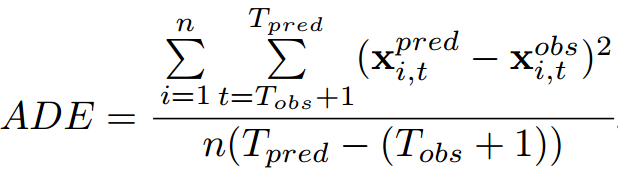
数据集整理

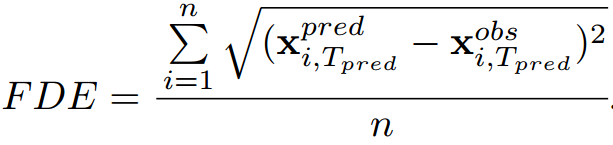
## 轨迹预测

### 评价指标

1. Average displacement error (ADE)



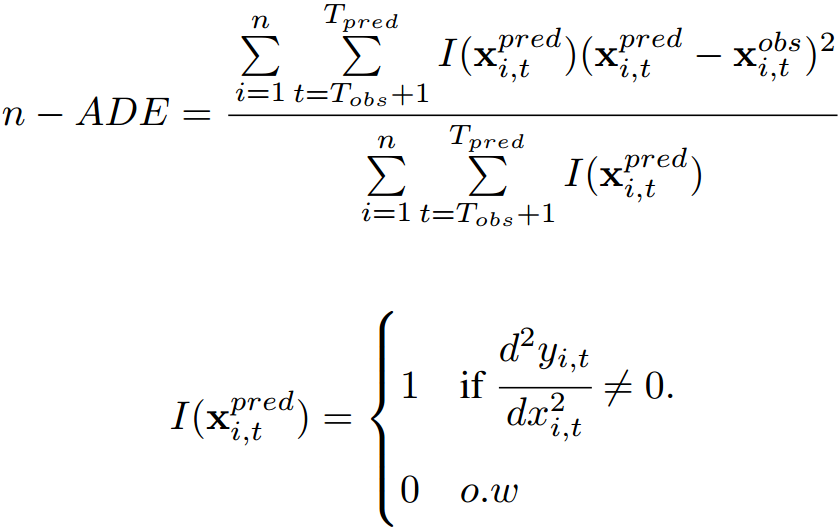
2. Final displacement error (FDE)



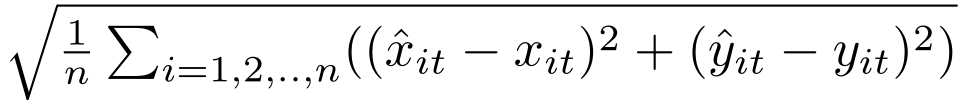
3. Average non-linear displacement error (n-ADE)

Non-linear区域的ADE

The average displacement error for the non-linear regions of the trajectory,



4.Root Mean Square Error in meters



5.不同类别的ADE加权求和

类别：行人+车辆+骑自行车的人

Apollo中这样使用

WSADE/WSFDE: weighted sum of ADE/FDE among different agents types

6.多个轨迹中最小ADE和最小FDE

minADE/minFDE: is the minimum ADE/FDE among multiple predictions (up to K=6)

7.DAC

可行驶区域之内的比重

DAC: is the ratio of the predicted positions inside the drivable area

8.碰撞检测

Calculates the percentage of collisions of primary pedestrian with neighbouring pedestrians in the scene. The model prediction of neighbouring pedestrians is used to check the occurrence of collisions. Lower is better.

9.多模态的测试指标

K个output中最小ADE

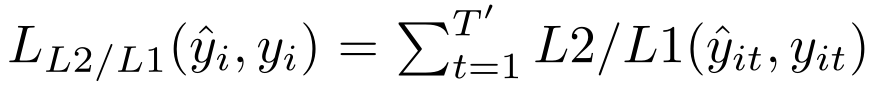
K个output中最小FDE

N个output，计算平均negative log-likelihood of groundtruth，越高越好

10.negative log-likelihoods

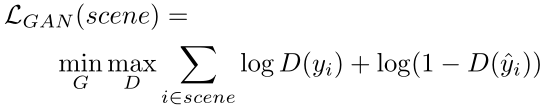
### loss

**L1/L2 loss**

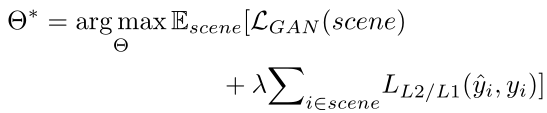


**Variety loss**

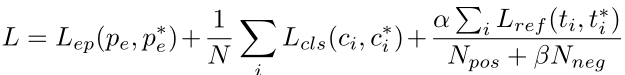
**GAN loss**



**组合loss**



**Multi-task loss**



终点回归损失+分类回归损失+refine损失

### 部分策略

帧率0.4，观察8帧，预测8/12帧；

Leave-one-out，留一个测试，在其余的数据集上训练；

### ETH和UCY

ETH和UCY中，涵盖了具有挑战性的群体行为，例如，两人走在一起，群体间彼此交叉以及在某些场景中形成和分散群体的过程。

包含行人碰撞，碰撞避免，群组活动。

Top-view的图像，摄像机静止。

世界坐标。



#### ETH

ETH论文

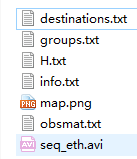
You’ll never walk alone: Modeling social behavior for multi-target tracking.

（有论文引用：Improving data association by joint modeling of pedestrian trajectories and groupings.）

两个场景ETH和Hotel

每个场景750个不同的行人

##### 第一个：seq\_eth.avi



在ETH main building的上方采集得到

8分39秒

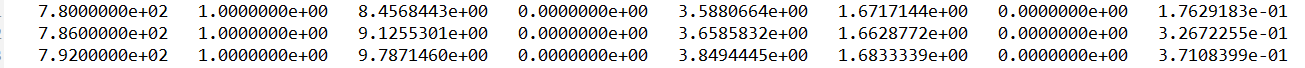
分辨率：640\*480

标注是2.5fps，即0.4s

尽量避免标注在场景边缘的人，因为其轨迹受到视野外行人/障碍物的影响

共8908帧

**1.obsmat.txt**



每行：帧号，行人ID，位置x，位置y，位置z，速度x，速度y，速度z

位置z和速度z实际未使用，值都是0

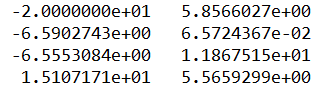
**2.H.txt**

位置和速度以米为单位，两者都是通过H.txt中的单应性矩阵计算得到

**3.groups.txt**

每行包含一组ID的列表，表示group

**4.destination.txt**



一个简化的假设，是假定的目的地，共4个目的地

**5.map.png**

存储了障碍物的信息，是图像坐标，需要根据H.txt转换为真实世界坐标

##### 第二个：seq\_hotel.avi

720\*576

12分54秒

视频是25fps，按照2.5fps标注，即0.4秒标注一帧

共6544帧

24个目的地

#### UCY

UCY论文

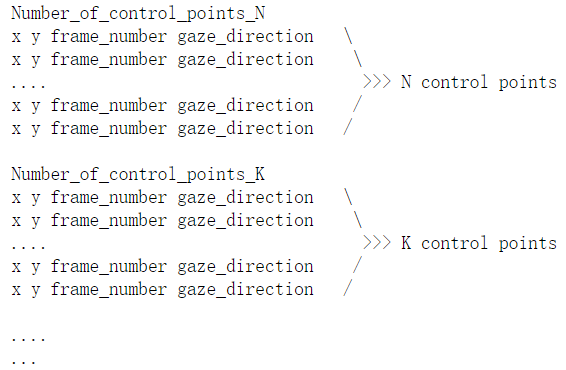
A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In Computer Graphics Forum, volume 26, pages 655–664. Wiley Online Library, 2007.

