计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(9):382−388 [doi: 10.15888/j.cnki.csa.008690]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合 2D 激光雷达的室内单目深度估计①

杨　瑞, 朱　明

(中国科学技术大学 信息科学技术学院, 合肥 230027)
通信作者: 杨　瑞, E-mail: yangrui0@mail.ustc.edu.cn

摘　要: 在室内单目视觉导航任务中, 场景的深度信息十分重要. 但单目深度估计是一个不适定问题, 精度较低. 目前, 2D 激光雷达在室内导航任务中应用广泛, 且价格低廉. 因此, 本文提出一种融合 2D 激光雷达的室内单目深度估计算法来提高深度估计精度. 本文在编解码结构上增加了 2D 激光雷达的特征提取, 通过跳跃连接增加单目深度估计结果的细节信息, 并提出一种运用通道注意力机制融合 2D 激光雷达特征和 RGB 图像特征的方法. 本文在公开数据集 NYUDv2 上对算法进行验证, 并针对本文算法的应用场景, 制作了带有 2D 激光雷达数据的深度数据集. 实验表明, 本文提出的算法在公开数据集和自制数据集中均优于现有的单目深度估计.

关键词: 2D 激光雷达; 单目深度估计; 通道注意力机制; 跳跃连接; 深度学习

引用格式: 杨瑞,朱明.融合 2D 激光雷达的室内单目深度估计.计算机系统应用,2022,31(9):382–388. http://www.c-s-a.org.cn/1003-3254/8690.html

Indoor Monocular Depth Estimation by Fusing 2D LiDAR

YANG Rui, ZHU Ming

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: The depth information of a scenario is very important in indoor monocular vision navigation tasks. However, monocular depth estimation is an ill-posed problem with low accuracy. At present, 2D LiDAR is widely used in indoor navigation tasks, and the price is low. Therefore, we propose an indoor monocular depth estimation algorithm by fusing 2D LiDAR to improve the accuracy of depth estimation. Specifically, the feature extraction of 2D LiDAR is added to the encoder-decoder structure, and skip connections are used to acquire more detailed information of monocular depth estimation. Additionally, a method using channel attention mechanisms is presented to fuse 2D LiDAR features and RGB image features. The algorithm is verified on the public dataset NYUDv2, and a depth dataset with 2D LiDAR data for the application scenarios of the algorithm is established. Experiments indicate that the proposed algorithm outperforms the state-of-art monocular depth estimation on both public dataset and self-made dataset.

Key words: 2D LiDAR; monocular depth estimation; channel attention mechanism; skip connection; deep learning

场景的深度信息广泛应用于 SLAM[1], 3D 目标检测[2] 等算法中, 这些算法是导航任务中的关键算法, 因此深度估计任务在导航中至关重要. 本文工作聚焦于室内导航场景下的单目深度估计算法. 在室内简单场景下, 常用的深度传感器: 激光雷达、深度相机、双目相机等, 因造价昂贵应用并不广泛, 而造价低廉的单目相机和 2D 激光雷达成为室内导航机器人的基本配置. 但 2D 激光雷达提供的深度信息有限, 仅有 2D 平面的局部深度信息, 而导航中, 全局的稠密深度信息才更具有价值. 近期很多研究聚焦于单目深度估计, 即仅通过单目图像估计深度. 但是, 单目深度估计精度低, 对于导航场景并不适用. 因此, 本文提出一种融合 2D 激光雷达的单目深度估计算法, 来提高单目深度估计精度, 使其能够应用于室内导航任务中. 该算法融合 2D 激光

① 基金项目: 科技创新特区计划 (20-163-14-LZ-001-004-01)
收稿时间: 2021-12-24; 修改时间: 2022-01-24; 采用时间: 2022-01-30; csa 在线出版时间: 2022-06-16

雷达的尺度信息和单目图像的纹理结构信息, 使得深度估计的精度得到较高的提升.

