上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

基于长期时序信息的视频时空行为定位识别

林巍峣

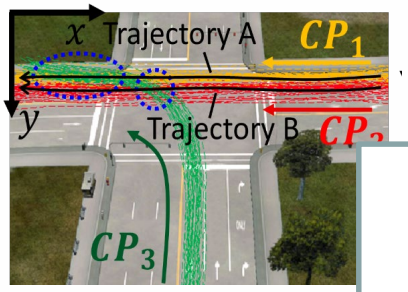
上海交通大学

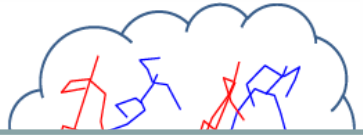
2020. 7. 28

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

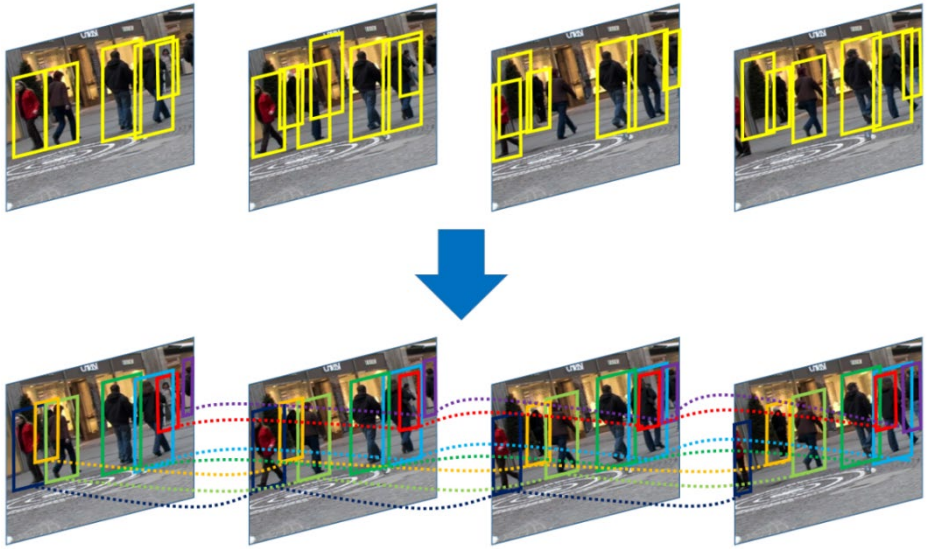
行为识别领域



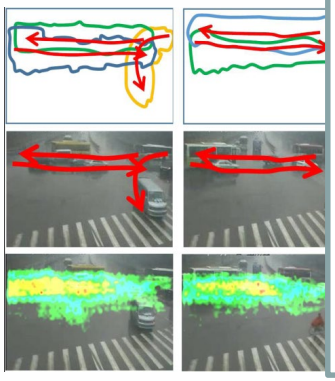


基于轨

的行为识别



目标感知、估计、与跟踪



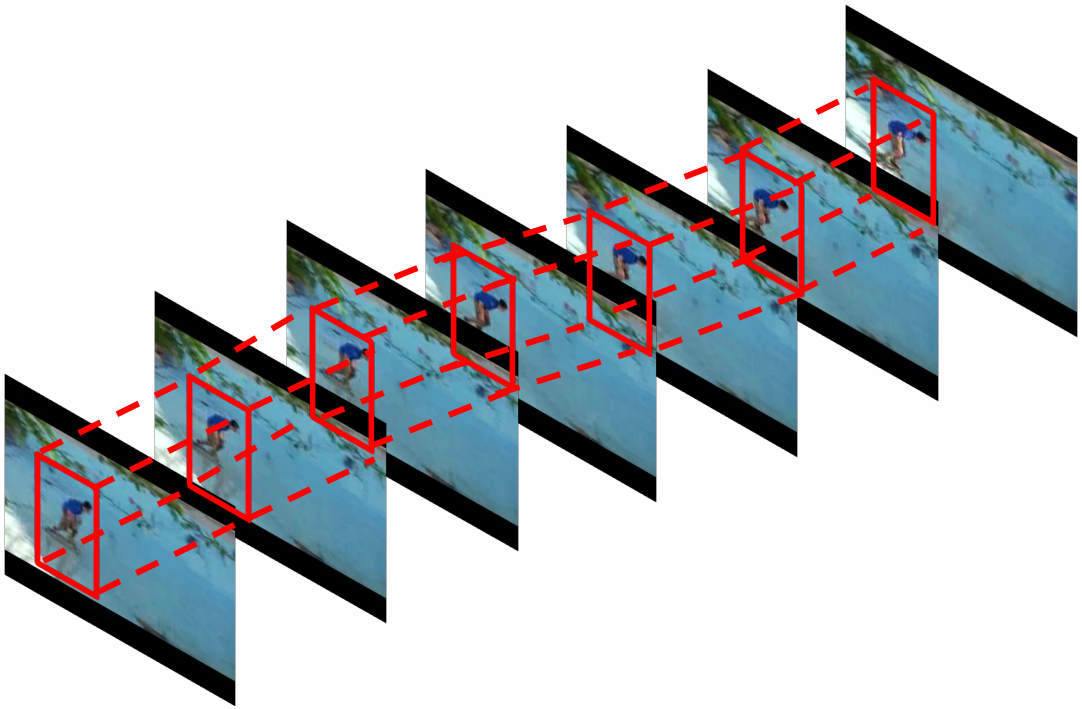
密集场景行为、场景分析　一般监控场景行为识别　异常行为识别

面向监控视频的行为识别

Introduction

- Spatial-Temporal Action Detection

Video Input ⇒ Tube Output



Introduction

Drawback of previous works

- Lack of long-term information
- High Complexity
- Vulnerable to noisy detections