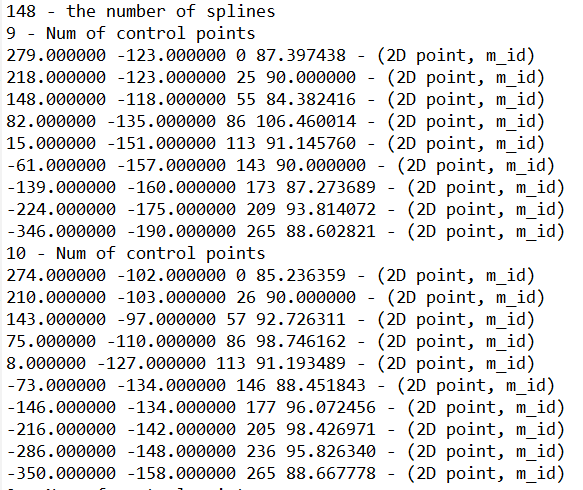
（有论文引用：Learning an image-based motion context for multiple people tracking.）

两个场景，每个场景有786个不同的行人

3部分组成：ZARA-01, ZARA-02 and UCY（Univ）.

UCY中有头部角度信息





文件中注释以“-”开头

每个样例的数量在第一行中

位置xy是图像坐标，图像中心是（0,0）

Gaze\_direction是人观察方向，单位是度，0度表示正在向上看（具体正负值没说含义）

##### Zara01

720\*576

6分1秒

视频帧率25fps

##### Zara02

720\*576

7分1秒

##### Students

720\*576

3分36秒

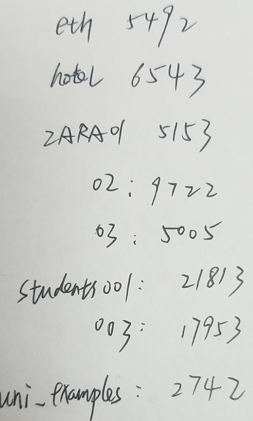
#### Social GAN对ETH和UCY的处理

在social GAN的数据集中，只使用了帧号，行人ID，位置x和位置y（且位置是保留两位小数）

每一个子场景都按照3:1划分train和val

训练某一子场景时，用其他所有的数据训练和验证，用该场景测试

每个子场景的片段数量（多少行，不代表人数也不代表帧数）



### SDD

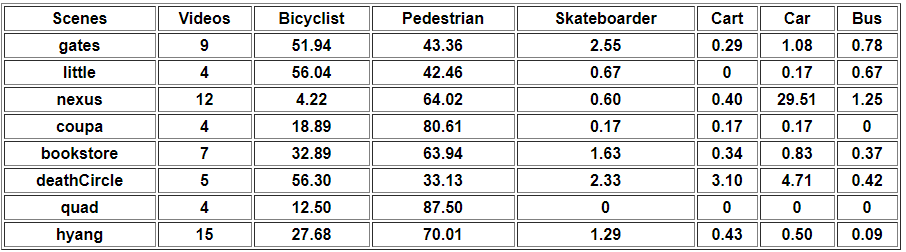
论文：Learning social etiquette: Human trajectory understanding in crowded scenes.

bird’s-eye view，8个不同的场景；

行人，自行车，车辆在大学校园环境内同时navigate；

图像坐标

8个场景：bookstore，coupa，deathCircle，gates，hyang，little，nexus，quad，每个场景都有视频，但视频边界是抖动的，连续帧之间的区域是不同的



**注释文件**

每行有10列，空格分开

ID，bb左上角x，bb左上角y，bb右下角x，bb右下角y，帧号，是否丢失（1表示超出屏幕范围），是否遮挡（1表示存在遮挡），生成方式（1表示标注是自动生成的），类别标签

论文将环境分为若干类：

Road：公路

Roundabout：环岛，环岛交叉路口

Sidewalk：人行道

Tree：树木

Building：建筑物

Bike rack：自行车停放架

### Town Center

论文：Stable multi-target tracking in realtime surveillance video（2011年CVPR）

参考：<https://blog.csdn.net/qq_35975447/article/details/106544475>

**视频文件+标注数据**

数据解析之后大概7000-8000帧，前4500帧是由标注的，在top文件中；后2998帧是没有标注的。数据基本是以下的街景照片，判断是墙角摄像头俯拍的，所以在人员分布基本能够清晰的看出来，调研发现很多人会使用这个数据做行人跟踪项目。

**优缺点**

优点：视频（可以做跟踪项目）、数据场景简单（适合初入行人检测）、标注完整等等、画面清晰干净；缺点：场景单一、数据量少等、俯拍角度人小（相对整图）。因为以上缺点，如果想训练一个可以迁移到其他场景的模型，可能效果上于测试集对比，可能会一定出入；数据量少，模型训练过度拟合到该场景，使得在普通场景下，难以适用。当然如果做跟踪的，这个会一个很好很干净的数据集。

### The CUHK Crowed

**论文**

Jing Shao, Chen Change Loy, and Xiaogang Wang. "Scene-Independent Group Profiling in Crowd." in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

Jing Shao, Chen Change Loy, and Xiaogang Wang. "Learning Scene-Independent Group Descriptors for Crowd Understanding." IEEE Transaction on Circuits and Systems for Video Technology (TCSVT), 2016.

**介绍**

在许多环境中包含许多不同密度和视角尺度的人群视频。

215个拥挤场景，共474个视频。每个视频都删除了过短轨迹，静止不动的点，其他错误。

数据集的详细信息见dataset\_info, 它包含视频名称，视频长度，视频大小，视频源，视频t0（用于组检测评估的帧），组检测（在cvpr论文中使用的300组检测），video\_gt（groundtruth视频类）和场景编号。

### ~~The subway station~~

来自论文：Socially-aware Large-scale Crowd Forecasting

地铁车站数据集是在纽约中央车站采集的30分钟序列，每个序列包含超过40,000个关键点轨迹。

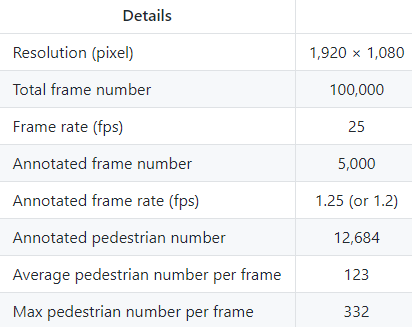
### the New York Grand Central

[22] S. Yi, H. Li, and X. Wang, “Understanding Pedestrian Behaviors from Stationary Crowd Groups,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3488–3496.

超过12000条准确的行人轨迹，从一个小时的拥挤视频标注得到

视频原始帧率为23fps，标注时每20帧标注一次。

33:20分钟，50010帧，25fps，分辨率720\*480



### the Edinburgh Informatics Forum datasets.

[23] B. Majecka, “Statistical Models of Pedestrian Behaviour in the Forum,” Master’s thesis, School of Informatics, University of Edinburgh, 2009

检测在爱丁堡大学信息学院主楼内走过的行人，几个月的观察，每天产生约1000条观察到的轨迹，总共有超过92000条轨迹。

视频原始帧率为9fps，标注时逐帧标注。

### VIRAT

现实环境中，自然场景，人没有受到控制，经常有偶发时间和活动；

多样性，在遍布美国的多个站点收集数据，包括各种相机视角和分辨率，包含许多人；

数量，包含各类人类动作和交互，每个动作类别都有大量样例（>30）；

广泛的分辨率和帧速率，2-30HZ，人身高在10-200像素，数据集既提供高清质量的原始视频，也提供时空上降采样的版本；

地面视频+航拍视频，包含两类视频；

包含两大类活动（单个对象和两个对象），涉及到人和车辆。每个版本包含不同的活动标注；

超过250小时的地面摄像头视频，并

The dataset and annotations are available at the mevadata.org site.

### ATC

论文：D. Brscic, T. Kanda, T. Ikeda, T. Miyashita, "Person position and body direction tracking in large public spaces using 3D range sensors", IEEE Transactions on Human-Machine Systems, Vol. 43, No. 6, pp. 522-534, 2013

日本大阪的ATC购物中心，建立的一个跟踪环境，该系统由多个3D距离传感器组成，占地面积900平方米。

提供了完整的人员跟踪数据集以及原始传感器测量值。

此处提供的数据是在2012年10月24日至2013年11月29日之间收集的。通常，数据收集是在每周的周三和周日各一次，从早到晚（9:40-20:20）。数据集共92天。

每天一个csv文件，csv文件的每一行都对应某一时刻的某个被跟踪人员，包括字段如下。

时间，人员ID，位置x（mm），位置y（mm），位置z（高度，mm），速度（mm/s），运动角度（rad），facing angle（rad）

### Daimler

论文：N. Schneider and D. M. Gavrila. Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study.

