控制与决策
Control and Decision

基于多引导结构感知网络的深度补全

孙虎, 金宇强, 张文安, 付明磊

引用本文:
孙虎,金宇强,张文安,付明磊. 基于多引导结构感知网络的深度补全[J]. 控制与决策, 2024, 39(2): 401−410.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0759

基于多引导结构感知网络的深度补全

孙 虎[1,2], 金宇强[1,2], 张文安[1,2], 付明磊[1,2†]

(1. 浙江工业大学 信息工程学院，杭州 310023；2. 浙江省嵌入式系统联合重点实验室，杭州 310023)

摘 要: 针对三维场景深度信息观测稀疏问题,提出一种融合彩色图像的多引导结构感知网络模型以补全稀疏深度. 首先,利用三维平面法向量与场景梯度信息之间的映射关系,设计一种两分支主干网络框架,结合图像特征和几何特征进行深度预测,以充分提取空间位置信息的特征表示;然后,考虑到大范围场景下不同物体的结构差异性,基于网络通道注意力机制设计一种自适应感受野的结构感知模块,以对不同尺度的信息进行表征;最后,在网络采样的过程中,以不同尺寸图像为指导对预测子深度图进行滤波并修复物体的边缘细节. 公开数据集上的实验结果表明,所设计的深度补全算法可以获得准确的稠密深度,同时通过两个下游感知任务进行深入评估,表明利用所提出方法能够有效提升其他感知任务的效果.

关键词: 稀疏场景；深度补全；结构感知；多传感器融合；图像引导滤波；自适应感受野

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0759

引用格式: 孙虎,金宇强,张文安,等. 基于多引导结构感知网络的深度补全[J]. 控制与决策, 2024, 39(2): 401-410.

Depth completion method based on multi-guided structure-aware networks

SUN Hu[1,2], *JIN Yu-qiang*[1,2], *ZHANG Wen-an*[1,2], *FU Ming-lei*[1,2†]

(1. College of Information Engineering，Zhejiang University of Technology，Hangzhou 310023，China；2. Zhejiang Provincial United Key Laboratory of Embedded Systems，Hangzhou 310023，China)

Abstract: Aiming at the problem of sparse depth information observation in 3D scenes, this paper proposes a multi-guided structure-aware network model fused with color images to complement the sparse depth. Using the mapping relationship between the 3D plane normal vector and the scene gradient information, we design a two-branch backbone network framework and combine image features and geometric features for depth prediction to fully extract the feature representation of spatial location information. Secondly, considering the structural differences of different objects in large-scale scenes, a network channel attention mechanism is designed. A structure-aware module with an adaptive receptive field is used to characterize information at different scales. Finally, in the process of network upsampling, the predicted sub-depth maps are filtered and the edge details of objects are repaired with the guidance of images of different sizes. The experimental results on public datasets show that the designed depth completion algorithm can obtain accurate dense depth. At the same time, through the in-depth evaluation of two downstream sensing tasks, the results show that the propsed method can effectively improve the effect of other sensing tasks.

Keywords: sparse scene；depth completion；structure awareness；multi-sensor fusion；image guide filtering；adaptive receptive field

0 引 言

深度补全是视觉场景理解中的基础性问题,且受到机器人导航、无人机控制以及无人系统规划等领域的广泛关注. 对于很多机器人控制中的计算机视觉任务,如目标检测[1]、语义分割[2]、姿态估计[3]、3D场景重建[4]等,丰富且精确的场景深度信息能够极大地简化任务难度,有效提升算法性能. 然而,现有采集设备能够观测到的深度十分稀疏,如商用64线Velodyne激光雷达的方位角与垂直视场分辨率分别为0.08°和0.4°,采集到的三维点云映射到彩色图上生成对应的深度图仅有约6%的有效深度信息[5]. 因此,如何从仅表达极少部分区域信息的稀疏深度图中

收稿日期: 2022-05-04；**录用日期:** 2022-10-10.

基金项目: 国家自然科学基金项目(62173305, 62111530299).

†通讯作者. E-mail: fuml@zjut.edu.cn.

恢复高精的稠密深度图像成为当前3D视觉领域的研究热点之一.

深度补全算法主要包括传统补全方法、无引导的深度学习方法以及基于引导的深度学习方法3类. 传统的深度补全方法大多应用于室内小场景下相对稠密的深度观测,如基于Kinect深度相机采集到的深度图,一般只有约10%～30%的深度信息缺失,并且这些缺失像素往往以空洞的形式集中在一起. 早期的传统方法基于图像修复理论,如基于图像引导的边界对齐[6]、低秩矩阵完备化[7]、压缩感知[8]、形态学方法[9]等可以在一定程度上补全缺失的像素深度信息,但是这类方法依赖于先验假设,且对较大空洞会丢失较多场景信息,难以应对激光雷达采集得到的极度稀疏的室外场景.

得益于神经网络强大的高维函数逼近能力与其在图像处理领域中的应用基础,近年来出现了许多基于学习方法的补全模型. 但仅将稀疏深度图作为输入,对标准卷积神经网络而言非常具有挑战性,因为输入中的无效像素数量庞大,标准卷积不足以向网络提供充分的信息. 针对该问题,Uhrig等[5]提出了一种具有稀疏不变性的卷积神经网络,可以在卷积操作中显式地考虑无效像素的数量与位置关系,并且在输入深度图的稀疏程度发生变化时,网络可以得到一致的特征表示. Ren等[10]利用稀疏输入的二值掩码引导卷积,基于掩码的稀疏结构提出了基于分块平铺式的稀疏卷积算子. Zhang等[11]将稀疏输入分为多个群,采用空间群卷积加速训练,在卷积操作中仅考虑有效的空间点. 这类方法的重点在于设计合适的卷积算子(如稀疏不变卷积)来处理稀疏输入并将有效信息传递到整个图像. 由于稀疏深度图输入缺乏足够的场景信息且稀疏程度不一致,虽然效果上这类方法较传统方法有所改进,但是模型输出结果仍存在过度模糊、泛化能力差等问题.

