文章编号 1004-924X(2018)05-1231-11

三维语义场景复原网络

林金花[1*]，王延杰[2]

(1. 长春工业大学 应用技术学院，吉林 长春 130000；
2. 中国科学院 长春光学精密机械与物理研究所，吉林 长春 130031)

摘要：从不完整的视觉信息中推断出物体的三维几何形状是机器视觉系统应当具备的重要能力，而识别出场景中物体的语义是机器视觉系统的核心。传统方法通常将二者分离实现，本文将场景复原与目标语义紧密结合，提出了一种三维语义场景复原网络模型，仅以单一深度图作为输入，实现对三维场景的语义分类和场景复原。首先，建立一种端到端的三维卷积神经网络，网络的输入是深度图，使用三维上下文模块来对相机视锥体内的区域进行学习，进而输出带有语义标签的三维体素；其次，建立了带有密集体积标签的合成三维场景数据集，用于训练本文的深度学习网络模型；最后通过实验表明，与现有的语义分类和场景复原方法相比，语义场景的复原接收区域增加了 2.0%。结果表明：三维学习网络的复原性能良好，语义标注的准确率较高。

关 键 词：机器视觉；场景复原；深度图；语义分类；卷积神经网络

中图分类号：TP391.41 **文献标识码**：A **doi**：10.3788/OPE.20182605.1231

Three-dimentional reconstruction of semantic scene based on RGB-D map

LIN Jin-hua[1*], WANG Yan-jie[2]

(1. *School of Application Technology, Changchun University of Technology, Changchun* 130000, *China*;
2. *Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun* 130031, *China*)
* *Corresponding author, E-mail: ljh3832@163.com*

Abstract: Reconstruction of 3D object is an important part in machine vision system, and the semantic understanding of 3D object is a core function for the machine vision system. In this paper, 3D restoration was combined with the semantic understanding of 3D object, a 3D semantic scene recovery network was proposed. The semantic classification and scene restoration of 3D scene were achieved only by using a single RGB-D map as input. Firstly, an end-to-end 3D convolution neural network was established. The input of the network was a depth map. The 3D context module was used for learning the region within the camera view, then the 3D voxels with semantic labels were generated. Secondly, a synthetic data set with dense volume labels was established to train the depth learning network. Finally, the experimental results showed that the recovery performance w improved by 2.0%

收稿日期：2017-10-10；**修订日期**：2017-11-06.

基金项目：国家 863 高技术研究发展计划项目资助(No. 2014AA7031010B)；吉林省“十三五”计划科研项目资助(No. 吉教字[2016]345)

compared with the state-of-art. It can be seen that the 3D learning network plays well in 3D scene restoration, it owns high accuracy in semantic annotation of object in the scene.

Key words: machine vision; scene restoration; RGB-D map; semantic classification; convolution neural network

1 引　言

在三维世界中，物体的物理存在性是由该物体所处空间区域是否被占用来决定的。为了与客观世界进行交互，通常依赖于对场景的三维几何和语义的理解。同样，对于一个机器人，能够从不完整的视觉信息中推断出目标物体的三维形状是机器人视觉系统应具备的重要能力，可以支持机器人完成抓取、躲避障碍物等任务；能够推断出场景中物体的语义，可以支持机器人完成检索目标的任务。传统方法通常将这两项任务分离，深度分割方法仅对二维可视表面操作，而不考虑三维几何形状[1-2]；形状复原方法仅考虑从上下文中分离出单一目标，仅对三维几何进行复原操作而不考虑语义。本文建立了一种语义场景重建模型，能够从单一深度图推断出空间体积占用和目标分类情况。

考虑到目标的语义和周围场景是紧密相关的，因此对空间体素占用的预测，以及对目标语义的识别，二者同样是相互关联的。也就是说，当场景中的物体语义被准确识别，即使物体对于视觉系统不是完全可见，也可以预测出该物体的三维几何形状。反之，当已知物体的几何形状，可以预测出物体的语义类别。

本文采用监督式特征学习方式来训练深度神经网络，对三维场景进行语义分类和复原。取一个深度图作为输入，本文的深度学习网络对场景进行语义分类和复原重建，为视锥体内的体素生成各自的语义标签，每个体素被标注为占用或空闲，已被占用的体素被标注为属于 N 个对象类别中的一个。这种预测会超出单个深度图的视锥体投影面，为整个场景提供占用信息。为了实现语义场景的复原重建，本文需要解决以下两个问题：第一，当场景中的目标物体部分可见时，如何从物体的三维体积数据中有效地捕获场景的上下文信息；第二，由于 RGB-D 数据集仅对可见表面进行标注，因此需要获取带有完整表面标注的训练数据。

针对第一个问题，设计了一种三维上下文扩展模块，能够有效地扩大神经网络的可接收区域，实现对场景中上下文信息的建模。接收区域的范围大小直接影响目标物体的语义识别效果，如图 1 所示，图中椅子的下半部分区域被桌子遮挡，很难识别椅子的完整形状。因此需要考虑目标物体周围场景的上下文信息，例如椅子旁边的桌子和地面。通过获取周围场景的上下文信息可以有效推测物体的几何形状并对其进行语义标注。针对第二个问题，建立了一种带有密集体积注释的合成三维场景数据集，数据集中的三维场景由单独标注的三维对象网格组成，通过体素化对象网格，计算得出具有密集对象标签的三维场景数据。实验表明，通过本文的深度学习网络，获得了三维上下文模型，结合大规模合成训练数据，能够有效提升语义场景复原性能。