In Lecture Notes in Computer Science: Proc. of the German Conference on Pattern Recognition (GCPR), vol. 8142, Springer, 2013.

包含从固定和移动车辆收集的68个行人序列的集合。

四种行人运动类型：过马路，停车，开始行走和弯腰，行人没有遮挡。

We provide

* original stereo pairs (8 bit PGM, 1176x640)
* calibration data
* ground truth (GT) annotations,
* pedestrian detector measurements and
* vehicle data (speed, yaw-rate)
* event tags and time-to-event labels (TTE in frames).

特别的，提供轨迹数据以（u,d）格式，u is the lateral image coordinate and d is disparity，在地面是平的假设基础上，可以计算地平面上行人的位置（X,Z）

序列总共包含19612个立体图像对。12485张图像包含人工标注的bounding box，9366张检测出的bb。仅对5-50米范围内进行评估，并且需要视差。

For more information, see our above GCPR’13 publication or read the documentation

进一步用事件label和事件发生时间（帧中的TTE）进行标记。TTE的含义看原文档。

### TrajNet

### PIE（Pedestrian Intention Estimation）

论文：PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction

参考：<https://github.com/aras62/PIE>

车辆第一人称视角的

用于研究交通场景中行人行为的数据集。包含车载摄像头记录的6个小时以上的视频数据，还有与视频数据同步的OBD传感器数据提供的车辆信息（车速，航向角和GPS）。

丰富的空间和行为标注，可用于研究行人和车辆的交互，以及与基础设施的交互（交通信号灯，标志和斑马线）。

超过300K的视频帧和1842个行人样本，使之成为研究交通场景中行人行为的最大的公开数据集。

Bounding box，是否遮挡（未遮挡，部分遮挡，完全遮挡），帧ID

**行人：**

动作：walking或站立

手势：hand\_ack (pedestrian is acknowledging by hand gesture),hand\_yield (pedestrian is yielding by hand gesture), hand\_rightofway (pedestrian is giving right of way by hand gesture), nod, or other.

Look：是否在看

Cross：是否crossing，crossing时与当前车相关，crossing但与本车无关

**车辆**

类型：car，truck，bus，train，bicycle，bike

**红绿灯**

红绿灯类别，当前等颜色

**标志**

标志类型

**crosswalk**

**transit\_station**

### ForkingPaths

3D模拟器（CARLA）创建的一个新的轨迹预测数据集，基于游戏引擎-虚幻4，在这个数据集中，我们重建了真实世界的场景和动态事件，然后让标注者控制agents走到设置好的目标点，记录下这些能反映真实人类在同样情况下可能会走的路径。

多名人类标注者观察该场景4.8秒后就可以以第一人称或者第三人称控制agent走到目的地。我们希望在这种方式下，可以在同样的场景中，捕捉到人类真实的反应以及可能选择的路线。

标注完成后，我们在3D模拟器中选择多个摄像头位置和角度进行数据录取，可以模拟一般的45度角监控视频的角度，也有头顶的无人机视频角度。我们甚至可以使用不同的天气状况和光照条件。

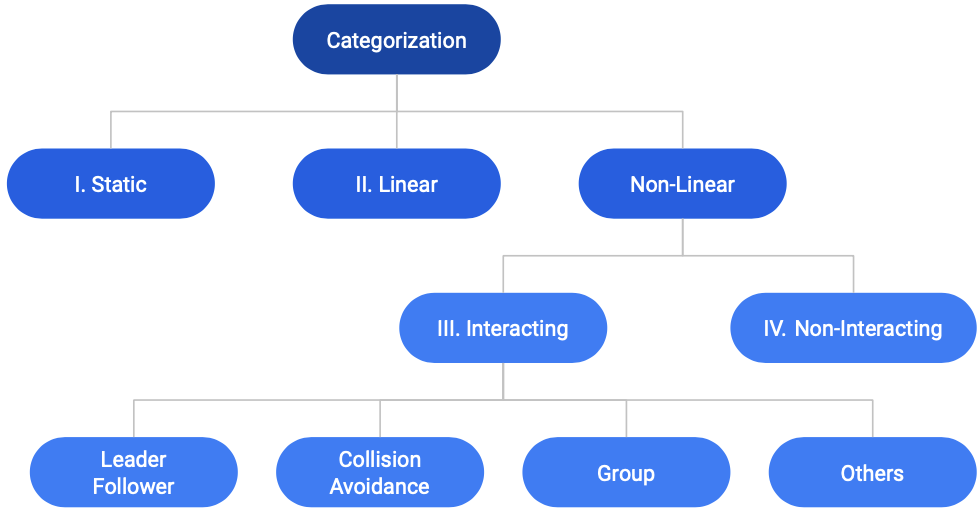
## Benchmark

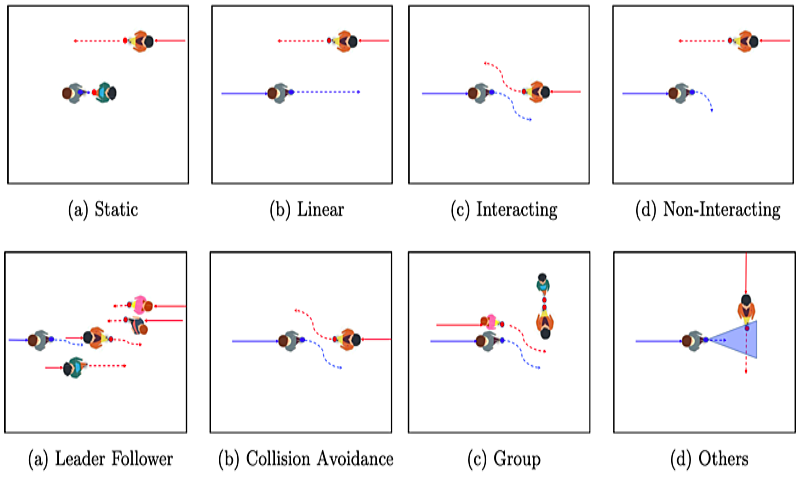
**Benchmark中提供了一系列工具，考虑直接拿来使用**

### TrajNet

### TrajNet++

（注意对轨迹进行了不同类别的分类）





### MOT-Challenge

Multiple Object Tracking Benchmark

### JackRabbot

Detection And Tracking Dataset and Benchmark

## 车辆-交通

Loss：RMSE

策略：8s一个片段，观察3s预测5s，帧率0.2s，即观察15帧预测25帧

### NGSIM driving dataset

#### 论文

J. Colyar and J. Halkias. Us highway dataset. Federal Highway Administration (FHWA), Tech. Rep. FHWA-HRT-07-030.

FHWA.NGSIM Interstate 80 Freeway Dataset. FHWA-HRT-06-137.

FHWA.NGSIM US 101 highway dataset. FHWA-HRT-07-030.

