单丰. 基于融合采样策略的轻量级 RGB-D 场景 3D 目标检测[J]. 智能计算机与应用,2024,14(4):68-75. DOI: 10.20169/j.issn.2095-2163.240409

基于融合采样策略的轻量级 RGB-D 场景 3D 目标检测

单 丰

(浙江理工大学 计算机科学与技术学院, 杭州 310018)

摘 要: 针对室内 RGB-D 场景中 3D 目标检测对复杂背景的适应性较差、难以进行有效采样,以及场景推断时间较长等问题, 本文提出一种基于融合采样策略的轻量级 RGB-D 场景 3D 目标检测方法。该方法以场景 RGB-D 数据作为输入, 首先通过深度相机投影将其转化为三维点云场景; 然后利用一种结合距离最远点采样和特征最远点采样的融合采样策略对场景点云进行采样, 有效保留了场景各实例代表点,将所有特征代表点汇集在一起形成场景的特征代表点云; 最后在代表点云中利用深度霍夫投票机制投票出场景各物体的中心, 并对各物体周围的相关特征进行聚类,从而实现场景的 3D 目标检测。实验结果表明, 与传统方法相比, 所提框架的目标检测准确率得到有效提升, 同一评估指标下的检测准确率平均提升 2.1%, 且同一环境下每个场景的推断速度 仅需要 57 ms, 远快于传统方法 2 倍多,从而保证了室内场景 3D 目标检测的准确性和高效性。

关键词: 3D 目标检测; 深度学习; RGB-D; 三维点云; 室内场景
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2024)04-0068-08

3D object detection in lightweight RGB-D scene based on fusion sampling

SHAN Feng

(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Aiming at the problems of 3D object detection in indoor RGB - D scenes such as poor adaptability to complex backgrounds, difficulty in effective sampling, and long scene inference time, this paper proposes a lightweight 3D object detection method for RGB-D scenes based on fusion sampling strategy This method takes scene RGB-D data as input, and first converts it into a 3D point cloud scene through depth camera projection; Then, a fusion sampling strategy combining farthest distance point sampling and feature farthest point sampling is used to sample the scene point cloud, effectively preserving the representative points of each instance of the scene, which are collectively referred to as feature representative points; Secondly, all feature representative points are gathered together to form a feature representative point cloud of the scene; Finally, a deep Hough voting mechanism is used to vote on the center of each object in the scene in the representative point cloud, and relevant features around each object are clustered to achieve 3D object detection in the scene The experimental results of network training on SUN RGB-D dataset show that the proposed framework effectively improves the accuracy of target detection compared to traditional methods, with an average improvement of 2.1% under the same evaluation index. Moreover, the inference speed for each scene in the same environment is much faster than traditional methods, which only requires 57 ms, thereby ensuring the accuracy and efficiency of 3D target detection in indoor scenes.

Key words: 3D target detection; deep learning; RGB-D; point cloud; indoor scenes

0 引 言

作为图形学与三维视觉中的重要问题, 目标检测得到工业界与学术界的普遍关注, 目前已成为视觉领域许多复杂高层视觉任务的基础, 如场景理解[1]、目标跟踪[2]、场景分割[3]等。然而,由于场景中不同目标物体外观、形状和姿态各不相同,以及成像过程中的光照、遮挡等因素干扰, 使得目标检测成为一个具有挑战性的难题。

近年来, 目标检测方法已经由传统基于手工设计特征的检测策略发展到基于深度神经网络的检测方法。基于卷积神经网络在特征提取中的有效性, Li 等[4]提出利用基于区域的卷积神经网络(Region-Based Convolutional Neural Network, R-CNN)检测

作者简介: 单 丰(1997-), 女,硕士研究生, 主要研究方向:计算机视觉与图像。Email:648389009@qq.com

收稿日期: 2023-03-17

场景图像中的目标对象，但由于检测时其对候选目标区域进行了局部缩放，而导致其对目标的检测精度受到一定限制。Peng 等[5]提出的基于空间金字塔池化的检测网络（Spatial Pyramid Pooling Network，SPP-Net），能够将任意大小的场景目标候选区域特征信息转换为固定长度的特征向量，从而能有效进行目标检测的多尺度训练。Ren 等[6]提出的快速区域卷积神经网络（Faster Regions with CNN Features，Faster R-CNN）是针对传统卷积神经网络的改进，其将场景图片输入卷积神经网络中提取其特征图，再使用建议框由特征图提取特征框，以减少卷积操作的重复计算，进一步提高目标检测精度与效率。

