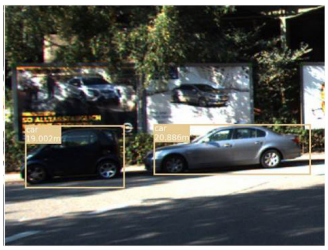
Opto-Electronic Engineering
光 电 工 程

Article

2021 年，第 48 卷，第 5 期

DOI: 10.12086/oee.2021.200418

# 基于 CNN 的点云图像融合目标检测



张介嵩，黄影平*，张　瑞

上海理工大学光电信息与计算机工程学院，上海 200093

**摘要：**针对自动驾驶场景中目标检测存在尺度变化、光照变化和缺少距离信息等问题，提出一种极具鲁棒性的多模态数据融合目标检测方法，其主要思想是利用激光雷达提供的深度信息作为附加的特征来训练卷积神经网络(CNN)。首先利用滑动窗对输入数据进行切分匹配网络输入，然后采用两个 CNN 特征提取器提取 RGB 图像和点云深度图的特征，将其级联得到融合后的特征图，送入目标检测网络进行候选框的位置回归与分类，最后进行非极大值抑制(NMS)处理输出检测结果，包含目标的位置、类别、置信度和距离信息。在 KITTI 数据集上的实验结果表明，本文方法通过多模态数据的优势互补提高了在不同光照场景下的检测鲁棒性，附加滑动窗处理改善了小目标的检测效果。对比其他多种检测方法，本文方法具有检测精度与检测速度上的综合优势。

**关键词：**数据融合；目标检测；卷积神经网络；滑动窗

**中图分类号：**TP391.4　　**文献标志码：**A

张介嵩，黄影平，张瑞. 基于 CNN 的点云图像融合目标检测[J]. 光电工程，2021，**48**(5): 200418

Zhang J S, Huang Y P, Zhang R. Fusing point cloud with image for object detection using convolutional neural networks[J]. *Opto-Electron Eng*, 2021, **48**(5): 200418

# Fusing point cloud with image for object detection using convolutional neural networks

Zhang Jiesong, Huang Yingping*, Zhang Rui

School of Optical-Electronic and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

**Abstract:** Addressing on the issues like varying object scale, complicated illumination conditions, and lack of reliable distance information in driverless applications, this paper proposes a multi-modal fusion method for object detection by using convolutional neural networks. The depth map is generated by mapping LiDAR point cloud onto the image plane and taken as input data together with the RGB image. The input data is also processed by the sliding window to reduce information loss. Two feature extracting networks are used to extract features of the image and the depth map respectively. The generated feature maps are fused through a connection layer. The objects are detected by processing the fused feature map through position regression and object classification. Non-maximal suppression is used to optimize the detection results. The experimental results on the KITTI dataset show that the proposed method is robust in various illumination conditions and especially effective on detecting small objects. Compared with other methods, the proposed method exhibits integrated advantages in terms of detection accuracy and speed.

收稿日期：2020-11-10； 收到修改稿日期：2021-01-22

基金项目：上海市自然科学基金项目(20ZR1437900)；国家自然科学基金面上项目(61374197)

作者简介：张介嵩(1996-)，男，硕士研究生，主要从事计算机视觉、机器学习的研究。E-mail：abcdta586351@126.com。

通信作者：黄影平(1966-)，男，教授，主要从事汽车电子、计算机视觉的研究。E-mail：huangyingping@usst.edu.cn。

**Keywords:** data fusion; object detection; convolutional neural networks; sliding window

## 1 引 言

无人驾驶汽车主要采用相机、激光雷达实现车辆、行人、骑行者等目标的探测。这两类传感器的数据模态不同，有着各自的优势和缺陷：激光雷达不受季节、光照条件的影响，探测距离长并且能够提供准确的三维位置信息，但雷达的点云数据是稀疏的，难以获得细节丰富的场景信息。相机能够提供稠密的纹理与色彩信息，但是被动传感器的特性使其容易受环境光照变化的影响。单一传感器难以提供满意的解决方案，两种传感器的融合可以利用多源数据的互补优势，弥补各自的缺陷，降低环境光照条件的影响，提高目标探测的鲁棒性和准确性，而且可以提供准确的目标位置信息。近年来，以卷积神经网络(CNN)为代表的深度学习技术在基于图像的目标检测方面取得巨大成功，也为多模态数据融合提供了一个非常有效的工具。

本文采用卷积神经网络，研究融合点云与图像数据的实时交通场景目标检测方法，有如下贡献：1) 提出了基于 CNN 的特征级点云与图像融合框架，设计了基于融合特征的目标检测网络，提高了不同光照条件下检测的鲁棒性；2) 采用滑动窗处理控制网络输入，平衡检测与数据采集时间，有效提升小目标的检测精度；3) 实现了对目标的精确检测及获取目标深度信息的多任务网络；4) 使用 KITTI 数据集进行实验评估，与多种检测算法的对比分析表明，本文方法具有检测精度和检测速度的综合优势。

## 2 相关工作

基于 CNN 的目标检测方法可以分为基于图像、激光雷达以及数据融合的方法。

### 2.1 基于图像的目标检测

Girshick 等人[1]借鉴滑动窗的思想提出了基于区域建议的卷积神经网络(RCNN)，成功地将神经网络由图像分类迁移到了目标检测中，大大提高了检测精度。然而这种需要在特征图上选择性搜索数千个候选框的方式也使得其检测速度较慢。在此框架下，SPP-Net[2]、Fast R-CNN[3]和 Faster R-CNN[4]等通过使用效率更佳的特征提取网络，优化模型结构，改进后处理方法等方式，力图提高基于候选区域的目标检测速度。Redmon 等人通过将目标检测视为回归问题，提出了一种单次检测网络 YOLO[5]，直接从图像中预测目标框和类别概率，大大提高了检测速度。然而检测时固定划分网格方式，降低了对小目标、彼此靠近目标等情况下的检测精度。随着 CNN 网络的不断发展，特征提取的性能也越来越强大，随后的 YOLO9000[6]、SSD[7]和 YOLOv3[8]等方法通过使用更优的特征提取网络以及引入残差网络的思想不断提高检测精度。总之，基于候选区域和单次检测网络在检测精度和速度上各有优势，难点在于同时取得精度和速度的最优。