RDE还提供了其他NGSIM数据环境（加利福尼亚州埃默里维尔的I-80，加利福尼亚州洛杉矶的Lankershim Boulevard和乔治亚州亚特兰大的Peachtree Street）的数据集和相应的元数据。

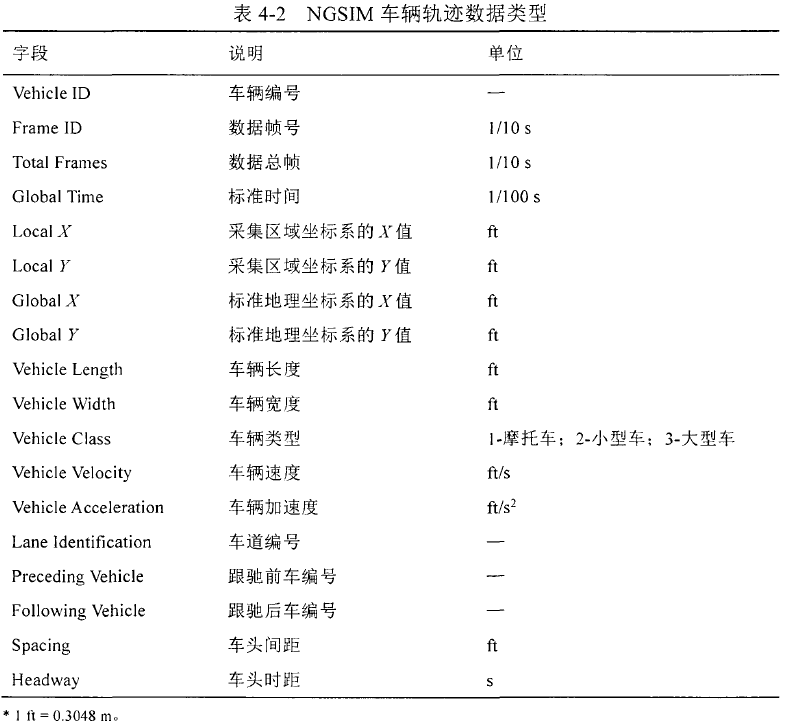
#### 介绍

２００４年，美国联邦高速公路管理局（ＦｅｄｅｒａｌＨｉｇｈｗａｙＡｄｍｉｎｉｓｔｒａｔｉｏｎ，ＦＨＷＡ）发起了“次时代仿真”（ＮｅｘｔＧｅｎｅｒａｔｉｏｎＳｉｍｕｌａｔｉｏｎ，ＮＧＳＩＭ）研究计划，通过共享高精度的车辆轨迹数据，为研究者提供了统一的模型研宄和参数标定平台，吸引了全世界研究机构和个人的参与。ＮＧＳＩＭ项目总共采集了两组高速公路的车辆轨迹数据。

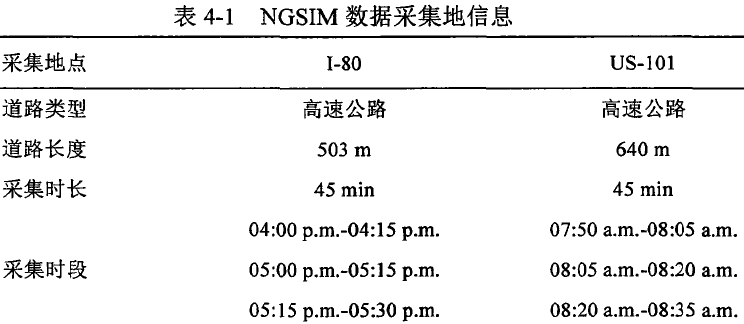
NGSIM项目通过在高楼640米架设固定摄像机的方式，以每秒10帧为间隔对上述高速公路中的车辆进行图像采集，最后通过图像识别技术提取并分析了车辆行驶过程中的相关数据，数据类型包括车辆位置、车辆速度、车辆类型等。数据集包含各种交通状况（轻度，中度和拥挤），总共包含约6000辆汽车。

考虑到ＮＧＳＩＭ车辆轨迹数据具有数据精度高、数据类型多样的特点，能够详细地描述车辆跟驰过程，适合用于数据驱动模型的研究。

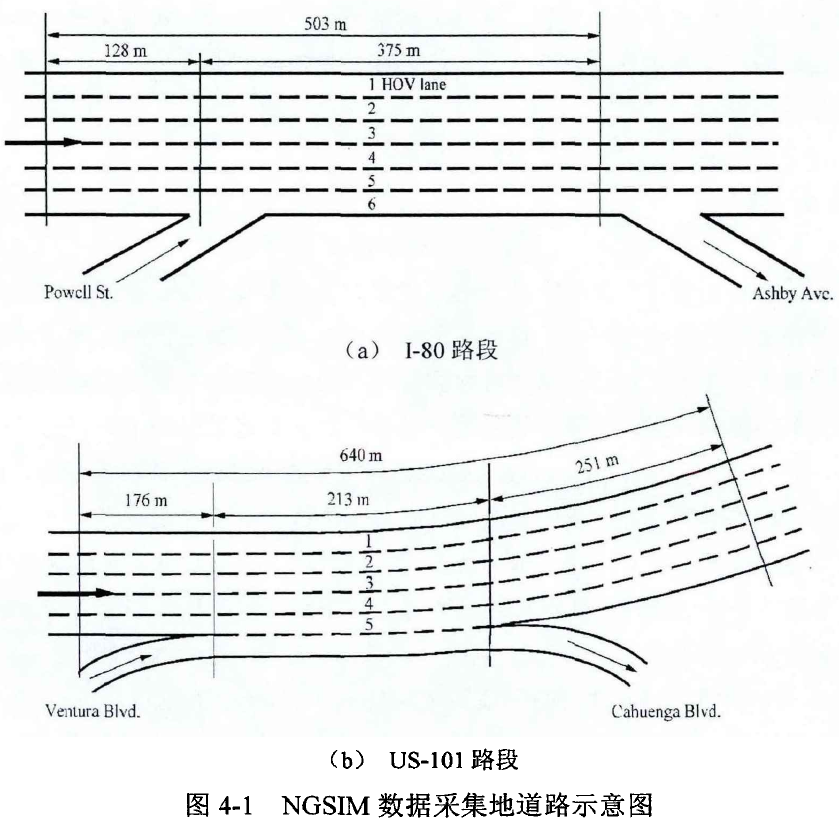
#### 字段介绍



#### 采集地信息-采集地道路示意



These periods represent the buildup of congestion, or the transition between uncongested and congested conditions, and full congestion during the peak period.



#### 其他

有边界的坐标

3个子集

数据正确性不保证，数据中可能存在空白

#### 数据清洗+预处理

A critical evaluation of the Next Generation Simulation (NGSIM) vehicle trajectory dataset

### Massachusetts Driving Dataset

数据集官网

https://www.cs.toronto.edu/˜vmnih/data/

### highD

#### 论文

#### 介绍1

数据集包括来自六个地点的11.5小时测量值和110 000车辆，所测量的车辆总行驶里程为45 000 km，还包括了5600条完整的变道记录。通过使用最先进的计算机视觉算法，定位误差通常小于十厘米。 虽然数据集创立的初衷是为自动驾驶而创建的，但它也适用于许多其他任务，例如交通模式分析或驾驶员模型的参数化

因车道数和速度限制分6个。

视觉检测+贝叶斯方法来平滑运动轨迹。

总共记录的数据总量是NGSIM方法记录的轿车数量的十倍，卡车的70多倍。

还有一个inD数据集和roundD数据集。

110500多辆车，44500KM，147个小时，6个地点，包括车辆轨迹，类型，尺寸和行为

不同交通状态，

#### 数据集构建

Map到第一帧，旋转平移第一帧，使得车道与图像边界平行

Unet检测得到bb

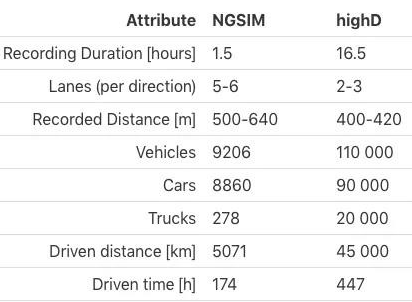
跟踪算法连接bb

使用[28]平滑轨迹：rauch-tung-striebel smoothing平滑

定位误差在像素级别

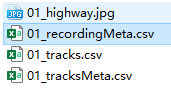
定义了几种行为（）

#### 与NGSIM对比



#### 数据解析

60个记录，每次记录有4个文件。



图像：不带车辆的道路图像，分辨率是降低之后的；

**RecordingMeta（整体的统计数据）**

当前记录的ID，视频帧率，位置ID（总共6个），限速，月份，weekday，开始时间，持续时间，所有车辆的总行驶距离，所有车辆总时间，车辆总数，小车总数，大车总数，upperLaneMarkings（8.51;12.59;16.43），lowerLaneMarkings（21.00;24.96;28.80）