除场景图像理解之外，三维场景目标检测对于大量真实场景应用而言至关重要，如自动驾驶、家用机器人等。针对 3D 目标检测的主流方法可以分为 2 类：基于全局特征的检测方法和基于局部特征的检测方法。关于基于全局特征的检测策略方面，如 Ru 等[7]提出的视点特征直方图（Viewpoint Feature Histogram，VFH）方法，该方法对场景表面噪声和丢失的深度信息具有鲁棒性，使其能快速检测出目标物体，但由于对遮挡环境的适应性较差，导致其检测准确率不高。基于局部特征的检测方法则首先提取场景中的关键点，并计算各关键点的特征描述符，再根据这些描述符进行目标检测。典型的关键点检测方法包括 3D Harris 检测[8]、LSP 检测[9]等，这些方法对复杂场景的背景适应性较差、计算耗时，且由于场景中三维目标的不规则数据格式与 6 个自由度的大搜索空间，导致在室内背景点云干扰下，其有效点云的特征难以得到有效利用，从而给目标检测带来较大挑战。

本文提出一种基于融合采样的轻量级 RGB-D 场景 3D 目标检测方法，且以场景中的 RGB-D 数据作为输入，提出一种结合距离最远点采样方法和特征最远点采样的融合采样策略，该策略能够在保留场景物体周围前景点的同时，亦能有效保留距离物体较远的场景前景点，并且留存一部分有助于物体分类的背景点，这些点统称为特征代表点；将具有大量特征信息的特征代表点通过投票网络中进行投票，最终产生出物体投票中心和 3D 目标建议框。实验表明，本文方法在目标检测中能够对点的特征进行充分采样，并通过融合采样策略保留下具有实例特征的代表点，该方法不仅排除了无关信息干扰，还加快了模型推断速度，从而能够高效、准确地从复杂室内场景中检测到 3D 目标物体。通过在 SUN RGB-D[10]数据集上的对比实验，验证了本文方法的有效性。相同实验环境下，本文框架在点云上的目标检测准确率提升至 59.8%且每个场景的推断时间只需 57 ms。

1　相关工作

针对三维物体的目标检测方法，主要包括传统的目标检测方法和基于深度神经网络的目标检测方法。

传统的目标检测算法通常依赖于各种定义的特征描述符。首先，通过不同大小的滑动窗口选择具有可能目标的多个图像区域；然后 通过特征提取，将区域中包含的信息转换为特征向量并对其进行分类。如：Lee 等[11]利用滑动窗口先检测出人脸特征在图像上的所有可能位置，并训练了一个能够用于检测 2 人人脸的检测器，再利用 AdaBoost 算法从潜在特征数据库中筛选出少量重要特征来建立分类器，从而实现实时人脸检测。此后，方向梯度直方图（Histogram of Oriented Gradient，HOG）特征算法[12]与支持向量机分类器相结合被广泛应用于目标检测任务，但由于其探测器计算量过大而导致检测效率低下。

目前，基于深度神经网络的目标检测方法得到普遍重视。基于卷积神经网络，Chen 等[13]提出一种 R-CNN 目标检测模型，其使用选择性搜索在图像上生成高质量候选区域，并使用 AlexNet 网络提取特征信息并利用支持向量机获得目标类别，最终实现检测框校准。借助输入的场景 RGB 图像和激光雷达点云数据，Li 等[14]提出的多视图 3D 目标检测方法先将三维点云投影到鸟瞰图与前视图，鸟瞰视图用于生成三维先验框，并将先验框投影到前视图和 RGB 图像中，再生成特征图并利用 ROI 池将其集成至同一维度，最终将集成数据经网络融合输出分类结果和包围盒。针对小目标遮挡和变形问题，Yan 等[15]提出一种基于快速 R-CNN 的检测方法，该方法对目标区域特征的遮挡与变形进行对抗处理以提升目标检测对遮挡和变形的鲁棒性。然而，由于场景环境的变化差异（如光线、视角等变化）通常会导致遮挡物体出现尺度不同现象，从而将导致目标检测效果的不理想。

利用二维目标检测驱动三维目标检测是目前图像与点云结合的目标检测方法中较为典型的方法，即二维目标框通过相机变换矩阵转到三维空间的视

锥中，然后在三维视锥中进行目标检测。例如 Qi[16] 提出的 F-PointNet 网络首先利用 FPN(Feature Pyramid Network)检测器在二维图像上提取目标二维检测框，并将检测框投影到三维空间，形成斜锥体，并对当前斜锥体内的点云进行语义分割以及三维边界框回归。但由于点云数据过于稀疏，需要适当增加图像分辨率提高候选框尺寸。借鉴 F-PointNet 的融合思想，Wang[17] 等提出 F-Convnet 网络，该网络借助 2D 候选区域生成视锥体，并将每个视锥体内点特征聚合为视锥体序列特征，再将这些特征排列为 2D 特征图送入全卷积层进行特征提取，最后利用检测头进行 3D 框的端到端估计。但是，由于该方法依赖于过少的前景点，容易误分割。

