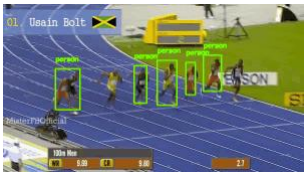
计算机视觉

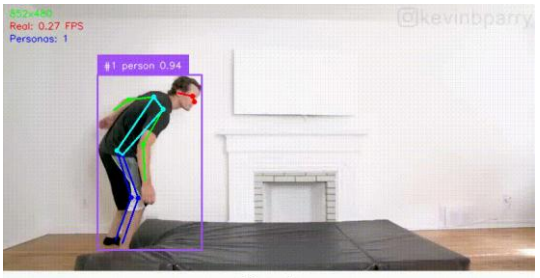
第8章 物体跟踪

陈飞: chenfei314@fzu.edu.cn

本章内容

- 物体跟踪概述
- 经典物体跟踪算法
- 物体跟踪的发展
- 常用数据集
- 评价标准

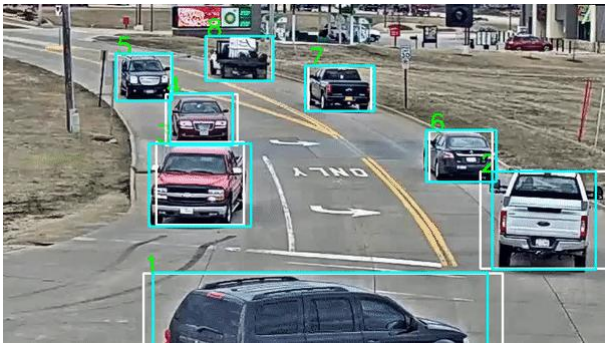




Drunk

8.1物体跟踪概述

- 目标跟踪是计算机视觉领域的一个重要问题，目前广泛应用在体育赛事转播、安防监控和无人机、无人车、机器人等领域.



3

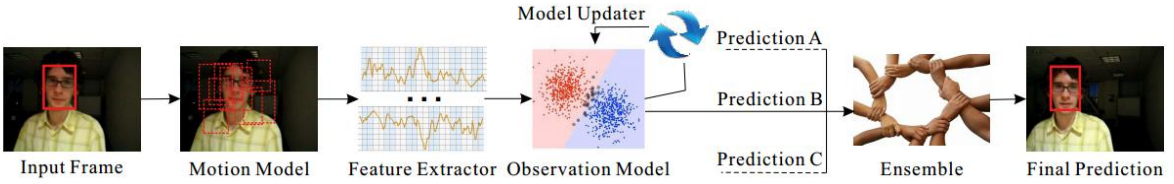
目标跟踪任务

- 单目标跟踪：给定一个目标，追踪这个目标的位置。
- 多目标跟踪：追踪多个目标的位置
- Person Re-ID：行人重识别，是利用计算机视觉技术判断图像或者视频序列中是否存在特定行人的技术。广泛被认为是一个图像检索的子问题。给定一个监控行人图像，检索跨设备下的该行人图像。旨在弥补固定的摄像头的视觉局限，并可与行人检测/行人跟踪技术相结合。
- MTMCT：多目标多摄像头跟踪（Multi-target Multi-camera Tracking），跟踪多个摄像头拍摄的多个人
- 姿态跟踪：追踪人的姿态

4

目标跟踪任务

- 按照任务计算类型又可以分为：
- **在线跟踪**：在线跟踪需要**实时处理**任务，通过过去和现在帧来跟踪未来帧中物体的位置。
- **离线跟踪**：离线跟踪是离线处理任务，可以**通过过去、现在和未来的帧**来推断物体的位置，因此准确率会比在线跟踪高。

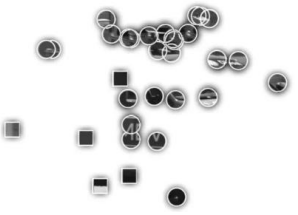


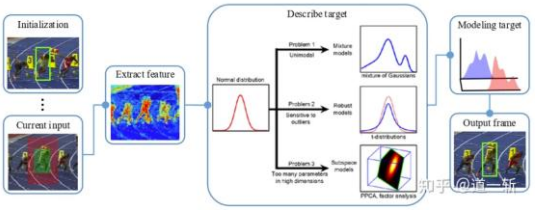
目标跟踪的困难点

- **形态变化**：运动目标发生姿态变化时，会导致它的特征以及外观模型发生改变，容易导致跟踪失败。
- **尺度变化**：当目标尺度缩小时，由于**跟踪框不能自适应跟踪**，会将很多背景信息包含在内，导致目标模型的更新错误:当目标尺度增大时，由于跟踪框不能将目标完全包括在内，跟踪框内目标信息不全，也会导致目标模型的更新错误。
- **遮挡与消失**：目标在运动过程中可能出现被遮挡或者短暂的消失情况。
- **图像模糊**：光照强度变化，目标快速运动，低分辨率等情况会导致图像模型，尤其是在运动目标与背景相似的情况下更为明显。

8.2经典物体跟踪算法

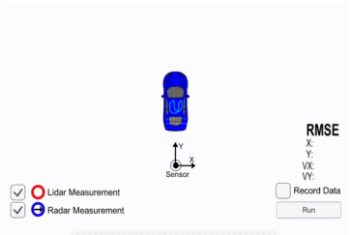
- 基于目标模型建模的方法：通过对目标外观模型进行建模,然后在之后的帧中找到目标.例如：区域匹配、特征点跟踪、基于主动轮廓的跟踪算法、光流法等.最常用的是特征匹配法,首先提取目标特征,然后在后续的帧中找到最相似的特征进行目标定位,常用的特征有:SIFT特征、Harris角点等。

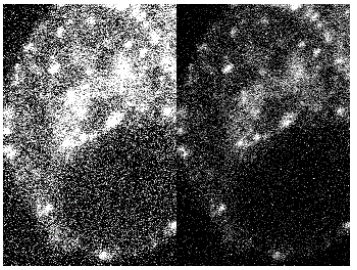




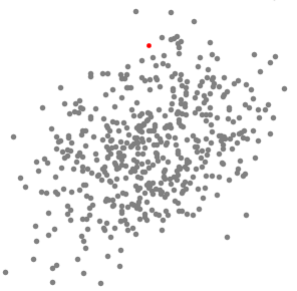
早期的目标跟踪算法

- 基于搜索的方法：将预测算法加入跟踪中,在预测值附近进行目标搜索,减少了搜索的范围.常见一类的预测算法有**Kalman滤波、粒子滤波方法**.另一种减小搜索范围的方法是内核方法:运用最速下降法的原理,向梯度下降方向对目标模板逐步迭代,直到迭代到最优位置.诸如,Meanshift算法。





Meanshift目标跟踪算法



Meanshift算法是一种在一组数据的密度分布中寻找**局部极值**的稳定的方法。就是找局部密度最大的位置，或者说找局部**“重心”**位置。

算法步骤如下：

1）选择**搜索窗口**。
- 窗口的初始位置；
- 窗口的类型（均匀、多项式、指数或者高斯类型）；
- 窗口的形状（对称的或歪斜的，可能旋转的，圆形或**矩形**）；
- 窗口的大小（超出窗口大小则被截去）。

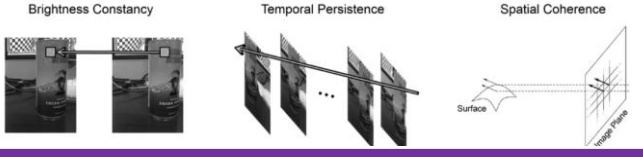
2）计算窗口（可能是带权重的）的重心。

3）将窗口的中心设置在计算出的重心处。

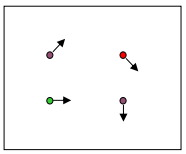
4）返回第2）步，直到满足停止条件（**最大迭代次数和最小移动距离**）。

光流法

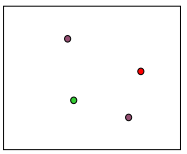
- 光流法(Lucas-Kanade)的概念首先在1950年提出，它是针对外观模型对视频序列中的像素进行操作.通过利用视频序列在相邻帧之间的像素关系，寻找像素的位移变化来判断目标的运动状态，实现对运动目标的跟踪.但是，光流法适用的范围较小，需要满足三种假设：
  - 图像的光照强度保持不变；
  - 空间一致性，即每个像素在不同帧中相邻点的位置不变；
  - 时间连续.
- 光流法适用于目标运动相对于帧率是缓慢的，也就是两帧之间的目标位移不能太大。



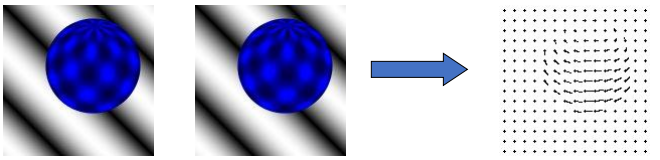
光流法

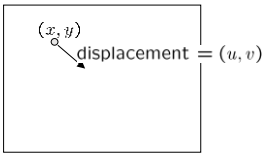


$H(x,y)$

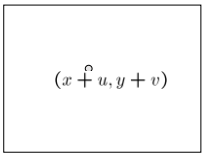


$I(x,y)$





$H(x,y)$



$I(x,y)$

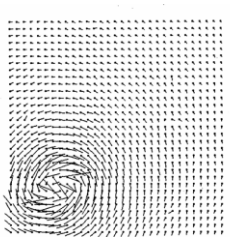
H(x,y) - I(x+u,v+y) = 0

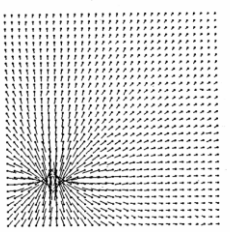
$$I(x+u, y+v) = I(x,y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \text{higher order terms}$$
$$\approx I(x,y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v$$

