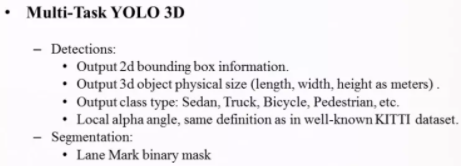
# Apollo 3D

### 单目摄像头下的物体检测网络

#### YOLO 3D for Object Detection in Apollo



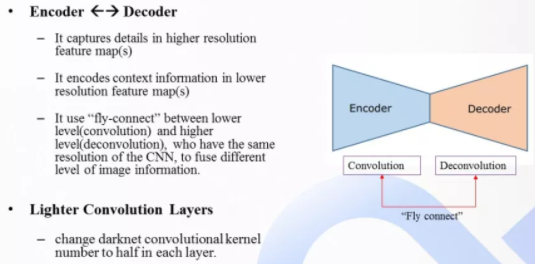
与原始YOLO V2的不同有如下几点

* **实现多任务输出**

* + **物体检测**，包括2D框(以像素为单位），3D真实物体尺寸（以米为单位），障碍物类别和障碍物相对偏转角（Alpha Angle，和KITTI数据集定义一致）

* + **物体分割**：车道线信息，并提供给定位模块

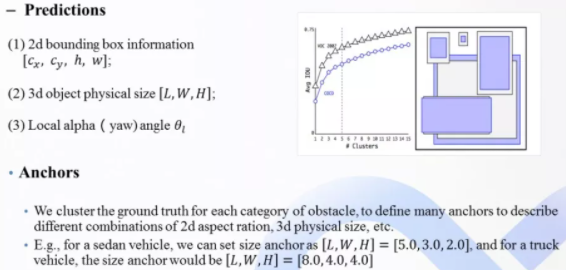
* **特征描述模块引入了类似FPN的Encoder和Decoder设计**：在原始Darknet基础上中，加入了更深的卷积层（Feature Map Size更小）同时添加反卷积层，捕捉更丰富图像上下文信息（Context Information）。高分辨率多通道特征图，捕捉图像细节（例如Edge，Corner），深层低分辨率多通道特征图，编码更多图像上下文信息。和FPN类似的飞线连接，更好的融合了图像的细节和整体信息



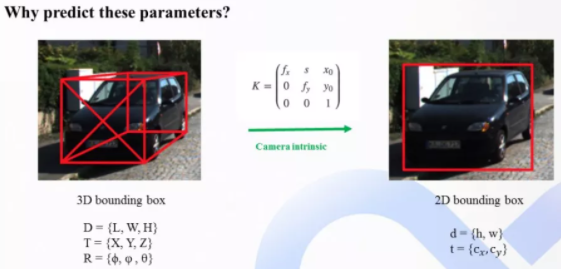
* **降低每层卷积核数目，加快运算速度。**例如我们发现卷积核数目减半，实验中准确率基本不变

如前所述，物体检测最终输出包括2D框(以像素为单位），3D真实物体尺寸（以米为单位），障碍物类别和障碍物相对偏转角（Alpha Angle，和KITTI数据集定义一致）等信息。

和YOLO V2算法一样， 首先在标注样本集中通过聚类，产生一定数目的“锚”模板，去描述不同类别、不同朝向、不同大小的障碍物。例如对小轿车和大货车，会定义不同的锚模板，去描述它们的实际物理尺寸

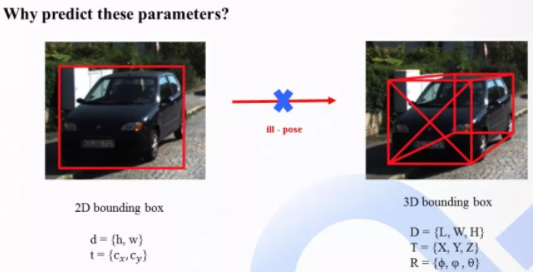


为什么要去训练、预测这些参数呢？以相机成像的原理来解释：针孔相机(Pinhole Camera)通过投影变换，可以将三维Camera坐标转换为二维的图像坐标。这个变换矩阵解释相机的内在属性，称为相机内参（Camera Intrinsic）K



对任意一个相机坐标系下的障碍物的3D框，可以用它的中心点 **T = {X, Y, Z}，长宽高 D = {L, W, H}**，以及各个坐标轴方向上的旋转角 R = {ϕ, φ , θ}来描述。这种9维的参数描述和3D框8点的描述是等价的，而且不需要冗余的8\*3个坐标参数来表示。

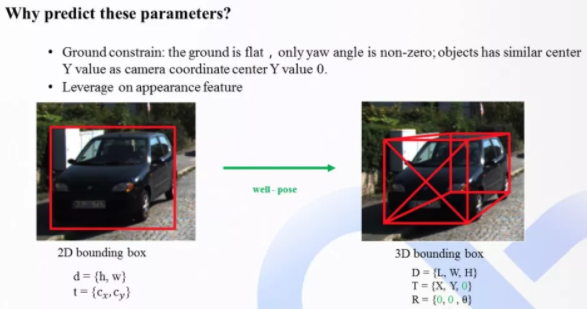
因此，对一个相机坐标系下3D障碍物，通过相机内参，可以投射到2D图像上，得到2D框[c\_x, c\_y, h, w]。从图中可以看到，一个障碍物在相机下总共有9维3D描述和4维2D描述，他们之间通过相机内参矩阵联系起来。



