

# 如何评价图森未来的AI Day?

[关注问题](#)[写回答](#)

自动驾驶 图森未来

关注者

96

被浏览

44,996

## 如何评价图森未来的AI Day?

昨天图森开放AI day，秀了一下肌肉，为什么网上没人讨论呀？个人感觉有些比较流行的话题还都是用上的，有大佬能从自己的专业角度分析一下嘛？

乘大势，驭未来——图森未来首届AI Day  
完整回放\_哔哩哔哩\_bilibili  
[www.bilibili.com/video/BV1Em4y1u7P7/?...](http://www.bilibili.com/video/BV1Em4y1u7P7/?...)

[关注问题](#)[写回答](#)[邀请回答](#)[好问题 2](#)[添加评论](#)[分享](#)[收起](#)[查看全部 3 个回答](#)

刘斯坦



绘画等 3 个话题下的优秀答主

[+ 关注](#)

305 人赞同了该回答

前一阵子图森未来搞了他们的第一次AI Day。大家都知道，自从特斯拉通过AI Day引领了技术潮流，很多公司都开始搞AI Day或者类似AI Day的活动，比如Cruise，小鹏，Comma AI。还有一些搞得比较差有一点骗投资性质的，就不点名了。。。每一次观看，都非常令人愉悦，很有启发。

现在又多了个图森。看完了才发现他们原来很多东西都发了论文了，去搜一下挂了Naiyan Wang名的论文，看得我目瞪口呆。发的论文和做的产品高度一致的，大概只有这一家。

这个AI Day看下来还是收获满满的，一是感知部分基本等于重要论文串讲，效率高，讲得还清楚。二是定位部分的那个兄弟，讲了很多卫星定位的原理，这个对我属于新知识，毕竟一直把GNSS当黑盒用的，他把里面的原理讲清楚了，舒适！三是演示了他们是如何用NeRF做闭环仿真<sup>Q</sup>的，这应该是业界第一家展示实际应用流程的。视频在这里：

乘大势，驭未来——图森未来首届AI Day  
完整回放\_哔哩哔哩\_bilibili  
[www.bilibili.com/video/BV1Em4y1u7P7/?...](http://www.bilibili.com/video/BV1Em4y1u7P7/?...)



## 完全稀疏的感知栈

整个感知栈是稀疏的，也是冗余的。就是雷达和相机两套系统都能独立开，



## 视觉的稀疏检测

完全稀疏的感知栈是最大亮点。基本思想是从二维检测出发，一步步的走向三维检测。

从二维检测出发是很聪明的做法。稀疏框架最大的风险就是漏掉目标，因为他是先提取一些“种子”目标作为稀疏的感知焦点，然后去搜集和这些感知焦点相关的信息最后回归出三维检测框。所以种子目标必须要覆盖所有目标，鉴于现在二维目标检测<sup>Q</sup>基本属于“已经解决”的问题，这些初始的种子目标是值得信任的。

得到二维检测的目标之后就是接入一个Transformer去生成query，通过稀疏注意力去各图像上收集信息并回归三维目标：

## 如何评价图森未来的AI Day?

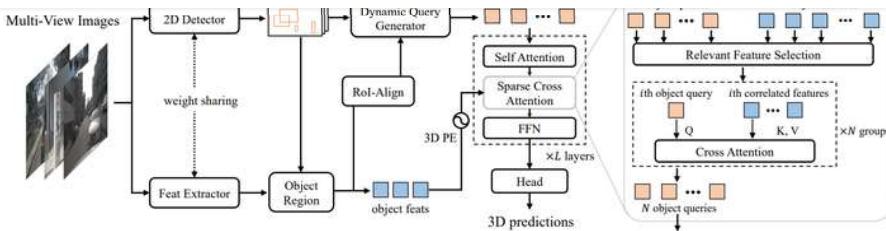
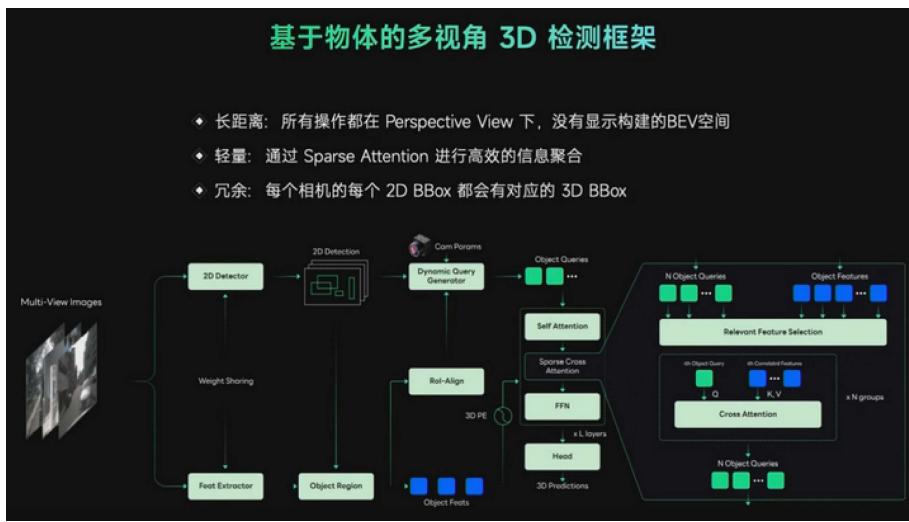
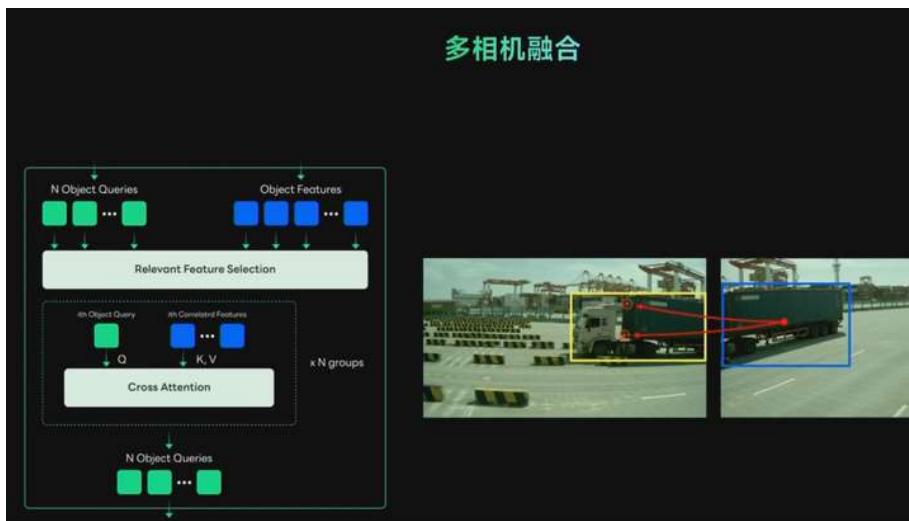


