物体区域信息引导下的RGB-D场景3D目标检测

缪永伟, 单丰, 杜思澄, 王金荣, 张旭东

引用本文

缪永伟, 单丰, 杜思澄, 王金荣, 张旭东. 物体区域信息引导下的RGB-D场景3D目标检测[J]. 计算机科学, 2023, 50(11A): 221200152-8.

MIAO Yongwei, SHAN Feng, DU Sicheng, WANG Jinrong, ZHANG Xudong. Object Region Guided 3D Target Detection in RGB-D Scenes [J]. Computer Science, 2023, 50(11A): 221200152-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200152

物体区域信息引导下的 RGB-D 场景 3D 目标检测

缪永伟[1,2]　单　丰[2]　杜思澄[3]　王金荣[1]　张旭东[4]
1 杭州师范大学信息科学与技术学院　杭州 311121
2 浙江理工大学计算机科学与技术学院　杭州 310018
3 伦敦国王学院自然科学学院　伦敦 N1C4BQ
4 浙江树人学院信息科技学院　杭州 310015
(ywmiao@hznu.edu.cn)

摘　要　针对室内场景 RGB-D 数据的 3D 目标检测是图形学与三维视觉中的重要问题。针对 RGB-D 场景中 3D 目标检测对复杂背景的适应性较差、目标检测中难以有效利用物体区域信息及场景点云特征信息等缺陷,基于物体区域信息引导,提出一种融合全局和局部点云特征并排除背景干扰的 3D 目标检测框架。该框架以场景 RGB-D 数据作为输入,首先提取彩色图像中待检测目标对象 2D 区域并为对象进行粗分类,再将对象区域二维边界框提升到三维斜锥体区域并转化形成点云数据;然后在斜锥体点云上利用物体区域分类信息进行特征提取,并利用特征变换与最大池聚合操作将点云全局特征和局部特征有效融合;接着利用融合特征以预测各采样点与前景背景相关程度的概率分数,依据此概率分数分割场景前景点与背景点,并通过场景背景点剔除以形成屏蔽性点云;最终在屏蔽性点云中投票产生物体中心点并借助物体区域信息提出建议及 3D 目标预测,此外,还加入了一个角点损失,对边界框精度进行优化。针对 SUN RGB-D 数据集进行网络训练,实验结果表明,与传统方法相比,所提框架的目标检测结果准确率得到有效提升,同一评估指标下的点云目标检测准确率达到 59.1%,并且在强遮挡或稀疏采样点区域下亦能够精确估计三维物体的边界框。

关键词:3D 目标检测;前景点云提取;点云分割;RGB-D;区域信息
中图法分类号　TP391

Object Region Guided 3D Target Detection in RGB-D Scenes

MIAO Yongwei[1,2],SHAN Feng[2],DU Sicheng[3],WANG Jinrong[1] and ZHANG Xudong[4]
1 School of Information Science and Technology,Hangzhou Normal University,Hangzhou 311121,China
2 School of Computer Science and Technology,Zhejiang Sci-Tech University,Hangzhou 310018,China
3 School of Natural Sciences,King's College London,London N1C4BQ
4 School of Information Science and Technology,Zhejiang Shuren University,Hangzhou 310015,China

Abstract　3D object detection for RGB-D scenes is an important issue in the literature of computer graphics and 3d vision. To overcome the poor adaptability to complex background of RGB-D scenes and it is hard to effectively combine the object region information and intrinsic feature of sampling points,a novel object region guided 3d detection framework is proposed,which can combine the global and local features of sampling points and also eliminate the background interference. Our framework takes the RGB-D data of 3Dscenes as input. First,the 2D regions of different objects in the underlying RGB image are be extracted and roughly be classified. These 2D boundary boxes of different objects can thus be lifted to their corresponding 3D oblique cone regions,and the RGB-D data located in the cone regions can also be converted to point cloud data. Furthermore,guided by the object region information,its feature of the sampling points located in each oblique cone can be extracted,and the global and local features of the sampling points are effectively fused by feature transformation and maximum pool aggregation operation. Moreover,these fused feature can be adopted to predict the probability score which reflect its correlation between each sampling point located in the foreground or background regions. According to this probability score,the sampling points of foreground and background regions can be segmented and the masked point cloud is thus generated by dividing the background sampling points from the underlying 3D scenes. Finally,the center point of the object is generated by voting in the shielded point cloud,and suggestions and 3D target prediction are made with the aid of object area information. In addition,a corner loss is added to optimize the accuracy of the bounding box. Using the public SUN RGB-D dataset,experimental results show that our proposed framework is effectively on 3D object detection. The accuracy rate of point cloud target detection under the same evaluation index reaches 59.1% if

基金项目:国家自然科学基金(61972458);浙江省自然科学基金(LZ23F020002);浙江省公益应用研究项目(LGF22F020006)
This work was supported by the National Natural Science Foundation of China(61972458),Natural Science Foundation of Zhejiang Province,China(LZ23F020002)and Zhejiang Public Welfare Application Research Project(LGF22F020006).
通信作者:张旭东(xdzhang@zjsru.edu.cn)

compared with the traditional method, and the boundary boxes of 3d objects can also be accurately estimated for different areas even with strong occlusion or sparse sampling points.

