

## 摘要

近年来，无人车自动驾驶与导航技术的蓬勃发展催生了无人车产业的蓬勃兴起。激光雷达传感器（LiDAR）能够为车辆提供详尽的周围环境感知，因此被广泛应用于无人驾驶的场景中。激光雷达产生的数据为点云，而影响点云分辨率的参数为激光雷达的线数。低线数激光雷达获得的点云的垂直分辨率较低，影响无人车环境感知的效果；而高线数激光雷达成本高昂，直接制约了无人驾驶技术的普及。因此，如何有效的利用低线数雷达提供的数据进行环境感知，成为了目前在无人车领域较为热门的话题。

为了解决上述问题，本文提出了一种新颖的三维感知传感器机构，该机构通过无刷电机驱动曲柄连杆，为低线数的激光雷达提供垂直方向上的往复旋转运动；在此基础上，融合配准激光雷达在每次往复运动中发布的点云，来提升点云垂直方向上的分辨率；针对由于激光雷达的运动而产生的点云的运动畸变，本文提出一种帧内多次线性插值的方法，通过对激光雷达的每帧点云进行时间上的分段线性插值，有效地矫正了因三维感知机构运动而导致的点云的运动畸变。

此外，为了能够充分利用无人车各个传感器的数据进行三维障碍物的检测，本文还通过标定激光雷达与相机的外参，融合了激光雷达与相机的数据，对三维障碍物进行了检测与定位；同时利用激光雷达的点云信息对相机的目标检测结果做了优化，提高了基于图像的目标检测方法的查准率。

为了验证本文提出的三维感知机构的可靠性，本文首先利用Gazebo对三维感知机构进行了仿真，随后在无人车实验平台上利用上述的三维感知机构进行了三维建图与点云的障碍物分割的实验，并且检验了相机与激光雷达的标定效果。实验结果表明，本文提出的三维感知机构能够有效地提高激光雷达的垂直分辨率。通过三维感知机构融合配准后的点云，其垂直分辨率可以达到64线激光雷达的三到四倍，对于三维建图以及无人车环境下的三维障碍物的检测都有着重要的意义。

**关键词：**无人车，障碍物检测，激光雷达，点云，多传感器融合



## ABSTRACT

The activity in the area of autonomous vehicles navigation in recent years has initiated a series of reactions that stirred the automobile industry. Due to the LiDAR sensor's ability of providing a detailed understanding of the environment surrounding the vehicle, it has been widely used in a plethora of autonomous driving scenarios. The data type of LiDAR is point cloud. A factor that affects the point cloud's vertical resolution is the number of LiDAR's beams. Point clouds achieved from a low-beams LiDAR can be very sparse and thus have a negative effect on the environmental perception, while high-beams LiDAR are extremely expensive, thus constrain the popularity of the technology of autonomous vehicles.

To solve the problem mentioned above, a novel mechanism is proposed in this paper. This mechanism reciprocate the low-beams LiDAR sensor in the heading direction through a crank-linker. Then, register all the point clouds during a period of the mechanism's motion to increase the vertical resolution of the point cloud. In addition, this paper proposes a method which rectifies the motion distortion of LiDAR by multiple linear interpolation in a single point cloud, thus improves the performance of multiple frame fusion.

Furthermore, to make the best of multiple sensors to detect obstacles in autonomous environment, camera image and LiDAR data are fused after calibrating the LiDAR and cameras' extrinsic parameters. This paper also uses point cloud to optimize the performance of objection detection accuracy based on camera image.

To evaluate the mechanism proposed in this paper, a simulation in Gazebo was performed to simulate the multiple frames fused cloud. Then the mechanism was placed in an autonomous vehicle to construct a 3d map and segment obstacles in point clouds. This paper also evaluates the result of camera-LiDAR's calibration. The experiment shows that the mechanism can effectively improve the vertical resolution of the point cloud, which is three to four times of the 64-beams LiDAR. It's very significant for 3d mapping and obstacle detection in autonomous driving scenarios.

---

## ABSTRACT

---

**Keywords:** autonomous vehicles, obstacle detection, LiDAR, point cloud, multiple sensors fusion

## 目 录

<b>第1章 绪论 .....</b>	<b>1</b>
1.1 研究工作的背景与意义 .....	1
1.2 国内外研究历史与现状 .....	2
1.2.1 基于相机的检测方法 .....	2
1.2.2 基于激光雷达的检测方法 .....	2
1.2.3 基于毫米波雷达的检测方法 .....	3
1.2.4 多传感器融合的检测方法 .....	3
1.3 本文的主要贡献与创新 .....	3
1.4 本论文的结构安排 .....	4
<b>第2章 三维感知传感器机构设计 .....</b>	<b>5</b>
2.1 三维感知传感器机构的机械结构设计 .....	5
2.1.1 机械结构运动原理 .....	6
2.1.2 曲柄连杆机构的设计 .....	7
2.2 三维感知传感器机构的电路设计 .....	9
2.2.1 驱动器 .....	9
2.2.2 执行机构 .....	10
2.2.3 传感器 .....	10
2.2.3.1 角度传感器 .....	10
2.2.3.2 激光雷达 .....	10
2.2.4 主控板 .....	10
2.2.5 电路拓扑 .....	12
2.3 三维感知传感器机构的软件设计与运动控制 .....	12
2.4 本章小结 .....	12
<b>第3章 点云和图像的多感知融合 .....</b>	<b>14</b>
3.1 激光雷达点云的多帧融合 .....	14
3.1.1 一种朴素的多帧融合策略 .....	14
3.1.2 点云的运动畸变的形成与矫正 .....	15

## 目录

---

3.1.2.1 点云的运动畸变 .....	15
3.1.2.2 运动畸变的矫正 .....	16
3.1.3 矫正运动畸变后的多帧融合策略 .....	18
3.2 激光雷达与相机的标定 .....	19
3.2.1 标定板 .....	20
3.2.2 相机坐标系中的三维特征点提取 .....	20
3.2.3 LiDAR坐标系中的三维特征点提取 .....	21
3.2.4 坐标系欧氏变换求解 .....	21
3.2.5 多帧坐标系变换结合 .....	24
3.3 点云与图像的融合 .....	25
3.3.1 相机成像原理 .....	25
3.3.2 激光雷达点云的投影 .....	27
3.4 本章小结 .....	27
<b>第4章 基于视觉与三维点云融合的三维障碍物检测方法 .....</b>	<b>29</b>
4.1 YOLO一种实时目标检测网络 .....	29
4.2 基于YOLO的视觉、三维点云结合的三维障碍物检测 .....	31
4.2.1 点云前景与背景的分割 .....	31
4.2.2 目标三维坐标的计算与检测结果的优化 .....	33
4.3 本章小结 .....	35
<b>第5章 实验验证与结果分析 .....</b>	<b>36</b>
5.1 Gazebo下的三维感知机构仿真 .....	36
5.2 三维感知机构结合里程计信息构建三维地图 .....	37
5.3 基于点云投影到深度图像上的物体分割 .....	39
5.3.1 地面点的去除 .....	39
5.3.2 点云到深度图像的投影 .....	41
5.3.3 基于深度图像的物体分割 .....	44
5.4 激光雷达与相机标定结果的验证 .....	46
5.5 多感知融合算法在KITTI数据集上的验证 .....	46
5.6 本章小结 .....	46
<b>第6章 全文总结与后续工作展望 .....</b>	<b>48</b>
6.1 全文总结 .....	48
6.2 后续工作展望 .....	49

## 目录

---

参考文献 .....	50
致 谢 .....	53
外文资料原文 .....	56
外文资料译文 .....	56



## 第1章 绪论

### 1.1 研究工作的背景与意义

近几年来，自动驾驶技术取得了长足的进步，而其中关键的技术就是多传感器的环境感知与融合。环境感知的一个重要环节便是障碍物检测。目前，虽然基于图像的障碍物检测已经取得了卓有成效的进步，然而相较于三维障碍物检测，二维障碍物检测有以下缺陷：

- (1) 基于单目相机的障碍物检测没有尺度信息，无法恢复出目标的三维坐标。
- (2) 基于双目相机的障碍物检测，当基线较短时，测量距离较长（5m以上）的物体时计算出来的距离信息很不准确，而当基线较长时，近处物体的检测又容易出现在两个相机的视野盲区之中，从而导致无法三角化而得出距离信息。

基于上述原因，越来越多的目光聚焦在了基于激光雷达（LiDAR）的三维障碍物检测。LiDAR是Light Detection And Ranging的缩写，中文译作“激光探测与测量”，一般指多线数的三维激光雷达传感器。相较于相机图像，激光雷达的点云拥有以下几点优势：

- (1) 测量范围广。目前的激光雷达的测量有效距离基本都在0.5-100米左右，远高于双目相机三角测距的适用范围。
- (2) 测量精度高。激光雷达的测距误差可达厘米级，同样优于双目相机的测距结果。目前最常见的旋转式激光雷达，其本质是多个激光束旋转后对每个时刻的测距结果进行保留与配准，最后再以点云的形式发布出去。决定激光雷达的竖直方向分辨率的参数为其激光束的数量，一般称之为激光雷达的线数。目前常用的激光雷达线数有16线、32线、64线等，其中由于低线数的激光雷达生成的点云在测量远距离物体时竖直方向分辨率较低。因此，低线数激光雷达点云的稀疏性较大地制约了三维障碍物检测任务的准确率。

通常在自动驾驶的无人车系统中会在车的四周装上多个16线的激光雷达进行点云融合，或者直接采用线数更高的激光雷达来做障碍物检测的任务。然而目前三维激光雷达造价不菲，无人车系统中光是64线激光雷达的成本就将近十万美金，如此高昂的成本在一定程度上限制了无人驾驶汽车的普及与推广。

鉴于上述存在在问题，本文希望能够提出一种基于多帧融合的低线数激光雷达感知机构，使其能够通过增加一个在垂直方向的往复运动，并将该机构上激光雷达的多帧点云融合发布来提高三维点云的稠密性，借而解决低线数激光雷达在障碍物检测问题上由于点云的稀疏性而造成的困难。并且，本文还希望通过融合上述机构发布的点云信息以及相机的图像信息，发挥各个传感器的优势从而提高三维障碍物检测任务的准确率与实现三维障碍物的多类别检测。

## 1.2 国内外研究历史与现状

根据本文的主要研究方向，下面将对面向无人车的三维障碍物检测方面的研究展开调研。目前，面向无人车的三维障碍物检测主要可以分为基于相机的检测方法、基于雷达的检测方法和多传感器融合的检测方法这三种类型。

### 1.2.1 基于相机的检测方法

基于相机的检测方法目前主要是以基于深度学习的物体检测为主。随着 ImageNet 数据集<sup>[1]</sup>的建立以及大规模目标检测竞赛的进行，深层卷积网络被成功运用在了图像识别与物体检测领域<sup>[2]</sup>，而通过深度学习来实现物体的识别与分类，正是目前无人车障碍物检测的主流方法。2017 年 Mask R-CNN<sup>[3]</sup>的提出，完成了针对目标的实例分割（Instant segmentation），随后大量的实例分割算法被提出<sup>[4, 5]</sup>，实例分割自身也成为了近年来计算机视觉的热门话题。

针对单目相机缺乏深度信息的问题，近些年来还有一些基于单目图像的端到端的深度估计的网络也被提出<sup>[6]</sup>，能够直接单目图像对图像的每个像素进行深度估计。不过目前这种方法得到的深度估计还不够精确，并且泛化能力不够强，因此直接利用这种深度估计进行障碍物检测的方法还比较少。

### 1.2.2 基于激光雷达的检测方法

激光雷达通过发射激光并测量其返回时间，得到距离信息，测量距离远、速度快、误差小并且分辨率高。基于激光雷达的障碍物检测一般可分为基于模型与无模型两种类别。基于模型的检测方法<sup>[7]</sup>通过先验知识构建模型来同时进行点云的检测与分类，但是计算量巨大并且很难做到实时。无模型的方法则首先采用模型估计地面<sup>[8]</sup>来去除地面点云，为了减少计算量，将剩余点云投影到地面<sup>[9]</sup>或者

一个虚拟的平面上生成图像（称之为深度图（Range image））<sup>[10]</sup>，之后再从图像上进行障碍物的检测与分割。

另外亦有国外研究提出3D检测网络将特征提取和边界框预测统一到单个阶段的端到端可训练深度网络中，通过深度学习同时解决点云的检测与分类问题<sup>[11]</sup>。其深度学习网络输入为点云，通过对点云预处理和卷积后得出障碍物的包围盒以及类别。随着一种稀疏矩阵卷积的方法的提出<sup>[12]</sup>，这些目标检测方法速度已经达到实时性的要求。然而，这种基于点云的端到端的方法仍然具有训练时间长、泛化能力不够强的缺点，只在公开的数据集诸如KITTI上进行了测试，而在实际场景中的检测效果还有待考究。

### 1.2.3 基于毫米波雷达的检测方法

毫米波雷达是对毫米波段的信号进行检测的雷达。相对于激光雷达其价格较为低廉，而且也能测得物体的位置信息。目前也有一些利用毫米波雷达来进行障碍物检测的工作<sup>[13, 14]</sup>。然而毫米波雷达的数据稳定性较差，对金属比较敏感，并且只能提供距离和角度信息，没有高度信息。由于这些缺点，单纯利用毫米波雷达的数据会给技术开发带来很多挑战。

### 1.2.4 多传感器融合的检测方法

由于无人车是一个多传感器的系统，对环境感知算法的鲁棒性有较高的要求，因而在无人车环境中，经常融合多个传感器的信息来提供对周围环境的感知。激光雷达与相机标定的工作<sup>[15]</sup>就是为了结合点云与图像来为三维障碍物进行检测与定位<sup>[16]</sup>。一些方法实现了激光雷达与相机的在线标定，并根据激光雷达与相机的信息进行目标检测<sup>[17, 18]</sup>；另外一些工作将激光雷达的点云投影到图像上，并将点云上采样得到带有深度信息的图像作为卷积神经网络的输入，进行道路的检测<sup>[19]</sup>。除此之外

## 1.3 本文的主要贡献与创新

针对低线数激光雷达在三维障碍物检测问题中因为点云的稀疏性与不一致性给三维目标检测带来的困难，本文提出了一种新颖的三维感知机构，通过给激光雷达提供偏航角方向上的往复运动，并且融合多帧点云来增加激光雷达点云在竖

直方向上的分辨率，同时通过帧内多次线性插值来矫正因机构运动而带来的点云的运动畸变。

同时，本文还通过对相机与激光雷达的标定，融合了相机图像与激光雷达的传感器数据信息，借由相机的目标检测算法来对物体进行检测，通过激光雷达的点云信息来对检测出的三维障碍物进行定位，并且利用点云信息优化了视觉的目标检测的结果。

## 1.4 本论文的结构安排

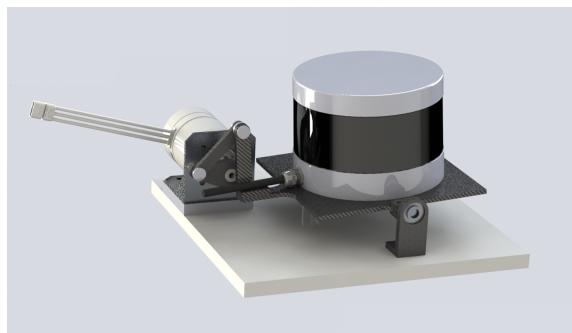
本文的结构安排如下：

- (1) 第一章为绪论，介绍面向无人车的三维障碍物检测的研究背景与本文的主要研究内容。
- (2) 第二章为三维感知机构的设计，介绍了本文提出的一种三维感知机构的机械、电路、软件方面的设计。
- (3) 第三章介绍了点云的畸变矫正以及相机与激光雷达的标定，为后文的相机与激光雷达的数据融合提供外参。
- (4) 第四章介绍了相机与激光雷达的融合，利用相机图像对障碍物进行检测与分类，同时利用激光雷达的点云信息来对障碍物进行定位。
- (5) 第五章根据前文提到的三维感知机构，设计了仿真实验与基于点云投影到深度图像的物体分割。

## 第2章 三维感知传感器机构设计

为了能够解决低线数激光雷达在障碍物检测问题上由于点云的稀疏性而造成的困难，本章提出了一种三维感知传感器机构，通过增加三维激光雷达在垂直方向的往复旋转，同时融合多帧激光点云，来增加激光雷达在铅垂方向上的分辨率，实现类似于高线数激光雷达的稠密点云。

### 2.1 三维感知传感器机构的机械结构设计



(a)



(b)

图 2-1 机构总图(a)solidworks渲染图;(b)实物图

### 2.1.1 机械结构运动原理

本章所述的机构结构如图2-1(b)所示，其中3508电机提供驱动转矩，曲柄连杆装置将电机的旋转运动转化为激光雷达底座在俯仰角(pitch)方向上的往复运动，同时绝对值磁编码器记录激光雷达在俯仰角上的角度变化，以供多传感器融合时使用。