1 概述

1.1 相关工作

近年来, 深度学习方法广泛应用于单目深度估计任务中. Eigen 等人[3] 首先将深度学习应用于深度估计任务中, 其后续的工作[4] 对上述工作进行了拓展, 在预测深度的同时完成表面法向预测和语义分割. Fu 等人[5] 将连续值回归问题转化为量化的序数回归问题. Hao 等人[6] 使用空洞卷积来提取多尺度信息, 并使用注意力机制来融合多尺度信息. Yin 等人[7] 从重建的三维场景中随机抽取 3 个点, 以 3 个点确定的虚拟法向作为几何约束来更精确恢复三维结构. Lee 等人[8] 在解码阶段使用局部平面引导层来得到原分辨率的深度图, 而不是标准的上采样层. Huynh 等人[9] 将非局部共面性约束与深度注意力体 (DAV) 合并到网络中, 通过平面结构引导深度估计. Bhat 等人[10] 提出了一种基于 Transformer 的结构块, 将深度范围划分为多个单元, 每个单元的中心值自适应估计每幅图像, 并将单元中心线性组合得到深度值估计. 该方法达到目前单目深度估计的最好效果.

从单幅图像估计深度缺乏绝对的尺度信息, 且精度较低, 因此, 通过稀疏的深度数据和单目图像融合来估计密集深度成为热门. Liao 等人[11] 首先提出使用 2D 激光雷达作为额外的深度输入, 比只使用 RGB 图像获得更高的精度. 与文献 [11] 不同的是, 文献 [12–14] 中使用的稀疏深度信息不具有方向性和局部性, 其使用的深度是从深度图全局随机采样的深度点或者是多线激光雷达的深度数据, 这类问题更准确地说是深度补全问题. Ma 等人[12] 使用全局随机采样的深度点和 RGB 图像作为输入, 通过简单的编解码结构得到了更高的精度. Cheng 等人[13] 提出了卷积空间传播网络, 为深度估计学习关联矩阵, 采用一个线性传播模型以循环卷积的形式传播, 并通过深度卷积神经网络学习邻近像素间的关联关系. Park 等人[14] 提出了一种端到端的非局部空间传播网络, 估计每个像素的非局部邻域及其关联矩阵.

1.2 本文工作

本文主要工作如下: (1) 提出了一种融合 2D 激光雷达数据的单目深度估计网络; (2) 提出了一种运用通道注意力机制融合 2D 激光雷达特征和 RGB 图像特征的方法; (3) 使用跳跃连接来获得更多的细节信息; (4) 制作了带有 2D 激光雷达数据的深度数据集. 实验表明, 本文算法对比单目深度估计和深度补全任务均取得更好的效果.

2 基于 2D 激光雷达的单目深度估计网络

本文基于上述研究提出了一种端到端的融合 2D 激光雷达 (以下简称雷达) 数据的单目深度估计网络. 输入为一张 RGB 图像以及一张映射到二维图像的雷达数据 (如图 1 所示). 输出为深度图 (RGB 图像中每个像素位置对应的深度值).

2.1 网络结构

本文的网络结构如图 1 所示. 首先输入的 RGB 图像和雷达数据分别通过特征提取网络提取出多尺度特征M_i^r, M_i^l, 其中$i=1,2,3,4$, 分别表示 4 次下采样后的特征. 设输入的图像尺寸为$W\times H$, 则每一层特征对应的尺寸为$[W/2^i, H/2^i]$. 接着将特征M_4^r和特征M_4^l通过通道注意力特征融合模块 (CAM) 融合. 得到的融合特征经过 ASPP[15] 增大感受野, 输出的特征通道数较大, 需要加入一层1×1卷积来改变特征通道数. 接着通过 3 层 CAM 模块以及跳跃连接上采样层将特征恢复到$[W/2, H/2]$. 最后对该特征上采样, 并经过一层3×3卷积得到$W\times H$的深度图.

本文为了融合 RGB 图像和雷达的特征, 根据文献 [16] 提出的通道注意力机制, 提出了一种通道注意力特征融合模块, 如图 1 所示. 对于提取的多尺度特征M_i^r, M_i^l, 并不是每个通道都具有相同的作用, 因此, 针对特征的每个通道引入一个权重, 通过损失函数学习每个通道的权重, 使重要的特征强化, 不重要的特征减弱, 使特征指向性更强. 而 RGB 和雷达的特征是有关联性的, 因此, 本文将 RGB 与雷达的特征相互融合, 即雷达特征的权重由 RGB 特征产生, RGB 特征的权重由雷达特征产生. 以下详细介绍该模块.

