由于我的轨迹预测是为了导航服务的，所以所有的2D图像上的轨迹预测统统不考虑，因为这些轨迹预测无法提供障碍物在三维空间中的位置。

要在三维空间上预测轨迹，只用视觉是不精确的，所以我的想法是激光雷达加视觉融合。阅读了《Human Motion Trajectory Prediction: A Surve》，选择了其中采取kitti数据集的几篇论文，因为这几篇论文采取了激光雷达加视觉的方式，和我的想法较为接近，所以深入学习这几篇论文。

1. DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents

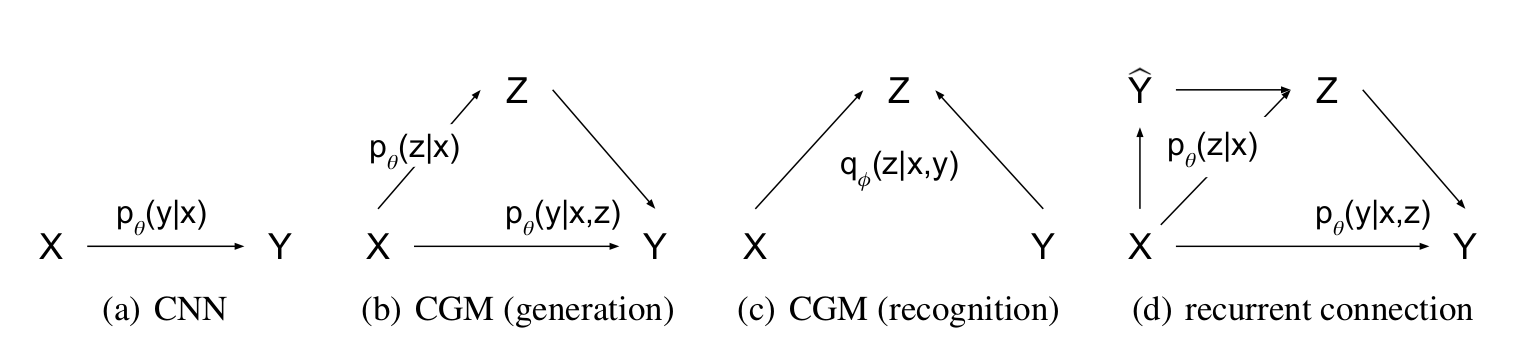
1.1辅助论文《Learning Structured Output Representation using Deep Conditional Generative Models》

这篇论文中的网络结构被DESIRE用来作对比，因此也了解一下。这篇的结构称为CVAE，是一种用来推断在状态s下出现x的概率的网络结构。这种网络通常用来预测。

CVAE是在VAE（variational auto-encoder）的基础上改进的，VAE是如下过程：

假设根据一个先验的分布概率p(z)产生了一组潜在的变量z，而在潜在变量z下，产生数据x的概率为p(x|z)。这里的理解是这样的，我们在观察中只能观察到数据x，但我们先验的知道这些数据大概和潜在的变量z有关，而且我们有一个先验的概率知道产生z的概率是多少。于是现在我们希望知道在z下产生x的概率分布。这样做有一个好处，就是x是很多的，但是z很少，从前一个状态的z过度到下一个状态的z是先验的，于是我们先得到z，在按照一定概率分布预测下一个时刻的x，这就是预测的过程。但是在整个过程中既不知道z的数量，也不知道通过z产生x的关系，甚至z的过渡过程都是不准确的，所以这个问题很困难。随机梯度变分贝叶斯（SGVB）的框架就是用来解决这个问题的。因为在这个问题中概率分布是很难学习的，所以SGVB训练了近似的分布q(z|x)。求分布q或者说p的方法是蒙特卡洛采样。这里原始的p可能是不可微的，离散的，近似的q是一个可微的函数，因此可以使用梯度下降。

整个模型的结构分为三类，如下图所示：



b的模型有一个好处，就是z->y的映射可以是一到多的。注意到b中x->y和z->y，可以近似的认为z->y。

CVAE的变化在于，认为z是由x和y产生的，即观察y和数据x双方构成了特征变量z，这里图c是一个基本结构，而d则是具体的方式，首先左上角的y是x经过一个CNN猜测的观察值，然后用这个x,y来产生z，在用z和x在真实的y中去学习分布函数。