### 2.2 基于激光雷达的目标检测

由于相机对光线和阴影较为敏感，不能提供准确和足够的位置信息，往往会影响系统的可靠性。相比之下，激光雷达可以探测目标的距离和三维信息。因此将激光雷达和深度学习相互结合的方法也获得了很大的发展。Qi 等人将原始的激光雷达产生的点云信息直接作为输入，提出了端到端的点云处理网络 PointNet[9]。但是由于对点云特征全部最大池化为一个特征，因此忽略了局部特征的表达导致精度欠佳。随后，PointNet++[10]提出了集合抽象模块和特征传播模块，改善了对局部特征的获取能力。不同于直接对无序的数据做处理，Zhou 等人[11]将激光雷达点云转换为具备一定规则分布的体素(Voxel)，在点云上建立三维网格来处理 LiDAR 点云。然而，这需要大量的计算来进行后续处理，无法达到实时性的需求。为了提升对点云的处理速度，Complex-YOLO[12]，BirdNet[13]和 LMNet[14]等提出多视图的投影方法将三维激光雷达点云数据投影到一个或多个二维平面上，以此视为二维图像。从转换视图的角度与前视图相比，鸟瞰图(BEV)上的每个对象都有较低的遮挡率，因此被广泛采用。Li 等人提出的 VeloFCN[15]将点云数据投影到图像平面坐标系，利用完全卷积神经网络(FCN)从深度数据中检测车辆，成为当时最快的基于点云的检测方法，但缺乏足够的纹理和色彩信息，检测精度较差。

### 2.3 基于融合方法的目标检测

点云数据具有精确的几何信息，但是数据非常稀疏。图像作为高分辨率的数据，具有丰富的纹理特征，可以逐个区分物体。最近越来越多的研究工作利用深度学习将点云和图像进行融合，主要分为目标级融合和特征级融合。1) 目标级融合采用 2D 候选区域与点

云的检测方法相结合，比如 F-PointNet[16]提出了一种两阶段的三维物体检测框架，采用基于图像的 2D 检测方法提取候选区域并以 PointNet 处理点云。与之相似的，FPC-CNN[17]采用 PC-CNN[18]检测 2D 包围框，将点云数据投影到图像平面上，并对 2D 包围框中的点云投影数据进行后续处理。这种方法由于传感器安装高度和遮挡不同，基于图像的候选区域往往会由于遗漏导致检测精度降低。2) 特征级融合将 3D 点云数据提取的深度特征与相应图像区域相互结合。例如，MV3D[19]从 BEV 生成 3D 候选区域，将其投影到激光雷达前视图和 RGB 图像上来获取三个视图上的区域特征，以此将所有视图的特征融合。AVOD[20]将 3D 锚框分别投影到 BEV 和 RGB 图像上获得对应候选区域的特征图，将特征图融合后进行目标检测。

## 3 本文方法

本文方法由两部分组成，数据预处理和融合检测网络，方法整体框架如图 1 所示。

首先，将激光雷达点云投影至图像平面得到稀疏的深度图，然后通过深度补全得到密集的深度图，与相机得到的 RGB 图像共同作为网络输入。将 RGB 图像与密集深度图进行滑动窗处理，得到近似为方形的数据切片送入融合目标检测网络。融合检测网络以 EfficientNet[21]作为特征融合网络，先分别对图像和点云深度图进行特征提取，然后将两组特征图进行级联融合，网络检测部分对融合产生的特征图进行塔式多尺度[22]处理，构建残差网络对多尺度特征进行目标预测，并通过非极大值抑制(NMS)优化提炼，最后输出包含目标类别、位置、置信度和距离的检测结果。

