利用稀疏点云偏序关系的半监督单目图像深度估计

曾一鸣, 胡　瑜*, 韩银和, 李晓维

[1] (中国科学院计算技术研究所智能计算机研究中心 北京 100190)
[2] (中国科学院大学 北京 100049)
(huyu@ict.ac.cn)

摘　要: 为了减少传统基于学习的深度估计方法对大量稠密深度数据的依赖, 提出了一种基于偏序关系的深度估计方法. 首先对 RGB 图像进行超像素划分, 根据稀疏点云在超像素图像上的投影生成超像素的深度, 进而在超像素中心之间建立了深度偏序关系, 结合稀疏点云的实际深度值作为监督信息, 训练卷积神经网络估计场景深度. 在 NYU Depth v2 数据集上的实验结果表明, 该方法仅需稀疏点云就可达到 0.262 的平均相对误差, 优于之前国际同类方法, 甚至超过部分使用稠密深度数据的监督方法.

关键词: 深度估计; 偏序关系; 稀疏点云
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2019.17731

Exploiting Partial Order Relationship of Sparse Point Cloud for Semi-Supervised Monocular Image Depth Estimation

Zeng Yiming, Hu Yu*, Han Yinhe, and Li Xiaowei

[1] (*Research Center for Intelligent Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[2] (*University of Chinese Academy of Sciences, Beijing* 100049)

Abstract: To reduce the dependency of dense depth data, we propose a partial-order-relationship based depth estimation method, which learns monocular depth maps from the sparse 3D point cloud. Firstly, the RGB image is divided into superpixels, and the depth values of superpixels are generated according to the projection of the sparse point cloud on the superpixel image. Then a set of depth partial order relationship is generated at the center of superpixels. Subsequently, the depth partial order relationship of the RGB image and the actual depth value of the sparse point cloud are combined to train a convolutional neural network to estimate the depth map of the scene. The experimental results on the sparse point clouds sampled from NYU Depth v2 show that our method achieves an average relative error of 0.262 with sparse point cloud, which is better than the conventional methods, and even better than some supervised methods using dense depth data.

Key words: depth estimation; partial order relationship; sparse point cloud

收稿日期: 2019-01-14; 修回日期: 2019-06-13. 基金项目: 空间智能控制技术实验室开放基金课题(HTKJ2019KL502003); 中国科学院计算技术研究所创新课题(20186090). 曾一鸣(1991—), 男, 博士研究生, 主要研究方向为自动驾驶感知; 胡　瑜(1975—), 女, 博士, 研究员, 博士生导师, CCF 会员, 论文通讯作者, 主要研究方向为自主导航、自动驾驶、算法加速; 韩银和(1980—), 男, 博士, 研究员, 博士生导师, CCF 会员, 主要研究方向为计算机体系结构和芯片、智能硬件、机器人芯片系统; 李晓维(1964—), 男, 博士, 研究员, 博士生导师, CCF 会员, 主要研究方向为集成电路测试与诊断、验证、可靠性设计、无线传感网络.

1 简　介

获得场景深度是机器人进行环境三维重建的基础, 场景深度的准确性直接影响到机器人所构建地图的好坏. 综合考虑需求与成本, 现有机器人, 如 PR2, AMIGO, Pioneer 和 UBR-1 等, 通常搭载 RGB 相机与低线数的激光雷达来协同探测周围的环境. 基于激光雷达的稀疏三维点云所生成的稀疏环境地图, 难以满足路径规划、物体识别和抓取等依赖稠密地图的任务需求. 为了获取稠密的深度图, 从单目图像中估计深度的方法引起了学者的强烈兴趣.

传统基于图模型的单目深度估计方法[1]通常包含以下步骤: (1) 预先使用超像素分割等方法将图像分割为若干块, 并假设此图像块中所有像素的深度值相同; (2) 分别选取图像块中的绝对深度特征和相对深度特征, 估计各图像块的绝对深度和相邻块的相对深度; (3) 构建后端求解模型, 如马尔可夫随机场(Markov random field, MRF)等, 通过后端模型建立局部特征和深度之间的相关关系及不同图像块之间的相关深度关系, 并用稠密深度数据训练模型; (4) 使用训练好的模型估计场景深度.

近年来, 深度神经网络在图像分类、物体识别、语音识别等领域取得了突破性进展, 深度神经网络也被用于单目图像的场景深度估计. Eigen 等[2]提出使用深度神经网络直接拟合从 RGB 图像到深度图的映射, 估计场景的深度, 并设计了一种多尺度的卷积神经网络(convolutional neural network, CNN)端到端地回归深度. 首先用粗尺度的网络估计场景大概的深度范围, 然后使用细尺度的网络微调各像素点的深度值. Laina 等[3]将全卷积网络(fully convolutional network, FCN)用于深度估计, 在 FCN 的基础上提出一种新的快速上采样结构, 然后结合深度估计和同时定位与地图构建(simultaneous localization and mapping, SLAM), 完成稠密三维建模. Liu 等[4]将 CNN 与图模型结合, 将 CNN 作为特征提取器, 提取图像中的深度特征, 然后结合条件随机场(conditional random fields, CRF)等模型建立深度特征之间的联系并估计深度. 上述从单目图像估计深度的方法都会受到尺度不变性问题的影响, 但通过结合 RGB 与稀疏的点云估计深度[5-6], 利用点云的绝对深度值, 能够解决尺度不变性问题.

