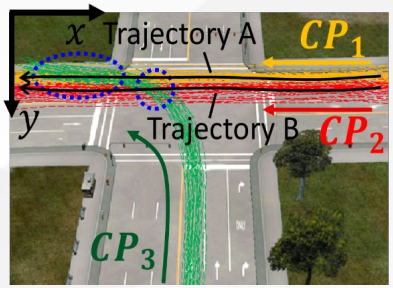
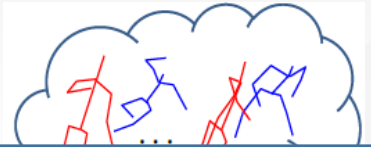
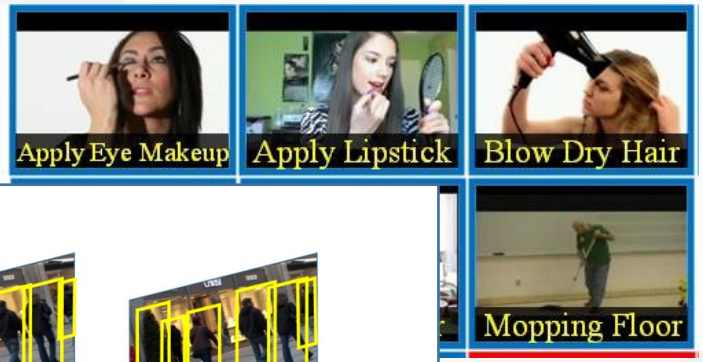
开放环境下的视频行为感知问题讨论

林巍峣
上海交通大学

行为识别领域

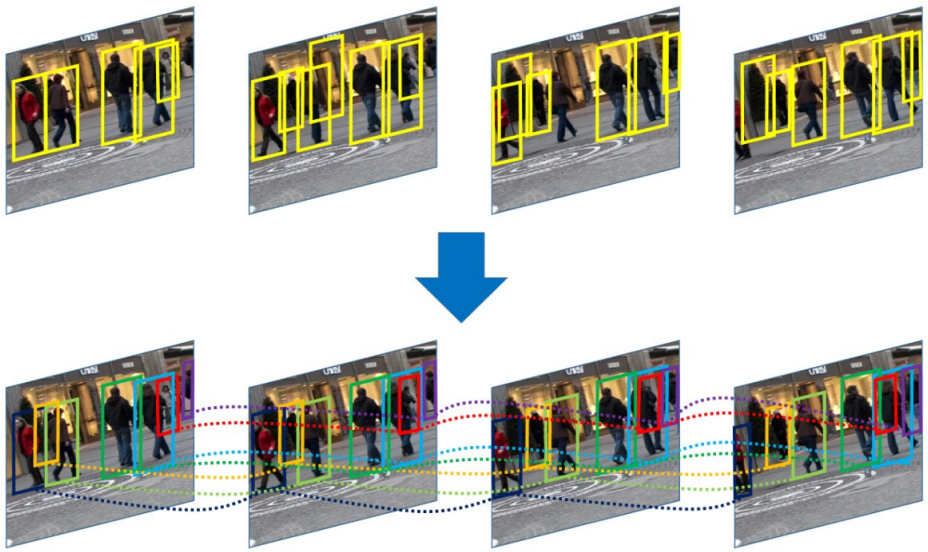




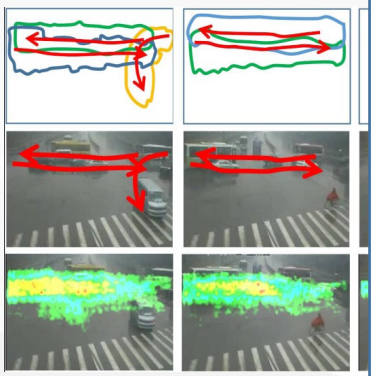


基于轨迹

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

主要思路和问题

技术路径

- 从下向上的感知（检测、跟踪、三维骨架）
- 少样本/无监督
- 多模态 （音频+视频、文本+视频）

行为类型

- 个人行为
- 组群行为
- 其他

主要思路和问题

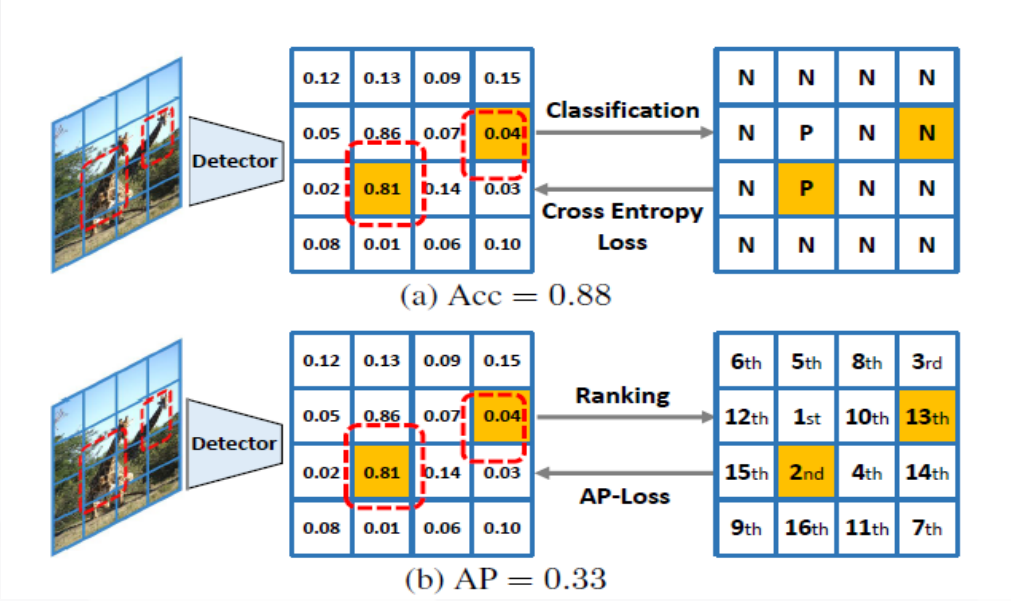
技术思路

- **从下向上的感知（检测、跟踪、三维骨架）**
- 少样本/无监督
- 多模态 （音频+视频、文本+视频）

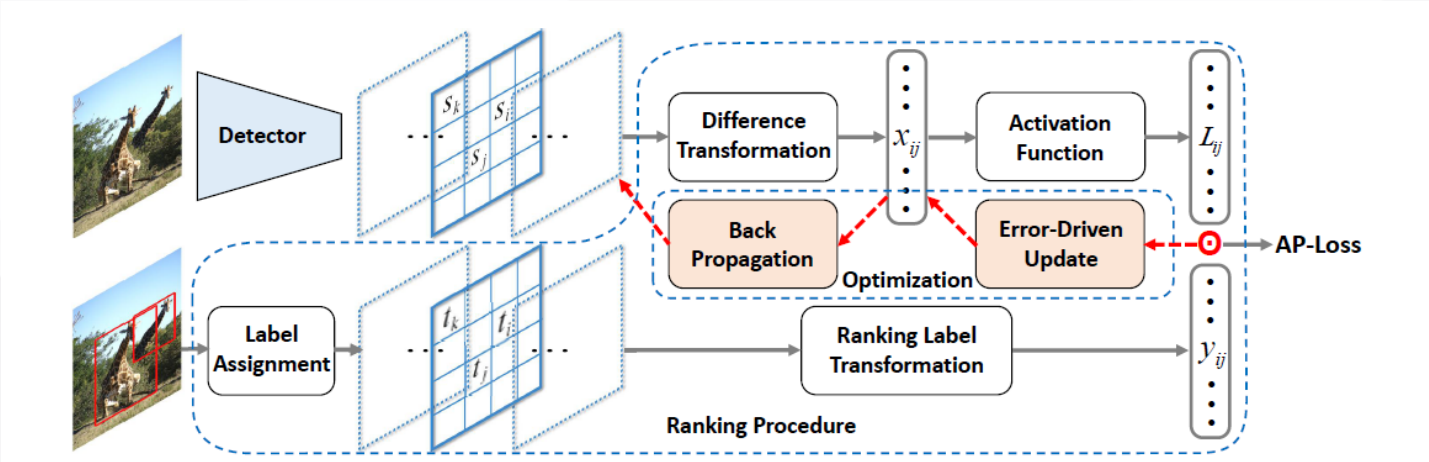
行为类型

- 个人行为
- 组群行为
- 其他

目标检测 (AP-Loss)