### 3.1 融合检测网络

融合检测网络基于 EfficientNet-B2 构建，网络架构如图 2 所示。

EfficientNet[21]是由 Google 的研究人员于 2019 年

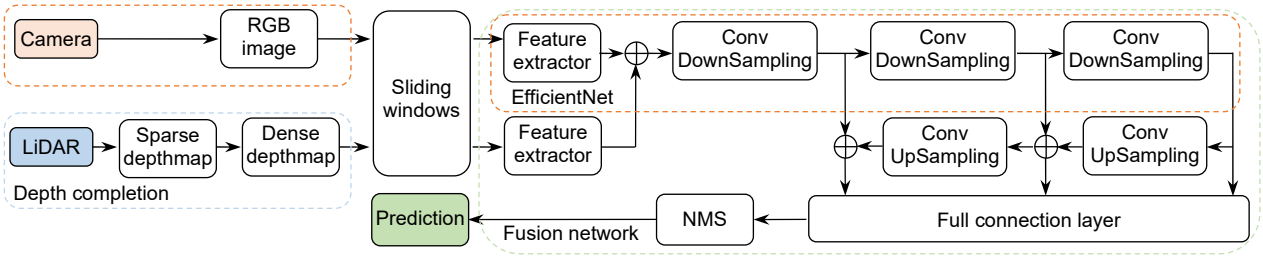



图 1 方法框图

Fig. 1 Framework of the proposed method

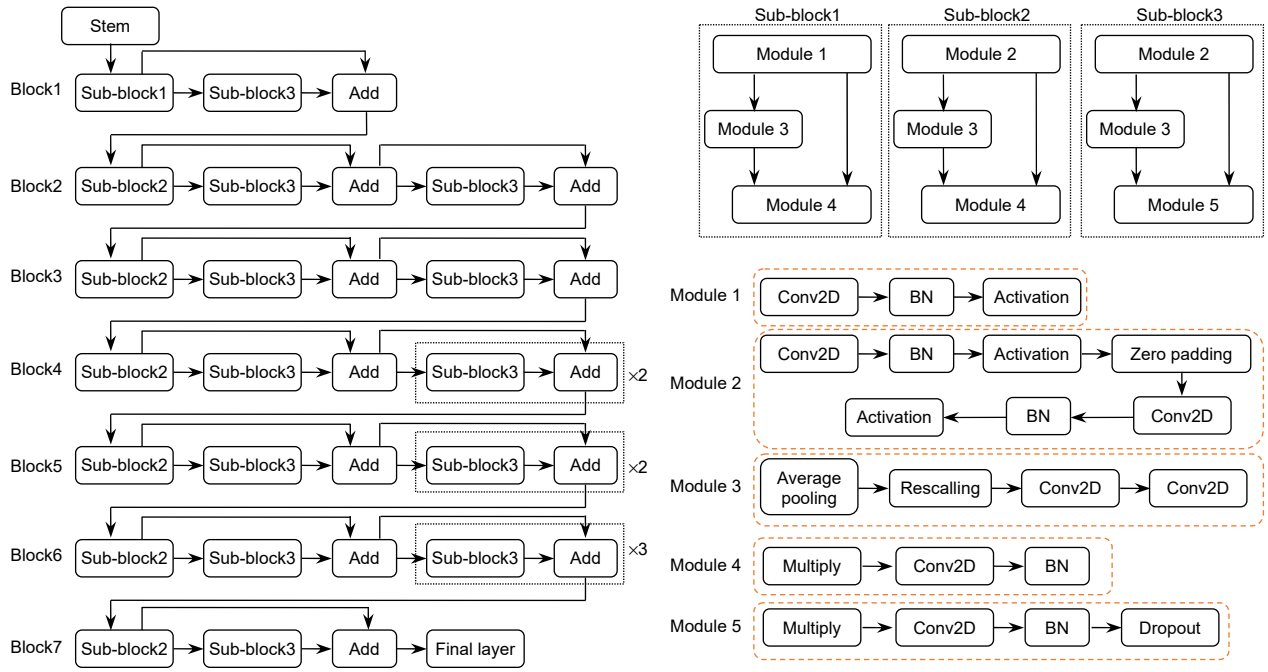



图 2 EfficientNet-B2 架构图

Fig. 2 EfficientNet-B2 architecture diagram

提出的一组深度、宽度和分辨率可变的复合卷积神经网络集合。通常卷积神经网络想提高检测精度，通过增加网络的深度(depth)、宽度(width)和输入分辨率(resolution)实现。但与此同时，网络占用的资源和计算量也会呈非线性的增长。Google 通过对深度、宽度和分辨率三个维度的研究，通过模块化的思想设计网络主体，选取合适的复合系数构建了 B0~B7 八种不同参数量的高效卷积神经网络，相比于 ResNet[23]、DenseNet[24]、Inception[25]、GPipe[26]等经典的主干网络，无论在分类精度和运算效率上都有显著提升。本文选取 EfficientNet-B2 作为基础网络，其具有 9.2 M 参数量，1.0 B 浮点计算操作，相比于 DarkNet-53，参数量减少 77.56%，浮点计算量减少 79.59%。其基本组成单元由 5 种 Module 构成，其中 Module1 和 Module2 实现基本的卷积操作和池化功能；Module3 用于跳跃连接不同的 Module；Module4 和 Module5 用于实现特征图的连接，与 Module3 共同构建残差网络。不同模块的组合构成 3 种不同的子模块，通过级联成为最终的网络主体。

融合网络利用 EfficientNet 的 Block1 和 Block2 作为特征提取器，分别对输入的 RGB 图像和密集深度图进行卷积和下采样，得到深度和尺度一致的特征图。通过图 3 所示的融合层进行特征图合并，使用 1×1 的卷积核，保持特征图尺度不变的前提下大幅增加特征的非线性特性，降低特征图维度实现跨通道信息交互，充分融合两种模态的数据特征。将融合后的特征图送入网络后续的 Block 中，对特征进行进一步提取和下采样，最终输出 13×13 的特征图。网络的参数设定如表 1 所示。

检测器部分采用特征金字塔网络(FPN)结构，在 32 倍降采样、16 倍降采样、8 倍降采样的三个特征图上进行多尺度目标预测，让网络同时学习浅层和深层特征，获得更好的表达效果，检测器结构如图 4 所示。

检测器通过神经网络的回归实现目标位置、类别和置信度的预测。因此，检测器的损失函数主要有三个部分构成：目标定位偏移量损失 $L_{\mathrm{loc}}(l,g)$，目标分

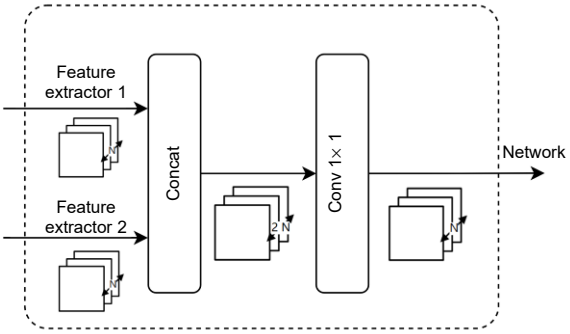



图 3 特征融合层

Fig. 3 Feature fusion layer

表 1 特征提取网络参数

Table 1 The parameters of the feature extraction network

| Stage | Kernel | Resolution | Channel |
|---|---|---|---|
| 1(Stem) | Conv,3×3 | 416×416->208×208 | 32 |
| 2(Block1) | Conv,3×3 | 208×208->208×208 | 16 |
| 3(Block2) | Conv,3×3 | 208×208->104×104 | 24 |
| 4(Block3) | Conv,5×5 | 104×104->52×52 | 48 |
| 5(Block4) | Conv,3×3 | 52×52->26×26 | 88 |
| 6(Block5) | Conv,5×5 | 26×26->26×26 | 120 |
| 7(Block6) | Conv,5×5 | 26×26->13×13 | 208 |
| 8(Block7) | Conv,3×3 | 13×13->13×13 | 352 |

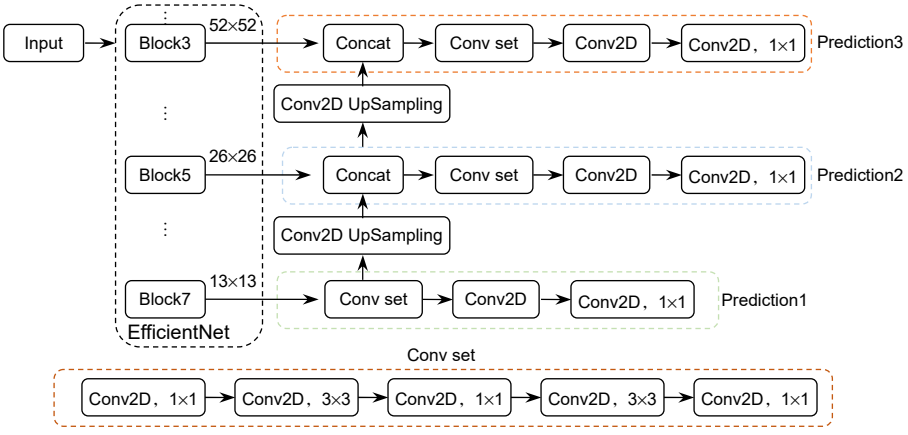



图 4 目标检测器结构

Fig. 4 Structure of the object detector

类损失 $L_{\rm cla}(O,C)$ 以及目标置信度损失 $L_{\rm conf}(o,c)$：

$$L(l,g,O,C,o,c)=\lambda_1 L_{\rm loc}(l,g)+\lambda_2 L_{\rm cla}(O,C)+\lambda_3 L_{\rm conf}(o,c)\ , \tag{1}$$

