计算机视觉基础：视频分析

高层语义理解篇

胡建芳，郑伟诗

https://isee-ai.cn/~hujianfang/

中山大学

机器智能与先进计算
教育部重点实验室

视频解析概述

- 研究背景：

  现实中的视觉数据大部分是有时序关联的视频数据

- 视频分析应用：

  安防监控，网络视频审核，机器人交互设计等

视频解析概述

- 突出研究团队：

Andrew Zisserman, Visual Geometry Group, University of Oxford

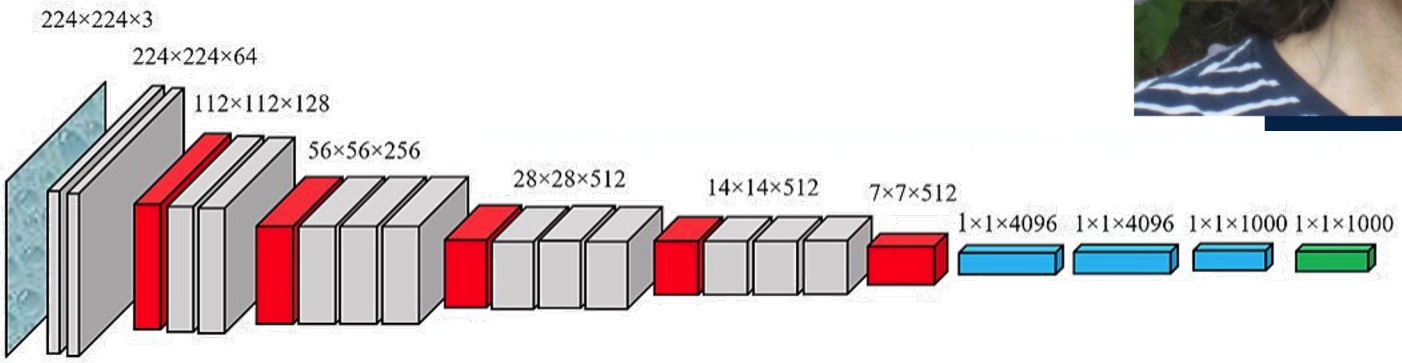
开启了深度学习视频识别先河，提出了two-stream 框架

UNIVERSITY OF OXFORD

Our Research / Artificial Intelligence

Showing articles associated with

Andrew Zisserman



VGG Network

视频解析概述

概述：简单的工作回顾，欲知详细细节，请看相关的论文文献

视频分析任务概述

- 基于手工设计的方法：

  两步法：特征提取 + 任务分析（对应于损失函数）

- 基于深度学习的方法：

  一步法：构建端到端的网络模型（网络设计+损失函数）

与图像任务相比：编程实现复杂，需要考虑计算问题（时效）

视频分析任务概述

- 单纯的视觉分析任务（**理解世界**）：

  面向**人**的分析（行为识别，行为预测，行为检测，行为分割）

  面向**物体**的分析（物体跟踪，对象分割）

- 视频与自然语言结合（**理解并服务世界**）：

  基于视频的问答

  视频描述

计算机视觉研究的目的：**理解世界**， **服务世界**

世界是运动的

计算机视觉（视频） **自然语言，与人交互**

视频分析任务概述：两个重要指标

- 效果（目前研究关注的多）：

  识别或预测的准确率

- 效率（实际应用中关注的多）：

  时间代价

天下没有免费的午餐：**效果**与**效率**之间的矛盾

模型复杂（网络更深，参数多）　　**模型简单（网络更浅，参数少）**

效果与效率之间，达到一个好的平衡

与人相关的视频分析任务：

行为动作

行为识别

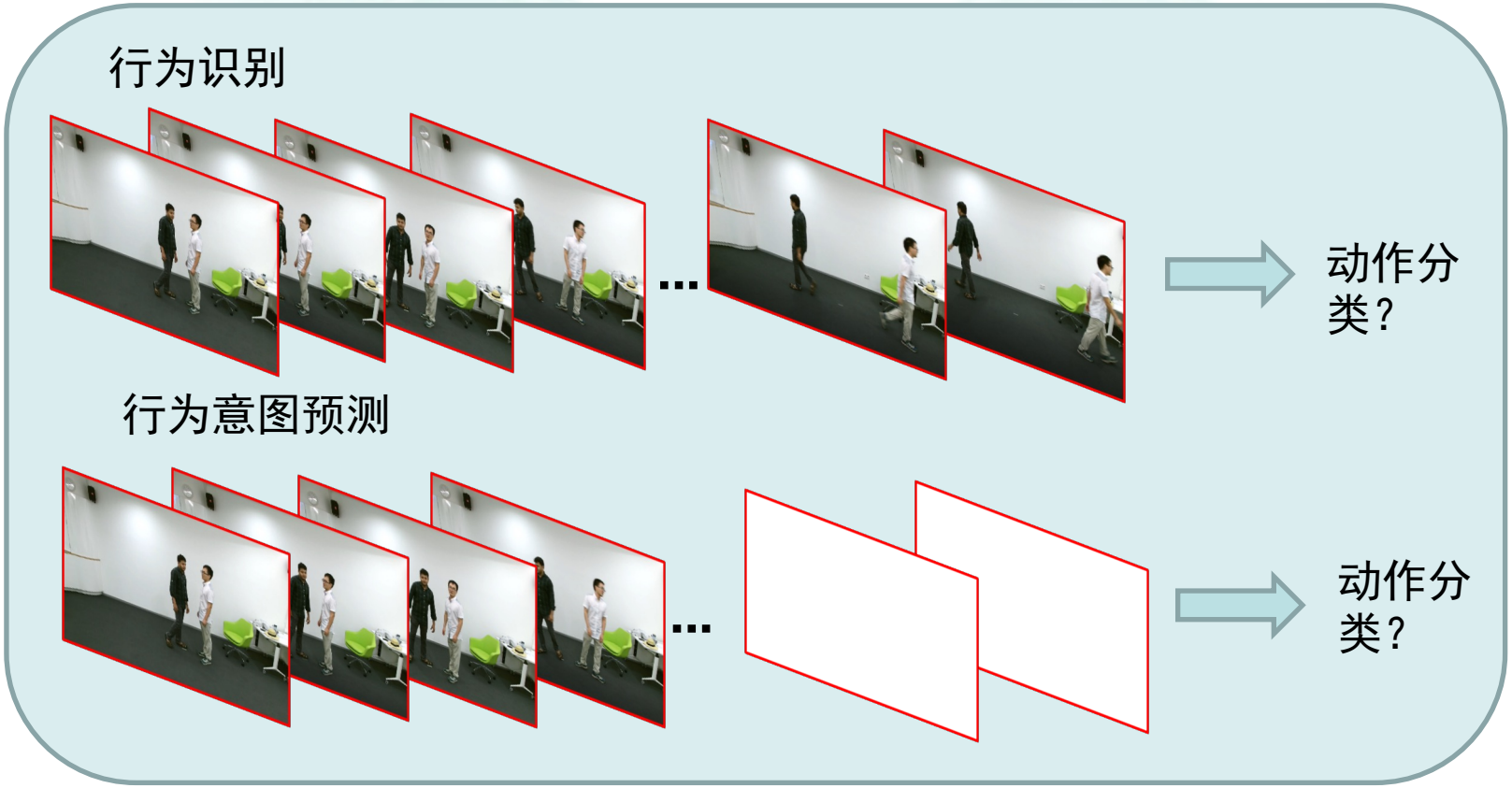
- 任务难点

给定视频片段，识别视频中的动作信息

视频特征 + 分类器；

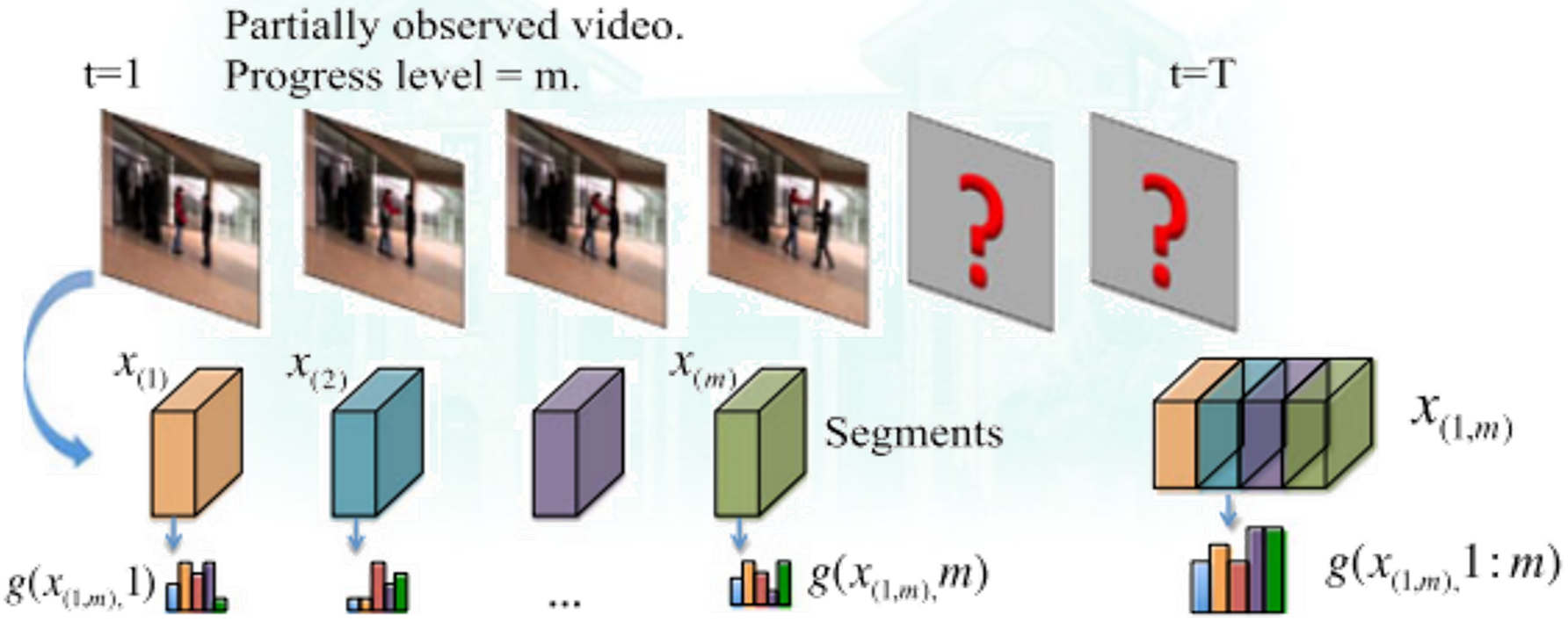
行为意图预测

- 任务难点
  1. 动作信息不完整
  2. 合理的时序建模



行为意图预测

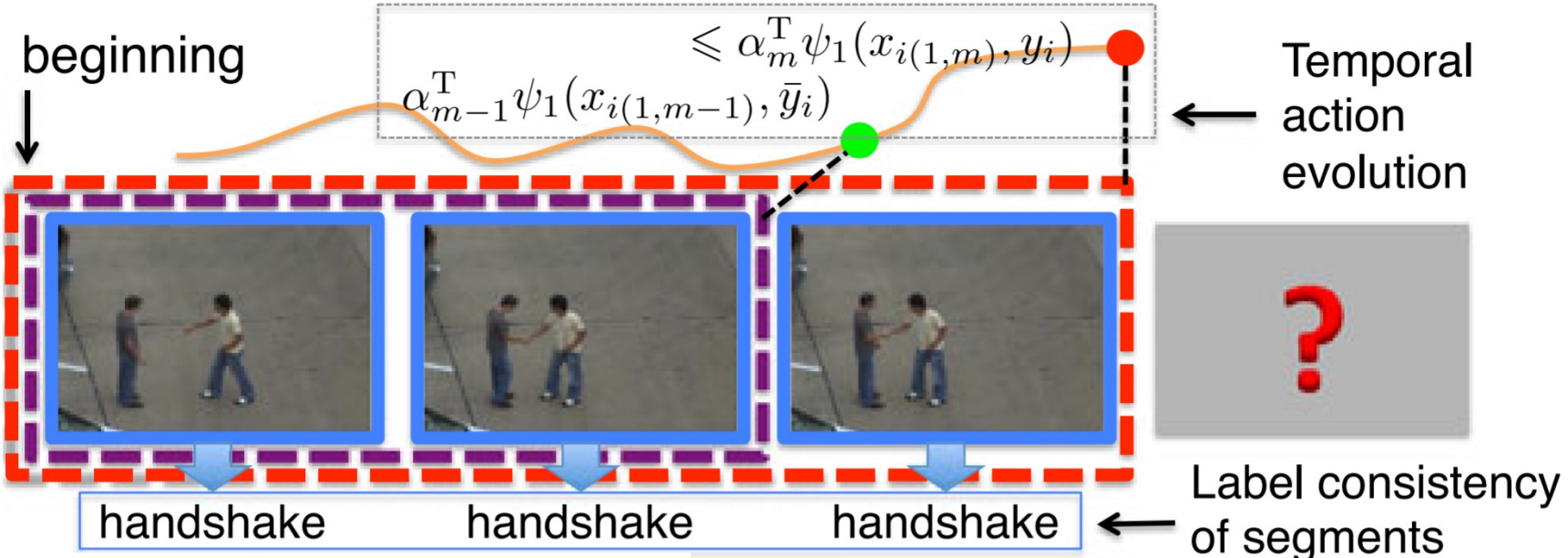
- 问题定义
  1. 将完整视频分成N等分
  2. 针对前m段子视频，进行识别