11

光流法

$$\begin{aligned} 0 &= I(x+u, y+v) - H(x,y) \\ &\approx I(x,y) + I_x u + I_y v - H(x,y) \\ &\approx (I(x,y) - H(x,y)) + I_x u + I_y v \\ &\approx I_t + I_x u + I_y v \\ &\approx I_t + \nabla I \cdot [u \ v] \end{aligned}$$





$$0 = I_t + \nabla I \cdot [\frac{\partial x}{\partial t} \ \frac{\partial y}{\partial t}]$$

$$0 = I_t(\mathbf{p_i}) + \nabla I(\mathbf{p_i}) \cdot [u \ v]$$

采用5x5 窗口, 每个像素25个方程

$$\underset{25\times2}{\begin{bmatrix} I_x(\mathbf{p_1}) & I_y(\mathbf{p_1}) \\ I_x(\mathbf{p_2}) & I_y(\mathbf{p_2}) \\ \vdots & \vdots \\ I_x(\mathbf{p_{25}}) & I_y(\mathbf{p_{25}}) \end{bmatrix}} \underset{2\times1}{\begin{bmatrix} u \\ v \end{bmatrix}} = - \underset{25\times1}{\begin{bmatrix} I_t(\mathbf{p_1}) \\ I_t(\mathbf{p_2}) \\ \vdots \\ I_t(\mathbf{p_{25}}) \end{bmatrix}}$$

A d b

12

Lukas-Kanade Flow

方程数量 > 未知数个数

$$\underset{25\times2}{A}\ \underset{2\times1}{d} = \underset{25\times1}{b} \longrightarrow \text{minimize } \|Ad-b\|^2$$

最小二乘法

$$(A^TA)\ d = A^Tb$$

$$\begin{bmatrix} \sum I_xI_x & \sum I_xI_y \\ \sum I_xI_y & \sum I_yI_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -\begin{bmatrix} \sum I_xI_t \\ \sum I_yI_t \end{bmatrix}$$

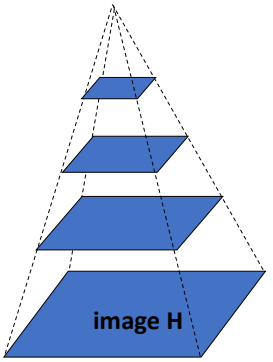
A^TA $\qquad$ A^Tb

$$A^TA = \begin{bmatrix} \sum I_xI_x & \sum I_xI_y \\ \sum I_xI_y & \sum I_yI_y \end{bmatrix} = \sum \begin{bmatrix} I_x \\ I_y \end{bmatrix} [I_x\ I_y] = \sum \nabla I (\nabla I)^T$$

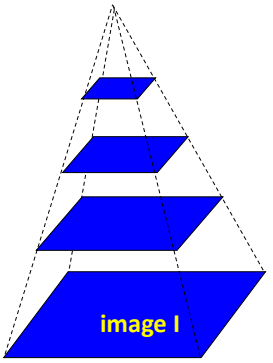
13

光流法

假设运动足够小，但是实际情况运动可能较大。



Gaussian pyramid of image H

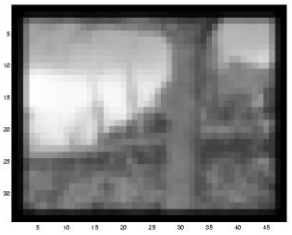
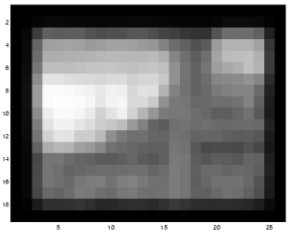


Gaussian pyramid of image I

14

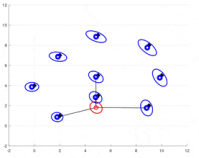
降低分辨率

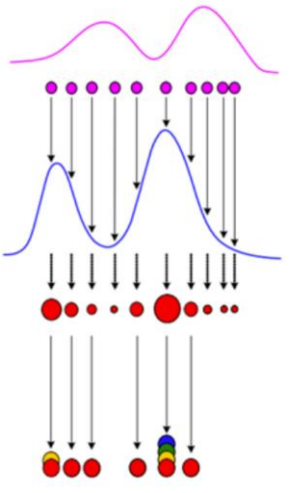


粒子滤波

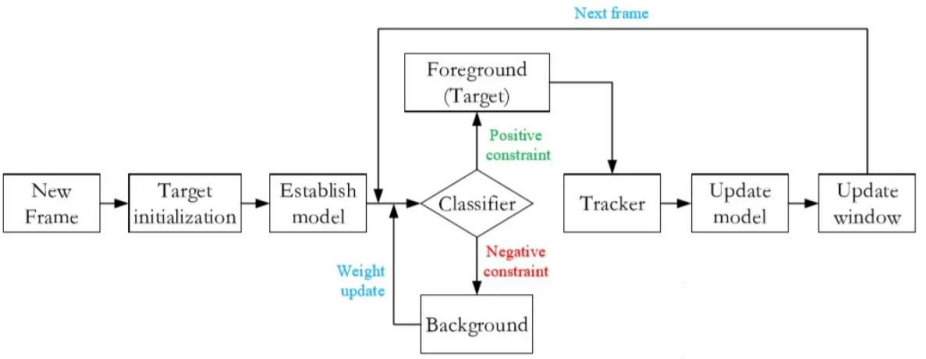
- 粒子滤波（Particle Filter）是一种基于粒子分布统计的方法。以跟踪为例，在目标搜索的过程中，它会按照一定的分布（比如均匀分布或高斯分布）撒一些粒子，统计这些粒子的相似度，确定目标可能的位置。在这些位置上，下一帧加入更多新的粒子，确保在更大概率上跟踪上目标。Kalman Filter 常被用于描述目标的运动模型，它不对目标的特征建模，而是对目标的运动模型进行了建模，常用于估计目标在下一帧的位置。





基于判别的跟踪算法

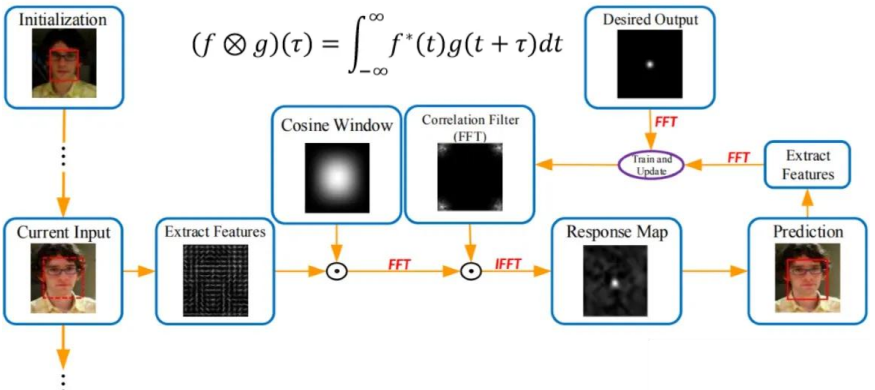
- 判别模型：将目标模型和背景信息同时考虑在内,通过对比目标模型和背景信息的差异,将目标模型提取出来,从而得到当前帧中的目标位置.使用机器学习/深度学习方法训练分类器。



17

基于核相关滤波的跟踪算法

- 将通信领域的相关滤波(衡量两个信号的相似程度)引入到了目标跟踪中。一些基于相关滤波的跟踪算法(CSK、KCF、BACF、SAMF)等, 也随之产生, 速度可以达到数百帧每秒, 可以广泛地应用于实时跟踪系统中.

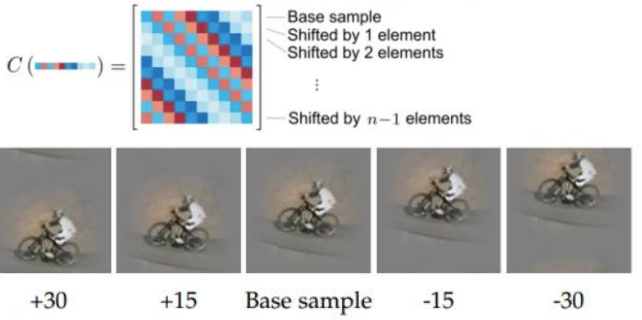


18

基于核相关滤波的跟踪算法

- 相关滤波器（CF），也称为判别相关滤波器（DCF），其原理是两个相关信号f和g的卷积响应大于不相关信号的卷积响应。

$$(f \otimes g)(\tau) = \int_{-\infty}^{\infty} f^*(t)g(t+\tau)dt \quad (f \otimes g)[n] = \sum_{-\infty}^{\infty} f^*[m]g[m+n]$$



+30 +15 Base sample -15 -30

```
function responses = detection(alphaf, x, z, sigma)
  k = dgk(x, z, sigma);
  responses = real(ifft2(alphaf .* fft2(k)));
end

function k = dgk(x1, x2, sigma)
  c = fftshift(ifft2(fft2(x1) .* conj(fft2(x2))));
  d = x1(:)'*x1(:) + x2(:)'*x2(:) - 2*c;
  k = exp(-1 / sigma^2 * abs(d) / numel(x1));
end
```