其中：定位偏移量损失采用平方差损失函数(MSE loss)，分类损失采用多分类交叉熵损失函数(cross entropy loss)；置信度损失采用二值交叉熵损失函数(binary cross entropy loss)；$l$ 为预测矩形框的坐标，$g$ 为真实值的坐标，$O$ 为预测框中是否存在目标的概率；$C\in\{0,1\}$，0 表示不存在，1 表示存在；$o_{ij}$ 为第 $i$ 个目标框中存在第 $j$ 类目标的概率，$c_{ij}\in\{0,1\}$，0 表示不存在，1 表示存在；$\lambda_1$，$\lambda_2$，$\lambda_3$ 为平衡系数。

## 3.2 数据预处理

### 3.2.1 点云深度图的产生与稠密化

坐标系标定是多传感器信息融合的首要条件，不同传感器有着不同的采集频率和独立的坐标系，必须统一数据采集频率进行时间配准才能把不同坐标系的数据转换到同一坐标系，实现数据的融合。激光雷达数据投影到像素坐标系的变换流程如图 5 所示。

根据常昕等人[27]的研究，由激光雷达坐标系到像素坐标系的转换关系为

$$Z_{\rm C}\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}\frac{1}{{\rm d}x}&0&u_0\\0&\frac{1}{{\rm d}x}&v_0\\0&0&1\end{bmatrix}\cdot\begin{bmatrix}f&0&0&0\\0&f&0&0\\0&0&1&0\end{bmatrix}\cdot\begin{bmatrix}\boldsymbol{R}_{\rm L\text{-}C}&\boldsymbol{T}_{\rm L\text{-}C}\\0^{\rm T}&1\end{bmatrix}\cdot\begin{bmatrix}X_{\rm L}\\Y_{\rm L}\\Z_{\rm L}\\1\end{bmatrix}\ , \tag{2}$$

其中：$(u, v)$为像素坐标，$(X_{\rm C}, Y_{\rm C}, Z_{\rm C})$为相机坐标，$(X_{\rm L}, Y_{\rm L}, Z_{\rm L})$为激光雷达坐标，$\boldsymbol{R}_{\rm L\text{-}C}$ 表示从激光雷达坐标系到相机坐标系的旋转矩阵，$\boldsymbol{T}_{\rm L\text{-}C}$ 表示从激光雷达坐标系到世界坐标系的三维平移向量，$u_0$，$v_0$ 是相机的内参，$f$ 为相机焦距。本文相机内参及投影矩阵由实验使用的 KITTI 数据集提供，根据式(2)可计算出激光点云在图像平面的坐标$(u, v)$，投影结果如图 6 所示。

激光雷达到图像的投影得到的深度图是稀疏的深度图，大量的空像素不足以描述物体的特征甚至干扰网络的计算，深度补全任务的目的是从稀疏的深度图生成密集的深度预测。

该问题可以被表述如下：

$$f(I,D_{\rm sparse})=D_{\rm dense}\ , \tag{3}$$

$$\min.\|\hat{f}(I,D_{\rm sparse})-f(I,D_{\rm sparse})\|_F^2=0\ , \tag{4}$$

其中：$I$ 为图像，$D_{\rm sparse}$ 为稀疏深度图，$D_{\rm dense}$ 为密集深度图。

相比于基于深度学习的方法，传统的图像处理算法在深度补全上具有更快的处理速度，同时不需要大量数据的训练也能保证较好的效果。

由于 KITTI 数据集中点云数据的深度范围在 0~80 m 之间，没有点云的像素区域深度值为零。若直接采用膨胀操作会导致大值覆盖小值，丢失目标边缘信息。因此在形态学处理之前将深度值反转：

$$D_{\rm inverted}=100.0-D_{\rm input}\ , \tag{5}$$

其中：$D_{\rm inverted}$ 为反转后的深度值，$D_{\rm input}$ 为输入的深度值。通过在有效值和空值之间建立了 20 m 的缓冲区，保证膨胀操作时更好地保留对象的边缘。本文所使用的深度补全算法流程图如图 7 所示。

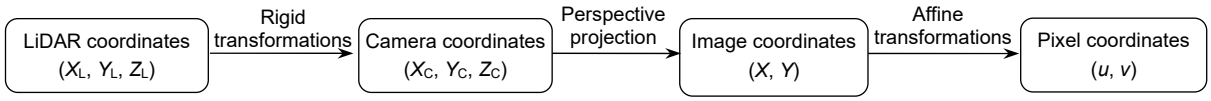



图 5 坐标系变换

Fig. 5 Transformation of coordinates

图 6 点云投影至图像平面

Fig. 6 Projection of LiDAR point cloud on the image plane

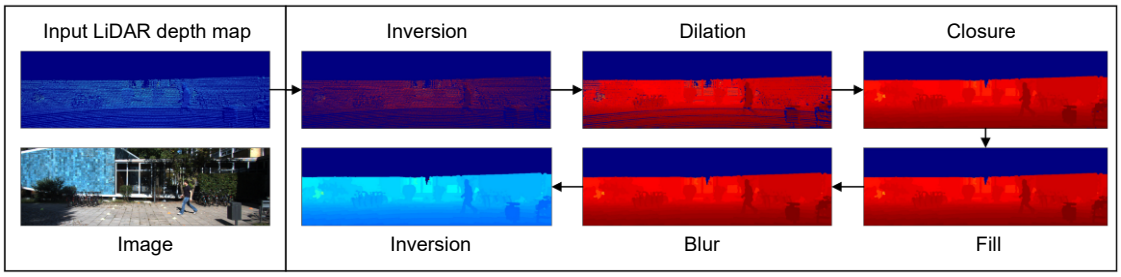



图 7 深度补全算法流程

Fig. 7 The formation of the dense depth map

### 3.2.2 滑动窗处理

在将数据输入网络之前，需要保证输入图像的分辨率与网络设定的参数一致。然而，目标检测不同于图像分类等其他神经网络，待检测对象的纹理、色彩、尺寸都是特征之一。因此不能通过拉伸原图进行分辨率的匹配。通常，网络在加载数据时都是以输入图像长边为标准进行缩放，对空余部分进行补零，如图 8(a)所示。

这种因长宽比例过大导致在加载数据时的信息丢失，在交通场景数据集中尤为明显。鉴于此，本文考虑使用长宽比例接近 1 的滑动窗口对原始图像进行扫描，采用保留重叠的方式对输入数据进行切分，将切分后的滑动窗口进行填充后再传入网络，同时把所有的滑动窗口的结果重新映射到原始图像对应的坐标，经过非极大值抑制(NMS)获得最终检测结果。

## 4 实验及结果分析

实验在 Intel Xeon(R) Silver 4110 CPU@ 2.10 GHz 处理器，32 G 内存，11 GB NVIDIA GeForce 1080Ti GPU，Ubuntu 20.04 操作系统的计算机上运行，融合检测网络基于 Pytorch 网络框架搭建。