Keywords 3D object detection, Foreground point cloud extraction, Point cloud segmentation, RGB-D, Regional information

1 引言

作为图形学与三维视觉中的重要问题，目标检测得到了工业界和学术界的普遍关注，其目的是检测二维图像或三维场景中的特定目标物体或感兴趣区域(ROI)[1]，并确定目标物体位置和类别。目标检测已经成为视觉领域中众多复杂、高层视觉任务的基础，如场景理解[2]、目标跟踪[3]、场景分割[4]等。然而由于各种目标物体外观、形状和姿态各不相同以及成像过程中受光照、屏蔽等因素干扰，目标检测成为了一个具有挑战性的课题。

近年来，目标检测方法由传统基于手工特征的检测方法发展到基于深度学习的检测方法[5-6]。Li 等[7]利用 R-CNN (Region-based Convolutional Network)网络检测场景图像中的目标对象，但检测时其对候选目标区域进行缩放导致检测精度受到了一定限制，同时该网络训练较复杂。Peng 等[8]提出了基于空间金字塔池化的检测网络 SPP-Net(Spatial Pyramid Pooling Network)，该网络能够将任意大小的目标候选区域特征信息转换为固定长度的特征向量，从而有效进行多尺度训练。针对传统 R-CNN 进行改进，Ren 等[9]提出的快速 R-CNN 网络则将原始图片输入卷积神经网络中获取其特征图，再使用建议框对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 2D 目标检测之外，3D 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，即基于全局特征方法和基于局部特征方法。基于全局特征的检测方法有 Ru 等[10]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，能快速检测出目标物体，但由于其对遮挡环境的适应性差导致检测准确率不高。基于局部特征方法则首先提取场景或模板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或位姿估计。典型的关键点检测方法包括 3D Harris 检测[11]、LSP 检测[12]以及描述符提取 ROPS 方法[13]等，这些方法对复杂场景的背景适应性较差，计算耗时。

针对复杂背景干扰以及模糊样本等问题，Wang 等[14]提出基于特征点的 Anchor Free 目标检测算法的改进算法，该算法充分利用 Center Ness 参数以缓解模糊样本对网络性能的影响并提高目标检测的准确率。Chen 等[15]提出一种基于局部梯度强度图的动态规划检测前跟踪(LIG-DP-TBD)算法，该算法可以在背景复杂度高且信噪比低的环境中对弱小目标图像进行检测。然而，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前景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。

针对室内 RGB-D 场景数据的 3D 目标检测问题，基于物体区域信息引导，本文提出一种结合全局和局部点云特征并排除背景干扰的 3D 目标检测框架。该框架以场景 RGB-D 数据作为输入，首先提取彩色图像中的物体对象二维区域并为对象进行粗糙分类，再将物体对象二维边界框扩展到三维斜锥体区域并转化形成点云数据；然后在斜锥体点云上利用物体区域信息进行特征融合，将点云的全局特征和每个点的局部特征连接在一起；利用融合后的总体特征预测各采样点的概率分数，并以此对场景前景点与背景点进行分割，通过场景背景点剔除来形成用于 3D 目标预测的屏蔽性点云；最终方法在屏蔽性点云中对物体中心点进行投票，依据先前物体区域的分类信息提出建议并进行 3D 目标预测。实验表明，本文方法在目标检测中能充分利用场景物体的区域信息，有效降低搜索空间，排除背景干扰，高效、准确地从复杂场景中检测到 3D 目标物体。

具体来说，本文的主要贡献如下：

1)利用物体的区域分类信息，降低目标对象搜索空间，为区分前背景点云和目标检测做准备；同时在分割后的点云中采用投票机制进行 3D 目标检测并加入了新的角点损失，使之对三维边界框精度进行优化，提升目标检测准确率。

2)通过池化获取全局点云特征，并将点云的全局特征和每个点的局部特征进行融合，预测各采样点与前背景点云的相关程度，分割出前景点并有效排除背景噪声干扰。

3)在 SUN RGB-D 数据集上设计了对比实验以验证本文方法的有效性。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 59.1%。

2 相关工作

针对三维物体的目标检测，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[16-18]。其过程如下：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。Lee 等[16]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 算法从一个庞大的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。方向梯度直方图(Histogram of Oriented Gradient, HOG)特征[17]与支持向量机分类器相结合的方式已被广泛应用于目标检测任务，但由于其探测器计算量过大，导致检测效率低下。Li 等[18]提出一种基于多尺度可变形零件模型混合的物体检测系统，该模型将目标对象分解为各个部件并分别进行训练，其在预测过程中将合并所有部件的预测结果以实现对目标对象的检测。然而，由于传统目标检测方法需要提取候选区域信息并手动设计特征，导致其应用范围受到较大限制。

近年来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[19]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，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。为了提高小目标在复杂场景中的检测性能，研究者尝试将目标周围背景进一步融合进深度神

经网络中。借助 RGB-D 传感器，Lugo 等[20]利用灰度信息识别无纹理对象，先将获得的 RGB 图像转换为灰度图像，并分割场景背景与前景区域，其对前景经噪声去除后应用 5 种分类模型进行特征提取，最终预测目标对象类别。借助输入的场景 RGB 图像和激光雷达点云数据，Li 等[21]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，其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；再结合这 3 个输入生成特征图并利用 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[22]提出一种基于快速 R-CNN[9]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，场景环境的变化（如光线、视角等变化）通常会导致遮挡物体出现尺度不同的现象，从而导致目标检测效果的不理想。