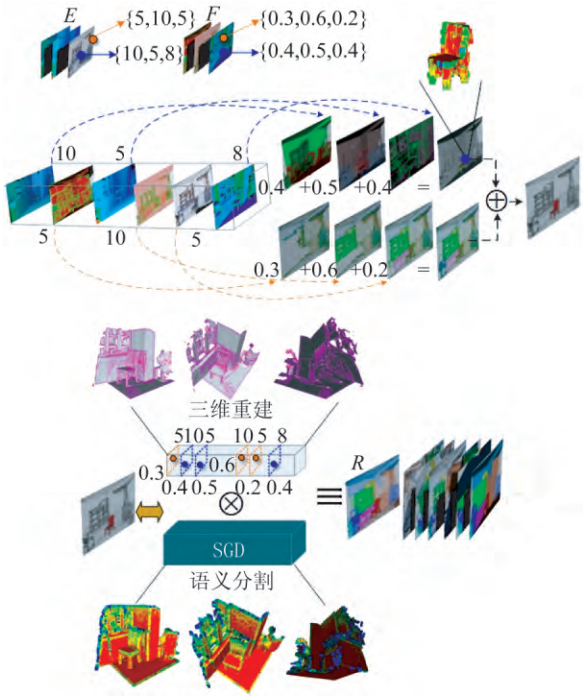

图 1　本文网络的三维语义场景复原过程

Fig. 1　3D processing of semantic scene completion of our neural network

本文的核心是构建了一种端到端的三维卷积神经网络模型，实现场景的语义标注和体积复原。本文设计了一个三维上下文扩展模块，对较大的接受区域内的场景上下文进行学习。为了提供更有效的训练数据，本文建立了一个大规模合成三维场景数据集，数据集中的场景具有空间占用情况和语义分类标注信息。

2　相关工作

下面从三个方面介绍相关领域的研究进展情况。主要包括 RGB-D 分类，三维形状复原和空间体素语义标注三个方面。

先前的工作大多集中在对 RGB-D 图像进行分类[3-4]，这些方法仅对可见像素区域进行语义标注，不考虑物体的完整几何形状，因此无法对可见区域以外的部分进行场景复原。这些方法针对单一物体进行形状复原[5]，且仅对单一物体有效，若要将这些方法用于场景的复原操作，则需要附加语义分类任务，复原的鲁棒性和实时性都无法保证。对于三维场景的复原，当不可见区域相对较小时，可以使用平面拟合或目标对称方法来填补孔洞[6]。然后，这些方法依赖于目标物体的规则几何形状，并且当缺损区域较大时，会导致复原失败。Firman 等[7]提出的不可见区域复原策略对三维场景的复原效果较好，这种复原方法仅对几何结构进行复原，无法对语义进行分类，所以当物体的几何形状较复杂时，会降低复原的准确性。科研人员通过检测和拟合层次三维网格模型来实现几何复原和语义分类[8-11]，这些方法的复原质量受检测到的网格模型影响，当检测到的模型本身带有缺损，复原效果通常是不理想的。科研人员使用三维包围盒方法来估计物体的几何形状[12-13]，但是包围盒方法通常无法标识物体的几何细节，导致预测结果不精准。科研人员开始通过建立单独的模块来完成特征提取和上下文建模。Zheng 等[14]通过物理推理来预测不可见区域的体素。Kim 等[15]通过训练一个体素模型来实现室内场景的语义标注与复原重建。Hane 等[16]和 Blaha 等[17]采用联合优化方法对场景进行多视角重建与分类。这些方法使用预定义特征来从上下文模型中分离出特征学习部分，然而使用 CRF 模型来复原长序列场景，这种方法时间复杂度较高，语义复原成本较高，影响了复原的实时性能。

针对上述问题，本文建立了端到端的、基于大规模场景数据的深度学习网络模型，能够联合学习底层特征表示和高层上下文信息，直接从较大的接收区域获取长序列场景的上下文线索。本文构建了一个大规模合成三维场景数据集，合成场景数据已经被用于二维图像的语义分类[18-20]，然而三维合成数据还未被广泛应用。现有的数据集大多集中在特定物体或小规模的室内场景[21]。本文建立的大规模合成数据集，包含 45 622 个建筑物，775 574 个房间，作为大数据样本来训练深度神经网络模型。

3　语义场景复原网络

给定一个三维场景的深度图，本文提出的语义场景复原网络的核心是将视锥体内的体素映射到一组类标签，即 $C=\{c_0,\cdots c_{N+1}\}$，其中 N 是对象类的个数，c_0 表示空体素。在训练阶段，从三维合成场景的虚拟视角来渲染深度图，并体素化带有对象标签的整个三维场景。在测试阶段，使用 RGB-D 相机来获取深度图，本文网络的语义场景复原过程，如图 1 所示。下面分 3 个部分介绍本文语义复原网络，主要包括体素数据编码，复原网络结构和生成训练数据。

本文构建了一种深度卷积神经网络，用于学习三维场景的语义信息，如图 1 所示。首先，网络的输入为 RGB-D 图像，经过卷积运算后，提取整副图像的特征语义，生成特征图；其次，特征图经过 Maxpooling 层，得到特征向量，作为分类器的输入数据；再次，分类器对特征图像中的候选区域进行语义分类，识别目标语义；最后，根据目标语

义，对候选框的位置进行反向估计，微调至损失值达到最小即可。

3.1　体素数据编码

为了实现场景复原，首先要对可见面的深度数据进行编码，并将编码后的数据作为复原网络的有效输入，进而实现语义场景的复原操作。理想的编码方式在视角方向不变的情况下，直接将二维可见信息集成到三维输出区域，使得复原网络能够更高效的学习目标几何形状和场景表示。为了实现这一目标，本文采用截断式带符号距离函数（Truncated Signed Distance Function，TSDF）来对三维场景编码，场景中的每个体素包含两个数据，分别是距离值 d（距离该体素最近的面）和符号变量（标志该体素是空闲体素或被遮挡体素）。为了适用于本文复原网络，这里对传统 TSDF 结构进行了改进：

