论文阅读笔记：Probabilistic 3D Multi-Modal, Multi-Object Tracking for Autonomous Driving

本页面由姚宇辰于2022/05/17迁移自wiki 羊野空间

一、前置信息：2D/3D MOT 调研（ https://www.i4k.xyz/article/xyq1212/107562167 ）

主流方式

- Kalman滤波器（预测、更新）+匈牙利算法
- GNN方法

基本步骤

①检测 ②特征提取、运动预测 ③相似度计算 ④数据关联

1. 检测：上游框架的检测质量对追踪质量的影响最大

2. 特征提取：
   a. 第t帧识别结果和前t-1帧轨迹特征的提取
      - 运动特征：匀速运动模型、LSTM(轨迹预测)&MLP(第t帧)
      - 表观特征：Re-ID(行人重识别)、CNN、光流
   b. 2D特征与3D特征的融合
3. 运动预测：
   - Kalman滤波、联合概率数据关联、GNN节点消息传递

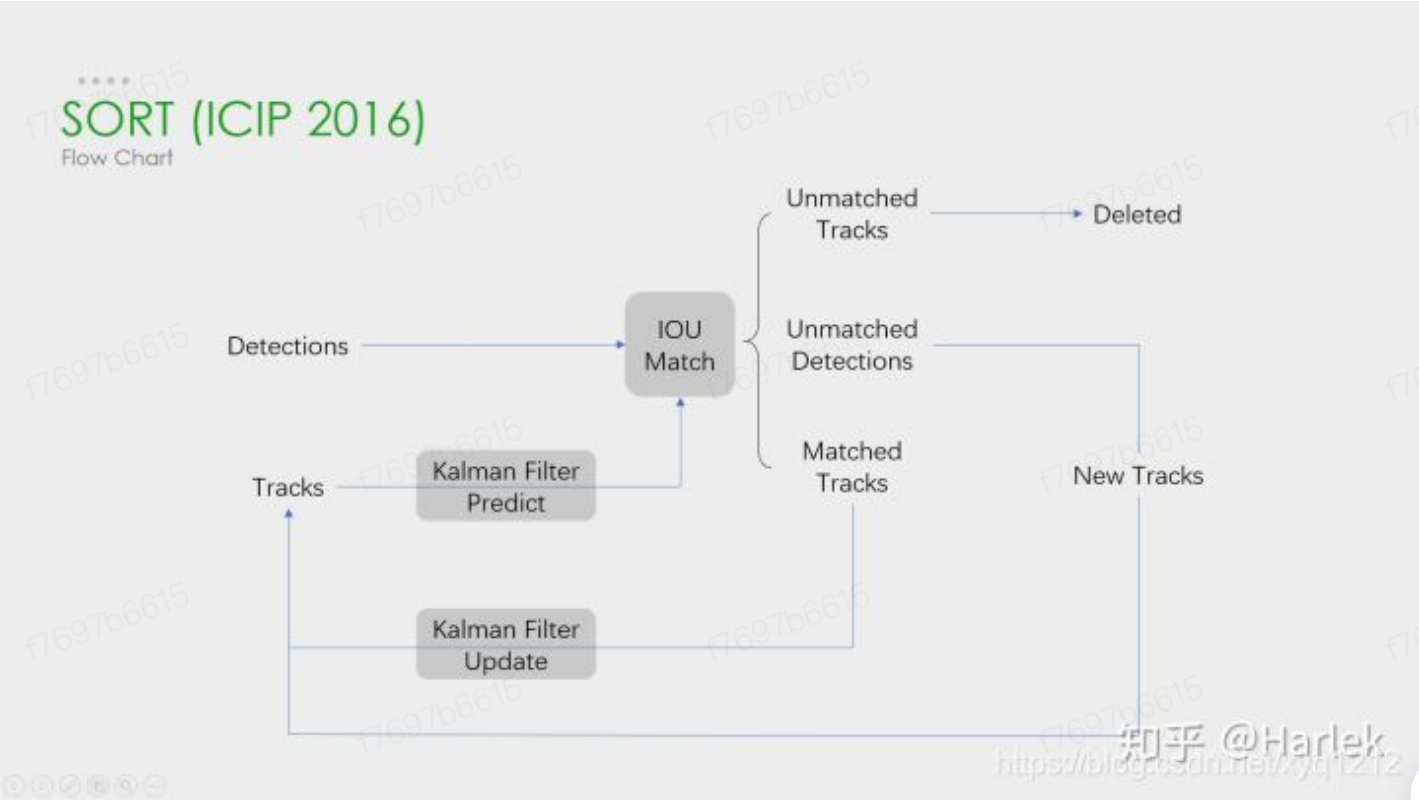
4. 相似度计算、数据关联：
   a. 相似度计算：
      i. 分别计算detection和tracking的运动相似度和表观相似度
      ii. 再将它们以某种方式融合起来得到最终的相似度矩阵

    - IoU、欧氏距离、余弦距离、马氏距离、网络回归
  b. 数据关联：根据相似度矩阵（代价矩阵）得到最终匹配结果，二部图匹配问题
    - 基于IoU的贪婪匹配，只使用运动模型，IoU作为代价矩阵进行贪婪匹配
      - 全局最优、局部最优
    - 匈牙利算法/KM算法
    - GNN