针对点云场景的目标检测，由于三维点云数据与二维图像数据不同，点云数据并没有呈规则格点分布，许多物体的质心都位于点云外面（如桌子、椅子），因此直接利用二维场景的目标检测思想在三维点云数据上很难实现。基于此，Qi 等[23]借鉴霍夫投票机制提出了 VoteNet 网络，其利用投票机制生成目标对象建议框并完成分类回归，从而解决了点云场景目标检测的有效性问题；然而，由于生成的投票点主要聚集在目标物体中心附近，因而检测中其难以很好地反映物体的整体空间形状。Cheng 等[24]提出的 BRNet 网络则通过聚类后得到的投票中心预测目标物体的方向角和目标空间位置，再在检测目标内均匀采样确定代表点，并选出与代表点距离小于一定阈值的采样点进行聚类，最后对建议框进行二次修正得到检测结果。然而，该方法中如果需要提高目标检测效果，那么在投票过程中就需要产生足够多的代表点并采样得到目标物体内均匀密集分布的采样点，该过程的有效解决思路是过滤场景的背景点云。

利用二维目标检测策略以驱动三维目标的检测成为目前图像与点云相结合的目标检测方法中较为典型的方法，即将二维目标框通过相机变换转到三维空间视锥中，并在三维视锥中进行目标检测。Qi 提出的 F-PointNet 网络[25]首先利用 FPN（Feature Pyramid Network）检测器在二维图像上提取目标的二维检测框，再利用相机内参进行投影变换，并将上一步得到的相机平面检测框投影到三维空间以形成一个斜锥体，再对当前斜锥体内的点云数据进行语义分割以及三维边界框回归。然而，由于点云数据过于稀疏，检测中需要适当增加图像分辨率以提高候选框尺寸，此外当斜锥体中存在不止一个实例时，其会使得语义分割结果非常混乱而导致检测效果不精确。Wang 等[26]提出的 F-Convnet 则借助二维候选区域生成视锥体，该方法先将各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二维特征图并经全卷积层进行特征提取，最后进行目标物体三维框的端到端估计。该方法在检测过程中依赖于过少的前景点，容易导致误分割，从而降低了目标检测的精确性。

以场景 RGB-D 数据作为输入，本文提出一种可以将场景点云全局特征和局部特征进行有效融合的网络，其结合物体的区域信息并排除背景点云干扰，从而能在前景点云中对场景目标进行有效检测。实验表明，该方法可以有效消除复杂场景背景数据或无关聚类数据对目标检测的干扰，从而提高室内场景三维目标检测的准确性和高效性。

3 本文方法

针对 RGB-D 场景中目标检测对室内复杂背景的适应性较差，难以在大规模场景中有效地定位目标对象，且在目标检测中难以很好地利用场景区域信息及全局/局部特征等缺陷，基于物体区域信息引导，本文提出了一种结合全局和局部点云特征并排除背景干扰的 3D 目标检测网络。该网络以场景 RGB-D 数据作为输入，首先提取出场景 RGB 图像中的二维目标对象区域并为其进行分类，再将每个物体的二维边界框提升到三维斜锥体区域形成斜锥体点云；在此基础上利用斜锥体区域内物体的分类信息对全局特征和局部特征进行融合，并将全局特征和各采样点局部特征进行连接，再根据融合特征进行分割，从而将背景点云剔除并形成屏蔽性点云；最终在屏蔽性点云中对每个物体的中心点进行投票，并依据先前物体区域信息中的分类结果进行 3D 目标检测。本文网络总体框架如图 1 所示。

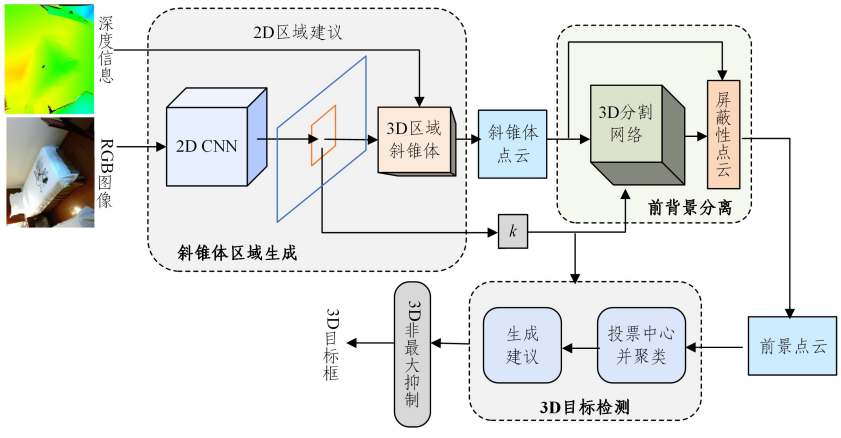


图 1 本文网络框架

Fig. 1 Framework of the proposed network

如图 1 所示，本文网络框架分为 3 个阶段：1）基于场景彩色图二维边界框生成三维斜锥体点云，并将 RGB-D 数据转换成点云数据；2）结合斜锥体点云全局特征与局部特征去除背景点云，并用于在场景点云中进行前背景分离；3）在斜锥体区域信息引导下进行场景前景点云 3D 目标检测，并在规范化坐标中细化建议以获得最终目标检测结果。具体地，网络