与稠密深度图像相比,包含场景纹理信息的高分辨率彩色图像更容易获取,且二者在信息上存在着相似性,因此许多学者使用同一场景下彩色图像所包含的先验信息来引导完成对稀疏深度观测的稠密化过程. 其中针对室内场景,Schneider等[12]结合图像像素的强度和场景结构化边缘信息引导深度上采样;Liao等[13]基于残差神经网络,同时以彩色图像和稀疏深度图作为网络输入,构建回归损失函数进行深度估计. 上述方法只针对室内小规模场景,难以解决室外大尺度场景的深度估计问题. 因此,Jaritz 等[14]利用大尺寸感受野的滤波器提取全局特征,省去了如稀疏不变卷积、归一化卷积等操作,使推理速度更快,更为简单实用;Tang等[15]采用编解码结构对局部特征进行检索,在编码的不同阶段将图像特征融合到稀疏深度特征中,并在解码过程融合对应尺度的局部信息,以引导稀疏深度的上采样. 这两种方法分别关注场景的全局特征和局部特征,但前者忽略了场景的结构细节,导致性能下降,而后者以固定尺寸的感受野提取特征,未能利用不同尺度特征间的关联性进行有效融合. 此外,Qiu 等[16]将图像和表面法向量的估计值使用经过学习得到的注意力图进行融合,提高了深度估计的精度;Xu等[17]利用深度与法向量之间的几何约束进行建模,通过扩散细化网络对预测结果进行优化;Zhu等[18]引入了不确定性驱动的损失函数提高深度补全的鲁棒性,即在概率框架下同时预测缺失的深度值及其不确定性,以对具有高不确定性的像素进行自适应修正;Jeon等[19]以三维点云为中心进行深度补全,在特征提取过程中充分考虑点云数据的空间邻接性,使用双边卷积层对包含任务关键信息的特定通道和区域进行有效编码;Chen等[20]使用几何感知嵌入的方法直接考虑对象的形状、大小和结构,对深度预测进行正则化的隐式辅助约束,将来自3D点的局部和全局几何结构信息进行编码,以指导稠密深度估计. 另外,对于场景中的多尺度细节,Huang等[21]和Zhao等[22]分别利用层次化结构和多模态图神经网络对不同尺度的场景结构进行感知;Liu等[23]则使用了一种基于学习的导向核函数对结构细节进行提取,尽管关注了不同尺度特征,但是均采用的是固定感受野方式进行感知;Huang等[21]通过稀疏不变操作整合低阶和高阶特征;Zhao等[22]通过图传播关联对应的深度值;Liu等[23]通过学习的核对稀疏深度测量值进行不同层次尺度的插值.

综上所述,现有的方法框架存在几个主要问题: 1) 针对大型场景中多尺度的目标信息,仅使用固定感受野的卷积进行特征提取,没有自适应地偏向更具信息性的特征表示; 2) 没有充分利用引导信息,而场景中同一对象的梯度存在相似性,这可以通过平面法向量来进行表征,对原结果进一步细化; 3) 没有考虑补全结果的实用性,补全的深度存在结构模糊、噪声等问题,不利于下游任务的执行. 针对上述问题,本文提出一种端到端的融合多引导结构感知网络(multi-Guided structure-aware network, MGSAN),设计相应的快速法向量估计模块、自适应感受野的结构感知模块以及图像引导的滤波上采样模块,整体框

架如图 1 所示. 首先,提出多引导的融合网络框架,用于提高网络的特征提取与学习能力. 由于平面法向量中包含场景的梯度信息,对场景深度估计具有较好的引导作用,且已经较好地应用于小范围室内深度补全[24],设计彩色纹理与平面法向量两分支的多引导网络融合策略,通过纹理-法向量的分级特征融合,搭建起对应尺度不同源特征间信息传递的桥梁,充分利用不同源信息;然后,结合自适应感受野的结构感知模块,根据多分支结构设置不同大小的空洞卷积得到不同大小邻域的信息,并通过通道注意力机制对多尺度信息进行融合,捕获图像中的完整信息. 此外,使用以纹理图像为引导的滤波平滑方法,在图像的局部轮廓细节指导下,对上采样重建过程中的子深度图进行快速平滑,以保留场景的边缘细节,其中图像引导滤波模块作用于网络传播过程中的采样深度图像,滤波输出是引导图像的局部线性变换,自适应感受野场景结构感知模块作用于网络结构中的残差块,使其带有自适应权重.

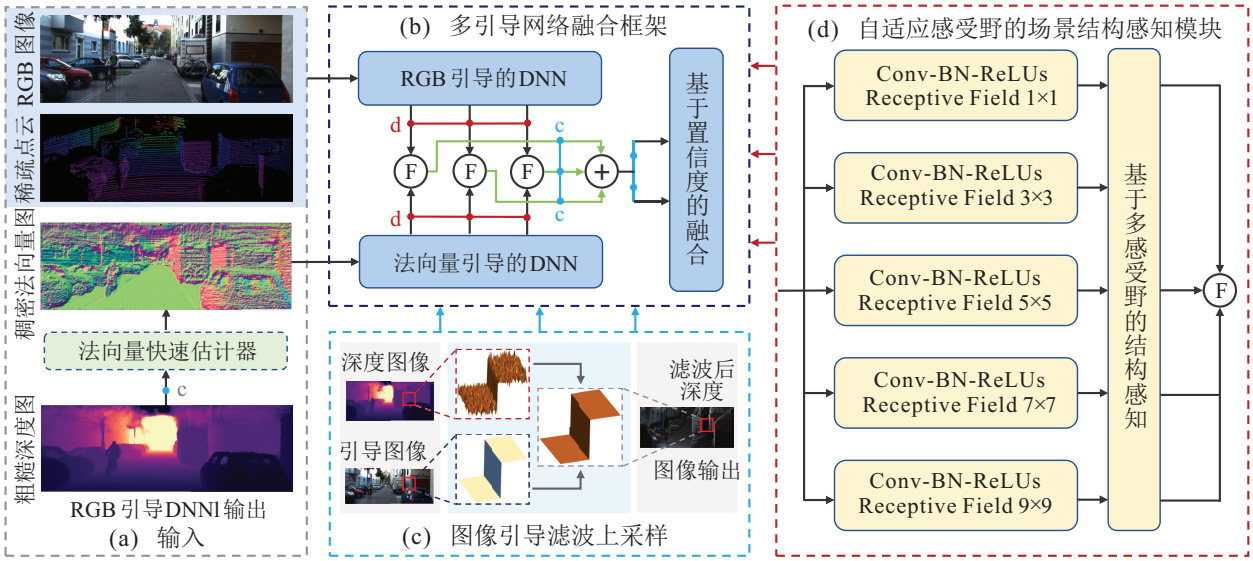


图 1　所提出方法整体框架

本文主要贡献总结如下:

1) 提出一种新的融合多引导结构感知网络框架 MGSAN,用于稀疏场景的三维深度补全. 通过融合两分支的纹理和法向量多尺度特征信息产生稠密的深度预测,并在上采样阶段使用图像引导的滤波方法平滑场景的边缘信息.