Figure 2. The framework of the proposed MV2D. Given the input multi-view images, image feature maps are extracted by a feature extractor. Meanwhile, a 2D detector is used to obtain per-view 2D detection results. Dynamic query generator takes object features, 2D detection boxes and camera parameters as input to initialize a set of object queries. RoI-Align is applied to the object regions to obtain the fixed length object features for query generator. All the features fallen in the object regions are decorated with 3D PE (3D position embedding) [26], then the object queries and object features are input to a decoder to update query features. Compared to vallina transformer decoder, the decoder in MV2D employs sparse cross attention where each object query only interacts with its relevant features. Lastly, a prediction head is applied to the updated object queries to generate 3D detection results.

相关论文插图

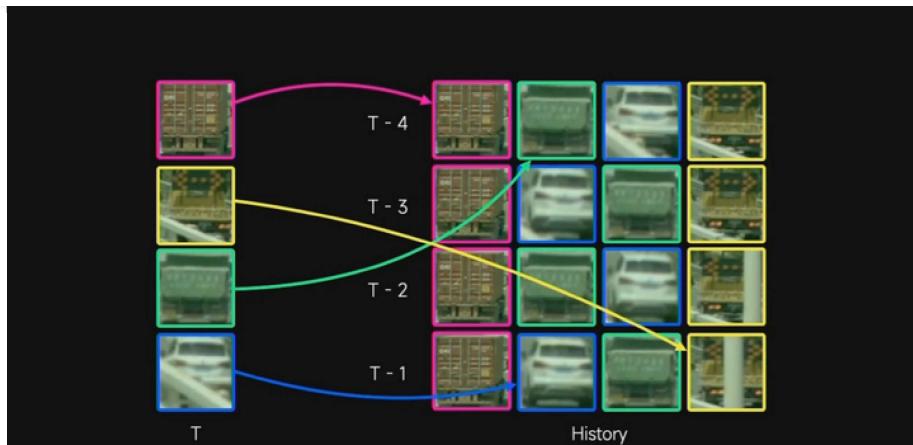


发布会插图



图森未来的整个技术栈还有一个特点是极度压榨目标跟踪带来的丰富信息积累。从感知栈，到 NeRF，到自动标注，目标跟踪都是核心。感知栈涉及的部分就是时序融合了。道理也很简单，就是对目标进行跟踪后收集同一目标在不同帧的信息进行融合：

## 如何评价图森未来的AI Day?



### 激光雷达的稀疏检测

二维图像的稀疏化，讲道理还比较方便，毕竟图像本身是致密信息。但激光雷达从一开始就是稀疏信息，所以不得不从一开始就引入稀疏计算，他们直接用稀疏卷积：



之后的模型是基于一种类似 clustering<sup>Q</sup> 的逻辑进行的，先把稀疏的点云<sup>Q</sup>按照普通的网格聚成团，然后加几层Attention根据新发现的语义特征重新聚类<sup>Q</sup>：

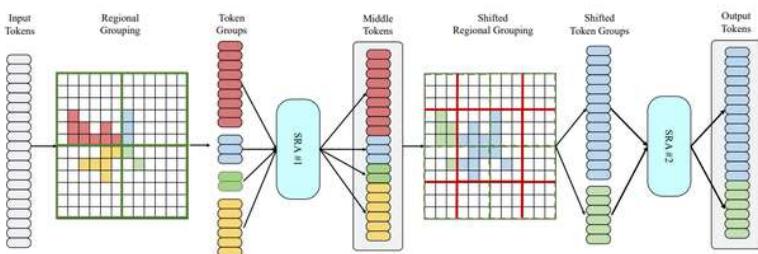
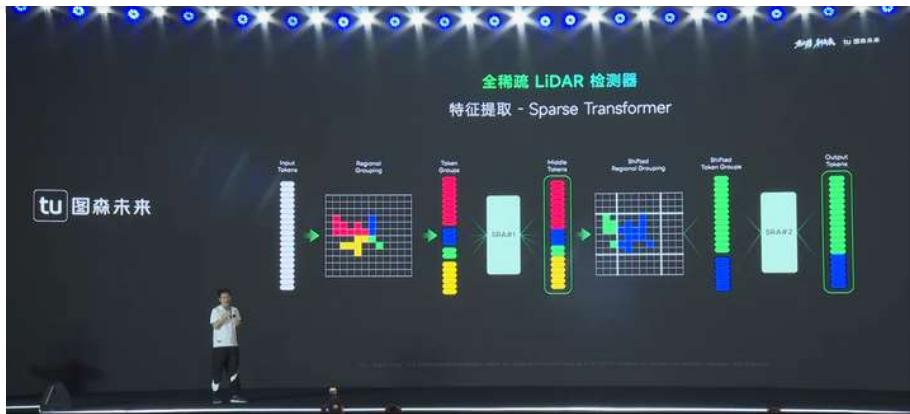


Figure 3. Computation of an example block in SST. For an incoming set of tokens, *Regional Grouping* first groups them according to the partitions of regions (in Sec. 4.2). Second, *Sparse Regional Attention* (SRA) deals with each group of tokens separately (Sec. 4.3). Third, the tokens are grouped another time according to *Region Shift*, and a second SRA processes the new groups of tokens (Sec. 4.4). These three steps complete the computation of a block.

