计算机科学
COMPUTER SCIENCE

**三维城市场景中的小物体检测**

陈佳舟，赵熠波，徐阳辉，马骥，金灵枫，秦绪佳

引用本文

陈佳舟，赵熠波，徐阳辉，马骥，金灵枫，秦绪佳. 三维城市场景中的小物体检测[J]. 计算机科学, 2022, 49(6): 238-244.

CHEN Jia-zhou, ZHAO Yi-bo, XU Yang-hui, MA Ji, JIN Ling-feng, QIN Xu-jia. Small Object Detection in 3D Urban Scenes[J]. Computer Science, 2022, 49(6): 238-244.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400174

# 三维城市场景中的小物体检测

陈佳舟[1]　赵熠波[1]　徐阳辉[1]　马　骥[1]　金灵枫[1,2]　秦绪佳[1]
1 浙江工业大学计算机科学与技术学院　杭州 310012
2 东南数字经济发展研究院数字空间技术研发中心　浙江 衢州 324000
(cjz@zjut.edu.cn)

**摘　要**　三维目标检测是三维城市场景语义分析的关键环节,但是现有的目标检测方法主要关注诸如建筑、道路等较大的物体,对路灯、井盖等小物体的检测误差较大。为此,提出了一种多视图的三维城市场景小物体检测方法,在倾斜摄影的基础上结合精准三维定位方法,提高了三维城市场景中小物体检测的精度。首先在无人机原片上利用深度学习方法检测城市小物体,然后将这些图像检测结果反投影到三维城市模型上,并通过聚类得到最终的三维检测结果。实验结果表明,所提方法能够在倾斜摄影测量得到的大规模三维城市模型上自动检测井盖、窗户等城市小物体,不受视线遮挡的影响,相对于正射图上的物体检测具有较高的准确性和稳定性。

**关键词**:三维城市模型;多视角;小物体;目标检测;聚类
**中图法分类号**　TP391

## Small Object Detection in 3D Urban Scenes

CHEN Jia-zhou[1],ZHAO Yi-bo[1],XU Yang-hui[1],MA Ji[1],JIN Ling-feng[1,2] and QIN Xu-jia[1]
1 College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310012,China
2 Digital Space Technology R&D Center,Southeast Digital Economic Development Institute,Quzhou,Zhejiang 324000,China

**Abstract**　3D object detection is the core of semantic analysis in 3D urban scenes,but the existing object detection methods mainly focus on large objects such as buildings and roads,while the detection accuracy of these methods for small objects such as street lamps and manhole covers is low. For this sake,a multi-view small object detection method for 3D urban scenes is proposed. It combines the oblique photogrammetry and 3D object localization,to improve the detection accuracy of small objects. Firstly,small objects are detected in the UAV images using a deep neural network. Then,detection results are back projected onto the three-dimensional urban model. Finally,the 3D detection results are obtained by clustering these 3D objects obtained by back projection. Experimental results show that the proposed method can automatically detect small objects such as manhole covers and windows on the large-scale 3D urban model reconstructedby oblique photogrammetry,it is free of spatial occlusion,and has high accuracy and stability compared with object detection on orthophoto maps.

**Keywords**　3D urban model,Multi-view,Small objects,Object detection,Clustering

## 1　引言

随着城市建设的不断推进,人们对如何高效、方便地进行城市规划进行了大量研究。三维城市模型作为一种能在计算机上展示并交互的工具,被越来越多的城市工作者使用。三维城市模型作为智慧城市技术[1],极大地改善了人们的生活质量,其未来的发展在一定程度上也依赖于三维城市模型的优化。根据现有的三维城市建模方法[2],我们可以轻松获取高精度的三维城市模型,例如,倾斜摄影测量技术利用图形学中的 Structure of Motion(SfM)和 Multi-View Stereo(MVS)方法[3-5],通过多组无人机航拍的照片进行高精度的三维城市模型重建。该技术作为一种高性能、大尺度的三维建模技术已经被广泛使用,极大地促进了智慧城市技术的发展。

尽管依照现有的技术来获取高精度的三维城市模型并不困难,当前相对成熟的建模方法对建筑、道路等有着较好的建模效果,但是以上技术对于一些相对较小的物体,如窨井盖、路灯、消防栓、门窗等的建模效果欠佳。由于建筑遮挡、模型拼接等原因,小物体三维建模过程中常常会出现信息丢失、位置不符等情况。与此同时,以上较小的建模对象在测绘、交通、消防等领域又有着极高的利用价值,是智慧城市技术应用

到稿日期:2021-04-17　返修日期:2021-08-09
基金项目:浙江省文物科技保护项目(2020014);国家自然科学基金(61902350);衢州市科技计划项目(2019K38)
This work was supported by the Science and Technology Protection Project of Cultural Relics in Zhejiang Province(2020014),National Natural Science Foundation of China(61902350) and Science and Technology Project of Quzhou(2019K38).
通信作者:马骥(maji@zjut.edu.cn)

