软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(4):1765−1778 [doi: 10.13328/j.cnki.jos.006399]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于动态门控特征融合的轻量深度补全算法*

孙海峰[1], 穆正阳[1], 戚 琦[1], 王敬宇[1], 刘 聪[2], 廖建新[1]

[1](网络与交换国家重点实验室 (北京邮电大学), 北京 100876)

[2](中国移动通信有限公司研究院, 北京 100053)

通信作者: 廖建新, E-mail: jxlbupt@gmail.com

摘 要: 稠密深度图在自动驾驶和机器人等领域至关重要, 但是现今的深度传感器只能产生稀疏的深度测量, 所以有必要对其进行补全. 在所有辅助模态中, RGB 图像是常用且易得的信息. 现今的许多方法都采用 RGB 和稀疏深度信息结合进行补全. 然而它们绝大部分都是利用通道拼接或逐元素求和简单的对两种模态的信息进行融合, 没有考虑到不用场景下不同模态特征的置信度. 提出一种以输入深度稀疏分布为指导, 结合双模态信息量的动态门控融合模块, 通过动态产生融合权重的方式对两个模态特征进行更高效的结合. 并且根据不同模态的数据特征设计了精简的网络结构. 实验结果表明所提出模块和改进的有效性, 提出的网络在两个有挑战性的公开数据集 KITTI depth completion 和 NYU depth v2 上, 使用了很少的参数量达到了先进的结果, 取得了性能和速度的优秀平衡.

关键词: 深度补全; 特征融合; 轻量模型; 图像处理; 自动驾驶

中图法分类号: TP18

中文引用格式: 孙海峰, 穆正阳, 戚琦, 王敬宇, 刘聪, 廖建新. 基于动态门控特征融合的轻量深度补全算法. 软件学报, 2023, 34(4): 1765−1778. http://www.jos.org.cn/1000-9825/6399.htm

英文引用格式: Sun HF, Mu ZY, Qi Q, Wang JY, Liu C, Liao JX. Fast and Accurate Depth Completion Method Based on Dynamic Gated Fusion Strategy. Ruan Jian Xue Bao/Journal of Software, 2023, 34(4): 1765−1778 (in Chinese). http://www.jos.org.cn/1000-9825/6399.htm

Fast and Accurate Depth Completion Method Based on Dynamic Gated Fusion Strategy

SUN Hai-Feng[1], MU Zheng-Yang[1], QI Qi[1], WANG Jing-Yu[1], LIU Cong[2], LIAO Jian-Xin[1]

[1](State Key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications), Beijing 100876, China)

[2](China Mobile Research Institute, Beijing 100053, China)

Abstract: Dense depth map is essential in areas such as autonomous driving and robotics, but today's depth sensors can only produce sparse depth measurements. Therefore, it is necessary to complete it. In all auxiliary modalities, RGB images are commonly used and easily obtained. Many current methods use RGB and sparse depth information in depth completion. However, most of them simply use channel concatenation or element-wise addition to fuse the information of the two modalities, without considering the confidence of each modalities in different scenarios. This study proposes a dynamic gated fusion module, which is guided by the sparse distribution of input sparse depth and information of both RGB and sparse depth feature, thus fusing two modal features more efficiently by generating dynamic weights. And designed an efficient feature extraction structure according to the data characteristics of different modalities. Comprehensive experiments show the effectiveness of each model. And the network proposed in this paper uses lightweight model to achieve advanced results on two challenging public data sets KITTI depth completion and NYU depth v2. Which shows our method has a good balance of performance and speed.

Key words: depth completion; feature fusion; light-weighted model; image processing; autonomous driving

---

* 基金项目: 国家重点研发计划 (2020YFB1807805); 国家自然科学基金 (62071067, 62001054, 61771068)
收稿时间: 2020-12-27; 修改时间: 2021-03-08, 2021-05-06; 采用时间: 2021-06-14; jos 在线出版时间: 2022-06-15
CNKI 网络首发时间: 2022-11-16

准确而稠密的深度感知作为计算机视觉的基础感知问题对许多应用都有着至关重要的作用, 深度信息可以直接从深度相机或激光雷达中读取出来, 不幸的是, 激光雷达虽然可以准确地测量深度信息, 但是有效深度值的数量很低, 并且价格昂贵. 例如, Velodyne HDL-64e 雷达捕获的投影深度图仅具有大约 5.6% 的有效像素, 但是售价接近 70 万. 在室内场景中, 诸如 Kinect 之类的 ToF 传感器会生成相对密集的深度图, 但是存在许多孔洞和不准确的测量值, 使得原始深度图很难直接用于某些任务, 例如精确的 3D 重建[1], 精确三维动作捕捉[2]等. 单目深度估计和立体匹配算法也可以生成密集的深度图, 但是前者是一个病态问题并且只能获取相对深度, 后者对 RGB 图像的质量十分敏感. 因此, 结合两种模态信息的深度补全问题对于学术和工业领域都具有巨大的价值, 深度补全任务的输入输出如图 1 所示.

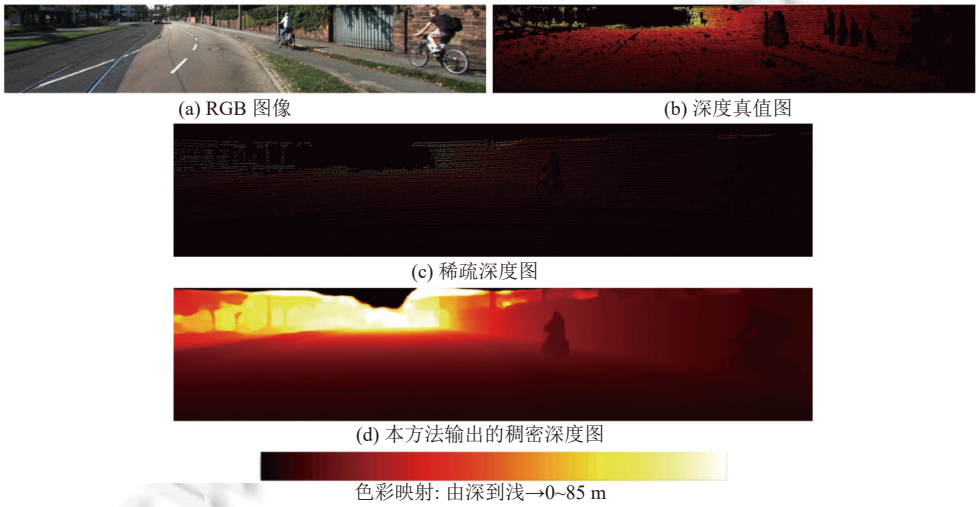


图 1 深度补全任务输入输出数据示意图

随着深度学习方法的演进, 越来越多基于卷积神经网络的深度补全方法被提出, 取得了相比于非学习传统方法[3–5]的巨大提升. 一个主流的方案是直接将稀疏深度图与 RGB 图输入进编码器解码器结构并回归深度, 称为前融合. 这类方法[6–11]以单一结构处理模态差异巨大的数据, 增加了网络学习的难度. 为了提升精度, 这类方法往往采用多任务学习的策略, 在估计深度图的同时产生法向量, 置信度, 语义分割等结果, 并且设计了较为复杂的后处理过程, 进一步增加了主干网络的学习难度. 所以这类方法往往需要较大的网络进行处理, 例如 ResNet34, ResNet50[12], 虽然准确度较高, 但是网络参数量很大, 并且推理速度较慢. 另一个主流的结构是采用双分支编码器分别处理 RGB 和稀疏深度模态的信息, 之后在解码器融合多模态的特征后回归深度, 称为后融合. 这类结构[13–20], 避免了用同一结构处理不同模态数据, 使得轻量化成为可能, 并且可以探索更加复杂的特征交互. 本文提出的方法也是遵循后融合的基础结构, 但是目前后融合方法存在两个问题: 第 1 是特征融合的策略比较简单, 大多是通道维度拼接, 逐元素相乘或者逐元素相加; 第 2 它们大多使用对称编码器结构, 没有根据不同模态特性设计紧凑的网络, 降低参数量和运行速度.