行为意图预测

- 最大分割边界模型

  假设：随着时间增多，越能确定视频动作信息



$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C(\xi_1 + \xi_2 + \xi_3)$$

$$\text{s.t. } \forall m, \forall(\bar{y}_1, \ldots, \bar{y}_N) \in \mathcal{Y}^N,$$

$$\frac{1}{N}\sum_{i=1}^{N}[\mathbf{w}^{\mathrm{T}}\Phi(x_{i(1,m)}, y_i) - \mathbf{w}^{\mathrm{T}}\Phi(x_{i(1,m)}, \bar{y}_i)]$$

$$\geqslant \frac{M}{N}\sum_{i=1}^{N}\delta(\bar{y}_i, y_i) - \frac{\xi_1}{u(m/M)},$$

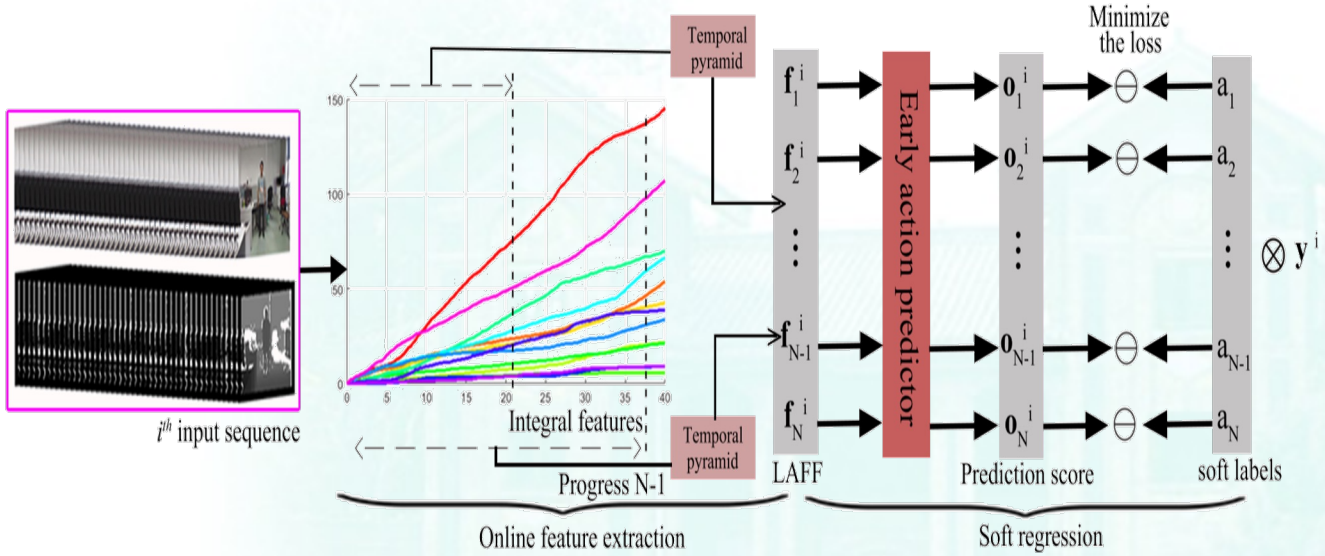
$$m = 2, \ldots, M, \forall(\bar{y}_1, \ldots, \bar{y}_N) \in \mathcal{Y}^N,$$

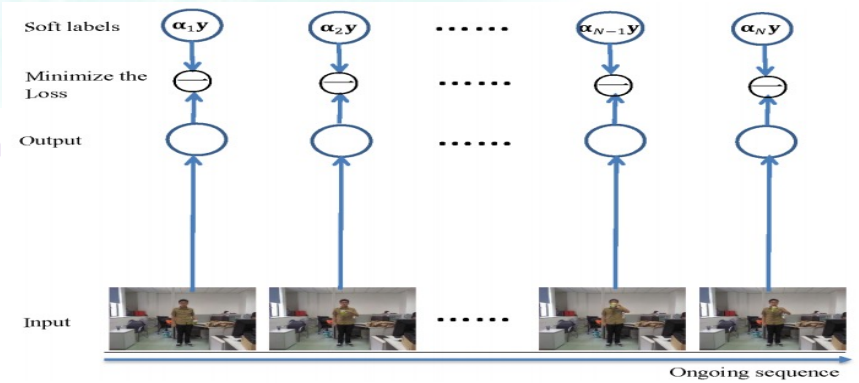
$$\frac{1}{N}\sum_{i=1}^{N}[\boldsymbol{\alpha}_m^{\mathrm{T}}\psi_1(x_{i(1,m)}, y_i) - \boldsymbol{\alpha}_{m-1}^{\mathrm{T}}\psi_1(x_{i(1,m-1)}, \bar{y}_i)]$$

$$\geqslant \frac{M}{N}\sum_{i=1}^{N}\delta(\bar{y}_i, y_i) - \frac{\xi_2}{u(m/M)},$$

行为意图预测

- 弱回归模型
  1. 每个子序列，学习一个弱标签





(a) soft linear regression (SLR)

Prediction loss　Consistency　Regularization

$$\min_{\mathbf{W},\boldsymbol{\alpha}} \sum_{i=1}^{I}\sum_{n=1}^{N} \left\|(\mathbf{W}^T\mathbf{X}_i - \mathbf{y}_i\boldsymbol{\alpha}^T)\mathbf{S}\right\|_{1,2} + \frac{\xi_1}{2}\left\|\nabla\boldsymbol{\alpha}\right\|_2^2 + \frac{\xi_2}{2}\left\|\mathbf{W}\right\|_F^2$$

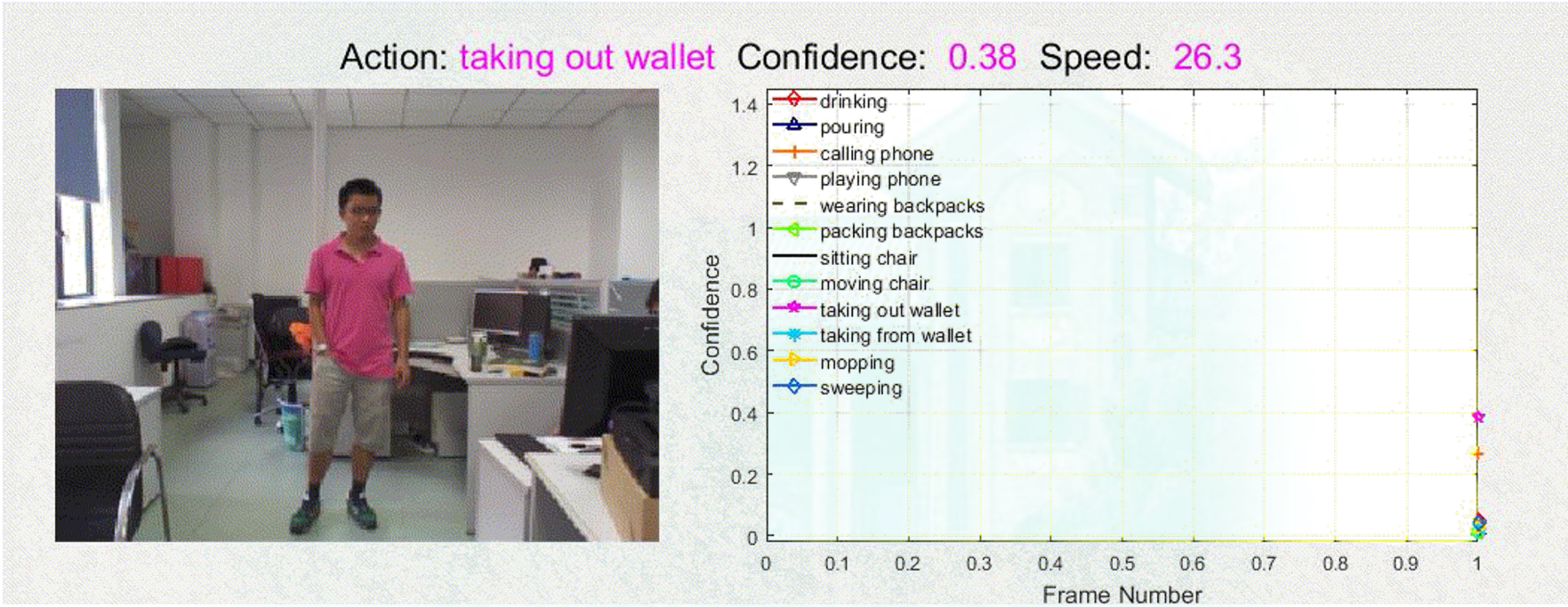
$$s.t., \boldsymbol{\alpha}^T\mathbf{e}_N = 1, 0 \le \boldsymbol{\alpha} \le 1.$$

