计算机视觉基础：视频分析

底层视觉特征篇

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

- 视频 vs. 图像：

  图像为静态的，视频为动态的，需要更多考虑**时间维度的关联**信息

- 视频分析任务分类：

过去，现在，将来

行为检测，行为对象分割

过去做了什么：行为时间、地点

过去

行为识别

现在在做什么

现在

行为意图预测

将来要做什么：理解行为意图

将来

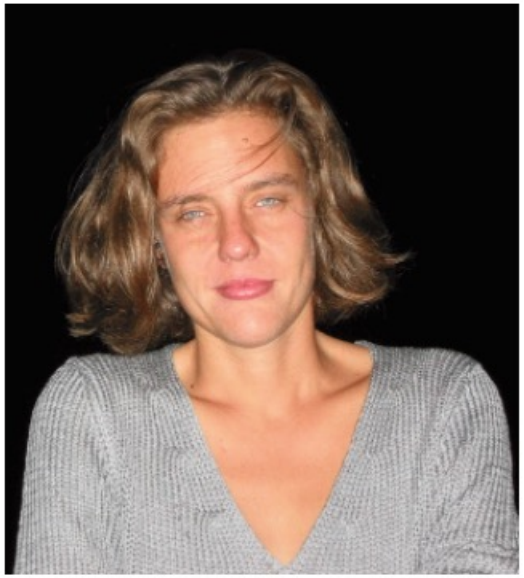
视频解析概述

- 突出研究团队：

  Cordelia Schmid，Inria（法国科学院）

  设计多种视频特征，并被广泛使用

Cordelia Schmid



INRIA Research Director, Head of the THOTH project-team
655, avenue de l'Europe
38330 Montbonnot, FRANCE
Tel: 04 76 61 54 47
E-Mail: Cordelia.Schmid "at" inria.fr

Research interests: image and video descripton, object and category recognition, machine learning

视频解析概述

概述：简单的叙述回顾，欲知详细细节，请看相关的论文文献

视频特征提取方法

- 基于手工设计的特征：

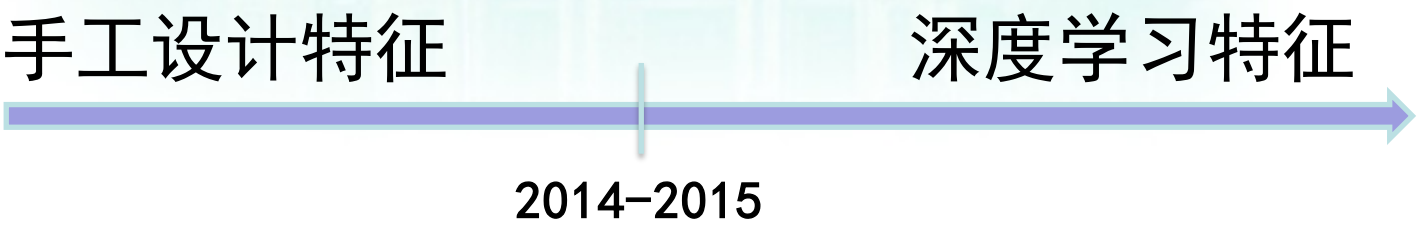
  根据人工经验，设计特征计算方法，需要较强的工程技术能力。

  手工设计很难，几十年来也就是那么几套方案

- 基于深度学习的特征：

  手工设计网络结构，以学习的方式，提取视频特征

  自动学习网络结构，提取网络特征



手工设计的特征

- 时空兴趣点方法 （space-time interest point）：

  提取大规模的小时空区域特征信息，基于统计意义

  细粒度（关注非常局部小区域的时空关联，如3帧）

- 稠密轨迹方法 （dense trajectory）：

  短时间的轨迹特征，长时间的轨迹特征

  中粒度（关注较大局部区域的时空关联，如10+帧）

- 模板匹配方法（不作介绍）

  粗粒度（空间区域大，不够鲁棒）

- 结构化特征

  利用人体的结构信息，姿态

时空兴趣点法

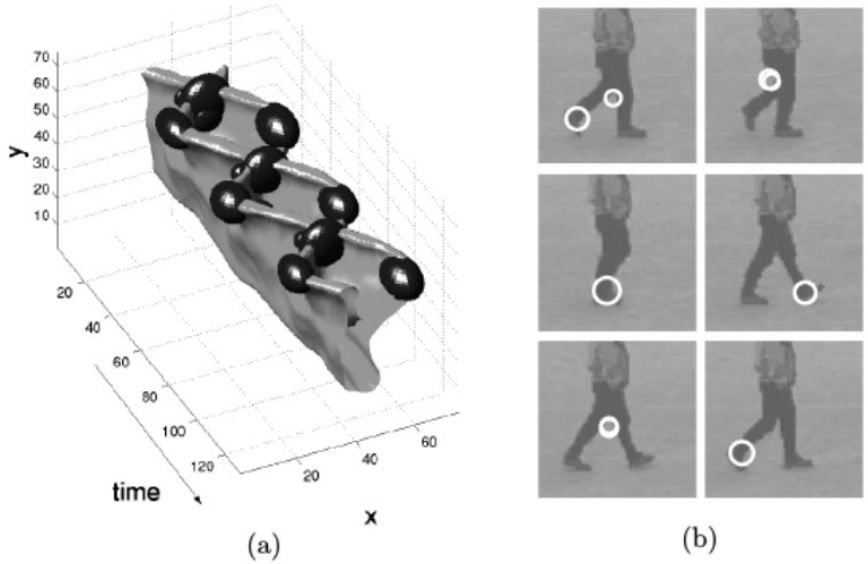
- 时空兴趣点方法 （space-time interest point）：
  1. 时空兴趣点检测
  2. 时空兴趣点区域表示
  3. 视频特征表示

时空兴趣点法

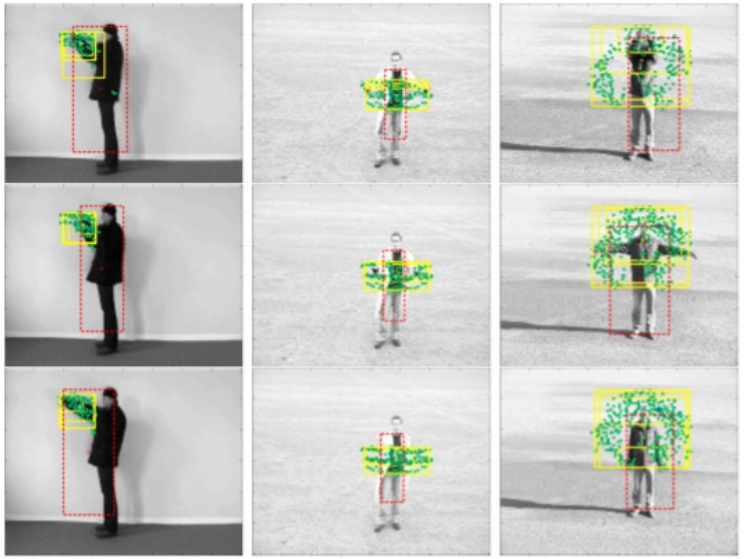
- 时空兴趣点检测：

  1. 3D Harris角点检测：检测运动方向变化的点，检测出来的点较少，统计结果不是很可靠

  2. 时空特征点检测：基于时域Gabor滤波器设计，可以检测出周期性运动点，检测的点较多



Harris 检测

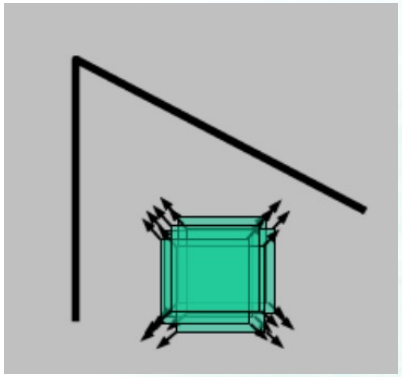


时空特征点检测

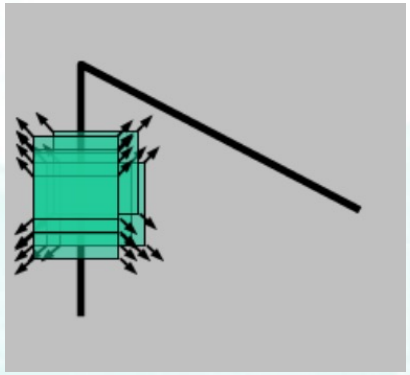
时空兴趣点法

- 2D图像角点检测：

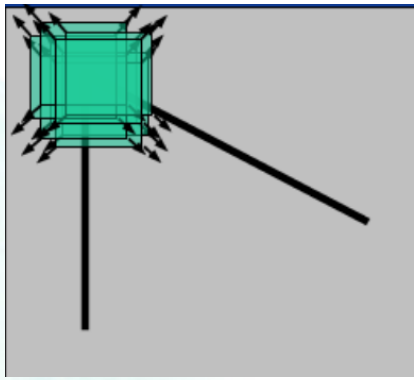
  小窗口向任意方向移动都导致图像灰度有明显变化



平坦区域：
任意移动，无变化



边缘：
边缘方向移动，无变化



角点：
任意方向移动，明显变化

$$E(u,v) = \sum_{(x,y)\in W} w(x,y)[I(x+u,y+v) - I(x,y)]^2$$

窗口

窗口函数

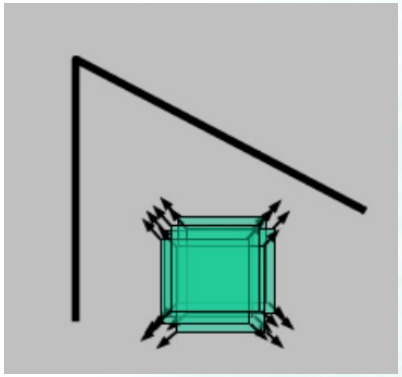
$$E(u,v) = \begin{bmatrix} u & v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix} \quad M = \sum_{(x,y)\in W} w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$$