过程中不可或缺的一部分。因此，亟需一种自动的、更为完善的方法对小物体进行检测和定位。

三维场景中的目标检测是三维城市场景语义分析的关键环节，人工智能技术的快速发展显著提高了三维目标检测的精度和效率。例如：基于点云的检测方法可以从三维城市点云模型中完成目标检测；而基于体素的检测方法则将三维城市模型表达为空间均匀剖分的体素，以便卷积等深度学习操作，从而达到三维目标检测的目的。但是，以上技术大多适用于建筑、道路等大面积物体的目标检测，对于小物体的检测效果不佳。同时，在模型上的目标检测并不能解决建模过程中的小物体丢失和位置不符的问题。

二维无人机航拍通过对一个城市场景进行多视角、多距离的拍摄，来得到全方位的城市场景信息，城市场景中几乎所有的小物体都能在无人机航拍图片中被捕获。基于以上发现，本文提出了一种应用于城市场景的二维无人机航拍检测小物体目标的方法，以进一步优化三维城市场景的小物体检测。同时，借助相机标定技术，利用二维航拍图片的检测结果完成目标物体在三维模型中的定位，最终将不同图片中检测到的所有小物体进行聚类（以下统称聚类），把多角度、多距离照片中检测到的同一小物体在空间中进行整合，以此完成三维城市场景中小物体的精确目标检测。

## 2 相关工作

现有的目标检测方法较多，大体上可以分为二维目标检测和三维目标检测两大类。

### 2.1 二维目标检测

在二维图像的目标检测中，基于 CNN 神经网络的检测算法已被广泛应用，并衍生出 R-CNN 神经网络、YOLO 神经网络、SSD 神经网络等诸多算法。

Girshick 等于 2014 年提出了 R-CNN 神经网络[6]，用于目标检测和语义分割。Girshick 于 2015 年提出了 Fast R-CNN 神经网络[7]，Fast R-CNN 神经网络在计算速度上比 R-CNN 神经网络提高了 9 倍。随后，Ren 等于 2015 年提出了 Faster R-CNN 神经网络[8]，该神经网络在 Fast R-CNN 的基础上融入了 RPN 候选框生成算法，使得目标检测速度得到进一步提升。He 等于 2017 年提出了 Mask R-CNN[9] 神经网络，该算法作为 Faster R-CNN 的拓展，尽管在速度上相比 Faster R-CNN 有所减慢，但其在检测目标的同时能对候选框使用 FCN 进行语义分割，同时实现了 mask 与 class 的解耦，极大地提高了检测精度。

由 Redmon 等于 2016 年提出的一种名为 YOLO[10] 的目标检测方法，能够将训练和检测整合在一个单独网络中进行。该方法借助一个单独的 end-to-end 网络，输入图像后经过一次推断就能完成检测。同年，Redmon 等提出了名为 YOLO9000[11] 的改进算法，因其能实现 9 000 多种物体的实时检测而取名为 YOLO9000，该算法在保持原有检测速度的基础上提高了检测精度。2018 年，Redmon 等又提出了名为 YOLOv3[12] 的检测方法，该方法调整了网络结构，利用多尺度特征进行对象检测，在保持速度优势的前提下，提升了预测精度，尤其加强了针对小物体的识别能力。2015 年，Liu 等提出了名为 SSD[13] 的检测方法，在准确率相同的情况下，其检测速度比 YOLOv3 检测方法的速度高出 3 倍。

### 2.2 三维目标检测

近年来，三维目标检测方法发展迅速。依据射影几何学，仅仅依赖一张图像无法准确恢复物体的三维位置，即使能得到目标物体的相对位置信息，也无法获得真实尺寸。因此，正确检测目标的三维位置需要借助多个相机、运动相机组成的立体视觉系统，或者由深度相机、雷达等传感器得到的 3D 点云数据。对于特定类型的目标，基于机器学习的方法使得通过单目相机进行物体三维模型重建成为可能，即基于单目图像的三维模型重建技术[14]。

近年来，三维目标检测的应用正朝着纵深方向发展。Mousavian 等于 2017 年提出了一种结合深度神经网络回归学习、几何约束的三维目标（主要针对车辆）检测和三维位置估计算法。同年，Tekin 等提出了一种基于 YOLO 的三维目标检测方法 YOLO-6D[15]。2019 年，Li 等提出了一种运用于无人驾驶技术的三维物体检测技术 GS3D[16]。