2) 提出一种自适应感受野的结构感知模块. 该方法能够收集不同大小区域的场景结构信息,并基于通道注意力机制对多尺度信息进行有效地融合.

3) 在 KITTI 数据集[25]上的实验表明,所提出的多引导网络结构和结构感知模块有助于在线深度补全效果的提升. 另外,两个下游感知任务的实验结果表明更好的深度估计能进一步改善下游任务的效果.

1　多引导结构感知网络 MGSAN

多引导结构感知网络(MGSAN)的目标是从观测到的稀疏深度数据 D_S 和对应彩色图像 X_I 中恢复稠密深度信息. 即对于传感器采集得到的场景中所有稀疏深度-彩色图像配对样本集合 $\{(D_S, X_N)\}_M$,给定一组配对样本 $\{(D_S, X_N)_k\}_{k=0}^{M-1}$,学习得到一个有效映射函数 $F(\cdot)$,其满足 $D = F(D_S, X_I)$. 其中: $D_S \in \mathbb{R}^{H\times W}$ 为采集得到的稀疏深度图,$X_I \in \mathbb{R}^{3\times H\times W}$ 为对应的彩色图像,$D \in \mathbb{R}^{H\times W}$ 为补全后的稠密深度图. 由此,首先描述 MGSAN 的整体结构及其内部每一个组件的构成;其次分别介绍快速平面法向量估计模块、结构感知模块和图像引导滤波模块,并应用于在线深度补全框架.

1.1　网络结构

所提出的融合框架(如图 2 所示)基于编解码网络,采用两分支主干结构,由彩色纹理图像和平面法向量引导的深度补全子网络构成,并在网络传播中对两种模态信息进行有效融合.

RGB 引导分支以配对样本 $\{(D_S, X_N)_k\}_{k=0}^{M-1}$ 作为输入预测稠密深度 D_I,其中编码器含有一个卷积层和 10 个 ResNet 残差块,解码器由 5 个反卷块构成,且与编码器端生成的特征图通过跳跃连接进行元素加操作. 将 RGB 分支预测的稠密深度 D_I 通过快速法向量估计器生成稠密法向量 X_N 作为网络中间变量,并与稀疏深度 D_S 构成新的配对样本 $\{(D_S, X_N)_k\}_{k=0}^{M-1}$ 作为法向量引导分支网络的输入. 采用多尺度分层融合策略考虑彩色图像中的特

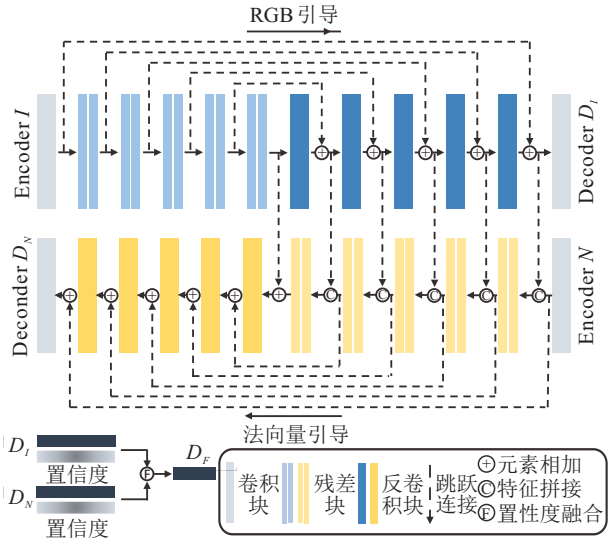


图 2 多引导融合网络结构

征,挖掘在不同尺度下深度、纹理与法向量特征之间的相关性,实现不同模态特征的多阶段融合. 需要注意的是,除了最后一层卷积,其余卷积层后都进行了归一化和ReLU激活操作. 另外,各分支编码器中的残差块均与自适应感受野的结构感知模块直接相连,在法向量分支解码上采样过程中,使用图像引导的平滑滤波器细化场景的边缘细节. 最后,基于文献[1]的后期融合策略,结合两个分支最终预测的深度 D_I 和 D_N 及其置信度 C_I 和 C_N 进行融合,得到最终预测深度为

$$D(i,j)=\frac{\mathrm{e}^{C_I(i,j)}\cdot D_I+\mathrm{e}^{C_N(i,j)}\cdot D_N}{\mathrm{e}^{C_I(i,j)}+\mathrm{e}^{C_N(i,j)}},\tag{1}$$

其中 (i,j) 表示图像像素坐标. 针对该两分支结构,分别考虑RGB引导分支损失、法向量引导分支损失以及基于置信度融合后深度预测损失,以同时提高图像和法向量引导部分深度预测的准确性和最终融合预测深度的准确性,总的损失函数定义为

$$L_{\text{total}}=L(D)+\lambda_I L(D_I)+\lambda_N L(D_N).\tag{2}$$

其中: λ_I 和 λ_N 为实验设置的超参数; $L(\cdot)$ 采用标准的 L_2 范数损失函数,即对于预测 $\hat{D}$,有 $L(\hat{D})=\|(\hat{D}-D_{\text{gt}})\cdot \text{II}(D_{\text{gt}}>0)\|^2$, $\text{II}(\cdot)$ 表示真值的有效深度区域, D_{gt} 表示相应的场景深度真值,即为官方数据集提供的由传感器采集得到的稠密深度图.

1.2 快速法向量估计模块

由于法向量引导分支的法向量输入是由图像分支得到的深度预测进一步计算所得,为了保证模型训练和推断的效率,使用一种快速有效的法向量计算模块. 基于相机投影模型,可以将三维点 $P_i=[X\ Y\ Z]^{\mathrm{T}}$ 转化为像素坐标中的像素点 $P_i=[u\ v\ 1]^{\mathrm{T}}$,即

$$Z\begin{bmatrix}u\\v\\1\end{bmatrix}=KP_i=\begin{bmatrix}f_x&0&u_0\\0&f_y&v_0\\0&0&1\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix}.\tag{3}$$

对于 P_i 及其邻域 $Q_i=[q_{i1},\dots,q_{ik}]$ 中的点 q_{ij},平面法向量 $n_i=[n_x,n_y,n_z]^{\mathrm{T}}$ 有性质

$$n_x x+n_y y+n_z z+b=0.\tag{4}$$

结合式(3)和(4),可以得到

$$\frac{1}{z}=-\frac{1}{b}\Big(\frac{n_x(u-u_0)}{f_x}+\frac{n_y(v-v_0)}{f_y}+n_z\Big).\tag{5}$$

对其分别关于 u 和 v 求偏导,有

$$\frac{\partial 1/z}{\partial u}=-\frac{n_x}{bf_x},\ \frac{\partial 1/z}{\partial v}=-\frac{n_y}{bf_y}.$$

该结果可以通过逆深度图像上分别执行水平和垂直图像滤波器来近似[26],将其写成关于 n_x 和 n_y 的形式并代入式(4),得到最终简化的法向量估计结果为

$$n_x=-bf_x\frac{\partial 1/z}{\partial u},\ n_y=\frac{\partial 1/z}{\partial v};\tag{6}$$

$$\hat{n}_z=-\phi\Big\{\frac{\Delta x_{ij}n_x+\Delta y_{ij}n_y}{\Delta Z_{ij}}\Big\},\ j=1,2,\dots,k.\tag{7}$$

其中 $\phi\{\cdot\}$ 为均值滤波器,用于对邻域内的法向量估计结果进行融合.