第一阶段以场景 RGB 图及其深度信息作为输入，再利用二维目标检测器提取出图像中的二维对象区域并对场景中物体进行分类，从而产生 k 个目标类别；同时将提取出的二维对象区域扩展到三维点云，从而使每一目标物体都有其对应的三维斜锥体，然后集合场景内所有目标物体的斜锥体点云及其携带的特征将其输入到第二阶段进行分割。在 3D 分割中，网络利用斜锥体区域内的全局特征和各采样点的局部特征将整个场景点云分割为前景点和背景点并把场景背景点剔除，最终集合所有前景点点云形成屏蔽性点云并输入到第三阶段。在网络第三阶段中对每一个目标物体中心进行投票以分别产生投票中心点，然后聚类中心点附近的采样点，同时利用第一阶段斜锥体区域内得到的 k 个目标类别分类信息生成高质量的目标建议，并经三维非最大抑制和加入的角点损失对边界框进行精度优化，最终生成 3D 目标框。在 SUN RGB-D 数据集[27]的 3D 目标检测基准上进行的大量实验表明，本文提出的 3D 目标检测框架能有效降低对象搜索空间并排除背景噪声干扰，可以准确、高效地从复杂场景中检测目标物体。

3.1 基于二维边界框的 3D 斜锥体点云生成

一般地说，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通常低于其 RGB 图像分辨率。因此，本文借鉴了传统 F-PointNet 方法[25]中提取斜锥体点云的方法：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，在三维斜锥体点云生成过程中，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斜锥体区域，使场景中各目标物体均能获取其对应的斜锥体区域，该斜锥体定义目标对象的三维搜索空间，如图 2 所示。接着收集场景中每一个斜锥体内的所有采样点以形成整个场景的斜锥体点云。然而，由于各目标对象的斜锥体朝向、大小不尽相同将导致各物体点云位置坐标范围发生较大变化，因此，本文将每一个物体斜锥体的朝向向场景的中心视图旋转，并使目标对象斜锥体中心轴与场景 RGB 图像平面正交以规范化各斜锥体，该归一化操作将有助于改进斜锥体建议生成中目标对象斜锥体点云提取的旋转不变性。

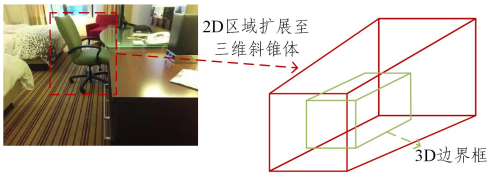


图 2 二维区域到三维斜锥体

Fig. 2 From 2D region to 3D oblique cone

3.2 结合斜锥体点云特征的背景点去除

在给定目标对象二维图像区域及其相应的三维斜锥体的前提下，常用的办法是利用二维卷积神经网络直接回归三维目标对象的位置或边界框。然而，由于场景中存在物体遮挡、背景噪声等原因，本文考虑使用一种基于 PointNet 的分割网络对目标对象进行分割[25]，因为目标对象在物理空间中为自然分离状态，所以其比在成像图像中的分割更自然、方便、准确。

如图 3 所示，以三维斜锥体点云为输入，本文方法利用权值共享的多层感知器（Multi-layer Perceptron，MLP）提取场景采样点局部特征，并与经最大池化操作得到的全局特征进行结合，再预测各采样点属于感兴趣目标对象的概率。需要注意的是，由于各斜锥体内仅包含一个感兴趣目标对象，因此其他非相关区域（如地面、墙壁等）的采样点或位于感兴趣目标对象后方的采样点与感兴趣目标对象无关，属于遮挡区域。与二维实例分割中的情况类似，由于场景中物体斜锥体的位置不同，一个斜锥体中的对象点可能会变得凌乱或遮挡另一个斜锥体中的点云，因此，本文的分割网络将会学习识别特定类别对象的几何体。

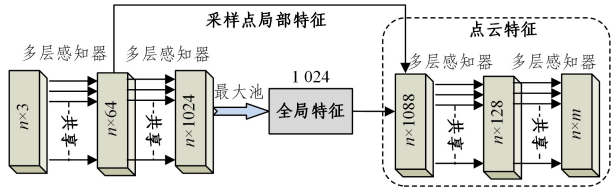


图 3 场景分割网络

Fig. 3 Scene segmentation network

对于一个 $n\times3$ 的点云，本文使用多层感知器（MLP）将每个采样点投影到 64 维空间，并使用线性全连接层扩大每个采样点的特征，以防止最大池化时采样点信息损失太多；之后再次利用 MLP 将 64 维空间映射到 1 024 维，在 1 024 维空间中由最大池化实现点云的对称性操作并得到全局特征。然而，仅利用斜锥体点云全局特征难以对各采样点进行前景背景分割，因而在得到斜锥体点云全局特征后需进一步将全局特征与各采样点局部特征进行连接，并将其反馈给各采样点特征，最终每个采样点的特征维度为 1 088 维，即每个采样点特征同时具备局部信息和全局信息，最后对每个采样点进行打分并输出 m 个分数。

若要使场景点云数据经过若干几何变换，其点云语义标记依然保持不变，通常的解决方案是在特征提取之前将所有输入集对齐到规范空间。Jaderberg 等[28]引入了空间变换的概念，通过采样和插值以对齐 2D 图像，这需要借助 GPU 上的特定模块来实现。然而，本文点云输入允许以更简单的方式实现此目标，具体地，借助一个小型变换网络（Transformation Net，亦称为 T-Net 网络[29]）以预测点云变换矩阵，并将此变换直接应用于输入模型各采样点坐标信息。该 T-Net 网络由与采样点无关的特征提取、最大池和全连接层等基本模块组成，如图 4 所示。