上述深度估计方法需要大量经过校准的(RGB, Depth)图像对用以监督深度神经网络的训练, 这就需要预先通过 RGB-D 相机采集大量 RGB 图像与稠密的深度图像. 而稀疏的点云既可以来自低成本激光雷达, 也可以来自视觉建图时得到的稀疏特征点, 比稠密深度数据更易获取, 因此, 通过稀疏的点云学习场景具有重要的研究意义.

本文提出了一种利用图像内深度偏序关系的单目深度估计方法, 将深度值的回归问题转化为图像中不同位置的远近分类问题. 根据点云的深度值, 在图像超像素中心生成深度偏序关系, 结合深度偏序关系与点云中的实际深度值训练 CNN, 用训练后的 CNN 估计场景深度. 本文方法在 NYU Depth v2 数据集测试所得相对平均误差为 0.262, 误差低于国际同类方法.

2 相关工作

根据训练集中包含深度数据的不同, 基于学习的深度估计方法可以分为监督学习、半监督学习和无监督学习的方法.

2.1 基于监督学习的深度估计方法

传统深度估计方法通常基于图模型. 人类视觉系统能够提取场景中存在的单目深度线索, 从二维图像推导出三维结构信息. 受此启发, Saxena 等[1,7]使用超像素分割算法将图像分割为多个图像块, 对各图像块分别选取绝对深度特征和相对深度特征以构建不同的深度值, 然后使用 MRF 建立不同图像块之间的深度关系, 求解全局深度与各图像块深度, 并将此方法扩展成二维图像重建三维场景的系统.

在基于深度神经网络的单目图像深度估计方法中, Eigen 等[2]率先提出了多尺度的 CNN 方法, 直接从输入的场景图像回归场景的深度图; 其网络中包含 2 个子网络, 粗尺度的网络估计大概范围, 细尺度网络调整场景的深度细节, 取得了远超传统方法的准确度. 随后, Eigen 等[8]提出一种针对像素级任务的 CNN 架构, 将粗细尺度网络[2]拓展为三尺度 CNN, 在深度估计、表面法向量计算和语义分割方面都取得了很好的效果. 继 Eigen 之后, Laina 等[3]将 FCN 引入到深度估计问题中, 设计了一种高效的上采样结构, 并引入 berHu 损失函数, 提高了场景深度值较小时的训练效率, 获得了更高的精度. Zeng 等[9]针对深度线索在神经网络传播

过程中损失的问题，建立了深度线索前递结构与局部相对几何结构损失项，在提高精度的同时保留了场景中更多的几何结构. 此外，直接使用 CNN 进行深度估计与表面法向量估计[10-11]、语义分割[12-13]等多种任务的协同训练可以有效地提高深度估计的精度.

在基于 CNN 与图模型协同建模的深度估计方法中，Liu 等[4]提出一种 CNN 和 CRF 的联合模型，首先使用超像素分割算法将图像分割成小的图像块，然后利用 CNN 建立超像素分割后图像块的势函数及相邻位置上图像块的势函数，并统一到 CRF 中，实现对此联合模型的学习. Roy 等[14]使用浅层 CNN 作为条件随机森林的叶子结点，随着森林的遍历浅层 CNN 叠加为多层 CNN，通过这样的形式可以利用图像块与其邻居的关系估计图像深度. 进一步地，Xu 等[15]使用连续 CRF 融合 CNN 中不同尺度的特征并建立序列模型实现稠密深度图的估计.

由于单目图像深度估计会受到尺度不变性问题的影响，Tateno 等[16]将深度神经网络估计的场景深度融合到单目 SLAM 中，提高了建图精度，并且缓解了单目 SLAM 中尺度不变性的问题. Ma 等[5]在 Laina 等[3]方法的基础上，将稀疏的点云编码，与 RGB 图像拼接后输入到 CNN 中进行深度估计，通过点云数据中的绝对深度值辅助确定尺度，提升深度估计的准确性. Liao 等[6]根据单线激光扫描得到的切面轮廓线，在切面的垂直方向生成一个尺度确定的深度参考平面，然后使用 CNN 估计相对深度参考平面的场景深度差，最后生成完整的深度图.

2.2 基于半/无监督的深度估计方法

上述监督学习方法需要使用大量的稠密深度数据做标签来训练模型，为了减少对稠密深度数据的依赖，近年来，学者提出了新方法来进行深度估计. Zhou 等[17]利用双目图像中的视差来训练深度估计网络，通过 CNN 估计双目图像中左图像的视差，通过视差将左图像映射到右图像，并计算与真实右图像间的光度损失，以此作为损失函数来训练 CNN；训练后的 CNN 可以估计单目图像的视差，进而估计左图对应的深度图. 在此基础上，Kuznietsov 等[18]提出了半监督的 CNN 估计方法，其损失函数中结合双目图像估计的视差项与三维点云对应像素点位置的深度差项，完成了深度估计 CNN 的训练.