相关论文插图

这是 @王峰 讲的：

## 如何评价图森未来的AI Day?



为了获得更精确的目标，还进行了二阶段的refinement<sup>Q</sup>。激光雷达目标只能看到靠近车辆的那层壳，看不到尾巴，这一直是个问题。这个论文就是hallucinated了一些内部的点云，神经网络<sup>Q</sup>最擅长的事情：hallucination，属于灵魂搭配了这是：

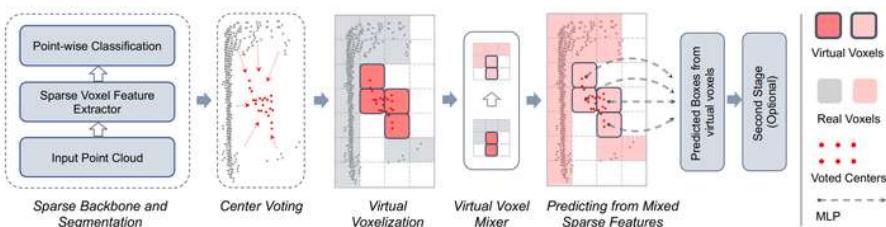
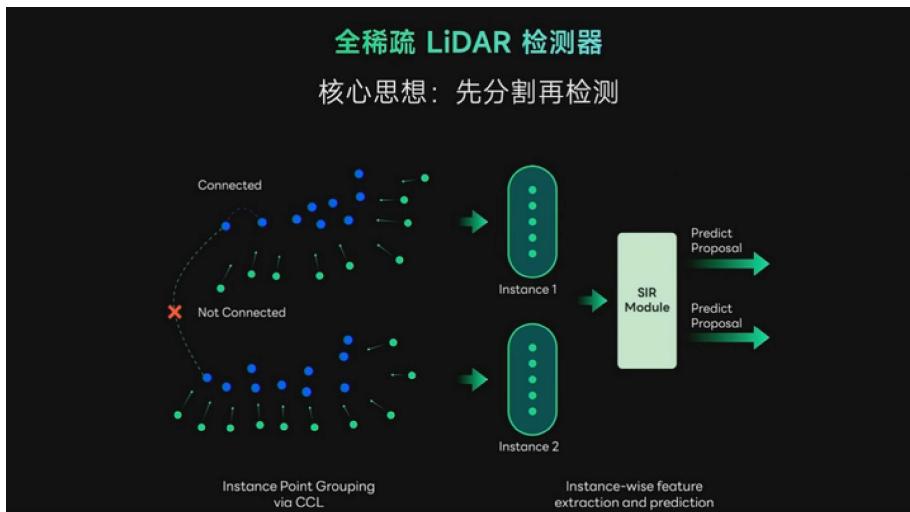


Fig. 2. Overview of FSDv2. The components before virtual voxelization is the same as FSDv1. Afterward, we employ virtual voxelization on both the original real points and the voted centers. Then we mix the features of virtual voxels (the four outlined red voxels) and real voxels (gray) via a Virtual Voxel Mixer for enhancement. We use light red to show the features after mixing. Eventually, predictions are made from the virtual voxels.



### 没有Corner Case和远距离感知

没有Corner Case的claim来自这里，对于没有Corner Case这个说法，理解可以有很多种。就比如这个高速公路上有人（或动物）横穿马路的“Corner Case”或者下雨什么的，我也觉得这些早就不能算Corner Case了。

图森未来对这种情况的处理算是另一种大力出奇迹吧，就是远距离感知。卡车的笨拙动力特性，确实需要对前方路况进行极早的预判，说实在的，其实卡车这种东西就不应该由由人类去开。这么庞大的物体要安全的跑高速，其实已经超过人类的能力范围了。我们姑且简单的理解为图森的做法是用望远镜远远地看前方500m吧。

我们的世界就这么点大，同一个异常物，你从多个角度长时间反复盯着看，总归能看到。一帧丢了算是个高概率事件，好几个摄像头加起来几百上千帧都检测不到，那是不可能的。所以高度的信息冗余+长时间的观察+望远镜放大，我估计就是只老鼠都逃不掉。

其实我觉得到了卡车这个级别的自动驾驶，以后一定会出现真正的超视距感知，就是放无人机，飞到前面后面把方圆<sup>Q</sup>一公里都看清楚。那就真的是高枕无忧了。