可以转换为特征值问题

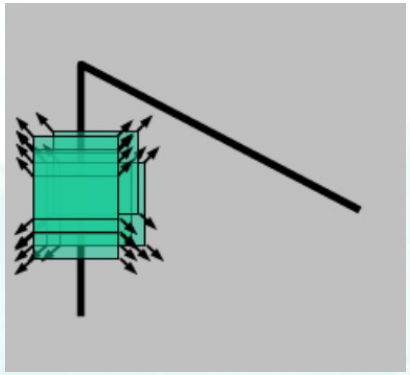
时空兴趣点法

- 2D图像角点检测：

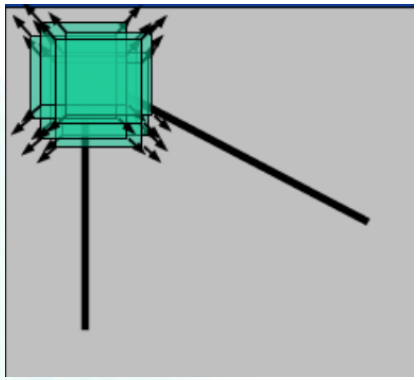
  小窗口向任意方向移动都导致图像灰度有明显变化



平坦区域：
任意移动，无变化

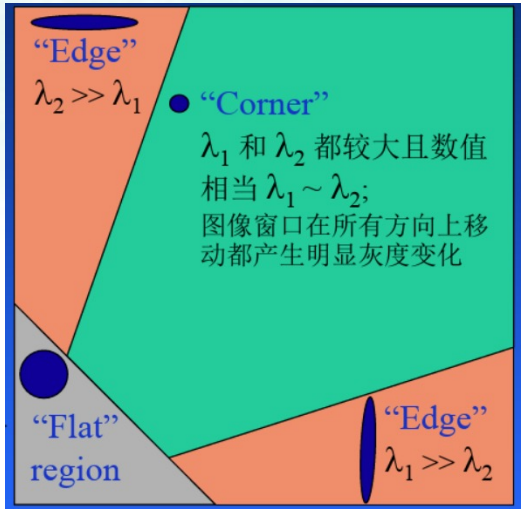


边缘：
边缘方向移动，无变化



角点：
任意方向移动，明显变化

$$M = \sum_{(x,y)\in W} w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$$



时空兴趣点法

- 时空特征点检测：

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

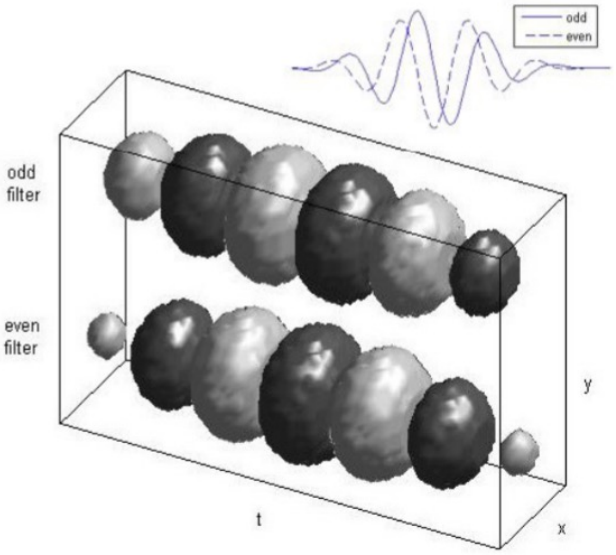
高斯函数，光滑作用

视频图像矩阵，m*n*T

Gabor滤波器

$$h_{ev}(t;\tau,\omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2}$$

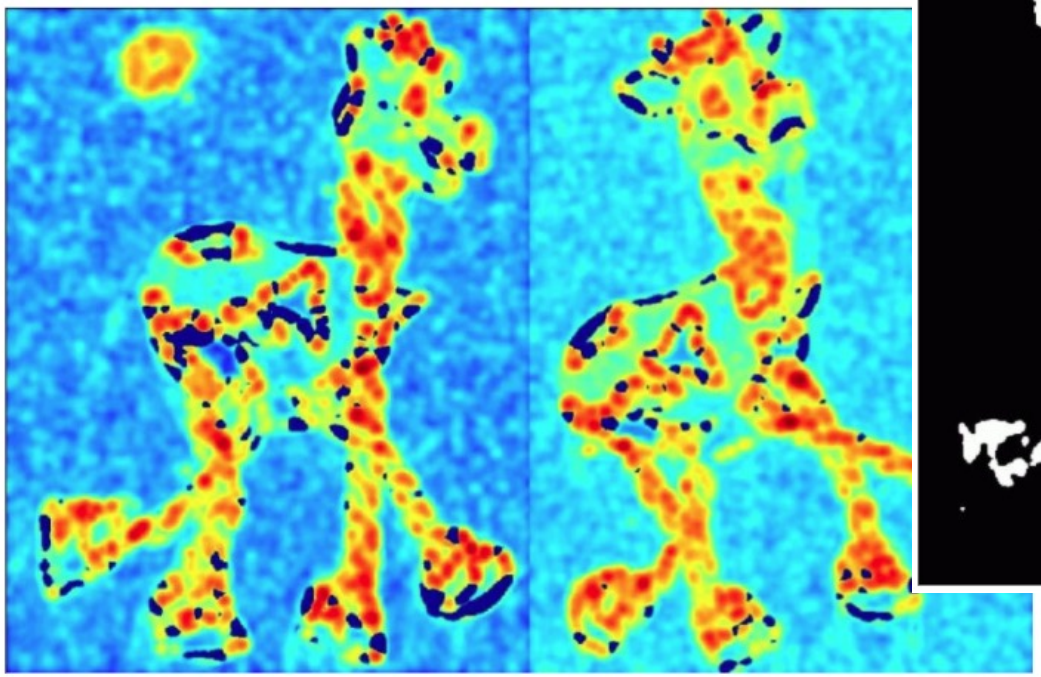
$$h_{od}(t;\tau,\omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2}$$



响应函数R的局部最大值，即为时空特征点

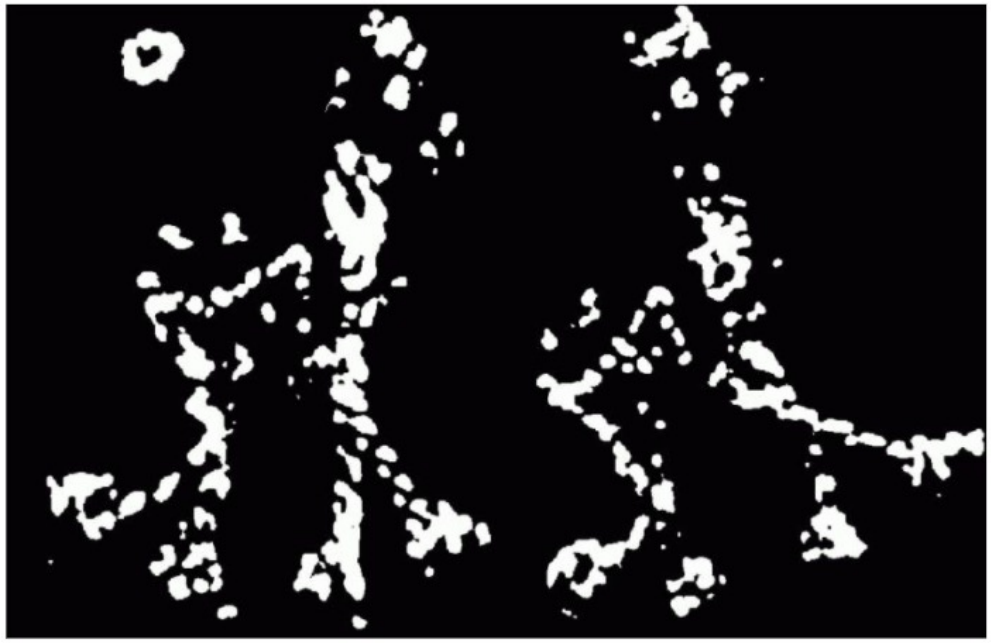
时空兴趣点法

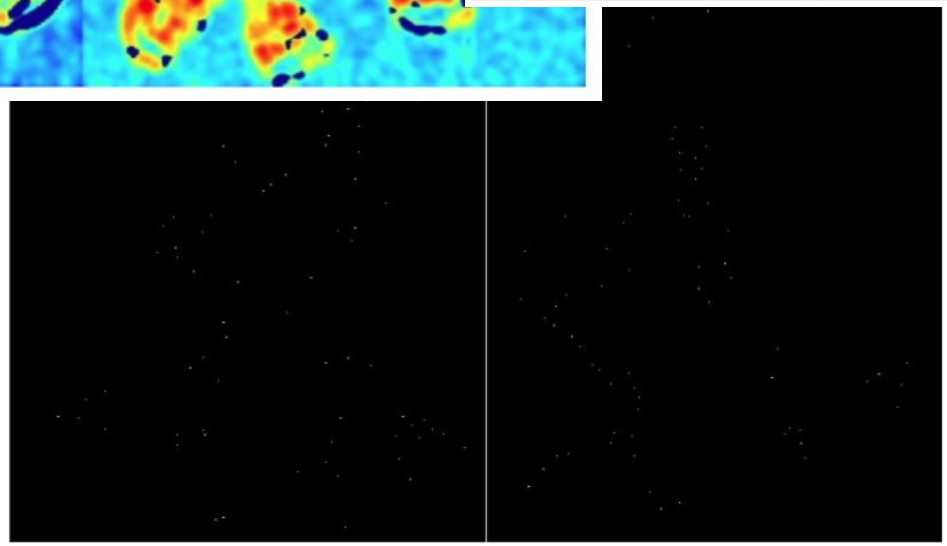
- 时空特征点检测：



响应 R

二值化的 R



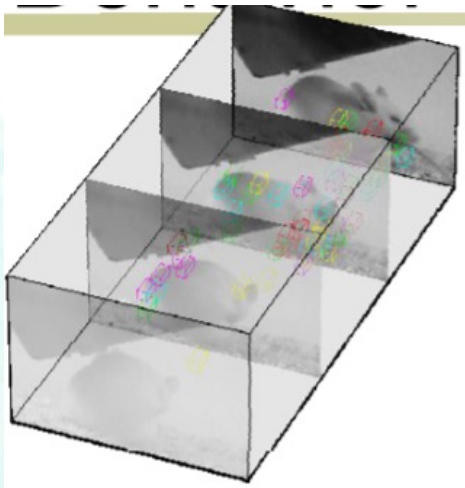


R局部最大值

时空兴趣点法

- 时空兴趣点特征表示：

  每个兴趣点周围提取局部立方块，该邻域包含显著的运动和动作外观信息



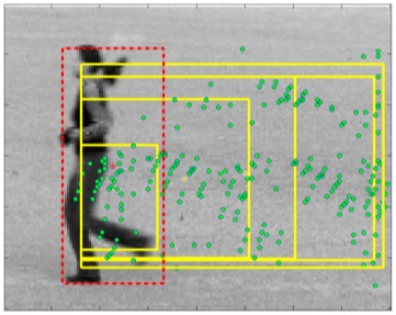
方法一：　**提取梯度特征，**
梯度能比较好刻画变化信息

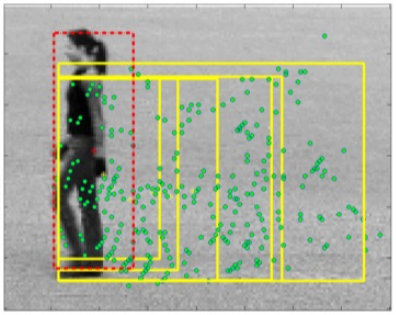
$$j = (L_x, L_y, L_t, L_{xx}, \ldots, L_{ttt})\Big|_{\sigma^2=\tilde{\sigma}_i^2, \tau^2=\tilde{\tau}_i^2}$$

