2021年5月，马斯克宣布Autopilot的视觉系统现在已经足够强大，可以在没有毫米波雷达的情况下单独使用。紧接着特斯拉官网做出更新，在其车辆的传感器介绍页面也取消了毫米波雷达的图示。

根据 Karpathy——CV界华人大佬的Fei-Fei Li的学生Andrej Karpathy博士的说法，相机比雷达好100倍，所以雷达不会给融合带来太多影响，有时雷达甚至会使情况变得更糟。

2021年8月特斯拉的Day上，Karpathy博士对Autopilot的视觉方案做了详细的讲解。其核心模板叫HydraNet，内部细节设计非常具有启发性，下面一起来看看。

要理解这个方案，首先需明确一下Autopilot视觉系统的输入和输出。

图1 Tesla视觉感知系统的输入和输出

Autopilot的视觉系统的传感器由环绕车身的8个摄像头组成，分别为前视3目：负责近、中远3种不同距离和视角的感知；侧后方2目，侧前方2目，以及后方1目，完整覆盖360度场景，每个摄像头采集分辨率为1280 × 960、12Bit、 36Hz的RAW格式图像，对车身周边环境的探测距离最远可达250m。

摄像头收集的视觉信息经过一系列神经网络模型的处理，最终直接输出用于规划和智驾系统3D场景下的 ”Vector Space”。

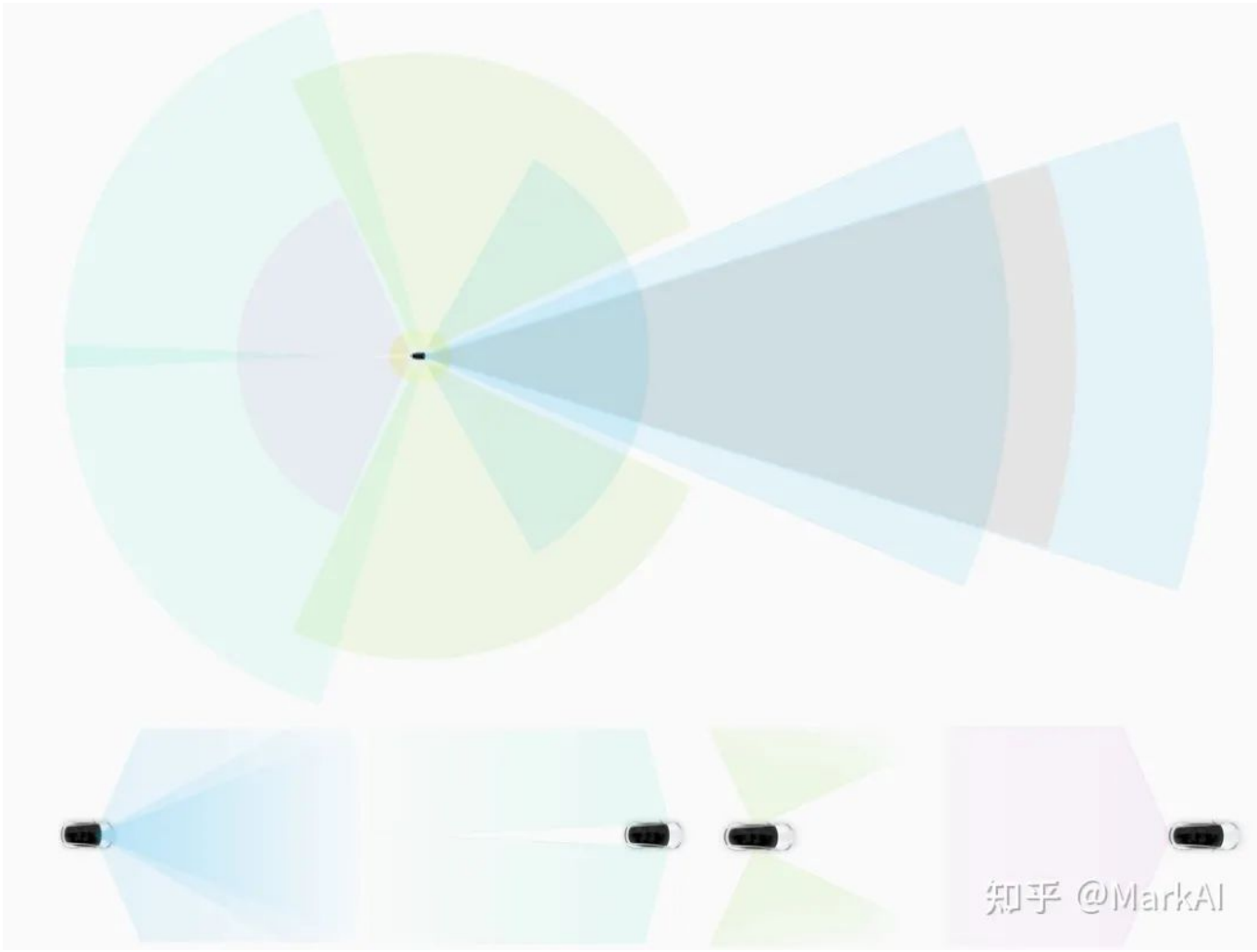

图2 Tesla车载相机布置方式

Tesla的自动驾驶感知算法经过了多个版本迭代，应用到了近期的FSD中。我们首先介绍一下最初的HydraNet。

HydraNet

HydraNet以分辨率为1280×960、12-Bit、36Hz的RAW格式图像作为输入，采用RegNet作为Backbone，并使用BiFPN构建多尺度feature map，然后再在此基础上添加task specific的Heads。