## 如何评价图森未来的AI Day?

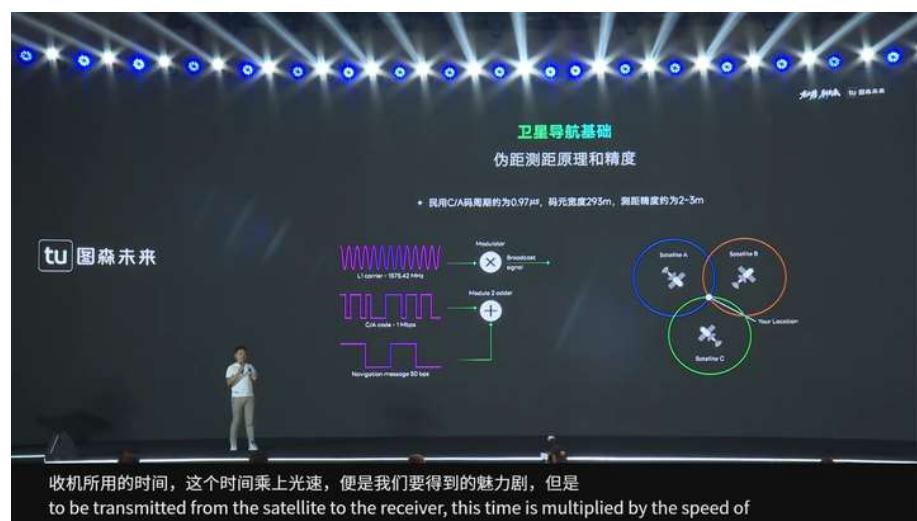


@Naiyan Wang 讲解各种冗余：

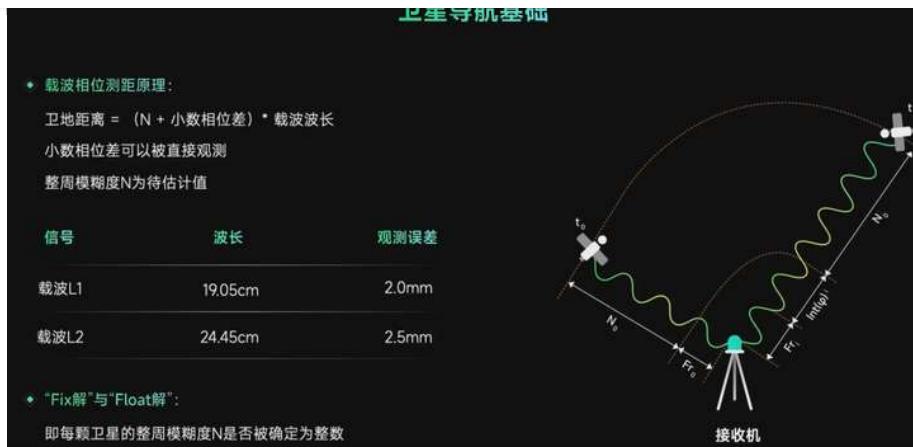


## 定位系统

讲定位系统的哥们儿等于是上了个卫星导航<sup>Q</sup>入门课的样子，本人听得很专心，以前只是大概知道差分GPS大概运作原理，这下子爽了：



## 如何评价图森未来的AI Day?



这里讲的是定位使用的粒子滤波<sup>Q</sup>，演讲没有讲细节。经乃岩提醒来自于这篇论文：

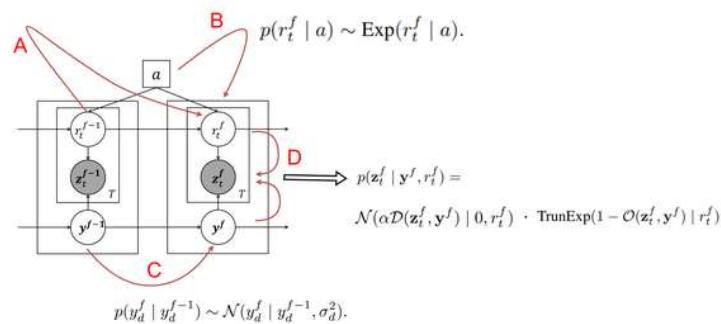
Naiyan Wang - Ensemble-Based Tracking:  
Aggregating Crowdsourced Structured Time Seri...  
[winsty.net/ebt.html](http://winsty.net/ebt.html)

稍微学习了一下这篇论文，整个概率框架总结如下：

use a Gamma distribution to model it:

$$p(r_t^f | r_t^{f-1}) \sim G\left(r_t^f | k, \frac{r_t^{f-1}}{k}\right), \quad (10)$$

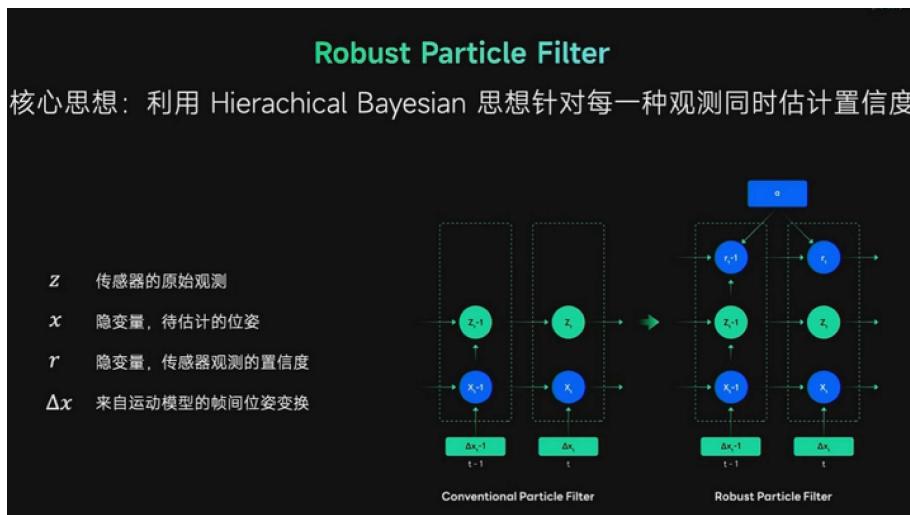
where  $k$  is a model parameter. The choice is deliberate since  $E[r_t^f] = r_t^{f-1}$ .



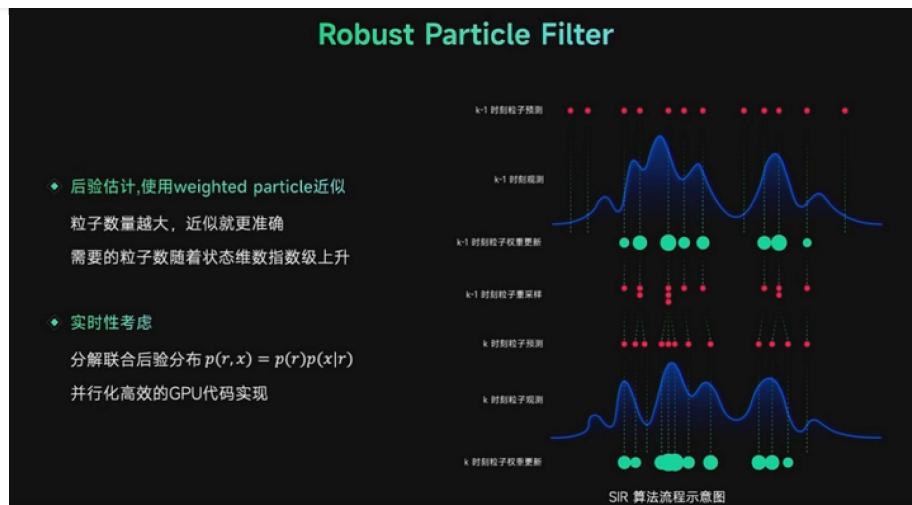
图中 $r$ 是不确定度， $z$ 是测量值， $y$ 是未知的真实值， $a$ 是一个先验概率， $f$ 是帧的编号。

A是不确定度的传播概率，用Gamma分布建模。B是不确定度的先验概率，用一个指数分布<sup>Q</sup>建模。C是真实值的传播概率，D是已知 $y$ 和 $r$ 的情况下测量得到 $z$ 的概率，用两个概率分布一起建模。

总之，对A, B, C, D建模完毕，现在情况是已知 $y$ ，能求 $z$ 的概率，但我们需要的是已知 $z$ 去估计 $y$ ，同时 $y$ 又是未知，这里就套用贝叶斯定律<sup>Q</sup>并使用采样的方法估计 $y$ ，同时考虑到采样结果的时序传播，使用粒子滤波对采样得到的 $y$ 进行时序传播：



## 如何评价图森未来的AI Day?



最后还谈了如何将GNSS因为相位关系导致的周期性, 多峰的概率分布整合到粒子滤波中去, 可见他们对GNSS设备Raw Signal的应用真是压榨到了极致。也算是开了眼界了。

## 路径规划

挺inspiring的, 我现在就在做基于神经网络的路径规划, 对我当前的工作很有帮助, 说多了容易暴露, 就不展开了, 自己去看吧。N阶交互也算是趋势了:



## NeRF

好, 重头戏来了, NeRF。NeRF最近非常热, 大家都知道, 在意大利的时候, 很多搞NeRF的博士都不知道为什么我一个搞自动驾驶<sup>Q</sup>的对NeRF感兴趣, 我跟他们解释完了才知道。但比较失望的是, NeRF用于自动驾驶, 几乎是至今唯一实用的地方? 但着眼于城市环境动态环境的NeRF极少, 都是在室内静态环境搞来搞去。说实在的, 那些NeRF对我而言都没法用啊, 汽车行驶在路面上, 很多动态物体, 特斯拉展示的NeRF里全是鬼影, 这没法用。。。而且NeRF对于缺乏信息的部分喜欢hallucinate<sup>Q</sup>一些乱七八糟的玩意, 和ChatGPT一本正经的胡说八道不同, NeRF产生的hallucination看起来根本就是错的。

