软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(1):444−462 [doi: 10.13328/j.cnki.jos.006488]

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

基于 RGB-D 图像的语义场景补全研究综述*

张 康, 安泊舟, 李 捷, 袁 夏, 赵春霞

(南京理工大学 计算机科学与工程学院, 江苏 南京 210094)

通信作者: 袁夏, E-mail: yuanxia@njust.edu.cn

摘 要: 近年来随着计算机视觉领域的不断发展, 三维场景的语义分割和形状补全受到学术界和工业界的广泛关注. 其中, 语义场景补全是这一领域的新兴研究, 该研究以同时预测三维场景的空间布局和语义标签为目标, 在近几年得到快速发展. 对近些年该领域提出的基于 RGB-D 图像的方法进行了分类和总结. 根据有无使用深度学习将语义场景补全方法划分为传统方法和基于深度学习的方法两大类. 其中, 对于基于深度学习的方法, 根据输入数据类型将其划分为基于单一深度图像的方法和基于彩色图像联合深度图像的方法. 在对已有方法分类和概述的基础上, 对语义场景补全任务所使用的相关数据集进行了整理, 并分析了现有方法的实验结果. 最后, 总结了该领域面临的挑战和发展前景.

关键词: 三维场景; 语义场景补全; 环境理解; 计算机视觉; 深度学习

中图法分类号: TP391

中文引用格式: 张康, 安泊舟, 李捷, 袁夏, 赵春霞. 基于RGB-D图像的语义场景补全研究综述. 软件学报, 2023, 34(1): 444–462. http://www.jos.org.cn/1000-9825/6488.htm

英文引用格式: Zhang K, An BZ, Li J, Yuan X, Zhao CX. Survey on Semantic Scene Completion Based on RGB-D Images. Ruan Jian Xue Bao/Journal of Software, 2023, 34(1): 444–462 (in Chinese). http://www.jos.org.cn/1000-9825/6488.htm

Survey on Semantic Scene Completion Based on RGB-D Images

ZHANG Kang, AN Bo-Zhou, LI Jie, YUAN Xia, ZHAO Chun-Xia

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: In recent years, with the continuous development of computer vision, semantic segmentation and shape completion of 3D scene have been paid more and more attention by academia and industry. Among them, semantic scene completion is emerging research in this field, which aims to simultaneously predict the spatial layout and semantic labels of a 3D scene, and has developed rapidly in recent years. This study classifies and summarizes the methods based on RGB-D images proposed in this field in recent years. These methods are divided into two categories based on whether deep learning is used or not, which include traditional methods and deep learning-based methods. Among them, the methods based on deep learning are divided into two categories according to the input data type, which are the methods based on single depth image and the methods based on RGB-D images. Based on the classification and overview of the existing methods, the relevant datasets used for semantic scene completion task are collated and the experimental results are analyzed. Finally, the challenges and development prospects of this field are summarized.

Key words: 3D scene; semantic scene completion; environment understanding; computer vision; deep learning

研究表明, 为了能够执行诸如导航、互动或物体检索等高级任务, 机器人需要具备对周围环境进行语义级别理解的能力. 同时, 为了进行抓取等任务, 机器人需要具备从单一视角图片中推断出完整三维场景的能力. 人类可以从单视角观察到的图像估计出对象的完整几何形状, 从而建立环境的三维模型, 目前机器人在这方面的能力还比较薄弱. 在语义场景补全提出之前, 研究者们通常分别使用语义分割方法和形状补全方法来使机器人获得语义

* 基金项目: 国家自然科学基金 (61773210)
收稿时间: 2020-09-16; 修改时间: 2021-02-21, 2021-05-31; 采用时间: 2021-08-29; jos 在线出版时间: 2021-10-20
CNKI 网络首发时间: 2022-11-16

理解能力和推断完整形状的能力. Song 等人[1]认为语义分割和形状补全实际上是高度耦合、相互促进的, 并由此提出了语义场景补全 (semantic scene completion, SSC) 这一概念. 语义场景补全是一项结合了三维形状补全与三维语义分割的计算机视觉任务, 可以帮助机器人感知三维世界, 并和环境交互.

无论是形状补全还是语义分割, 首先需要对三维空间进行有效的表示. 体素作为一种常用的三维数据表示方式, 经常作为三维空间的形状补全和语义分割的基本操作单元, 是语义场景补全常用的数据形式. 在体素表示的基础上, 场景补全的目标是将被遮挡的空间区分为被占据空间和空闲空间. 具体而言, 形状补全是依据场景的布局和物体的形状, 将三维栅格中的每个体素标记为空体素或实体素的二分类问题. 然而, 形状补全仅考虑场景的几何信息, 缺少了物体的语义类别. 在形状补全的基础上进行语义分割, 则能够实现对所有被物体占据的空间的语义类别分类. 也就是说, 在三维场景中, 无论是处于物体表面的实体素, 还是处于遮挡区域的实体素, 每个体素均对应于一个特定的语义类别.

为了将形状补全和语义分割在形式上统一, 语义场景补全的目标是为视野内的每个体素预测一个类别标签 $c=\{c_0,c_1,c_2,\ldots,c_N\}$. 其中, c_0 表示空体素, $c_1,c_2,\ldots,c_N$ 表示 N 个物体类别, 因而总类别数为 N+1. 换而言之, 语义场景补全是将空体素也作为一种类别, 并将包括遮挡区域在内的所有体素进行分类, 从而同时实现形状补全和语义分割.

语义场景补全对于许多计算机视觉和机器人应用都有重要意义, 例如: 机器人导航、自动驾驶、场景重建等. 对于室内导航任务, 语义场景补全可以从局部观察重建和理解三维场景, 有利于整个场景的构建, 从而指导导航任务. 在增强现实任务中, 语义场景补全可以进行有效的三维建模, 从而促进增强现实技术的改进. 该项研究虽然是一个出现时间不长的计算机视觉细分研究方向, 但是发展速度很快, 本文总结了近年来基于 RGB-D 图像的语义场景补全方法.

本文第 1 节对语义场景补全的相关背景进行介绍. 第 2 节对传统的语义场景补全方法进行总结. 第 3 节对基于深度学习的语义场景补全方法进行总结. 第 4 节对数据集和评价指标进行总结. 第 5 节对现有方法的性能进行分析. 第 6 节对面临挑战与发展前景进行分析. 最后总结全文.

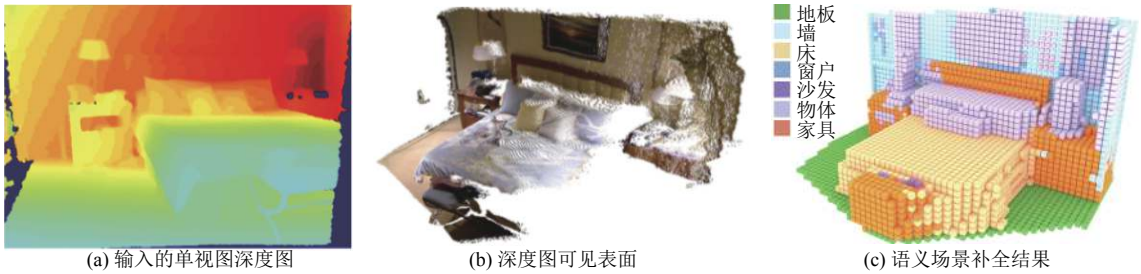
1 相关背景

场景理解是计算机视觉研究的核心问题之一, 而语义分割则是为场景理解铺平道路的一项高级任务, 其目的是对显示出来的图像依照不同目标存在的区域进行划分和标注. 其涉及的应用领域包括自动驾驶、人机交互、图像搜索、增强现实等. 因此, 语义分割是计算机视觉领域的一项关键任务. 已经有很多学者对 RGB-D 图像分割展开了研究[2−5]. 然而, 这些方法注重于获取观察到的像素的语义标签, 基本不考虑物体的完整形状, 因此不能预测可见表面以外的标签或完成场景补全.

形状补全是场景理解的一个核心问题, 其目的是从单一视觉图像推断出物体的完整形状. 到目前为止, 已经有很多关于形状补全的工作[6−9]. 不过大多数方法只针对单个物体的形状进行补全. 要将这些方法应用到场景中, 需要额外的分割或目标掩码. 对于场景补全, 当缺失区域相对较小时, 可以采用平面拟合[10]或物体对称的方法[11,12]来填充空洞. 然而, 这些方法严重依赖几何的规律性, 当缺失区域很大时往往会失败. Firman 等人[13]提出了一种基于几何信息的场景补全方法并取得了较好的实验结果. 由于该方法没有考虑语义信息, 因此当场景结构比较复杂时会产生不准确的结果.

鉴于以上方法不能很好地处理场景补全任务, 所以 Song 等人[1]于 2017 年提出了语义场景补全的概念, 将形状补全和语义分割相结合来解决这一问题. 以单一深度图像为例, 图 1 给出了语义场景补全任务的直观解释, 图 1(b) 颜色仅用于可视化, 为了三维显示效果更好, 此图的显示角度与深度图略有差异. 近些年随着深度学习技术的不断发展, 特别是其在机器视觉应用领域取得了巨大成功, 越来越多的基于深度学习的语义场景补全方法被提出, 并且性能也得到了较大提升.

以下主要对基于 RGB-D 图像的语义场景补全方法进行总结、分类和对其性能进行分析. 近年来提出的基于 RGB-D 图像的语义场景补全方法的时间线如图 2 所示[13−34].



(a) 输入的单视图深度图 (b) 深度图可见表面 (c) 语义场景补全结果

图 1 语义场景补全

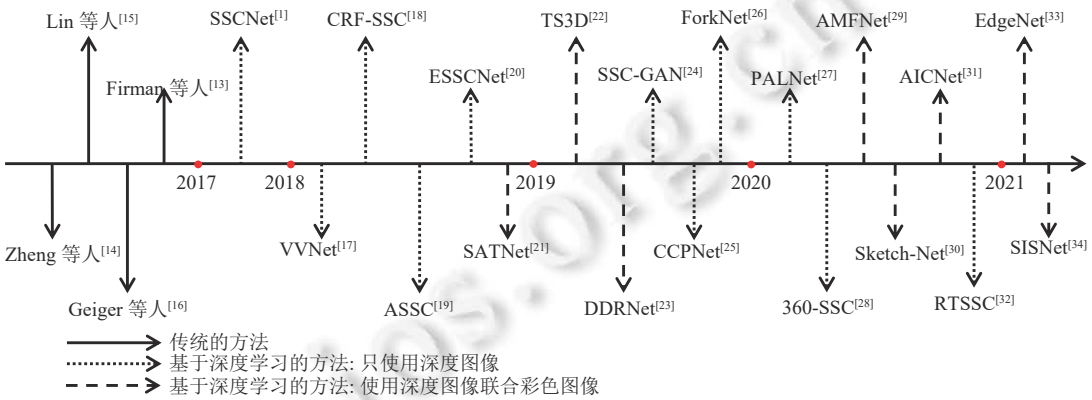


图 2 基于 RGB-D 图像的语义场景补全方法的时间线

2 传统的语义场景补全方法

传统的场景理解任务主要集中在二维图像的分割和目标识别上. 这样的表示缺乏重要的物理信息, 例如对象的三维体积、支持关系、稳定性和功能可见性, 而这些对于机器人进行抓取、操作和导航等应用来说是至关重要的. Zheng 等人[14]提出了一种通过从点云推理物体的物理稳定性来理解场景的方法. 作者将 RGB-D 数据转为点云, 并使用了一个简单的假设, 即: 受重力作用影响, 人工物体在静态场景中应该处于稳定状态. 这一假设适用于所有场景类别, 并为场景理解中似是而非的解释提出了有用的约束条件. 该方法包括两个主要步骤: (1) 几何推理: 从不完整的点云中恢复三维体积基元; (2) 物理推理: 通过优化稳定性和场景先验, 将不稳定的基元分组为物理稳定的对象. 作者提出使用一种新的不连通图[35]来表示能量绘景, 并使用 Swendsen-Wang Cut 方法[36]进行优化. 实验证明该算法在目标分割、场景的三维体积恢复和场景理解方面取得了较好的实验结果.

