LIDAR-Camera Fusion for Road Detection Using Fully Convolutional Neural Networks

这篇论文是Luca Caltagirone 在Fast LIDAR-based Road Detection Using Fully Convolutional Neural Networks 基础上进一步深入研究完善的。

Abstract：

本文采用深度学习算法，将Lidar point cloud数据和Camera images数据相结合来road detection。首先将非结构化的稀疏点云投影到相机成像平面然后进行上采样以获得一组编码空间信息的密集2D图像。然后训练若干完全卷积神经网络（FCN）以通过使用来自单个传感器的数据或通过使用三种融合策略来执行道路检测：early，late和新提出的cross fusion。Cross fusion FCN设计是直接从数据中学得完整的信息；这是通过在LIDAR和相机处理分支之间使用可训练的交叉连接来实现的。

在KITTI上获得了excellent performance。

Introduction：

单独使用Lidar和Camera均有优点缺点，基于这两种传感器的优缺点，作者在这篇论文中提出了一个将LiDAR point cloud和camera images结合来进行road detection的办法。

本文主要的贡献：

1. 提出一个新颖的Lidar-camera fusion FCN在KITTI road benchmark上获得outperforms
2. 从KITTI驾驶序列中提取的视觉上具有挑战性的场景的数据集，可用于进一步突出组合LIDAR数据和相机图像以进行道路分割的益处。