本文针对以上两个问题, 提出了高效门控深度补全算法 LGFN (light-weighted gated fusion network), 首先针对不同模态设计了紧凑的网络结构: 对于稀疏深度模态, 由于其信息量很少并且分散, 我们采用浅层网络进行特征提取, 并且在输入端使用拥有较大卷积核的卷积操作迅速扩增感受野, 使特征图中的有效值迅速膨胀, 可以使得后续卷积学习到更有意义的特征. RGB 模态的特征提取采用轻量语义分割中 ERFNet[21]的编码器设计. 其次, 通过对雷达数据和 ToF 传感器数据的观察发现, 稀疏深度图有效点的密度分布并不是均匀的, 简单的逐元素相加或相乘方式对每个位置的特征融合都是等权重的, 但是因为 RGB 特征和深度特征的域差异, 深度特征的融合权重应高于 RGB 特征. 因此我们提出了使用稀疏深度密度信息引导的动态门控特征模块, 进行高效的特征融合. 我们以后融

合结构的先进算法[19]作为基线, 验证所提出的两个设计的有效性, 相比于基线方法取得了 9.7 倍的参数量下降和 5.8% 的精度提升. 并在两个有挑战的公开数据集 KITTI depth completion 和 NYU depth v2 上相比于现有算法使用很少的参数量达到了先进的结果, 取得了精度和速度的更好权衡.

1 相关工作

1.1 深度补全

一些早期的研究将深度补全问题看作能量函数优化问题[3,22,23]; Hawe 等人[3]基于压缩感知理论重构密集的视差图; Liu 等人[4]使用小波-轮廓字典来直接重构深度图; Ku 等人[5]使用一系列人工定义的常规算子, 例如腐蚀膨胀, 模糊等将稀疏深度图转换为稠密深度图. 然而, 这些方法因为使用人工定义特征而在精度上不尽如人意.

近些年来, 基于卷积神经网络的学习方法成为深度补全领域的主导方法, 并且相比于非学习方法, 精度有着较大的提升. 具体来说, 根据输入模态的不同, 这些方法大体分为无 RGB 引导的和有 RGB 引导的. 无 RGB 引导的方法研究重点在于如何从稀疏深度图中获取更有效的特征. Uhrig 等人[24]提出了稀疏不变卷积来处理深度图的稀疏性; 遵循这篇工作, Eldesokey 等人[25,26]使用 normalize convolution 在网络中传播置信度来生成稠密的深度图; Huang 等人[27]拓展了稀疏不变卷积, 提出了稀疏不变下采样, 上采样等操作符, 并使用编解码器结构进行深度图补全; Chodosh 等人[28]将压缩感知与深度学习相结合, 提取多级特征产生稠密深度. 对于有 RGB 引导的方法, Ma 等人[6,7]提出了端到端的深度补全网络, 并提出了有监督和无监督两种训练策略, 通过构造相邻视频帧之间的光度一致性来作为监督信号; Jaritz 等人[13]提出了语义分割和深度补全的多任务框架. 此外, 有一些方法引入了额外模态的线索. Zhang 等人[29]通过一个网络学习室内场景的局部法线信息, 并在后续工作[11]中证明了法线线索对室外场景的有效性; Cheng 等人[8,30]提出了一种后处理方式, 通过网络输出深度图的同时输出相邻像素之间的相似性矩阵, 并使用此矩阵在深度图上空间传播进行细化; 遵循这两篇工作的思路, Park 等人[9]通过将局部传播拓展为非局部空间传播并且提出了可学习的相似矩阵归一化策略来进一步提升精度. 上述方法虽然取得了较高的精度, 但是由于多任务学习和引入额外辅助线索等因素, 它们普遍具有较高的参数量和较慢的推理时间. 本文提出的方法通过针对模态特性的轻量化设计和高效的双模态特征融合模块, 参数量相比于基线模型下降了 9.7 倍, 并且取得了性能提升.

1.2 高效神经网络设计

轻量网络的设计宗旨是在不损失太多精度的情况下减少网络的参数量, 减少运算时间, 使其更容易的部署在算力较弱的设备上, 提高算法的适用性. 所以设计高效神经网络结构[31,32]一直是一个活跃的领域. 分类任务的 SqueezeNet[32]通过在广泛的使用 1×1 卷积减少参数量; MobileNet[31]系列通过将普通卷积分解为深度可分离卷积来降低参数量; ShuffleNet[33]通过组卷积和通道洗牌的策略达到轻量化的目的. 目标分割领域内的 YOLO 系列[34]和 SSD 模型[35]通过精心设计的单阶段的网络结构, 在不牺牲较大精度的同时, 相比于双阶段的 Faster-RCNN[36]算法达到了 6.6 倍的性能提升. 然而在具有 RGB 引导的深度补全领域, 鲜有针对输入模态特性设计的高效网络, 本文针对 RGB 模态和稀疏深度模态数据特性设计了高效的编码器.

2 LGFN 模型

在本节中, 我们将讨论所提出的 LGFN 整体模型架构的设计, 双模态轻量化编码器的设计, 以及所提出的动态门控融合模块的细节设计.

2.1 整体结构

给定了稀疏雷达投影到像平面的图像 D 以及对应标定好的 RGB 图像 I, 深度补全网络通过结合两个模态的信息生成稠密深度图, 这个问题通常被建模为逐像素的回归问题. 现有的大部分工作[6,7,19]使用编码器解码器结构进行深度补全, 有些工作使用两个[37,38]甚至多个编解码器[11]进行并行处理之后融合或者多阶段处理, 出于轻量化的考虑我们采用单阶段编解码器的设计, 遵循 DDP[19]提出的框架进行重新设计与改进, 形成本文提出的网络, 并

以其作为基线模型. 提出的网络结构由: 双分支编码器, 动态门控特征融合模块和由反卷积构成的深度回归解码器构成, 具体如图 2 所示.