大多数的深度重建管线使用摄像机直线投影的方式来寻找距离最近的表面点。这种方法减少了 TSDF 的计算时间，但严重依赖于摄像机投射视角。本文选取可见面上的任意点来计算最近距离，降低了对摄像视角的严重依赖，提高了 TSDF 的计算效率。此外，沿着遮挡边界的空闲体素区域，会出现介于 $\pm d_{\max}$ 之间的强梯度变化。虽然可以通过移除符号标记来减少梯度变化，然而符号标记能够表示场景的遮挡区域，这对场景复原是至关重要的。为了改善这一问题，本文重新定义 TSDF 距离值 $d_{\text{next}}=\text{sign}(d)(d_{\max}-d)$，改进的 TSDF 距离在场景表面上具有强梯度变化，这对于复原网络学习场景的几何特征更加有效。图 2 给出了本文改进的 TSDF 编码过程，首先使用深度传感器获得场景的点云数据，即可见面。然后对输入的 RGB-D 图像计算得到标准 TSDF 体素编码数据。再次将点云信息与体素数据作为卷积网络的输入，分别用于生成特征图像和三维几何特征图，最后经过本文复原网络的分类和微调操作，实现的三维场景的语义复原操作。

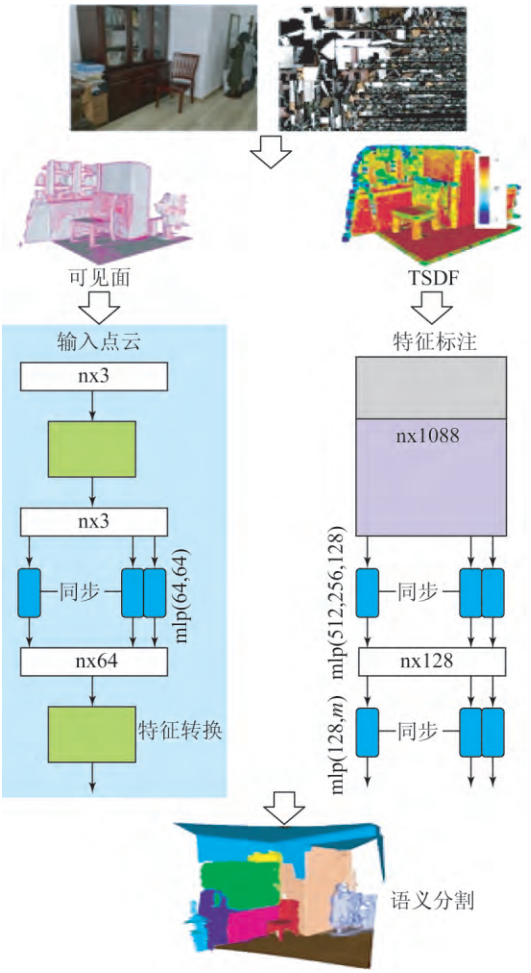

图 2　复原表面的几种 TSDF 编码方式

Fig. 2　Encoding of TSDF for Completion

3.2　复原网络结果

语义场景复原网络结构如图 3 所示。网络的输入是一个高分辨率三维体积卷，通过三维卷积层来学习本地几何表示。本文使用具有步长和聚集层的卷积网络将分辨率降低到初始输入的四分之一，并使用三维上下文模块来捕获更高层次的对象间上下文信息；接下来，不同尺度的网络响应被传送到两个卷积层，实现对多个尺度信息的聚合操作；最后，使用体素级卷积层来预测最终的体素标签。本文在卷积层添加了几个快捷连接来获取较好的梯度传播。下面介绍本文复原网络的几个核心部分。

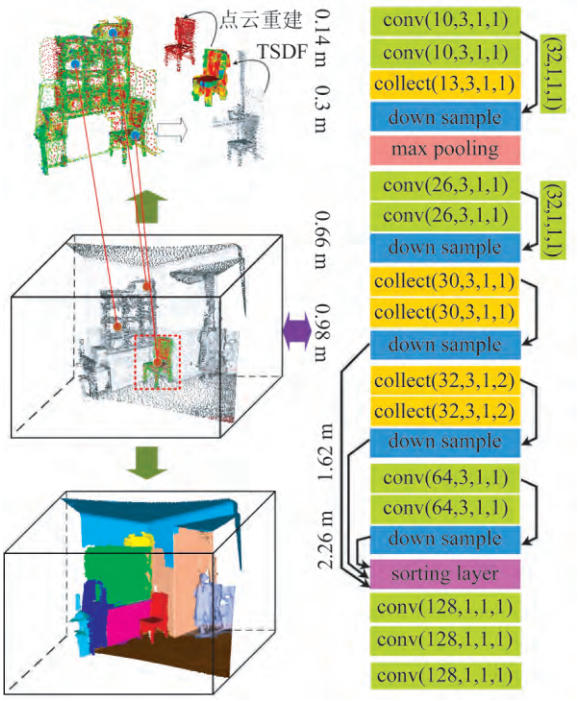

图 3　本文的语义场景复原网络

Fig. 3　Semantic scene completion of our neural network

3.2.1　生成输入体积卷

给定一个三维场景，假设三维空间尺寸分别为水平方向 4.8 m，垂直方向 2.88 m，深度为 4.8 m。本文将三维场景编码成 TSDF 格式，网格尺寸为 0.02 m，截断值为 0.24 m，生成一个 240 mm×144 mm×240 mm 的体积卷作为复原网络的输入。

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 $\{P_i \mid i=1,\cdots,n\}$，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。