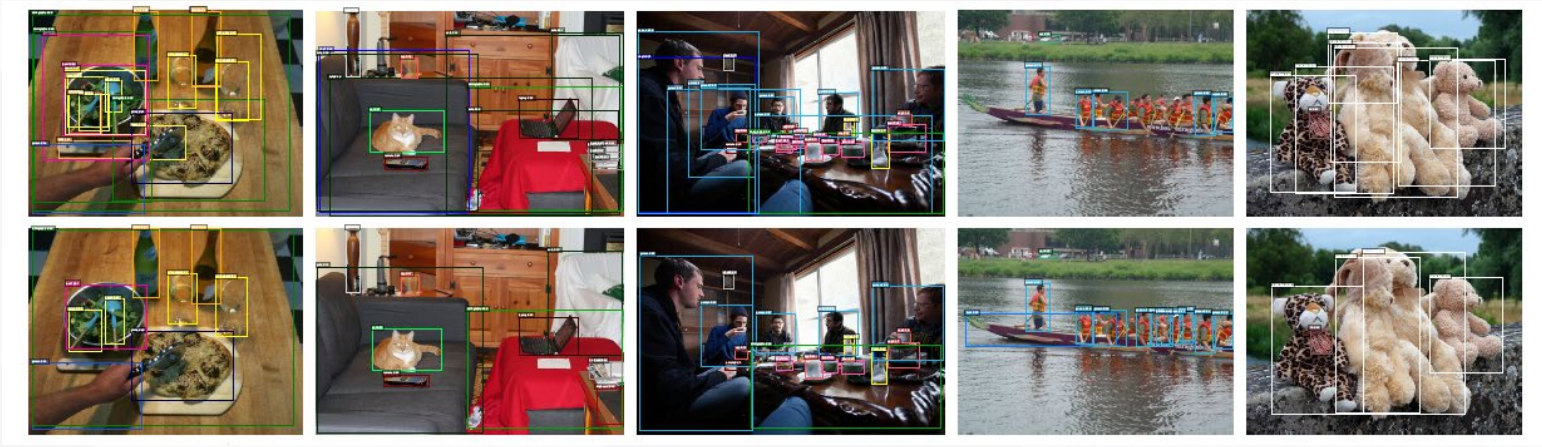


分类Loss受前景背景样本不平衡影响



AP loss for Accurate One-Stage Object Detection, TPAMI, 2020

目标检测 (AP-Loss)



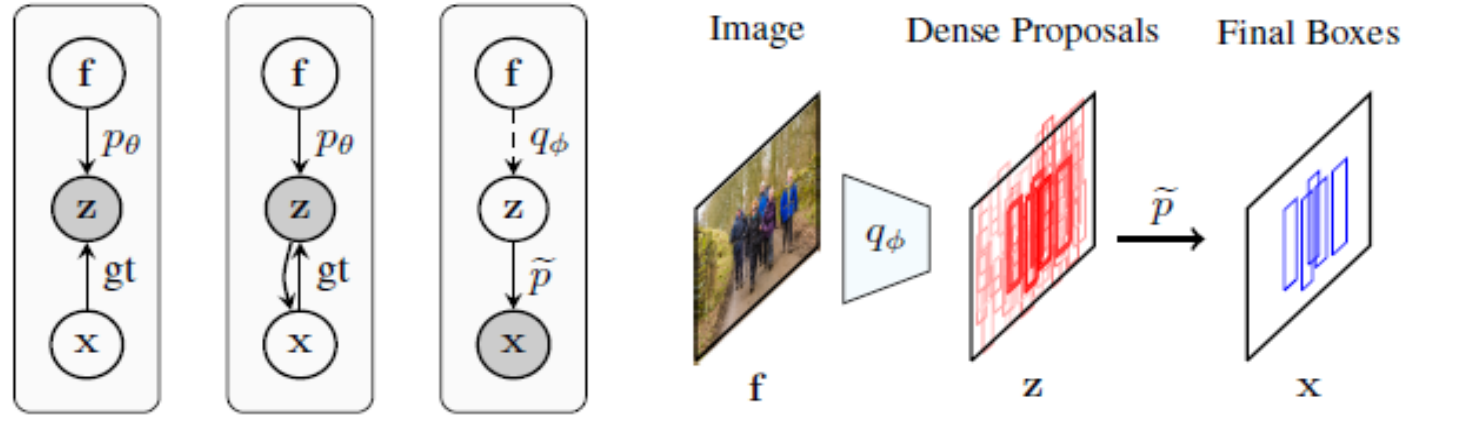
Method	Backbone	Multi-Scale	VOC07	VOC12	COCO					
			AP_{50}	AP_{50}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv2 [21]	DarkNet-19	✗	78.6	73.4	21.6	44.0	19.2	5.0	22.4	35.5
DSOD300 [26]	DS/64-192-48-1	✗	77.7	76.3	29.3	47.3	30.6	9.4	31.5	47.0
SSD512 [16]	VGG-16	✗	79.8	78.5	28.8	48.5	30.3	-	-	-
SSD513 [5]	ResNet-101	✗	80.6	79.4	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [5]	ResNet-101	✗	81.5	80.0	33.2	53.3	35.2	13.0	35.4	51.1
DES512 [34]	VGG-16	✗	81.7	80.3	32.8	53.2	34.6	13.9	36.0	47.6
RFBNet512 [15]	VGG-16	✗	82.2	-	33.8	54.2	35.9	16.2	37.1	47.4
PFPNet-R512 [9]	VGG-16	✗	82.3	80.3	35.2	57.6	37.9	**18.7**	38.6	45.9
RefineDet512 [33]	VGG-16	✗	81.8	80.1	33.0	54.5	35.5	16.3	36.3	44.3
RefineDet512 [33]	ResNet-101	✗	-	-	36.4	57.5	39.5	16.6	39.9	51.4
RetinaNet500 [13]	ResNet-101	✗	-	-	34.4	53.1	36.8	14.7	38.5	49.1
RetinaNet500+AP-Loss (ours)	ResNet-101	✗	**83.9**	**83.1**	**37.4**	**58.6**	**40.5**	17.3	**40.8**	**51.9**
PFPNet-R512 [9]	VGG-16	✓	84.1	83.7	39.4	61.5	42.6	25.3	42.3	48.8
RefineDet512 [33]	VGG-16	✓	83.8	83.5	37.6	58.7	40.8	22.7	40.3	48.3
RefineDet512 [33]	ResNet-101	✓	-	-	41.8	62.9	45.7	**25.6**	**45.1**	**54.1**
RetinaNet500+AP-Loss (ours)	ResNet-101	✓	**84.9**	**84.5**	**42.1**	**63.5**	**46.4**	**25.6**	45.0	53.9

Towards accurate one-stage object detection with AP-loss, CVPR, 2019

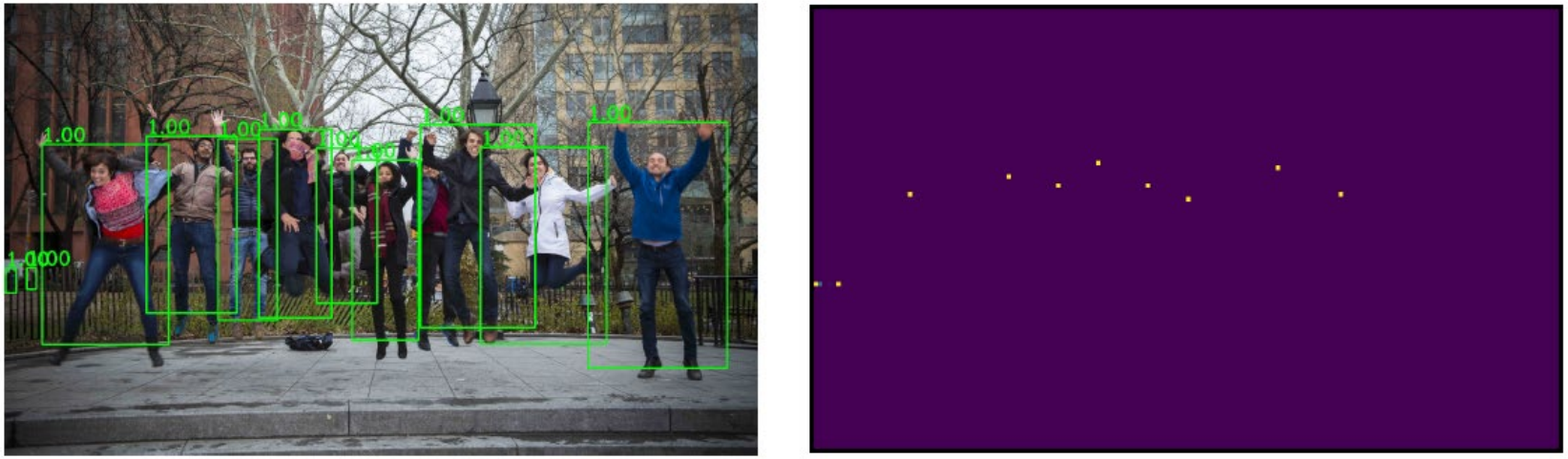
目标检测 (Patch-of-interest composition)

Kill two birds with one stone: boosting both object detection accuracy and speed with adaptive patch-of-interest composition, 2017.