此外，最近出现的基于点云场景的目标检测中，由于点云场景中离散采样点分布不具有规则网格，场景中的物体质心往往存在于表征物体的点云数据之外(如：桌子、椅子)，因而传统的二维图像目标检测策略难以直接应用于三维点云场景。

2 本文方法

本文以场景 RGB-D 数据作为输入，首先利用二维卷积神经网络提取出 RGB 图像中的二维目标对象区域，并为其进行分类；再利用相机投影矩阵，将二维区域转化为三维点云；然后在此点云上，本文算法舍弃了目前基于点的方法中不可或缺的特征传播层和细化层结构，从而直接对点进行采样。本文采用了一种结合距离最远点采样方法与特征最远点采样方法相融合的采样策略，不仅保留了物体周围有效的前景点，同时也保留了和物体语义信息相关的部分背景点，便于后续的回归和分类，本文称这些点为特征代表点；最后为了更好地利用保留下来的代表点，本文采用基于深度霍夫[18] 的投票检测算法进行检测，利用代表点中携带的丰富特征信息投票出物体中心，再对中心附近的特征进行聚类并提出 3D 建议。网络总体框架如图 1 所示。

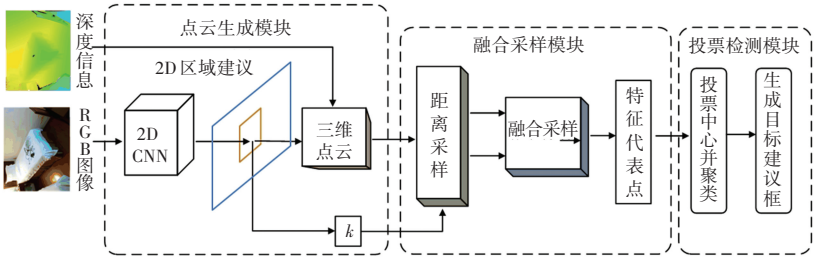


图 1 本文网络框架

Fig. 1 Network framework of this paper

2.1 基于 RGB-D 图像的三维点云生成

通常情况下，大多数 3D 传感器(尤其是深度传感器或深度相机)获取的场景数据深度信息的分辨率均低于 RGB 图像分辨率。因此，本文利用二维卷积网络提取出彩色图像中目标对象的二维区域，并为对象进行分类。具体地说，在三维点云生成过程中，利用已知的摄影机投影矩阵将待预测目标对象的二维边界框提升至三维区域，使场景中各目标物体均能获取其对应的三维点云；接着收集场景中每一个物体的三维点云，以形成整个场景的三维点云。

2.2 距离最远点采样

现有的两阶段 3D 目标检测方法大多数是利用语义分割获得了前景点后，以每个前景点为中心进行 3D 检测框的粗提议，再对这些粗提议检测框内部点进行特征提取与处理，最后微调检测框，以获得更精确的提议。但是这些方法中的特征传播层和细化过程在模型推断过程中会消耗一半以上的时间，如果简单地将这些模块删除，直接对点进行基于距离的最远点采样进而生成提议，则会造成检测精度降低。

基于距离的最远点采样法特点是：以空间距离最大为原则，不断迭代采样场景的点云，使采样后的点云基本覆盖整个场景。这种方法虽然避免了随机采样中会更着重于密度较高点云的问题，但是因为场景中背景点数量偏多，而有些较远目标中的前景点较少，所以这样的采样方式几乎会过滤掉距离较远的物体的所有前景点，导致这些具有少量前景点的物体不能被检测到，最终使检测性能下降。

2.3 特征最远点采样

由于特征传播层和细化阶段消耗了大量推断时间，因此移除特征传播层是非常重要的。但是如果没有特征传播层，网络将会直接利用距离最远点采样法，来选择点的子集作为下采样的代表点。这种

采样方法只考虑了点之间的相对位置，而没有注意到由于点的数量巨大，大部分幸存的代表点实际上是背景点这一问题。尤其对于远距离物体或者小型物体，由于总代表点的数量有限，因此极有可能会忽略这些物体，因为其前景点的数量远少于背景点的数量。