比如图森展示的这一段行驶环境, 道路两旁的树(应该是树吧?)变成了这幅像雨像雾又像风的样子:

## 如何评价图森未来的AI Day?

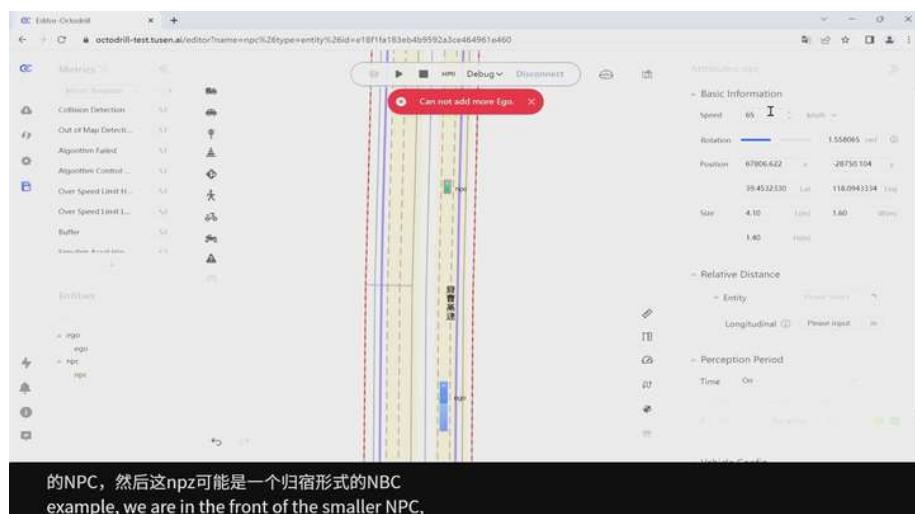


但他这个NeRF确实是可用的，因为其实汽车行驶的时候对两边的树其实并无所谓，只要路面结构清楚就够了，从展示的效果看，路面结构还是很清晰的。另外图森利用了他们在目标跟踪方面的技术优势，直接把动态目标抠出来了，然后基于目标跟踪+物体轨迹重建+基于surface的隐式表达<sup>Q</sup>重建物体，这些工作都是发表了论文的，太多了就不列举了。。。

用NeRF重建静态环境，用目标跟踪收集物体信息重建物体，这些重建出来的物体就变成asset<sup>Q</sup>存起来随时调用。

好了现在问题来了，为什么要用NeRF重建环境？主要目的还是闭环仿真。如果在运行过程中遇到奇特的场景出问题了，相关模块肯定要定位问题，改进算法。那么问题来了，出问题的场景一般是非常少见的，很多场景还非常的tricky，很难检测。那么最简单的做法，就是放到原来出问题的场景重新跑一遍。但原始数据只是传感器数据，如果根据改进的算法，车子的行为产生了偏差（几乎是一定的，否则我改进干嘛。。。），那么原始数据<sup>Q</sup>就没用了。所以要把车子穿越到原始场景重新跑一下，一般是用游戏引擎<sup>Q</sup>或者Carla之类的仿真平台重建然后跑一下仿真，这样的问题是太假了，毕竟是渲染。

现在图森的方案是静态环境的NeRF+这个环境编辑器<sup>Q</sup>：

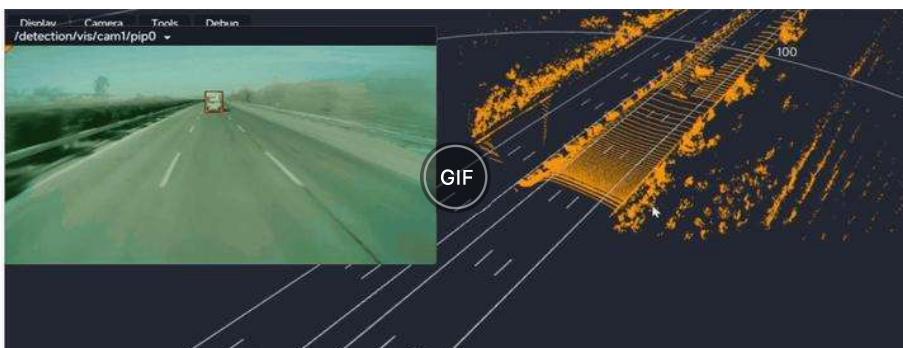


+动态物体资源库：

## 如何评价图森未来的AI Day?



设置好环境，把物体摆上去之后，进行仿真：



怎么说呢，能把这一套东西跑通，确实是牛逼，这里面需要投入的资源和精力着实难以估量。重建的效果，算是凑合着用吧，游戏引擎是假，但NeRF也有一些缺陷，不过今后如果有新的进展，应该可以很容易的整合进去，关键是这个infrastructure建起来了，之后的改进都好说。找几个博士合作，把树叶不规则物体之类的问题解决一下应该不是难事。

### 自动标注

自动标注，他们同样是把图森对目标跟踪的技术优势发挥到了极致，一切都在这篇论文里“一帧检测，永不丢失”，别细看了，反正就是**目标跟踪**<sup>9</sup>给你搞到极致：

