基于深度图及分离池化技术的场景复原及语义分类网络

林金花[1,2] 姚 禹[1] 王 莹[1]

摘 要 在机器视觉感知系统中, 从不完整的被遮挡的目标对象中鲁棒重建三维场景及其语义信息至关重要. 目前常用方法一般将这两个功能分开处理, 本文将二者结合, 提出了一种基于深度图及分离池化技术的场景复原及语义分类网络, 依据深度图中的 RGB-D 信息, 完成对三维目标场景的重建与分类. 首先, 构建了一种 CPU 端到 GPU 端的深度卷积神经网络模型, 将从传感器采样的深度图像作为输入, 深度学习摄像机投影区域内的上下文目标场景信息, 网络的输出为使用改进的截断式带符号距离函数 (Truncated signed distance function, TSDF) 编码后的体素级语义标注. 然后, 使用分离池化技术改进卷积神经网络的池化层粒度结构, 设计带细粒度池化的语义分类损失函数, 用于回馈网络的语义分类重定位. 最后, 为增强卷积神经网络的深度学习能力, 构建了一种带有语义标注的三维目标场景数据集, 以此加强本文所提网络的深度学习鲁棒性. 实验结果表明, 与目前较先进的网络模型对比, 本文网络的重建规模扩大了 2.1 %, 所提深度卷积网络对缺失场景的复原效果较好, 同时保证了语义分类的精准度.

关键词 机器视觉感知系统, 池化技术, 深度图, 深度学习, 卷积神经网络

引用格式 林金花, 姚禹, 王莹. 基于深度图及分离池化技术的场景复原及语义分类网络. 自动化学报, 2019, **45**(11): 2178−2186

DOI 10.16383/j.aas.2018.c170439

Scene Restoration and Semantic Classification Network Using Depth Map and Discrete Pooling Technology

LIN Jin-Hua[1,2] YAO Yu[1] WANG Ying[1]

Abstract In the machine vision perception system, it is very important to robustly reconstruct the 3D scene and recognize target semantics. At present, commonly used methods generally deal with these two functions separately. In this paper, we propose a scene restoration and semantic classification network using the depth map. Based on the RGB-D information in the depth map, reconstruction of a 3D target scene is completed along with classification. Firstly, a deep convolutional neural network model from the CPU end to the GPU end is constructed, which takes depth samples as input from sensor and deeply learns contextual target scene information in the camera projection area. The output of the network comes from the improved truncated signed distance function (TSDF) coding voxel-level semantic annotation. Secondly, in order to enhance the deep learning ability of the convolutional neural network, a three-dimensional target scene dataset with semantic annotation is constructed to enhance the robustness of the proposed network. Experimental results show that compared with the current advanced network model, the reconstruction scale of this network model expands by 2.1 %. The proposed convolutional network has good reconstruction effect on the missing scene and the accuracy of semantic classification is also guaranteed.

Key words Machine vision perception system, pooling technology, depth map, deep learning, convolutional neural network

Citation Lin Jin-Hua, Yao Yu, Wang Ying. Scene restoration and semantic classification network using depth map and discrete pooling technology. *Acta Automatica Sinica*, 2019, **45**(11): 2178−2186

在客观物质世界中, 目标实体的客观存在形式通常取决于其所占用的三维空间位置. 机器系统识别客观实体的语义及其拓扑存在性需要精准的神经网络模型. 在机器视觉感知系统中, 鲁棒重建三维场景以及识别目标语义至关重要, 能够实现机器系统对目标区域信息的有效捕捉与精准定义, 有效地识

收稿日期 2017-08-01 录用日期 2017-12-14
Manuscript received August 1, 2017; accepted December 14, 2017
国家高技术研究发展计划 (863 计划) (2014AA7031010B), 国家自然科学基金 (51705032), 吉林省教育厅 “十三五” 科学技术研究项目 (20 16345) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2014AA7031010B), National Natural Science Foundation of China (51705032), and Jilin Province “Thirteenth Five” Science and Technology Research Project (2016345)
本文责任编委 黄庆明
Recommended by Associate Editor HUANG Qing-Ming
1. 长春工业大学应用技术学院 长春 130012 2. 中国科学院长春光学精密机械与物理研究所 长春 130031
1. School of Application Technology, Changchun University of Technology, Changchun 130012 2. Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130031

别出目标场景形状及其语义信息, 语义识别与场景重建相互作用以确保机器视觉系统能够鲁棒识别并复原目标场景. 传统方法一般分别完成这两项工作, 例如, 二维识别方法一般仅对二维图像进行分类处理, 不会重建目标拓扑结构[1−2]; 相反, 几何重建方法仅复原三维结构信息, 而不识别目标语义. 针对这一问题, 本文构建了一种场景重建与语义识别相互结合的深度卷积神经网络模型, 同时实现了对三维场景的重建与语义分类功能.

为了高效训练本文的模型, 使用监督式学习方法完成卷积神经网络的训练过程, 进而实现场景重建与语义识别功能. 本文方法对深度数据进行重新表示, 使用截断式带符号距离函数 (Truncated signed distance function, TSDF) 编码方式对目标场景进行三维体素重定义, 每个体素包含: 被占用体素与空闲体素两种含义. 如何从不完整的目标场景中识别其语义以及不可见区域的语义标注问题是本文需要解决的关键问题.

针对上述问题, 本文构造了一种上下文区域拓展网络, 增加了接收区域场景的体素信息, 使得目标语义识别面更广. 另一方面, 本文构建了一种有效的用于深度学习的数据集, 并对其完成了体素标注.

1 相关工作

在机器视觉系统中, 鲁棒完成三维场景的语义分割任务至关重要, 常用任务包括机器人路径规划、人员协调辅助以及智能监控等. 近年来, 为了满足视觉系统需求, 实现对目标场景的语义分割任务, 深度学习神经网络得到广泛应用, 深度神经网络通过学习大规模场景数据, 生成训练标签, 进而实现目标场景理解任务. 然而, 对于大多数的视觉处理任务, 真实场景数据是有限的, 并且受深度感知技术和语义分类方法的限制, 使得构建高效的深度学习网络并不容易.