$\mathbf{W}$ Activity predictor

$\mathbf{X}_i$ Features of the *i*-th video

行为意图预测

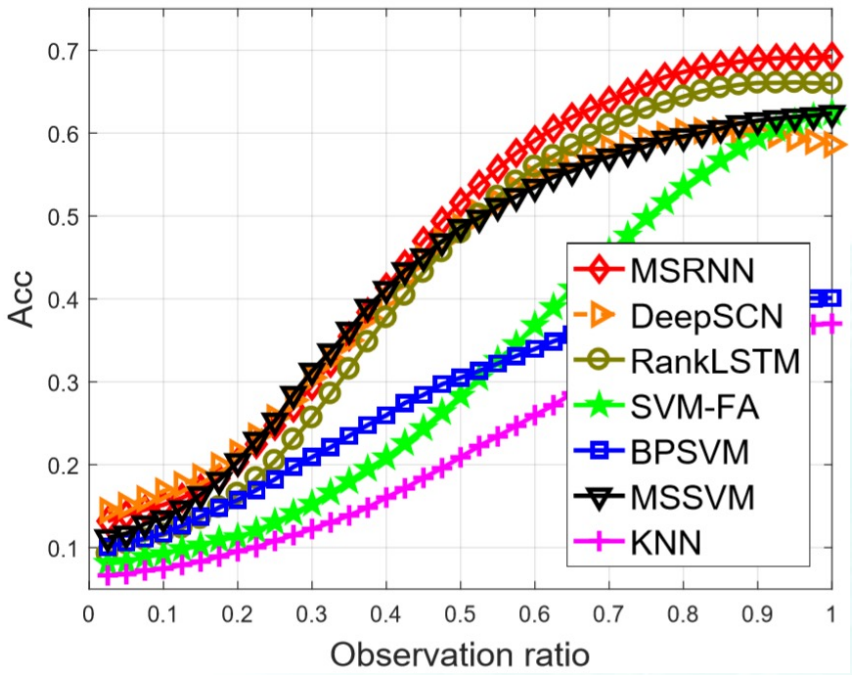
- 弱回归模型
  1. 每个子序列，学习一个弱标签

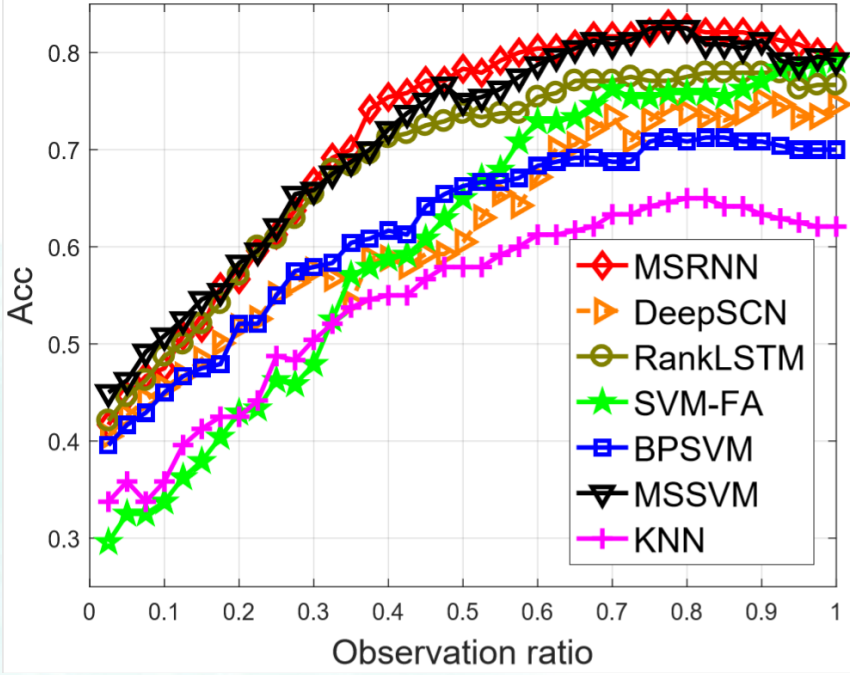


行为意图预测

- 弱回归模型

1. 每个子序列，学习一个弱标签

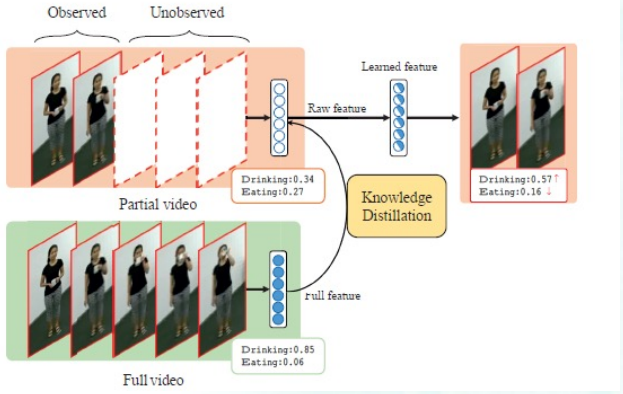


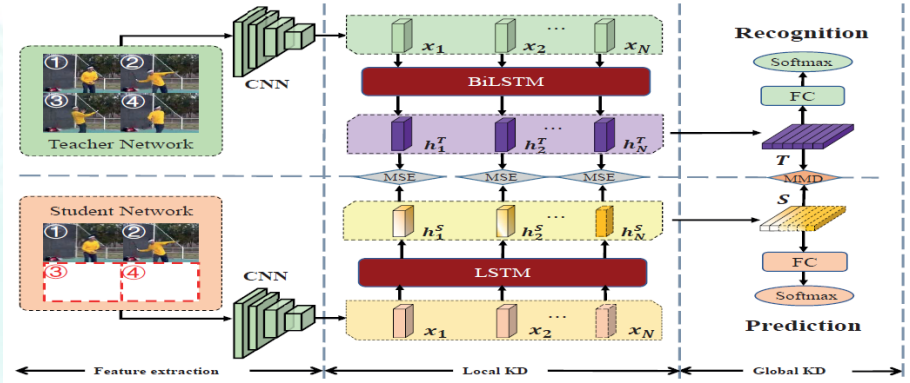


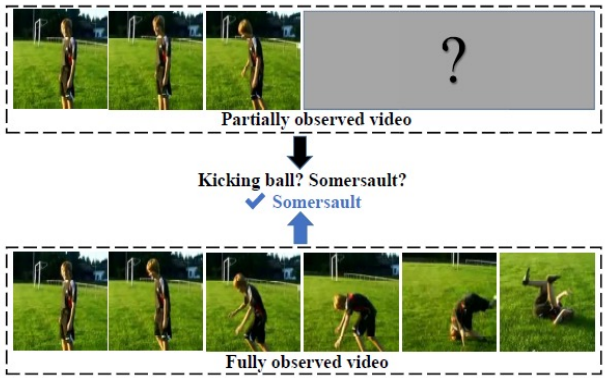
行为意图预测

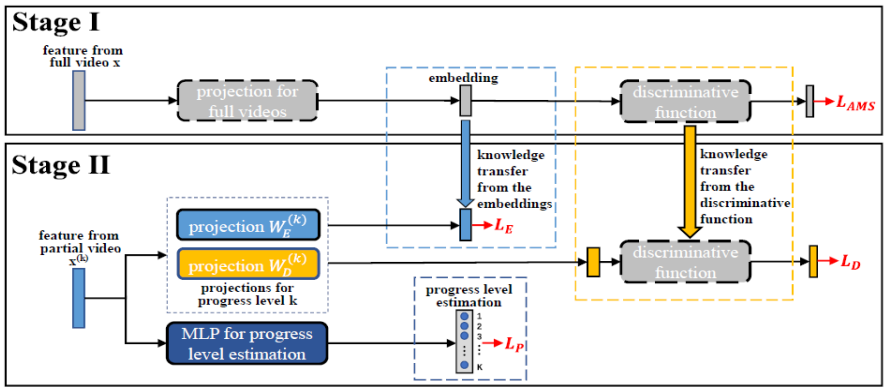
- 基于知识迁移的建模方法

  通过搭建合理的行为识别模型，并利用合理的方法将知识迁移到行为意图预测模型，提升模型的信息挖掘能力，例如蒸馏学习、迁移学习等方法。





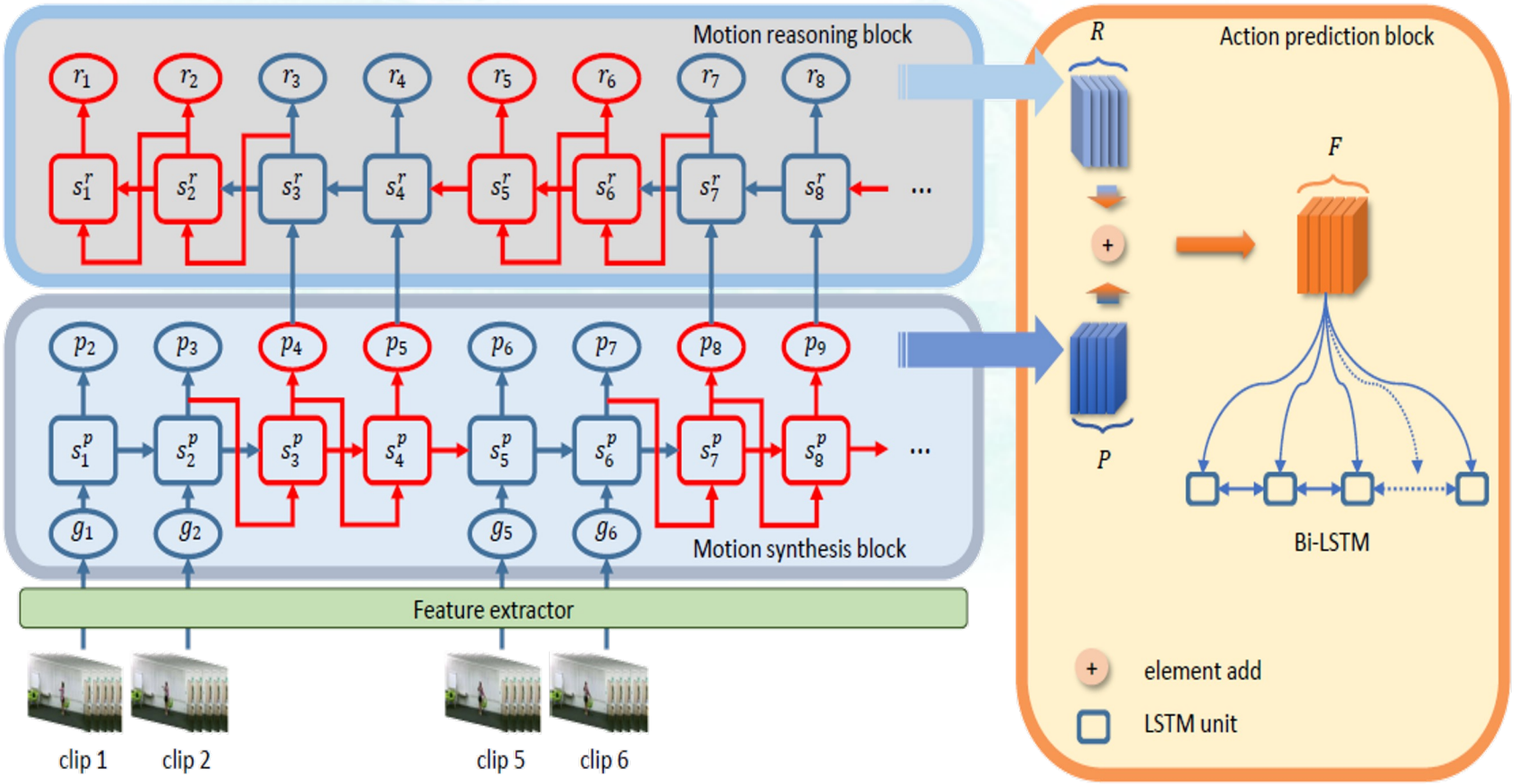




行为意图预测

- 基于深入挖掘动作模式信息

  通过对已知动作信息的深入挖掘，进而生成信息对缺失的动作信息进行补充

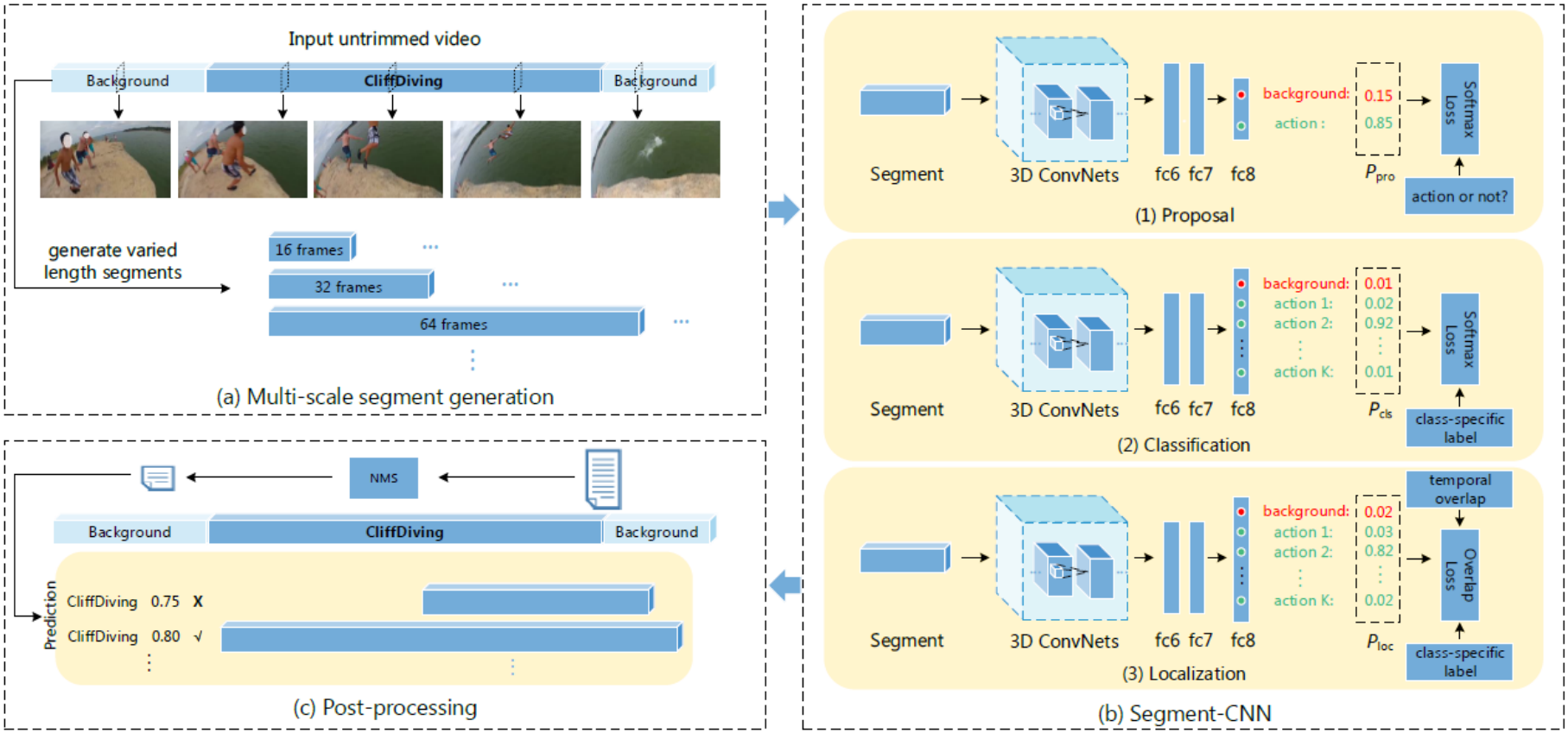


行为检测

Action detection