Related work：

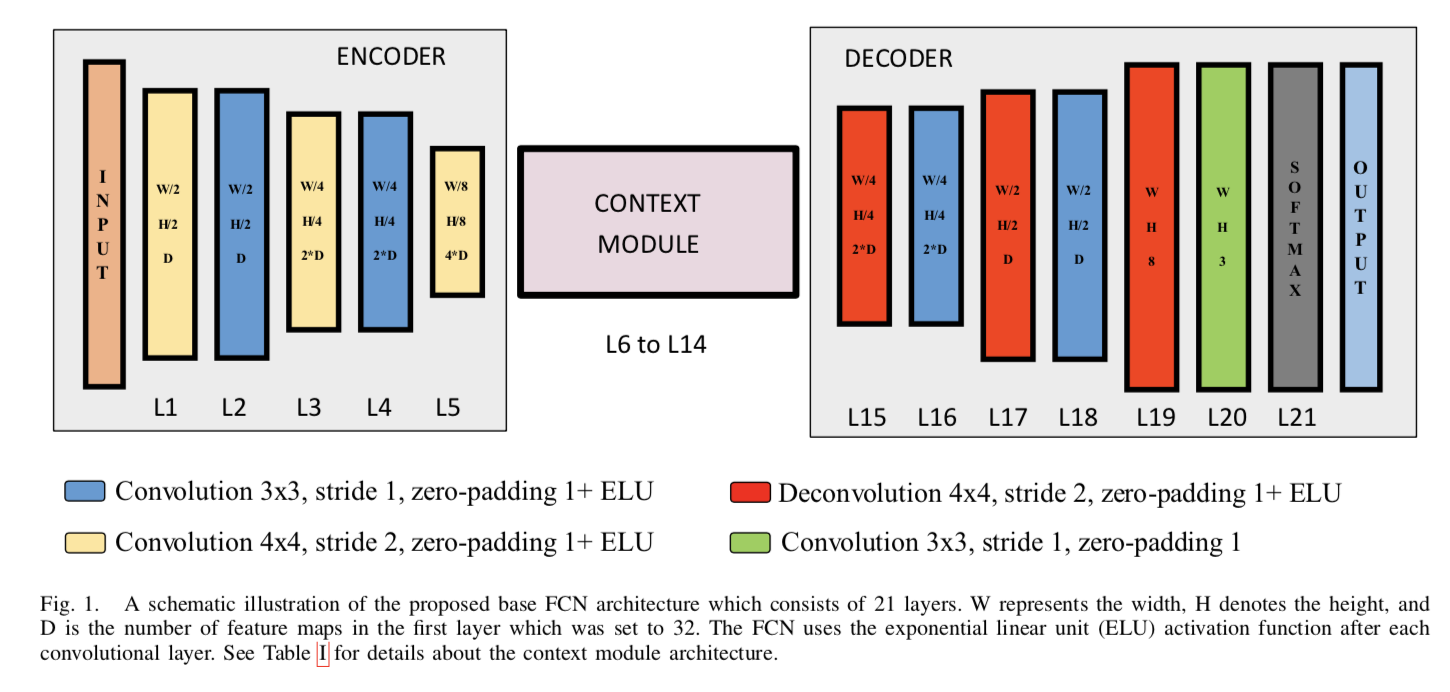
相关的一些文献：

1. [6]  M. Bertozzi and A. Broggi, “Gold: a parallel real-time stereo vision system for generic obstacle and lane detection,” *IEEE Transactions on Image Processing*, vol. 7, no. 1, pp. 62–81, Jan 1998.
2. [7]  M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, “Multinet: Real-time joint semantic reasoning for autonomous driv- ing,” *arXiv preprint arXiv:1612.07695*, 2016.
3. [8]  Z. Chen and Z. Chen, “Rbnet: A deep neural network for unified road and road boundary detection,” in *International Conference on Neural Information Processing*. Springer, 2017, pp. 677–687.
4. [9]  L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde, “Fast lidar-based road detection using fully convolutional neural networks,” in *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE, 2017, pp. 1019–1024.
5. [10]  P. Y. Shinzato, D. F. Wolf, and C. Stiller, “Road terrain detection: Avoiding common obstacle detection assumptions using sensor fusion,” in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 687–692.
6. [11]  L. Xiao, B. Dai, D. Liu, T. Hu, and T. Wu, “Crf based road detection with multi-sensor fusion,” in *2015 IEEE Intelligent Vehicles Symposium (IV)*, June 2015, pp. 192–198.
7. [12]  L. Xiao, R. Wang, B. Dai, Y. Fang, D. Liu, and T. Wu, “Hybrid conditional random field based camera-lidar fusion for road detection,” *Information Sciences*, 2017.

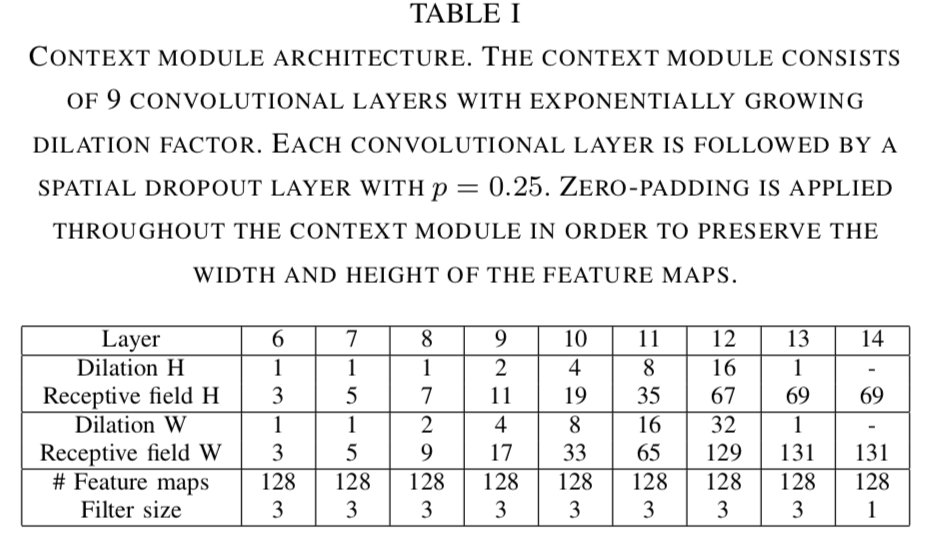
Network Architectures：

1. Base FCN

本文的神经网络采用全卷积编码解码器，中间包含一个中间上下文模块。



The encoder consists of 5 convolutional layers: 4 × 4 convolutions with stride 2 are used in order to downsample the input tensors thus reducing memory requirements. The context module consists of 9 convolutional layers with 3 × 3 kernels and exponentially growing dilation [21]. This makes it possible to quickly grow the network’s receptive field while limiting the number of layers.