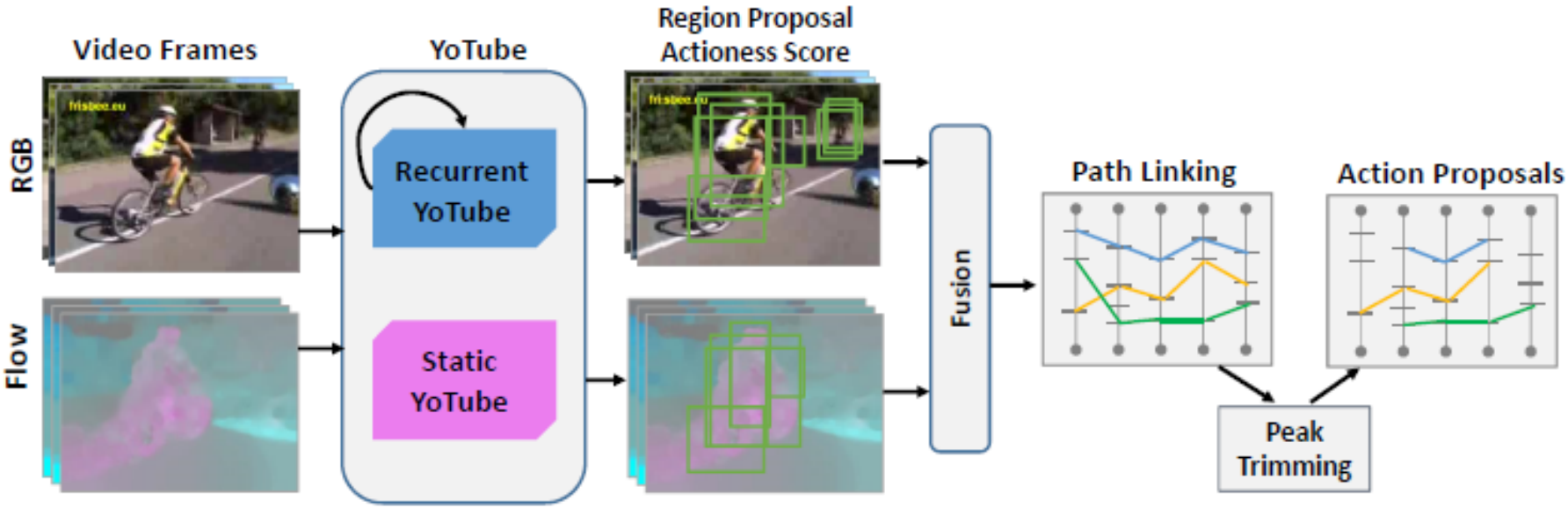
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

Related works

YoTube: Searching Action Proposal via Recurrent and Static Regression Networks

Hongyuan Zhu*, Romain Vial*, Shijian Lu,
Yonghong Tian, Xianbin Cao

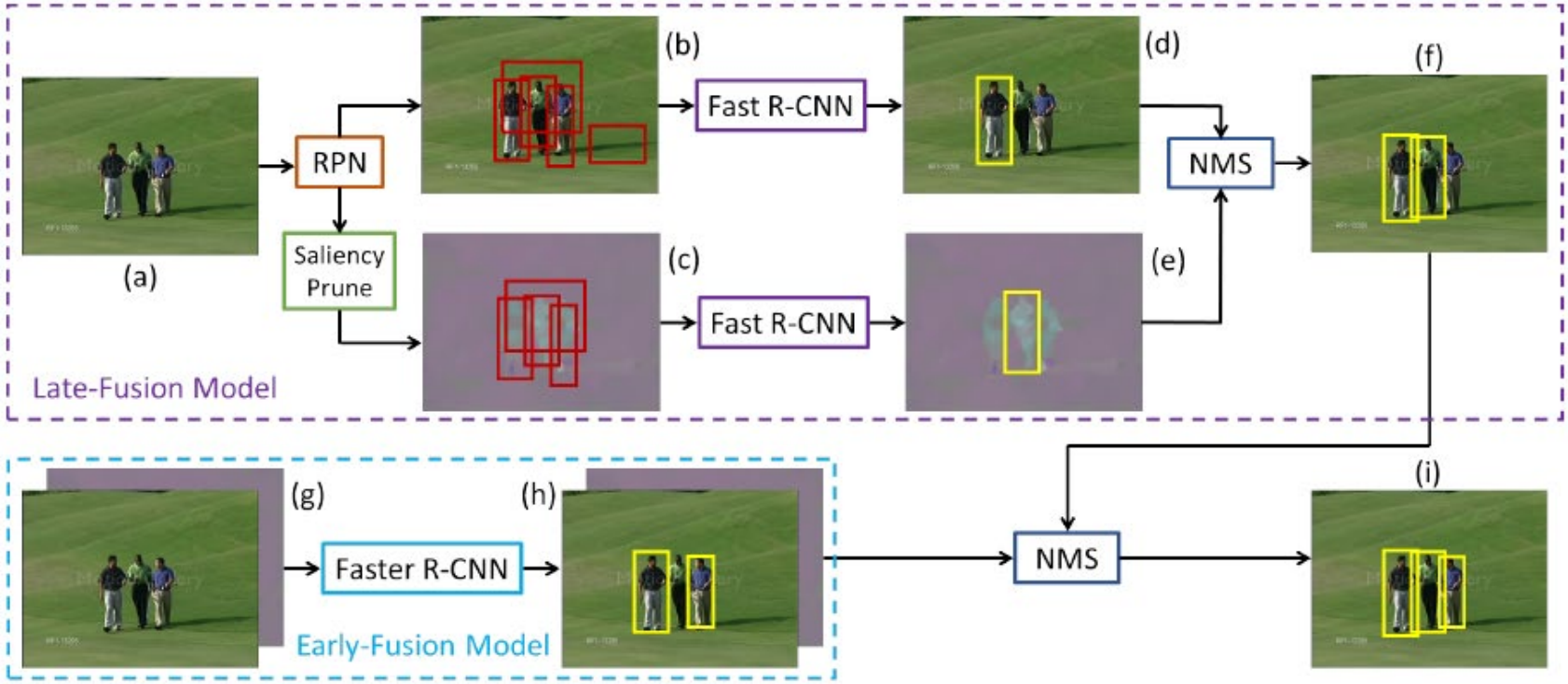


Related works

Discovering Spatio-Temporal Action Tubes

Yuancheng Ye[a], Xiaodong Yang[b], YingLi Tian[a,*]