Henriques J F, Caseiro R, Martins P, et al. ***Exploiting the circulant structure of tracking-by-detection with kernels*** *[C]// ECCV, 2012*

19

讨论：生成式模型 vs. 判别式模型

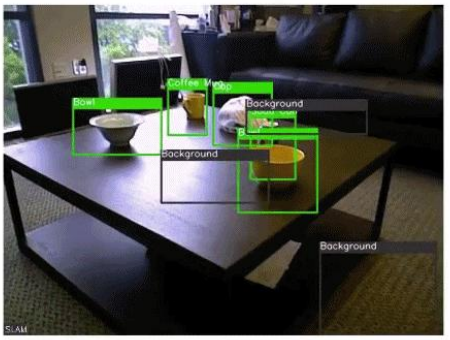
- 比较生成式跟踪器（如均值漂移、卡尔曼滤波）与判别式跟踪器（如相关滤波、Siamese网络）在遮挡、形变、光照变化下的优缺点，并举例说明各自的最佳适用场景。

- **生成式模型**：学习目标自身的表观特征，然后在图像中寻找与该模型最匹配的区域。
  → 代表方法：均值漂移（Mean Shift）、卡尔曼滤波（Kalman Filter）、粒子滤波（Particle Filter）。
- **判别式模型**：将跟踪转化为二分类问题（目标/背景），在线或离线学习目标和背景的决策边界。
  → 代表方法：相关滤波（KCF、DCF）、Siamese网络（SiamFC、SiamRPN）。

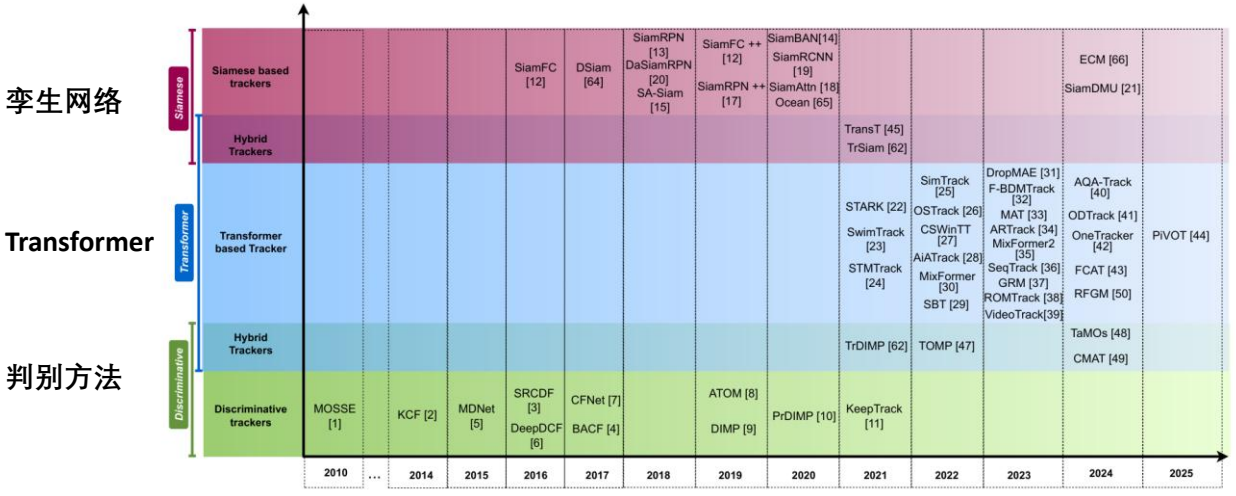
20

8.3物体跟踪的发展

- 用深度学习建立全新的跟踪框架，进行目标跟踪。在大数据背景下，利用深度学习训练网络模型，得到的卷积特征输出表达能力更强，但同时也带来了计算量的增加。

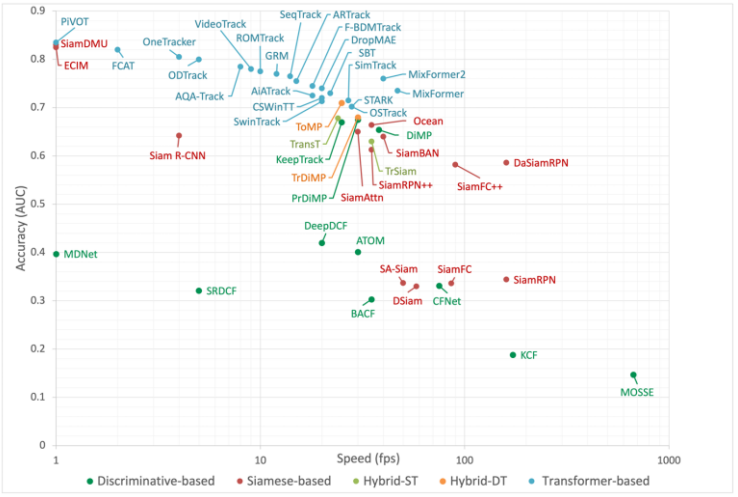


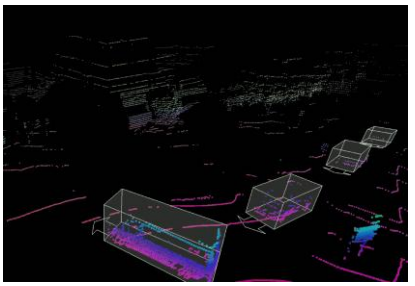
物体跟踪的发展



A Deep Dive into Generic Object Tracking: A Survey

物体跟踪的发展

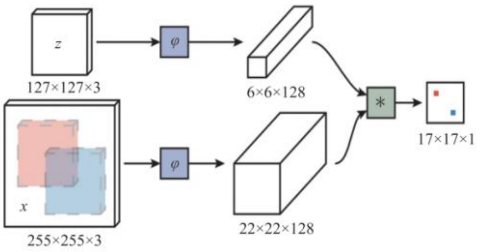




A Deep Dive into Generic Object Tracking: A Survey

SiamFC孪生网络跟踪模型

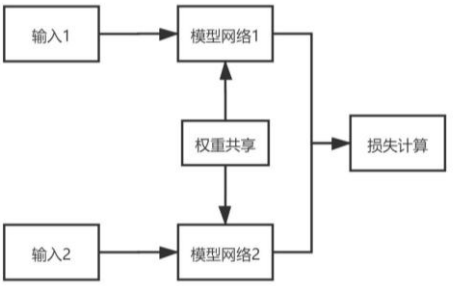
- 孪生网络结构通常具有俩个输入分支：
  模板分支和搜索区域分支。
  ①模板是在初始时刻确定的跟踪对象。
  ②搜索区域分支是后续时刻对应的图像帧
- 孪生网络的目标是在后续图像帧中
  通过相似度计算确定与模板最相似的候选区域。



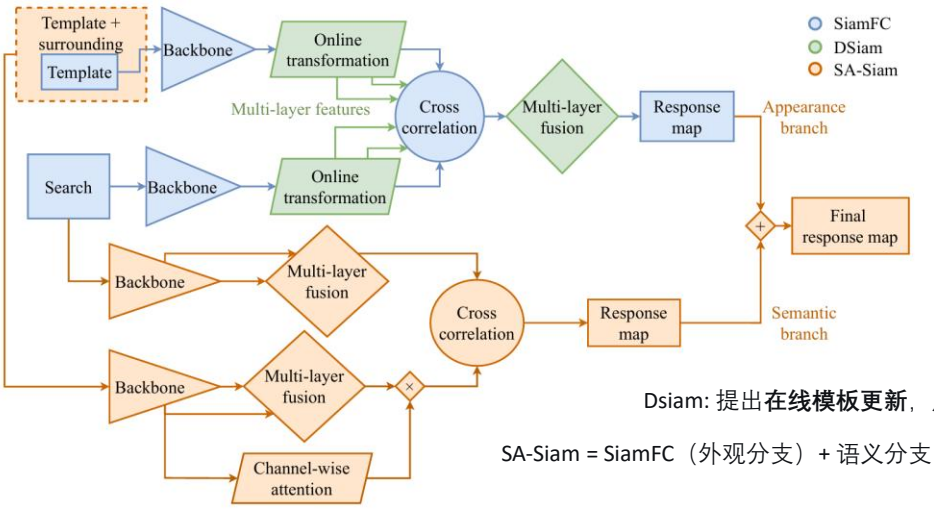
- 特点：（1）基于参数共享的原理，网络整体参数规模得到优化。（2）该结构具有相同的参数与权重信息，有助于在语义特征信息生成阶段实现相同的映射模式，便于评估双分支输入的相似性

孪生网络

- 在计算机视觉领域的应用中，通常利用卷积神经网络提取图像特征，采用相似性度量函数进行双分支特征信息的相似性计算
- 基本的计算方法包括欧式距离法和余弦距离法，在后续的发展过程中也提出互相关操作的概念，即将一分支的输出结果作为卷积核在另一分支输出结果上进行卷积操作从而生成对应的互相关响应图