密集遮挡下的检测



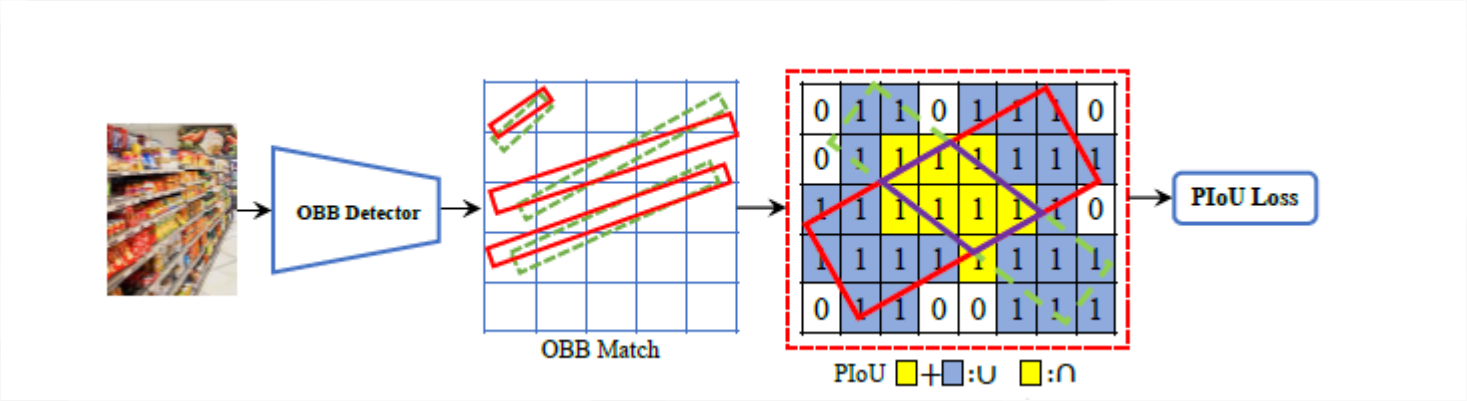
(a) Offline. (b) Online. (c) Ours. (d) The three variables in single-stage object detection.



(a) Detection result (no NMS). (b) Score map on FPN P_3.

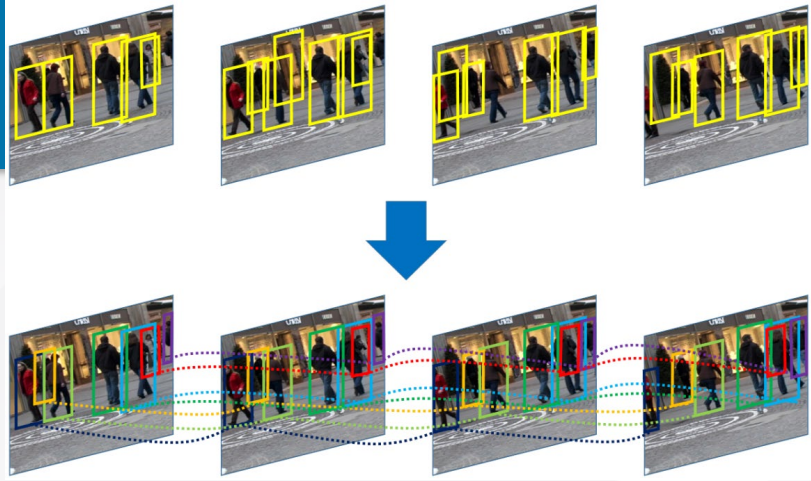
Variational Pedestrian Detection, CVPR, 2021.

旋转目标检测



PIoU Loss: Towards Accurate Oriented Object Detection in Complex Environments, ECCV, 2020.

多目标跟踪



$$(X^*, Y^*) = \arg\max_{X,Y} \sum_{k=1}^{n_h} \frac{\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} S_{ij} x_{ik} y_{jk} + \sum_{i=1}^{n_t}\sum_{j=1}^{n_t} S_{ij} x_{ik} x_{jk} + \sum_{i=1}^{n_t}\sum_{j=1}^{n_t} S_{ij} y_{ik} y_{jk}}{\sqrt{\sum_{i=1}^{n_t} x_{ik} + \sum_{j=1}^{n_t} y_{jk} + 1}}$$

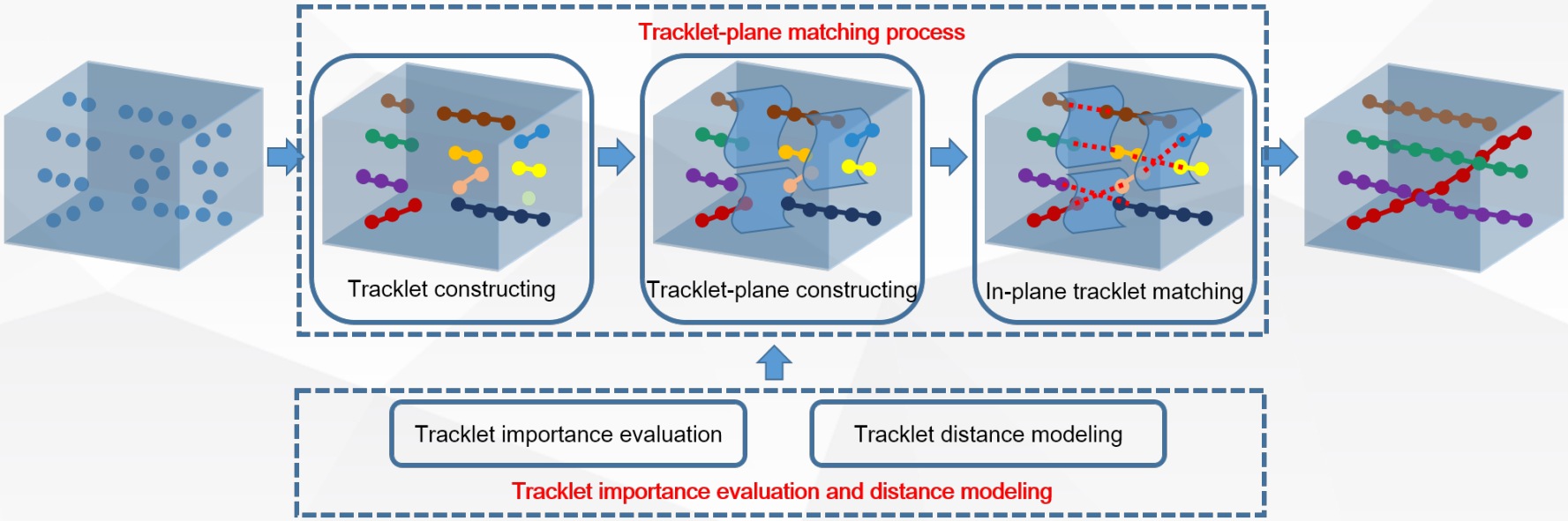
$$s.t. \begin{cases} x_{ik}, y_{jk} \in \{0,1\} \\ \sum_{i=1}^{n_h} x_{ik} \le 1 \\ \sum_{j=1}^{n_h} y_{jk} \le 1 \end{cases}$$