$$L_{x^m y^n t^k} = \sigma^{m+n}\tau^k (\partial_{x^m y^n t^k} g) * f,$$

方法二：

$$[C_s^r, C_s^{Sp}, C_s^D, C_s^{Vd}, C_s^{Hd}, C_s^{Hr}, C_s^{Wr}, C_s^{Or}]$$





提取点云特征，
点云的密度，窗口大小等

鲁棒性更强，对个别噪声点不敏感

时空兴趣点法

- 视频特征表示：

  每个视频可以提取到很多个兴趣点特征，注意，不同视频特征点个数可能不一样。怎么去生成一个维度统一的视频特征？

1）pooling 方法：均值，和，最大值

2）词袋模型:直方图表示

时空兴趣点法

- 视频特征表示：词袋模型（Bag of words）

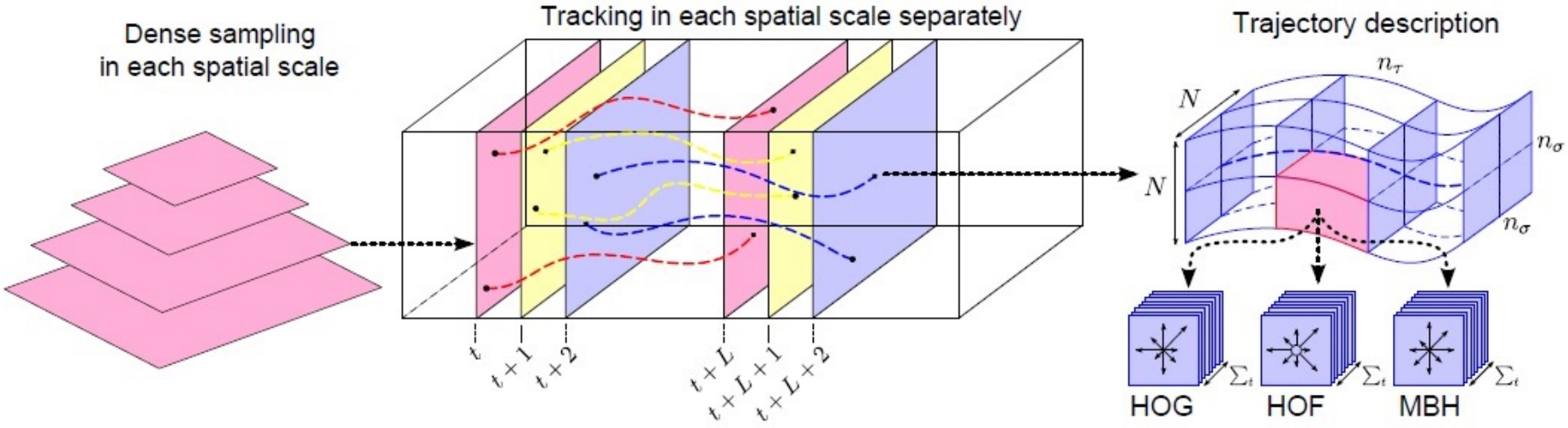
  构造词典，k-means 聚类（用卡方距离，曼哈顿距离）

  计算视频直方图特征

为了更好地学习词典和直方图特征：**词典学习+稀释表达模型**

稠密轨迹方法

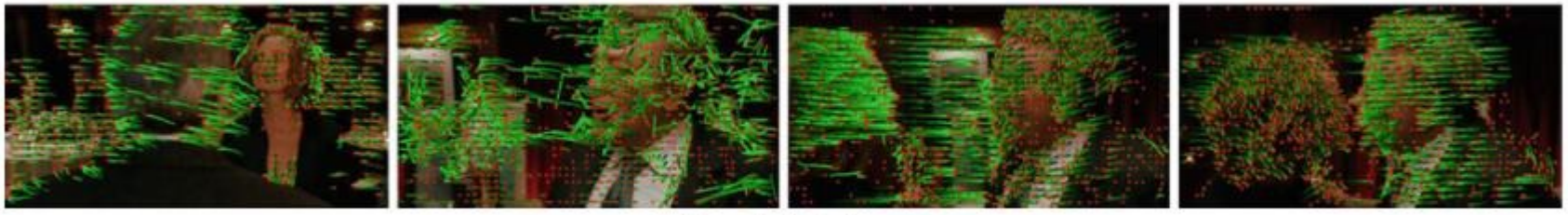
- 视频特征表示：
  1）将视频通过网格划分的方式，多个尺度上分别密集采样特征点
  2）跟踪这些特征点，形成轨迹
  3）轨迹描述子，提取HOG,HOF,MBH特征
  4）特征编码，Bag of features



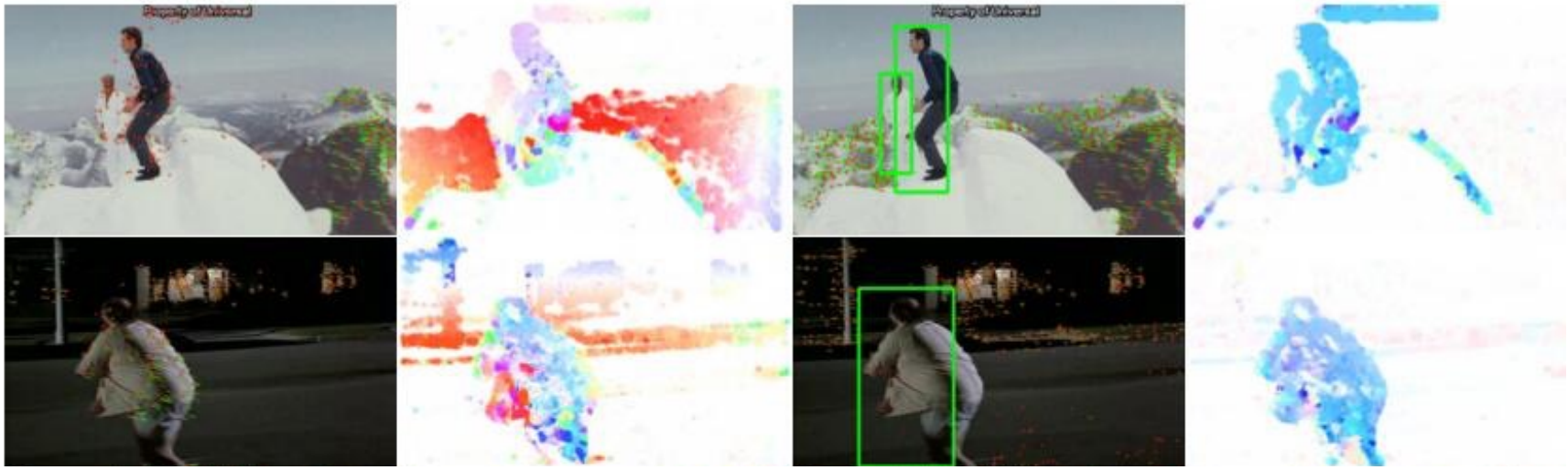
稠密轨迹方法

- 改进稠密轨迹方法：

  1）估计相机运动估计来消除背景上的光流以及轨迹减少与内容运动无关的噪声，特征更加鲁棒



Dense trajectories



稠密轨迹方法

- 光流（optical flow）计算：

  基本假设：1)亮度恒定不变；2)时间连续或运动幅度比较小

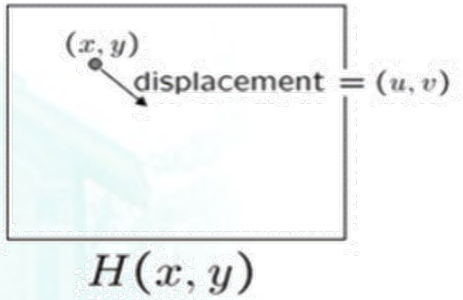
$$I(x,y,t) = \mathrm{I}(\mathrm{x}+\mathrm{dx},\mathrm{y}+\mathrm{dy},\mathrm{t}+\mathrm{dt})$$

$$I(x,y,t) = \mathrm{I}(\mathrm{x},\mathrm{y},\mathrm{t}) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \varepsilon$$

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t}\frac{dt}{dt} = 0$$

$$I_x u + I_y v + I_t = 0$$

光流



约束条件：LK 光流法约束方程；

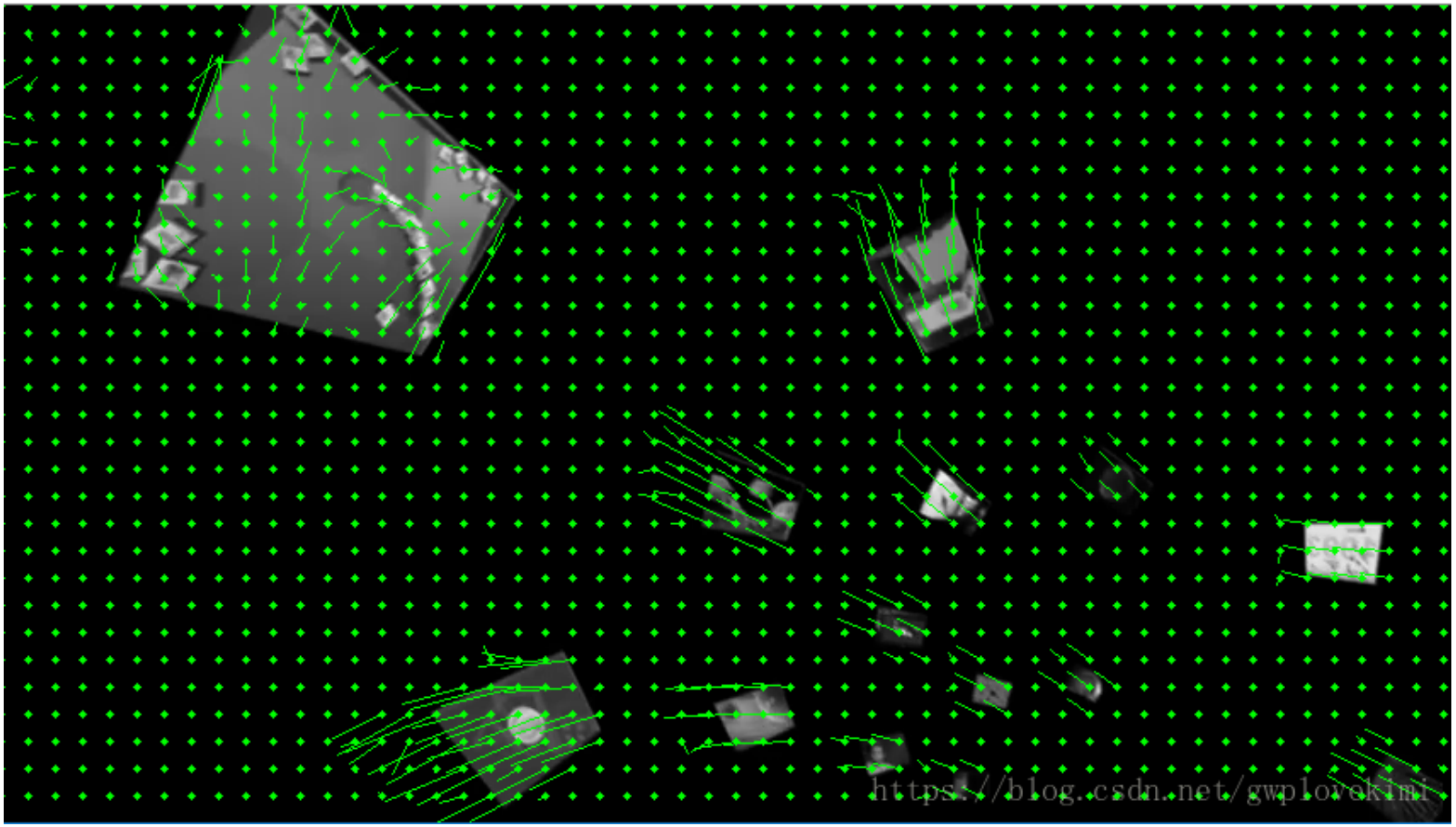
$$\sum_{(\mathrm{x},\mathrm{y}\in\Omega)} W^2(x)\left(I_x u + I_y v + I_t\right)^2$$