孪生网络



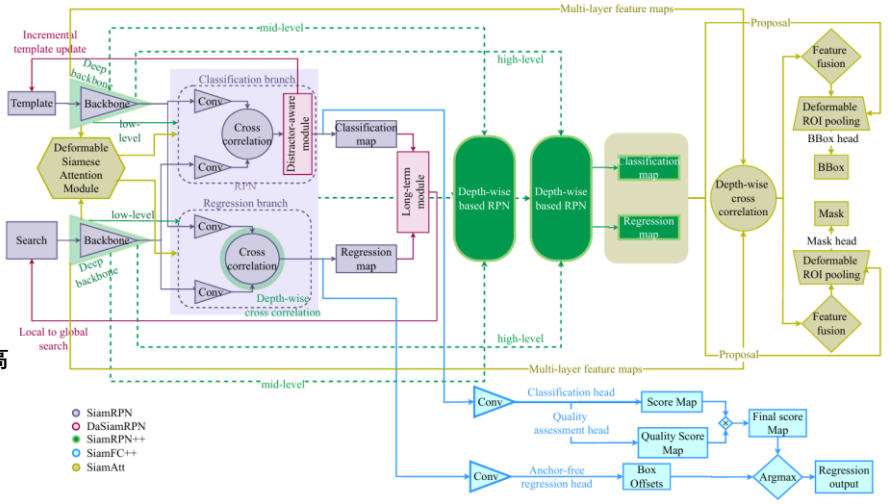
Dsiam: 提出**在线模板更新**，应对目标形变

SA-Siam = SiamFC（外观分支）+ 语义分支 + 注意力机制

孪生网络

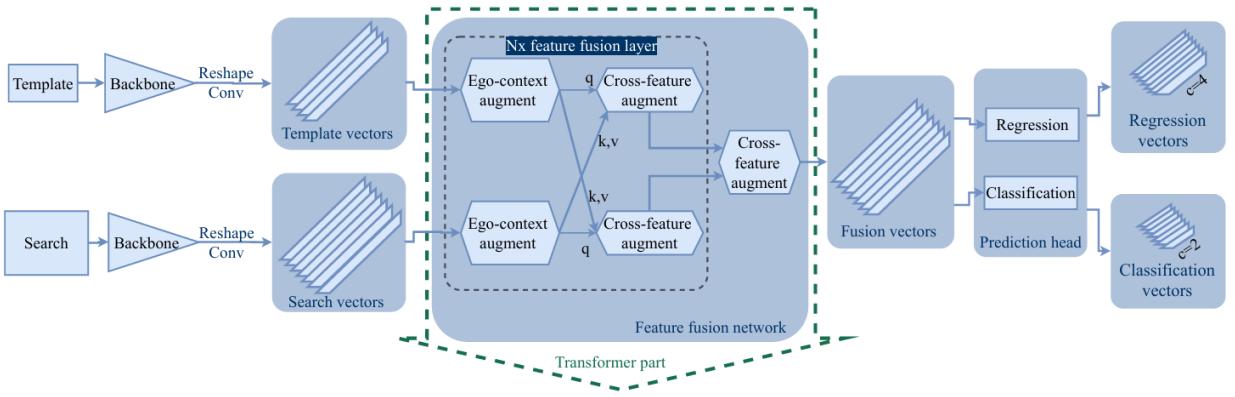
SiamFC（基础孪生网络）与Faster R-CNN中的RPN（区域建议网络）的创造性结合

证明了“检测”与“跟踪”在本质上高度统一——都是“定位+分类”



Transformer网络

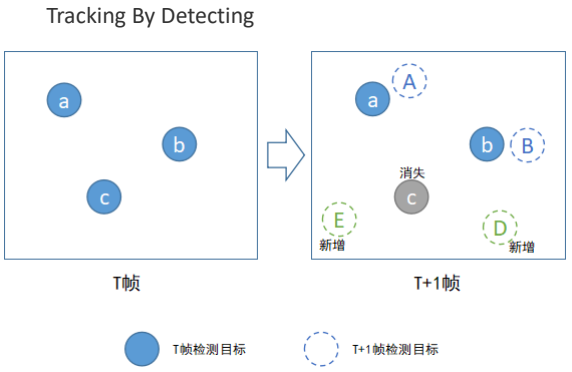
如何让跟踪器像人类一样，理解目标与周围环境全域的、长距离的依赖关系，而不仅仅是局部的相似性.

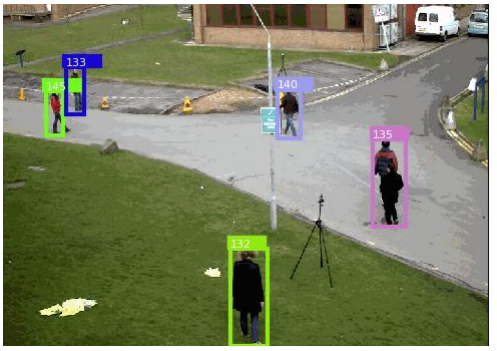


B. Cheng, X. Wang, W. Zhang, C. Zhang, H. Li, J. Sun, P. Luo, Transtrack: Multiple object tracking with transformer, arXiv preprint arXiv:2012.15460 (2020).