首先, 分别对 RGB 和雷达特征进行3×3卷积, 记得到的特征分别为U^r, U^l, 尺寸分别为$C_r\times W'\times H'$, $C_l\times W'\times H'$. 这里C_r, C_l为通道数, W', H'表示每个通道的长和宽. 再分别对U^r, U^l进行全局池化:

$$z_c^r = F_{sq}(U_c^r) = \frac{1}{W'\times H'}\sum_{i=1}^{W'}\sum_{j=1}^{H'}U_c^r(i,j) \tag{1}$$

$$z_c^l = F_{sq}(U_c^l) = \frac{1}{W' \times H'} \sum_{i=1}^{W'} \sum_{j=1}^{H'} U_c^l(i,j) \tag{2}$$

其中, 下标c表示第c个通道, z^r, z^l为对应的输出, 尺寸分别为$C_r \times 1 \times 1$, $C_l \times 1 \times 1$.

全局池化屏蔽了特征图的空间分布信息, 同时获取全局信息, 能够更加准确的计算通道的权重. 接下来通过两层全连接层, 即1×1卷积以及非线性层 (Sigmoid) 学习每个通道的系数:

$$s^l = F_{ex}(z^r) = \sigma\left(W_2^r \delta\left(W_1^r z^r\right)\right) \tag{3}$$

$$s^r = F_{ex}\left(z^l\right) = \sigma\left(W_2^l \delta\left(W_1^l z^l\right)\right) \tag{4}$$

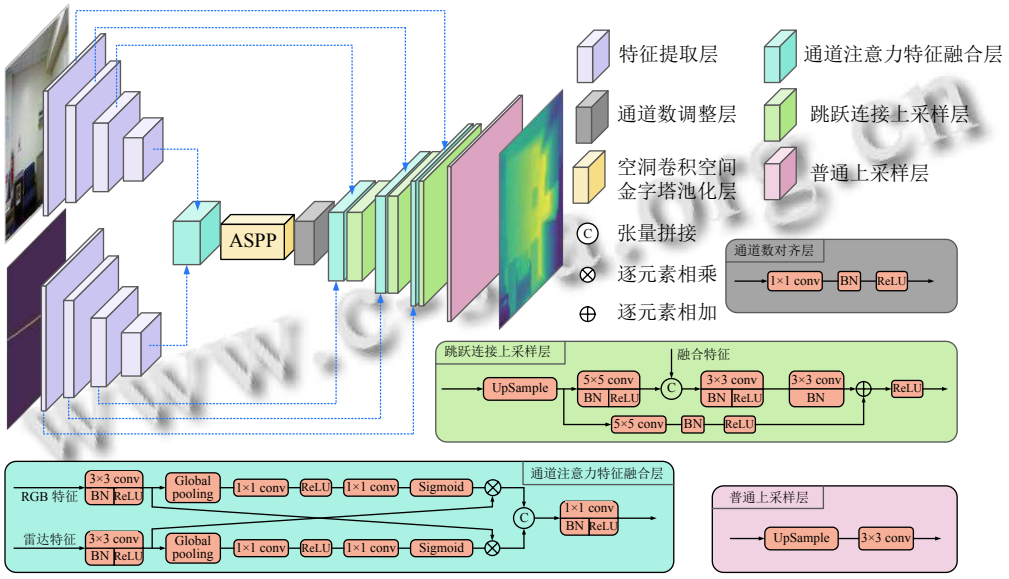


图 1 网络结构图

这里δ为 ReLU 函数, σ为 Sigmoid 函数, W_1^r, W_1^l分别为第 1 层全连接层参数, 尺寸分别为$\frac{C_r}{r} \times C_r$, $\frac{C_l}{r} \times C_l$. W_2^r, W_2^l分别为第 2 层全连接层参数, 尺寸分别为$C_l \times \frac{C_r}{r}$, $C_r \times \frac{C_l}{r}$. 这里的$r = 16$, 是为了减少通道个数, 计算量, s^r, s^l分别对应U^r, U^l的通道权重. 得到权重后, 通过式 (5) 和式 (6) 乘上每个通道的权重:

$$\tilde{U}_c^r = F_{\text{scale}}(U_c^r, s_c^r) = s_c^r U_c^r \tag{5}$$

$$\tilde{U}_c^l = F_{\text{scale}}\left(U_c^l, s_c^l\right) = s_c^l U_c^l \tag{6}$$

这里得到的$\tilde{U}^r$, $\tilde{U}^l$即为加权后的特征.

最后, 将得到的特征$\tilde{U}^r$, $\tilde{U}^l$通过张量拼接操作拼接, 再经过一层1×1卷积得到最终的融合特征.