实验数据来自 KITTI 的 Object Detection Evaluation 2012 数据集，包含训练数据和测试数据两部分。本文选取其中双目相机的彩色左视图，激光雷达点云，激光雷达与相机的标定数据和训练标签进行实验分析。训练集中包含 7481 张训练图像和 51865 个带有标签的目标，标签中目标被分为 9 类，包括一个 DonCare 类。本文将 Pedestrian 和 Person_sitting 归为一类 Pedestrian；将 Car、Truck 和 Van 归为一类 Car；Cyclist 独立作为一类，另外 Misc 和 Tram 因为数据太少被舍弃。

KITTI 图像序列包含三种场景：Easy、Moderate、Hard。Easy 为最小边框高度大于 40 pixels，无遮挡，截断不到 15%的目标障碍物；Moderate 为最小边框高度大于 25 pixels，部分遮挡，截断不到 30%的目标障碍物；Hard 为最小边框高度大于 25 pixels，多遮挡，截断不到 50%的目标障碍物。

实验结果采用 KITTI 的评价方法，如果目标的检测框与标签的边框的重叠度(IoU)达到 50%以上，则将该对象视为已正确检测到。选取准确率(Precission)、召回率(Recall)和平均准确度(Average precision，AP)作为性能评价指标。

准确率：

$$P=\frac{T_{\rm P}}{F_{\rm P}+T_{\rm P}}\ , \tag{6}$$

召回率:

$$R=\frac{T_{\rm P}}{F_{\rm N}+T_{\rm P}}\ , \tag{7}$$

平均准确度：

$$AP=\int_0^1 P(R){\rm d}R\ , \tag{8}$$


图 8 数据加载方式。(a) 缩放与填充；(b) 滑动窗处理

Fig. 8 Methods of data loading. (a) Resizing and padding; (b) Sliding windows

其中：$T_P$ 为真正例(true positive)，$F_P$ 为假正例(false positive)，$F_N$ 为假负例(false negative)，$P(R)$ 为不同召回率所对应的准确率。

实验主要分为三个部分：第一，通过对比使用滑动窗前后的实验结果，验证滑动窗对小目标检测效果的改善。第二，通过实验结果评估本文方法在复杂光照条件下对障碍物的检测能力，验证本文方法的鲁棒性。第三，通过实验结果比较本文方法和多种目标检测方法对交通场景中障碍物的检测效果，验证多模态数据融合方法对目标检测性能的提升。

### 4.1 输入数据的构建结果

网络输入数据为相机采集的 RGB 图像和激光雷达得到的 3D 点云，在数据预处理中，实现对激光雷达点云的图像平面投影、深度补全、滑动窗数据拆分工作。

利用 KITTI 数据集提供的点云和图像数据，通过式(2)将雷达点云投影至图像平面，得到稀疏的深度图，结果如图 9(a)所示。稀疏的深度图中存在大量的零值，以空洞的形式表现在图像中。在送入网络之前需要进行深度补全。深度补全使用基于 OpenCV 的形态学操作，实现膨胀、闭运算、空值填充、模糊处理，补全结果如图 9(b)所示，整个运算过程不依赖于神经网络和 RGB 数据的引导，在 CPU 上运算耗时 11 ms。

KITTI 数据集的图像分辨率为 1242×375，通过对 Car、Pedestrian、Cyclist 三种类别的标签统计，Car 的平均目标尺寸为 111×66，Pedestrian 的平均目标尺寸为 43×103，Cyclist 的平均目标尺寸为 55×76，本文方法单帧数据处理时间为 0.017 s，激光雷达采样间隔为 0.1 s。为匹配网络输入选择方形滑动窗，保证重叠区域大于最大平均目标宽度，总处理时间小于采样间隔。因此，实验采用 375×375 的滑动窗，滑动窗口次数为 4，步长为 217，相邻窗口保留 158 个像素宽度的重叠区域，减少因滑动窗的截断导致的漏检。

### 4.2 特征提取结果

采用 EfficientNet-B2 的 Block1 和 Block2 作为特征提取器对 RGB 图像和点云深度图提取特征，其结果与 DarkNet-53 的特征提取结果对比如图 10 所示。