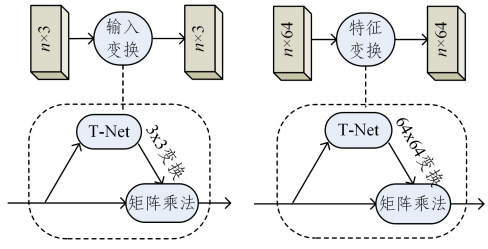


图 4 空间中的矩阵变换

Fig. 4 Matrix transformation in 3D space

针对多类别物体检测，本文进一步利用二维语义辅助进行实例分割。例如，如果感兴趣对象为场景中的一张床，则本文分割网络将在检测类似床的物体之前先使用该方法识别其语义信息以辅助目标分割。具体而言，在本文网络中将语义类别编码为一个分类向量（对于预定义的 k 类则为 k 维向量），并将该分类向量连接到中间点云特征。

场景斜锥体点云经分割后可以提取分类为感兴趣目标对象的采样点形成前景点云,同时剔除场景背景点以形成用于3D目标预测的屏蔽性点云。为提升算法的鲁棒性和平移不变性,本文方法利用点云质心将屏蔽性点云坐标转换为局部坐标,使坐标规范化。实验中,我们发现上述坐标变换和斜锥体的规范化旋转对 3D 检测结果至关重要,如表 1 所列。

表 1　归一化对目标检测的性能影响

Table 1　Influence of normalization on target detection performance

斜锥体规范化	掩码集中	T-Net 网络	准确度 AP/%
—	—	—	12.5
√	—	—	48.1
√	√	—	71.5

3.3　斜锥体区域信息指导的前景点云目标检测

在场景斜锥体前景点云中,考虑到点云的不规则性、稀疏性等特点,受 VoteNet[23] 启发,本文采用基于投票/聚类的 3D 目标检测网络进行物体检测。具体地,针对输入的前景点点云,本文方法提取出经分割网络融合后的各采样点特征并根据其特征生成投票,其投票目的是预测对象中心。因此,投票聚类将会出现在对象中心附近,然后依据先前物体区域的分类信息进行聚合并生成 3D 建议进而检测物体边界框,如图 5 所示。

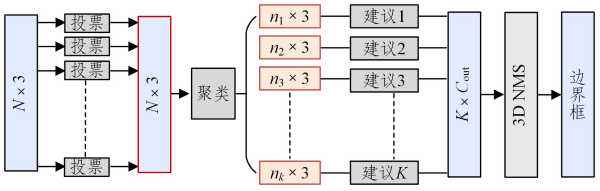


图 5　3D 目标检测网络

Fig. 5　3D target detection network

给定具有(x,y,z)坐标的 N 个采样点作为输入点云,此时这些采样点被视为种子点。每个种子点通过投票模块独立生成投票,投票为对象不同部分的上下文聚合创建规范的"集合点"。随后将投票分组聚类,这里本文选择了一种简单策略,即根据空间接近度进行统一抽样和分组,对给定投票$\{\boldsymbol{v}_i=[\boldsymbol{y}_i;\boldsymbol{g}_i]\in R^{3+C},i=1,2,\cdots,M\}$使用基于$\{\boldsymbol{y}_i\}$的最远点采样策略对 K 个投票子集采样从而得到$\{\boldsymbol{v}_{ik},k=1,\cdots,K\}$,然后找到每个 $\boldsymbol{v}_{ik}$ 的相邻投票形成 K 簇 $c_k=\{\boldsymbol{v}_i^{(k)}\parallel|\boldsymbol{v}_i-\boldsymbol{v}_{ik}\parallel\leqslant r\},k=1,\cdots,K$。该方法虽然简单,但由于这种集群技术容易集成到端到端的通道中,从而在实验中取得了较好的聚类效果。

经上述采样点聚类后将由建议模块生成对象建议。与传统的 Hough 投票[23]识别物体边界的回溯步骤相比,利用深度学习的 Hough 投票过程允许从部分观测值提出框架边界以及预测其他参数(如方向、类别等)。因而本文方法使用一个共享 PointNet 模块[30]进行投票聚合并给出集群建议。具体地,给定投票簇 $C=\{\boldsymbol{w}_i,i=1,2,3,\cdots,n\}$ 及其集群中心 $\boldsymbol{w}_i=[\boldsymbol{z}_i;\boldsymbol{h}_i],\boldsymbol{z}_i\in R^3$ 为投票位置,$\boldsymbol{h}_i\in R^c$ 为投票特征。为使用局部投票的几何特征,本文方法利用 $\boldsymbol{z}_i'=(\boldsymbol{z}_i-\boldsymbol{z}_j)/r$ 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 $\boldsymbol{p}(C)$对象建议,如式(1)所示:

$$\boldsymbol{p}(C)=MLP_2\{\max_{i=1,\cdots,n}\{MLP_1([\boldsymbol{z}_i';\boldsymbol{h}_i])\}\} \tag{1}$$

其中,建议 $\boldsymbol{p}$ 表示为一个多维向量,该向量包含对象性得分、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。

4　损失函数

4.1　投票损失

在传统 Hough 投票[23]中,投票结果与局部关键点的偏移通常利用预先计算得到的码本以查找确定。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。