深度神经网络被广泛用于解决对象分类和目标检测问题[3−4]. 然而受数据规模、存储介质和计算能力的限制, 深度神经网络的复杂程度也随之提高, 限制了深度神经网络的适用范围. 这种限制主要出于两个方面: 1) 随着模型尺度的增大, 网络的复杂度也随之增加, 例如 Googlenet 数据集的 50 MB 模型, Resnet-101 的 200 MB 模型, Alexnet 的 250 MB 和 VGG-net 的 500 MB 模型; 2) 复杂神经网络通常需要超高性能的处理器的支持, 即高配置的 GPU 高速并行处理单元的支持, 这使得研究人员致力于模型的压缩, 以减小神经网络的内存和处理单元占用率[5−6]. 例如, Ren 等[7] 对遮挡目标场景鲁棒地完成重建过程, 将大权重矩阵分解为几个可分离小矩阵来减少冗余, 重建效果较好, 但无法实现语义识别功能. 对于神经网络的完全连接层, 这种方法已被证明非常有效. 科研工作者给出了多种基于连接限幅的语义重建方法, 删除了预训练和再训练模型的冗余连接. 这些方法将模型参数的数量减少了一个数量级, 而不会对分类精度造成重大影响, 但三维重建精度会随着降低[8−11]. 另一种语义重建策略是限制模型本身的架构. 例如, 去除完全连接的层, 使用小尺寸的卷积滤波器等, 目前较先进的深层网络, 如 Nin、Googlenet 和 Resnet 都采用这种架构. 然而这种方法对重建场景的几何拓扑细节表示不佳影响了重建分辨率[12−13]. Zheng 等[14] 使用固定点表示来量化预训练神经网络的权重, 以加快网络在 CPU 上的运行时间, 同时使用空间预测方法来推断遮挡场景信息. Kim 等[15] 提出了替代量化方法来减小模型尺寸, 在保证最小精度损失的情况下, 使用 k 级均值矢量量化实现了 4 ∼ 8 倍的重建精度, 然而引起网络训练时间的增加. Hane 等[16] 和 Blaha 等[17] 使用绑定更新优化策略来保证重建视觉的多样性, 以此加强网络的重建精度.

针对上述问题, 本文给出了一种适用于大尺度场景重建与语义识别的深度卷积神经网络模型, 将目标几何信息与目标上下文语义信息相结合, 进而完成对目标场景的鲁棒重建与识别. 另外, 本文建立了一种用于三维场景学习的数据集, 可用于对 RGB 图像的语义分割过程[18−21].

2 语义场景复原网络

本文的深度卷积神经网络由多个层次的处理单元组成, 关键核心是完成摄像机视锥体划分范围里的空间体素分配到一系列语义类别标注, 假设 $C = \{c_0, \cdots, c_{N+1}\}$, 其中, N 表示目标场景包含的类别总数, c_0 代表未被占用的体素. 每个神经单元的激励函数如下: $z = g(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$, 其中, $\boldsymbol{w} \in \mathbf{R}^{c\times w\times h}$ 为权重向量, $\boldsymbol{x} \in \mathbf{R}^{c\times w\times h}$ 为输入向量, $g(\cdot)$ 为非线性函数. 本文卷积神经网络实现了由这些单元构成的多个层, 并用张量 $W \in \mathbf{R}^{c\times w\times h}$ 来表示权重. c, w 和 h 分别用来定义滤波器通道的数量, 宽度和高度. 由于这种基本运算在整个网络中重复, 且神经网络通常由繁多的处理单位组成, 式 (1) 的表示方式是决定整个网络模型复杂程度的主要因素. 网络的复杂程度主要与两个因素有关: 1) 存储权重 $\boldsymbol{w}$ 需要巨大的内存开销; 2) 大量的点积运算 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ 需要高成本的计算开销. 当权重和点积运算为浮点值时, 上述两个方面的开销会剧增, 导致实际应用能力差[22]. 因此本文所提的低精度卷积神经网络更适用于解决实际三维重建与语义分类问题. 本文网络的场景重建与语义识别过程如图 1 所示. 下面分节阐述本文网络模型的构造与重建过程.

图 1 本文深度卷积神经网络的场景重建与语义分类过程

Fig. 1 3D reconstruction and semantic classification of our depth convolutional neural network

2.1 体素数据编码及分离池化方法

首先, 对三维场景的语义分类原理进行分析, 构建基于改进的 TSDF 编码以及细粒度池化特性的深度卷积神经网络模型; 其次, 提出估计算法对三维语义感知特性参数进行估计, 解决 TSDF 编码下具有细粒度池化层的深度卷积神经网络的模型优化问题; 最后, 建立考虑改进的 TSDF 编码下三维语义场景的语义分类性能评价体系, 预测网络对三维场景的语义分类性能, 改善机器系统对三维场景的语义感知性能, 为具有三维语义感知能力的机器视觉系统在军用和民用上的应用提供理论依据.

本文对 TSDF 进行了改进, 使之适应于场景重建与语义分类的混合卷积神经网络模型. 一般情况下, 深度卷神经网络模型使用距离相机位置最近投影直线的方式来获取场景关键点. 然而, 该方法在节省重建投影视觉的同时, 却以关键点捕获精度为代价, 影响了对三维场景的最终重建精度以分类性能. 为提高重建精度及语义分类性能, 本来采用了一种随机选取池化层内部表面点的方式来提取关键点, 改善了 TSDF 距离的计算时间, 同时保证了重建与分类精度. 分离池化后的特征区域本身具有细粒度空间几何拓扑结构的特性, 当随机采用发生时, 平均池化粒度值基本保持不变, 因此确保了随机采用的平均精准度, 以此构建的 TSDF 的精度也随之增加. 本文采用池化技术的体素编码方式如图 2 所示.