针对上述问题，本文注意到物体的语义信息能够被深度神经网络很好地捕获，因此利用特征距离作为特征采样中的准则，可以去除背景上的许多相似的无效点，同时又可以保证与对象相关的远处前景点不被过滤掉。因为来自不同对象点的语义特征各不相同，因此为了保留有效代表点并删除位于背景上的无用点，本文采用了一种基于特征距离的采样策略。然而，仅将语义特征距离作为唯一标准将在一个物体中保留相当多的点，这会引起前景点冗余。因此，为了减少冗余和增加多样性，本文将空间距离和语义特征距离共同作为最远点采样的标准，本文将这种采样方法称为特征最远点采样，其公式如下：

$$C(A,B) = \lambda L_d(A,B) + L_f(A,B) \tag{1}$$

其中，L_d 为空间距离计算函数；L_f 为特征距离计算函数；λ 为权重。

2.4　融合采样

由于本文采用了特征最远点采样方法对点进行采样，不同物体中的大量有效前景点得以保留。然而，由于总代表点的数量有限，会造成背景点被丢弃，这虽然有利于回归，但是不利于分类。因为在采样后的特征聚合中，需要对一个点及其这个点周围一定范围内的点进行局部特征的提取。如果这个点周围的背景点数量过少，那么对于该点分类的敏感度也会受限，因此在分类任务中效果较差。

由上述分析可知，不仅应该尽可能多地采样前景点，还需要收集足够且有效的背景点，以便进行更可靠的分类。基于此，本文提出了一种结合距离最远点采样和特征最远点采样的融合采样策略，该策略要求上述两种采样方法分别对一半的点进行采样，以保留更多用于定位的前景点和足够的用于分类的背景点。融合采样过程如图 2 所示：

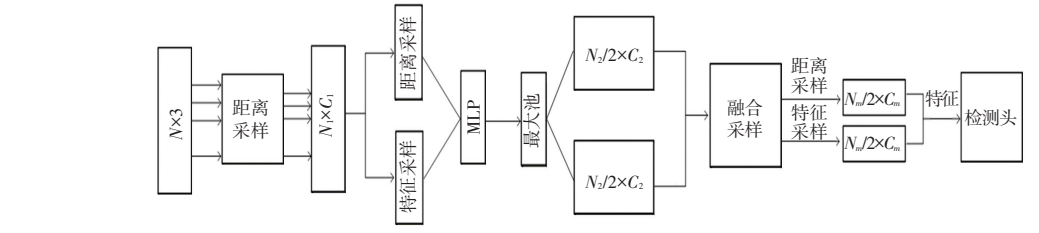


图 2　融合采样过程

Fig. 2　Fusion sampling process

以场景点云数据作为该模块输入，首先通过距离最远点采样对场景各个采样点进行采样得到新的采样点特征；然后结合距离最远点采样和特征最远点采样分别采样场景中一半采样点，并通过多层感知器 MLP 和最大池化操作分别得到含有新特征的采样点；其次将 2 部分含新特征的场景采样点进行融合并重复一次上述操作以完成融合采样；最后将得到的含有丰富特征信息的采样点，即特征代表点，输入到检测网络中进行检测。

2.5　投票检测

为了更好地利用保留下来的特征代表点，本文采用基于深度霍夫的投票检测算法进行检测，利用代表点中携带的丰富特征信息投票出物体中心进而提出 3D 建议，检测过程如图 3 所示。

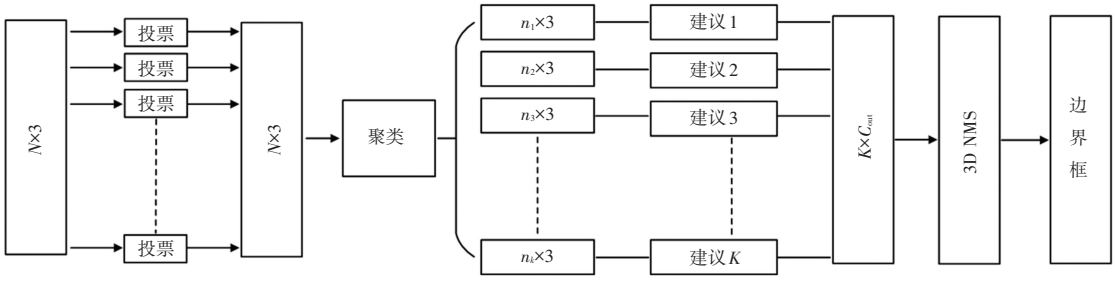


图 3　3D 目标检测网络

Fig. 3　3D object detection network

给定具有 (x,y,z) 坐标的 N 个特征代表点作为场景点云输入，此时这些代表点被视为种子点。每个种子点通过投票模块独立生成投票，随后将投票结果进行分组聚类。为简单起见，本文根据投票结果的空间接近程度进行统一抽样与分组。具体地说，给定投票 $\{v_i=[y_i;g_i]\in R^{3+C},i=1,2,\cdots,M\}$，在三维欧氏空间中使用基于 $\{y_i\}$ 的最远点采样，对 K 个投票的子集分别进行采样得到 $\{v_{ik},k=1,\cdots,K\}$，然后找到各个 v_{ik} 的相邻投票，形成 K 簇 $c_k=\{v_i^{(k)}\mid\|v_i-v_{ik}\|\leqslant r\},k=1,\cdots,K$，从而完成聚类。

