# 工作内容

## 论文阅读SuperDepth\_Self-Supervised, Super-Resolved Monocular Depth Estimation

SuperDepth\_Self-Supervised, Super-Resolved Monocular Depth Estimation  
论文提出superdepth，获取depth及egomotion.

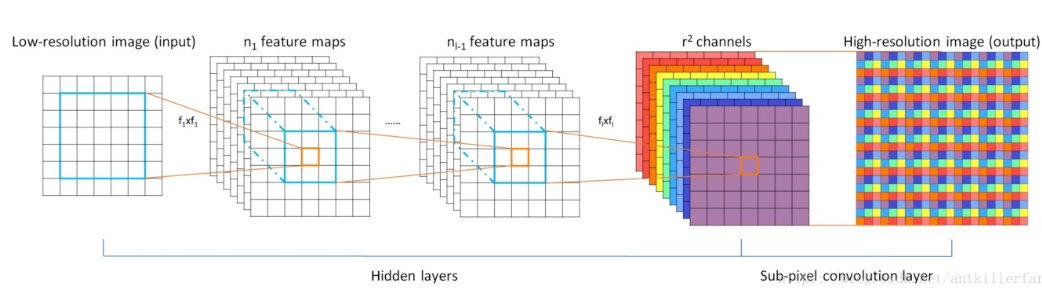
### 网络结构优化

整体框架与Unsupervised Monocular Depth Estimation with Left-Right Consistency一文（下简称monodepth）相似。Depth获取网络设计基于Dispnet网络，分为encoder和Decoder两部分。

论文在结构的主要贡献是decoder部分优化：

decoder后半段网络结构的作用是提高分辨率，形成一个视差图（Dispmap）的金字塔，以此来计算损失函数。本论文受 超分辨率 启发，借鉴ESPCN的网络结构进行修改上采样（或称反卷积）结构。

ESPCN基本结构sub-pixel convolutional layer：



上图很直观得表达了sub-pixel convolution的做法，前面就是一个普通的CNN网络，到后面彩色部分就是sub-pixel conv的操作了。首先，如果我想对原图放大2倍，那么我需要生成出2^2=4个same size的特征图。将4个same size的特征图拼成一个X2的大图，这就是sub-pixel convolution的操作了。

这是一种抽样的反思想，如果把一张x2的大图，每隔2个点抽样一个，那就会得到4张低分辨率的图像。于是，如果我们可以通过CNN来获得4张符合分布的低分辨率图像，那么就可以组成一张高分辨率的大图。

问题：

论文中提到了sub-pixel convolution 是由4个连续的2D卷积层组成分别是32, 32, 32, 16 layers，如果输出层层数是16，那么放大倍率是4，在DispNet的原本结构(4 pyramid scales in the disparity network)中,相邻的层是以2倍放大的，这里或许有其他的考虑。

### 2损失函数优化

使用公式代替monodepth中R-L consistency。

优势：引入相机的物理参数K（内参），以及相机之间的转移矩阵x（t-s）

本质上讲，R-L consistency是本论文方法的一个特例（光轴平行，光心在同一垂直光轴平面中），这样做引入物理本质

在论文zhou的论文Unsupervised learning of depth and ego-motion from video中，也使用了该方法。

简单来说，相机像素坐标系到世界坐标系据有以下关系（不考虑畸变问题）

丆 
0 
0 
0 
丆 
0 
0 
0 
X 
0 
0 
Z 
0 
这 样 ， 可 以 使 用 矩 阵 的 乘 法 统 的 处 理 X ， 丫 ， Z 这 三 个 坐 标 。 
坐 标 三 ） 就 是 像 点 p = @0） 的 齐 次 坐 标 ， 其 中 

