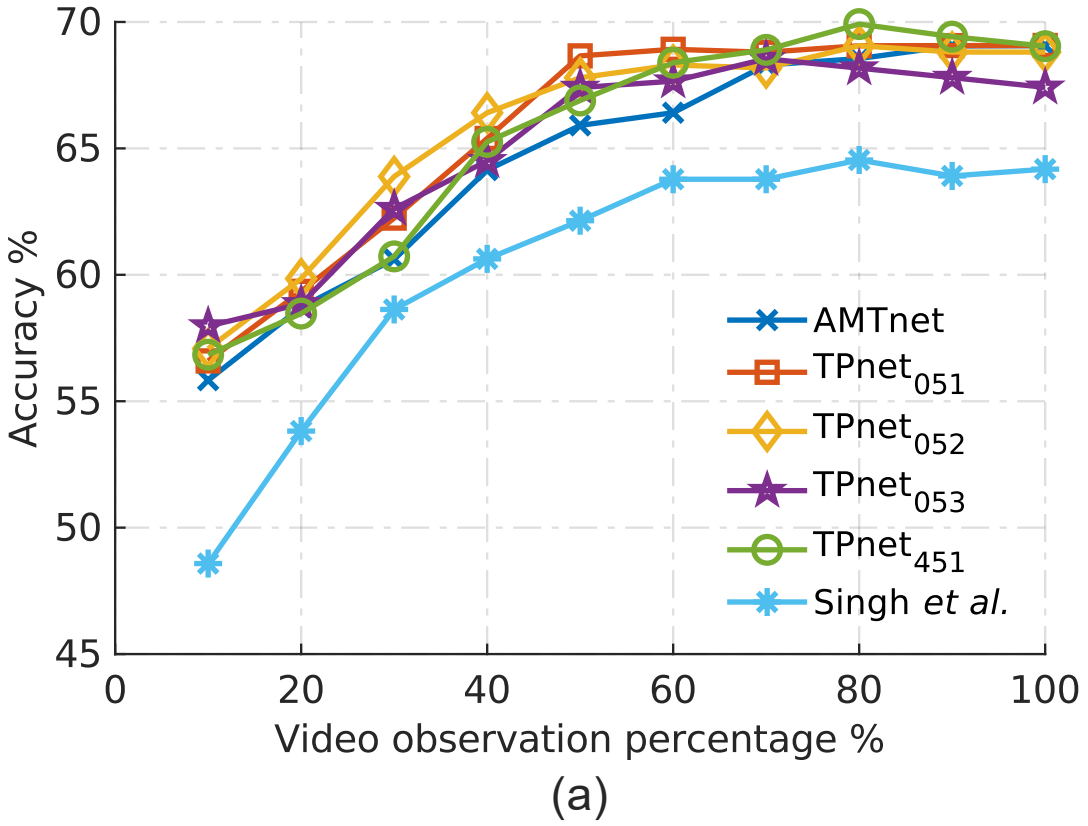
Method \\ threshold δ	0.2	0.5	0.75	0.5:0.95	Accuracy %
AMTnet Saha *et al. rgb ICCV2017*	57.7	55.3	--	--	--
ACT Kalegoton *et al.* [8]	74.2	73..7	52.1	44.8	61.7
T-CNN Hou *et al.* [9]	**78.4**	**76.9**	--	--	67.2
AMTnet-LateFusion	71.7	71.2	49.7	42.5	65.8
AMTnet-FeatFusion-Concat	73.1	72.6	**59.8**	**48.3**	68.4
AMTnet-FeatFusion-Sum	73.5	72.8	59.7	48.1	**69.6**

Action detection on JHMDB-21 (prediction aspect)