**XX\_tracksMeta.csv（每辆车的统计数据，1047辆）**

车辆ID（升序），车辆长，车辆宽，当前车开始的帧，当前车总帧，车类型，行驶方向（Either 1 for the left direction (upper lanes) or 2 for the right direction (lower lanes)），行驶距离，最小速度，最大速度，平均速度，前车最小距离DHW（-1表示没有前车），THW最小值（-1表示没有前车），最小碰撞时间TTC（-1表示没有前车或没有有效TTC），变道次数

**XX\_tracks.csv**

按车辆ID排序，每辆车内按帧ID升序排序

帧ID，车ID，bb左上角x坐标，bb左上角y坐标，width，height，x速度，y速度，x加速度，y加速度，frontSightDistance（从当前车中心到当前record末尾的距离），backSightDistance（来向，距离记录开始边界的距离）

DHW（距离，0表示没有前车），THW（碰撞时间，0表示没有前车或没有有效时间），TTC，

precedingXVelocity（图像坐标系中前车的纵向速度）

前车ID（0表示没有），后车ID（0表示没有），左前方车ID，左侧并行车辆（坐标有重合即算作重叠），左后方车ID，右前方车ID，右侧并行车ID，右后方车ID，

车道ID

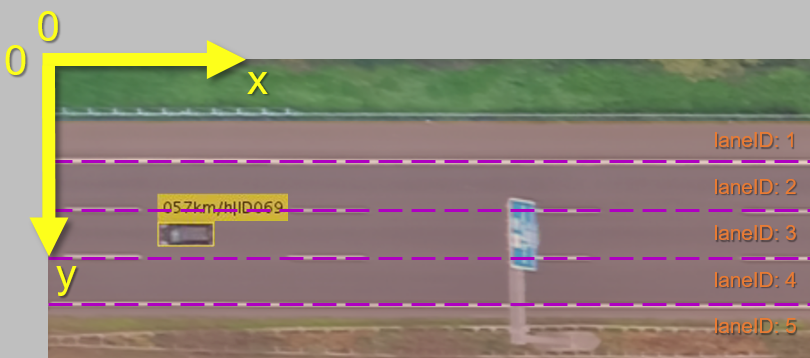
#### 坐标系

全局坐标即图像坐标系

不使用像素坐标，因为所有都已经转化为SI坐标

车道在水平方向延伸，位置可以由y坐标表示

上方车道的车向左行驶，因此速度在X方向是负的



### InD数据集

#### 数据集构建

都map到第一帧

检测：标注400张用于训练

两个基于U-NET的网络分别检测大目标和小目标

用分割，再进一步处理得到bb，这样包含语义信息

用阈值和形态学计算去除异常值

最后人工审核和标注有歧义的

Tracking之后，再贝叶斯平滑处理

#### 介绍

十字路口

4个路口，10个小时数据

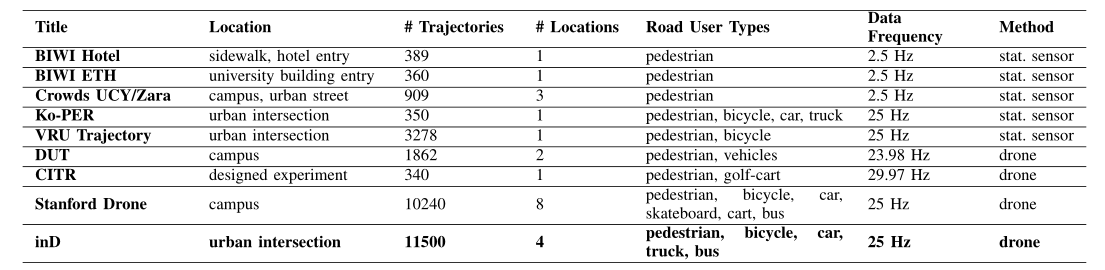
Camera-equipped drones

丢失了高度信息

无交通信号信息

#### 与其他数据集对比

ETH 和 UCY，规模小，只有行人；SDD有多个类别，但车辆很少；CITR中，受控的人在场景内行走；highD是针对高速路的，没有行人；



### INTERACTION数据集

论文

目前学术界和工业界一致认为，行为预测（Prediction: 如轨迹、动作、意图）是自动驾驶领域最具挑战性的问题之一，它是阻碍全自动驾驶实现的一大因素。而要解决这一问题，有两个条件不可或缺：一是包含很多车辆和行人交互的真实场景运动数据的收集和积累，二是可以通过这些数据对各种预测算法进行正确而有效的评价。但不幸的是，目前没有一个可以公平比较不同预测模型（或算法）性能的基准，尤其是在考虑到有规划在环的情况下（模型集成了预测和规划两个模块），公平基准是个大难题。

#### 介绍

地图信息的重要性和交互实体的完整性；

交互驾驶场景的多样性，不同文化国家，场景和行为的复杂性，关键情况会更有价值，地图信息很重要。

车载传感器VS鸟瞰，各自的优缺点。

高精地图：lanelet2模式，厘米级别精度的高精地图

关于数据集的各种分析，详见原文

#### 构建过程

视频稳定和对齐，都map到第一帧；

Faster RCNN检测；

卡尔曼滤波用于数据关联和跟踪；

用[27]平滑车辆运动轨迹

### Argoverse

论文

Argoverse: 3D Tracking and Forecasting with Rich Maps

<https://www.jiqizhixin.com/articles/2019-07-26-8>

<https://blog.csdn.net/taifengzikai/article/details/104096337>

#### 介绍

Argoverse是三种不同类型研究数据的集合。第一种数据是包含无人驾驶车队采集的113个场景的传感器数据，以及所有物体的3D目标标注的追踪信息。第二种数据是一个包含有30多万个场景的所有物体的运动轨迹。第三种数据是几个匹兹堡和迈阿密及周边地区的高精地图，对以上两种数据提供了丰富的环境信息。

Argoverse运动预测数据集。对无人车而言，跟踪物体已经运动的轨迹（3D目标跟踪）固然很重要，但预测物体将会怎么运动也很重要。同人类驾驶员一样，无人车需要去评估：“这辆车会不会并入我的车道”，“我前面的这个司机会左转吗”等等问题。为此，我们挖掘出了超过1000小时的有价值的驾驶记录。这里，有价值表示车辆处于十字路口，或者因为前方有车要并入而减速，或者在转弯之后加速，或者因为路上有行人而停下来，以及其他类似的情况。我们筛选除了超过30万个这样的场景，每个场景包括五秒10赫兹的序列，和每一帧的每个被跟踪物体的2D鸟瞰视图的中心。每个序列有一个有价值的运动轨迹作为我们运动预测的衡量标准。研究人员可以根据前两秒观测到的轨迹去预测接下来三秒某辆车的运动轨迹。

333K的5s ，211K训练，41K验证，80K测试；10HZ，0-2s观察，2-5s预测，每个轨迹都map到地图上。

### In-house

行为预测，包含地图，bb和跟踪相关注释。

感知结果是自动获取，轨迹是手动标注。

2.2M训练，0.55M测试，每条轨迹4s，0-1观察，1-4预测。

轨迹从真实世界场景中提取，包括静止，转向，变道，掉头，保留了自然驾驶场景分布。

地图：包括车道线，停止/避让标志，人行道和速度bump？

### KITTI

### Cityscapes

50个不同城市的街道场景，双目视频序列。高质量的pixel level的标注有5000帧，weakly标注的有20000帧。

目的：像素级别的场景理解任务，实例级，全景语义标注，弱标签数据训练网络。

**标注信息**

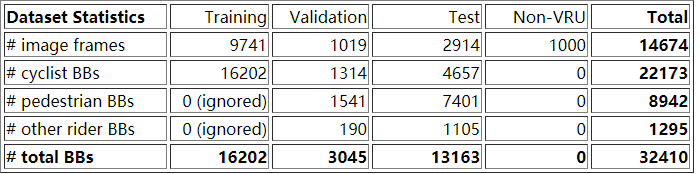
30个类别，多样性，



### Cyclist

论文：A New Benchmark for Vision-Based Cyclist Detection.