经上述投票结果分组聚类后，将由建议模块生成对象建议，然而由于投票结果本质上是一组高维点，故本文进一步采用点集学习网络以聚合投票，并生成三维目标对象建议。与传统的 Hough 投票[18]识别场景物体边界的回溯过程相比，采用深度学习的 Hough 投票过程，则允许从部分观测值得到场景物体框架边界，以及其他预测参数（如方向/类别等）。本文采用共享 PointNet 模块[19]进行投票聚合，并给出集群中的建议。具体地说，给定投票簇 $C=\{w_i,i=1,2,3,\cdots,n\}$ 及其集群中心 $w_i=[z_i;h_i]$，$z_i\in R^3$ 为投票位置，$h_i\in R^c$ 为投票特征。为使用局部投票的几何特征，本文方法利用 $z_i'=(z_i-z_j)/r$ 将投票位置转换为局部规范化坐标系，并输入投票聚合模块以生成该集群 $p(C)$ 对象建议，如(2)式所示：

$$\boldsymbol{p}(C)=\mathrm{MLP}_2\{\max_{i=1,\cdots,n}\{\mathrm{MLP}_1([z_i';h_i])\}\} \quad (2)$$

式中：建议 $\boldsymbol{p}$ 表示为一个多维向量，该向量包含对象得分、边界框参数和语义分类得分。其中，来自各个聚类的投票由 MLP_1 独立处理，然后按通道最大化汇集到单个特征向量，并传递给 MLP_2，使得来自不同投票的信息进一步组合。

3 损失函数

3.1 投票损失

在传统的 Hough 投票中，投票与局部关键点的偏移通常在预先计算码本中查找确定的投票结果，而本文使用基于深度神经网络的投票模块生成投票，该模块可以与网络其余部分联合训练并使训练过程更高效、检测结果更准确。

具体地说，给定种子点 $\{s_1,s_2,\cdots,s_M\}$（其中 $s_i=[x_i\in R^3;f_i\in R^c]$），各种子点将通过投票模块独立生成投票。投票模块通过多层感知器网络 MLP 实现，该模块具有完全连接层、ReLU 和批处理规范化。MLP 采用种子特征 f_i 并输出欧氏空间偏移量 $\Delta x_i\in R^3$ 和特征偏移 $\Delta f_i\in R^c$，由此得到从每一种子点 s_i 生成的投票 $v_i=[y_i;g_i]$，其中 $y_i=x_i+\Delta x_i$、$g_i=f_i+\Delta f_i$。通过投票模块预测得到的三维偏移量 Δx_i，则由式(3)的投票损失进行监督学习。

$$L_{\mathrm{reg}}=\sum_i\|\Delta x_i-\Delta x_i^*\|/M_{\mathrm{pos}} \quad (3)$$

其中，M_{pos} 表示对象曲面上的种子总数，Δx_i^* 为从种子位置 x_i 到其所属对象的边界框中心真实位移。

3.2 分类和回归损失

为保证目标检测网络的高效性和准确性，本文在融合采样模块中引入分类和回归损失，对网络进行监督。公式如下：

$$L=\frac{1}{N_c}\sum_i L_c(s_i,u_i)+\lambda\frac{1}{N_P}\sum_i[u_i>0]L_r \quad (4)$$

其中，N_c 表示场景中所有采样点的数目；N_p 表示场景前景点数目；s_i 和 u_i 分别表示预测得分与中心性得分；分类损失 L_c 采用交叉熵损失计算；回归损失 L_r 包括距离回归损失 L_{dist}、尺度回归损失 L_{size}、角度回归损失 L_{angle} 和角损失 L_{corner}。对于距离回归损失和尺度回归损失，采用 $Smooth-L_1$ 损失函数计算[20]；角度回归损失包括分类损失与残差预测损失，如式(5)所示：