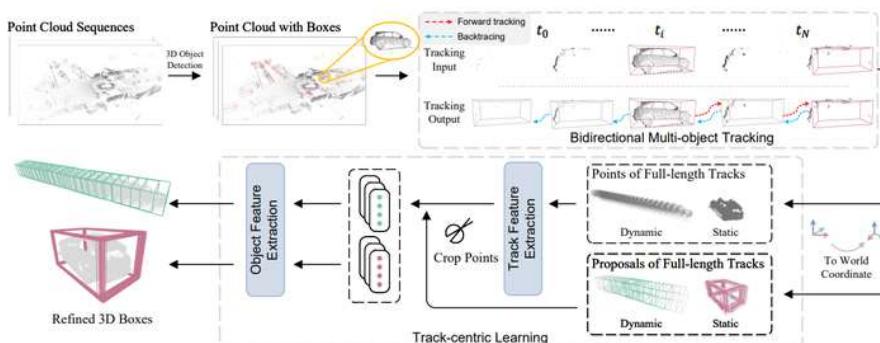
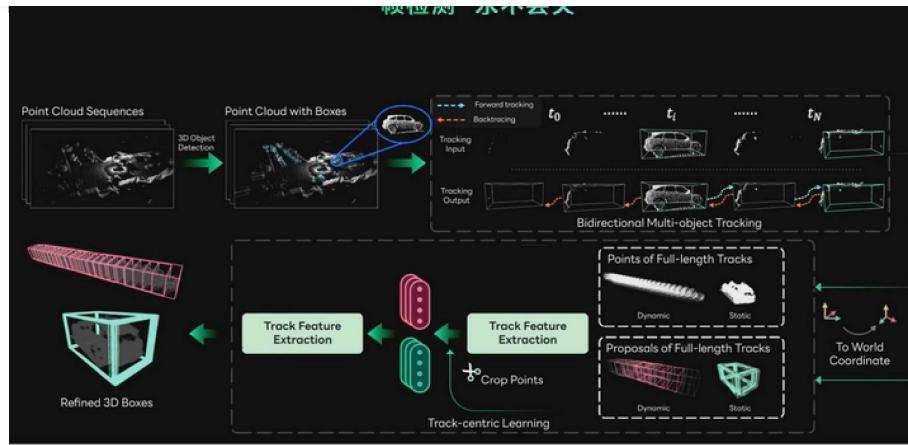


Figure 2: Overall architecture of CTRL. (1) A base detector is utilized to generate basic detection results. (2) In the bidirectional tracking module, a forward tracking process is applied first to fill missing boxes and extend the track to the future, indicated by the red arrow. Then we backtrace to the start frame and extend the track to the past, indicated by the blue arrow. (3) The bidirectional extended tracks are sent into the track-centric learning module for refinement.

## 如何评价图森未来的AI Day?



最后他们的成果是标注效果超越人类，这个claim看起来很吓人啊，但其实没有的。根据我实际观察，就目标检测而言，大型神经网络的检测能力超越人类并不奇怪：



主要是这是目标检测任务，相对简单，算是“已经解决”的问题，如果你考虑语义分割<sup>Q</sup>任务，可能还是需要人类参与的。

### 一点碎碎念

这个AI Day我觉得比小鹏的有诚意，展示了一个非BEV<sup>Q</sup>网络全稀疏的做法。爽是蛮爽的，不过这套做法不太通用。全稀疏对加速器不友好，不是标准模型<sup>Q</sup>，部署起来比较费劲，你看图森还专门得写插件，中间的稀疏注意力也麻烦得很。当然，有一个工程能力强的团队，这都不是事。

此外全稀疏也意味着对全局信息的舍弃，虽然第二个阶段的检测主要目的是确定目标的具体形状和三维位置，可能和周围的信息关系不大，但毕竟还是丢掉了一些全局信息。众所周知，神经网络比较依赖于context<sup>Q</sup>，比如，回归物体的尺寸和位姿是不是可以从四周的context中找到一些hint？当然，我们也可以将聚焦于局部能让神经网络从真正有意义的信息中回归出结果，而不是通过环境的context“作弊”。

我估计全稀疏还有一个原因是货车太大了，传感器多，覆盖空间广，这导致现在流行的致密信息前融合变得不现实。稀疏化可能是不得已而为之。

如此All in NeRF，风险挺大。不过那一整套闭环仿真系统挺牛逼的，如果能把接口做好，可以做到随时把NeRF换成传统三维重建<sup>Q</sup>或仿真器，那其实NeRF并不是关键，那套仿真系统的Infra才是关键。倒是可以一开始设计的时候就想好可替换的问题，应该不难。

有一些重要的信息都missing了，没谈到。比如车道线啥的都没谈到，TuSimple Lane数据集还是个Benchmark呢，我还玩过，可惜没有车道线识别的相关信息……比如数据闭环是怎么做的也没谈到，虽然大家做的都差不多，但围观友商找到的奇奇怪怪的corner case<sup>Q</sup>也是很有娱乐性的……

## 如何评价图森未来的AI Day?

了，估计特斯拉来了也做不到这么好。

Object as Query: Lifting any 2D Object  
Detector to 3D Detection  
[arxiv.org/abs/2301.02364](https://arxiv.org/abs/2301.02364)



Embracing Single Stride 3D Object  
Detector with Sparse Transformer  
[arxiv.org/abs/2112.06375](https://arxiv.org/abs/2112.06375)



FSD V2: Improving Fully Sparse 3D  
Object Detection with Virtual Voxels  
[arxiv.org/abs/2308.03755](https://arxiv.org/abs/2308.03755)



Super Sparse 3D Object Detection  
[arxiv.org/abs/2301.02562](https://arxiv.org/abs/2301.02562)



Once Detected, Never Lost: Surpassing  
Human Performance in Offline LiDAR...  
[arxiv.org/abs/2304.12315](https://arxiv.org/abs/2304.12315)



编辑于 2023-09-03 22:46 · IP 属地德国

▲ 赞同 305 ▾ 17 条评论 分享 收藏 喜欢 ... 收起 ^

更多回答

 **yeyan**    
中国地质大学 工程硕士

+ 关注

谢邀 @somewhere

54 人赞同了该回答

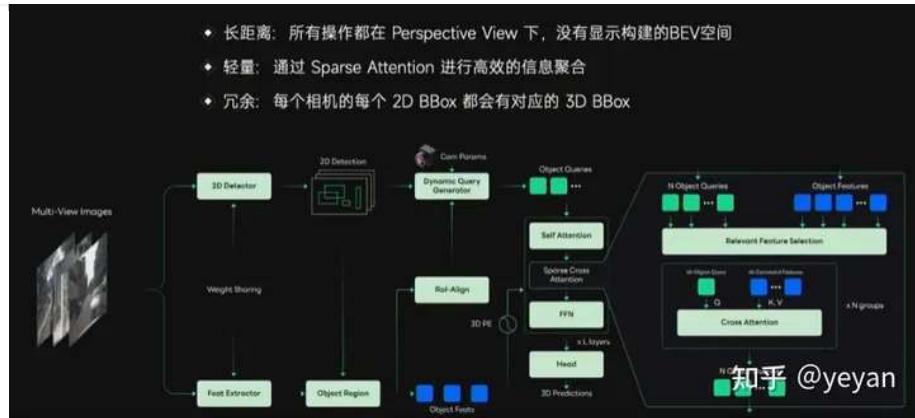
早上看了视频回放，主要是由图森的算法专家介绍他们一些技术问题的解决方案，覆盖了自动驾驶常见的方向，包含感知/定位/规划/数据闭环等，干货满满。