Firman 等人[13]从 Kim 等人[37]的工作中获得灵感, 使用从不同的对象类中学习到的轮廓去分割图像中的对象. 其证明形状可以超越类别, 使形状预测不需要语义理解. 由于作者关心形状, 独立于语义理解, 所以可以自由地使用与测试时呈现对象不同的训练对象. Firman 等人[13]假设具有不同语义类别的对象通常共享类似的三维形状组件, 从而使有限的数据集能够建模大量对象的形状, 进而估计它们隐藏的几何形状. 为了探究这一假设, 提出了一种算法, 该算法可以基于已有的体积元素训练的监督模型来补全未观察到的桌面大小物体的几何形状. 模型将单一深度图像的局部观察映射到周围邻域表面形状的估计上. 并且在一系列室内对象集和真实场景上定性和定量地验证了方法的性能.

Lin 等人[15]利用 RGB-D 数据处理室内场景的理解问题. 为了实现这一目标, 提出了一种利用二维分割、三维几何以及场景和对象之间的上下文关系的方法. 具体地说, 将 CPMC[38]框架扩展到三维以生成候选长方体, 并利用条件随机场来整合多源信息以对长方体进行分类. 该方法将场景分类与三维物体识别结合起来, 通过概率推理共同解决这一问题. 在具有挑战性的 NYUv2 数据集上测试了方法的有效性. 实验结果表明, 通过有效的证据整合和

整体推理, 方法取得了显著的改进.

之前的三维场景理解方法通常只推断物体[39,40]或将布局估计作为预处理步骤[15], 而 Geiger 等人[16]的方法将三维物体和场景布局结合起来考虑. 利用 Kinect 摄像头捕获的单一 RGB-D 图像推断三维对象和室内场景的布局是一项具有挑战性的任务. 为了实现这一目标, Geiger 等人[16]提出了一个高阶图形模型, 并对图像中的布局、对象和超像素进行了联合推理. 与之前的方法相比, 其模型通过使用可逆图形得到了详细的三维几何信息, 并为了充分考虑到场景属性和投影几何学, 显式地加强了遮挡和可见性约束. 作者将此任务转换为因子图中的映射推理, 并使用消息传递有效地解决此问题. 通过 NYUv2 室内数据集上的几个基线评估该方法. 实验结果表明, 该方法能够较好地推断出含有大量噪声和遮挡的场景.

综上所述, 前两种方法只考虑了几何信息, 没有考虑语义信息, 只能完成场景补全的任务; 后两种方法可以完成语义场景补全任务, 使用传统的数学方法进行处理, 得到的补全结果精度有待提高. 随着数据的海量化和深度学习的迅速发展, 语义场景补全领域也涌现出大量的深度学习方法, 并且取得了不错的成绩.

3 基于深度学习的语义场景补全方法

近年来, 深度学习成功应用于计算机视觉的很多领域, 并取得了骄人的成绩. 针对基于 RGB-D 图像的语义场景补全任务, 目前基于深度学习的方法可分为两类: 基于单一深度图像的方法和基于深度图像联合彩色图像的方法. 本节将详细介绍这两类方法.

3.1 基于单一深度图像的方法

Song 等人[1]提出了一个直接的解决方法, 即用三维卷积神经网络来提取上下文特征. 该方法使用单张深度图作为输入, 并使用 flipped-TSDF 编码将其编码为一个三维体积. 其中 TSDF 表示截断符号距离函数 (truncated signed distance function), 是一种常见的编码三维空间的方法. 其作用是在每个体素中储存该体素到其最接近的物体表面的距离 d (含符号), 并用正负符号来表示该体素是位于表面前方还是位于表面后方. 之前常用的 TSDF 编码有普通的 TSDF 编码和投影 TSDF 编码. 而这两种编码都有各自的缺点. 普通的 TSDF 编码会使得三维网格中的空白体素上出现强梯度; 投影 TSDF 编码有严重的视角依赖性. Song 等人[1]提出的 flipped-TSDF 编码改进了这两个缺点. flipped-TSDF 编码的计算公式如下:

$$d_{\text{flipped}} = sign(d)(d_{\max} - d) \tag{1}$$

其中, $d_{\max}$ 为规定的最远距离, $sign(d)$ 表示 d 的符号.

在 flipped-TSDF 编码基础上, 将编码后的三维网格输入到三维卷积神经网络中, 该网络提取并聚合局部几何和上下文信息, 并生成相机视图截锥内所有体素的占用率和对象类别的概率分布. 具体地, 网络以一个高分辨率的三维体积作为输入, 首先使用多个三维卷积层来学习局部几何表示. 使用卷积层和池化层来降低分辨率到原始输入的 1/4. 然后, 使用一个基于膨胀的三维上下文模块来捕获更高层次的对象间上下文信息. 然后将来自不同尺度网络的输出特征图连接并输入到另外两个卷积层中, 以聚合来自多个尺度的信息. 最后, 使用基于体素的 Softmax 层来预测最终的体素标签. 另外, 为了更好地传播梯度, 添加了几个短连接. 此网络命名为 SSCNet, 其网络结构如图 3 所示. 其主要贡献有两个: 一是首次将场景补全和深度图的语义标注两个任务结合起来进行处理; 二是构建了人工合成的带有密集标注的三维场景数据集——SUNCG. 自此之后, 在计算机视觉和机器视觉领域, 语义场景补全任务引起了很多学者的兴趣. 该网络的性能限制在于没有用到颜色信息, 并且该网络对 GPU 显存的较大依赖限制了输出分辨率和神经网络的深度, 此外该三维网络的计算量也非常庞大.

为了减少网络的计算量, Guo 等人[17]提出使用二维卷积神经网络代替部分三维网络. 这不仅能够减少网络的计算量, 还能从输入的图像中计算出多个特征图来作为三维投影的输入. 该网络称为 VVNet (view-volume network), 网络结构如图 4 所示. 通过将二维神经网络和三维神经网络的结合, VVNet 有效地降低了计算成本, 实现了从多通道高分辨率输入中提取特征, 从而显著提高了结果的准确性. 其具体的做法是首先输入单一深度图, 经过数个二维神经网络后, 再将输出的特征图投影为三维网格. 相比之下, SSCNet 直接将深度图投影为三维网格, 然

后将三维网格输入到三维神经网络中. VVNet 通过这种方式将网络最前面的几个三维卷积层和三维池化层替换为二维卷积层和二维池化层, 大大减少了网络的计算量. 此外, 作者还设计了一种新的扩大感受野的主干网 (见图 5), 新主干网加入了一个新的池化层, 对提取的三维特征进行采样. 在合成数据集 SUNCG 和真实数据集 NYU 上的实验证明了该方法的有效性.

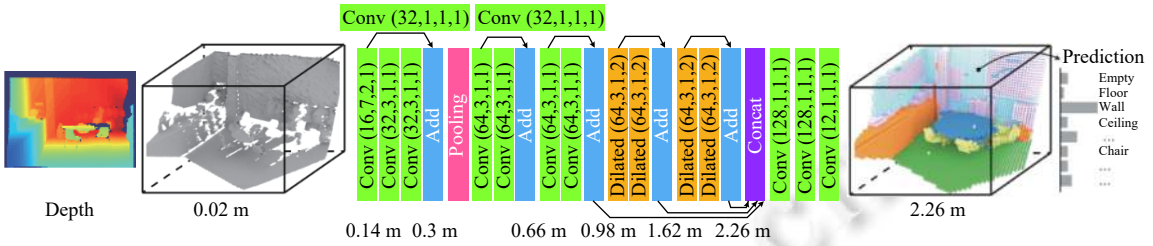


图 3 SSCNet 网络结构

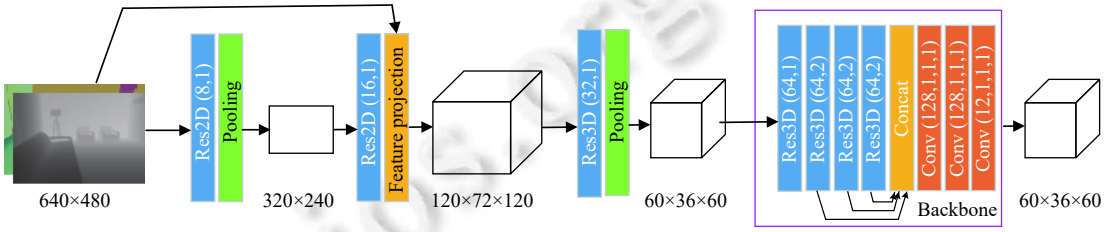


图 4 VVNet-120 网络结构, 其中, 数字“120”代表投影后三维特征图的分辨率

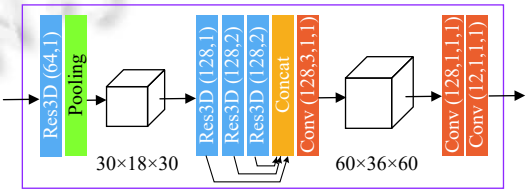


图 5 新的主干网结构

Zhang 等人[18]将密集的条件随机场 (conditional random field, CRF) 引入语义场景补全模型. 其主要思想是将 SSCNet 的输出概率图与处理后的深度图像相结合, 提出的模型称为 VD-CRF (volume data-CRF). 其主要步骤如下: 首先, 使用 TSDF 或 flipped-TSDF 来表示输入的深度图像. 然后, 将来自 SSCNet 的输出概率图和下采样的体积数据相结合, 以构建 VD-CRF 模型的势函数. 最后, 选择一种可靠的推理算法对预测进行推理. 为了方便之后的实验结果分析, 我们将该方法命名为 CRF-SSC.

受到对抗学习思路的启发, Wang 等人[19]提出了一种使用对抗学习来进行语义场景补全的方法. 我们将其命名为 ASSC. 该方法使用了多个对抗损失函数, 通过将两种潜在特征相关联 (一种是工作在部分 2.5 维数据上的编码器的潜在特征, 另一种是从训练重建完整语义场景的变分三维自动编码器中提取的潜在特征), 加强了与真值有关的实际输出的同时, 还有效地嵌入了内部特征. 此外, 在测试时, 作者在下采样时保留了原始的 2.5 维输入结构, 以提高模型内部表示的有效性. 具体地, 作者使用两个鉴别器训练框架将深度信息反投影到带有语义标签的三维体积空间中. 其中, 一个鉴别器用来将重构的语义场景与真值进行比较, 从而优化整个框架. 另一个鉴别器用来优化已学习到的潜在特征. 作者提出的网络是第一个以语义场景补全为目标的生成对抗网络, 并证明了对抗方法对于当前任务是一种有意义的选择. 该方法的缺点是深度图像的编码器与体素化真值的编码器不同, 使得编码器为了匹配两种不同的编码, 丢弃了太多的信息, 从而造成了大量的信息损失.

