**Time 3D: End-to-End Joint Monocular 3D Object Detection and Tracking for Autonomous Driving**

**--related reading(CVPR 2022)**

**学号：162010312 姓名：王康 班级：1620103**

**自行选择近几年CVPR/ICCV/ECCV/SIGGRAPH/IEEE TVCG上关于3D vision的论文 （或者近两年在arxiv上预发表的关于3D vision论文）(例如RGB-D数据处理（例如跟踪，补全，物体检测）、点云数据处理、可微分渲染、曲面重建、PointNet, PointCNN, Point transformer,MeshCNN, ShapeConv, 清华“计图“开源框架上实现的一些算法，或者从RBG图像中估算depth与normal信息，等等)**

提交**纸质版（双面打印，用订书机订好）**与**电子版**！！！

**正文：**参照以下提纲撰写，要求内容翔实、清晰，层次分明，标题突出。**请勿删除或改动下述提纲标题及括号中的文字。**

**1．本论文的研究意义和研究现状及发展动态分析，附主要参考文献目录）；20分**

**黑色字体写。（可以图文并茂，建议自己思考画图）**

**研究意义：**虽然单独利用单目 3D 对象检测和 2D 多对象跟踪可以直接以逐帧方式应用于序列图像，但独立跟踪器切断了从 3D 检测器到跟踪的不确定性传输，从而无法将跟踪误差差返回到 3D 检测器。在这项工作中，作者建议以端到端的方式仅从单目视频中联合训练 3D 检测和 3D 跟踪。关键组件是一个新颖的时空信息流模块，它聚合几何和外观特征，以预测当前和过去帧中所有对象的稳健相似性分数。具体来说，作者利用了transformer的注意力机制，其中self-attention聚合了特定帧中的空间信息，而cross-attention利用了序列帧时域中所有对象的关系和亲和力。然后监督亲和力以估计轨迹并引导相应 3D 对象之间的信息流。此外，作者提出了一种时间一致性损失，它明确地将 3D 目标运动建模纳入学习，使 3D 轨迹在世界坐标系中平滑。

**研究现状及发展动态分析：**由于缺乏深度信息，单目3D目标检测是一个自然 的不适定问题，因此很难估计3D目标的准确和稳定状态。典型的解决方案是通过2D多对象跟踪器（MOT）平滑先前和当前的状态，但目前大多数模型仍然是人为设计的，因此相应的跟踪器只能独立跟踪探测器。而本论文提出的Time3D 在 nuScenes 3D 跟踪基准上实现了 21.4% AMOTA、13.6% AMOTP，超过了所有已发布的竞争对手，运行速度为 38 FPS，在 nuScenes 3D 检测基准上实现了 31.2% mAP、39.4% NDS。Time3D可能会激励未来的研究人员将3D跟踪和三维检测结合在一个统一的框架中，并编码更多的3D信息，使基于视觉的自动驾驶更加实用。

**2．本论文的motivation（动机）、研究内容，以及要解决的关键问题（此部分为重点阐述内容）；20分**

**（可以图文并茂，建议自己思考画图）**

**动机：**三维目标检测是[自动驾驶](https://so.csdn.net/so/search?q=%E8%87%AA%E5%8A%A8%E9%A9%BE%E9%A9%B6&spm=1001.2101.3001.7020" \t "https://yongqi.blog.csdn.net/article/details/_blank)的一项重要任务。与激光雷达系统相比，单目摄像头价格便宜、稳定、灵活，深受批量生产汽车的青睐。然而，由于缺乏深度信息，单目3D目标检测是一个自然的不适定问题，因此很难估计3D目标的准确和稳定状态。典型的解决方案是通过2D多对象跟踪器（MOT）平滑先前和当前的状态。目前，大多数模型仍然是人为设计的，因此相应的跟踪器只能独立跟踪探测器，而这些方法在自动驾驶场景中存在三个缺点。因此，作者提出了将3D单目目标检测和3D MOT结合到一个具有端到端训练方式的统一架构中。