1.3 场景结构感知模块

由于稀疏点云分布不均匀的特性,多引导网络得到的稠密深度图仍存在较大误差. 这是因为在网络传播过程中,神经元的感受野范围通常是固定的,而对于场景中如行人、路标以及车辆等尺度较小的目标,深度观测值较少,大尺寸感受野捕获的相关信息较少. 相反地,小尺寸感受野无法高效地编码如车道、建筑等大尺度场景结构,见图3.

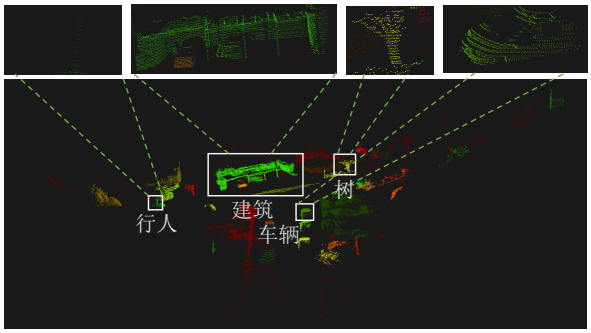


图 3 稀疏点云示例

针对该问题,提出一种自适应感受野的结构感知模块,根据输入点云自适应地调整感受野的尺寸,以收集多尺度的空间结构信息. 感受野的调节使用注意力机制动态选择实现,其本质是一种通过网络自主学习出权重系数,并以动态加权的方式强调所感兴趣的区域同时抑制不相关背景区域的机制. 通过引入通道注意力机制,对融合后的多分支信息流,使用网

络学习不同尺度下的特征图注意力,使网络能够更加侧重于重要的尺度特征,最后将学习到的权重与原来的卷积块相乘,得到在空间维度加权后的特征. 其中,一般的通道注意力机制基于非线性全连接层计算所有通道之间的依赖关系[27],但高复杂度$\mathcal{O}(n^2)$增加了模型的计算负担. 鉴于此,本文使用关于通道数量的自适核$\mathcal{O}(n)$获取邻域内的局部通道交互信息[28],以替换传统通道注意力模块中全连接层.

如图4所示,首先通过空洞卷积增大感受野尺寸,并采用多分支结构计算不同感受野下提取得到的特征;然后使用元素相加融合各分支多尺度特征,通过全局平均池化(gobal average pooling, GAP)聚合各通道信息;接着基于通道维数的自适应核模块计算局部的通道依赖关系;最后使用Sigmoid函数得到每个通道的归一化权重,并计算最终的自适应特征. 具体而言,由于通道维数通常为2的幂次方,且高维通道往往拥有更大的依赖范围,使用关于通道维数C的非线性函数$S(\cdot)$自适应地获得局部核的大小K,即

$$K = S(\cdot) = \mathrm{Odd}\left(\left|\frac{\log_2(C)}{\gamma} + b\right|\right). \tag{8}$$

其中: $\mathrm{Odd}(t)$为与t最接近的奇数; $\gamma = 2, b = 0.5$为固定参数. 因此,自适应核的大小由通道数C确定. 对于每一个通道,根据K邻域内通道间的相关性计算其权重,并通过权重共享策略减少参数量进一步提高效率,最终各通道权重ω_c可以表示为

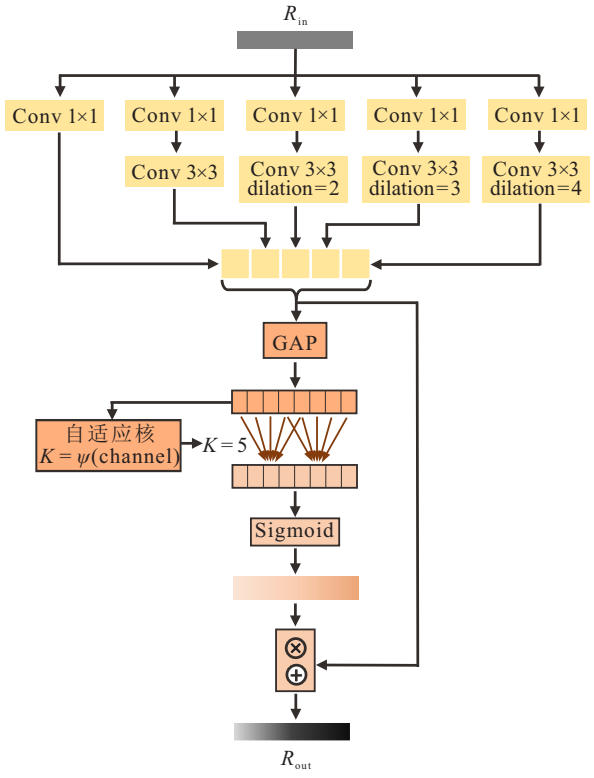


图 4　稀疏点云示例

$$\omega_c = \sigma\left(\sum_{k=1}^{K} \omega^k y_c^k\right). \tag{9}$$

其中: y_c为整个特征块中第c个卷积层的特征, y_c^k为局部核大小为k时得到的邻域通道, ω为原始卷积层的每个权重, σ为Sigmoid激活函数. 如图4所示,式(8)中的权重分配策略可以看作是一个卷积核大小为K的卷积操作[29],其仅包含了K个参数. 另外,为了让该模块更高效,采用分组卷积的方式进一步减少模型计算量,即将感受野大于3×3的分支通过堆叠内核大小为3×3的卷积和不同空洞率的空洞卷积来实现.