另外, Zhang 等人[20]引入空间分组卷积 (spatial group convolution, SGC) 来加速三维密集预测任务的计算.

SGC 将输入体素分成不同的组. 显然, 分组策略在 SGC 中起着重要的作用. 作者提出了两种不同的划分策略: 随机均匀分组和固定模式分组 (采用模运算), 然后对这些分离的组进行三维稀疏卷积. 由于在进行卷积时只考虑有效体素, 计算量显著减少, 但精度略有下降. 在语义场景补全任务上验证了所提操作的有效性. 但该方法可视化效果不尽如人意. 为了方便之后的实验结果分析, 我们将该方法命名为 ESSCNet.

Chen 等人[24]针对 Wang 等人方法[19]的缺陷, 提出了一种可以与任何三维卷积网络相结合的方法, 该方法不存在信息丢失的问题. 我们将其命名为 SSC-GAN. 特别地, 其提出了一个条件生成对抗网络来预测三维空间的语义标签和占用率. 实验证明, 条件生成对抗网络表现比标准的生成对抗网络要好, 但如果真值与深度数据没有像 NYU Kinect 那样很好地对齐, 条件生成对抗网络的表现也会差强人意.

为了提高标签一致性, Zhang 等人[25]提出了一个新的深度学习框架, 称为级联上下文金字塔网络 (cascaded context pyramid network, CCPNet), 以从单一深度图像联合推断出三维立体场景的占用率和语义标签. 其提出的 CCPNet 通过级联上下文金字塔提高了标签的一致性. 同时, 基于底层特征, 利用引导残差细化模块逐步恢复对象的精细结构. 其提出的框架有 3 个突出的优势: (1) 显式地建模三维空间上下文, 以提高性能; (2) 生成结构细节保持完好的全分辨率三维体积; (3) 提出内存需求低的轻量级模型, 具有良好的可扩展性. 具体地说, CCPNet 是一个自级联的金字塔结构, 连续聚合多尺度的三维上下文和局部几何细节, 以实现全分辨率的场景补全. 网络结构如图 6 所示, 它由 3 个关键组件组成, 即: 三维扩张卷积编码器 (dilated convolution encoder, DCE), 级联上下文金字塔 (CCP) 和引导残差细化 (guided residual refinement, GRR). 在功能上, DCE 采用核分离的多个膨胀卷积从单视图深度图像中提取三维特征表示. 然后, CCP 执行由全局到局部的上下文聚合, 以提高标签的一致性. 在上下文聚合之后, 引入 GRR, 以利用浅层学习到的低层特征来细化目标对象. 实验表明, CCPNet 显著提高了语义补全的准确性, 降低了计算成本, 并提供了高质量的全分辨率的补全结果.

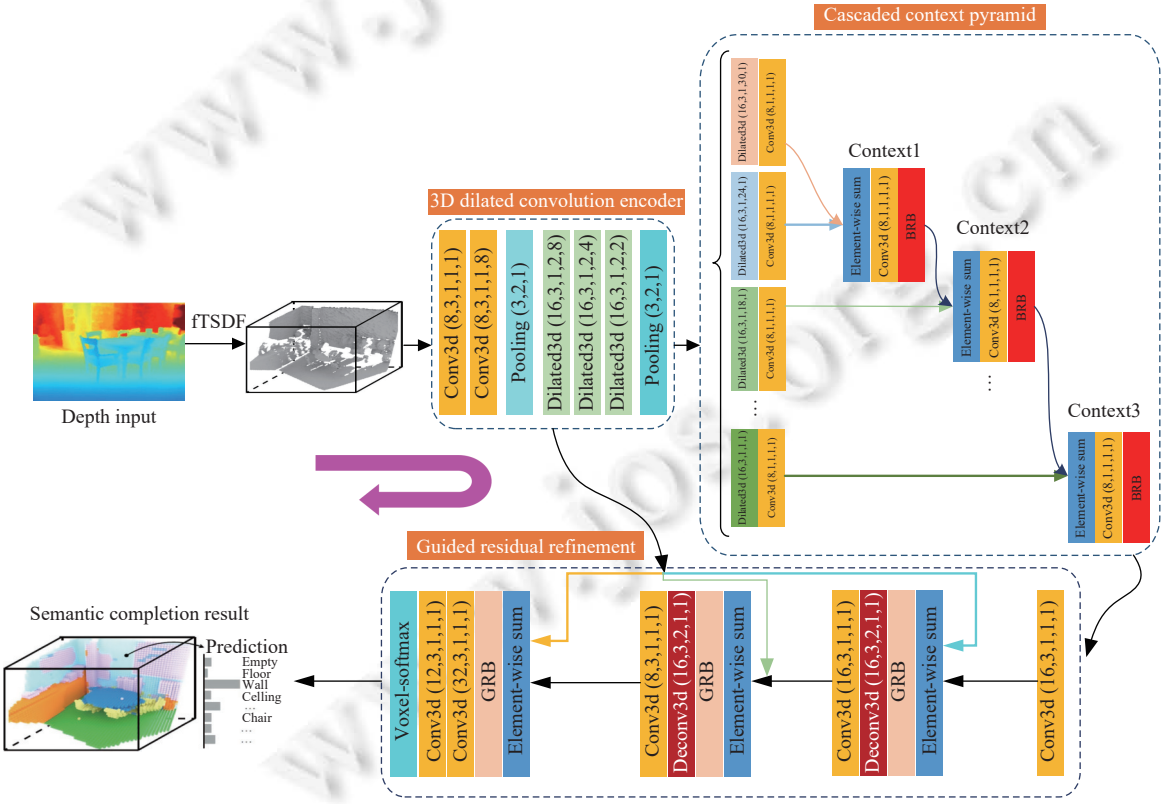


图 6 CCPNet 网络结构

Wang 等人[26]提出了一种基于单个编码器和 3 个独立生成器的单一深度图像的语义场景补全模型——ForkNet (网络结构见图 7). 该方法在下采样操作之后平行设置了 3 个生成器, 分别用于生成不完整的表面几何形状 ($\hat{x}$), 完整的几何体积 (g) 和完整的语义体积 (s). 在生成器之间添加特定的连接来使几何信息成为约束条件, 并通过经常不精确的真值注释来推动补全过程. 还利用多个鉴别器来提高重建的准确性和真实性. 作者针对语义三维场景补全和三维对象补全在标准基准上进行了测试. 对于前者, 使用 SUNCG 和 NYU 进行测试. 对于后者, 使用 ShapeNet 和 3D-RecGAN 进行测试. 此方法在真实数据集上的场景重建和对象补全均达到先进水平. 其主要贡献总结为 3 点: ① 提出一种新的架构, 基于建立在相同的共享潜在空间上的 1 个编码器和 3 个生成器, 有助于生成额外的配对训练样本; ② 利用生成器之间的特定连接, 在经常不精确的真值标注上实现几何信息约束, 并驱动补全过程; ③ 使用多个鉴别器来回归细节信息和真实的补全结果.

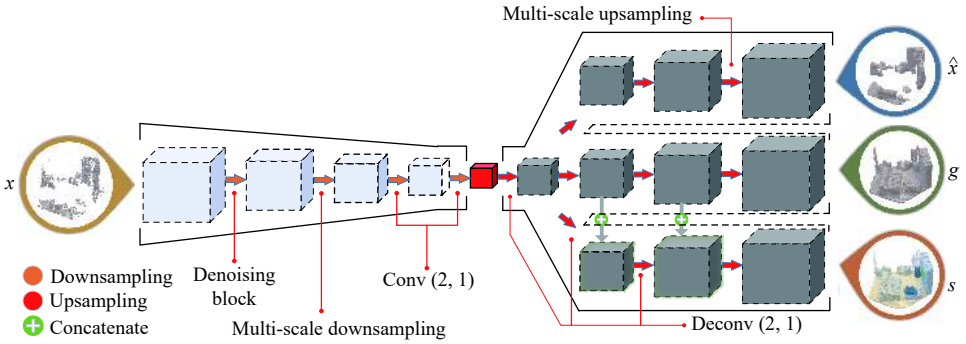


图 7 ForkNet 网络结构

Li 等人[27]在以 TSDF 编码的三维网格作为输入的基础上, 增加了一路网络分支, 并且以二维的深度图像作为输入. 该分支先采用二维卷积神经网络从深度图像中提取特征, 然后将提取的特征投影至三维空间. 随后, 在三维空间中, 对两路分支的信息进行汇合. 该网络称为 PALNet, 其结构如图 8 所示. 二维-三维特征投影算法是该网络结构的重要组成, 在 DDRNet、AICNet 等工作中, 同样采用了该算法进行特征投影. 该网络结构使用的 TSDF 分支能够保持人工设计特征的优点. 直接从深度图像提取特征的好处在于能够通过网络学习的方式获取深度图中的细粒度信息. 此外, 该工作提出了局部几何各向异性 (local geometric anisotropy, LGA) 的概念来表示体素的几何信息, 即邻域体素与当前体素的类别差异. 并根据 LGA 构建了一种新型的损失函数 PA-Loss (position aware loss). 该损失函数通过构建体素的六邻域结构, 并根据 LGA 来度量该体素的重要性. PA-Loss 的计算公式如下.

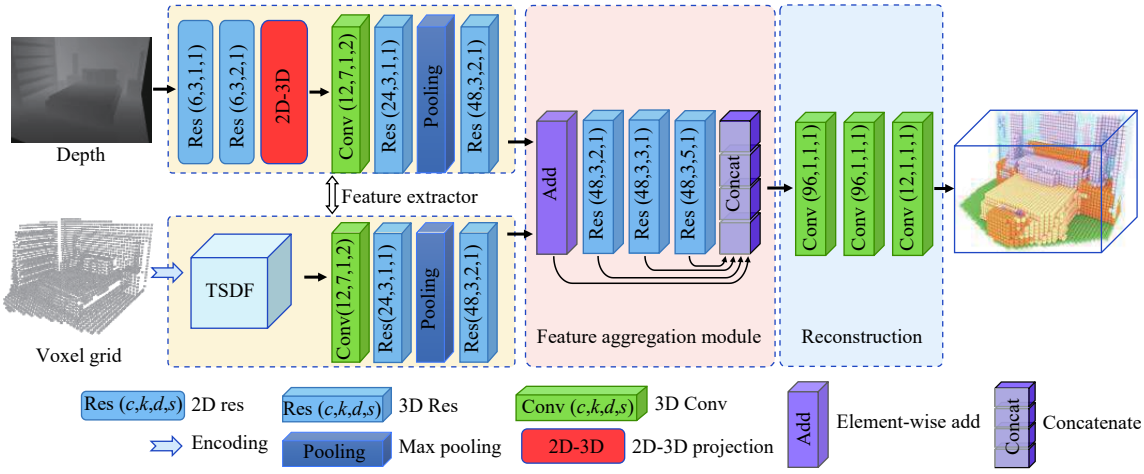


图 8 PALNet 网络结构

$$L_{PA}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}I_n y_{nc}\log\hat{y}_{nc} \tag{2}$$

其中, N 是总体素数, C 是总类别数. y_{nc}, $\hat{y}_{nc}$ 是类别 c 对应的真值标签和对应的预测结果组成的 one-hot 向量, I_n 是体素 n 的 LGA 重要性因子 (详见文献 [27]). 随后, 依次构建每个体素的训练权重, 并结合传统的交叉熵损失函数, 来提升训练过程中对几何信息更丰富的体素的训练效果.