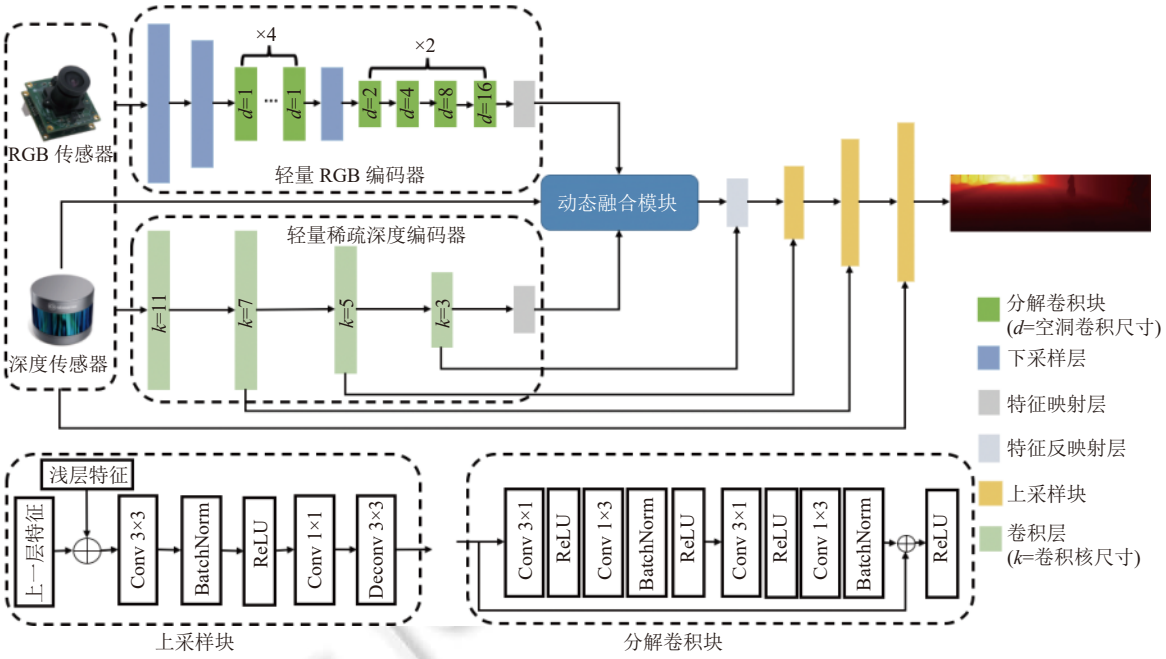


图 2 LGFN 的整体结构

双分支编码器分别提取 RGB 模态和稀疏深度模态的特征, 之后使用转换层将两个模态的特征映射到同一空间进行特征后融合, 相比于在输入端对特征在通道维度进行拼接的前融合策略[6–9]或者使用共享权值的暹罗网络[39]进行两个模态的特征提取, 双分支编码器可以降低编码器的学习难度, 允许我们针对不同模态特征设计不同的轻量网络. 两个编码器产生的特征向量使用所提出的动态门控特征融合模块进行动态融合, 之后使用由卷积核为 3×3 的反卷积和普通卷积构成的上采样块恢复到原始分辨率并进行深度值回归. 值得注意的是, 不同于之前工作将输入图像降采样到 1/16 分辨率 (最宽处宽度为 256 通道) 甚至是 1/32 (网络最宽处为 512) 分辨率再进行恢复, 我们将输入降采样到 1/8, 网络最宽处为 128 通道, 从而大大降低了网络参数量. 在实验中发现, 采用这样的设计相比于基线网络精简参数量的同时提升了精度. 这是因为深度补全这类像素层次问题, 编码器下采样带来的分辨率损失使其在对空间精度敏感的任务上很难取得准确的预测结果[40]. 本文采用的编解码网络结构中, 最终的高分辨表征主要来源于两个部分: 一是原本的高分辨率表征 (从深度分支的跳跃连接), 但是由于只经过了少量的卷积操作, 其本身只能提供低层次的表达, 并且因为深度分支信息量较低, 有效表征非常少; 第二是低分辨率表征 (对应本文双分支网络特征融合后的特征块) 通过上采样得到的高分辨率表征, 其本身虽然拥有很好的语义表达能力, 但是上采样本身并不能完整地弥补空间分辨率的损失. 所以, 这类网络的最终效果很大程度上受限于语义表达力强的表征所对应的分辨率, 也就是编码器降采样的程度. 所以我们采用的少降采样的浅层设计可以在轻量化模型的同时提升精度.

2.2 轻量双分支提取特征模块设计

2.2.1 轻量 RGB 编码器

现有的方法往往使用 ResNet[12]提出的残差块作为 RGB 特征提取的基础模块, 使用 ResNet 骨干网络 layer4 的输出作为提取到的 RGB 特征进行融合, 提取到的特征向量的大小为 $H/16\times W/16\times 256$. 深层的特征向量具有丰富的语义信息, 但是对空间信息进行了进一步的压缩, 针对深度补全任务来说, 这一步操作消耗大量的计算反而带

来了对任务无用的信息. 所以我们采用 ResNet layer3 的输出维度 $H/8\times W/8\times 128$ 作为 RGB 编码器输出向量维度, 通过压缩网络的深度和宽度降低了网络参数量和推理时间. 受到语义分割领域轻量级网络 ERFNet[21]的启发, 我们使用其编码器部分作为本文方法的 RGB 特征编码器, 编码器的输出通过 1×1 卷积进行每个像素特征的映射, 便于之后的特征融合. 该编码器使用 3 种策略: 通过将 3×3 二维卷积分解为两个卷积核大小分别为 3×1 和 1×3 的一维卷积, 使用 Bottleneck[12]的模块结构设计以及先降采样再进行连续卷积的整体编码器设计. 在精度下降不多的情况下大大降低了计算量和参数量.

2.2.2 轻量稀疏深度编码器

现有的稀疏深度编码器设计没有考虑稀疏深度图本身的特征, 往往设计得比较冗余. 但是由图 1(c) 可以看出, 稀疏深度图的有效值点个数很少, 不会超过图像像素总数的 6%, 蕴含很少的信息量, 所以可以使用较浅的卷积神经网络对其进行编码. 并且有效值点的分布往往非常分散, 需要具有较大感受野的卷积对其进行特征提取. 基于这两点目标, 我们设计了如表 1 所示的针对稀疏深度的编码器, 不同于 RGB 分支全部使用小卷积核, 在 conv1 到 conv3 使用大卷积核迅速扩增感受野, 之后使用卷积核为 3×3 的卷积进行细节的特征提取, 最后使用卷积核为 1×1 卷积将每个像素的特征转化为便于融合的特征, 在融合时提升数值稳定性. 值得注意的是, 我们发现网络设计中常用的 BN 层在稀疏深度编码器分支中会因为图中有效值过少导致数值紊乱, 不利于训练, 所以 BN 层没有在本模块使用. 与 RGB 编码器相同, 本模块的输出特征图尺寸同样为 $H/8\times W/8\times 128$.

表 1 轻量稀疏深度编码器的网络结构

网络层	输入特征图尺寸	操作符 (卷积核, 输出通道, 步长)
Conv1	$H\times W\times 1$	$\left\{\begin{array}{c}11\times 11, 16, \text{stride } 1\\ \text{ReLU()}\end{array}\right\}$
Conv2	$H\times W\times 16$	$\left\{\begin{array}{c}7\times 7, 32, \text{stride } 2\\ \text{ReLU()}\end{array}\right\}$
Conv3	$H/2\times W/2\times 32$	$\left\{\begin{array}{c}5\times 5, 64, \text{stride } 2\\ \text{ReLU()}\end{array}\right\}$
Conv4	$H/4\times W/4\times 64$	$\left\{\begin{array}{c}3\times 3, 128, \text{stride } 2\\ \text{ReLU()}\end{array}\right\}$
Transform	$H/8\times W/8\times 128$	$1\times 1, 128, \text{stride } 1$
编码器输出	$H/8\times W/8\times 128$	

2.3 动态门控特征融合模块

在提取到 RGB 和稀疏深度的特征后, 现有方法通常采用加法融合和卷积融合策略进行特征融合. 其中, 加法融合将编码器产生的两个特征向量进行逐元素相加得到最终的融合向量表示. 这种融合方式将两个模态的特征等权重的进行融合, 忽略了任务本身的特性, 在深度补全任务中, 深度特征应该占据更高的权重. 例如在深度信息稠密的区域, 如果等权重的进行特征融合, RGB 特征由于域差异问题反而会影响融合后的效果.

卷积融合将两个模态的特征向量在通道维度上进行拼接后通过 1×1 卷积进行加权融合操作. 这种融合方式依靠可学习的卷积核对每个像素的两个模态特征进行加权融合, 但是缺乏对每个样本的动态调整和邻域信息的感知. 例如图像上的某个固定区域, 由于场景和传感器的原因, 该区域内不同样本产生的深度信息量可能完全不同, 卷积融合因为样本不可知并且在推理时权重固定, 可能导致相比于加法融合更不合理的融合权重分配. 在实验部分我们证明了这一观点.