**研究内容：**

本论文提出一个将3D单目目标检测和3D MOT结合到一个具有端到端训练方式的统一架构，该架构可以：

（1）仅从单目图像预测2D框、3D框、Re-ID特征，而无需任何额外的合成数据、CAD模型、 实例掩码或深度图。

（2）对这些线索的兼容特征表示进行编码。

（3）学习差异关联，通过跨时间同时组合异质线索来生成轨迹。

（4）引导流经所有对象的信息，以生成具有时间一致性的目标状态。

并期望作出以下贡献：

（1）提出了一个统一的框架，通过端到端的方式组合不同线索，共同学习3D对象检测和3D多 对象跟踪。

（2） 提出了一种嵌入提取器，通过将二维和三维长方体转换为统一表示，使几何和外观信息兼容。

（3） 提出了一种时间一致性损失，通过约束时间拓扑使轨迹更加平滑。

（4）在nuScenes 3D跟踪基准测试中的实验表明，该方法在实时（26FPS）运行时，与其他竞 争对手相比，取得了最佳的跟踪精度

**关键问题：**

由于缺乏深度信息，单目3D目标检测是一个自然的不适定问题，因此很难估计3D目标的准确和稳定状态。典型的解决方案是通过2D多对象跟踪器（MOT）平滑先前和当前的状态。

目前，大多数模型仍然是人为设计的，因此相应的跟踪器只能独立跟踪探测器。

而这些方法在自动驾驶场景中存在三个缺点：

1）分别处理检测和关联，其中独立的跟踪模块切断了不确定性从3D探测器到跟踪的传输，而不 能将误差差分传递回3D探测器。

2）来自同一类别的对象通常具有相似的外观信息，并且在自动驾驶场景中经常发生遮挡和不同的 速度变化。未能将这些异构线索整合到一个统一的网络中。

3）不直接约束网络中外观和几何信息流的情况下估计轨迹，这对于轨迹平滑度、速度估计和运动 属性（例如停车、移动或停车）至关重要。

因此，作者提出了将3D单目目标检测和3D MOT结合到一个具有端到端训练方式的统一架构中。

**3．研究方案（技术）；（此部分为重点阐述内容）；30分**

**（可以图文并茂，建议自己思考画图）**

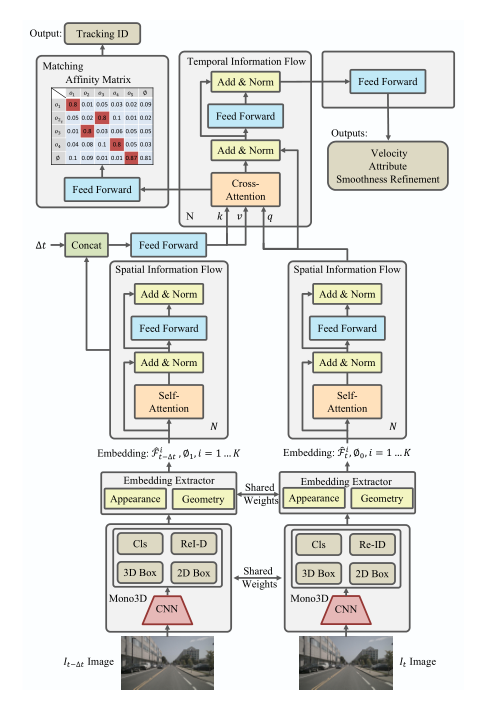
Time 3D仅将单目视频图像作为输入，包括以下步骤：

1）在JDE模式下，设计了一种快速准确的单目3D目标检测器，用于获取每帧的2D框、3D框、类别和Re-ID嵌入。

2）一个异构线索嵌入模块，将外观和几何特征编码为兼容的特征表示

3）一个时空信息流模块，将所有对象的信息跨帧相互传播，估计相似度以生成三维轨迹，并聚合世界坐标系中的几何相对关系以估计速度、属性和框平滑度优化。

图1 展示了Time 3D的整个架构细节。



*图1 Time 3D的架构细节*

****1. 单目三维目标检测****

采用KM3D作为单目3D检测器，该检测器根据可微几何推理模块（GRM）进行位置估计，预测尺寸、方向和九个透视角。FairMOT之后，添加了一个与其他检测头平行的Re-ID头，其重点是生成不同对象的区分特征。与其他头部实现相同的卷积层，但输出256维向量以在每个对象的2D中心提取Re-ID特征。

****2. 异构线索嵌入****

理想的数据关联可以在很长一段时间内进行多个线索（例如外观和几何）的嵌入。但是，外观特征（Re-ID特征）在向量空间中，几何特征（例如位置、尺寸和方向）在欧几里德空间中，使得它们很难在统一的网络中组合。因此，在本文中，作者对外观、几何和运动信息的兼容表示进行了编码。

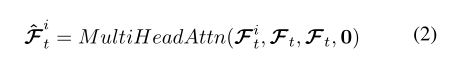
对于图像中的每个2D框和3D框，首先将其参数转换为2D角点和3D角点。这些角点被展平，然后输入到轻量级的PointNet中，该网络仅由3层MLP和MaxPooling组成，以生成具有d维特征的几何特征。