其中是像素坐标的齐次坐标表示，满足：丆 
0 
0 
0 
丆 
0 
0 
0 
X 
0 
0 
Z 
0 
这 样 ， 可 以 使 用 矩 阵 的 乘 法 统 的 处 理 X ， 丫 ， Z 这 三 个 坐 标 。 
坐 标 三 ） 就 是 像 点 p = @0） 的 齐 次 坐 标 ， 其 中 可以看出，

我们知道，Z即该像素点的深度标量，也就是网络的最终输出，便可以从输出的深度图绘制出世界坐标系下的视场内物体，再通过坐标旋转的方式建立起两平面照片的关系。

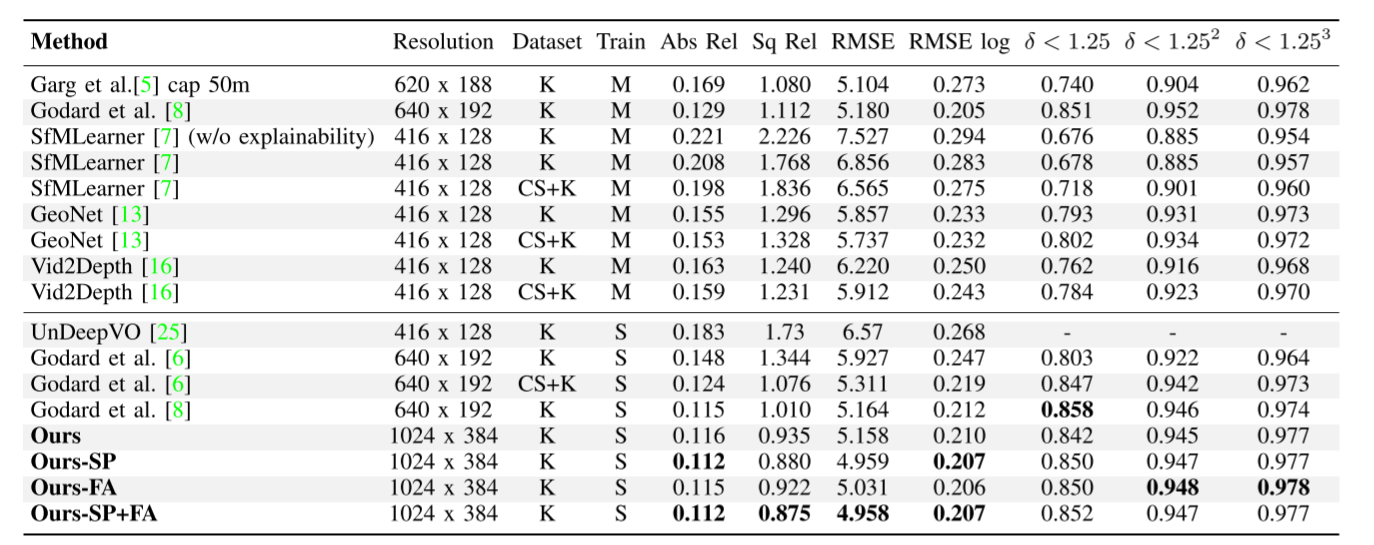
对于双目相机来说，是左右相机之间的坐标矩阵，可以通过联合标定来计算，是预先获取的值。而对于前后帧的关系来说，需要通过PoseNet获取ego-motion，再反馈给深度网络DispNet。

### 3.训练细节与实验比较

数据集对比：

|  |  |
| --- | --- |
| 名称 | 数据集以及细节 |
| Monodepth | 双目数据训练，前后帧之间没有互动关系 |
| Sfmleaner++ (Zhou, Google brain) | 单目视频流，loss函数通过帧与帧之间的关系构建 |
| Superdepth | 双目数据训练，前后帧之间没有互动关系 |

结果比较



是现有所有对稠密深度图获取最优的方式。

## 下周工作

1.尝试在monodepth的基础上复现Superdepth

2.深度阅读Depth Prediction Without the Sensors\_ Leveraging Structure for Unsupervised Learning from Monocular Videos论文