给定种子点$\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_M\}$(其中 $\boldsymbol{s}_i=[\boldsymbol{x}_i;\boldsymbol{f}_i]$),$\boldsymbol{x}_i\in R^3$,$\boldsymbol{f}_i\in R^c$,各种子点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,其具有全连接层、ReLU 和批处理规范化。多层感知器采用种子特征 $\boldsymbol{f}_i$ 并输出偏移量 $\Delta\boldsymbol{x}_i\in R^3$ 和特征偏移 $\Delta\boldsymbol{f}_i\in R^c$,由此可以得到投票 $\boldsymbol{v}_i=[\boldsymbol{y}_i;\boldsymbol{g}_i]$,$\boldsymbol{y}_i=\boldsymbol{x}_i+\Delta\boldsymbol{x}_i$,$\boldsymbol{g}_i=\boldsymbol{f}_i+\Delta\boldsymbol{f}_i$,代表从每一种子点 $\boldsymbol{s}_i$ 生成的投票。预测得到的偏移量 $\Delta\boldsymbol{x}_i$ 由式(2)所示的回归损失进行监督学习:

$$L_{\text{reg}}=\sum_i\parallel\Delta\boldsymbol{x}_i-\Delta\boldsymbol{x}_i^*\parallel/M_{\text{pos}} \tag{2}$$

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

4.2　边界框优化的角点损失

虽然本文方法预测的目标物体边界框参数较完整,但学习过程中并没有针对边界框精度进行优化,即边界框中心、大小和朝向角度均需具有单独的损失项。如果目标物体边界框中心和大小得到了准确预测,但其朝向角度没有预测时将会导致最终含场景真实框的 3D IoU 将由角度误差控制。为了解决该问题,本文加入正则化损失(即角点损失)以预测目标物体边界框与真实框角点的距离损失之和,如式(3)所示:

$$L_{\text{cor}}=\sum_{i=1}^{NS}\sum_{j=1}^{NH}\delta_{ij}\min\{\sum_{k=1}^{8}\parallel\boldsymbol{P}_k^{ij}-\boldsymbol{P}_k^*\parallel,\sum_{i=1}^{8}\parallel\boldsymbol{P}_k^{ij}-\hat{\boldsymbol{P}}_k^*\parallel\} \tag{3}$$

该损失表示预测框的 8 个角点与场景真实框角点之间的距离损失。由于角点位置由边界框中心、大小和朝向角度共同确定,因此角点损失能对网络参数进行多任务训练。为计算角点损失,本文先从所有尺寸的边界框模板与朝向构造 $NS\times NH$ 个"锚"框。实验中 NS 为不同大小的锚个数,每个锚有 4 个维度并分别表示锚的置信度和长宽高残差回归;NH 为不同朝向的锚个数,每个锚有 2 个参数并分别表示锚的置信度和朝向角 θ。$\boldsymbol{P}_k^{ij}$ 表示锚框角,其中 i,j 和 k 分别表示锚大小类、朝向角度和预定义角顺序;$\boldsymbol{P}_k^*$ 为真实边界框的第 k 个角的位置矢量;$\hat{\boldsymbol{P}}_k^*$ 为真实 3D 边界框经翻转角度后得到的第 k 个角位置矢量,引入 $\hat{\boldsymbol{P}}_k^*$ 是因为实验中需要翻转朝向角度以增强数据集,其为了避免翻转朝向角度带来的预测损失而需考虑角矢量与翻转矢量之间的损失。式(3)中 δ_{ij} 为二维掩码,其表示在 $NS\times NH$ 个锚中具有正确大小与朝向角的锚框将会计算其损失,其余锚框 $\delta_{ij}=0$。

4.3　多任务损失

为保证目标检测网络的高效性和准确性,本文采用如式(4)所示的多任务损失以同时优化 3 个网络(分割网络、T-Net 网络和边界框预测网络)。

$$L=L_{\text{seg}}+L_{\text{reg}}+\lambda(L_T+L_{hc}+L_{hr}+L_{sc}+L_{sr}+\gamma L_{\text{cor}}) \tag{4}$$

其中,L_{seg}为分割网络的语义分割损失,L_{reg}为投票损失,L_T 为 T-Net 网络的质心回归损失,L_{hc} 和 L_{hr} 分别为网络模型预测

的边界框朝向角分类损失和朝向角语义分割损失，L_{sc} 和 L_{sr} 分别表示网络模型预测目标物体边界框大小的分类损失和语义分割损失，L_{cor} 则为角点损失。同时，本文在多任务损失中利用同方差不确定性特点[31]以自动调整损失函数中权重系数并提升目标物体的检测效率和准确性。

5 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。

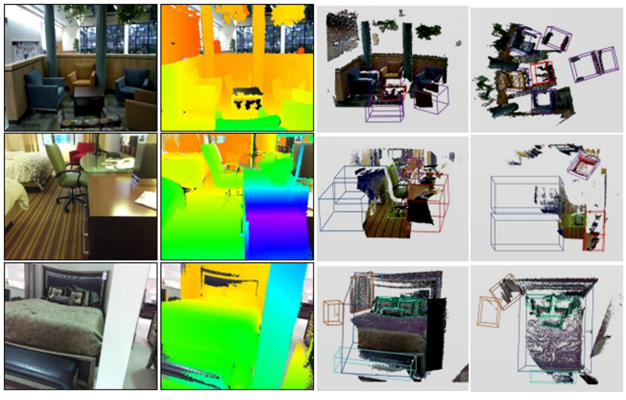
基于物体区域信息的引导，本文提出结合全局和局部点云特征并排除背景噪声干扰的 3D 目标检测网络。该方法以场景 RGB-D 数据作为输入，首先利用 2D CNN 预估出 RGB 图像中的 2D 目标对象区域并对其进行分类，再将目标对象的二维边界框提升至三维斜锥体区域以形成场景斜锥体点云；接着结合场景斜锥体点云的全局特征和各个采样点的局部特征，并利用先前斜锥体区域内的对象分类信息对场景点云进行前背景分割，从而形成仅含有前景点的屏蔽性点云，在屏蔽性点云中进行 3D 目标检测，并利用角点损失优化 3D 目标检测框，最终得到各场景物体对应的 3D 边界框。实验结果表明，与传统方法相比，本文方法能够充分利用物体的区域信息并降低目标对象的搜索空间，同时可以有效地排除背景噪声干扰，从而高效、准确地从复杂场景中检测到不同目标物体。