每个tracklet的一端只和一个超平面关联**

超平面两端之间的轨迹相似度

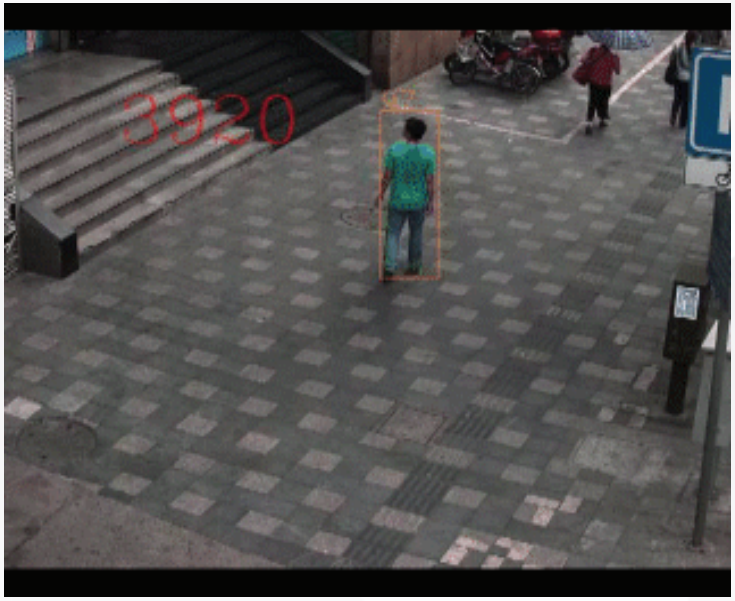
超平面同一端之间的轨迹相似度

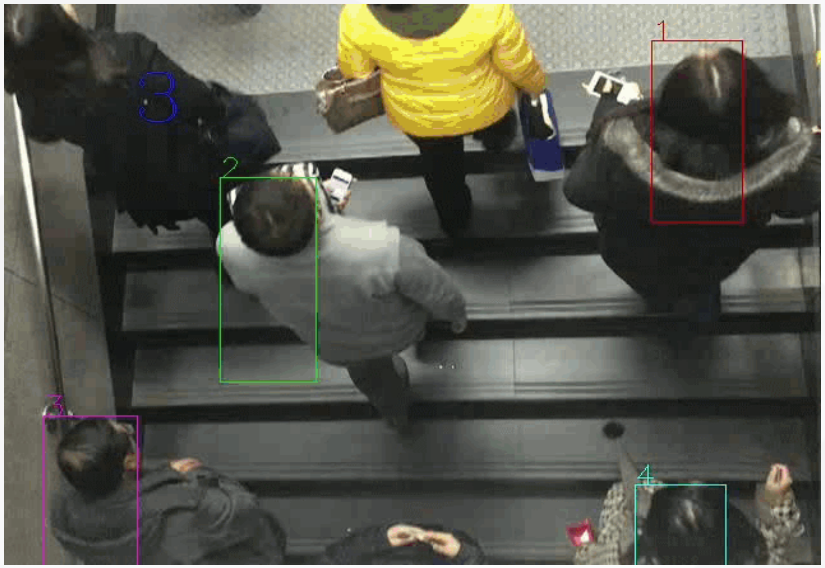
高置信度短轨迹数： n_t　超平面数： $n_h = n_t / 16$



Spatio-Temporal Point Process for Multiple Object Tracking, TNNLS, 2020.

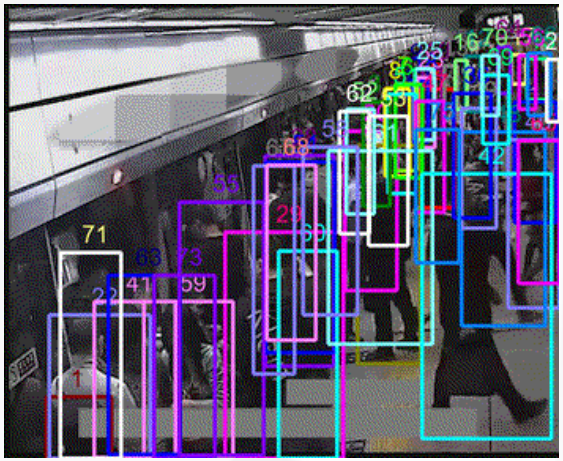
复杂场景实时目标（人、车）检测、识别、跟踪

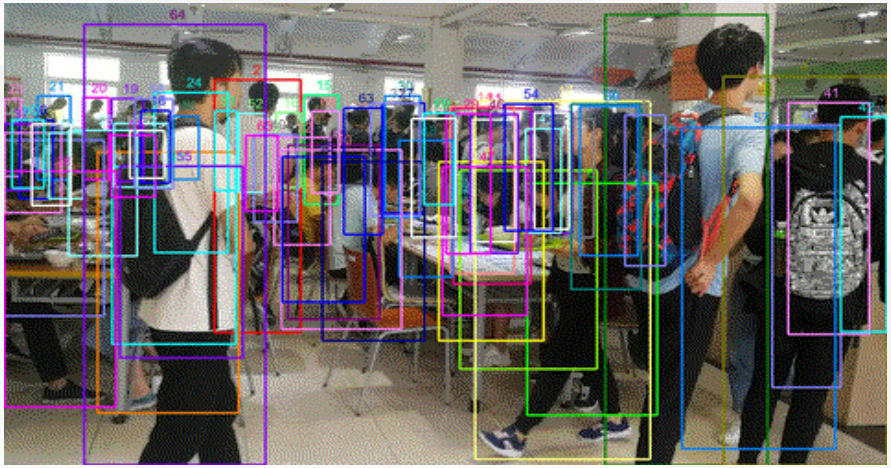


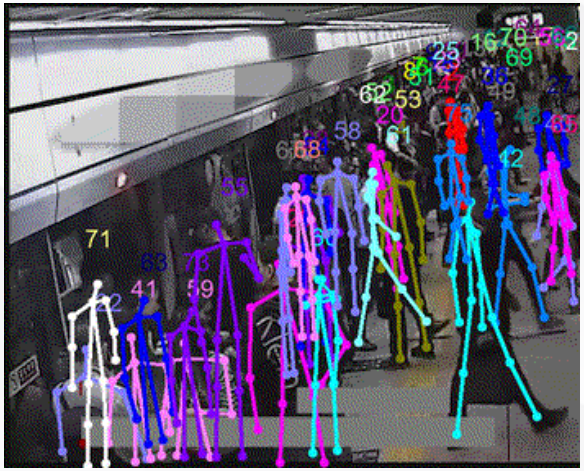


Challenge on Large-scale Human-centric Video Analysis in Complex Events (HiEve， ACM MM’ 20 Grand Challenge)

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





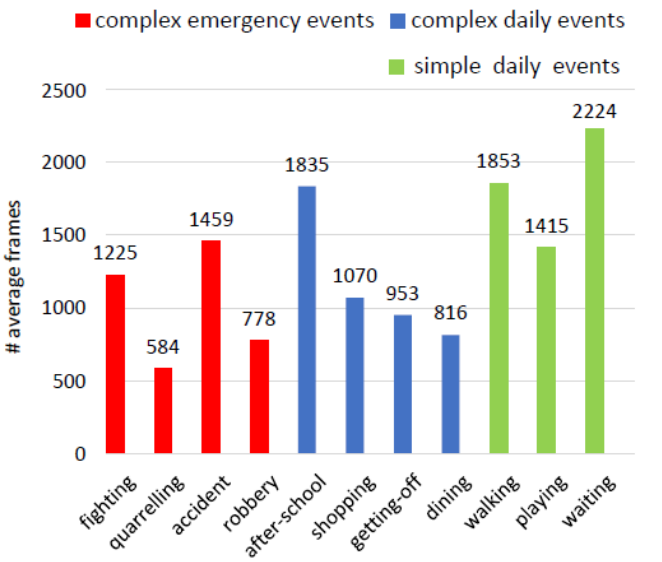


Challenge on Large-scale Human-centric Video Analysis in Complex Events (HiEve， ACM MM’ 20 Grand Challenge)