然而，只通过2D框[c\_x, c\_y, h, w]，是没有办法还原成完整的3D障碍物信息。

### 训练预测参数的设计

从几何学角度降低需训练的参数

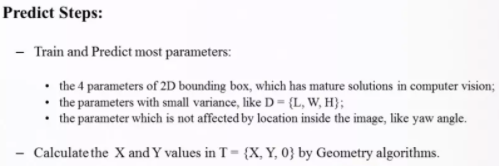


**首先利用地面平行假设，可以降低所需要预测的3D参数。**

例如：（1）假设3D障碍物只沿着垂直地面的坐标轴有旋转，而另外两个方向并未出现旋转，也就是只有yaw偏移角，剩下的Pitch Roll均为0。（2）障碍物中心高度和相机高度相当，所以可以简化认为障碍物的Z=0。

**第二，可以利用成熟的2D障碍物检测算法，准确预测出图像上2D障碍物框(以像素为单位）。**

**第三，对3D障碍物里的6维描述，可以选择训练神经网络来预测方差较小的参数**，例如障碍物的真实物理大小，因为一般同一类别的障碍物的物理大小不会出现量级上的偏差（车辆的高度一般在2-5米之间，很少会出现大幅变化）。而yaw 转角也比较容易预测，跟障碍物在图像中的位置关系不大，适合通用物体检测框架来训练和预测。实验中也多次证明此项。

所以现在唯一没有训练和预测的参数就是**障碍物中心点相对相机坐标系的偏移量X分量和Y分量**。需要注意的是障碍物离相机的物理距离Distance=sqrt(X^2+Y^2)。所以得到X和Y，我们自然就可以得到障碍物离相机的真实距离，这是单目测距的最终要求之一。

综上，我们可以合理的推断出, 实现单目摄像头的3D障碍物检测需要两部分：

**1. 训练网络，并预测出大部分参数：**

（1）图像上2D障碍物框预测，因为有对应的大量成熟算法文献；

（2）障碍物物理尺寸，因为同类别内方差较小；

（3）不被障碍物在图像上位置所影响，并且通过图像特征（Appearance Feature）可以很好解释的障碍物yaw偏转角。

**2. 通过图像几何学，来计算出障碍物中心点相对相机坐标系的偏移量X分量和Y分量。**

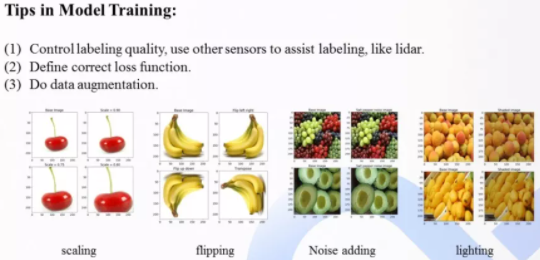
### 模型训练与距离测算

模型训练上，需要注意一些潜在的细节：

1) 确保标注质量，尤其是3D障碍物框。可以借助激光雷达等来辅助标注障碍物尺寸，偏转角等等；

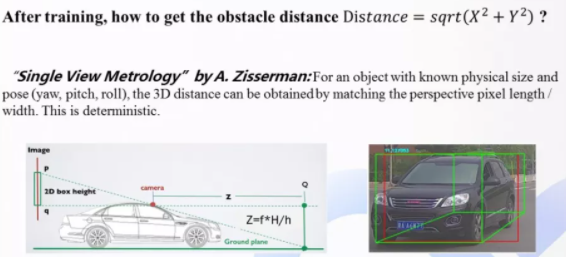
2) 定义合适的损失函数，参考Zoox的paper **《3D Bounding Box Estimation Using Deep Learning and Geometry》**；

3) 做好数据增强，避免过拟合, 图中简单描绘了一些Data Augmentation的方式。对于无人车，可以尝试更多的方法。

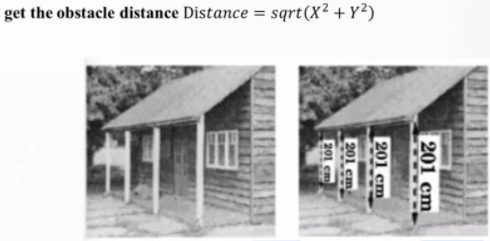


当训练好相应的神经网络，输出需要的各个参数之后，需要考虑的是如何计算出障碍物离摄像头的距离。根据之前介绍，通过内参和几何学关系，可以链接起图像中3D障碍物大小（单位为像素）和真实3D坐标系下障碍物大小（单位为米）。

采用**单视图度量衡**（ Oxford教授 A. Zisserman的论文《Single View Metrology》）来解释这个几何关系：任一物体，已知它的长宽高、朝向和距离，则它在图像上的具体形状大小等可唯一确定；反之亦然。



如图中房屋的支撑柱，大小高度完全相同，但是处于图像的不同位置，所占用的像素、长宽都有差别。



基于单视图度量衡，可以建立一个哈希查询表，去根据物体图像尺寸，物理尺寸，朝向角来查询物体的距离。

对于每种障碍物，根据它的平均（或单位）尺寸，去建立查询表，覆盖360度yaw 角的变化，来映射不同的距离。（例如2D框的25像素高，yaw角为30度，则它的距离为100米等等）。图中示例了一个小轿车在不同距离下、不同偏转角yaw angle情况下，在图像上的显示。