2.2 复原网络结构

三维场景语义分类问题是机器视觉领域的热点

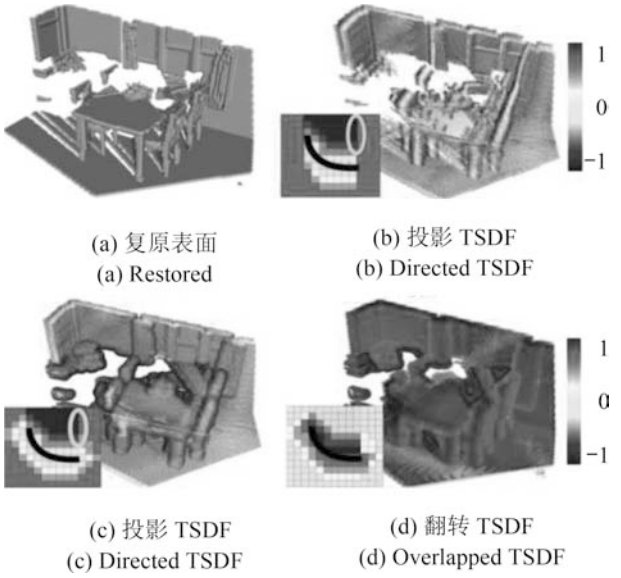


图 2 常用的 TSDF 编码可视化结果

Fig. 2 Visualization of several encoding TSDF

研究问题. 本文考虑结合 TSDF 编码与分类池化技术的三维场景重建与语义分类网络模型如图 3 所示. 下面分五个方面阐述本文深度卷积神经网络的场景复原与语义分类过程.

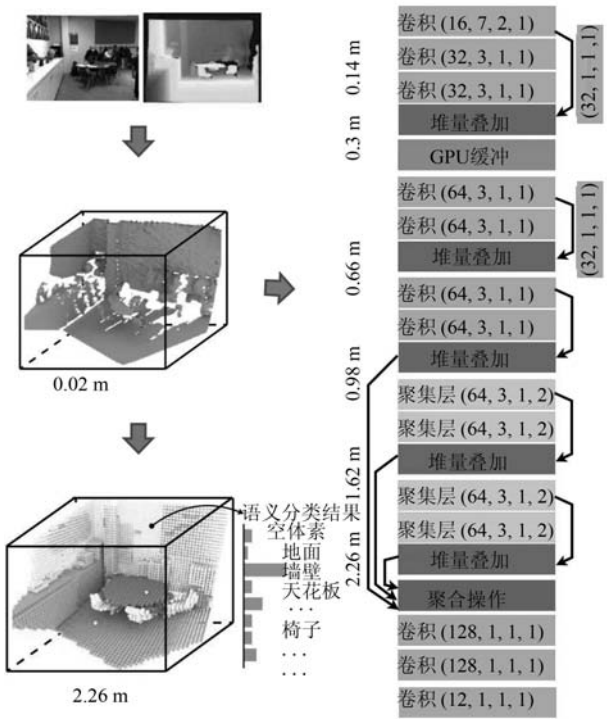


图 3 本文所提深度卷积神经网络模型

Fig. 3 Our depth convolutional neural network

1) 本文构建了一种以 RGB-D 深度图作为输入的深度学习网络框架. 一个点云由一组三维点数据构成, 即 $\{P_i \,|\, i = 1, \cdots, n\}$, 每个三维点 P_i 由五维

向量表示. 对于对象分类任务, 输入点云直接从目标形状采样, 或者从一个场景点云预分割得到. 对于语义分割, 输入可以是用于部分区域分割的单个对象, 或者用于对象区域分割的三维场景子体积. 本文网络将为 n 个点和 m 个语义子类别中的每一个输出 $n \times m$ 个分数. 图 4 给出了本文语义分类网络架构. $T1$ 和 $T2$ 是输入点和特征的对称转换网络. FC 是完全连接的层在每个点上操作. MLP 是每个点上的多层感知器. ***vec*** 是大小为 16 的向量, 指示输入形状的类别. 本文网络能够预测体素数量, 如图 4 中的左下角曲线图所示, 这表明本文复原网络能够从本地邻域获取信息, 对区域分割具有鲁棒性.

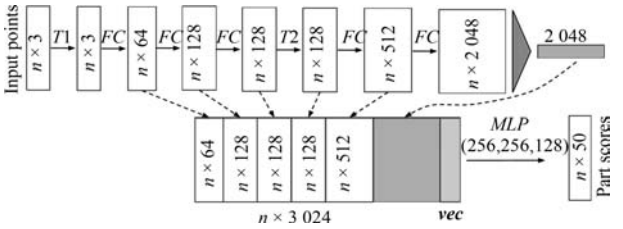


图 4 本文语义分类的卷积流程

Fig. 4 Convolutional streamline of our semantic classification

2) 本文语义复原网络从训练 LS_3DDS 合成数据集中, 直接学习接收域信息来获取条件概率矩阵, 即在三维场景语义分类中, 条件概率 $p(A_i\,|C_n)$ 表示在语义类别 C_n 中出现的语义对象 A_i 的比率来计算概率分布

$$\vartheta_{ni}^{A} = p(A_i|C_n) = \frac{\sum\limits_{I\in C_n} a_i(I)}{\sum C_n} \tag{1}$$

其中, $\sum C_n$ 表示 LS_3DDS 数据集中属于类别 C_n 的场景个数, 且 $\sum_i p(A_i\,|\,C_n) = 1$. 本文的三维场景语义类别个数 N, 对象个数为 M, 语义对象条件概率矩阵为 $N \times M$ 阶矩阵, 即 $\Theta = [\vartheta_{ni}^{A}]_{N\times M}$. 这里通过计数随机事件的出现频率来估计概率分布, 需要大量的真实观测数据. 使用本文构建的 LS_3DDS 数据集训练语义神经网络模型, 由于合成数据集规模较大且手动标记标签精准, 使得计算得出的条件概率较准确, 保证了本文语义场景复原网络的精准度, 如图 5 所示, 接收区域的增大提高了本文网络的上下文语义识别精准度.