经典算法（6 examples）

SORT：最base的算法，用于2D追踪

- Kalman滤波+匈牙利匹配，IoU为代价矩阵
- 只使用运动模型，Kalman预测值作为运动特征
- AB3DMOT与该框架相似，但是做的是3D匹配

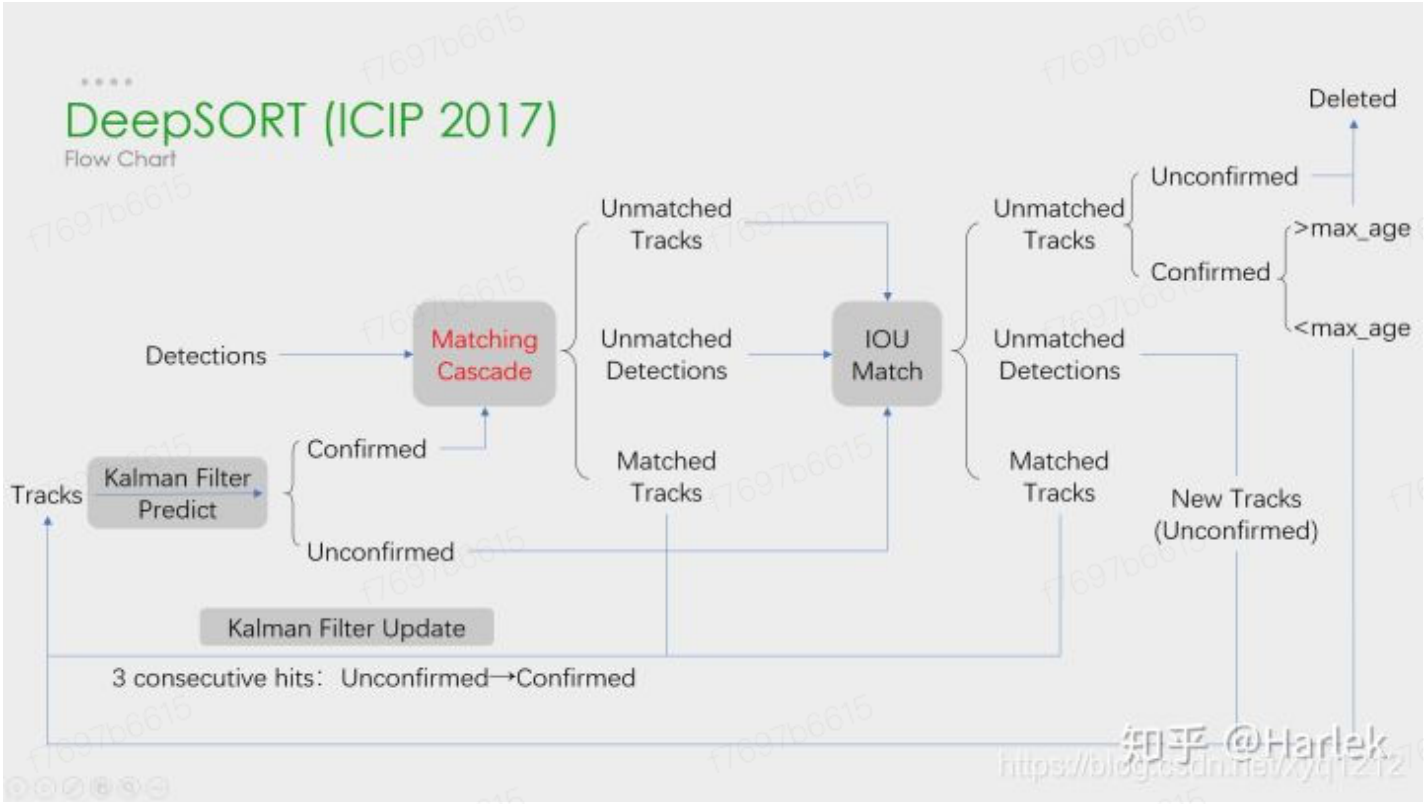


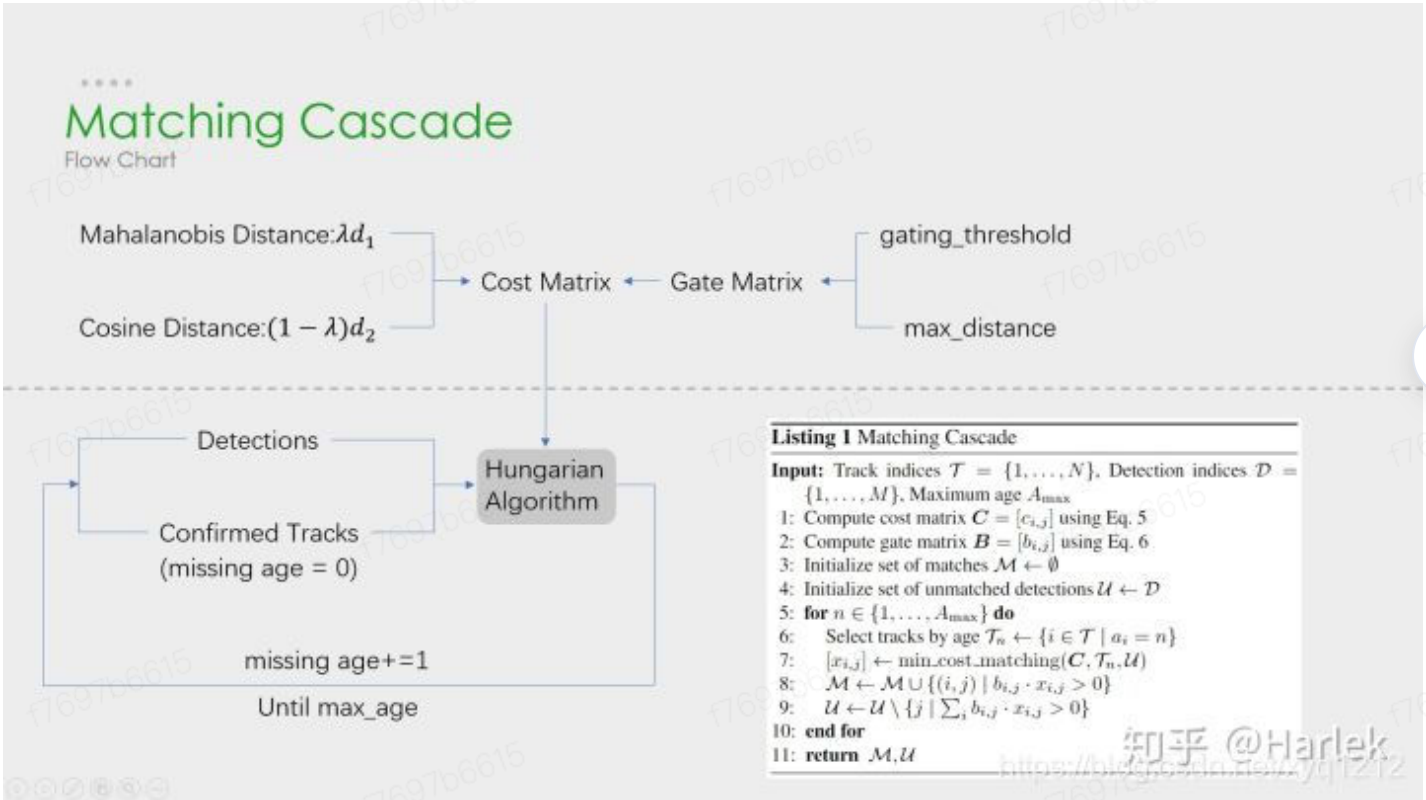
DeepSORT：SORT的改进，代价函数中加入了表观特征

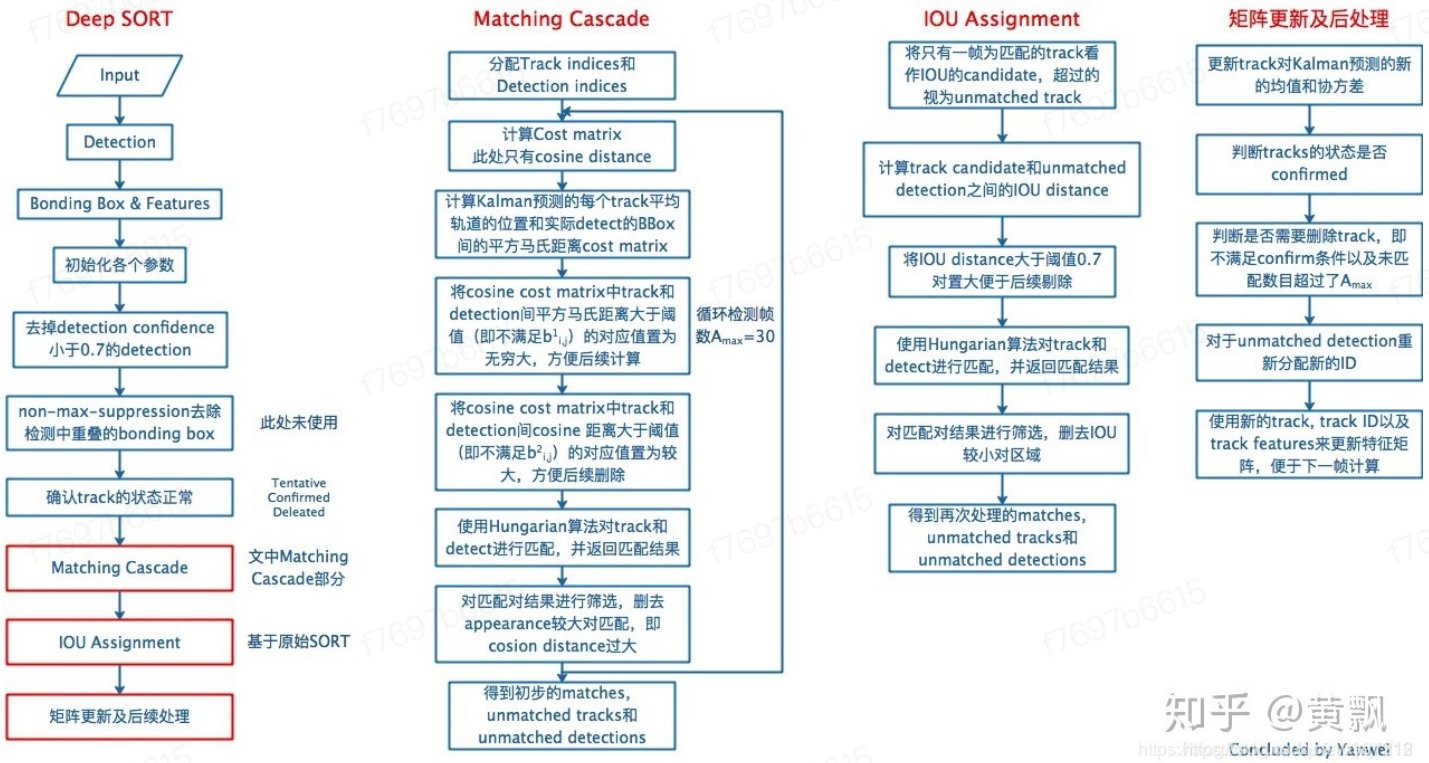
源码解读：https://zhuanlan.zhihu.com/p/90835266