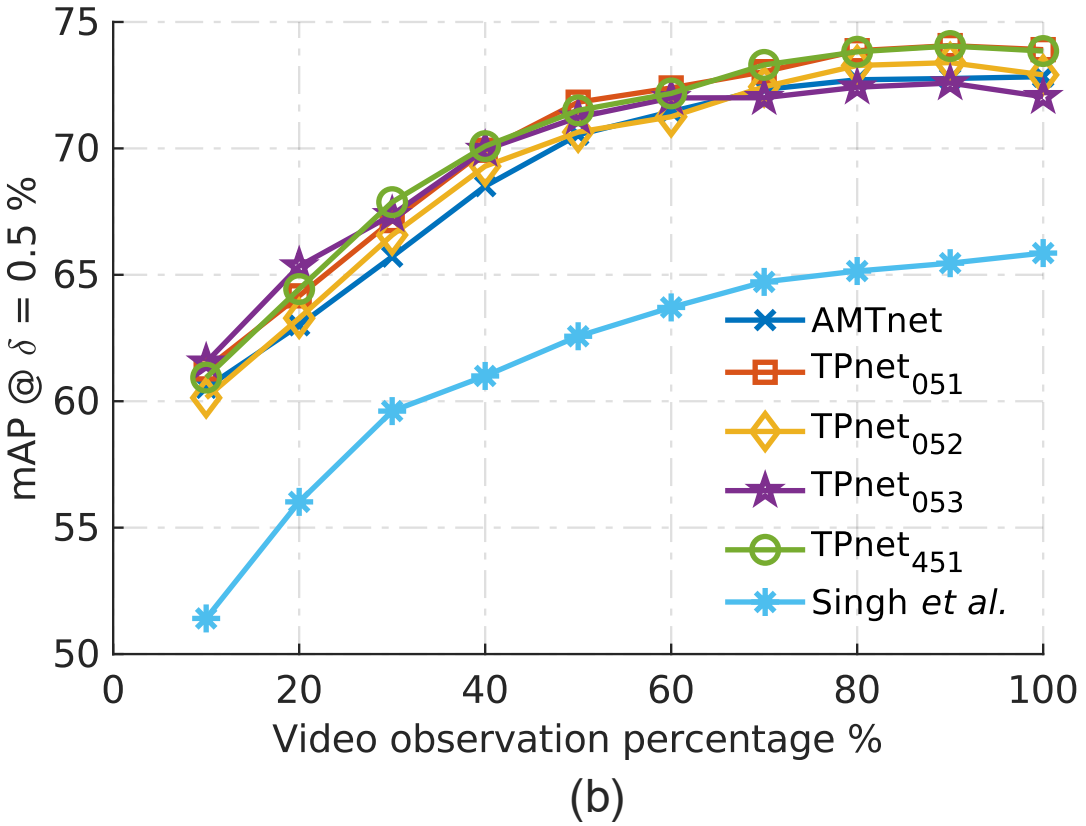
Method \\ threshold δ	0.2	0.5	0.75	0.5:0.95	Accuracy %
ACT Kalegoton *et al. ICCV2017*	74.2	73..7	52.1	44.8	61.7
AMTnet-FeatFusion-Sum	73.5	72.8	59.7	48.1	69.6
Ours TPnet – 053	72.6	71.2	58.0	42.5	65.8
Ours TPnet – 453	73.8	73.0	59.1	48.3	68.4
Ours TPnet – 051	74.6	73.1	60.5	49.0	**69.8**
Ours TPnet – 451	**74.8**	**74.1**	**61.3**	**49.1**	68.9

TPnet - abc represents our TPnet, where a = Δp, b = Δf and c = n.

Action prediction and online detection



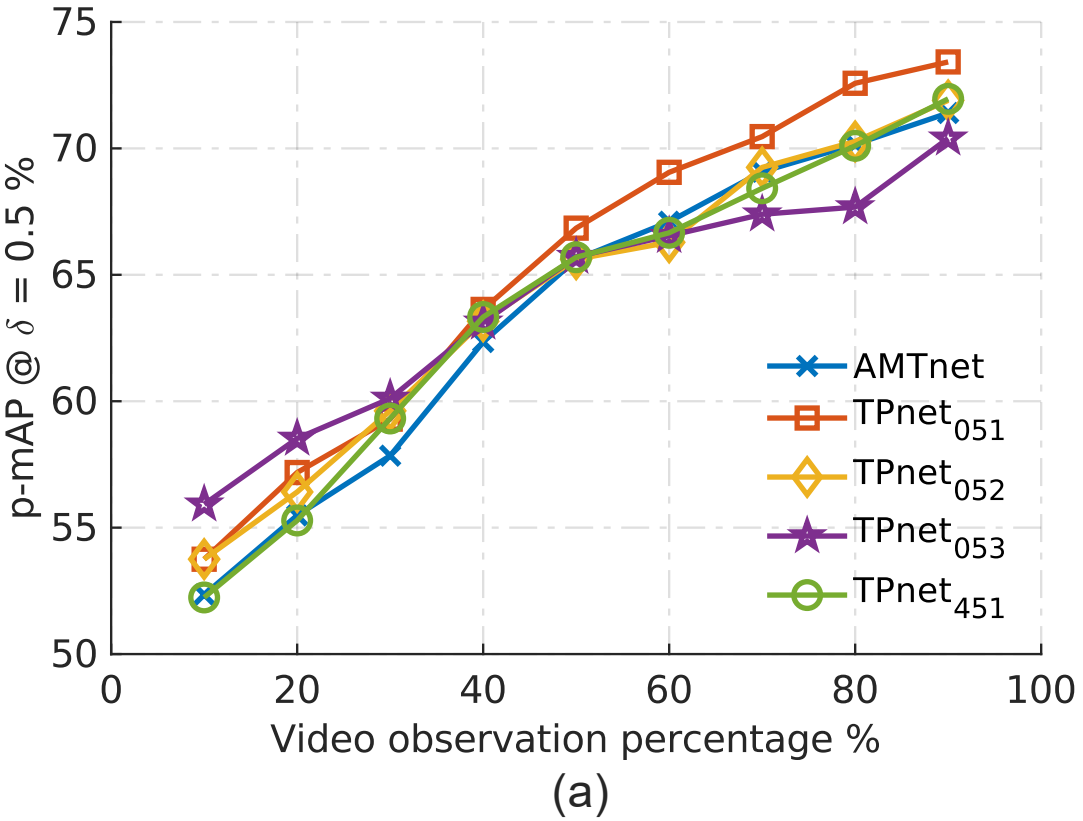
(a)