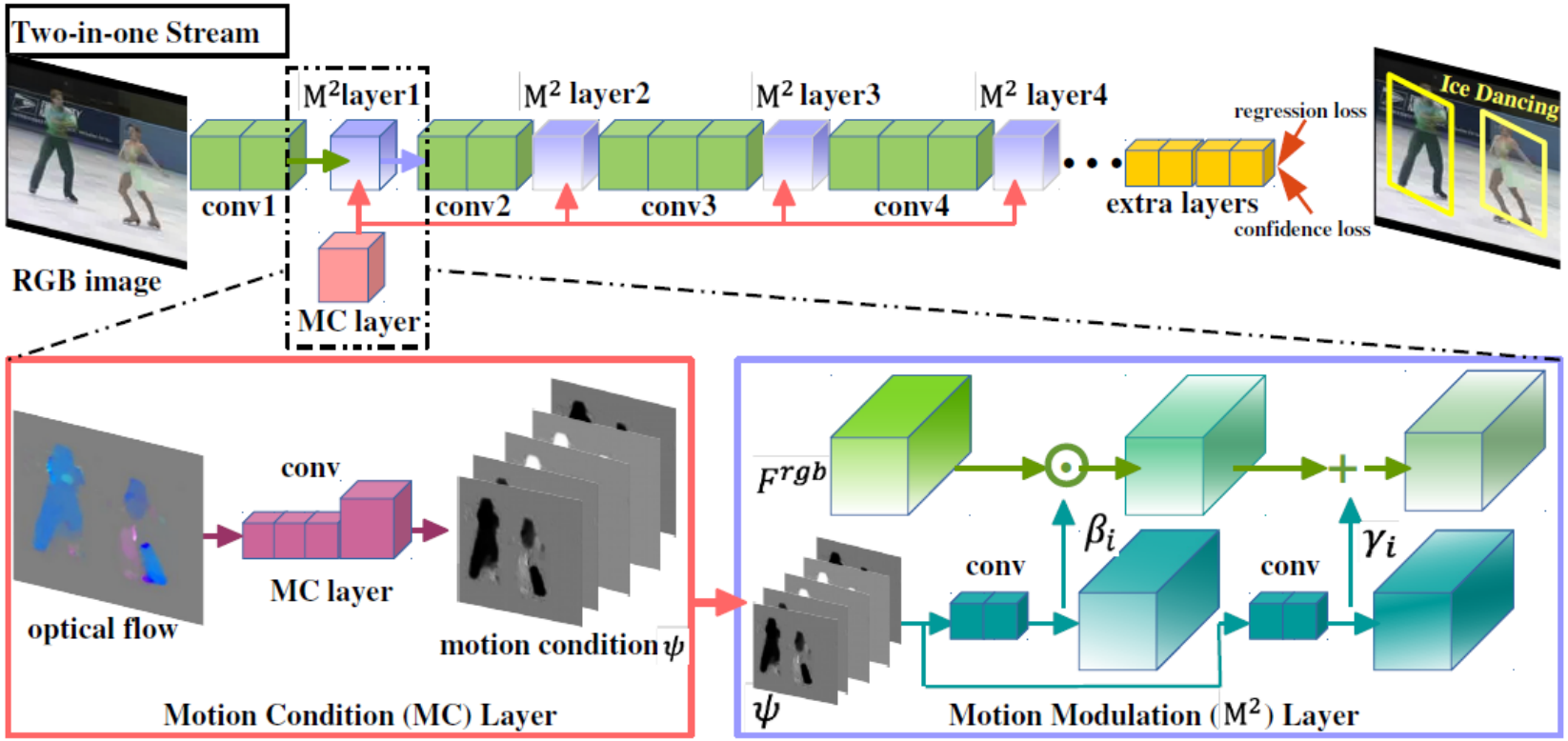
[a]*The City College and Graduate Center, City University of New York*
[b]*NVIDIA Research*