为了将 360°图像应用在该任务中, Dourado 等人[28]提出了一种对 360°RGB 图像和相应深度图进行完整室内语义场景补全的方法. 最近的 SSC 研究仅对所使用传感器的视场所覆盖的房间小区域进行占用预测, 这意味着需要多幅图像覆盖整个场景, 这是一种不适用于动态场景的方法. 作者仅使用一张 360°图像及其对应的深度图来推断整个房间的占用率和语义标签. 具体来说, 由输入的全景深度图生成的不完整体素网格被自动划分为 8 个重叠视图, 分别提交给 3D-CNN. 结果预测由 8 个单独的预测 (文中使用了 EdgeNet[33]) 自动集合产生. 该工作是首个使用 360°成像传感器或立体球形摄像机来扩展 SSC 任务以完成场景理解的工作. 为了方便之后的实验结果分析, 我们将该方法命名为 360-SSC.

为了满足实时性要求, Chen 等人[32]提出了一种基于特征聚合策略和条件预测模块的实时语义场景补全方法, 我们称之为 RTSSC, 框架图如图 9 所示. 具体来说, 该方法采用一个基于 ResNet 的网络作为主干网络. 在神经网络的后期, 其利用扩张卷积来扩大感受野. 作者提出了全局聚合模块来融合全局上下文特征和局部特征, 并基于此提出了一种多级特征聚合策略以将全局上下文和具有不同感受野的特征结合起来. 此外, 还采用了两步预测方案来挖掘体积占用率与语义类别之间的条件关系, 首先生成二值预测, 并根据提取的特征和预测结果进行语义预测. 这样可以为语义预测提供结构信息, 进而提高性能. 该方法在一个 GTX 1080 Ti GPU 上以每秒 110 帧的速度实现了具有竞争力的性能.

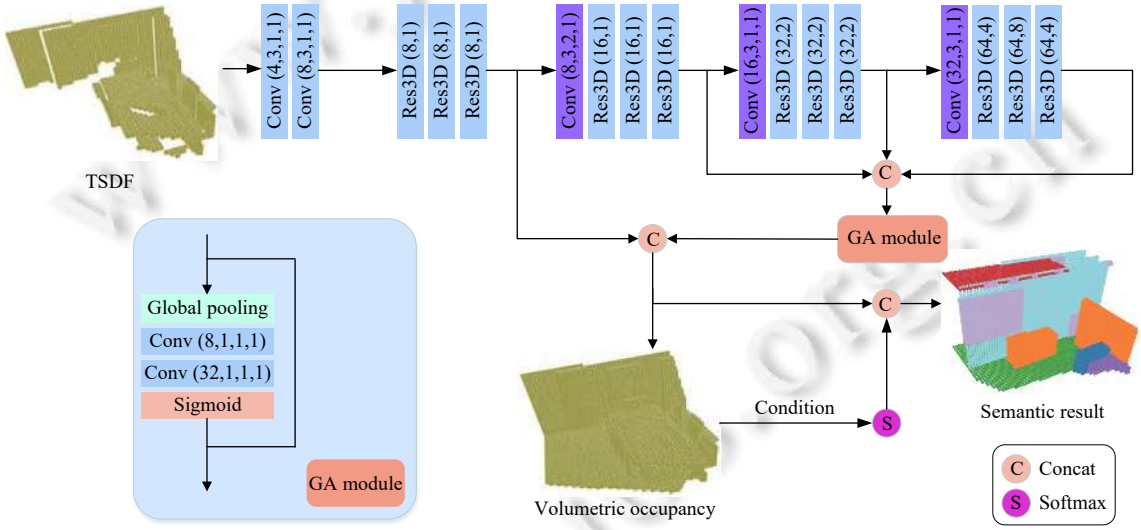


图 9 RTSSC 网络结构

3.2 基于深度图像联合彩色图像的方法

彩色图像提供了场景的颜色等信息, 有利于整个场景的理解与分析, 所以将彩色图像和深度图像作为输入来处理语义场景补全任务是非常有帮助的. 近年来也涌现出很多不错的工作.

Liu 等人[21]提出了一个特别的方法 SATNet, 与其他的直接将二维图像作为输入不同, 该方法首先对二维图像进行二维语义分割, 然后再将语义分割的结果投影成三维图像, 最后进行三维语义场景补全. 此框架具有 3 个优点: (1) 显式的语义分割显著提高了性能; (2) 传感器数据融合方式灵活, 可扩展性好; (3) 任何子任务的进展都会促

进整体效果. 具体来说, SATNet 的核心思想是借助低层语义分割来帮助实现高层语义场景补全. 它包含 3 个模块: 二维语义分割子网络 SNet, 通过单个彩色图像或深度图像来估计语义信息; 二维-三维重投影层, 将二维语义转换为三维空间; 三维语义场景补全子网络-TNet, 对体素进行处理以实现整个场景的语义补全. 对于单个深度图像或 RGB-D 输入, SATNet 能够以单分支的方式高效地完成语义场景补全. 此外, 对于 RGB-D 输入, 还提出了一种更高效的双分支结构来融合这两种信息, 并且得到了较好的补全结果.

Garbade 等人[22]提出了一种利用深度信息和从 RGB 图像中推断出的语义信息的双流方法 (TS3D) 来完成这项任务. 该方法构造了一个不完整的三维语义张量, 该语义张量采用紧凑的三通道编码方法对推导出的语义信息进行编码, 并利用三维 CNN 对完整的三维语义张量进行推导. 具体来说, 对 SSCNet 方法进行了扩展, 保留了其有益的上下文合并策略和端到端可训练性, 并同时对其进行了修改, 以便在输入阶段和损失阶段能够利用从 RGB 图像推断出的语义信息. 对于单个 RGB-D 图像, 首先使用二维 CNN 从 RGB 数据中推断语义标签, 并构造一个不完整的三维语义张量. 为此, 将推断出的语义标签映射到三维空间中, 并通过推断出的类别标签标记每个可见的表面体素. 此三维语义张量是不完整的, 因为它只包含可见体素的标签, 而不包含闭塞体素的标签. 其次, 将深度图像进行三维投影生成其体积表示. 最后, 该张量被用作三维 CNN 的输入, 从而推断出一个完整的三维语义张量, 其中包括所有体素的占用率和语义标签.

Li 等人[23]从三维卷积核的设计角度出发, 提出了一种简单高效的轻量化三维卷积模块, 即维度解耦卷积 (DDR) 模块. 图 10 展示了 DDR 模块与残差模块的不同结构. 维度解耦卷积通过在卷积层和连接层上进行尺寸拆分, 能够有效地减少网络参数. 同时, 基于 DDR 还构建了轻量化的 ASPP 模块, 利用多尺度信息提升了网络对不同大小的物体的适应能力. 进一步, 对于语义场景补全任务, 通过多尺度融合深度信息和颜色信息, 提出了一种轻量化深度网络-DDRNet, 该方法与 SSCNet 相比, 仅使用了极少的网络参数 (参数量减少了约 83%), 但补全结果得到了提升. 比较结果见表 1.

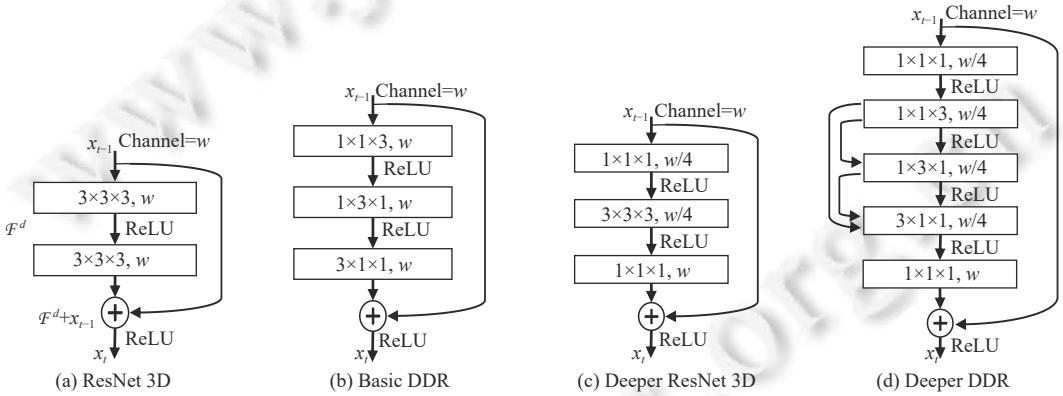


图 10 残差块和 DDR 块

表 1 DDRNet 与其他方法的参数量, FLOPs 和性能比较结果

方法	参数量 (k)	FLOPs (G)	SC-IoU (%)	SSC-IoU (%)
SSCNet	930.0	163.8	55.1	24.7
ESSCNet	—	22.0	56.2	26.7
DDRNet-D	155.0	20.6	59.0	28.9
DDRNet-RGBD	195.0	27.2	61.0	30.4

最近, Li 等人[29]提出了一种基于注意力的端到端三维卷积网络 (AMFNet) 用于语义场景补全任务, 该方法通过基于二维语义分割的多模态融合架构和基于残差注意块的三维语义补全网络来实现. 该方法不仅利用了颜色信息和深度信息, 而且借鉴了深度网络学习二维语义分割的经验. 该方法的整体框架图见图 11. 图 11 以 RGB-D 图

像 (分别为 RGB 和 HHA 图像) 为输入, HHA 编码方式是指使用 3 个通道对深度图重新编码, 这 3 个通道分别是水平视差、高于地面的高度和像素的局部表面与推断重力方向的倾角. 具体来说, 训练了一个使用二维语义分割信息的模型, 并用其指导 SSC 任务中的三维补全和语义分割. 该模型假设对二维语义分割任务有效的图像特征对语义场景补全同样有效. 基于这一假设, 提出了一种新的双分支多模态融合网络, 该网络在二维分割的基础上增强了三维分割信息, 从而可以有效地指导三维补全和语义标注. 此外, 由于三维数据固有的稀疏性 (在三维场景中大多数体素是空的), 其认为一些与空体素相关的特征 (例如颜色和纹理) 是没有价值的, 而注意力机制能够使得模型具有聚焦于重要部分的能力. 因此, 提出了一种基于残差注意块 (residual attention block, RAB) 的三维网络, 以充分利用空间划分中可靠的深度线索来完成语义场景补全. RAB 的结构如图 12 所示.