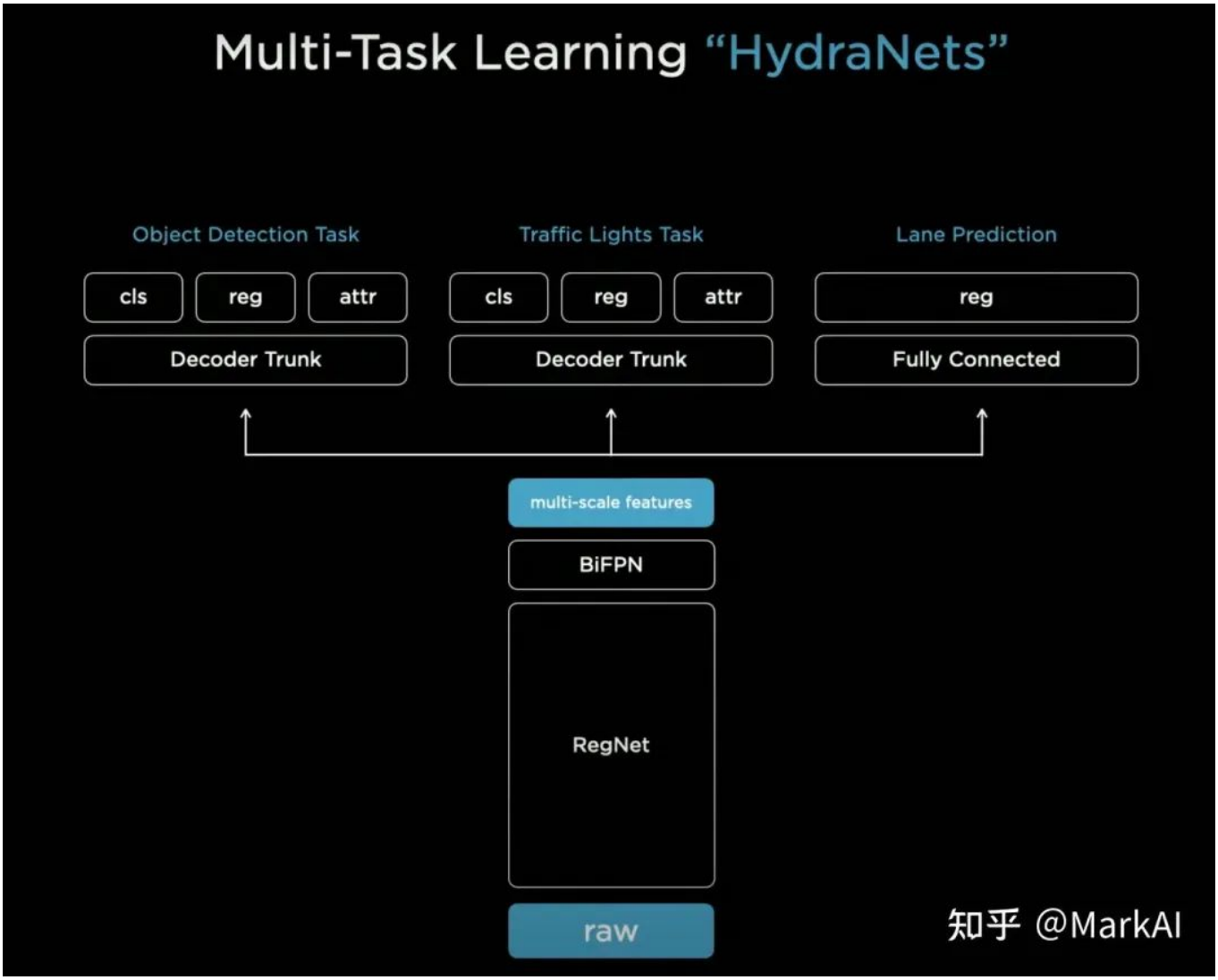

图3 HydraNet模型结构

熟悉目标检测或是车道线检测的读者会发现，初代HydraNet的各部分都比较普通，共享Backbone和BiFPN可以在很大程度上节省在部署时的算力需求，也算是业界比较常见的。

但是，Tesla却把这样的结构玩出了花来，主要带来以下三点好处：

1、预测非常高效：共享特征避免了大量重复的计算；

2、解耦每个子任务：每个子任务可在backbone的基础上进行fine-tuning，或是修改，而不影响其他子任务。

3、加速：训练过程中可将feature缓存，这样fine-tuning时可以只使用缓存的feature来fine-tune模型的head，而无需重复计算。

从上述可以看出，HydraNet的实际训练流程为先端到端的训练整个模型，然后使用缓存的feature分别训练每个子任务，再端到端地训练整个模型，以此循环迭代。

就这样，一个普通的模型就被Tesla挖掘到了极致，模型训练中一切可以共享的都进行共享，减少不必要的计算开销。

同时我们也容易发现，该模型如果仅使用单相机图像作为输入，会有很大盲区，只能用于很简单的辅助驾驶任务，例如车道保持等，并且这还需要借助其他传感器(超声波雷达、毫米波雷达等)来降低风险。

更复杂的自动驾驶任务(城市辅助、自主变道等)需要多相机的图像作为感知系统的输入，同时感知系统的预测结果需要转换到三维空间中的车体坐标系下，才能输入到规划和控制系统用以规划驾驶行为。

相比单相机，多相机输入不是简单的针对多个相机的图像输入分别预测，然后投影到车体坐标系下就可以，而需整合多个相机的感知结果，再投影到车体坐标系。这是一个很复杂的工程问题。

针对这一问题，Tesla给出了一个很好的解决方案。

进化一：多相机输入

我们知道不能简单的单独使用每个相机图像的感知结果来进行车辆的规划控制，要精确的知道每个交通参与者的位置，道路的走向，需要车体坐标下的感知结果。要达到这种效果有以下三种可能的方案：