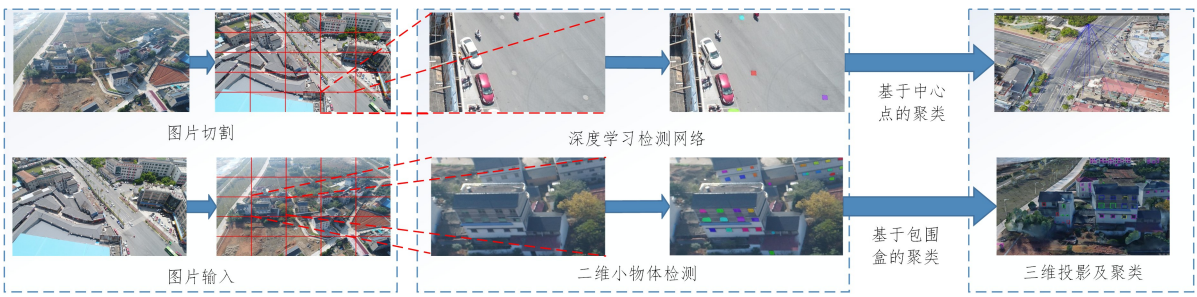


图 1 本文方法概要图

Fig. 1 Overview of ourmethod

同时，基于双（多）视图图像的三维目标检测方法也颇为丰富。Mousavian 等于 2017 年提出了一种基于 3DOP 的三维目标检测方法[17]。Li 等于 2019 年提出的基于立体视觉 R-CNN 的三维目标检测算法[18] 是在扩展 Faster R-CNN 网络框架的基础上，进行双目立体视觉检测的优化算法。

基于点云的目标检测方法同样被大量运用于真实场景中。Qi 等于 2017 年提出 Frustum-PointNet 方法[19]，将 PointNet 的应用拓展到了 3D 目标检测上。Shi 等于 2019 年提出的 PointRCNN 算法[20] 实现了纯粹使用点云数据完成三维目标检测任务。由香港中文大学联合商汤科技发表的

Part-A2 Net 方法[21]和海康威视的 Voxel-FPN 方法[22]都在点云的目标检测方法中有较高的精确度。

而对于三维模型中一些较小的物体，Liu 等于 2020 年提出了一种 IPG-NET 算法[23]。该算法在特征金字塔网络(Feature Pyramid Network，FPN)的基础上构建而成，利用一种名为图像金字塔的引导网络来构建新的神经网络。其主要内容包括图像金字塔引导子网、基于 ResNet 的骨干网络和融合模块。图像金字塔引导子网络从图像金字塔中接收一组图像，并提取图像金字塔特征进行融合。子网络的功能是提取浅层特征以提供空间信息和详细信息。图像金字塔特征用于引导骨干网络保持空间信息和小物体的特征。该算法使用融合模块执行引导，融合骨干网中的深层特征和图像金字塔引导子网中的浅层特征。融合模块的思想是将两种类型的特征进行转换，然后将它们组合在一起，以实现目标检测(尤其是小目标检测)的增强效果。

基于以上目标检测方法，本文将尝试通过物理切割等方法将一张较大的航拍图形切分为几张较小的航拍图。利用航拍图的目标检测结果进行三维空间映射及聚类，以得到最终的小物体的三维检测还原，在目标检测效率与精度方面都有所提升。

## 3 算法设计

本文主要尝试借助无人机拍摄的照片来完成对一些小物体在三维场景中的定位以及优化，具体包括：无人机照片识别的小物体、小物体在三维模型中的定位以及小物体的三维聚类。

### 3.1 目标检测

针对小物体的检测，需要寻找一种相对准确的方法来识别无人机照片中的小物体模型。由于检测的目标是一些小型的物体，其在图片中的像素占比相对较小，因此检测的精度要求比建筑等大目标检测更高。通过比较 Faster R-CNN，Mask R-CNN，YOLOv2，YOLOv3，SSD 等目标检测方法发现，SSD 和 YOLO 的检测方法虽然在速度上相比 Mask R-CNN 明显的加快，但是检测精度明显不如 Mask R-CNN。因此，本文最终决定以 Mask R-CNN 为目标检测方法来进行无人机照片中小物体的训练检测及识别工作。

#### 3.1.1 训练集标注与模型训练

本文将 LABELME 作为标注工具对无人机照片中的小物体进行标注。对于城市中的小物体，本文在标注过程中统一使用四边形，即用一个尽量小的四边形贴合图片中的小物体，以此得到小物体的训练标注。同时，为了快速增加训练集的数量，本文利用测绘公司现有的 CAD 图纸，来完成一些在水平面的特定物体标注。通过转化这些图纸文件并与场景正射图结合，可以直接生成有标记的图片，在扩充数据集的同时缩短了标注的时间。

Mask R-CNN 检测法在处理过程中会将图片缩放至 1 000 * 1 000 的像素，而输入的训练图片和检测图片大多超过了该像素。为解决因缩放而导致的特征丢失问题，本文对图片和标注文件进行分割，得到了长宽都小于 1 000 像素的图片和标注文件，使其能够更好地被训练。同时，对图像进行分割也能增加目标在照片中的面积占比，提高了检测的精度。

无人机照片中存在众多影响小物体检测的干扰项，如马路边的路灯、屋顶的中央空调室外机、井盖，以上小物体面积较小、形状相似，很容易被误检测。为了提高检测精度，可以通过训练和检测相结合的方法来实现去除干扰项的目的。即在训练过程中，通过完整标记所有的小物体，来达到减少误检测的目的。