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**	√	√	√

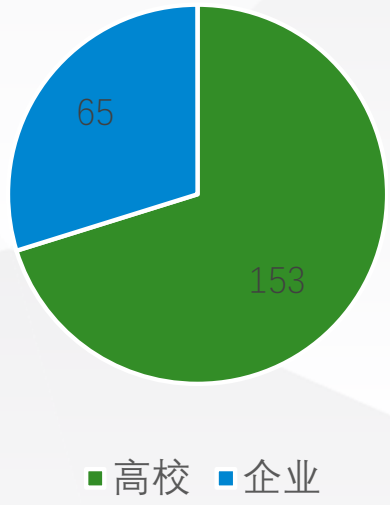


报名参赛队：218

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/，

相关论文：https://arxiv.org/abs/2005.04490

已于8月重新开放！

主要思路和问题

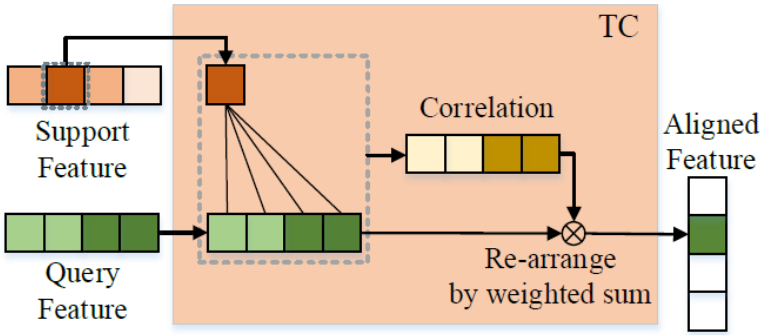
技术思路

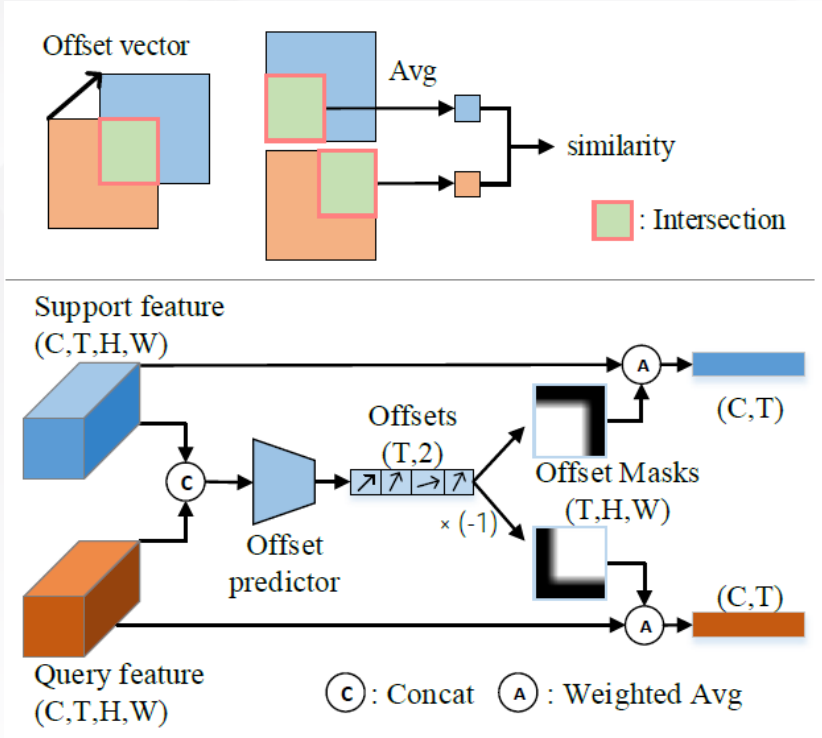
- 从下向上的感知（检测、跟踪、三维骨架）
- **少样本/无监督**
- 多模态 （音频+视频、文本+视频）

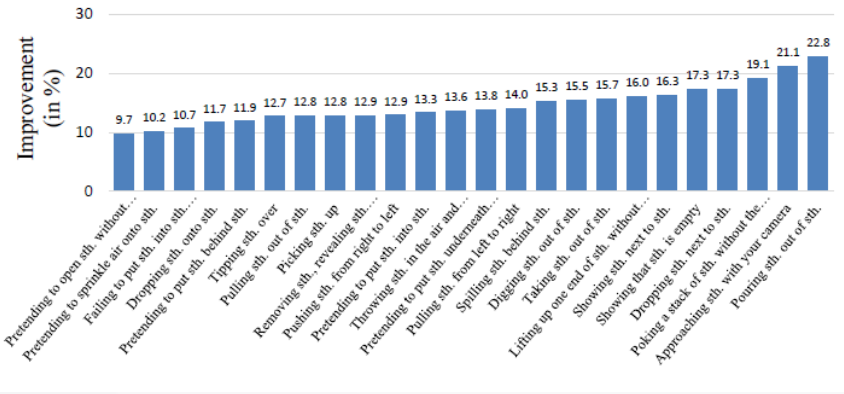
行为类型

- 个人行为
- 组群行为
- 其他

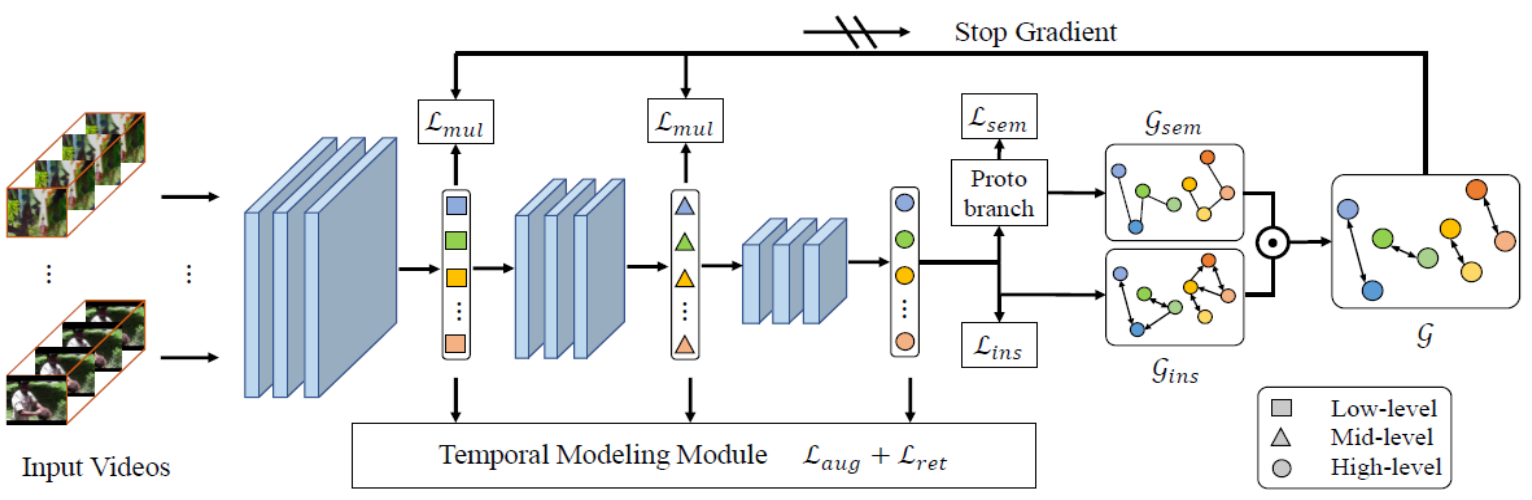
少样本：时空对齐

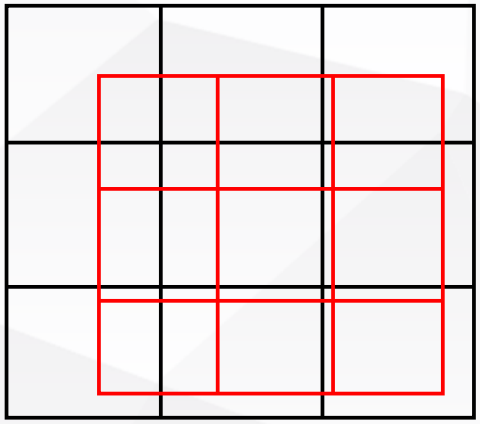

TC
Support Feature
Query Feature
Correlation
Re-arrange by weighted sum
Aligned Feature


Offset vector
Avg
similarity
: Intersection
Support feature (C,T,H,W)
Query feature (C,T,H,W)
Offset predictor
Offsets (T,2)
× (-1)
Offset Masks (T,H,W)
(C,T)
(C,T)
C : Concat
A : Weighted Avg


Improvement (in %)
30
20
10
0
9.7
10.2
10.7
11.7
11.9
12.7
12.8
12.8
12.9
12.9
13.3
13.6
13.8
14.0
15.3
15.5
15.7
16.0
16.3
17.3
17.3
19.1
21.1
22.8
Pretending to open sth. without...
Pretending to sprinkle air onto sth.
Failing to put sth. into sth....
Dropping sth. onto sth.
Pretending to put sth. behind sth.
Tipping sth. over
Pulling sth. out of sth.
Picking sth. up
Removing sth., revealing sth....
Pushing sth. from right to left
Pretending to put sth. into sth.
Throwing sth. in the air and...
Pretending to put sth. underneath...
Pulling sth. from left to right
Spilling sth. behind sth.
Digging sth. out of sth.
Taking sth. out of sth. without...
Lifting up one end of sth. without...
Showing sth. next to sth.
Showing that sth. is empty
Dropping sth. next to sth.
Poking a stack of sth. without the...
Approaching sth. with your camera
Pouring sth. out of sth.