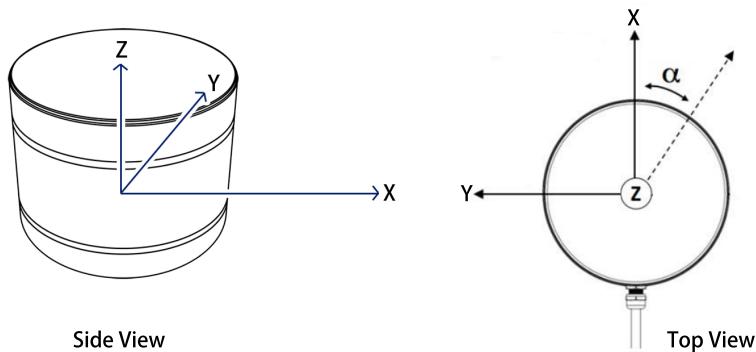


图 2-2 激光雷达坐标系

激光雷达自身的坐标系如图2-2所示，为右手系，其坐标原点在激光雷达的体心。由于激光雷达的传感器性质，导致其在Z轴方向上的点云十分稀疏，因此，本文希望通过机构让激光雷达绕Y轴（也就是上文所述的俯仰角方向）做往复旋转运动，并通过配准多帧激光雷达的点云来实现激光雷达点云在Y轴方向上的稠密化。

在文献<sup>[20]</sup>中，作者将二维激光雷达固联到电机上，控制电机进行俯仰角方向上的往复运动，并通过电机上的编码器来读取电机的旋转角度值，然后融合多帧二维激光雷达的点云为新的点云后发布，如图2-3所示。

该机构的主要优点为结构简单，这也是本文初次采用的机构设计。然而该机构有以下几个缺点：

(1) 在往复运动中，当运动方向发生改变时，由于舵机控制精度的问题，很难做到平滑换向，并且经常伴随有较大的震动，给之后的传感器融合算法带来了困难。

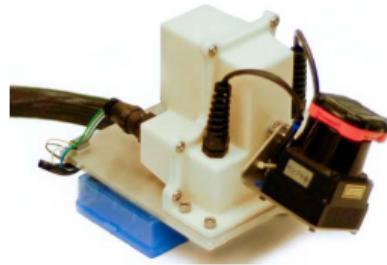


图 2-3 LOAM中的机构设计

(2) LOAM中采用的是二维激光雷达，重量较轻，而本文需要带动三维激光雷达进行往复运动，重量较重（近1kg），长时间使用舵机带动会使舵机产生较为明显的回程间隙，影响角度的测量与后续的传感器融合的效果。

因为上述原因，我们没有采用这种结构设计，而是采用了之前提到的曲柄连杆机构来针对三维激光雷达进行往复运动，相较于上述机构，曲柄连杆结构有以下几个优点：

(1) 换向平滑。执行电机只需要一直向同一方向旋转，曲柄连杆机构就能够自动换向，并且输出的角度曲线近似正弦曲线。

(2) 对执行机构负担小。仅需要较小并且较为恒定的转矩就能够驱动较大的负载做往复运动。

(3) 对执行机构的控制要求低。在该机构中，无刷电机只需要输出恒定的转矩就能够完成三维激光雷达在偏航角方向上的往复运动，并且经过验证，其角度输出近似正弦曲线，而若采用上述的舵机机构，要想得到相近的角度曲线，则对舵机的软件控制提出了较高的要求。

综上，本文选择曲柄连杆机构作为该机构的驱动机构。

### 2.1.2 曲柄连杆机构的设计

该三维感知机构的一个难点在于如何设计与电机相连的曲柄连杆机构。这里参照《机械设计基础》<sup>[21]</sup>一书中的相应章节对曲柄四连杆机构的连杆长度进行求解。如图2-4所示，假设已知该铰链四杆机构两连架杆AB和CD所形成的角度 $\psi_1$ 和 $\phi_1$ 在三个不同位置下的角度，要求连杆a,b,c,d的尺寸。则将连杆视为向量，向x、y轴投影，有

$$a \cos \phi + b \cos \delta = d + \cos \phi \quad (2-1)$$

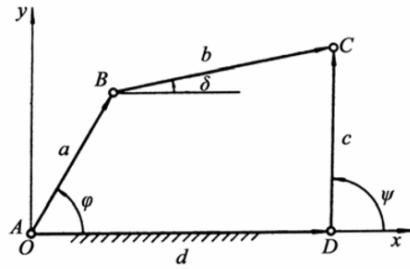


图 2-4 四杆机构的数学模型

$$a \sin \phi + b \sin \delta = c \sin \psi \quad (2-2)$$

将式2-1与式2-2先进行移项，然后作平方和相加，从中消去 $\delta$ 后整理可得

$$b^2 = a^2 + c^2 + d^2 + 2cd \cos \psi - 2ad \cos \phi - 2ac \cos(\phi - \psi) \quad (2-3)$$

设

$$\begin{cases} R_1 = (a^2 + d^2 + c^2 - b^2) \\ R_2 = d/c \\ R_3 = d/a \end{cases}$$

代入，则式2-3可以化简为

$$R_1 - R_2 \cos \phi + R_3 \cos \psi = \cos(\phi - \psi) \quad (2-4)$$

该式即为铰链四连杆机构的角位置方程，该方程有三个待定参数 $R_1$ 、 $R_2$ 、 $R_3$ 。故应有三组对应的 $\psi$ 和 $\phi$ 角才能得出这个方程的解。将三组 $\psi$ 和 $\phi$ 角代入求解该方程后，可以得到四个构件之间的长度关系为

$$\begin{cases} a = d/R_3 \\ c = d/R_2 \\ b = \sqrt{a^2 + c^2 + d^2 - 2acR_1} \end{cases}$$

则根据机构的具体设置情况，知道 $a, b, c, d$ 中的任何一条边的长度后，便可知剩下四条边的长度。

在实际设计中，我们已知 $\psi_1$ 和 $\phi_1$ 的三组对应角度为

$$\begin{cases} \psi_1 = 30^\circ \quad \phi_1 = 36.3^\circ \\ \psi_1 = 60^\circ \quad \phi_1 = 43.87^\circ \\ \psi_1 = 120^\circ \quad \phi_1 = 35.75^\circ \end{cases}$$

并且根据我们的机构设置，构件d的长度为105.72mm。将这些已知量代入公式中可得

$$\begin{cases} a = 31.6mm \\ b = 49.18mm \\ c = 108.37mm \end{cases}$$

由此，便得到了曲柄机构的连杆构件设计参数。

## 2.2 三维感知传感器机构的电路设计

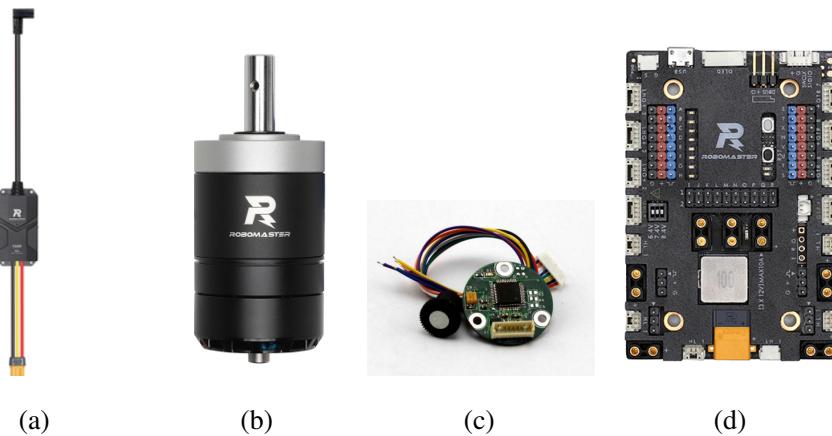


图 2-5 驱动器与传感器(a)C620 电调;(b)M3508无刷电机;(c)磁编码器;(d)RM A型开发板

### 2.2.1 驱动器

该三维感知机构采用的驱动器为DJI C620电调，如图2-5(a)所示。该电调支持50-500Hz的PWM（脉宽调制）信号控制以及CAN总线指令控制，最高支持20A的

持续电流，支持对CAN总线上的电调快速设置ID，支持通过CAN总线获取电机温度、转子位置和转子速度等信息，切换电机时可无需进行位置传感器的参数校准。

### 2.2.2 执行机构

该三维感知机构采用的执行机构为DJI M3508无刷电机，如图2-5(b)所示。该电机可搭配上文所述C620电调实现正弦驱动，相比传统方波驱动具有更高的效率、机动性和稳定性。其最高可持续输出力矩为2.8Nm，满足驱动曲柄四连杆机构的需求。

### 2.2.3 传感器

#### 2.2.3.1 角度传感器

该三维感知机构采用的角度传感器为傲蓝13线磁编码器，如图2-5(c)所示。该编码器采用RS485方式通信，其单圈分辨率为8192cpr，精度为±0.1度。该编码器为绝对值式编码器，其相对于增量式编码器不同点在于，增量式编码器以上电时的位置为零点，每次使用都要机械对位；而绝对值式编码器能够记录机构的唯一位置，即单圈内编码器的每一个示值，都唯一对应了空间中机构的位置与角度。考虑到我们曲柄连杆机构的特性，显然绝对值式编码器更加符合我们的要求。

#### 2.2.3.2 激光雷达

该三维感知机构采用的激光雷达为速腾聚创的RS-LiDAR-16，如图2-6所示。该激光雷达为16线激光雷达，其测距范围为50cm-150m，精度误差为±2cm。垂直视场角为30度，其角分辨率为2度；水平视场角为360度，其角分辨率为0.09-0.36度（对应的点云频率为5Hz-20Hz）。

### 2.2.4 主控板

该三维感知机构采用的主控板为DJI Robomaster A型开发板，如图2-5(d)所示。该开发板具备类型丰富的接口，包括12V、5V、3.3V电源接口、CAN接口、UART接口、可变电压PWM接口、SWD接口等。同时该开发板拥有电源输入的防反接、过压保护、缓启动、12V电源输出过流保护、PWM端口的ESD等多重保护。



图 2-6 速腾16线激光雷达

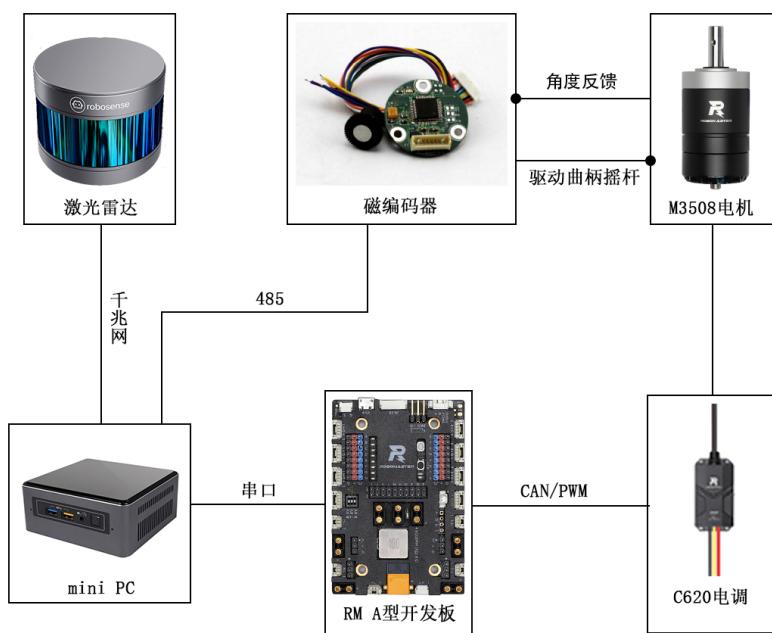


图 2-7 电路拓扑

### 2.2.5 电路拓扑

该三维感知机构的电路拓扑如图2-7所示。激光雷达通过千兆网接口将点云传输到mini PC上，磁编码器通过485转USB与mini PC通信，同时主控通过PWM控制电调输出，调节M3508电机的转速，M3508电机提供曲柄四连杆机构的驱动力矩，而磁编码器又将曲柄机构作用在激光雷达底座上的旋转通过485通信输出到mini PC。

## 2.3 三维感知传感器机构的软件设计与运动控制

根据上文所述的机械设计以及电路设计，该三维感知机构的软件设计主要实现了以下几个任务：

1. 实现了各个传感器、主控到Mini PC的通信，同时将数据以ROS（Robot Operating System）话题的方式发布出去，以供第二章节提到的多帧融合算法使用。
2. 实现了电机的多档调速功能。为了应对不同的场景，在主控中实现了多档调速功能，以调节曲柄机构的往复运动频率。
3. 实时检测电调的温度信息，提供了基于温度检测的堵转保护（温度过高自动切断控制）。

此外，本文还记录了在电机输出恒定转速情况下的曲柄连杆机构的输出的角度信息，如图2-8所示。该图纵坐标为角度制的输出角度。从图中可以看出，本章所设计的三维感知机构其输出角度近似正弦曲线，并且没有较大的换向震动，相比起上文所提到的舵机的结构拥有稳定可靠的优势。

## 2.4 本章小结

本章提出了一种新的三维感知机构，并从该机构的机械设计、电路设计以及软件设计和运动控制三个方面介绍了该机构。在机械方面，本文提出使用无刷电机+曲柄摇杆机构来代替简单的舵机给三维激光雷达提供一个竖直俯仰角方向上的往复运动，这种结构的优势在于机构换向流畅、控制简单以及对机构负载小，能够为本文后续章节提到的激光雷达的多帧融合提供结构上的稳定与可靠性。在电路方面，本章利用绝对值式磁编码器对曲柄机构运动的角度进行了记录与输出，相较于增量式编码器，磁编码器不需要保证每次上电时机构都在同一个位置，

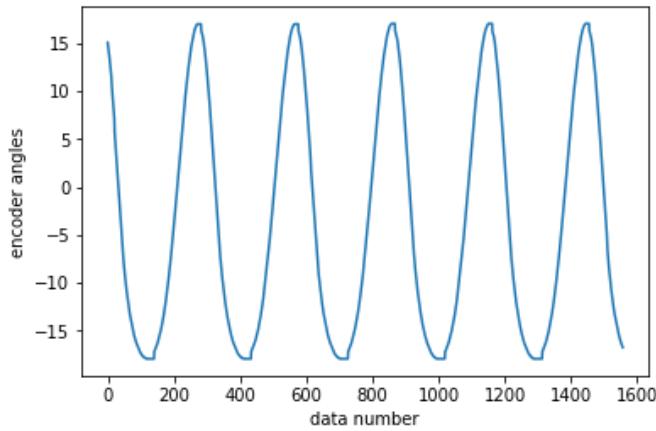


图 2-8 曲柄连杆机构输出角度

为机械结构的设计提供了便利。在软件方面，本章实现了各个传感器与主控以及mini PC的通信，主控对无刷电机的多档控制以及对电机的堵转保护，并且绘制了输出角度，验证了机构的可行性。本文的后续章节将利用该机构输出的点云信息以及角度信息来进行点云的多帧融合稠密化，并且在融合的点云上进行三维障碍物的检测与分类。

## 第3章 点云和图像的多感知融合

根据本文第二章所提到的机构，能够将三维激光雷达在其俯仰角方向上提供一个有规律的正弦往复运动。本文提出该机构的主要目的为将激光雷达在时间轴上的多帧点云进行配准，进而增加激光雷达在竖直方向的分辨率。本章将对第二章所述的三维感知机构得到多帧激光雷达点云进行融合，并且在目前机构的基础上进行激光雷达与相机的标定，从而使得融合后的点云能够应用于后文提到的基于视觉与激光融合的三维障碍物检测方法。

### 3.1 激光雷达点云的多帧融合

#### 3.1.1 一种朴素的多帧融合策略

点云的配准（registration）是指将有重合部分的点云进行对齐的一项技术。其关键核心为求出给定点云的坐标系相对于目标点云所在坐标系的旋转与平移，借以将给定点云转换到目标点云坐标系中。

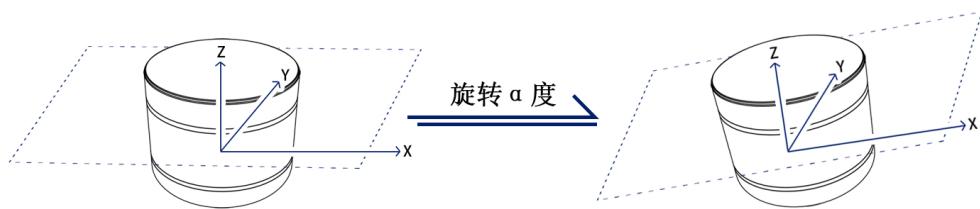


图 3-1 三维感知机构的旋转

如图3-1所示，本文认为当激光雷达俯仰角与地面平行时，其激光雷达坐标系为参照系 $f$ ，第二章所述的三维感知机构目的就在于在激光雷达做往复运动时，

将激光雷达每帧点云在参照系 $f$ 上进行配准，最后在曲柄机构的一个运动周期后将点云融合输出。当机构运动时，激光雷达自身坐标相对于 $f$ 发生了旋转，因而在激光雷达坐标系中的点云相对于 $f$ 有一个 $\alpha$ 角度的旋转量。因而要将此时的点云配准到 $f$ 坐标系中时，要抵消因机构旋转而造成的坐标系的旋转。最为直观的策略就是，读取绝对值编码器返回的角度 $\alpha$ ，将激光雷达每帧点云沿着俯仰角方向旋转 $-\alpha$ 的角度，然后配准多帧的激光雷达点云并发布。

在实际的实现过程中，编码器返回角度的频率约为30Hz，而点云发布的频率约为10Hz。为了对角度信息进行线性插值，将角度信息保存到一个队列 $Q$ 中，同时在接收到新的点云信息时，根据点云的时间戳 $t_{lid}$ ，在队列 $Q$ 中寻找相邻两帧角度信息，记这两帧角度信息的时间戳为 $t_{enc}^0, t_{enc}^1$ ，则这两帧信息应满足 $t_{enc}^0 \leq t_{lid} \leq t_{enc}^1$ 。那么 $t_{lid}$ 时刻的角度值由下式线性插值得到：

$$\alpha = \alpha_0 + (\alpha_1 - \alpha_0) \times \frac{t_{lid} - t_{enc}^0}{t_{enc}^1 - t_{enc}^0} \quad (3-1)$$