- Kalman滤波+级联匹配+IoU匹配
- 级联匹配中代价矩阵由表观相似度和运动相似度加权得到
  - Re-ID行人重识别 提取表观特征，轨迹表观特征是过去100帧的简单平均，余弦距离计算相似度
  - Kalman过滤器预测值作为运动特征，马氏距离计算相似度
  - 匈牙利算法得到初步匹配结果
  - 丢帧最少的轨迹优先匹配，把轨迹置信度考虑进来了

- IoU匹配
  - IoU作为代价矩阵
  - 匈牙利算法进行匹配







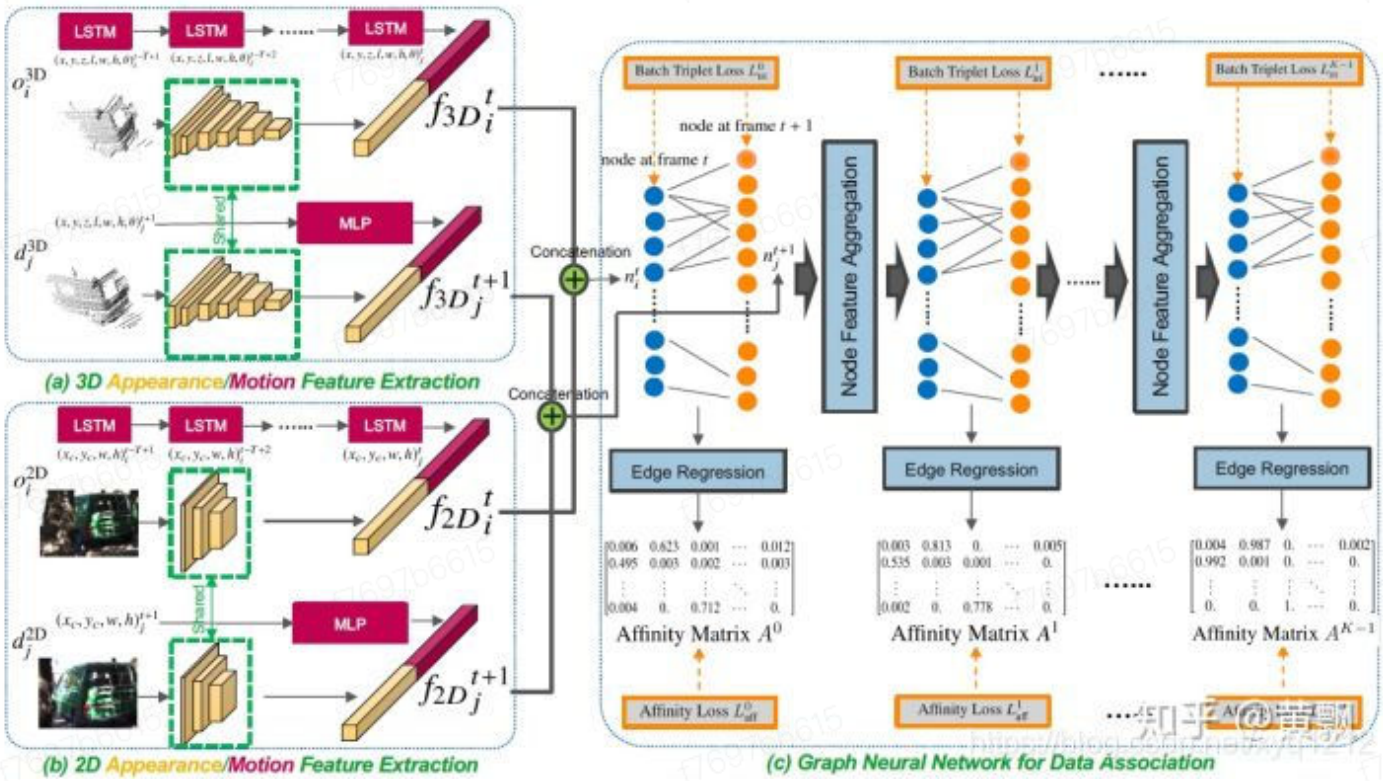
AB3DMOT：（3D Multi-Object Tracking: A Baseline and New Evaluation Metrics）