为了更好地融合两个模态的特征, 本文提出了动态门控特征融合模块, 结构图如图 3 所示. 编码器输出的两个模态特征与经过下采样对齐的稀疏深度图在通道维度进行拼接后送入小型神经网络来生成融合权重, 我们使用

maxpooling 进行下采样. 该权重表示对哪些位置的哪些特征进行抑制或增强, 使用逐元素乘的方式用来对 RGB 特征进行过滤, 最后通过逐元素加的方式对稀疏深度特征进行补全和增强. 小型神经网络由两个卷积核分别为 3×3 和 1×1 的卷积层和在其中的 LeakyReLU 非线性层组成, 该网络可以根据输入特征的信息量结合每个样本输入深度的空间稀疏分布动态的生成融合方式, 并且依靠 3×3 卷积进行邻域感知和学习. 该模块可公式化为:

$$X = E_{sD}(D) + w \times E_{\mathrm{RGB}}(I),\ \text{where}\ w = Net(E_{sD}(D), E_{\mathrm{RGB}}(I), downsample(D)) \tag{1}$$

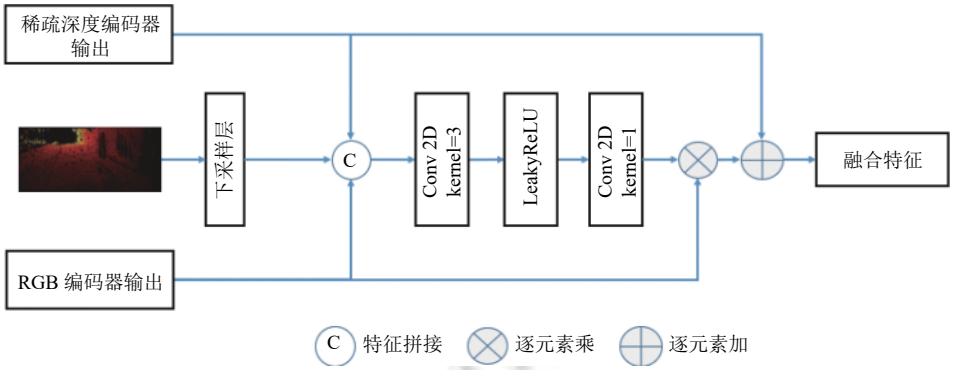


图 3 动态门控特征融合模块示意图

经过上述设计, 该模块以样本可知的方式动态的生成适应于当前场景的融合策略并对特征进行融合, 使得融合后的表征更加准确和丰富. 实验部分证明, 本模块使用较少的参数量和计算代价增长带来了较大的精度提升, 且适用于不同的网络架构.

2.4 损失函数设计

我们结合 l_1 和 l_2 损失作为我们训练阶段的损失函数, 计算方式如下:

$$l_\rho\left(D^{\mathrm{gt}}, D^{\mathrm{pred}}\right) = \frac{1}{\|M\|}\sum_{x\in X}\sigma\left(d_x^{\mathrm{gt}}\right)\times\left|d_x^{\mathrm{gt}} - d_x^{\mathrm{pred}}\right|^\rho \tag{2}$$

其中, D^{gt} 表示数据集提供的真值图, D^{pred} 表示网络输出的深度图, d_x 表示对应图像在坐标 x 处的取值. 对于现实世界的数据, 深度真值图往往是半稠密的, 所以需要函数 $\sigma(x)$ 进行过滤, 如果当前点在真值中的值小于等于 0, 则返回 0, 如果大于 0, 则返回 1. $\|M\|$为真值图中有效点的数量. 整体的损失函数可以表示为:

$$L_{\mathrm{train}} = \alpha\times l_1\left(D^{\mathrm{gt}}, D^{\mathrm{pred}}\right) + \beta\times l_2\left(D^{\mathrm{gt}}, D^{\mathrm{pred}}\right) \tag{3}$$

其中, α 和 β 分别为控制两个损失项比例的权重, 在第 3.4 节 NYU depth v2 数据集中将就该权重对指标的影响进行具体讨论.

3 实验与分析

3.1 训练细节

所提出的方法使用 PyTorch 框架[41]进行实现, 并在装有 Intel(R) Xeon(R) CPU E5-2660 和一块 NVIDIA Tesla M40 GPU 的机器上进行训练和测试. 对于所有的实验, 使用 Adam 优化器进行优化, 优化器参数为: β_1=0.9, β_2=0.999, 初始学习率使用 0.001. 使用 ReduceLROnPlateau 进行学习率调度, 损失连续 5 个 epoch 不下降时降低学习率为原来的 0.5 倍, 并设置最小学习率为 0.000 05. 对于 KITTI 数据集, 损失函数只使用 l_2 训练 60 个 epoch 左右达到收敛; 对于 NYU depth v2 数据集, 训练 50 个 epoch 左右达到收敛, 损失函数中的 α 和 β 全部设置为 1.

3.2 数据集和评价标准

3.2.1 KITTI depth completion[42]

它是室外场景深度补全的权威数据集, 数据集由超过 90 000 组数据 (包括左右目 RGB 图, 激光雷达投影到左

右目的稀疏深度图) 和半稠密的真值深度图构成, 我们将其中的 85 898 组数据用来构成训练集, 其余 6 852 组为验证集. 另外官方还提供无真值的 1 000 组数据作为测试集. 我们使用训练集进行训练, 验证集表现作为超参数调整的指标, 最后将预测好的测试集输出提交到官方服务器进行测试. 在训练过程中, 我们将 375×1242 尺寸的数据, 裁剪到 256×1216 (雷达有效值的最大区间), 并使用随机水平翻转作为数据扩增方案. 我们使用官方的误差指标进行评价: 均方根误差 (RMSE, 单位为 mm, KITTI 数据集主要的指标), 平均绝对值误差 (MAE, 单位 mm), 深度倒数的均方根误差 (iRMSE, 单位为 1/km), 深度倒数的平均绝对值误差 (iMAE, 单位为 1/km). RMSE 指标对大误差敏感, 通常代表整体深度图的准确性, iRMSE 和 iMAE 通常关注临近深度传感器的深度的准确性. 本数据集的所有指标均为越小越好.

3.2.2 NYU depth v2[43]

该数据集包括采集自 464 个不同室内场景的由摄像头采集的 RGB 图和由 Microsoft Kinect 采集的深度图组成. 遵循之前工作[44,45]的划分, 我们将 249 个场景用来训练, 使用其中的约 50K 组数据, 654 组数据用来评估最终的表现. 采用与之前工作[7,9]类似的数据预处理策略, 图片被降采样到 320×240, 之后中心裁剪到 304×228. 采用随机色彩变化, 随机水平翻转, 随机旋转的数据扩增方式. 稀疏深度的生成由稠密深度图随机采样 500 个点得到. 使用均方根误差 RMSE (单位 m), 相对误差 REL (单位 m) 和一定阈值内的相对误差百分比 (δ_t, $t\in\{1.25,1.25^2,1.25^3\}$) 作为评价指标. 其中 RMSE 和 REL 指标越小越好, δ_t 指标越大越好.

3.3 LGFN 模型各个部分的有效性验证

在本部分, 我们使用 NYU depth v2 数据集进行了一系列实验, 验证本文 LGFN 方法中各个部分设计的有效性, 我们使用 RMSE 和 REL 两个指标作为精度度量, 参数量作为模型轻量化的度量. 实验设置及结果如表 2 所示, 序号为实验序号. 我们使用 DDP[19]作为基线模型, 并在其上进行文中所提改进的应用. 基线方法的结构如表 2 的实验 1 所示, 采用截取到 layer4 的 ResNet18 作为两个模态分支的编码器, 送入解码器的特征大小为 $H/16\times W/16\times 512$, 使用简单的加法融合结合两个模态的信息.