卷积神经网络中, 下采样是为了扩大感受野, 使每个卷积输出都包含较大范围的信息, 但在这个过程中图像的分辨率不断下降, 导致细节信息会逐渐丢失, 这对于要恢复和原图像相同尺寸的深度估计任务来说并无益. 因此本文引入 ASPP[15]. ASPP 可以在扩大感受野的同时, 而不进行下采样, 减少下采样带来的信息丢失.

对于深度估计任务, 空间域的信息非常重要. 而网络下采样中的池化操作丢失了部分空间域信息. 因此本文借鉴文献 [10,13] 等的方法, 将特征提取网络提取的特征M_i^r, M_i^l通过跳跃连接的方式融入上采样过程中, 来丰富空间域的信息, 提升结果的细节信息. 如图 1 所示, CAM 模块将M_i^r, M_i^l融合得到融合特征, 并通过图 1 中的跳跃连接上采样结构将融合特征通过张量拼接的方式, 融入上采样过程中.

2.2 损失函数

不同的损失函数对于最终的预测结果影响很大. 深度估计任务中常用的损失函数是计算预测深度$\hat{y}$和真实深度y的绝对值误差. 而单纯使用绝对值误差缺乏深度图结构信息. 因此本文为了平衡结构损失, 引入了以下损失函数:

$$L = \lambda L_{\text{depth}} + \alpha L_{\text{grad}} + \beta L_{\text{normal}} + \gamma L_{SSIM} \tag{7}$$

L_{depth}为预测深度$\hat{y}$和实际深度y之间的 L1 损失:

$$L_{\text{depth}} = \frac{1}{N} \sum_i |y_i - \hat{y}_i| \tag{8}$$

L_{grad}表示深度图梯度的 L1 损失:

$$L_{\text{grad}} = \frac{1}{N}\sum_i |g_x(y_i,\hat{y}_i)| + |g_y(y_i,\hat{y}_i)| \tag{9}$$

其中, g_x表示深度图的x方向梯度, g_y表示深度图的y方向梯度.

L_{normal}为实际深度图的表面法向n与预测深度图表面法向$\hat{n}$的 L1 损失, 表面法向可以通过 Sobel 算子估计得到$n=(-\nabla_x(y),-\nabla_y(y),1)$. 则表面法向的损失表示为:

$$L_{\text{normal}} = \frac{1}{N}\sum_i |1 - n_i \cdot \hat{n}_i| \tag{10}$$

L_{SSIM}使用了结构相似性 *SSIM*[17], 用于衡量两幅图片的相似度指标, 因 *SSIM* 的范围是 [0, 1] 且为 1 时两幅图像一样, 所以这里损失函数定义为式 (11) 形式:

$$L_{SSIM} = \frac{1 - SSIM(y,\hat{y})}{2} \tag{11}$$

SSIM 为计算两幅图像的结构相似性操作, 在计算时需将两幅图片归一化操作.

这里的系数$\lambda,\alpha,\beta,\gamma$用于平衡各项损失, 本文在后续实验中取$\lambda=0.3, \alpha=0.4, \beta=1.8, \gamma=2.0$.

3 实验分析

3.1 数据集

本文分别在公开数据集 NYUDv2[18] 和自制数据集上进行了算法的验证实验.

NYUDv2 数据集是使用 Kinect V1 深度相机采集的室内场景数据集. 数据集包含 120k 对 RGB 图和深度图样本, 654 对测试样本. 原始数据的 RGB 图和深度图的分辨率为640×480. 本文训练数据选取 50k 的训练样本, 使用文献 [18] 提供的方法填充缺失的深度值. 因本文算法需要 2D 激光雷达数据, 而公开数据集中没有提供, 所以本文从已有深度图中模拟雷达数据. 分为以下几个步骤: (1) 使用文献 [18] 提供的方法对齐深度图和 RGB 图并补全缺失的深度; (2) 使用文献 [18] 中的算法估计每张图片的重力方向; (3) 将深度图转化为 3D 点云图; (4) 以深度图正中心位置对应的点云为基准, 垂直重力方向作平面, 将平面截取的所有点云作为模拟的雷达数据; (5) 将模拟的雷达数据重新映射回 2D 图像.