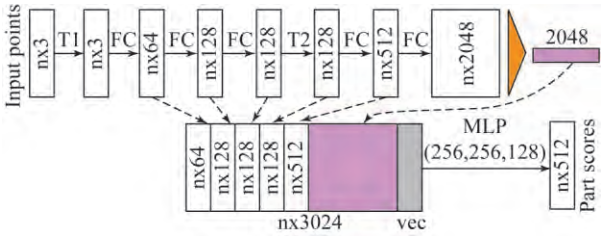

图 4　语义分类网络架构

Fig. 4　Semantic classification network architecture

3.2.2　捕获较大区域的三维上下文信息

上下文信息为理解目标场景提供有价值的线索。在三维空间中，由于缺乏高频信号，上下文信息更为重要。例如，桌面，床面和地板的几何形状相似，仅通过几何形状很难对其有效分类。因此，物体在场景中的相对位置可以作为场景分类的有效信息，实现对目标物体的有效分类。为了学习上下文信息，本文的网络需要具有足够大的接收区域。本文对增量卷积方法进行扩充，使之适用于三维场景区域。在内核卷积之前，增量卷积通过从输入卷中提取值时加入步长来扩展正常卷积。因此，在保证不损失分辨率或覆盖范围，同时保证参数数量不变的情况下，本文实现了对接收场的指数级扩展。图 5 给出了本文复原网络和传统三维卷积网络的接收区域尺寸对比，红色线框为本文网络（彩图见期刊电子版）。

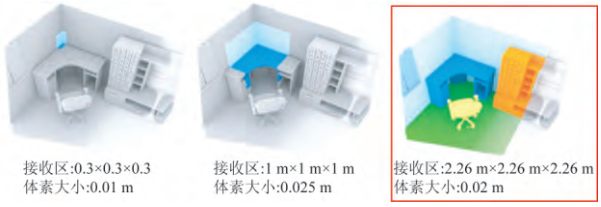

图 5　接收区尺寸对复原网络的影响

Fig. 5　Influence of receiving size on restoration network

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 $p(A_i \mid C_n)$ 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：

$$\vartheta_{mi}^{A} = p(A_i \mid C_n) = \Big[\sum_{I\in C_n} a_i(I)\Big] \div \sum C_n, \quad (1)$$

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 $\sum_i p(A_i \mid C_n) = 1$。

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 $\Theta=[\vartheta_{mi}^{A}]_{N\times M}$。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。

本文采用多任务损失函数来调节反向传递过程，即微调语义识别框的位置，使之达到像素级精准度：

$$L=L_{\mathrm{cls}}(p,u)+\lambda L_{\mathrm{reg}}(t,v), \tag{2}$$

其中 $L_{\mathrm{cls}}(p,u)=-\log\ p_u$ 是地面真值 u 的对数损失。$L_{\mathrm{reg}}(t,v)=\sum smooth_{L_1}(t_i-v_i)$，且

$$smooth_{L_1}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & |x|\geqslant 1\end{cases}.$$

3.2.3　多尺寸上下文聚合

不同的目标物体具有不同的三维物理尺寸，因此为了有效识别物体，网络需要捕获不同尺度的信息。例如，为了识别小物体，需要更多的本地信息；为了识别大物体，需要更多的全局信息。为了聚合不同尺寸的信息，本文设计了一个层次，将网络响应与不同的接收域连接起来。再将聚合的特征图映射到两个卷积层，这使得本文复原网络能够跨不同尺寸的响应来传递信息。

3.2.4　均衡体素项

由于三维数据具有稀疏特性，空闲体素和被占用体素的比率为 9∶1。为了均衡体素分布，本文对训练数据进行取样，保证每个抽样区域都具有一个相对均衡的体素分布。对于每个包含 N 个被占用体素的训练体素卷，为被占用区域随机抽取 $2N$ 个空闲体素，同时忽略视锥体外的空闲体素。

3.2.5　计算体素损失值

网络的损失函数是体素级损失之和，即：

$$L(p,y)=\sum_{i,j,k}w_{ijk}L_{\mathrm{sm}}(p_{ijk},y_{ijk}),$$

其中：L_{sm} 是体素损失，y_{ijk} 是地面真值标签，p_{ijk} 表示在 $N+1$ 个分类下，在坐标 (i,j,k) 处的体素预测概率，N 表示物体分类个数，空闲体素属于分类 0。权值 w_{ijk} 基于上述抽样算法，取值等于 0 或 1。

3.2.6　训练数据

本文使用 Caffe 深度卷积神经网络[22]来训练复原标签。在 LS_3DDS 数据集上预训练本文的语义场景复原网络，需要大约 15 d，GPU 型号是 Nvidia GTX1070。在训练期间，每个子块包含一个三维视图卷，需要 8 GB 的 GPU 内存。为了获得更稳定的梯度估计，本文累积了 4 次迭代梯度，然后再更新权重。

3.3　合成训练数据

训练复原网络的难点之一是缺乏体素级别的大规模密集语义注释数据集。现有的 RGB-D 数据集通常会被遮挡或部分可见，无法为整个空间提供体素级别的体积占用和语义标签。Firman 等使用 KinectFusion 方法收集带有重建 RGB-D 视频的桌面数据集[23]。但是该数据集不带有语义标签，仅包含简单的桌面场景。本文给出了一种大规模合成三维场景数据集 LS_3DDS，并从中获取大量的具有合成深度图像和地面实况的训练数据。

本文给出的 LS_3DDS 数据集包含 45 622 个不同的场景，这些场景是真实的带有家具装饰的室内房间，并使用专业图形软件来设计。经过筛选，删除数据集中的重复和空白场景，保证了场景数据集的质量。最后，场景共分 84 类，有效的地面场景 49 884 个，有效的目标物体 5 697 217 个，并手动标记物体，得到带有分类标签的场景数据集。