- Kalman滤波+匈牙利匹配，只使用匀速运动模型，代价矩阵为IoU
  - Kalman更新模块涉及到了方向校正的问题：当预测方向和识别方向相差大于π，那么修正预测结果（识别结果出错概率更大还是预测结果出错概率更大，修正哪个会更好一些？）
  - 更新模块根据贝叶斯规则，将匹配成功的识别结果和轨迹加权平均
- 创新
  - 将Kalman滤波器扩展到3D领域
  - 提供了3D MOT的评估工具
  - 提出了新的评估指标，考虑不同的轨迹置信度阈值
- 实验结论
  - 3D detector：影响最大的模块，对比3D 和2D识别器，结果是3D远优于2D，因此可以看出识别质量是追踪过程中最重要的
  - 2D&3D卡尔曼过滤器：3D优于2D，深度信息对追踪有帮助，对平坦的路段来说有必要吗?
  - 角速度：有角速度反而会降低效果，可能是因为车辆行驶过程中没有明显的角速度变化，因此加入这个维度的信息反而会带来噪声（那么在路口场景下，一定会有很多转弯、掉头等情况，是不是加入角速度会更好一些？）
  - 方向校正：加入后也会提高性能，如果考虑修改detection的角度会怎样?
  - IoU(min)：越大性能降低越明显
  - 新轨迹判定参数*min：小一些会提高精度并且减少FN数量，但是会加大IDS；大一些会会减少遮挡带来的问题，但是精度和准确的都会下降
  - 旧轨迹删除参数Agemax ： 它对IDS是没有影响的，但是与遮挡是有直接联系的
- 优势：速度