Temporal action localization in untrimmed videos via multi-stage cnns

仿照物体检测，搭建了三阶段的动作检测框架



行为检测

Action detection

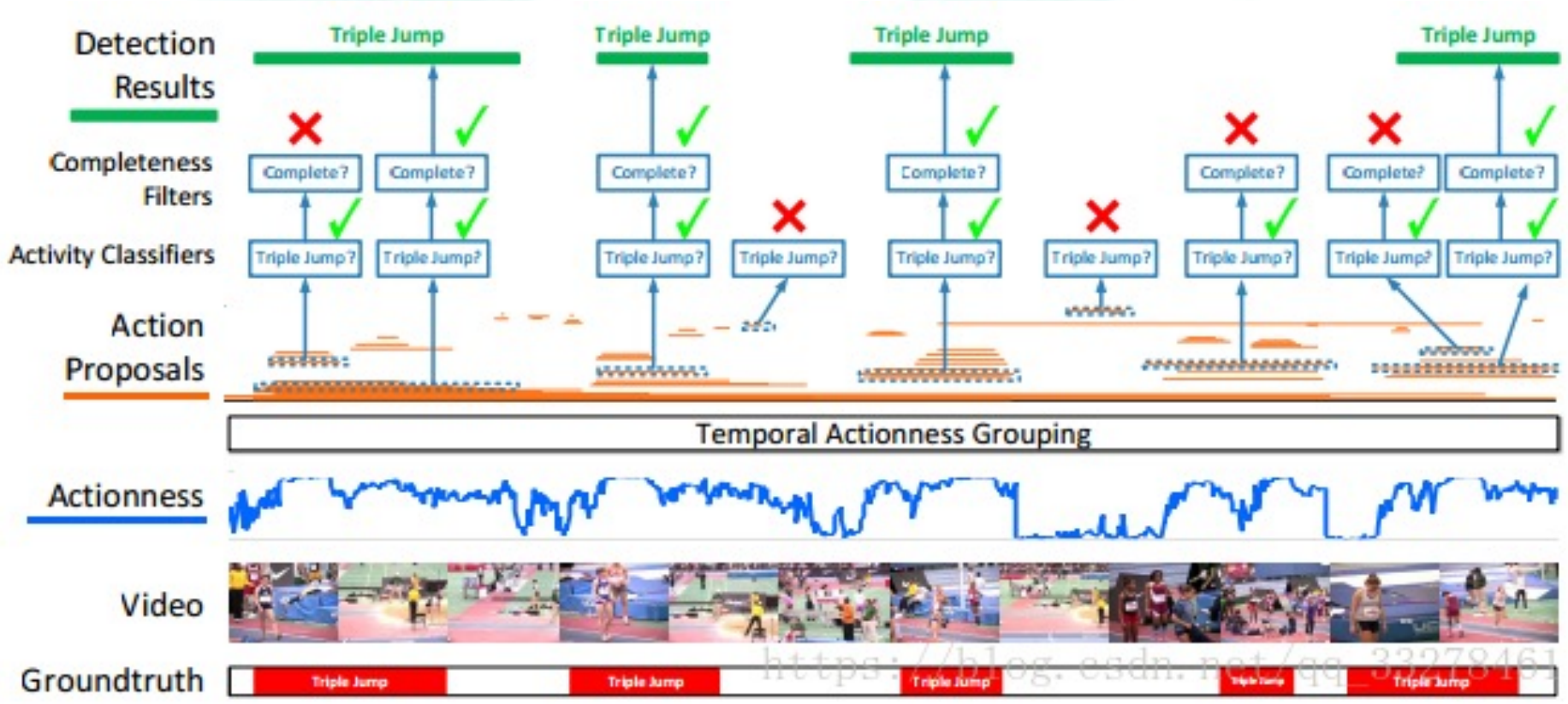
Temporal action localization in untrimmed videos via multi-stage cnns

行为检测

Action detection

A Pursuit of Temporal Accuracy in General Activity Detection

引入 actionness,减少不必要的proposal，自底向上合并



视频解析概述

与物相关的视频分析任务：

跟踪，物体分割

目标跟踪

- 目标跟踪是计算机视觉中的一个重要研究方向，有着广泛的应用，如：视频监控、公共安全、人机交互等等。

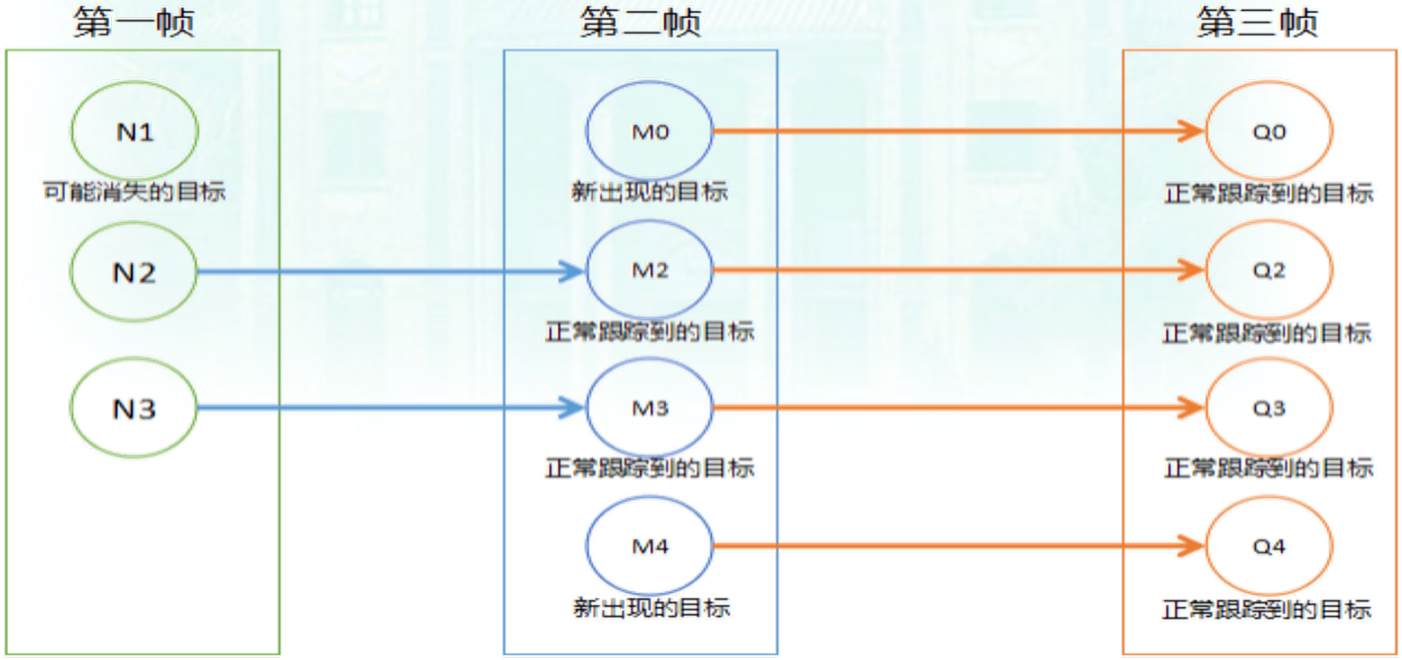
目标跟踪

- 过去二三十年，视觉目标跟踪技术取得了长足的进步，特别是最近几年，利用深度学习的目标跟踪方法取得了令人满意的效果，使目标跟踪技术获得了突破性的进展。

目标跟踪算法

- 问题定义：给定视频的初始帧的物体位置，得到后续视频中物体的位置
- 基于相关性（模板匹配，相关计算）
- Tracking as detection:

  用目标检测的方法逐帧检测物体位置，根据相关性，把前后帧目标串接起来

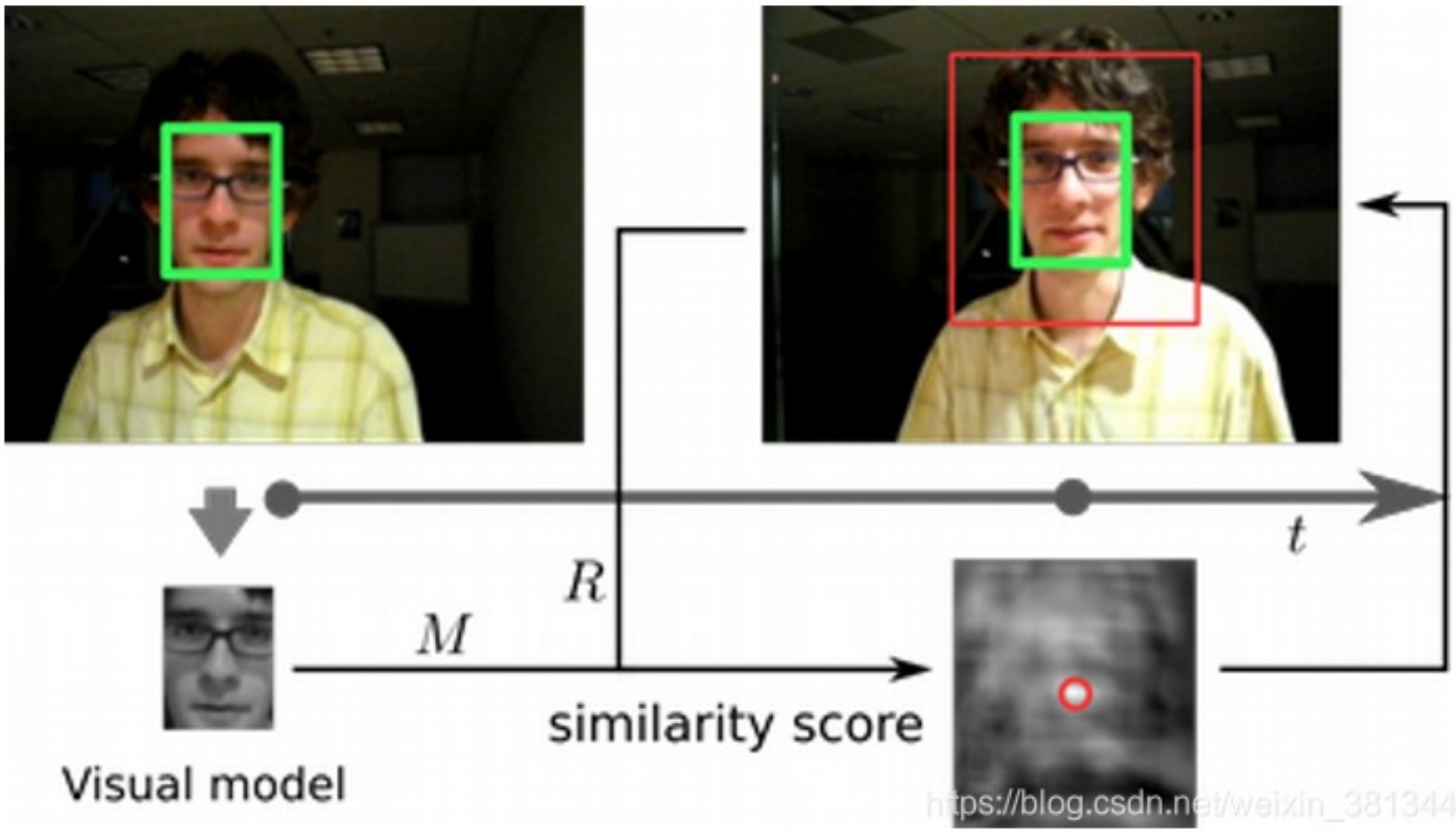


目标跟踪算法

- 模板匹配

  将输入框，与一定区域内的候选框计算相似度

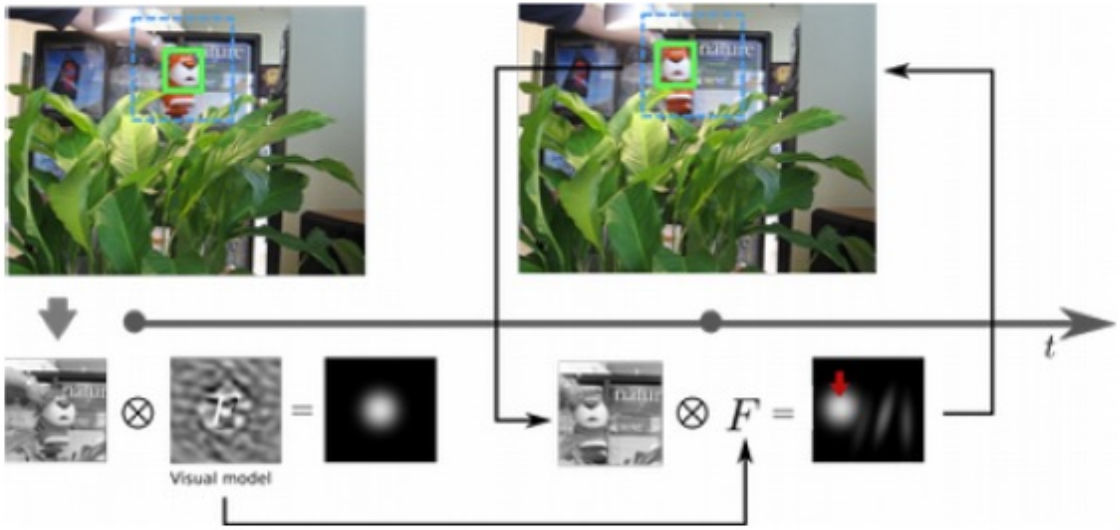
  返回相似度最高的框

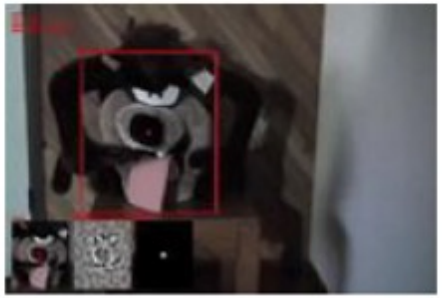


目标跟踪算法

- 相关性

  利用傅里叶变换，加速匹配（相关性）计算





$$\arg\min_{\mathbf{F}} \|\mathbf{T} \star \mathbf{F} - \mathbf{G}\|^2$$

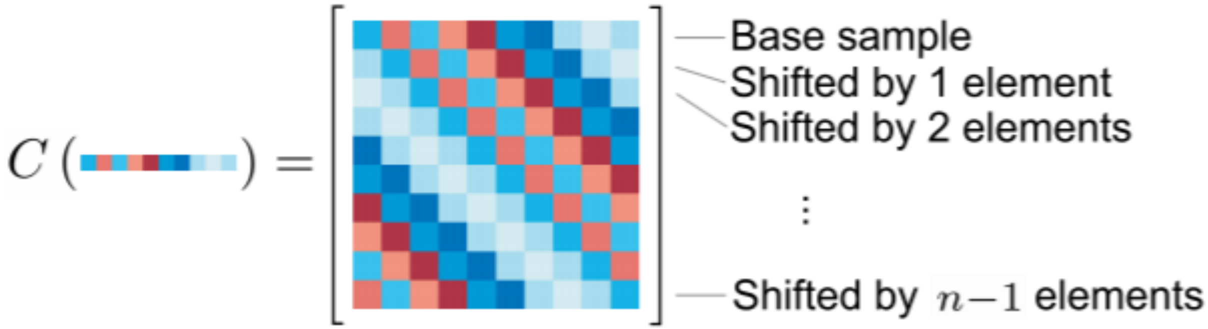
Closed-form solution
Fast computation via FFT

+ Fast
+ Discriminative
– Limited range
– Features matter

目标跟踪算法

- 相关性

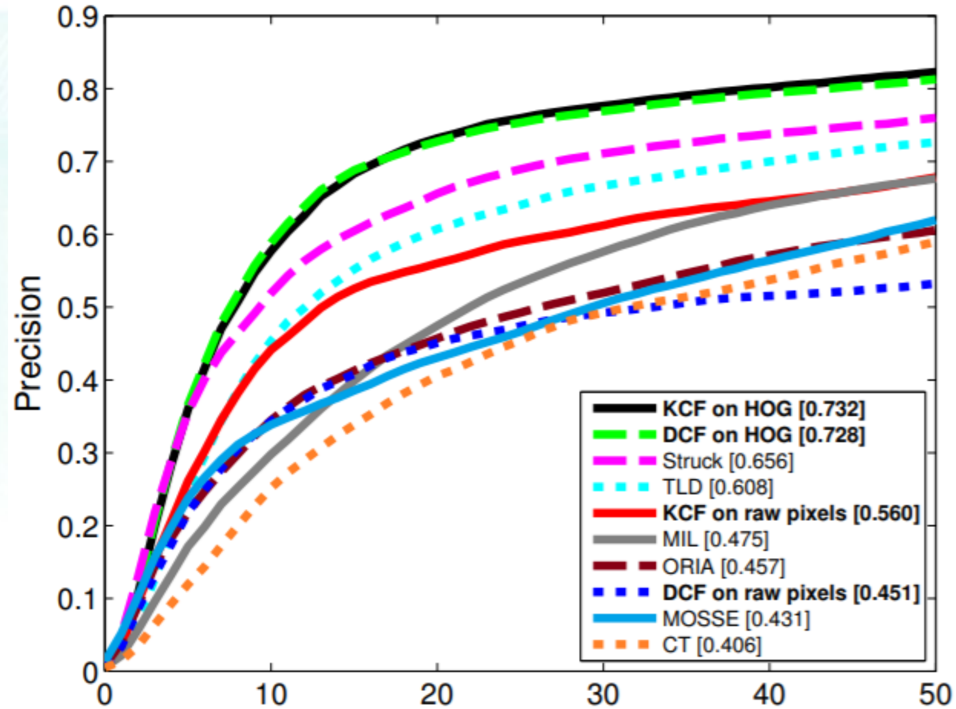
  KCF，循环矩阵



循环矩阵可以用离散福利叶变换进行对角化

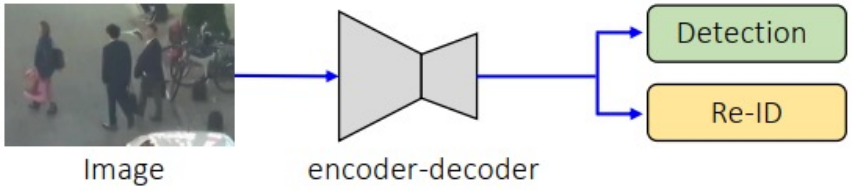
$$\hat{\mathbf{w}} = \frac{\hat{\mathbf{x}}^* \odot \hat{\mathbf{y}}}{\hat{\mathbf{x}}^* \odot \hat{\mathbf{x}} + \lambda}.$$

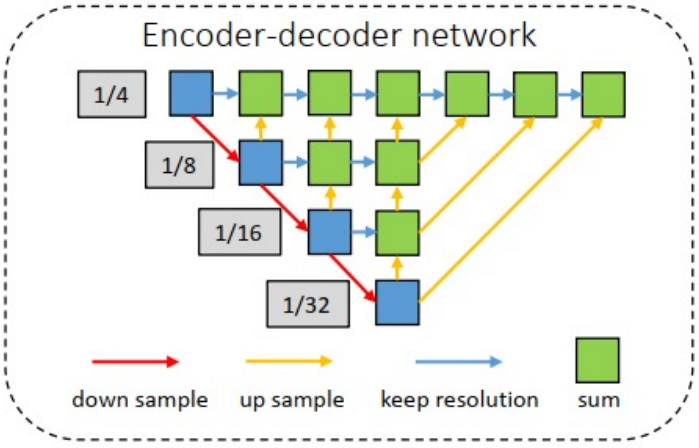
速度非常快，达到 172 FPS

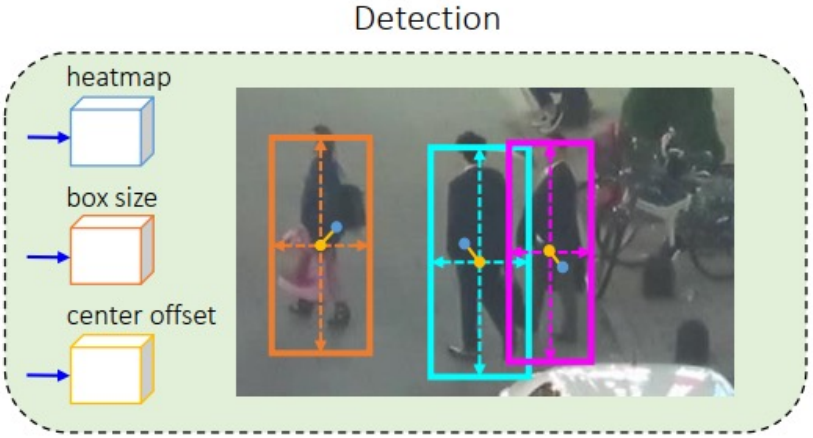


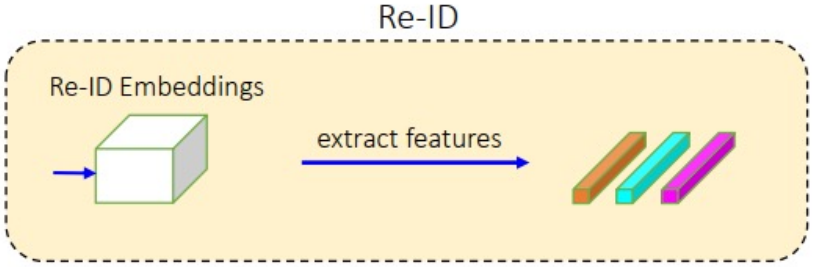
目标跟踪算法：FairMOT（2020）

- 采用anchor-free的CenterNet作为检测模型（Detection），检测速度更快
- 沿用Re-ID关联模型，继承其能够跟踪较长时间被遮挡目标的优点
- 将检测模型和关联模型联合在一起，显著减少处理时间