另一种减少稠密深度数据依赖的方法是使用相对位置关系. Zoran 等[19]使用输入图像中不同点对的相对位置关系来训练网络模型，使之能判断不同位置之间的相对远近；估计深度时使用训练后的模型对不同图像块进行排序，并使用全局优化方法解决排序冲突，得到各点的深度顺序，最后在各点之间的像素上插值以产生所有像素对应的深度. Chen 等[20]制作了一个室外的相对深度数据集，数据集中每幅 RGB 图像包含人工标记的相对远近关系，并通过人工标记的相对远近关系训练 CNN 来估计场景深度.

本文利用深度偏序关系学习场景深度，在 RGB 图像中产生一系列深度偏序关系，并结合 RGB 图像中的深度偏序关系与稀疏三维点云的绝对深度值来训练 CNN，使用训练后的 CNN 估计场景深度.

3 本文方法

本文方法有 2 个主要步骤：(1) 从 RGB 图像中选择适合进行深度比较的点，并根据稀疏三维点云的绝对深度值产生 RGB 图像上点间的深度偏序关系；(2) 将单目深度值的回归问题转化为图像内不同位置间的远近分类问题，结合 RGB 图像中的深度偏序关系与稀疏三维点云的绝对深度值来训练 CNN.

3.1 深度比较位置关系生成

如果直接使用稀疏三维点云中的绝对深度值来训练 CNN，尽管思路简单，但所得深度图的精度差，甚至训练过程不收敛，难以应用于实际的机器人任务. 受 Chen 等[20]方法的启发，本文将深度值的回归问题转化成 RGB 图像上不同点对间的分类问题，通过 CNN 判断 RGB 图像中不同位置的远近，并结合稀疏的三维点云数据生成准确的深度图，以此减少对大量稠密深度数据和人工标注的依赖. 首先，从 RGB 图像中选择 N 个点. 为便于说明，用边 $E_{i,j}$ 表示需要比较的一对位置点 p_i，p_j. 对于选点策略，考虑到场景深度不连续通常会导致对应 RGB 图像中像素光度的变化，从而产生剧烈的梯度变化[19]. 因此，进行远近比较的点应远离这些梯度剧烈变化的区域，尽量处于相对同质的区域中心；而连接这些点的边 $E_{i,j}$ 应该尽量穿过这些梯度剧烈变化的区域，从而能够更好地观察到深度变化，使得边的顶点所在区域的深度

有明显不同.

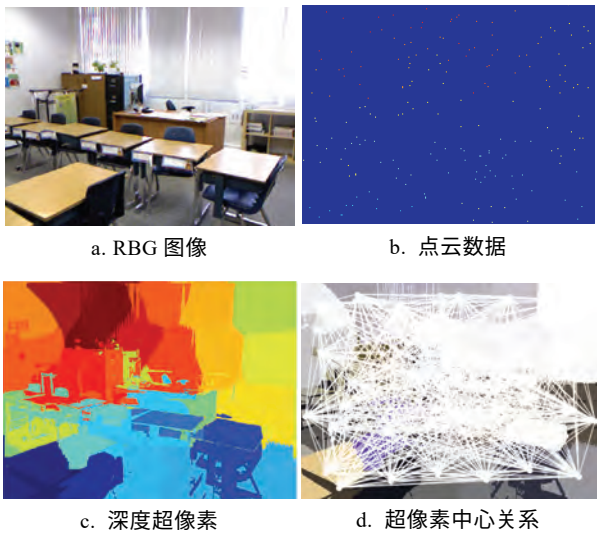
满足上述要求的点是超像素的中心. 超像素由一系列位置相邻且颜色、亮度、纹理等特征相似的像素点组成，其区域内的像素较为相似，且图像块梯度变化较小，一般不会破坏图像中物体的边界信息. 超像素的中心处于较为同质的区域内，与附近的像素较为相近，远离剧烈变化的边界，因此适合用于远近比较. 将 RGB 图像进行超像素分割后，选取所有超像素中心，两两组合作为比较深度远近时使用的点对.

下一步通过三维点云的深度值来比较点对之间的远近，产生 RGB 图像上的深度偏序关系. 由于超像素内的像素同质，可视做超像素内所有像素的深度相近[4]. 对于 RGB 图像上所有的超像素区域 $\{S_i\}$，将三维点云中的点 $P(x,y,z)$ 投影到 RGB 图像平面上. 若 $P(x,y,z)$ 在 RGB 图像平面上的投影 $P'(x',y')$ 正好在某一超像素区域 S_i 内，则此超像素对应的深度等于点云中此数据点的深度，即 $\exists P'(x',y')\in S_i$ 则 $z_{S_i}=z_p$；如果此超像素区域内没有点云数据的投影，则丢弃此块超像素区域，并且将此超像素的深度值设置为零；如果此超像素区域内有多个点云数据的投影，则取最靠近中心的点. 最后，两两比较各个超像素的远近关系，并打上相对位置标签 $F_{\text{ordinal}}(i,j)$，若 $p_i\in S_i$, $p_j\in S_j$，则 $E_{i,j}=E_{p_i,p_j}=F_{\text{ordinal}}(i,j)$，其中，

$$F_{\text{ordinal}}(i,j)=\begin{cases}1, & (p_i-p_j)>\varepsilon\\ 0, & |p_i-p_j|\leqslant\varepsilon\\ -1, & (p_j-p_i)>\varepsilon\end{cases}.$$

考虑到深度传感器本身存在的测量偏差，此处取阈值 ε=0.1.