$$L_{\mathrm{angle}}=L_c(d_c^a,t_c^a)+D(d_r^a,t_r^a) \quad (5)$$

其中，d_c^a 和 d_r^a 分别为预测的角度等级与残差，而 t_c^a 和 t_r^a 分别是其预期达到的目标值。

角损失 L_{corner} 定义为预测得到的 8 个角点与真实值之间差异，计算如下：

$$L_{\mathrm{corner}}=\sum_{m=1}^{8}\|P_m-G_m\| \quad (6)$$

其中，P_m 和 G_m 是点 m 的真实值和预测位置。

4 实验结果与分析

本文网络实现基于 Ubuntu 系统，GPU 为 RTX 2080 Ti，Python 版本 3.7，Tensorflow 版本是 1.14。

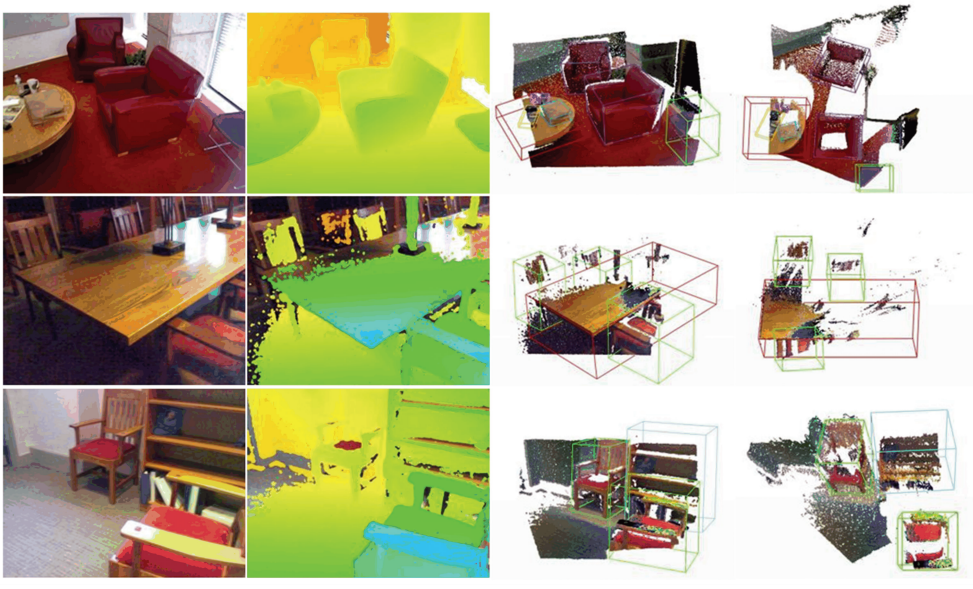
4.1 场景数据集及数据处理

实验使用 SUN RGB-D 数据集对网络进行预训练。SUN RGB-D 数据集是广泛应用于室内场景理解中的单视角数据集，分别由 4 个不同传感器捕获，其包含 10 335 张场景图像。该数据集具有大量注释信息，其中包含 146 617 个 2D 多边形标注和 58 657个具有准确朝向的 3D 边界框。为了网络在 SUN RGB-D 数据集上进行有效训练，本文使用实际摄像机参数，将场景 RGB-D 数据转换为场景点

云数据，并分别提供了网络在 10 个常见物体类别上的目标检测性能。

4.2　目标检测效果

图 4 给出了本文方法对不同场景进行目标检测的代表性示例。图 4(a)为待检测的不同场景，图 4(b)为不同场景所对应的场景深度图，图 4(c)和图 4(d)分别为利用本文方法得到的目标检测效果，并分别以不同视角显示的效果图。从中可以看出，针对给定的不同场景，虽然含有物体分布杂乱、物体遮挡、物体显示不全等缺陷，但经本文方法进行 3D 目标检测，仍然能检测得到鲁棒的检测效果。例如：图 4 第一行场景中，桌子上摆放了一些杂乱物体(杂志、水杯等)，有的被遮挡，有的过小，但利用本文方法仍可以检测出其完整的三维边界框并为其进行分类。同理，图 4 第二行场景中最左边和最里面的椅子明显显示不全，并且只残留了部分在画面中(如椅背、椅子把手等)；又如第三行场景中的书架，仅显示少量信息。对于上述问题采用本文方法，仍可以检测出其边界框，从而表明本文方法进行 3D 目标检测的鲁棒性和有效性。



(a) 场景彩色图　(b) 场景深度图　(c) 检测效果 1　(d) 检测效果 2

图 4　本文方法的目标检测效果

Fig. 4　Target detection effect of this method

4.3　消融性实验

对于消融性实验，本文分别采用不同子采样方法中的点召回率(点召回率表示采样代表点所涵盖的物体数量占物体总数的百分比)和准确率 *AP*(*AP* 表示检测网络所提出建议的准确度)进行比较，以衡量目标物体误检程度。实验结果见表 1。