Related works

Dance with Flow: Two-in-One Stream Action Detection

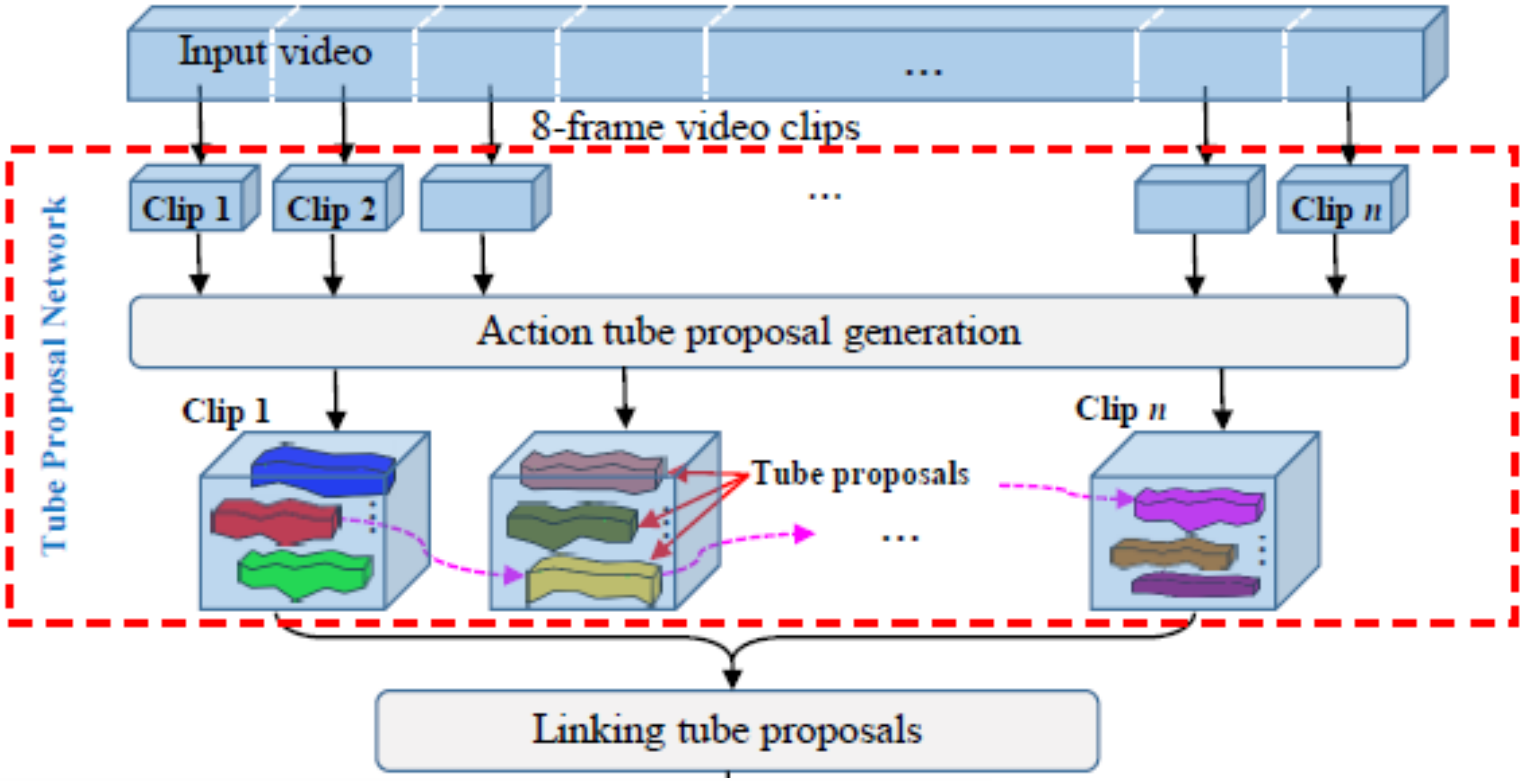
Jiaojiao Zhao and Cees G. M. Snoek
University of Amsterdam



Related works

An End-to-end 3D Convolutional Neural Network for Action Detection and Segmentation in Videos

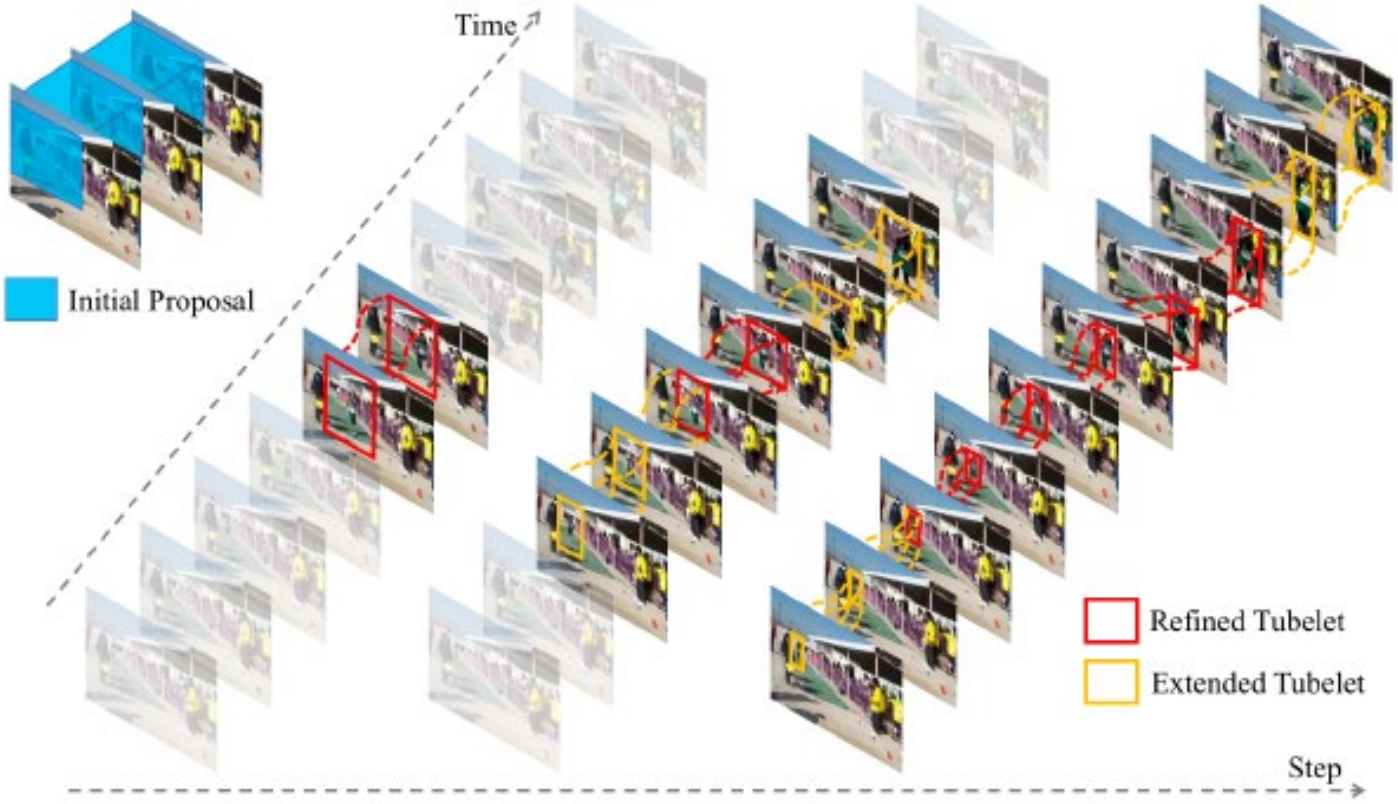
Rui Hou, *Student Member, IEEE,* Chen Chen, *Member, IEEE,* and Mubarak Shah, *Fellow, IEEE*