GNN3DMOT

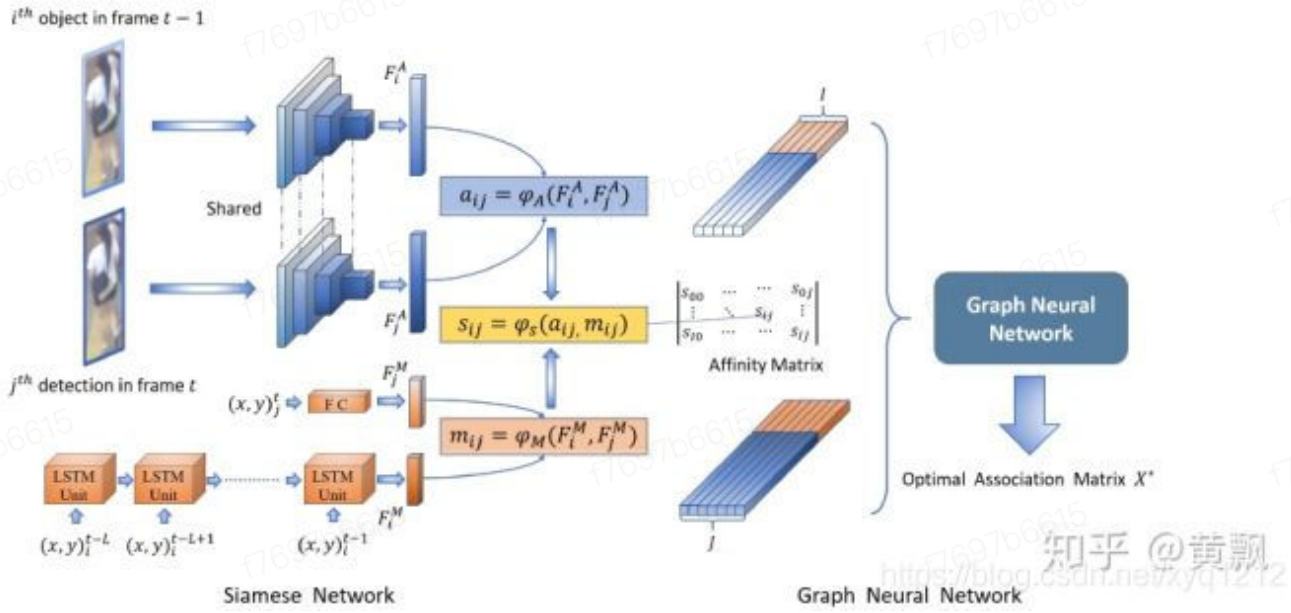
https://zhuanlan.zhihu.com/p/149244248

《Graph Neural Network for 3D Multi-Object Tracking with Multi-Feature Learning》



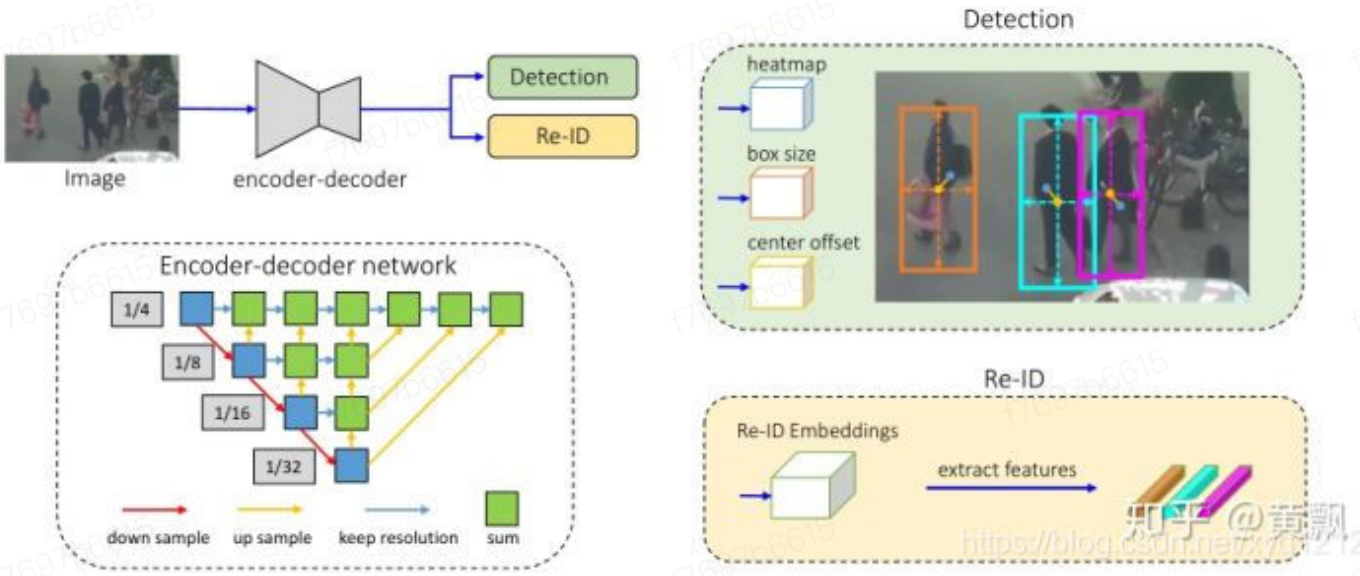
- LSTM/MLP获取运动特征、CNN获取表观特征
- 2D特征与3D特征融合
- 使用网络回归计算相似度
- 预测方面：结合关联矩阵和节点差异
- 损失函数考虑匹配对之间距离更小，非匹配对之间距离更大
- 网络更新：边-节点-全局更新（所有节点的特征均值和边权均值）-边更新，加入了一个全局变量在整个更新过程中进行调节

EDA_GNN



- 孪生网络获取表观相似度、LSTM预测位置，得到运动相似度，两个相似度结合构建相似度矩阵
- 基于消息传递机制，使用GNN网络更新节点特征，节点特征由表观特征和位置信息拼接得到

FairMOT



- 感觉和JDE类似，detection和Re-ID的结合，detection更注重class，而Re-ID是对相同class中个体的identity，所以联合训练还是存在一定的问题
- anchor-free框架（貌似Tracktor++和CentreTrackor也是anchor-free）；Re-ID更关注底层特征，因此使用特征层次融合的方式；Re-ID特征维度不宜过高的问题

二、论文链接：https://arxiv.org/pdf/2012.13755.pdf

三、论文解读：https://blog.csdn.net/qq_42575422/article/details/120

768722

1、为什么要做这个研究（理论走向和目前缺陷）？

之前的3D多目标跟踪的缺陷:

- 相似度计算基本都不考虑目标的几何和外观特征
- 也很少会把点云和图像特征融合在一块做
- 生命周期管理无一例外都是基于经验来设置一个固定参数

2、他们怎么做这个研究 （方法，尤其是与之前不同之处）？

主要创新就是加了三个可训练的模块：

1）特征融合模块：融合图像（maskrcnn）和点云（centerpoint）的特征，计算检测和跟踪的特征相似度。

2）距离组合模块：组合融合的 深度特征距离 和 马氏距离 作为相似度度量。

3）跟踪初始化模块：基于融合的深度（几何+表观）特征，决定新出现的目标是否加入跟踪队列。

3、发现了什么（总结结果，补充和理论的关系）？

效果比center point要好，但是用的可训练网络太多，文中未报告时间消耗。

四、论文详情：

摘要：提高跟踪准确度的关键点就是：数据关联以及生命周期管理。

本文提出由不同的可训练模块组成的概率、多模态、多目标跟踪算法。
首先，设计了一个可学习的**融合2D图像和3D激光点云表观特征**和几何信息的模块。
其次，设计了一个可学习的衡量检测与跟踪的 相似度（混合了马氏距离和特征距离）的模块。
再者，提出了一个可学习的“为没有匹配的检测 决定何时进行初始化”的模块。
效果在NuScenes数据集上很好。

1 引言

数据关联

之前的3D多目标跟踪的（数据关联）相似度度量都是

- 中心点的欧式距离
- 或3D框的马氏距离，

只是根据距离差异或者3D框的尺寸朝向差异来决定是否关联，完全不考虑表观特征和几何信息，这就导致卡尔曼滤波预测的位置信息等如果不准的话，准确率就大打折扣。

生命周期管理

跟踪的生命周期管理也很重要，会很影响FP和IDSwitch指标：
之前的方法：连续跟踪几帧后再加入跟踪队列，连续丢几帧后就从跟踪队列总删除。
本文提出利用即几何和表观特征决定是否加入跟踪队列。