1.4　引导滤波模块

由于深度真值的稀疏性,网络预测的深度图会丢失部分结构细节,尤其是在场景目标轮廓边缘处,本文在MGSAN网络的每一个上采样卷积层应用图像引导滤波器,一方面可以通过彩色图像中的纹理、轮廓和结构信息对第l层网络的预测深度$\hat{D}_l$进行优化;另一方面,对于未执行上采样操作的预测图像,该模块复杂度仅为$\mathcal{O}(n/s^2)$,其中s为采样比例,满足模型运行的高效要求. 图像引导滤波器是一种局部线性滤波器,其使用邻域窗口范围内像素的均值和方差信息作为局部估计,通过引导图像内容自适应调整输出权重. 与高斯滤波核函数相比,能够较好地分割复杂的图像前景与背景,并在实现图像平滑的同时保持良好的边缘细节. 本文将引导图像X_I和网络估计的场景深度图$\hat{D}$作为输入,输出图像$\hat{D}^{\mathrm{post}}$对应窗口中的第i个像素点,可以表示为

$$\hat{D}_i^{\mathrm{post}} = \sum_j W_{ij}(X_I)\hat{D}_j. \tag{10}$$

其中: $W_{ij}(X_I)$为核函数,可表示为

$$W_{ij}(X_I) = \frac{1}{|\omega|^2}\sum_{k:(i,j)\in\omega_k}\left(1 + \frac{(I_i - \mu_k)(I_j - \mu_k)}{\sigma_k^2 + \varepsilon}\right), \tag{11}$$

i和j为像素索引, $\sum_j W_{ij}(I) = 1$, ω_k为第k个核函数滑动窗口, $|\omega|$为窗口大小, μ_k和σ_k^2分别为引导图像I在窗口ω_k内的均值和方差, ε为人工设置的平滑因子.

由于点云输入的稀疏性,通过原始网络直接输出的深度场景中总存在不连续预测、噪声干扰等问题. 而基于图像输入对深度图进行滤波后,可以对场景中的目标轮廓、结构边缘等梯度变化较大的区域进行优化,并且对场景中的预测噪声也有一定的抑制作用. 滤波结果示例见图5. 由输入图像(a)进行引导,

对比未经过引导滤波的深度图像(b)与滤波过后的深度图像(c),物体的轮廓更加清晰,边缘特征更为明显.

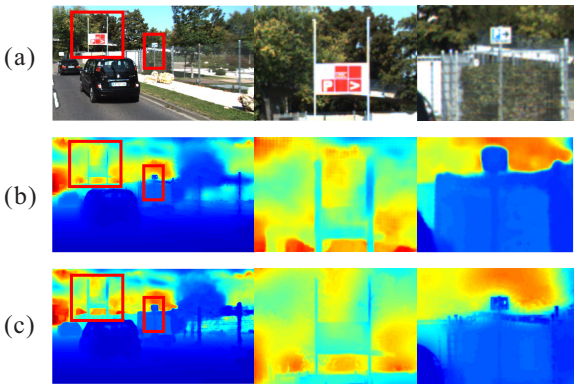


图 5 深度估计的结果滤波前后对比

2 实验分析

2.1 实验细节与数据集

为了验证所提出的融合多引导结构感知网络深度补全方法的有效性,使用KITTI数据集[25]进行实验. 该数据集是基于立体RGB相机、立体灰度相机和64线激光雷达等多传感器的自动驾驶场景相关数据集,本文采用激光-相机对齐后的86 898个稀疏点云与对应RGB图像作为训练集,将KITTI提供的半稠密场景深度信息作为真值,并设置包含1 000个数据样本的测试集和验证集用于模型评估. 模型基于PyTorch实现,并在搭载4块Tesla V 100 GPU的集群服务器上采用多阶段训练策略进行训练. 训练中$\lambda_I = 0.2, \lambda_N = 0.3$,批处理大小为10,初始学习率为0.001,权重衰减系数为10e-6,在第1阶段持续迭代30轮,然后分别以0.02和0.002的学习率再迭代15轮完成第2、第3阶段的训练,保存模型参数.

2.2 深度补全实验结果

使用如下4个评估指标评估模型性能:

1) 均方根误差(RMSE)

$$\text{RMSE} = \frac{1}{|D_{\text{gt}}|}\left(\sum_{u,v\in \text{II}(D_{\text{gt}}>0)} |D_{\text{gt}}(u,v) - D(u,v)|^2\right)^{\frac{1}{2}}. \quad (12)$$

2) 平均绝对误差(MAE)

$$\text{MAE} = \frac{1}{|D_{\text{gt}}|}\sum_{u,v\in \text{II}(D_{\text{gt}}>0)} |D_{\text{gt}}(u,v) - D(u,v)|. \quad (13)$$

3) 反深度的均方根误差(IRMSE)

$$\text{IRMSE} = \frac{1}{|D_{\text{gt}}|}\left(\sum_{u,v\in \text{II}(D_{\text{gt}}>0)} \left|\frac{1}{D_{\text{gt}}(u,v)} - \frac{1}{D(u,v)}\right|^2\right)^{\frac{1}{2}}. \quad (14)$$

4) 反深度的平均绝对误差(IMAE)

$$\text{IMAE} = \frac{1}{|D_{\text{gt}}|}\sum_{u,v\in \text{II}(D_{\text{gt}}>0)} \left|\frac{1}{D_{\text{gt}}(u,v)} - \frac{1}{D(u,v)}\right|^2. \quad (15)$$

式(12)~(15)中u、v为图像的像素坐标. RMSE和MAE直接评估预测深度图的误差;为了进一步补充,使用倒置的深度值IRMSE和IMAE作为测量指标. MAE指标反映了预测深度图的平均误差,RMSE对预测深度图中误差较大的像素更敏感,较低的MAE、较高的RMSE意味着平均误差比较小,但有部分像素有比较大的误差. IRMSE和IMAE通过取深度的倒数衡量近处深度估计的效果,这两个指标更加关注于靠近激光雷达和相机区域深度值的准确性.

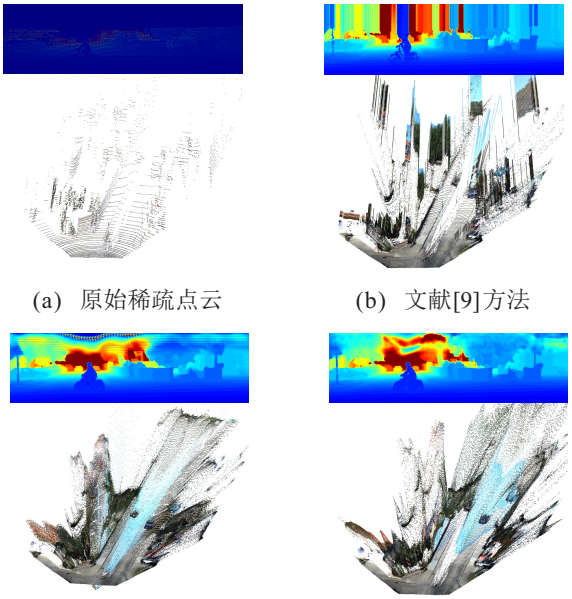
在KITTI验证集上,将所提出的MGSAN模型与文献[9, 16, 18-20, 30-33]进行比较,对比结果见表1. 从均方根误差(RMSE)指标看,本文方法与文献[16]方法基本相当,而其他方法的误差偏大;在反深度的均方根误差(IRMSE)和平均绝对误差(MAE)这两项指标上,本文方法均取得了较好的结果,表明本文方法平均误差更小,在靠近激光雷达和相机区域的深度值预测更为准确;对于指标反深度平均绝对误差(IMAE),本文方法比这一指标下表现最好的文献[20]方法更好,表中数据展示的算法精度误差表明本文算法鲁棒性更好,具有更加均衡的效果.