#### 3.1.2 小物体目标检测

在 Mask R-CNN 的检测过程中，目标物体在所检照片中的占比直接影响了检测结果的精度。为了保持结果的高检出率，需要对检测图片进行切割，以保持目标物体在图片中的占比与训练图片的占比相近。为此，本文首先随机选取其中一张检测图片，对其进行手工标注，计算窗户在被检测图片中所占的像素比例，再计算训练图片中窗户所占的像素比例，以决定图片切割的大小。同时，为了防止出现因为切割导致一个物体被切成多块而无法识别的情况，本文在切割过程中将每张图片在正常切割的基础上，向上、下、左、右分别延展 30 个像素，以保证每个小物体至少在一张分割图片上是完整的。然后，利用训练完的 Mask R-CNN 神经网络对切割后的图片进行二维目标检测。为了方便后续的三维模型投影，本文将切割后图片的二维检测结果在原图上进行汇总。由于本文提出的切割方法在切割过程中将图片外扩，因此可能出现一个物体被多次检测的问题。为解决这一问题，本文利用检测结果 mask 的最小包围框来计算中心点，如果中心点距离边界小于 30 像素，则该检测结果在另一张图片中会因重复检测而被去除。同时，Mask R-CNN 神经网络能检测出物体所占的具体像素点，生成物体在二维图像上对应的 mask，提高之后的三维空间投影的精细度。

### 3.2 三维模型投影

#### 3.2.1 图像的三维高度信息计算

首先利用三维模型生成软件 Smart3D 将倾斜摄影照片组生成为三维模型以及照片相机姿态信息 XML 文件，再根据倾斜摄影照片名称，从 XML 文件中读取相机中心位置坐标 $C$、相机旋转的 3×3 旋转矩阵 $\boldsymbol{R}$、3×3 的相机轴矩阵 $\boldsymbol{O}$，并利用投影公式进行投影转换。

$$\boldsymbol{F}\cdot D(\Pi(\boldsymbol{O}\cdot\boldsymbol{R}(X-C)))=x-x_0 \tag{1}$$

其中，$X$ 为投影后的三维坐标，$x$ 为输入像素点的二维坐标，$x_0$ 为照片的主点，$\boldsymbol{F}$ 为相机矩阵。进一步进行公示推演：

$$\boldsymbol{F}=\begin{bmatrix} f & s \\ 0 & pf \end{bmatrix} \tag{2}$$

其中，$f$ 为相机的焦距，$s$ 为倾斜参数，$p$ 为像素比率，$\Pi$ 为透视投影函数，函数的定义如下：

$$\Pi(u,v,w)=\left(\frac{u}{w},\frac{v}{w}\right) \tag{3}$$

其中，$u,v,w$ 分别为函数中的 3 个参数，$D$ 为畸变方程，方程的定义为：

$$\begin{pmatrix}(1+k_1r^2+k_2r^4+k_3r^6)u+2P_2uv+P_1(r^2+2u^2)\\(1+k_1r^2+k_2r^4+k_3r^6)v+2P_1uv+P_2(r^2+2u^2)\end{pmatrix}$$

其中，$r^2=u^2+v^2$，$u$，$v$ 分别为方程中的两个参数，$k_1$，$k_2$，$k_3$，$P_1$，$P_2$ 为 XML 文件中记录的相机畸变参数。

根据投影结果，可以得到二维像素在三维模型中的一个三维坐标点，连接相机中心与该交点得到一条射线，该射线与三维模型相交得到实际的三维坐标点$(x,y,z)$。依次重复计算每一像素的三维坐标，保存得到图片的三维高度信息。

#### 3.2.2 检测结果的三维投影

利用 3.2.1 节中得到的三维高度信息，逐像素查找检测结果的三维坐标信息，根据三维坐标信息结合原始照片进行三维投影，以此得到目标检测方法所得 mask 在三维空间中的位置。至此，求得各个检测到的物体在三维空间中的坐标位置。

### 3.3 三维检测结果聚类

在 3.2 节中得出了航拍图片中的小物体所对应的空间位置，但是同一物体在多张照片中都被检测到时，则需要对该物体进行聚类操作。通常情况下，只有目标物体多次检测的标签一致且距离非常接近时，才会被认定为检测到同一物体。

下文针对不同的小物体提出了两种聚类方法，一种是基于中心点的聚类方法，适用于一些大小固定的物体检测定位；另一种则是基于包围盒的聚类方法，其适用于需要精确还原目标在三维检测中的几何特性的物体。

#### 3.3.1 基于中心点聚类

对于一些城市中的小物体，我们只需要知道其空间位置，而无需在意其实际模型形状，如路灯等；对于一些形状相对固定的物体，如城市井盖等，我们只需要知道这个物体的中心点在空间中的位置即可完成聚类。针对以上几类物体，在检测过程中得到该物体在航拍图片上的 mask，即可以利用求得的 mask 计算该 mask 的最小外界包围框。利用包围框的几何特性，本文将包围框的中心点作为井盖在二维图像中的位置表示。然后，将此中心点映射回三维模型，并得到其在三维模型中的空间坐标，以此作为该小物体在三维空间中的位置信息。