$$(A^T W^2 A)^{-1} A^T W^2 b.$$

稠密轨迹方法

- 光流（optical flow）计算：

  基本假设：1)亮度恒定不变；2)时间连续或运动幅度比较小

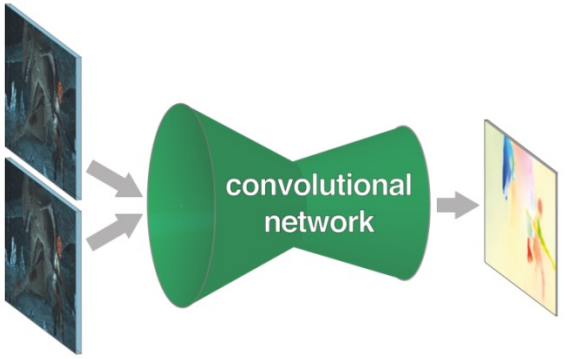


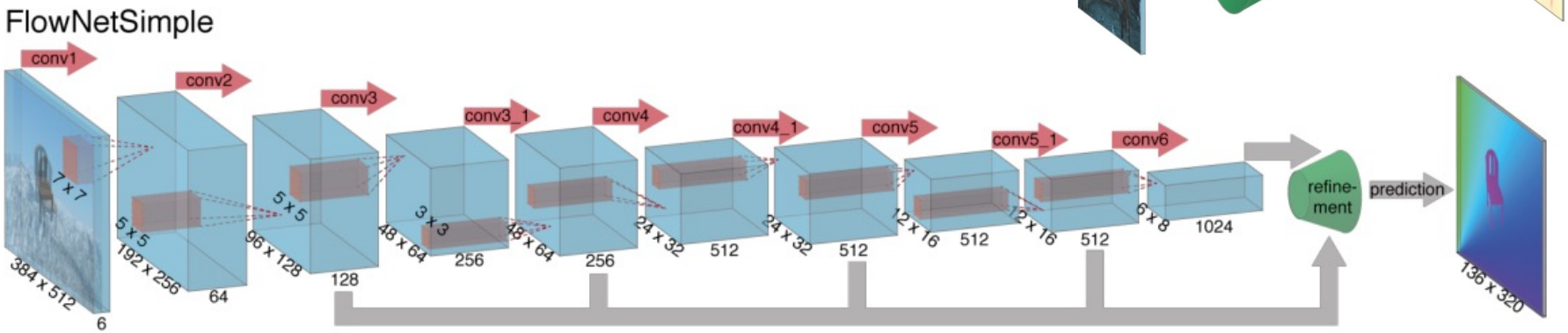
稠密轨迹方法

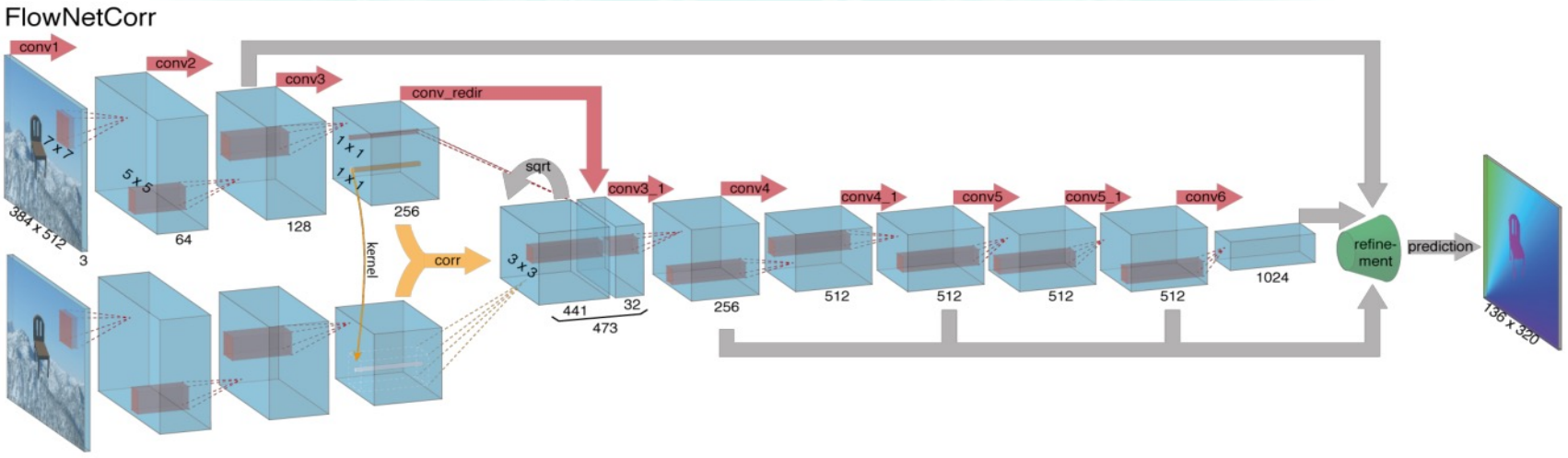
- 光流（optical flow）计算：

  光流计算耗时间和资源。

  基于深度学习的光流估计，FlowNet

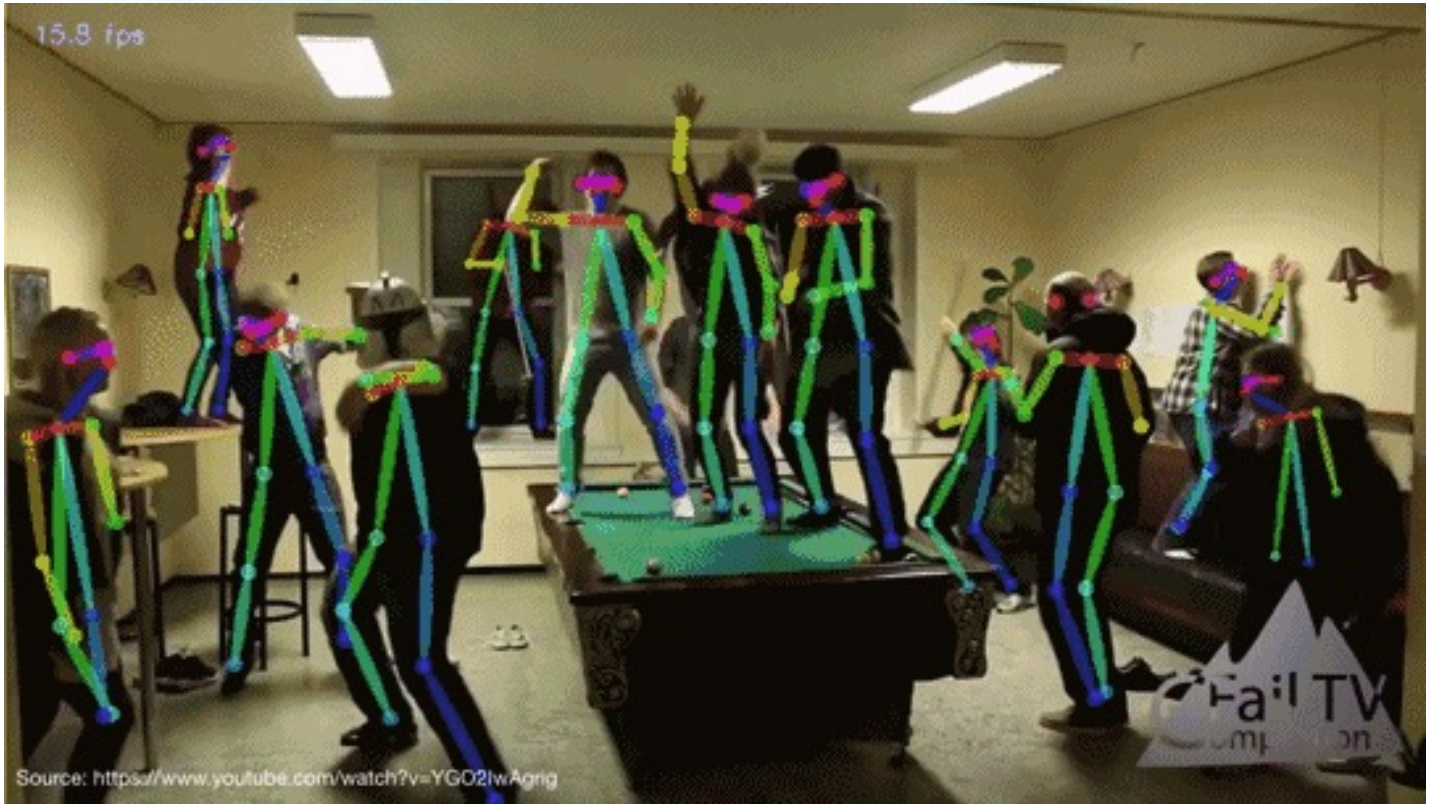






结构化特征

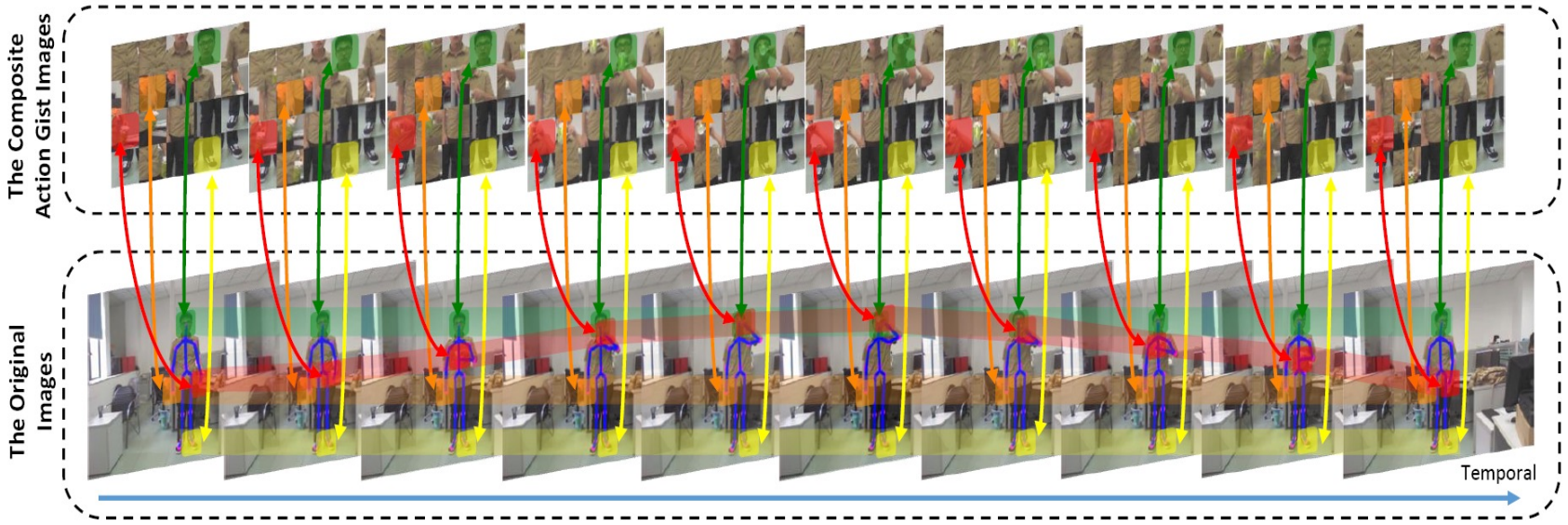
- 人体结构信息：
  1）人体具有复杂的结构信息，如姿态和轮廓
  2）通过观察结构信息，可以有效发现人体的运动