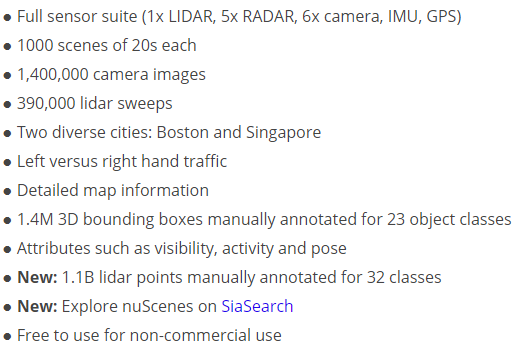
用来检测骑自行车人的数据集



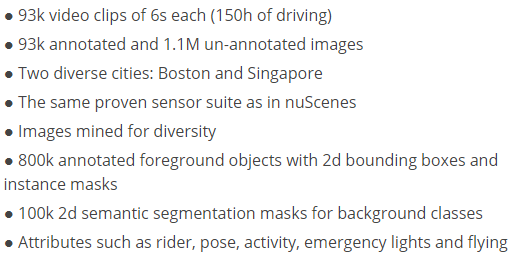
### NuScenes

面向自动驾驶的数据集。

**数据集总体情况**



**关于图像**



### BDD100K dataset

论文：BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning

面向自动驾驶，第一人称视角的，迄今为止最大的驾驶视频数据集。

超过1100个小时高清视频数据，100000个视频序列，覆盖一天中不同的时间点，天气情况和驾驶场景。所有视频都包含GPS，IMU和时间。

标注信息：在100000张图片上标注2d bounding box，类别包括bus, traffic light, traffic sign, person, bike, truck, motor, car, train, and rider.

超过10000张图像，像素级别的语义标注。

可行驶区域（不知道是标注的还是模型检测得到的）

车道线标注。

### Apollo数据集

3D点云，图像，GPS/IMU，地图数据。

Pose信息，传感器的内外参

多种时间，多种天气条件，多种场景。

### Udacity

摄像头，用crowdAI标注，类别：小车，卡车，行人。

Csv文件中：Bounding box位置，帧号

Dataset2中，多了Street Lights，多了遮挡信息的标注

### TRAF(The Dense and Heterogeneous Traffic Dataset)

论文：TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions, CVPR 2019.

密集且异构的交通视频数据集。包括各类交通参与物。

总体而言，每帧含有13辆机动车，5个行人和2辆自行车。

标注包括像素坐标，agent的ID和agent类型。

视频根据摄像头角度（front-facing/top-view），运动/静止，时间，难度级别进行分类。

### Lyft Level 5

数据采集设备：类似apollo

高精度地图：4000多个人工标注的语义元素，包括车道段，人行道，停车标志，停车区，减速带，所有的语义地图均对应到了几何地图中。

**感知数据集**

与nuScenes format一样

**预测数据集**

来自激光雷达，雷达，摄像头采集的数据，处理得到，所有agent的运动数据，是训练运动预测模型的理想选择。

17万个场景，

数据集以zarr格式提供，

ego pose, time, and multiple agent states（the position, orientation, bounds, and type）

### HDD dataset

“Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and

Causal Reasoning,”