同时，根据计算曲柄机构的角度是增加还是减少，来判断曲柄机构的运动方向，并且将曲柄机构一个往复运动周期内的点云融合为一帧新的点云输出。

然而这种朴素的融合策略在实际中效果不好，体现在融合后的点云所显示的物体轮廓失真严重，如图3-4(a)所示，原因是没有考虑激光雷达的运动对激光雷达点云生成的影响。

### 3.1.2 点云的运动畸变的形成与矫正

在激光雷达点云的多帧融合中，如果只是进行简单的历史点云叠加（如上文所述），那么融合后的点云相较于真实情况会有很严重的失真，其原因就在于第二章所提到的三维感知机构在给激光雷达在偏航角方向上的往复运动时，点云会产生不可忽视的运动畸变。本章节首先介绍什么是激光雷达点云的运动畸变，然后提出一种通过插值的方式矫正激光雷达的运动畸变。

#### 3.1.2.1 点云的运动畸变

激光雷达的点云的形成本质上是由激光雷达内部的多个激光测距器将一个旋转周期内的各个测量值记录下来并同时发布后得到的。因此点云中的每个点并不是在同一时刻被测量出来的。如果激光雷达在测量的过程中也在运动，那么激光雷达的点云可能会发生畸变<sup>[22]</sup>。

下面以二维激光雷达为例，介绍激光雷达点云运动畸变的形成。图3-2(a)中

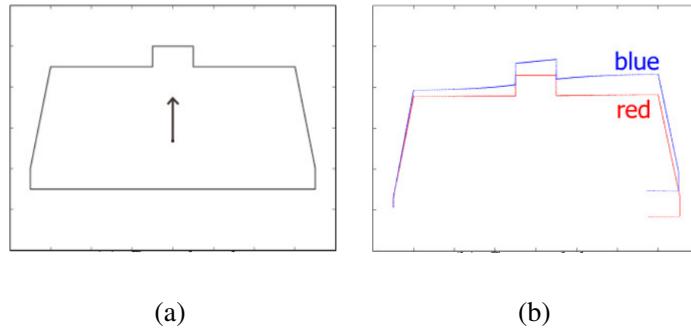


图 3-2 激光雷达运动畸变(a)ground truth;(b)采集得到的数据

的黑色的线条表示二维激光雷达处在的真实环境的轮廓图，箭头表示二维激光雷达的运动方向。图3-2(b)中的蓝色的线条表示二维激光雷达的原始数据。注意到其已经发生了畸变，因为二维激光雷达内的激光测距器通过逆时针的方向旋转，使得右上角的数据优先得到，而左上角的数据在激光雷达向箭头方向运动了一段距离之后才进行测量，自然导致了运动畸变的产生。

值得一提的是上图表示的二维激光雷达发生的运动畸变是当激光雷达运动方向为水平运动方向时造成的，而当激光雷达在空间中有垂直方向的旋转运动时，其造成的运动畸变远比水平运动严重。这是因为激光雷达旋转一周的时间普遍在0.1秒左右，其水平运动的距离往往很小可以忽略不计。而当激光雷达有竖直方向的旋转时，即使在0.1s内只有2度的俯仰角的旋转（这在本文的机构运动中并不算很快），在测量20m处的物体时，其运动造成的点云畸变可使得点云的同一线上的第一个点与最后一个点的垂直距离相差将近70cm。

综上所述，对于第二章所述的机构，由于其施加了在激光雷达俯仰角方向上的旋转，因此导致其在竖直方向上的运动畸变不可忽视，从而简单的叠加点云会导致在做激光雷达的物体检测时的失真。

### 3.1.2.2 运动畸变的矫正

对于本文所提到的三维感知机构在俯仰角方向上的旋转所产生的运动畸变的矫正，一个较为直观的方法是，依次遍历激光雷达每帧点云中的每个点，计算其产生的时间戳 $t$ ，对磁编码器的角度进行插值，计算出在时间戳 $t$ 上的角度 $\alpha$ ，然后将该点绕原点在俯仰角方向旋转 $-\alpha$ 的角度。然而激光雷达每帧点云高达数十万个

点，如果对每个测量得到的点进行插值，则在一秒内要进行近百万次的插值与旋转操作，显然对于无人驾驶汽车上的移动处理器平台来说这是不现实的。

因此本文提出一个折中的假设，设点云中第一个点产生的时间戳为 $t_0$ ，最后一个点产生的时间戳为 $t_k$ ，则将 $t_0$ 到 $t_k$ 之间的时间均匀分为 $n$ 份，每份长为 $\Delta t$ 。本文假设每个 $\Delta t$ 时间段内的激光雷达的测距点的产生的时间都是相同的，为其第一个点的产生时间。根据这个假设，每个时间段内的所有点都只要进行相同角度变换即可进行运动畸变的矫正。遵循该假设，则每帧点云只需要进行 $n$ 次角度插值即可，极大地减小了运算量。同时虽然该假设认为同一时间段内的点云是同一时间产生的，每个时间段内的点仍有运动畸变的影响，然而在实验过程中发现，只要当 $n$ 取一个不太小的值 ( $n \geq 50$ )，则该假设的所产生的时间段内的运动畸变产生的影响很小，可忽略不计。

在实际的程序实现中，由于激光雷达每帧点云是分段传输的，如图3-3所示。RS-LiDAR-16激光雷达采用UDP协议向PC传输点云信息，而UDP协议相较于TCP协

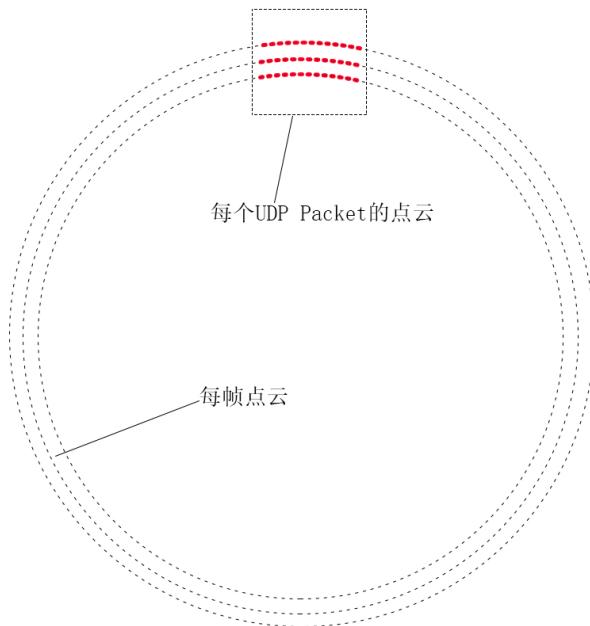


图 3-3 点云的传输

议，发送数据之前不需要双方建立链接，并且发送的数据没有校验，也没有丢包的检测，因此不适合一次性发送大量数据给PC。该激光雷达将一帧点云分为84个UDP包(UDP Packet)，每个包中的点云都是激光雷达旋转 $360/84 = 4.28$ 度后得到

的16线的点云的集合。当激光雷达的驱动程序接收到84个UDP Packet之后，将这84个Packet合并成一帧点云输出。

本文直接对每个Packet（而不是每帧）分别进行编码器角度的插值与点云的旋转，最后再将旋转后的84个Packet进行合并发布。这样相对于之前提到的朴素的融合方法，每帧帧内多进行了84次线性插值，从而大大减小了相机运动畸变带来的影响。

### 3.1.3 矫正运动畸变后的多帧融合策略

如上文所言，对单帧点云进行多次插值之后的多帧融合算法流程如Algorithm 1所示。

---

#### **Algorithm 1** Improved multiple pointcloud fusion algorithm

---

```

1: for packet  $\in$  point cloud do
2:    $\alpha_1, \alpha_2 = \text{FindClosestAngles}(\text{encoderAngles}, \text{packet.timestamp})$ 
3:    $\alpha = \alpha_1 + (\alpha_2 - \alpha_1) \times \frac{\text{packet.timestamp} - \alpha_1.\text{timestamp}}{\alpha_2.\text{timestamp} - \alpha_1.\text{timestamp}}$ 
4:   rectifiedPacket = RotatePointCloudByYaw(packet,  $-\alpha$ )
5:   rectifiedPointcloud += rectifiedPacket
6: return rectifiedPointcloud

```

---

在Algorithm 1中，`FindClosestAngles`函数查找在所有的编码器角度中，时间上离`packet`的时间戳最近的两帧编码器的角度值，随后对这两个角度进行线性插值得到`packet`时间戳下的机构的角度。`RotatePointCloudByYaw`函数将特定的点云绕激光雷达坐标系原点旋转指定的角度。最后将每个矫正后的`packet`合并到一个新的矫正后的点云中并发布出去即得到了矫正因机构而产生的运动畸变后的点云。

在图3-4(b)中展示了消除运动畸变后的激光雷达测得的轿车的点云图案。相较于图3-4(a)，其点云图案没有出现明显的失真，并且通过机构进行多帧融合后的点云，能够明显的看出轿车的轮廓细节信息，包括车的反光镜、前挡风玻璃等。而矫正畸变前的轿车点云则很难分辨出这些细节，并且由于运动畸变的作用，其体积明显比矫正后的点云更大一些。由此证明了本文提出的多帧融合算法拥有较好的矫正畸变的效果。

为了显示畸变矫正后的对比效果，本文还挑选了离激光雷达较远的建筑物上的窗户的点云的矫正机变的前后对比图，如图3-5(a), 3-5(b)所示。当距离较远时，

运动畸变的作用会被放大，然而从图中可以看出，经过畸变矫正后的点云，仍然拥有较为不错的细节丰富程度。

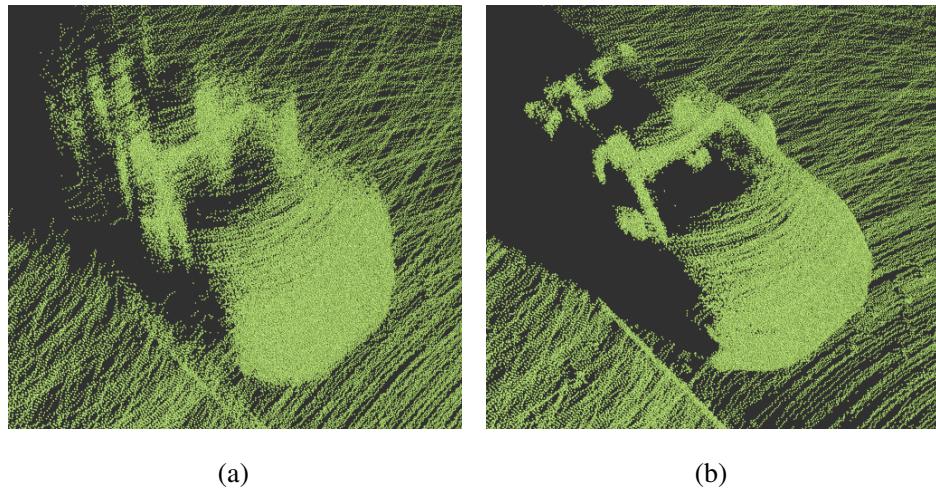


图 3-4 多帧融合后的轿车点云(a)矫正畸变前;(b)矫正畸变后

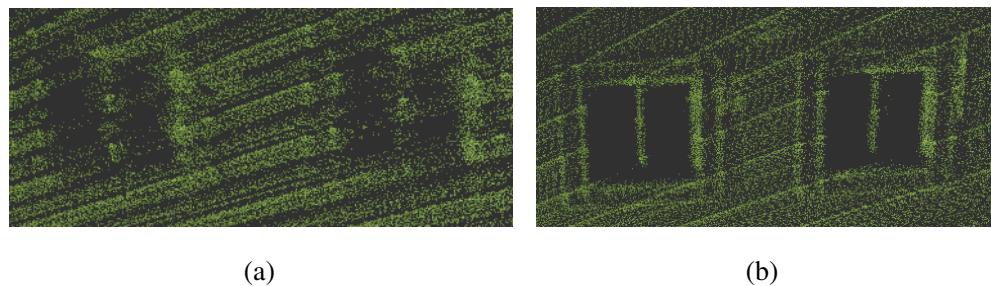


图 3-5 多帧融合后的窗户点云(a)矫正畸变前;(b)矫正畸变后

## 3.2 激光雷达与相机的标定

自动驾驶无人机是一个多传感器的系统，多传感器信息的融合可以使整个无人机系统的决策更加智能。激光雷达虽然能够获取较为精确的点云信息，然而点云信息只包含了三维距离信息。而相机可以通过图像获得大量信息诸如颜色、纹理信息等，但是其受光照与天气条件影响严重，并且从单目图像中无法获取三维结构信息。为了同时收集三维信息与物体的颜色与纹理信息，激光雷达与相机经

常进行传感器的数据融合，来为多传感器系统提供更稳定的数据支持。为了进行数据的融合，首先得知道相机坐标系与激光雷达坐标系之间的旋转与平移关系，因此，激光雷达与相机的标定就显得尤为重要了。

本章节后续将介绍一种文献<sup>[23]</sup>提到的，利用两张贴有ArUco Marker<sup>[24]</sup>的标定板所提供的3d-3d特征匹配的方法，来进行相机与激光雷达的标定。

### 3.2.1 标定板

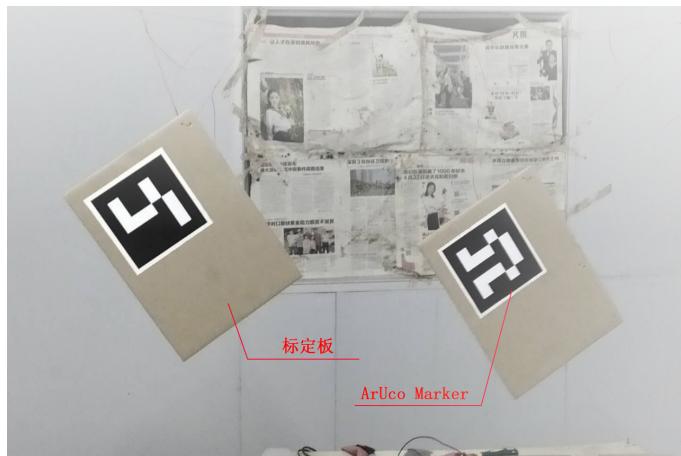


图 3-6 标定环境

本文参照文献<sup>[23]</sup>提到的方法，制作了两块长约40cm，宽约27cm的长方形硬纸板材质的标定板，并且将两块ArUco Marker粘在硬纸板的固定位置上，如图3-6所示。虽然一块标定板已经可以得到四组3d-3d匹配点来解决标定问题，本文仍然采用了两块标定板，目的是构造多于四对的匹配点来减小标定误差。

### 3.2.2 相机坐标系中的三维特征点提取

ArUco markers是一种经过特定编码的二维码图案，用以实现对二维码自身的定位与畸变矫正。更多细节可以参考文献<sup>[24]</sup>。该文献提出，通过特定的机器视觉算法检测到marker的四个角点后，可以对marker上的二维码进行解码运算，进而求得二维码的id与四个角点的顺序。而通过输入marker的边长后，还能够通过PnP<sup>[25]</sup>求解出相机坐标系到marker自身坐标系的转换。

在本文中，将ArUco marker粘在硬纸板的矩形标定板上，并且测量得出硬纸板的四条边长以及marker在硬纸板中的位置，即可得到硬纸板的四个角点在marker坐标系中的位置。而本文通过ROS中aruco\_ros以及aruco\_mapping<sup>[26]</sup>两个程序包可

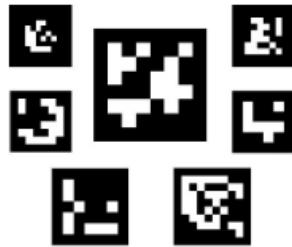


图 3-7 ArUco markers

以检测ArUco marker的位置，进而得到相机坐标系到marker坐标系的转换，从而得到相机坐标系下的硬纸板的四个角点的位置。相机坐标系下角点位置的计算公式为

$$\begin{bmatrix} x_{camera} \\ y_{camera} \\ z_{camera} \\ 1 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & t_x \\ R_{21} & R_{22} & R_{23} & t_y \\ R_{31} & R_{32} & R_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{aruco} \\ y_{aruco} \\ z_{aruco} \\ 1 \end{bmatrix} \quad (3-2)$$

其中， $x_{camera}$ 是相机坐标系中角点的坐标， $x_{aruco}$ 是ArUco marker坐标系中的角点坐标，而上式中的 $[R|t]$ 矩阵则是相机坐标系相对于marker坐标系的欧氏变换矩阵。

### 3.2.3 LiDAR坐标系中的三维特征点提取

本文所参照的标定方法，其在LiDAR坐标系中的三维特征点提取是通过直线拟合的方法进行的。如图3-8所示。图中显示的点为激光雷达的点云投影到相机图像上之后进行边缘检测后所形成的点。在得到该幅图像后，需要手动框选标定板的每条边上的点，随后标定程序会对这些点进行直线拟合，每两条直线的交点即为所求的LiDAR坐标系中的三维点。

### 3.2.4 坐标系欧氏变换求解

在得到两个坐标系下的三维特征匹配点之后，两个坐标系之间的 $[R|t]$ 欧氏变换 Closest Point, ICP) [27]算法求得。假设 $P, Q$ 为 $\mathbb{R}^3$ 中的一组对应点，ICP算法尝