5.1 场景数据集及数据处理

本文方法使用 SUN RGB-D 数据集[27]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 10335 张场景图像。该数据集具有大量标注信息，包含 146617 个 2D 多边形标注和 58657 个具有准确朝向的 3D 边界框。为了使本文网络在 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，并分别提供了本文网络在 10 个常见物体类别上的目标检测性能。

5.2 本文方法的目标检测效果

图 6 给出了本文方法对不同场景进行目标检测的代表性示例。



(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 6 本文方法的目标检测效果

Fig. 6 Target detection results of the proposed method

图 6(a)为待检测的不同场景，图 6(b)为不同场景所对应的场景深度图，图 6(c)和图 6(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，给定的这些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 6 第一行场景中最靠近视点的带有花纹的沙发被明显截断，但利用本文方法仍可以检测出其完整的 3D 边界框；图 6 第二行场景中最左边的床以及图 6 第三行场景中最左边的床头柜和窗帘，虽仅显示了少量信息(如一个角)，但采用本文方法仍可以检测出其边界框。此外，图 6 第三行场景中床上的大量枕头放置凌乱、有覆盖和遮挡，但从利用本文方法检测得到的俯视视角中可以清晰看到其场景中的 7 个枕头并预测出各自的完整边界框，表明了本文方法进行 3D 目标检测的鲁棒性和有效性。

5.3 方法的消融性实验

本文方法在目标检测过程(特别是斜锥体区域形成以及场景背景点云去除等)中对场景点云进行了多次的归一化操作。为了进行消融性实验，本文给出了各归一化步骤对最终的场景 3D 目标检测效果的影响，并采用 IoU 阈值为 0.7 时的 3D 目标边界框的检测准确度(Average Precision，AP)来衡量，如表 1 所列。其中，斜锥体规范化是指使场景中每个目标对象斜锥体之间具有相似的坐标分布，而掩码集中是指使目标对象中的采样点具有规范坐标。从表 1 中可以看出，斜锥体规范化和掩码集中对最终的目标检测结果具有重要影响。此外，通过 T-Net 网络将对象点云与对象中心对齐亦能显著提高目标检测性能。

表 2 列出了本文方法不同模块对最终目标检测结果影响的消融性实验，其评估指标采用 IoU 阈值为 0.25 时 3D 目标边界框的平均检测准确度(Mean Average Precision，mAP)。从表 2 中可以看出，目标检测时不预先进行前背景点云分割或投票聚类的效果远不及聚集所有模块进行目标检测的效果。同时，点云分割模块本身对最终的目标检测效果的影响较大，与未进行点云分割相比，在斜锥体生成后加入点云分割模块最终检测结果的平均准确度提升了 10.3%。

表 2 各模块对目标检测结果的影响

Table 2 Influence of each module on target detection results

斜锥体生成	点云分割	投票聚类	mAP/%
√	×	×	43.7
√	√	×	54.0
√	√	√	59.1

5.4 不同目标检测方法的比较

表 3 给出了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，其评估指标是采用 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[32]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[33]和 2D-driven 方法[34]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 方法[23]为基于深度霍夫投票机制的点云目标检测网络。Frustum VoxNet[35]算法先检测 2D 对象，再对这些 2D 对象的斜锥体进行体素化，从而实现目标检测。类似方法还有 F-PointNet 网络[25]，但由于 F-PointNet 网络在稀疏点云情形下需增加图像分辨率以精确生成建议框并同时进行分割与检测，其对被遮挡物体及不规则物体(如梳妆台、椅子等)的检测精度产生了较大影响，因而本文采用投票方式检测目标对象，

并将第一阶段的 2D 分类结果与投票产生的 3D 目标建议框进行匹配,从而可以在不增加图像分辨率的前提下生成精确建议框,其有效提升了目标检测结果的准确度。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 1.4%,和其他方法相比均提高了 5%以上。此外,本文方法在待检测的十大目标物体类别中有三大物体类别的检测准确度均优于其他方法,具有明显优越的性能。

表 3 SUN RGB-D 数据集上的目标检测结果

Table 3 Object detection results on SUN RGB-D dataset

(单位:%)

方法	类别准确度										平均准确度
	bed	bookshelf	chair	bathtub	desk	dresser	nightstand	sofa	table	toilet	mAP
DSS[32]	78.8	11.9	61.2	44.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG[33]	63.7	31.8	62.2	58.3	**45.2**	15.5	27.4	51.0	51.3	70.1	47.6
2D-driven[34]	64.5	31.4	48.3	43.5	27.9	25.9	41.9	50.4	37.0	80.4	45.1
VoteNet[23]	**83.0**	28.8	**75.3**	**74.4**	22.0	29.8	**62.2**	**64.0**	47.3	90.1	57.7
Frustum VoxNet[35]	79.5	19.1	49.1	44.6	12.5	19.6	36.2	40.8	27.5	84.6	41.4
F-PointNet[25]	81.8	33.3	64.2	43.3	24.7	32.0	58.1	61.1	51.1	**90.9**	54.0
ours	82.3	**36.4**	74.8	64.9	27.9	**35.7**	61.8	62.7	**53.5**	90.6	**59.1**