3) 本文神经网络的池化器采用分段常值函数, 定义为

$$Q(x) = q_i, \quad x \in (t_i, t_{i+1}] \tag{2}$$

该池化器将量化间隔 $(t_i, t_{i+1}]$ 内的所有 x, 并将其映射为量化级别 $q_i \in \mathbf{R}$, 其中, $i = 1, \cdots, m$, 且 $t_1 = -\infty$, $t_{m+1} = +\infty$. 这将泛化符号函数, 将其看作是 1 位池化器. 一个均匀池化器需要满足以下条件:

$$q_{i+1} - q_i = \Delta, \quad \forall i \tag{3}$$

其中, Δ 是恒定量化步长. 受精度降低的约束, 量化级 q_i 作为激励 x 的重构值. 因为对于任意 x, 该池化器足以存储式 (2) 的量化索引 i 以恢复量化级别 q_i, 所以非均匀池化器需要 $\log_2 m$ 比特的存储空间来存放激励 x. 然而, 在算术运算过程中, 通常需要超过 $\log_2 m$ 比特来表示 x, 并使用 q_i 代替索引 i. 对于均匀池化器, Δ 是通用缩放因子, 通常以 $\log_2 m$ 比特来存储激励 x 而不存索引. 本文在卷积运算中也同样采用这种存储策略.

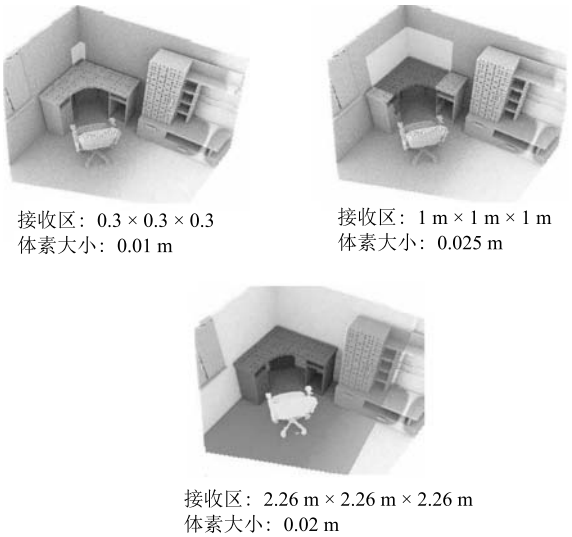


图 5 本文摄像头接收范围直接影响网络性能

Fig. 5 Our camera receiving range directly affects performance of network

4) 设计最优池化器以保证三维重建精度与语义分类准确率, 需要将池化器定义在均值误差范围内, 即

$$Q^*(x) = \arg\min_{Q} \mathrm{E}_x\left[(Q(x)-x)^2\right] = \arg\min_{Q} \int p(x)(Q(x)-x)^2 \mathrm{d}x \tag{4}$$

其中, $p(x)$ 是 x 的概率密度函数. 因此, 式 (2) 中点积的最优池化器取决于它们的统计值. 虽然式 (4) 的最优解 $Q^*(x)$ 通常是不均匀的, 但通过将式 (3) 的约束代入式 (4), 可以得到均匀解 $Q^*(x)$. 给定点积样本, 式 (4) 的最优解可以通过劳埃德算法获得. 这是一个迭代算法. 由于每个网络单元必须设计不同的池化器, 并且该池化器随反向传播迭代而改变, 因此该过程的直接计算实现是较繁琐且有难度的.

5) 本文使用半波高斯池化器来实现反向近似操作, 通过利用深层网络激励的统计结构来的克服池化器随反向传播迭代而改变的问题. 文献 [23−24] 证明了点积近似具有接近高斯分布的对称、非稀疏分布特性. 考虑到 ReLU 是半波整流器, 本文使用半波高斯池化器 (Half wave Gauss pool, HWGP) 来实现反向近似操作, 定义如下:

$$Q(x)=\begin{cases} q_i, & x\in(t_i,t_{i+1}] \\ 0, & x\le 0 \end{cases} \tag{5}$$

其中, $q_i\in\mathbf{R}^+$, $i=1,\cdots,m$, $t_i\in\mathbf{R}^+$, $i=1,\cdots,m+1$, $t_1=0$, $t_{m+1}=\infty$; q_i 和 t_i 是高斯分布的最优量化参数. SGNN 保证了这些参数仅取决于点积分布的均值和方差. 然而, 因为这些参数在不同的单元之间变化, 所以无法消除网络上劳埃德算法的重复使用.

这个问题可以通过批量归一化方法来缓解, 这迫使网络的每个层的响应都具有零均值和单位方差. 本文将这种归一化操作应用于点积运算, 结果如图 6 所示. 尽管点积分布不完全符合高斯分布, 它们之间存在微小差异, 但二者都接近高斯分布, 且平均值和单位方差为零. 因此, 最佳量化参数 q_i^* 和 t_i^* 在神经网络的单元、层和反向传播迭代过程中大致相同. 因此, 劳埃德算法在整个网络上仅使用一次即可. 实际上, 由于所有分布都近似于零均值和单位方差的高斯分布, 因此可以从该分布的样本中设计池化器. 本文从零均值和单位方差的标准高斯分布中抽取了 10^6 个样本, 并通过劳埃德算法获得了最优量化参数. 在点积批量归一化之后, 再将所得到的参数 q_i^* 和 t_i^* 用于参数化在所有层中使用的 SGNN.

3 实验结果与分析

为测试本文卷积神经网络的重建精度与语义分类性能, 本节采用摄像机捕获的三维场景数据以及合成数据对网络进行训练与测试.