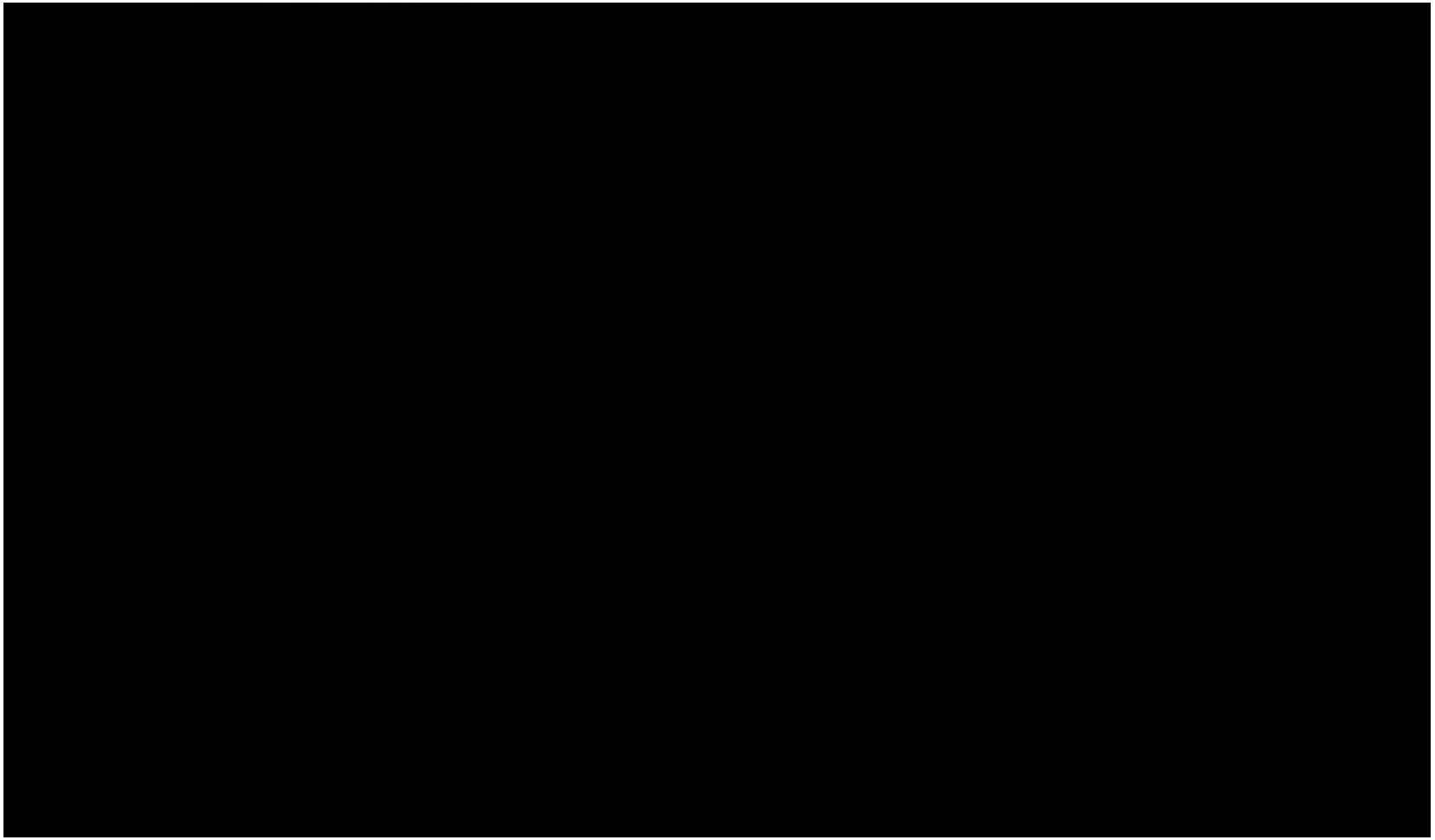






目标跟踪算法：FairMOT（2020）

- 采用anchor-free的CenterNet作为检测模型（Detection），检测速度更快

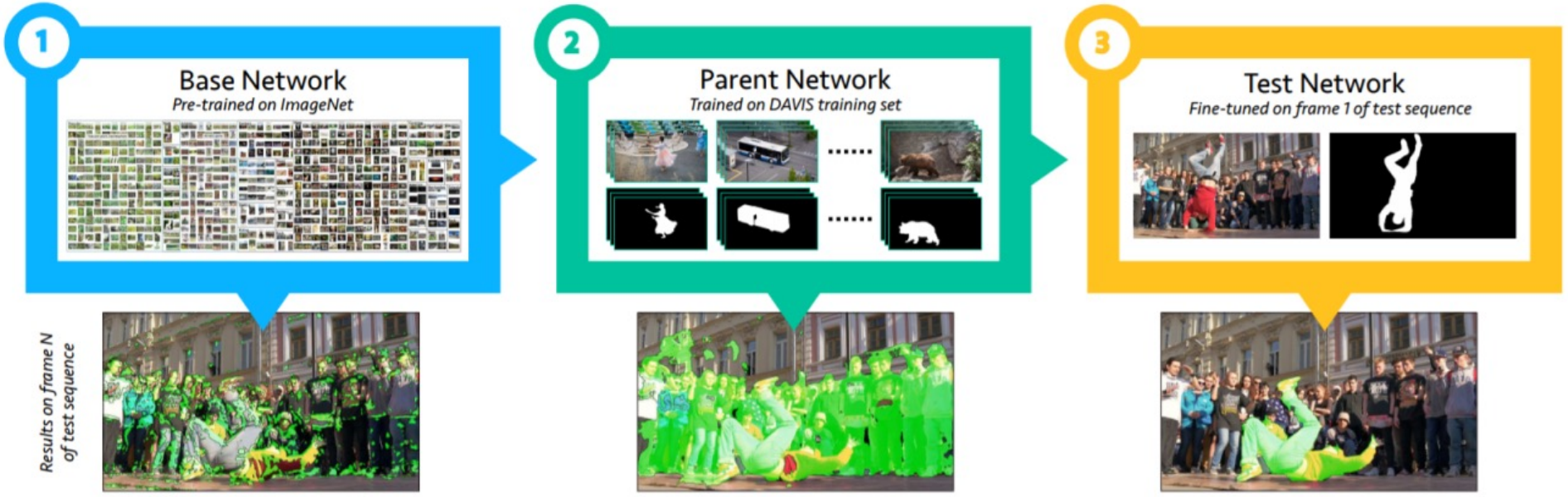


视频目标分割

- 给定视频和要分割的目标物体在第一帧的掩膜，算法需在剩余帧中将该目标物体分割出来

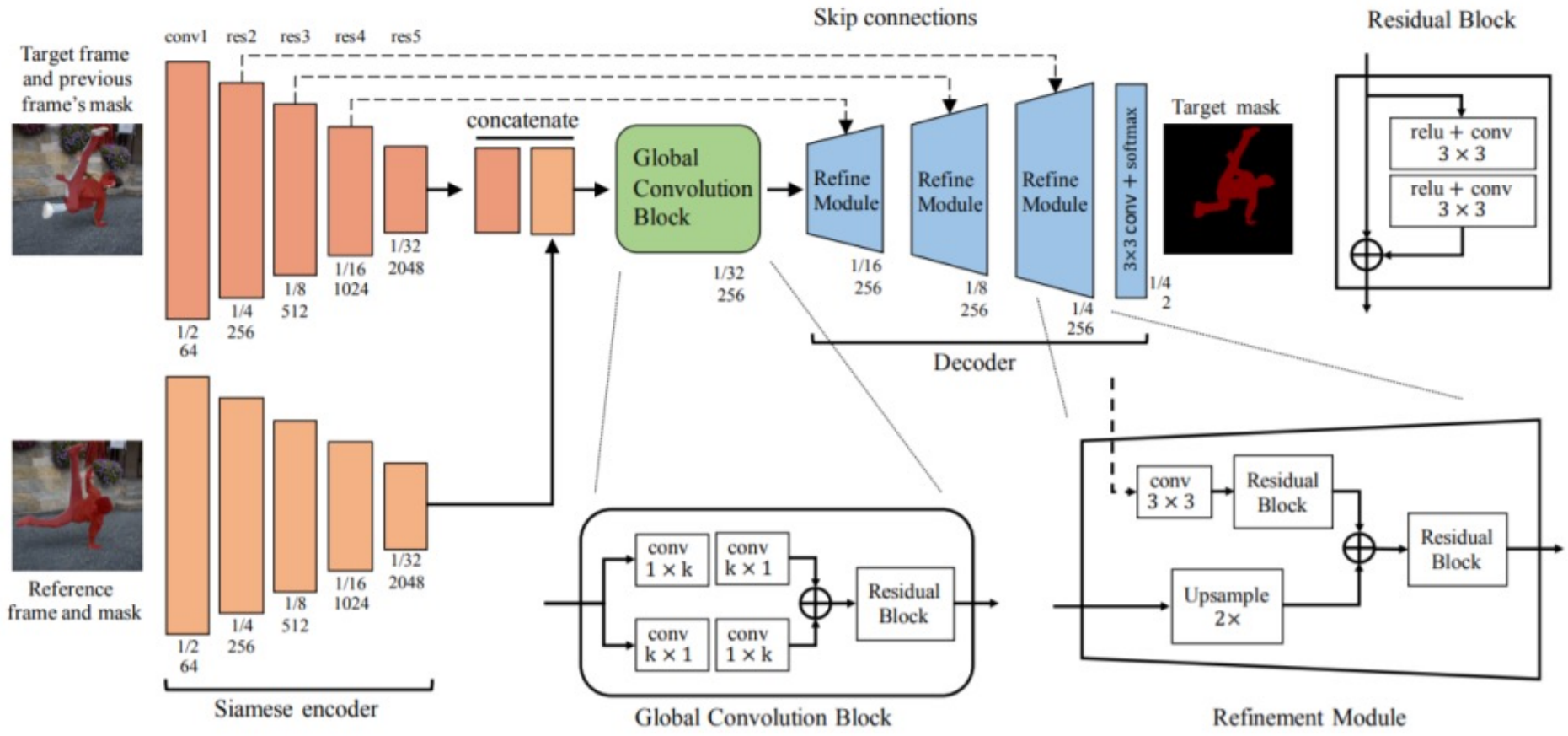
视频目标分割

- OSVOS算法(CVPR2017)
  - 预训练基础网络->在视频目标分割数据集的训练集训练
  - 测试时，使用第一帧的标注微调模型
- 缺点：测试时需使用第一帧标注微调模型，十分耗时