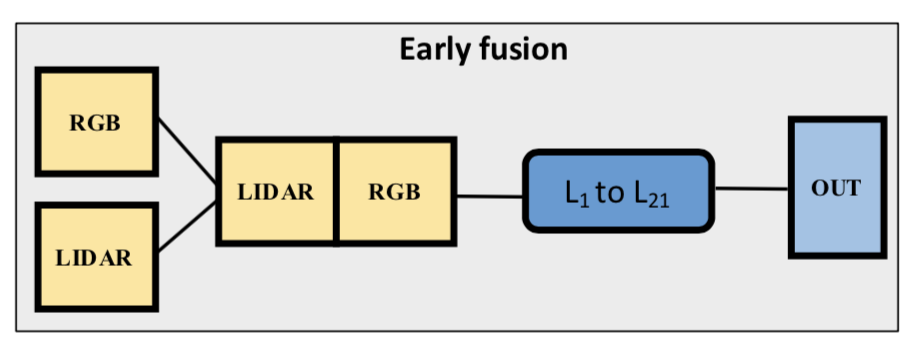


Decoder包含6个卷积层，用来upsample，通过使用具有4×4内核和步幅2的3个跨步卷积层来实现上采样。每层卷积层的激活函数是ELU，context module中的每一个卷积后都加上了p=0.25的dropout layer。

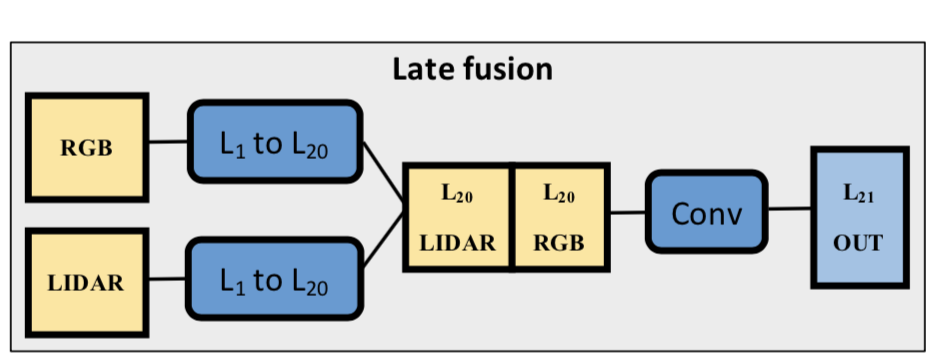
1. Early and late fusion

LIDAR point clouds转化成2D images要和camera images空间大小一致。

early fusion approach ：输入LIDAR和相机张量简单地在深度维度上连接，从而产生具有6个通道的张量（RGBZYX）

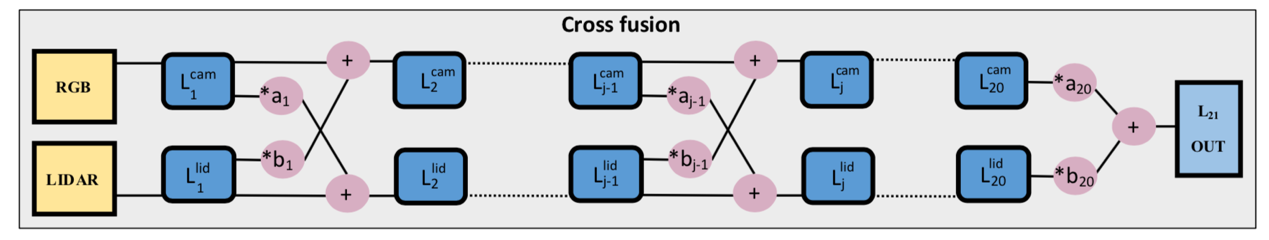


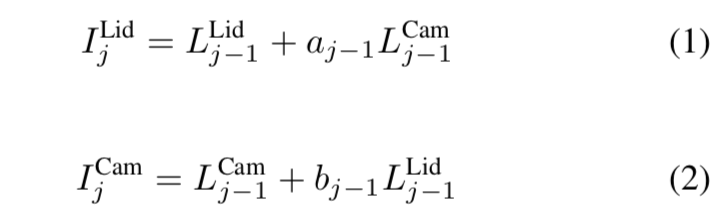
late fusion approach : 这里，LIDAR和摄像机信息的集成在两个独立处理分支的最末端进行