方案1：在各相机上分别做感知任务，然后投影到车体坐标系下进行整合；

方案2：将多个相机的图像直接变换和拼接到车体坐标系下，再在拼接后的图像上做感知任务；

方案3：直接端到端处理，输入多相机图像，输出车体坐标下的感知结果；

对于方案1，实践发现效果不理想。比如图4，图像空间显示很好的车道线检测结果，投影到车体坐标之后，就变得不太能用。

原因在于这种实现方式需要精确到像素级别的预测， 才能够比较准确地将结果投影到车体坐标，而这一要求过于严苛。

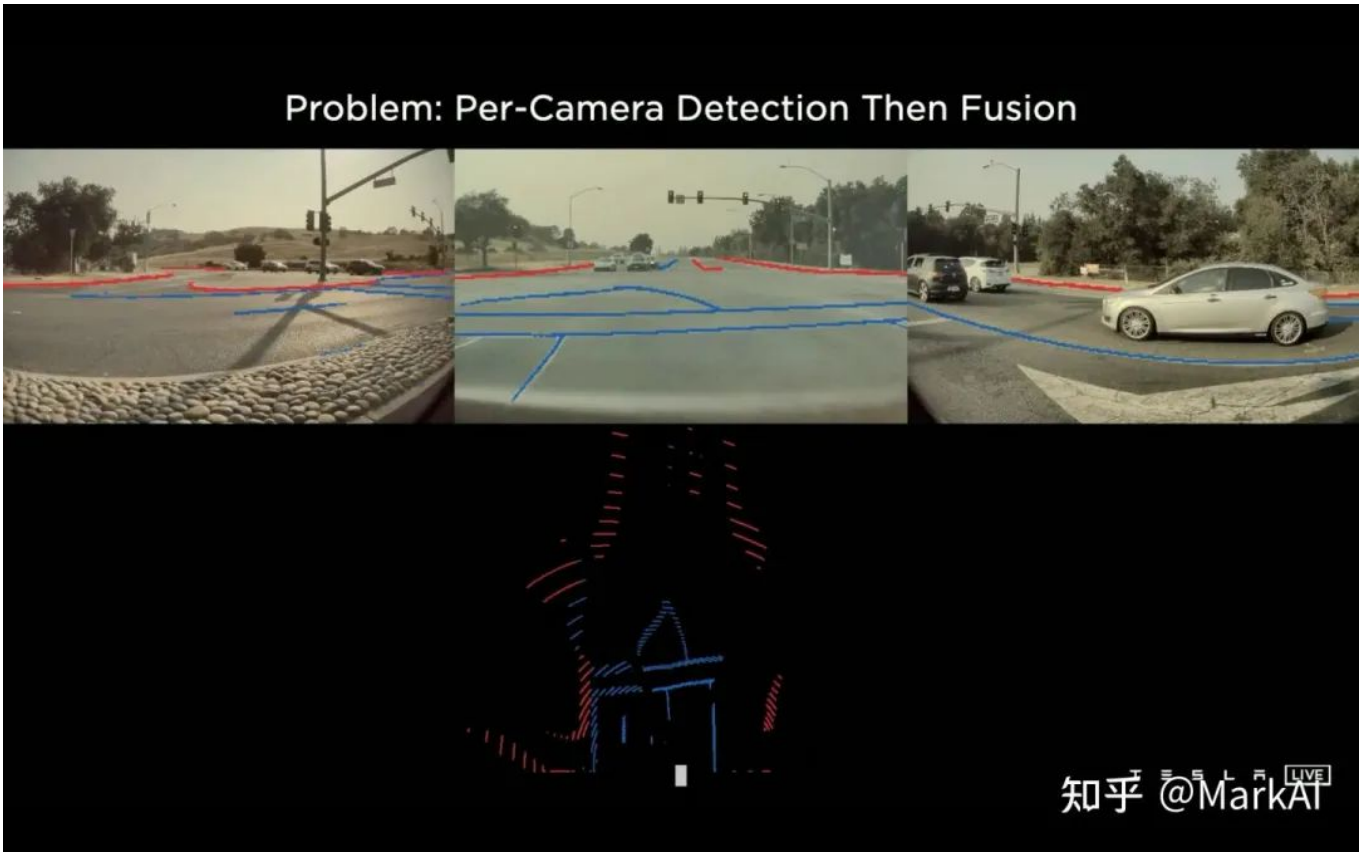

图4 图像空间下很好的车道线结果在Vector Space下不太理想

另外在多相机的目标检测中，会遇到一些问题，当一个目标同时出现在多个相机视野时，投影到车体坐标后会出现重影。此外，对于一些比较大的目标，一个相机的视野不足以囊括整个目标，每个相机都只能捕捉到局部，整合这些相机的感知结果就会变成非常困难。

对于方案2，图像完美拼接本就是一件非常困难的事情，同时拼接还受到路平面、遮挡的影响。

于是Tesla最终选择了方案3。方案3会面临两个问题，一个是如何将图像空间的特征转换到车体坐标，另一个是如何获得车体坐标下的标注数据。下面主要讨论第一个问题。

关于将图像空间的到车体坐标的特征转换， Tesla 使用一个 Multi-Head Attention 的transformer，来表示这个转换空间，而将每个相机的图像转换为key和value。

这是一个很精妙的方案，完美地运用了Transformer的特点，将每个相机对应的图像特征转换为Key和value，然后训练模型以查表的方式自行检索需要的特征用于预测。