深度图能够模拟典型图像的捕获过程，为了生成合成深度图，本文使用一种简单的方法来获得相机视点。给定一个三维场景，首先从一个均匀的网格开始，间隔 1 m，并且该网格不被物体所占据。然后，根据高斯采样分布来选择摄像头姿态，即相机高度采用 $\mu=1.5$ m，$\sigma=0.1$ m 的高斯分布进行采样；相机倾斜角度采用 $\mu=-10°$ 和 $\sigma=5°$ 的高斯分布进行采样。其次，本文使用深度传感器来绘制得到深度图，同时剔除不佳相机视点。即当视图满足以下 3 点时，表示该视图是有效视图：(1)有效深度区域占用图像的 70%以上；(2)除了墙壁，天花板和地板之外，包含两种以上的物体；(3)除了墙壁，天花板和地板之外，物体占用图像面积的 30%以上。为了减少数据冗余，本文从每个房间挑选目标语义最丰富的 5 幅图像，共生成 130 269 个有效视图来训练本文的语义复原网络。

由于 LS_3DDS 三维场景数据集由有限数量的对象实例组成，本文首先对库中的每个对象进行体素化，然后根据每个场景的配置和视点来转

换各自的标签，进而加速体素化过程。例如，本文将每个对象体素化为一个 128×128×128 的体素网格，并设置体素大小 s，使得对象的最大尺寸紧紧贴在对象边界框上；由于对象尺度不同，s 也随之变化。

给定相机视图，本文定义了世界坐标中大小为 240×144×240 体素网格，场景体素大小等于 2 cm。对于场景中的每个对象，通过转换，旋转和缩放操作来转换对象体素网格。然后，在变换对象边界框内，对场景体素网格中的每个体元进行迭代处理，并计算其到最近邻居对象体素的距离。如果距离小于对象体素大小 s，则该场景体素将被标记为属于该对象类别。类似地，本文标记墙壁，地面和天花板场景的所有体素，并将它们当做厚度等于一个场景体素大小的平面来标记。所有剩余体素被标记为空白体素。

4　实验结果与分析

为了评估本文方法的鲁棒性与实时性，使用真实数据集与合成数据集，将本文方法与其它几种方法进行性能对比。

对于真实场景，本文采用 NYU 数据集[3]，包含 1 449 个深度图。对 Guo 等[24-25]的三维网格注释进行体素化处理，获得了地面真实场景。基于 Handa 等方法获得对象类别[18]。注释由 7 个类别、33 个对象网格组成，其他分类使用三维方框或平面近似。在某些情况下，由于标签错误和有限的网格集，使得网格注释与深度数据的对应关系不完全一致。为了解决这种不一致性，Firman 等使用渲染深度图进行测试。然而，由于渲染过度简化的网格，导致几何细节丢失，尤其在对象被表示为三维方框的区域丢失严重[26]。因此，本文使用渲染深度图和源网格注释图进行测试。

对于合成数据，本文创建了一个几何细节完备的对象测试集 LS_3DDS，其深度图和地面真实体素完全对齐。LS_3DDS 测试集包括从 184 个场景渲染得到的 500 个深度图像。

本文将预测体素标签的体素级交叉点(Intersection over Union，IoU)作为评价指标。对于语义场景复原操作，本文计算可见和遮挡体素上每个对象类的 IoU。对于场景重建操作，本文将所有非空对象类视为一个类别，并计算遮挡体素的 IoU。这里不评估视图外的体素或房间。

表 1 和表 2 给出了定量比较结果，图 6 与图 7 给出了定性比较结果。在表 1 中，本文将语义场景复原操作与 Lin 等[12]，以及 Geiger 和 Wang 等[10]的复原方法进行比较，以下简称 L 复原方法与 GW 复原方法。这两种算法都将 RGB-D 帧作为输入，并在三维场景中产生对象标签。L 复原方法使用三维边界框和平面近似所有对象。GW 复原方法在测试时间内检测观察到的深度图，并整合成三维网格模型。用于检索的网格模型库是用于地面真实注释的子集模型库。因此，他们可以在一个小数据库中确定精确的网格模型，进而实现精准对齐。相比之下，本文算法仅基于深度图像，并且在测试时不使用额外的网格模型。尽管如此，本文的深度复原网络产生更准确的体素级预测，本文方法的复原百分比为 30.5%，GW 复原方法为 19.6%。从图 8 中可以看出(彩图见期刊电子版)，这两种复原网络都无法识别电脑显示器(橙色虚线方框标记)，而本文网络(红色方框标记)正确检测了显示器对象，并将其重建。此外，由于本文方法不需要整合模型，节省了每个场景的复原时间。

表 1　本文方法与 L/GW 复原方法的性能对比

Tab. 1　Comparison of three algorithm(ours, L and GW) with different models

		L	GW	本文 NYU	本文 LS_3DDS	本文 NYU+LS_3DDS
复原	闭环率	58.5	65.7	57.0	55.6	**59.3**
	IoU	36.4	44.4	55.1	53.2	**56.6**
语义场景复原	天花板	0	10.2	15.1	5.8	**15.1**
	地面	11.7	62.5	**94.7**	81.8	94.6
	墙壁	13.3	19.1	24.4	19.6	**24.7**
	窗	**14.1**	5.8	0	5.4	10.8
	椅子	9.4	8.5	12.6	12.9	**17.3**
	皮箱	29.0	40.6	32.1	34.4	**53.2**
	花盆	24.0	27.7	35.0	26.0	**45.9**
	桌子	6.0	7.0	13.0	13.6	**15.9**
	显示器	7.0	6.0	7.8	6.1	**13.9**
	家具	16.2	22.6	27.1	9.4	**31.1**
	物品	1.1	5.9	10.1	7.4	**12.6**
	平均值	12.0	19.6	24.7	20.2	**30.5**