其中图 10(a)、图 10(c)、图 10(e)分别为场景一、

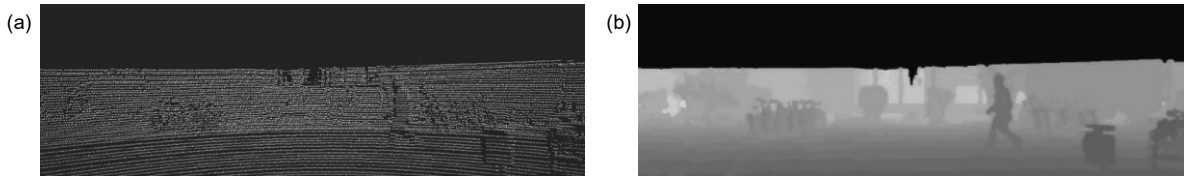



图 9 深度补全前后对比。(a) 稀疏的深度图；(b) 密集的深度图

Fig. 9 Comparison of depth maps. (a) Sparse depth map; (b) Dense depth map

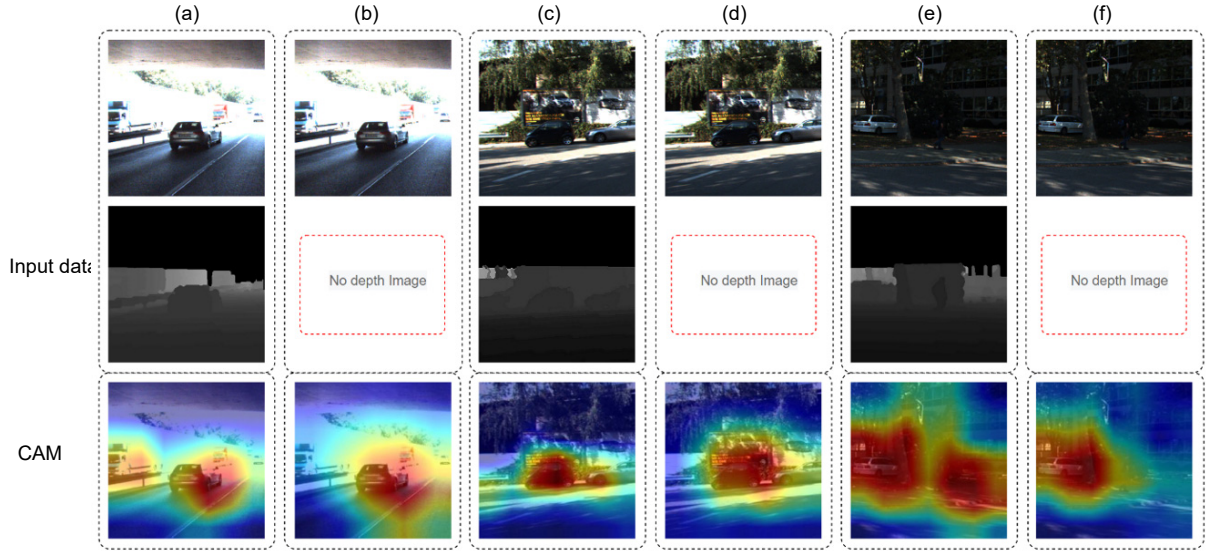



图 10 EfficientNet-B2 与 DarkNet-53 特征提取效果对比

Fig. 10 The feature extraction comparison of EfficientNet-B2 and DarkNet-53

二、三的 EfficientNet 融合网络特征提取结果；图 10(b)、图 10(d)、图 10(f)分别为场景一、二、三的 DarkNet 特征提取结果。采用 Grad-CAM++[28]对于网络特征提取的结果可视化，通过将特征提取网络的最后一个卷积层的特征图加权映射到原始图像平面，以热图的形式表征特征提取的效果。结果表明，相比于单模态特征提取方法，融合网络对复杂光线场景中目标所在区域有更准确的响应，如图 10(a)，10(e)所示；同时，引入深度信息的融合网络对平面广告牌上的假目标没有错误响应，如图 10(c)所示。

## 4.3 检测结果及分析

### 4.3.1 定性分析

检测结果如图 11 所示，其中第一行是网络输入的 RGB 图像，第二行是网络输入的密集深度图，第三行为进行对比的 YOLOv3 网络检测结果(输入仅为 RGB 图像)，第四行为未采用滑动窗的融合方法，第五行为采用滑动窗的融合方法，图 11(a)~图 11(d)为四个不同场景。

对比仅采用 RGB 图像数据的 YOLOv3 算法，本文方法采用多模态数据作为输入，对图像和点云数据进行特征级的融合，综合利用图像数据的密集纹理信息和点云数据的深度信息，有效降低了目标检测的误检率和漏检率，同时获取目标的距离信息；引入滑动窗口的处理方式，显著提升了小目标的检测效果。如图 11 所示，在明暗反差剧烈的场景(a)中，本文方法准确地识别出了远近的三辆汽车，以及汽车前阴影中的人，而在 YOLOv3 的检测结果中，仅仅检测到了一辆汽车；在隧道场景(b)中，包含了深度信息的本文方法准确检测出了阴影中的车辆和远处过曝的车辆，图像的方法仅检测到了纹理清晰的目标；在曝光不足的场景(c)中，阴影中的行人在图像中难以区分，在深度图中清晰可辨；在存在虚假目标的场景(d)中，图像的方法将广告牌中的车辆误认为目标车辆，在深度图中仅真实车辆与背景之间存在深度差异，广告牌为平面，本文方法没有发生误检。

### 4.3.2 定量评估

将本文方法与 Faster-RCNN、YOLOv3、VoxelNet、MV3D、F-PointNet 以及 AVOD 进行比较，这些方法分别对应的输入数据为 RGB 图像、雷达点云和融合数据。各种方法在 Easy、Moderate、Hard 三种场景中分别对 Car、Pedestrian 和 Cyclist 三类目标检测性能见表 2，表中的 mAP 是在 Easy、Moderate、Hard 三种场景中对所有目标统计的平均检测精度。