表 4 列出了在相同数据集下采用不同检测网络模型和目标检测时间的比较结果。如表 4 所列,在数据集 SUN RGB-D 上进行不同检测网络的方法评估时,本文方法的检测处理时间比采用 H3DNet[36] 方法的检测效率快 2.5 倍,且网络模型参数比 H3DNet 要少。这是由于本文方法在目标检测中充分利用了待检测物体的区域信息并有效结合了场景点云的全局特征和局部特征,省去了不必要的参数使用,同时在检测中将遮挡物体分离出去,排除了背景点云的干扰,实验验证了本文方法的有效性。

表 4 不同模型的效率统计

Table 4 Efficiency statistics of different methods

方法	模型参数大小/MB	检测时间/s
VoteNet[23]	11.2	0.076
H3DNet[36]	56.0	0.241
Ours	40.7	0.092

结束语 针对 RGB-D 场景中 3D 目标检测,以物体区域信息引导,本文提出了一种结合全局和局部点云特征并排除背景噪声干扰的 3D 目标检测方法。该方法首先从输入的场景 RGB 图像中预测目标物体的二维边界框,并将二维对象区域扩展到三维斜锥体区域;然后利用生成的斜锥体区域的对象特征信息,将点云全局和局部特征信息进行融合并剔除无关的场景背景点,从而生成仅含有前景点的屏蔽性点云;在屏蔽性点云中再利用物体斜锥体区域中产生的分类信息进行目标检测。该方法在目标检测中不仅较好地利用了场景点云全局与局部特征,排除了复杂场景中背景点云的干扰,同时降低了目标物体检测的搜索空间,从而能够实现更可靠、更灵活的目标对象定位与边界框检测。

未来将进一步考虑使用目标检测过程更加便捷且规模较小的前景背景分离方法,同时考虑适用于大规模场景目标检测的轻量级网络,此外面向复杂未知场景的增量式目标检测策略也是值得探索的一个方向。

参 考 文 献

[1] ARNOLD E, AL-JARRAH O Y, DIANATI M, et al. A survey on 3d object detection methods for autonomous driving applications [J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3782-3795.

[2] 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.

[3] CHENG B, SHENG L, SHI S, et al. Back-tracing representative points for voting-based 3D object detection in point clouds[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8963-8972.

[4] YANG W K, YUAN X P, CHEN X F, et al. Multi feature segmentation of 3D lidar point cloud space [J]. Computer Science, 2022, 49(8): 143-149.

[5] DENG Z, LATECKI L J. Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in RGB-depth images [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5762-5770.

[6] HOU J, DAI A, NIEβNER M. 3D-SIS: 3D semantic instance segmentation of RGB-D scans[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4416-4425.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] AO S, GUO Y, GU S, et al. SGHs for 3D local surface descrip-

tion [J]. IET Computer Vision, 2020, 14(4): 154-161.
[14] WANG C, LIU Y J, XIE Q, et al. Anchor free target detection algorithm based on soft label and sample weight optimization [J]. Computer Science, 2022, 49(8): 157-164.
[15] CHEN Y, HAO Y G, WANG H Y, et al. A dynamic programming pre detection tracking algorithm based on local gradient intensity map [J]. Computer Science, 2022, 49(8): 150-156.
[16] 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.
[17] DOUMA A, SENGUL G, SALEM F, et al. Applying the histogram of oriented gradients to recognize arabic letters[C]// IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA. IEEE, 2021: 350-355.
[18] LI G, YU Y. Contrast-oriented deep neural networks for salient object detection [J]. IEEE Transactions on Neural Networks & Learning Systems, 2018, 29(12): 6038-6051.
[19] 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.
[20] LUGO G, HAJARI N, REDDY A, et al. Textureless object recognition using an RGB-D sensor[C]// Proceedings of International Conference on Smart Multimedia. Cham: Springer, 2019: 13-27.
[21] 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.
[22] 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.
[23] 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.
[24] CHENG B, SHENG L, SHI S, et al. Back-tracing representative points for voting-based 3d object detection in point clouds[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8963-8972.
[25] QI C R, LIU W, WU C, et al. Frustum pointnets for 3d object detection from rgbd data[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Rrecognition. 2018: 918-927.
[26] WANG Z, JIA K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection [C]// 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 1742-1749.
[27] 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.
[28] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[C]// Proceedings of Advances in Neural Information Processing Systems. 2015: 2017-2025.
[29] KOSSAIFI J, BULAT A, TZIMIROPOULOS G, et al. T-Net: Parametrizing fully convolutional nets with a single high-order tensor[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 7822-7831.
[30] QI C R, SU H, MO K, et al. PointNet: deep learning on point set for 3d classification and segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 77-85.
[31] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7482-7491.
[32] 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.
[33] REN Z, SUDDERTH E B. Three-dimensional object detection and layout prediction using clouds of oriented gradients[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1525-1533.
[34] LAHOUD J, GHANEM B. 2d-driven 3d object detection in rgb-d images[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4622-4630.
[35] SHEN X, STAMOS I. Frustum VoxNet for 3D object detection from RGB-D or Depth images[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 1698-1706.
[36] ZHANG Z, SUN B, YANG H, et al. H3DNet: 3d object detection using hybrid geometric primitives[C]// Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 311-329.

MIAO Yongwei, born in 1971, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include computer graphics, computer vision and deep learning.

ZHANG Xudong, born in 1982, Ph.D, associate professor. His main research interests include computer graphics and computer vision.