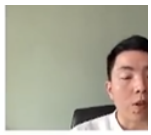
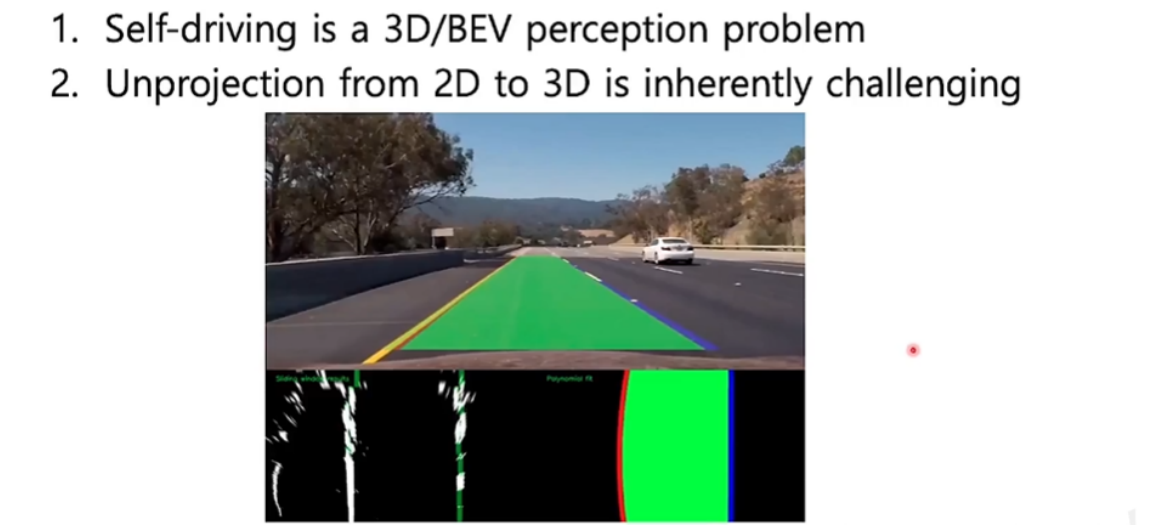
BEV

一 .why bev perception

1. self-driving is a 3D/BEV perception problem
2. 2D to 3D

直接人为用2D检测投影到3D，由于较远处像素信息较少，噪声较多，容易失真。





3. make it easy for sensor fusion(多模态融合)

二 .LSS

将单视图的检测扩展到多视图为什么不可行：具体来说，针对来自n个相机的图像数据，我们使用一个单视图检测器，针对每个相机的每张图像数据进行检测，然后将检测结果根据对应相机的内外参数，转换到车辆本体参考下，这样就完成了多视图的检测。

这样简单的后处理无法data-driving，因为上面的转换是单向的，也就是说，我们无法反向区分不同特征的坐标系来源，因此我们无法轻易的使用一个端到端的模式来训练改善我们的自动感知系统

Method

1. **Lift: Latent Depth Distribution**：将2D图像特征生成3D特征

   对于每个像素，生成一系列离散的深度值（alpha，alpha为D*1矩阵），在模型训练的时候，由网络自己选择合适的深度。

   为什么要对每个像素定义一系列离散的深度值？因为2D图像中的每个像素点可以理解成一条世界中某点到相机中心的一条射线，现在不知道的是该像素具体在射线上位置(也就是不知道该像素的深度值)。本文是在距离相机5m到45m的视锥内，每隔1m有一个模型可选的深度值(这样每个像素有41个可选的离散深度值)

   然后将alpha与feature c做外积，得到一个D*C的矩阵，作为per-pixel outer product，下面这个图，由于第三个深度下特征最为显著，因此该位置的深度值为第三个

经过此操作，则估计出了pixel对应的depth信息，将2D图像特征生成3D特征

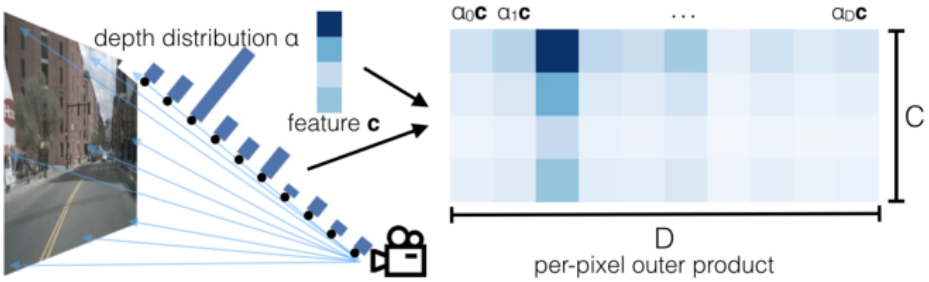


Fig. 3: We visualize the "lift" step of our model. For each pixel, we predict a categorical distribution over depth $\alpha \in \triangle^{D-1}$ (left) and a context vector $\mathbf{c} \in \mathbb{R}^C$ (top left). Features at each point along the ray are determined by the outer product of α and $\mathbf{c}$ (right).

2. **Splat: Pillar Pooling**：将3D特征通过相机内外参，投影到一个统一的坐标系中

   将多个相机中的像素点投影在同一张俯视图中，先过滤掉感兴趣域(以车身为中心200*200范围)外的点。然后需要注意的是，在俯视图中同一个坐标可能存在多个特征，这里有两个原因:1是单张2D图像不同的像素点可能投影在俯视图中的同一个位置,2是不同相机图像中的不同像素点投影在俯视图中的同一个位置，例如不同相机画面中的同一个目标。对于同一个位置的多个特征，作者使用了sum-pooling的方法计算新的特征，最后得到了200x200xC的feature

   最后接个一个BevEncode的模块将200x200xC的特征生成200x200x1的特征用于loss的计算

Conclusion

纯视觉bev检测的难点主要在于单目深度估计困难，本文提供了一种解决方法，但是由于对每一个pixel生成D*C的feature，导致计算量较大。