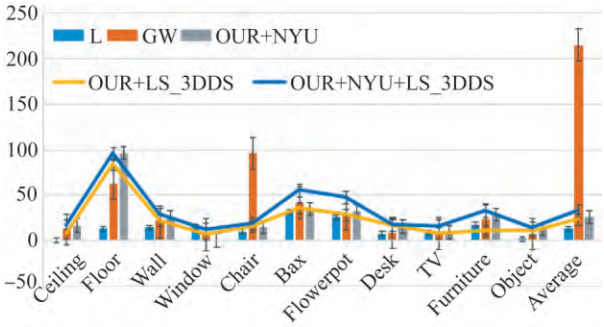

图 6　几种网络的语义分类性能对比

Fig. 6　Comparison of several method for performance of semantic classification

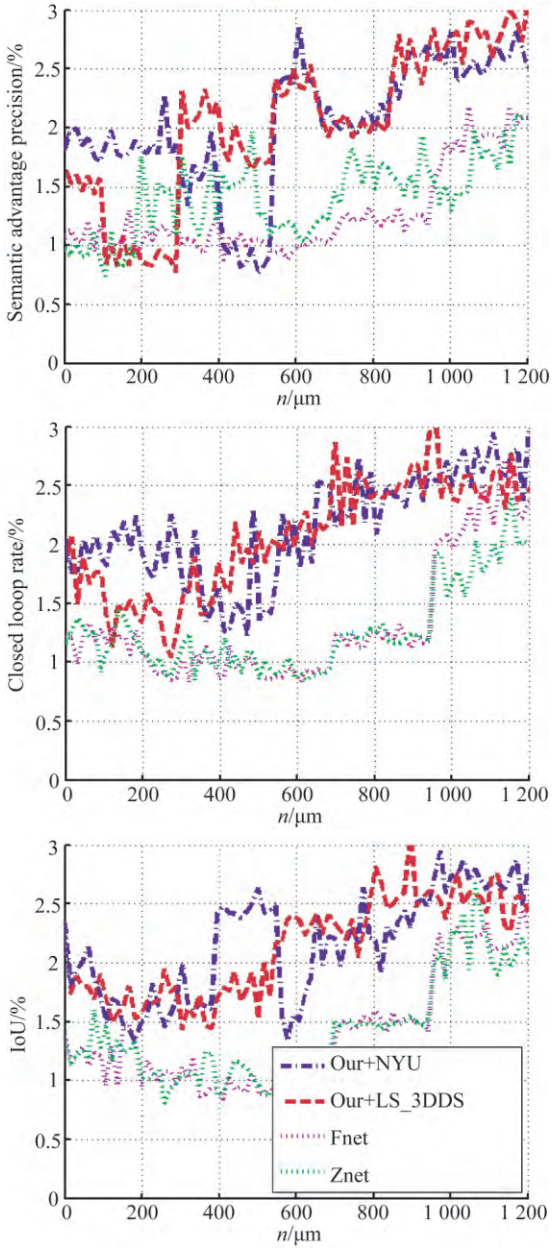

图 7　几种方法的语义识别准确率对比曲线图

Fig. 7　Comparison of several method for accuracy of semantic identification

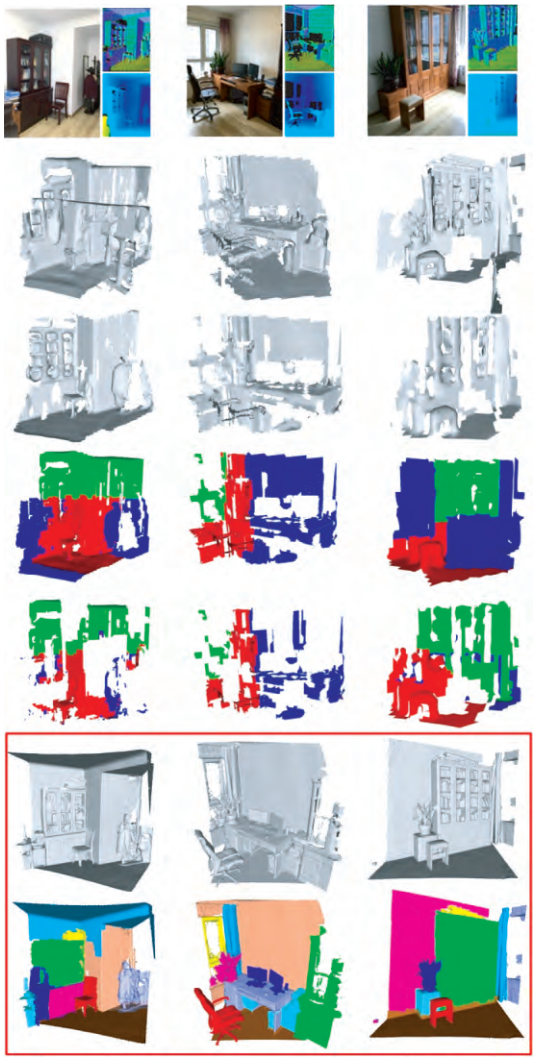

图 8　几种复原网络的语义重建结果对比图

Fig. 8　Comparison of several network for semantic reconstruction

本文通过训练模型来预测每个体素的占用情况，并对每个体素进行二进制分类，0 表示空白体素，1 表示被占用体素，在表 2 中对训练模型的性能进行了比较，将本文方法与 Firman 等[7] 和 Zheng 等[14] 的方法进行比较，以下简称 F 复原方法和 Z 复原方法，这两种方法都是基于单一深度图来预测二进制体素占用情况，并无场景语义分类。本文方法将 *F* 复原方法与 *Z* 复原方法结合，其中 F 复原方法仅实现了复原操作。本文从测试数据集中随机选取 200 副图像，使用 NYU 评价标准来测试，F 复原方法对于大多数场景的复原效果良好，但是当场景较复杂时，会出现复原失败现象。例如，在图 8 中第 3 列的书柜复原失败