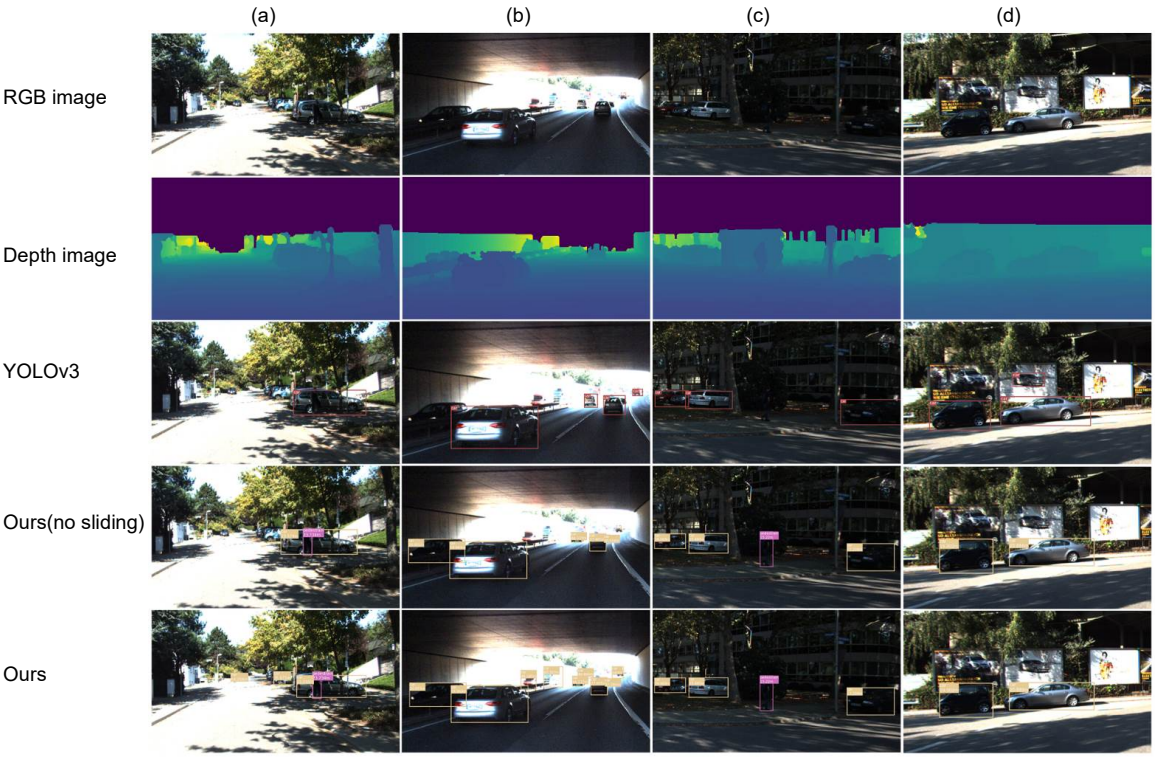



图 11 不同场景下的检测结果

Fig. 11 Detection results in different scenarios

表 2 与其他方法在 KITTI 数据集上的性能对比

Table 2 Performance comparison of different algorithms on the KITTI dataset

| Methods | Data | mAP/% | Car | | | Pedestrian | | | Cyclist | | | Times/(s/F) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Easy | Moderate | Hard | Easy | Moderate | Hard | Easy | Moderate | Hard | |
| Faster-RCNN | Image | 70.26 | 87.90 | 79.11 | 70.19 | 71.41 | 62.81 | 55.44 | 78.35 | 65.91 | 61.19 | 0.23 |
| YOLOv3 | Image | 56.75 | 88.71 | 74.40 | 65.58 | 67.23 | 49.47 | 44.99 | 50.88 | 36.89 | 32.64 | 0.027 |
| VoxelNet | LiDAR | 50.99 | 77.47 | 65.11 | 57.73 | 39.48 | 33.69 | 31.51 | 61.22 | 48.36 | 44.37 | 0.23 |
| MV3D | Image+LiDAR | 62.85 | 71.09 | 62.35 | 55.12 | - | - | - | - | - | - | 0.36 |
| F-PointNet | Image+LiDAR | 82.94 | 95.85 | 95.17 | 85.42 | 89.83 | 80.13 | 75.05 | 86.86 | 73.16 | 65.21 | 0.16 |
| AVOD | Image+LiDAR | 62.15 | 95.17 | 89.88 | 82.83 | 50.90 | 39.43 | 35.75 | 66.45 | 52.60 | 46.39 | 0.08 |
| Ours (No sliding) | Image+LiDAR | 60.29 | 90.01 | 82.40 | 73.58 | 69.23 | 48.45 | 44.95 | 55.18 | 41.59 | 37.32 | 0.017 |
| Ours | Image+LiDAR | 70.12 | 92.51 | 86.90 | 79.52 | 81.91 | 66.37 | 48.45 | 73.22 | 58.93 | 43.30 | 0.087 |

由表中数据可以看出，与 YOLOv3、VoxelNet、MV3D 以及 AVOD 相比较，融合深度图特征和滑动窗口处理的本文方法(最后一行)在精度上有全面提升。在取得与 Faster-RCNN 接近的检测精度的同时，检测速度大幅提升。与多模态目标级融合方法 F-PointNet 比较，检测精度上稍逊，但检测速度有较大提升。综上所述，本文方法取得了检测精度与检测速度的平衡。

通过计算可知，本文方法在 Easy、Moderate、Hard 场景中对 Car、Pedestrian 和 Cyclist 的平均检测精度分别是 82.55%、70.73%和 57.09%，单帧计算耗时 0.087 s，基本满足实时性要求。对比速度最快的单次目标检测方法 YOLOv3，在三种场景中对于 Car、Pedestrian 和 Cyclist 的平均检测精度分别提升 13.6%、17.15%和 9.37%；对比基于候选区域的 Faster-RCNN，检测精度分别提升 3.32%、1.46%和-5.17%；对比基于激光点云的目标检测方法 VoxelNet，检测精度分别提升 23.15%、21.68%和 12.56%；对比多模态数据融合的检测方法 AVOD，检测精度分别提升 11.70%、10.10%和 2.12%。

表中后两行分别为不使用滑动窗预处理的检测结果和使用滑动窗预处理的检测结果，相较于前者，附加滑动窗处理的方法对小目标的检测精度提升明显，但单帧计算时间有所增加。

通过对比本文方法与 YOLOv3 对汽车、骑行者、行人三类障碍物 P-R 曲线下的包围面积可见，本文方法对小目标检测效果提升显著，如图 12 所示。

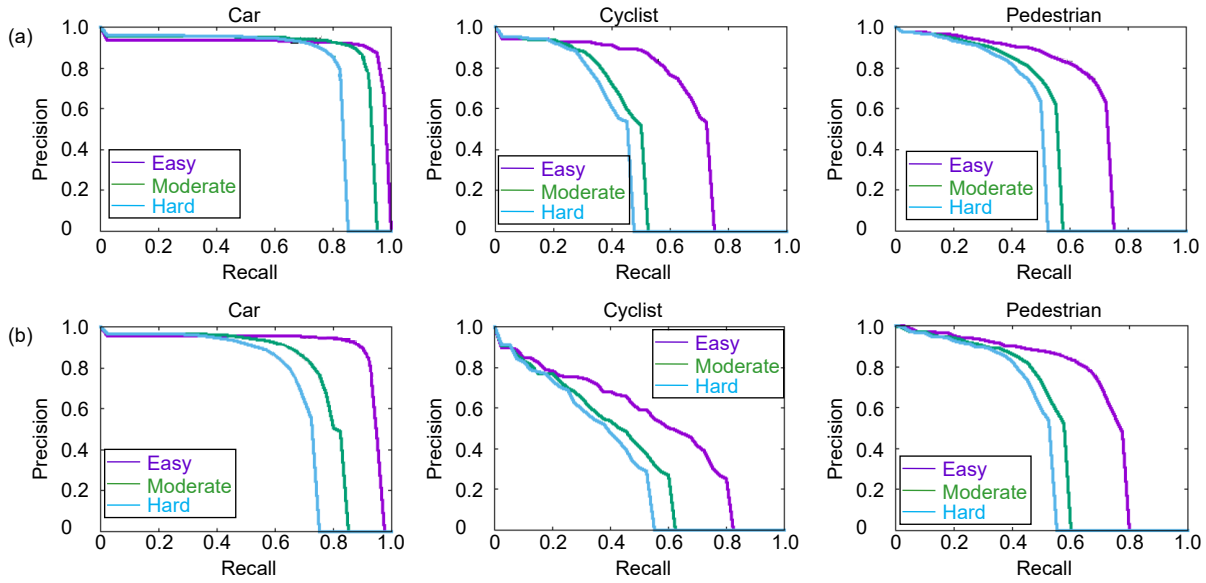



图 12 本文方法与 YOLOv3 在 KITTI 上的 P-R 曲线对比。(a) 本文方法的结果；(b) YOLOv3 的结果

Fig. 12 Comparison of the P-R curve between our method and YOLOv3. (a) Our method; (b) YOLOv3

## 5 结 论

本文提出了一种基于卷积神经网络的融合激光雷达点云与相机图像的目标检测方法，设计实现了一种点云与图像特征级融合的网络框架，并针对输入图像长宽比过大导致的信息损失提出了一种滑动扫描窗口的数据处理方法。采用 KIITI 数据集进行实验验证，对比其他多种检测方法，本文方法具有检测精度与检测速度上的综合优势，并能同时获取目标的距离信息。这些结果表明，本文方法借助多模态数据的优势互补提高了在不同光照场景的检测鲁棒性和准确性，附加滑动窗处理改善了对小目标的检测效果。

## 参考文献


[1] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014: 580–587.