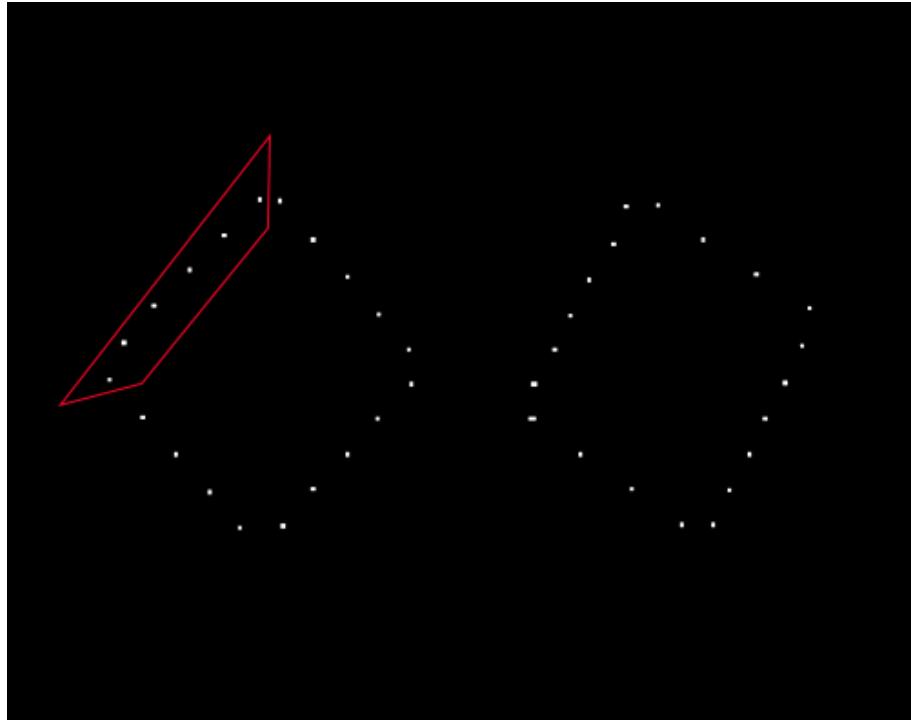


图 3-8 直线拟合提取特征点

试使经过欧氏变换后的P点集与Q点集的重合误差最小。其求解问题可以表述为式3-3所示。

$$\arg \min_{R \in SO(3), t \in \mathbb{R}^3} \|(RP + t) - Q\|^2 \quad (3-3)$$

一般来说，ICP问题认为对于点集P中的每个点，在点集Q中与之对应的点为其距离最近的点，依次确认匹配关系后，该算法通过减小两个点集的欧氏距离来对齐两个点集。

ICP的方法找到两组点的匹配关系，在没有良好的初始值的情况下，极易发生误匹配，导致算法最后没办法收敛到一个正确的解。考虑到在本文的标定环境下，两组点的对应关系的是已知的，则两组点之间的 $[R|t]$ 即存在一个闭式解。Kabsch算法<sup>[28][29]</sup>提供了找到两组对应点间的旋转矩阵的闭式解的方法，而两组点间的平移量可以在进行旋转对齐后求得。下面参照文献<sup>[29]</sup>，介绍Kabsch算法的主要步骤。

首先假设旋转已知，求两个点集P与Q之间的平移量，设

$$F(t) = \sum_{i=1}^n \|(RP_i + t) - Q_i\|^2 \quad (3-4)$$

对上式进行求导，令两边等于0，得

$$\frac{\partial F(t)}{\partial t} = 2 \sum_{i=1}^n ((RP_i + t) - Q_i) = 0 \quad (3-5)$$

因为

$$\frac{\partial F(t)}{\partial t} = 2R \sum_{i=1}^n P_i + 2nt - 2 \sum_{i=1}^n Q_i \quad (3-6)$$

可得

$$t = \frac{1}{n} \sum_{i=1}^n Q_i - R \frac{1}{n} \sum_{i=1}^n P_i \quad (3-7)$$

$$t = \bar{Q} - R\bar{P} \quad (3-8)$$

将式3-8的结果代入式3-4中，有

$$R = \arg \min_{R \in SO(3)} \sum_{i=1}^n \| (R(P_i - \bar{P}) - (Q_i - \bar{Q})) \|^2 \quad (3-9)$$

令

$$X = P_i - \bar{P}, X' = RX, Y = Q_i - \bar{Q}$$

则目标函数可化简为

$$\sum_{i=1}^n \|X'_i - Y_i\|^2 = Tr((X' - Y)^T (X' - Y)) \quad (3-10)$$

其中  $X'_i, Y_i$  是点集  $X', Y$  中的点。使用矩阵的迹的性质，上式可化简为

$$Tr((X' - Y)^T (X' - Y)) = Tr(X'^T X') + Tr(Y^T Y) - 2Tr(Y^T X) \quad (3-11)$$

考虑到  $R$  是一个旋转矩阵，旋转矩阵必定正交且行列式为1，也就是说， $\|X'_i\|^2 = \|X_i\|^2$ ，则有

$$Tr((X' - Y)^T (X' - Y)) = \sum_{i=1}^n (|X_i|^2 + |Y_i|^2) - 2Tr(Y^T X) \quad (3-12)$$

可知  $\sum_{i=1}^n (|X_i|^2 + |Y_i|^2)$  项与旋转矩阵  $R$  无关，将其从目标函数中消去，有

$$R = \arg \min_{R \in SO(3)} Tr(Y^T X') \quad (3-13)$$

将  $X' = RX$  代入，并利用矩阵迹的性质，有

$$Tr(Y^T X') = Tr(Y^T RX) = Tr(XY^T R) \quad (3-14)$$

对 $XY^T$ 进行SVD分解，有 $XY^T = UDV^T$ ，则

$$Tr(XY^T R) = Tr(UDV^T R) = Tr(DV^T RU) = \sum_i^3 d_i v_i^T R u_i \quad (3-15)$$

令 $M = V^T RU$ ，则 $M$ 由正交矩阵相乘而得，故易知其亦为正交矩阵，并且 $\det(M) = \pm 1$ 。因此 $M$ 的每个列向量的模均为1，列向量的每个元素均小于等于1。则有

$$Tr(XY^T R) = \sum_i^3 d_i M_{ii} \leq \sum_i^3 d_i \quad (3-16)$$

令上式值最大，则得到 $M_{ii} = 1$ ，因此 $M = I$ ，为单位矩阵。有

$$M = I \Rightarrow V^T RU = I \Rightarrow R = VU^T \quad (3-17)$$

需要注意的是， $R$ 应当是一个旋转矩阵，也就是说， $R \in SO(3)$ ，所以应当确保 $\det(R) = +1$ 。如果由上式得到的 $R$ ，其特征值为-1，则其为不满足条件的解，需要找到使 $Tr(Y^T X')$ 第二大的 $R$ ，即

$$Tr(Y^T X') = d_1 M_{11} + d_2 M_{22} + d_3 M_{33} \text{ where } d_1 \geq d_2 \geq d_3 \text{ and } |M_{ii}| \leq 1 \quad (3-18)$$

在上式中，当 $M_{11} = M_{22} = 1$ 且 $M_{33} = -1$ 时，该式值第二大。将上述情况考虑进去，则 $R$ 的闭式解为 $R = UCV^T$ ，其中 $C$ 为一个矫正矩阵，

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & sign(\det(UV^T)) \end{bmatrix}$$

### 3.2.5 多帧坐标系变换结合

在实验中发现，由相机图像求得的三维特征点比较稳定，而由于激光雷达本身存在传感器误差，其通过直线拟合求交点得到的三维角点不够稳定，在相机与激光雷达位置均固定的情况下，每次测得的激光雷达测得的三维特征角点在会有一个较小的位移，而这样的位移会对用Kabsch算法得到的解产生较大的误差。

为了减小因激光雷达传感器误差而造成的三维特征角点不准确的问题，本文在相机与激光雷达位置固定的情况下，针对多帧图像与点云进行了匹配运算，对 $N$ 次结果进行求取平均值。对平移量的平均值求取公式为

$$\bar{t} = \frac{1}{N} \sum_{i=1}^N tvec_i$$

而对于旋转量的求取平均值，先将上文求得的旋转转为四元数 $rvec_i$ 形式，再对四元数求取平均值。

$$r = \frac{1}{N} \sum_{i=1}^N rvec_i$$

$$\bar{r} = \frac{r}{\|r\|}$$

通过对多帧激光雷达点云与相机进行三维点匹配得到的欧氏变换求取平均值，能够有效的抑制因激光雷达的传感器误差而造成的标定误差。

### 3.3 点云与图像的融合

本章所述的传感器设置如图3-9所示。其中激光雷达坐标系到相机坐标系的欧氏变换已由第二章介绍的标定方法得到。在介绍如何将激光雷达点云投影到相机图像上之前，本文将先介绍相机的成像原理，这是点云投影的理论基础。

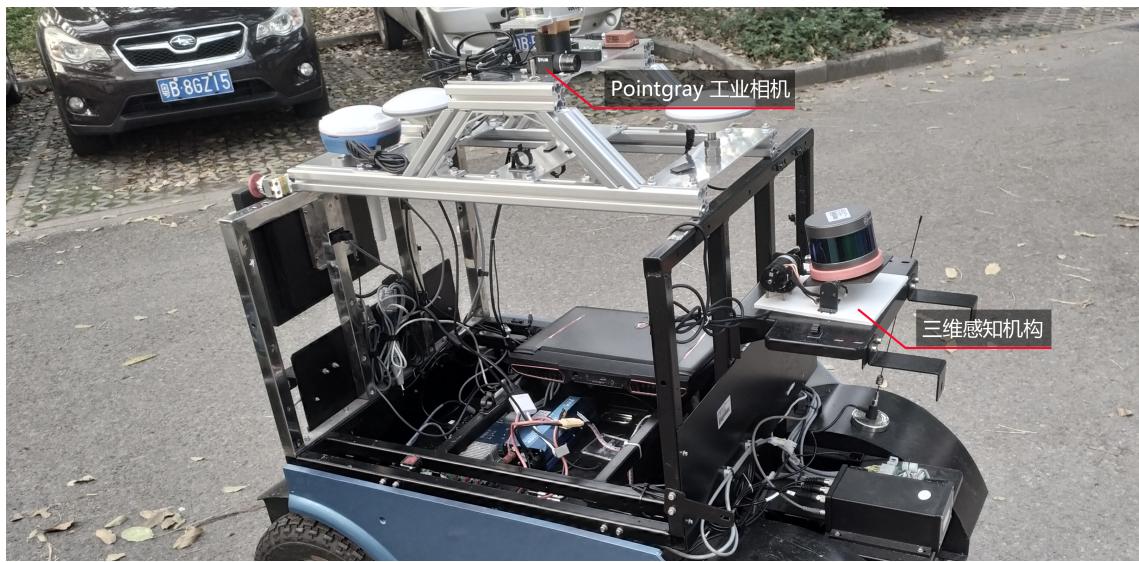


图 3-9 传感器设置

#### 3.3.1 相机成像原理

所谓的相机成像，就是相机将三维空间中的坐标映射到二维图像平面的过程。这一映射可以有许多数学模型去解释，最简单也最为常用的就是针孔成像模型，如图3-10所示。现在对该模型进行建模。设图中 $O - x - y - z$ 为相机坐标系。通

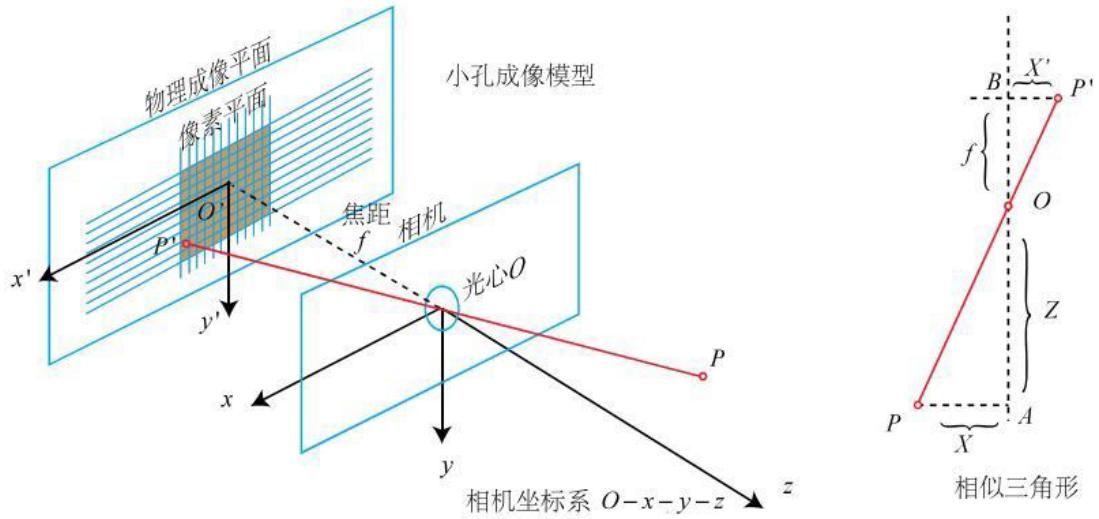


图 3-10 相机针孔成像原理

常认为相机坐标系  $Z$  轴指向相机的前方,  $X$  轴指向右方,  $Y$  轴指向下方。设真实空间中有一三维点  $P$ , 经过相机的小孔  $O$  成像后, 成像点落在相机的成像平面  $P'$  处, 设  $P$  点的坐标为  $[X, Y, Z]^T$ ,  $P'$  点的坐标为  $[X', Y', Z']^T$ , 并且设相机的成像平面到模型中的针孔的距离为  $f$  (焦距), 则根据三角形的相似关系, 有

$$\frac{Z}{f} = -\frac{X}{X'} = -\frac{Y}{Y'} \quad (3-19)$$

整理得

$$\begin{cases} X' = f \frac{X}{Z} \\ Y' = f \frac{Y}{Z} \end{cases} \quad (3-20)$$

实际上, 在相机中最后得到的是一个个的像素, 设在相机成像平面上存在一个像素平面  $O-u-v$ ,  $P'$  在像素平面的坐标为  $[u, v]^T$ 。像素坐标系通常定义其原点  $O'$  在图像的左上角, 其  $u$  轴与  $v$  轴的方向与相机坐标系中的  $X$  轴与  $Y$  轴相同。像素坐标系相对于相机坐标系, 相差了一个平移与缩放。假设相机坐标系在  $u$  轴上缩放了  $\alpha$  倍, 在  $v$  轴上缩放了  $\beta$  倍, 同时像素坐标系相对于相机坐标系的原点平移了  $[c_x, c_y]^T$ 。则  $P'$  在成像平面上的坐标与其像素坐标  $[u, v]^T$  之间的关系为

$$\begin{cases} u = \alpha X' + c_x \\ v = \beta Y' + c_y \end{cases} \quad (3-21)$$

将上式代入式3-20中，并设 $\alpha f = f_x$ ,  $\beta f = f_y$ , 则有

$$\begin{cases} u = f_x \frac{X}{Z} + c_x \\ v = f_y \frac{Y}{Z} + c_y \end{cases} \quad (3-22)$$

将上式写为矩阵的形式，则有

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z} \begin{bmatrix} fx & 0 & cx \\ 0 & fy & cy \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \triangleq \frac{1}{Z} K P \quad (3-23)$$

其中矩阵  $\begin{bmatrix} fx & 0 & cx \\ 0 & fy & cy \\ 0 & 0 & 1 \end{bmatrix}$  称之为相机的内参，通常，相机的标定就是为了得出相机的内参矩阵。本文在融合激光与相机之前已通过ROS的标定程序得到了相机的内参 $K$ 。

### 3.3.2 激光雷达点云的投影

式3-23中， $P$ 为相机坐标系内的一点，所以在根据第二章得到的标定方法得出激光雷达与相机坐标系的变换矩阵 $T$ 后，先将激光雷达点云转换到相机坐标系中，再通过内参矩阵投影到图像上，即

$$P_{uv} = \frac{1}{Z} K T P_{LiDAR} \quad (3-24)$$

其中 $P_{LiDAR}$ 表示激光雷达坐标系中的点云中的三维点。

本文利用前文提到的三维感知机构进行了多帧点云的融合配准，并将融合后的点云投影到了相机的图像上，点云投影效果如图3-11所示。其中投影点的颜色越深，代表其距离相机的位置越近；颜色越浅，代表离相机的距离越远。在将激光雷达点云投影到图像之后，便可以利用YOLO的检测结果对点云进行分割与分类。

## 3.4 本章小结

本章介绍了什么是激光雷达点云配准与运动畸变，并且提出了插值矫正运动畸变的方法：在对点云进行多帧融合时，利用点云传输的特性，将点云在时间上为84个Packet单独发布，同时线性插值得到这些Packet的时间戳上的编码器



图 3-11 激光雷达点云的投影

角度值，将这些Packet点云依据编码器的角度配准到参考坐标系中，并且将所有的Packet点云进行叠加配准后发布为新的点云。由于对每帧点云多进行了84次旋转角度的插值，从点云图案上看，本文提到的方法极大地改善了运动畸变对点云配准的影响，完成了矫正运动畸变的目的。同时，本章还介绍了一种利用特制的标定板在相机与激光雷达坐标系中寻找三维特征匹配点的方法来进行相机的标定，并详细介绍了如何利用三维匹配点来计算出两个点集所在的坐标系之间的欧氏变换（Kabsch算法）。本文还根据得到的欧氏变换矩阵，将激光雷达的点云与相机的图像进行了融合。在后续的章节，本文将基于传感器融合后的结果进行三维障碍物的分割与分类。