除了在外观特征中重新标识外，作者还添加了类别线索，可以进一步用于约束同一对象在不同帧之间的相似性。

****3. 时空信息流****

将transformer体系结构扩展到时空信息，其中的自我注意力机制在某个时间传播对象信息，交叉关注跨时间聚合对象信息。空间信息流的结构如图1底部所示。

首先从3D探测器的主中心头提取图像中的中心点，并索引其相应的外观特征和几何特征，然后与MLP层连接以生成其输入嵌入。空间信息流可以概括为：

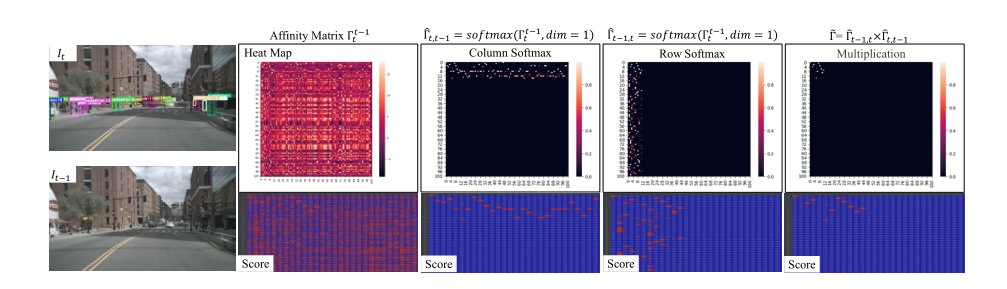


空间信息流模块过程严格遵循对传播信息的自我注意力机制，并对空间拓扑进行编码。

时间信息流模块从成对的帧中聚合信息和在residual下使用多头交叉注意力机制。时间流模块的结构如图1顶部所示。在交叉注意力机制中，点积权重探索了不同帧中成对检测对象之间的关系。它表示由softmax归一化的0-1的概率一致性，其中0定义为不同的对象，1是相同的对象。这种概率可以用作直接跟踪的相似性分数。

为了防止ID转换，作者将Time 3D设计为半全局关联，因此还需要捕获时间信息。

考虑到图像对中的对象可能没有对应关系，作者在DAN之后学习了未识别目标的额外行和列。简单地将FFN添加到交叉注意力机制的权重中，并估计适应性矩阵Γ，将其训练为单峰适应性矩阵进行一对一匹配，如图2所示。



*图2 单峰适应性矩阵*

时间信息流分支可以生成跟踪信息，并指导外观和几何信息的时间聚合。然后，时间聚合模块对目标的时间过渡进行建模，以预测框平滑度细化和时间相关变量（例如速度、运动属性）。因此，时间信息流的机制可以总结为：