为了验证本文算法的有效性, 本文制作了带有 2D 激光雷达数据的数据集. 数据采集使用如图 2 所示的小车, 配备 Kinect V1 以及 2D 激光雷达. 采集数据前首先对深度相机进行标定, 并测量雷达和深度相机之间的相对位置. 为避免相机和雷达位姿校准, 在安装时使用水平仪进行调校, 使相机和雷达之间保持水平以及相对位姿一致.

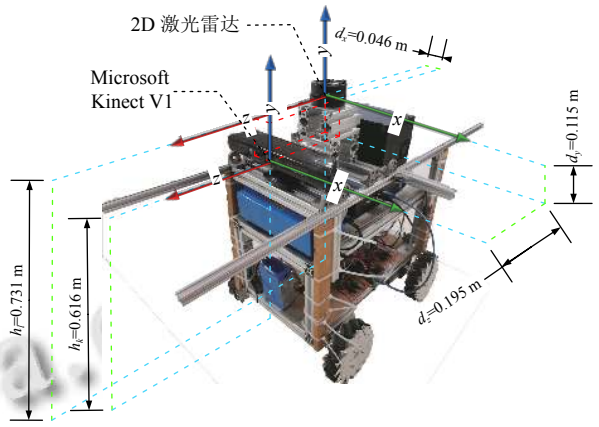


图 2 数据集采集平台

采集完数据后需对原始数据做预处理, 处理过程如图 3 所示. (1) 通过时间戳同步采集的原始 RGB 图, 深度图和雷达数据; (2) 使用文献 [18] 的算法对齐深度图和 RGB 图并补全深度图; (3) 将深度图转化为 3D 点云图; (4) 使用式 (12) 将雷达数据与深度图对齐并映射到 2D 图像, 映射到图像时需要去除超出图像范围的点, 以及深度值小于 0 的点 (2D 激光雷达可测量 360°, 所以会出现小于 0 的值).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{r\cos\theta + d_z} \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r\sin\theta + d_z \\ d_y \\ r\cos\theta + d_z \end{bmatrix} \tag{12}$$

其中, r, θ是雷达数据极坐标形式的距离和夹角. u为映射后二维图像x轴坐标, $u\in[0,640), u\in N$, v为映射后二维图像y轴坐标, $v\in[0,480), v\in N$. f_x, f_y, c_x, c_y为相机内参. d_x, d_y, d_z为相机与雷达的相对位置, 如图 2 所示.

最终, 共采集 24k 组数据. 选取其中的 16k 组数据作为训练集, 800 组作为测试集. 对比图 4 和图 5 中的雷达数据, 可以看出, 实际雷达数据比模拟的雷达数据更加稀疏.

3.2 与现有方法的对比实验

本文算法使用深度学习框架 PyTorch 1.7 实现. 训练和测试使用 11 GB 显存的 NVIDIA GeForce GTX 1080 Ti GPU. 本文做了多组特征提取网络的对比实验, 最终, RGB 图像的特征提取网络选择 ResNet50[19], 雷达数据的特征提取网络选择 ResNet18 达到最好的效果. 特征提取网络的参数使用 ImageNet 数据集上的预训

练模型初始化. 采用的训练优化器是 SGD, 起始的学习率设置为 0.01, 并且当连续 5 个 epoch, *rel* 指标 (见下文) 没有降低, 则将学习率调整为原学习率 10%. 在大约 40 个 epoch 达到稳定.

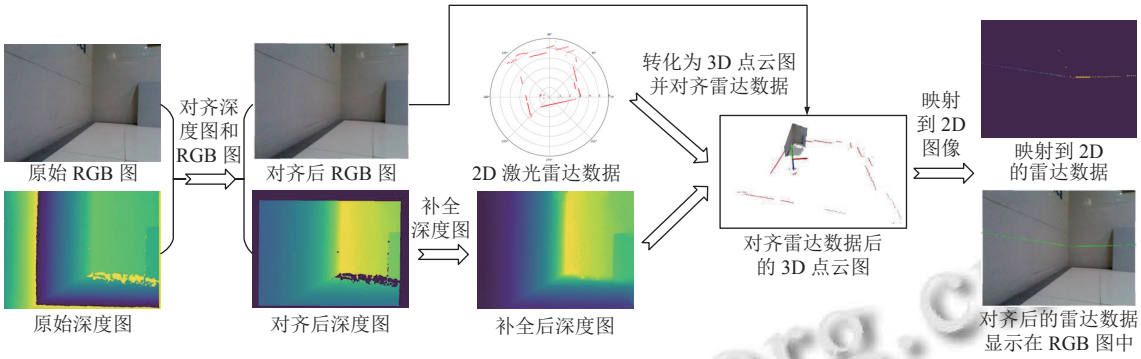


图 3 自制数据集制作流程

对于 NYUDv2 和本文自制数据集, 深度图和 RGB 图分辨率均为640×480. 为了与文献 [3] 保持一致, 首先对深度图和 RGB 图均下采样到320×240, 再以中心为基准裁剪图像, 得到304×228的输入数据. 网络输出的深度图大小也为304×228.