（黑色原点处），第一列的箱子也复原失败（黑色原点处）。相比之下，本文的复原网络利用三维上下文的语义来更好地完成几何场景复原。实验结果表明理解对象语义有利于提高场景复原的有效性，降低失败概率。

本文预训练了一个深度神经网络模型，仅在可见表面上估计语义标签，并将其与联合训练模型进行比较。即使仅在可见表面进行评估，使用场景复原并附加监督训练的模型胜过仅具有表面标记训练的模型，仅带有表面标记的语义场景复原比率为 51.2%，联合训练模型的复原率为 54.2%。这表明理解完整的对象几何和三维上下文信息有利于识别对象语义。

表 2　本文方法与 F、Z 复原方法的性能对比

Tab. 2　Comparison of three algorithm(ours, F and Z) with different models

	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

为了验证合成训练数据的有效性，本文比较了仅使用 NYU 训练的模型和在 LS_3DDS 上进行预训练之后的模型，如表 1 和表 2 所示。实验结果表明合成数据有助于提高网络的复原性能，对于语义场景复原操作，IoU 值提高了 10.3%。

图 9 给出了输入深度图和预测结果。即使桌子周围的其他对象的深度信息不可见，本文复原网络也会根据观察到的对象来预测上下文对象，接收域从 1.16 m 增加到 2.26 m。图 9 中的第一幅图像为根据深度图重建的点云模型，其他图像为场景中对象的预测结果。虽然没有观察到任何其他对象的信息，本文网络仍然可以根据观察到的对象和学习到的三维上下文来预测对象位置，扩大接受域使网络具有更丰富的上下文信息，网络复原性能从 42.0%提高到 44.3%。

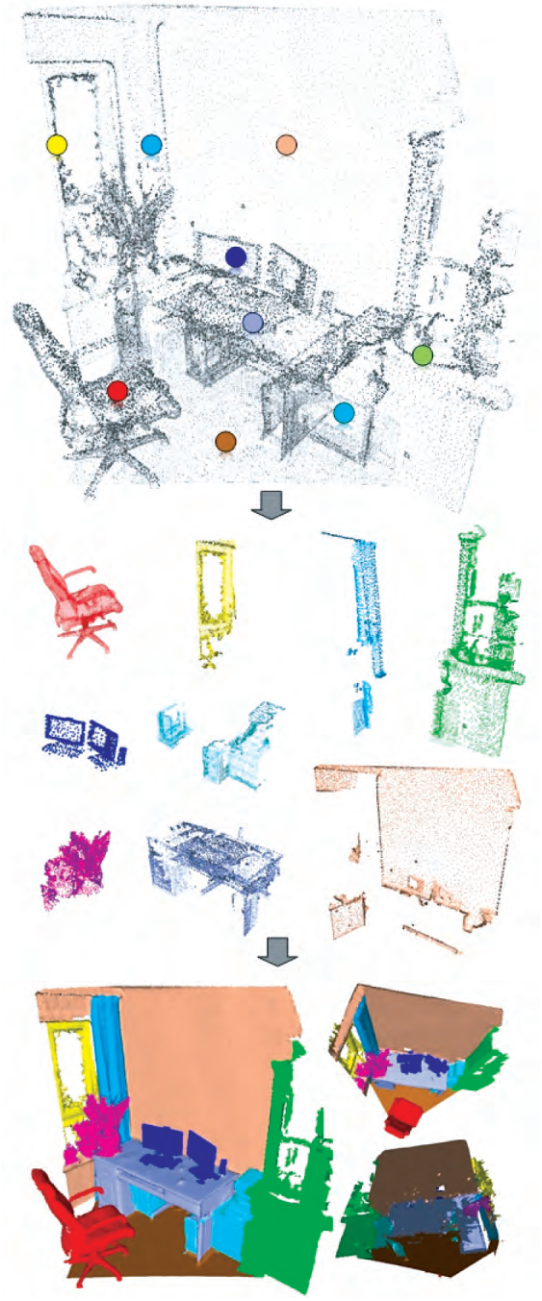

图 9　本文网络的三维对象语义分类图

Fig. 9　Semantic prediction of 3D object by our network

5　结　论

本文给出了一种语义场景复原网络模型，这是一种三维深度神经网络模型，用于预测三维场景的空间占用和语义标签，实现了对目标场景的语义复原操作。构建了一种深层卷积神经网络，用于学习三维场景的对象语义信息，并对三维对象进行语义分类和反向微调操作，进而实现像素

级的精准语义场景复原。给出了一个大规模合成三维场景数据集，用于训练语义场景复原网络。实验结果表明，三维语义复原网络与单一模式复原网络相比，复原性能提高了 2.0%；本文网络使用了三维上下文信息和合成数据集来训练神经网络模型，提高了复原鲁棒性。

参考文献：

[1] GUPTA S, ARBELÁEZ P, MALIK J. Perceptual organization and recognition of indoor scenes from RGB-D images [C]. *Proceedings of* 2013 *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2013: 564-571.

[2] REN X F, BO L F, FOX D. RGB-(D) scene labeling: features and algorithms [C]. *Proceedings of* 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012: 2759-2766.

[3] SILBERMAN N, HOIEM D, KOHLI P, *et al*.. Indoor segmentation and support inference from RGBD images [C]. *Proceedings of the* 12*th European Conference on Computer Vision*, Springer, 2012: 746-760.

[4] LAI K, BO L F, FOX D. Unsupervised feature learning for 3D scene labeling [C]. *Proceedings of* 2014 *IEEE International Conference on Robotics and Automation*, IEEE, 2014: 3050-3057.