这样的设计的好处是，无需显式地在特征空间上做一系列几何变换，也不受路平面等因素影响，很顺畅的将输入信息过渡到了车体坐标。

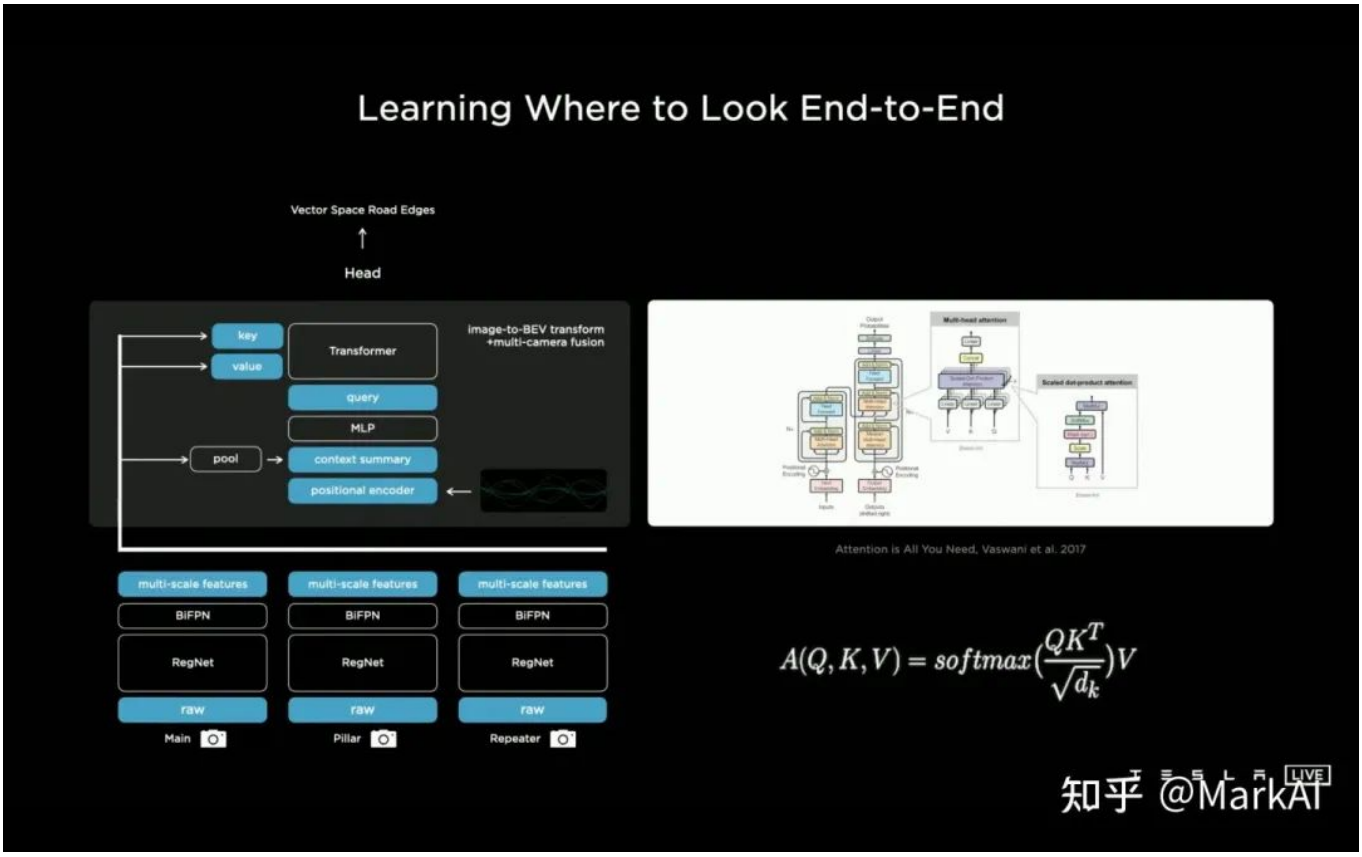

图5 使用Transformer整合多相机信息

加入这一优化后，车道线识别更加准确清晰，目标检测的结果更加稳定，同时不再有重影，效果如图6所示。

图6 使用Transformer整合多目信息后，感知效果明显提升

进化二：时间和空间信息

经上述优化后，感知模块虽然可以在多相机输入的情况下得到车体坐标准确且稳定的预测结果，但是是针对单帧的处理，没有时序信息。

而在自动驾驶场景中，需要对交通参与者的行为进行预判，同时视觉上的遮挡等情况需要结合多帧信息进行处理，因此需要考虑时序信息。

为了解决此问题，Tesla在网络中添加了特征队列模块用于缓存时序上的特征，以及视频模块用来融合时序上的信息。此外，还给模型加入了IMU等模块带来的运行学信息，比如车速和加速度。