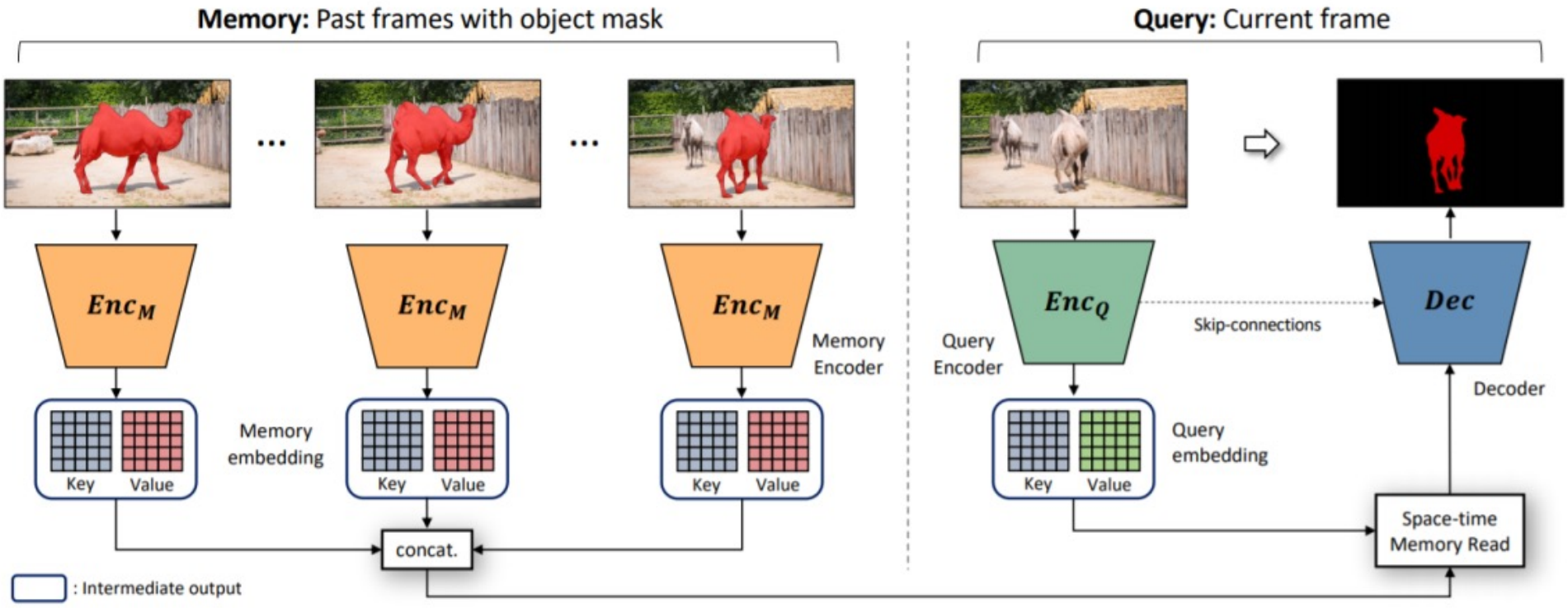
视频目标分割

- RGMP算法(CVPR2018)
  - 将第一帧、当前帧图片以及第一帧、前一帧掩膜作为孪生网络的输入，直接输出当前帧掩膜



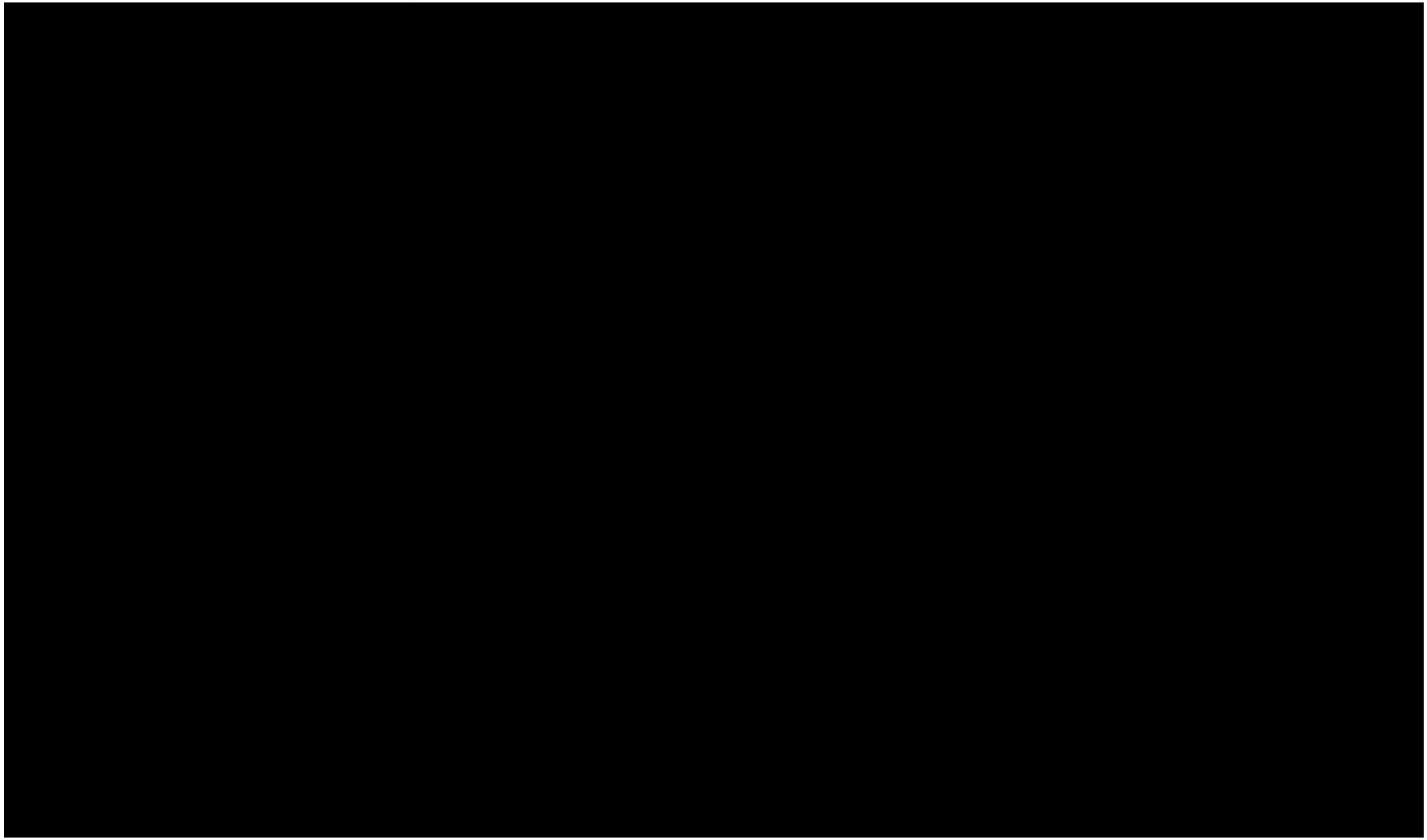
视频目标分割

- 时空记忆网络(ICCV2019)
  - 使用记忆模块存储历史信息，对当前帧图片通过编码器得到查询（Query），用于在记忆模块中查询得到所需的特征以完成分割



视频目标分割

- 时空记忆网络(ICCV2019)



与自然语言结合的视频分析任务：

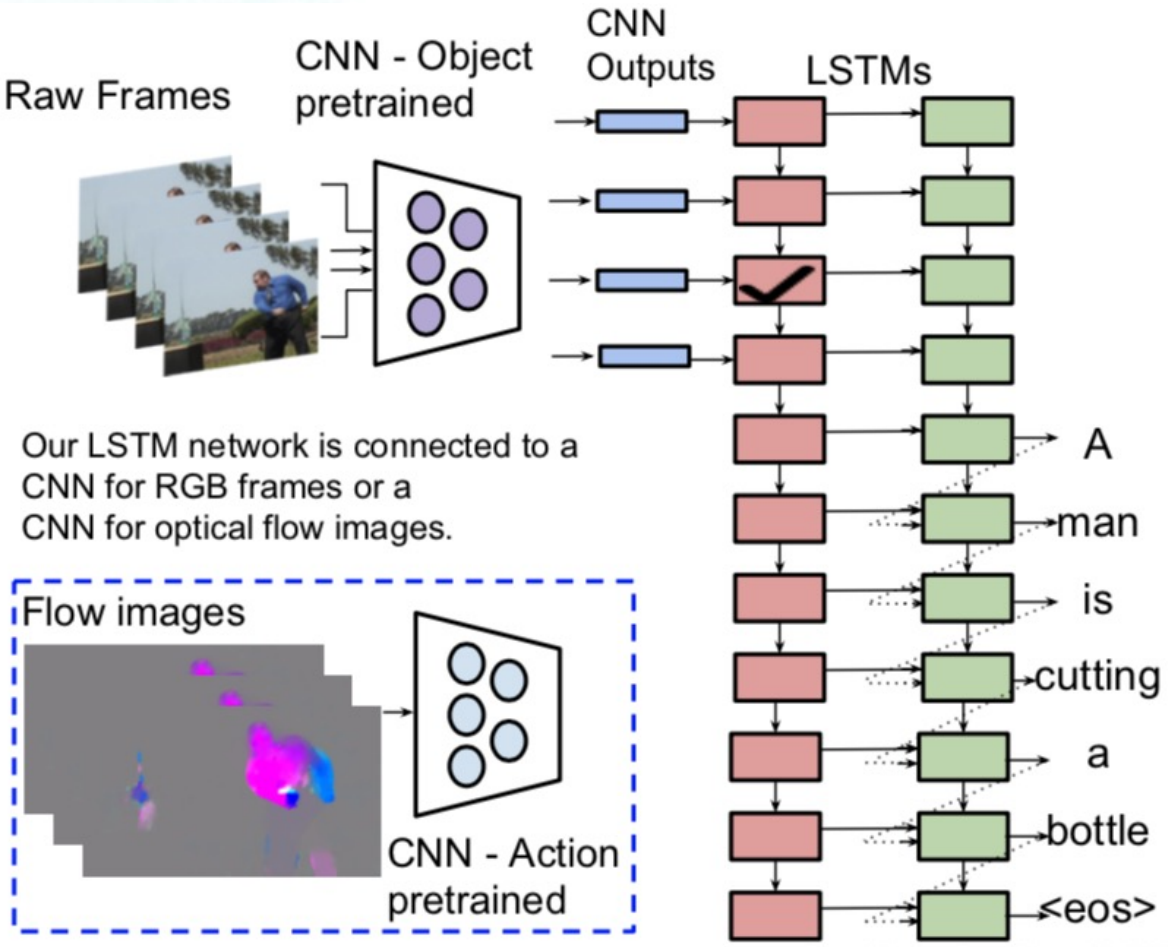
视频标注

视频与自然语言结合

- Video caption

  对视频内容进行语言描述（对象+动作）

视觉理解模型

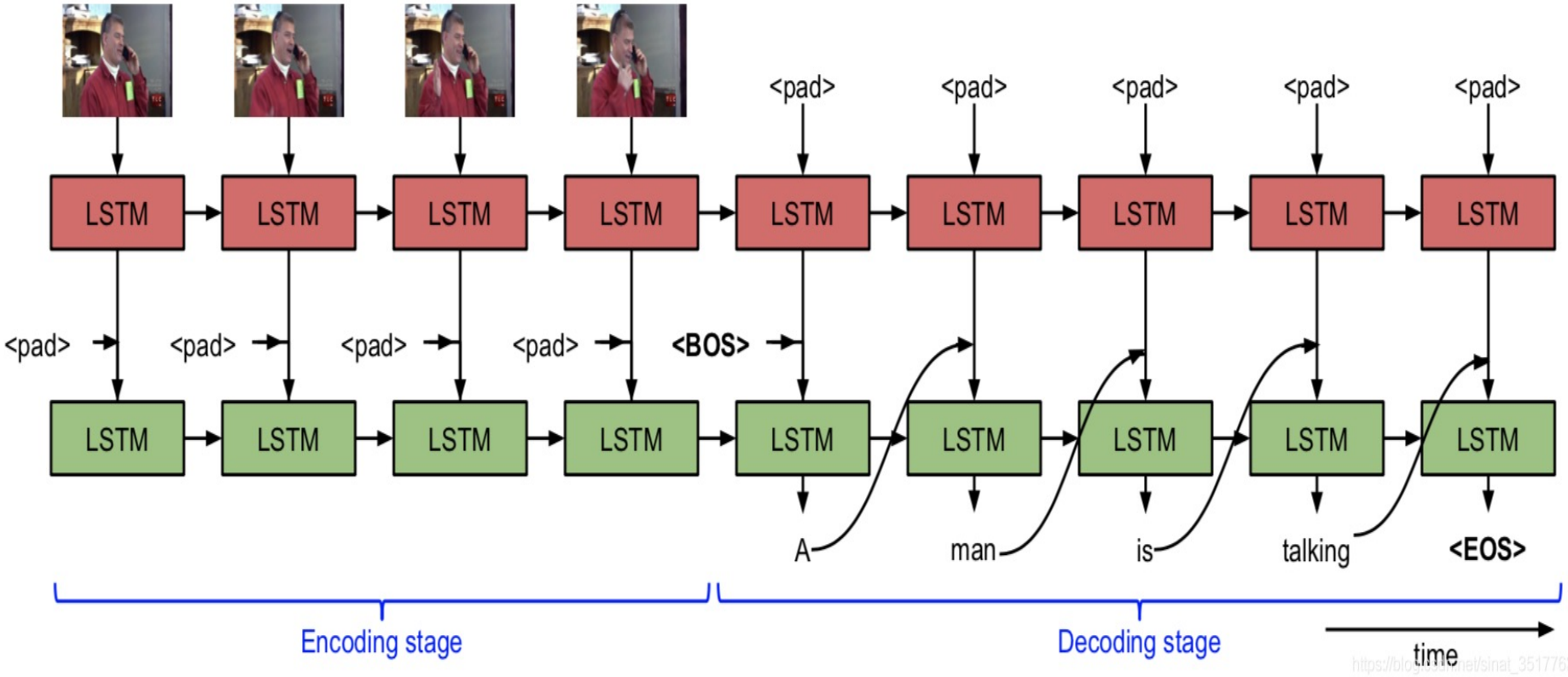


语言解析模型

视频与自然语言结合

- Video caption

  对视频内容进行语言描述（对象+动作）



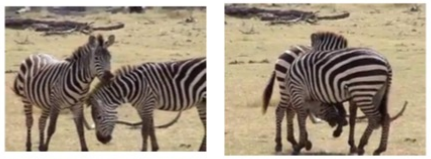
视频与自然语言结合

- Video caption

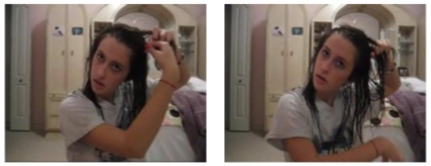
  对视频内容进行语言描述（对象+动作）

Correct descriptions.