本文采用以下几种方法对训练数据进行增强: (1) 对 RGB 图和深度图随机旋转, 旋转角度为[−5°, 5°]; (2) 对 RGB 图和深度图随机水平翻转, 概率为 0.5; (3) 对 RGB 图的亮度、对比度、饱和度分别随机调整, 三者的范围均为 [0.6, 1.4]; (4) 将 RGB 图标准化, 使用的均值为 [0.485, 0.456, 0.406], 标准差为 [0.229, 0.224, 0.225].

本文使用以下几个评价指标来评价深度估计算法的性能:

(1) 平均相对误差 (*rel*): $\dfrac{1}{N}\sum_i \dfrac{|y_i-\hat{y}_i|}{y_i}$

(2) 均方根误差 (*rms*): $\dfrac{1}{N}\sqrt{\dfrac{1}{N}\sum_i (y_i-\hat{y}_i)^2}$

(3) 精确度 (δ^k): $\max\left(\dfrac{y_i}{\hat{y}_i},\dfrac{\hat{y}_i}{y_i}\right)<\delta^k$, $\delta=1.25$, $k=1,2,3$

其中, y_i是实际深度, $\hat{y}_i$是预测的深度, N是总像素个数.

为了比较已有的单目深度估计算法, 本文在公开数据集 NYUDv2 上测试了算法的效果. 与本文相关的深度估计算法主要有纯单目深度估计算法 (仅使用 RGB 图像) 和深度补全算法. 对于深度补全算法[12–14], 因其原文的稀疏深度信息是全局的, 本文不与其原始效果比较. 本文将文献 [12–14] 的输入换成本文处理后的 NYUDv2 数据集训练测试. 表 1 为本文算法与现有算法的比较结果. 其中, *表示使用雷达数据, 其余仅使用 RGB 图, ↑表示数值越大越好, ↓表示数值越小越好

从表 1 中可以看到, 本文提出的算法多项指标超过了目前最好的纯单目深度估计算法[10], 并且相较于深度补全算法也有一定的提升. 相较于同样使用 2D 激光雷达数据的[11], 本文在各个指标上均有较大的提升. 图 4 展示了部分预测结果, 从结果中看出, 本文算法能够较准确的预测图像的深度.

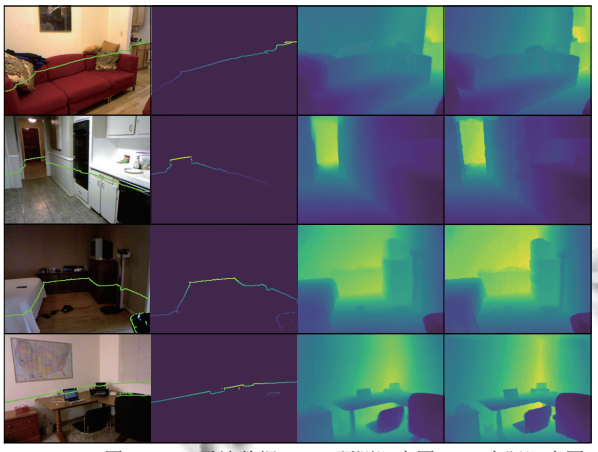
表 1 NYUDv2 数据集效果对比

方法	*rel*↓	*rms*↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Eigen等人[4]	0.158	0.641	0.769	0.950	0.988
Fu等人[5]	0.115	0.509	0.828	0.965	0.992
Hao等人[6]	0.127	0.555	0.841	0.966	0.991
Yin等人[7]	0.108	0.416	0.875	0.976	0.994
Lee等人[8]	0.110	0.392	0.885	0.978	0.994
Huynh等人[9]	0.108	0.412	0.882	0.980	0.996
Bhat等人[10]	0.103	0.364	0.903	0.984	**0.997**
*Liao等人[11]	0.104	0.442	0.878	0.964	0.989
*Ma等人[12]	0.088	0.369	0.914	0.981	0.995
*Cheng等人[13]	0.070	0.336	0.941	0.985	0.995
*Park等人[14]	0.074	0.323	0.927	0.985	0.996
*本文	**0.062**	**0.293**	**0.949**	**0.987**	0.996