经上述处理后，在Heads中进行解码得到最终的输出。

特征队列模块将时序上多个相机的特征，运动学的特征，以及特征的position encoding concat到一起，处理后的特征将输入至视频模块，如图7所示。

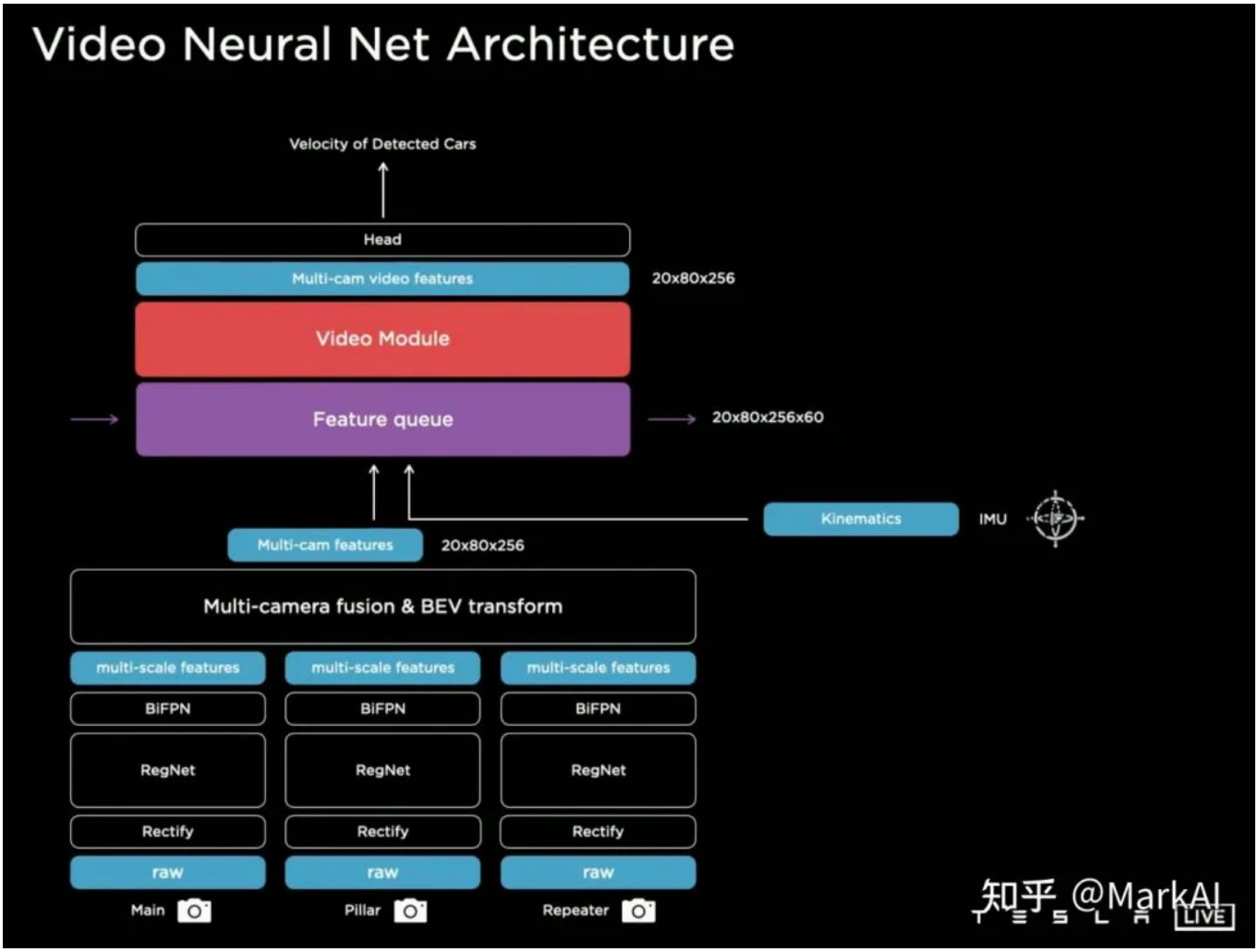

图7 在模型中加入特征队列，视频模块，以及运动信息作为进一步优化

特征队列模块按照队列的数据结构组织特征序列，其可分为时间特征队列和空间特征队列，如图8所示。

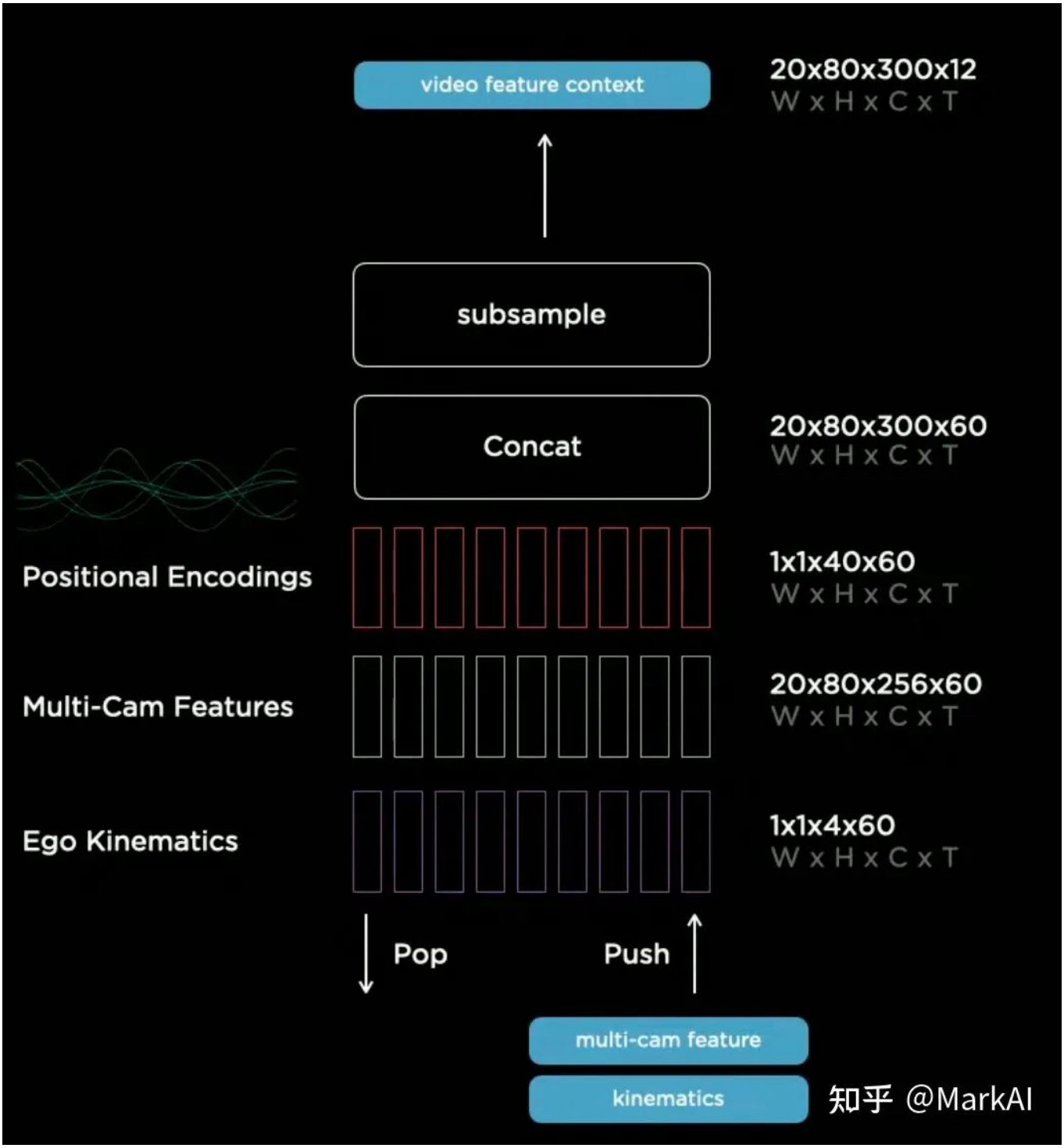

图8 特征队列

时序特征队列：每隔27ms将一个新的特征加入队列。时序特征队列可以稳定感知结果的输出，比如运动过程中发生的目标遮挡，模型可以找到目标被遮挡前的特征来预测感知结果。