[5] ROCK J, GUPTA T, THORSEN J, *et al*.. Completing 3D object shape from one depth image [C]. *Proceedings of* 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2015: 2484-2493.

[6] MONSZPART A, MELLADO N, BROSTOW G J, *et al*.. RAPter: rebuilding man-made scenes with regular arrangements of planes [J]. *ACM Transactions on Graphics*, 2015, 34(4): 103.

[7] FIRMAN M, AODHA O M, JULIER S, *et al*.. Structured prediction of unobserved voxels from a single depth image [C]. *Proceedings of* 2016 *IEEE Computer Vision and Pattern Recognition*, IEEE, 2016: 5431-5440.

[8] GUPTA S, ARBELÁEZ P, GIRSHICK R, *et al*.. Aligning 3D models to RGB-D images of cluttered scenes [C]. *Proceedings of* 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2015: 4731-4740.

[9] SONG S R, XIAO J X. Sliding shapes for 3D object detection in depth images [C]. *Proceedings of the* 13*th European Conference on Computer Vision*, Springer, 2014: 634-651.

[10] GEIGER A, WANG CH H. Joint 3D object and layout inference from A single RGB-D image [M]//GALL J, GEHLER P, LEIBE B. *Pattern Recognition*. Cham: Springer, 2015: 183-195.

[11] NAN L L, XIE K, SHARF A. A *search-classify* approach for cluttered indoor scene understanding [J]. *ACM Transactions on Graphics*, 2012, 31(6): 137.

[12] LIN D H, FIDLER S, URTASUN R. Holistic scene understanding for 3D object detection with RGBD cameras [C]. *Proceedings of* 2013 *IEEE International Conference on Computer Vision*, IEEE, 2013: 1417-1424.

[13] SONG S, XIAO J. Deep sliding shapes for amodal 3D object detection in RGB-D images [J]. *Computer Science*, 2015, 139(2): 808-816.

[14] ZHENG B, ZHAO Y B, YU J C, *et al*.. Beyond point clouds: scene understanding by reasoning geometry and physics [C]. *Proceedings of* 2013 *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2013: 3127-3134.

[15] KIM B S, KOHLI P, SAVARESE S. 3D scene understanding by voxel-CRF [C]. *Proceedings of* 2013 *IEEE International Conference on Computer Vision*, IEEE, 2013: 1425-1432.

[16] HÄNE C, ZACH C, COHEN A, *et al*.. Joint 3D scene reconstruction and class segmentation [C]. *Proceedings of* 2013 *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2013: 97-104.

[17] BLÁHA M, VOGEL C, RICHARD A, *et al*.. Large-scale semantic 3D reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling [C]. *Proceedings of* 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2016: 3176-3184.

[18] HANDA A, PATRAUCEAN V, BADRINARAYANAN V, *et al*.. SceneNet: understanding real world indoor scenes with synthetic data [J]. *Computer Science*, 2015: 4077-4085.

[19] 吕朝辉，沈萦华，李精华. 基于 Kinect 的深度图像修复方法 [J]. 吉林大学学报(工学版), 2016, 46(5): 1697-1703.

LÜ CH H, SHEN Y H, LI J H. Depth map inpainting method based on Kinect sensor [J]. *Journal of Jilin University (Engineering and Technology Edition)*, 2016, 46(5): 1697-1703. (in Chinese)

[20] **刘迎，王朝阳，高楠，等. 特征提取的点云自适应精简 [J]. 光学 精密工程**, 2017, 25(1): 245-254.
LIU Y, WANG CH Y, GAON, *et al*.. Point cloud adaptive simplification of feature extraction [J]. *Opt. Precision Eng.*, 2017, 25(1): 245-254. (in Chinese)

[21] **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 [J]. 自动化学报**, 2017, 43(5): 814-821.
HU CH SH, ZHAN SH, WU C ZH. Image super-resolution based on deep learning features [J]. *Acta Automatica Sinica*, 2017, 43(5): 814-821. (in Chinese)

[22] CHANG A X, FUNKHOUSER T, GUIBAS L, *et al*.. ShapeNet: an information-rich 3D model repository [J]. arXiv:1512.03012, 2015.

[23] JIA Y Q, SHELHAMER E, DONAHUE J, *et al*.. Caffe: convolutional architecture for fast feature embedding [C]. *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, 2014: 675-678.

[24] NEWCOMBE R A, IZADI S, HILLIGES O, *et al*.. KinectFusion: real-time dense surface mapping and tracking [C]. *Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality*, IEEE, 2011: 127-136.

[25] GUO R Q, ZOU CH H, HOIEM D. Predicting complete 3D models of indoor scenes [J]. arXiv: 1504.02437, 2015.

[26] **蔡强，郝佳云，曹健，等. 结合局部特征及全局特征的显著性检测 [J]. 光学 精密工程**, 2017, 25(3): 772-778.
CAI Q, HAO J Y, CAO J, *et al*.. Salient detection via local and global feature [J]. *Opt. Precision Eng.*, 2017, 25(3): 772-778. (in Chinese)

作者简介：

林金花(1980－)，**女，吉林长春人，博士，讲师**，2004 年、2008 年于西安交通大学分别获得学士、硕士学位，2017 年于中国科学院长春光机所获得博士学位，**主要从事数字图像处理与目标识别方面的研究**。E-mail: ljh3832@163.com

王延杰(1963－)，**男，吉林长春人，研究员，博士生导师**，1988 年于吉林工业大学获得学士学位，1998 年于中国科学院长春光机所获得硕士学位，**主要从事数字图像处理，信息处理，自动目标识别等方面的研究**。E-mail: wangyj@ciomp.ac.cn