图 1 是由 RGB 图像与稀疏三维点云生成深度超像素与超像素中心偏序关系的示意图. 其中，图 1a 为场景的 RGB 图像，图 1b 为所测量到的稀疏点云，将图 1a 使用超像素分割并且使用图 1b 中的点云绝对深度值填充，可以得到图 1c；进而得到超像素中心位置的相对远近，生成 RGB 图像上不同点的深度偏序关系. 确定图像内不同位置的偏序关系后，结合 RGB 图像中的深度偏序关系与稀疏三维点云的绝对深度值来训练 CNN.



a. RBG 图像　　b. 点云数据

c. 深度超像素　　d. 超像素中心关系

图 1　点对选择及远近比较示意图

3.2　CNN 模型设计

为了得到像素级的稠密深度图，本文设计了一种带跳跃连接结构的自编码-解码 CNN，如图 2 所示. 其中编码器基于 ResNet-50[21]，用于提取不同 RGB 图像的特征. 解码器聚合多尺度特征，学习深度细节并避免插值的不准确性. 解码器的上采样操作采用 pixel shuffle 方式[22]，上采样后的特征经过残差卷积模块后与浅层的特征相结合[9]，最

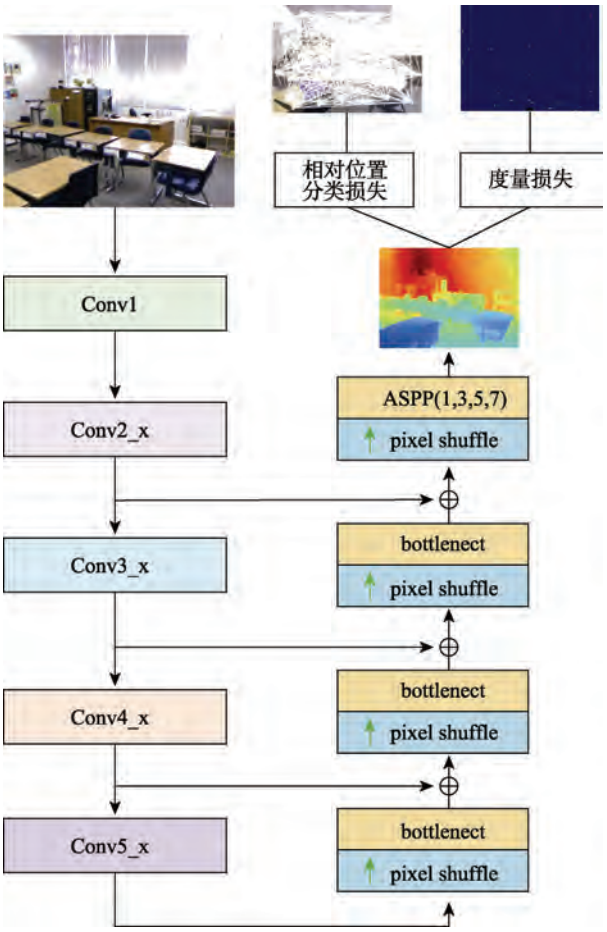


图 2　网络结构设计

后使用带孔空间池化金字塔(atrous spatial pyramid pooling, ASPP)模块聚合不同尺度的特征，经由1×1的卷积层输出稠密深度图.

其中，每个 pixel shuffle 模块将特征图上采样2倍，同时特征图通道数下降为原来的1/4，如图3所示. 使用 pixel shuffle 聚合不同通道上的特征以产生更高分辨率的特征图，无需引入额外的数值，并且 pixel shuffle 的运算比插值更快，同时减少了特征的通道数，进而减少了后续卷积运算次数，在训练与测试时速度更快. 因此, pixel shuffle 更适合用于计算存储资源相对紧张且有实时性要求的自主移动无人系统中.

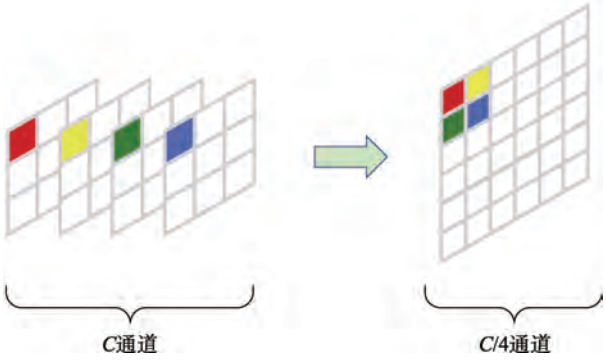


图 3 pixel shuffle 示意图

在解码器网络中使用 ASPP 聚合不同感受野下的特征，多种尺度特征使得在恢复深度时能结合不同尺度的信息，以产生更加丰富的形状细节. 本文解码器的 ASPP 设计了 4 种不同扩张率的带孔卷积，其扩张率其分别为(1, 3, 5, 7)，每个卷积核大小都统一为3×3，如图4所示，其中Conv(kernel_size, rate)中的2个参数分别表示卷积核的大小及其扩张率.