2 相关研究

A.3D目标检测: 本文用的CenterPoint检测器

B.3D多目标跟踪：AB3DMOT（3DIOU），ProbabilisticTracking(马氏距离），CenterPoint（中心点欧氏距离）

C.结合特征和表观特征进行关联的方法：GNN3DMOT，PnPNet。

3 本文方法

三个可训练模块：

1）特征融合模块：融合lidar和图像特征

2）距离组合模块：组合深度特征距离和马氏距离作为相似度度量

3）跟踪初始化模块：基于其深度（几何+表观）特征决定新出现的目标是否加入跟踪队列

A.卡尔曼滤波

状态空间模型（同ProbabilisticTracking，匀速运动模型）：

$$\mathbf{s}_t = (x, y, z, a, l, w, h, d_x, d_y, d_z, d_a)^T,$$

但是量测用的是centerpoint的检测结果，跟一般的不同的是，这个检测结果除了正常的中心点3D位置+朝向+3D尺寸外还包含了两个额外参数（dx, dy），

可认为是相对上一帧同一目标中心点的位移，通过计算 中心点速度x两帧时间差 获取。

观测噪声矩阵和过程噪声矩阵同ProbabilisticTracking，也是从训练集中统计获取的。

跟踪模型的总体架构如下图：

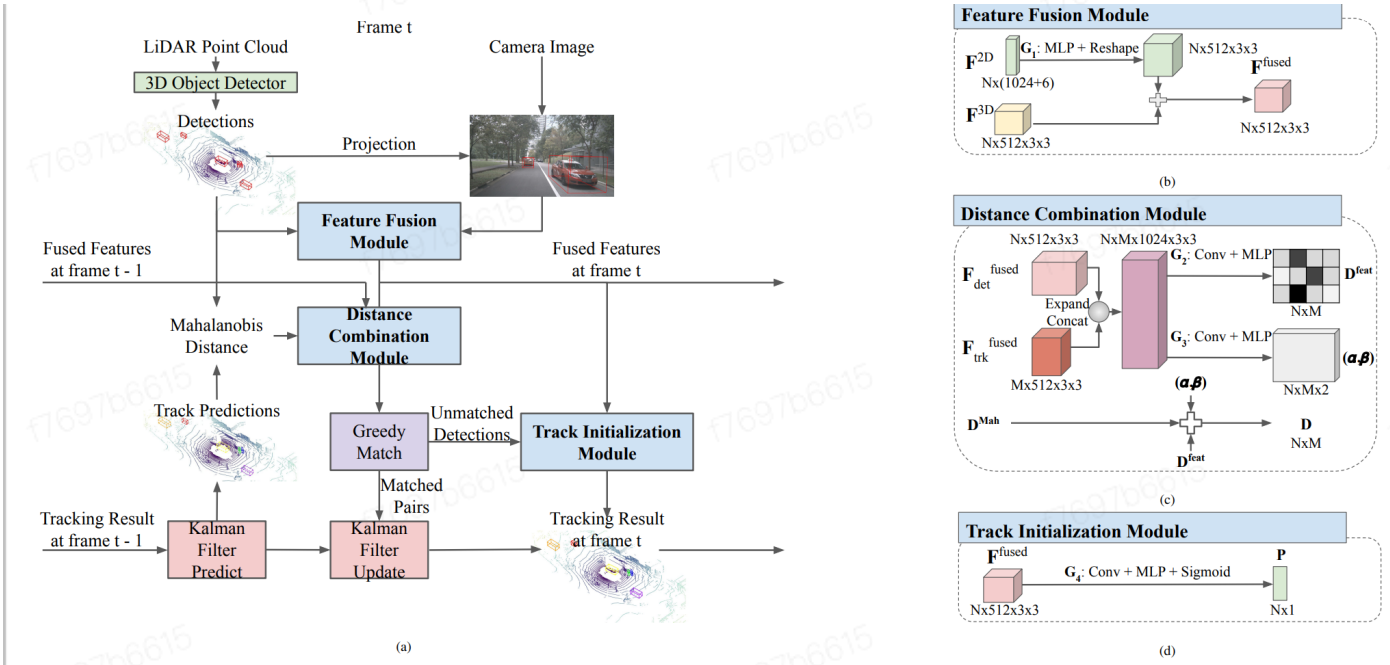


Fig. 1: Algorithm Flowchart. Sub-figure (a) depicts the high level overview of our proposed architecture, and (b)(c)(d) on the right illustrate the details of each neural network module. At frame t, We use a 3D object detector and extract the LiDAR and image features for each detected object. These features are fused by the **Feature Fusion Module**. The fused features of detections from time t and tracks from time $t-1$ are used in the **Trainable Distance Combination Module** to learn the combination of the deep feature distance and the Mahalanobis distance. We apply a greedy match algorithm on the combined distance for data association. Matched pairs are further processed by the Kalman Filter to refine the final object states. The **Track Initialization Module** determines whether to initialize a new track for each unmatched detection.

B.2D和3D特征的融合

①对于每个检测结果：
将其在世界坐标系下的2D位置（x,y）
转换到3D目标检测器的中间2D特征图的位置(xmap, ymap) in，
从这个特征图上提取到这个实例的512x3x3的点云特征。

②然后把3D检测框 映射到2D图像上，
再用预训练的Mask RCNN主干网络提取特征，
再经RoIAlign后得到一个1024维的向量，

这个向量再连接一个6维的one-hot向量（用于3D框映射到了哪个相机平面？），得到最终的2D特征。
最后用一个MLP将两个特征向量融合，得到最终的Nx512x3x3维的融合特征，其中N代表检测个数，
G1代表MLP+reshape, 公式如下：

$$\mathbf{F}^{fused} = \mathbf{G}_1(\mathbf{F}^{2D}) + \mathbf{F}^{3D}$$

C.距离组合模块

马氏距离和融合特征的距离如何权衡问题。如下公式：

$$\mathbf{D} = \mathbf{D}^{Mah} + \boldsymbol{\alpha}(\mathbf{D}^{feat} - (0.5 + \boldsymbol{\beta}))$$

alpha和beta是两个系数矩阵， Dfeat是前后帧融合特征距离，它们是用神经网络计算得到的，这个神经网络见框架图中的©。

1）融合特征距离计算公式表达：

$$\mathbf{D}^{feat.} = \mathbf{G}_2(\mathbf{F}_{det}^{fused}, \mathbf{F}_{trk}^{fused})$$

G2是一个卷积神经网络，通过将其视为一个二分类问题进行监督。
由于对于前后帧预测没有真实ground truth，只能选择离其中心点最近的gt框作为匹配的标注框，
如果前后帧的这个两个标注框的id相同并且两个检测中心的欧式距离小于2m，二分类的监督K = 0，否则K = 1。
监督过程的交叉熵损失公式：

$$L^{dist} = \mathbf{BCE}(\mathbf{D}^{feat}, \mathbf{K}),$$

2）组合系数alpha和beta

G3也是一个卷积神经网络：

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{G}_3(\mathbf{F}_{det}^{fused}, \mathbf{F}_{trk}^{fused})$$

训练这个网络的损失函数参考了PnPNet，用了max-margin（最大边缘）+contrastive对比损失训练。
对于一对组合系数（alpha,beta），训练的结果应使：
利用这对组合系数计算出来的 检测和正例跟踪的组合距离di
小于和负例跟踪的组合距离dj：

$$L_{i,j}^{contr} = \mathbf{max}(0, C^{contr} - (d_i - d_j))$$

每一帧训练的这个总的损失公式表达：

$$L^{contr} = \frac{1}{|\text{Pos}||\text{Neg}|} \sum_{i \in \text{Pos}, j \in \text{Neg}} L_{i,j}^{contr}$$