自监督、无监督





$S[0,:] = [0.3, 0.3, 0, 0.2, 0.2, 0, 0, 0, 0]$

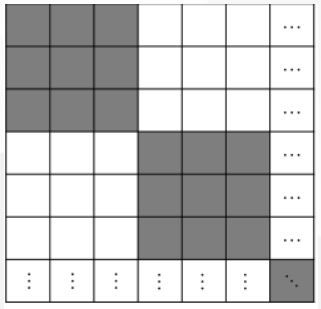
$[S_0, S_1, \ldots, S_{n-1}]$

$G = zeros(nthw, nthw)$

$G[thwi:thw(i+1), thwi:thw(i+1)] = S_i$

$F_l \in R^{nthw \times c},\ F_g \in R^{nthw \times c}$

$matmul(F_l, F_g^T) \rightarrow G$



Enhancing Self-supervised Video Representation Learning via Multi-level Feature Optimization, ICCV, 2021

主要思路和问题

技术思路

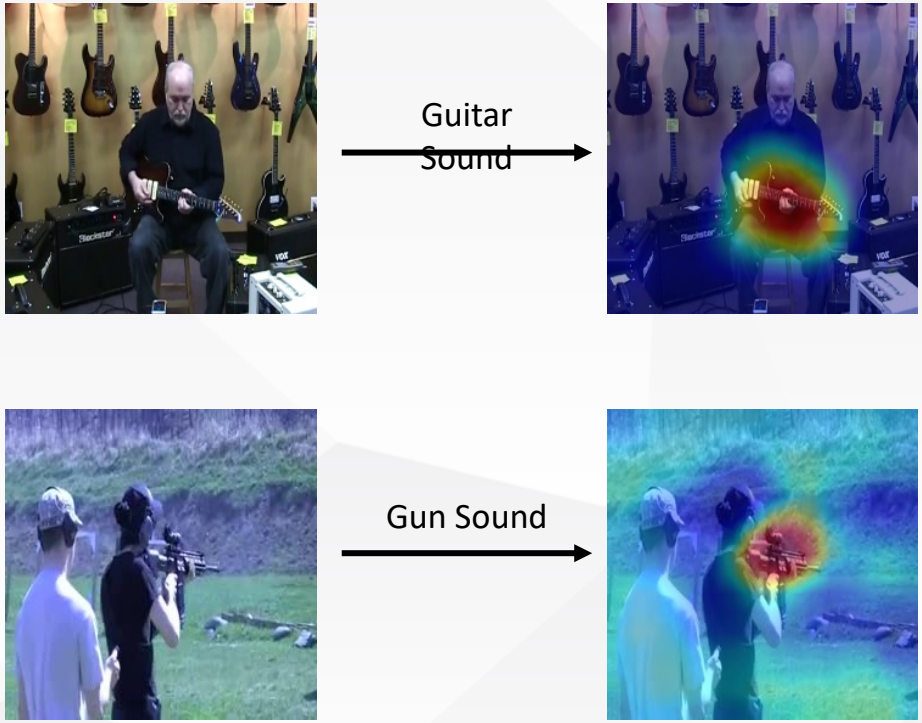
- 从下向上的感知（检测、跟踪、三维骨架）
- 少样本/无监督
- **多模态 （音频+视频、文本+视频）**

行为类型

- 个人行为
- 组群行为
- 其他

音视频联合识别

Multi-Source Sound Localization

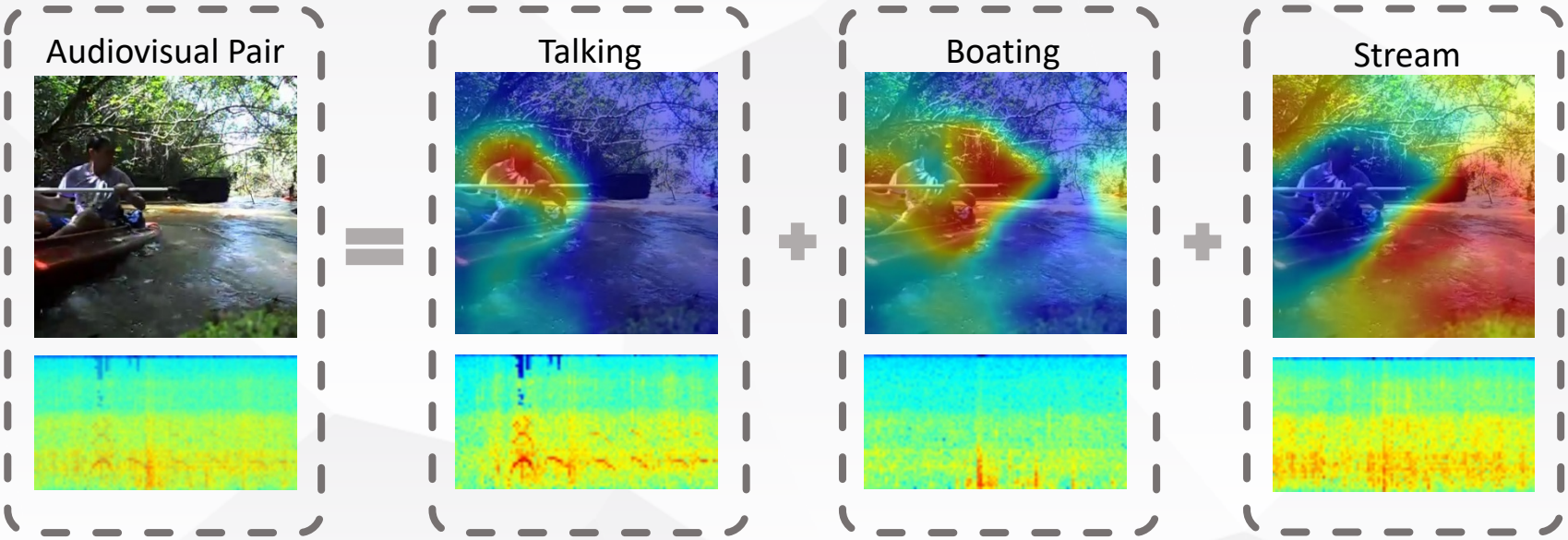


- Multiple audiovisual components
- Associate sound-object pairs without one-to-one annotations

[R. Qian et al, Multiple Sound Sources Localization from Coarse to Fine, ECCV 2020]

音视频联合识别

Multi-Source Sound Localization



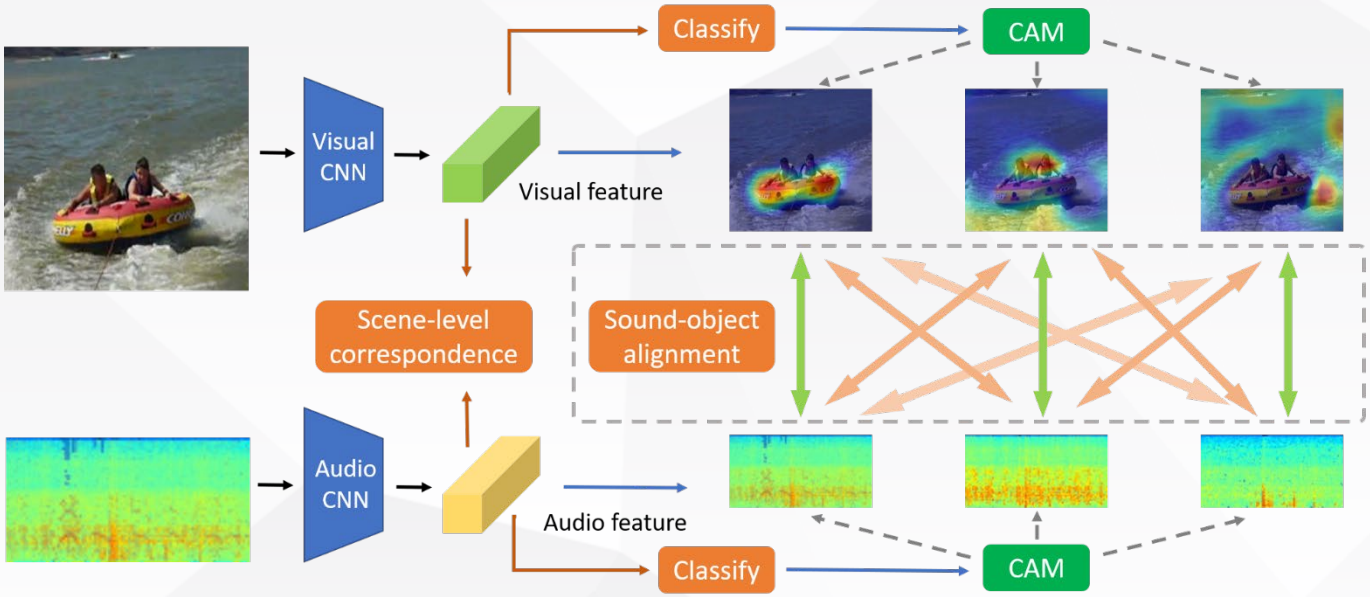
Decompose the challenging problem into two easy-to-solve subtasks:

- Audiovisual content learning
- Category- and video-based feature alignment

[R. Qian et al, Multiple Sound Sources Localization from Coarse to Fine, ECCV 2020]

音视频联合识别

Multi-Source Sound Localization



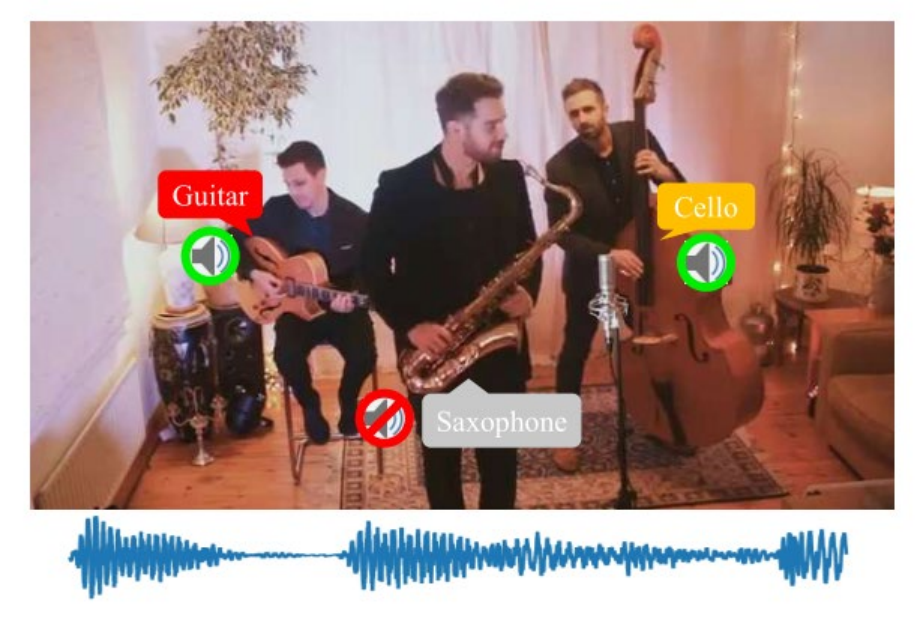
Two stage framework for coarse-to-fine multiple sound sources localization:

- Coarse-grained scene-level correspondence and category-level discrimination
- Fine-grained feature alignment between disentangled audiovisual components

[R. Qian et al, Multiple Sound Sources Localization from Coarse to Fine, ECCV 2020]

音视频联合识别

Discriminative Sound Localization



Cocktail-party scenario consisting of mixed sound and multiple sounding objects as well as silent ones

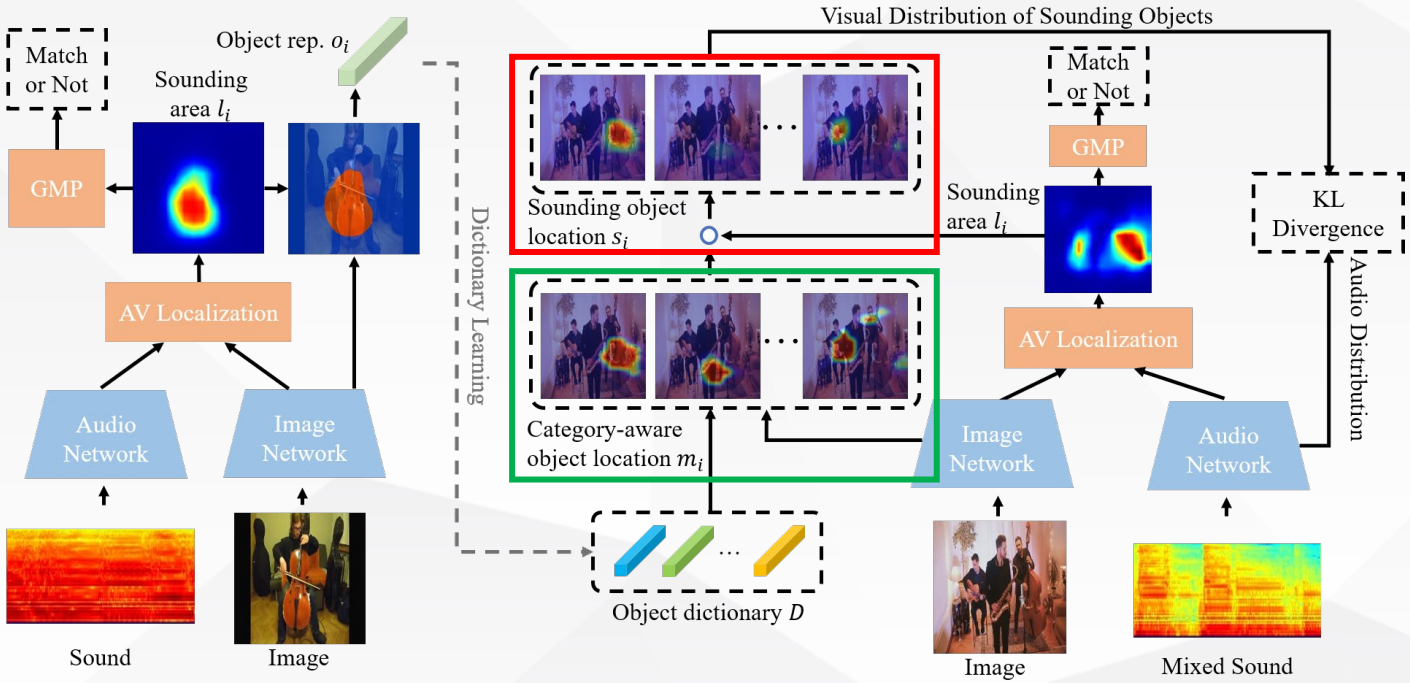
Contributions of our work:

- Discriminatively localize objects without resorting to semantic annotations of objects
- Determine whether specific class is sounding or not, and filtering out silent ones from the mixed sound

[D. Hu et al, Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching, NeurIPS 2020]

音视频联合识别

Discriminative Sound Localization



Two-stage framework to first learn object knowledge from single-source videos and then apply to cocktail-party scenarios and employ audiovisual consistency to align audio and visual distribution

[D. Hu et al, Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching, NeurIPS 2020]