为了验证算法的有效性, 本文制作了带有 2D 激光雷达数据的深度数据集, 并对比了其他算法在本文自制数据集上的表现. 本文自制数据集主要为室内导航场景, 存在很多墙面等低纹理场景, 以及不同光照的场景, 并且雷达数据与深度数据也存在噪声误差, 所以整体难度比较大. 表 2 展示了不同方法在本文数据集上表现. 其中, 文献 [10] 不使用雷达数据. 从表中的数据可以看到, 本文的算法依旧取得了较好的效果, 并且对噪声具有鲁棒性. 图 5 为部分预测结果.

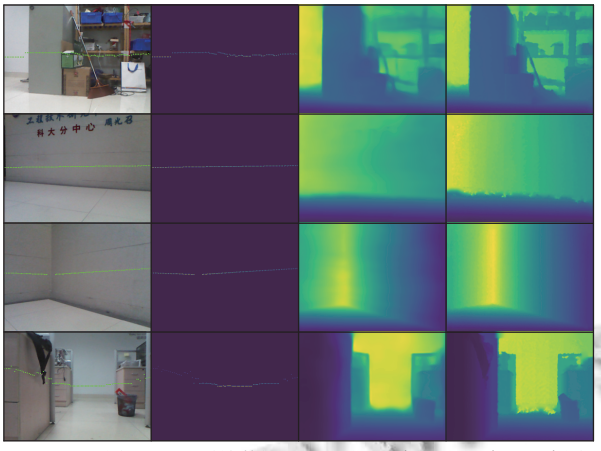
表 2 自制数据集效果对比

方法	*rel*↓	*rms*↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Bhat等人[10]	0.105	0.417	0.883	0.964	0.983
*Ma等人[12]	0.090	0.373	0.911	0.964	0.982
*Cheng等人[13]	0.074	0.337	0.925	0.970	0.984
*Park等人[14]	0.067	0.315	0.928	0.972	0.986
*本文	**0.064**	**0.298**	**0.944**	**0.979**	**0.990**



(a) RGB 图 (b) 雷达数据 (c) 预测深度图 (d) 实际深度图

图 4 NYUDv2 数据集预测结果



(a) RGB 图 (b) 雷达数据 (c) 预测深度图 (d) 实际深度图

图 5 自制数据集预测结果

3.3 消融实验

为了验证雷达数据的加入带来的效果, 以及不同结构和特征提取网络带来的影响, 本文针对 NYUDv2 数据集做了一系列消融实验.

为了比较加入雷达数据的影响, 本文去掉了雷达特征提取部分和相关联的 CAM 模块, 其余保持不变, 做了对比试验, 结果如表 3. 从表中可以看出, 雷达数据使 *rel*, *rms* 分别降低 55.4%, 46.1% (越低越好), δ_1, δ_2, δ_3分别提高了 18.9%, 3.60%, 0.91% (越高越好).

表 3 2D 激光雷达数据的影响

特征	*rel*↓	*rms*↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
RGB	0.139	0.544	0.798	0.954	0.987
RGB+LiDAR	**0.062**	**0.293**	**0.949**	**0.987**	**0.996**
提升 (%)	55.4	46.1	18.9	3.60	0.91

另外, 为比较 CAM 模块作用, 本文将 CAM 模块替换为普通的张量拼接, 实验结果如表 4. CAM 模块使 *rel*, *rms* 分别降低 12.7%, 7.28% (越低越好), 而δ_1, δ_2, δ_3三个指标因本来就比较高, 提升不明显.

表 4 注意力特征融合模块 (CAM) 的影响

模块	*rel*↓	*rms*↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
无CAM	0.071	0.316	0.946	0.986	0.996
CAM	**0.062**	**0.293**	**0.949**	**0.987**	**0.996**
提升 (%)	12.7	7.28	0.32	0.10	0

针对不同的特征提取网络, 本文也做了几组对照实验, 如表 5 所示. 这里分别对比了 ResNet[19], DenseNet[20], MobileNetv2[21] 的影响, 其中, R 表示 ResNet, D 表示 DenseNet, M 表示 MobileNet, 前者表示 RGB 特征提取网络后者为雷达特征提取网络. 从数据中可以看到, DenseNet 与 ResNet 结果差距不大, 且参数量较大. 而在轻量级网络 MobileNetv2 中, 本文的方法在大量减少参数量的同时, 精度降低较小, 且超越了已有的单目深度估计算法. 在实际运用中, 可以考虑更轻量的 MobileNetv2. 后续工作将基于轻量级网络, 并结合知识蒸馏等方法压缩加速模型, 以达到导航场景的实时性要求.