表 2 LGFN 模型各部分改进在 NYU depth v2 验证集上的实验对比

实验序号	所用方法					输出深度的精度与模型参数量		
	基线编码器	降采样到1/8	轻量RGB编码器	轻量稀疏深度编码器	特征融合模块	RMSE	REL	参数量 (M)
1	√ (双分支)	× (c=512)	×	×	add	0.145 4	0.020 3	28.438
2	√ (双分支)	√ (c=256)	×	×	add	0.136 1	0.019 5	7.01
3	√ (双分支)	√ (c=256)	×	×	动态门控	0.130 7	0.018 4	8.06
4	√ (单分支)	√ (c=128)	√	×	add	0.147 9	0.023 8	4.528
5	√ (单分支)	√ (c=128)	×	√	add	0.147 1	0.022 5	3.300
6	×	√ (c=128)	√	√	add	0.147 0	0.022 6	2.374
7	×	√ (c=128)	√	√	concat	0.164 4	0.026 7	2.555
8 (LGFN)	×	√ (c=128)	√	√	动态门控	0.136 0	0.019 3	2.687

3.3.1 轻量化设计的有效性

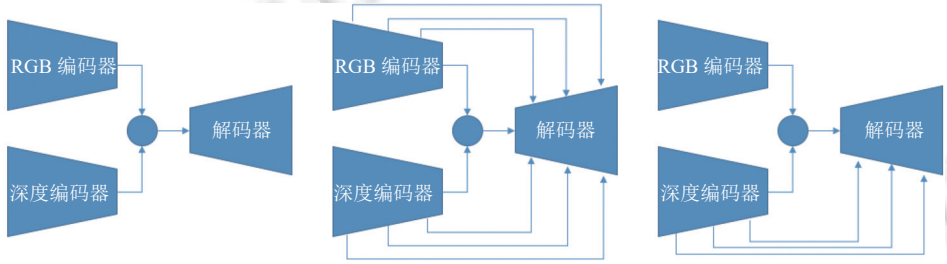
实验 2 将网络变浅, 取用 ResNet18 中 layer3 的输出作为每个模态提取出来的特征, 送入解码器的维度为 $H/8\times W/8\times 256$. 参数量取得了 4 倍的下降, 而精度提升了 6.3%. 这验证了我们在第 2.1 节中的论述. 进一步我们进行了双分支编码器轻量化设计的实验, 实验 4、5 分别将稀疏深度和 RGB 轻量编码器替换为由 ResNet 块构成的编码器, 与实验 6 的对比可知, 我们的轻量化设计在维持了编码器表征能力的同时精简了参数量. 如实验 6 所示, 在实验 2 的基础上对网络的宽度进行压缩, 输入解码器的通道维度为 128, 并且将基于 ResNet18 的双编码器设计替换成本文所提出了轻量化设计. 对比可知, 采用轻量化设计后, 参数量下降了 2.7 倍, 仅带来了微弱的精度损失. 这些损失之后可以被本文提出的高效特征融合方式弥补回来, 见实验 2 与实验 8 的对比, 在没有精度损失的基础上参数量下降了 2.4 倍.

3.3.2 动态门控特征融合模块的有效性

实验 8 (轻量网络+动态融合), 实验 6 (轻量网络+加法融合), 实验 7 (轻量网络+卷积融合), 证明了所提模块的有效性. 其中, 卷积融合的效果最差, 因为所使用的 1×1 卷积的融合权重是拟合了训练集的融合方式, 在验证集进行推理测试时, 融合权重固定, 会存在较大的泛化问题. 加法融合因为等权重融合, 泛化性效果好于卷积融合, 精度提升了 10.5%. 并且因为没有可学习参数, 参数量相比于卷积融合下降了 0.18 M. 本文提出的动态融合模块, 精度相比于加法融合在 RMSE 指标上提升了 7.5%, REL 指标提升了 14.6%, 参数量仅提升了 0.3 M, 验证了本文 LGFN 中的融合方式要优于其他两种. 同时, 我们将该模块应用在了基线模型的结构上, 同样提升了一定的精度, 如实验 3 和实验 2 的对比所示. 证明了该模块可以适用于不同的网络结构.

3.3.3 跳跃连接的设计

输出模糊是编码器解码器结构的公认问题, 因为编码器会有损的降低输入的分辨率, 这种损失很难被解码器恢复. 为了缓解这个问题, 很多基于该结构的方法[15,17,19]使用类似于 U-Net[46]风格的跳跃连接, 将同一分辨率的编码器层的输出, 连接到解码器对应的分辨率上. 常见的有通道拼接和逐元素加法的连接方式, 出于轻量化设计考虑, 本文实验使用逐元素加法. 我们实验了 3 种跳跃连接方式, 如图 4 所示, 从左到右依次为: (a) 不使用跳跃连接; (b) 同时使用来自 RGB 和深度编码器的跳跃连接; (c) 只使用来自深度编码器的跳变连接, 3 种方式的计算量和参数量几乎没有区别. 实验结果发现, (a) 方式由于输出模糊问题取得了最差的精度. 本文所使用的 (c) 方式取得了最好的精度. 基线模型使用 (b) 连接方式精度不如 (c), 是因为 RGB 编码器的浅层特征与深度补全任务的深度输出有着较大的领域差异, RGB 编码器提取的浅层纹理细节和深度图需要的几何细节有着较大区别, 使同 RGB 的跳跃连接反而会干扰深度图的生成.



(a) RMSE: 0.143 9 MAE: 0.060 4 REL: 0.022 4

(b) RMSE: 0.140 1 MAE: 0.056 8 REL: 0.020 9

(c) RMSE: 0.136 0 MAE: 0.054 8 REL: 0.019 9

图 4 不同跳跃连接设计示意图以及在 NYU depth v2 验证集上的精度对比

总之, 相比于基线模型, 通过对网络整体宽度深度和架构的优化, 针对不同模态数据特征的轻量编码器设计和高效融合方式的引入, 如实验 1 和实验 8 的对比所示, LGFN 在 RMSE 指标提升了 6.5%, REL 指标提升了 5% 的基础上, 参数量仅为基线模型的 1/10 左右.

3.4 与现有方法比较

3.4.1 KITTI 数据集

表 3 展示了 KITTI 数据集 (室外场景) 上主流算法和本文提出模型的精度和对应的参数量比较 (其中的“—”符号代表无法获取到该方法的参数量信息). 值得注意的是 DeepLiDAR[11]使用了通过自动驾驶模拟器系统渲染的额外 50K 组训练数据进行表面法向量子网络的训练. PwP 方法[10]使用了额外的法向量标签进行训练. CSPN++[8]和 NLSPN[9]使用了复杂的后处理方法, 它们在表 2 中展示的参数量没有包括后处理部分. 从表中我们可以看出, 基于大数据学习的神经网络方法的精度要远高于非学习的传统方法. 本文的 LGFN 整体在精度没有明显下降的情况下, 参数量有了显著的下降. 相比于相似参数量量级的 Spade-RGBsD 方法[13], LGFN 通过高效的融合方式显著提升了精度. 相比于 MS-Net[LF] 方法[25], LGFN 虽然使用了较多的参数, 但是精度有了较大的提升, 并且参数没有