表 1 深度补全方法精度比较

method	iRMSE	iMAE	RMSE	MAE
文献[9]	3.76	1.21	1 239.84	298.30
文献[16]	2.56	1.15	758.38	226.50
文献[18]	**2.29**	1.07	764.61	220.86
文献[19]	2.29	1.08	773.90	231.29
文献[20]	2.44	1.05	799.31	232.98
文献[30]	2.80	1.07	922.93	249.11
文献[31]	3.21	1.35	954.36	288.64
文献[32]	4.67	1.52	1 268.22	360.28
文献[33]	3.37	1.05	960.05	251.77
MGSAN	2.36	**1.00**	**742.02**	**216.02**

本文的深度补全模型与其他模型对道路场景的稀疏点云进行深度补全后的结果对比如图6所示. 可以看到,经过深度补全算法后点云的密度明显增加,其中文献[9]使用几何计算方法得到的深度图仅在稠密区域效果较好,点云稀疏的地方有较大的误差;文献[16]基于深度学习的方法相较于传统几何方法整

体效果更好,但忽视了边缘细节方面,存在一些误差较大的区域;本文方法得到的三维场景更为准确,细节上也有更好的结果.



(a) 原始稀疏点云　(b) 文献[9]方法　(c) 文献[16]方法　(d) 本文方法

图 6　深度补全效果对比

为了研究不同模块对算法的影响,在KITTI验证集上对各个部分进行消融实验,结果如表2所示. 图像引导分支对颜色或纹理的变化较为敏感,在这些区域会产生更大的误差,法向量和稀疏深度采集得到的是场景的结构信息,引入的法向量引导分支实现了更高的精度;同时由于彩色图像具有清晰的对象边界,使用图像引导滤波的方式进一步平滑了稀疏深度在对象边界周围的噪声. 由于尺度特征的感受野大小对场景估计有明显的影响,在图像下采样过程中引入的自适应感受野提高了场景深度估计的准确性.

表 2　消融实验比较

method	iRMSE	iMAE	RMSE	MAE
RGB	3.05	1.31	872.10	266.78
RGB+法向量	2.59	1.16	797.17	235.90
RGB+法向量+图像引导	2.51	1.12	790.91	230.58
RGB+法向量+图像引导+感受野	**2.36**	**1.00**	**742.02**	**216.02**

本文方法在KITTI数据集上深度预测结果以及与其他方法的具体对比如图7所示,自上而下分别为原始输入的RGB图像,文献[9,16,18-20,30-33]模型结果与本文的方法预测得到的深度图对比,右侧对应的图像是其误差图,颜色越深表示误差越大,本文方法保留了更加清晰和平滑的物体边界,为下游任务提供了更加有效的深度信息.

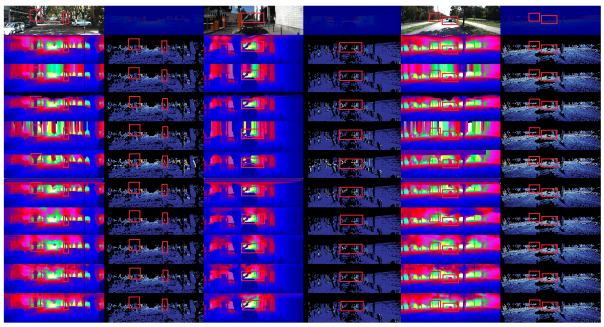


图 7　深度估计方法补全结果及误差

2.3　下游任务实验结果与分析

前文评估依据的定量指标依赖于场景的深度真值,但对于室外场景往往难以得到稠密且准确的真实深度信息,并且KITTI数据集提供的真值数据分布不均匀,无法较全面地评估预测模型. 因此,本文从实际应用场景出发,具体研究深度补全模型对于下游任务的影响. 例如,对于视觉SLAM,特征点的匹配与跟踪是影响精度的主要因素之一,而场景的深度信息可以带来绝对尺度的、更精确的位姿估计结果;对于场景流估计,旨在对场景中的动态物体进行预测,而稠密的静态场景与动态目标三维信息是分割前后景的重要参照;对于语义分割而言,稠密的三维信息能够带来更准确的分割结果,尤其是在物体间交互的轮廓边缘. 综上所述,为了进一步评估和分析本文的补全结果,将所提出的MGSAN模型直接应用于RGB-D SLAM和道路语义分割两个不同的下游机器人任务. 下游任务实验在4×2.90 GHz 8 GB Intel Xeon CPU、NVIDIA GeForce RTX 2080 GPU上实现.

2.3.1　RGBD 视觉里程计

本文使用目前应用广泛的ORB-SLAM2[34]作为视觉里程计实验对象. ORB-SLAM2的RGB-D模式需要输入同一时刻下的RGB图像和与之对应的深度图,一般只能应用于室内使用深度相机的场景,而使用深度补全的方法得到每张RGB图片对应的逐像素深度图后,可以在KITTI户外数据集上使用RGB-D模式.

分别使用文献[16,19]以及本文方法生成每个序列对应的稠密深度信息,对于10个里程计序列,分别在单目以及RGB-D模式运行10次,表3展示了完整序列在不同模式、不同补全方法下的平均旋转误差和平均平移误差对比结果. 由于序列01具有大量无纹理场景和不同尺度大小的动态物体,所有方法均跟踪失败. 另外可见,单目相机模式下的里程计效果较差,而使用深度补全算法得到相应的稠密深度图后,RGB-D模式可以获得良好的定位效果,与文献

[16] 和文献 [9] 方法相比, 本文方法在不同序列中表现较为稳定, 在大部分序列中获得了最好的效果. 此外, 所提出算法在所有序列上的平均误差最小, 与单目相比降低了约 79 % 的误差, 对于里程计定位精度的提升更为明显.