记所有检测得到的小物体的中心点的空间坐标为集合 $P=\{K_1,K_2,\cdots,K_n\}$，从 $i=0$ 开始遍历集合 $P$，对于任意的 $K_i$，当存在 $j<i$ 使得 $K_i$ 与 $K_j$ 的距离小于某一值时，则认为 $i$ 和 $j$ 两个检测结果为同一小物体，归为一类。通过以上步骤可以得到所有小物体的检测结果，对于被不同图像多次检测的同一小物体，求该小物体的所有检测结果的 $XYZ$ 坐标的平均值，并将其作为该井盖的空间坐标信息，由此完成基于中心点的小物体检测聚类。

#### 3.3.2 基于三维矩形包围盒聚类

本文通过在 3.2.1 节中得到的空间三维投影信息对二维图像上检测到的一个 mask 中的所有像素点进行空间三维投影计算，以得到所有属于该 mask 中所有的像素点的空间坐标信息。得到 mask 中所有像素的三维坐标信息后，求此 mask 在空间 $XY$ 平面上的 OBB 有向包围盒，并将 OBB 包围盒按最大最小 $Z$ 值拉伸得到三维最小空间包围盒。

本文对目标空间包围盒进行聚类，以删除重复的目标。对于不同角度拍摄到的同一个小物体，首先判断这两个包围盒在各自平行于地面的平面是否有相交部分，以此定位其在 $XY$ 平面中的位置，再对有相交面积的两个包围盒求 $Z$ 方向上的最大最小值范围是否有重合部分，如果在 $Z$ 方向上依然有重合部分，则确定为同一个小物体，对该小物体的计次加一，最终得到检测到的目标物以及每个目标物被检测到的次数。二维航拍图像对同一物体进行了多视角的拍摄，在其中一个角度中，某些物体可能被误检测为目标物体，本文将只检测到一次的目标物认定为单一视角的误检测结果。本文通过可视化的结果表示目标物体在二维航拍图像中被检测的次数，如图 2 所示，其中红色框表示该小物体只在一张二维航拍图片中被检测到，结合三维场景不难发现红色框的物体多为误检测结果，本文将红色的检测结果作为误检结果去除；绿色框表示在两张图像中被检测到，蓝色框表示在 3 张图像中被检测到，黄色框表示在 4 张图像中被检测到，紫色框表示该小物体在 5 张及以上图像中被检测到。

图 2 不同检测次数可视化(电子版为彩图)

Fig. 2 Visualization based on the detection number from different views

同时需要选择一个最接近实际模型的包围盒作为最终的聚类结果。对于某一个三维空间中的真实目标 $K$，记聚类结果中目标 $K$ 包含的包围盒集合为$\{B_1,B_2,\cdots,B_n\}$，其中 $n$ 表示包围盒的数量。对于集合中的每一个包围盒 $B_i$，计算其他包围盒的重叠面积之和 $\sum_{j=0,j\neq i}^{n} Area(B_i,B_j)$，记为 $A_i$。最终得到一个集合$\{A_1,A_2,\cdots,A_n\}$，求得集合中的最大值 $A_j$，将包围盒 $B_j$ 作为该真实目标 $K$ 的包围盒。

## 4 应用案例与实验结果

分别以井盖和窗户作为两种聚类方法的典型代表对本文方法进行实验。本文方法在 3.30 GHz Intel(R)Core(TM)PC 机上利用 python 实现。本文采用的航拍图像分辨率为 5 472 * 3 648 像素。由于本文是针对小物体进行研究，其对分辨率的敏感度更高，因此本文在实验部分对不同分辨率的图片进行检测来论证分辨率对实验结果的影响。

### 4.1 城市井盖的三维检测结果

城市井盖的铺设在城市建设中是不可或缺的一环，其铺设位置和密度直接影响了居民的日常生活。本文以城市井盖模型为例，完成其在空间中的三维目标检测。同时，为提高检测的精确度，本文在被检测的目标图片中选取一张，计算检测目标在图片中的像素占比，在此基础上对二维图像进行切割，使其与训练图片中的检测目标在图片中的像素占比保持一致，随后利用训练后的模型对其进行目标检测，找出井盖在二维航拍图片上的 mask 位置。利用求得井盖的最小外包围框，即可轻松得到此井盖在二维图像上的中心点的位置，如