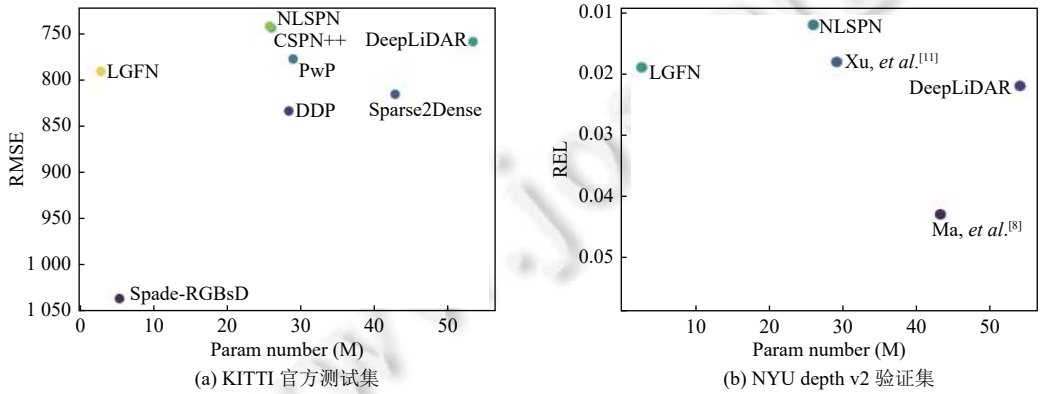
带来推理时间的延长, 如表 4 所示. 相比于 DDP[19] (本文的基线方法), 通过方法部分所述的一系列改进, 在参数量仅为基线 1/10 的基础上, RMSE 精度上取得了较多的提升. 相比于 PwP[10]和 DeepLiDAR[11], LGFN 没有使用额外的数据和标签的情况下, 在 MAE, iRMSE, iMAE 指标上取得了小幅提升, 参数量取得了较大的压缩. CSPN++[8]和 NLSPN[9]虽然参数量与 DDP[19]相当, 但是因为需要进行多轮后处理, 所以推理时间较长, 根据 KITTI benchmark 的数据, 它们的单张深度图推理时间分别为 0.2 s 和 0.22 s. 本文方法在精度有着较低损失的情况下, 参数量仅为它们的 1/9, 推理速度为 0.02 s, 提升了 10 倍. 如表 4 所示, 与其他现有方法相比, 本文方法有着显著的速度优势. 同时维持了相似的精度表现. 图 5 可视化了各个方法精度和速度的权衡. 从图 5(a) 可以看出 LGFN (本文提出方法) 相比于其他方法取得了更好的平衡.

表 3 在 KITTI 深度补全官方测试集上与主流算法的比较, RMSE 为主要指标

方法类型	方法名称	参数量(M)	RMSE	MAE	iRMSE	iMAE
非学习方法	Fast[47]	—	3 548.87	1 767.8	26.48	9.13
	Bilateral[43]	—	2 989.02	1 200.56	9.67	5.08
	TGV[22]	—	2 761.29	1 068.69	15.02	6.28
神经网络方法	SparseConvs[24]	—	1 601.33	481.27	4.94	1.78
	MorphNet[48]	—	1 045.45	310.49	3.84	1.57
	Spade-RGBsD[13]	5.4	1 035.29	248.32	2.6	0.98
	HMSNet[27]	—	841.78	253.47	2.73	1.13
	DDP (基线方法)[19]	28.4	832.94	203.96	2.1	0.85
	MS-Net[LF]-L2[25]	0.356	829.98	233.26	2.60	1.03
	Sparse2Dense[7]	42.82	814.73	249.95	2.8	1.21
	PwP[10]	28.99	777.05	235.17	2.42	1.13
	DeepLiDAR[11]	53.44	758.38	226.5	2.56	1.15
	CSPN++[8]	26.1	743.69	209.28	2.07	0.90
	NLSPN[9]	25.84	741.68	199.59	1.99	0.84
	LGFN	2.68	790.13	226.10	2.49	1.03

表 4 在 KITTI DC 测试集上与主流算法的时间比较 (s)

方法名称	DDP[19]	Sparse2Dense[7]	DeepLiDAR[11]	CSPN++[8]	NLSPN[9]	MS-Net[LF]-L2[25]	LGFN
推理速度	0.08	0.08	0.07	0.2	0.22	0.02	0.02



(a) KITTI 官方测试集 (b) NYU depth v2 验证集

图 5 精度/参数量权衡图

3.4.2 NYU depth v2 数据集

与主流算法在室内数据集的比较, 因为该数据集没有提供稀疏深度输入, 所以为了公平比较, 我们采取遵循主

流方法[7,9–11]的方式, 如表 5 所示, 与室外数据集共享相似的结论. LGFN 在室内场景下, 在维持先进性能的基础上, 显著降低了参数量, 提升了运算速度. 又如图 5 右侧所示, 相比于其他方法, 本文方法在精度和参数量上取得了很好的平衡, 具有广泛的适用性. 我们对损失函数中的调制系数进行实验, 只使用 L2 损失会对难学习区域有较大惩罚, 使得模型在 RMSE 指标上有较好表现, 但是会降低 REL 指标. 只使用 L1 损失会关注大部分像素的准确性, 使得模型在 REL 指标表现较好. 同时使用则会取得均衡的结果.

表 5　在 NYU depth v2 测试集上与主流算法的比较

采样	方法名称	参数量(M)	RMSE	REL	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
采样500稀疏深度点输入	Bilateral[43]	—	0.479	0.084	92.4	97.6	98.9
	TGV[22]	—	0.635	0.123	81.9	93.0	96.8
	Zhang, *et al.*[29]	—	0.228	0.042	97.1	99.3	99.7
	Sparse2Dense[7]	42.82	0.204	0.043	97.8	99.6	99.9
	EncDec-Net[EF][25]	0.484	0.123	0.017	99.1	99.8	100.0
	DeepLiDAR[11]	53.44	0.115	0.022	99.3	99.9	100.0
	PwP[10]	28.99	0.112	0.018	99.5	99.9	100.0
	NLSPN[9]	25.84	0.092	0.012	99.6	99.9	100.0
	LGFN ($\alpha=0,\beta=1$)	2.68	0.118	0.020	99.3	99.9	100.0
	LGFN ($\alpha=1,\beta=0$)	2.68	0.127	0.013	99.2	99.9	100.0
	LGFN ($\alpha=1,\beta=1$)	2.68	0.122	0.017	99.3	99.9	100.0
采样200稀疏深度点输入	Sparse2Dense[7]	42.82	0.230	0.044	97.1	99.4	99.8
	EncDec-Net[EF][25]	0.484	0.171	0.026	98.3	99.6	99.9
	LGFN	2.68	0.169	0.023	99.0	99.8	99.9
不均匀采样500稀疏深度点输入	EncDec-Net[EF][25]	0.484	0.140	0.024	0.989	0.997	0.999
	LGFN	2.68	0.133	0.022	0.991	0.998	0.999

图 6 展示了本文方法相比于其他方法在输入深度图采样点数变化时的鲁棒性 (模型均在 500 采样点情况下进行训练). 可以看出本文方法在 200 个采样点及以上的稀疏情况下均具有优势. 200 个采样点以下本方法表现较差, 可能是因为过于稀疏的深度输入导致动态门控模块的退化. 进一步在表 5 的最后两行, 进行了不均匀采样下的对比实验, 验证动态门控模块对于不同空间位置稀疏程度不均匀时的鲁棒性. 实验设置为: 都使用在 500 均匀采样点下的预训练模型, 在采样后的深度图下随机将一个面积为 2 500 像素的矩形区域置为 0 (模拟传感器在某些区域失效的场景), 结果显示本文 LGFN 方法具有优势.