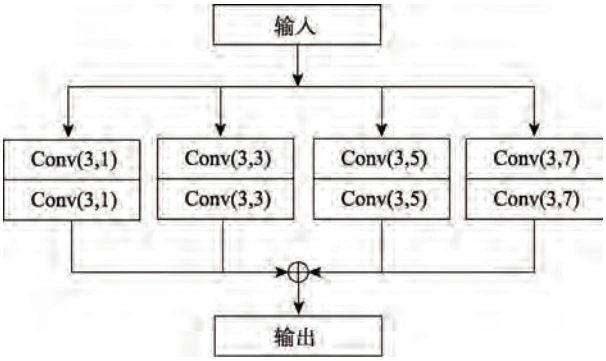


图 4 ASPP 结构图

3.3 损失函数

由 CNN 生成深度图之后，还需要设计损失函数来约束深度图上偏序关系与 RGB 图上深度偏序关系的一致性，并且使深度图像素的具体数值接近稀疏三维点云数据的绝对深度值. 本文设计的损失函数由点对的相对位置分类损失与深度值的度量损失2项组成. 首先考虑点对的相对位置分类损失，对于每一组点对 p_i, p_j，定义

$$L_{\text{rank}}(i,j)=\begin{cases}\log\left(1+e^{-E_{i,j}(z_i-z_j)}\right), & E_{i,j}\neq 0\\ (z_i-z_j)^2, & E_{i,j}=0\end{cases}.$$

其中，z_i, z_j 表示点 p_i, p_j 处的估计深度；边 $E_{i,j}$ 表示 p_i, p_j 的相对位置关系. 当深度相等时鼓励小的误差，否则鼓励大的误差，以此训练 CNN 判断位置的远近. 然后将所有点对的损失求和构成相对位置分类损失项 $L_{\text{ordinal}}=\sum_i\sum_j L_{\text{rank}}(i,j)$.

其次考虑深度值的度量损失，对于 CNN 估计出的深度 z_i 与稀疏三维点云中的绝对深度值 z_i^*，使用 berHu 损失函数

$$L_{\text{berHu}}(z_i,z_i^*)=\begin{cases}\left|z_i-z_i^*\right|, & \left|z_i-z_i^*\right|\leqslant c\\ \dfrac{\left(z_i-z_i^*\right)^2+c^2}{2c}, & \left|z_i-z_i^*\right|>c\end{cases}.$$

其中，$c=\dfrac{1}{5}\max\left(\left|z_i-z_i^*\right|\right)$.

该 L_{berHu} 函数在 L_1 损失与 L_2 损失之间表现出良好的平衡. L_2 项对具有较大差值的像素赋予高权重，而 L_1 对差值较小的像素的梯度影响要大于 L_2. 三维点云中所有的点可在 CNN 估计的深度图中找到对应的点，则深度的度量损失项为

$$L_{\text{metric}}=\sum_i L_{\text{berHu}}\left(z,z^*\right).$$

最后2项相加得到训练网络使用的损失函数

$$L_{\text{total}}=L_{\text{ordinal}}+\beta L_{\text{metric}}.$$

其中，β 是一个经验常数. 上述损失函数结合了点对之间的深度偏序关系与绝对深度值，鼓励估计的深度图上偏序关系与 RGB 图像上偏序关系一致；同时，估计的深度图在数值上尽可能地接近三维点云中的深度值.

CNN 训练完成后，可直接用于深度估计，无需对 RGB 图像进行分割产生点对，仅需要将 RGB 图像传入网络，网络输出即为场景的深度图.

4 实验结果与分析

本节将对本文方法进行全面分析，评估本文方法在室内深度数据集 NYU Depth v2[23]上的性能表现，以及稀疏的三维点云中点云的数量对深度

估计的影响, 定量和定性地评价模型的深度估计结果. 因本文方法需要在 RGB 图像的不同位置判断远近, 本节还将讨论不同的位置选择策略对深度估计的影响.

4.1　实验环境与数据集

本文方法在开源深度学习框架 PyTorch 上实现, 所用计算平台搭载 Intel i7-6850k CPU 与 NVIDIA GeForceGTX 1080Ti GPU. 在训练过程中, 使用在 ImageNet 上预训练的 ResNet-50 网络模型参数来初始化本文模型的编码器, 并使用随机梯度下降(stochastic gradient descent, SGD)方法训练整个模型, 其中, 设置 β=0.5, batch size 设为 8, momentum 设为 0.9, 学习速率初始化为 10^{-3}, 每遍历数据集 10 次后将学习率降为原来的 10%, 并且在验证集误差上升时, 提前停止.

本文所用训练集与测试集与文献[2,5,19]相同. 模型先在原始数据上训练, 为了改善数据集的多样性和可变性并获得更多训练样本, 对样本进行缩放、裁剪、翻转处理. 为了得到稀疏三维点云, 本文实验从 RGB-D 图像中随机采样不同数目的数据生成三维点云.