Early和late fusion需要人工决定fusion在哪一个stage被执行，而cross fusion就可以通过FCN自己学习。

1. Cross fusion



假设第j层的输入张量是和，要分别feed给和层，那么有如下的表达式：  


aj,bj ∈ R

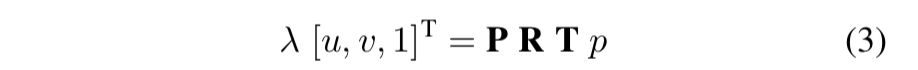
交叉融合参数被初始化为零，训练过程中会自动调整。

Data Preprocessing：

每一个LiDAR point cloud转化层一组3个2D images，以直接建立颜色强度和3D信息之间的correspondences。

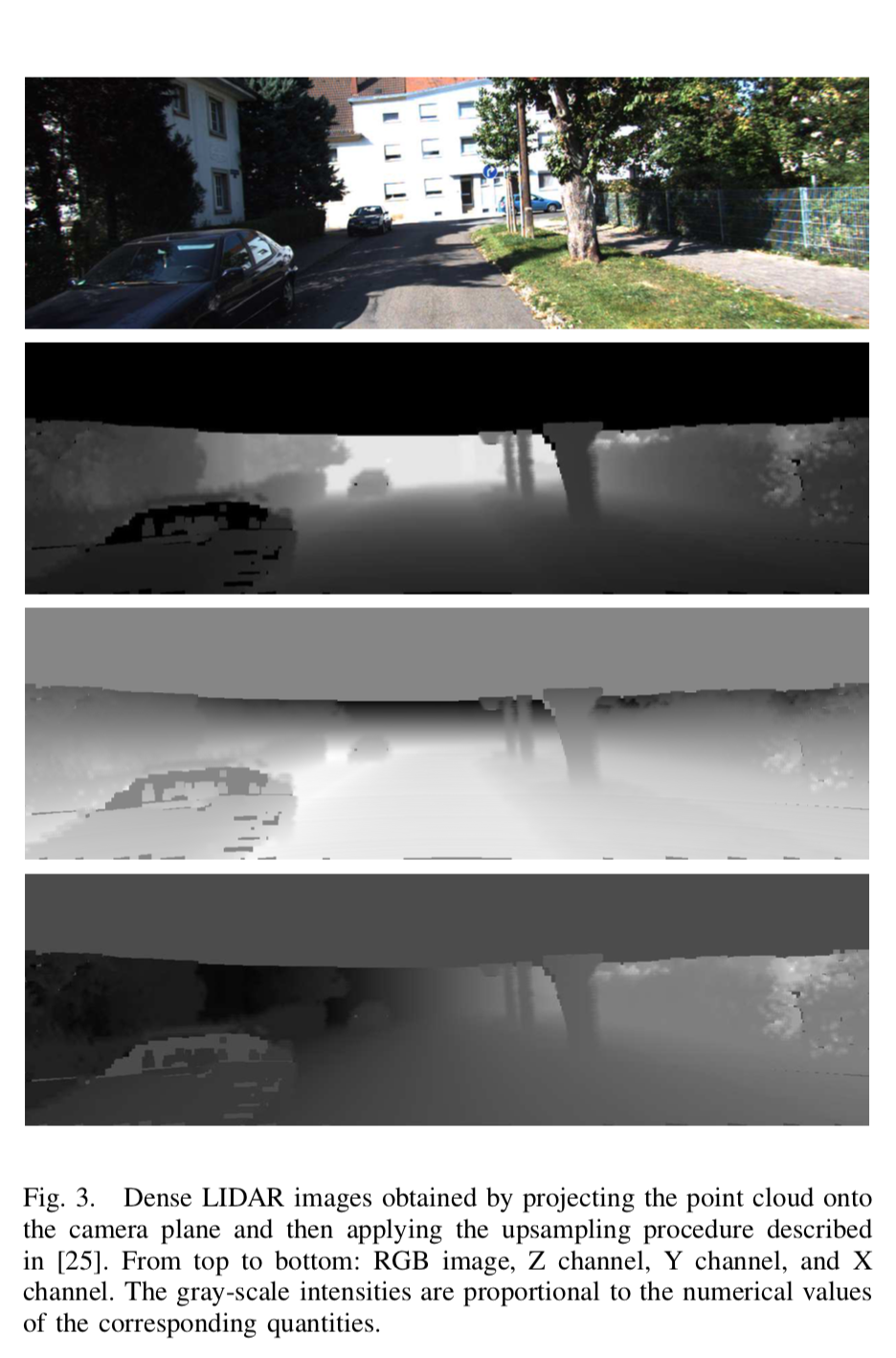
用Velodyne HDL64生成的一个point cloud有大约100kpoints，每一个点p在LiDAR坐标系统中有指定的坐标p = [x, y, z, 1]T