### 1. 感知

#### 1.1 视觉

目前自动驾驶感知算法多是bev算法，之前也看过bevfusion,bevdepth<sup>Q</sup>,bevformer等bev算法，首先分析了目前大部分bev算法的缺陷：长距离bev<sup>Q</sup>需要大量的计算和空间资源；缺乏对环视相机的互补冗余，针对该问题提出了基于物体的多视角3D检测框架。

## 如何评价图森未来的AI Day?



对于每个相机先检测2d box<sup>Q</sup>，然后利用检测的box再去query，做attn计算，减少了attn的计算消耗；直接在图像上检测也能检测到长距离的目标。

### 为什么要从 2D BBox 出发？

- 高效：成熟的检测器框架以及高质量易获取的标注数据
- 冗余：每个相机进行2D BBox检测，实现相机之间的冗余互补
- 准确：图像丰富的语义信息，可以实现高准确度以及高召回



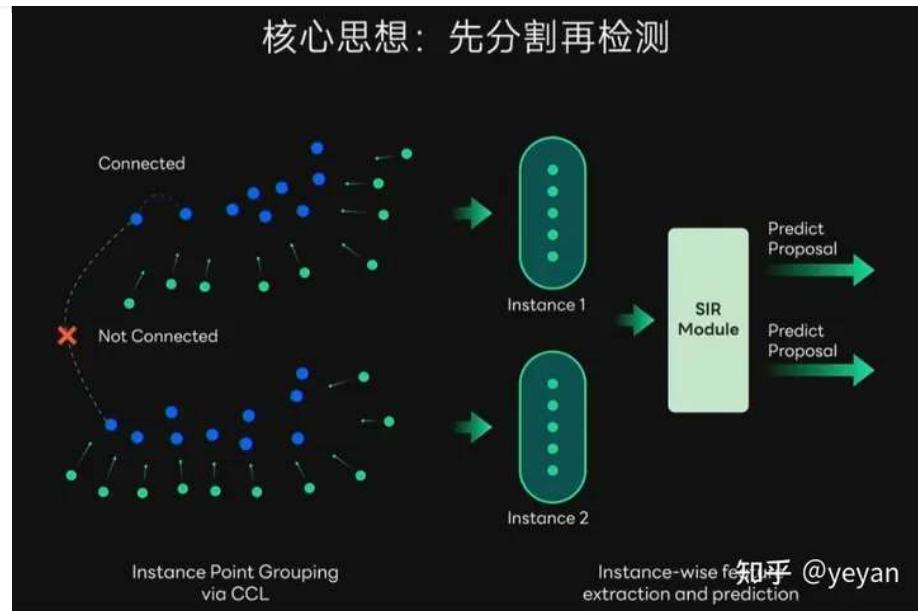
多视角融合<sup>Q</sup>sparse attention，多相机就是通过上述2d box限制atten的范围，**多模态<sup>Q</sup>**思路类似，将图像2d box投影到点云中，也可以限制点云atten的范围；时序融合采用了track的思路，track时每个目标都有一个id，对同一id的物体才去交互，也减少了atten<sup>Q</sup>计算的复杂度。



## 1.2 点云感知

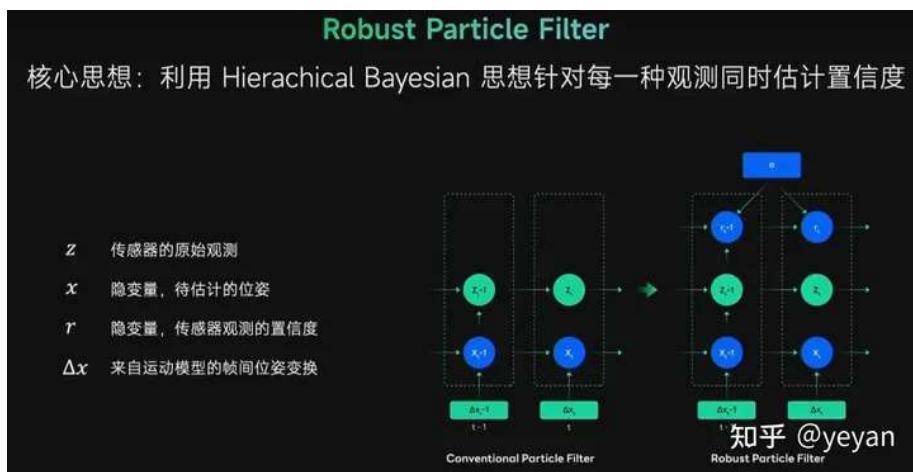
点云感知采用稀疏卷积的思路，在空间上，目前比如mink,spconv都实现了稀疏卷积，加速对点云数据的推理，考虑到**点云物体<sup>Q</sup>**只有表面的数据，中心缺乏特征，他们采用先分割再检测的方案。在时间上，基于点云的多帧数据，由于点云大部分都是背景，在检测时先去除背景（应该是类似**图像2dbox<sup>Q</sup>**的方法，主要还是减少了atten计算的范围），提高检测的效率。

## 如何评价图森未来的AI Day?



## 2. 定位

定位就是根据多传感器进行位置估计，在测绘领域，车载惯导系统<sup>Q</sup>一般就是利用gps,imu,轮速计<sup>Q</sup>进行pos解算，做绝对定位，但问题是在高架等场景，当gps信号丢失时，pos位置就会飘，影响定位精度。在自动驾驶时，由于多了视觉/lidar<sup>Q</sup>/radar等传感器，还可以利用这些传感器感知得到的目标与高精度地图或局部地图做匹配，提高定位的准确性。



## 3. 规划

采用联合预测规划<sup>Q</sup>思路，增加Contingency分支，评估规划安全性。



### 关于作者



刘斯坦

爱因斯坦的斯坦。。。

绘画、艺术、深度学习 (Deep Learning) 话题的优秀答主

王健飞、杨军、Xun Huang 也关注

## 如何评价图森未来的AI Day?



回答 914 茅草 77 天汪者 115,485

关注他

发私信

被收藏 80 次

- |                 |       |
|-----------------|-------|
| 有用的             | 3 人关注 |
| 陆荏苒 创建          |       |
| 技术              | 1 人关注 |
| vdgs 创建         |       |
| 机器学习            | 1 人关注 |
| Gtesla 创建       |       |
| 智驾-感知技术         | 1 人关注 |
| 樟木君Armstrong 创建 |       |
| 智能车             | 1 人关注 |
| Arthas 创建       |       |