## 第4章 基于视觉与三维点云融合的三维障碍物检测方法

随着深度学习领域内的卷积神经网络在目标检测任务中广泛运用，基于视觉的目标检测的深度学习算法大行其道。在大规模目标检测挑战赛ILSVRC<sup>[1]</sup> 中，基于深度学习的目标检测算法连续六年（2012-2017）取得了优越的表现。2017年，ILSVRC的夺冠深度学习网络SENet在目标检测问题上的错误率仅为2.251%，亦同时宣告了基于图像的目标检测任务基本被攻克。

然而，基于视觉的目标检测无法得到物体的三维位置信息，并且受光照影响严重。为了解决这些问题，本章后续部分将提出一种激光雷达与相机的多模态传感器融合技术，其能够在基于视觉的目标检测算法检测到目标时，能够根据激光雷达的点云得到障碍物的三维位置信息，同时根据激光雷达的点云信息反馈于目标检测任务的识别上，提升视觉在光照条件不够良好的情况下的检测准确率。

### 4.1 YOLO—一种实时目标检测网络

在介绍相机图像与激光雷达点云的融合之前，本文将先介绍本文使用的视觉目标检测算法。本文使用YOLO(You Only Look Once)<sup>[30]</sup>算法作为视觉图像上的目标检测算法。

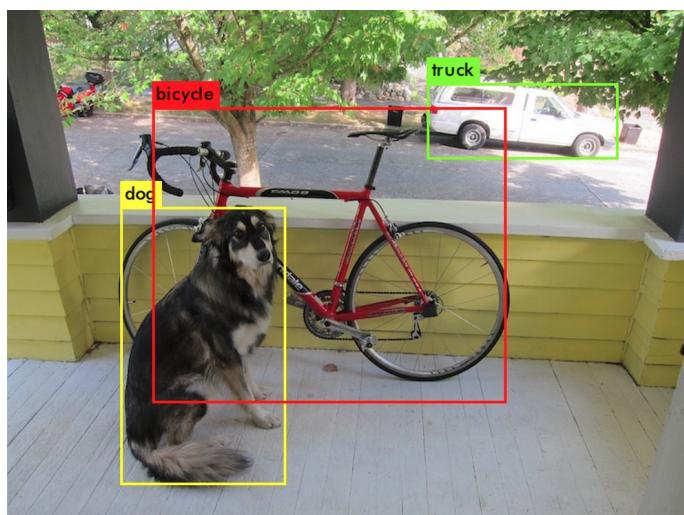


图 4-1 YOLO—一种实时目标检测网络

早期的目标检测算法通过提取图像的一些特征（例如Haar、SIFT、HOG等），运用DPM（Deformable Parts Model）模型，并且通过滑动窗口来预测具有较高得分的区域的策略来进行目标检测。这种传统的方法不仅相当耗时，而且检测准确率也不是很高。

随后出现了基于object proposal策略的方法，相比于滑动窗口这种类似于穷举的方法，该方法大大减少了运算量，同时定位精度也得到了很大的提高。结合13年兴起的卷积神经网络后，Object detection的性能得到了质的飞跃。

然而，诸如R-CNN、Faster R-CNN这样的网络，因为候选区域较多、网络运算量较大，因而很难在GPU性能不够强劲的移动机器人场景中做到实时检测。而本文采用的YOLO网络将目标检测问题转化为一个回归问题。给定图像 $I$ ，YOLO网络直接在图像的多个位置上回归出识别目标的边界框（bounding box）以及其类别。

YOLO没有选择滑动窗口或者提取proposal的策略来训练网络，而是直接将整张图作为网络的输入，既极大地提升了运算速度，亦很好的区分了图像的前景与背景区域，而使用proposal策略的Fast R-CNN则经常将背景区域误识别为目标区域。

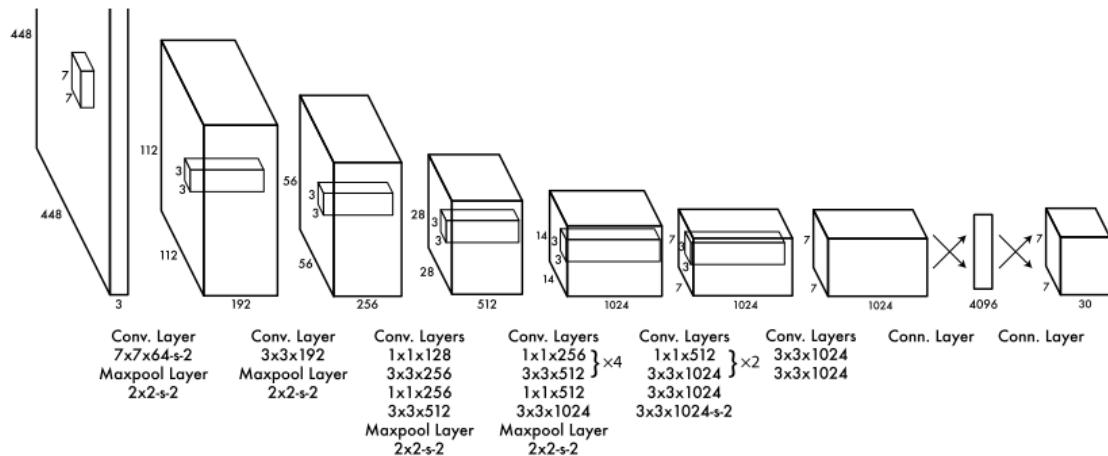


图 4-2 YOLO网络结构

图4-2为YOLO的网络结构，从图中可以看出，该目标检测网络拥有24层卷积层，随后用两层全连接层对结果进行回归输出。

由于其简介的网络结构设计，使得其在实时性能上表现卓越。YOLO能够在每秒推断（inference）45帧的情况下仍能够保证和Faster R-CNN同样的准确率。基

于其良好的实时性与较为优秀的目标检测准确率，本文将使用YOLO作为视觉与三维点云融合的目标检测基准方法，并利用激光雷达的点云信息为其检测结果提供三维位置信息并提高准确率。

## 4.2 基于YOLO的视觉、三维点云结合的三维障碍物检测

### 4.2.1 点云前景与背景的分割

本文首先利用YOLO对相机图像进行了目标检测的推断。YOLO的检测效果如图4-3所示，其输出为多个边界框与识别的物体的类别，以及YOLO对推断结果的置信度（confidence）。图4-3为设置YOLO检测的置信度阈值为0.8时输出的结果（即不输出置信度低于0.8的检测结果）。



图 4-3 YOLO检测结果

将点云按照上文投影到图像中，则点云中的每个点 $P'$ 都有一个唯一的像素坐标 $[u, v]^T$ 与其对应。剔除像素坐标不在YOLO检测的边界框内的点，便得到了基于YOLO检测结果的点云分割与分类结果，如图4-4(a)所示。

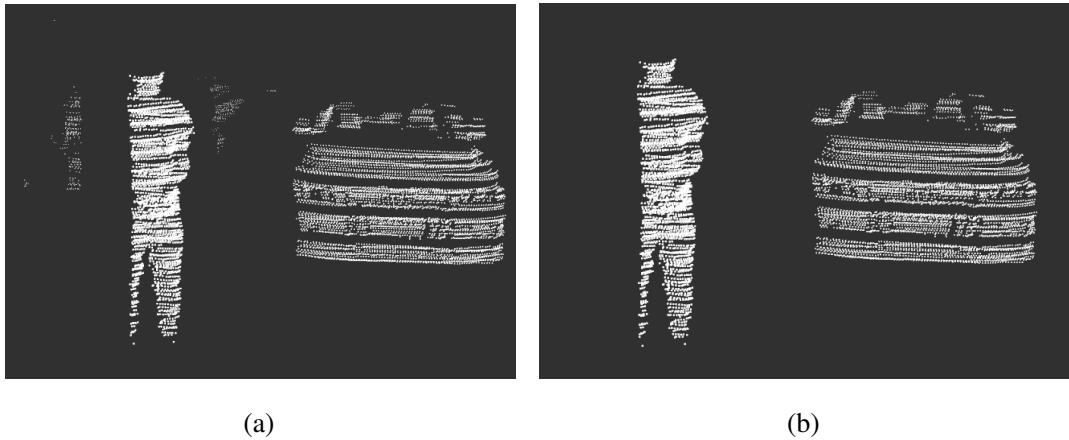


图 4-4 点云前景与背景的分割(a)原始点云;(b)Kmeans分类后的结果

从图4-4(a)中可以看出，尽管按照上述结果的确将人与汽车的点云分割了出来，然而由于YOLO得出的边界框内除了识别的目标，还有一些背景物体，体现在点云中就是，除了表示人与轿车的点云，还有一些人与汽车后的灌木的点云，其在图像上的投影也在候选框里。这些点云如果不加以去除，会对目标物体的三维位置的确定造成较大的不利影响。

考虑到目标点云与背景点云转换到相机坐标系后，在Z轴方向上有较大的距离，本文采用K-means算法<sup>[31]</sup>，对投影后在同一边界框内的点云进行聚类分割。

K-means算法的目的是，把 $n$ 个点划分到 $K$ 个聚类中，使得每个点都属于离它最近的均值（称之为聚类中心）对应的聚类，以之作为聚类标准。其算法流程为：

- (1) 假设分为 $K$ 类；
- (2) 任意随机出 $K$ 个点，认为该 $K$ 个点为聚类的中心点；
- (3) 遍历每一个点 $x$ ，计算其与上一步得出的 $K$ 个点的欧式距离 $D(x)$ ，将该点归为使得 $D(x)$ 最小的聚类中；
- (4) 计算每个聚类内所有点的平均位置，认为其为新的聚类中心；
- (5) 重复3和4，直到 $K$ 个聚类中心被选出来；
- (6) 利用得到的新的 $K$ 个点，继续重复3-5的步骤，直到每个点的分类结果不变，或者循环达到迭代次数上限。

对于图4-4(a)中的每个物体边界框内的点云，假设其可以被分为两类：前景与背景。设原始三维点集为 $P_3$ ，将所有点映射到相机坐标系的Z轴上，成为一个新

的一维点集 $P_1$ ，并对 $P_1$ 进行K-means聚类，其中 $K = 2$ 。则聚类得到的两类点中，中心点Z轴坐标较小的即为前景，也就是YOLO检测的物体的点云。图4-4(b)为K-means聚类提取得到的结果。从图中可知，采用K-means算法较好的将前景与背景分割开来，剔除了无关的背景点，只保留了与YOLO检测到的目标有关的点。

## 4.2.2 目标三维坐标的计算与检测结果的优化

在得到了与YOLO的检测结果相关的目标物体的点云后，物体的三维坐标可以由求点集的中心点坐标而得。同时，为了更好的表示三维物体的检测结果，本文还将检测得到的物体用长方体包围盒去拟合，如图4-6所示。包围盒的顶点由点集的 $X_{min}, Y_{min}, Z_{min}$ 以及 $X_{max}, Y_{max}, Z_{max}$ 决定，其中 $X_{min}$ 表示点集中X轴坐标最小的点的X坐标， $X_{max}$ 表示点集中X轴坐标最大的点的X坐标。

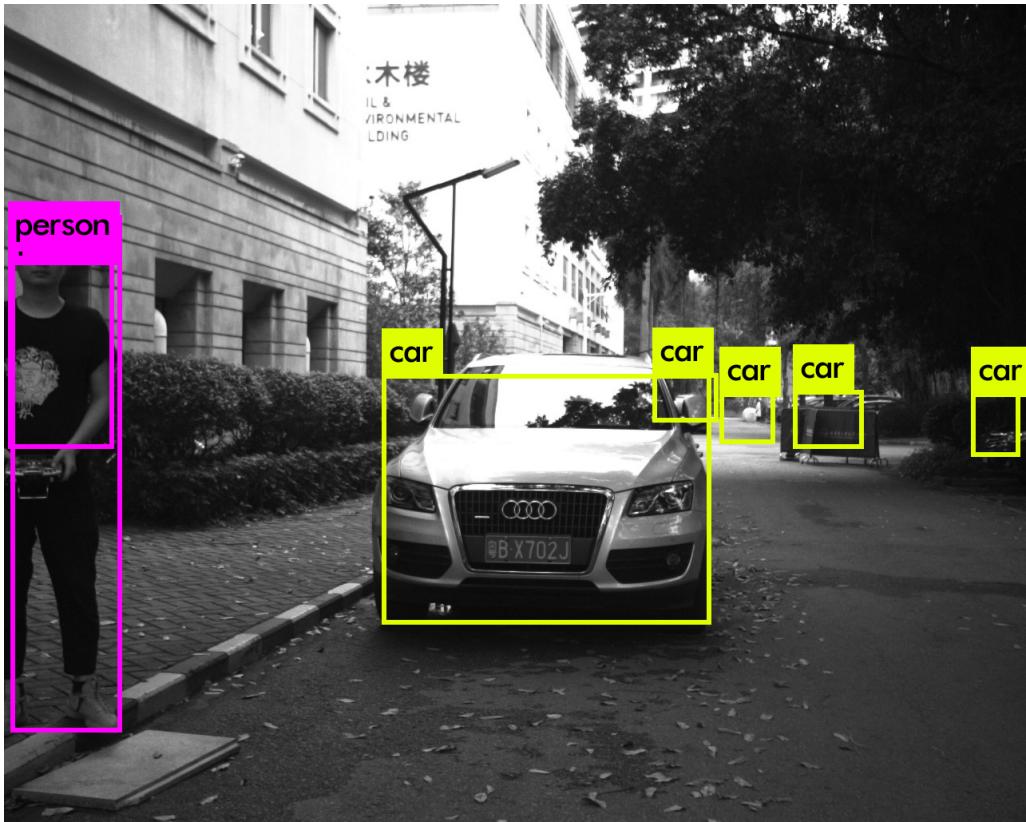


图 4-5 YOLO在低置信度阈值下的误检测

上文提到，图4-3中的结果是将YOLO检测结果的置信度阈值设为0.8后的检测结果。而将YOLO检测结果的阈值设为0.4后，YOLO就出现了许多误检测，如图4-5所示。由于光照等情况的影响，很多时候YOLO的检测结果置信度都不会高

于0.8，而降低置信度阈值又会造成误检测的结果增多。本文提出了一种在YOLO低置信度阈值下，融合激光雷达点云位置信息来去除误检测结果的方法。

从图4-5中可以看出，大部分误检测的区域，其边界框对应的三维空间区域可能没有点云，或者其三维位置信息与边界框面积不相符合。为了去除误检测结果，本文首先假设已知每个待检测的目标物体的最小投影矩形。举例来说，对图4-5来说，认为汽车在相机坐标系中的最小投影矩形为 $2.4m \times 1.8m$ ，即对于轿车而言，至少有 $2.4m \times 1.8m$ 的面积在相机视野范围内，那么，去除误检测的策略流程可以表述为：

- (1) 如果YOLO检测得到的边界框内没有点云，则认为发生了误检测。
- (2) 如果YOLO检测得到的边界框内有点云，则计算点云中所有点的平均坐标 $X_{mean}, Y_{mean}, Z_{mean}$ 。
- (3) 在相机坐标系中构建长方形平面，其顶点坐标为

$$\begin{aligned}
 & (X_{mean} - X_{hypo}, Y_{mean} - Y_{hypo}, Z_{mean}) \\
 & (X_{mean} + X_{hypo}, Y_{mean} - Y_{hypo}, Z_{mean}) \\
 & (X_{mean} - X_{hypo}, Y_{mean} + Y_{hypo}, Z_{mean}) \\
 & (X_{mean} + X_{hypo}, Y_{mean} + Y_{hypo}, Z_{mean})
 \end{aligned} \tag{4-1}$$

其中， $X_{hypo}, Y_{hypo}$ 为上文提到的最小投影矩阵的长与宽。

- (4) 将该长方形平面通过内参矩阵投影到相机图像中，得到一个估计的长方形框选面积 $S_{hypo}$ 。
- (5) 将 $S_{hypo}$ 与YOLO检测出的边界框的面积 $S_{yolo}$ 做比较，认为不满足约束条件

$$0.5 \times S_{hypo} \leq S_{yolo} \leq 1.5 \times S_{hypo} \tag{4-2}$$

的检测结果为误检测。

经过该点云图像融合去除误检测策略后，三维物体检测结果如图4-6所示。其中误检测的部分用红色矩形标出，而正确的检测部分利用上文提出的长方体包围盒表示出来。可以看出，通过该策略能够有效的排除YOLO在低置信度阈值下的误检测，并且利用激光雷达的信息，能够很好的计算出检测的目标的三维位置。

### 4.3 本章小结

本章介绍了YOLO，一种实时的视觉目标检测方法，该方法最大的优点在于其实时性，能够在较高的识别准确率下，达到每秒推断45帧的速度。并且之后融合了相机与激光雷达的信息，对三维障碍物进行了分割与分类。利用YOLO的推断结果，首先对激光雷达点云做了前景的提取与分割；随后，根据点云信息，计算得出了检测的目标的三维位置信息，并计算出物体的三维包围盒；最后，针对YOLO在低置信度阈值下容易出现误检测的问题，本文利用点云信息来验证YOLO的推断结果，并且甄别与去除了误检测的输出。从结果中可以看出，融合视觉与三维点云的信息，能够较好的对三维物体进行检测与识别，为无人驾驶技术中的环境感知与路径规划提供了更多的信息。