Related works

STEP: Spatio-Temporal Progressive Learning for Video Action Detection

Xitong Yang[1*] Xiaodong Yang[2] Ming-Yu Liu[2]
Fanyi Xiao[3*] Larry Davis[1] Jan Kautz[2]
[1]University of Maryland, College Park [2]NVIDIA [3]University of California, Davis



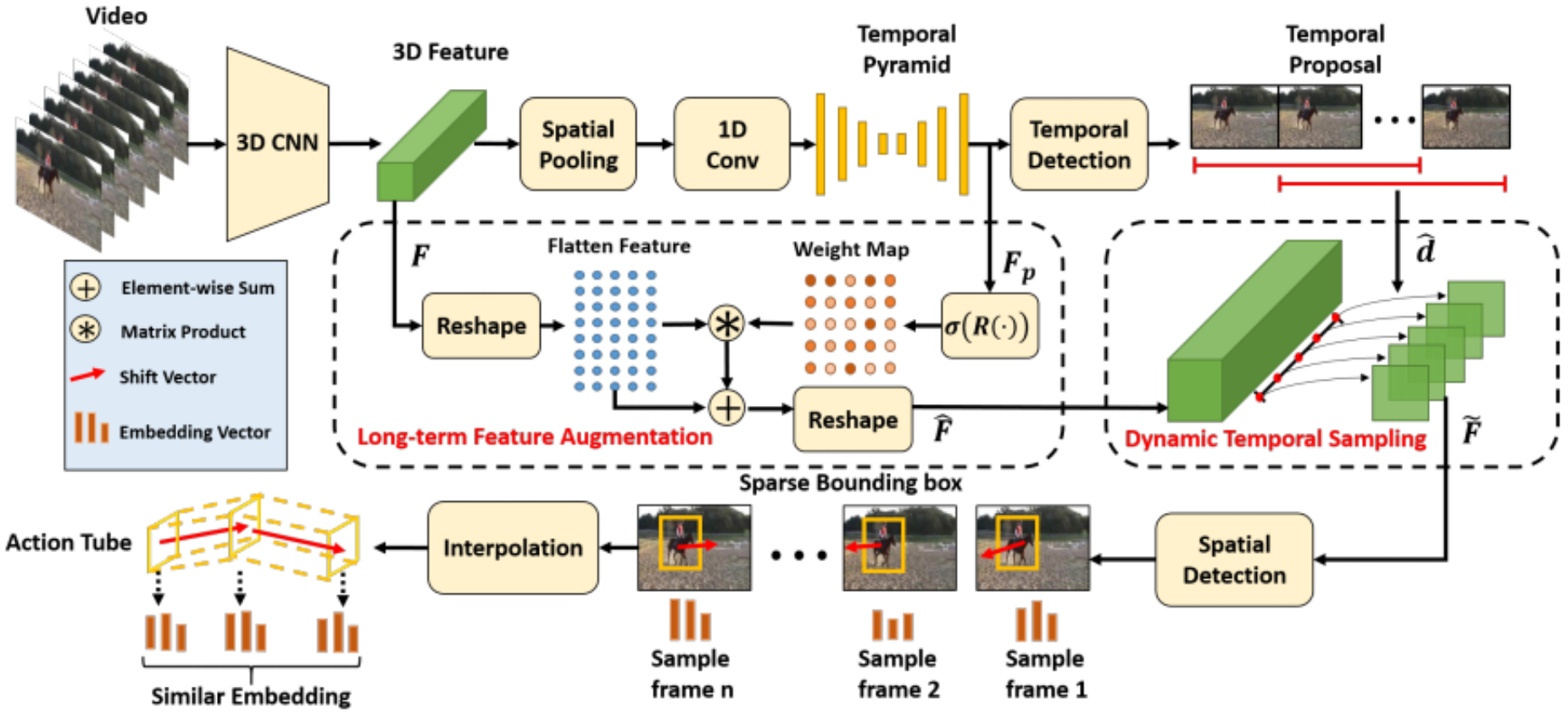
Our Solutions

- Integrate long-term information
- Select key frames to process
- Reduce the impact of noisy detections

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

Proposed Method



Feature Augmentation

3D feature extraction:

$$F = CNN(Video; \theta_C)$$

1D temporal feature:

$$F_p = TemporalPyramid(F; \theta_t)$$

Obtaining reweighting coefficients (attention):

$$S = \sigma(R(F_p; \theta_p))$$

Long-term feature augmentation:

$$F^a_{c,t,h,w} = F_{c,t,h,w} + \frac{1}{T_f}\sum_{k=1}^{T_f} S_{t,k} F_{c,k,h,w}$$

Dynamic Temporal Sampling

- Generate dynamicity ground-truth for learning

Algorithm 1: Ground-truth dynamic level generation

Input: Tube Φ, threshold ϵ
Output: dynamic level d

get bounding box set $\mathcal{B} = \{b_i | i = 1 \dots T\}$ from tube Φ ;
for $d = 1$ *to* T **do**
  $r = 0$;
  uniformly sample d bounding boxes from $\mathcal{B}$ and obtain sampled box set $\mathcal{P} = \{b_{t_k} | t_k = \lfloor kT/d \rfloor, k = 1 \dots d\}$;
  Interpolate over $\mathcal{P}$ to obtain $\hat{\mathcal{B}} = \{\hat{b}_i | i = 1 \dots T\}$;
  for $k = 1$ *to* T **do**
    $r = r + IOU(b_k, \hat{b}_k)$;
  end
  $r = r/T$;
  if $r \geq \epsilon$ **then**
    return d;
  end
end