https://v.qq.com/x/page/l08258ssxie.html

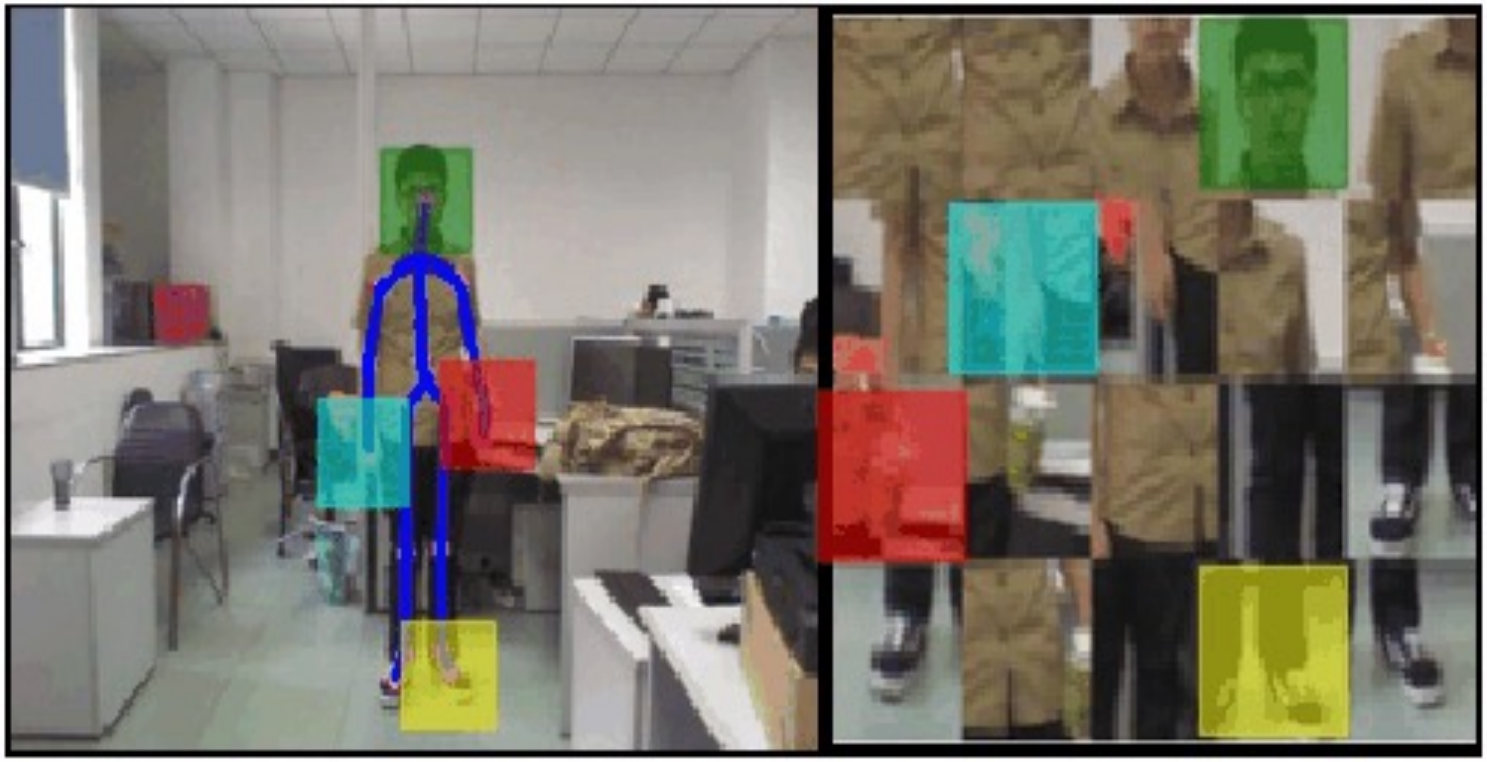
结构化特征

- 人体结构信息：
  1）结合人体结构，可以获得比光流更稳定可靠的轨迹
  2）可以排除掉背景的干扰
  3）多种特征可以互补，提升效果



结构化特征

- 人体结构信息：
  1）结合人体结构，可以获得比光流更稳定可靠的轨迹
  2）可以排除掉背景的干扰
  3）多种特征可以互补，提升效果



深度学习特征

- 卷积神经网络 （CNN）

  双流神经网络：RGB 和 光流两种模态数据

  3D 卷积神经网络

- 递归神经网络（RNN,LSTM）

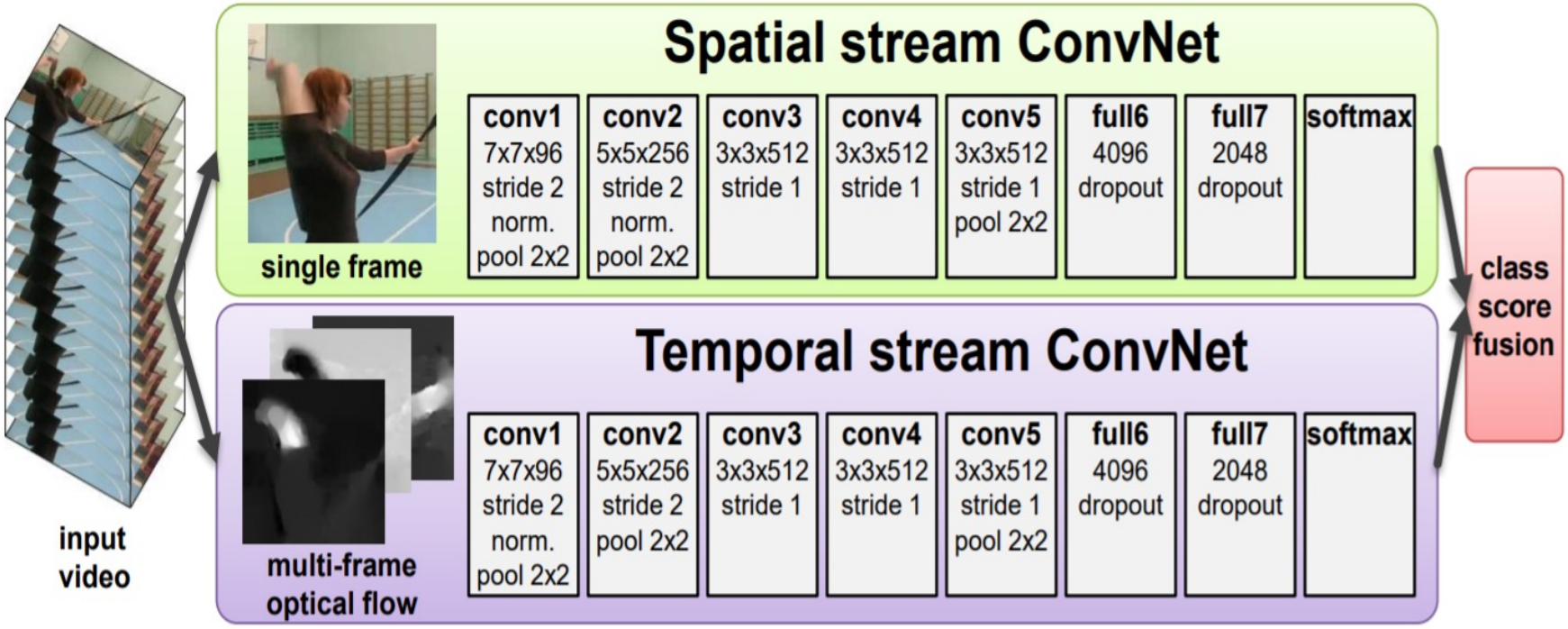
深度学习特征

- 双流神经网络

双流神经网络：RGB 和 光流两种模态数据

图像卷积神经网络技术，直接迁移应用过来

可以方便进行直接拓展，变成3流，4流，N流神经网络

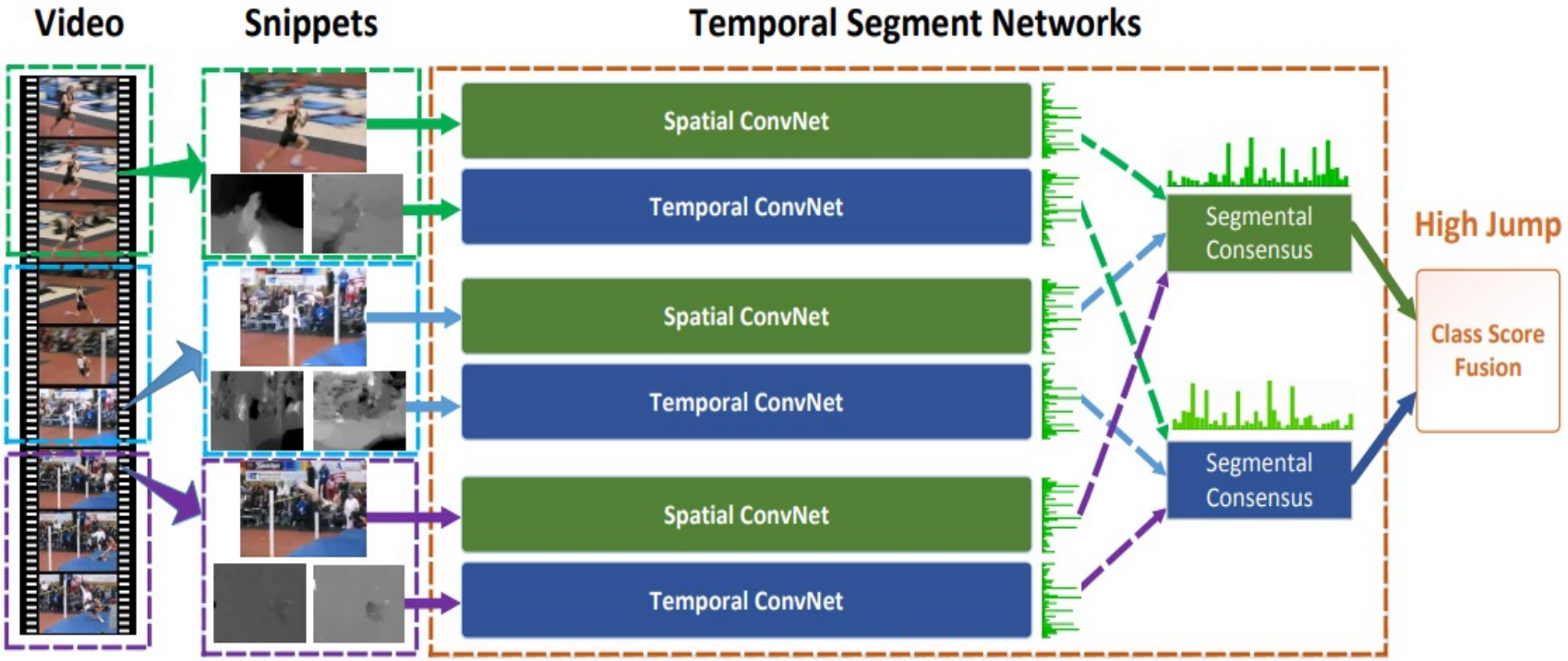


双流神经网络

- 时间分割网络
  （Temporal Segment Network TSN）