在使用真实场景数据进行训练时, 本文使用 NYU 数据集训练深度卷积神经网络模型, 该数据集由 1 449 个 RGB-D 深度图. 本文针对由 Guo 等[25] 提出的带有几何标注的三维体积模型, 捕获了大量的三维真实场景数据信息. 另外, 同时采用了 Sun 等[26] 的采样策略捕获了多种三维场景对象数据. 通常情况下, 当语义标注信息与实际网络拓扑信息不完全对应时, 数据集中的深度信息与几何信息也会出现不匹配的现象. 针对这一问题, Silberman 等[3] 等采用绘制 RGB-D 图的方式对目标三维场景的三维物理位置信息进行标记. 然而在标记的过程中不可避免的影响原有三维拓扑结构, 使得三维重建场景的本地特性未能较好地保留. 为此, 本文结合了上述几种重建数据集的构造方式, 对本文神经网络进行测试.

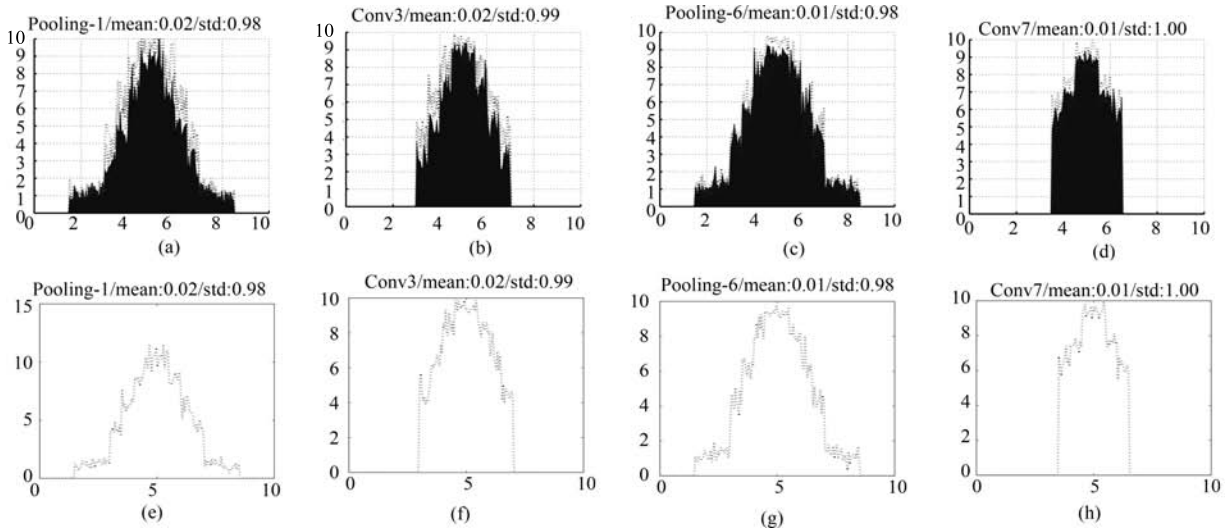


图 6 带有二进制权值和量化激励的网络层点积分布图. (a), (b), (c), (d) 分别为下采样层 1、卷积层 3、下采样层 6、卷积层 7 的点积分布图 (具有不同的均值和标准偏差); (e), (f), (g), (h) 分别为下采样层 1、卷积层 3 、下采样层 6、卷积层 7 对应的点积误差分布曲线

Fig. 6 Dot product distribution of network with binary weights and quantitative activation. (a), (b), (c) and (d) are the point product distribution maps of the pooling layer 1, the convolution layer 3, the pooling layer 6 and the convolution layer 7, respectively, they share a different mean and standard deviation; (e), (f), (g) and (h) are the dot product error distribution curves corresponding to the pooling layer 1, the convolution layer 3, the pooling layer 6 and the convolution layer 7, respectively.

本文在表 1 和表 2 中展示了对神经网络性能的定量分析, 同时在图 7 中给出了网络的定性分析结果. 在表 1 中, 将本文网络模型与 Lin 等[12] 以及 Gupta 等[8] 和 Wang 等[21] 提出的网络模型展开对比, 为方便引用, 文中下述段落将上述几种网络重命名为 L 网、GW 网. 这两种网络模型采用深度输入帧为神经网络的输入数据, 同时在目标场景的体素级网络上生成语义标注. L 网采用包围盒以及超平面近似的方式标记全部体素网格. GW 网对场景进行测试的同时搜索 RGB-D 数据信息, 进而完成对全局场景的重建测试. 以上两种网络能够在较小的训练数据集上重构精准的三维场景几何结构模型, 对应关系的匹配方法较精准. 与之不同, 本文网络采用单一深度图作为输入, 同时结合分离池化技术对特征采用关键点进行优化处理, 生成细节丰富的 TSDF 编码方法, 另外无需附加网络来协调测试过程, 提高了重建性能. 因此, 本文深度卷积神经网络模型能够生成更加精准的重建模型, 同时保证了语义分类精度. 本文深度卷积神经网络的三维场景复原精度值为 30.5 %, GW 网的精度百分比为 19.6 %. 由图 7 给出的重建对比图可知, 这两种网络模型同时将沙发对象语义标记为床, 然而, 本文网络模型能够准确识别目标对象语义, 并采用虚线方框来标记,

表 1 本文网络与 L、GW 网络的复原与分类性能比较 (%)

Table 1 Comparison of three networks for performance of reconstruction and semantic classification (%)

		L	GW	本文 NYU	本文 LS_3DDS	本文 NYU+LS_3DDS
复原	闭环率	59.6	66.8	57.0	55.6	**69.3**
	IoU	37.8	46.4	59.1	58.2	**58.6**
语义场景复原	天花板	0	14.2	17.1	8.8	**19.1**
	地面	15.7	65.5	**92.7**	85.8	94.6
	墙壁	16.7	17.1	28.4	15.6	**29.7**
	窗	**15.6**	8.7	0	7.4	18.8
	椅子	9.4	4.5	15.6	18.9	**19.3**
	床	27.3	46.6	37.1	37.4	**53.6**
	沙发	22.9	25.7	38.0	28.0	**47.9**
	桌子	7.2	9.3	18.0	18.7	**19.9**
	显示器	7.6	7.0	9.8	7.1	**12.9**
	家具	15.6	27.7	28.1	10.4	**30.1**
	物品	2.1	8.3	15.1	6.4	**11.6**
	平均值	18.3	26.8	32.0	27.6	**37.3**