图 8 为应用本文深度补全方法与原始里程计方法的轨迹对比. 由于图像特征点提取对物体轮廓、场景边缘非常敏感, 本文设计的图像引导滤波模块可以较好地修复轮廓细节, 提高深度信息质量, 有助于后续里程计的位姿估计.

表 3 里程计定位误差对比

Frames	ATE(in m) Transl. Mean / ATE(in m) Transl.			
	Monocula	文献 [16] 算法	文献 [9] 算法	本文算法
4 541	8.45/9.64	1.44/1.52	1.45/1.54	**1.30/1.37**
1 101	/	/	/	/
4 661	16.32/22.95	**5.39/6.67**	4.72/5.91	5.85/7.44
801	0.79/1.09	0.38/0.42	0.41/0.42	**0.40/0.45**
271	0.81/0.91	0.28/0.33	0.45/0.52	**0.27/0.27**
2 761	5.50/5.80	**0.48/0.54**	0.57/0.63	0.61/0.66
1 101	13.03/14.97	2.38/2.60	0.82/0.91	**2.11/2.29**
1 101	2.30/2.52	0.26/0.29	0.37/0.39	0.26/**0.28**
4 071	/	4.67/6.03	4.30/5.64	**3.92/4.84**
1 590	46.12/53.86	**2.52/3.26**	3.46/4.39	3.01/4.01
1 200	5.98/7.73	2.39/2.62	2.48/2.74	**2.14/2.41**
平均误差	9.93/11.95	2.02/2.43	2.48/2.74	**1.99/2.40**
性能提升	0/0	0.79/0.79	0.78/0.78	**0.80/0.79**

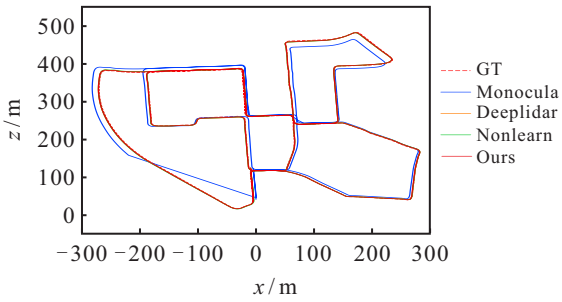


图 8 里程计定位效果对比

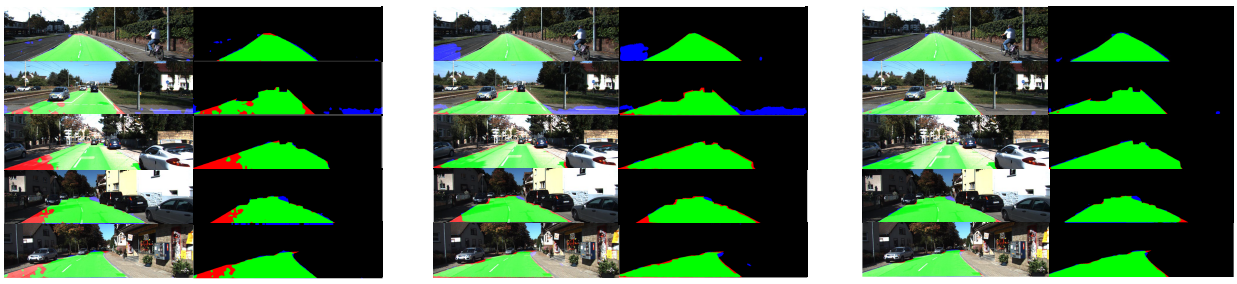
2.3.2 道路语义分割

可通行区域检测 (free space detection) 即道路检测, 是自动驾驶环境感知的重要一环, 对于车辆运动规划与控制具有重要意义. 本文使用目前最先进的 SNE-RoadSeg 道路分割算法[35] 进一步验证模型补全深度的应用效果. 对于 KITTI 的道路分割数据, 使用得到的 3 种不同深度图对原始道路分割模型重新进行训练. 表 4 展示了对标注道路、标注多车道线道路和无标注道路 3 种测试集上的定量对比结果. 可见, 在官方给出的最大 F_1 得分 MaxF、平均精度 AvgPrec、准确率 PRE、召回率 REC、假阴率 FPR、假阳率 FNR 等 6 个指标中, 基于本文方法得到的稠密深度进行道路分割获得了更好的结果.

表 4 使用不同深度进行道路分割结果

性能指标	文献 [16] 算法			文献 [9] 算法			本文算法		
	um	umm	uu	um	umm	uu	um	umm	uu
MaxF	95.75	96.48	94.44	93.12	94.47	91.83	**97.54**	**97.43**	**96.71**
AvgPrec	**92.41**	93.14	91.86	91.87	92.72	91.09	92.50	**93.22**	**92.15**
PRE	96.05	96.10	94.56	93.33	93.16	91.58	**97.94**	**97.33**	**96.57**
REC	95.45	96.86	94.32	92.91	95.82	92.08	**97.14**	**97.52**	**96.85**
FPR	0.85	1.36	0.94	1.44	2.44	1.46	**0.44**	**0.93**	**0.59**
FNR	4.55	3.10	5.68	**7.09**	4.18	7.92	**2.86**	**2.48**	**3.15**

将基于稠密深度的道路分割结果进行定性对比分析, 如图 9 所示. 其中红色部分表示没有被检测出来的路面区域, 蓝色部分表示错误标记为路面区域的部分, 绿色表示路面识别正确的区域. 从各个部分检测结果的对比可以看到, 使用本文方法得到的深度图进行标记误识别率非常低, 并且由于图像引导滤波对深度图的边缘进行了保护, 识别得到的道路边缘更加完整.



(a) 文献[16]方法 (b) 文献[9]方法 (c) 本文方法

图 9 道路分割效果对比

3 结 论

本文提出了一种融合多引导结构感知网络框架实现稀疏的三维深度补全,包含结构感知模块、自适应感受野模块和图像引导滤波模块. 首先通过两分支主干网络结构融合多个模块的优势有效地提高了深度估计的准确性和鲁棒性,其中结构感知模块充分利用图像RGB信息和法向量几何信息在空间上的互补特性,提高了深度估计的效果;然后,通过基于通道注意力机制的自适应感受野模块有效区分场景中不同尺度的目标深度;最后使用图像引导滤波模块有效保留物体的边缘细节并抑制场景噪声. 实验表明,所提出方法能够有效应对复杂大范围的自动驾驶场景,并且可以应用于其他下游感知任务中,如RGB-D的里程计、道路语义分割等. 所提出的深度补全方法还有进一步提升空间,充分考虑场景中的几何轮廓信息或者结合多任务训练策略使用无监督方式提升精度等将是未来的研究方向.