  针对每一流的数据，沿着时间维度，切割成多个子视频

  对每个子视频用双流网络做分类，最后结果融合

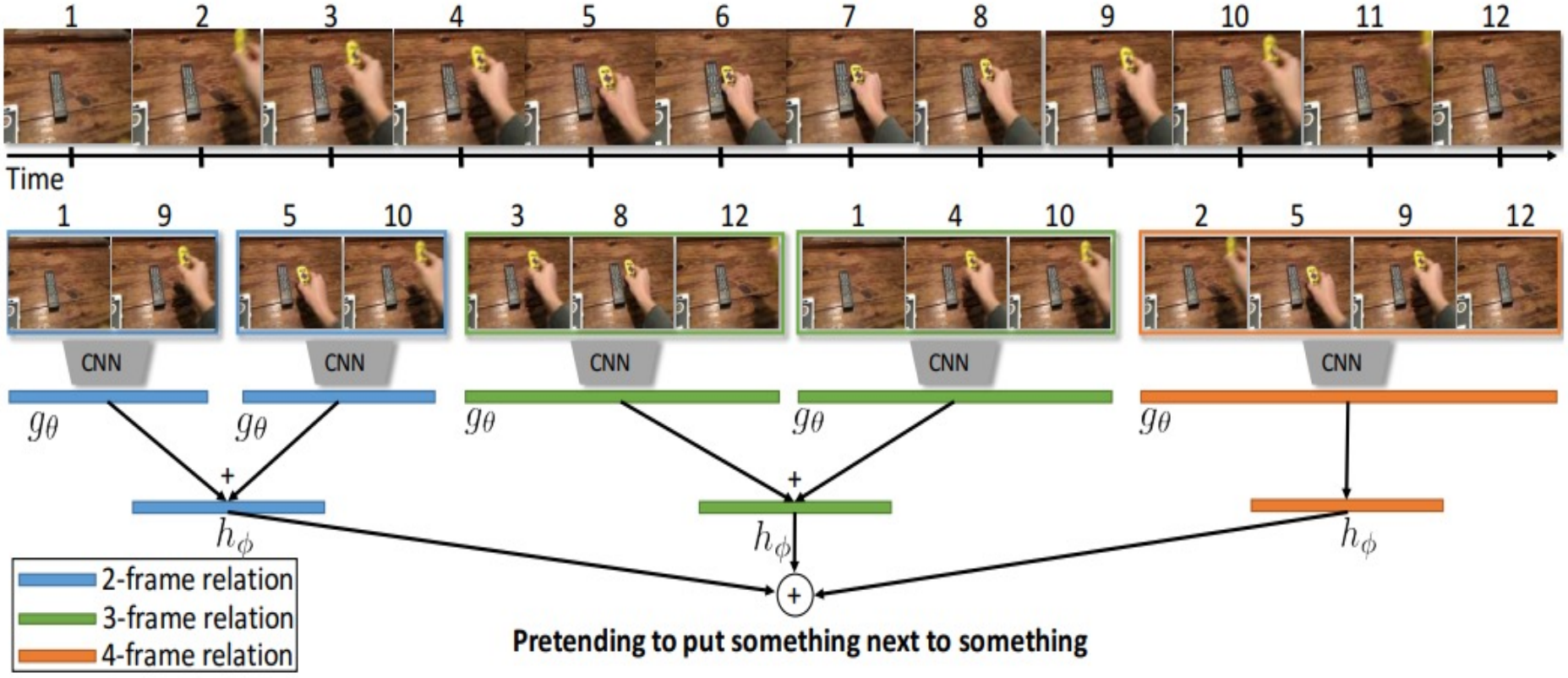


双流神经网络

- 时间分割网络

（Temporal Relational Reasoning in Videos）

在TSN的基础上，多个时间尺度，再进行动作分类



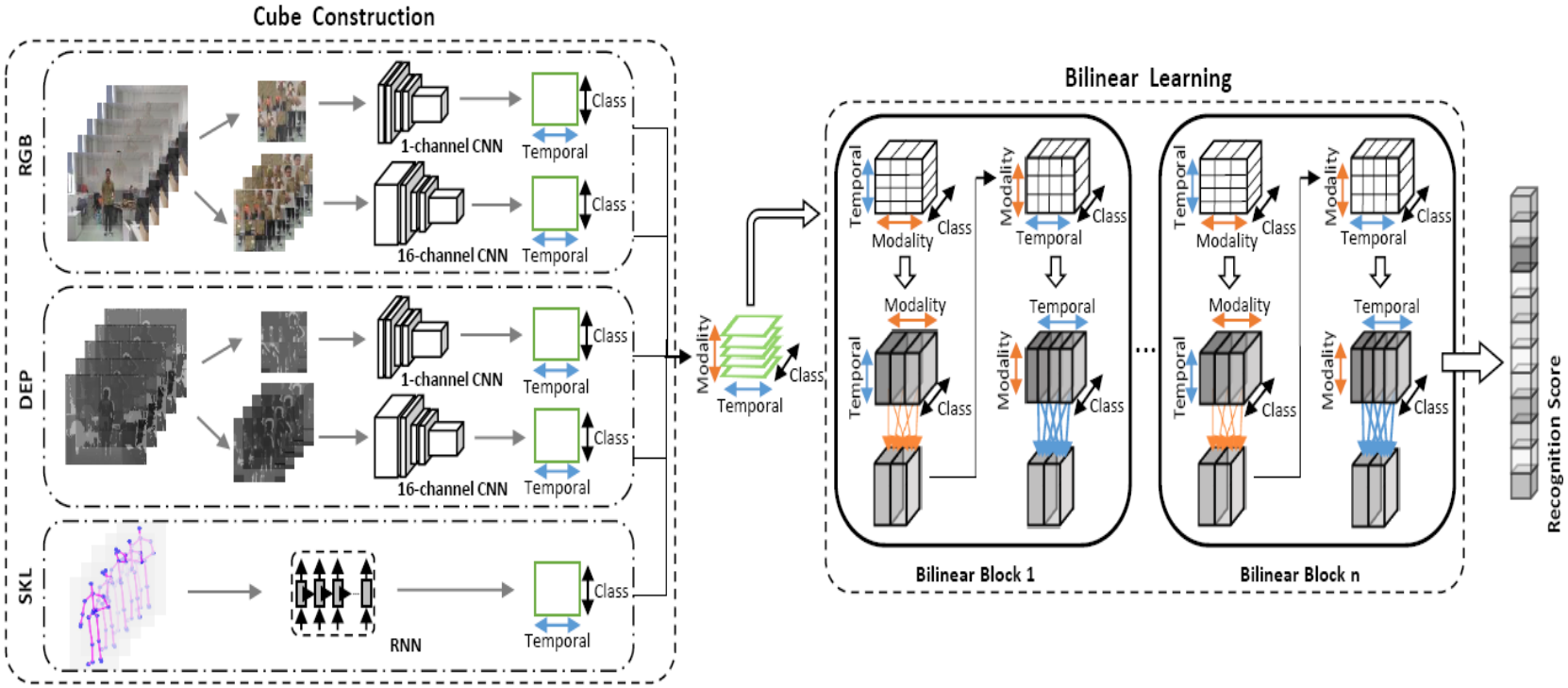
双流神经网络

- 深度双线性网络

对每一流的数据，提取时空特征

构建双线性特征融合模块：时间维度融合不同子视频序列特征；流（模态）维度，融合不同流视频特征

端到端的训练



深度双线性神经网络

- 深度双线性模块

来源于数学上的双线性映射

Bilinear Map Revisited:

$$f(x, y) = x^T A y \Rightarrow f(X, Y) = X^T A Y$$

$$L = X^T A \qquad f(X, Y) = LY$$

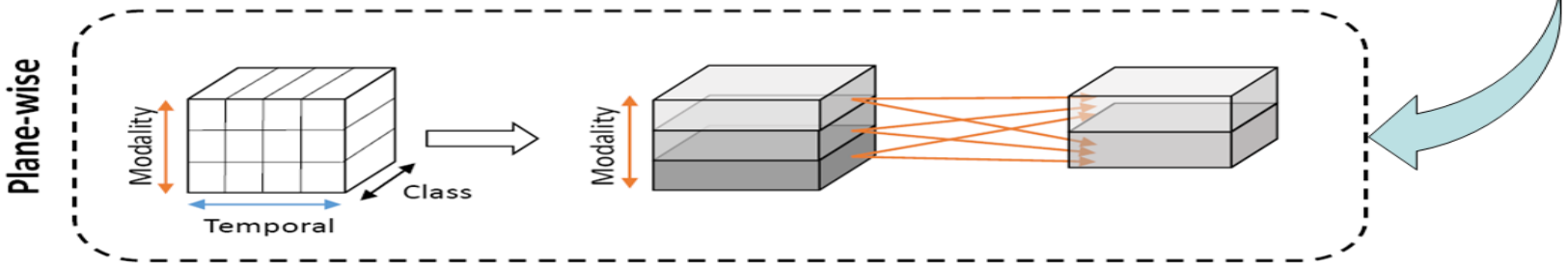
Our Bilinear learning:

$$L(:, :, c) = X^T A(:, :, c), c = 1, 2, ..., C$$

Modality pooling

$$Z(:, :, c) = L(:, :, c)Y, c = 1, 2, ..., C$$

Temporal pooling



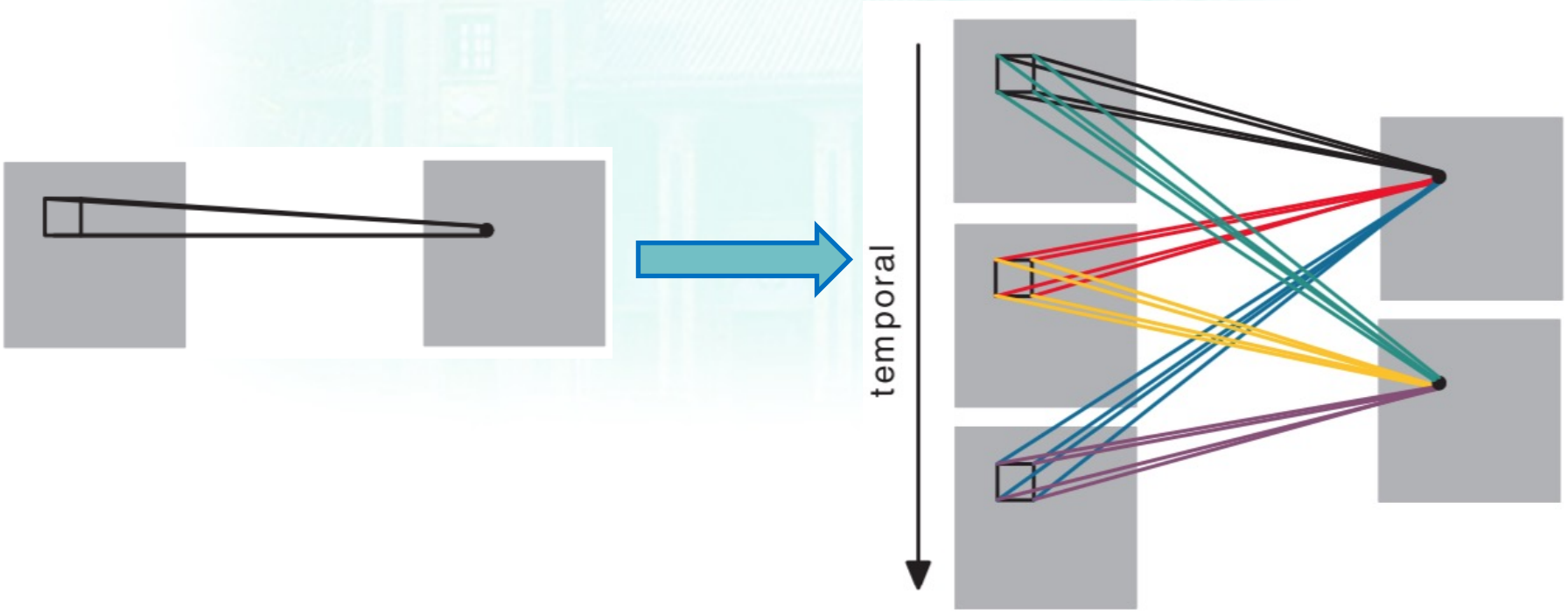
深度学习特征

- 3D卷积神经网络

  将图像中的2D卷积，拓展成带时间维度的3D卷积

  引入时间卷积，可以提取时间和空间维度上的关联信息

  缺点：**计算量大，耗计算资源，需要大规模数据训练**

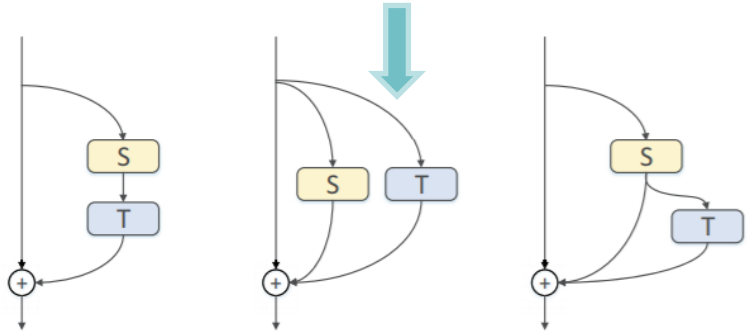


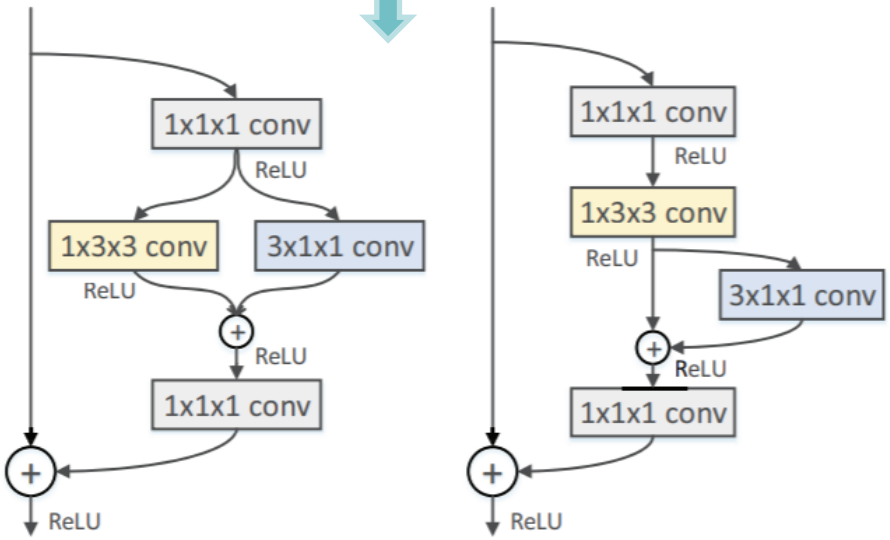
3D卷积神经网络

- 伪3D卷积神经网络

  将三维卷积，拆分成两个不同维度的独立卷积（空间和时间）

  基于残差网络结构





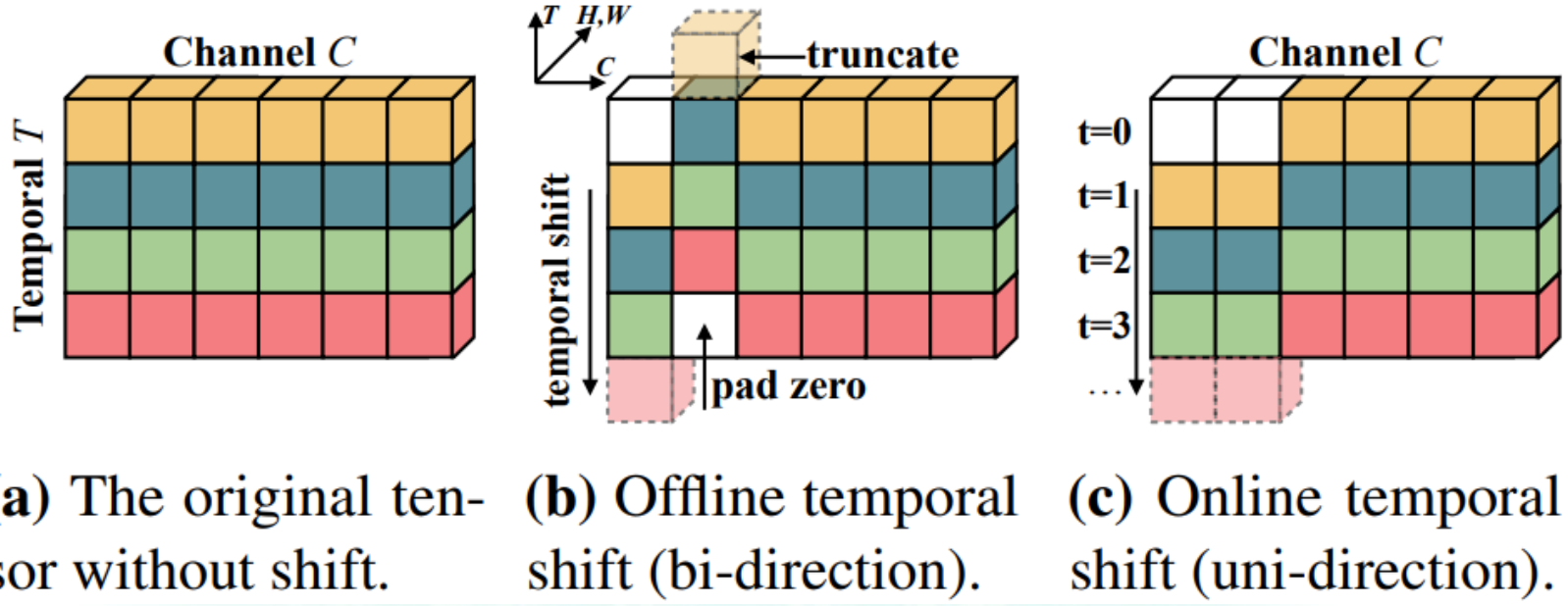
Method	Depth	Model size
C3D	11	321MB
ResNet	152	235MB
P3D ResNet	199	261MB

参数量大规模减少

2D卷积神经网络

- TSM(Temporal Shift Module)

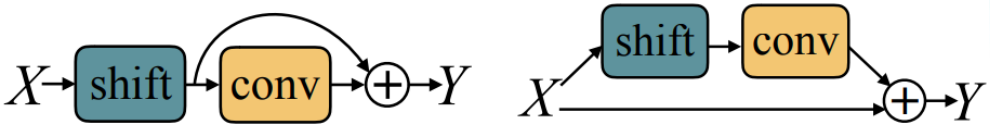
  将时间移位，2D卷积实现时空联合建模



(a) The original tensor without shift.

(b) Offline temporal shift (bi-direction).

(c) Online temporal shift (uni-direction).



(a) In-place TSM.

(b) Residual TSM.