表 2 本文网与 F 网、Z 网的重建性能对比数据 (%)

Table 2 Comparison of our network reconstruction performance with F and Z networks (%)

	训练数据集	复原准确率	闭环率	IoU 值
F 复原方法	NYU	66.5	69.7	50.8
Z 复原方法	NYU	60.1	46.7	34.6
本文复原	NYU	66.3	96.9	64.8
本文语义复原	NYU	75.0	92.3	70.3
	LS_3DDS	**75.0**	**96.0**	**73.0**

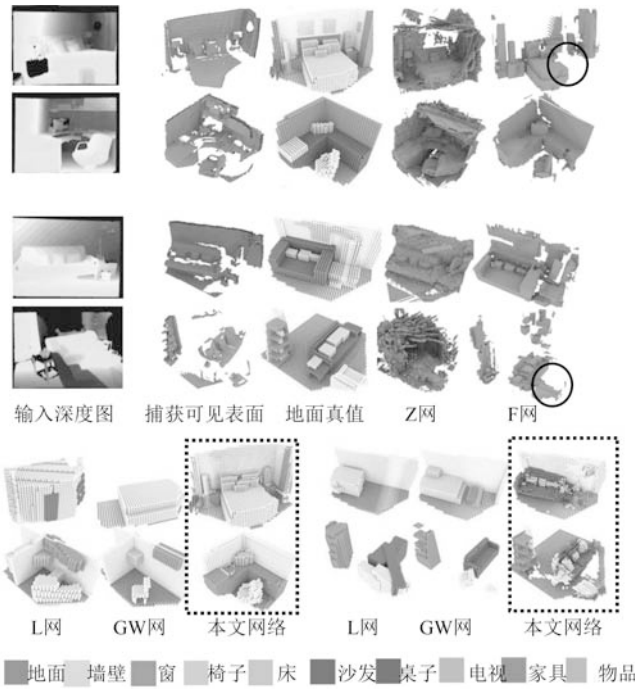


图 7 几种复原网络的可视化性能对比图

Fig. 7 Visualization performance comparison for several completion neural networks

本文方法的语义标记精准度更高, 同时, 本文网络无需对目标场景进行预处理, 三维场景复原与语义分类同时完成, 在保证重建精度的同时, 节省了对三维目标场景的重建时间已经语义分类开销.

本文对卷积神经网络进行训练, 进而检测三维体素的空间占用比率, 首先将单个体素数据进行编码, 未被占用的体素用二进制字符 “0” 来标记, 已经被占用的体素项目用字符 “1” 来标记. 表 2 给出了使用以上数据集训练完成的网络模型的性能对比数据. 使用本文网络对场景进行重建复原, 同时使用 Silberman 等[3] 和 Zheng[14] 等提出的网络对场景重建复原, 为方便引用, 文中下述段落将上述几种网络重命名为 F 网、Z 网. 这两种方法采用 RGB-D

图作为网络的输入, 实现对三维场景的复原处理, 然而二者皆不具有语义分类标注功能. 本文网络针对上述两种网络完成了整合改进, 将场景复原与语义分类融合到统一的深度卷积神经网络模型中. 本文网络首先在测试阶段, 采用 200 张输入深度图, 同时采用 NYU 体系来平均本文网络的重建与分类性能. F 网实现了对大规模场景的三维重建过程, 并且重建的精度较高, 然而, 当场景的目标语义较复杂, 遮挡现象严重时, 网络的重建精度受到限制, 三维场景的重建效果受到影响. 例如, 在图 7 中第 4 行的椅子复原失败 (如图中蓝色圆圈所示). 然而, 使用本文网络来重建目标场景时, 由于结合了上下文语义评价体系, 改善了语义重建的精准度. 从本组实验结果可以看出, 本文的将重建与语义分类相结合的方法, 在提高三维重建精度的同时, 避免了不必要的语义检测失效问题.

本文训练了一种用于三维重建与语义分类的统一架构深度卷积神经网络模型, 本文对未被遮挡的场景表面几何进行具体的语义标注, 同时采用联合策略对目标网络模型进行训练, 并对比起重建结果. 然而, 当本文网络对未被遮挡的表面进行测试是, 采用三维场景重建结合语义分类来训练三维神经卷积神经网络模型的效果由于仅使用几何表面语义训练的网络模型, 实验结果表明带有几何标注的三维场景重建精准度为 52.3 %, 然而, 联合两种网络得到的三维场景重建精准度为 55.3 %. 因此, 本文提出的重建与语义分类相互结合的网络模型, 具有互相协作相互促进的优势.

在图 8 中, 本文网络对未知区域的场景语义及几何形状进行了预测. 当桌子场景周边的目标场景未出现在摄像头捕获范围内时, 使用本文网络仍然能够较精准的预测出目标场景的上下文语义信息, 从预测结果可见, 本文网络的重建精度较好, 语义分类预测出的对象标注信息较准确. 例如, 在图 8 中出现的第 1 张深度图中, 该图中的周边对象均不可见, 然而, 即便信息被完全遮挡, 依据本文的池化技术仍然能够精准的预测出上下文语义, 扩大了语义识别的目标场景面积, 本文网络的重建性能从 39.0 % 提高到 45.3 %.

图 9 给出了不同体素编码方式对复原网络性能的影响. 无增量卷积和带增量卷积网格具有相同数量的参数, 而在带增量卷积网络结构中, 三个卷积层被增量卷积取代 (如图 3 所示), 将接收域从 1.62 m 增加到 2.26 m (如图 5 所示). 增加接收区域使网络能够获得更丰富的上下文信息, 并将网络性能从 38.0 % 提高到 44.3 %. 将带有和不带有聚合层的两种网络进行性能比较, 如图 9 所示, 结果表明带有聚合层的模型对场景复原和语义分类都产生较高的 IoU 值, 分别增涨 3.1 % 和 2.1 %.