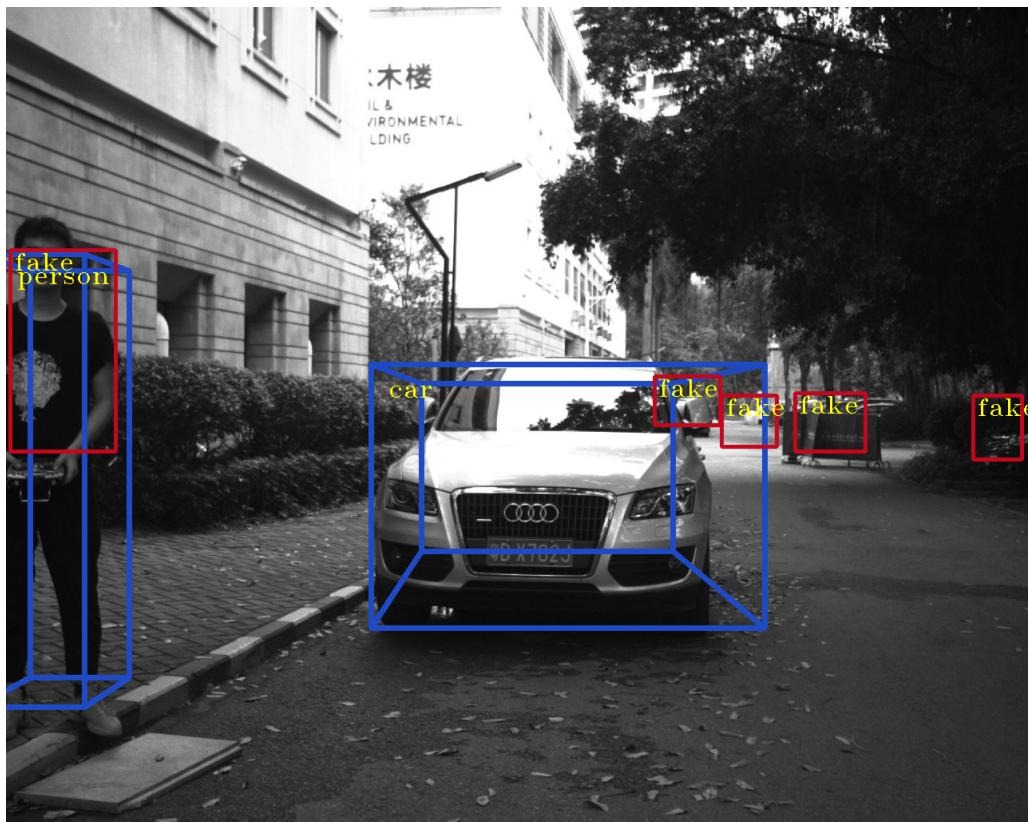


图 4-6 进行多传感器融合增强后的YOLO检测结果

## 第5章 实验验证与结果分析

为了验证前文提出的三维感知传感器机构对点云的融合配准效果，本章首先针对该机构进行了Gazebo仿真实验，随后针对融合配准后的点云进行了三维建图与三维障碍物分割的实验，并且针对激光雷达与相机标定的实验结果进行了评估。

### 5.1 Gazebo下的三维感知机构仿真

Gazebo是一个功能强大的三维物理仿真平台，具备强大的物理引擎、高质量的图形渲染、方便的编程与图形接口，最重要的还有其具备开源免费的特性。Gazebo支持显示逼真的三维环境，包括光线、纹理、影子等。它还支持传感器数据的仿真，同时可以仿真传感器噪声。

本文利用Gazebo对三维感知机构的运动以及点云融合进行了仿真，如图5-1(a)以及图5-1(b)所示。

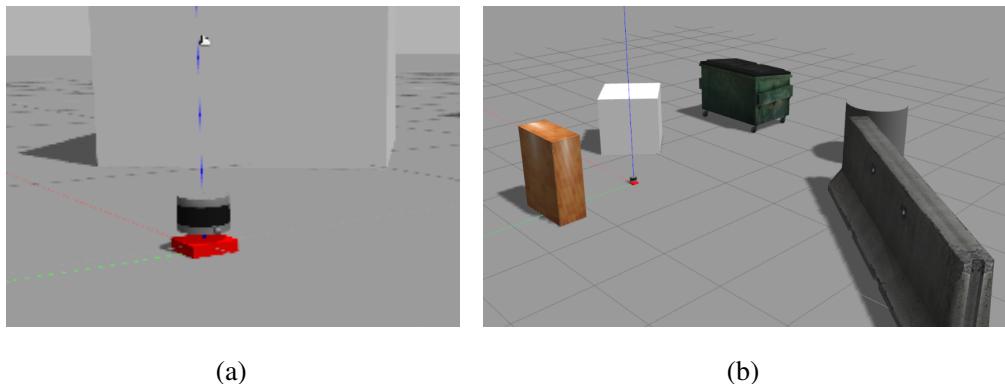


图 5-1 Gazebo仿真(a)三维感知机构的模型;(b)仿真环境

三维感知机构的Gazebo模型如图5-1(a)所示。由于在加上了ROS control的控制器后，利用单个link连接雷达，控制link做旋转运动就能够使得激光雷达在俯仰角上做来回的正弦运动，而进行Gazebo仿真的目的主要是为了仿真出激光雷达在往复运动情况下的点云配准情况，因此本文没有在Gazebo中设计曲柄摇杆机构。在

仿真实验中，激光雷达传感器使用的是Velodyne 的VLP-16型激光雷达。Gazebo支持传感器数据的仿真，按照第三章的配准融合策略融合后的点云如图5-2所示。

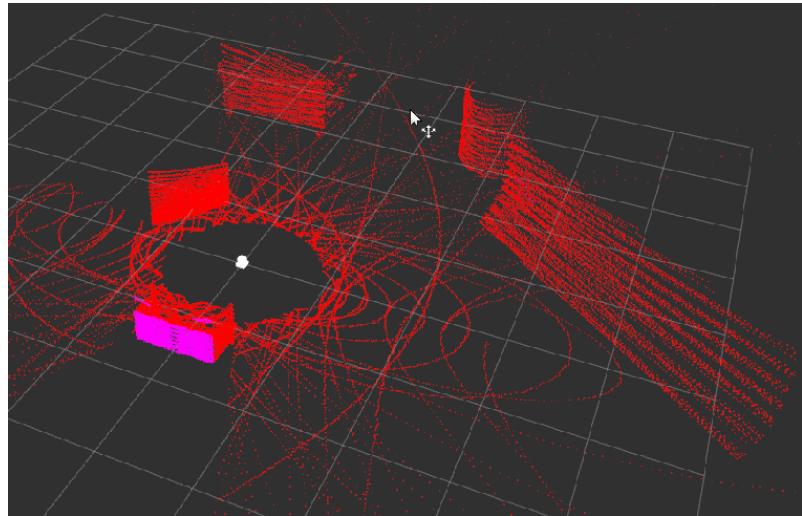


图 5-2 Gazebo仿真得到的融合点云

从图中可以看出，经仿真得到的点云基本能够反应仿真环境下障碍物的分布情况，并且仿真得到的点云分辨率比低线数激光雷达直接获得的点云分辨率要高得多。

## 5.2 三维感知机构结合里程计信息构建三维地图

前文提出的三维感知机构，其实验场景都是在机构位置静止不变的情况下得到的。而当机构架设在小车上时，由于小车相对于世界坐标系有一个运动，如果仍然采用之前的融合策略进行点云的输出，则输出的点云会有小车运动方向上的畸变。因此，在运动的小车上进行三维感知时，需要结合里程计的信息，计算出机构在小车运动方向上的位移，借此消除点云在小车运动方向上的畸变。同时，可以根据小车自身坐标系到里程计坐标系的变换，利用该三维感知机构构建基于小车里程计坐标系的三维地图。

本文结合图3-9所示的传感器设置，利用小车的里程计对小车行驶过程中的点云进行了采集、融合与构建了三维地图。本文利用了ROS的TF坐标变换<sup>[32]</sup>的设计，在结合里程计信息的同时，避免了繁琐的插值运算。

ROS的TF是一种这样的设计：TF是一个让使用者随时间跟踪多个坐标系的功能包。其数据结构类型为树状结构，能够根据时间缓冲并维护多个坐标系之间的

坐标变换关系，可以帮助使用者在任意的时间点请求任意的坐标系之间的变换。

在本章的实现中，首先认为机构自身的坐标系为 $baselink$ ，而激光雷达所在的坐标系为 $lidar$ ，根据磁编码器返回的角度 $\alpha$ ，发布 $baselink$ 坐标系到 $lidar$ 坐标系的变换，认为其只有绕y轴旋转，即偏航角的变化，而机构的往复运动没有造成两个坐标系之间的位移，故其欧氏变换矩阵为

$$\begin{bmatrix} \cos \alpha & 0 & \sin \alpha & 0 \\ 0 & 1 & 0 & 0 \\ -\sin \alpha & 0 & \cos \alpha & 0 \end{bmatrix}$$

随后，根据小车底层的编码器的信息，发布 $odom$ 坐标系到 $baselink$ 坐标系的变换。该欧氏变换的旋转与位移皆通过小车的编码器与阿克曼转角的角度积分得到。那么在TF树中， $odom$ 坐标系的子节点为 $baselink$ 坐标系， $baselink$ 坐标系的子节

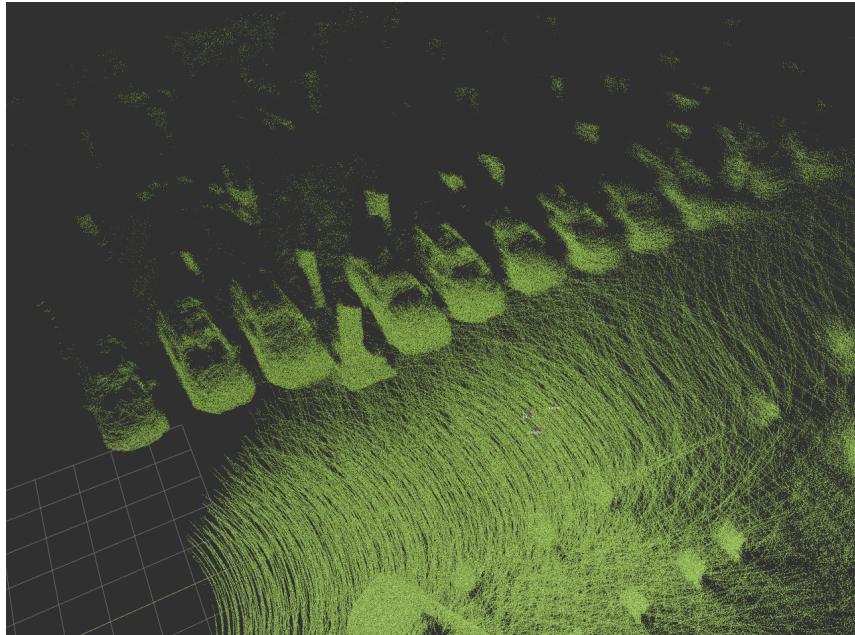


图 5-3 结合里程计生成的三维地图

点为 $lidar$ 坐标系。则当激光雷达点云产生时，可以根据激光雷达的产生的点云信息的时间戳 $t$ 来请求 $odom$ 坐标系到 $baselink$ 坐标系之间的欧氏变换。尽管有可能在时间戳 $t$ 上没有发布 $odom$ 到 $baselink$ 、 $baselink$ 到 $lidar$ 坐标系的欧氏变换，但是ROS的TF可以利用在时间戳 $t$ 前后时刻已发布的欧氏变换来进行插值得出 $t$ 时刻的变换矩阵。有了 $t$ 时刻的变换矩阵，可以很容易的将激光雷达坐标系下的点云变换到世界坐标系中。将每帧点云在世界坐标系中进行配准，则可得基于三维感知

机构生成的三维点云地图，如图5-3所示。

### 5.3 基于点云投影到深度图像上的物体分割

在得到了基于三维感知机构融合与配准的点云后，一个重要的步骤就是进行点云的分割，进而求得点云中可能为障碍物的部分。本章节的实验环境如图5-4所示。



图 5-4 实验环境

在该实验中，首先进行三维感知传感器机构对点云的融合，融合结果如图5-5(a)所示。关于点云融合配准的细节，本文已经在第三章中详细阐述，此处便不再赘述。

#### 5.3.1 地面点的去除

有关点云的聚类，实际上就是指将三维空间点中的相近的点认为是同一个物体，将其归为同一类。而从点云图像中可以看出，地面点构成了点云的绝大部分。在聚类的过程中，地面点会对聚类的结果造成较大的影响，体现在地面点将两个不同物体的点云连接了起来，从而使得聚类算法会将两个不同的物体归为一类。本文采用文献<sup>[8]</sup>提出的算法，首先对地面进行平面拟合，求出近似平面的方程，随后根据方程将平面方程以下的点去除。

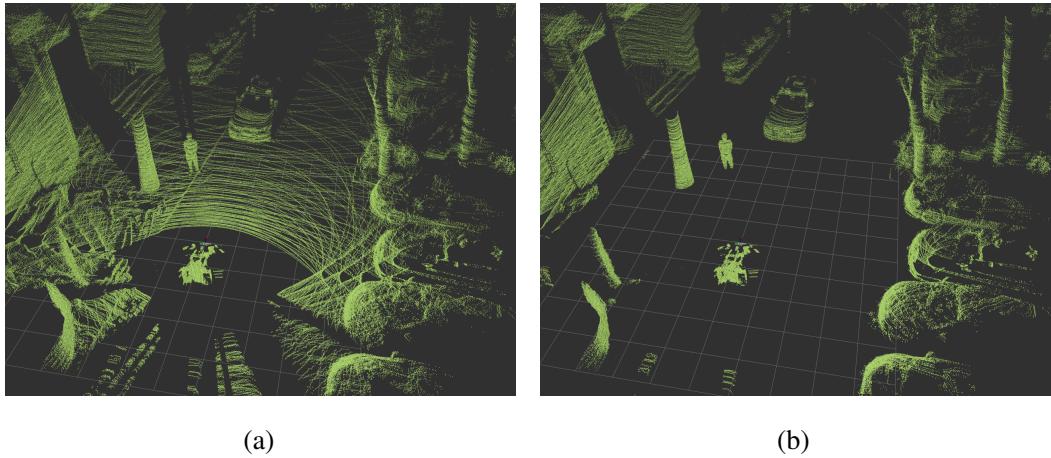


图 5-5 多帧融合后的点云(a)原始点云;(b)去除地面点后的点云

在去除地面点之前，该算法首先提出了两个假设：

- (1) 认为地面可以被近似为一个平面，即认为环境中地面是较为平整而不是有曲率的。
- (2) 认为地面点应该是点云中高度最低的一些点。

算法流程如Algorithm 2所示。该算法使用最低点作为代表平面的点（lowest point representative）。首先，该算法先提取地面点的“种子（seed）”点，即认为最低的 $N_{LPR}$ 个点中，高度小于这些点的平均值+阈值 $Th_{seeds}$ 的点为地面点。这个阈值有效的预防了传感器噪声的影响，即避免了因传感器噪声而导致的过低的点对地面的平面拟合造成影响。

为了估计地面的方程，该算法使用了一个简单的线性方程：

$$\begin{aligned} ax + by + cz + d &= 0 \\ n^T x &= -d \end{aligned} \tag{5-1}$$

其中， $n = [a, b, c]^T$ ,  $X = [x, y, z]^T$ ，该方程首先求得关于种子点的协方差矩阵，通过协方差矩阵求得平面的法向量 $n$ 。协方差矩阵 $C$ 通过下式

$$C = \sum_{i=1:|S|} (s_i - \hat{s})(s_i - \hat{s})^T \tag{5-2}$$

而得。其中 $\hat{s} \in R^3$ 是所有 $s_i \in S$ 的平均值。

协方差矩阵 $C$ 表征了种子点的离散程度，通过对对其进行奇异值分解（singular value decomposition）可得其奇异向量（singular vectors），奇异向量描述了其在三

个方向上的离散程度。考虑到地面较为平坦，其种子点在竖直方向上的离散程度比较小，则与地面垂直的法向量 $n$ 即为三个奇异向量中模最小的一个。

在得到法向量 $n$ 之后，在式5-1中，令 $x = \hat{s}$ ，求得 $d$ 。如此便得到了对地面进行拟合的平面方程。对于点云中的每个点 $p(x, y, z)$ ，求 $n^T x + d$ 的值，若大于0，则在平面之上（为非地面点）；若小于0，则在平面之下（为地面点）。

### 5.3.2 点云到深度图像的投影

正常的16线激光雷达点云每帧拥有三万个三维点，而由于本文提到的三维感知机构融合配准了多帧点云，其发布的点云每帧高达几十万个三维点。如此巨大的数据规模，如果采用常规的三维点云聚类方法，其每帧耗时将相当可观（在三维点云上的聚类算法通常时间复杂度在 $O(n\log(n))$ 以上）。因此需要一些对数据进行降采样的方法来提高算法的效率。

第二种方法是将去除地面点后的三维点云投影到地面的栅格平面上，这种方法也称为生成点云的鸟瞰图（bird's eye view）<sup>[33]</sup>。随后在二维图像上进行物体的分割。这种方法运算速度很快，适合实施运算。然而这种方法有可能不能充分分割障碍物，如果多个物体彼此比较接近，它们有可能被认为是同一个物体。这取决于给地面划分栅格时栅格的大小，所以在不同的环境下，有可能要调整不同的栅格大小来改进算法。