S2VT: A man is doing stunts on his bike.



S2VT: A herd of zebras are walking in a field.



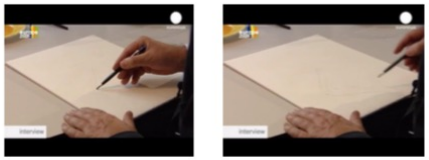
S2VT: A young woman is doing her hair.

S2VT: A man is shooting a gun at a target.

(a)

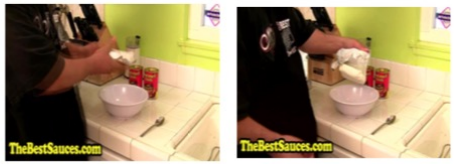
Relevant but incorrect descriptions.

S2VT: A small bus is running into a building.



S2VT: A man is cutting a piece of a pair of a paper.

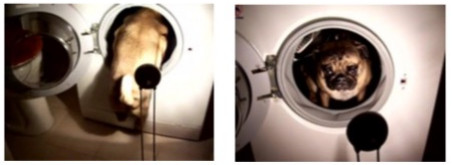
S2VT: A cat is trying to get a small board.



S2VT: A man is spreading butter on a tortilla.

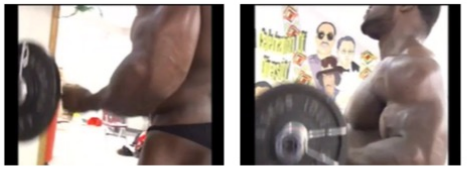
(b)

Irrelevant descriptions.



S2VT: A man is pouring liquid in a pan.

S2VT: A polar bear is walking on a hill.



S2VT: A man is doing a pencil.

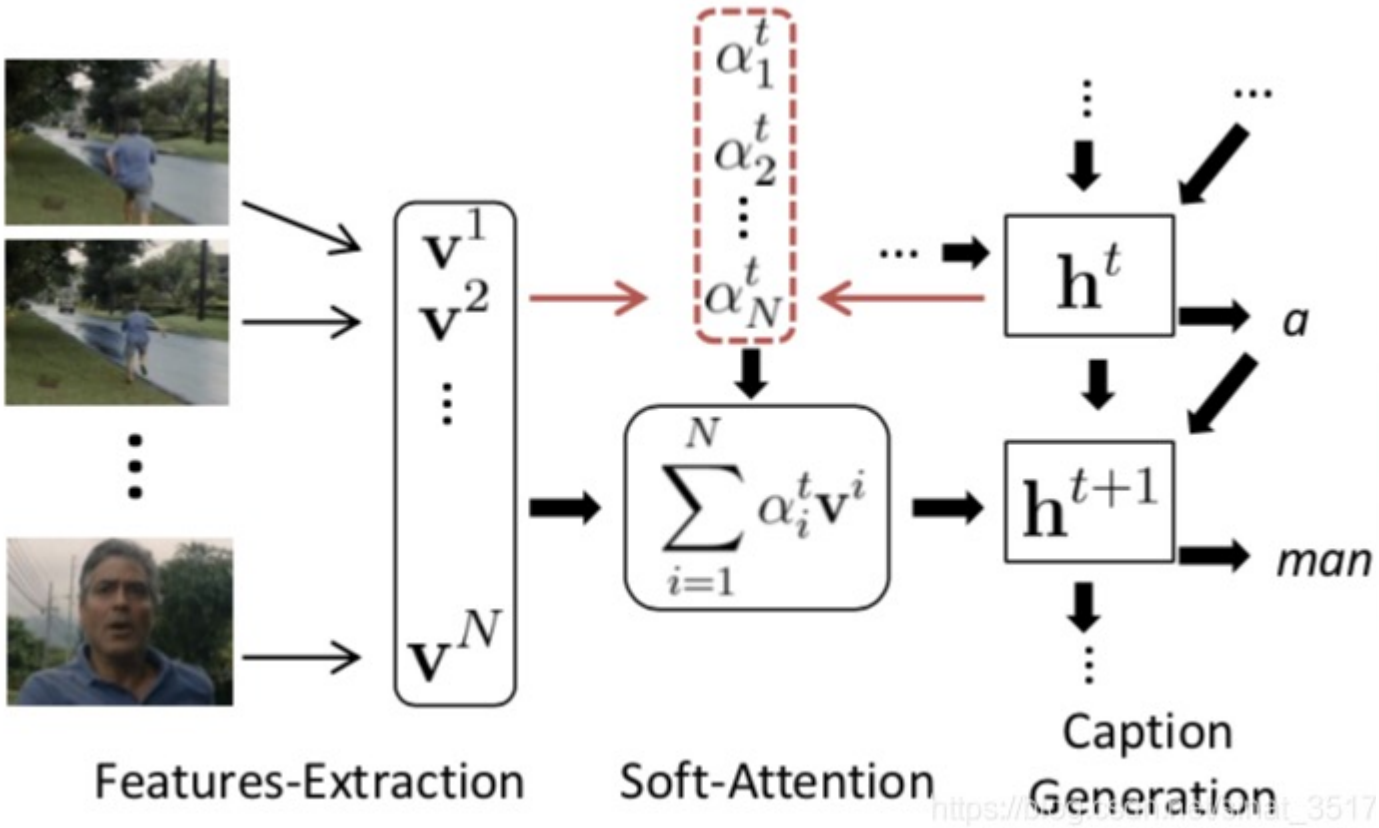
S2VT: A black clip to walking through a path.

(c)

视频与自然语言结合

- Video caption

  引入attention机制，对视频分析有帮助



视频与自然语言结合

- Video caption

引入attention机制，对视频分析有帮助

Model	Feature	Bleu 1	2	3	4	mb	Meteor	Perplexity
non-attention	GNet	77.3	60.7	49.3	39.1	38.6	28.68	33.09
	GNet+3DConv$_{non-att}$	76.1	60.2	49.2	39.0	38.7	27.65	33.42
soft-attention	GNet	79.1	63.2	51.2	40.6	40.3	29.00	27.89
	GNet+3DConv$_{att}$	**80.0**	**64.7**	**52.6**	**42.2**	**41.9**	**29.60**	**27.55**
(Thomason et al., 2014)						13.68	23.9	
(Venugopalan et al., 2014)	No Pretraining					31.19	26.87	
	Pretraining					33.29	29.07	

Corpus:
She rushes out.
Test_sample:
The woman turns away.

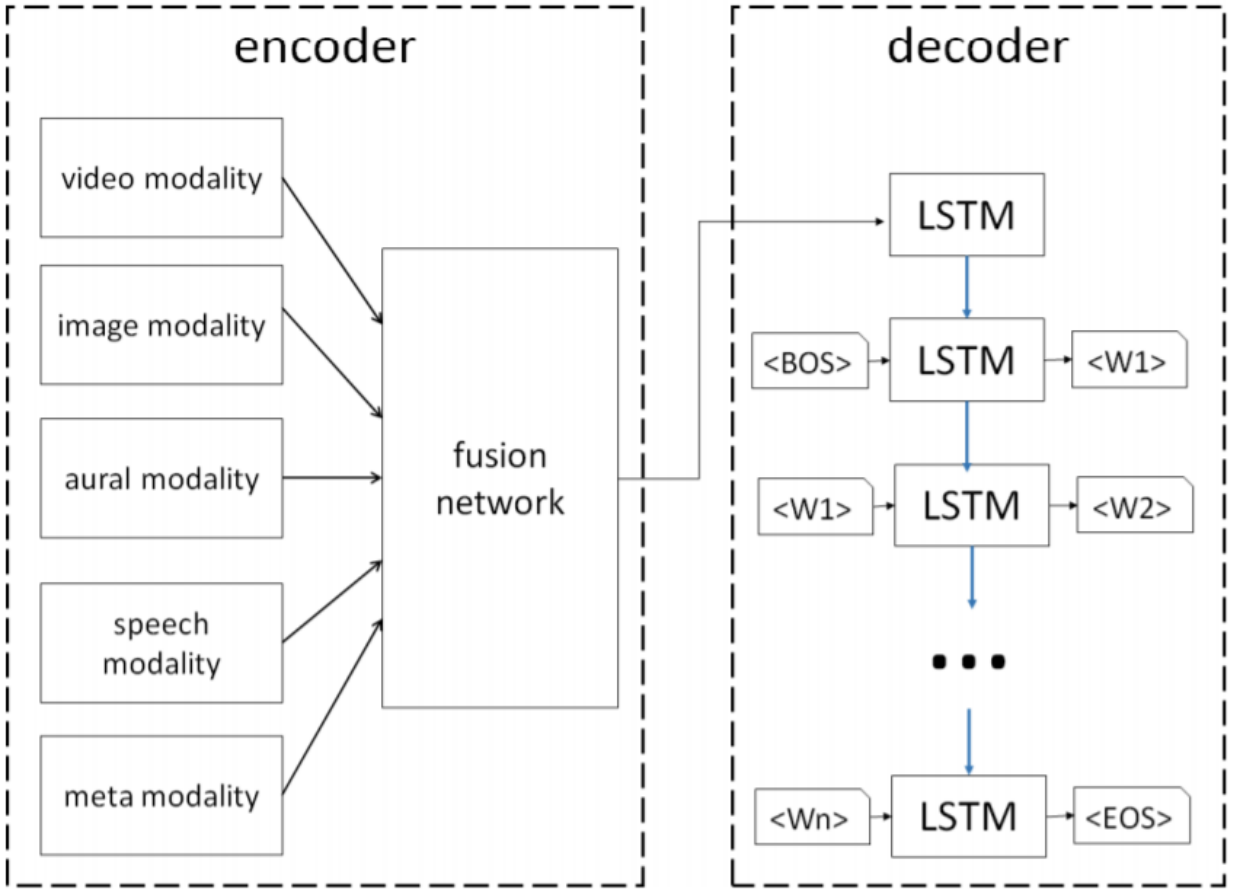
Corpus:
SOMEONE sits with his arm around SOMEONE.
He nuzzles her cheek, then kisses tenderly.
Test_sample:
SOMEONE sits beside SOMEONE.

Corpus:
SOMEONE shuts the door.
Test_sample:
as he turns on his way to the door , SOMEONE turns away.

视频与自然语言结合

- Video caption

  引入多模态融合机制，对视频分析有帮助



视频模态
图像模态
语音模态
等

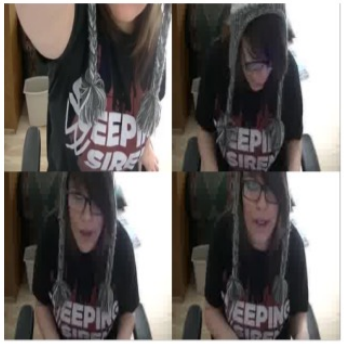
视频与自然语言结合

- Video caption

  引入多模态融合机制，对视频分析有帮助

Table 1: Performance of Multi-Modality Fusion

modality	model	BLEU@4	METEOR	ROUGE	CIDEr
video	c3d	36.94	27.27	58.40	41.85
	c3d+iDT	34.96	26.59	57.58	36.51
video+aural	c3d+mfccbow	39.80	27.89	60.02	41.07
	c3d+mfccfv	40.07	27.76	60.02	39.60
	c3d+mfccbow+mfccfv	41.32	28.21	60.45	43.66
video+aural+image	c3d+mfccbow+mfccfv+vgg19	41.81	28.67	60.41	43.35
video+aural+speech	c3d+mfccbow+mfccfv+asr	40.55	27.98	60.10	42.28
video+aural+meta	c3d+mfccbow+mfccfv+category	43.70	28.95	61.35	45.74



(a)

(1) A woman is talking to a camera.
(2) **A woman is singing a song.**
GT: A girl is sitting at a piano playing and singing.

(b)

(1) A man is talking to a woman.
(2) There is a suit man is talking with a man.
GT: A man in a white shirt and dark suit jacket is talking about merging living and retail space.

不是特别准确，但是捕捉到关键词

视频解析概述

还有很多视频分析任务：

光流计算，行人重识别等

篇幅有限，不一一作介绍

视觉分析的未来是视频分析

视频分析的未来是结合自然语言

The End