多目标跟踪

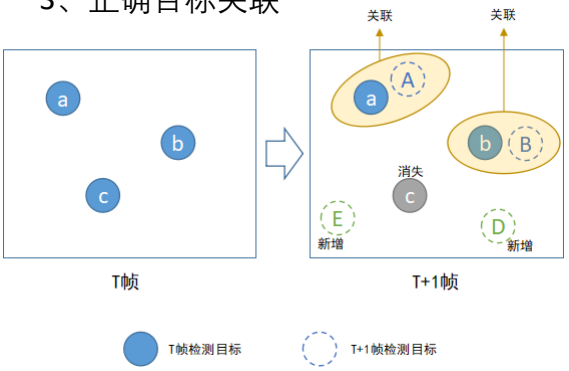
- 跟踪的本质是关联视频前后帧中的同一物体（目标），并赋予唯一TrackID。





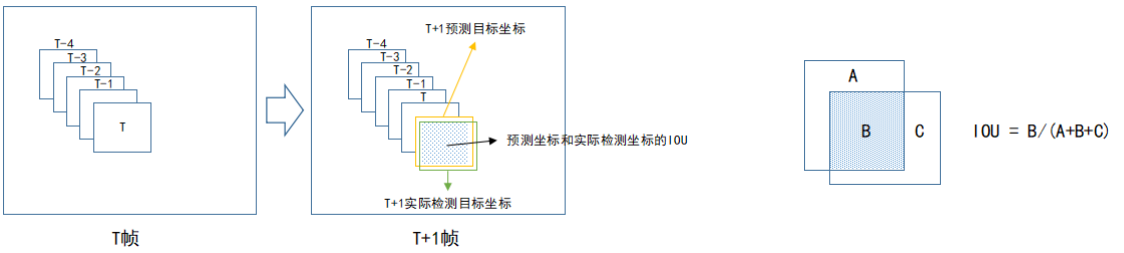
目标关联

1、如何处理中途出现的新目标
2、如何处理中途消失的目标
3、正确目标关联



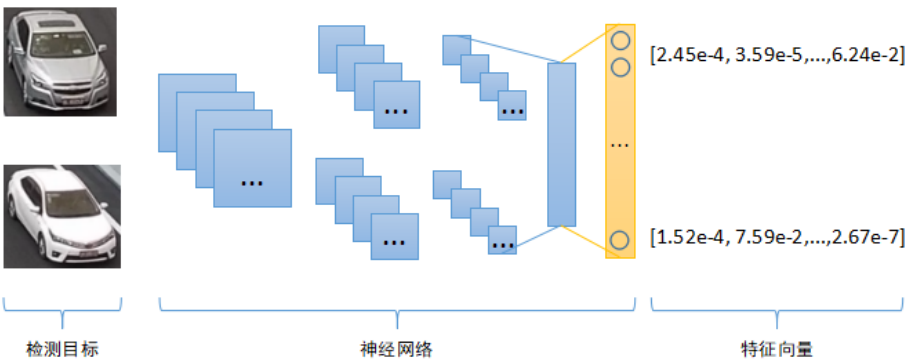
基于坐标的目标关联

因为目标密集，相邻目标的坐标（left、top、width、height）重合度比较高。

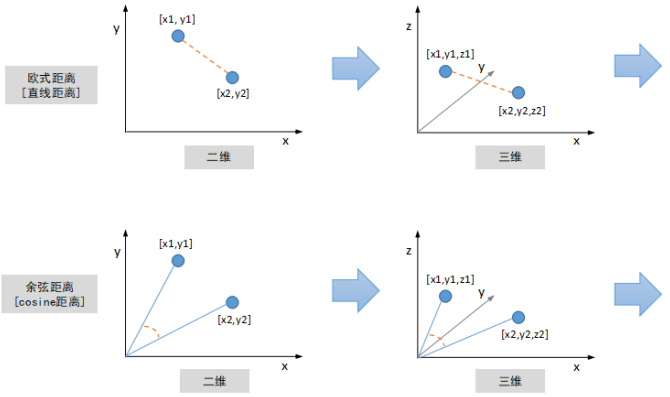


基于特征的目标关联

前后两帧中挨得近的物体且外观长得比较像的物体为同一目标。

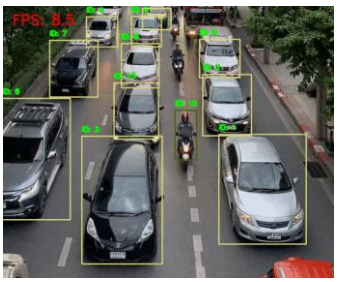


计算两个图像特征的相似度

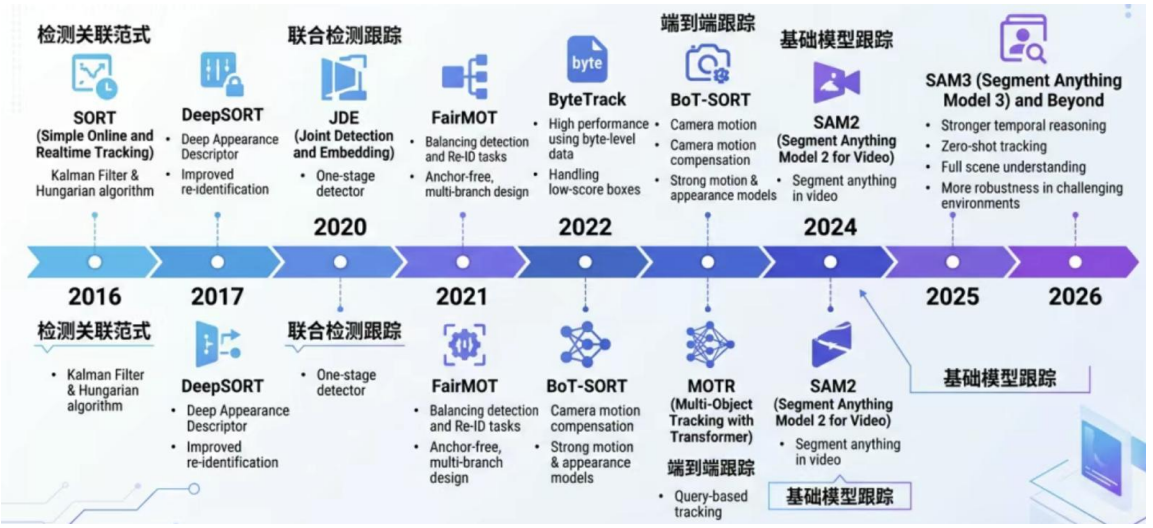


视觉目标跟踪与重识别统一技术发展

- 2016年DeepSORT（Simple Online and Realtime Tracking）首次将深度外观特征引入跟踪关联；
- 2020年FairMOT实现了检测与ReID的公平联合学习；
- 2021年ByteTrack用极简策略刷新了MOT基准；
- 2022年BoT-SORT融合运动与外观达到新高度；
- 2023年MOTR系列推动端到端Transformer跟踪；
- 2024年SAM 2将分割与视频跟踪统一；
- 2025年SAM 3引入概念级跟踪；
- 2026年多模态大模型开始重塑整个跟踪范式。



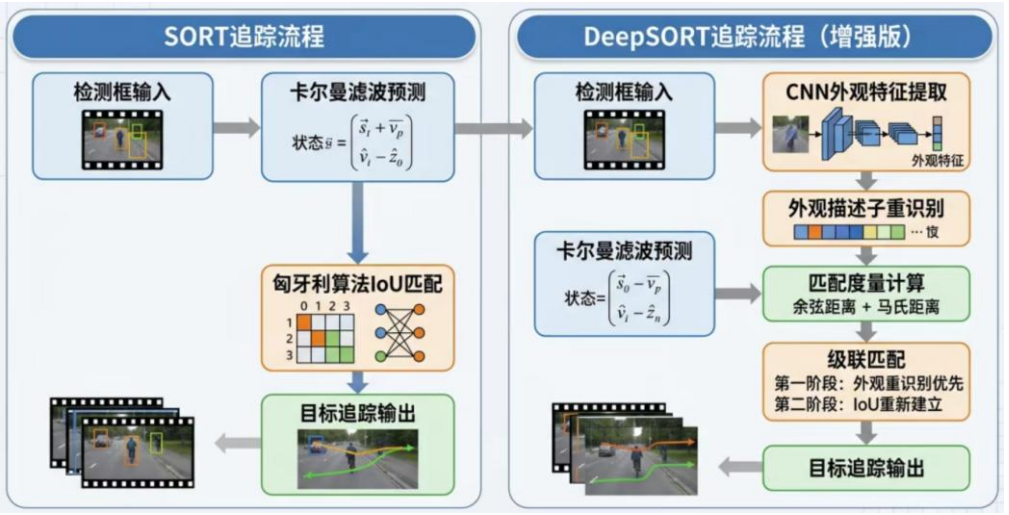
视觉目标跟踪与重识别统一技术发展



SORT与DeepSORT：经典跟踪范式的奠基

- 2016年，SORT（Simple Online and Realtime Tracking），确立了"检测-关联"（Tracking-by-Detection）的经典范式。
  - 用卡尔曼滤波预测目标在下一帧的位置，用匈牙利算法将预测框与检测框进行最优匹配，匹配代价仅使用IoU（交并比）距离。
- SORT的弱点:身份切换（ID Switch）频繁——当目标被遮挡或检测丢失时，重新出现后会被分配新的ID。
- 2017年，DeepSORT引入深度外观特征。训练轻量级CNN，为每个检测框提取128维外观描述子，并维护每个轨迹最近100帧的外观特征库。关联时同时使用马氏距离(运动信息)和余弦距离(外观信息)。
- DeepSORT将SORT的ID Switch从1423降低到781（MOT16），证明了外观特征对跟踪稳定性的关键作用。

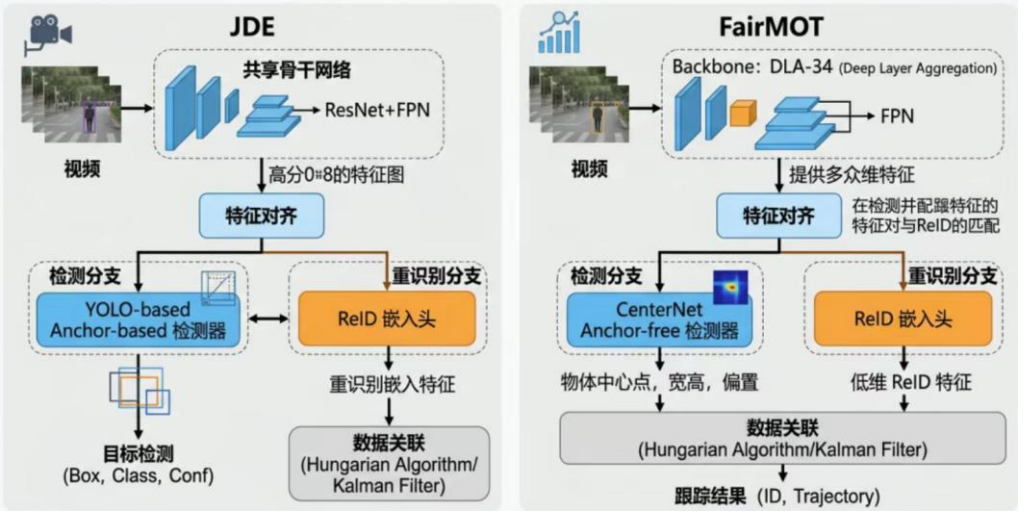
SORT与DeepSORT架构对比



JDE与FairMOT：联合检测与跟踪的统一

- DeepSORT（先检测再提取外观特征）在速度上存在瓶颈。
- 2020年，JDE（Joint Detection and Embedding），首次在单个网络中同时输出检测框和外观嵌入。JDE基于YOLOv3骨干网络，在检测头旁边并行添加了一个嵌入头，共享特征提取的计算。速度达到22.2 FPS，比DeepSORT快近一倍。
- 问题：检测和ReID之间存在竞争。检测需要学习类别无关的特征来定位目标，而ReID需要学习实例级的判别特征来区分不同个体。
- 2021年，提出FairMOT（A Simple Baseline for Multi-Object Tracking with Fair motivated Design）。(1)使用CenterNet作为anchor-free检测器，避免anchor导致的特征对齐问题；(2)检测和ReID分支使用相同分辨率的特征图，确保公平性；(3)采用多层特征聚合（DLA-34骨干），为两个任务提供更丰富的特征。

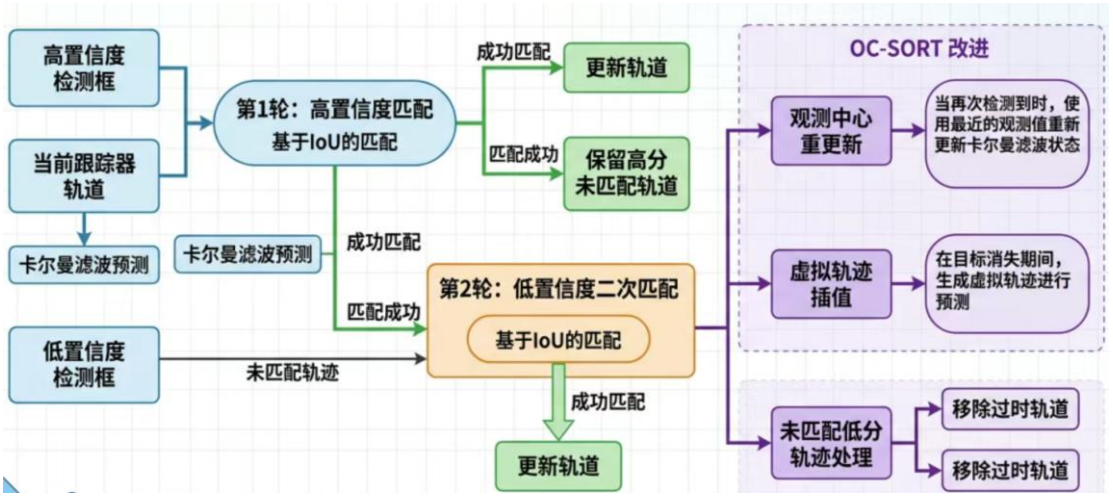
JDE与FairMOT联合检测跟踪框架对比



ByteTrack与OC-SORT：简洁高效的关联策

- 2022年ByteTrack：不要丢弃任何检测框，包括低置信度的。BYTE关联算法——第一轮用高置信度检测框与现有轨迹匹配（基于IoU），第二轮用剩余的低置信度检测框与未匹配的轨迹。反思：在检测器足够强大的情况下，复杂的外观模型是否真的必要？
- 2022，OC-SORT（Observation-Centric SORT），引入三个关键技术：(1)观测中心的重更新（OCR），在目标重新出现时用观测值修正累积的预测误差；(2)观测中心的动量（OCM），利用目标的运动方向一致性辅助关联；(3)虚拟轨迹恢复，为丢失的轨迹生成虚拟观测以维持运动模型。在遮挡严重的场景中表现尤为突出。