空间特征队列：每前进1m将一个新的特征加入队列。主要用于需要长时间静止等待的场景，比如等红绿灯之类。因为在该状态下一段时间后，之前的时序特征队列中的特征会因出队而丢失。因此需要用空间特征队列来记住一段距离之前路面的信息，包括箭头、路边标牌等交通标志信息。

上述的特征队模块仅用于增加时序信息，而视频模块主要用来整合这些时序信息。Tesla采用RNN结构来作为视频模块，并将其命名为空间RNN模块，如图9所示。

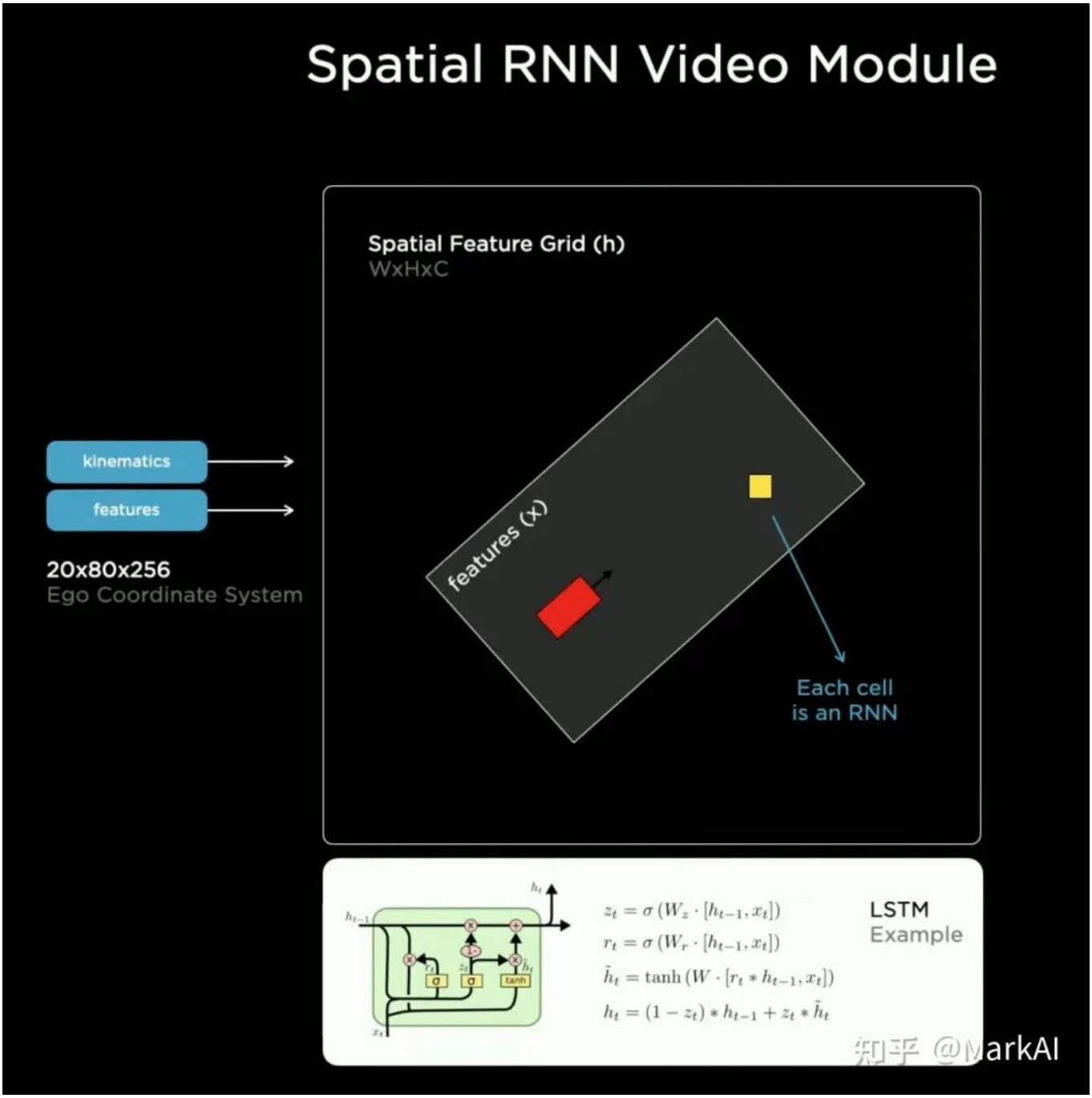

图9 空间RNN作为视频模块

因为车辆在二维平面上前进，所以可以将隐状态组织成一个2D的网格。当车辆前进时，只更新网格上车辆附近可见的部分，同时使用车辆运动学状态以及隐特征(hidden features) 更新车辆位置。

在这里，Tesla相当于采用了一个2D的feature map作为局部地图，在车辆前进过程中，不断根据运动学状态以及感知结果更新这个地图，避免因为视角和遮挡带来的不可见问题。同时在此基础上，可以添加一个Head用来预测车道线，交通标志等，以构建高精地图。