给定LIDAR-camera转换矩阵T、矫正矩阵R和相机投影矩阵P，可以计算纵坐标u,横坐标v，其中p的投影与相机平面相交：



上述转换运用于point cloud中的每一个point，但丢弃使得λ<0或当[u，v]落在图像之外时的point。

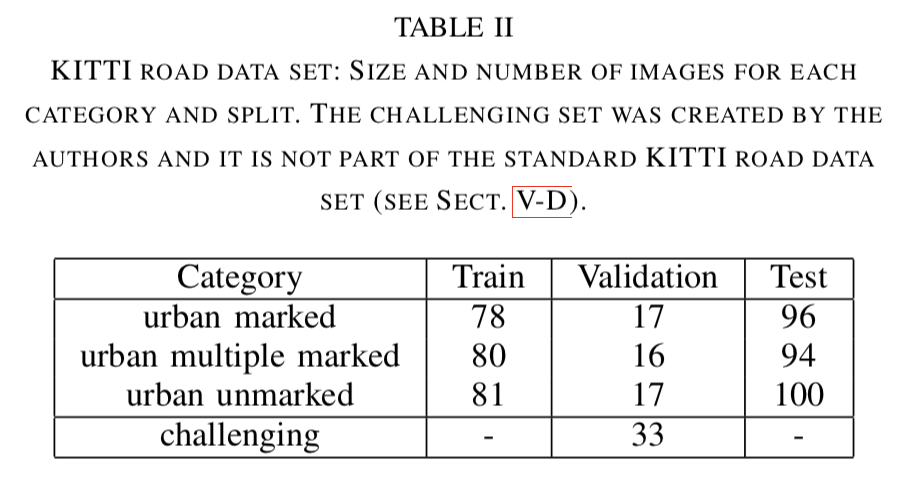
鉴于一个RGB图像每一个像素包含red、green和blue三个信息，因此上述转换过程生成三个images-X/Y/Z,其中每一个像素分别包含投影到XYZ平面上的3D point的x,y,z坐标。（就是将Points分别投影到XYZ平面生成3张images）相机和LIDAR之间的重要区别在于RGB图像对于每个像素具有有效值，而在LIDAR图像中，许多像素被设置为默认零值，因为没有激光束到达相应区域。因此出于这个原因，通常的做法是[14]，[25]，[26]在使用机器学习算法处理它们之前对LIDAR图像进行上采样。 这项工作利用了Premebida等人提出的方法。 [25]要做到这一点。 图3显示了通过应用此过程获得的密集LIDAR图像的示例。



Experiments and discussion

A.Data set

在这项工作中，考虑了五种不同的FCN：ZYX，RGB，早期融合，晚期融合和交叉融合。 ZYX表示仅在LIDAR图像上训练的基础FCN（参见Sect.III-A）。类似地，RGB是仅在相机图像上训练的基础FCN。早期，晚期和交叉融合是实施同名融合策略的FCN（参见第III-B节）。每个FCN都使用KITTI道路数据集进行训练，该数据集包括289个训练图像和290个在城市，乡村和高速公路环境中拍摄的测试图像。值得一提的是，大多数训练样例都是在相当理想的天气和光照条件下捕获的，这可能会模糊将摄像机图像与其他传感模式相结合的好处。因此，正如将在Sect。中描述的那样。 V-C是一个更具挑战性的场景的附加数据集，用于性能评估。表II提供了有关数据集拆分的更多信息。鉴于RGB图像由于整流过程而具有不同的尺寸，因此应用零填充以确保每个训练示例具有384×1248像素的相同尺寸。



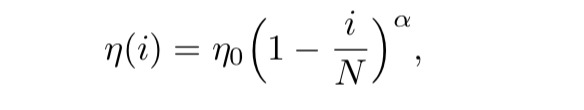
B. Training procedure

迭代轮数N=100k

Adam优化

每1000轮在验证集上比较一次FCN的performance保存权重参数

学习率如下定义：



where i denotes the current iteration number, η0 is the starting learning rate which was set to 0.0005, and α = 0.9. The batch size was set to 1. Given the small size of the data set, data augmentation was also carried out by applying random rotations in the range [−20◦ , 20◦ ] about the center of the images. The FCNs were implemented in PyTorch and trained using an Nvidia GTX1080 GPUs. The evaluation measures used in the following comparisons are the pixel- wise maximum F-measure (MaxF), precision (PRE), recall (REC), and average precision (AP) [29].