ByteTrack与OC-SORT关联策略



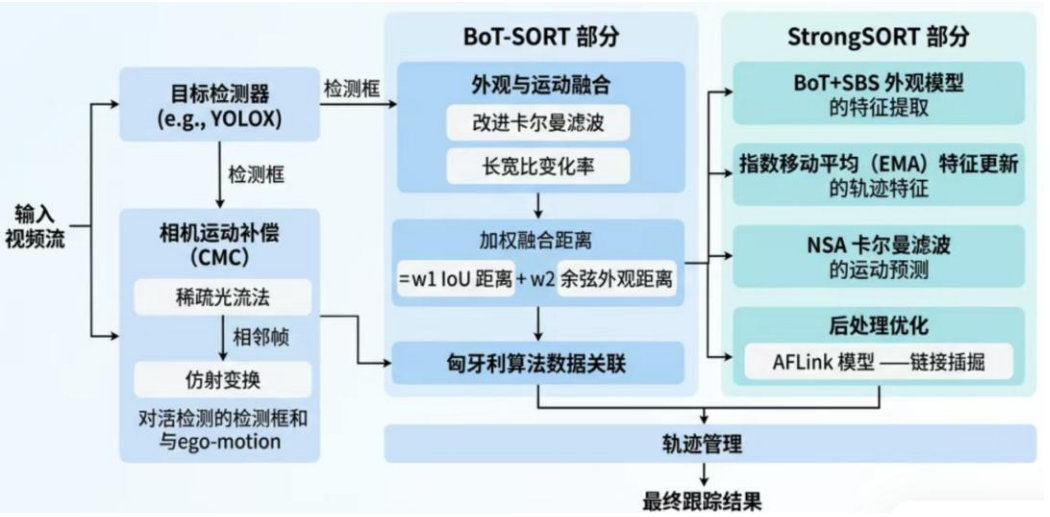
41

BoT-SORT与StrongSORT：运动与外观的深度融合

- 2022年，BoT-SORT（Robust Associations Multi-Pedestrian Tracking），将运动和外观信息进行了更精细的融合。创新包括：(1)相机运动补偿（CMC），使用稀疏光流估计相机运动并补偿卡尔曼滤波的预测；(2)改进的卡尔曼滤波状态向量，加入宽高比变化率；(3)IoU与外观余弦距离的加权融合作为匹配代价。
- 2023，Deep OC-SORT加入深度外观特征和动态外观模型更新
- 2024，Hybrid-SORT提出了混合关联策略，根据场景复杂度自适应地调整运动和外观线索的权重。
- 2024，LITE(**L**ightweight **I**ntegrated **T**racking-feature ex**T**raction)则提出了轻量级集成跟踪特征提取范式，消除了独立ReID模型的推理开销，在保持精度的同时大幅提升速度。

42

BoT-SORT 与 StrongSORT架构比较:外观与运动融合



行人重识别(Person re-identification)

- 行人重识别Person ReID

1)利用计算机视觉技术判断图像或者视频序列中是否存在特定行人的技术；

2)行人重识别是指在已有的可能来源与非重叠摄像机视域的视频序列中识别出目标行人。

智能安防

智能商业—无人超市

存在的问题

(a)低分辨率 (b)遮挡 (c)视角、姿势变化 (d)光照变化 (e)视觉模糊性

存在着无正脸照、配饰、服装搭配、穿衣风格以及由于不同的数据集之间存在着域的偏移问题，使得在源数据集下训练的模型在目标数据集下很难取得好的性能，泛化性能不强。

类内差异增大，类间差异减少

思考：如何解决？

1. 能不能用人脸识别做重识别？
2. 有些人靠衣服的颜色就可以判断出来了，还需要行人重识别么？
3. 使用图像检索的指标来衡量行人重识别的结果是否合适？

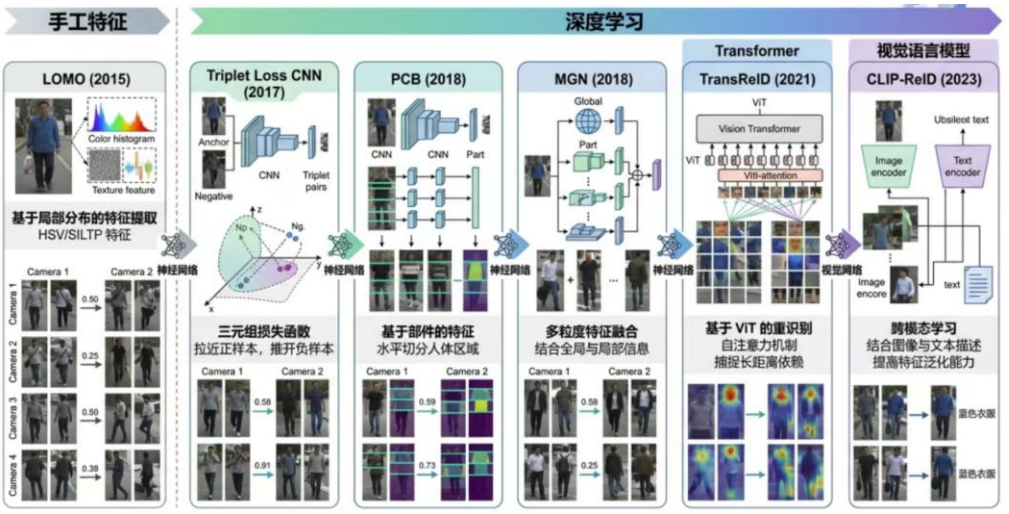
(a)低分辨率 (b)遮挡 (c)视角、姿势变化 (d)光照变化 (e)视觉模糊性

行人重识别ReID：从手工特征到深度学习

- 手工特征：LOMO（Local Maximal Occurrence, 2015）提取多尺度局部最大出现特征，结合XQDA度量学习；ELF（Ensemble of Localized Features）使用颜色直方图和纹理特征的组合。这些方法在小规模数据集上有效，但泛化能力有限。
- 2017年，深度ReID框架，使用ResNet-50提取全局特征，通过三元组损失拉近同一身份、推远不同身份的特征距离。PCB（Part-based Convolutional Baseline, 2018）将特征图水平切分为多个条带，分别提取局部特征再拼接。MGN（Multiple Granularity Network, 2018）同时学习全局和多粒度局部特征。
- 2020年后，Transformer架构进入ReID领域。TransReID（2021）首次将ViT应用于ReID，通过侧信息嵌入（SIE）编码相机和视角信息。CLIP-ReID（2023）利用CLIP的视觉语言对齐能力。SOLIDER（2023）提出语义引导的自监督预训练。2024年，LUPerson-T基于大规模无标注行人数据的Transformer预训练。

47

行人重识别技术演进



48

行人重识别目前所采用的方法

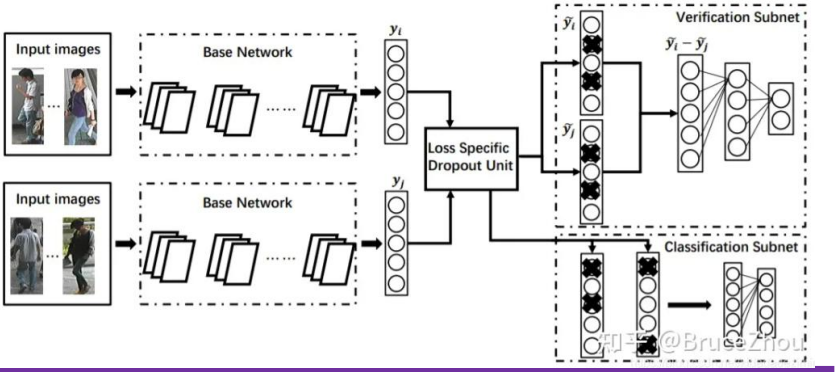
1. 基于表征学习的ReID方法:

- 分类(Classification/Identification)问题或者验证(Verification)问题：

（1）分类问题是指利用行人的ID或者属性等作为训练标签来训练模型；

（2）验证问题是指输入一对（两张）行人图片，让网络来学习这两张图片是否属于同一个行人。

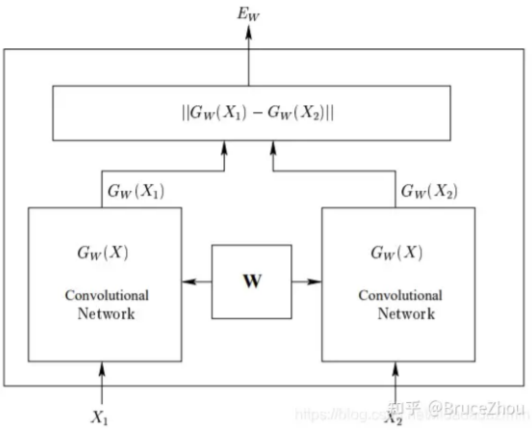
作者	年份	图像特征	时间信息
D. Gray等[4]	2008	颜色、纹理	无
A.Krizhevsky等[38]	2012	CNN颜色、形状	无
Zhao R等[7]	2013	颜色	无
B. Ma等[47]	2014	外观、纹理 生物激励特征	无
Xiang Li等[48]	2015	颜色、形状、纹理	无
Gou M等[14]	2016	颜色、局部、纹理、轨迹	有
T. Matsukawa等[49]	2016	局部、形状、颜色、梯度	无
McLaughlin等[10]	2016	颜色、轨迹、CNN	有



基于度量学习的ReID方法

- 度量学习旨在通过网络学习出两张图片的相似度。在行人重识别问题上，具体为同一行人的不同图片相似度大于不同行人的不同图片。最后网络的损失函数使得相同行人图片（正样本对）的距离尽可能小，不同行人图片（负样本对）的距离尽可能大。
- 常用的度量学习损失方法有：
  - **对比损失(Contrastive loss)**
  - **三元组损失(Triplet loss)**
  - **四元组损失(Quadruplet loss)**
  - **难样本采样三元组损失(Triplet hard loss with batch hard mining, TriHard loss)**。

对比损失(Contrastive loss)

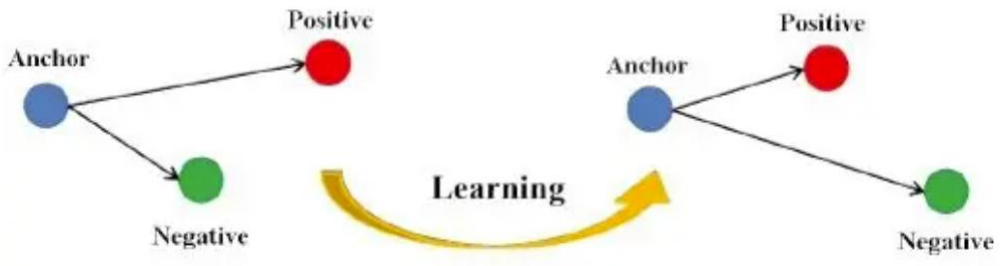


对比损失用于训练孪生网络(Siamese network),其结构图如上图6所示。孪生网络的输入为一对（两张）图片，这两张图片可以为同一行人，也可以为不同行人。每一对训练图片都有一个标签y，其中y = 1，表示两张图片属于同一个行人（正样本对），反之y = 0表示它们属于不同行人（负样本对）。

51

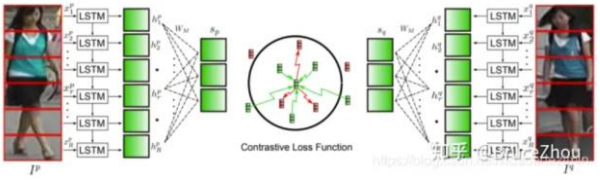
三元组损失(Triplet loss)

- 三元组可以拉近正样本对之间的距离，推开负样本对之间的距离，最后使得相同ID的行人图片在特征空间里形成聚类，达到行人重识别的目的。

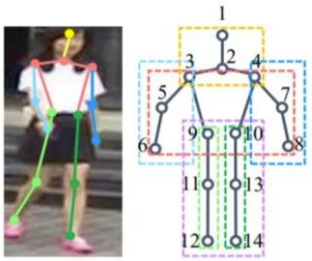


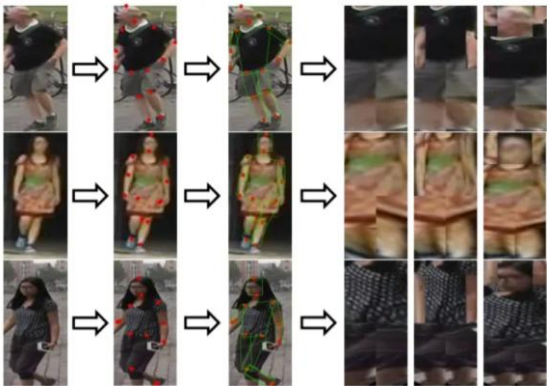
52

基于局部特征的ReID方法



切块是一种很常见的提取局部特征方式。





一个行人通常被分为14个关键点，这14个关键点把人体结果分为若干个区域。为了提取不同尺度上的局部特征，作者设定了三个不同的PoseBox组合。

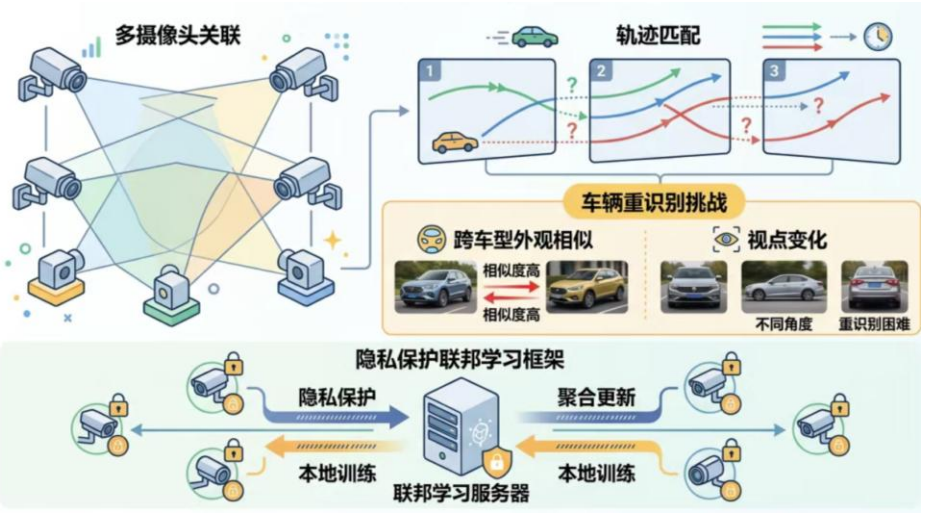
53

跨摄像头跟踪与车辆重识别

- CityFlow（2019）提供了首个大规模城市交通场景的MTMCT基准，包含40个摄像头、超过200个车辆身份。
- 同一型号的车辆外观高度相似，不同视角下同一车辆的外观变化巨大。VeRi-776（2016）和VehicleID（2016）是两个经典基准数据集。
- TransReID-SSL（2022）将自监督预训练引入车辆ReID，利用大量无标注车辆图像学习通用表示。
- CLIP-VReID（2024）借助CLIP的跨模态能力，通过车辆属性文本描述增强视觉特征的判别力。
- 2024-2025年，跨摄像头跟踪开始与大模型融合。多模态融合方法结合视觉特征、车牌识别、时空约束和交通拓扑信息，实现更鲁棒的跨摄像头匹配。

54

跨摄像头跟踪与车辆重识别



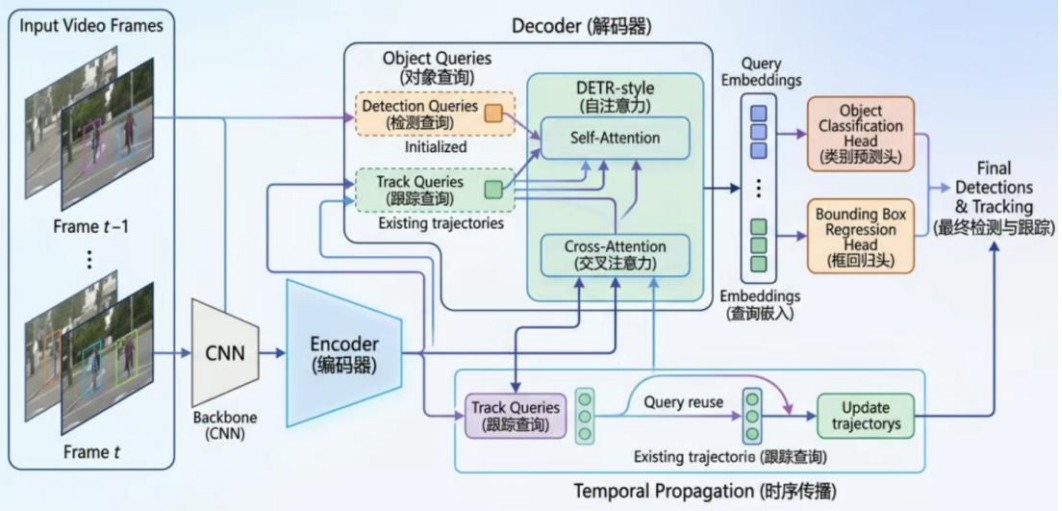
55

Transformer跟踪器： 从TransTrack到MOTRv3

- 2021年，**TransTrack**首次将Transformer应用于MOT，使用两组query——检测query负责发现新目标，跟踪query负责关联已有轨迹。
- **TrackFormer**（2022, CVPR），在DETR框架中引入track query，每个track query编码一个被跟踪目标的时空信息，通过自注意力与其他query交互，通过交叉注意力从当前帧提取特征。
- MOTR（2022, ECCV）由旷视提出，引入了track query的时序聚合机制。
- MOTRv2（2023）"检测器+Transformer关联"的混合架构。
- MOTRv3（2024）可学习的query交互模块和改进的训练策略
- CO-MOT（2024）提出协同查询机制，让检测query和跟踪query之间进行信息交换。