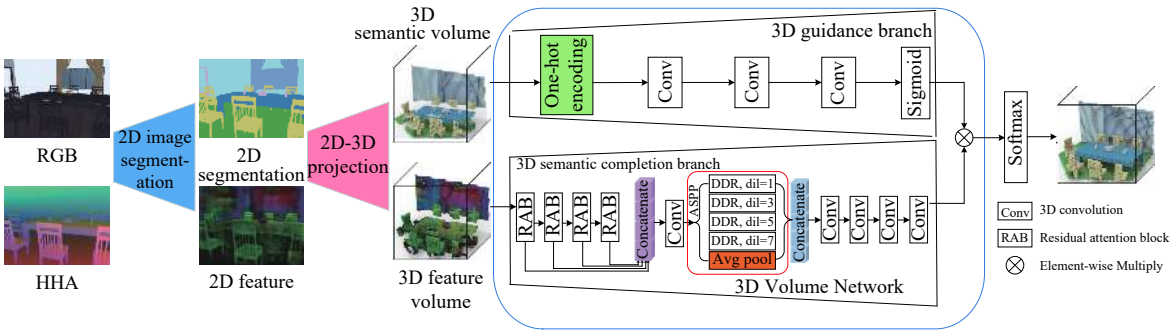


图 11 AMFNet 的网络结构

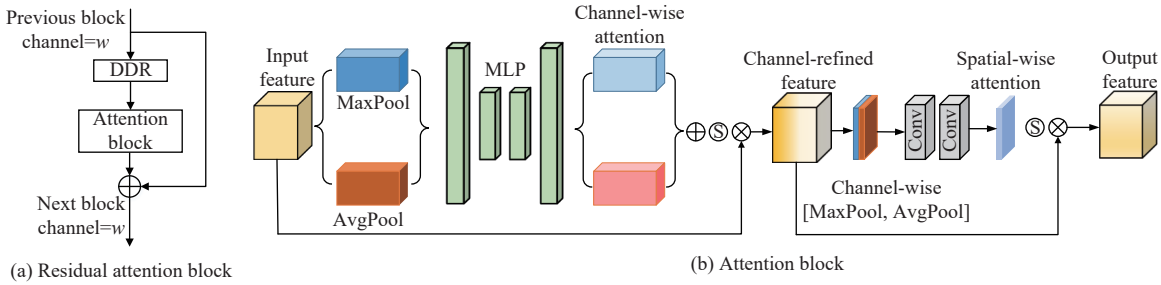


图 12 RAB 的结构说明. RAB 的结构类似于 DDR 块[23], 但同时注入了通道和空间注意

在之前的研究中, 体素分辨率是导致性能瓶颈的关键困难之一. 为了解决这一问题, Chen 等人[30]设计了一种新的几何策略用来将低分辨率的体素表示嵌入到深度信息中. 为此, 作者首先提出了一种新的三维草图感知特征嵌入方法来显式有效地编码几何信息. 随后进一步设计了一个简单而有效的语义场景补全框架 (Sketch-Net), 该框架包含一个轻量级的三维草图幻觉模块, 其通过条件变分自编码器 (CVAE) 的半监督结构先验性来指导空间占用和语义标签的推理. 所提方法的框架图见图 13. 作者证明了其提出的几何嵌入比从常用的 SSC 框架中学习到的深度特征更有效. 该方法的实验结果在 3 个公共数据集上都超过了最先进的水平, 而这只需要三维体积的输入和输出分辨率为 60×36×60.

语义场景补全任务面临的关键挑战是如何有效地利用三维上下文来建模各种形状、布局和视觉特性有严重变化的物体. 为此, Li 等人[31,41]基于 DDRNet, 对卷积核的可变性进行了研究, 并提出了一种新型模块结构, 即各向异性卷积 (AIC) 模块 (示例见图 14). 和以往的模块不同, AIC 模块拥有可变形状的卷积核, 并且计算量小, 参数效率高. 它还可以作为一个即插即用模块来代替标准的三维卷积单元. 具体来说, 该方法的基本思想是将三维卷积运算分解为 3 个连续的一维卷积, 并为每个此类一维卷积配备不同内核大小的混合器, 随后沿着每个一维卷积智能地学习此类内核的组合权重. 因此, 基本上可以通过连续执行此类自适应一维卷积来建模各向异性三维上下文. 此外, 进一步提出了一种新的用于语义场景补全的各向异性卷积网络-AICNet, 实验结果证明了其有效性.

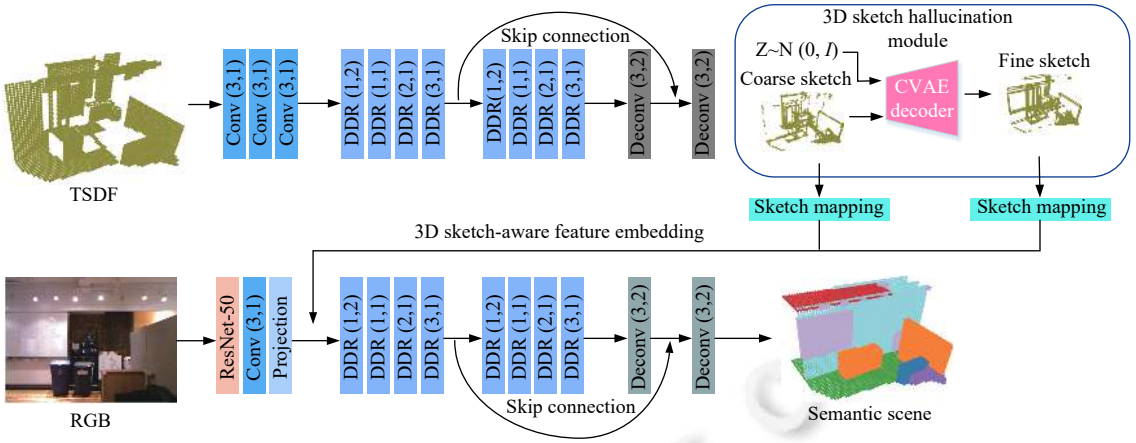


图 13 Sketch-Net 框架图

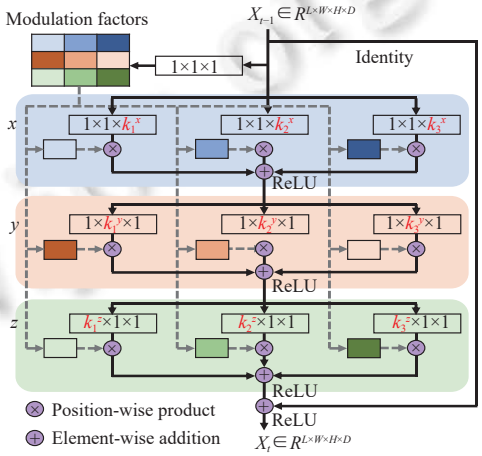


图 14 各向异性卷积模块

近来, Cai 等人[34]提出了一种基于实例级和场景级语义信息的场景-实例-场景网络 (SISNet) 框架. 该方法能够推断细粒度的形状细节以及邻近的对象, 这些对象的语义类别容易混淆. 关键是该方法将实例从一个粗略补全的语义场景中分离出来, 而不是用一个原始的输入图像来指导实例和整个场景的重建. SISNet 进行了场景到实例 (SI) 和实例到场景 (IS) 的迭代语义补全. 具体来说, SI 能够编码物体的周围环境, 以便有效地将实例从场景中解耦, 并且每个实例都可以被体素化为更高的分辨率, 以捕捉更精细的细节. 而使用 IS, 细粒度的实例信息可以集成到 3D 场景中, 从而导致更准确的语义场景补全. 这个设计的灵感来自物体是场景的主要组成部分, 它们与场景紧密相关. 例如, 当窗户重建得很好时, 可以很容易地推断出周围的墙. 利用这种迭代机制, 场景和实例补全相互受益, 以获得更高的补全精度.

4 数据集和评价指标

4.1 数据集

近年来语义场景补全任务在科研中的关注逐渐得到提高, 而解决这一任务, 大量的数据是不可缺少的. 以下将详细介绍常用的 3 个数据集: NYUv2, NYUCAD 和 SUNCG. 数据集的基本信息见表 2.

NYUv2 数据集由 Silberman 等人[4]于 2012 年提出. NYUv2 数据集使用 Microsoft Kinect 的 RGB 和 Depth 摄

像机从个不同的美国城市的大量商业和住宅建筑中收集数据, 它由 1 449 组密集标记的 RGB 和深度图像对组成, 包含 26 个场景类别的 464 个不同的室内场景. 使用 Amazon Mechanical Turk 对每张图像进行密集的逐像素标记. 如果一个场景包含一个对象类的多个实例, 那么每个实例都会收到一个唯一的实例标签, 例如同一个图像中的两个不同的杯子会被标记为: 杯子 1 和杯子 2, 以唯一地标识它们. 该数据集包含 35 064 个不同的对象, 跨越 894 个不同的类别. 在训练时, 将其分为 795 个训练样本和 654 个测试样本. 由于室内场景的复杂性和 Kinect 数据采集造成的深度图像测量误差, NYUv2 是一个具有挑战性的数据集.

表 2　语义场景补全数据集信息

数据集	训练集样本数量	测试集样本数量	下载地址
NYUv2[4]	795	654	https://github.com/shurans/sscnet
NYUCAD[13]	795	654	
SUNCG-D[1]	139 368	470	
SUNCG-RGBD[21]	13 011	499	https://github.com/ShiceLiu/SATNet

由于 NYU 数据集中一些手工标记的体积及其对应的深度图像没有很好地对齐, 所以 Firman 等人[13]提出了 NYUCAD 数据集, 其深度图是根据标记的体积绘制的.

SUNCG 数据集是一个大规模的合成场景数据集, 其包含 45 622 个不同的真实房间和家具布局的场景, 这些都是通过 Planner5D 平台手工创建的. 整个数据集有 49 884 个有效楼层, 其中包含 404 058 个房间和来自 2 644 个唯一对象网格的 5 697 217 个对象实例, 这些对象网格覆盖了 84 个类别. 作者手动标记库中的所有对象, 以分配类别标签. SUNCG 数据集包括两类: SUNCG-D 和 SUNCG-RGBD. 其中 SUNCG-D 由 Song 等人[1]于 2017 年提出, 由 139 368 个训练样本和 470 个测试样本组成; SUNCG-RGBD 由 Liu 等人[21]在 2018 年提出, 该数据集由 13 011 个训练样本和 499 个测试样本组成.

4.2 评价指标

对于语义场景补全算法的性能, 通常使用真值与预测结果之间的交并比 (intersection over union, *IoU*) 来进行评价. 同时, 也考察了形状补全子任务的性能评价.

对于形状补全子任务, 将预测结果中所有的物体均视为一个类别, 使其成为二分类任务. 这个二分类任务的目标是将视野范围内体素划分为空体素与被物体占据的已占用体素. 如果体素属于任何语义类别, 则将其视为已占用的体素. 其中, 已占用体素不仅包括从深度图对应的视角下可见的物体表面所对应的体素, 而且包括被遮挡的部分. 在这个二分类任务中考察已占用体素的预测结果. 将预测的已占用体素结果中的预测正确的样本设置为“正阳性”(*TP*), 预测错误的样本设置为“假阳性 (*FP*)”, 而未被检测到的已占用体素设置为“假阴性 (*FN*)”. 由此便可分别得到形状补全预测结果的精度 (precision, *P*) 和召回 (recall, *R*) 如下:

$$P = \frac{TP}{TP+FP} \tag{3}$$

$$R = \frac{TP}{TP+FN} \tag{4}$$

同时, 可以计算 *IoU* 作为形状补全这个子任务的综合评价指标. *IoU* 是预测结果与真值的交集比上它们的并集, 其定义如下:

$$IoU = \frac{\text{预测结果} \cap \text{真值}}{\text{预测结果} \cup \text{真值}} = \frac{TP}{TP+FP+FN} \tag{5}$$

可以简单地理解为, *IoU* 就是模型的预测结果和真值的重叠率. 比值越高则预测结果越准确, 最理想情况是预测结果与真值完全重叠, 即比值为 1.