4.2　实验结果

为衡量本文方法的效果, 在测试集上进行评估实验, 评估所用指标包括平均相对误差(average relative error, REL)、平方相对误差(square relative error, SQR-REL)和均方根误差(root mean squared error, RMSE), 各指标计算方法如下:

$$\mathrm{REL}=\frac{1}{N}\sum_{i\in P}\frac{\left|d_i-d_i^*\right|}{d_i^*},$$

$$\mathrm{SQR\text{-}REL}=\frac{1}{N}\sum_{i\in P}\frac{\left(d_i-d_i^*\right)^2}{d_i^*},$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i\in P}\left(d_i-d_i^*\right)^2}.$$

4.2.1　深度准确性评估

为了衡量本文方法深度估计的准确性, 使用上述指标进行评估, 并与目前最具代表性的几种深度估计方法进行对比, 结果如表 1 所示. 其中, 本文(Dense)表示本文模型使用稠密深度数据为标签, 并只使用 L_{metric} 作为损失函数来监督网络训练; 本文(Noise)表示在稠密深度数据中加入了随机噪声; 本文(Sparse)表示本文模型使用稀疏点云作为标签, 并使用 L_{total} 作为损失函数来监督网络训练.

表 1　不同方法结果对比

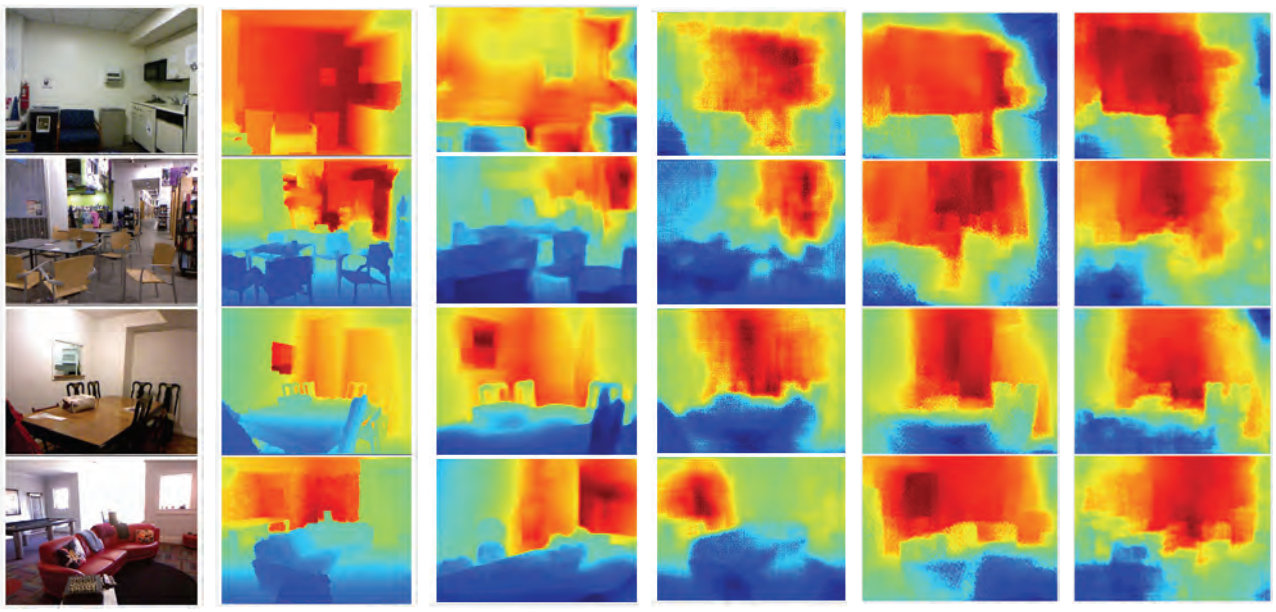
方法	标签	REL	SQR-REL	RMSE
Make3D[7]	稠密深度	0.349		1.214
Eigen 等[8]	稠密深度	0.158	0.121	0.641
Liu 等[4]	稠密深度	0.230		0.824
Laina 等[3]	稠密深度	0.127		0.573
本文(Dense)	稠密深度	0.153	0.117	0.532
本文(Noise)	稠密深度	0.159	0.122	0.579
Chen 等[20]	人工标记	0.360	0.460	1.130
Zoran 等[19]	稀疏深度	0.400	0.540	1.200
本文(Sparse)	稀疏深度	0.292	0.264	0.844

在使用稠密深度信息进行训练的情况下, 所得深度值均为实际值. 由深度估计结果可以看出, 本文方法深度估计的误差小于其他方法, 并且依赖稠密数据的方法易受噪声影响. 在使用稀疏三维点云进行训练的情况下, 依靠稠密深度数据方法的性能会严重退化. Chen 等[20]使用 CNN 学习人工标注的对称点远近来估计深度, Zoran 等[19]通过对不同超像素块进行排序的方法恢复深度; 并且为了避免单目深度的尺度问题, 都将所得深度正则化. 本文方法在图像内不同位置鼓励场景深度相近处产生相同深度, 鼓励相差较远处产生深度落差, 并结合三维点云的绝对深度值, 确定深度的尺度, 可从稀疏点云中学习场景深度. 由上述 3 项评估指标结果可见, 本文估计场景深度的误差小于其他方法.