图 3 中红点的位置即为井盖在图像上的中心点位置。同时利用已知的相机参数通过图 4 所示的方法将二维图像中的井盖投影到三维城市模型中，图 5 给出了最终在三维场景中的结果。城市井盖通常情况下都在路面的固定位置上，井盖的海拔高度可以被看作与地面在同一水平面上，本文通过高度信息的校验，将井盖检测结果的高度与地面高度进行比较，当井盖的高度与地面高度相差超过一定距离 $h$ 时，则认为该检测结果不在地面上，将其视为错误检测结果删去，以达到去除部分干扰项的目的。本文考虑到井盖有可能出现在马路边的人行道上等干扰因素，根据国内道路施工规范，人行道和马路的落差在 0.2 m 以内。因此，本文将 $h$ 设置为 0.2 m，这样既不会将人行道的井盖漏检，也能很好地排除非道路上的目标物体的误检结果。针对 $h$ 参数的对比实验将在 4.3 节中进行详细介绍。

图 3 井盖的二维检测结果(电子版为彩图)

Fig. 3 2D detection result of manhole covers

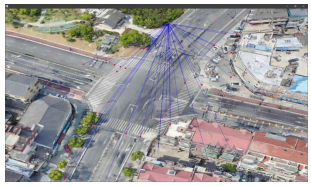

图 4 井盖的三维映射

Fig. 4 3D-projection of manhole covers

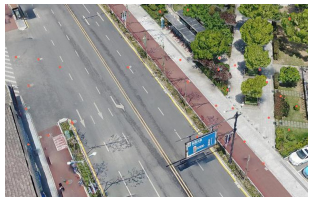

图 5 井盖的聚类结果

Fig. 5 Clustering result of manhole covers

### 4.2 窗户的三维检测结果

窗户设计是城市建设中的一个重要元素，窗户无法单独通过中心点进行检测，并且其形状相对不固定。因此，为了完成窗户的三维检测，需要对窗户在空间中的具体三维形状进行重建。为了达成这一目标，本文利用 Mask R-CNN 神经网络进行训练并完成二维图片中窗户的检测工作。如图 6 所示，在得到窗户在二维图像上的 mask 后，对 mask 中的所有像素点进行三维投影，以计算出所有像素点在三维空间中的空间坐标位置。然后，在空间中对离散的三维空间点进行包围盒的求取工作，图 7 中框线中的点即为其模型上的点云。

图 6 窗户的二维检测结果

Fig. 6 2D detection result of windows

图 7 计算最小包围盒

Fig. 7 Computation of the minimum bounding boxes

本文通过 3.3.2 节中提及的方法，利用两种方式去除了大部分的误检项。窗户玻璃只有在墙面上的两面的面积较大，而其他面的面积基本可以忽略，利用这个特性计算出包围盒所有的 6 个面的面积然后进行比较，当一个包围盒在水平平面的面积比其他几面的面积都大时，则表示该包围盒所表示的并不是窗户。

去除干扰项后，对所有的窗户进行聚类，得到最后的检测结果，如图 8 所示。同时，为了提高检测的精度和准确率，本文对同一个窗户被检测到的次数进行了统计，对于只被检测到一次的窗户，本文在实际聚类过程中将其作为一个误检项去除。

图 8 窗户的聚类结果

Fig. 8 Clustering result of windows

### 4.3 结果比较

本文将正射图的检测结果和二维航拍图形中的检测结果进行比较分析，从而验证本文基于中心点聚类方法的有效性和可行性。利用三维建筑对窗户的实际位置进行标定，结合正射图的建筑墙面标定得到的标准窗户与窗户的检测结果进行比较分析，从而验证本文基于包围盒聚类方法的有效性和可行性。

#### 4.3.1 基于中心点聚类结果的分析

本文通过地方测绘局对良渚文化村拍摄的二维航拍图片进行模型重建，以构建良渚文化村的三维模型。结合三维模型生成的正射图、实地二维图像的信息，对实际井盖的位置进行标注，以此作为标准，并将该标准分别与基于正射图的三维场景、基于多视角二维航拍图片的三维场景这两种不同的方式进行井盖检测结果对比。当检测结果中的井盖与标准结果中的井盖在水平平面内距离小于 1 m 时，认为该井盖已被正确识别。为了比较这两种方式的准确性，本文在正射图上对

二维航拍图拍摄范围进行标注，以保证二维航拍图像对应城市场景的被检测区域与正射图对应城市场景的被检测区域的大小一致。检测结果如表 1 所列。

表 1 两种井盖检测方法的结果比较

Table 1 Comparison between two methods of manhole cover's detection

| Algorithm | Standard | Detection | Accuracy | Detection Rate/% | Accuracy Rate/% |
|---|---|---|---|---|---|
| Orthograph | 249 | 302 | 193 | 77.51 | 63.91 |
| UAV image | 249 | 336 | 217 | 87.15 | 64.58 |

表 1 中，standard 指以正射图上标注的场景内所有井盖为标准，detection 为通过神经网络检测到的井盖数量，accuracy 为正确检测的井盖数量。从结果中不难发现，本文方法与基于正射图的方法在检测到的井盖的正确率上相差较小，但二维航拍图能够检测到更多正确的井盖，检出率比正射图提升了近 10%，说明本文方法真实有效。虽然基于正射图的检测方法也能较为准确地找到目标，但当井盖被上方物体（如树木等）遮挡时难以识别。而基于二维航拍图片的方法通过多视角的图像能够减少检测死角，增加检测精度，因此有相对较高的检出率。在 4.1 节中提到利用检测结果与地面的高度信息来比较去除干扰项，实验过程中本文选取 $h=0.1$ m，$h=0.2$ m，$h=0.3$ m 分别进行计算，计算结果如表 2 所列。