对于语义场景补全任务, 评估每个类别的预测体素标签和真值标签之间的交并集 (*IoU*). 这一指标反映了每个类别的语义分割结果的准确性. 在获得每个类别的 *IoU* 后, 通过计算所有类别的平均 *IoU* (mIoU) 作为整体性能的评价指标. 因此, mIoU 是语义场景补全任务中最为重要的指标.

5 现有方法的性能评价及分析

我们总结了现有方法在 3 个常用数据集上的实验结果[1,13−34], 即 NYU、NYUCAD 和 SUNCG. 其中包括传统的方法, 如 Zheng 等人、Lin 等人、Geiger 和 Firman; 基于深度图像的方法, 如 SSCNet、VVNet、ESSCNet、ASSC、CRF-SSC、SSC-GAN、CCPNet、PALNet、ForkNet、360-SSC 和 RTSSC; 基于深度图像联合彩色图像的方法, 如 SATNet、TS3D、DDRNet、EdgeNet、Sketch-SSC、AICNet、AMFNet 和 SISNet. 实验结果总结在表 3–表 7 中, 最优性能指标用粗体表示. 表 3 和表 4 分别展示了现有方法在 NYU 和 NYUCAD 数据集上的结果, 表 5 和表 6 展示了现有方法在 SUNCG 数据集上的结果, 表 7 梳理了当前关注到时间效率的方法的实验结果. “*”表示基于深度图像联合彩色图像的深度学习方法. “(SUNCG)”表示使用 NYU/NYUCAD+SUNCG 来训练模型.

表 3 NYU 数据集结果 (%)

方法	场景补全			语义场景补全											
	精确率	召回率	*IoU*	天花板	地板	墙	窗户	椅子	床	沙发	桌子	电视	家具	物体	平均值
Lin等人[15]	58.5	49.9	36.4	0	11.7	13.3	14.1	9.4	29	24	6.0	7.0	16.2	1.1	12.0
Geiger等人[16]	65.7	58	44.4	10.2	62.5	19.1	5.8	8.5	40.6	27.7	7.0	6.0	22.6	5.9	19.6
SSCNet[1]	57.0	**94.5**	55.1	15.1	94.7	24.4	0.0	12.6	32 .1	35.0	13.0	7.8	27.1	10.1	24.7
SSCNet (SUNCG)[1]	59.3	92.9	56.6	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	13.9	31.1	12.6	30.5
ESSCNet[20]	71.9	71.9	56.2	17.5	75.4	25.8	6.7	15.3	53.8	42.4	11.2	0	33.4	11.8	26.7
VVNetR-120[17]	69.8	83.1	61.1	19.3	94.8	28.0	12.2	19.6	57.0	50.5	17.6	11.9	35.6	15.3	32.9
CRF-SSC[18]	—	—	60.0	18.1	92.6	27.1	10.8	18.8	54.3	47.9	17.1	15.1	34.7	13.0	31.8
PALNet[27]	68.7	85.0	61.3	23.5	92.0	33.0	11.6	20.1	53.9	48.1	16.2	24.2	37.8	14.7	34.1
ASSC[19]	—	—	—	49.6	42.7	51.2	24.2	23.0	28.1	30.4	29.9	—	22.0	11.5	31.3
SSC-GAN[24]	63.1	87.8	57.8	—	—	—	—	—	—	—	—	—	—	—	22.7
ForkNet[26]	—	—	63.4	36.2	93.8	29.2	18.9	17.7	61.6	52.9	23.3	19.5	45.4	20.0	37.1
CCPNet[25]	74.2	90.8	63.5	23.5	96.3	35.7	20.2	25.8	61.4	56.1	18.1	28.1	37.8	20.1	38.5
CCPNet (SUNCG)[25]	78.8	94.3	67.1	25.5	98.5	38.8	27.1	27.3	64.8	58.4	21.5	30.1	38.4	23.8	41.3
360-SSC[28]	—	—	—	15.6	92.8	50.6	6.6	26.7	—	35.4	33.6	—	32.2	15.4	34.3
RTSSC[32]	—	—	73.4	—	—	—	—	—	—	—	—	—	—	—	34.4
DDRNet[23]*	71.5	80.8	61.0	21.1	92.2	33.5	6.8	14.8	48.3	42.3	13.2	13.9	35.3	13.2	30.4
TS3D[22]*	—	—	60.0	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
SATNet[21]*	67.3	85.8	60.6	17.3	92.1	28.0	16.6	19.3	57.5	53.8	17.2	18.5	38.4	18.9	34.4
NoSNet[21]*	67.3	84.4	59.5	19.8	94.5	28.1	0.7	15.4	50.8	43.2	15.7	11.0	31.9	7.7	29.0
Sketch-Net[30]*	85.0	81.6	71.3	43.1	93.6	40.5	24.3	30.0	57.1	49.3	29.2	14.3	42.5	28.6	41.1
EdgeNet[33]*	76.0	68.3	56.1	17.9	94.0	27.8	2.1	9.5	51.8	44.3	9.4	3.6	32.5	12.7	27.8
EdgeNet (SUNCG)[33]*	79.1	66.6	56.7	22.4	95.0	29.7	15.5	20.9	54.1	53.0	15.6	14.9	35.0	14.8	33.7
AICNet[31]*	62.4	91.8	59.2	23.2	90.8	32.3	14.8	18.2	51.1	44.8	15.2	22.4	38.3	15.7	33.3
AMFNet[29]*	67.9	82.3	59.0	16.7	89.2	27.3	19.2	20.2	56.1	50.4	15.1	13.5	36.8	18.0	33.0
AMFNet (w/o-Attn)[29]*	64.5	86.5	58.6	21.3	90.3	26.1	7.7	18.0	53.8	48.4	13.0	0	36.7	16.3	30.1
SISNet[34]*	**90.7**	84.6	**77.8**	53.9	93.2	51.3	38.0	38.7	65.0	56.3	37.8	25.9	51.3	36.0	**49.8**

对于传统的方法, Zheng 等人[14]和 Firman 等人[13]的方法只能解决场景补全任务, 并且从表 4 可以看出, 这两个方法在 NYUCAD 上的 *IoU* 分别仅为 34.6% 和 50.8%, 远远低于后来提出的方法. 而 Lin 等人[15]和 Geiger 等人[16]的方法都以 RGB-D 帧为输入, 在三维场景中生成对象标签. Lin 等人[15]使用三维边框和平面来近似所有对象. Geiger 等人[16]对测试时观测到的深度图检索并拟合三维网格模型. 用于检索的网格模型库是用于真值标注模型的超集. 因此, 他们可以通过在一个小的数据库中找到精确的网格模型来实现完美的对齐. 这两种传统方法都可以完成语义场景补全任务, 但得到的效果并不好. 从表 3 可以看出, Lin 等人[15]和 Geiger 等人[16]的方法在 NYU 上的平均 *IoU* 值分别仅为 12% 和 19.6%. 基于此, 研究者们开始关注深度学习方法在此任务中的应用.

表 4 NYUCAD 数据集结果 (%)

方法	场景补全			语义场景补全											
	精确率	召回率	*IoU*	天花板	地板	墙	窗户	椅子	床	沙发	桌子	电视	家具	物体	平均值
Zheng等人[14]	60.1	46.7	34.6	—	—	—	—	—	—	—	—	—	—	—	—
Firman等人[13]	66.5	69.7	50.8	—	—	—	—	—	—	—	—	—	—	—	—
SSCNet[1]	75.4	**96.3**	73.2	32.5	92.6	40.2	8.9	33.9	57.0	59.5	28.3	8.1	44.8	25.1	39.2
VVNetR-120[17]	86.4	92.0	80.3	—	—	—	—	—	—	—	—	—	—	—	—
CRF-SSC[18]	—	—	78.4	35.5	92.6	52.4	10.7	39.9	60.0	62.5	34.0	9.4	49.2	26.5	43.0
PALNet[27]	87.2	91.7	80.8	54.8	92.8	60.3	15.3	43.1	60.7	59.9	37.6	8.1	48.6	31.7	46.6
SSC-GAN[24]	81.0	91.0	74.9	—	—	—	—	—	—	—	—	—	—	—	42.3
CCPNet[25]	91.3	92.6	82.4	56.2	94.6	58.7	35.1	44.8	68.6	65.3	37.6	35.5	53.1	35.2	53.2
CCPNet (SUNCG)[25]	93.4	91.2	85.1	58.1	95.1	60.5	36.8	47.2	69.3	67.7	39.8	37.6	55.4	37.6	55.0
RTSSC[32]	—	—	82.2	—	—	—	—	—	—	—	—	—	—	—	44.5
TS3D[22]*	—	—	76.1	25.9	93.8	48.9	33.4	31.2	66.1	56.4	31.6	38.5	51.4	30.8	46.2
DDRNet[23]*	88.7	88.5	79.4	54.1	91.5	56.4	14.9	37.0	55.7	51.0	28.8	9.2	44.1	27.8	42.8
Sketch-Net[30]*	90.6	92.2	84.2	59.7	94.3	64.3	32.6	51.7	72	68.7	45.9	19.0	60.5	38.5	55.2
AICNet[31]*	88.2	90.3	80.5	53.0	91.2	57.2	20.2	44.6	58.4	56.2	36.2	9.7	47.1	30.4	45.8
SISNet[34]*	**94.2**	91.3	**86.5**	65.6	94.4	67.1	45.2	57.2	75.5	66.4	50.9	31.1	62.5	42.9	**59.9**

表 5 SUNCG-D 数据集结果 (%)

方法	场景补全			语义场景补全											
	精确率	召回率	*IoU*	天花板	地板	墙	窗户	椅子	床	沙发	桌子	电视	家具	物体	平均值
SSCNet[1]	76.3	95.2	73.5	96.3	84.9	56.8	28.2	21.3	56.0	52.7	33.7	10.9	44.3	25.4	46.4
SATNet[21]	80.7	96.5	78.5	97.9	82.5	57.7	58.5	45.1	78.4	72.3	47.3	45.7	67.1	55.2	64.3
ESSCNet[20]	92.6	90.4	84.5	96.6	83.7	74.9	59.0	55.1	83.3	78.0	61.5	47.4	73.5	62.9	70.5
VVNetR-120[17]	90.8	91.7	84.0	98.4	87.0	61.0	54.8	49.3	83.0	75.5	55.1	43.5	68.8	57.7	66.7
ASSC[19]	—	—	—	41.4	37.7	45.8	26.5	21.8	25.4	23.7	20.1	—	16.2	5.7	26.4
SSC-GAN[24]	83.4	92.4	78.1	—	—	—	—	—	—	—	—	—	—	—	55.6
CRF-SSC[18]	—	—	74.5	95.4	84.3	57.7	24.5	28.2	63.4	55.3	34.5	19.6	45.8	28.7	48.9
ForkNet[26]	—	—	86.9	95.0	85.9	73.2	54.5	46.0	81.3	74.2	42.8	31.9	63.1	49.3	63.4
CCPNet[25]	**98.2**	**96.8**	**91.4**	99.2	89.3	76.2	63.3	58.2	86.1	82.6	65.6	53.2	76.8	65.2	**74.2**
RTSSC[32]	—	—	84.8	—	—	—	—	—	—	—	—	—	—	—	63.5