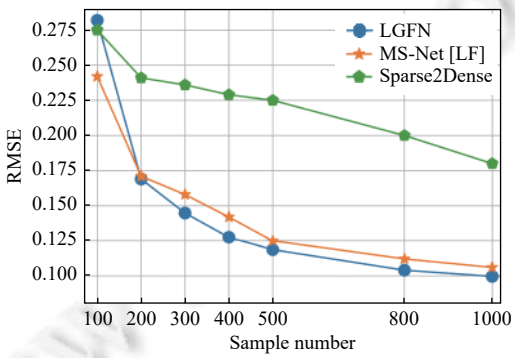


图 6　在 NYU depth v2 数据上对于输入稀疏程度的鲁棒性测试以及与 MS-Net[LF][25]和 Sparse2Dense[7]的对比

3.4.3　结果可视化

图 7 展示了 LGFN (图 7 中 Ours) 与 DDP (基线模型)[19], Sparse2Dense[7], NLSPN[9]在 KITTI 测试集上的深度

图输出可视化比较. 本文的方法因为动态融合了两个模态的信息, 所以对于物体边缘和输入深度线索稀疏甚至缺乏的困难场景下估计得较为准确. 总体来看, LGFN 的深度图输出与现在最为准确的 NLSPN 方法可视化效果接近, 优于基线模型和较早的 Sparse2Dense[7]方法. 具体来说图 7 左边红框中, 轿车顶的稀疏深度输入因为传感器硬件限制有所缺乏, 导致了 DDP[19]和 Sparse2Dense[7]方法没能在深度图输出中恢复合理的轿车顶边缘, 但是 LGFN 和 NLSPN 可以恢复出较为正确的轮廓. 如图 7 右侧红框所示, 因为深度的不均匀和信息缺失, DDP[19]和 Sparse2Dense[7]对于路牌杆的形状恢复较差, NLSPN[9]和 LGFN 形状恢复较好, 在较远处深度大面积缺乏的区域, 由于动态融合模块的高效信息融合, LGFN 对于框中车辆的形状恢复及内部深度一致性要优于其他 3 种方法.

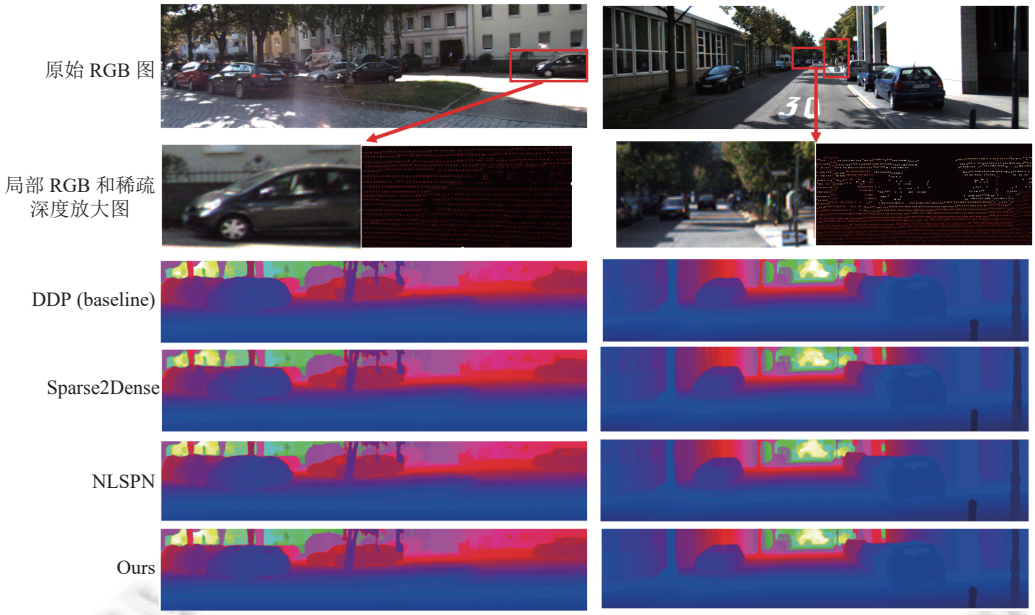


图 7 在 KITTI DC 测试集上与不同方法的结果可视化比较[7,9,19]

4 结束语

本文提出了一种针对深度补全问题, 端到端的高效卷积神经网络 LGFN, 可以从稀疏的 LiDAR 数据和稠密的 RGB 数据中恢复图像中每个像素点深度. 我们首先分析了每个模态的数据特征, 并对应的设计了轻量化的编码器, 在精度不产生明显下降的基础上大大降低了参数量, 提高了运行速度. 进一步针对深度补全的关键问题 (如何从多模态的数据中有效的利用观测到的空间环境) 提出了动态门控特征融合模块, 根据输入样本和两个模态提取的特征动态生成融合权重. 相比于基线模型的对比实验证明了所提出方法的有效性, 在精度提升的基础上取得了较大的参数量下降. 相比于其他算法, 该方法在保留先进网络精度的同时, 参数量和运算速度有着显著提升, 取得了速度和精度的均衡, 简单高效, 具有更加广泛的应用前景.

References:

[1] Li SR, Li Q, Li HY, Hou PH, Cao WG, Wang XD, Li H. Real-time accurate 3D reconstruction based on Kinect v2. Ruan Jian Xue Bao/Journal of Software, 2016, 27(10): 2519−2529 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5089.htm [doi: 10.13328/j.cnki.jos.005089]

[2] Su L, Chai JX, Xia SH. Local pose prior based 3D human motion capture from depth camera. Ruan Jian Xue Bao/Journal of Software, 2016, 27: 172−183 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16032.htm

[3] Hawe S, Kleinsteuber M, Diepold K. Dense disparity maps from sparse disparity measurements. In: Proc. of the 2011 Int'l Conf. on Computer Vision. Barcelona: IEEE, 2011. 2126−2133. [doi: 10.1109/ICCV.2011.6126488]

[4] Liu LK, Chan SH, Nguyen TQ. Depth reconstruction from sparse samples: Representation, algorithm, and sampling. IEEE Trans. on

Image Processing, 2015, 24(6): 1983–1996. [doi: 10.1109/TIP.2015.2409551]

[5] Ku J, Harakeh A, Waslander SL. In defense of classical image processing: Fast depth completion on the CPU. In: Proc. of the 15th Conf. on Computer and Robot Vision, Toronto: IEEE, 2018. 16–22. [doi: 10.1109/CRV.2018.00013]

[6] Ma FC, Karaman S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation. Brisbane: IEEE, 2018. 4796–4803. [doi: 10.1109/ICRA.2018.8460184]

[7] Ma FC, Cavalheiro GV, Karaman S. Self-supervised sparse-to-dense: Self-supervised depth completion from LiDAR and monocular camera. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation. Montreal: IEEE, 2019. 3288–3295. [doi: 10.1109/ICRA.2019.8793637]

[8] Cheng XJ, Wang P, Guan CY, Yang RG. CSPN++: Learning context and resource aware convolutional spatial propagation networks for depth completion. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(7): 10615–10622. [doi: 10.1609/aaai.v34i07.6635]

[9] Park J, Joo K, Hu Z, Liu CK, So Kweon I. Non-local spatial propagation network for depth completion. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 120–136. [doi: 10.1007/978-3-030-58601-0_8]

[10] Xu Y, Zhu XG, Shi JP, Zhang GF, Bao HJ, Li HS. Depth completion from sparse LiDAR data with depth-normal constraints. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 2811–2820. [doi: 10.1109/ICCV.2019.00290]

[11] Qiu JX, Cui ZP, Zhang YD, Zhang XD, Liu SC, Zeng B, Pollefeys M. DeepLiDAR: Deep surface normal guided depth prediction for outdoor scene from sparse LiDAR data and single color image. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3308–3317. [doi: 10.1109/CVPR.2019.00343]

[12] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[13] Jaritz M, De Charette R, Wirbel E, Perrotton X, Nashashibi F. Sparse and dense data with CNNs: Depth completion and semantic segmentation. In: Proc. of the 2018 Int'l Conf. on 3D Vision. Verona: IEEE, 2018. 52–60. [doi: 10.1109/3DV.2018.00017]

[14] Zhong YQ, Wu CY, You SY, Neumann U. Deep RGB-D canonical correlation analysis for sparse depth completion. In: Proc. of the 33rd Conf. on Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 5332–5342.