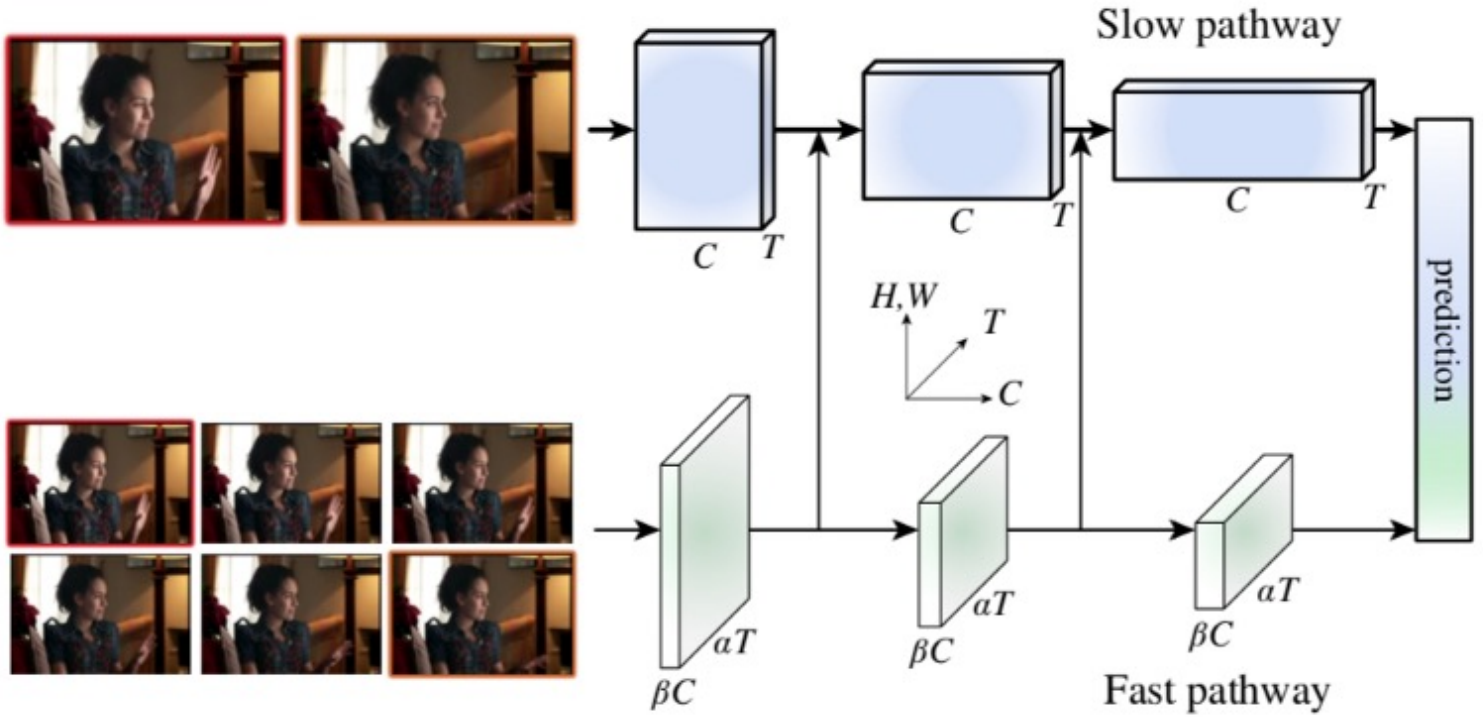
使用残差效果更好

3D卷积神经网络

- SlowFast

  慢速高分辨率CNN(Slow通道)分析视频中的静态内容

  快速低分辨率CNN(Fast通道)分析视频中的动态内容



捕捉不同时间尺度的特征

深度学习特征

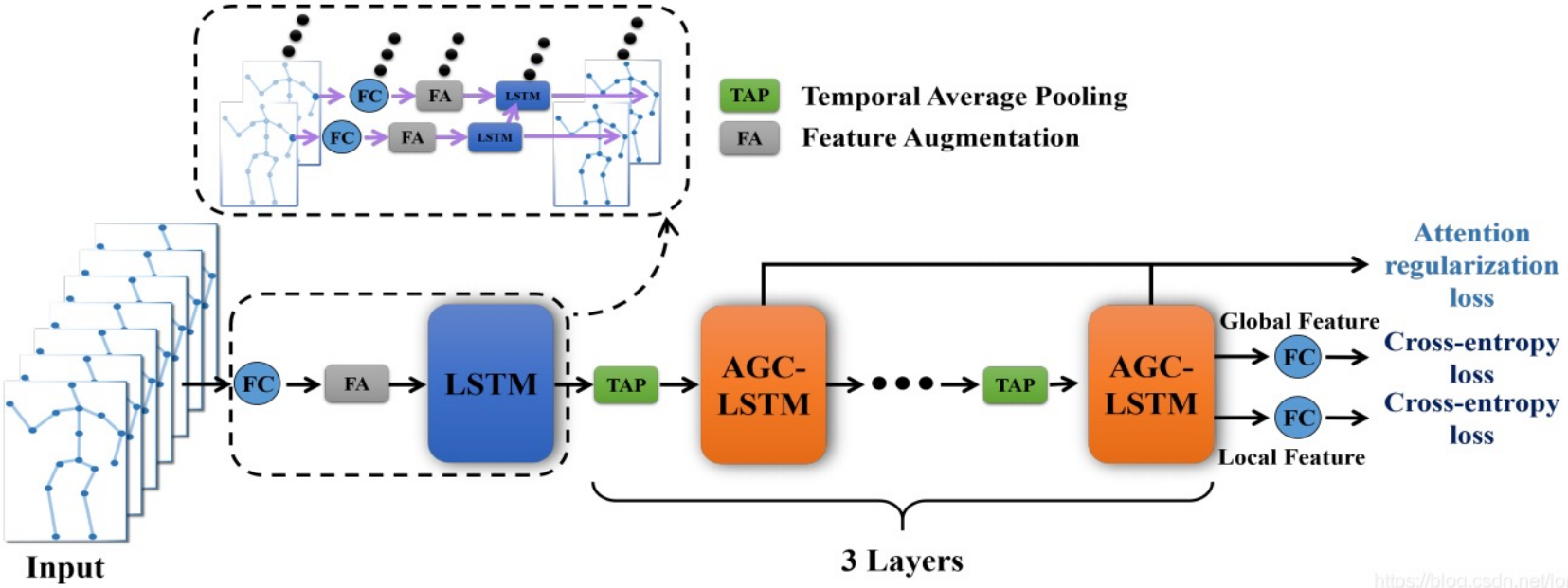
- 递归神经网络（RNN, LSTM）

  递归神经网络对时间上的关联进行直接建模

  挖掘视频帧在长时间上的关联性（CNN一般只能短时间）

  目前来看，**对于冗余度较高的视频帧，效果不是太好，**

  对于高度结构化的数据，效果较好

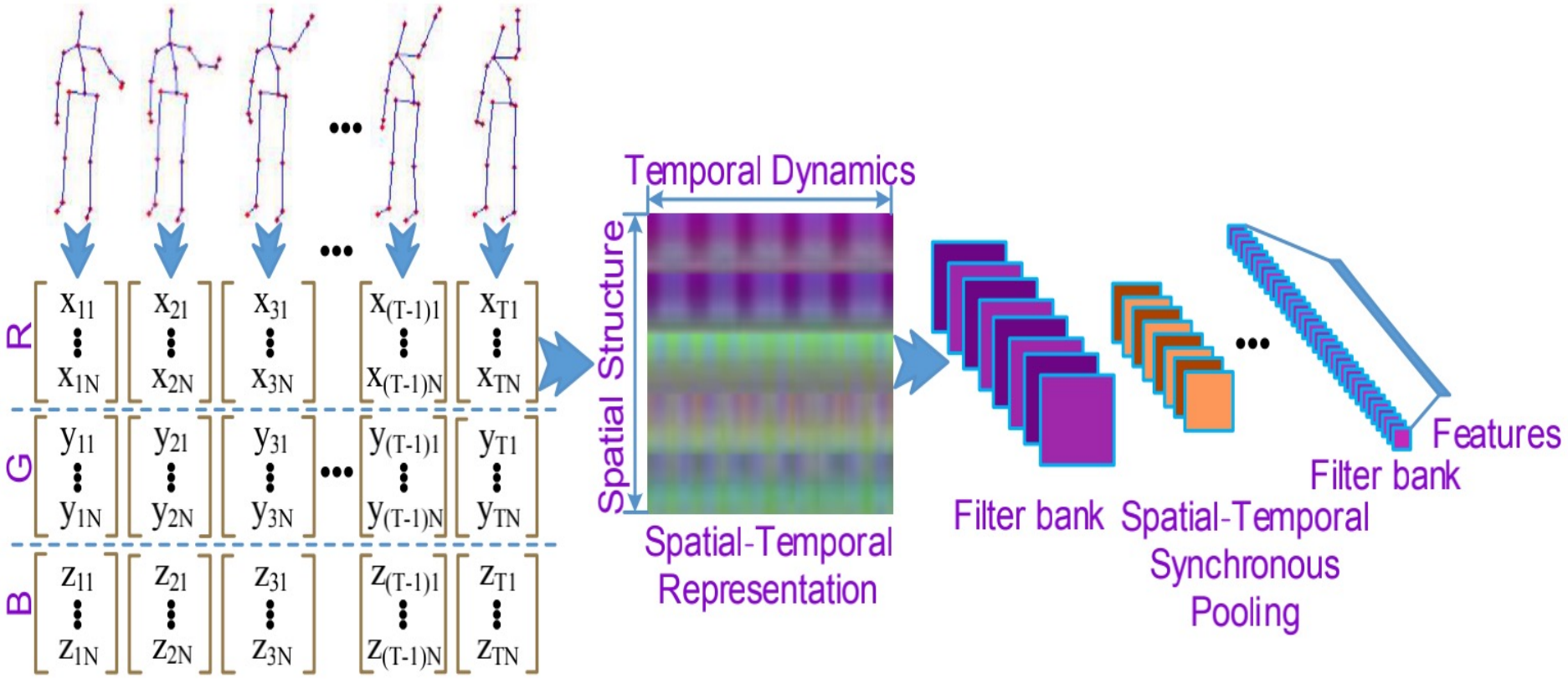


深度学习特征

- 卷积神经网络：结构化数据

  卷积神经网络提取结构化数据的特征

  结构化数据转成可视化图像数据

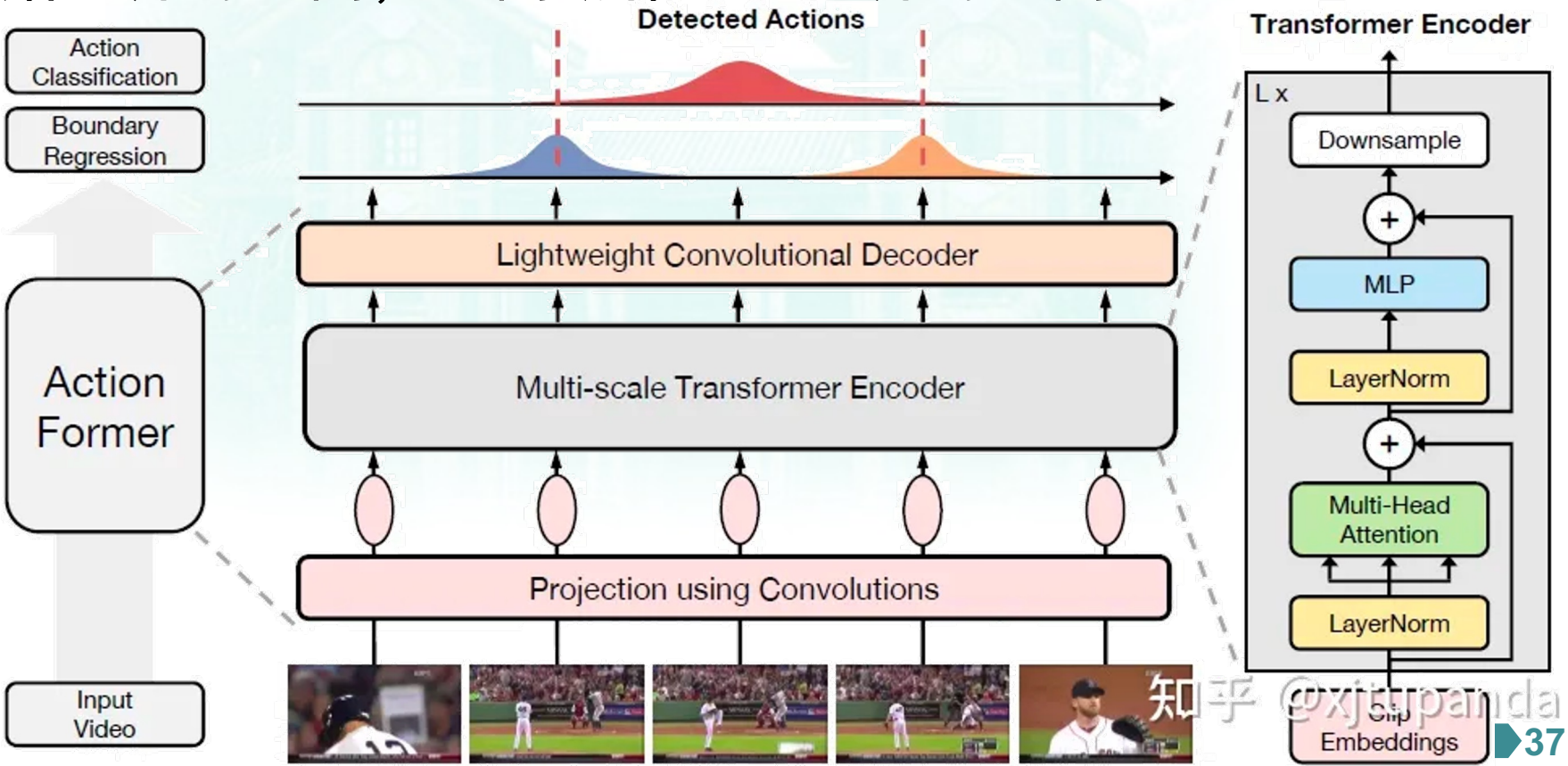


深度学习特征

- Transformer架构：actionformer

  模型的输入：视频提取出的特征序列

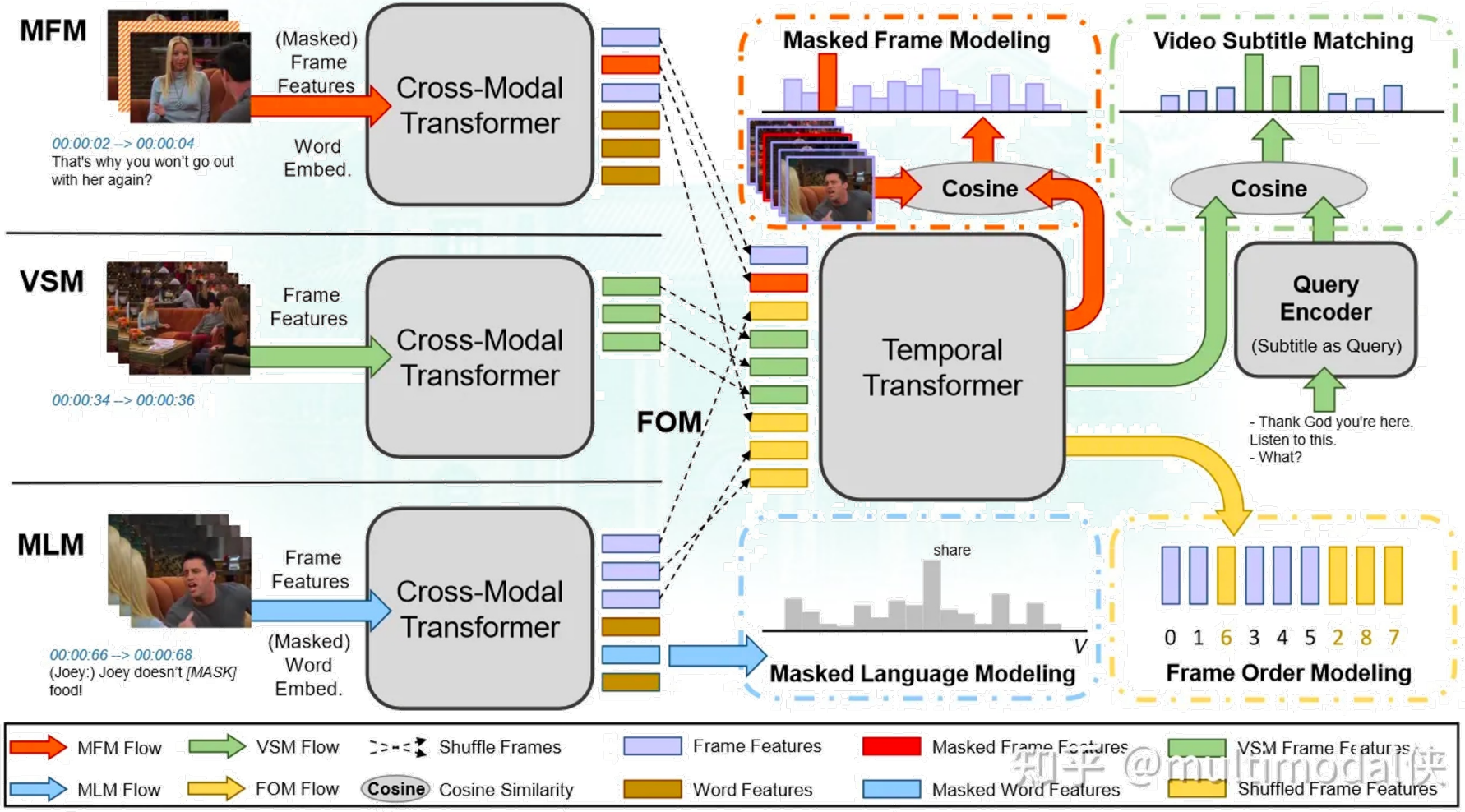
  模型输出：三元组，各个动作类别的置信度，距离动作开始边界的距离，距离动作结束边界的距离



深度学习特征

- Transformer架构：多任务预训练

  视频语言预训练模型HERO



视频特征提取方法

- 还有其它方法

图匹配

图卷积神经网络

异或图

模板匹配

……

方法是多样的
目标是一致的

提取时间与空间关联信息