表 2 不同 $h$ 值下的检出率与准确率

Table 2 Detection rate and accuracy rate with different $h$

| $h$/m | Standard | Detection | Accuracy | Detection Rate/% | Accuracy Rate /% |
|---|---|---|---|---|---|
| 0.1 | 249 | 226 | 149 | 59.84 | 65.93 |
| 0.2 | 249 | 336 | 217 | 87.15 | 64.58 |
| 0.3 | 249 | 482 | 218 | 87.55 | 45.23 |

可以发现，当 $h=0.1$ m 时有很多井盖未被正确识别，导致井盖的检出率大大降低；而当 $h=0.3$ m 时，又有较多误检测的井盖被计入，导致检测的正确率大大降低。通过大量实验比较，当 $h$ 值为 0.2 m 时，本文方法的检出率和误检率均较高，符合本文对 $h$ 值这一参数的理论预期。

4.3.2 基于包围盒聚类结果分析

本文利用倾斜摄影测量技术重建的三维模型对窗户进行标定，每个窗户在空间中点出其左上角与右下角的空间坐标点，同时利用场景的正射图对建筑的外立面轮廓进行标注，得到每个建筑在正射图中的一个多边形，而多边形的每一条边则为建筑在 $XY$ 平面中的外轮廓位置。考虑到墙面的标定在正射图上完成，大部分乡村建筑存在屋檐结构，实际墙面位置与标定位置可能存在误差，可能出现墙体和相应窗户之间没有相交区域的问题。为此，本文将标记的线段向多边形内部方向平移一定距离 $d$，将平移后的线段与原线段组成一个矩形，以包含正确的墙面位置。记窗户左上角点 $A(x_1,y_1,z_1)$，点 $B(x_2,y_2,z_2)$，去除高度信息 $z$ 值得到点 $a(x_1,y_1)$ 与点 $b(x_2,y_2)$，如果 $a$ 和 $b$ 两点均在上文中得到的矩形范围内，则视该标定的窗户在此墙面上，求得点 $a$ 和 $b$ 在标定线段的垂点为 $c$ 和 $d$，与平移后线段的垂点为 $c_1$ 和 $d_1$。得到的 $c_1$，$c$，$d_1$，$d$ 这 4 个点则为标准窗户在 $XY$ 平面的四边形的 4 个顶点，同时给 4 个点分别加上高度信息 $z_1$ 和 $z_2$，从而得到标准窗户的空间包围盒。

对于平移距离 $d$ 的取值通常需要考虑墙面与外轮廓的距离误差。由于国内的建筑多存在屋檐结构，因此平移的距离需要超过屋檐的长度，才能正确包含墙面位置。为了合理地选取 $d$ 值，本文分别取不同的 $d$ 值进行计算，在三维模型中本文一共标记了 193 个窗户作为标准窗户，在不同的长度 $d$ 下计算得到的标准窗户的准确率（准确率＝计算得到的标准窗户数量/标记的标准窗户数量）如图 9 所示。可以发现，当 $d$ 从 0 不断增大时，计算得到的标准窗户的准确度也在不断提高，当 $d=0.3$ m 时窗户的准确度达到 100%，因此本文将 $d=0.3$ m 作为计算标准窗户的参数，同时也将 0.3 m 作为计算得到的标准窗户的厚度，用于评估检测结果。

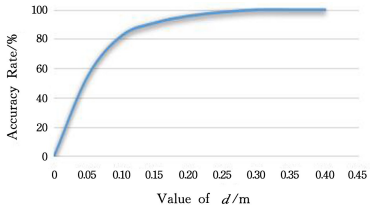


图 9 不同 $d$ 值下的标准窗户计算准确率

Fig. 9 Calculation accuracy of standard-window with different $d$

计算每个标准窗户的包围盒与检测结果的包围盒的体积交集与并集的比值及两个包围盒的交并比，如果交并比大于 0.1，则认为该标准窗户被成功检测，如果所有检测结果与其交并比都小于 0.1，则认为该窗户未被成功检测。具体结果如表 3 所列。

表 3 窗户的检测结果分析

Table 3 Analysis of window detection result

| Standard | Detection | Accuracy | Detection Rate/% | Accuracy/% |
|---|---|---|---|---|
| 193 | 205 | 184 | 95.34 | 89.76 |

通过表 3 不难发现，本文对于空间中的窗户具有较高的准确率。需要指出的是，由于正射图上基本看不到垂直于墙上的窗户，因此基于正射图的检测方法难以适应，这也是本文多视图方法的优势之一。