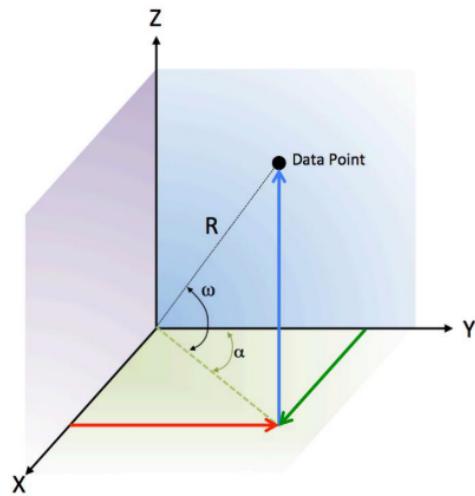


图 5-6 三维点的球坐标表示

而本文参考文献<sup>[34]</sup>提出的方法，将点云投影到深度图像（Range Image）上，并且不需要在深度图像中提取特征，而是通过一种图搜索的算法将物体进行分割与定位。也因此，其时间复杂度为O(n)级别，能够满足实时性的要求。

将点云投影为深度图像，首先要将点云中的点的表示由笛卡尔坐标系转换为球坐标系，如图5-6所示。球坐标系中的三维点 $p = (\alpha, \omega, r)$ ，其中 $\alpha$ 与 $\beta$ 的表示如图所示， $r$ 为三维点到坐标系原点的距离。则显然，点的球坐标系与笛卡尔坐标系坐标转换的关系为

$$\begin{cases} r = \sqrt{x^2 + y^2 + z^2} \\ \omega = \arcsin \frac{z}{r} \\ \alpha = \arccos \frac{y}{r \cos \omega} \end{cases} \quad (5-3)$$

求得三维点的球坐标后，就要根据三维点的 $\alpha$ 与 $\omega$ 来进行到深度图像的投影。假设深度图像的长与宽为 $w$ 与 $h$ ，并且空间中有一三维点 $p_{3d} = (\alpha, \omega, r)$ ，首先设融合后的点云中的点的最大 $\alpha$ 角为 $\alpha_{max}$ ，最大 $\omega$ 角为 $\omega_{max}$ ，则三维点 $p_{3d}$ 映射到二维点 $p_{2d} = (u, v)$ 的关系为

$$\begin{cases} \alpha_{percol} = 2\alpha_{max}/w \\ \omega_{perrow} = 2\omega_{max}/h \\ u = (\alpha/\alpha_{percol} + w/2) \mod w \\ v = -\omega/\omega_{perrow} + h/2 \end{cases} \quad (5-4)$$

其中 $\alpha_{percol}$ 与 $\omega_{perrow}$ 是深度图像中每行与每列代表的角度，而有关 $u, v$ 的变换是为了让激光雷达坐标系中正前方的点云能够投影到深度图像的中心。经过投影后得到的深度图像如图5-7所示，其中像素点的像素值为三维点的 $x$ 坐标值，像素值越大，颜色越亮，表示该像素所表征的三维点离相机越远。



图 5-7 由点云投影得到的深度图像

---

**Algorithm 2** ground plane fitting methodology for one segment of the point cloud

---

**Output:**

$P_g$ : points belonging to ground surface

$P_{ng}$ :points not belonging to ground surface

**1: Initialization:**

2:  $P$  : input point cloud

3:  $N_{iter}$  : number of iterations

4:  $N_{LPR}$  : number of points used to estimate the LPR

5:  $Th_{seeds}$  : threshold for points to be considered initial seeds

6:  $Th_{dist}$  : threshold distance from the plane

**7: Main Loop:**

8:  $P_g = \text{ExtractInitialSeeds}(P, N_{LPR}, Th_{seeds})$ ;

9: **for**  $i = 1 : N_{iter}$  **do**

10:     model = EstimatePlane( $P_g$ );

11:     clear( $P_g, P_{ng}$ );

12:     **for**  $k = 1 : \|P\|$  **do**

13:         **if**  $model(p_k) < Th_{dist}$  **then**

14:              $P_g \leftarrow p_k$ ;

15:         **else**

16:              $P_{ng} \leftarrow p_k$ ;

**17: ExtractInitialSeeds:**

18:  $P_{sorted} = \text{SortOnHeight}(P)$ ;

19:  $LPR = \text{Average}(P_{sorted}(1 : N_{LPR}))$ ;

20: **for**  $K = 1 : \|P\|$  **do**

21:         **if**  $p_k.height < LPR.height + Th_{seeds}$  **then**

22:              $seeds \leftarrow p_k$ ;  
**return**  $seeds$

---

### 5.3.3 基于深度图像的物体分割

投影后的深度图像相当于将具有相同 $\alpha, \omega$ 而 $r$ 不同的点用同一个位置的像素点来表示，因而相当于通过损失有限的信息对点云信息进行了压缩。在实际无人车的行驶情况中，由于当 $\alpha, \omega$ 相同时，可以先关注 $r$ 比较小的点，这些点表示同一方向上近处的物体，这对无人驾驶时的路径规划是十分重要的，因而这样的信息损失是完全可以接受的。

对于深度图像上的物体分割，本文采用广度优先搜索（Breadth First Search, BFS）的思想，对深度图像中每个像素值不为0的点，搜索其邻域内的点，若其像素值（即实际的距离）与当前像素值之差不超过某个阈值，则认为其为同一类物体。算法流程如Algorithm 3所示。

---

#### Algorithm 3 Range Image Labelling

---

```

1: LabelRangeImage(R)
2: Label  $\leftarrow 1, L \leftarrow zeors(R_{rows} \times R_{cols})$ 
3: for  $r = 1 : R_{rows}$  do
4:   for  $c = 1 : R_{cols}$  do
5:     if  $L(r, c) == 0$  then
6:       LabelComponentBFS( $r, c, \text{Label}$ );
7:       Label = Label + 1;
8:     LabelComponentBFS( $r, c, \text{Label}$ )
9:     queue.push( $\{r, c\}$ )
10:   while queue is not empty do
11:      $\{r, c\} = \text{queue.top}();$ 
12:      $L(r, c) = \text{Label}$ 
13:     for  $\{r_n, c_n\} \in \text{Neighbourhood}\{r, c\}$  do
14:       if  $abs(R(r_n, c_n) - R(r, c)) < Thres$  then
15:         queue.push( $\{r_n, c_n\}$ )
          queue.pop()

```

---

在实际的实现中， $Thres$ 取值为0.5，分割的结果如图5-8所示。本文将每个不同聚类的物体标记为不同的颜色，从图中可以看出，这种方法很好的将空间位置不同的物体分割了出来。

在深度图像上将物体分割后，一个重要的步骤就是根据图像中物体的像素坐标以及深度值反推出物体在世界坐标系中的坐标。假设该物体所有像素的平均坐标为 $p(\bar{u}, \bar{v})$ ，其平均深度通过求所有像素的平均值而得，设为 $\bar{r}$ 。则可以根据三维点云投影深度图的算法倒推得其三维坐标：

$$\left\{ \begin{array}{l} \alpha = (\bar{u} - w/2) * \alpha_{percol} \\ \omega = -(\bar{v} - h/2) * \omega_{perrow} \\ x = \bar{r} \cos(\alpha) \cos(\omega) \\ y = -\bar{r} \cos(\alpha) \sin(\omega) \\ z = \bar{r} \sin(\omega) \end{array} \right. \quad (5-5)$$



图 5-8 基于深度图像的聚类分割结果

由此，便实现了基于点云投影到深度图像上的实时物体分割与定位。

表 5-1 人为测量，ICP以及Kabsch计算的相机与激光雷达标定结果

	Tape measurement	ICP	Kabsch
X(m)	-0.22 to -0.24	-0.222	-0.222
Y(m)	0.25 to 0.29	0.292	0.287
Z(m)	0.60 to 0.62	0.633	0.631
Roll(degree)	unmeasurable	5.16	5.15
Pitch(degree)	unmeasurable	2.29	2.29
Yaw(degree)	unmeasurable	-3.35	-3.36
RMSE(m)	-	0.0144	0.0130

## 5.4 激光雷达与相机标定结果的验证

为了检验激光雷达与相机的标定结果，本文较为粗略的人工卷尺测量了相机与激光雷达的坐标系之间的平移量，而由于旋转量较难获得，因此在这里并没有进行测量。标定结果与测量值的比较如表5-1所示。本文使用了ICP与Kabsch两种方法来根据三维点间的匹配求两坐标系之间的变换，并且对两种算法的结果计算了均方根误差（RMSE）。从表中可以看出，Kabsch算法得到的均方根误差相比ICP算法更小。而两种算法算出来的平移量都比较接近人工测量的结果，验证了利用本文提到的标定方法进行标定，具有较为可靠的标定效果。

## 5.5 多感知融合算法在KITTI数据集上的验证

为了验证本文提出的激光雷达与视觉融合的算法的有效性，本文在KITTI<sup>[35]</sup>数据集上进行了验证，如图5-9(a), 5-9(b)所示。图5-9(a)验证了融合算法的三维包围盒以及三维位置的准确性。而图5-9(b)则显示了融合算法帮助去除误检测的作用。红框为YOLO检测为汽车的结果，然而根据上文提到的去除误检测策略将其判定为误检测，从而验证了上文提到了多感知融合算法具有良好的去除误检测的效果。

## 5.6 本章小结

本章主要对本文第三章提出的三维感知机构的点云融合配准的策略进行了Gazebo仿真验证，同时拓展了融合后的点云的应用，将多帧融合后的点云应用于三维稠密地图的构建以及投影到深度图像中进行分割。在Gazebo仿真实验中可以发现第三章提出的融合策略有效地将点云进行了融合配准，融合后的点云能够较为清晰地反应仿真环境中障碍物的位置与形状；而将三维感知机构的信息与无人车里程计的信息结合后，又能够将多帧融合后的点云用于大规模三维建图之中，准确地反映出对环境的感知；最后，利用多帧融合后的点云投影到深度图像中，并对深度图像进行聚类分割，能够实现无人车环境下的障碍物分割与定位，对无人车环境下的环境感知与自主导航具有较大的意义。

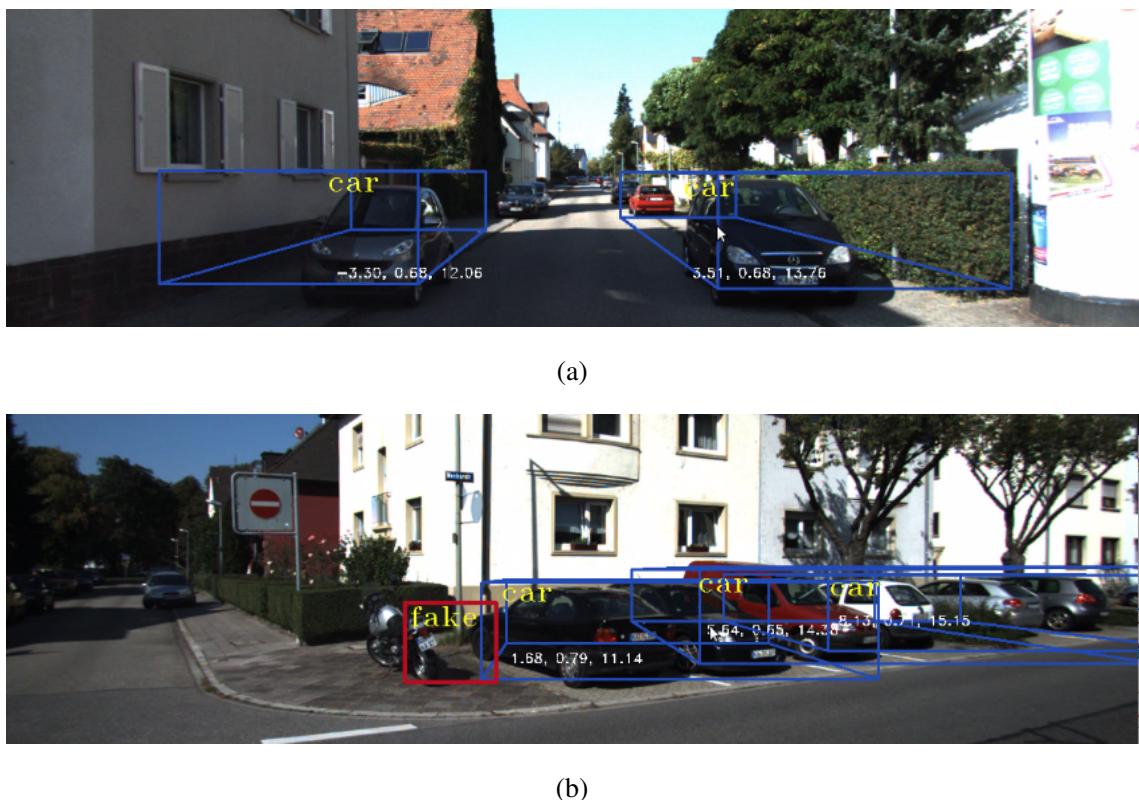


图 5-9 KITTI上的融合算法验证(a)三维包围盒的验证;(b)去除YOLO误检测的验证

## 第6章 全文总结与后续工作展望

### 6.1 全文总结

在无人车自动驾驶领域，如何利用好激光雷达来进行环境感知与障碍物检测一直是一个热门的话题。低线数激光雷达在环境感知时返回的信息较少，很难直接根据点云信息对物体进行识别，而高线数激光雷达又造价高昂，极大限制了无人驾驶的成本。因此，本文立足于解决低线数激光雷达由于线数较少而较难根据点云信息检测障碍物的问题，设计了一种新颖的三维感知机构，能够通过给低线数激光雷达提供俯仰角上的往复运动来增加激光雷达竖直方向上的分辨率；同时，本文亦融合了相机与激光雷达点云的数据信息，利用图像对障碍物进行了分类与分割，结合点云完成了对障碍物的定位。本文完成的主要研究工作有：

(1) 提出一种新颖的三维感知机构，该机构通过无刷电机驱动曲柄摇杆机构来给底座上的激光雷达提供俯仰角上的往复运动，并通过融合多帧激光雷达的点云来增加低线数激光雷达在竖直方向上的点云的分辨率。在融合的过程中，本文通过帧内多次线性插值来矫正因激光雷达往复运动而造成的运动畸变，使得融合后的点云能够更加真实地反应环境的情况。

(2) 融合了相机图像与激光雷达点云的信息，对三维障碍物进行了分割、分类以及定位。首先利用目标检测网络YOLO来对图像进行推断，对障碍物进行了分割与分类；随后，根据投影到相机上的点云信息，计算出了待检测目标的三维位置信息；最后，利用点云的三维位置信息来解决YOLO在低置信度阈值下的误检测问题，提高了YOLO在低执行度阈值下的目标检测的查准率。

综上，本文面向无人车领域，提出了一种新颖的利用低线数激光雷达多帧融合来进行环境感知的方式。通过融合多帧点云，本文提出的机构能够提供类似于高线数激光雷达的丰富的点云信息，同时利用该点云信息，结合相机图像信息进行了三维障碍物的识别与定位。

## 6.2 后续工作展望

有关无人车自动驾驶环境感知的研究近几年发展迅速，在本文研究工作的基础上，仍有以下方向值得进一步研究：

1. 在本文提出的多帧融合三维感知机构中，如果障碍物运动速度较快，则由于多帧融合，在融合后的点云中会有由于障碍物运动而产生的拖影。在三维建图时，如何消除运动物体产生的点云来建立只有静态物体的地图是一个值得深入研究的问题。一个可行的办法是引入八叉树地图（OctoMap），对于传感器观测到的障碍物，归入八叉树地图，利用贝叶斯滤波来滤除动态物体；随后再用八叉树地图来恢复出没有动态障碍物的点云。
2. 在相机与传感器的融合中，目前使用的YOLO卷积网络的输入为图像，其为 $m \times n \times 3$ 的有序矩阵信息。而通过将融合后的点云投影，则能够为图像的像素提供第四维信息，即深度信息，如果将四维像素信息输入卷积网络，则有希望大大提高YOLO目标检测的查准率与查全率。而且由于本文所得的激光雷达进行了竖直方向的分辨率的提升，使得投影到图像上的点云密度能够较好的和像素相匹配。

总而言之，本文提出的三维感知机构实现了低线数激光雷达点云在竖直方向上的稠密化，其对于三维稠密地图的建立以及三维目标检测的任务都有较为重要的意义，基于此平台可以未来可以开展许多有关三维环境感知感知、三维建图等方面的工作。

## 参考文献

- [1] F. Li. Imagenet Large Scale Visual Recognition Challenge (ILSVRC)[EB/OL]. <http://www.image-net.org/challenges/LSVRC/>
- [2] R. Girshick, J. Donahue, T. Darrell, et al. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1):142–158
- [3] K. He, G. Gkioxari, P. Dollár, et al. Mask r-cnn[C]. 2017, 2961–2969
- [4] M. Teichmann, M. Weber, M. Zoellner, et al. Multinet: Real-time joint semantic reasoning for autonomous driving[C]. 2018, 1013–1020
- [5] S. Liu, L. Qi, H. Qin, et al. Path aggregation network for instance segmentation[C]. 2018, 8759–8768
- [6] R. Garg, V. K. B.G., G. Carneiro, et al. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue[C]. Springer International Publishing, Cham, 2016, 740–756
- [7] B. Douillard, J. Underwood, V. Vlaskine, et al. A pipeline for the segmentation and classification of 3D point clouds[C]. 2014, 585–600
- [8] D. Zermas, I. Izzat, N. Papanikolopoulos. Fast segmentation of 3D point clouds: A paradigm on LiDAR data for autonomous vehicle applications[C]. 2017
- [9] J. Behley, V. Steinhage, A. B. Cremers. Laser-based segment classification using a mixture of bag-of-words[C]. 2013, 4195–4200
- [10] D. Korchev, S. Cheng, Y. Owechko, et al. On real-time lidar data segmentation and classification[C]. 2013, 1
- [11] Y. Zhou, O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]. 2018, 4490–4499
- [12] Y. Yan, Y. Mao, B. Li. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10):3337
- [13] S. Futatsumori, N. Miyazaki. Concept of Helicopter All-Around Obstacle Detection Using Millimeter-Wave Radar Systems Experiments with a beam-switching radar system and a multicopter[C]. 2018, 510–511