IMG_259

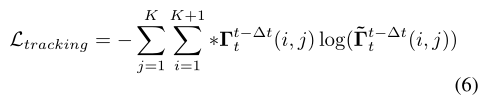
****4. 训练损失****

作者将多任务损失分为三部分：单目目标3D检测损失LMono3D、跟踪损失Ltracking和时间一致性损失LCons。

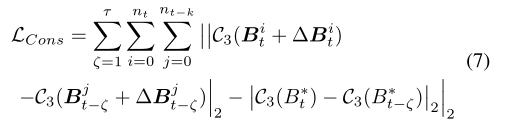
1) 单目目标3D检测损失LMono3D

IMG_260

2) 跟踪损失Ltracking

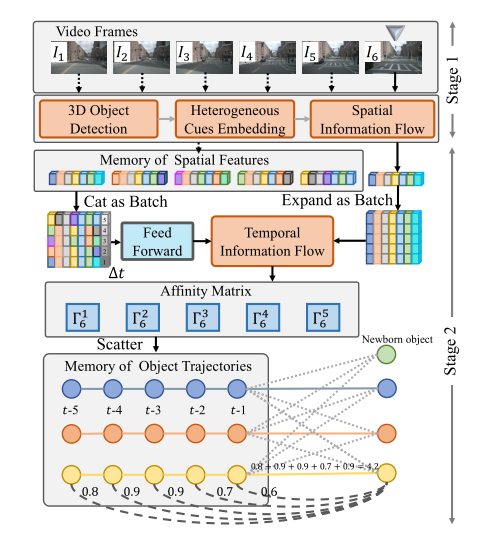


3) 时间一致性损失LCons



****5. 跟踪推断****

如图3所示。作者首先在每次tamp时按顺序执行3D对象检测、异构线索提取和空间信息流。然后，存储空间特征及其时间戳。给定当前帧图像及其空间特征，通过向前传递时间信息流来计算适应性矩阵。为了减少ID切换，作者按照DAN计算当前帧中对象的适应性矩阵和存储在所有轨迹中的空间特征，并将它们相加为对象和轨迹之间的相似性分数。最后，采用Hungarian算法获得最优分配。在逐帧运行分配过程时，生成对象轨迹。因此，每个帧图像只通过一次高权重的3D检测、嵌入提取器和空间信息网络，而存储的空间特征通过轻量级的时间信息流被多次使用以计算相似度得分。因此，Time 3D可以实时运行。



*图3：跟踪推断*

**4．本论文的创新之处；10分**

本文提出了一种新的框架，在实时运行的情况下，仅从单目视频中联合学习三维目标检测和三维多目标跟踪；提出了一个基于transformer的模块，可以预测当前和过去帧中所有目标的鲁棒相似性；还提出了一个损失函数。

作者的框架将异构线索（包括类别、2D框、3D框和Re-ID特征）编码为兼容嵌入。基于transformer的体系结构将时空信息流传递到估计轨迹，并通过时间一致性损失进行优化，使其更加平滑。在nuScenes数据集上，所提出的Time 3D在实时运行的同时实现了最先进的跟踪性能。

为了为不同线索设计兼容的特征表示，作者提出将2D框和3D框的不同大小的参数转换为统一表示、2D角点和3D角点，其中几何信息可以通过广泛使用的PointNet结构从角点原始坐标中提取为高维特征。

作者采用了多方面完备的实验，包括3D跟踪和3D物体检测性能测试集、Time 3D的定性结果、各方面的消融实验、时间一致性损失以及与最先进的产品进行比较。

1. **本论文不足之处与可能改进的地方；10分**

本论文在实验的实施细节上可加以深入。Time3D训练没有用KM3D预训练模型，使用了ImageNet预训练模型作为初始化。将图像的大小从900×1600调整到448×800，以进行快速训练。作者训练Time3D 200个epoch，起始学习率为1.25e−4，在epoch 90和120时将学习率降低×10。为了速度快，只把过去的五帧和当前帧关联起来。由上，在条件允许的情况下，可以花费更多的时间成本以获取更精确的实验数据。

本论文的训练损失由作者将多任务损失分为三部分：单目目标3D检测损失LMono3D、跟踪损失Ltracking和时间一致性损失LCons。需要考虑更多因素。

由于对3D计算机视觉方面相关知识了解较少，本论文更多的可改进之处仍有待发掘。

**6. 读完本英文论文或者上完智能计算机图形学这门课的心得体会（可以从各个方面写）；10分**

本论文是今年刊登在CVPR上的一篇论文，提出了将3D单目目标检测和3D MOT结合到一个具有端到端训练方式的统一架构。本论文是从现实生活出发，基于自动驾驶的三维目标检测而来，从单目摄像头上做文章，是一种新奇的角度。文章不仅提出了一个统一的框架，也提出了一种嵌入提取器和一种时间一致性损失，以完备的过程让我了解了TIME 3D 的可取之处，为单目3D目标检测作出了极大贡献。可以说这是我第一篇仔细阅读的论文，经此一役，也向我打开了新世界的大门，对一篇完整的论文的内容与结构有了清晰的认知，为我以后对更多论文的阅读甚至写作打下了基础。

在上学期选课时，就对计算机图形学这一课程产生了兴趣，在初步了解后便选择了这门课。经过一学期的学习，虽然课程时间不长，但接触了不少计算机视觉方面的知识，初步学习了OpenGL语言的编写，此外还听了很多研究生学长学姐的学习分享，对研究生阶段的知识学习有了第一次的接触。但由于我个人的知识储备不足，在听这门课的部分内容（尤其论文分享）时存在较大困难，接受起来不是那么容易，这也是我对这门课唯一遗憾的地方，但同时也是这门课存在的意义，不仅仅传授了图形学的相关知识，也让我发现了自己大量的知识漏洞以及要去学习的方向和目标。

比如在写完3D room的大作业时会有一种自豪感，在阅读论文时会感到困难，在课上演示3D视觉知识或播放相关视频时会感到新奇。所以上完整个课时后，我既对学习到的图形学领域的知识充满了好奇，也对自己没有全部弄懂课上讲过的知识感到遗憾，同时也找清了未来的学习动力。这就是计算机图形学这门课教给我的东西，始于课堂又超出课堂。