表 6 SUNCG-RGBD 数据集结果 (%)

方法	场景补全			语义场景补全											
	精确率	召回率	*IoU*	天花板	地板	墙	窗户	椅子	床	沙发	桌子	电视	家具	物体	平均值
SSCNet[1]	43.5	90.7	41.5	64.9	60.1	57.6	25.2	25.5	40.4	37.9	23.1	29.8	45.7	4.7	37.7
SATNet[21]*	56.7	91.7	53.9	65.5	60.7	50.3	56.4	26.1	47.3	43.7	30.6	37.2	44.9	30.0	44.8
NoSNet[21]*	50.8	92.8	48.8	66.6	56.7	41.7	30.6	22.7	47.1	36.4	22.1	25.2	30.7	13.4	35.7
Sketch-Net[30]*	**94.1**	86.2	81.8	77.9	82.3	68.4	57.9	35.7	71.8	63.7	45.1	12.8	64.2	32.0	55.6
EdgeNet[33]*	93.7	90.3	85.1	97.2	94.9	78.6	57.4	49.5	80.5	74.4	55.8	51.9	70.1	62.5	70.3
AMFNet[29]*	57.5	91.6	54.5	80.4	69.1	55.0	60.4	27.0	42.2	46.7	32.5	42.3	36.9	27.4	47.3
AMFNet(w/o-Attn)[29]*	54.8	94.7	53.2	79.6	66.9	51.7	60.2	26.5	38.2	45.5	24.5	27.9	38.1	28.7	44.3
SISNet[34]*	93.3	**96.1**	**89.9**	85.2	90.0	83.7	80.8	60.0	83.5	80.9	68.6	77.3	86.7	70.1	**78.8**

对于基于单一深度图像的深度学习方法, 从表 3 可以看出, SSCNet[1]作为首个语义场景补全方法, 其在 NYU 上的平均 *IoU* 达到了 24.7%, 使用 SUNCG 训练后更是达到了 30.5%, 相比 Lin 等人[15]和 Geiger 等人[16]的方法分

别提升了 18.5% 和 10.9%. 这说明将语义分割和场景补全结合的方法的确比单独做其中一个任务的性能更好, 此外, 从表 4–表 6 可以看出, SSCNet 的平均 *IoU* 分别为 39.2%、46.4% 和 37.7%. 从表 3 和表 5 可以看出, Zhang 等人提出的 ESSCNet[20]在 NYU 上相比 SSCNet 提升了 2%, 而在 SUNCG-D 上足足提升了 24.1%, 说明稀疏卷积也能提高此任务的性能. 从表 3 和表 5 看出, Guo 等人[17]提出的 VVNet 将性能进一步提升到了 32.9% 和 66.7%, 这表明适当地使用二维卷积代替三维卷积可以在不影响性能的情况下有效地减少内存占用和训练时间. 从表 3 看出 PALNet[27]在 NYU 数据集上的平均 *IoU* 达到 34.1%, 虽然与 VVNet 相比没有太大优势, 但是 PALoss 解决了体素位置的重要性的问题, 使得网络更加关注边缘和角部. 从表 3–表 5 可以看出, 使用 GAN 的 ASSC[19]和 SSC-GAN[24]虽然可以解决语义场景补全问题, 但是得到的结果却不尽如人意, 说明 GAN 对于语义场景补全任务的应用价值还有待商榷. 另外, 从表 3–表 5 可以看出, 利用 CRF 的 CRF-SSC[18]得到的平均 *IoU* 分别为 31.8%、43.0% 和 48.9%, 表明条件随机场的使用对于该任务是有帮助的. 从表 3 和表 5 可以看出, ForkNet[26]的平均 *IoU* 分别达到了 37.1% 和 63.4%, 表明此叉子结构可以很好地解决该问题. 从表 3–表 5 可以看出, CCPNet[25]在 NYU 上的平均 *IoU* 达到了 41.3%, 在 NYUCAD 上达到了 55.0%, 在 SUNCG-D 上达到了最高值 74.2%, 这是由于该方法使用级联金字塔网络, 充分融合了各个不同尺寸的特征. 从表 3 可以看出, 将 360°图像应用在该任务中的 360-SSC[28]方法也可以得到不错的结果, 平均 *IoU* 值为 34.3%. 从表 3 可以看出, RTSSC[32]的 *IoU* 值为 73.4%, 其平均 *IoU* 也有所提升, 并且从表 7 中可以看出 RTSSC 的模型参数和推理速度皆达到了最佳, 表明所提方法对于语义场景补全任务的实时性和准确性都有所提升.

表 7　语义场景补全方法的时间效率对比

方法	参数 (k)	FLOPs (G)	方法	参数 (k)	FLOPs (G)
SSCNet[1]	930.0	163.8	SATNet[21]	1 200	187.5
DDRNet[23]*	195.0	27.2	PALNet[27]	223.0	78.8
AICNet[31]*	847.0	113.7	CCPNet[25]	89	11.8
VVNet[17]	685.0	119.2	RTSSC[32]	**65**	**1.6**
ESSCNet[20]	160.0	22.0			

对于基于彩色图像联合深度图像的深度学习方法, 从表 3 和表 4 可以看出, TS3D[22]的平均 *IoU* 在 NYU 和 NYUCAD 上分别达到了 34.1% 和 46.2%, 表明其提出的双流方法对于此任务是有帮助的, 并且构造的三维语义张量采用紧凑的三通道编码方法对推导出的语义信息进行编码, 更加提高了预测准确度. 从表 3 和表 6 中可以看出, SATNet[21]在 NYU 和 SUNCG-RGBD 上的平均 *IoU* 值分别为 34.4% 和 44.8%, 此外, 对于每一类, 还得到较高的 *IoU*, 并且比不使用语义分割的 NoSNet[21]的值要高. 图 15 表示使用二维语义分割初始化时得到的补全损失比未初始化时下降得更快, 进一步证明了深度图或彩色图的二维语义分割可以促进最终的补全. 从表 3 和表 4 可以看出, DDRNet[23]的平均 *IoU* 在 NYU 和 NYUCAD 上分别达到了 30.4% 和 42.8%, 表明了设计的新架构, 利用了来自多层次和多模态的鲁棒特征和数据融合, 有效地补充了从彩色图像到无纹理对象的细节. 从表 3 可以看出, EdgeNet[33]在 NYU 上的平均 *IoU* 值为 33.7%, 表明此网络能够处理由深度信息和边缘信息融合而成的特征. 从表 3、表 4 和表 6 可以看出 Sketch-Net[30]的性能较之前的方法有明显提升, 这一改进是由于新颖的两阶段结构充分利用了结构先验知识, 所提供的结构先验能准确推断出场景中不可见的区域, 且细节结构保持良好. 这一点可以从表 8 看出, 草图是建模结构先验的最佳表示 (最优指标用粗体表示), 因为它可以推断出细节结构保持良好的不可见区域. 并且值得注意的是, 尽管一些工作使用了比 Sketch-Net 更大的输入和输出分辨率, 但输入和输出分辨率为 60×36×60 的 Sketch-Net 仍优于它们. 从表 3 和表 4 可看出, AICNet[31,41]的平均 *IoU* 在 NYU 和 NYUCAD 上分别达到了 33.3% 和 45.8%, 虽然此方法的性能不是最好的, 但是此网络使用各向异性卷积代替标准的三维卷积. 克服了三维卷积的缺点: (1) 固定的感受域对于处理物体变化并不理想; (2) 三维卷积是资源密集型的, 限制了三维网络的深度, 从而牺牲了建模能力. 从表 3 和表 6 可以看出 AMFNet[29]的平均 *IoU* 在 NYU 和 SUNCG-RGBD 上分别达到了 33% 和 47.3%, 比不使用注意力机制时的 AMFNet(w/o-Attn) 的值要高, 其中在 SUNCG-RGBD 上达到了目前最高

水平, 表明其设计的双分支多模态融合结构和残差注意模块可以有效地提取重要信息, 从而完成语义场景补全任务. 从表 3 和表 4 观察到, SISNet[34]比所有现有的方法有很大的优势, 更具体地说, 该方法的 *IoU* (场景补全) 和平均 *IoU* (语义场景补全) 在 NYU 上均达到了最高值, 分别为 77.8% 和 49.8%, 在 NYUCAD 上也均达到了最高值, 分别为 86.5% 和 59.9%. 并且从表 6 也可以看出 SISNet 在 SUNCG-RGBD 上的性能也达到了目前最优水平. 实验结果充分验证了场景到实例 (SI) 和实例到场景 (IS) 的迭代语义补全模型可以提升补全精度.

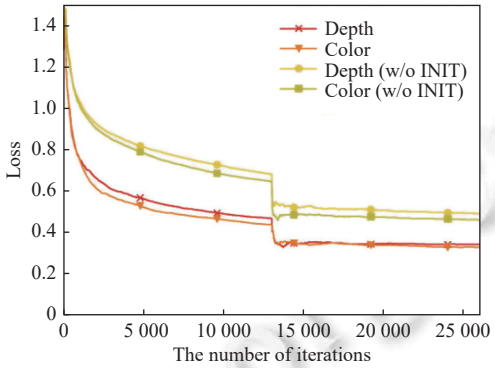


图 15 有无语义分割初始化的收敛速度对比

表 8 NYUCAD 数据集上不同的结构先验表示的实验结果

输入	形状	语义标签	草图	SC-IoU (%)	SSC-mIoU (%)
TSDF+RGB	√	—	—	83.1	52.5
TSDF+RGB	—	√	—	82.6	53.2
TSDF+RGB	—	—	√	**84.2**	**55.2**

6 面临挑战与发展前景

研究语义场景补全问题, 不仅具有理论价值, 还有很高的应用价值. 对该领域的探索能够帮助机器人技术更好地拓展到现实应用中. 比如该技术可应用于室内机器人, 并为其室内导航提供高价值的语义地图; 应用于三维测量, 为室内装修设计和家具设计提供高效的测量方式; 应用于三维游戏, 尤其是 AR 类游戏, 为三维空间中的物体检测、替换、新增等提供依据. 总之, 该方法的研究方兴未艾, 其应用价值有着很大的潜力空间.

语义场景补全任务虽然在近年来得到了一定的关注, 并且取得了不错的成绩, 但该领域仍然有一些研究问题亟待解决.

(1) 对于计算资源的庞大需求: 计算资源主要包括计算量和显存消耗. 尽管有些方法, 诸如 DDRNet, 显著地减少了网络参数量, 从而减少了模型的计算量, 并提升了运算速度. 然而三维数据的运算量依旧非常庞大. 而且, 对于显存的消耗更是巨大. 这不仅限制了语义场景补全在实时应用中的部署, 也限制了场景范围的大小和分辨率, 使得网络预测的精细程度难以提升.

(2) 真实场景下的高质量标注数据不足: 由于三维数据的标注非常困难, 使得人工标注的结果与真实的场景存在较多的不对齐, 这对网络的训练过程造成困扰. 采用人工合成的数据, 能够提供较为精确的三维真值数据. 例如 SUNCG 数据集提供大量的仿真数据, 然而, 该合成数据的彩色图像与真实场景存在较大的差异, 依旧缺乏真实场景下的彩色图像作为训练数据.