通过可视化该RNN的feature，可以更加明确该RNN具体做了什么：不同channel分别关注了道路边界线、车道中心线、车道线、路面等等，如图10所示。

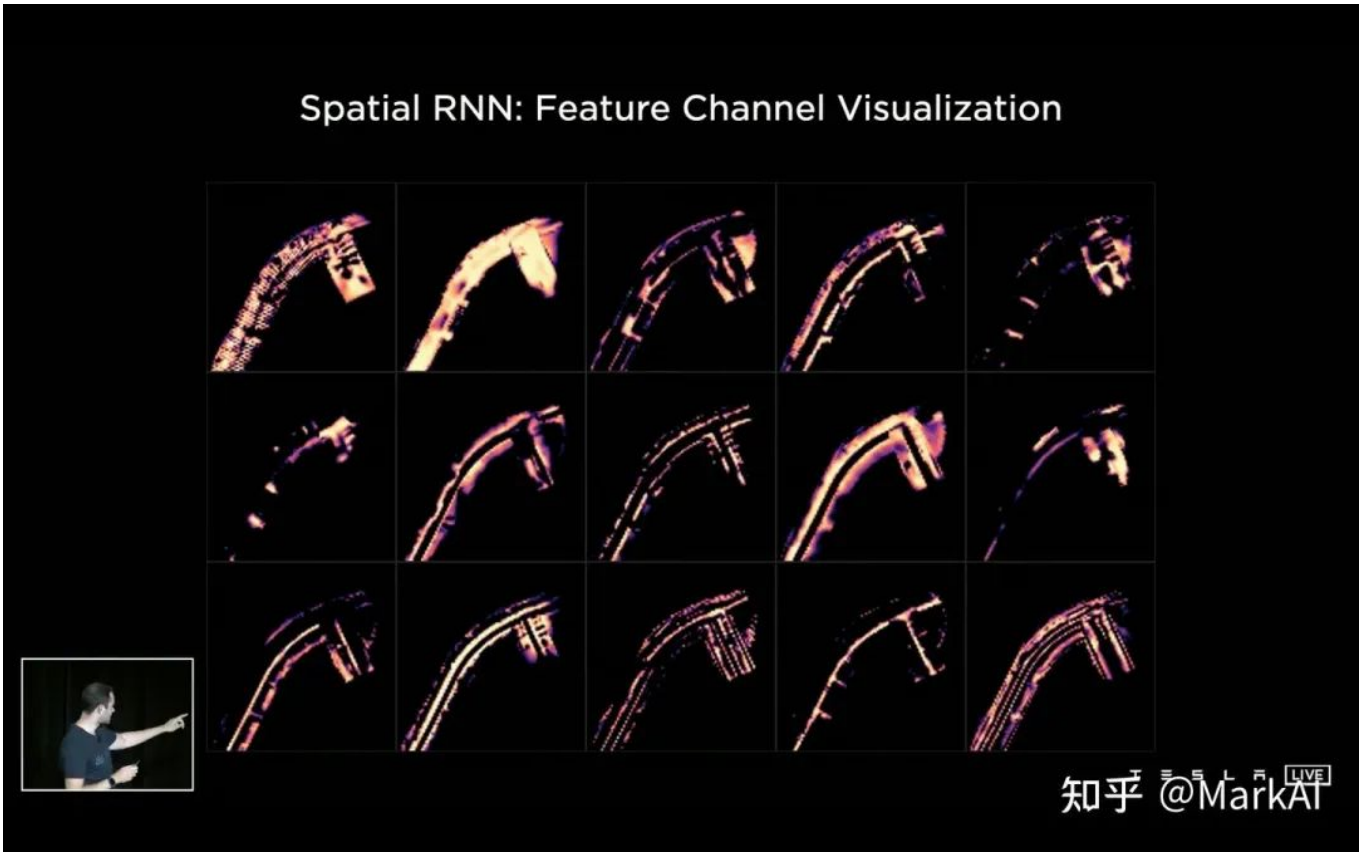

图10 空间RNN学到的特征可视化

添加了视频模块后，能够提升感知系统对于时序遮挡的鲁棒性，以及对于距离和目标移动速度估计的准确性，如图11所示。

图11 加入视频模块可以改善对目标距离和运动速度的估计，绿线为激光雷达的GT，黄线和蓝线分别为加入视频模块前后模型的预测值

最终的模型

在初版HydraNet的基础上，使用Transformer整合了多个相机的特征，使用Feature Queue维护一个时序特征队列和空间特征队列，并且使用Video Module对特征队列的信息进行整合，最终接上HydraNet各个视觉任务的Head输出各个感知任务。