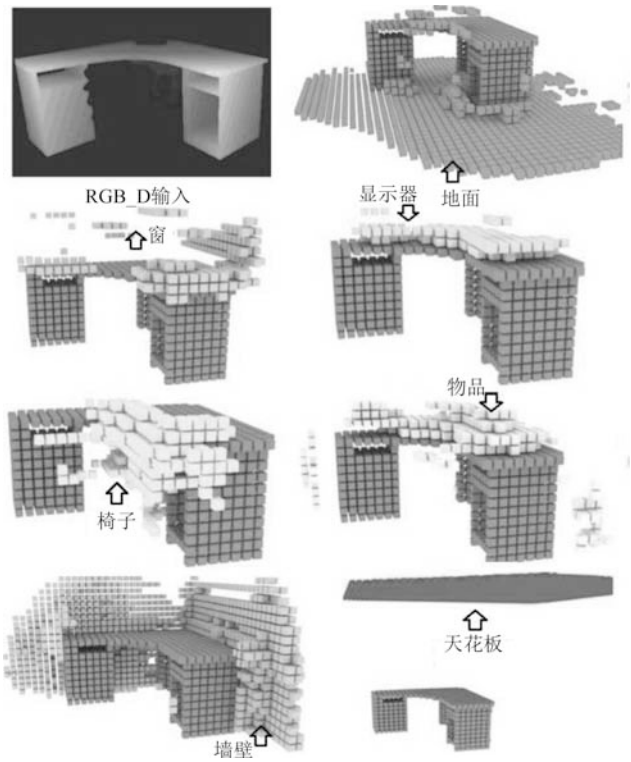


图 8 本文网络预测出的周围对象

Fig. 8 Prediction of surrounding object by our network

图 9 中给出了采用不同体素编码方式的网络性能, 即投影 TSDF, 标准 TSDF 和翻转 TSDF (改进后) 三种编码方式的比较. 实验结果显示, 使用标准 TSDF 可以消除摄像机视角的依赖性, 并使得 IoU 值提高了 2.4 %; 而使用翻转 TSDF 时, 梯度变化集中在表面上, IoU 值比标准 TSDF 提高了 10.1 %, 比投影 TSDF 提高了 12.5 %.

4 结束语

本文提出了一种基于深度图与分离池化技术的深度卷积神经网络模型, 将深度图作为输入并使用分离池化方法提取深度特征, 进而完成对三维场景的几何结构重建及语义分类任务. 同时, 构建了一种用于训练本文网络模型的三维合成数据集, 增强了神经网络的学习能力. 实验结果表明, 本文网络兼具复原与分类功能为一体, 与单一形式的网络模型相比, 本文网络的重建精度提高了 2.1 %. 本文网络采用分离池化技术及语义丰富的训练数据集, 优化了传统单一类型网络的性能, 实现了对三维场景的鲁棒重建与分类.

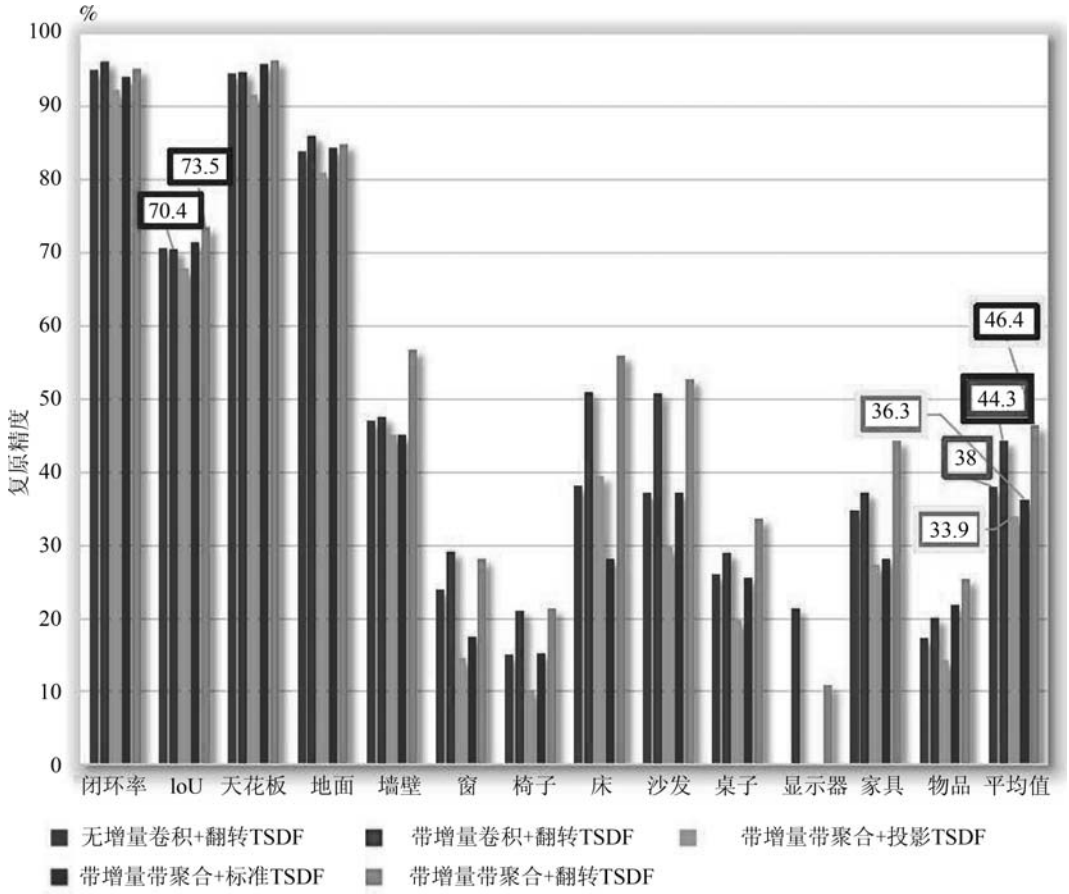


图 9 改进的 TSDF 编码对语义场景复原性能的影响

Fig. 9 Effect of improved TSDF on semantic scene completion

References

1 Gupta S, Arbeláez P, Malik J. Perceptual organization and recognition of indoor scenes from RGB-D images. In: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013. 564−571

2 Ren X F, Bo L F, Fox D. RGB-(D) scene labeling: features and algorithms. In: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2012. 2759−2766

3 Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGBD images. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 746−760

4 Lai K, Bo L F, Fox D. Unsupervised feature learning for 3D scene labeling. In: Proceedings of 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014. 3050−3057

5 Rock J, Gupta T, Thorsen J, Gwak J Y, Shin D, Hoiem D. Completing 3D object shape from one depth image. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 2484−2493.