Training:

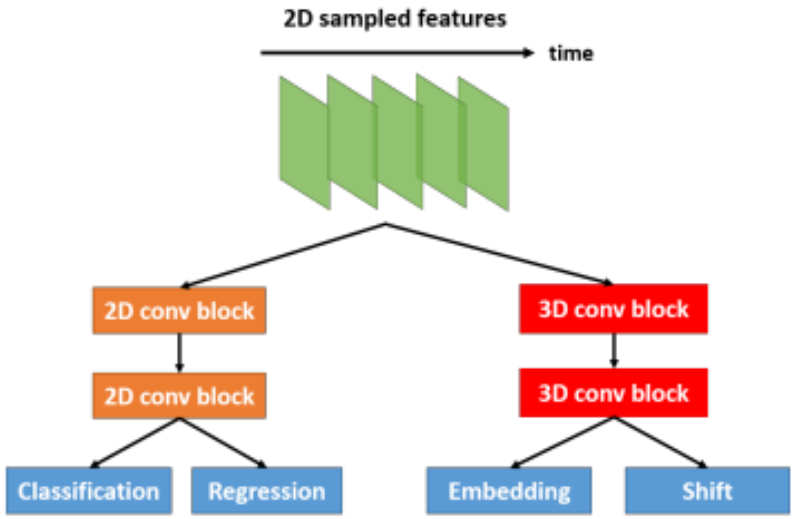
$$L_d = smooth\_L_1(d - \hat{d}; \gamma)$$

Inference:

$$\widetilde{\mathcal{F}}^i_{c,h,w} = \sum_{k=1}^{T_f} \hat{\mathcal{F}}_{c,k,h,w} \max\left(0, 1 - \left| s + \frac{e-s}{n-1} i - k \right|\right)$$

Detection & Linking

Detection



classification loss
box regression loss

clustering loss
shift regression loss

$$L_a = \sum_i C_i \mathbf{d}_2(\bar{\mathbf{f}}, \mathbf{f}_i) + (1 - C_i)[\alpha - \mathbf{d}_2(\bar{\mathbf{f}}, \mathbf{f}_i)]_+$$

$$L_m = \sum_{k=2}^{n} \sum_i C_i \mathbf{d}_2(\delta_i, \delta_{gt,k})$$

Linking

Similarity between boxes on adjacent sampled frames:

$$D_{a,tij} = ||\mathbf{f}_{ti} - \mathbf{f}_{(t+1)j}||_2$$
$$D_{s,tij} = ||\delta_{ti} - \bar{\delta}_{tij}||_2$$

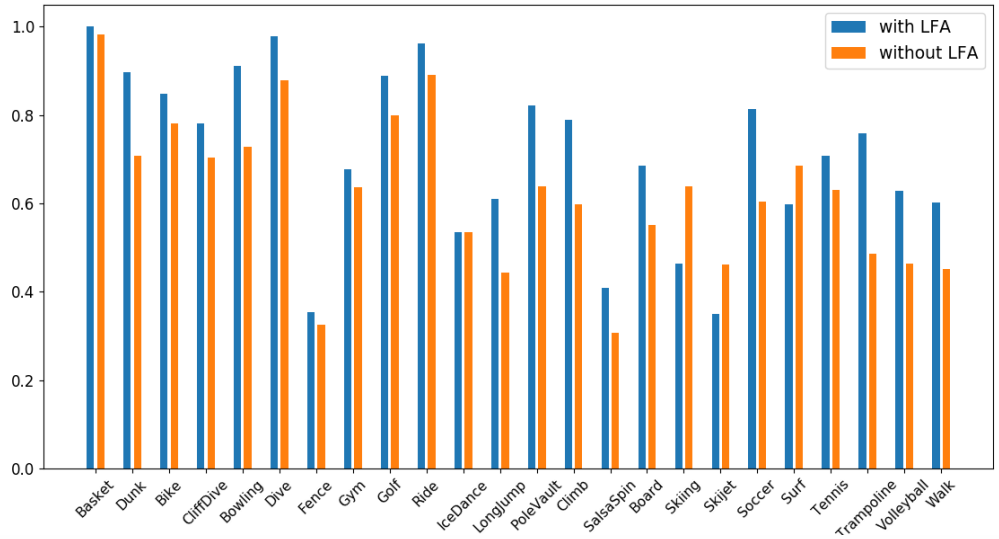
$$\mathcal{C}\left(\hat{b}_{ti}, \hat{b}_{(t+1)j}\right) = \exp\left(-\frac{D_{a,tij} + D_{s,tij}}{2}\right)$$

Link adjacent boxes according to similarity:

$$\hat{b}_{(t+1)} = \arg\max_{\hat{\mathcal{B}}_{(t+1)}} \mathcal{C}\left(\hat{b}_t, \hat{b}_{(t+1)j}\right)$$

Experiment

- Datasets: UCF101-24, JHMDB-21, UCFSports
- Ablation Study:

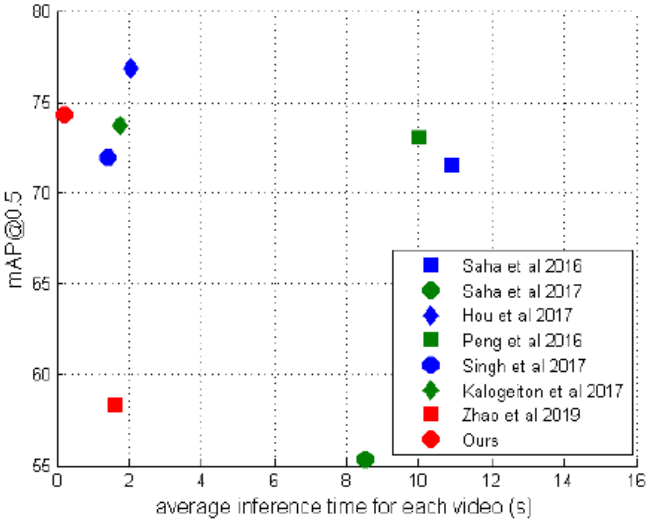


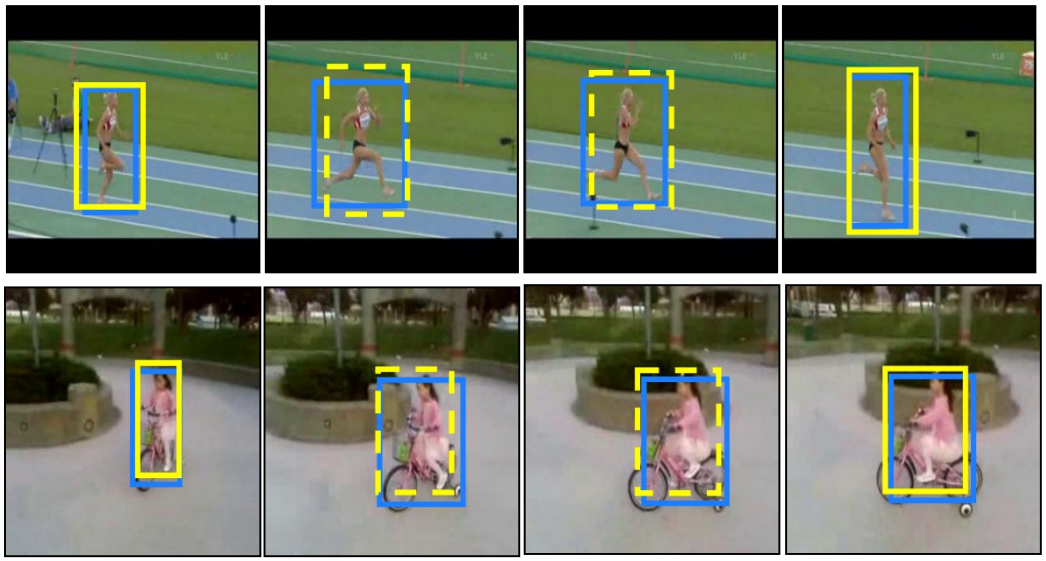
LFA		✓	✓	✓	✓
Embedding	✓			✓	✓
Shift	✓		✓		✓
v-mAP@0.3	62.2	67.4	68.9	69.6	**71.1**

scheme	v-mAP@0.3	per-video time (s)
fixed point-avg	67.4	0.551
fixed point-4	63.3	0.412
fixed point-10	67.5	0.654
fixed step-avg	66.9	0.571
DTS	**71.1**	**0.569**

Comparison with SOTA

method	input	JHMDB-21			UCFSports			UCF101-24		
Δ		0.2	0.5	0.75	0.2	0.5	0.75	0.3	0.5	0.75
(Saha et al. 2016)	RGB+Flow	72.6	71.5	-	-	-	-	54.9	35.9	-
(Peng and Schmid 2016)	RGB+Flow	74.3	73.1	48.2	94.8	94.7	47.3	65.7	32.1	2.7
(Kalogeiton et al. 2017)	RGB+Flow	74.2	73.7	52.1	92.7	92.7	78.4	-	49.2	19.7
(Hou, Chen, and Shah 2017)	RGB+Flow	78.4	76.9	-	95.2	95.2	-	69.4	-	-
(Singh et al. 2017)	RGB+Flow	73.8	72.0	44.5	-	-	-	-	46.3	15.0
(Yang, Gao, and Nevatia 2017)	RGB+Flow	-	-	-	-	-	-	60.7	37.8	-
(Song et al. 2019)	RGB+Flow	74.1	73.4	52.5	-	-	-	-	52.9	21.8
(Zhao and Snoek 2019)	RGB+Flow	-	58.0	42.8	-	92.7	**83.4**	-	48.3	**22.1**
(Li et al. 2018)	RGB+Flow	**82.3**	**80.5**	-	**97.8**	**97.8**	-	**70.9**	-	-
(Saha et al. 2016)	RGB	52.9	51.3	-	-	-	-	48.3	30.7	-
(Saha, Singh, and Cuzzolin 2017)	RGB	57.8	55.3	-	-	-	-	51.7	33.0	0.5
(Li et al. 2018)	RGB	-	61.7	-	-	87.6	-	-	-	-
Ours	RGB	**76.1**	**74.3**	**56.4**	**94.3**	**93.8**	79.5	**71.1**	**54.0**	21.8