4.3.3 最低分辨率要求

本文通过同一场景不同分辨率的无人机航拍图像进行窗户的目标检测及三维聚类，分别将场景的无人机航拍图像照片缩放为原图像分辨率的 80%，60%，40%，20%，10%，5%，4%，3%，分别计算不同缩放比例下窗户的检出率和准确率，并通过折线图来展示，如图 10 所示，橙色线表示不同缩放比例对窗户的检出率的影响，蓝色线表示不同缩放比例对窗户检出结果的准确率的影响。通过该结果可以发现，当缩放比例大于 20%时，缩放比例对图片的检测结果几乎没有影响。本文方法要求的最低分辨率为 1094 * 730 像素。

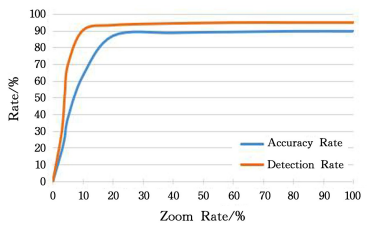


图 10 不同分辨率下检出率与准确率的变化（电子版为彩图）

Fig. 10 Changes of detection rate and accuracy rate with different resolutions

**结束语** 本文提出了一种基于多视图的三维城市场景小物体检测方法，首先利用深度学习实现对多视角、多距离的二维航拍图片中的小物体进行目标识别，然后通过相机的空间信息对三维城市场景中的小物体进行初步定位，并通过聚类算法实现最终的精确定位。通过井盖和窗户两个小物体检测实验，证明了本文方法的有效性，它不受视线遮挡的影响，相对于正射图上的物体检测具有较高的准确性和稳定性。

本文方法也具有一定的局限性，例如对于一些非矩形的小物体，利用包围盒重建后发现无法维持其本身的形状，会给后续模型建构带来困难。同时，利用中心点定位的方法也容易出现因为相机标定的误差造成定位不准确的情况。将来可以尝试通过多边形拟合等方法更好地对这些小物体进行识别和重建。由实验结果可知，本文提出的基于中心点的聚类方法能很大程度地提高目标物体的检出率，但同时会产生误检，需要在未来的识别过程中加入更多限制条件，以获取更好的检测结果。

## 参 考 文 献


[1] ZHU Q. Three-dimensional GIS and its Application in Smart City[J]. Geo-Information Science, 2014, 16(2): 151-157.

[2] LI S L, XIE W J, LI L, et al. A review of computer rapid building modeling methods[J]. Journal of Geographical Sciences, 2019, 42(9): 1966-1990.

[3] AGARWAL S, FURUKAWA Y, SNAVELY N, et al. Building Rome in a day[J]. Communicationsof the ACM, 2011, 54(10): 105-112.

[4] FURUKAWA Y, PONCE J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(8): 1362-1376.

[5] WU C, AGARWAL S, CURLESS B, et al. Multicore bundle adjustment[C]// Proceedings of Computer Vision and Pattern Recognition. IEEE, 2011: 3057-3064.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014: 580-587.

[7] GIRSHICK R. Fast r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[8] 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, 2016, 39(6): 1137-1149.

[9] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.

[10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[11] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7263-7271.

[12] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. arXiv: 1804. 02767, 2018.

[13] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]// Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[14] CHEN J, ZHANG Y Q, SONG P, et al. Application of Deep Learning in 3D Reconstruction of Objects Based on Single Image [J]. IEEE/CAA Journal of Automatica Sinica (JAS), 2019, 45(4): 657-668.

[15] TEKIN B, SINHA S N, FUA P. Real-time seamless single shot 6d object pose prediction[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 292-301.

[16] LI B, OUYANG W, SHENG L, et al. Gs3d: An efficient 3d object detection framework for autonomous driving [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1019-1028.

[17] MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3d bounding box estimation using deep learning and geometry[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017: 7074-7082.

[18] LI P, CHEN X, SHEN S. Stereo r-cnn based 3d object detection for autonomous driving [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7644-7652.

[19] 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 (CVPR). 2018: 918-927.

[20] SHI S, WANG X, LI H. Pointrcnn: 3d object proposal generation and detection from point cloud [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019: 770-779.

[21] SHI S, WANG Z, WANG X, et al. Part-$a^2$ net: 3d part-aware and aggregation neural network for object detection from point cloud [J]. arXiv: 1907. 03670, 2019.

[22] KUANG H, WANG B, AN J, et al. Voxel-FPN: Multi-Scale Voxel Feature Aggregation for 3D Object Detection from LIDAR Point Clouds [J/OL]. https://arxiv. org/abs/1907. 05286.

[23] LIU Z, GAO G, SUN L, et al. IPG-Net: Image Pyramid Guidance Network for Small Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPR). 2020: 1026-1027.


**CHEN Jia-zhou**, born in 1984, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. His main research interests include computer graphics and visual analysis.

**MA Ji**, born in 1985, Ph.D, lecturer, is a member of China Computer Federation. His main research interests include data visualization and so on.

（责任编辑：柯颖）