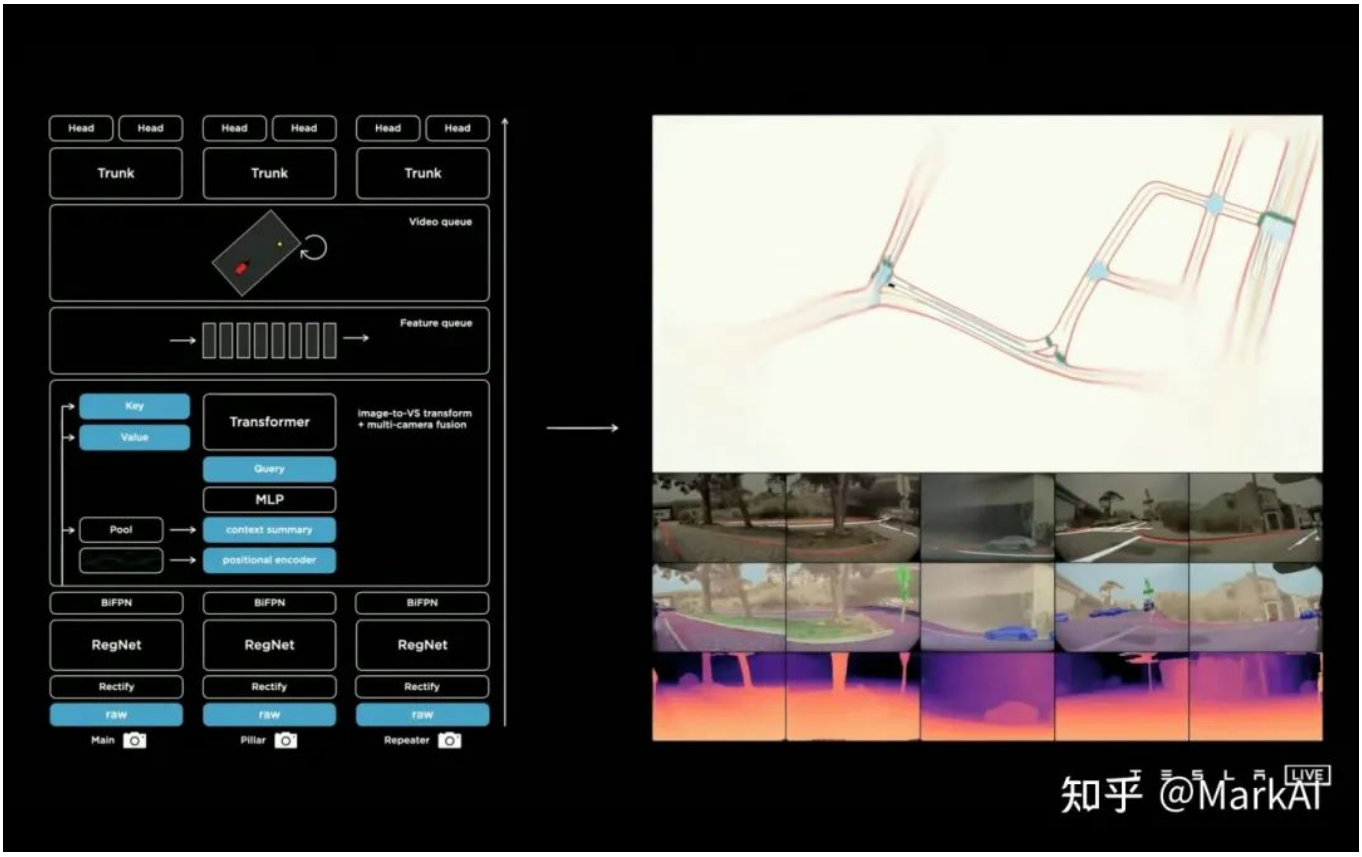

图12 最终完整的模型结构以及对应感知结果

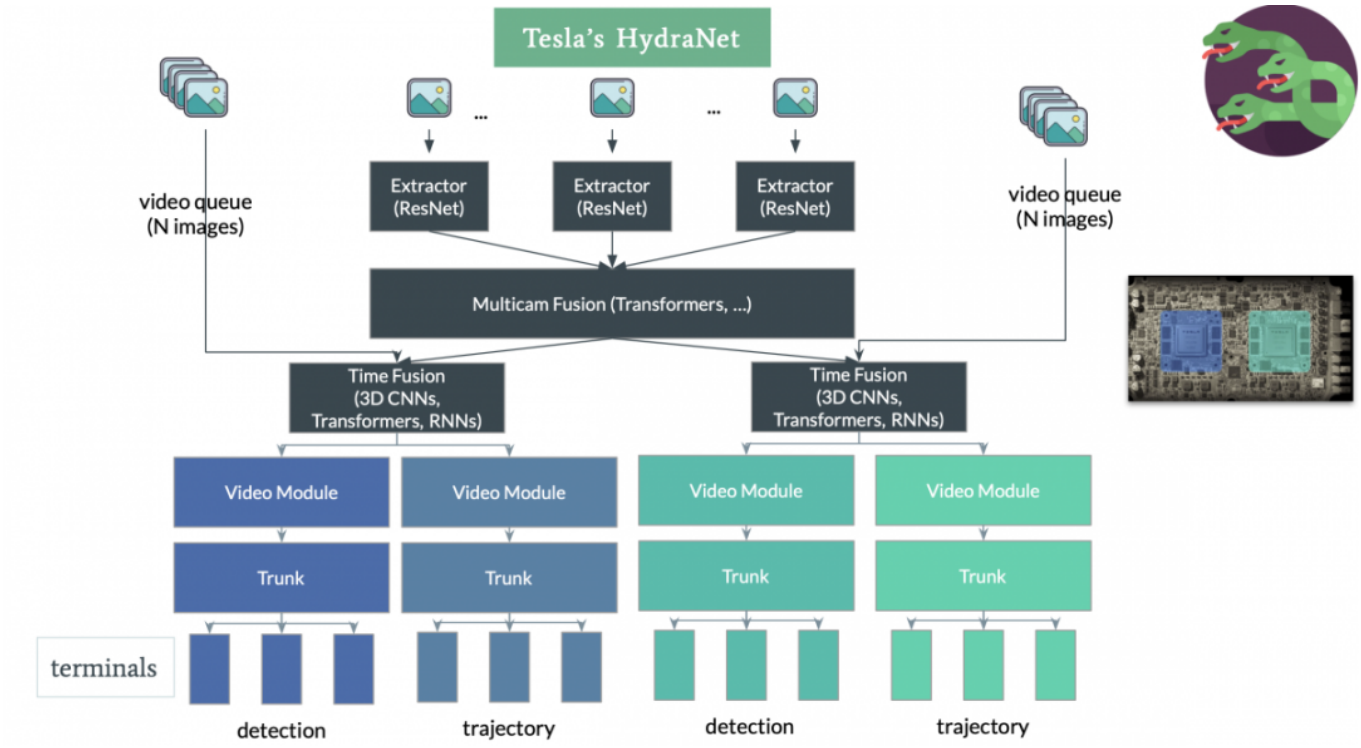

图13 最新HydraNet模型结构

最新的HydraNet模型如图13所示，简单展示了图像提取、多相机图像融合、时间融合，以及最后拆分为不同的HEAD。

整个感知系统使用一个模型进行整合，融合了多个相机时序上和空间上的信息，最终直接输出所有需要的感知结果，一气呵成，非常干净和优雅，可以当做教科书一般。

赞叹该系统的精妙之外，也可以看到Tesla团队强大的工程能力，背后强大的算力和数据标注系统是支持这一切的前提，当然，那啥，本质上还是有钱啦……

此外，该系统也并不是最终版的自动驾驶感知系统，还会一直不断迭代升级，国内的同行们要加油了！！