（这里感觉损失需要取反？）

其中Pos和Neg分别代表正确的和错误的跟踪匹配对。

（下面这一块没看太懂，这两个其它的最大边缘损失作用是什么？如何发挥作用？）

为了同时使用学习的融合距离D来拒绝在推断中匹配失败的外点。定义两个其它的最大边缘损失如下：

$$L^{pos} = \frac{1}{|\text{Pos}|} \sum_{i \in \text{Pos}} \mathbf{max}(0, C^{pos} - (T - d_i)),$$
$$L^{neg} = \frac{1}{|\text{Neg}|} \sum_{j \in \text{Neg}} \mathbf{max}(0, C^{neg} - (d_j - T)),$$

其中，C_pos和C_neg定义为固定边缘，以及T是固定的阈值来拒绝匹配失败的外点。

总的损失函数定义如下：

$$L^{coef} = L^{contr} + L^{pos} + L^{neg}$$

实际中，选择T = 11，C_contr = 6，C_pos = C_neg = 3。

最后关联时用的贪心算法进行匹配。

D.跟踪初始化模块

本文将是否跟踪初始化（将检测加入跟踪队列）视为一个二分类任务，用一个卷积神经网络（G4）预测是否进行跟踪初始化：

$$\mathbf{P} = \mathbf{G}_4(\mathbf{F}^{fused}).$$

训练此网络用交叉熵损失:

$$L^{init} = \mathbf{BCE}(\mathbf{P}, \mathbf{P}^{target})$$

如果有GT和检测匹配的话，监督标签P_target=1，否则为0。推断阶段，P>0.5就进行初始化。

4 实验

A.数据集：NuScenes

B.评估指标：AMOTA(在不同召回率下的平均跟踪准确率，参考NuScenes挑战赛）

C.量化结果：只在NuScenes验证集上做了实验,实验结果如下：

和单模态跟踪模型的对比：

TABLE I: Evaluation results on the NuScenes [9] validation set: evaluation in terms of overall AMOTA and individual AMOTA for each object category, in comparison with the baseline methods. In each column, the best-obtained results are typeset in boldface. (*Our implementation by using [2]'s open-source code and [1]'s object detection results.)

Tracking method	Modalities	Input detection	Overall	bicycle	bus	car	motorcycle	pedestrian	trailer	truck
AB3DMOT [7]	3D	MEGVII [13]	17.9	0.9	48.9	36.0	5.1	9.1	11.1	14.2
ProbabilisticTracking [2]	3D	MEGVII [13]	56.1	27.2	74.1	73.5	50.6	75.5	33.7	58.0
CenterPoint [1]	3D	CenterPoint [1]	65.9	43.7	80.2	84.2	59.2	**77.3**	51.5	**65.4**
ProbabilisticTracking [2]*	3D	CenterPoint [1]	61.4	38.7	79.1	78.0	52.8	69.8	49.4	62.2
Our proposed method	3D	CenterPoint [1]	67.7	47.0	81.9	84.2	66.8	75.2	**53.5**	**65.4**
Our proposed method	2D + 3D	CenterPoint [1]	**68.7**	**49.0**	**82.0**	**84.3**	**70.2**	76.6	53.4	**65.4**

目前在Nuscenes排行榜排名（本文方法第11名）截图：

Method					Metrics									
Date	Name	Modalities	Map data	External data	AMOTA	AMOTP (m)	MOTAR	MOTA	MOTP (m)	RECALL	GT	MT	ML	FAF
		Any	All	All										
2020-11-11	AlphaTrack	Camera, Lid	no	no	0.693	0.585	0.800	0.576	0.306	0.723	17081	5560	1744	58.126
2021-05-12	MLPMOT	Camera, Lid	no	no	0.683	0.490	0.775	0.554	0.311	0.728	17081	5600	1627	66.521
2021-05-20	CBMOT+	Camera, Lid	no	no	0.681	0.528	0.772	0.550	0.305	0.720	17081	5479	1563	66.649
2020-08-20	Noah Octopus Trac	Lidar	no	no	0.679	0.562	0.808	0.572	0.305	0.709	17081	5630	1619	54.020
2020-10-31	EagerMot	Camera, Lid	no	no	0.677	0.550	0.793	0.568	0.335	0.727	17081	5303	1842	56.849
2021-03-05	CB-MOT	Camera, Lid	no	no	0.676	0.518	0.765	0.539	0.300	0.711	17081	5420	1654	67.734
2020-07-18	MCMOT	Camera, Lid	no	yes	0.666	0.644	0.816	0.556	0.333	0.693	17081	5533	1581	51.065
2021-05-11	ccca	Lidar	no	no	0.661	0.555	0.835	0.555	0.290	0.668	17081	5388	1610	44.123
2021-05-26	OGR3MOT	Lidar	no	no	0.656	0.620	0.797	0.554	0.303	0.692	17081	5278	2094	56.002
2020-07-03	Tracker	Lidar	no	no	0.656	0.570	0.774	0.543	0.298	0.699	17081	5383	1832	54.699
2021-05-26	StanfordIPRL-TRI-M	Camera, Lid	no	no	0.655	0.617	0.785	0.555	0.318	0.707	17081	5494	1557	57.399

Team: StanfordIPRL-TRI-Probabilistic

Authors: Hsu-kuang Chiu, Jie Li, Rares Ambrus, Jeannette Bohg

Affiliation: Stanford University, Toyota Research Institute

Description: [ICRA2021][Probabilistic 3D Multi-Modal, Multi-Object Tracking for Autonomous Driving] with 2D object detection enhancement

Project url: n/a

Paper url: https://arxiv.org/pdf/2012.13755.pdf

More:

Object Class	AMOTA	AMOTP (m)	RECALL	MOTAR	GT	MOTA	MOTP (m)	MT	ML	FAF	TP	FP	FN	IDS	FRAG	TID (s)	LC
bicycle	0.469	0.895	0.553	0.716	2186	0.392	0.286	77	65	23.533	1199	341	977	10	6	0.912	1
bus	0.713	0.677	0.733	0.914	1701	0.668	0.311	61	29	7.441	1243	107	455	3	5	0.341	0
car	0.83	0.388	0.853	0.824	68518	0.698	0.257	3532	803	175.184	58072	10236	10093	353	254	0.268	0
motorcycle	0.631	0.561	0.706	0.772	1945	0.542	0.299	86	37	23.037	1365	311	572	8	9	0.589	0
pedestrian	0.741	0.507	0.786	0.821	34010	0.629	0.295	1398	375	92.114	26079	4672	7278	653	415	0.418	0
trailer	0.657	0.753	0.683	0.828	2566	0.563	0.423	86	47	29.395	1746	301	813	7	12	0.729	1
truck	0.546	0.54	0.637	0.619	8639	0.394	0.354	254	201	51.086	5495	2093	3135	9	16	0.302	0

和多模态跟踪模型的对比：

TABLE II: Evaluation results on the NuScenes [9] validation set: evaluation in terms of overall AMOTA and the AMOTA for car, in comparison with GNN3DMOT [5], PnPNet [4]. GNN3DMOT [5] only reports the overall AMOTA, and PnPNet [4] only reports the AMOTA for cars. Note that each method uses a different 3D object detector, and that could affect the tracking accuracy significantly. (*GNN3DMOT [5] renames AMOTA as sAMOTA in their paper.)