## 参考文献

---

- [14] N. Long, K. Wang, R. Cheng, et al. Unifying obstacle detection, recognition, and fusion based on millimeter wave radar and RGB-depth sensors for the visually impaired[J]. Review of Scientific Instruments, 2019, 90(4):044102
- [15] R. Ishikawa, T. Oishi, K. Ikeuchi. LiDAR and Camera Calibration Using Motions Estimated by Sensor Fusion Odometry[C]. 2018, 7342–7349
- [16] F. Zhang, D. Clarke, A. Knoll. Vehicle detection based on lidar and camera fusion[C]. 2014, 1620–1625
- [17] K. Banerjee, D. Notz, J. Windelen, et al. Online Camera LiDAR Fusion and Object Detection on Hybrid Data for Autonomous Driving[C]. 2018, 1632–1638
- [18] D. Xu, D. Anguelov, A. Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation[C]. 2018, 244–253
- [19] L. Caltagirone, M. Bellone, L. Svensson, et al. LIDAR–camera fusion for road detection using fully convolutional neural networks[J]. Robotics and Autonomous Systems, 2019, 111:125–131
- [20] J. Zhang, S. Singh. LOAM: Lidar Odometry and Mapping in Real-time.[J]. 2014, 2:9
- [21] 王良才等. 机械设计基础[M]. 北京: 北京大学出版社, 2007, 72–73
- [22] S. Hong, H. Ko, J. Kim. VICP: Velocity updating iterative closest point algorithm[C]. 2010, 1893–1898
- [23] A. Dhall, K. Chelani, V. Radhakrishnan, et al. LiDAR-camera calibration using 3D-3D point correspondences[J]. arXiv preprint arXiv:1705.09785, 2017
- [24] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, et al. Automatic generation and detection of highly reliable fiducial markers under occlusion[J]. Pattern Recognition, 2014, 47(6):2280–2292
- [25] V. Lepetit, F. Moreno-Noguer, P. Fua. EPnP: An Accurate O(n) Solution to the PnP Problem[J]. International Journal of Computer Vision, 2008, 81(2):155
- [26] J. Bacik. aruco\_mapping[EB/OL]. [http://wiki.ros.org/aruco\\_mapping](http://wiki.ros.org/aruco_mapping), Dec 16, 2016
- [27] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces[J]. International Journal of Computer Vision, 1994, 13(2):119–152
- [28] W. Kabsch. Kabsch\_algorithm[EB/OL]. [https://en.wikipedia.org/wiki/Kabsch\\_algorithm](https://en.wikipedia.org/wiki/Kabsch_algorithm)
- [29] O. Sorkine. Least-squares rigid motion using svd[J]. Technical notes, 2009, 120(3):52
- [30] J. Redmon, S. Divvala, R. Girshick, et al. You only look once: Unified, real-time object detection[C]. 2016, 779–788

- [31] J. MacQueen, et al. k-means clustering[EB/OL]. [https://en.wikipedia.org/wiki/K-means\\_clustering](https://en.wikipedia.org/wiki/K-means_clustering)
- [32] T. Foote. tf: The transform library[C]. 2013, 1–6
- [33] Z. Wang, W. Zhan, M. Tomizuka. Fusing Bird’s Eye View LIDAR Point Cloud and Front View Camera Image for 3D Object Detection[C]. 2018, 1–6
- [34] I. Bogoslavskyi, C. Stachniss. Efficient online segmentation for sparse 3d laser scans[J]. PFG—Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 2017, 85(1):41–52
- [35] A. Geiger, P. Lenz, C. Stiller, et al. Vision meets Robotics: The KITTI Dataset[J]. International Journal of Robotics Research (IJRR), 2013
- [36] X. Han, J. Lu, Y. Tai, et al. A real-time LIDAR and vision based pedestrian detection system for unmanned ground vehicles[C]. 2015, 635–639

## 致 谢

我的大学本科的学习生活就要结束了，而研究生生涯即将开始。值此辞旧迎新之际，有很多想说的话。

最感谢的集体是电子科技大学，为我在本科的四年里的学习生活提供了无尽的支持，无论是物质还是精神层面上。学校总是给予学生最大的资源与帮助，每次遇到困难时，学校永远是最坚强的后盾。

感谢的人有许多。首先，最感谢的是我即将到来的研究生生涯的导师，陈浩耀副教授。老师在为人与学术上都是一面旗帜，指引着学生的前进方向。能够到一个新的城市，去接触新的环境，去做自己想做的研究，都离不开老师物质与精神上的支持。十分有幸，能够成为老师的研究生。

其次，我要感谢实验室的师兄与师姐。感谢李四林与苏鹏鹏两位师兄，毕业设计的机械结构他们提供了极大的帮助；感谢所有师兄与师姐，对我的问题耐心解答，帮助我更快的融入了NRSL的大家庭。

我也要感谢我的女朋友。感谢她在我迷茫困顿的时候给我的精神上的慰藉。

最后，希望我能记住母校电子科大的校训，“求实求真，大气大为”，不负母校的期望。

## You Only Look Once: Unified, Real-Time Object Detection

Joseph Redmon\*, Santosh Divvala\*†, Ross Girshick¶, Ali Farhadi\*†

University of Washington\*, Allen Institute for AI†, Facebook AI Research¶

<http://pjreddie.com/yolo/>

### Abstract

We present YOLO, a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

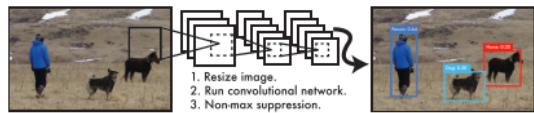
Our unified architecture is extremely fast. Our base YOLO model processes images in real-time at 45 frames per second. A smaller version of the network, Fast YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on background. Finally, YOLO learns very general representations of objects. It outperforms other detection methods, including DPM and R-CNN, when generalizing from natural images to other domains like artwork.

### 1. Introduction

Humans glance at an image and instantly know what objects are in the image, where they are, and how they interact. The human visual system is fast and accurate, allowing us to perform complex tasks like driving with little conscious thought. Fast, accurate algorithms for object detection would allow computers to drive cars without specialized sensors, enable assistive devices to convey real-time scene information to human users, and unlock the potential for general purpose, responsive robotic systems.

Current detection systems repurpose classifiers to perform detection. To detect an object, these systems take a classifier for that object and evaluate it at various locations and scales in a test image. Systems like deformable parts models (DPM) use a sliding window approach where the classifier is run at evenly spaced locations over the entire image [10].

More recent approaches like R-CNN use region proposal



**Figure 1: The YOLO Detection System.** Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to  $448 \times 448$ , (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model’s confidence.

methods to first generate potential bounding boxes in an image and then run a classifier on these proposed boxes. After classification, post-processing is used to refine the bounding boxes, eliminate duplicate detections, and rescore the boxes based on other objects in the scene [13]. These complex pipelines are slow and hard to optimize because each individual component must be trained separately.

We reframe object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. Using our system, you only look once (YOLO) at an image to predict what objects are present and where they are.

YOLO is refreshingly simple: see Figure 1. A single convolutional network simultaneously predicts multiple bounding boxes and class probabilities for those boxes. YOLO trains on full images and directly optimizes detection performance. This unified model has several benefits over traditional methods of object detection.

First, YOLO is extremely fast. Since we frame detection as a regression problem we don’t need a complex pipeline. We simply run our neural network on a new image at test time to predict detections. Our base network runs at 45 frames per second with no batch processing on a Titan X GPU and a fast version runs at more than 150 fps. This means we can process streaming video in real-time with less than 25 milliseconds of latency. Furthermore, YOLO achieves more than twice the mean average precision of other real-time systems. For a demo of our system running in real-time on a webcam please see our project webpage: <http://pjreddie.com/yolo/>.

Second, YOLO reasons globally about the image when

图 -1 yolo原文

making predictions. Unlike sliding window and region proposal-based techniques, YOLO sees the entire image during training and test time so it implicitly encodes contextual information about classes as well as their appearance. Fast R-CNN, a top detection method [14], mistakes background patches in an image for objects because it can't see the larger context. YOLO makes less than half the number of background errors compared to Fast R-CNN.

Third, YOLO learns generalizable representations of objects. When trained on natural images and tested on artwork, YOLO outperforms top detection methods like DPM and R-CNN by a wide margin. Since YOLO is highly generalizable it is less likely to break down when applied to new domains or unexpected inputs.

YOLO still lags behind state-of-the-art detection systems in accuracy. While it can quickly identify objects in images it struggles to precisely localize some objects, especially small ones. We examine these tradeoffs further in our experiments.

All of our training and testing code is open source. A variety of pretrained models are also available to download.

## 2. Unified Detection

We unify the separate components of object detection into a single neural network. Our network uses features from the entire image to predict each bounding box. It also predicts all bounding boxes across all classes for an image simultaneously. This means our network reasons globally about the full image and all the objects in the image. The YOLO design enables end-to-end training and real-time speeds while maintaining high average precision.

Our system divides the input image into an  $S \times S$  grid. If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object.

Each grid cell predicts  $B$  bounding boxes and confidence scores for those boxes. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts. Formally we define confidence as  $\text{Pr}(\text{Object}) * \text{IOU}_{\text{pred}}^{\text{truth}}$ . If no object exists in that cell, the confidence scores should be zero. Otherwise we want the confidence score to equal the intersection over union (IOU) between the predicted box and the ground truth.

Each bounding box consists of 5 predictions:  $x, y, w, h$ , and confidence. The  $(x, y)$  coordinates represent the center of the box relative to the bounds of the grid cell. The width and height are predicted relative to the whole image. Finally the confidence prediction represents the IOU between the predicted box and any ground truth box.

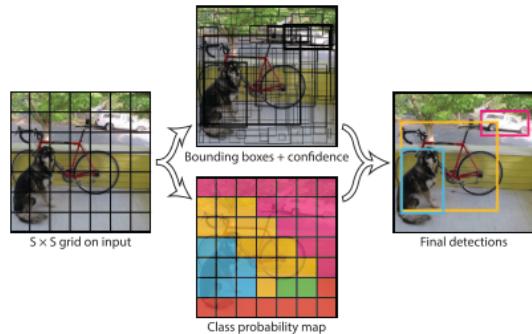
Each grid cell also predicts  $C$  conditional class probabilities,  $\text{Pr}(\text{Class}_i | \text{Object})$ . These probabilities are conditioned on the grid cell containing an object. We only predict

one set of class probabilities per grid cell, regardless of the number of boxes  $B$ .

At test time we multiply the conditional class probabilities and the individual box confidence predictions,

$$\text{Pr}(\text{Class}_i | \text{Object}) * \text{Pr}(\text{Object}) * \text{IOU}_{\text{pred}}^{\text{truth}} = \text{Pr}(\text{Class}_i) * \text{IOU}_{\text{pred}}^{\text{truth}} \quad (1)$$

which gives us class-specific confidence scores for each box. These scores encode both the probability of that class appearing in the box and how well the predicted box fits the object.



**Figure 2: The Model.** Our system models detection as a regression problem. It divides the image into an  $S \times S$  grid and for each grid cell predicts  $B$  bounding boxes, confidence for those boxes, and  $C$  class probabilities. These predictions are encoded as an  $S \times S \times (B * 5 + C)$  tensor.

For evaluating YOLO on PASCAL VOC, we use  $S = 7$ ,  $B = 2$ . PASCAL VOC has 20 labelled classes so  $C = 20$ . Our final prediction is a  $7 \times 7 \times 30$  tensor.

### 2.1. Network Design

We implement this model as a convolutional neural network and evaluate it on the PASCAL VOC detection dataset [9]. The initial convolutional layers of the network extract features from the image while the fully connected layers predict the output probabilities and coordinates.

Our network architecture is inspired by the GoogLeNet model for image classification [34]. Our network has 24 convolutional layers followed by 2 fully connected layers. Instead of the inception modules used by GoogLeNet, we simply use  $1 \times 1$  reduction layers followed by  $3 \times 3$  convolutional layers, similar to Lin et al [22]. The full network is shown in Figure 3.

We also train a fast version of YOLO designed to push the boundaries of fast object detection. Fast YOLO uses a neural network with fewer convolutional layers (9 instead of 24) and fewer filters in those layers. Other than the size of the network, all training and testing parameters are the same between YOLO and Fast YOLO.

图 -2 yolo原文

## You Only Look Once:统一的实时目标检测方法

### 1.1 介绍

人只需要瞄一眼图像就能立刻知道图像上的物体是什么，物体在哪儿。人类的视觉系统快速又精确，允许我们处理复杂的任务，例如几乎不需要太多思考地驾驶汽车。快速，准确的目标识别算法能够让电脑自动去驾驶汽车而不需要别的特定的传感器，给驾驶员提供实时的传感器信息，提供实现响应式的机器人系统的可能。

目前的检测系统是对分类器的改进。为了检测一个目标，这些系统训练出一个专门针对这个目标的分类器，并在图像中的不同位置、不同大小处去检验分类器的值。一些系统如deformable parts models（DPM）使用滑动窗口的方法来检测障碍物，其中的分类器在图像的每一个相等大小的子图像块上做分类任务。

更多最近的算法如R-CNN使用区域候选（region proposal）方法来在图像中快速产生可能的边界框，然后在这些候选的框内使用分类方法。在分类后，对这些边界框进行处理以使得边界框更准确，并去除面积重叠的边界框，对框中的物体重新评估其准确率与置信度。这些复杂的方法运算速度较慢，并且很难优化，因为每个分类器都需要分开来单独训练。

我们重新构架了目标检测任务，认为它只是一个单纯的回归问题，从像素中能够直接提取边界框的四个顶点以及分类的概率值。使用我们的系统，你只需要看一次（you only look once）图片就能够预测物体是什么，它们在图像的什么位置。

YOLO不可思议的简单：如图1所示。只用一个卷积网络来同时预测多个边界框以及分类的置信度。YOLO使用整个图像来进行训练，并且直接以训练的结果作为优化。这个统一的模型相对于传统的目标检测算法有许多优势。

首先，YOLO非常快。因为我们将目标检测任务构建为一个回归问题，我们不需要一个复杂的管道。我们只需要在测试时将我们的神经网络跑在一张新的图片上来预测分类结果。我们的网络，在Titan X GPU环境下，基础版本的YOLO可以达到45帧每秒，而快速版本的则可以达到超过150帧每秒。这就意味着我们可以

实时处理视频流，并且延迟低于25毫秒。另外，YOLO的平均测量精度是别的系统的两倍。如果想看我们的系统的实时演示的demo，请点击这个网站：

[https://pjreddie.com/darknet/yolo/。](https://pjreddie.com/darknet/yolo/)

其次，YOLO在做预测时，考虑到了整个图像。不像滑动窗口和候选区域的方法，YOLO在训练和测试时考虑到了整个图像，所以（训练出的网络）间接地包含了物体的上下文信息。Fast R-CNN，一个顶尖的目标检测方法，经常将背景误认为是目标，因为它不能看到更多的上下文信息。YOLO则不然。将背景误识别的概率，YOLO是Fast R-CNN的一半。

最后，YOLO学习物体的表示与生成方法。当在自然图像上训练，并且在人工的艺术品上检测时，YOLO的检测效果远比顶尖的目标检测算法，例如DPM和R-CNN要好。因为YOLO是高度生成的，所以当存在一个新领域的未知的输入时，YOLO崩溃的可能性更小。

尽管有这么多优点，YOLO仍然比目前精度最高的目标检测算法精度低。虽然它能够迅速地识别图像中的物体，但是它在精确定位上有一些困难，尤其是一些较小的物体。我们会在之后的实验中检验速度与精度的关系。

我们的所有训练与测试的代码都是开源的。一些预训练模型也提供下载。

## 1.2 统一检测

我们将目标检测的多个任务统一进了一个神经网络。我们的网络使用整个图像的特征来预测每个边界框。该网络还同时对每个边界框里的物体进行分类。这就意味着我们的网络感知整个图像以及图像中的物体。YOLO的设计实现了端到端的训练和实时的检测速度，同时还保有较高的准确率。

我们的系统将输入的图像分割为 $S \times X$ 个栅格。如果物体的中心落入了栅格中，那么这个栅格就负责检测这个物体。

每个栅格都预测B个边界框以及这些边界框的置信度。这些置信度分数反映了模型有多确信这些边界框内包含物体，以及这些边界框有多准确。我们定义置信度为 $Pr(\text{Object}) * IOU_{pred}^{truth}$ 。如果在栅格中没有物体存在，则该置信度应该为0。否则其置信度应该等于预测边界框与真值的边界框的重叠面积占比（intersection over union）。

每个边界框需要预测的参数有五个： $x, y, w, h$ 和置信度。 $(x, y)$ 坐标表示边界框的中心相对于栅格的边界的位置。宽度和高度则是相对于整个鱼香而言。最后，预测的置信度表示预测的边界框与真实的边界框之间的IOU得分。

每个栅格同时预测C个可能类别， $Pr(Class_i|Object)$ 。我们只计算那些有物体的栅格的概率，我们也对每个栅格预测一系列类别，而不管边界框B的数目。