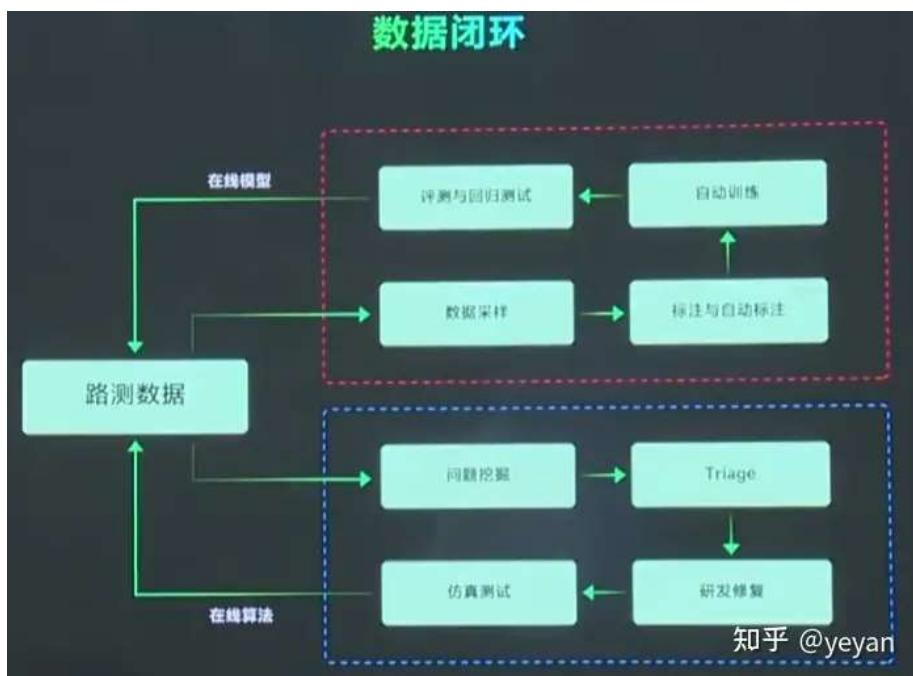
### 相关问题

- 你是如何进入「图森未来」的？有哪些值得分享的经验？ 0 个回答
- 图森未来要讲述一个什么样的未来呢？ 0 个回答
- 2023 年图森未来的发展会迎来哪些变化？ 0 个回答
- 图森未按时提交三季报被警告，这起到了哪些警示作用？ 0 个回答
- 卡车里的图森未来，何时窥见商业化曙光？ 0 个回答

## 4. 数据

自动驾驶数据非常重要，昨天马斯克在做特斯拉 fsdV12自动驾驶的直播时，在45分钟内也接管了一次，说明自动驾驶对全场景覆盖的重要性。

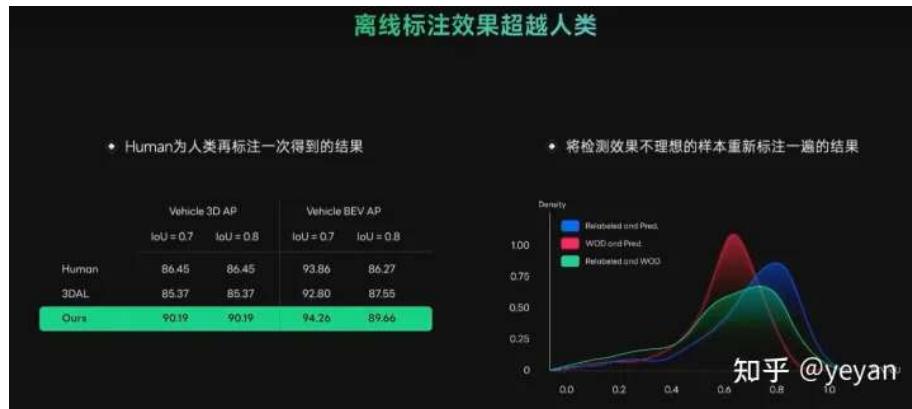
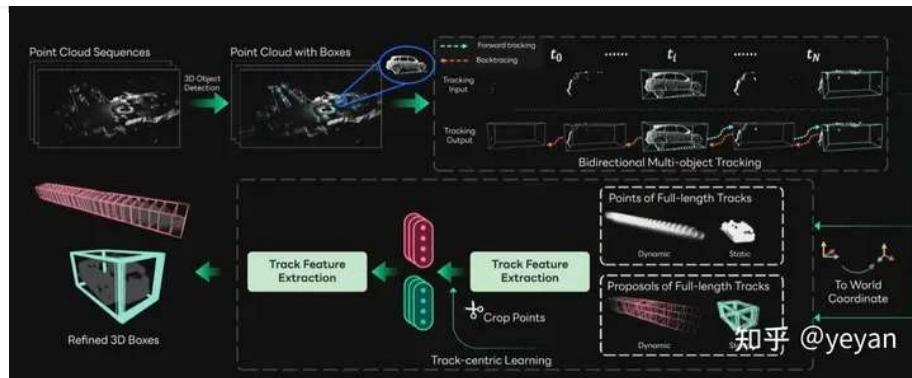
主要介绍图森数据闭环的工作，包括仿真，自动标注等。



数据标注这块提到了离线预标注，利用跟踪等算法，做到一帧检测，永不丢失，去辅助标注，从而达到离线标注效果超过人类。



## 如何评价图森未来的AI Day?



## 参考

以上技术基本都是图森论文的成果转化，可以参考图森的官方代码【[github.com/tusen-ai/sst](https://github.com/tusen-ai/sst)】，里边有以下几篇论文的实现。

- [Embracing Single Stride 3D Object Detector with Sparse Transformer \(CVPR 2022\).](#)
- [Fully Sparse 3D Object Detection \(NeurIPS 2022\).](#)
- [Super Sparse 3D Object Detection \(TPAMI 2023\).](#)
- [Once Detected, Never Lost: Surpassing Human Performance in Offline LiDAR based 3D Object Detection \(ICCV 2023, Oral\).](#)
- [FSD V2: Improving Fully Sparse 3D Object Detection with Virtual Voxels.](#)

编辑于 2023-09-01 21:36

真诚赞赏，手留余香

赞赏

还没有人赞赏，快来当第一个赞赏的人吧！

▲ 赞同 54 ▾ ● 1条评论 ↗ 分享 ★ 收藏 ❤ 喜欢 ... 收起 ^

 Jhin

+ 关注

发布于 2023-08-31 11:27

▲ 赞同 ▾ ● 添加评论 ↗ 分享 ★ 收藏 ❤ 喜欢 ...

## 如何评价图森未来的AI Day?