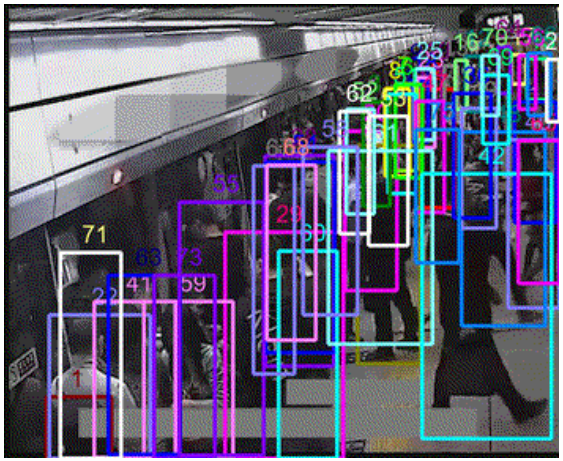


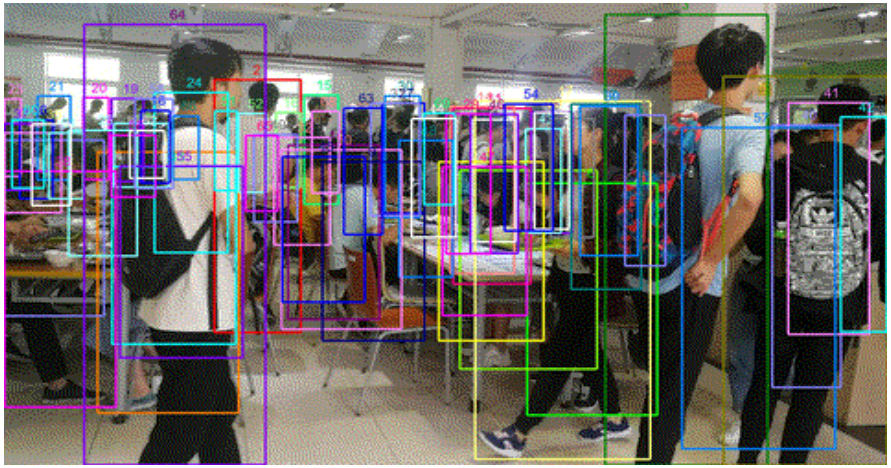
Reference

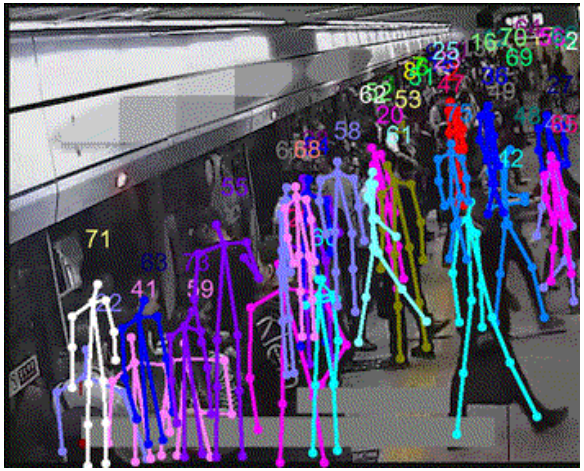
- Paper
  - ✓ Y. Li, W. Lin, T. Wang, J. See, R. Qian, N. Xu, L. Wang, S. Xu. “Finding Action Tubes with a Sparse-to-Dense Framework,". Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), 2020.

Challenge on Large-scale Human-centric Video Analysis in Complex Events (ACM MM' 20 Grand Challenge)

Challenge网页，http://humaninevents.org/





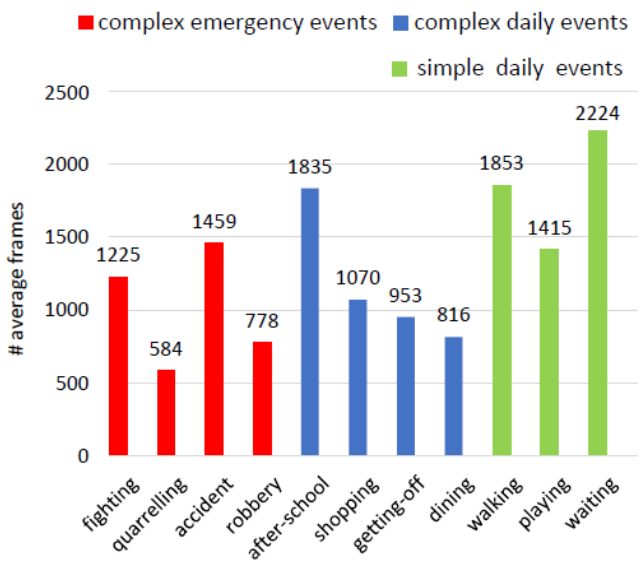


Challenge on Large-scale Human-centric Video Analysis in Complex Events (HiEve，MM’ 20 Grand Challenge)

Challenge网页，http://humaninevents.org/ 8月重新开放！

Challenge on Large-scale Human-centric Video Analysis in Complex Events (HiEve)

Dataset	# pose	# box	# traj.(avg)	# action	pose track	surveillance	complex events
MSCOCO [1]	105,698	105,698	NA	NA	×	×	×
MPII [4]	14,993	14,993	NA	410	×	×	×
CrowdPose [5]	~80,000	~80,000	NA	NA	×	×	×
PoseTrack [2]	~267,000	~26,000	5,245(49)	NA	√	×	×
MOT16[6]	NA	292,733	1,276(229)	NA	×	√	×
MOT17	NA	901,119	3,993(226)	NA	×	√	×
MOT20 [7]	NA	1,652,040	3457(478)	NA	×	√	×
Avenue [8]	NA	NA	NA	15	×	√	×
UCF-Crime [3]	NA	NA	NA	1,900	×	√	√
Ours	**1,099,357**	1,302,481	2,687(485)	**56,643**	√	√	√

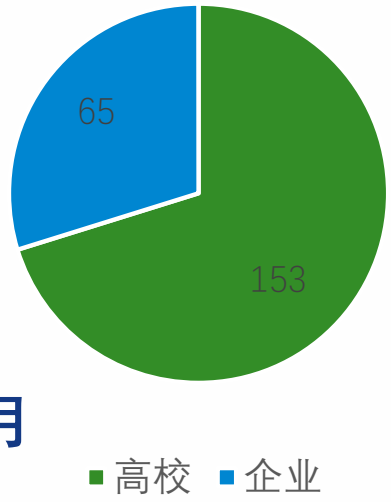


报名参赛队：218

有效提交结果队：125

结果公布：6月30日

Challenge重新开放：8月



Challenge 网址： http://humanineventsorg/ ， 将于8月重新开放！

相关论文：https://arxiv.org/abs/2005.04490

For More Information:

https://weiyaolin.github.io/

Thank You!