参考文献(References)

[1] Gu J Q, Xiang Z Y, Ye Y W, et al. DenseLiDAR: A real-time pseudo dense depth guided depth completion network[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1808-1815.

[2] Vu T H, Jain H, Bucher M, et al. DADA: Depth-aware domain adaptation in semantic segmentation[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 7363-7372.

[3] Borghi G, Fabbri M, Vezzani R, et al. Face-from-depth for head pose estimation on depth images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(3): 596-609.

[4] 郑太雄, 黄帅, 李永福, 等. 基于视觉的三维重建关键技术研究综述[J]. 自动化学报, 2020, 46(4): 631-652. (Zheng T X, Huang S, Li Y F, et al. Key techniques for vision based 3D reconstruction: A review[J]. Acta Automatica Sinica, 2020, 46(4): 631-652.)

[5] Uhrig J, Schneider N, Schneider L, et al. Sparsity invariant CNNs[C]. International Conference on 3D Vision. Qingdao, 2018: 11-20.

[6] Zuo Y F, Wu Q, Zhang J, et al. Explicit edge inconsistency evaluation model for color-guided depth map enhancement[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(2): 439-453.

[7] Xue H Y, Zhang S M, Cai D. Depth image inpainting: Improving low rank matrix completion with low gradient regularization[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2017, 26(9): 4311-4320.

[8] Ma F C, Carlone L, Ayaz U, et al. Sparse depth sensing for resource-constrained robots[J]. The International Journal of Robotics Research, 2019, 38(8): 935-980.

[9] Zhao Y M, Bai L, Zhang Z M, et al. A surface geometry model for LiDAR depth completion[J]. IEEE Robotics and Automation Letters, 2021, 6(3): 4457-4464.

[10] Ren M Y, Pokrovsky A, Yang B, et al. SBNet: Sparse blocks network for fast inference[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8711-8720.

[11] Zhang J H, Zhao H, Yao A B, et al. Efficient semantic scene completion network with spatial group convolution[C]. Computer Vision — ECCV 2018: 15th European Conference. Munich, 2018: 749-765.

[12] Schneider N, Schneider L, Pinggera P, et al. Semantically guided depth upsampling[M]. Cham: Springer International Publishing, 2016: 37-48.

[13] Liao Y Y, Huang L C, Wang Y, et al. Parse geometry from a line: Monocular depth estimation with partial laser observation[C]. IEEE International Conference on Robotics and Automation. Singapore, 2017: 5059-5066.

[14] Jaritz M, Charette R D, Wirbel E, et al. Sparse and dense data with CNNs: Depth completion and semantic segmentation[C]. International Conference on 3D Vision. Verona, 2018: 52-60.

[15] Tang J, Tian F P, Feng W, et al. Learning guided convolutional network for depth completion[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2021, 30: 1116-1129.

[16] Qiu J X, Cui Z P, Zhang Y D, et al. DeepLiDAR: Deep surface normal guided depth prediction for outdoor scene from sparse LiDAR data and single color image[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 3308-3317.

[17] Xu Y, Zhu X G, Shi J P, et al. Depth completion from sparse LiDAR data with depth-normal constraints[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 2811-2820.

[18] Zhu Y F, Dong W S, Li L D, et al. Robust depth completion with uncertainty-driven loss functions[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 3626-3634.

[19] Jeon Y, Kim H, Seo S W. ABCD: Attentive bilateral convolutional network for robust depth completion[J]. IEEE Robotics and Automation Letters, 2022, 7(1): 81-87.

[20] Chen H, Yang H, Zhang Y , et al. Depth completion using geometry-aware embedding[C]. IEEE International Conference on Robotics and Automation. Philadelphia, 2022: 8680-8686.

[21] Huang Z X, Fan J M, Cheng S G, et al. HMS-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2019: 29: 3429-3441.

[22] Zhao S S, Gong M M, Fu H, et al. Adaptive context-aware multi-modal network for depth completion[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2021, 30: 5264-5276.

[23] Liu L N, Liao Y Y, Wang Y, et al. Learning steering kernels for guided depth completion[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2021, 30: 2850-2861.

[24] Zhang Y D, Funkhouser T. Deep depth completion of a single RGB-D image[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 175-185.

[25] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. IEEE Conference on Computer Vision and Pattern Recognition. Providence, 2012: 3354-3361.

[26] Fan R, Wang H L, Xue B H, et al. Three-filters-to-normal: An accurate and ultrafast surface normal estimator[J]. IEEE Robotics and Automation Letters, 2021, 6(3): 5405-5412.

[27] Li X, Wang W H, Hu X L, et al. Selective kernel networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 510-519.

[28] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 11531-11539.

[29] 王松, 纪鹏, 张云洲, 等. 自适应感受野网络的行人重识别[J]. 控制与决策, 2022, 37(1): 119-126.
(Wang S, Ji P, Zhang Y Z, et al. Adaptive receptive network for person re-identification[J]. Control and Decision, 2022, 37(1): 119-126.)

[30] van Gansbeke W, Neven D, de Brabandere B, et al. Sparse and noisy LiDAR completion with RGB guidance and uncertainty[C]. The 16th International Conference on Machine Vision Applications. Tokyo, 2019: 1-6.

[31] Ma F C, Cavalheiro G V, Karaman S. Self-supervised sparse-to-dense: Self-supervised depth completion from LiDAR and monocular camera[C]. International Conference on Robotics and Automation. Montreal, 2019: 3288-3295.

[32] Eldesokey A, Felsberg M, Khan F S. Propagating confidences through CNNs for sparse data regression[J/OL]. 2018, arXiv: 1805.11913.

[33] Eldesokey A, Felsberg M, Holmquist K, et al. Uncertainty-aware CNNs for depth completion: Uncertainty from beginning to end[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 12011-12020.

[34] Mur-Artal R, Tardós J D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[35] Fan R, Wang H L, Cai P D, et al. SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 340-356.

作者简介

孙虎 (1998−), 男, 博士生, 从事无人车感知与导航、视觉融合感知等研究, E-mail: hsun@zjut.edu.cn;

金宇强 (1998−), 男, 博士生, 从事无人车感知与导航、视觉融合感知等研究, E-mail: yqjin@zjut.edu.cn;

张文安 (1982−), 男, 教授, 博士生导师, 从事无人车定位导航、 网络化控制、 多传感器信息融合等研究, E-mail: wazhang@zjut.edu.cn;

付明磊 (1981−), 男, 副教授, 博士生导师, 从事移动机器人定位导航技术、 无人系统智能感知技术等研究, E-mail: fuml@zjut.edu.cn.