56

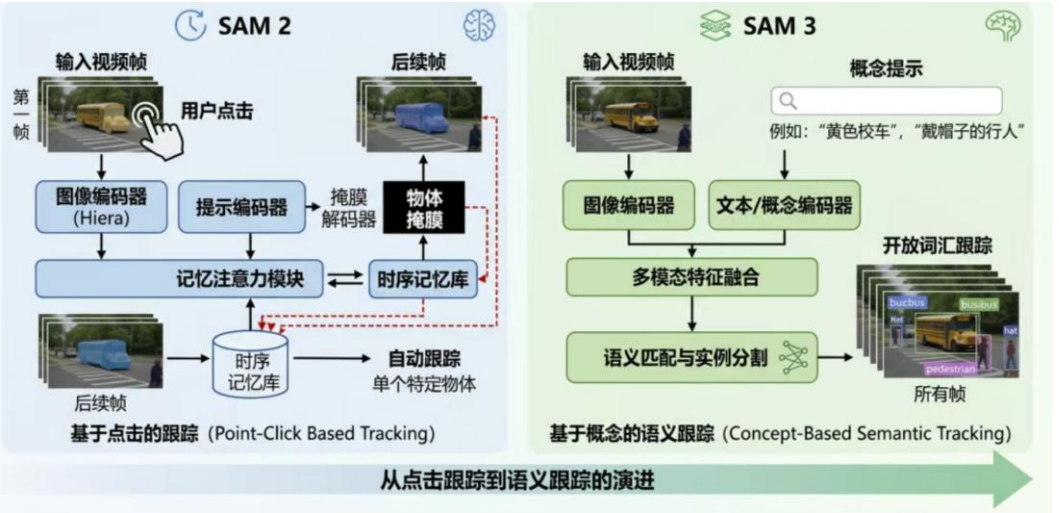
Transformer端到端跟踪(TransTrack, TrackFormer, MOTR)



SAM 2视频跟踪与SAM 3概念跟踪

- 2024，Meta发布SAM 2（**Segment Anything Model 2**），将图像分割基础模型扩展到视频领域，实现了分割与跟踪的统一。
- SAM 2在SA-V（Segment Anything Video）数据集上训练，该数据集包含50.9K视频和642.6K掩码序列，是当时最大的视频分割数据集。
- 2025年底，**Meta发布SAM 3**，引入"概念提示"。与SAM 2需要在特定帧上点击指定目标不同，SAM 3允许用户用自然语言描述目标类别（如"黄色校车"、"行走的行人"），模型自动在视频中发现、分割和跟踪所有匹配实例。
- 视频跟踪**从"指哪跟哪"进化到"说什么跟什么"**，为开放世界视频理解开辟了全新范式。

SAM 2与SAM 3视频跟踪架构演进



开放世界跟踪与TAO

- 2020年，**TAO（Tracking Any Object）**数据集，包含2907个视频、833个类别、17287条轨迹，首次将MOT扩展到大规模开放类别场景。TAO的评估协议要求跟踪器能够处理长尾分布的类别，包括训练时从未见过的目标。
- OVTrack（2023）利用CLIP的开放词汇能力，将文本描述作为类别提示来指导跟踪。
- MASA（Matching Anything by Segmenting Anything, 2024）基于SAM的分割能力，先分割出所有可能的目标，再通过学习到的匹配模型进行帧间关联。
- Grounding DINO + SAM 2的组合方案在**2024年成为开放世界跟踪的强基线**：Grounding DINO负责根据文本描述检测目标，SAM 2负责在视频中跟踪和分割。
- UniTrack（2024）提出统一的跟踪框架，用单一模型同时处理SOT、MOT和VOS任务。
- 2025年，随着SAM 3的发布，**概念级跟踪**成为可能。

开放世界跟踪技术演进与主要方法

61

8.4常用数据集

Dataset	# V	# F	Multi-view	GT	Indoor	Outdoor
MOT16	14	11K	×	✓	✓	✓
KITTI	50	–	✓	✓	×	✓
PETS 2016	13	–	✓	✓	×	✓
PETS 2009	3	–	✓	✓	×	✓
CAVIAR	54	–	✓	✓	✓	×
TUD Stadtmitte	1	179	×	✓	×	✓
TUD Campus	1	71	×	✓	×	✓
TUD Crossing	1	201	×	✓	×	✓
Caltech Pedestrian	137	250K	×	✓	×	✓
UBC Hockey	1	≈100	×	×	×	✓
ETH Pedestrian	8	4K	✓	✓	×	✓
ETHZ Central	3	13K	×	✓	×	✓
Town Centre	1	4.5K	×	✓	×	✓
Zara	4	–	×	×	×	✓
UCSD	98	–	×	×	×	✓
UCF Marathon	3	1.3K	×	✓	×	✓
ParkingLOT	3	2.7K	×	✓	×	✓

65

目标跟踪任务中常用数据集对比

数据集	视频数量	平均帧长	帧率	目标类别	属性	训练/测试集划分	总时长
OTB2013[27]	51	578	30	10	11	无	16.4 分钟
OTB2015[17]	100	590	30	16	11	无	32.8 分钟
TColor-128[28]	128	429	30	27	11	无	30.7 分钟
VOT[6]	60	不定	30	不定	不定	无	不定
NFS[29]	100	3830	240	17	9	无	26.6 分钟
UAV123[30]	123	915	30	9	12	无	62.5 分钟
GOT-10K[31]	10K	150	10	563	6	有	–
TrackingNet[7]	31K	451	30	27	–	有	–
OxUvA[32]	366	–	–	22	–	有	14.3 小时
LaSOT[8]	1400	2,506	30	70	14	有	32.5 小时

66

8.5评价标准

- MOTA (Multiple Object Tracking Accuracy):

  MOTA衡量的是跟踪算法在多目标跟踪任务中的准确性。计算MOTA时，主要考虑的因素是误报(false positives, FP)、漏报(false negatives, FN)和身份切换次数(ID switches, IDsw)。

- MOTA = 1 - (FN + FP + IDS) / GT
  FN（False Negatives）：表示实际存在但未检测到的**目标数**。
  FP（False Positives）：表示实际不存在但错误检测为目标的目标数。
  IDS（Identity Switches）：表示跟踪过程中**目标身份的错误切换次数**。GT（Ground Truth）：表示实际存在的目标总数。

- MOTA的取值范围为[-∞, 1]，值越接近1表示跟踪性能越好。

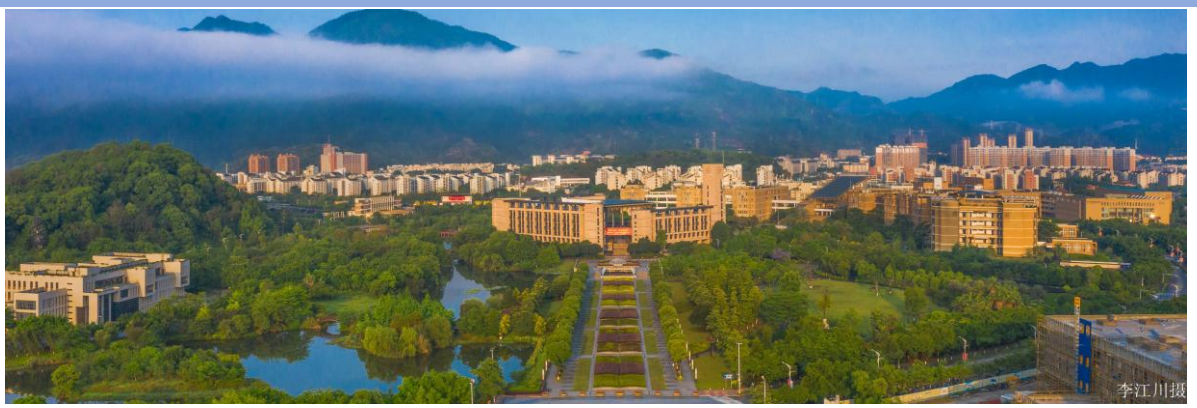
67

8.5评价标准

- IDF1 (ID F1 Score):
  IDF1用于衡量跟踪算法在识别目标身份方面的性能。计算IDF1时，主要考虑的是真正例(true positives, TP)、假正例(false positives, FP)和假反例(false negatives, FN)。
- IDF1 = 2 * IDTP / (2 * IDTP + IDFP + IDFN)
  IDTP（Identified True Positives）：表示正确识别的目标检测数。
  IDFP（Identified False Positives）：表示错误识别的目标检测数。
  IDFN（Identified False Negatives）：表示未正确识别的目标检测数。
- IDF1的取值范围为[0, 1]，值越接近1表示目标身份识别性能越好。

8.5评价标准

- HOTA (Higher Order Tracking Accuracy):
  HOTA是一个更综合的指标，考虑了跟踪的准确性（A）和身份识别（ID）两个方面。
  HOTA的计算公式为：$HOTA=((1-A)\times(1-ID))0.5$
  A 是跟踪的准确性，计算公式与MOTA类似，但是不考虑身份切换。
  ID 是身份识别的准确性，与IDF1相关。



Thank you!