CITR

DUT

Ko-PER:固定式激光扫描仪和相机，德国十字路口

VRU：同一路口的另外一个版本

## 其他部分数据集

### PETS 2009

论文：PETS2010: Dataset and Challenge

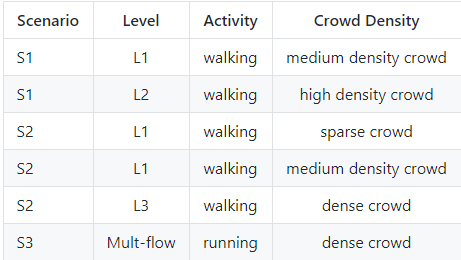
多台摄像机拍摄，大约40名演员。

人群计数，密度估计，跟踪人群中的个人，检测流量和人群事件。

包含相机参数，内外参。

轨迹：行人的bb，关键帧之间插入bb，平滑轨迹得到。每个进入场景内的行人有唯一ID。

注意：标注是有误差的。

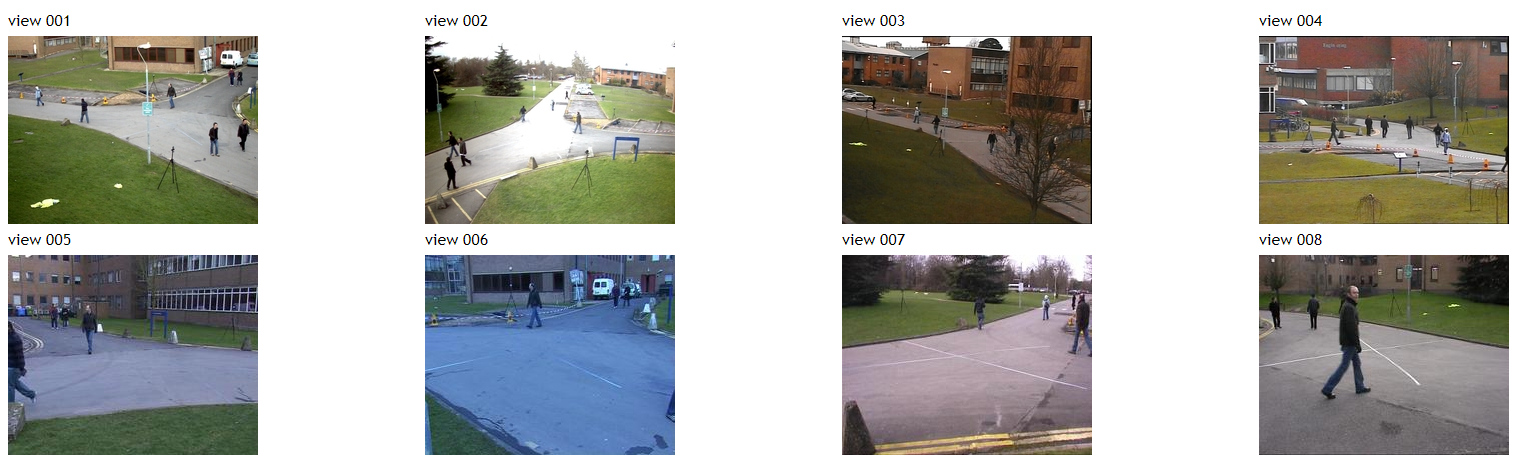


S1：人群计数和密度估计

S2：跟踪

S3：Flow Analysis and Event Recognition

8个相机



### HERMES(Pedestrian Dynamics) Dataset

详见：<https://ped.fz-juelich.de/database/doku.php>

在实验室条件下进行，以关注单个变量的影响。

包含鸟瞰视频和轨迹文件。

每行文件：ID，帧号，XYZ坐标（3d）

有单向流动的实验，双向流动的实验

封闭走廊，不同长度不同宽度，不同人数下的实验

### Waymo

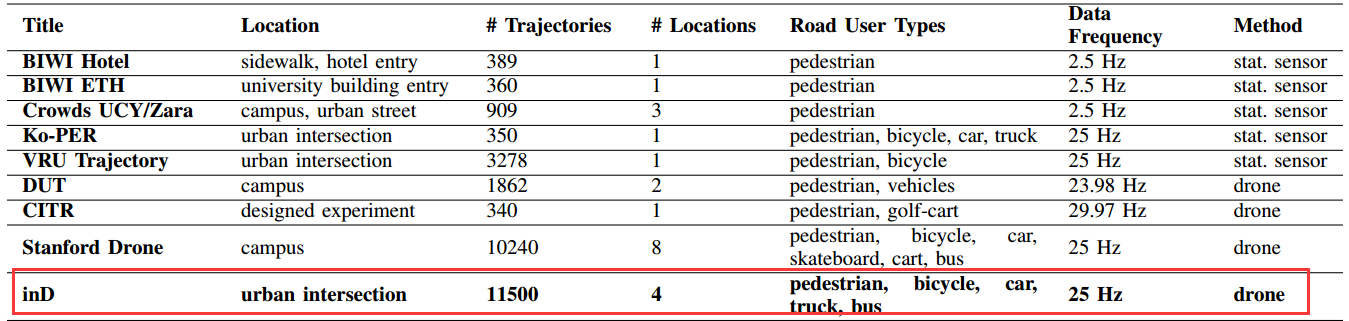
1950个细致划分，每个包含20s，10HZ的数据。

原始数据包含：中距离激光雷达，短距离激光雷达，5个摄像头（正面和侧面），传感器之间时间同步，传感器之间的旋转平移，车辆姿态数据。

标注信息：4个物体类别，雷达数据高质量标注，3D框，视频数据标注，2D框和跟踪ID

### InD

城市交叉路口，无人机采集，11500条轨迹，4个地点采集，25HZ



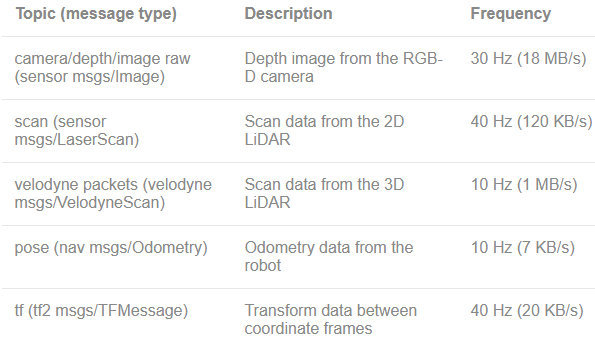
### L-CAS

论文：Multisensor Online Transfer Learning for 3D LiDAR-based Human Detection with a Mobile Robot

采集装备：四个轮子的机器人，odometry sensor, a Velodyne VLP- 16 3D LiDAR (mounted at 80 cm from the floor), an ASUS Xtion PRO LIVE RGB-D camera (55 cm from the floor), and a HOKUYO UTM-30LX 2D LiDAR (11.9 cm from the floor).

分静止和移动两个场景，

数据包括：



### Wild Track

论文：MegaPixels: Origins and Endpoints of Datasets Created "In The Wild"（2020年）

真实世界场景，不是演员，在苏黎世联邦理工大学主楼外采集，7个35分钟的视频，数千名学生。

对超过1000个人进行标注，标注位置，

Go pro采集的

#### DUT

论文：Top-view Trajectories: A Pedestrian Dataset of Vehicle-Crowd Interaction from Controlled Experiments and Crowded Campus （2019四月）

（The 30th IEEE Intelligent Vehicles Symposium in Paris.）

校园场景，人轨迹，有车辆影响，车辆行人都没有优先通行权；

两个位置，一个是没有信号灯的交叉路口的人行道区域，一个是一个较大的共享空间；

轨迹用卡尔曼滤波refine；

还有一个sister数据集，CITR，是在受控实验环境下进行；

无人机采集，视频分辨率4K，23.98fps；

共1793条轨迹，

行人轨迹：ID，帧号，label，xy位置，xy速度

车辆轨迹：ID，帧号，label，xy位置，heading angle，速度

### CITR

停车场，受控实验中收集，无人机采集，

### VRU

1068个行人，464个骑自行车的人，城市交叉路口，用相机和激光雷达采集。

广角双目摄像头，记录行人轨迹，跟踪行人头部生成3D位置，50HZ；激光雷达跟踪cyclist，12.5HZ。

ID，时间戳，x位置，y位置（in meter）

### Ko-PER

论文：The Ko-PER Intersection Laserscanner and Video Dataset（2014年）

公共路口，摄像机+激光雷达

## 数据集构建

轨迹提取框架---基于视觉的各类算法

行为等的半自动标注技术

半自动标注技术---能力评估，审核，任务分发机制

框架主要用于交通摄像头，也可以用于无人机等设备的轨迹提取（需要加视频稳像）

轨迹是最精细的，在此之上，我们有不同尺度的grid，表示可以到达的区域

行为的标注有起始时间，具体的行为的集合

未来考虑跨摄像头，集成更长时间范围更广空间内的运动和行为信息

### Large-Scale extraction of accurate vehicle trajectories for driving behavior learning

由于与其他车辆和行人的强烈互动，城市环境仍然是自动驾驶汽车所面临的巨大挑战。越来越多地探索机器学习方法来解决这些情况，但是它们的性能在很大程度上取决于车辆轨迹数据集的可用性。此类数据集通常是使用特定的基础结构和设备创建的，这既昂贵又不可扩展。结果，当前只有少量的车辆轨迹数据集可用，代表非常具体的情况，例如高速公路行驶，并且包含有限数量的轨迹。

我们认为，要释放自动驾驶车辆行为学习方法的潜力，我们需要准确的车辆轨迹的大型数据集，以表示在非常不同的情况下交互的车辆。在本文中，我们介绍了一种全自动且可扩展的框架，用于从单个固定的单眼交通摄像头中准确提取车辆轨迹。我们利用了这样一个事实，即交通摄像头通常位于交通拥挤且发生大量互动（例如交叉路口）的地方，因此它是非常多样化且具有成本效益的高度多样化车辆轨迹的来源。通过这项工作，我们旨在为大规模轻松，准确地创建车辆轨迹数据集创建一个框架。

#### 介绍

现实世界中的测试非常昂贵且危险，因此诸如Carla[2]或Waymo CarCraft [3]之类的仿真环境被广泛用于开发，测试和验证模型。但是，这些环境受周围交通和行人编码行为模型的限制，因为它们不能反映真实人类行为的复杂性

一类方法学习直接用于控制车辆的驾驶策略[456]，另一类方法学习预测意图和未来运动信息，然后使用预测结果来计划和控制车辆。方法的性能都依赖于庞大而多样的数据集的统计代表性。

移动车辆上布设传感器收集，视野有限，或专用的感应基础设施，成本高且不易扩展，且区域固定。交通摄像头有高度多样化的车辆轨迹，成本低，交通摄像头也很多，利用视觉领域最新进展，可以有一种自动框架，从单目摄像头中提取车辆轨迹。

框架可以精确提取轨迹，基于真实仿真环境的性能评估方法。

#### 方法

可以在一系列流程之前加一个视频稳像的操作

**摄像头参数校正**

在线视频流，无法访问摄像头信息，或者其他无法获取摄像头参数的信息的情况。

成像模型，包含内参，外参……

方法一：确定若干组匹配点之间的距离（常用先验知识或来自在百度地图上测量距离），最小化目标函数，直到误差小于阈值

方法二：确定四组匹配点

方法三：灭点及其推导

在摄像头有失真的情况下，扩展[24]，加入失真系数

作为这部分的一个副产品，可以进行流速测算

**目标检测**

MaskRCNN这种提取mask的和YOLO提取bb的

具体的轨迹用区域内的哪个坐标呢（必须要落地的？）

拿bb的中心投影到平面坐标，在俯视图情况下效果好，但在交通摄像头这种有倾斜的摄像头效果不好

**真实世界坐标的确定**

目标检测是同时确定类别，由类别确定其长宽高；

真实坐标（x,y）是地面上，3D框的几何中心，同时假定航向角，根据成像矩阵，可以从3D世界到图像坐标上，最小化目标函数



A与B的补集的交集 + B与A的补集的交集，

W≥1，表示目标检测的mask要在3D框投影之后的mask之内

最终可以得到优化后的真实轨迹和航向角

**跟踪算法**

将同一辆车的轨迹串联起来。

用最大IOU的方法，基于以下假设（视频帧率高于车辆动态，车辆很少被完全遮挡，因为摄像头一般放置在ROI区域的上方）

**平滑**

RTS[21]进行平滑，直接对2D像素坐标进行平滑（原因是测量的不确定性最终都集成在了像素坐标中，平滑可以尽可能的消除误差）

#### 实验

所提出框架的通用性和适应性，跟踪的准确性，轨迹是否精确？

通用性：框架应用在不同的摄像头数据上，都可以在不知道任何信息的情况来提取

跟踪的准确性：手工标注来验证，三个指标。（需要注意到长轨迹有时会被分为若干短轨迹集合，而非从头跟到尾）

跟踪覆盖率：实际跟踪的占车辆总的轨迹的比例

**跟踪精度**

跟踪车辆，从而可以估算其位置速度和航向，需要有带有车辆真实信息的交通视频来进行验证。

使用仿真环境CARLA[2],生成视频并同时保存车辆状态信息，用本框架提取与gt进行对比。

用RMSE指标，三种对比，本文是3D框地面的中心，一种是目标检测2D框的中心，一种是目标检测2D框的下边界中点。

位置平均精度是62cm。

数据源：在线实时交通数据流，部分机构，每个人都可以的合作伙伴

### 构建方法细化

画一个流程图

#### 摄像头参数校正

对于许多已经布设安装好的交通摄像头，或者其他在线交通视频流，我们很难获得其精确完整的内外参等信息，导致无法对成像过程进行解析。因此，本文提出一种用于计算内外参的校准方法，该方法支持摄像头硬件参数完全已知或部分已知两种情况下的校准。