[15] Shivakumar SS, Nguyen T, Miller ID, Chen SW, Kumar V, Taylor CJ. DFuseNet: Deep fusion of RGB and sparse depth information for image guided dense depth completion. In: Proc. of the 2019 IEEE Intelligent Transportation Systems Conf. Auckland: IEEE, 2019. 13–20. [doi: 10.1109/ITSC.2019.8917294]

[16] Wu CY, Neumann U. Scene completeness-aware lidar depth completion for driving scenario. In: Proc. of the 2021 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021. 2490–2494. [doi: 10.1109/ICASSP39728.2021.9414295]

[17] Xiang R, Zheng F, Su HP, Zhang Z. 3dDepthNet: Point cloud guided depth completion network for sparse depth and single color image. arXiv:2003.09175, 2020.

[18] Lee S, Lee J, Kim D, Kim J. Deep architecture with cross guidance between single image and sparse LiDAR data for depth completion. IEEE Access, 2020, 8: 79801–79810. [doi: 10.1109/ACCESS.2020.2990212]

[19] Yang YC, Wong A, Soatto S. Dense depth posterior (DDP) from single image and sparse range. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3348–3357. [doi: 10.1109/CVPR.2019.00347]

[20] Lee BU, Jeon HG, Im S, Kweon IS. Depth completion with deep geometry and context guidance. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation. Montreal: IEEE, 2019. 3281–3287. [doi: 10.1109/ICRA.2019.8794161]

[21] Romera E, Álvarez JM, Bergasa LM, Arroyo R. ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation. IEEE Trans. on Intelligent Transportation Systems, 2018, 19(1): 263–272. [doi: 10.1109/TITS.2017.2750080]

[22] Ferstl D, Reinbacher C, Ranftl R, Ruether M, Bischof H. Image guided depth upsampling using anisotropic total generalized variation. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 993–1000. [doi: 10.1109/ICCV.2013.127]

[23] Herrera CD, Kannala J, Ladický L, Heikkilä J. Depth map inpainting under a second-order smoothness prior. In: Proc. of the 18th Scandinavian Conf. on Image Analysis. Espoo: Springer, 2013. 555–566. [doi: 10.1007/978-3-642-38886-6_52]

[24] Uhrig J, Schneider N, Schneider L, Franke U, Brox T, Geiger A. Sparsity invariant CNNs. In: Proc. of the 2017 Int'l Conf. on 3D Vision. Qingdao: IEEE, 2017. 11–20. [doi: 10.1109/3DV.2017.00012]

[25] Eldesokey A, Felsberg M, Khan FS. Confidence propagation through CNNs for guided sparse depth regression. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2423–2436. [doi: 10.1109/TPAMI.2019.2929170]

[26] Eldesokey A, Felsberg M, Holmquist K, Persson M. Uncertainty-aware CNNs for depth completion: Uncertainty from beginning to end. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12011–12020. [doi: 10.1109/CVPR42600.2020.01203]

[27] Huang ZX, Fan JM, Cheng SG, Yi S, Wang XG, Li HS. HMS-Net: Hierarchical multi-scale sparsity-invariant network for sparse depth

completion. IEEE Trans. on Image Processing, 2020, 29: 3429–3441. [doi: 10.1109/TIP.2019.2960589]

[28] Chodosh N, Wang CY, Lucey S. Deep convolutional compressed sensing for LiDAR depth completion. In: Proc. of the 14th Asian Conf. on Computer Vision. Perth: Springer, 2018. 499–513. [doi: 10.1007/978-3-030-20887-5_31]

[29] Zhang YD, Funkhouser T. Deep depth completion of a single RGB-D image. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 175–185. [doi: 10.1109/CVPR.2018.00026]

[30] Cheng XJ, Wang P, Yang RG. Depth estimation via affinity learned with convolutional spatial propagation network. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 108–125. [doi: 10.1007/978-3-030-01270-0_7]

[31] Howard AG, Zhu ML, Chen B, Kalenichenko D, Wang WJ, Weyand T, Andreetto M, Adam H. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.

[32] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360, 2016.

[33] Zhang XY, Zhou XY, Lin MX, Sun J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856. [doi: 10.1109/CVPR.2018.00716]

[34] Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934, 2020.

[35] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot MultiBox detector. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 21–37. [doi: 10.1007/978-3-319-46448-0_2]

[36] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

[37] Giannone G, Chidlovskii B. Learning common representation from RGB and depth images. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019. 408–415. [doi: 10.1109/CVPRW.2019.00054]

[38] Kuznietsov Y, Stückler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2215–2223. [doi: 10.1109/CVPR.2017.238]

[39] Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS. Fully-convolutional siamese networks for object tracking. In: Proc. of the 2016 European Conf. on Computer Vision. Amsterdam: Springer, 2016. 850–865. [doi: 10.1007/978-3-319-48881-3_56]

[40] Sun K, Xiao B, Liu D, Wang JD. Deep high-resolution representation learning for human pose estimation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5686–5696. [doi: 10.1109/CVPR.2019.00584]

[41] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin ZM, Desmaison A, Antiga L, Lerer A. Automatic differentiation in PyTorch. In: Proc. of the 31st Conf. on Neural Information Processing Systems. Long Beach: NIPS, 2017.

[42] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3354–3361. [doi: 10.1109/CVPR.2012.6248074]

[43] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGBD images. In: Proc. of the 12th European Conf. on Computer Vision. Florence: Springer, 2012. 746–760. [doi: 10.1007/978-3-642-33715-4_54]

[44] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: NIPS, 2014. 2366–2374. [doi: 10.5555/2969033.2969091]

[45] Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N. Deeper depth prediction with fully convolutional residual networks. In: Proc. of the 4th Int'l Conf. on 3D Vision. Stanford: IEEE, 2016. 239–248. [doi: 10.1109/3DV.2016.32]

[46] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[47] Barron JT, Poole B. The fast bilateral solver. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 617–632. [doi: 10.1007/978-3-319-46487-9_38]

[48] Dimitrievski M, Veelaert P, Philips W. Learning morphological operators for depth completion. In: Proc. of the 19th Int'l Conf. on Advanced Concepts for Intelligent Vision Systems. Poitiers: Springer, 2018. 450–461. [doi: 10.1007/978-3-030-01449-0_38]

附中文参考文献:

[1] 李诗锐, 李琪, 李海洋, 侯沛宏, 曹伟国, 王向东, 李华. 基于Kinect v2的实时精确三维重建系统. 软件学报, 2016, 27(10): 2519–2529. http://www.jos.org.cn/1000-9825/5089.htm [doi: 10.13328/j.cnki.jos.005089]

[2] 苏乐, 柴金祥, 夏时洪. 基于局部姿态先验的深度图像3D人体运动捕获方法. 软件学报, 2016, 27: 172−183. http://www.jos.org.cn/1000-9825/16032.htm

孙海峰(1989－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为人工智能, 机器视觉, 自然语言处理, 深度学习.

王敬宇(1978－), 男, 博士, 教授, 博士生导师, 主要研究领域为智能网络, 机器学习, 边缘计算等.

穆正阳(1997－), 男, 硕士, 主要研究领域为深度学习, 计算机视觉, 深度估计与补全.

刘聪(1980－), 男, 博士, 高级工程师, 主要研究领域为人工智能, 物联网.

戚琦(1982－), 女, 博士, 副教授, 博士生导师, 主要研究领域为智能边缘计算, 轻量级神经网络, 业务网络智能化.

廖建新(1965－), 男, 博士, 特聘教授, 博士生导师, 主要研究领域为移动通信网络, 业务网络化, 人工智能, 多媒体业务.