本文方法所得深度图也有更清晰的场景结构和更丰富的场景细节, 如图 5 所示. 这主要是因为本文方法中的解码器结构较 SegNet[24]与 Laina 等[3]方法中解码器结构更优, 本文解码器在上采样过程中没有引入额外的数值, 通过混淆相邻通道上的特征值以产生通道数更少、分辨率更高的特征图, 并且通过 ASPP 聚合不同感受野的特征使得最后的结果具有更丰富的细节. 此外, 在本文解码器上采样过程中, CNN 中的特征图的尺寸每放大 1 倍, 特征图的通道数就降为原来的 1/4, 后续卷积层的参数数量也减少到原来的 1/4, 这样的解码器形式参数少、运算快, 更适合资源受限的应用.

4.2.2　点云稀疏性对深度估计的影响

为了验证点云数目对 CNN 深度估计的影响, 本文随机采样了不同数目的点云数据, 以这些数据点的深度值作为 CNN 的标签, 在损失函数中仅使用 L_{metric} 来训练网络, 其余实验设置与之前保持一致. 训练所得 CNN 在 NYU Depth v2 的测试集上进行深度估计, 结果如表 2 所示.



a. RGB 图像 b. 场景深度图 c. Eigen 等[2] d. 本文(Full) e. Chen 等[20] f. 本文(PC)

图 5 深度结果对比图

表 2 基于稀疏点云训练结果

点云数目	REL	SQR-REL	RMSE
300	失败	失败	失败
1 000	失败	失败	失败
3 000	0.990	2.695	2.905
6 000	0.978	2.639	2.876
10 000	0.965	2.572	2.840
30 000	0.896	2.237	2.655
60 000	0.792	1.789	2.383
100 000	0.646	1.246	1.999
本文(Dense)	0.153	0.117	0.532

可以看出，当点云很稀疏时，CNN 训练失败，随着点云数目的增加 CNN 开始收敛，能估计出一部分深度值，但是误差很大；随着点云数目进一步增加，深度值的误差出现下降趋势. 这是因为点云数目较少时，三维点云基本无法建立场景的环境模型，仅仅表示为场景中若干离散的点，没有明显的几何结构. CNN 估计的深度图中只有少量像素点对应这些离散的点云，导致 RGB 图像中潜在结构与深度图中结构的对应关系是缺省的，RGB 图像到深度图的映射不完整，使得模型训练不充分；而当点云较为稠密时，对环境的描述能力增加，RGB 图像中的结构与深度图的中结构对应，模型能学到 RGB 图像到深度图的映射. 但并不是所有点都有对应的深度值，相比使用完整深度图训练的模型，其结果仍有很大差距.

为了使得所有深度值都有对应的标签，假设 RGB 图像中一个超像素内所有像素的深度值都是相同的，并用三维点云中的深度值填充对应的超像素，详见第 3.1 节，这样得到一幅近似完整的深度图用于训练模型，所得深度估计结果如表 3 所示. 其中，本文(SparseDS)表示使用深度偏序关系与深度超像素训练模型所得结果.

表 3 基于深度超像素训练结果

超像素数目	REL	SQR-REL	RMSE
300	0.374	0.530	1.324
1 000	0.363	0.511	1.298
3 000	0.361	0.496	1.278
6 000	0.367	0.500	1.280
本文(Sparse)	0.292	0.264	0.844
本文(SparseDS)	0.297	0.366	1.020

可以看出，基于深度超像素所得结果较之稀疏的点云有所提升，但是该方法基于超像素内部深度相同的假设，超像素边缘的深度误差大，缺少深度的细节. 而本文将深度的回归问题转化成图像内不同位置的远近分类问题，判断不同超像素中心的远近，不关注超像素深度的细节，并结合点云的深度，取得了更小的估计误差. 在训练时联合深度偏序关系与深度超像素可以加速网络收敛的速度，减少训练时间，但是深度超像素的近似值降低了深度的准确性.

4.2.3 深度距离分析

为了衡量本文方法在不同深度距离范围的准

确性, 本文按场景深度距离由近到远划分为 5 级, 对不同范围的深度估计结果分别进行评估, 如表 4 所示. 随着场景深度的增加, 深度估计的相对偏差减小, 但是绝对偏差增大. 这与深度本身的性质相关, 具有较大景深的场景投影时会比具有较小景深的场景压缩更多, 恢复时更易产生较大偏差; 并且在距离较远的时, 传感器本身的噪音对数据的质量也有干扰. 在测试过程中深度距离为 8~10 m 时的误差较小, 这是由于该范围已接近传感器量程上限, 并且由于遮挡等原因, 传感器能采集到的数据较深度距离 8 m 以内的深度值少很多, 因此, 该范围的测试结果有较大偏差.