6 Shah S A A, Bennamoun M, Boussaid F. Keypoints-based surface representation for 3D modeling and 3D object recognition. *Pattern Recognition*, 2017, **64**: 29−38

7 Ren C Y, Prisacariu V A, Kähler O, Reid I D, Murray D W. Real-time tracking of single and multiple objects from depth-colour imagery using 3D signed distance functions. *International Journal of Computer Vision*, 2017, **124**(1): 80−95

8 Gupta S, Arbeláez P, Girshick R, Malik J. Aligning 3D models to RGB-D images of cluttered scenes. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, Massachusetts, USA: IEEE, 2015. 4731−4740

9 Song S R, Xiao J X. Sliding shapes for 3D object detection in depth images. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 634−651

10 Li X, Fang M, Zhang J J, Wu J Q. Learning coupled classifiers with RGB images for RGB-D object recognition. *Pattern Recognition*, 2017, **61**: 433−446

11 Nan L L, Xie K, Sharf A. A *search-classify* approach for cluttered indoor scene understanding. *ACM Transactions on Graphics (TOG)*, 2012, **31**(6): Article No. 137

12 Lin D H, Fidler S, Urtasun R. Holistic scene understanding for 3D object detection with RGBD cameras. In: Proceedings of 2013 IEEE International Conference on Computer Vision (ICCV). Sydney, NSW, Australia: IEEE, 2013. 1417−1424

13 Ohn-Bar E, Trivedi M M. Multi-scale volumes for deep object detection and localization. *Pattern Recognition*, 2017, **61**: 557−572

14 Zheng B, Zhao Y B, Yu J C, Ikeuchi K, Zhu S C. Beyond point clouds: scene understanding by reasoning geometry and physics. In: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013. 3127−3134

15 Kim B S, Kohli P, Savarese S. 3D scene understanding by voxel-CRF. In: Proceedings of 2013 IEEE International Conference on Computer Vision (ICCV). Sydney, NSW, Australia: IEEE, 2013. 1425−1432

16 Häne C, Zach C, Cohen A, Angst R. Joint 3D scene reconstruction and class segmentation. In: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013. 97−104

17 Bláha M, Vogel C, Richard A, Wegner J D, Pock T, Schindler K. Large-scale semantic 3D reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 3176−3184

18 Handa A, Patraucean V, Badrinarayanan V, Stent S, Cipolla R. Understanding real world indoor scenes with synthetic data. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 4077−4085

19 Lv Chao-Hui, Shen Ying-Hua, Li Jing-Hua. Depth map inpainting method based on Kinect sensor. *Journal of Jilin University (Engineering and Technology Edition)*, 2016, **46**(5): 1697−1703
(吕朝辉, 沈萦华, 李精华. 基于 Kinect 的深度图像修复方法. 吉林大学学报 (工学版), 2016, **46**(5): 1697−1703)

20 Hu Chang-Sheng, Zhan Shu, Wu Cong-Zhong. Image super-resolution based on deep learning features. *Acta Automatica Sinica*, 2017, **43**(5): 814−821
(胡长胜, 詹曙, 吴从中. 基于深度特征学习的图像超分辨率重建. 自动化学报, 2017, **43**(5): 814−821)

21 Wang P S, Liu Y, Guo Y X, Sun C Y, Tong X. O-CNN: octree-based convolutional neural networks for 3D shape analysis. *ACM Transactions on Graphics (TOG)*, 2017, **36**(4): Article No. 72

22 Yücer K, Sorkine-Hornung A, Wang O, Sorkine-Hornung O. Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction. *ACM Transactions on Graphics (TOG)*, 2016, **35**(3): Article No. 22

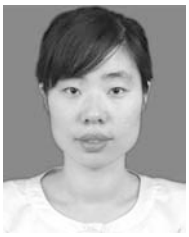
23 Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Networks*, 2000, **13**(4−5): 411−430

24 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 3431−3440

25 Guo R Q, Zou C H, Hoiem D. Predicting complete 3D models of indoor scenes. arXiv:1504.02437, 2015.

26 Sun Xu, Li Xiao-Guang, Li Jia-Feng, Zhuo Li. Review on deep learning based image super-resolution restoration algorithms. *Acta Automatica Sinica*, 2017, **43**(5): 697−709
(孙旭, 李晓光, 李嘉锋, 卓力. 基于深度学习的图像超分辨率复原研究进展. 自动化学报, 2017, 43(5): 697−709)

林金花 博士, 长春工业大学讲师. 主要研究方向为数字图像处理, 目标识别与跟踪. 本文通信作者.
E-mail: linjinhua@ccut.edu.cn
(**LIN Jin-Hua** Ph. D., lecturer at Changchun University of Technology. Her research interest covers digital image processing, target recognition, and tracking. Corresponding author of this paper.)



姚 禹 博士, 长春工业大学讲师. 主要研究方向为复杂机电系统建模、滤波与控制. E-mail: yaoyu@ccut.edu.cn
(**YAO Yu** Ph. D., lecturer at Changchun University of Technology. Her research interest covers complex electromechanical system modeling, filtering and control.)

王 莹 博士, 长春工业大学讲师. 主要研究方向为数字图像处理.
E-mail: wangying@ccut.edu.cn
(**WANG Ying** Ph. D., lecturer at Changchun University of Technology. Her main research interest is digital image processing.)