[2] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *IEEE Trans Pattern Anal Mach Intell*, 2015, **37**(9): 1904–1916.

[3] Girshick R. Fast R-CNN[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015: 1440–1448.

[4] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2017, **39**(6): 1137–1149.

[5] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016: 779–788.

[6] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017: 7263–7271.

[7] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot MultiBox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016: 21–37.

[8] Redmon J, Farhadi A. YoLOv3: an incremental improvement[Z]. arXiv:1804.02767, 2018.

[9] Qi C R, Su H, Mo K C, *et al*. PointNet: deep learning on point sets for 3D classification and segmentation[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017: 652–660.

[10] Qi C R, Yi L, Su H, *et al*. PointNet++: deep hierarchical feature learning on point sets in a metric space[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017: 5099–5108.

[11] Zhou Y, Tuzel O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 4490–4499.

[12] Simon M, Milz S, Amende K, *et al*. Complex-YOLO: Real-time 3D object detection on point clouds[Z]. arXiv:1803.06199, 2018.

[13] Beltrán J, Guindel C, Moreno F M, *et al*. BirdNet: a 3D object detection framework from LiDAR information[C]//*Proceedings of 2018 21st International Conference on Intelligent Transportation Systems*, Maui, HI, USA, 2018: 3517–3523.

[14] Minemura K, Liau H, Monrroy A, *et al*. LMNet: real-time multiclass object detection on CPU using 3D LiDAR[C]//*Proceedings of 2018 3rd Asia-Pacific Conference on Intelligent Robot Systems*, Singapore, 2018: 28–34.

[15] Li B, Zhang T L, Xia T. Vehicle detection from 3D lidar using fully convolutional network[Z]. arXiv:1608.07916, 2016.

[16] Qi C R, Liu W, Wu C X, *et al*. Frustum PointNets for 3D object detection from RGB-D data[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 918–927.

[17] Du X X, Ang M H, Karaman S, *et al*. A general pipeline for 3D detection of vehicles[C]//*Proceedings of 2018 IEEE International Conference on Robotics and Automation*, Brisbane, QLD, Australia, 2018: 3194–3200.

[18] Du X X, Ang M H, Rus D. Car detection for autonomous vehicle: LIDAR and vision fusion approach through deep learning framework[C]//*Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vancouver, BC, Canada, 2017: 749–754.

[19] Chen X Z, Ma H M, Wan J, *et al*. Multi-view 3D object detection network for autonomous driving[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017: 1907–1915.

[20] Ku J, Mozifian M, Lee J, *et al*. Joint 3D proposal generation and object detection from view aggregation[C]//*Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Spain, 2018: 1–8.

[21] Tan M X, Le Q V. EfficientNet: rethinking model scaling for convolutional neural networks[Z]. arXiv:1905.11946, 2020.

[22] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017: 2117–2125.

[23] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016: 770–778.

[24] Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017: 4700–4708.

[25] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016: 2818–2826.

[26] Huang Y P, Cheng Y L, Bapna A, *et al*. GPipe: efficient training of giant neural networks using pipeline parallelism[C]//*Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019: 103–112.

[27] Chang X, Chen X D, Zhang J C, *et al*. An object detection and tracking algorithm based on LiDAR and camera information fusion[J]. *Opto-Electron Eng*, 2019, **46**(7): 180420.
常昕, 陈晓冬, 张佳琛, 等. 基于激光雷达和相机信息融合的目标检测及跟踪[J]. 光电工程, 2019, **46**(7): 180420.

[28] Chattopadhay A, Sarkar A, Howlader P, *et al*. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks[C]//*Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision*, Lake Tahoe, NV, USA, 2018: 839–847.

# Fusing point cloud with image for object detection using convolutional neural networks

Zhang Jiesong, Huang Yingping*, Zhang Rui

School of Optical-Electronic and Computer Engineering,
University of Shanghai for Science and Technology, Shanghai 200093, China

Result of multimodal data fusion detection method (Right side)

Overview: Autonomous vehicles are equipped with cameras and light detection and ranging (LiDAR) for obstacle detection. Cameras can provide dense texture and color information, but the nature of the passive sensor makes it vulnerable to variations in ambient lighting. LiDAR can get accurate spatial information and is not affected by seasons and lighting conditions, but the point cloud data is sparse and distributed non-uniformly. Single sensor cannot provide satisfactory solution for task of environment perception. Fusion of the two sensors can take advantage of the two modalities of data, improving the robustness and accuracy. In recent years, convolutional neural networks (CNNs) have achieved great success in vision-based object detection and also provide an efficient tool for multi-modal data fusion. This paper proposes a novel multi-modal fusion method for object detection by using CNNs. The depth information provided by LiDAR is used as additional features to train CNNs. Disordered and sparse point cloud is projected onto the image plane to generate the depth map which is processed by a depth completion algorithm. The dense depth map and the RGB image are taken as the input of the network. The input data is also processed by sliding the window to reduce information loss caused by resolution mismatch and inappropriate aspect ratio. We adopt EfficientNet-B2 as backbone network of feature extraction, data fusion, and detection. The network extracts respectively the features of the RGB image and the depth map and then fuses the feature maps together through a connection layer. Followed by $1\times1$ convolution operation, the detection network uses feature pyramid to generate three scales of feature maps and estimates objects through position regression and object classification. Non-maximum suppression is used to optimize the detection results for all sliding windows. The output of the network contains information about location, class, confidence and distance of the target. The experiments were conducted on the KITTI benchmark dataset by using a workstation equipped with 4-core processor and 11 GB NVIDIA 1080Ti GPU and Pytorch neural network framework. By quantitatively analyzing the single-frame inference time and average precision (mAP) of different data modality detection methods, the experimental results show that our method achieves a balance between detection accuracy and detection speed. By qualitatively analyzing the performance of different detection methods under various scenarios, the results show that the proposed method is robust in various illumination conditions and especially effective on detecting small objects.

Zhang J S, Huang Y P, Zhang R. Fusing point cloud with image for object detection using convolutional neural networks[J]. *Opto-Electron Eng*, 2021, **48**(5): 200418; DOI: 10.12086/oee.2021.200418

Foundation item: Shanghai Nature Science Foundation of Shanghai Science and Technology Commission, China (20ZR1437900) and National Nature Science Foundation of China (61374197)

* E-mail: huangyingping@usst.edu.cn