表 **4**　各范围深度误差表

距离/m	本文(Dense)			本文(Sparse)		
	REL	SQR-REL	RMSE	REL	SQR-REL	RMSE
0~2	0.193	0.118	0.345	0.457	0.408	0.868
2~4	0.129	0.093	0.452	0.319	0.504	1.180
4~6	0.098	0.122	0.526	0.229	0.883	1.586
6~8	0.074	0.169	0.534	0.120	0.556	0.957
8~10	0.047	0.161	0.427	0.059	0.240	0.408

4.2.4　点对位置的影响

为了将深度值的回归问题转化为图像中不同位置处远近的分类, 本文使用超像素分割算法 SLIC(simple linear iterative cluster)对图像进行划分, 取超像素的中心作为判断深度远近的位置点. 为了验证图像内位置对深度估计的影响, 本文也使用随机采样的方式从三维点云中选择数目相当的点对, 并投影到 RGB 图像上作为判断深度远近的位置, 并训练同样的网络模型, 所得结果对比如图 6 所示.

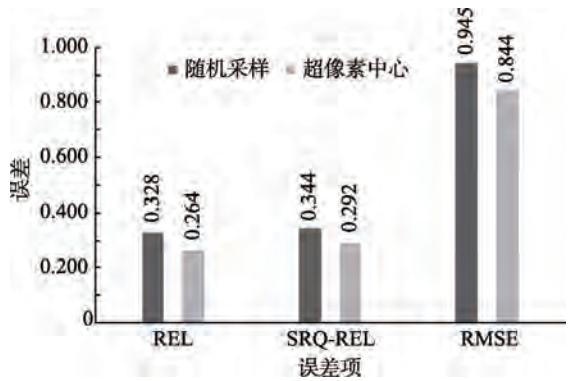


图 6　图像内位置选择结果对比图

使用超像素中心点作为图像内深度对比的位置比随机采样位置更优. 因为在场景的深度不连续处通常会导致对应 RGB 图像中像素光度的变化, 超像素自然地跟随图像中的梯度, 其中心处于相对较为同质的区域内, 内部的变化小, 而超像素中心之间的变化大; 这样能够更好地反映出场景深度变化, 不同点所在区域深度有明显不同, 更有利于深度神经网络学习不同位置间的远近关系.

5　结　语

本文针对从稀疏点云中学习场景深度的问题展开研究, 提出了一种利用图像内深度偏序关系的单目图像深度估计方法, 本文方法仅需要室内机器人采集的稀疏点云数据, 减少了对稠密深度图的依赖, 实验结果优于对比方法. 本文方法在室内深度估计上取得了较好的效果, 但是对于更为复杂的室外环境存在一定的限制, 特别是对于深度传感器无法测量的区域, 如天空等. 在未来的工作中, 将考虑在模型中导入语义信息, 进一步优化对场景远近的分类判断.

参考文献(References):

[1] Saxena A, Chung S H, Ng A Y. Learning depth from single monocular images[C] //Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2005: 1161-1168

[2] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014, 2: 2366-2374

[3] Laina I, Rupprecht C, Belagiannis V, *et al*. Deeper depth prediction with fully convolutional residual networks[C] // Proceedings of the 4th International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2016: 239-248

[4] Liu F Y, Shen C H, Lin G S. Deep convolutional neural fields for depth estimation from a single image[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 5162-5170

[5] Ma F C, Karaman S. Sparse-to-dense: depth prediction from sparse depth samples and a single image[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2018: 1-8

[6] Liao Y Y, Huang L C, Wang Y, *et al*. Parse geometry from a line: monocular depth estimation with partial laser observation[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2017: 5059-5066

[7] Saxena A, Sun M, Ng A Y. Make3D: learning 3D scene struc-

ture from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840

[8] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 2650-2658

[9] Zeng Y M, Hu Y, Liu S C, *et al*. GeoCueDepth: exploiting geometric structure cues to estimate depth from a single image[C] //Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press , 2017: 17-22

[10] Li B, Shen C H, Dai Y C, *et al*. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1119-1127

[11] Wang X L, Fouhey D F, Gupta A. Designing deep networks for surface normal estimation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 539-547

[12] Wang P, Shen X H, Lin Z, *et al*. Towards unified depth and semantic prediction from a single image[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 2800-2809

[13] Jiao J B, Cao Y, Song Y B, *et al*. Look deeper into depth: monocular depth estimation with semantic booster and attention-driven loss[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 55-71

[14] Roy A, Todorovic S. Monocular depth estimation using neural regression forest[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 5506-5514

[15] Xu D, Ricci E, Ouyang W L, *et al*. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 161-169

[16] Tateno K, Tombari F, Laina I, *et al*. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6565-6574

[17] Zhou T H, Brown M, Snavely N, *et al*. Unsupervised learning of depth and ego-motion from video[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6612-6619

[18] Kuznietsov Y, Stückler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2215-2223

[19] Zoran D, Isola P, Krishnan D, *et al*. Learning ordinal relationships for mid-level vision[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 388-396

[20] Chen W F, Fu Z, Yang D W, *et al*. Single-image depth perception in the wild[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: Curran Associates Press, 2016: 730-738

[21] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[22] Wang P Q, Chen P F, Yuan Y, *et al*. Understanding convolution for semantic segmentation[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2018: 1451-1460

[23] Silberman N, Hoiem D, Kohli P, *et al*. Indoor segmentation and support inference from RGBD images[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2012: 746-760

[24] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495