(3) 采用三维栅格的空间表示方式通常需要较高的存储空间和内存占用, 需要考虑更高效的表示方式来进行语义场景补全任务, 诸如采用稀疏表示等方式和八叉树等数据结构来存储, 或者采用点云、mesh 网格来进行三维空间的表示和处理.

(4) 三维场景的补全和实例分割是值得探索的后续研究方向之一. 目前仅进行语义分割, 然而许多交互任务需要区分每一个物体实例, 例如餐饮服务机器人端盘子, 需要明确操作哪一个盘子. 实例分割将会拓宽该三维场景补全和分割的应用领域, 使其具有更高的应用价值.

7 总 结

本文总结了近年来出现的基于 RGB-D 图像的语义场景补全方法, 主要分为传统方法、基于单一深度图像的深度学习算法以及基于彩色图像联合深度图像的深度学习算法, 涵盖了该领域最先进和最新的工作, 并为读者提供了有关此类任务的必要背景知识. 我们收集了和该领域有关的 3 个数据集和 23 种方法, 并以表格形式提供了数据集和方法的实验比较结果. 最后, 我们就未来的研究方向和该领域未解决的问题提供了有用的见解.

References:

[1] Song SR, Yu F, Zeng A, Chang AX, Savva M, Funkhouser T. Semantic scene completion from a single depth image. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 190–198. [doi: 10.1109/CVPR.2017.28]

[2] Gupta S, Arbeláez P, Malik J. Perceptual organization and recognition of indoor scenes from RGB-D images. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 564–571. [doi: 10.1109/CVPR.2013.79]

[3] Ren XF, Bo LF, Fox D. RGB-(D) scene labeling: Features and algorithms. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 2759–2766. [doi: 10.1109/CVPR.2012.6247999]

[4] 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]

[5] Lai K, Bo LF, Fox D. Unsupervised feature learning for 3D scene labeling. In: Proc. of the 2014 IEEE Int'l Conf. on Robotics and Automation (ICRA). Hong Kong: IEEE, 2014. 3050–3057. [doi: 10.1109/ICRA.2014.6907298]

[6] Varley J, DeChant C, Richardson A, Ruales J, Allen P. Shape completion enabled robotic grasping. In: Proc. of the IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017. 2442–2447. [doi: 10.1109/IROS.2017.8206060]

[7] Rock J, Gupta T, Thorsen J, Gwak J, Shin D, Hoiem D. Completing 3D object shape from one depth image. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 2484–2493. [doi: 10.1109/CVPR.2015.7298863]

[8] Wu ZR, Song SR, Khosla A, Yu F, Zhang LG, Tang XO, Xiao JX. 3D ShapeNets: A deep representation for volumetric shapes. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 1912–1920. [doi: 10.1109/CVPR.2015.7298801]

[9] Nguyen DT, Hua BS, Tran MK, Pham QH, Yeung SK. A field model for repairing 3D shapes. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vega: IEEE, 2016. 5676–5684. [doi: 10.1109/CVPR.2016.612]

[10] Monszpart A, Mellado N, Brostow GJ, Mitra NJ. RAPter: Rebuilding man-made scenes with regular arrangements of planes. ACM Trans. on Graphics, 2015, 34(4): 103. [doi: 10.1145/2766995]

[11] Kim YM, Mitra NJ, Yan DM, Guibas L. Acquiring 3D indoor environments with variability and repetition. ACM Trans. on Graphics, 2012, 31(6): 138. [doi: 10.1145/2366145.2366157]

[12] Mattausch O, Panozzo D, Mura C, Sorkine-Hornung O, Pajarola R. Object detection and classification from large-scale cluttered indoor scans. Computer Graphics Forum, 2014, 33(2): 11–21. [doi: 10.1111/cgf.12286]

[13] Firman M, Aodha O M, Julier S, Brostow GJ. Structured prediction of unobserved voxels from a single depth image. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 5431–5440. [doi: 10.1109/CVPR.2016.586]

[14] Zheng B, Zhao YB, Yu JC, Ikeuchi K, Zhu SC. Beyond point clouds: Scene understanding by reasoning geometry and physics. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 3127–3134. [doi: 10.1109/CVPR.2013.402]

[15] Lin DH, Fidler S, Urtasun R. Holistic scene understanding for 3D object detection with RGBD cameras. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 1417–1424. [doi: 10.1109/ICCV.2013.179]

[16] Geiger A, Wang CH. Joint 3D object and layout inference from a single RGB-D image. In: Proc. of the 37th German Conf. on Pattern Recognition. Aachen: Springer, 2015. 183–195. [doi: 10.1007/978-3-319-24947-6_15]

[17] Guo YX, Tong X. View-volume network for semantic scene completion from a single depth image. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI, 2018. 726–732. [doi: 10.24963/ijcai.2018/101]

[18] Zhang L, Wang L, Zhang XD, Shen PY, Bennamoun M, Zhu GM, Shah SAA, Song J. Semantic scene completion with dense CRF from a single depth image. Neurocomputing, 2018, 318: 182–195. [doi: 10.1016/j.neucom.2018.08.052]

[19] Wang YD, Tan DJ, Navab N, Tombari F. Adversarial semantic scene completion from a single depth image. In: Proc. of the 2018 Int'l Conf. on 3D Vision (3DV). Verona: IEEE, 2018. 426–434. [doi: 10.1109/3DV.2018.00056]

[20] Zhang JH, Zhao H, Yao AB, Chen YR, Zhang L, Liao HE. Efficient semantic scene completion network with spatial group convolution. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 749–765. [doi: 10.1007/978-3-030-01258-8_45]

[21] Liu SC, Hu Y, Zeng YM, Tang QK, Jin BB, Han YH, Li XW. See and think: Disentangling semantic scene completion. In: Proc. of the 2018 Int'l Conf. on Neural Information Processing Systems. Montréal: NeurIPS, 2018. 261–272.

[22] Garbade M, Chen YT, Sawatzky J, Gall J. Two stream 3D semantic scene completion. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE, 2019. 416–425. [doi: 10.1109/CVPRW.2019.00055]

[23] Li J, Liu Y, Gong D, Shi QF, Yuan X, Zhao CX, Reid I. RGBD based dimensional decomposition residual network for 3D semantic scene completion. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 7685–7694. [doi: 10.1109/CVPR.2019.00788]

[24] Chen YT, Garbade M, Gall J. 3D semantic scene completion from a single depth image using adversarial training. In: Proc. of the 2019 IEEE Int'l Conf. on Image Processing (ICIP). Taipei: IEEE, 2019. 1835–1839. [doi: 10.1109/ICIP.2019.8803174]

[25] Zhang PP, Liu W, Lei YJ, Lu HC, Yang XY. Cascaded context pyramid for full-resolution 3D semantic scene completion. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 7800–7809. [doi: 10.1109/ICCV.2019.00789]

[26] Wang YD, Tan DJ, Navab N, Tombari F. ForkNet: Multi-branch volumetric semantic completion from a single depth image. In: Proc. of the 2019 IEEE Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 8607–8616. [doi: 10.1109/ICCV.2019.00870]

[27] Li J, Liu Y, Yuan X, Zhao CX, Siegwart R, Reid I, Cadena C. Depth based semantic scene completion with position importance aware loss. IEEE Robotics and Automation Letters, 2020, 5(1): 219–226. [doi: 10.1109/LRA.2019.2953639]

[28] Dourado A, Kim H, De Campos T, Hilton A. Semantic scene completion from a single 360-degree image and depth map. In: Proc. of the 2020 Int'l Conf. on Computer Vision Theory and Applications. Valetta: VISAPP, 2020. 36–46.

[29] Li SQ, Zou CQ, Li YP, Zhao XB, Gao Y. Attention-based multi-modal fusion network for semantic scene completion. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 11402–11409. [doi: 10.1609/aaai.v34i07.6803]

[30] Chen XK, Lin KY, Qian C, Zeng G, Li HS. 3D sketch-aware semantic scene completion via semi-supervised structure prior. In: Proc. of the 2020 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 4192–4201. [doi: 10.1109/CVPR42600.2020.00425]

[31] Li J, Han K, Wang P, Liu Y, Yuan X. Anisotropic convolutional networks for 3D semantic scene completion. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 3348–3356. [doi: 10.1109/CVPR42600.2020.00341]

[32] Chen XK, Xing YJ, Zeng G. Real-time semantic scene completion via feature aggregation and conditioned prediction. In: Proc. of the 2020 IEEE Int'l Conf. on Image Processing (ICIP). Abu Dhabi: IEEE, 2020. 2830–2834. [doi: 10.1109/ICIP40778.2020.9191318]

[33] Dourado A, De Campos TE, Kim H, Hilton A. EdgeNet: Semantic scene completion from a single RGB-D image. In: Proc. of the 25th Int'l Conf. on Pattern Recognition (ICPR). Milan: IEEE, 2021. 503–510. [doi: 10.1109/ICPR48806.2021.9413252]

[34] Cai YJ, Chen XS, Zhang C, Lin KY, Wang XG, Li HS. Semantic scene completion via integrating instances and scene in-the-loop. In: Proc. of the 2021 IEEE Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 324–333.

[35] Jellinek J. Energy landscapes: With applications to clusters, biomolecules and glasses. Physics Today, 2005, 58(6): 63–64. [doi: 10.1063/1.1996481]

[36] Barbu A, Zhu SC. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1239–1253.

[37] Kim J, Grauman K. Shape sharing for object segmentation. In: Proc. of the 12th European Conf. on Computer Vision. Florence: Springer, 2012. 444–458. [doi: 10.1007/978-3-642-33786-4_33]

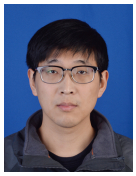
[38] Carreira J, Sminchisescu C. CPMC: Automatic object segmentation using constrained parametric Min-Cuts. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1312–1328. [doi: 10.1109/TPAMI.2011.231]

[39] Jia ZY, Gallagher A, Saxena A, Chen T. 3D-based reasoning with blocks, support, and stability. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 1–8. [doi: 10.1109/CVPR.2013.8]

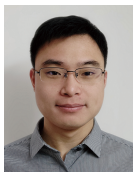
[40] Jiang H, Xiao JX. A linear approach to matching cuboids in RGBD images. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 2171–2178. [doi: 10.1109/CVPR.2013.282]

[41] Li J, Wang P, Han K, Liu Y. Anisotropic convolutional neural networks for RGB-D based semantic scene completion. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021. [doi: 10.1109/TPAMI.2021.3081499]

张康(1993－), 男, 博士生, 主要研究领域为深度学习, 图像处理, 场景补全.



安泊舟(1994－), 男, 博士生, 主要研究领域为深度学习, 语义场景补全.



李捷(1988－), 男, 博士, 主要研究领域为计算机视觉, 机器人视觉中的场景理解, 语义分割, 关键点检测, 点云.

袁夏(1981－), 男, 博士, 副教授, 主要研究领域为智能机器人环境理解与自主导航, 视觉显著性分析, 场景语义分割.

赵春霞(1964－), 女, 博士, 教授, 博士生导师, 主要研究领域为模式识别与计算机视觉, 人工智能, 移动机器人.