摄像机通过透镜成像，将三维的现实场景投影到二维的图像平面上，而对应的要通过二维图像提取信息并计算出相应对象在三维空间中的几何信息，就需要确定摄像机的成像几何模型，即通过实验计算得到相机自身的参数和相机相对于世界坐标系的方位参数。

参考[张正友]，我们确定成像模型如下

其中，m= 和M= 分别是图像坐标系中的二维坐标和真实世界坐标系中的三维空间坐标，s是尺度因子，RT分别是旋转和平移矩阵。C是摄像头内参数矩阵，f表示，cx表示。H是图像和真实世界坐标之间的单应性矩阵，直接表示投影过程。

……

由于实际测量中存在误差，数据包含噪声，因此采用极大似然估计来优化下列目标函数：

在已知摄像头部分信息时，我们可以将已知参数直接带入内参矩阵，当摄像头信息完全未知时，我们只需要直接估计描述成像过程的单应性矩阵H即可，这两种策略实际上分别对应着摄像头参数的软校正和硬校正两种策略。

上述计算流程是在假设摄像头没有失真的情况下进行的，如果存在失真，通过在公式（1）中添加失真模型，并在目标函数（2）中添加这些参数即可。

方法一：确定若干组匹配点之间的距离（常用先验知识或来自在百度地图上测量距离），最小化目标函数，直到误差小于阈值

方法二：确定四组匹配点

方法三：灭点及其推导

在摄像头有失真的情况下，扩展[24]，加入失真系数

本文提出的摄像头参数校准方法，同样可以用于其他需要对成像过程进行解析的计算过程，如基于单目视觉对车辆流速的测算等。

通过现实世界中的部分带距离的物理量，我们可以得到表征成像过程的已知的公式（1）中的相机参数。

#### 目标检测

目标检测的作用是得到图像坐标中，表示目标对象位置的矩形框，许多算法可以同时得到目标的类别，如小汽车，公共汽车，卡车，行人，骑自行车/电动车的人等。

还有一类算法诸如maskRCNN，得到的是描述目标对象位置和轮廓的图像掩码，相比于传统的目标检测，还涵盖了目标轮廓信息，两者在一定程度上可以相互转化。

#### 真实世界坐标估计

从交通摄像头提取车辆轨迹的一个关键挑战是从图像中对于车辆的检测计算出车辆在真实世界中的位置。

一种最直接的处理方法是直接将矩形框的几何中心或下边界中心作为车辆位置的近似值，然后依据摄像头参数校正得到的单应性矩阵，计算得到真实世界坐标。这种简单的解决方案在俯视视角的视频流中效果很好，这是因为俯视视角下，图像各个位置单位像素表示的实际面积近似，但大多数交通摄像头是倾斜的，在倾斜的情况下，图像近端区域和远端区域的单位像素表示的实际面积差异很大，由此产生了不可忽略的误差。

对于检测到的每种车辆类型，预定义了通用的车辆参数（长宽高），假设车辆真实世界坐标为3D框地面部分的几何中心，因此可以通过航向角确定的车辆角度和单应性矩阵，可以将3D检测框投影到图像坐标上，得到关于车辆区域的多边形区域。因此可以通过最大化

通过最大化3D框在图像上的投影和maskRCNN得到的目标对象掩码，可以计算得到航向角和车辆真实世界坐标。即：



其中，A表示maskRCNN检测得到的目标对象区域，B表示3D矩形框在图像上的投影区域。超参数w≥1可以保证由maskRCNN得到的掩码在投影区域范围内。

如图所示……

真实世界坐标表示：矩形框几何中心，3D矩形框地面部分几何中心，矩形框下边界中点

#### 跟踪

现有跟踪算法

#### 轨迹平滑

考虑到摄像机校准过程中和目标检测过程中引入的误差，计算的不确定性都直接集成在了图像坐标中，因此本文直接对代表车辆位置的图像坐标进行运动平滑，

with a point-mass constant velocity model

#### 后处理

**标注grid**

除了坐标层面的信息，本文还通过不同尺度的网格大小，将坐标映射到所属的网格中，这种处理方式以更高维度的语义信息对轨迹进行了编码。

**标注行为**

不同于预测，数据在训练时的时间范围都是已知的，预训练模型先给出动作识别结果，由人工进行校验，修改错误识别的标签。

~~在海量地质滑坡数据生产过程中，将采用人工标注与智能标注的迭代、交互式的数据标注方式，将人和智能系统融为一体，来大幅度提高标注效率。半自动标注实现流程如下~~

构建识别网络

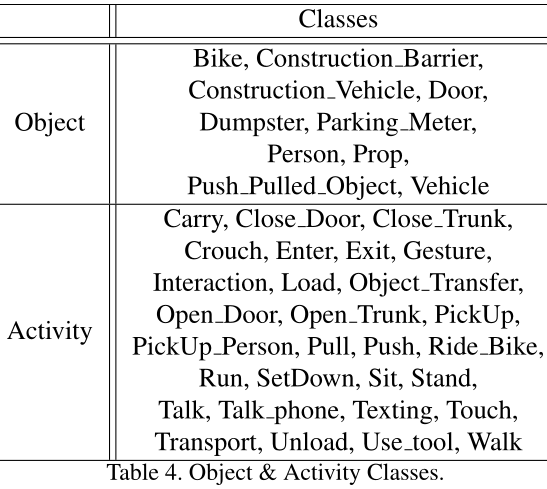
公开数据集和人工标注的数据进行识别模型预训练

预训练模型进行行为识别

交互式标注工具，人工审核标注结果

（标注者能力评价）

#### 行为+意图



确定行人行为集合

确定车辆意图集合

#### 半自动标注

意图行为相关标签的采取半自动标注的方式获取。

一方面

#### 评估方法

人工标注gt进行验证

RMS error

在虚拟环境CARLA [2]中，生成交通视频并保存车辆的运动信息，然后应用本框架处理原始视频得到车辆轨迹，与gt比较来评估准确性。

## Sport Players



## 行为识别

## 时空动作检测

## 暂未分类数据集

The statistics of the three datasets Something-Something dataset (Something-V1 [9] and Something-V2 [28] where the Something-V2 is the 2nd release of the dataset in early July 2018) [9,28], Jester dataset [10], and Charades dataset [11] are listed in Table 1.

human-object interaction recognition

比如UCF101[2]、Sport1M[3]和THUMOS[4],这些数据集中包含很多不需要对长期时间关系的推理也能被识别的行为：静止帧和光流法足以识别这些带有标签的行为。

## 论文中看到暂未分类

JAAD and PIE include egocentric videos of pedestrians in traffic scenes and dynamic conditions. JAAD contains 2,800 trajectories, while PIE’s includes 1,835 trajectories. All videos are captured 1920 × 1080 pixels at 30fps. We follow [30] and sample tracks with an overlap ratio of 0.5,discard tracks with length below 2 sec, and use 0.5 sec of observed data as input to predict future trajectories of length 0.5, 1.0, and 1.5 sec.

HEV-I contains 230 first-person videos of vehicles at 10 fps. Following [45], we split the data into 40,000 train (70%) and 17,000 test (30%) samples. HEV -I does not include human annotations, so we follow [45] and use Mask-RCNN [15] and Sort [2] to generate “ground truth” trajectories. We use 1.6 sec of observed data as input to predict future trajectories of length 0.5, 1.0, and 1.5 sec.