表 5 特征提取网络的影响

方法	参数量	*rel*↓	*rms*↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
R50+R18	78 M	0.062	0.293	0.949	0.987	0.996
D169+D121	96 M	0.063	0.297	0.948	0.987	0.996
Mv2+Mv2	14 M	0.071	0.317	0.941	0.986	0.996

4 结论与展望

本文提出了一种融合 2D 激光雷达的单目深度估计网络, 使用跳跃连接提高了上采样的效果, 并提出一种通过通道注意力机制融合 RGB 特征和雷达特征的方法. 相较于现有的单目深度估计以及深度补全方法, 均取得了更好的效果. 另外, 针对不同的结构和特征提取网络, 本文也做了一系列对照试验. 下一步研究工作将着重于将本文深度估计算法应用到 SLAM 以及 3D 目标检测中, 构建单目视觉导航系统.

参考文献

1 Endres F, Hess J, Sturm J, *et al*. 3-D mapping with an RGB-D camera. IEEE Transactions on Robotics, 2014, 30(1): 177–187. [doi: 10.1109/TRO.2013.2279412]

2 Song SR, Xiao JX. Deep sliding shapes for amodal 3D object detection in RGB-D images. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 808–816. [doi: 10.1109/CVPR.2016.94]

3 Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. arXiv: 1406.2283, 2014.

4 Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 2650–2658. [doi: 10.1109/ICCV.2015.304]

5 Fu H, Gong MM, Wang CH, *et al*. Deep ordinal regression network for monocular depth estimation. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2002–2011. [doi: 10.1109/CVPR.2018.00214]

6 Hao ZX, Li Y, You SD, *et al*. Detail preserving depth estimation from a single image using attention guided networks. Proceedings of 2018 International Conference on 3D Vision (3DV). Verona: IEEE, 2018. 304–313. [doi: 10.1109/3DV.2018.00043]

7 Yin W, Liu YF, Shen CH, *et al*. Enforcing geometric constraints of virtual normal for depth prediction. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 5683–5692. [doi: 10.1109/ICCV.2019.00578]

8 Lee JH, Han MK, Ko DW, *et al*. From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv: 1907.10326, 2019.

9 Huynh L, Nguyen-Ha P, Matas J, *et al*. Guiding monocular depth estimation using depth-attention volume. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 581–597. [doi: 10.1007/978-3-030-58574-7_35]

10 Bhat SF, Alhashim I, Wonka P. Adabins: Depth estimation using adaptive bins. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4008–4017. [doi: 10.1109/CVPR46437.2021.00400]

11 Liao YY, Huang LC, Wang Y, *et al*. Parse geometry from a line: Monocular depth estimation with partial laser observation. Proceedings of 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 5059–5066. [doi: 10.1109/ICRA.2017.7989590]

12 Ma FC, Karaman S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 4796–4803. [doi: 10.1109/ICRA.2018.8460184]

13 Cheng XJ, Wang P, Yang RG. Learning depth with convolutional spatial propagation network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2361–2379. [doi: 10.1109/TPAMI.2019.2947374]

14 Park J, Joo K, Hu Z, *et al*. Non-local spatial propagation network for depth completion. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 120–136. [doi: 10.1007/978-3-030-58601-0_8]

15 Chen LC, Papandreou G, Kokkinos I, *et al*. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

16 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

17 Wang Z, Bovik AC, Sheikh HR, *et al*. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]

18 Silberman N, Hoiem D, Kohli P, *et al*. Indoor segmentation and support inference from RGBD images. Proceedings of the 12th European Conference on Computer Vision. Florence: Springer, 2012. 746–760. [doi: 10.1007/978-3-642-33715-4_54]

19 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

20 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269. [doi: 10.1109/CVPR.2017.243]

21 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520. [doi: 10.1109/CVPR.2018.00474]

(校对责编: 孙君艳)