表 1　不同采样方法对比

Table 1　Comparison of different sampling methods　%

	点召回率	*AP*
距离采样	92.37	70.1
特征采样	**98.43**	75.3
融合采样	98.25	**78.6**

实验表明，尽管仅具有特征采样的模型比仅具有距离采样的模型产生了更高的点召回率和更好的准确率，但其错误地将部分背景点视为前景点，导致分类精度下降，而融合采样策略的检测准确度更高，因此必须将距离采样和特征采样相结合才能保证室内场景 3D 目标检测的准确性和高效性。

表 2 给出了 SUN RGB-D 数据集上不同采样策略、不同数量的场景采样点和不同采样权重对点召回率结果的影响。表中“4096”、“1024”和“512”分别表示子集中代表点的数量。如表 2 第二列所示，在只有 512 个代表点时，距离采样的点召回率仅为 53.2%，这意味着几乎有近一半的实例未被检测到；而从特征采样方法中使用不同 λ 的结果看来，λ 为 1 时，点召回率最高。

表 2　各因素对点召回率结果的影响

Table 2　Influence of various factors on the result of point recall rate　%

代表点数	距离采样	特征采样		
		$\lambda=0$	$\lambda=1$	$\lambda=2$
4 096	99.3	99.3	**99.3**	99.3
1 024	67.1	81.7	**85.6**	83.9
512	53.2	67.5	**74.8**	72.1

4.4 不同目标检测方法比较

利用不同目标检测方法对 SUN RGB-D 数据集不同场景进行检测的实验效果见表 3。评估指标采用 *IoU* 阈值为 0.25 时,3D 目标边界框的平均检测准确度(*mAP*)。其中,DSS 方法[21]是直接基于 3D CNN 的目标检测方法,该方法提出将物体几何与 RGB 线索进行结合,对 3D 目标物体进行建议与分类;COG 方法[22]和 2D-driven[23]方法则使用场景房间布局的上下文信息以提高目标检测的性能;而 VoteNet[18]方法为基于深度霍夫投票机制的点云目标检测网络。Frustum VoxNet[24]算法是先检测 2D 对象,再对这些 2D 对象的斜锥体进行体素化,从而实现检测。类似方法还有 F-PointNet,但该方法在点云稀疏的情况下需要适当增加图像分辨率来精确建议框。值得注意的是,在数据集训练样本较多的情况下,本文检测方法相比 VoteNet 方法对目标物体的平均检测准确度提高了 2.1%,和其他方法相比均提高了 5%以上;在待检测的 10 大目标物体类别中有 4 大物体类别的检测准确度均优于其他方法,具有明显优越性能。

表 3 SUN RGB-D 数据集上的目标物体检测结果

Table 3 Target object detection results on SUN RGB-D data set

方法	类别准确度/%										平均准确度
	bed	bookshelf	chair	bathtub	desk	dresser	nightstand	sofa	table	toilet	*mAP*/ %
DSS[21]	78.8	11.9	61.2	44.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG[22]	63.7	31.8	62.2	58.3	**45.2**	15.5	27.4	51.0	51.3	70.1	47.6
2D-driven[23]	64.5	31.4	48.3	43.5	27.9	25.9	41.9	50.4	37.0	80.4	45.1
VoteNet[18]	83.0	28.8	**75.3**	**74.4**	22.0	29.8	**62.2**	**64.0**	47.3	90.1	57.7
Frustum VoxNet[24]	79.5	19.1	49.1	44.6	12.5	19.6	36.2	40.8	27.5	84.6	41.4
F-PointNet[16]	81.8	33.3	64.2	43.3	24.7	32.0	58.1	61.1	51.1	**90.9**	54.0
ours	**83.4**	**37.9**	74.5	68.1	31.4	**35.6**	60.9	62.3	53.7	90.3	**59.8**

此外,在相同数据集下本文模型的推断时间仅需 57 ms,与 PointRCNN[25]和 F-PointNet 相比快了 2~3 倍左右。在数据集 SUN RGB-D 上进行不同检测网络的方法评估时,PointRCNN 模型的推断时间为 93 ms,而 F-PointNet 模型的推断时间为 168 ms。本文模型的检测推断速度之所以加快,是因为在目标检测中放弃了特征传播层并且减少了细化模块,同时采用了融合采样策略进行采样,既缩短了模型推断时间,又保证了 3D 目标检测准确率。

5 结束语

针对 RGB-D 场景中 3D 目标检测对复杂背景的适应性较差、难以有效采样以及模型推断时间较长等问题,本文提出一种基于融合采样的轻量级 RGB-D 场景 3D 目标检测方法。该方法以场景中的 RGB-D 数据作为输入,利用二维卷积神经网络和相机投影矩阵将其转化为三维点云后,利用融合采样策略对点云中的各点进行采样,保留与物体相关的特征代表点,并将这些代表点输入到投票检测模块进行最终的 3D 目标检测。该方法在目标检测中不仅有效地对点进行了采样,还缩短了模型推断时间,使目标检测在准确性和高效性之间达到平衡,从而能够实现更可靠、更灵活的目标定位和检测。

未来将进一步考虑适用于大规模场景目标检测的轻量级 3D 目标检测网络,此外,面向复杂未知场景的开放式目标检测策略也是值得探索的一个方向。

参考文献

[1] CUI Q, SUN H, YANG F. Learning dynamic relationships for 3D human motion prediction [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 6519-6527.