Tracking method	Modalities	Overall	car
GNN3DMOT [5]*	2D + 3D	29.84	-
PnPNet [4]	2D + 3D	-	81.5
Our proposed method	2D + 3D	68.7	84.2

D.消融研究

围绕三个新加的可训练模块展开对比分析：

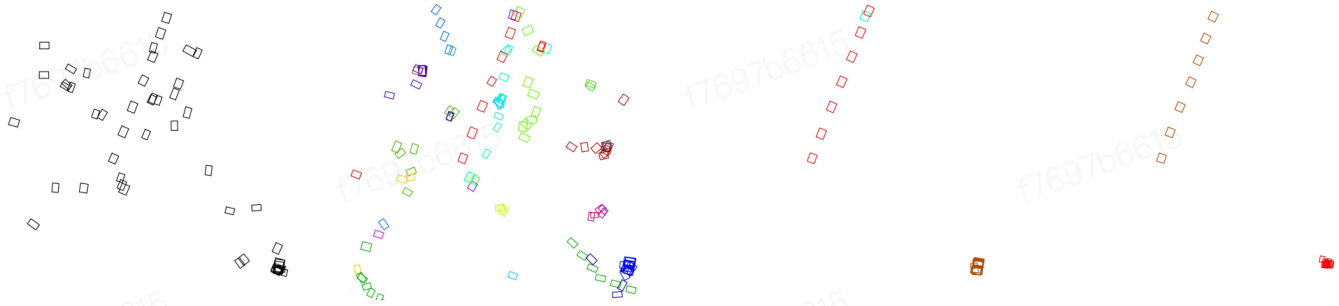
TABLE III: Ablation results for the validation set of NuScenes [9]: evaluation in terms of overall AMOTA and individual AMOTA for each object category, in comparison with variations of our proposed method. All variations use CenterPoint [1]'s object detection results as input. In each column, the best-obtained results are typeset in boldface.)

Tracking method	Modalities	Overall	bicycle	bus	car	motorcycle	pedestrian	trailer	truck
Distance Combination Module only	3D	67.1	46.3	81.9	84.2	63.8	74.9	**53.5**	**65.4**
Track Initialization Module only	3D	66.2	45.1	78.4	84.2	66.6	75.1	52.7	61.2
Our proposed method	3D	67.7	47.0	81.9	84.2	66.8	75.2	**53.5**	**65.4**
Distance Combination Module only	2D + 3D	67.6	46.5	**82.0**	**84.3**	65.4	76.3	53.1	**65.4**
Track Initialization Module only	2D + 3D	67.4	48.6	80.4	81.6	68.4	75.3	53.3	64.5
Our proposed method	2D + 3D	**68.7**	**49.0**	**82.0**	**84.3**	**70.2**	**76.6**	53.4	**65.4**

2D对于3D是锦上添花

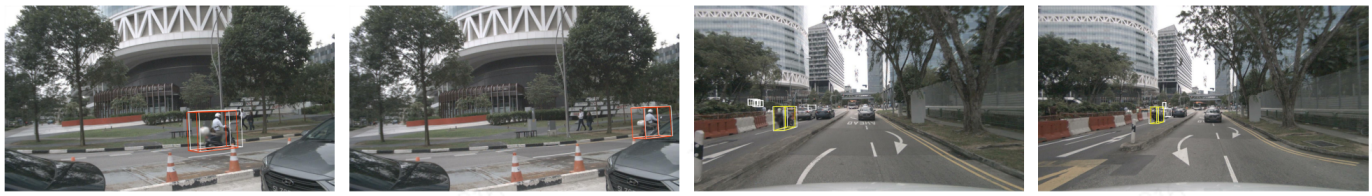
精细的数据关联起决定性作用

比较好的结果可视化：



(a) Input detection (b) CenterPoint [1] (c) Our proposed method (d) Ground-truth

Fig. 2: Bird-eye-view tracking visualization of motorcycles. We plot the bounding boxes from every frame of the same driving sequence in each sub-figure. Different colors represent different tracking ids in the tracking results, and indicate different instances of objects in the ground-truth annotation. (a): input detection bounding boxes provided by CenterPoint [1]'s object detector. (b): CenterPoint [1]'s tracking result. (c): our proposed method's tracking result. (d): ground-truth annotation. Our tracking result has significantly fewer false-positive bounding boxes compared with CenterPoint [1]'s result. Our tracking result is also closer to the ground-truth annotation.



(a) Sequence 0, Frame 1 (b) Sequence 0, Frame 2 (c) Sequence 1, Frame 28 (d) Sequence 1, Frame 29

Fig. 3: Tracking visualization of motorcycles projected to camera images. (a), (b) are two consecutive frames in Sequence 0. (c), (d) are from Sequence 1. The color boxes are the tracking results. Different colors indicate different tracking ids. The white boxes represent the detections. Our model can accurately track the motorcycles in the red bounding boxes in Sequence 0 and the yellow bounding boxes in Sequence 1. In Sequence 0, our **Distance Combination Module** learns to generate a larger positive $\alpha = 2.594$ value, because the appearance features seem to provide strong information to match the detected motorcycles across these consecutive frames. In Sequence 1, our model generates a smaller $\alpha = 1.802$, potentially because the bounding boxes are smaller and the image is more blurred. Our **Track Initialization Module** also correctly decides not to initialize new tracks for the false-positive detections in (c) Sequence 1 Frame 28 .

5 结论

未来的工作：

1）考虑加入更多的模态信息生成融合特征，比如地图信息。

2）测试更好的检测算法。

3）改进运动模型。

4）考虑利用可微分的滤波框架，利用递归滤波的算法先验，对运动和观测模型进行端到端的微调。

6 补充

作者提出了曼哈顿距离作为关联距离和数据驱动计算卡尔曼滤波的协方差初始值，提高了6个点，他觉得是这两个的作用。

但把距离改成x和y方向的L2距离，协方差用默认的，用对匈牙利算法，AMOTA提高了5个点，所以数据驱动的协方差矩阵有一定作用，1个点左右。