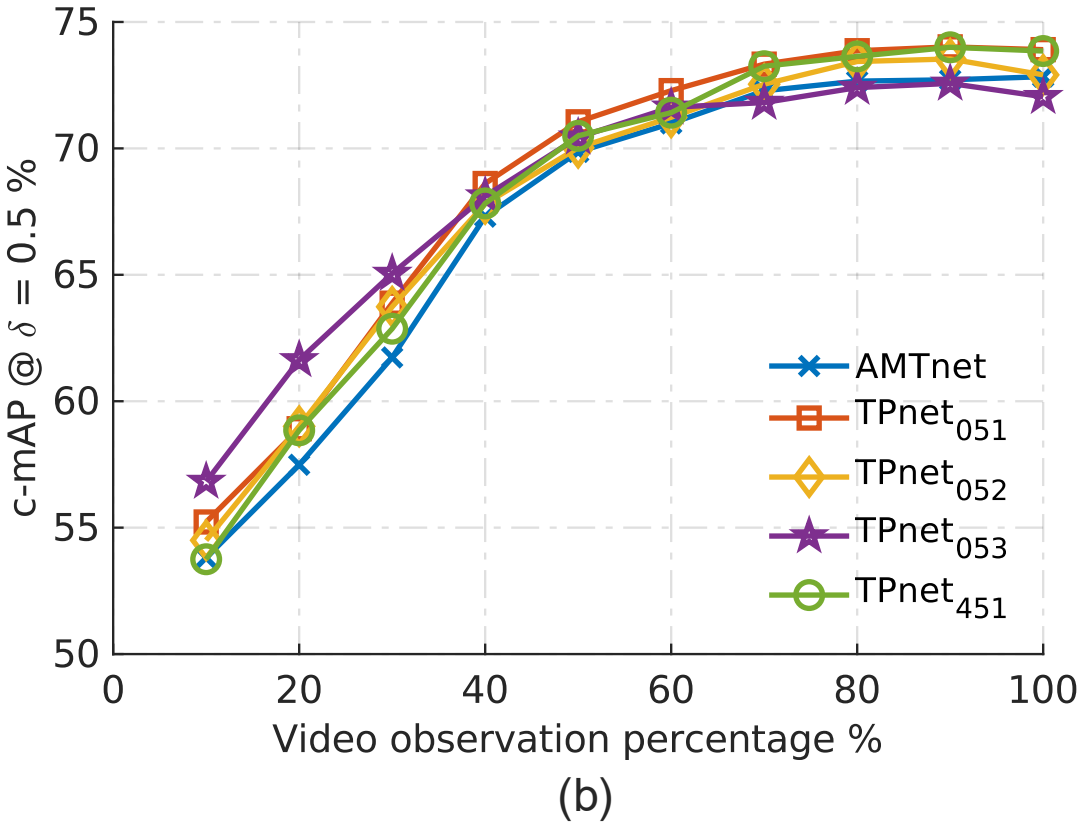
(b)

TPnet - abc represents our TPnet, where a = Δp, b = Δf and c = n.

Tube prediction results



(a)



(b)

TPnet - abc represents our TPnet, where a = Δp, b = Δf and c = n.

Limitations of the current method on J-HMDB

- Temporally trimmed videos
- Only short videos tried so far
- Single action per video BUT it can be easily extend to handle multiple action instances
- Path variation is small in the dataset, need for a more realistic dataset
- Lot of the trajectories are ‘linear’
- Only uses two frames for prediction; does not utilise all the information from the current tube

Conclusions

- Tube prediction is a holistic problem (detection, label and trajectory prediction)
- Regression can be useful for future prediction
- Past predictions are useful to regularise the shape of the action tubes
- Feature-level fusion is essential for better localisation

Thank you

Questions