[2] 黄月平, 李小锋, 卢瑞涛,等. 基于自适应标签和稀疏学习相关滤波的红外目标跟踪算法研究 [J]. 仪器仪表学报, 2022, 43(12): 199-208.

[3] 姚绍华,贺松,涂园园.基于改进欧式聚类的三维激光雷达点云目标分割方法[J].智能计算机与应用,2021,11(10):73-76.

[4] LI J, WONG H C, LO S L, et al. Multiple object detection by a deformable part-based model and an R-CNN[J]. IEEE Signal Processing Letters, 2018, 25(2): 288-292.

[5] PENG C, MA J. Semantic segmentation using stride spatial pyramid pooling and dual attention decoder[J]. Pattern Recognition, 2020, 107(1): 182-196.

[6] REN S, HE K, GIRSHICK R,et al.Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149 .

[7] RU C, WANG F, LI T,et al. Outline viewpoint feature histogram: An improved point cloud descriptor for recognition and grasping of workpieces[J].Review of Scientific Instruments, 2021, 92(2): 1095-1101.

[8] LI Y, LI Q, HUANG Q, et al. Spatiotemporal interest point detector exploiting appearance and motion-variation information [J]. Journal of Electronic Imaging, 2019, 28(3): 348-361.

[9] DIETRICH P I, BLAICHER M, REUTER I, et al. In situ 3D nanoprinting of free-form coupling elements for hybrid photonic integration[J]. Nature Photonics, 2018, 12(4):241-247.

[10] SONG S R, LICHTENBERG S P, XIAO J X. SUN RGB-D: A RGB-D scene understanding benchmark suite [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 567-576.

[11] LEE C, MOON J H. Robust lane detection and tracking for real-time applications [J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(12): 4043-4048.

[12] DOUMA A, SENGUL G, SALEM F, et al. Applying the histogram of oriented gradients to recognize arabic letters[C]//Proceedings of IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA. IEEE, 2021: 350-355.

[13] CHEN M, YU L, ZHI C, et al. Improved faster R-CNN for fabric defect detection based on Gabor filter with genetic algorithm optimization[J]. Computers in Industry, 2022, 134(1): 207-214.

[14] LI F, JIN W, FAN C, et al. PSANet: Pyramid splitting and aggregation network for 3D object detection in point cloud[J]. Sensors, 2020, 21(1):136-149.

[15] YAN D, LI G, LI X,et al. An improved faster R-CNN method to detect tailings ponds from high-resolution remote sensing images [J]. Remote Sensing, 2021, 13(11):2052-2063.

[16] QI C R, LIU W, WU C, et al. Frustum pointnets for 3D object detection from RGB-D data [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 918-927.

[17] WANG Z, JIA K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection [C]// Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 1742-1749.

[18] QI C R, LITANY O, HE K,et al. Deep hough voting for 3D object detection in point clouds[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9277-9286.

[19] 龚国栋, 李耀斌, 花向红,等. 一种探讨点云深度学习决策的PointNet++解析网络 [J]. 测绘地理信息, 2022, 47 (6): 50-54.

[20] LIU C, YU S, YU M, et al. Adaptive smooth L1 loss: A better way to regress scene texts with extreme aspect ratios [C]// Proceedings of 2021 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2021: 1-7.

[21] SONG S, XIAO J. Deep sliding shapes for amodal 3D object detection in RGB-D images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 808-816.

[22] 王亚东, 田永林, 李国强,等. 基于卷积神经网络的三维目标检测研究综述 [J]. 模式识别与人工智能, 2021, 34 (12): 1103-1119.

[23] 张鹏, 宋一凡, 宗立波, 等. 3D 目标检测进展综述 [J]. 计算机科学, 2020, 47 (4): 94-102.

[24] SHEN X, STAMOS I. Frustum VoxNet for 3D object detection from RGB-D ordepth images[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 1698-1706.

[25] SHI S, WANG X, LI H. Point RCNN: 3D object proposal generation and detection from point cloud[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.IEEE, 2019: 770-779.