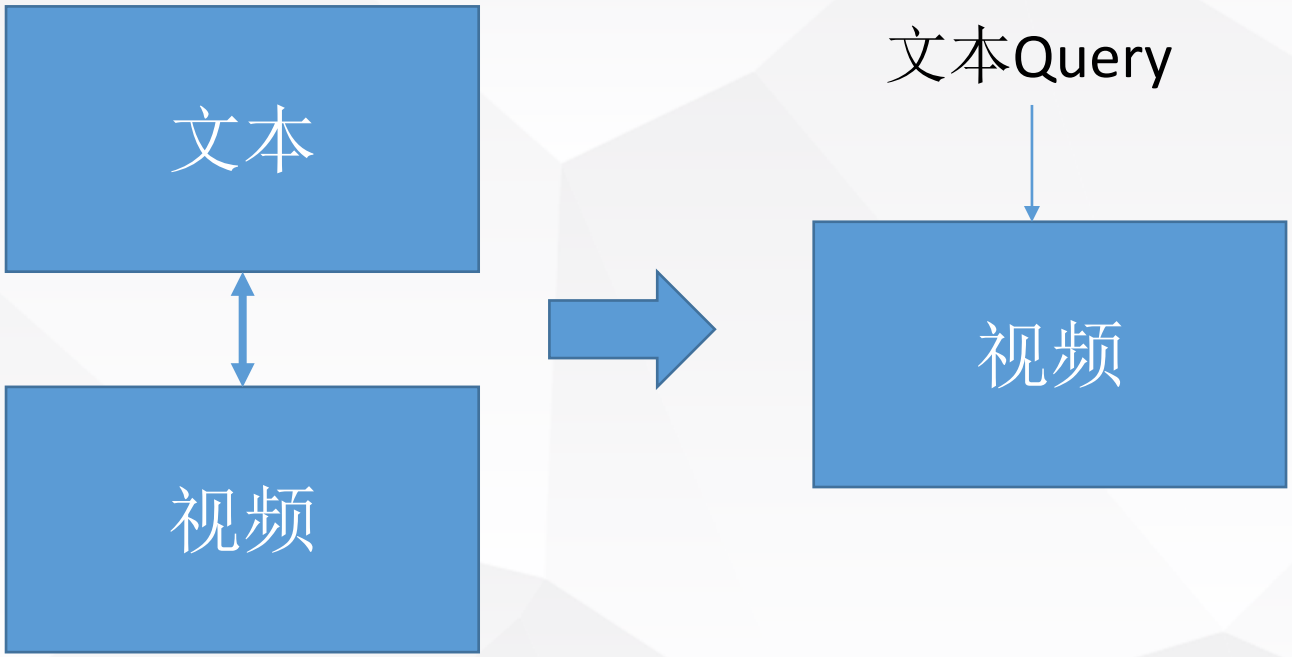
音视频联合识别

Discriminative Sound Localization

Localization results on realistic and synthetic cocktail-party videos. The green box indicates target sounding object area, while the red box means this class of object is silent and its activation value should be low.

[D. Hu et al, Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching, NeurIPS 2020]

文本-视频联合



海量数据

主要思路和问题

技术路径

- 从下向上的感知（检测、跟踪、三维骨架）
- 少样本/无监督
- 多模态 （音频+视频、文本+视频）

行为类型

- **个人行为**
- 组群行为
- 其他

其他成果

从下到上的事件检测，细粒度事件检测

主要思路和问题

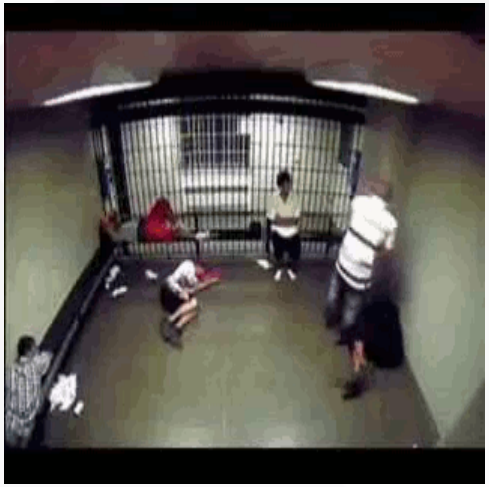
技术思路

- 从下向上的感知（检测、跟踪、三维骨架）
- 少样本/无监督
- 多模态 （音频+视频、文本+视频）

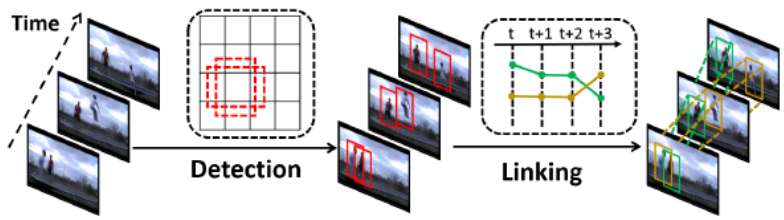
行为类型

- 个人行为
- **组群行为**
- 其他

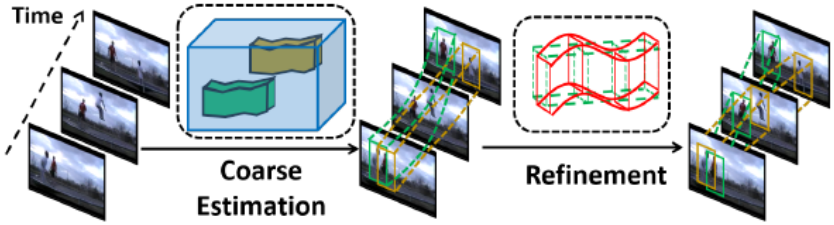
实时行为识别与检测（2）

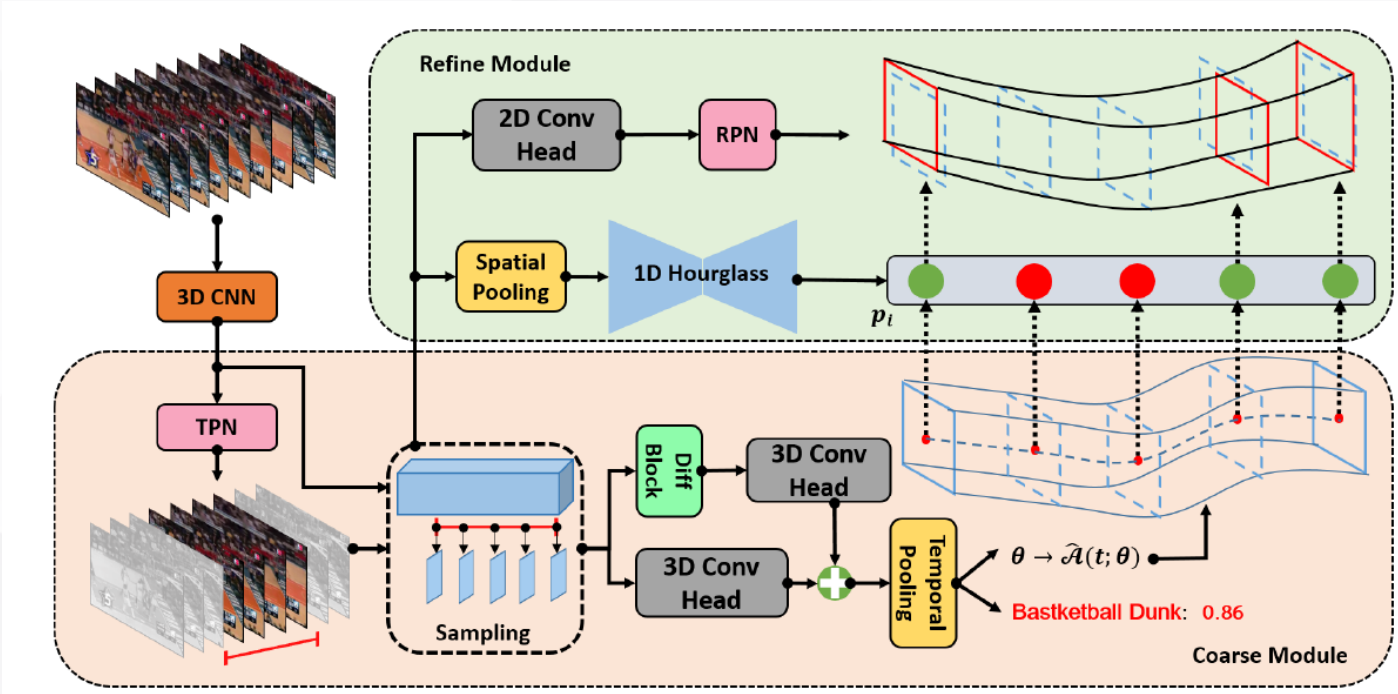


行为识别（时空行为定位）



(a) detection and linking method





CFAD: Coarse-to-Fine Action Detector for Spatiotemporal Action Localization, ECCV, 2020.

主要思路和问题

技术思路

- 从下向上的感知（检测、跟踪、三维骨架）
- 少样本/无监督
- 多模态 （音频+视频、文本+视频）

行为类型

- 个人行为
- 组群行为
- **其他**

For More Information:

https://weiyaolin.github.io/

http://humaninevents.org/

Thank You!